ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE INFORMÁTICA Y ELECTRÓNICA

ESCUELA DE INGENIERÍA EN SISTEMAS

"PROPUESTA METODOLÓGICA PARA LA GESTIÓN DE LA CALIDAD DE DATOS EN PROYECTOS DE INTEGRACIÓN. CASO PRÁCTICO: SII-ESPOCH"

TESIS DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS

PRESENTADO POR:

Margarita Isabel Solís Velasco

RIOBAMBA-JULIO

2011

AGRADECIMIENTO

Mi más sincero agradecimiento a mi directora de tesis Ing. Ivonne Rodríguezpor brindarme su amistad y toda su confianza, paciencia y orientación para la realización y culminaciónde esta tesis.

A la Ing. Gloria Arcos por todo su apoyo y sugerencias para el desarrollo de la misma.

DEDICATORIA

A mis padres, pilares fundamentales de mi vida quienes siempre han velado por mi bienestar y educación, y que gracias a su confianza y apoyo en cada momento, he podido ver alcanzado mi meta.

A mi hermana Marthy, y a mis hermanos Rafael y Hernán por ser mi ejemplo de superación y triunfo en la vida.

FIRMAS RESPONSABLES Y NOTAS

NOMBRES	FIRMAS	FECHA
Ing. Iván Menes **DECANO DE LA FACULTAD INFORMÁTICA Y ELECTRÓNICA**	_____	_____
Ing. Raúl Rosero **DIRECTOR DE ESCUELA INGENIERÍA EN SISTEMAS**	_____	_____
Ing. Ivonne Rodríguez **DIRECTOR DE TESIS**	_____	_____
Ing. Gloria Arcos **MIEMBRO DE TESIS**	_____	_____
Tlgo. Carlos Rodríguez **DIRECTOR DEL CENTRO DE DOCUMENTACIÓN**	_____	_____

Nota: _____

RESPONSABILIDAD DEL AUTOR

“Yo, Margarita Isabel Solís Velasco,soy responsable de las ideas, doctrinas y resultados expuestos en esta tesis, y el patrimonio intelectual de la misma pertenecen a la Escuela Superior Politécnica de Chimborazo”.

_____

FIRMA

ÍNDICE DE ABREVIATURAS

DQ	Data Quality
SII	Sistema de Información Institucional
ESPOCH	Escuela Superior Politécnica Chimborazo
OLAP	Proceso Analítico en línea
OLTP	Proceso transaccional en línea
ROLAP	Proceso Relacional Analítico en línea
MOLAP	Proceso Multidimensional Analítico en línea
BI	Business Intelligence
ETL	Extracción, Transformación,Carga
CRM	Customer Ralationship managment
ERP	Enterprise Resource Planning
DWH	Datawarehouse
DSS	Sistema de soporte de decisiones
ODS	Object Data Store
EII	Enterprise Information Integration
EIA	Enterprise Application Integration
B2B	Business To Business
SCM	Software Configuration Management
CMM	Modelo de Madurez de Calidad
XML	Extensible Markup Language
SOA	Arquitectura Orientada a Servicios
ODS	Operational Data Store

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

INDICE DE MODELOS DE APOYO

CAPITULO I

MARCO REFERENCIAL

1.1 Antecedentes

Desde que existen los datos informatizados, siempre ha existido la preocupación de que sean correctos. Inicialmente las correcciones únicamente se podían realizar editando manualmente los datos, o desarrollando programas que corregían algún dato de forma manual.

Los primeros sistemas de calidad de datos surgieron para resolver problemas con las direcciones postales. Las iniciativas surgieron desde gobiernos o entidades financieras, que perdían millones de dólares en envíos postales que no llegaban a su destino.

Las empresas u organizaciones se encuentran hoy en día bajo una fuerte presión para invertir en tecnologías que impulsen la ventaja competitiva y mejoren los resultados operacionales. Un despliegue exitoso de un proyecto de integración puede ayudar a valorar la salud de una empresa, establecer los indicadores de rendimiento oportunos y monitorizar las operaciones del día a día con un ojo puesto en el crecimiento global. Por consiguiente, la demanda de datos precisos para

realizar las tareas de integración continúa creciendo tanto en el lado de la demanda como en el del suministro de información.

La intensa demanda de datos está impulsando la adopción generalizada del uso de Business Intelligence (BI) en toda la base de usuarios, desde el grupo de directivos a los usuarios en los puntos finales. Esta amplia adopción ha propiciado que el BI avanzase más allá de las tradicionales funciones de query, reporting analítico y procesamiento analítico online (OLAP), para incluir ahora cuadros de mando operacionales, tablas de resultados personalizables y avanzadas técnicas de visualización. Desde la perspectiva de la cadena de suministro de la información, esto implica que los datos necesitan estar accesibles y ser agregados y racionalizados para poder consumirse por el BI, independientemente del formato, donde sea que el usuario lo necesite. Y cada día, la apuesta es más alta.

Frente a las tradicionales aplicaciones de BI centradas en las queries y las analíticas, muchos nuevos usuarios de BI se centran en las decisiones operacionales y las consiguientes acciones. Esto significa que toda acción que los usuarios emprenden basándose en la fortaleza de los informes y las alertas está influida por la precisión de los datos utilizados para los informes.

Por esta razón, un creciente número de organizaciones están emprendiendo iniciativas de calidad de datos como principio central de sus iniciativas de BI en la empresa.

Existen diferentes plataformas business Intelligence que actualmente están abordando el tema de calidad datos tales como Business Objects, Oracle, Microsoft SQL Server, Informática Data Integration, SAS, IBM, Sybase, Teradata.

Anteriormente se han realizado las siguientes investigaciones acerca de Business Intelligence.

- Estudio y uso de los métodos y técnicas de datawarehouse y datamining aplicado en al ámbito académico de la Escuela Superior Politécnica de Chimborazo.

- Estudio de la tecnología OLAP e implementación de un sistema de soporte decisional para Fundación Marco.
- Estudio de la tecnología Business Intelligence y su aplicación en un Modelo de Sistema de Información Gerencial en Petroproducción.
- Propuesta Metodológica para aplicar Business Intelligence en COHERVI S.A.
- Estudio Comparativo de Herramientas Open Source para Análisis Multidimensional. Caso Práctico: PROASETEL S.A, Análisis Multidimensional del Rub-Ecuador.
- Estudio de Herramientas Business Intelligence para la implementación de un Sistema de Información General en el Departamento de Planificación.
- Guía Práctica para Implementar Balanced Scorecard, Caso Práctico: Construcción de un Sistema de Gestión de Rendimiento en la Subgerencia Financiera de Petroproducción.

En las investigaciones anteriormente mencionadas se ha tratado los procesos que involucra una solución Business Intelligence pero no contienen una forma de trabajo que permita cerciorarse que los datos con los que se trabajen sean de calidad. Razón por la cual se ve la necesidad de contar con una metodología que permita gestionar la calidad de datos en este tipo de soluciones.

1.2 Justificación

- **Justificación Teórica**

La calidad de datos es una de las principales cuestiones que afectan al análisis y soporte para la toma de decisiones en las empresas. Por un lado, la proliferación en un proyecto de integración, con datos extraídos de sistemas y aplicaciones dispares, puede empeorar la calidad de los datos y provocar una pérdida de confianza en el reporting. Por otro lado, un proyecto de integración desplegado con datos de calidad puede ayudar a una organización a competir más eficaz y decisivamente. En otras palabras, la calidad de los datos puede colocar a una organización a la defensiva o a la ofensiva, dependiendo de lo bien que pueda gestionarla.

Las organizaciones toman decisiones basadas en los datos disponibles en ese momento. Si una organización puede mejorar la calidad de estos datos, puede mejorar la calidad de las decisiones aumentando su efectividad y eficacia. Esto permitirá a la organización mejorar más lucrativamente.

Estamos en la era de la calidad, y es esta la que nos dará una ventaja competitiva, en un mundo globalizado.

- **Justificación Metodológica**

Gran parte del éxito de un proyecto de integración de datos se basa en su forma de trabajo, el uso de una metodología sirve para que cada miembro del equipo sepa que hacer y cuando.

Una metodología nos da mayor grado de certidumbre que nuestro proyecto cumplirá los objetivos trazados y en el tiempo convenido además su utilización nos puede orientar a definir correctamente los objetivos de negocio y definir las metas del proyecto.

- **Justificación Práctica**

La metodología desarrollada se aplicara en el Sistema de Información Institucional de la Escuela Superior Politécnica de Chimborazo (SII-ESPOCH) para los sistemas de la Unidad de Educación a Distancia y Académico Institucional, el cual permitirá gestionar información de relevancia que sirva de soporte para la toma de decisiones dentro de la institución.

Esto se implementara con un nuevo modelo de gestión basado de la calidad en todos su procesos y en la disponibilidad oportuna y exacta de la información por esta razón se requiere contar con la aplicación de una metodología para la gestión de calidad de datos mediante el cual se convertirán en útiles con mayor rapidez, siendo los datos la base primordial para obtener información de calidad.

Al incrementar la confianza en los datos, el sistema puede reconocer y actuar inmediatamente ante nuevos patrones y tendencias, con una granularidad y precisión más elevadas. Identificar los sobrecostes y otras oportunidades para reducir gastos y

ahorrar puede ser resultados directo de la capacidad de los usuarios al utilizar datos precisos y verificables procedentes del reporting.

Además contar con la aplicación de una metodología para la gestión de la calidad de datos incrementa la auditabilidad y la visibilidad, un aspecto especialmente valioso para los propósitos de conformidad y gestión de riesgos.

1.3 Objetivos

1.3.1 Objetivo General

- ✓ Desarrollar una propuesta metodológica para la gestión de calidad de datos en proyectos de integración aplicado a la integración de datos para SII-ESPOCH

1.3.2 Objetivos Específicos

- ✓ Estudiar acerca de la calidad de datos y su influencia en proyectos de integración.
- ✓ Analizar los procesos existentes que se utilizan para obtener calidad de datos en las empresas.
- ✓ Aplicar la metodología desarrollada para la integración de datos en los sistemas de la Unidad de Educación a Distancia y Académico Institucional que son parte de SII-ESPOCH utilizando herramientas software para la medición de calidad de datos.

1.4 Hipótesis

La metodología propuesta permitirá asegurar la calidad de datos en proyectos de integración.

CAPITULO II

MARCO TEÓRICO

2.1 INTRODUCCIÓN

Los datos oportunos y precisos utilizados en proyectos de integración son críticos en el trabajo de muchas organizaciones. Sin las estructuras para suministrar y actuar de forma consistente sobre datos de confianza y alta calidad, pueden verse amenazados y las organizaciones ven socavadas su capacidad para valorar el estado real de la organización y emprender las acciones apropiadas para dirigir su negocio y competir eficazmente.

Las soluciones de calidad de datos pueden aprovechar las sinergias con los procesos y soluciones de integración de datos existentes en la empresa, con capacidad para acceder y gestionar todo tipo de datos en un enfoque dirigido por métricas. Los resultados pueden ser muy superiores frente al uso de una tecnología tradicional de calidad de datos, normalmente limitada a la limpieza de los datos de clientes.

El despliegue exitoso de la calidad de datos a este nivel realmente empresarial ayuda a una organización a maximizar los retornos sobre sus inversiones, mediante la mejora de su capacidad para aprovechar el BI para impulsar la ventaja competitiva y el liderazgo de mercado.

2.2 DEFINICIÓN

El TDWI[1] (The Datawarehouse Institute) define la calidad de datos como la calidad del contenido y estructura de los datos (en función de una serie de criterios variables) y las prácticas empresariales y tecnológicas estándar que mejoran los datos, como las acciones de limpieza, correspondencia, agrupación, deduplicación y estandarización de datos de nombre y direcciones, y el enriquecimiento con fuentes externas [Phillip Russom, "Taking Data Quality to the EnterpriseThrough Data Governance" (Calidad de datos en toda la empresa a través del gobierno de datos),Data Warehousing Institute (TDWI), marzo de 2009].

Los servicios para calidad de datos permiten incrementar la precisión, puntualidad, relevancia y consistencia de la información, a nivel corporativo o en múltiples unidades de negocios dentro de la empresa, asegurando de esta forma que las decisiones se tomen con base en información consistente y precisa, y que los sistemas transaccionales operen con la información que se apegue a estándares de calidad. [2]

2.3 BENEFICIOS

A través de la gestión de la calidad de datos será posible contar con los siguientes beneficios:

- Implantación de mecanismos y estándares de calidad para suprimir la entrada de datos incorrectos.
- Identificación de elementos duplicados y elaboración de cruces de información entre dos o más fuentes de datos (internas o externas).
- Optimización en procesos de mensajería a través de la normalización de datos.
- En soluciones de Business Intelligence, permite contar con información confiable para la correcta toma de decisiones.

[1]TDWI(El Data WarehousingInstitute)ofreceeducación, capacitación, certificación, noticias, e investigaciónpara los ejecutivos yprofesionales de la tecnología de la información(TI) en todo el mundo. Fundadaen1995, TDWIeselprimer institutode educaciónde business Intelligencey datawarehouse.

- Eficiente operación en sistemas transaccionales como ERP's, CRM's, SCM, Call Centers, B2B, Aplicaciones Web y Sistemas Legacy, entre otros.
- Operaciones Eficientes: Se obtienen menores tasas de error en aplicaciones ERP Y CRM, páginas de e-commerce y otros sistemas transaccionales.
- Reporting Preciso: Desde los departamentos que se ocupan de la operación del día hasta las áreas directivas de planificación estratégica reciben informes que reflejan fielmente la realidad del negocio.
- Análisis Intuitivo: Resultados mejores y más fiables de las herramientas predictivas, de presupuestación, de campañas de marketing y de inteligencia de negocio.
- Mejor Servicio al Cliente: Una visión unificada de la información de contacto, incluyendo el historial de compras y preferencias, proporciona a los empleados que trabajan de cara al cliente la información que necesitan para ofrecerles el mejor servicio para incrementar su satisfacción.
- Aumento de Ingresos: Se pueden identificar nuevas oportunidades de upsell y cross-sell, gestionar las cuentas de forma eficaz, entender y anticipar los patrones de compras del cliente a través de en una visión uniforme del cliente.
- Cumplimiento Regulatorio Fiable: Disponer de informes precisos y establecer procesos de negocio relacionados con la gestión de datos fiables y replicables permite que se cumplan normativas como ISO, blanqueo de capitales, etcétera.[3]

2.4 IMPACTOS EN EL NEGOCIO

En esta área, la calidad de los datos impacta en el negocio ya que éstas se esfuerzan por:

- Distribuir las alertas a un amplio rango de puntos de usuario desde cualquier fuente de datos.
- Garantizar un alto movimiento para una gran variedad de tipos de suscripción sobre conjuntos de datos estandarizados y no conflictivos.
- Permitir a los usuarios abrir archivos adjuntos o entrar en vínculos mientras presentan datos consistentes e integrados.

- Mitigar el riesgo de distribuir alertas y notificaciones incorrectas con una calidad de datos pre-definida y aprobada.
- Permitir desencadenar alertas en tiempo real cuando múltiples datos de eventos cumplen umbrales específicos
- Aprovechar los datos autentificados para la personalización del contenido y la filiación de grupos.

Pero también tiene un lado negativo, según se desprende de una serie de estudios llevados a cabo.

El impacto que puede producir la mala calidad de datos puede ser devastador. Estas son algunas de las consecuencias del impacto:

- Desmesurada cantidad de tiempo y de recursos de IT para investigar, limpiar y conciliar datos.
- Carga de trabajo operacional adicional para recopilar y corregir datos para el análisis.
- Pérdida de credibilidad en los sistemas y en la cadena de suministro de BI en su conjunto.
- Toma de decisiones más lenta y errónea que impacta negativamente en la satisfacción de los clientes y en el rendimiento empresarial.
- Errores o tardanza en la consecución de los propósitos de cumplimiento y riesgos.[4]

Riesgos:

- Riesgo en el cumplimiento normativas
- Sistema de gestión del riesgo
- Sistema de integración del riesgo
- Riesgo en la inversión
- Riesgo competitivo
- Detección del fraude
- Riesgos legales

Pérdida de Ingresos:

- Cobro ineficiente
- Mala relación con el cliente
- Pérdida de oportunidades

Incremento de costos:

- Detección y corrección
- Prevención
- Reingeniería de procesos
- Penalizaciones
- Sobrepagos
- Recursos incrementados
- Retrasos
- Cargas de trabajo

Baja confianza:

- Falta de credibilidad
- Temor en toma decisiones
- Menor predictibilidad
- Forecasting incorrecto
- Reporting ineficiente

En la siguiente figura se muestra en cantidades porcentuales algunos de los impactos que sobresale en el negocio:

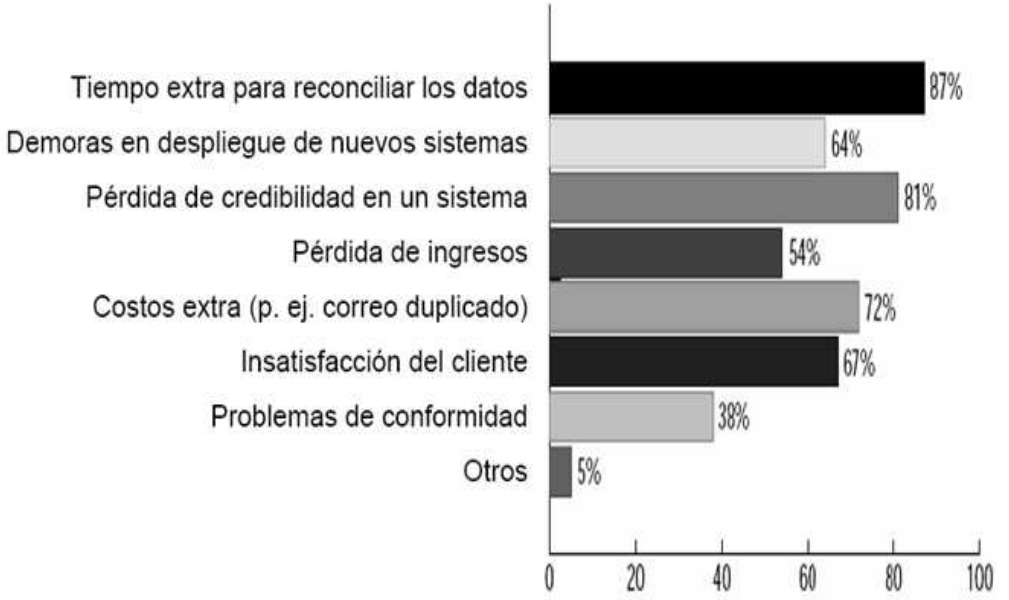


Figura II.1 Impacto en el negocio

Fuente :TDWI Data Quality Survey

2.5 CICLO DE VIDA DE LA INFORMACIÓN Y LA CALIDAD DE DATOS

El ciclo de vida no es un proceso lineal, sino iterativo. Existen cuatro componentes importantes que impactan en la calidad de los datos a través de este ciclo. Datos (qué)- Conocer hechos u otros aspectos de interés para la organización. Procesos (cómo) - Son las funciones, actividades, acciones, tareas o procedimientos que hurgan los datos.

Organización y personas (quiénes) – Equipos, roles, responsabilidades individuales que afectan o usan los datos, o están involucrados con los procesos. Tecnología (cómo)- Ventanas, aplicaciones, bases de datos, programas, almacenamiento o manipulación de datos que están involucrados con los proceso. [5]

Tabla II.1 Ciclo de vida de la información

Fases del ciclo de vida de la información	Definición	Ejemplo de actividades
Planificar	Preparar el recurso.	Identificar objetivos, arquitectura, desarrollar estándares y definiciones. Cuando diseñamos desarrollamos aplicaciones, bases de datos, procesos.
Obtener	Adquirir el recurso.	Cargar datos de ficheros externos, crear registros.
Almacenar	Conservar la información del recurso electrónicamente.	Almacenar datos electrónicamente en bases de datos o ficheros.
Mantener	Asegurarse que el recurso trabaja apropiadamente.	Actualizar, manipular, estandarizar, validar, transformar o consolidar registros.
Aplicar	Uso del recurso para alcanzar los objetivos.	Recuperar datos; usar la información para reportes, toma de decisiones, procesos automáticos.
Disponer	Descartar el recurso cuando no será utilizado.	Archivar, eliminar datos o registros.

2.6 FUENTES DE PRODUCCIÓN DE ERRORES EN LOS DATOS

- **Entrada de datos**: La mayor fuente de errores son las entradas de información manual, producido por ruido en la comunicación, errores tipográficos o equivocaciones, o por otros factores externos, como por ejemplo: en la entrada de datos de un contacto, este desconoce su código postal y la persona que está realizando la entrada, como no conoce y no tiene herramientas para localizar esta codificación, deja en blanco ese ítem.
- **Datos externos:** Frecuentemente se incorporan datos externos de forma automática, en los sistemas de información de las organizaciones, sin tomar las precauciones oportunas, y esto provoca que generen multitud de problemas de calidad de datos. Por ejemplo si se quiere incorporar un nuevo listado de productos, en la base de datos de una compañía, y no se ha asimilado previamente, aquellas referencias que el proveedor a actualizado (no se trata de un nuevo producto, quizás le ha cambiado: el nombre, el formato, etc) tendremos duplicidades en nuestro catálogo de productos.

- **Errores de carga de los sistemas transaccionales**: los múltiples errores que suelen ocurrir durante la carga en los sistemas transaccionales, provoca una deficiencia de la calidad de los datos.

- **Migraciones:** cuando se realiza una migración de datos, sin haber analizado en profundidad los cambios que hay que aplicar a la información, una de las muchas consecuencias, será la ausencia de calidad de datos, existencia de valores obsoletos o en un formato distinto al esperado en el nuevo sistema y duplicidades.

  En la figura II.2 se presenta de forma porcentual las principales fuentes de errores:

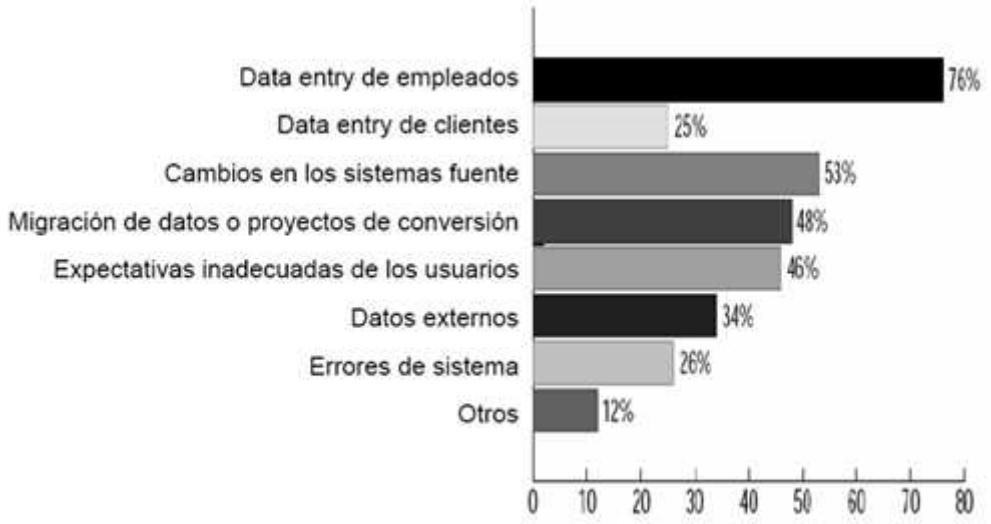


Figura II.2 Fuentes de errores en los datos

Fuente: TDWI Data Quality Survey

2.7 CAUSAS DE LA CRECIENTE MALA CALIDAD DE DATOS

- **Cada vez hay más datos de más fuentes en más sistemas:** ERPs, fuentes externas, web, call centres, datawarehouses, etc. todos los procesos se informatizan, los volúmenes aumentan, las aplicaciones se migran, los sistemas se comunican.

- **Los datos que eran introducidos para un propósito ahora está siendo aplicados a otras aplicaciones:** La Calidad de Datos puede ser relativamente bueno para los sistemas transaccionales pero no para sistemas BI o CRM. Una factura incorrecta en el transaccional afecta sólo a un cliente. En el sistema de BI, esta factura puede tener un impacto muy grande.
- **Mayores niveles de calidad de datos requerido para procesos automatizados:** La mala calidad de datos lleva a problemas de pagos en sistemas ERP, SCM, etc. Todos los procesos se automatizan y cada vez hay menos intervención humana. Por ejemplo, una persona detectaría y cancelaría por sentido común un proceso erróneo, como por ejemplo, el envío de una carta a la calle “nnnnnnn”, mientras que un sistema automatizado de mailing, no.
- **Mayor sensibilidad del público:** Los clientes esperan un mejor servicio. Los datos defectuosos llevan a una pobre gestión del cliente que cada vez es más exigente y cada vez conoce mejor sus derechos.

- **Falta de motivación:**

  “Atajos” adoptados por las personas responsables de la entrega de datos, debido a la falta de incentivos individuales a la producción de datos con calidad.
- **Cambio organizativo:**
  a. Nuevos requerimientos de datos que hacen que contingencias ocultas se manifiesten en la forma de problemas más serios.
  b. Algunos comprenden o reconocen que los datos son un activo estratégico importante, pero la mayoría no emplea procedimiento alguno de control de calidad, ni hace uso de las herramientas necesarias para garantizar la coherencia y precisión de los datos.
- **Falta de normas de calidad y controles:**
  a. Carencia de un procedimiento de calidad de datos.
  b. Ausencia de normas de calidad de datos, o no conformidad si las normas existen.
  c. Mala conversión de datos y tratamiento inconsistente desde los sistemas fuentes.
  d. Falta de documentación y evaluación periódica de los datos.

Mayoritariamente, se delega la responsabilidad de la calidad de los datos en el departamento de informática, el cúal, aunque está implicado, no tiene la capacidad de modificar procesos del negocio o comportamientos fuera de su área de influencia, que permitan la mejora sustancial de la calidad de los mismos. Un principio fundamental en la gestión de calidad de datos es detectar y corregir los errores lo más cerca posible del origen de los mismos, con el objetivo de minimizar los costos asociados.[6]

2.8 REQUERIMIENTOS DE CALIDAD DE DATOS

- **Eficacia del contacto:** Tradicionalmente la calidad de datos se ha centrado en la mejora de datos personales, como nombres y direcciones, con el objetivo de mejorar la eficacia del contacto. Generalmente para la mejora de procesos de marketing o cualquier proceso que requiera el análisis o el contacto del cliente.
- **Identificación de relaciones:** Generalmente para la búsqueda de duplicados, relación de dos fuentes de datos o la detección de unidades familiares o corporativas.
- **Calidad de Datos General:** Mejora de datos de cualquier dominio, como puede ser datos de producto, finanzas, tráfico, activos, etc.
- **Análisis de Calidad de Datos:** perfilado, detección, medición, análisis, cuantificación del impacto, monitorización de problemas de calidad de datos.[7]

Figura II.3 Requerimientos de Calidad de Datos

2.9 EVOLUCIÓN DE LA CALIDAD DE DATOS Y LA INTEGRACIÓN

Las herramientas disponibles del mercado tradicionalmente eran diseñadas para un solo propósito, se estructuraba para un estilo muy simple de integración de datos, como las ETL solo abordaban la replicación de datos, o una parte de la calidad de datos (por ejemplo perfilado de datos, o limpieza de datos).

El costo y la limitada funcionalidad de las herramientas han hecho que las organizaciones desplegaran cada una de las herramientas tácticamente, mientras continuaban desarrollando código. A medida que las organizaciones reconocen la importancia estratégica de la integración y la calidad de los datos, empiezan a buscar soluciones para enfrentarse de una forma simple y centralizada a todo el rango de necesidades y requerimientos.

El cambio en la demanda y las presiones de la competencia en el mercado, ha provocado la consolidación de los proveedores y la aparición de herramientas con múltiples propósitos. Además la madurez de este tipo de herramientas, ha permitido a las organizaciones empezar a reducir la cantidad de código a medida y comenzar a apoyarse en los metadatos.

Actualmente estamos en el punto de la convergencia, dónde ya se ha comenzado a desarrollar herramientas que solucionan de manera unificada todas las necesidades de integración y calidad de datos. Herramientas muy versátiles y que de una forma muy centralizada, pueden dar soporte a muchos proyectos dentro una organización.

El futuro pasa por soluciones integradas en las que no existirá el código a medida y estén erigidas sobre metadatos.[8]

Figura II.4 Evolución de la Calidad de Datos

2.10 DIMENSIONES DE LA CALIDAD DE DATOS

A continuación se presenta las dimensiones de calidad con los cuales se puede determinar la calidad de datos

Tabla II.2 Dimensiones de la Calidad de Datos

Dimensión	Descripción
Exactitud	1) El grado en el cúal el dato tiene atributos que correctamente representan el valor correcto del atributo intencionado de un concepto o evento en un contexto especifico de empleo 2) Lamedidao elgradode concordanciaentre elvalor delos datos(o conjuntodevalores) y una fuentesupone que escorrecto.
Completitud	El grado en el cuál el dato esta asociado con una entidad que tiene valores para todos los atributos esperados e instancias de entidad relacionadas en un contexto específico de uso.
Consistencia	El grado en el cuál el dato tiene los atributos que son libres de contradicción y son coherentes con otros datos en un contexto específico de uso.

	1)Los datossemantienenpor loqueestánlibresdevariaciónocontradicción. 2) Lamedidaogrado en queun conjunto dedatoscumple conun conjunto derestricciones.
Credibilidad	El grado en el cual el dato tiene atributos que son considerados como verdaderos y creíbles por usuarios en un contexto especifico de uso
Actualidad	El grado en el cuál el dato tiene los atributos que son del periodo correcto en un contexto específico de uso
Accesibilidad	El grado en el cual se puede accesar al dato en un contexto específico de uso ,en particular por la gente que necesita el soporte de tecnología o una configuración especial debido a alguna inhabilidad (incapacidad)
Conformidad	El grado en el cuál el dato tiene atributos que se adhieren a normas ,convenciones o regulaciones vigentes y reglas similares relacionadas con la calidad de datos en un contexto específico de uso
Confidencialidad	El grado en el cuál el dato tiene los atributos que se aseguran que este es solo accesible e interpretable por usuarios autorizados en un contexto especifico de uso.
Eficiencia	El grado en el cúal el dato tiene los atributos que pueden ser procesados y proporciona los niveles esperados de funcionamiento (desempeño) usando las cantidades y los tipos de recursos apropiados en un contexto especifico de uso
Precisión	El grado en el cuál el dato tiene atributos que son exactos o que proporcionan la discriminación en un contexto específico de uso
Trazabilidad	El grado en el cuál el dato tiene atributos que proporcionan un rastro de auditoría de acceso a los datos y de cualquier cambio hecho a los datos en un contexto especifico de uso
Entendibilidad	El grado en el cual el dato tiene atributos que le permiten ser leído e interpretado por usuarios ,y es expresado en lenguajes apropiados ,símbolos y unidades en un contexto especifico de uso
Disponibilidad	El grado en el cuál el dato tiene atributos que le permiten ser leído e interpretado por usuarios autorizados y/o aplicaciones en un contexto especifico de uso
Portabilidad	El grado en el cuál el dato tiene los atributos que le permiten ser instalado ,substituido o movido de un sistema a otro conservando la calidad existente en un contexto especifico de uso
Pertinencia	Los datosdebenserpertinentesa los finesparalos que se utiliza.

	Estoimplicauna revisión periódicade losrequisitosparareflejarlas necesidades cambiantes. Puede que seanecesariopara capturardatosen elpuntodeactividadquesólo es relevante paraotros fines, en lugar dela intervenciónactual. Los procesos deaseguramiento de la calidady la retroalimentaciónson necesariospara garantizarlacalidadde estos datos.
Oportunos	Los datosdebensercapturadotan prontocomosea posibledespuésdeleventoo actividadydebe estardisponible paraelusoprevisto, dentro deun plazorazonable. Los datosdebenestardisponiblesde forma rápida ycon la frecuencia suficientepara satisfacer las necesidadesde información ypara influir enel niveladecuado deservicio delas decisioneso de gestión.
Validez	Los datos deben ser registrados y utilizados de conformidad con los requisitos pertinentes, incluída la correcta aplicación de las normas o definiciones. Esto garantizará la coherencia entre los períodos y con otras organizaciones similares. Unacondición en la quelos valores dedatospasantodas las edicionesdeaceptabilidad, produciendolos resultados deseados.
Fiabilidad	Los datosdebenreflejarlos procesosde datosestable yconsistentede recogidaa través delos puntos de recogidayen el tiempo, ya seamanual omediantesistemasbasados en computadora,o una combinación.Los gestores y losinteresadosdebenestar seguros deque el progresohacialos objetivos de rendimientoreflejalos cambiosrealesen lugardelas variacionesen los métodosde recopilación de datoso métodos.
Integridad	Las necesidades de datos se deben especificar claramente basadas en las necesidades de información de la organización y recopilación de datos procesos coincidentes con estos requisitos. Los registros de monitoreo faltantes, incompletos, o no válido puede dar una idea de la calidad de los datos y también puede apuntar a problemas en la grabación de elementos de datos ciertos.
Totalidad	Medición que refleje el grado en que las bases de datos cuentan con toda la información critica para el negocio
Oportunidad	Medición de que la información este disponible cuando se requiere para

	tomar una decisión.
Integridad	El grado enquelos valoresestánpresentesenlos atributosque lo requieran.
Singularidad	1) La capacidad deestablecerla singularidaddeun registrode datos(y los datos devalores de clave).
Duplicidad	Mide la duplicidad no deseada que existe en los campos, registros, o conjunto de datos
Exactitud	Una medida de la corrección del contenido de los datos exige que una fuente autorizada de la referencia sea identificada y accesible.
Consistencia y Sincronización	Una medida de la equivalencia de la información almacenada para hacer los datos equivalentes.
Facilidad de uso y conservación	Una medida del grado en que los datos pueden ser accedidos y usados, y el grado de conservación en que los datos pueden ser actualizados, mantenidos, y manipulados
Cobertura de datos	Una medida de la disponibilidad y el alcance de los datos comparados con el universo total de datos o población del interés
Presentabilidad	Formato y apariencia que respaldan el uso apropiado de la información

2.11 CATEGORÍAS DE LA CALIDAD DE DATOS

La calidad de datos puede ser dividida en las siguientes 8 categorías:

1) Dominios
2) Definiciones
3) Integridad
4) Validación
5) Reglas de negocio
6) Integridad estructural
7) Transformaciones
8) Flujo de datos

- **Dominios**

Los dominios describen el rango y tipos de valores presentes en un conjunto de datos. Los típicos errores que pueden ocurrir relacionados a dominios son:

a. **Inesperado valor de dominio:** La documentación para un sistema indica que los valores para una columna será (A, B, C) pero los datos actuales contienen(A, B, C, d, e, f) Esto puede conducir a una variedad de problemas fatales.

b. **Cardinalidad.** Indica el número de valores únicos encontrados dentro de un conjunto de datos. Para una clave primaria, la cardinalidad esperada puede ser igual al total de números de registros, mientras para un campo de SI/NO la cardinalidad esperada puede ser dos.

c. **Singularidad.** El grado de singularidad en los datos puede ser un problema en la calidad de datos. Un campo que es 98% único puede indicar "basura" en un campo de clave primaria.

d. **Constantes.** Indican que el mismo valor está presente en cada registro. Las aplicaciones tienden a descuidar constantes creando de esta manera problemas de integridad .Un cambio en una constante usualmente indica un cambio en la lógica de un programa desde la producción de datos.

e. **Valores Atípicos.** Algunos datos pueden tener valores inesperados, tales como número de miembros de familia= -3.Los valores atípicosson famososporla generación defallos del sistemadesolicitudes.

f. **Longitud.** Esto se refiere al tamaño de los datos. Cambios pequeños tales como un movimiento desde un identificador de 8 dígitos a uno de 6 dígitos, puede interrumpir todas las aplicaciones en toda la empresa.

g. **Precisión.** Los errores de redondeo y truncamiento son frecuentemente introducidos durante el movimiento o acceso de datos.

h. **Escala.** Son los datos expresados en porcentaje, un factor o un periodo de tiempo.No es posible comparar 100 y 1.00 como porcentajes

i. **Internalización.**Pueden haber códigos postales, el tiempo, o formatos de fecha inesperadas.

- **Definiciones.**

Las definiciones indican como las entidades son referenciadas a lo largo de la empresa. Mientras palabras simples como "ingresos" pueden tener un significado muy diferente

en las ventas, marketing, inventarios y finanza, las diferencias siempre son más ocultas entre *Empleado->Id* y *Factura->Id* y *Factura->Empleado->Id* .La definición de problemas son divididos en estas categorías:

a. **Sinónimos.** Sean o no los mismos nombres en las entidades son en realidad las mismas.Por ejemplo los campos EMP_ID, EMPID, y EM01 todas pueden o no pueden actualmente referirse al mismo tipo de datos.

b. **Homónimos.** Estos indican los campos que se escriben igual, pero en realidad no son los mismos. El nombre de la variable común "Id" puede significar muchas diferentes cosas en diferentes contextos.

c. **Relaciones.** Solo porque un campo sea nombrado FK_INVOICE no significa que esté realmente sea un campo foráneo para invocarlo.

- **Integridad**

La integridad indica si es que todos los datos están realmente presentes. Si bien esto parece elemental, campos y valores faltantes son quizás el problema más común de calidad de datos .La integridad examina estas áreas:

a. **Integración.** Es importante considerar si los datos actuales coinciden con la descripción de los datos En otras palabras si nuestra metadata es precisa. Un campo sin usar es un buen ejemplo de problema de integración.

b. **Precisión.** A menudo es necesario examinar las categorías de acuerdos entre los valores de los datos y una fuente que se supone escorrecta. La comparación de las tablas de resumen de las cantidades reales generados por operaciones a menudo conduce a resultados inesperada. Igualmente muchas fuentes de datos pueden ser comparados con orígenes externos, igualación y limpieza, estandarización de nombres, e igualación demográfica son ejemplos de controles de exactitud.

c. **Valores Actuales.** Esto se refiere a las categorías en los cuales los valores están presentes en los atributos que los requieren. Si el 50% de los registros de los clientes no tienen una dirección email, después en una campaña de marketing por correo no es probablemente el rumbo deseado para llegar a nuestros clientes.

d. **Fiabilidad.** Existe la posibilidad de la desconfianza en los datos debido a su contexto .Por ejemplo un código postal debe coincidir con la ciudad y el estado.

e. **Redundancia.**Se debe considerar si hay duplicados en los datos

f. **Consistencia.** Se debe preguntar si existen conflictos en los datos. Por ejemplo si el mismo número de factura es referenciado con dos diferentes cantidades.

- **Validez**

Como su nombre lo dice la validez indica si un dato es válido o no. Sorpresivamente muchas bases de datos de negocios están plagadas con datos que no pueden ser correctas. Control en la validez consiste en:

a. **Aceptabilidad.**Se debe preguntar si los datos pasan un conjunto de pruebas de aceptabilidad. Por ejemplo la parte de un número podría consistir de 7 dígitos, alfanumérico string comenzando con dos dígitos alpha y 5 números.

b. **Fiabilidad.** Esto se refiere a la probabilidad que los datos realicen la función indicada en un periodo de tiempo determinado.

c. **Anomalías.**Son los hechos en los que los datos son claramente imposibles. Por ejemplo un abogado no puede contar con 48 horas de un día o un auto no puede estar en diferentes avenidas al mismo tiempo

d. **Puntualidad**. Se debe preguntar si están los datos actualizados. Una medida en tiempo real de una fuente de datos que recientemente fue completado con éxito hace tres meses no es muy útil.

- **Reglas de Negocio**

Las reglas de negocio es la medida de cumplimiento entre los datos reales yla producción y consumo de los datos faltantes. Desde una perspectiva de calidad de

datos, las reglas de negocio son importantes ya que son las únicas que pueden ser medidas objetivamente .Las reglas de negocio pueden ser medidas como:

a. **Restricciones** Se debe preguntar si los datos cumplen con un conjunto conocido de limitaciones. Cualquier cosa que puede ser descrita matemáticamente o algorítmicamente puede llegar a ser una restricción.

b. **Reglas de Calculo.** Cercamente relacionados a las restricciones, las reglas de cálculoverifican los valores a través de registros por ejemplo asegurarse que una cantidad equivalga al precio.

c. **Comparaciones.** Estas reglas de negocio verifican las relaciones existentes entre campos de un registro .Por ejemplo la fecha de envío nunca puede ser menor que la fecha de pedido.

d. **Condiciones.** Estas reglas indican la regla lógicasi-entonces ciertos objetos de datos .Por ejemplo si un empleado es de un nivel 2 luego él/ella debe recibir un aumento del 5% a menos que el pago total de los empleados es mayor que $55000/año.

e. Dependencias funcionales. Estas reglas miden las invariancias a través de columnas de datos. Por ejemplo para cada número de cliente los datos deben siempre contener el nombre del mismo cliente

- **Integridad Estructural**

La integridad estructural examina si los datos están completos en un nivel macro. Este asegura que se tome los datos como un todo, con esto estaremos consiguiendo los resultados correctos. La integridad estructural se compone de:

a. **Integridad referencial.** Si esperamos una relación uno a uno entre dos elementos de datos, es la existencia de un elemento A siempre implica la existencia de B? Si esperamos una relación uno-a-muchos, siempre hay por lo menos una B para cada A? Muchas aplicaciones y puestos de trabajo de ETL desactivan los controles de integridad referencial a fin de acelerar la carga de bases de datos (o utilizar alguna otra lógica para comprobar la integridad referencial). Como resultado, los datos que acaba en una base de datos que

supuestamente proporciona integridad referencial a menudo pueden ser incorrectos.

b. **Vínculos.** Por ejemplo están en nuestra factura registros que no están en nuestro catalogo de piezas?Hay órdenes que se han marcado como entregado que no se puede comparar en una factura?

c. **Claves Primarias.** Son las claves primarias únicas?

d. **Chequeo de cardinalidad.** Ciertas relaciones indican que la cardinalidad de las columnas seráequivalente. Por ejemplo la cardinalidad de un campo de búsqueda en una tabla principal debe coincidir con la cardinalidad de un campo de búsqueda en la tabla de búsqueda. Esto puede considerarse como el punto de vista macro de análisis de dominio.

- **Transformaciones.**

Las transformaciones examinan el impacto de las trasformaciones de datos como movimientos de datos desde un sistema a otro sistema. La lógica para una transformación de datos puede estar defectuosa, pero la única manera de comprobarlo es comparar el conjunto de datos origen y la fuente y verificar que la transformación se lleve a cabo correctamente. Los controles de la transformación incluyen:

a. **Filtración.** Esto verifica que los registros transferidos estándestinadospara ser transferido .Por ejemplo una carga de un datawarehouse podría requerir la transferencia de los últimos registros de la semana.

b. **Fusión .**Algunas transformaciones requieren que múltiples fuentes de datos son fusionadas juntas para formar un único objetivo. Por ejemplo algunos registros de clientes podrían ser fusionados .La fusión indica si o no el todo es igual a la suma de sus partes.

c. **Mapas de transformación.** Este verifica que transformaciones simples esperadas han tomado se ha llevado a cabo tales como A->1,B->2 etc. Una función de transformación se aplica tanto a los registrosdel origen y del destino para verificar que los resultados generados sean los esperados.

d. **Cálculos.** Algunos campos objetivos son el resultado de cálculos en la fuente.Por ejemplo total de ventas podría ser el resultado de ventas 1+ventas

2.Este control calcula los resultados esperados desde una fuente y compara esto en el resultado en la meta.

- **Flujo de datos.**

Esto es concerniente a los resultados agregados de los movimientos de datos desde la fuente a los destinos. Muchos problemas de calidad de datos pueden ser localizados en cargas incorrectas, cargas perdidas, o fallas de sistema que pasan desapercibidos. La transferencia de datos automatizados que no son inmediatamente verificados puede multiplicarse en sistemas destino con datos defectuosos de una manera desapercibida. Trazar la historia de estos controles es muy valioso en el estudio de los problemas de rendimiento. Los problemas de flujo de datos son:

a. **Datos perdidos a través de los sistemas .**Este conjunto de controles determina que hacen los registros que existes en un sistema origen en sistemas destinos. Por ejemplo porque hay un cliente en el sistema de servicios al cliente que es desconocido en el sistema de entradas de pedidos.

b. **Cuentas de registro.** El registro de cuenta verifica que el número de registros producidos en el sistema destino es lo esperado. Las claves primarias duplicadas pueden resultar en menos registros en el destino que el esperado.

c. **Sumas de comprobación (Checksums).** Cuando se realiza la transferencia uno por uno de una tabla o columna, simples checksums pueden verificar que los datos se escriban como lo esperado.

d. **Marcas de tiempo (Timestamps).** Es la marca de tiempo de los datos en el origen de la zona de carga lo esperado?Hemos cargado previamente un registro con esta marca de tiempo?

e. **Tiempo de Procesamiento.** Esta la transferencia tomando una cantidad excesiva de tiempo para completarla?Rápidamente descubrir que una transferencia se está desacelerando gradualmente cada noche permite la gestión de los recursos y la prevención de incendios.[10]

2.12 PERSONAL DE GESTIÓN DE CALIDAD DE DATOS

Ungrupo degestiónde datosdebe serestablecida anivelde empresa, que debenestar dotadoscon los administradoresde datos, administradores delos metadatosylos administradores de calidad de datos.

- **Administradores de datos.** Estas personasson responsablesdel modelado lógico delaempresa, para establecery mantenerestándaresde nomenclatura, yparala captura dereglas de negociorelacionadas con los datos.

- **Administradores delos metadatos.** Estas personas son responsablesdela carga,la vinculación, la gestiónydifusión demetadatospara facilitar lacomprensióncomún de datosypara fomentar la reutilización de datosLos metadatos sonelcontextoinformación sobre los datos.

  Los metadatosincluyenlos nombres deloscomponentesde datos, definiciones dedatos, reglas de negocio, el contenido delos datos(dominios), tipo de datos, los datos longitud, propietario de los datos, transformaciones de datos, el grado delimpieza,y así sucesivamente.

- **Administradoresde calidad de datos**Estas personas se encargan dela prevenciónde la propagaciónde datosde baja calidaden toda laempresa, ypor lo tanto,latoma de decisiones. Por lo tanto, essuresponsabilidadpara llevar a cabola auditoria periódica de los datos, metadatos ymodelos dedatos, y participarenlos esfuerzos dereconciliaciónde datos, ayudando aidentificar yresolver lascausasdelos problemasde calidad de datos.
  Losresultadosde lasauditorías y los esfuerzos dereconciliacióndebenalimentarde nuevo enun ciclo continuo de mejora dela calidadde datos. [11]

2.13 NIVELES DE MADUREZ DE LA CALIDAD DE DATOS

Una manera fácildedeterminarel nivelde madurez decalidad de los datos en una organización es buscarsus actividadesde mejora de calidadde datosactual.

La FiguraII.5muestra las actividades comunes de mejora de calidadde datosencada uno deloscinconiveles datosde madurezde calidad basadosenla adaptacióndel modelo de

capacidad de madurez (CMM) de calidadde datos de Larry English[2].Loscinconivelesson

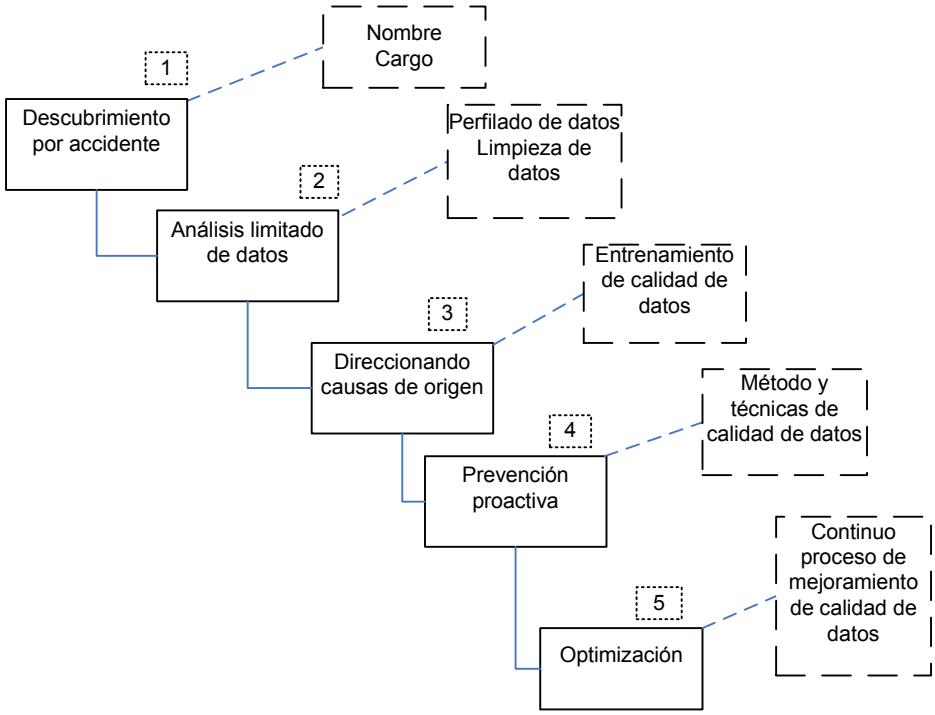


Figura II.5 Niveles de madurez de la Calidad de Datos

Nivel 1:Descubrimiento. En el nivel 1, la organización está tropezando condefectosde datoscomobloqueo de programasose quejanlos consumidoresde la información. Nohay datosde calidadproactivade mejora de procesos, ningún grupode calidad de datos, yno existe financiación.

Laorganizaciónniega tenerproblemasgravesde calidadde datos, y considera el análisis de datos una pérdida detiempo. Básicamente, laorganizaciónestádormidaynoquiere despertar.

Nivel 2: Análisis limitado de datos. En el nivel 2, la organización lleva a cabo algunas actividades limitadas de análisis y corrección de datos, tal como perfilado y depuración de los datos.

Todavía no hay soporte para toda la empresa para mejorar la calidad de datos, grupo de calidad, y la financiación. Sin embargo, algunos individuos aislados reconocen

---

Larry English es uno de las principales autoridades en Gestión de la Información y de Gestión de Calidad Total de la Información. Es un futurista que ha identificado las nuevas normas y regulaciones para permitir a las organizaciones ser eficaces en el manejo de su información en la emergente era de la información.

sus datos erróneos y quieren incorporar disciplinas de calidad de datos en sus proyectos.
Estas personas pueden ser administradores de datos, administradores de bases de datos, desarrolladores, o gente de negocios.

Nivel 3: Direccionando causas de origen. En el nivel 3, la organizaciónempiezaaabordar lascausasdesusdatoserróneosa través deprogramas de ediciónyla formaciónen cuanto a calidad de datos. Se crea grupode calidad de datos yexiste financiación para proyectos de mejora de calidad de datos.
Elgrupo decalidad delos datosinmediatamenterealizauna evaluación de calidad de datos desusarchivoscríticosy bases de datos toda la empresa, prioridad a las actividades de mejora de calidad de datos. Este grupotambiénpone en marcha unprogramaintegral de calidadde datosa través dela formación y disciplinas de información en laorganización

Nivel 4: Prevención proactiva. En el nivel 4, la organizacióntrabaja activamenteen laprevencióndefectos dedatos en el futuromediante la adición demásdisciplinas de mejora en programas decalidad de los datos. Los administradoresde toda la organizaciónaceptanla responsabilidad personal para lacalidad de los datos. Se han establecidométricas paramedir elnúmero dedefectosde datosproducidospor el personal, yestosparámetrosseconsideranen las evaluaciones desempeñodelpersonal de trabajo.
Incentivosparamejorar lacalidad delos datoshan reemplazado alos incentivospara arranquedelos sistemasala velocidadde la luz.

Nivel 5: Optimización. En el nivel 5, la organizaciónestá enunciclo deoptimización continuayla mejora desusprocesos de prevención de datos deerróneosla Calidad delos datoses unaparteintegral detodos losprocesos de negocio.
Cadadescripción de las funcionesrequiere atencióna lacalidad delos datos, lanotificación de defectosde datos, la determinación delas causas, la mejora delacalidad delos datosafectadosprocedimientos de eliminación deraíz de las causas,ymonitoreo de los efectos causados.Básicamente, la culturade laorganizaciónha cambiado. [11]

2.14 CALIDAD DE DATOS EN PROYECTOS DE INTEGRACIÓN

2.14.1 Integración de datos

En entornos distribuidos, las fuentes de datos se caracterizan por diversostipos de heterogeneidades que se pueden clasificar en general, (i) heterogeneidad tecnológica, (ii) el esquema de heterogeneidades y (iii) la heterogeneidad a nivel de instancia.

- **Heterogeneidades Tecnológicas** se deben a la utilización de productos por diferentes proveedores, empleados en las distintas capas de información y la infraestructura de comunicación Un ejemplo de la heterogeneidad tecnológica es el uso de dos diferentes sistemas de bases de datos relacionales por ejemplo DB2 de IBM vsSQL Server de Microsoft.
- **Esquema de Heterogeneidades** son principalmente causadas por eluso de los modelos de datos diferentes, como una fuente que adopta modelo relacional de datos y una fuente diferente que adopta el modelo de datos XML, ylasdiferentes representaciones de datos, como una fuente que almacena direcciones,un solo campo y otra fuente que almacena las direcciones con campos separadospara la calle, el número de civiles, y la ciudad.
- **Heterogeneidades a nivel de instancia** son debido a datos diferentes, contradictorios valores proporcionados por fuentes distintas del mismoobjeto. Este tipo de heterogeneidad puede deberse a errores de calidad, tales como laprecisión,errores de integridad, vigencia y consistencia; errores, pudieronprovocarsepor ejemplo, de procesos independientes que se alimentan las fuentes de datos diferentes.

Hoy en día, hay muchos ejemplos de situaciones en las que los datos que residen en distintas fuentes se deben acceder de forma unificada, superando tales heterogeneidades.

La integración de datos es un área de investigación importante que hapermitido a un usuario acceder a datos almacenados por la heterogeneidad de fuentes de datos a través de la presentación de una visión unificada de los datos.

Aunquela integración de datos debe enfrentar todos los tipos de heterogeneidades mencionadas anteriormente, en estainvestigación se centrará especialmente en las

heterogeneidades de instancia, donde los problemas de calidad de datos se vuelven muy importantes.

De hecho, el nivel de heterogeneidades de instancia puede afectar fuertemente el procesamiento de consultas en los sistemas de integración de datos. En concreto,la actividad de procesamiento de consultas se puede realizar al considerar que diferentesfuentes de datos pueden presentar diferentes niveles de calidad de los datos.

2.14.2 Proyectos de Integración de Datos

En las organizaciones de hoy es frecuente encontrar lo que se denominan islas de información, que no son más que sistemas de información desarrollados para satisfacer necesidades especificas de la organización, muchas veces usando plataformas, lenguajes, bases de datos diferentes y lo más importante, no se encuentran directamente comunicadas.

Hay una serie de factores tecnológicos para conducir proyectos de integración de datos dentro de las organizaciones en este momento. Arquitectura orientada a servicios SOA, modernización o sustitución de los sistemas heredados , Administración de datos maestros y complejos cambios o adición de datos de todos los proyectos requieren la integración de datos.

Por lo tanto, es fácil de ver la integración de datos como uno de los componentes técnicos de una iniciativa tecnológica en general. Se trata de hacer accesibles los datos de un sistema o aplicación a otra.[13]

Los principales proyectos de integración de datos en los que es necesario gestionar calidad de datos es:

- Business Intelligence-Datawarehouse
- Gestión de datos maestros
- Consolidación de aplicaciones
- Sincronización de datos
- Data Mining
- Migración

En la siguiente figura se muestra como se parte de necesidades empresariales conduciendo a iniciativas de TI para finalmente encontrar soluciones en proyectos de integración.

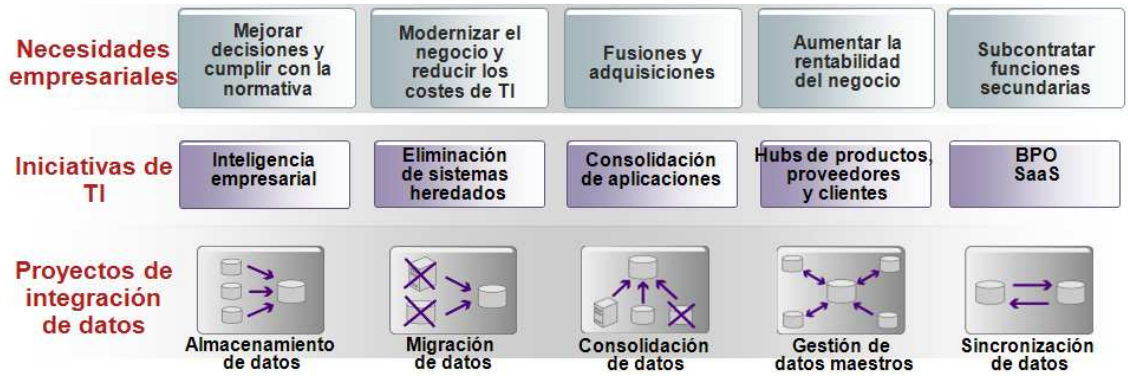


Figura II.6 Proyectos de Integración de Datos

2.14.2.1 Gestión de Datos Maestros

Los datos maestros describen a la gente, lugares y cosasque están involucrados en los negocios de una organización.

Las personaspuede ser clientes, empleados, vendedores, proveedores, los lugares por ejemplo territorios de ventas,oficinas, y las cosas por ejemplo,cuentas, productos, activos, juegos de documentos).

Debido a que estosdatos tienden a serutilizado pormúltiplesprocesos de negocioyde TI sistemas, la estandarización de formatosde datos maestrosy los valoresde sincronización es fundamental el éxito de la integración.

Los datos maestrostienden aagruparse enlos registros maestros, que podrán incluirse los datos de referencia.Un ejemplo de datosde referencia asociadoes uncampo de estado dentro deuna dirección enun registro maestro decliente.

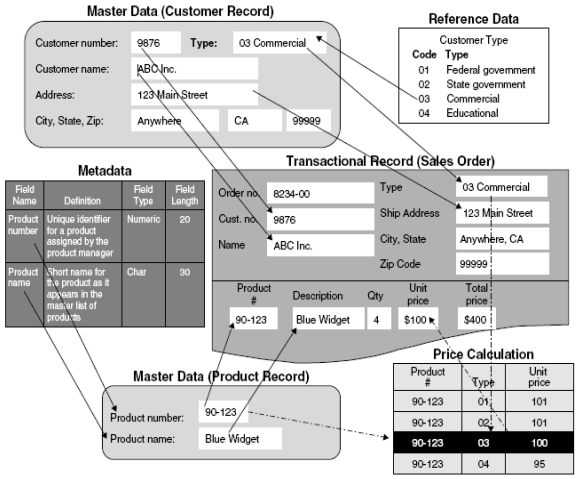


Figura II.7 Datos Maestros

2.14.2.2 Consolidación de aplicaciones

Las organizaciones de TI necesitan mejorar los niveles de servicio y el rendimiento de las aplicaciones y, a la vez, reducir costos. Una de la estrategia más lógica es reducir la complejidad y aumentar la capacidad de gestión del activo más estratégico en materia de TI: el centro de datos.

La consolidación de centros de datos tiene como objetivo simplificar el entorno de TI y facilitar su gestión. Una vez completada, la consolidación ayudará a reducir los costos de infraestructura y mantenimiento y a mejorará los niveles de servicio y el rendimiento de las aplicaciones.

2.14.2.3 Sincronización de datos

Existen muchos casos en los sistemas de información los datos se administran de forma independiente por múltiples aplicaciones o bases de datos. Sin embargo, es necesario mantener la coherencia entre dichos sistemas. La necesidad de la sincronización de datos puede ser permanente (sincronización entre sistemas operativos) o temporal (por ejemplo, durante una migración). La sincronización puede ser monodireccional o bidireccional.

La sincronización de datos incluye todos los procesos que mantienen sincronizados los datos entre las aplicaciones y las bases de datos.

2.14.2.4 Data Mining

La minería de datos (DM, *Data Mining*) consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

2.14.2.5 Migración

Al actualizar a una nueva versión de una base de datos o de una aplicación, o al cambiar a un nuevo sistema, los datos necesitan ser preservados en este nuevo sistema. El propósito de la migración de datos es transferir datos existentes al nuevo ambiente. Necesita ser transformado a un formato conveniente para el nuevo sistema, mientras que se preserva la información presente en el viejo.

2.14.2.6 Business Intelligence

Business Intelligence es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.

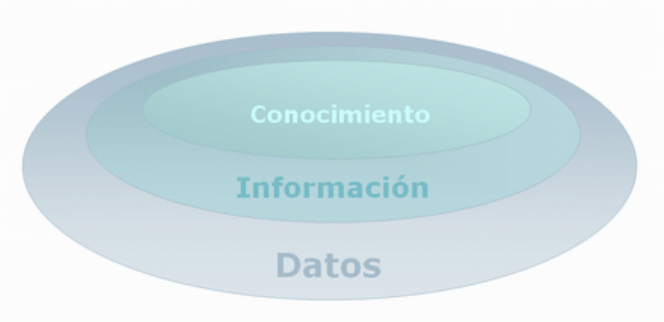


Figura II.8 Business Intelligence

Desde un punto de vista más pragmático, y asociándolo directamente con las tecnologías de la información, podemos definir Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting,

análisis OLTP / OLAP, alertas...) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

La inteligencia de negocio actúa como un factor estratégico para una empresa u organización, generando una potencial ventaja competitiva, que no es otra que proporcionar información privilegiada para responder a los problemas de negocio: entrada a nuevos mercados, promociones u ofertas de productos, eliminación de islas de información, control financiero, optimización de costes, planificación de la producción, análisis de perfiles de clientes, rentabilidad de un producto concreto, etc...

Los principales productos de Business Intelligence que existen hoy en día son:

- Cuadro de Mando Integrales (CMI)
- Sistemas de Soporte de Decisiones(DSS)
- Sistemas de Información Ejecutiva(EIS)

Por otro lado, los principales componentes de orígenes de datos en el Business Intelligence que existen en la actualidad son:

- Datamart
- Datawarehouse

Los sistemas y componentes del BI se diferencian de los sistemas operacionales en que están optimizados para preguntar y divulgar sobre datos. Esto significa típicamente que, en un datawarehouse, los datos están desnormalizados para apoyar consultas de alto rendimiento, mientras que en los sistemas operacionales suelen encontrarse normalizados para apoyar operaciones continuas de inserción, modificación y borrado de datos. En este sentido, los procesos ETL (extracción, transformación y carga), que nutren los sistemas BI, tienen que traducir de uno o varios sistemas operacionales normalizados e independientes a un único sistema desnormalizado, cuyos datos estén completamente integrados.

En definitiva, una solución BI completa permite:

- Observar ¿qué está ocurriendo?
- Comprender ¿por qué ocurre?
- Predecir ¿qué ocurriría?
- Colaborar ¿qué debería hacer el equipo?

Decidir ¿qué camino se debe seguir?

- **Arquitectura de una solución de Business Intelligence**

Una solución de Business Intelligence parte de los sistemas de origen de una organización (bases de datos, ERPs, ficheros de texto...), sobre los que suele ser necesario aplicar una transformación estructural para optimizar su proceso analítico.

Para ello se realiza una fase de extracción, transformación y carga (ETL) de datos. Esta etapa suele apoyarse en un almacén intermedio, llamado ODS, que actúa como pasarela entre los sistemas fuente y los sistemas destino (generalmente un datawarehouse), y cuyo principal objetivo consiste en evitar la saturación de los servidores funcionales de la organización.

La información resultante, ya unificada, depurada y consolidada, se almacena en un datawarehouse corporativo, que puede servir como base para la construcción de distintos datamarts departamentales. Estos datamarts se caracterizan por poseer la estructura óptima para el análisis de los datos de esa área de la empresa, ya sea mediante bases de datos transaccionales (OLTP) o mediante bases de datos analíticas (OLAP).

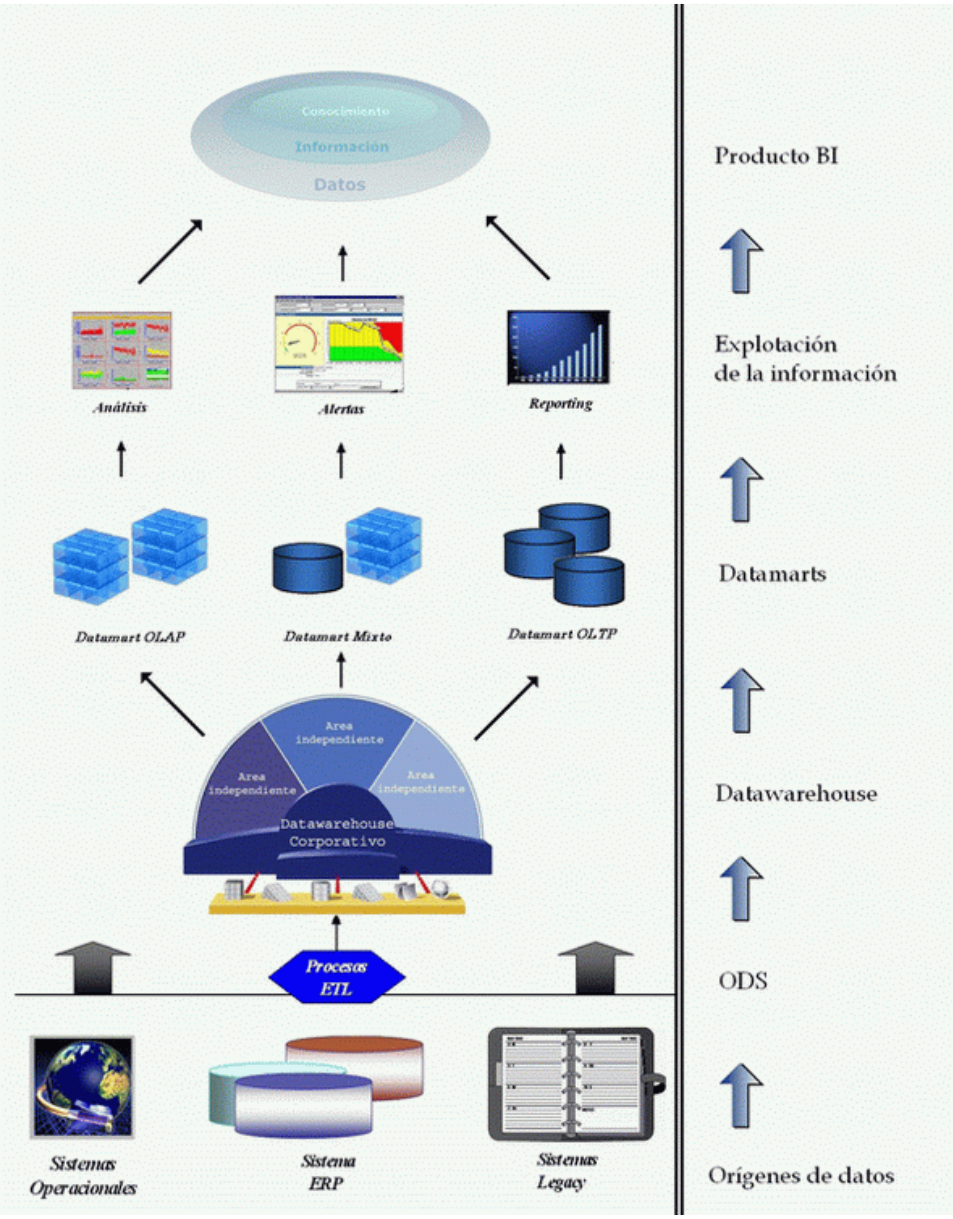


Figura II.9 Arquitectura Business Intelligence

Los datos albergados en el datawarehouse o en cada datamart se explotan utilizando herramientas comerciales de análisis, reporting, alertas... etc. En estas herramientas se basa también la construcción de productos BI más completos, como los sistemas de soporte a la decisión (DSS), los sistemas de información ejecutiva (EIS) y los cuadros de mando (CMI) o Balanced Scorecard (BSC).

- **Donde actúa la Calidad de Datos en BI**

En la implementación de una solución business Intelligence para la construcción de un DWH, La calidad de datos nos ayuda antes de la carga (ETL) a explorar, perfilar y medir el nivel de calidad de los datos.

Durante la carga gracias a una herramienta de calidad de datos, se estandarizan y codifican los datos (data cleanse), para posteriormente des-duplicarlos y enriquecerlos; en el tratamiento de des-duplicación los sistemas de Calidad de datos, cuentan con procedimientos que implementan lógica difusa, utilizados para establecer relaciones, que a través de otros procedimientos (sentencias de SQL p.e.) no son posibles.

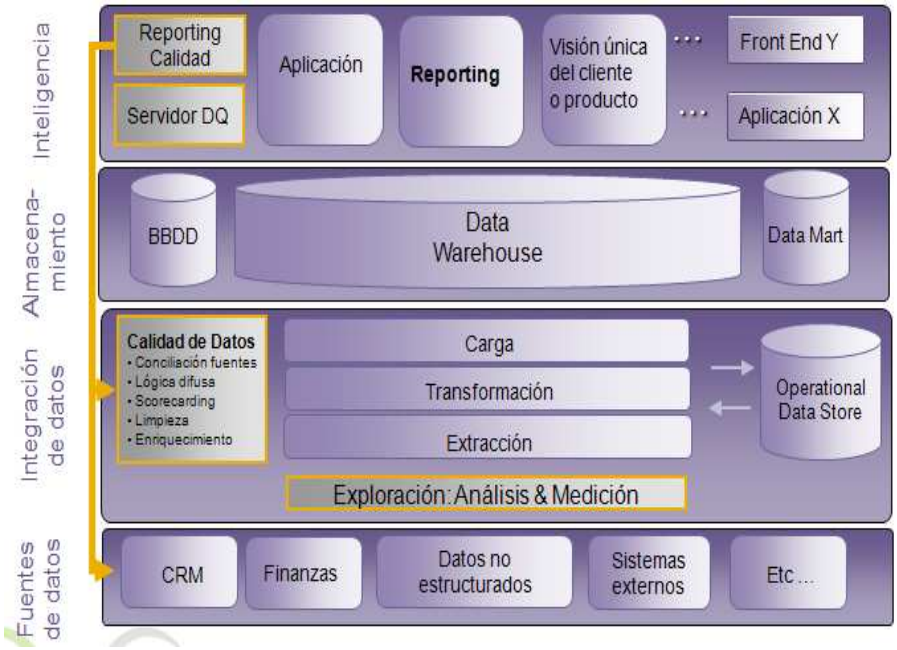


Figura II.10 Calidad de Datos en BI

- **Impacto de la mala calidad de datos en BI:**

En BI el impacto de la mala calidad de datos se aplica a sus cinco estilos:

- **Cuadros de mando y tablas de resultados**

La calidad de datos impacta en los usuarios de los cuadros de mando y tablas de resultados, que deben ser capaces de:

a. Consumir y actuar, rápidamente, sobre datos completos en los indicadores y marcadores de los cuadros de mando.
b. Lograr una visión integrada y colaborar utilizando datos estandarizados.
c. Aprovechar una metodología formal de tablas de resultados con datos consistentes.
d. Profundizar para ver datos precisos sobre el rendimiento a nivel grupal e individual.
e. Identificar procesos de negocio que estén generando tendencias relevantes, con la mínima duplicación de los datos.
f. Derivar linajes y realizar análisis interrelacionados a través de los datos validados.

- **Reporting Empresariales**

El reporting empresarial proporciona a los individuos, a todos los niveles de la organización, una amplia gama de reporting operacional y de todo tipo procedente de los sistemas ERP, CRM, PRM, facturación, etc., a lo largo de toda la empresa. La distribución de los informes es extensa, y la compensación y otros programas de incentivos suelen estar ligados a los resultados incluidos en estos informes.

La calidad de los datos impacta en el reporting empresarial en que las organizaciones deben:

a. Navegar por múltiples informes e imprimirlos en múltiples formularios que agregan los datos procedentes de fuentes dispares.
b. Seleccionar una variedad de parámetros y personalizar los informes para los usuarios con datos normalizados
c. Presentar múltiples tablas y gráficos con datos reconciliables a lo largo de una variedad de métricas de rendimiento.
d. Permitir que los usuarios de negocio creen sus propios informes sin implicación de IT con datos de alta fidelidad.
e. Reducir las comprobaciones y auditorías manuales con datos limpiados y comparados para la gestión de la conformidad.

f. Expedir facturas y declaraciones contables directamente desde el reporting BI, utilizando datos contables íntegros.

- **Análisis de cubos/OLAP**

OLAP permite a los usuarios “recortar” conjuntos interrelacionados de datos, o “cubos”, interactivamente . Por ejemplo, los usuarios pueden preconstruir las ventas por región para períodos de tiempo específicos, rendimiento por producto, rendimiento por personal de ventas, etc. Las funcionalidades de OLAP, como la navegación up/down, por clase, filtro y página pueden utilizarse para suministrar detalles subyacentes sobre el rendimiento.

Los cubos de análisis pueden ayudar a los usuarios a conducir análisis de partida y compartir los conocimientos con el grupo para una valoración completa. Esto ayuda a poner en marcha una valoración en profundidad utilizando el acceso a los datawarehouse y otros repositorios en lugar de recurrir a funcionalidades de análisis más avanzadas.

La calidad de datos impacta en el análisis OLAP puesto que los usuarios y sus organizaciones necesitan:

a. Navegar por cualquier dimensión para una investigación en profundidad con un completo acceso a los datos “target”.
b. Sencillas manipulaciones OLAP para cualquier subconjunto de dimensiones con datos bien formateados y conformes.
c. Minimizar el reporting conflictivo y garantizar la interactividad con objetos de datos consistentes y subyacentes.
d. Realizar análisis oportunos, dirigidos por el usuario, con datos correctos en múltiples dimensiones.
e. Suministrar datos actualizados y sincronizados para manejar los datos a nivel transaccional en el análisis de cubos.
f. Garantizar la seguridad de los datos cuando se permite a los usuarios crear y mantener los datos de cubos a lo largo de los datawarehouses.

- **Análisis avanzado/predictivo**

El análisis avanzado y predictivo capacita a los usuarios más experimentados a investigar y descubrir los detalles detrás de información específica sobre el rendimiento del negocio, excediendo probablemente los límites típicos del análisis OLAP.

El enfoque puede implicar análisis estadístico avanzado y capacidades de minería de datos. Para dirigir las decisiones proactivas y mejorar las posturas frente a amenazas potenciales del negocio, el análisis predictivo puede incluir la prueba de hipótesis, previsión de incidencias, predicción del suministro y la demanda y calificación de clientes. El modelado predictivo puede utilizarse para anticipar diversos eventos de negocio y los resultados asociados.

La calidad de datos impacta en el análisis avanzado y predictivo, puesto que los usuarios buscan:

a. Crear criterios de filtrado de informes en cualquier elemento de los datos para elaborar informes personalizados.
b. Buscar patrones y conocimientos predictivos por formatos de datos estandarizados para promover la toma de decisiones proactiva.
c. Lograr confianza en el hallazgo de tendencias interdependientes y los resultados esperados gracias a datos consistentes.
d. Emplear regresión de múltiples variantes y otras técnicas sobre datos precisos para lograr mejores predicciones
e. Personalizar las agrupaciones de datos con los mínimos conflictos sin duplicación de datos.
f. Probar hipótesis y usar funciones estadísticas, financieras y matemáticas con datos certificados.

- **Notificaciones y alertas**

Utilizando el correo electrónico, los navegadores, los servidores e impresoras en red, PDAs o portales, las notificaciones y alertas son utilizadas para compartir información de forma proactiva a lo largo de una amplia variedad de puntos de contacto del usuario.

Con el suministro oportuno de la información objetiva, los principales accionistas y responsables de la toma de decisiones pueden identificar áreas potenciales de oportunidades y detectar áreas problemáticas sobre las que actuar.

Este mecanismo de suministro de BI “de primera línea” mantiene a la organización alineada y al corriente de los riesgos y oportunidades de negocio, mientras que los eventos están todavía recientes y son significativos para justificar las respuestas oportunas.[12]

CAPITULO III

PROCESOS ACTUALES PARA LA GESTIÓN DE CALIDAD DE DATOS

3.1 Introducción

Para llevar a cabo una gestión de calidad de datos es necesario contar con diferentes procesos que permita orientar correctamente hacia la obtención de la misma.

A continuación se presenta los procesos de gestión que actualmente tienen organizaciones o empresas para gestionar calidad de datos

Los procesos seleccionados para su estudio corresponden a los propuestos por las empresas:

- POWERDATA
- INFORMATICA
- ADASTRA
- DATACTICS
- DMAIC

Además se expone el proceso según el autor J.Orli.

3.2 PROCESOS DE GESTIÓN

3.2.1 POWERDATA

Según la empresa POWERDATA un proceso de Calidad de Datos cuenta con las siguientes fases fundamentales:

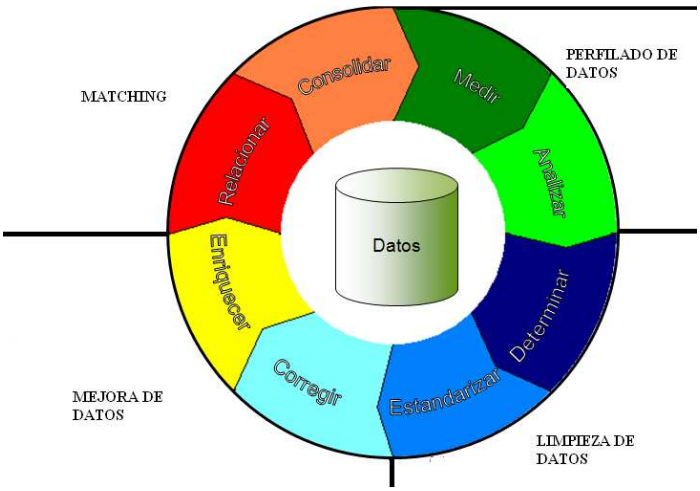


Figura III.1Fases según POWERDATA

3.2.1.1Perfilado De Datos

Un proceso que se lleva a cabo paramejorar lacalidad delos datosesdescubrirlos datos defectuosos a través de perfiles de datos, a veces llamadaarqueología de datos, que es elproceso deanálisiscorrección,integridad, singularidad, la consistencia yrazonabilidad de los datos.

En lugar de hacer una tareadifícilytediosaque requieredocenasdeSQL y programas debúsqueda de anomalías a través detodos losregistrosencadaarchivoobase de datos, existen herramientas de limpieza quetienen ahoralacapacidad deperfilar los datos.

Del mismo modo, se puede aprovechar algunas de las funciones de las herramientas para minería de datos para evaluar la calidad de los datos. Por ejemplo, la función de análisis de valores identifica las características de los valores de datos, tales como ceros, NULL, y el número de valores únicos, mientras que las funciones de análisis de solapamiento identifica el número de claves superpuestas que la cuota de tablas, lo cual es útil para la consolidación de datamart.
Histogramas y gráficos de dispersión permiten detectar visualmente los valores extremos. Además, el SQL generado por la herramienta se puede ejecutar en contra de la base de datos para diferenciar rápidamente el valor de las desviaciones aberrantes de normas.

Además con el perfilado de datos podemos localizar, medir, monitorizar y reportar problemas de calidad de datos

Existen dos tipos de perfilado:

a. Perfilado de estructura.
b. Perfilado de contenido, también llamado análisis de datos.

a. Perfilado de estructura

El perfilado de estructura consiste en el análisis de los datos sin tener en cuenta su significado. Se analiza la información desde un punto de vista estructural. Por ejemplo, un dato que contiene el nombre “Juan J Gómez” no tiene en cuenta si el nombre es válido, simplemente lo analiza y lo identifica como una cadena de caracteres de tamaño de 12 caracteres.

El análisis se realiza de forma semi-automática y masiva. Las soluciones especializadas en este tipo de perfilado pueden analizar cientos de tablas sin apenas necesidad de parametrización.

Tipos de análisis del Perfilado de Estructura:

- **Perfilado de columnas**: análisis de atributos que puede tener una columna de una tabla: tipo de datos, longitud, número de nulos, número

de valores únicos, frecuencias de valores, patrones de caracteres, máximos, mínimos, medias, etc.

- **Perfilado de dependencias:**Análisis de columnas dependientes de otras. Típicamente usado para validación de claves primarias y/o candidatas.
- **Perfilado de redundancias:** Búsqueda de relaciones entre las tablas. Generalmente usado para validación de claves foráneas y/o joins. Análisis de valores “huérfanos”, es decir, valores de una tabla referenciada no existentes en una tabla de referencia (“Child Orphan”), o viceversa (“Parent Orphan”).

b. Perfilado de contenido

El perfilado de contenido consiste en el análisis de la información contenida en los datos. Se analiza la información desde un punto de vista sintáctico y semántico. Por ejemplo, un dato que contiene el nombre “Juan J Gómez”, podría indicar que contiene un nombre válido, una letra que podría ser una inicial de un segundo nombre y un apellido válido.

Este tipo análisis requiere una configuración para cada campo para aplicar las reglas de negocio pertinentes en cada dato. Se deben combinar componentes específicos para tratamientos de cadenas, diccionarios o listas de valores e incluso diccionarios de patrones válidos.

Con el perfilado de contenido también puede obtenerse las mismas conclusiones que el perfilado de columnas del perfilado de estructura (es decir: tipo de datos, longitud, número de nulos, número de valores únicos, frecuencias de valores, patrones de caracteres, máximos, mínimos, medias, etc.), sin embargo sería necesaria una configuración para cada campo.

- **Etapas del perfilado de datos**

A continuación se muestra las etapas que se requiere para realizar el perfilado de datos:

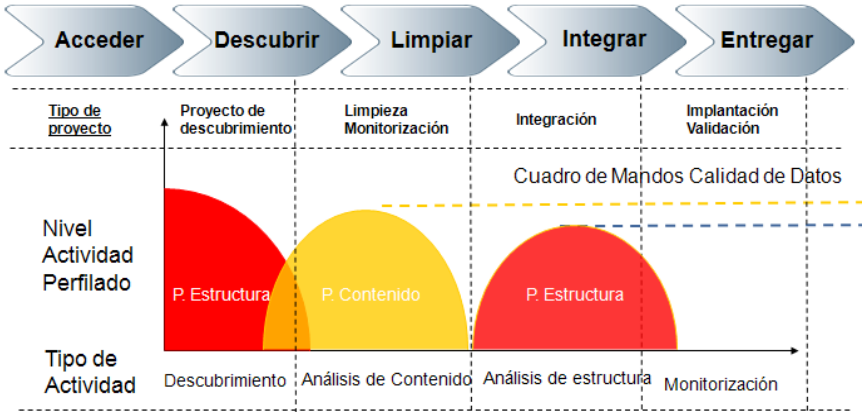


Figura III.2 Etapas del perfilado de datos

- **Indicadores de calidad del perfilado de datos**

Durante el perfilado, es importante no sólo la documentación de los datos, sino la detección de errores y su clasificación.

Por lo general se pueden distinguir seis categorías principales de indicadores de calidad de datos que se obtiene mediante el perfilado de datos como se muestra en la siguiente figura:

Figura III.3 Indicadores de Calidad

- **Existencia:** Datos omitidos (nulos, en blanco), o que contienen información no útil (valores por defecto, como “N/D”).
- **Conformidad:** Adecuación del formato de un dato con respecto a un estándar establecido. Por ejemplo, que los teléfonos de Ecuador estén todos en el formato de nueve dígitos, sin prefijo internacional, ni guiones, ni espacios.
- **Consistencia:**Nivel de coherencia entre dos o más conjuntos de datos. Por ejemplo, una columna indica que el cliente es varón y tiene por nombre “Luisa”.
- **Precisión:** Existencia de datos incorrectos u obsoletos. Generalmente se comprueba comparando con datos de referencia. Por ejemplo: fecha de nacimiento “1/3/1890”
- **Duplicados:** Datos referentes a la misma entidad repetidos y que no aportan información añadida. Por ejemplo: una misma persona que aparece dos o más veces en la misma tabla.
- **Integridad:**Problemas tanto de integridad referencial (claves primarias y foráneas), como de datos relacionados no unidos por un campo común. Por ejemplo, empresas del mismo grupo de empresas no relacionadas entre sí en una base de datos.

3.2.1.2Limpieza y Enriquecimiento de Datos

La fase de limpieza cuenta con las siguientes etapas:

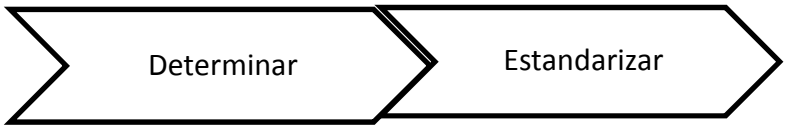


Figura III.4 Etapas de la Fase Limpieza de Datos

Después de medir datos erróneos conocidos, el lugar más fácil para empezar el proceso de mejora de la calidad de los datos es por la limpieza de datos operacionales en el momento en que se mueve en las bases de datos donde se utiliza para informar a toda la organización.

Sin embargo, lalimpieza de datos es un proceso laborioso, lento y costoso, y la limpieza de todos los datos por lo general no justifica los costos. Por otro lado, ninguna limpieza de los datos es igualmente inaceptable. Por ello es importante

analizar cuidadosamente los datos de origen y clasificar los elementos de información, crítica, importantes o insignificantes para la empresa.
A continuación, se concentran en la limpieza todos los elementos de datos críticos, y como el tiempo lo permite, limpiar la mayor cantidad de elementos importantes de datos como sea posible, dejando los elementos de datos no significativos.
En otras palabras, no es necesario limpiar todos los datos, y no hay necesidad de hacerlo todo a la vez.

Otro factor que influirá en la capacidad de limpiar los datos es si los datos correctos todavía existen o si pueden ser recreados con una cantidad mínima de esfuerzo manual o automatizado. En ese caso, podría ser mejor dejar sólo los datos únicos.

La limpieza de datos permite:

- Determinar y separar (parsing) elementos de un campo situándolo en su lugar correspondiente. Por ejemplo: “c/ Juan Bravo 34, 1ºB” separarlo en: tipo vía: “Calle”, nombre vía: “Juan Bravo”, número portal: “34”, piso: “1º”, puerta: “B”
- Estandarizar formatos. Por ejemplo el teléfono: +34.609.039.049 convertirlo en 609039049
- Corregir errores en los datos. Por ejemplo, corrección de códigos postales en base a la vía y a la localidad

- **Determinación y Separación de Datos**

La determinación y separación de datos consiste en la descomposición de los distintos elementos que componen los datos.

Existen múltiples modos para separar los datos. Desde métodos simples por medio de subcadenas (usado para referencias de productos, números de cuenta, separación del prefijo del teléfono), hasta métodos más complejos que usan diccionarios de valores y componentes que tienen en cuenta los patrones, la posición de un elemento, el tipo de dato, etc.

Ejemplos de determinación: separación de una descripción de producto en familia, tamaño, marca, etc (por ejemplo: “Pantecta 500 ml cápsulas” en: familia: Pantecta, posología: 500 ml, presentación: Cápsulas), un nombre completo en nombre, apellido 1 y apellido 2, una dirección en tipo de vía, nombre de la vía, número, etc.

- **Estandarización**

La estandarización es la adecuación de un dato a un formato esperado.

Es importante que un dato con formato esté estandarizado, pues facilita las búsquedas por campos indexados, mejora la visualización y se obtienen mejores resultados en la fase de matching posterior.

Las operaciones necesarias para la estandarización dependen de la naturaleza del dato. En el ejemplo, para calcular el dígito de control es necesario un pequeño “script” o código. Otros elementos clásicos de estandarización es el reemplazo de caracteres, sustitución de elementos a partir de diccionarios de traslación y componentes de tratamiento de mayúsculas y minúsculas.

3.2.1.3 Mejora de Datos

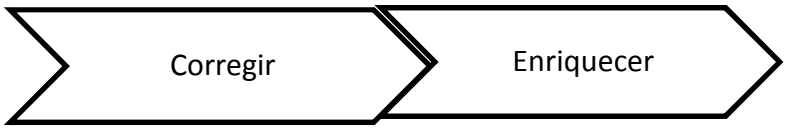


Figura III.5 Etapas de la Mejora de Datos

- **Corrección**

La corrección consiste en el reemplazo de un elemento erróneo por uno correcto.

Generalmente se realiza a través del reemplazo de los datos originales por los existentes en una fuente de datos de referencia o con diccionarios. Evidentemente estas fuentes deben de ser de elevada fiabilidad y calidad.

La relación con fuentes externas o diccionarios se realiza a través de la relación directa con otros campos o a través de un proceso de matching. Esta fase puede ser, por lo tanto, posterior a una fase de matching intermedia.

- **Enriquecimiento**

El enriquecimiento consiste en añadir datos a la fuente que no existían originalmente.

Generalmente se realiza a través de la incorporación fuente de datos de referencia o con diccionarios.

Técnicamente es muy similar al proceso de corrección, en la que en lugar de un dato incorrecto, lo que se reemplaza es un dato vacío.

- **Pregrouping**

El “pre-grouping” o preagrupación es un paso previo al matching, no obligatorio, pero muy conveniente.

En el matching los registros se comparan por parejas, es decir, si tenemos una tabla con 4 registros, no se compararán todos al mismo tiempo sino en grupos de dos en dos: el primero con el segundo,el primero con el tercero, el primero con el cuarto, el segundo con el tercero, etc.

En volúmenes grandes el número de combinaciones puede ser muy elevado. Por ejemplo, para 1.000.000 de registros el número de combinaciones de dos a dos es de: 499.999.500.000.

Por ello, a menos que se trate de un volumen pequeño de registros (menos de 10.000), es conveniente pre-agrupar los registros.

3.2.1.4Matching

El matching de datos se utiliza para:

- Detección de duplicados

- Relación entre dos fuentes de datos que no tienen campos de unión entre sí
- Detección de unidades familiares y corporativas (Householding)

Se pueden aplicar múltiples criterios para las relaciones, que posteriormente se pueden asociar entre sí

Existen dos métodos de matching:

a. Determinístico
b. Probabilístico

a. Matching Determinístico

En el Matching Determinístico se comparan los diferentes atributos asociados a la entidad a comparar. El resultado de la comparación puede terminar con uno de los siguientes resultados: positivo o negativo.

Generalmente se trata de comparaciones por igualdad, aunque se pueden realizar transformaciones, normalizaciones, codificaciones y limpiezas para una comparación más adaptada al mundo real.

Este método, sin embargo, es muy sensible a posibles errores tipográficos no contemplados en la estandarización. Por ejemplo: a la hora de comparar “Teresa” con “Theresa” dará como resultado un match negativo en el nombre.

b. Matching Probabilístico

En el Matching Probabilísco se comparan los diferentes atributos asociados a la entidad a comparar con algoritmos específicos. Estos algoritmos no realizan comparaciones de igualdad, sino que devuelven un porcentaje de similitud entre los dos atributos comparados.

Los algoritmos de comparación deberán ser adecuados para el tipo de datos, puesto que no es lo mismo comparar una cadena de texto libre (como un nombre, razón social,

descripción de producto, etc.) que un código (por ejemplo, teléfono, código postal, número de referencia, etc.).

Al igual que con el matching determinístico, es conveniente realizar transformaciones, normalizaciones, codificaciones y limpiezas previas para una comparación más adaptada al mundo real.

Finalmente, se toman todos los porcentajes obtenidos de las diferentes comparaciones y se realiza una media ponderada. Ciertos atributos pueden tener un mayor peso que otros, por ejemplo, al comparar empresas tendrá más peso la razón social que el teléfono.

El resultado final no será positivo-negativo, sino que será un valor porcentual con el nivel de semejanza. Pueden ajustarse uno o varios umbrales por los que se considere relacionado un par de registros si tiene un rango porcentual determinado. Por ejemplo, entre 90% y 100% puede considerarse un match positivo, entre 82% y 90% un match probable, y entre 75% y 82% un match dudoso.

Previo a la comparación puede realizarse un muestreo de todos los valores de uno o varios atributos, de modo que los valores más comunes tienen un peso inferior en la comparación final y los menos comunes tienen un peso mayor.

Por ejemplo, a la hora de comparar personas, puede tener más peso la coincidencia del apellido “LUCENA” que “SÁNCHEZ” por ser más infrecuente. Incluso agrupado por otro atributo: por ejemplo: el nombre “JORDI” es común en Ecuador, pero no lo es en Colombia. Los muestreos más complejos pueden ser realizados con una solución de datamining y usarse como fuente del proceso de matching probabilístico.

- **Consolidación**

Los duplicados detectados durante la fase de matching pueden tratarse de muy diversas maneras.

Por ejemplo, para los clientes de un banco o de una empresa aseguradora simplemente se crean códigos de relación entre los diferentes duplicados y se mantienen los datos

originales asociados a cada cuenta o cada póliza. Existirá un código de cliente y debajo de este código de cliente, distintas versiones del cliente por cada instancia duplicada.

Sin embargo en muchos otros casos: entornos de marketing, CPG, DWH, etc, es deseable que los diferentes registros duplicados desaparezcan físicamente de la tabla. Por ejemplo, si un cliente está triplicado, se desea suprimir los dos registros redundantes y almacenar un único registro.

Puede ser una problemática compleja, pues requiere re-asignar todas las transacciones (facturas, albaranes, pedidos, reclamaciones, etc) asociadas a la entidad a suprimir a la nueva. Y además existe la dificultad de elegir qué registros deben ser los borrados y qué datos deben permanecer sobre esa entidad.

Para este último aspecto como se indica en la Figura III.6 existen dos técnicas de consolidación:

- Registro superviviente (“survivorship”)
- Mejor registro (“Best Record”)

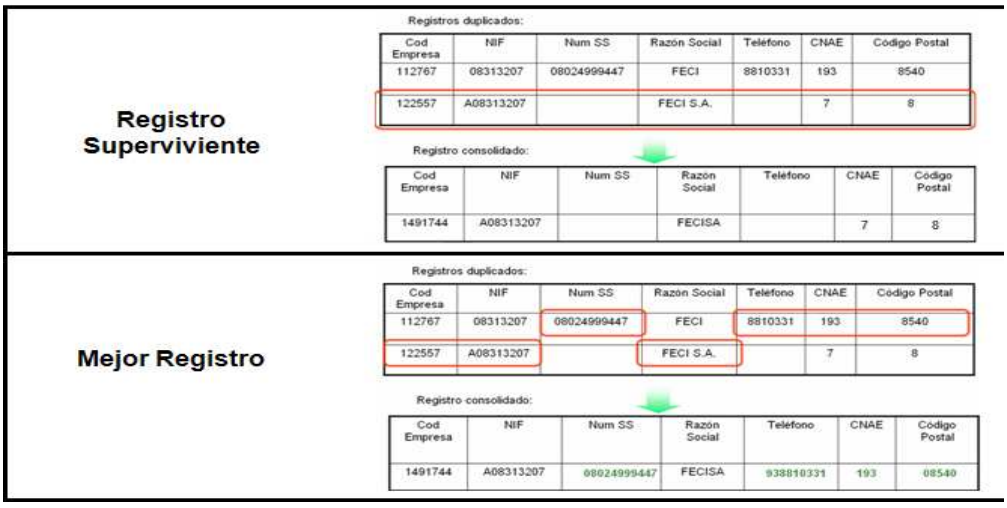


Figura III.6 Técnicas de consolidación

Consolidación por registro superviviente:

Consiste en la selección de uno de los registros del conjunto de registros duplicados. El criterio de selección depende de la naturaleza del dato. Con frecuencia se usa el registro más reciente o el más completo.

Consolidación por mejor registro:

Consiste en la combinación de diferentes datos de cada conjunto de registros duplicados, para componer el registro más completo y mayor calidad posible.

Se establece un criterio de selección para cada dato dependiendo de su naturaleza. Generalmente es el nivel de calidad el que decide el dato más adecuado, aunque pueden usarse otros criterios alternativos.

Por ejemplo, para la clave primaria: puede seleccionarse la clave primaria que tenga asociadas más registros en sus tablas referenciadas. Si tenemos un cliente duplicado, y uno de los registros de cliente tiene asociada 20 facturas y la otra una sola, es más lógico utilizar el código de cliente del que tiene asociada las 20 facturas, pues se requerirá menos actualizaciones en el sistema de facturación.

Pueden existir datos acumulables. Por ejemplo, en el caso del teléfono: una persona o empresa puede tener varios teléfonos, y puede que nos interese almacenar todos los teléfonos diferentes de cada conjunto de duplicados en lugar de seleccionar únicamente uno. Evidentemente el modelo de datos de destino deberá haber tenido en cuenta este hecho.[13]

3.2.2 INFORMATICA

La calidad de datos se gestiona mejor como parte de una arquitectura de integración de datos empresariales y, como resultado, el control y la gestión de la calidad de datos se complementa con el ciclo de vida de acceso, integración, transformación y entrega de los datos.

Como parte del programa de calidad de datos, las organizaciones necesitan establecer o restablecer procesos de calidad de datos como se muestra a continuación

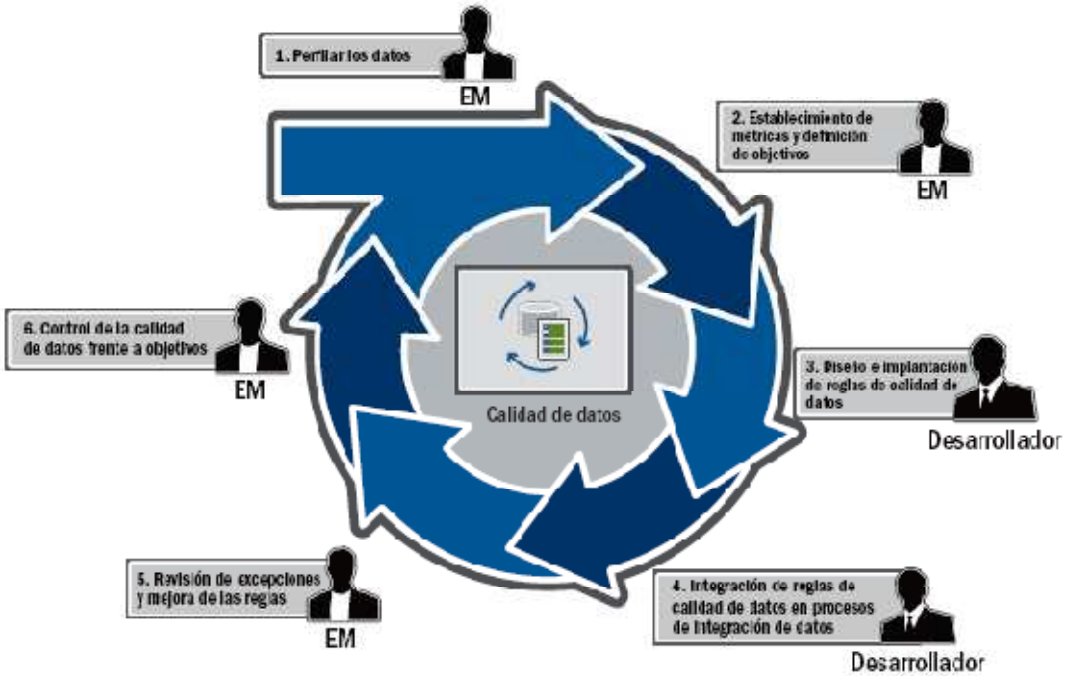


Figura III.7 Fases INFORMATICA

3.3.2.1 Perfilado de Datos

Es un elemento clave en la planificación de las iniciativas de calidad de datos, ya que permite determinar el contenido, la estructura y la calidad de estructuras de datos muy complejas, así como descubrir incoherencias ocultas e incompatibilidades entre las fuentes de datos y las aplicaciones de destino.

3.3.2.2 Establecer métricas y definir los objetivos

Ayuda a los equipos de IT y a las empresas a medir los resultados obtenidos gracias a los esfuerzos realizados para garantizar la calidad de datos como parte de la iniciativa de BI.

3.3.2.3 Diseño e implementación de reglas de calidad de datos

Ayudan a definir y medir los objetivos y los criterios de la calidad de datos.

3.3.2.4 Integración de reglas y actividades de calidad de datos

Creación de perfiles, limpieza/correspondencia, solución automatizada y gestión con los procesos de integración de datos es fundamental para mejorar la precisión y el valor de los activos de datos.

3.3.2.5 Revisión de excepciones y la mejora de las reglas

Se realizan de forma más eficaz como un esfuerzo conjunto que implica a miembros del equipo principal y a interesados de BI. En muchos casos, éstos últimos tienen un control limitado sobre los procesos empresariales y los sistemas operativos y esto hace que los datos sean de mala calidad.

Por este motivo, es importante que los principales interesados y los ejecutivos de una organización participen en la documentación de los problemas de calidad de datos y en la ejecución de un programa de calidad de datos formal.

3.3.2.6 Control proactivo de calidad de datos

Este control en cuadros de mando y notificaciones en tiempo real también se está convirtiendo en una de las mejores prácticas estándar. Los propios interesados de BI que participan activamente en el proceso de calidad de datos, pueden contar con las herramientas necesarias para ejercer esta tarea, ya que son los que mejor conocen cuál es el nivel de calidad que deben tener los datos.[14]

3.2.3 ADASTRA

La calidad de datos no es un problema de TI, es un problema de toda la empresa y un activo fundamental que depende en gran medida en los procesos de negocio. También está claro que la calidad de datos no es una simple tarea de una sola vez,es un complejo proceso iterativo y cíclico que emplea personal, herramientas y conocimiento. Usando este enfoque permite la identificación de problemas de calidad de datos en sus fuentes y conduce a eliminar las causas de los problemas en lugar de sus consecuencias

Según la empresa ADASTRA hay cuatro fasesbásicas en la gestión del ciclo de calidad de los datos:

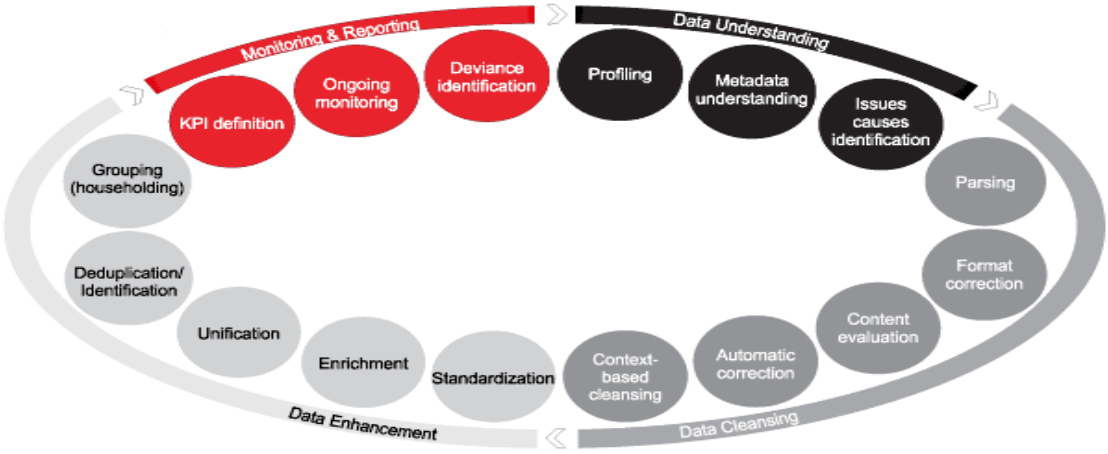


Figura III.8 Fases y Etapas ADASTRA

3.2.3.1 Comprensión de datos

Esun conjunto de actividades entre ellas se encuentran los procesos de perfilado de datos, metadatos e identificación de posibles causas

3.2.3.2 Limpieza de datos

Es el proceso mediante el que la calidad de datos se ha mejorado y los problemas de datos junto con sus causas se resuelven y los procesos relativos son corregidos o mejoran

3.2.3.3 Mejora de los datos

Este proceso le permite agregar valor a los datos existentes (mediante la adición de la información disponible de otros-fuentes de datos externas como la codificación geográfica, la dirección detallada y-o información sobre productos, etc.).

3.2.3.4 Monitoreo de los datos y elaboración de informes

Comprueba constantemente los datos y se identifican problemas nuevos datos o pérdida de calidad

Proporciona a los equipos de aplicación y los clientes de datos con la información sobre el éxito de las mejoras anteriores actividades

Estas fases traen los siguientes beneficios:

- Rápida y clara identificación de los problemas de calidad de datos junto con la importancia de la evaluación de las perdida de comerciales y por tanto de un solo tema

- Rápida identificación de las necesidades técnicas y de negocio
- Selección de la solución ,los procesos de negocios rápida y el diseño de arquitectura de soluciones basadas en las mejores prácticas [15]

3.2.4 DATACTICS

Este proceso de gestión de calidad de datos combina la experiencia de los analistas de datos y proporciona tecnología de generación de excelencia en calidad de datos en cada proyecto.

Ha sido desarrollada a partir de experiencias en situaciones de la vida real de calidad de datos, proporcionando un marco coherente.

Según DATACTICS el proceso de gestión de calidad de datos está compuesto de las siguientes fases:

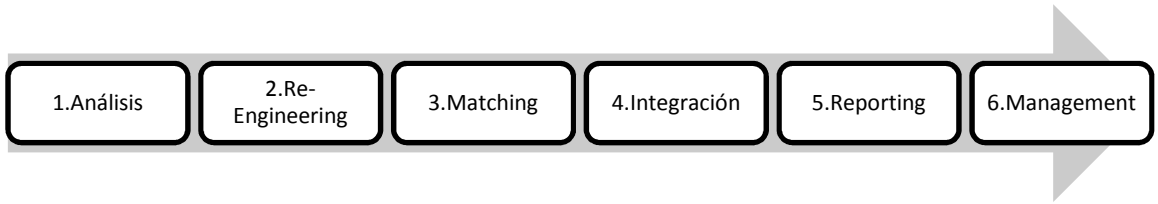


Figura III.9 Fases de DATACTICS

3.2.4.1Análisis

El paso inicial de cualquier estrategia de gestión de datos integral de calidad tiene que ser el descubrimiento de lo que exactamente contiene los datos que es importante y que no.

3.2.4.2Re-Engineering

Una vez determinado el contenido de los datos, la fase de re-ingeniería permite corregir errores, transformar los datos a las normas requeridas; mejorar los datos con información adicional y si es necesario, extraer elementos clave y de valor de la misma.

En resumen, asegurarse de que los datos están "aptos para el propósito".

3.2.4.3Matching

Tener datos reestructurados a un formato y nivel adecuado se puede llevar a cabo en su nivel más eficaz.

3.2.4.4Integración

La calidad de los datos no puede ser visto como un proceso aislado, por lo que la capacidad de integrar una metodología de gestión de calidad de los datos en los procesos empresariales existentes es fundamental.

3.2.4.5Reporting

Después de que las fases anteriores se han terminado, una gran cantidad de conocimientos e información sobre los datos han sido recogidos. La posibilidad de revisión, auditoría y compartir esta información es vital para una verdadera cultura de calidad de datos va a crecer y mejorar iterativamente en toda la organización.

3.2.4.6Management

La gestión de todos los procesos anteriores dentro de un único y simple ambiente proporciona un gran beneficio. Facilita el aumento de la productividad mediante la racionalización del flujo de trabajo y también proporciona una capa totalmente transparente desde el que profesionales de la calidad de los datos puede tanto monitorear y ejecutar los procesos de calidad de datos. [16]

3.2.5 PROCESO DE GESTIÓN ORLI

Según J.Orli las fases son básicas: Definiciónde los problemas, seguido de problemasdeidentificacióndedatos, Análisis y Mejorade los pasos para cada problema.

A continuación se ilustra las fases mencionadas:

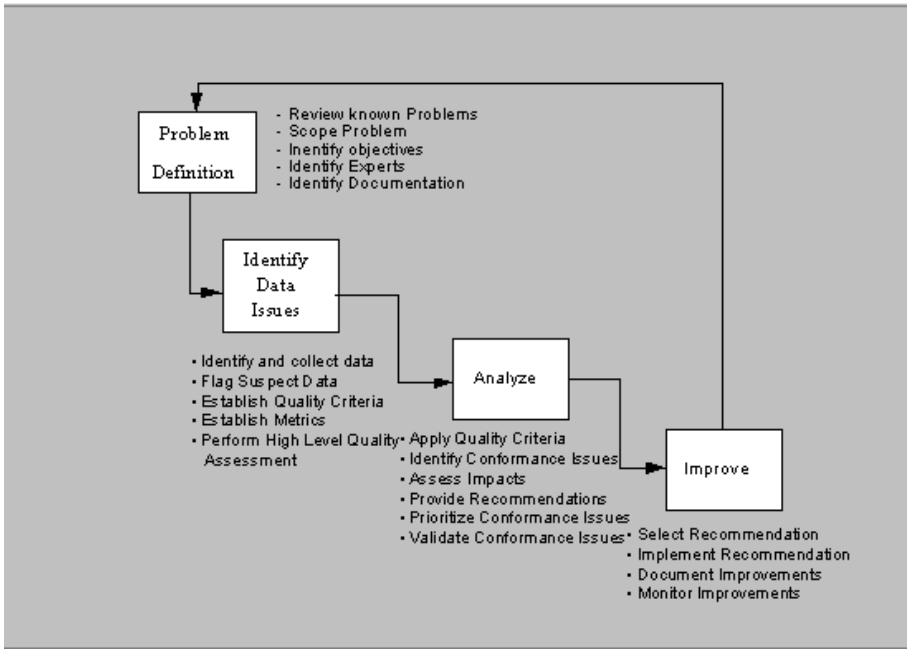


Figura III.10 Fases ORLI

Para cada área del problema, los problemas se identifican en detalle.Una parte importante del proceso se identifica exactamente en qué parte del ciclo de vida de los datos se origina el problema.A menudo, un problema de calidad de los datos requiere dos esfuerzos separados, un proyecto para corregir los datos que ya existe, y un proyecto para corregir la causa detrás de los problemas de datos.

3.2.5.1Definición del Problema.

Esta fase consta de:

- Revisión de problemas conocidos
- Ámbito del problema
- Identificación de objetivos
- Identificación de expertos
- Identificación de documentación

3.2.5.2Identificación de problemas de datos.

Esta fase consta de:

- Identificación y almacenamiento de datos
- Datos señalados como sospechosos
- Estableciendo criterios de calidad

- Estableciendo métricas
- Realizar una evaluación de alto nivel de calidad

3.2.5.3 Análisis.
Esta fase consta de:

- Aplicar criterios de calidad
- Identificar problemas de calidad
- Evaluación de impactos
- Entregar recomendaciones
- Priorizar problemas de conformidad
- Validar problemas de conformidad

3.2.5.4Mejoramiento.
Esta fase consta de:

- Seleccionar recomendaciones
- Implementar recomendaciones
- Documentar mejoras realizadas
- Monitorear mejorasrealizadas [17]

3.2.6PROCESO DE GESTIÓN DMAIC

Según la empresa DMAIC un proceso de gestión de calidad de datos consta de las siguientes fases:

Figura III.11 Fases DMAIC

3.2.6.1 Definir

La fase Definir consta de la siguiente etapa:

1. **Crearun diccionario de datos**

Inclusosiunesfuerzo porelaborarnose ha hechopara ejecutarunrepositorio demetadatos central, latareainicialparael equipo degestión dedatos debeserlosiguiente:

Identificartodos losarchivos deproducciónde los sistemas fuente

- campos importantes
- valoresesperados

Obtenerinformaciónde los repositoriosde metadatos, si se dispone de:

- repositorios de la empresa
- diccionarios de datos

Extraer datos desde definición de datos de objetos

- COBOL copybooks
- datos PL / 1
- catálogos de base de datos

El resultado de esto podría ser almacenado en un repositorio como MS Excel. Unejemplo de la puesta a punto ideal de la arquitectura de datoses la siguiente. Eldiccionario de datos es un documento clave que se utilizaen todas las fases de una aplicación.

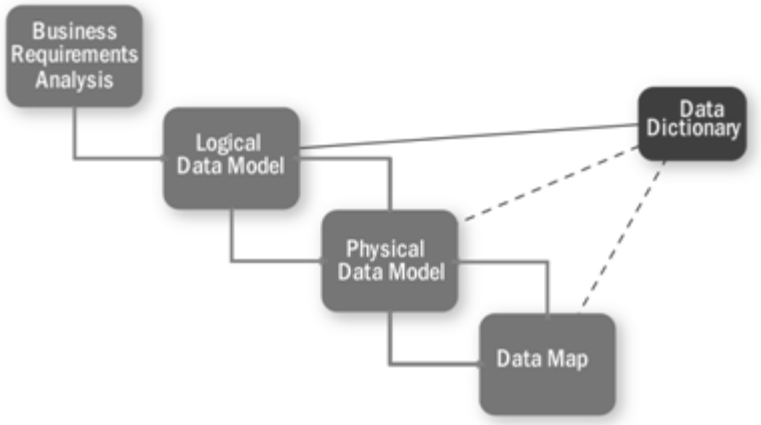


Figura III.12 Diccionario de datos

3.2.6.2 Medir

Creación deinformesde datospara definir unalínea basede calidadde datos. Elequipodebebuscar cosascomo:

Inspeccionarlos valores de datos

- relación calidad/lista defrecuencia
- puntoslimites(valores altos / bajos), (valores corto /largo)

- los valoresconcaracteresespeciales

Tomar en cuenta los valores que son:

- nulosovalores perdidos
- valores únicos
- los valoresconstantesovalores de los códigos

Buscar el usodecolumnas

- columnas de bajo uso(alto porcentaje deespacios en blanco oceros)
- valores de los datossin usar desdetablas de códigos
- indicacionesdeunnulocomo"sin valor", en blanco, "n / a", y otros
- inspeccionarla estructura
- los registroshuérfanosointegridadreferencial

El reconocimiento de patroneses decir,buscarformatospredefinidos y lafrecuencia decadamodelo:

- números deteléfono-con ysinguionesy paréntesis
- número de Seguro Social-con ysinguiones
- códigopostal-con ysinel PLUScampo de 4

Valores derivados

- cálculoscomoRESULTADO=INGRESOS-GASTOS
  -Las reglas de negociocomoifCLIENTE_TIPO="preferido" then descuento=10%
- tercera forma normal
  Lasalida deentregapodríaseren formadeun informeintegrado de unalistade losdefectosanotadosconinformaciónclave, talescomoimpacto en elnegocio, nivel crítico, frecuencia yresoluciones posibles. Porejemplo:

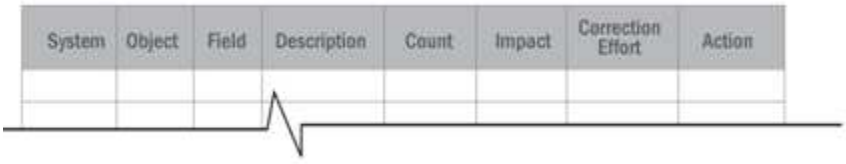


Figura III.13 Reporte de salida

Sistema-nombredelsistemaen elqueelerroro defectose encuentra

Objeto-nombredelobjeto(persona, tabla) donde elerror/defectoseencontró

Campo-nombredelcampo(atributo, etiqueta) en relación conel error

Descripción delerror-descripción detalladadel defecto

Conteo- el númerodeerroresqueseencontrarondeestetipo deerror

Impacto-en funcióndelacomprensióndelimpactoen lafuncionalidadde negocioy suimportancia, clasificarel defectocomocrítica, alta,mediaobaja

Esfuerzode corrección-se trata deexplicareltiempoyel esfuerzoquepodríanadoptar paracorregir unerrordeesta categoría

Acción-estoesparaexplicarcómo estetipode errorpodría sermanejado,por ejemplo,paracrearscriptsautomatizadosde actualización, de forma manualen lalimpiezasistema deorigen, no de acción, etc

3.2.6.3 Analizar

Comparación de informes de datos generadoscon el proyecto documentación, si esqueexiste(mapa de datos, diccionario de datos, altonivel dediseños, dibujos y modelosde bajonivel).

Identificar lasdiferenciasenlos valores esperadosy los informese identificardonde las transformacionesson requeridas.

- Encontrar con configuraciones y diseños de interfaz cuandovaloresinesperadosy /otransformacionesindocumentadossonnecesarios ycrear unplandeacciónparaestos.

  Lasalida entregablepodría serun documento que indique la causaraízque tenga los siguientes elementos:
- Con baseenlos resultados delacalidad de los datosinicialesevaluación, las cuestiones de grupoespecíficode datosenmás genéricoobservaciones.
- Identificarlacausaraízpreguntando a expertos en materias de negocios (PYME), desarrolladores y expertostécnicos
- Por último, formularrecomendacionessobrecómoabordartemasespecíficos, después de hablar conlas pymes, los desarrolladores,yexpertos entecnología.

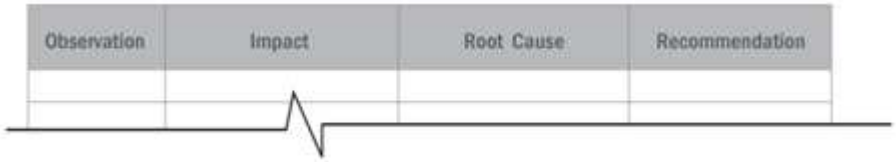


Figura III.14 Reporte Causas de origen

3.2.6.4 Mejorar

Trabajar con los equipos de desarrollo y equipos de sistemas fuentespara que realicen actualizaciones en:

- Su diccionario de datos con valores esperados para el código de valor de campo
- definiciones de datos con cualquier atributo nuevos descubierto acerca de los datos
- Mapas de los datos para indicar el tipo de datos. Agregar tipo de datos columna de atributos y rellenar con el tipo de datos para cada atributo para cada sistema
- Facilitar el consenso sobre nuevas transformaciones necesarias y / o aspectos únicos de atributos de datos

3.2.6.5 Controlar

Debe haberunseguimientocontinuode losdatoscríticos con métricasde calidadparaasegurarquelos datoslimpiadosno se caigan por debajo deunumbralde calidad. Las personasa menudocreen queuna vez que ha sido solucionado identificado yaplicado, un determinado problemaseguirá estará resueltopara siempre. Si bien estoesa menudo el caso, las cosasrara vez son así de sencillocuandolas personasylos procesos están involucrados. Es aconsejable medidas continuas para lacalidad de los datosalmacenados. [18]

3.3 Análisis de los Procesos

En la siguiente tabla se muestra las fases en común que contiene cada una de las empresas seleccionas lo cuál permite tener una visión de que fases se destacan para la gestión de calidad de datos.

Tabla III.1 Fases en común de los procesos

	POWERDATA	INFORMATICA	ADASTRA	DATACTICS	ORLI	DMAIC
Perfilado	x	x	x			
Limpieza	x		x			
Mejora de datos	x		x		x	
Matching	x			x		
Establecimiento de métricas y objetivos de calidad de datos		x				
Diseño e implementación de reglas de calidad		x				
Integración de reglas de calidad de datos en procesos de integración de datos		x				
Revisión de excepciones y mejora de las reglas		x				
Control de la calidad de datos frente a objetivos		x				
Comprensión de datos			x			
Monitoreo			x		x	
Análisis de datos				x		x
Integración				x		
Reporting	x			x		
Management				x		
Definición del problema					x	x
Identificación de problemas de datos					x	
Medir						x
Control						x
Re-Engineering				x		

Al tener en cuenta cuáles fases tienen en común los procesos expuestos anteriormente se ha realizado un análisis de las ventajas y desventajas que poseen estos como se muestra a continuación:

Tabla III.2 Ventajas y Desventajas de los procesos

Proceso	Ventajas	Desventajas
POWERDATA	Cuenta con fases básicas para la calidad de datos Explicación eficiente	No cuenta con fases orientadas al estudio del negocio Falta de fases y etapas definidas No contiene una planificación especifica
INFORMATICA	Orientado a proyectos de integración Formación de equipos de calidad de datos	No cuenta con fases o etapas definidas No contiene una planificación especifica
ADASTRA	Estructura clara y definida Cuenta con etapas por cada fase	No cuenta con fases de estudio del negocio No contiene una planificación especifica
DATACTICS	Cuenta con fases básicas para la calidad de datos	No cuenta con una fase de estudio del negocio No cuenta con fases ni etapas definidas No contiene una planificación especifica
ORLI	Cuenta con fases básicas para calidad de datos	No cuenta con una fase de estudio del negocio No cuenta con fases ni etapas definidas No contiene una planificación especifica
DMAIC	Estructura clara y definida Cuenta con etapas por cada fase	No cuenta con una fase de estudio del negocio No contiene una planificación especifica

CAPITULO IV

DESARROLLO DE LA METODOLOGÍA PROPUESTA

4.1 INTRODUCCIÓN

En este capítulo se desarrolla la propuesta metodológica basándose en los procesos de gestión de calidad de datos estudiados anteriormente y además construyendo un proceso metodológico que ayude al negocio a tener una idea clara y ordenada de los pasos a seguir para una correcta gestión de calidad de datos.

La metodología esta compuesta por siete fases, en la primera fase se realizar un reconocimiento del negocio su ámbito de trabajo y se planifica el curso que seguirá el proyecto, en la segunda fase se analiza la información actual que tiene el negocio lo que permite saber como es la forma de trabajo del negocio,en la tercera fase se realiza un análisis inicial de los datos para su posterior evaluación y con esto determinar los problemas existentes, con esta información en la cuarta fase se procede a la limpieza de datos, luego de la limpieza de datos se pasa a la quinta fase donde se realiza un análisis de los datos para saber el estado final de los mismos, con la información obtenida se procede a la fase 6 de mejoramiento y prevención donde se indica un plan de mejoramiento y prevención para evitar caer en los mismo problemas para finalmente realizar en la fase siete un seguimiento y control de lo realizado.

4.2 CONCEPTO METODOLÓGICO

Algunas organizaciones esperan mejorar la calidad de los datos en movimientos de datos desde sistemas heredados a planificadores de recursos empresariales(ERP) y paquetes de administración de relación con el cliente(CRM) .Otras organizaciones usan herramientas para perfilado de datos y limpieza de datos para desenterrar datos incorrectos, y luego limpiarlos con una herramienta de extracción, transformación y carga ETL para aplicaciones de datawarehouse .

Todo este esfuerzo para mejoramiento de la calidad de datos orientados a la tecnología son recomendables y definitivamente un paso en la dirección correcta .Sin embargo solo las soluciones tecnológicas no pueden erradicar las causas que originan la pobre calidad de datos debido a que la pobre calidad de los datos no es un problema IT es un problema de negocio.

Otras disciplinas de grandes empresas deben ser desarrolladas, instruidas e implementadas y se esfuerzan para mejorar la calidad de datos de una manera holística. Debido a que el mejoramiento de la calidad de datos es un proceso y no un evento la empresa u organización debe optar por una metodología de gestión de calidad de datos .

Los sistemas de información actuales necesitan integrar grandes cantidades de información de múltiples fuentes de datos para resolver requerimientos complejos de los usuarios. Un desafío en este tipo de sistemas es proveer al usuario con información adaptada a sus requerimientos de calidad. La calidad se expresa como un conjunto de

factores de calidad que miden ciertos aspectos relevantes de los resultados, como la actualización de los datos, la completitud o el tiempo de respuesta.

Debido a la heterogeneidad de las fuentes de datos resulta difícil evaluar la calidad de los datos para brindar a los usuarios respuestas uniformes y de alta calidad.

Esta investigación trata el problema de gestionar la calidad de la información producida por un proyectodeintegracióndedatos en el cual la calidad de la información devuelta al usuario depende principalmente de la calidad de las fuentes de datos y de las características del procesodecálculoque construye dicha información a partir de las fuentes. Más concretamente, la calidad depende de la calidad interna de las fuentes (la coherencia, la completitud, la actualización, etc.), de la confianza sobre quién produce los datos de esas fuentes, y también de la forma de producir la información devuelta al usuario.

A continuación se presenta el desarrollo de una metodología la cual proporciona un conjunto de decisiones centradas en la organización los objetivos de calidad de datos que determinan los datos para mejorar los procesos, las soluciones a implementar, y la gente a participar.

4.3 CARACTERÍSTICAS TÉCNICAS

La metodología propuesta se caracteriza por:

- Contiene una estructura organizada que permite al equipo de gestión de la calidad de datos encontrar fácilmente la información deseada.
- Metodología de tipo activo es decir los procesos se centran en el equipo de calidad de datos

- Gestiona la calidad de datos en proyectos de integración de manera integrada con la finalidad de obtener resultados exitosos en beneficios de la empresa u organización

4.4 ARQUITECTURA DE LA METODOLOGÍA PROPUESTA

De acuerdo a la investigación realizada que se lo detalló en el marco teórico, se observa que procesos y estrategias relacionados con la calidad de datos generalmente son abstractos y de alto nivel, no contienen la suficiente narrativa, no son funcionales o les falta integración con los objetivos de la organización.

Una de las características de esta metodología es que en cada etapa se describen procesos; a su vez, cada proceso se describe en términos de entradas, salidas, herramientas, técnicas y herramientas software. Las entradas y salidas son documentos; las herramientas y técnicas constituyen mecanismos aplicados a las entradas para crear salidas. Las herramientas software proporcionan un método de ayuda para la rápida generación de resultados

Figura IV.1 Arquitectura de la Metodología Propuesta

- **Planificación**. Este nivel se refiere a todas las fases relacionadas con la definición de una estrategia y el planteamiento para desarrollar las oportunidades que una empresa/organización desea llevar a cabo. Todas las

fases de este nivel ayuda en los planes de organización para el camino hacia el logro de las metas de la organización internos y externos.

- **Ejecución y análisis.**En este nivel se realiza la ejecución práctica de los procesos para la gestión de calidad de datos dentro de la organizacióny las áreas de atención que deben ser considerados para apoyar la uniformidad en el cumplimiento detareas de las diferentes áreas.
- **Observación.**Se relaciona con el control contínuo, la medición de los resultados y el impacto de las acciones adoptadas. También es compatible con la evaluación de las posibilidades de mejora.

4.5 ESTRUCTURA DEL PROCESO METODOLÓGICO PROPUESTO

El proceso metodológico que se propone se basa en una arquitectura por niveles los cuáles abstraen los principales aspectos de la gestión de calidad de datos, cada nivel cuenta con fases y estos a su vez cuentan con actividades lo que permite ir teniendo una visión clara y ordenada de lo que se va a realizar.

Para una mejor comprensión toda la estructura se muestra en la siguiente tabla:

Tabla IV.1 Estructura del proceso metodológico

NIVEL	FASES	ETAPAS	ACTIVIDADES	BASADA EN:
Nivel 1. Planificación	**Fase I**. Estudio y preparación	Etapa 1.1 Planificar	▪ Estructura del proyecto	-
		Etapa 1.2 Identificar el negocio	▪ Información general acerca del negocio ▪ Análisis del entorno y contexto del negocio ▪ Diagrama de flujo de trabajo ▪ Personal Involucrado ▪ Tecnología involucrada ▪ Problemas Iniciales detectados ▪ Impacto en el negocio ▪ Necesidades del negocio ▪ Priorización de las necesidades del negocio	-
Nivel 2. Ejecución y Análisis	Fase II.Análisis de la información	Etapa 2.1 Plan de captura de datos	▪ Captura de datos	-
		Etapa 2.2 Datos Disponibles	▪ Ciclo de vida de la información aplicado en el negocio ▪ Diagrama de flujo de datos	ADASTRA
		Etapa 2.3 Especificación de Datos	▪ Ámbito de Especificaciones de Datos ▪ Nivel de madurez de la Calidad de datos	INFORMATICA
	Fase III.Evaluación y Análisis Inicial de los datos	Etapa 3.1 Establecimiento de Calidad	▪ Requerimientos de la calidad de datos ▪ Total de datos para analizar	INFORMATICA
		Etapa 3.2 Medición de datos	▪ Perfilado de datos	POWERDATA INFORMATICA ADASTRA DMAIC
		Etapa 3.3 Análisis de la calidad de datos inicial	▪ Evaluación Inicial de las Fuentes de datos ▪ Dimensiones de calidad de datos afectaos ▪ Resultados de la	ADASTRA

			evaluación inicial	
		Etapa 3.4 Políticas internas de calidad	▪ Políticas	ORLI
	Fase IV.Limpieza de datos	Etapa 4.1 Limpieza	▪ Ejecución de limpieza	POWERDATA INFORMATICA ADASTRA DATACTICS ORLI DMAIC
	Fase Evaluación y Análisis Final de la Calidad de datos	Etapa 5.1 Evaluación Final de los datos	▪ Evaluación Final de los datos ▪ Área del negocio en que se realizo la limpieza de datos	-
		Etapa 5.2 Análisis de la evaluación final	▪ Resultado Final de la Calidad de datos	-
			▪ Resultado Final	
			▪ Resultados Finales	
Nivel 3.Observacion	Fase VI. Mejoramiento y Prevención	Etapa 6.1Analizar causas de origen	▪ Causas de Origen ▪ Personal tecnología involucrada	-
		Etapa 6.2 Diseñar Plan de mejoramiento y prevención	▪ Plan de mejoramiento y prevención	POWERDATA ADASTRA ORLI DMAIC
	Fase VII.Seguimiento y Control	Etapa 7.1 Diseñar plan de seguimiento y control	▪ Plan de seguimiento y control	POWERDATA ADASTRA ORLI DMAIC

4.6 CICLO DE VIDA DE LA METODOLOGÍA

El ciclo de vida de la metodología propuesta esta compuesta de siete fases cada una las cuales se compone de objetivos, entradas para la realización de la fases, un propósito establecido, los resultados que tendrá cada fase y las herramientas tanto técnicas como software que se utilizará .La siguiente gráfica indica las fases que se proponen para llegar gestionar la calidad de datos del negocio.

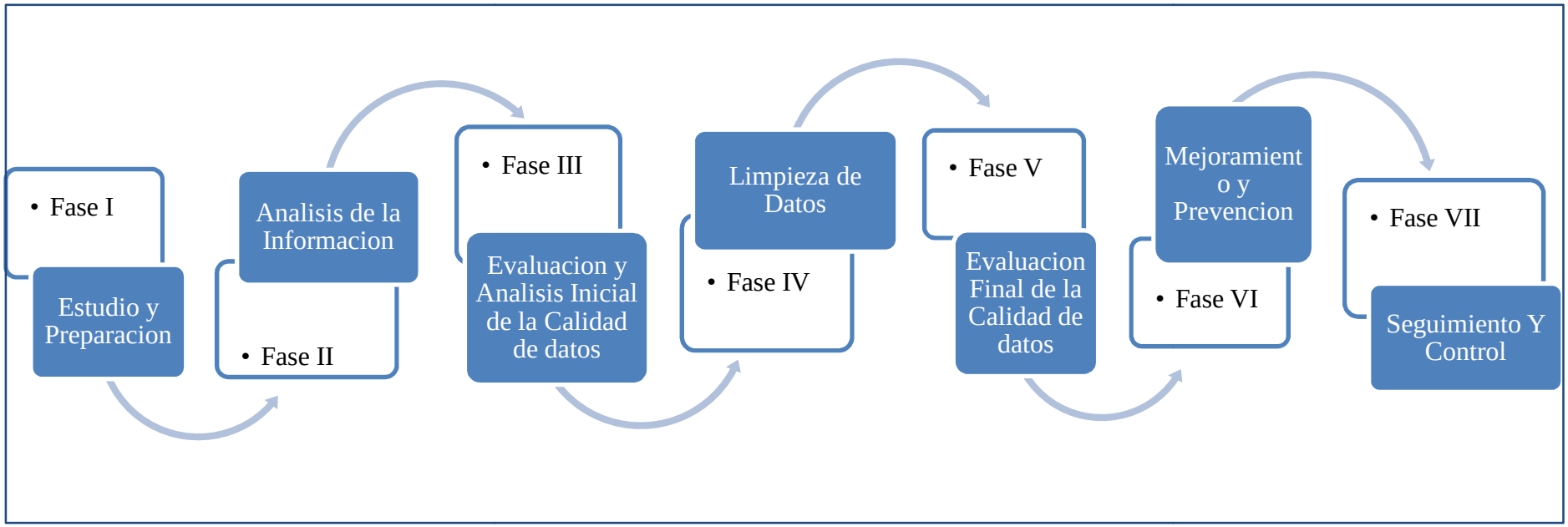


Figura IV.2 Ciclo de Vida de la Metodología Propuesta

4.7 TÉCNICAS Y HERRAMIENTAS SOFTWARE SUGERIDAS

Para llevar a cabo una metodología de gestión de la calidad de datos en proyectos de integración en las mejores condiciones posibles, es necesario contar con el apoyo de algunas técnicas que ayuden a su desarrollo.

Algunas de estas técnicas sirven para detectar problemas con la participación del personal, mientras que otras parten de mediciones o datos obtenidos del proceso a controlar y, a partir del análisis de estos datos, se obtienen los resultados buscados.

4.7.1 Técnicas para calidad de datos

Las técnicas que se utiliza para la calidad de datos son variadas entre ellas tenemos:

- **Técnicas de impacto en el negocio**

  En la tabla IV.2 se muestra las diferentes técnicas de impacto que se puede tener en el negocio como resultado de una mala gestión de calidad de datos:

Tabla IV.2 Técnicas de Impacto en el negocio

Técnicas de impacto	Definición
Anécdotas	Ponen en orden ejemplos o historias sobre el impacto de la mala calidad.
Tratamiento y uso.	Documenta cómo la información es actualmente utilizada y cuáles son los planes futuros.
Matriz de beneficio vs. Costo.	Analiza y evalúa la relación entre beneficios y gastos de los problemas, recomendaciones o mejoras.
Clasificación y prioridad.	Clasifica el impacto de los datos faltantes e incorrectos.

Cinco"porqués" del Impacto en el negocio	Pregunta "¿Por qué?" cinco veces para llegar al impacto real en el negocio.
Proceso del impacto	Ilustralosefectosdel dela mala calidad de datosen el proceso de negocio Almostrarel impactodela mala calidad datossobrelos procesos, la empresapuedetomar decisiones sobre lamejora delos problemasqueantes eranpoco claras
Costo de la baja calidad de datos	Cuantificalos costosy el impacto delos ingresosdelos datosde malacalidad.
Análisis costo-beneficio	Compara los beneficiospotencialesde invertirencalidad de datos conprevisión de costes, a través deunaevaluaciónenprofundidad. Estoincluye retorno dela inversión(ROI) sin fines de lucrode unainversióncomoporcentaje de la cantidadinvertida. Esta técnicayretorno de la inversiónsonlos enfoquesestándarde gestiónde toma dedecisiones financieras.Estainformacióndetalladapuede sernecesario antes deconsiderarode realizar cualquierinversiónfinancierasignificativa-ylas inversionesenla calidad dela informacióna menudoconsiderable.

4.7.2 Técnicas y herramientas para la resolución y análisis de problemas

A continuación se presenta las diferentes técnicas y herramientas que son utilizadas para la resolución y análisis de problemas:

Tabla IV.3 Técnicas de resolución de problemas

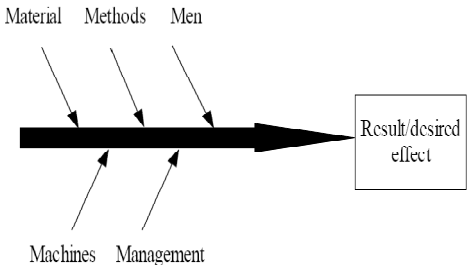
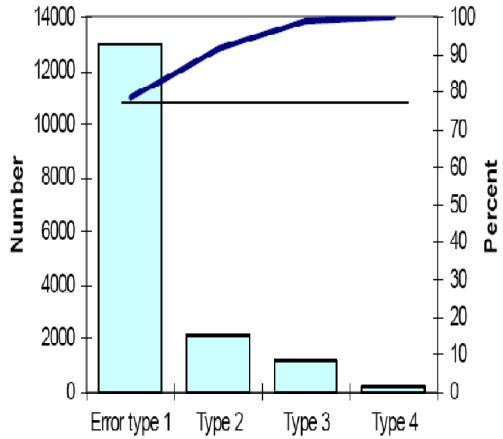
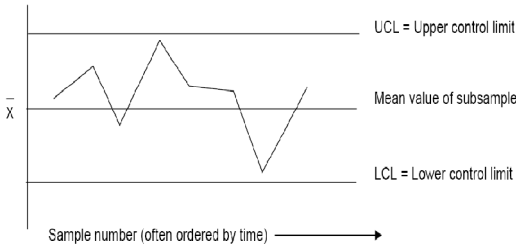
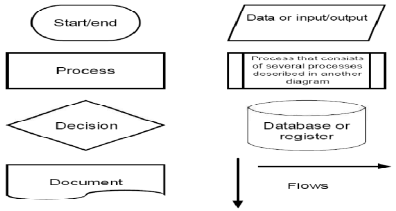
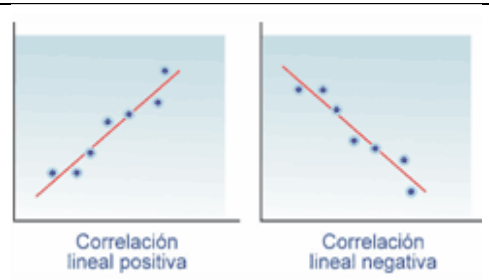
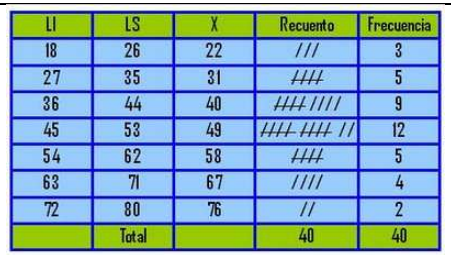
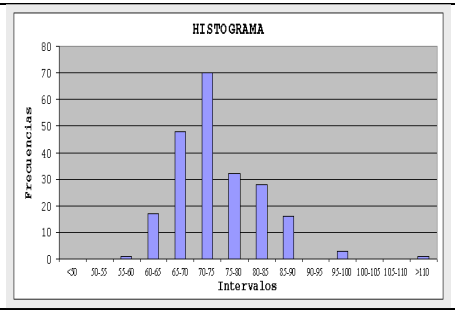
	Técnica		Gráfica	Descripción
	Diagrama Causa Efecto	Llamado también diagrama de Ishikawa, o "espina de pescado"		Utilizado durante lluvia de ideas para la estructuración de una tarea con los objetivos o los resultados que dependen de diferentes factores. Es una herramienta conveniente para la identificación de variables clave del proceso. A menudo, las variables están relacionadas con lo que indican las cinco "M": Materiales, Métodos, Hombres (= personas), máquinas y Manejo.
	Diagrama de Pareto	El diagrama de Pareto es una gráfica que muestra las distintas fuentes "o grupos" que contribuyen a un efecto total o error.		Las fuentes se agrupan según su importancia. A menudo se dice que el 20 por ciento de la causas contribuyen al 80 por ciento de los errores (o los efectos). Esto, por supuesto, varían, pero el mensaje refleja un punto importante, para distinguir los lo vital de las pequeñas contribuciones. Hay muchos ejemplos de uso del diagrama de Pareto en el trabajo de control de calidad, como en edición de las estadísticas. Aquí a menudo el número de errores se registra y se agrupan según sus causas.

	Diagrama de Control			Un gráfico de control es una herramienta para establecer un proceso estable que varía dentro de los límites conocidos, y para el seguimiento del proceso para detectar y eventualmente evitar imprevistos. También se puede utilizar para comparar las capacidades (nivel y la variación de las variables clave del proceso) de un nuevo proceso con el antiguo después de los cambios que se han hecho. El gráfico de control muestra el desarrollo de calidad de datos, incluso si el trabajo con mejoras de calidad está más allá del alcance de las evaluaciones de calidad de datos.
	Diagrama de Flujo	Un diagrama de flujo es una herramienta para los procesos de mapa y documento, que muestra las dependencias entre los procesos y las responsabilidades respectivas.		Puede ser utilizado para la asignación y análisis de los procesos con posibles cuellos de botella, las responsabilidades poco claras o detectar despidos en el flujo de trabajo. Proporciona una base para entender los procesos y para la identificación de variables clave del proceso. No existe un estándar internacionalmente acordado para diagramas de flujo del proceso. Sin embargo, hay un conjunto estándar de símbolos que se ejecuta a través distintos software
	Diagrama de Correlación	El diagrama de dispersión o de correlación nos permite estudiar si existe una relación entre dos variables		Este diagrama puede resultar de gran utilidad para la solución de problemas en un proceso, ya que nos permite comprobar qué causas (factores) están influyendo o perturbando la dispersión de una característica de calidad o variable del proceso a controlar.

	Hoja de Registro	También llamada “hoja de registro”, consiste en un documento donde se pueda recoger de forma fácil y estructurada todo tipo de datos para su posterior análisis. En función de los datos a recoger, se diseña la hoja y se apuntan los datos indicando la frecuencia de observación	 LI | LS | X | Recuento | Frecuencia 18 | 26 | 22 | /// | 3 27 | 35 | 31 | //// | 5 36 | 44 | 40 | //// ///// | 9 45 | 53 | 49 | //// //// // | 12 54 | 62 | 58 | //// | 5 63 | 71 | 67 | //// | 4 72 | 80 | 76 | // | 2 Total | 40 | 40	Las hojas de recogida de datos pueden servir para recoger datos de: •Localización de defectos •Causas de los defectos. •Clasificación de datos defectuosos. •Variación de las características de los datos Y permite observar: •Número de veces en el que sucede algo. •Tiempo necesario para que algo suceda. •Costo de una determinada actividad, a lo largo de un cierto periodo de tiempo. •Impacto de una actividad a lo largo de un período de tiempo.
		Se usa para recentar rápidamente la frecuencia con que algo sucede, conjuntando y presentando los datos de acuerdo a su ocurrencia, con lo cual se puede apreciar el conjunto y su variabilidad.	 HISTOGRAMA Frecuencias Intervalos	Se utiliza para mostrar la tendencia de los datos medidos de un factor relevante

4.7.3 Herramientas Software

Si bien una metodología de gestión de calidad de datos es una apropiada administración de personas y procesos, las herramientas tecnológicas tienen un papel importante. Muchas empresas realizan tareas de limpieza de datos con herramientas caseras, programas en SQL o herramientas limitadas incluidas en productos de ETL. El mercado de herramientas de calidad de datos es aun pequeño, pero se encuentra en expansión. Aproximadamente un tercio de las empresas tienen actualmente herramientas específicas de calidad de datos.

La funcionalidad esperable de las herramientas de calidad de datos consiste de:

- Profiling de datos
- Parsing de datos
- Estandarización o normalización
- Verificación
- Matching
- Consolidación

Las herramientas software con las que se trabajara son:

Tabla IV.4 Herramientas Software para Calidad de Datos

Herramienta	Tipo	Versión	Proveedor
Sql Server 2008 –Data Profile Task	Propietario	2008	Microsoft
Sql Power DQGuru	Open Source	0.9.7	Sql Power
Data Cleaner	Open source	2.0	Human Inference
BayCastle DataSlave	Propietario	2.2.2.91	BayCastle

4.7.3.1 SQL Server 2008 Data Profile Task

El Data Profiling Task nos permite realizar los siguientes tipos de análisis:

- **Solicitud de perfil de claves candidatas**
  Para determinar si un campo puede servir como identificador mostrando el porcentaje de valores únicos.
- **Solicitud de perfil de distribución de longitud de columnas**

  El mínimo y máximo tamaño de variables de tipo cadena. Muestra también la distribución de los valores según el tamaño de los mismos

- **Solicitud de perfil de columnas nulas**

  Muestra el porcentaje de valores nulos de un campo

- **Solicitud de perfil de distribución de valores de columnas**

  Para identificar la distribución de los datos, número de valores únicos, cantidad y porcentaje de los valores.

- **Solicitud de perfil de estadísticas de columnas**

  Perfil estadístico entre las columnas solicitadas

- **Solicitud de perfil de inclusión de valores**

  Para verificar si los valores del campo en una primera tabla (hija), están contenidos en una segunda tabla (padre). Esto también nos sirve para validar la calidad de los datos como por ejemplo la existencia de cada producto vendido en la tabla de [Ventas] dentro de nuestra tabla de [Productos].

- **Solicitud de perfil de patrón de columnas**

  Para identificar la distribución de nuestro datos según patrones en su contenido

- **Solicitud de proporción de columnas nulas**
  Para identificar valores nulos

- **Solicitud de perfil de dependencia funcional**

  Para identificar relaciones de dependencia entre distintos campos y detectar posibles violaciones. Por ejemplo la relación entre los campos País y Capital: Ecuador Quito,España Madrid, Puerto Rico San Juan, si se detecta un campo que no siga la dependencia, por ejemplo Ecuador Riobamba, esto se mostrará como una violación y disminuirá el índice de dependencia entre estos 2 campos.

4.7.3.2 Sql Power DQGuru

Es una herramienta DataCleansing que la empresa**SQLPower** liberó convirtiendo la licencia en OpenSource.

Características

- Intuitiva interfaz grafica que permite una rápida adopción y uso por analistas de datos.
- El proceso de interfaz intuitiva permite rápidamente crear y desarrollar flujos de trabajo de conversión de datos.
- Los usuarios pueden definir sus propias coincidencias de criterios de datos
- Verificación de duplicados a través de la interfaz
- Fusionar duplicados y relacionar datos.
- Puede ser usado para limpiezas de datos iniciales o periódicas
- Herramienta de administración de datos maestros
- Ejecuta con las bases de datos
  - ✓ Oracle
  - ✓ Postgres
  - ✓ MySql
  - ✓ Sybase
  - ✓ Sql Server
- Basada en Java-Plataforma independiente

4.7.3.3 Baycastle DataSlave

Es un galardonado producto de windows diseñado para validación, de-duplicación y transformación de datos. Mueve rápidamente los datos dentro y fuera de aplicaciones empresariales.

Características

- Migra datos de un sistema a otro
- Importa datos
- Valida los datos
- Valida y corrige los datos clave. Incluye herramientas de transformación integral de datos.
- De-duplicación de datos

4.7.3.4 Data Cleaner

Es una herramienta open source que está orientada a preparar los datos para cualquier proyecto en el que se deban aplicar técnicas de Calidad de datos. Es también multiplataforma dado que está desarrollada en Java. Incluye múltiples funcionalidades:

- Profiler: para determinar la calidad de los datos.
- Validator: para validar datos contra reglas que deben verificarse bajos la política de calidad establecida.
- Comparator: para comparar la información de diferentes fuentes de origen.
- Monitor: para establecer un seguimiento de la calidad de los datos.
- Dictionary: permite crear un repositorio de datos maestros y correctos contra los que validar nuestros datos.

4.8 METODOLOGÍA PROPUESTA

Para tener una idea clara de lo que se quiere obtener se propone revisar en cada etapa de las diferentes fases el siguiente esquema:

Tabla IV.5 Esquema de explicación de cada fase

Objetivo	¿Que se trata de lograr? Objetivos o los resultados previstos
Propósito	**¿Porque se debería hacer?** Porque la actividad es importante?
Entradas	**¿Que se necesita para realizar esta fase?** Información necesaria paraejecutar la fase, aportaciones deotras fases
Técnicas y Herramientas	**¿Qué técnicasayudarána completarel proceso?** Técnicas, herramientas yprácticaspara apoyarofacilitar elproceso.
Herramientas Software	**¿Qué software ayudara en esta fase?** Herramientas software para calidad de datos que ayudaran en la fase
Salidas	**¿Que se producecomo resultadodeestepaso?** Los resultadosdela fase

4.8.1 FASE 1. ESTUDIO Y PREPARACIÓN

4.8.1.1 Introducción

Esta fase es de vital importancia ya que es donde las metas del negocioylas estrategiasdeben unir todas las acciones ydecisiones. La informaciónrelacionada con una gestiónsiempredebecomenzarconla pregunta: "¿Por qué es importante parala empresa/organización?".

Todo lohecho conla informacióndebe apoyar a laempresa con el cumplimiento desus objetivos, y estepaso aseguraque se estátrabajando ensituaciones de importanciaparael negocio.

Figura IV.3 Etapas de la Fase 1

4.8.1.2 Etapa 1.1 Planificar

Iniciarelproyectopara hacer frentealosproblemaselegidos ya sea como un equipo o individual (dependiendo del tamaño del negocio) esta etapaesfundamental ya que más de unproyecto fracasa debido alaincomprensiónentrelos implicados(gerencia, miembros del equipo, empresas, TI, etc.).

Para evitar tener problemas a lo largo del proyecto se debe evitar la falta de claridad en cuanto a lo que se llevara a cabo. Una planificación efectiva es esencial para el éxito y ejecución de cualquier proyecto de gestión.

Tomar suficiente tiempo de planificación asegura estar observando los problemas o las oportunidades en los que vale la pena invertir una buena gestión.

a. Objetivo

- Definir las actividades que se desarrollara en el proyecto de gestión de calidad de datos

b. Propósito

- Asegurar queelproyecto tenga un lineamiento de trabajo especifico de gestión lo que permitirá optimizar tiempo y trabajo, un valor importante parael negocio.

c. Técnicas y/o Herramientas

- Anécdotas
- Matriz costo beneficio
- Tratamiento y uso

d. Herramientas Software

- Microsoft Office Project

e. Salidas

- Plan deproyecto, incluida laestructuración, cronología, estimaciones de recursos.

f. Actividades

- Crear la estructura del proyecto

4.8.1.3 Etapa 1.2 Identificar el Negocio

Esta etapa se centra en la comprensión de los objetivos del negocio y los requisitos desde una perspectiva del mismo, a continuación, se convierte este conocimiento en una definición de una solución y un plan preliminar para lograr los objetivos de calidad de datos del negocio.

En esta etapa se describe todo lo que se conoce sobre el negocio su misión, visión objetivos etc.

a. Objetivo

- Tener un buen conocimiento de las actividades que realiza el negocio
- Priorizarycompletarlos asuntosqueson objetodel negocio.

b. Propósito

- Comprender completamente todo el funcionamiento de la empresa y su perspectiva de desarrollo en el futuro para aportar con el cumplimiento del mismo.

c. Entradas

- Problemas yoportunidadesde negociodondela calidadde datoses un componente vital.
- Conocimiento o sospecha deproblemasde calidad de datos.

d. Técnicas y/o Herramientas

- Entrevistaseinvestigación.
- Organigramas

e. Salidas

- Acuerdoclaroydocumentación sobre el estudio del negocio
- Una descripción de los problemas de calidad dedatos detallados, además de las personas y tecnología involucrada.

f. Actividades

- Definir información general acerca del negocio.

- Analizar el entorno y contexto del negocio
- Construir un diagrama de flujo de trabajo
- Definir personal involucrado
- Definir tecnología involucrada
- Problemas iniciales detectados
- Definir impacto en el negocio
- Definir las necesidades del negocio
- Priorizar las necesidades del negocio

4.8.1.4 Modelos de apoyo sugeridos

Análisis del entorno			
Fortalezas	**Oportunidades**	**Debilidades**	**Amenazas**

Modelo IV.1 Análisis del entorno

Cargo	**Tareas que realiza**

Modelo IV.2 Estudio del personal involucrado

Tecnologia			
Herramienta	**Descripción**	**Estado**	**Función**

Modelo IV.3 Estudio de la tecnología Involucrada

Base de problemas						
No.	**Problema**	**Datos**	**Procesos**	**Personal/ Área**	**Herramientas/ Tecnología**	**Comentarios**
1						
2						
3						

Modelo IV.4 Captura y categorización de problemas

Nombre del Proyecto [Ingresar el nombre del proyecto]
Fecha Realizado por:

Recursos del Proyecto [Incluir información pertinente, como nombre, cargo, departamento o equipo].	
Promotor ejecutivo: Promotores del proyecto : Parte Interesada : Jefe del Proyecto : Equipo del Proyecto:	
Visión General del Proyecto :	
Resumen y antecedentes del proyecto	▪ Incluye : Antecedentes ▪ Breve descripción de los objetivos del proyecto y el propósito ▪ Activación del problema declaración o descripción de la situación principal al proyecto ▪ Justificación empresarial o la justificación del proyecto Esta información se coloca en un párrafo breve (resumen) que cualquiera pueda leer y comprender fácilmente
Beneficios :	Incluir los beneficios esperados del proyecto
Ámbito del Proyecto :	
Metas y Objetivos : 1. 2. 3.	
Principales entregas : 1. 2. 3.	
El proyecto es :	Incluye (en un alto nivel) Datos : Procesos: Personas/Organizaciones: Tecnología:
El proyecto no es :	Incluye (en un alto nivel) Datos : Procesos: Personas/Organizaciones: Tecnología:
Condiciones del proyecto :	
Factores críticos de éxito :	Aquellas cosas que deben estar en su lugar para que el proyecto tenga éxito
Hipótesis, problemas, dependencias, restricciones	Elementos que pueden impactar el ámbito del proyecto, horarios, o calidad de los entregables
Riesgos	Elementos con la posibilidad de un impacto negativo en el proyecto. Para cada riesgo, la probabilidad de que ocurra y qué medidas se tomarán si

	lo hace.
Métricas/rendimiento	Lo que hará un seguimiento para medir el éxito. Estos pueden ser desarrollados como el proyecto avanza.
Medidas /Objetivos	
Línea de Tiempo	Resumen de la línea de tiempo y los principales hitos
Costos	Costos estimados

Modelo IV.5 Estructura del proyecto

Priorización			
Prioridad No.	**Necesidad Gestionar calidad de datos en :**	**Razón**	**Comentarios**
1			
2			
3			

Modelo IV.6 Priorización de las necesidades del negocio

4.8.2 FASE 2. ANÁLISIS DE LA INFORMACIÓN

4.8.2.1Introducción

Analizar la información del negocio proporciona una base de entendimiento que se utiliza en todo el proyecto asegurando que se están evaluando los datos de relevancia asociados a los problemas de negocios.

Proporciona una comprensión de los requisitos y especificaciones contra el que la calidad de datos se compara.

Permite obtener un contexto para entender los resultados de las evaluaciones de datos y ayuda al análisis de las causas origen. Cuanto más se entienda el contexto y entorno que afectan a los datos, mejor se entiende lo que se va a evaluar de los datos. Además esta fase proporciona una comprensión de los procesos, personas, y tecnología que afecta la calidad de los datos.

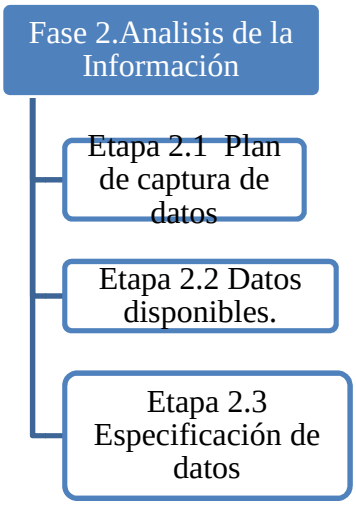


Figura IV.4 Etapas de la Fase 2

4.8.2.2 Etapa 2.1 Plan de captura de datos

La captura de los datos se refiere a cualquier extracción de ellos o acceder a ellos de alguna manera (por ejemplo, a través de una conexión directa a una base de datos). Desarrollar el plan de captura de datos incluye:

- Datos, método de acceso y las herramientas
- El formato de salida (por ejemplo, extraer a un archivo plano, copiar tablas de un servidor de pruebas.)

a. Objetivo

- Establecer la forma de acceso a los datos

b. Propósito

- Proporcionar información acerca de como los datos del negocio serán extraídos mediante herramientas software.

c. Entradas

- Salidas de la Fase 1

d. Técnicas y/o Herramientas

- Histograma

e. Herramientas Software

- DBMS utilizados por el negocio.

f. Salidas

- Plan de captura de datos del negocio

g. Actividades

- Captura de datos

4.8.2.3 Etapa 2.2 Datos disponibles

Esta etapa nos permite conocer el proceso de manipulación de datos que tiene el negocio y las personas que son responsables de los mismos.

a. **Objetivo**

- Reunir, compilar y analizar la información sobre el entorno de la información, apropiado nivel de detalle para los requisitos, datos y especificaciones, procesos, personas /organizaciones, y la tecnología asociada a la cuestión de negocios.

b. **Propósito**

- Proporcionarun estadoinicial delentorno de la información.

- Asegurar que los datos que deben evaluarse son los datos asociados con el problema de negocios.

c. **Entradas**

- Acuerdoclaroydocumentación sobre el estudio inicial del negocio
- Una descripción de losdatos detallado en alto nivel, procesos, personas y tecnología relacionada con las necesidades del negocio.
- Necesidades empresariales, objetivos, estrategias, problemasy oportunidades(Cualquier conocimiento para ayudar a describirelentorno dela informaciónactual,comoorganigramas, la aplicación, arquitectura, los modelos dedatos).

d. **Técnicas y/o Herramientas**

- Diagrama de flujo

e. **Herramientas Software**

- DBMS con las que trabaja el negocio.

f. **Salidas**

- Captura de datos y un plan de los resultados del análisis del entorno de información, como el impacto potencial de calidad de los datos y / o el negocio, posible causas y recomendaciones iniciales en ese momento.

- Especificaciones detalladas de lista de datos, cartografía de datos si la evaluación es de más de una de fuente de datos, las asignaciones iníciales de origen a destino si la migración de datos.
- Los modelos de datos con el detalle necesario para entender la estructura y las relaciones para que los datos puedan ser captados y analizados correctamente.

g. **Actividades**

- Ciclo de vida de la información aplicado al negocio
- Diagrama de flujo de datos

4.8.2.4 Etapa 2.3 Especificación de datos

La especificación de los datos mide la existenciay documentación de estándares de datos, modelos de datos, metadatos, datos de referencias.

a. Objetivo

- Definir especificación de los datos actuales

b. Propósito

- Proporcionar una base para el análisis de la calidad de datos.

Entradas

- Salidas de las etapas de la Etapa 2.2
- Problemas y oportunidades de negocio en que se sospeche la calidad de los datos como un componente.

c. Técnicas y/o Herramientas

- Clasificación y prioridad
- Impacto en el negocio

d. Salidas

- Medidas de los datos actuales en el negocio

e. Actividades

- Definir ámbito de especificaciones de datos
- Definir el nivel de madurez de la calidad de datos del negocio.

4.8.2.5 Modelos de apoyo sugeridos

Datos	Método de Acceso	Herramientas

Modelo IV.7 Plan de captura de datos

Alto nivel	Detalle	Detalle Completo

Modelo IV.8 Nivel de Detalle Flujo de Datos

Alto nivel	Detalle	Detalle Completo

Modelo IV.9 Nivel de detalle personas-organizaciones

Nombre del departamento/negocio	Quien colecciona los datos	Que datos son coleccionados	Quien usa los datos	Donde están los datos almacenados	Quien es propietario de los datos	Como están siendo actualizados los datos	Con que frecuencia se actualizan los dato

Modelo IV.10 Información para el flujo de datos

Especificación	Existe especificación? (Si-No)	Agregar o crear otras evaluaciones de calidad de datos (Si-No)	Evaluar la calidad de especificación? (Si-No)	Notas

Modelo IV.11 Especificación de datos

4.8.3 FASE 3. EVALUACIÓN Y ANÁLISIS INICIAL DE LA CALIDAD DE DATOS

4.8.3.1Introducción

Se han introducido a la calidad de datos dimensiones, aspectos o características de la información y una forma de clasificar la calidad de la información y necesidades de datos.

Las dimensiones se utilizan para definir, medir, y gestionar la calidad de los datos y de la información

El beneficio más gratificante de la evaluación cualitativa de los datos se concreta en la evidencia de los problemas que subyacen en el negocio, problemas identificados en la primera fase. Los resultados de la evaluación también proporcionan información de referencia necesaria para investigar las causas de origen, corregir errores de datos, y evitar los errores de datos en el futuro.

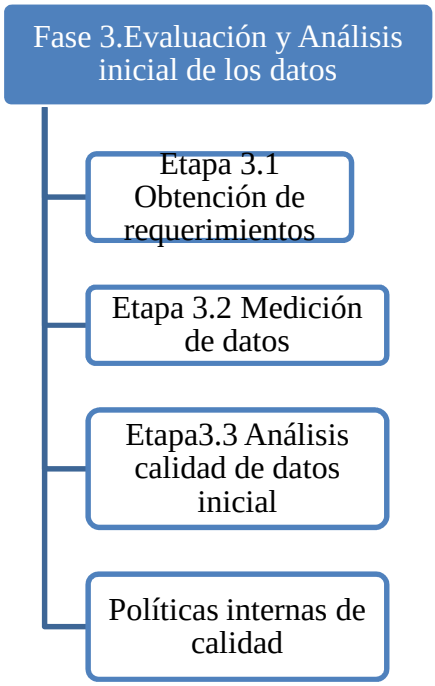


Figura IV.5 Etapas de la Fase 3

4.8.3.2Etapa 3.1 Obtención de Requerimientos

Antes de realizar la medición de datos del negocio es necesario contar con los requerimientos claros y definidos para que en la medición centrarse en los requerimientos planteados.

a. Objetivo

Obtener los diferentes requerimientos de calidad de datos que requiere el negocio

b. Propósito

Contar con requerimientos definidos para proceder a la medición de datos

c. Entradas

Salidas de la Fase 2

d. Salidas

Requerimientos de datos definidos

e. Actividades

Requerimientos de la calidad de datos

Total de datos a analizar

4.8.3.3 Etapa 3.2 Medición de datos

El proceso de perfilado de datos examina los datos existentes en el negocio y recopila información sobre los mismos.

a. Objetivo

- Conseguir métricas (dimensiones) de calidad de datos que incluyen si los datos cumplen los objetivos de la organización.

b. Propósito

- Determinar qué datos pueden ser usados para distintos propósitos.
- Reducir el riesgo de integrar información a nuevas aplicaciones dado que conocemos su estado.
- Entender problemas derivados de los datos en proyectos que hagan uso intensivo de los mismos.
- Tener una visión global de los datos de la organización para desplegar políticas de administración de datos.

c. Entradas

- Salidas de la etapa 3.1 de la presente fase
- Salidas de las etapas de la fase 2

d. Técnicas y/o Herramientas

- Clasificación y prioridad

e. Herramientas Software

- Sql Server 2008 –Data Profile Task
- Data Cleaner

f. Salidas

- Estructura de la calidad actual de los datos

g. Actividades

- Perfilado de datos

4.8.3.4 Etapa 3.3 Análisis de la calidad de datos Inicial

Unaevaluacióniniciales elprimer conjuntode pruebasrealizadasa losquelas evaluacionesposteriores sepueden comparar.

a. Objetivo

- Estableceruna evaluación inicial del negocio en cuanto a los problemas de calidadde datos utilizando medidas tanto cualitativa como cuantitativa.

b. Propósito

- Obtener el apoyo de la parte administrativa del negocio para las inversiones de calidad de los datos
- Determinar las inversiones adecuadas en los recursos de información como las correcciones de datos necesarias.

c. Entradas

- Salidas de la Etapa 3.2

d. Herramientas Software

- Sql Server 2008 –Data Profile Task
- Data Cleaner

e. Salidas

- Resultados de la evaluación inicial

f. Actividades

- Evaluación inicial de las fuentes de datos
- Definir dimensiones de calidad de datos afectados
- Definir los resultados de la evaluación inicial

4.8.3.5 Etapa 3.4 Políticas internas de calidad

a. Objetivo

- Presentar una visión de conjunto de la organización para su adecuada organización.
- Precisar expresiones generales para llevar a cabo acciones que deben realizarse en el negocio.
- Facilitar la descentralización, al suministrar a los niveles intermedios lineamientos claros a ser seguidos en la toma de decisiones.
- Tener una organización en cuanto a los datos del negocio.

b. Propósito

- Proporcionar guías generales para canalizar el pensamiento administrativo en direcciones específicas

c. Entradas

- Lista de políticas

d. Técnicas y/o Herramientas

- Anécdotas

e. Salidas

- Políticas establecidas y aprobadas

f. Actividades

- Discutir la lista de políticas con los responsables de cada función operacional.
- Determinar una lista de las políticas que realmente se requieren definir.
- Precisar los límites a que llegarán las políticas.
- Determinar una prioridad de políticas para ser aplicadas.
- Presentar un borrador de las políticas y discutirlas con los responsables del área correspondiente para su aceptación o modificación respectiva.
- Aprobación de las políticas por la dirección superior.

4.8.3.6 Modelos de apoyo sugeridos

Requerimientos de calidad de datos de Negocio							
No.	Problema	Tabla(s)	Columna(s)	Requerimiento	Acción(es)a tomar	Herramienta(s)a utilizar	Comentarios

Modelo IV.12 Requerimientos de la calidad de datos

No.	Problema	Dimensión de calidad Afectada
1		
2		
3		
4		
5		
6		

Modelo IV.13 Dimensiones de calidad afectados

4.8.4 FASE 4. LIMPIEZA DE DATOS

4.8.4.1 Introducción

Después de tener una visión clara de la calidad actual de los datos podemos empezar ya con la limpieza de datos. La corrección de errores en los datos es un importante avance en la información y en el proceso de mejora de la calidad de datos

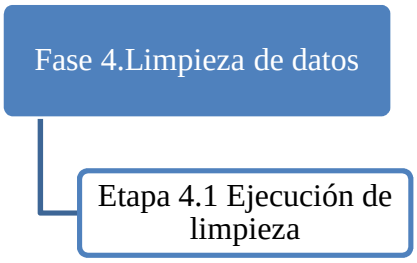


Figura IV.6 Etapas de la Fase 4

4.8.4.2 Etapa 4.1 Ejecutar Limpieza

El proceso de ejecución de limpieza de datos permite obtener finalmente datos útiles y actualizados lo que permitirá entender mejor el entorno de negocio, permitiendo maximizar los beneficios.

a. Objetivo

- Corregir erroresde los datosexistentes

b. Propósito

- Realizar la limpieza de los datosexistentesqueestáncausandoproblemaspara el negocio.

c. Entradas

- Salidas de lafase 3

d. Herramientas Software

- Sql Power DQGuru
- BayCastle DataSlave

e. Salidas

- Datos corregidos de acuerdo a las especificaciones.

f. Actividades

- Identificar los datos para ser corregidos
- Identificación de los registros a ser cambiado y las modificaciones esperadas.
- Establecer como se realizara la limpieza de datos
- Realizar la limpieza de datos teniendo en cuenta los requerimientos y las políticas establecidas anteriormente.

4.8.5 FASE 5. EVALUACIÓN Y ANÁLISIS FINAL DE LA CALIDAD DE DATOS

4.8.5.1 Introducción

La evaluación de la calidad de datos final proporciona proporciona información vital para indicar el estado final de calidad en los datos analizados y corregidos.

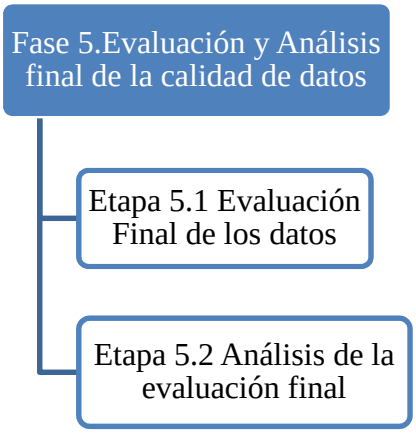


Figura IV.7 Etapas de la fase 5

4.8.5.2 Etapa 5.1 Evaluación Final de los datos

Al haber desarrollado una limpieza de los datos, esta etapa permite realizar un análisis de los mismos, lo que servirá para determinar los beneficios para el negocio y saber si se cumplieron con las necesidades del mismo.

a. Objetivo

- Analizar las correcciones realizadas en el conjunto de datos establecidos de mala calidad.

b. Propósito

- Obtener información detallada de lo que se ha realizado y si las correcciones que se realizaron satisfacen las necesidades del negocio.

c. Entradas

- Salidas de las etapas de la fase 3.
- Salidas de las etapas de la fase 4.

d. Salidas

- Recomendaciones para la acciónbasadaenlos resultados.
- Resultados de los análisis realizados.

e. Actividades

- Definir evaluación final de las fuentes de datos
- Definir áreas del negocio en los que se realizó la limpieza de datos.
- Resultados de la evaluación final

4.8.5.3 Etapa 5.2Análisis de la evaluación final

Esta etapa define el estado final en los cuales se entregan los datos después de todos los procesos realizados determinando con esto como se aporta con el negocio y cúal es el beneficio obtenido.

a. Objetivo

- Estableceruna evaluación final en cuanto a los problemas de calidad

b. Propósito

- Dar una visión clara de lo que se ha logrado

c. Entradas

- Salidas de la Fase 1
- Salidas de la Fase 3
- Salidas de la Fase 4

d. Salidas

- Resultados de la evaluación final

e. Actividades

- Definir el resultado final de la calidad de datos

- Definir resultados finales

4.8.6 FASE 6. MEJORAMIENTO Y PREVENCIÓN

4.8.6.1 Introducción

La correcciónde errores en los datos es un importante avance en el proceso de mejora de la calidad de información y los datos. Sin embargo, para la mejora continua es importante no sólo corregir los datos actuales, sino también prevenir los futuros

Este es un punto crítico en la gestión de calidad de datos en el que la comunicación es clave para asegurar que las recomendaciones finales se apliquen y para prevenir cometer futuros errores.

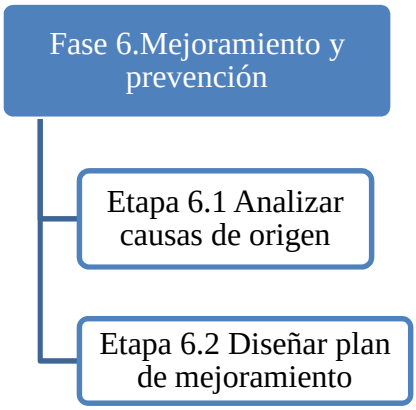


Figura IV.8 Etapas de la Fase 6

4.8.6.2 Etapa 6.1 Analizar causas de origen

Todos los problemas que surgen a partir de calidad de datos requieren diferentes niveles de tiempo, dinero, y recursos humanos. Hay una tendencia asaltar a una solución que parece ser la másconveniente para hacer frente con rapidez a una situación es el análisis de las causas de origen el cual ve en todo lo posible las causas de un problema, asunto o condición para determinarsu causa real. A menudotiempo y esfuerzo se gastan en el tratamiento de los síntomas de un problema sin la determinación de sus causas reales,

yevitar que el problema vuelva a ocurrir, es por esta razón que esta etapa tiene este objetivo.

a. Objetivo

- Encontrar las causas reales de los problemas de calidad de datos en el negocio.

b. Propósito

- Identificar ypriorizarlas verdaderas causasde losproblemas decalidad de datos, desarrollando recomendaciones para abordarlas causas fundamentales.

c. Entradas

- Salidas de la Fase 1
- Salidas de la Fase 2
- Salidas de la Fase 3
- Salidas de la Fase 5

d. Salidas

- Recomendaciones específicas para abordar las causas profundas de los problemas de calidad de datos

e. Actividades

- Definir Causas de origen
- Definir personal y tecnología involucrados con las causas de origen

4.8.6.3 Etapa 6.2Diseñar plan de mejoramiento

Esta etapa se enfoca en poner en práctica soluciones adecuadas que aborden las causas fundamentales de los problemas de calidad de los datos.

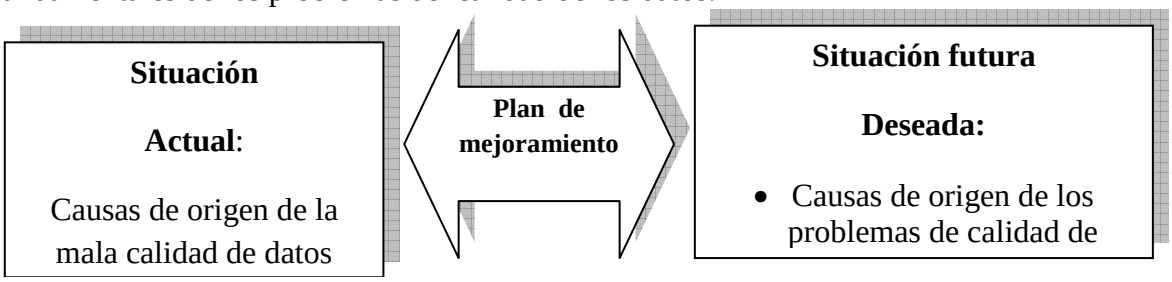


Figura IV.9 Plan de Mejoramiento

a. Objetivo

- Desarrollar un plan de acción basado en el análisis de la calidad de los datos y / o los resultados de la evaluación inicial y final de calidad de datos en el negocio.

b. Propósito

- Garantizar que la evaluación de la calidad de datos, resultados de negocios de impacto y recomendaciones se convierten en planes de mejora.
- Evitar que ocurran futuros errores en los datos tratando las causas de los errores.
- Asegurar que la inversión en la limpieza o la corrección de errores actuales no esta desperdiciado.

c. Entradas

- Salidas de las etapas de la fase 3
- Salidas de las etapas de la fase 5
- Ideas a pequeña escala o adicionales para implementar los cambios

d. Técnicas y/o Herramientas

- Utilizar modelo de plan de mejoramiento y prevención

e. Salidas

- Soluciones para abordar las causas fundamentales y la prevención de errores de datos en el futuro.
- Cambios en el negocio presente como resultado de las mejoras implementadas
- El personal afectado por los cambios de formación y con una comprensión coherente de cambios, las expectativas, las nuevas funciones y responsabilidades, nuevos procesos, etc
- Los cambios y sus resultados documentados para los futuros usuarios

f. Actividades

- Diseñar plan de mejoramiento y prevención

4.8.6.4 Modelos de apoyo sugeridos

Problema	Posibles acciones	Valoración de la viabilidad de cada acción	Importancia de cada acción para superar la debilidad	Responsable de verificación

Modelo IV.14 Plan de mejoramiento y prevención

4.8.7 FASE 7. SEGUIMIENTO Y CONTROL

4.8.7.1 Introducción

Realizar un seguimiento y control es fundamental para la utilización eficaz de los datos. Esto permite comprobar constantemente la calidad de los datos y se identifican problemas nuevos o pérdidas de calidad.

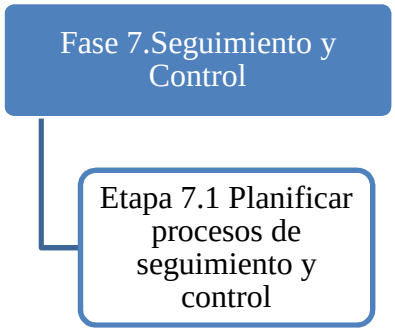


Figura IV.10 Etapas de la Fase 7

4.8.7.2 Etapa 7.1 Diseñar plan de seguimiento y control

a. Objetivo

- Implementar controles constantes de calidad.
- Asegurar que las nuevas soluciones se han adecuado a los controles de calidad de datos
- Vigilar y verificar las mejoras que se llevaron a cabo.

b. Propósito

- Determinar si las acciones de limpiezahan logrado el efecto deseado.

c. Entradas

- Salidas de las etapas de la fase 5.
- Salidas de las etapas de la fase 6
- Dependientesdelos controlesimplementados

d. Técnicas y/o Herramientas

- Utilizar modelo de diseño del plan de seguimiento y control

e. Herramientas Software

- Sql Server 2008 –Data Profile Task
- DataCleaner

f. Salidas

- Plan de seguimiento y controles recomendados.

g. Actividades

- Plan de seguimiento y control

4.8.7.3 Modelos de apoyo sugeridos

Modelo IV.15 Plan de seguimiento y control

Problema	Causas de origen	Proceso de control sugerido	Responsable	Frecuencia

CAPITULO V

APLICACION DE LA METODOLOGIA PROPUESTA EN EL SISTEMA DE INFORMACION INSTITUCIONAL ESPOCH (SII-ESPOCH) PARA LOS SISTEMAS DE LA UNIDAD DE EDUCACIÓN A DISTANCIA Y ACÁDEMICO INSTITUCIONAL

5.1 INTRODUCCIÓN

El presente capítulo tiene como objetivo aplicar la metodología propuesta anteriormente poniendo en uso sus fases, etapas y actividades en los sistemas académico y de la unidad de educación a distancia

Se partirá por el estudio y preparación del proyecto para lo cual se tendrá todos las fuentes de datos preparados para luego pasar al proceso de evaluación de los mismos.

Luego se procederá a la evaluación y análisis de los datos actuales de los sistemas académico y de la unidad de educación a distancia, mediante una reunión con los promotores del proyecto se obtendrá los requerimientos necesarios de los problemas encontrados con los datos para proceder al establecimiento de políticas.

Enseguida se procede a la limpieza de datos y posteriormente a la evaluación y análisis de los resultados para de esta manera establecer planes de mejora y control.

5.2 APLICACIÓN DE LA METODOLOGÍA PROPUESTA

5.2.1 FASE I. ESTUDIO Y PREPARACIÓN

5.2.1.1 Etapa 1.1 Planificar

En la siguiente tabla se muestra la estructura que tomará el proyecto para la gestión de calidad de datos

Tabla IV.1 Estructura del Proyecto

Nombre del Proyecto :DQSII-ESPOCH

Fecha:01 de Enero 2011

Realizado por: Margarita Isabel Solís Velasco

Recursos del Proyecto:

Recurso Humano :			
Necesidad	**Recurso**	**Cantidad**	**Estado**
Jefe del Proyecto	Ing. Ivonne Rodríguez	1	Asignada
Desarrollador	Isabel Solís	1	Asignada
Recursos Primarios:			
Estación de Trabajo para desarrollo	PC 800MHz 2GB RAM	1	Cumple
Herramientas de desarrollo para calidad de datos	Sql Server 2008-Integration Services(propietario) Data Cleaner(Open Source) SQL Power DQGuru(Open source) Baycastle Dataslave MapEditor(Shareware)	4	Cumple

Promotor ejecutivo: Ing. Alejandra Oñate
Promotores del proyecto : Ing. Ivonne Rodríguez
Parte interesada : Unidad Técnica de Planificación de la ESPOCH
Jefe del proyecto : Ing. Ivonne Rodríguez
Equipo del proyecto: Margarita Isabel Solís Velasco

Visión General del Proyecto

Resumen y antecedentes del proyecto	**Antecedentes** La ESPOCH administra gran cantidad de datos generando información importante, por esta razón decide realizar un sistema de información institucional la cual requiere inicialmente una importante gestión de calidad de datos para obtener información de calidad.

<table>
<tr><td></td><td>Objetivo:
Gestionar Calidad de datos para SII-ESPOCH
Propósito:
Obtener datos de calidad en las fuentes para luego ser enviada a un proceso de ETL.
Justificación del proyecto:
Los datos históricos con los que cuenta la institución se encuentran almacenados en varias fuentes de datos, al ser estas diferentes son generadas con una serie de errores lo cual causa datos de mala calidad por lo que es necesario gestionar estos datos para que no cause problemas en el momento de obtener información en el SII-ESPOCH</td></tr>
<tr><td>Beneficios :</td><td>Contar con información de calidad para la correcta toma de decisiones.

Resultados mejores y más fiables</td></tr>
<tr><td colspan="2">Ámbito del Proyecto :
Es un hecho claro que la información es uno de los activos más importantes de un negocio y, cada vez más, acceder a información de calidad de una manera eficaz resulta una necesidad.
El presente proyecto se centra en la gestión de la calidad de datos de la ESPOCH que permita a las personas interesadas que laboran en la institución acceder a información precisa y confiable.</td></tr>
<tr><td colspan="2">Metas y Objetivos :
1.Lograr una correcta gestión de calidad de datos
2.Beneficiar a la institución de contar con información de calidad
3. Brindar calidad de datos que genere información de calidad para la correcta toma de decisiones.</td></tr>
<tr><td colspan="2">Principales entregas :
1. Informe Final del proyecto</td></tr>
<tr><td>El proyecto es :</td><td>Datos :
Trabajar en los datos de las fuentes que requiere la ESPOCH para el proyecto de integración SII ESPOCH
Procesos:
• Gestionar calidad de datos para el proyecto de integración SII ESPOCH
Personas/Organizaciones:
Personal involucrada con la manipulación de datos de la institución.
Tecnología:
Herramientas para la gestión de calidad de datos.</td></tr>
<tr><td>El proyecto no es :</td><td>Datos :
Corregir diseño de base de datos
Procesos:
Diseños nuevos en las bases de datos
Personas/Organizaciones:</td></tr>
</table>

<table>
<tr><td></td><td>No involucra personal de sistemas de la ESPOCH</td></tr>
<tr><td>Condiciones del proyecto :</td><td>-Plazo máximo de ejecución del proyecto 6 meses
-Información constante de la evolución del proyecto</td></tr>
<tr><td>Factores críticos de éxito :</td><td>Contar con disponibilidad de recursos en el momento adecuado.</td></tr>
<tr><td>Hipótesis, problemas, dependencias, restricciones</td><td>-No contar con los recursos necesarios
-Falta de atención en el proyecto por parte de los interesados
-Disponibilidad de tiempo</td></tr>
<tr><td>Riesgos</td><td>
<table>
<tr><th>Riesgo</th><th>Probabilidad de ocurrencia</th><th>Acción a Tomar</th></tr>
<tr><td>Los recursos no están disponibles en su momento</td><td>Baja</td><td>Solicitar a los promotores del proyecto los recursos necesarios.</td></tr>
<tr><td>No se cumple con el cronograma de trabajo especificado</td><td>Media</td><td>Control de las actividades a desarrollar</td></tr>
<tr><td>Problemas de entrega de las fuentes de datos</td><td>Media</td><td>Solicitar a los promotores del proyecto las fuentes de datos al inicio del proyecto</td></tr>
<tr><td>Retraso en la entrega del proyecto</td><td>Media</td><td>Control continuo del cronograma de trabajo</td></tr>
</table>
</td></tr>
<tr><td>Métricas/rendimiento</td><td rowspan="2">El control de cumplimiento de objetivos se lo realizara mediante el cronograma de trabajo.</td></tr>
<tr><td>Medidas /Objetivos</td></tr>
<tr><td>Línea de Tiempo</td><td>Ver cronograma de trabajo</td></tr>
<tr><td>Costos</td><td>Debido a que se trata de un proyecto de tesis de grado el proyecto no tendrá costo para la institución.</td></tr>
</table>

- **Equipo de trabajo**

Tabla V.2 Equipo de trabajo

Miembro	Nombre	Responsabilidad
Jefe del Proyecto	Ing. Ivonne Rodríguez	▪ Planificación y control del proyecto. ▪ Supervisar el avance del proyecto
Desarrollador	Isabel Solís	▪ Desarrollo total de la gestión de calidad de datos para el proyecto de Integración SII-ESPOCH siguiendo los lineamientos establecidos.

- **Diagrama de contexto**

A continuación se muestra el diagrama de contexto de los datos del negocio:

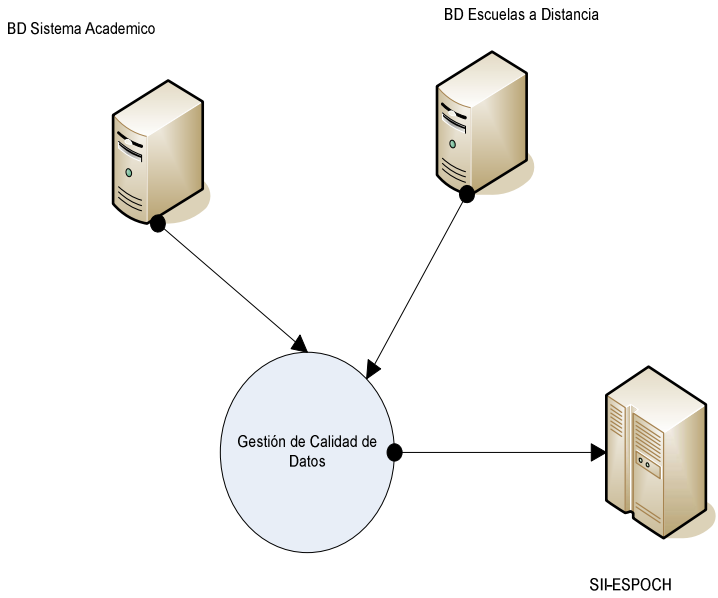


Figura V.1 Diagrama de contexto

- **Plan y cronograma de trabajo**

A continuación se indica el cronograma de trabajo fijado para el desarrollo del proyecto aplicando el tiempo para cada fase:

	Nombre de tarea	Duración	Comienzo	Fin
4	Estudio y Preparacion	21 días	lun 03/01/11	lun 31/01/11
5	Analisis de la Informacion	21 días	mar 01/02/11	mar 01/03/11
6	Evaluacion y Analisis Inicial de la calidad de datos	23 días	mar 01/03/11	jue 31/03/11
7	Limpieza de datos	23 días	vie 01/04/11	mar 03/05/11
8	Evaluacion y analisis final de la calidad de datos	15 días	mié 04/05/11	mar 24/05/11
9	Mejoramiento y Prevencion	8 días	mar 24/05/11	jue 02/06/11
10	Seguimiento y Control	8 días	vie 03/06/11	mar 14/06/11

Figura V.2 Cronograma de trabajo

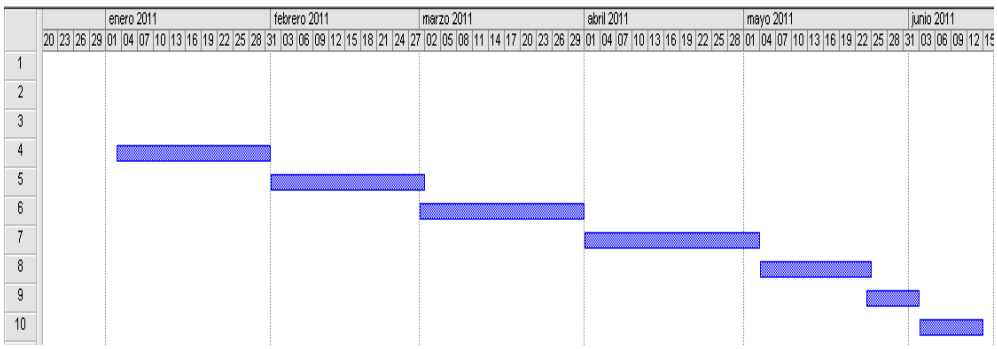


Figura V.3 Diagrama Gantt

5.2.1.2 Etapa 1.2 Identificar el Negocio

Entre las principales tareas que se debe realizar es el estudio del negocio para saber de esta manera en que rumbo se encuentra y de que manera se puede aportar a cumplir con los objetivos deseados.

- **Información General acerca del Negocio**

a. Nombre del Negocio.

Escuela Superior Politécnica de Chimborazo

b. Misión.

"Ser una institución universitaria líder en la Educación Superior y en el soporte científico y tecnológico para el desarrollo socioeconómico y cultural de la provincia de Chimborazo y del país, con calidad, pertinencia y reconocimiento social".

c. Visión.

"Formar profesionales competitivos, emprendedores, concientes de su identidad nacional, justicia social, democracia y preservación del ambiente sano, a través de la generación, transmisión, adaptación y aplicación del conocimiento científico y tecnológico para contribuir al desarrollo sustentable de nuestro país".

d. Objetivos.

Lograr una administración moderna y eficiente en el ámbito académico, administrativo y de desarrollo institucional.

Establecer en la ESPOCH una organización sistémica, flexible, adaptativa y dinámica para responder con oportunidad y eficiencia a las expectativas de nuestra sociedad.

Desarrollar una cultura organizacional integradora y solidaria para facilitar el desarrollo individual y colectivo de los politécnicos.

Fortalecer el modelo educativo mediante la consolidación de las unidades académicas, procurando una mejor articulación entre las funciones universitarias.

Dinamizar la administración institucional mediante la desconcentración de funciones y responsabilidades, procurando la optimización de los recursos en el marco de la Ley y del Estatuto Politécnico.

Impulsar la investigación básica y aplicada, vinculándola con las otras funciones universitarias y con los sectores productivos y sociales.

Promover la generación de bienes y prestación de servicios basados en el potencial científico-tecnológico de la ESPOCH.

e. Actualidad

La Escuela Politécnica se halla actualmente en unos de sus más altos estándares de calidad de educación superior a nivel nacional.

Fue catalogada por el Conea como una de las once universidades calificadas como clase A, que la define con la excelencia en educación superior. Mientras que el Conesup en su estudio lo ubico como la tercera universidad del país con una calificación de sobresaliente.

La ESPOCH se ha convertido en una universidad pionera en la educación a nivel nacional y con un alto auge de demanda de bachilleres por continuar sus estudios en dicha institución cada año. De ahí que la mayoría de los estudiantes que se encuentran en las diversas carreras provienen de lugares ajenos a la ciudad de Riobamba en donde se encuentra ubicada, de hecho mas de las tres quintas partes son de estudiantes de otras provincias y extranjeros, principalmente estos últimos de Colombia y Perú.

Sus actividades se resaltan a nivel externo tanto nacional como internacionalmente, debido a convenios, concursos y demás que han ayudado a su alto reconocimiento educativo y académico.

f. Proyectos de integración de datos

La institución maneja una cantidad muy considerable de información pero el no poder acceder a ella adecuadamente es un problema eminente que dificulta la adecuada toma de decisiones.

Por esta razón se realizo una propuesta presentada por el DESITEL que se denominó “Integración de los sistemas Financiero, RRHH, SAI de la ESPOCH mediante un data Warehouse para obtener información ágil y veraz que permitirá una acertada toma de decisiones y una mejora en la gestión de procesos internos de la ESPOCH”.

La UTP retoma esta propuesta y la amplia para lograr una solución integral frente a la notoria necesidad de contar con un Sistema de Información Estadística, que permita gestionar información institucional de relevancia, de forma ágil, confiable, oportuna, que sirva de soporte para una adecuada toma de decisiones dentro de la ESPOCH.

El Objetivo es Implementar un Sistema de Información Institucional ("SII-ESPOCH"), que permita gestionar información de relevancia, de forma ágil, confiable, precisa, oportuna, que sirva de soporte para la toma de decisiones dentro de la institución y lograr una administración moderna y eficiente en el ámbito académico y administrativo. Para realizar esto se requiere contar con una gestión de calidad de datos para el proyecto de integración que se quiere realizar de esta manera se asegurara que la información sea confiable y precisa.

- **Análisis del entorno y contexto del negocio**

A continuación se indica como se encuentra el entorno y el contexto de la institución en la actualidad:

Tabla V.3 Análisis del entorno y contexto del Negocio

Análisis del Entorno			
Fortalezas	**Oportunidades**	**Debilidades**	**Amenazas**
Importante Infraestructura Tecnológica	Aumento de demanda de Universidades y Politécnicas calificadas	Falta de Control Interno	Incursión de universidades extranjeras en el mercado nacional
Recursos Humanos Capacitados	Cambio en el perfil demográfico		
Misión, Visión, Objetivos, Metas bien definidos	Consideración de la Educación Superior como un factor determinante		

- **Diagrama de Flujo de Trabajo**

Para tener una idea clara como los datos se conducen en las labores diarias de la institución a continuación se presente el siguiente diagrama:

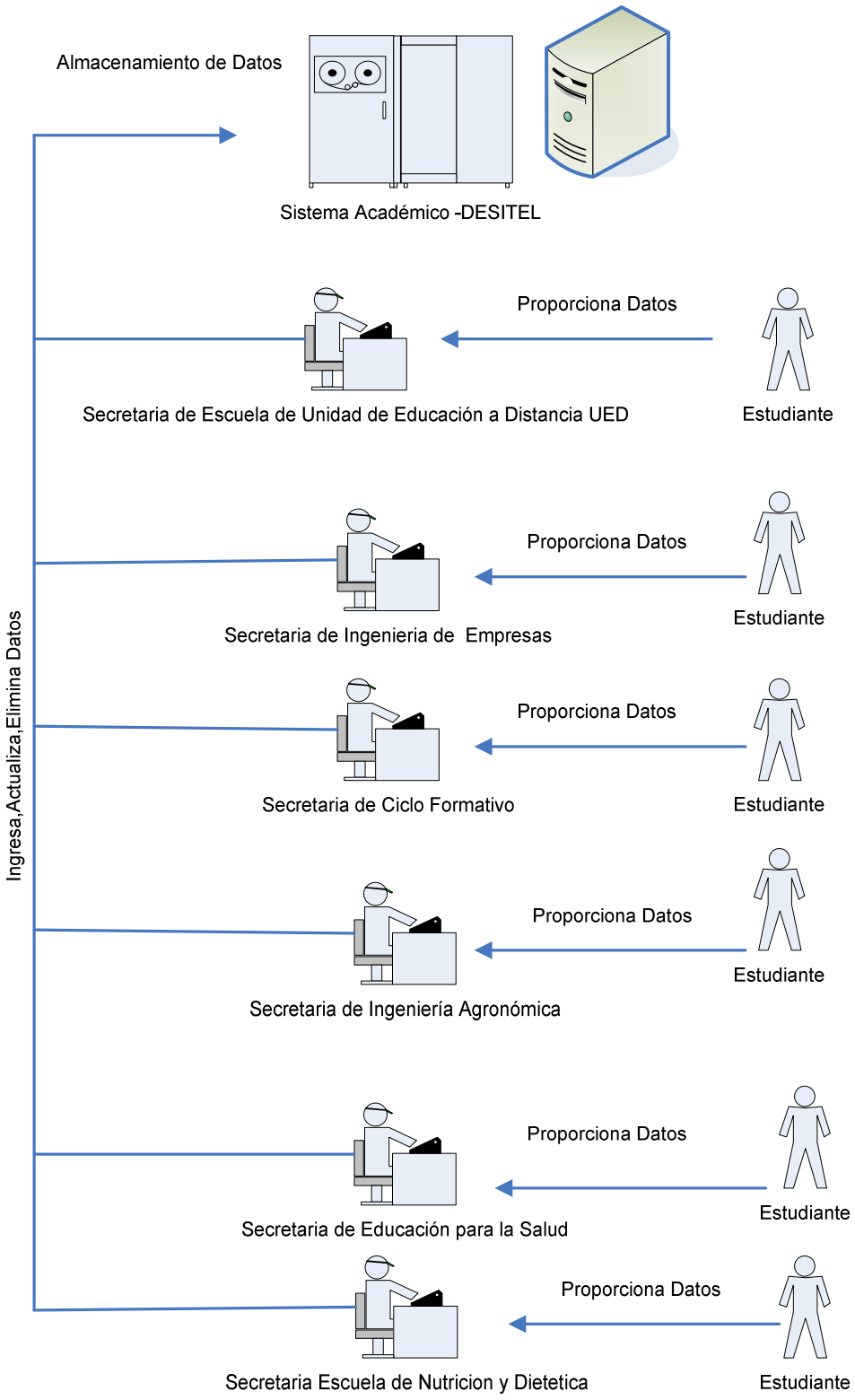


Figura V.4 Flujo de trabajo

- **Personal involucrado**

El personal que se involucra directamente con la manipulación de los datos se muestra en la siguiente tabla:

Tabla V.4 Personal Involucrado

Cargo	Tareas que realiza
Secretaria Escuela de Unidad de Educación a Distancia –UED Dirección de Escuela	Ingreso, Actualización y Modificación de Datos
Secretaria Escuela Ingeniería en Empresas Dirección de Escuela	Ingreso, Actualización y Modificación de Datos
Secretaria Escuela de Educación para la Salud Dirección de Escuela	Ingreso, Actualización y Modificación de Datos
Secretaria de Escuela Nutrición y Dietética Dirección de Escuela	Ingreso, Actualización y Modificación de Datos
Secretaria Escuela Ingeniería Agronómica Dirección de Escuela	Ingreso, Actualización y Modificación de Datos
Secretaria Escuela Ciclo Formativo Dirección de Escuela	Ingreso, Actualización y Modificación de Datos

- **Tecnología Involucrada**

La tecnología que utiliza la institución para la administración y almacenamiento de datos es:

Tabla V.5 Tecnología involucrada

Tecnología			
Herramienta	**Descripción**	**Estado**	**Función**
Sql Server 2008	Almacenamiento de la Información de la Institución	Buena Condición	Servidor de BD

- **Problemas Iniciales Detectados**

Algunos problemas con los datos que el personal ha tenido problemas se detalla a continuación:

Tabla V.6 Captura y Categorización de problemas

Base de problemas						
No.	Problema	Datos	Procesos	Personal/ Área	Herramientas/ Tecnología	Comentarios
1	Inconsistencias	Fechas	Reportes	Secretaria	Sql Server 2008	
2	Nulos	Nombres y Apellidos	Reportes			
3	Sin formato	Cedulas	Reportes			

- **Impacto en el negocio**

A continuación se presenta un gran impacto que ha tenido la mala calidad de datos en la institución principalmente en época de matriculación donde se requiere que la información sea rápida y de calidad, para esto se lo ha realizado mediante la técnica de anécdotas:

Tabla V.7 Impacto en el negocio

Dato: Cedulas de estudiantes
Proceso: Matriculas
Escenario: La secretaria no tiene datos validos para poder buscar información del estudiante
Impacto: Mala atención Mala información en los reportes Perdida de tiempo
Tecnología: Sistema Académico

- **Necesidades del negocio:**

Se requiere gestionar calidad de datos para el proyecto de Integración SII-ESPOCH en las siguientes escuelas:

- Escuela de Ingeniería Agronómica.
- Escuela de Ciclo Formativo
- Escuela de Ingeniería en Administración de Empresas
- Escuela de Educación para la salud
- Escuela de Nutrición
- Escuelas de Educación a distancia
- Priorización de las necesidades del Negocio

A continuación se presenta las necesidades del negocio en cuanto a calidad de datos priorizándolas cuales han tenido mayores problemas la institución.

Tabla V.8 Priorización de la Necesidades del Negocio

Priorización			
Prioridad No.	**Necesidad Gestionar calidad de datos en :**	**Razón**	**Comentarios**
1	Escuelas de Educación a distancia	Mayores problemas de calidad de datos	Se analizaran las escuelas deeducación adistanciaúnicamente de la matriz Riobamba
2	Escuela de Ingeniería en Administración de Empresas	Mayor cantidad de estudiantes	
3	Escuela de Ciclo Formativo	Mayor cantidad de estudiantes	
4	Escuela de Ingeniería Agronómica	Cantidad considerable de estudiantes	
5	Escuela de Educación para la salud	Cantidad considerable de estudiantes	
6	Escuela de Nutrición	Cantidad considerable de estudiantes	

5.2.2 FASE II. ANÁLISIS DE LA INFORMACIÓN

5.2.2.1Etapa 2.1 Plan de captura de datos

- **Captura de datos**

La captura de datos se procedió a realizar de la siguiente manera:

Tabla V.9 Plan de captura de datos

Datos	Método de Acceso	Herramientas
BD-Escuelas de Educación a	Importación de datos	Sql Server 2008

distancia		
Escuela de Ingeniería en Administración de Empresas	Restauración de backup de BD	Sql Server 2008
Escuela de Ciclo Formativo	Restauración de backup de BD	Sql Server 2008
Escuela de Ingeniería Agronómica	Restauración de backup de BD	Sql Server 2008
Escuela de promoción para la salud	Restauración de backup de BD	Sql Server 2008
Escuela de Nutrición	Restauración de backup de BD	Sql Server 2008

5.2.2.2 Etapa 2.2 Datos Disponibles

- **Ciclo de vida de la información aplicado en el negocio**

En la institución el ciclo de vida de la información esta de la siguiente manera:

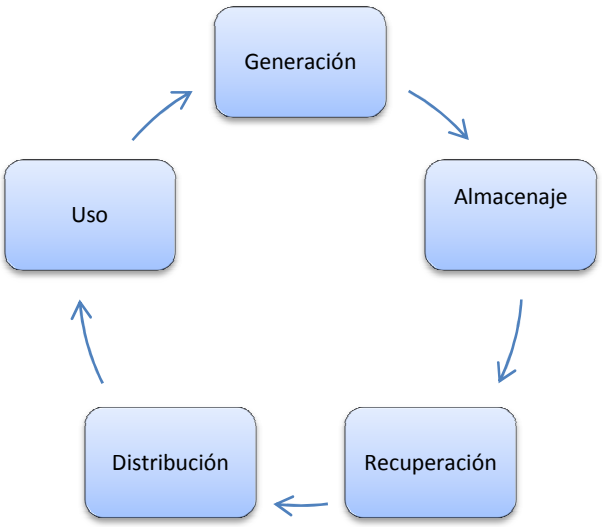


Figura V.5 Ciclo de vida la información

- **Diagrama de flujo de datos**

En la figura V.6 se muestra como fluyen los datos a través de las diferentes entidades de la institución mediante un DFD de nivel 1:

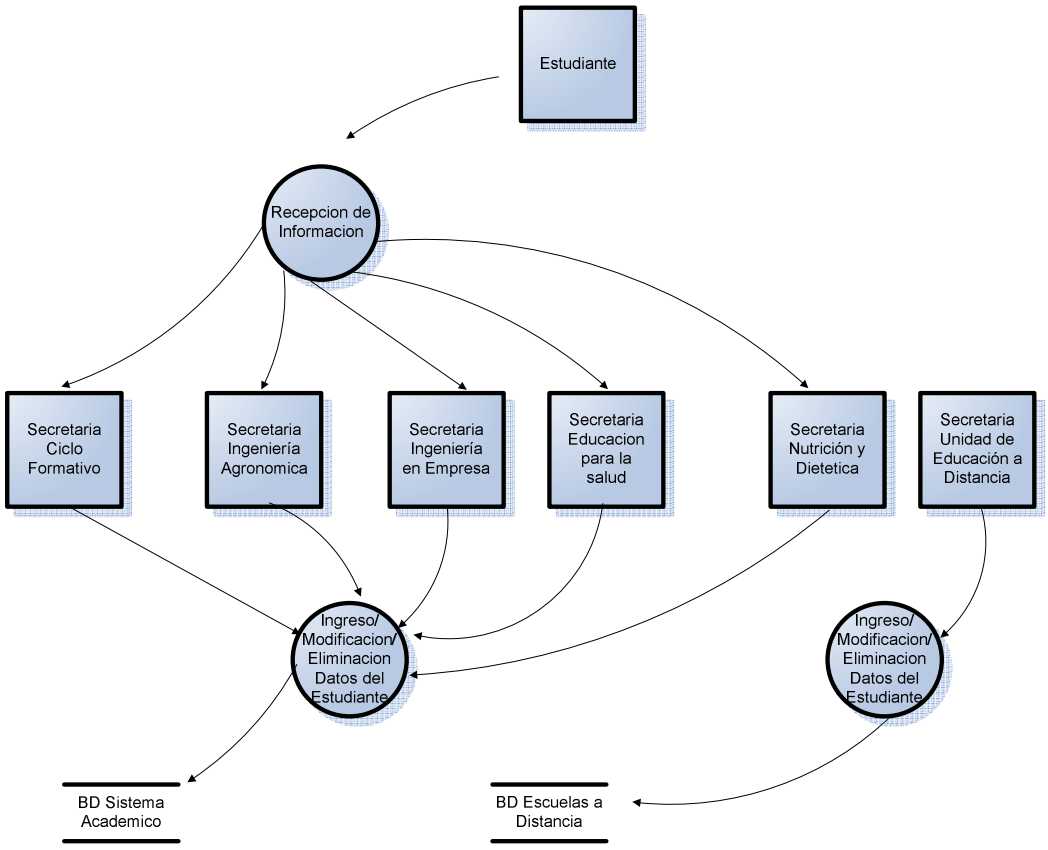


Figura V.6 Diagrama de flujo de datos

En las tablas que se muestran a continuación se indica el nivel de detalle del diagrama de flujo de datos expuesto anteriormente:

Tabla V.10 Nivel de detalle para el proceso de flujo de datos

Alto nivel	Detalle	Detalle Completo
Datos	Ingreso,Modificacion,Eliminacion	Utilización de los datos de estudiantes

Tabla V.11 Nivel de detalle para personal

Alto nivel	Detalle	Detalle Completo
Secretaria	Manipulación de datos	Utilización de los datos para procesos que realiza la institución que puede inscripción o matriculación

Tabla V.12 Información para el flujo de datos

Nombre del departamento/negocio	Quien colecciona los datos	Que datos son coleccionados	Quien usa los datos	Donde están los datos almacenados	Quien es propietario de los datos	Como están siendo actualizados los datos	Con que frecuencia se actualizan los dato
ESPOCH	Administradores de bases de datos	Estudiantes Docentes Periodos Notas Semestres Cursos Inscritos Estados Evaluaciones Exámenes Horarios Pensums Permisos Requisitos Materias Proyectos	Secretarías de la Institución	Servidores	ESPOCH	Procesos de actualización	Semestral

5.2.2.3 Etapa 2.3 Especificación de datos

- **Ámbito de Especificaciones de datos**

En la siguiente tabla se indica las especificaciones de datos con los que cuenta la institución actualmente:

Tabla V.13 Ámbito de especificaciones de datos

Especificación	Existe especificación? (Si-No)	Agregar o crear otras evaluaciones de calidad de datos (Si-No)	Evaluar la calidad de especificación? (Si-No)	Notas
Estándares de datos	No	Si	Si	
Modelos de datos	Si	Si	Si	
Reglas de negocio	No	Si	Si	
Metadatos	Si	Si	Si	
Referencia de datos	No	Si	Si	

- **Nivel de madurez de la calidad de datos**

La institución en cuanto a calidad de datos se encuentra en el nivel 2, ya que no existe un proyecto anterior de mejorar de calidad de datos lo cual causa problemas en los procesos administrativos .Pero para el personal que labora en la institución es de vital importancia contar de inmediato con una gestión de calidad de datos eficiente.

5.2.3 FASE III. EVALUACIÓN Y ANÁLISIS INICIAL DE LOS DATOS

5.2.3.1Etapa 3.1Obtención de requerimientos

Para obtener los requerimientos de la calidad de datos del negocio se procedió a solicitar un taller de trabajo con todo el personal involucrado en el proyecto

- **Requerimientos de la calidad de datos**

Los requerimientos que se solicito en el taller de trabajo con la Ing. Alejandra Oñate promotora ejecutiva del proyecto, Ing. Ivonne Rodríguez jefe del proyecto se resume en la siguiente tabla:

Tabla V.14 Requerimientos de Calidad de datos

No.	Problema	Tabla(s)	Columna(s)	Requerimiento	Acción(es)a tomar	Herramienta(s)a utilizar	Comentarios
1	Datos NULL y blancos	CESTUD	CEDIDE FECNAC	Medir cantidad de valores nulos y blancos, en el caso de existir nulos y blancos sustituir por un valores referencial	-Perfilamiento de datos -Sustitución: CEDIDE=000000000-0 FECNAC=1900-01-01	-Data Cleaner -BayCastle Map editor -Sql Server 2008(Data Profiling Task)	
2	Datos incompletos	CESTUD	CEDIDE	Medir el tamaño del campo en el caso de existir cedulas con longitudes menores o iguales a 9 sustituir por un valor referencial	-Perfilamiento de datos -Valor referencial en caso de cedula incompleta: 000000000-0		
3	Datos duplicados	CESTUD	CEDIDE	Medir cantidad de datos duplicados	-Perfilamiento de datos		
4	Datos sin formatos necesarios	CESTUD	CEDIDE	Medir cuales datos se encuentran con guión y sin guión, y definir un formato único	-Perfilamiento de datos Formato definido: Cedulas con guion		
5	Datos sin un estándar especifico	CESTUD	NOMEST APEEST	Medir que datos contienen mayúsculas y minúsculas ,establecer un estándar único	-Perfilamiento de datos -Estándar definido: Todos con mayúsculas		
			SEXO	Estándar : MAS y FEM	-Perfilamiento de datos -Estandarizar valor actual según corresponda a MAS y FEM		
			ECUEST	Estándar: ECUATORIANO	-Perfilamiento de datos -Estandarizar valor actual según corresponda a ECUATORIANO		
6	Inconsistencias en los datos	CESTUD	FECING	Medir la cantidad de inconsistencias ,en caso de inconsistencias definir un valor referencial de inconsistencia	-Perfilamiento de datos -Valor referencial de inconsistencia: 1900-01-01		

- **Total de Datos para Analizar**

Total de Datos de Escuelas de Educación a Distancia:

Tabla V.15 Total Datos UED

Base de Datos	**CESTUD$**
FADE_FASE_1IC	33
FADE_FASE_2IC	24
FADE_FASE_6	358
FADE_FASE_7	447
FADE_FASE_8	46
FADE_FASE_9	49
FADE_FASE_10	51
FADE_GGSBA	37
FADE_GGSES	22
Total de Datos a Analizar	1067

Total de Datos del Sistema Académico:

Tabla V.16 Total Datos Sistema Académico

Base de Datos	**Estudiantes**
OAS_CicloFormativo_db	3219
OAS_IngAgronomica	703
OAS_IngEmpresas_db	1402
OAS_NatPromSalud_db	454
OAS_Nutricion_db	758
Total de Datos a Analizar	6536

5.2.3.2 Etapa 3.2 Medición de datos

Perfilado de datos

Para realizar el perfilado de los datos se utilizo las siguientes herramientas software:

- Microsoft Sql Server 2008 Integration Services(SSIS)-Data Profile Task
- Data Cleaner2.1.1(herramienta Open Source)

- **Data Cleaner**

Instalación y Configuración:

Requerimientos Software

DataCleaner requiere Java Runtime Environment (JRE) versión 5.0 o posterior.

Drivers para Bases de Datos

Debido a que DataCleaner fue desarrollado utilizando la Plataforma Java, se necesitará los controladores de base de datos basada en Java, llamados JDBC.En Data Cleaner vienen instalados los drivers JDBC para los principales tipos de base de datos.

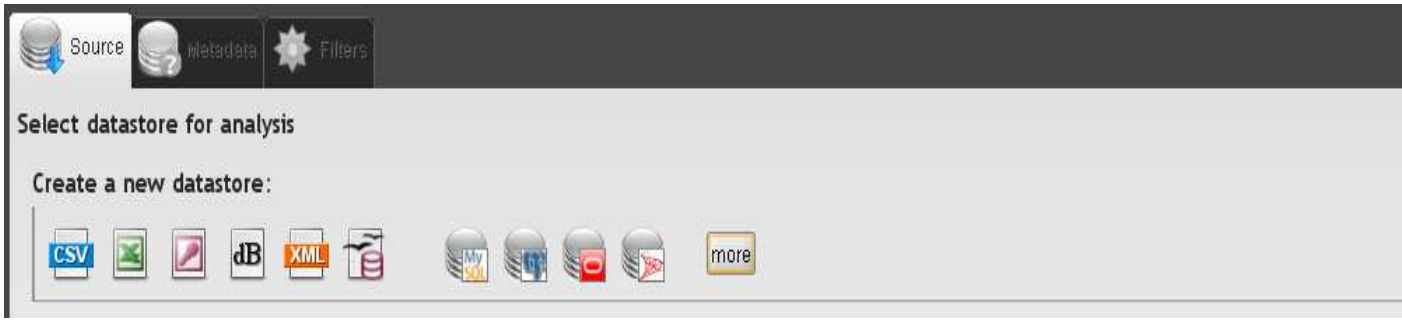


Figura V.7 Bases de Datos disponibles en Data Cleaner

En el caso de no estar disponible en la lista la base de datos que necesitamos procedemos a hacer clik en **more** para escoger la base de datos que necesitamos y agregarle el driver correspondiente.

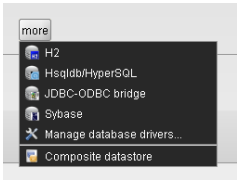


Figura V.8 Administración de drivers

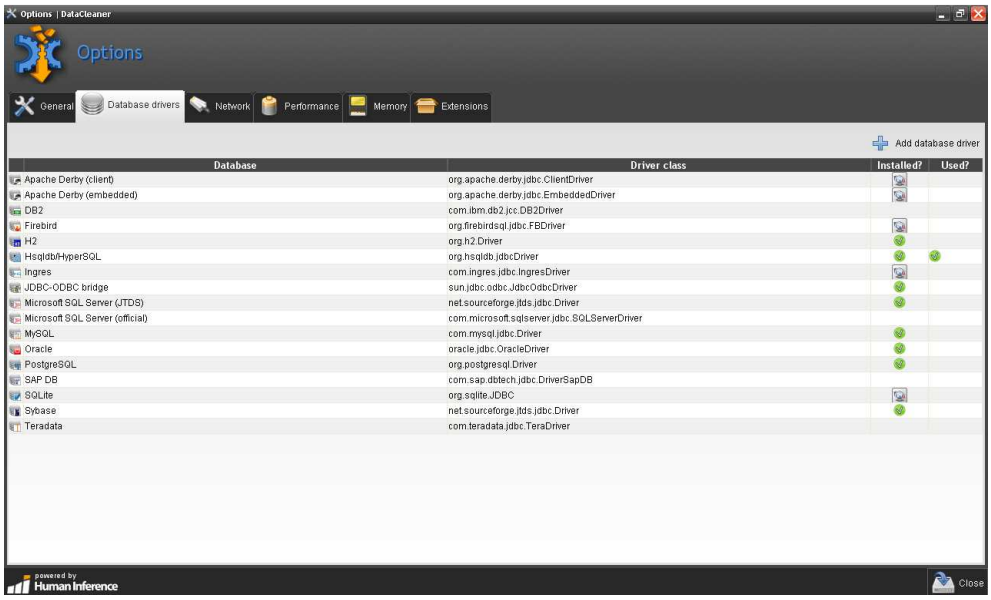


Figura V.9 Drivers de Bases de datos

Podemos agregar el driver de la siguiente manera:

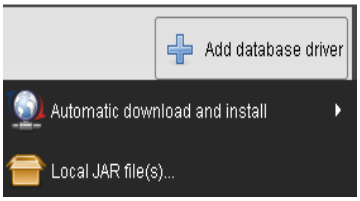


Figura V.10 Descargas de Drivers

Descarga e Instalación Automática: Si el tipo de base de datos de tipo esta representado en el submenú de esta opción, entonces DataCleaner automáticamente descargara el controlador y lo instalara.

Archivo JAR Local: Cuando el archivo ha sido descargado por uno mismo simplemente se ubica el driver en nuestro equipo y se instala.

Configuración:

Para empezar se necesita hacer una conexión a la base de datos con la que vamos a trabajar

En el caso de las bases de datos de las Escuelas a distancia tienen la extensión .dbf por lo que se utiliza la base de datos DBase:

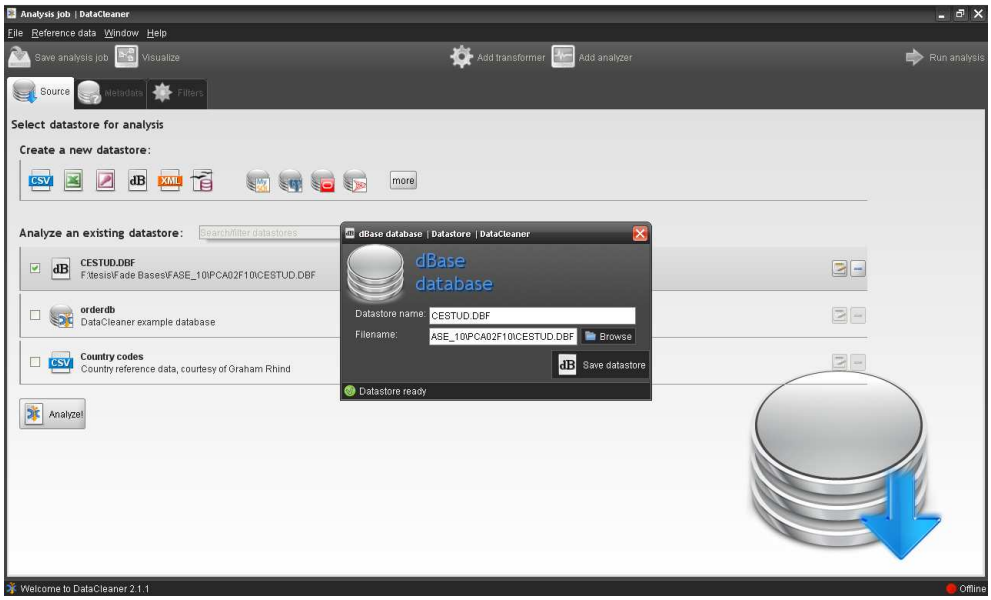


Figura V.11 Conexión a la base de datos

Para las bases de datos del Sistema Académico se utiliza la Base de datos Sql Server 2008:

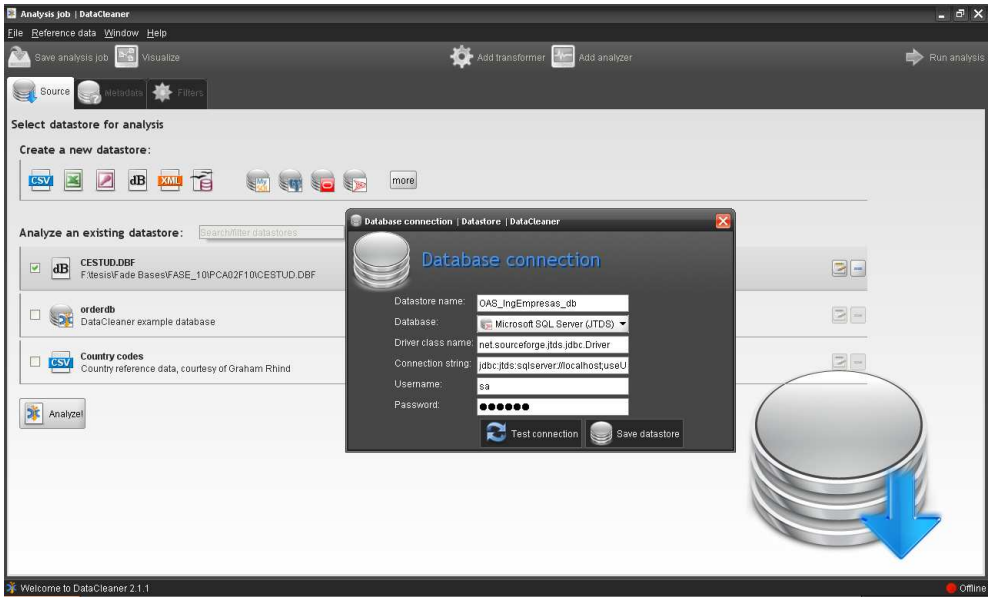


Figura V.12 Cadena de Conexión

Ejecución:

Para realizar el análisis de los datos se accede a **Analizer** y procedemos a escoger las columnas que analizaremos

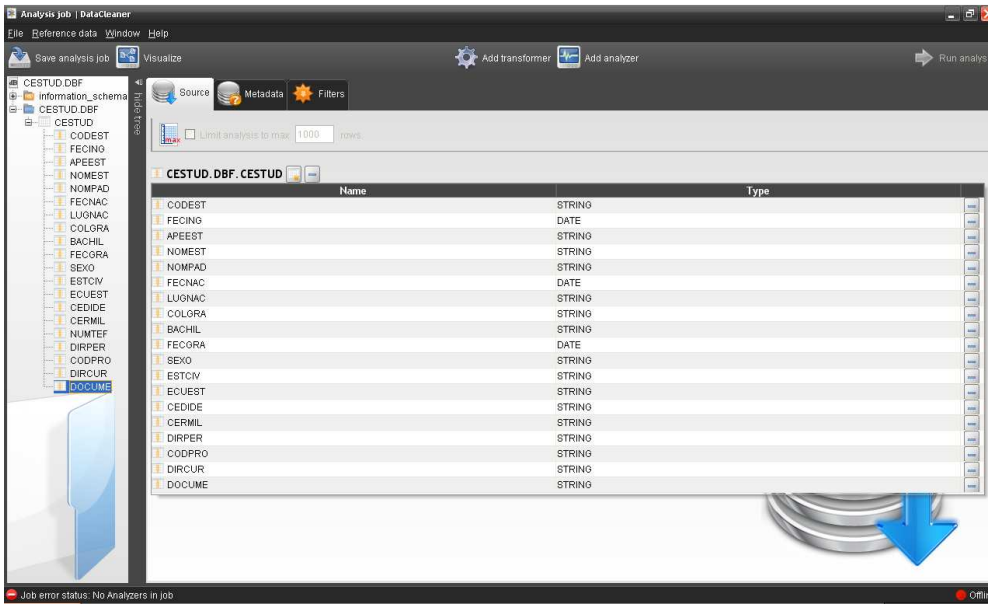


Figura V.13 Columnas a Analizar

Luego se procede al análisis de la metadata de la tabla:

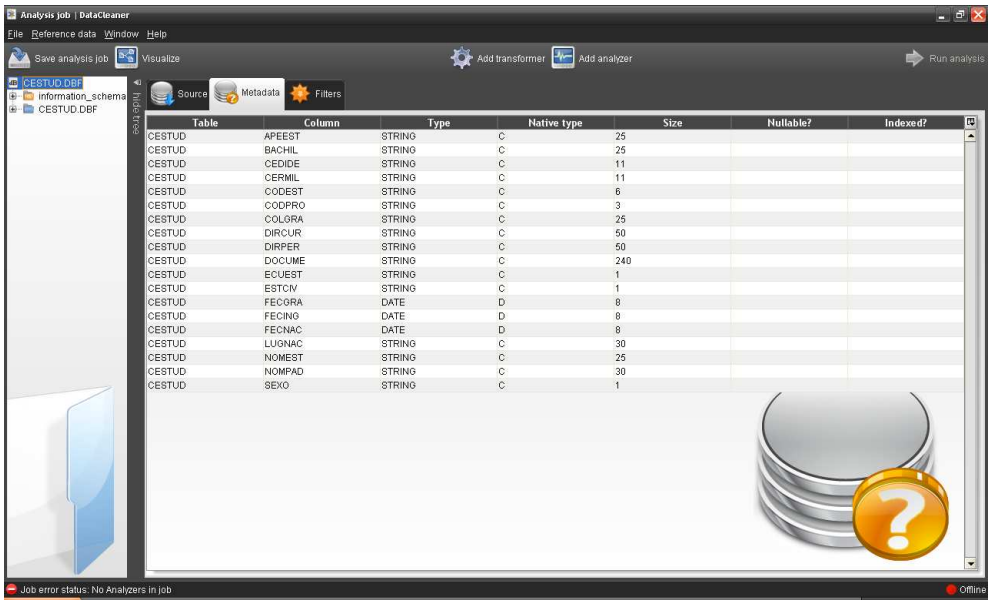


Figura V.14 Metadata de la Tabla

Luego se procede a un filtrado de datos pero para el proceso de perfilamiento de datos no lo se lo hará ya que se necesita saber los errores que se encuentran en los datos

Procedemos a realizar el análisis de los datos string y date que son los únicos tipos de datos que tenemos según la metadata:

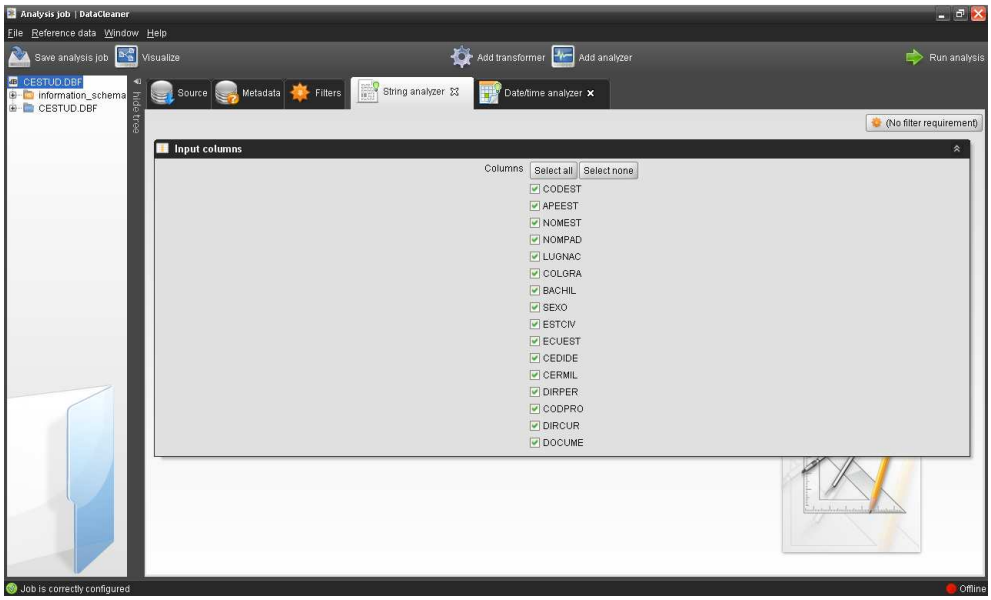


Figura V.15 Análisis de string

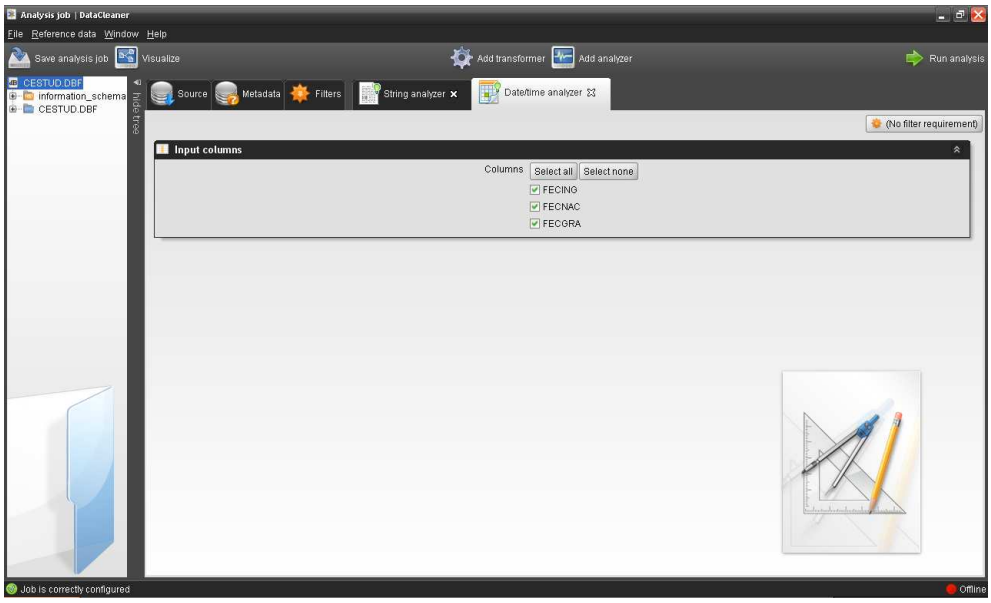


Figura V.16 Análisis de tiempo

Luego procedemos a la ejecución del análisis para eso hacemos clik en Run Analysisy nos mostrara la siguiente pantalla:

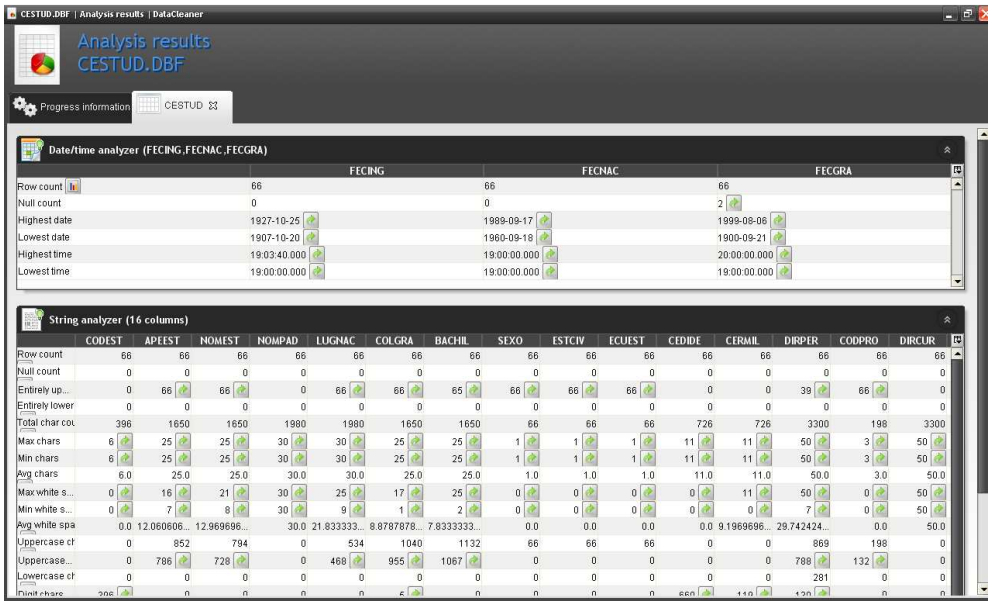


Figura V.17 Resultados del Perfilado

- **Microsoft Sql Server 2008 Integration Services(SSIS)-Data Profile Task**

Parala etapa de perfilados de los datos también se utilizo el componente data profile task que se encuentra dentro de Integration services:

SQL Server proporciona un único programa de instalación para instalar alguno de sus componentes o todos, incluido Integration Services. Mediante el programa de instalación puede instalar Integration Services con o sin otros componentes de SQL Server en un único equipo.

Instalación y Configuración

Al momento de instalar Sql Server 2008 tenemos que instalar todos los servicios para Business Intelligence development Studio

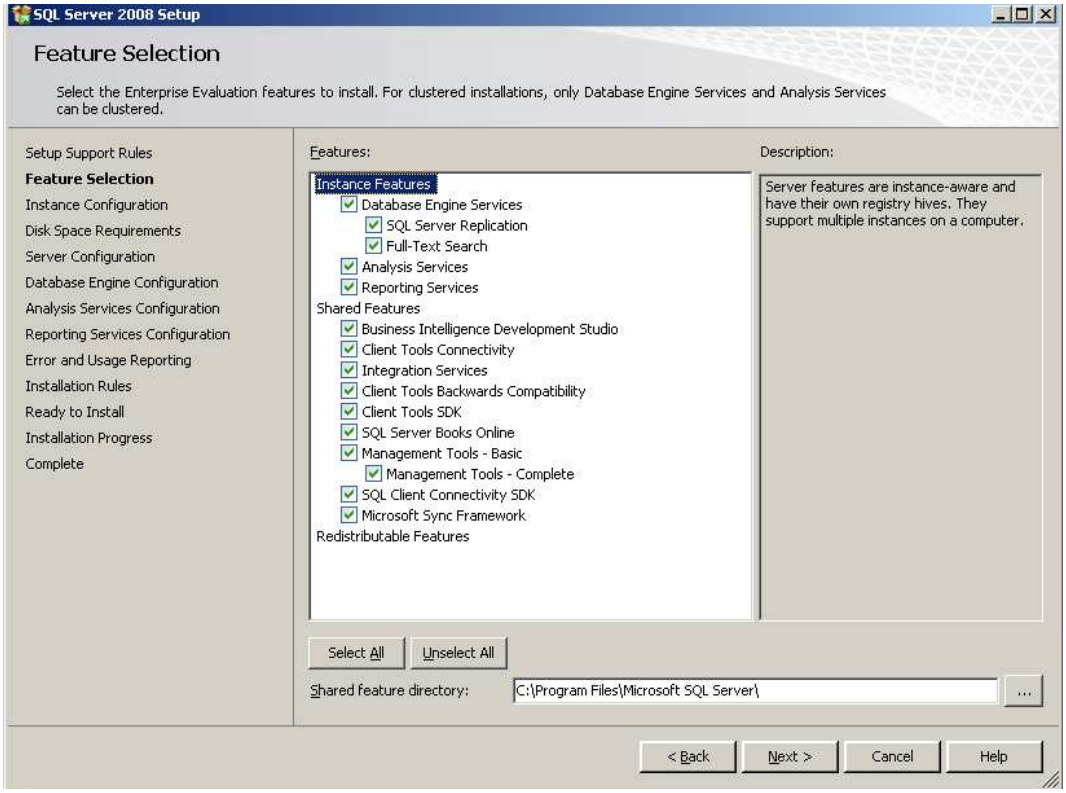


Figura V.18 Instalación Business Intelligence Studio

Ejecución

Para realizar el perfilado de datos Abrimos el programa Business Intelligence Development Studio que se encuentra en Inicio->Todos los programas ->Microsoft Sql Server 2008

Se agrega un nuevo proyecto de Integration Services

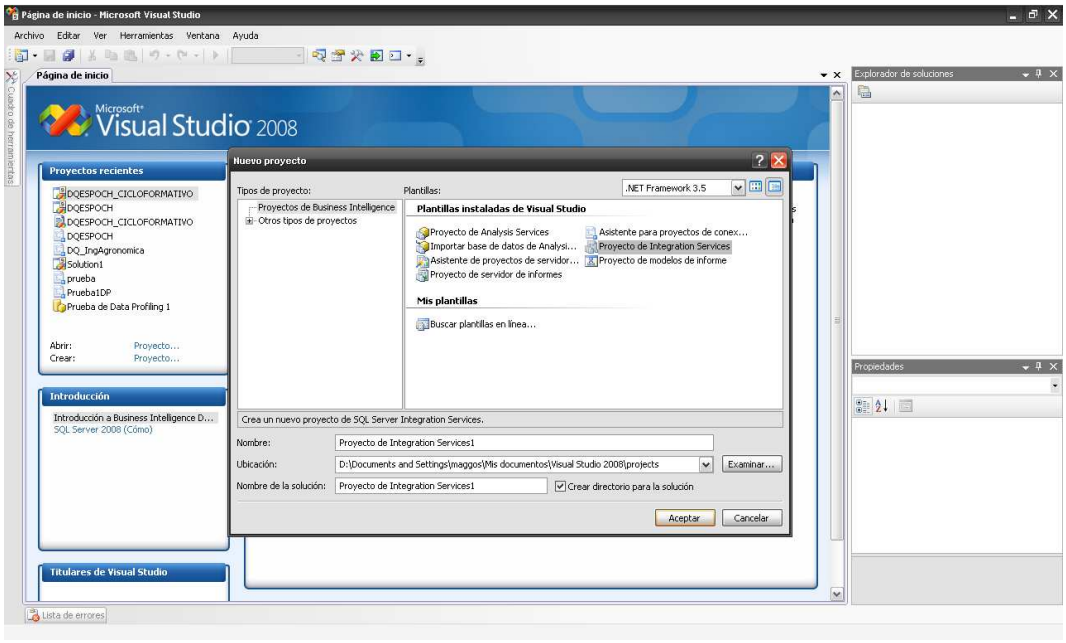


Figura V.19Selección de proyectos de integración de datos

Desde el cuadro de herramientas agregamos una tarea de generación de perfiles de datos

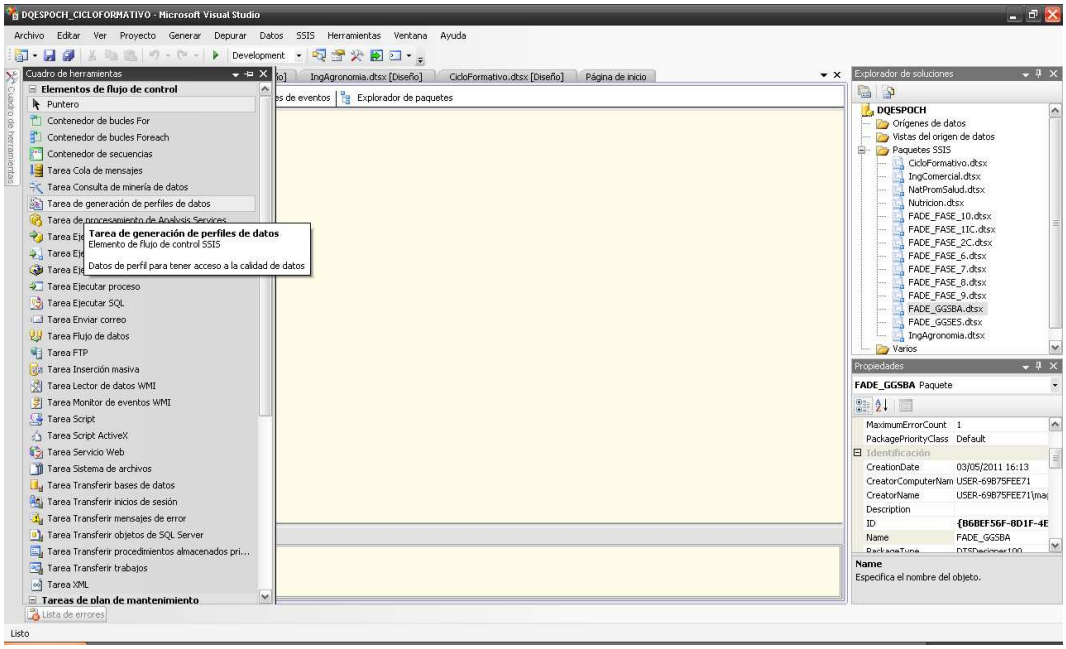


Figura V.20 Tarea de Generación de perfiles de datos

Para realizar la conexión a la fuente de datos agregamos una nueva conexión de ADO.NET y Agregamos una nueva conexión:

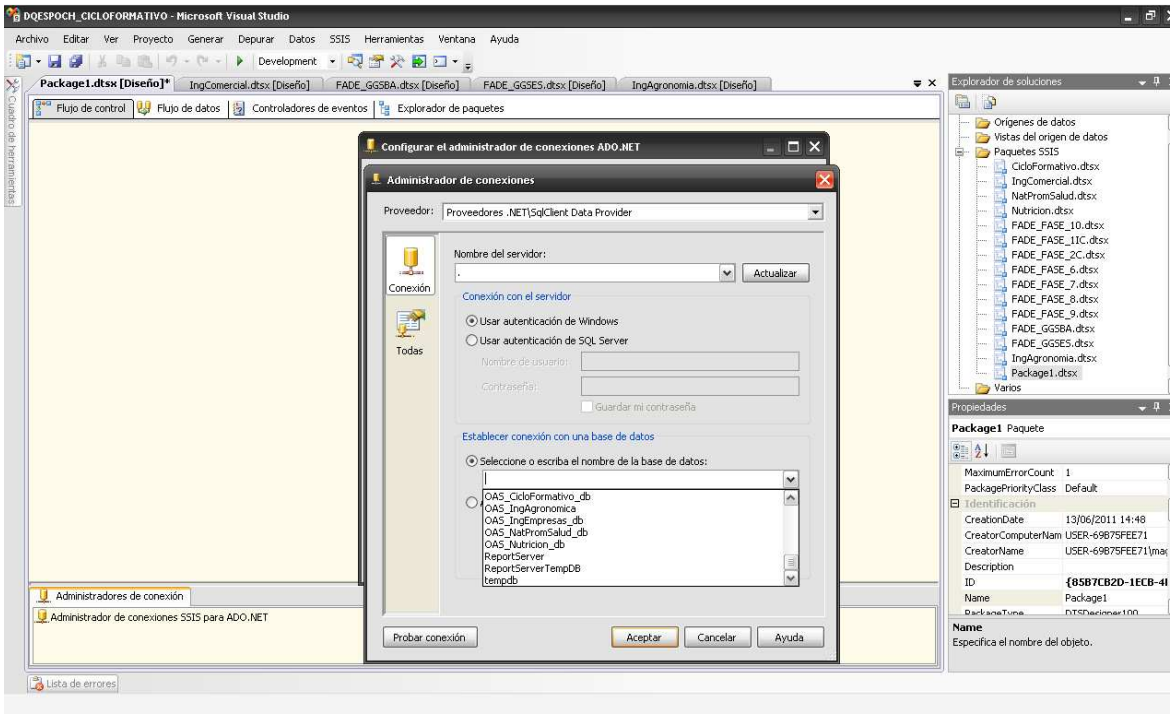


Figura V.21 Conexión al Servidor

Luego de realizar la conexión hacemos doble clik en la tarea de generación de perfiles de datos para la configuración del perfilado de datos:

En **General** configuramos el destino del perfilado de datos

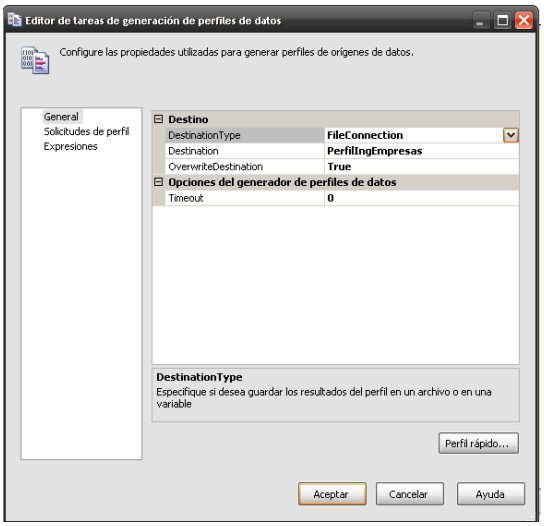


Figura V.22Opciones de Perfilado

En este caso en destination se indica la ruta donde se guardara el archivo de perfilado y Aceptamos.

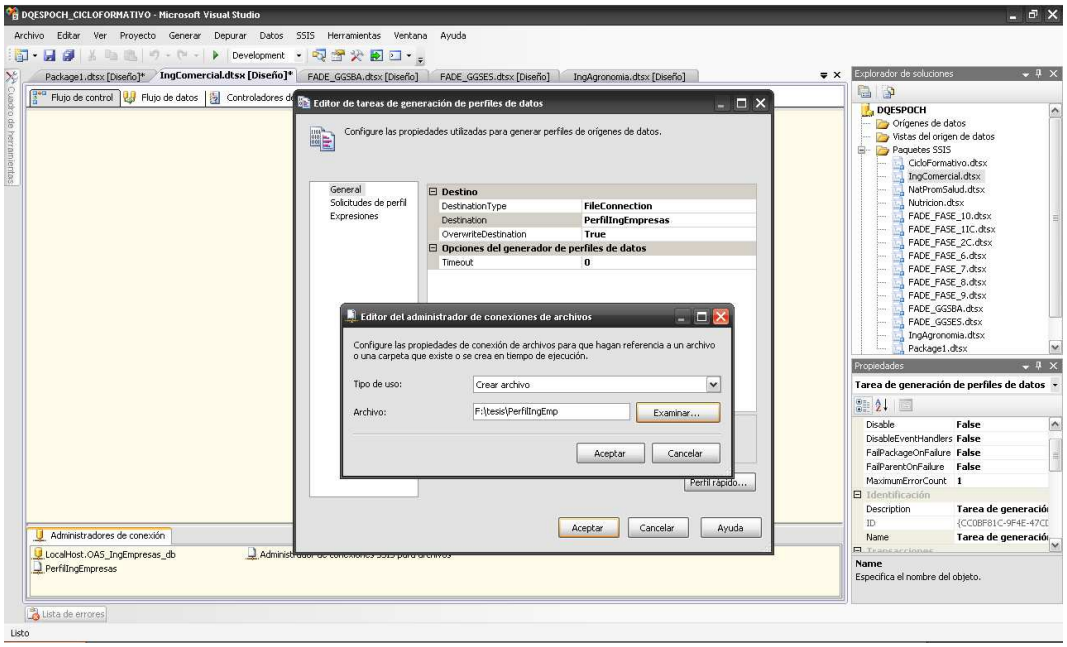


Figura V.23 Ubicación del archivo de perfilado

En **solicitudes de perfil**indica todas las solicitudes de perfilado que se puede realiza:

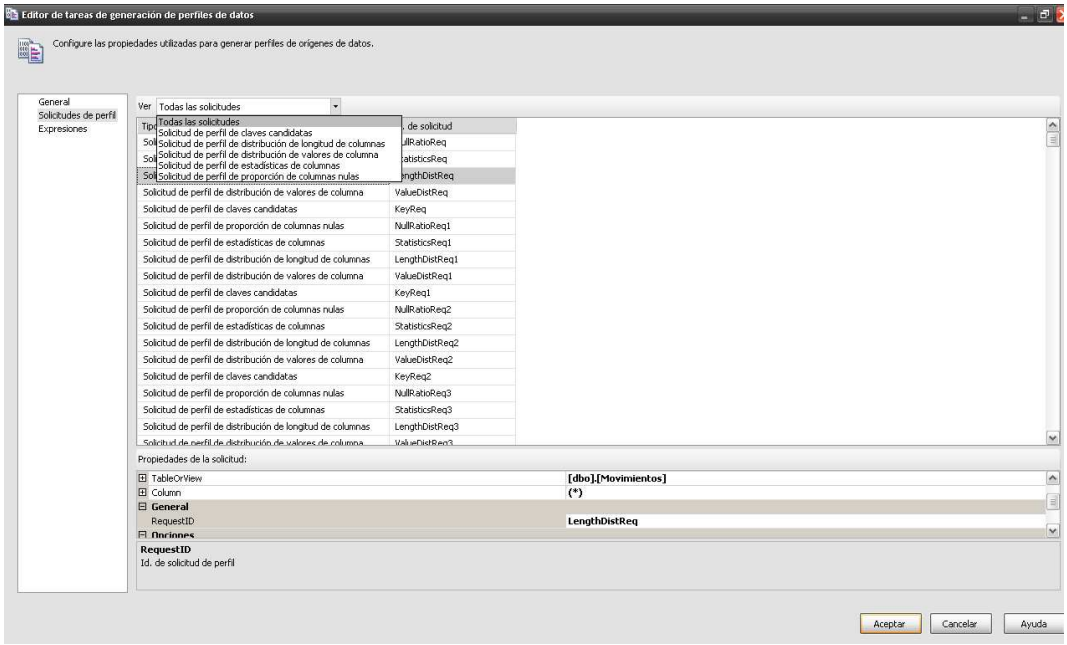


Figura V.24 Solicitudes de Perfilado de datos

Se hace clik en abrir y nos ubicamos donde se guardó el archivo de perfilado anteriormente y mostrara los resultados como se muestra en la siguiente pantalla:

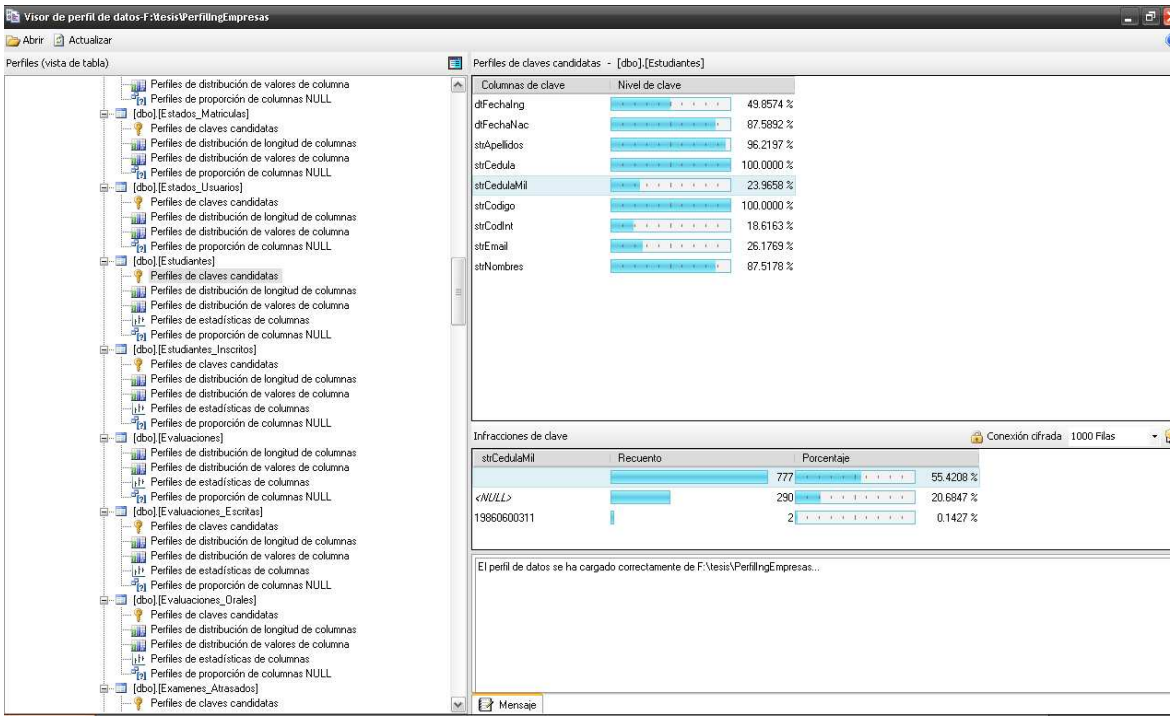


Figura V.25 Resultados del Perfilado

Podemos revisar todos los resultados de los distintos perfilados que hemos realizado para el respectivo análisis.

Todos los perfilados se encuentran en la sección de anexos

5.2.3.3 Etapa 3.3 Análisis de la calidad de datos inicial

- **Evaluación inicial de las fuentes de datos**

Escuelas de Educación a Distancia

Tabla CESTUD$-Base De Datos FADE_FASE_1IC

- **Valores NULL**

Tabla V.17 Valores NULL FADE_FASE_1IC

Columna	Cantidad de NULL	Porcentaje de NULL
APEEST	2	6%
BACHIL	2	6%
CEDIDE	3	9%
CERMIL	24	73%
CODEST	1	3%
CODPRO	2	6%
COLGRA	2	6%
DIRCUR	32	97%
DIRPER	5	15%
DOCUME	33	100%
ECUEST	2	6%
ESTCIV	2	6%
FECGRA	11	33%
FECING	2	6%
FECNAC	2	6%
LUGNAC	2	6%
NOMEST	2	6%
NOMPAD	33	100%
NUMTEF	8	24%
SEXO	2	6%

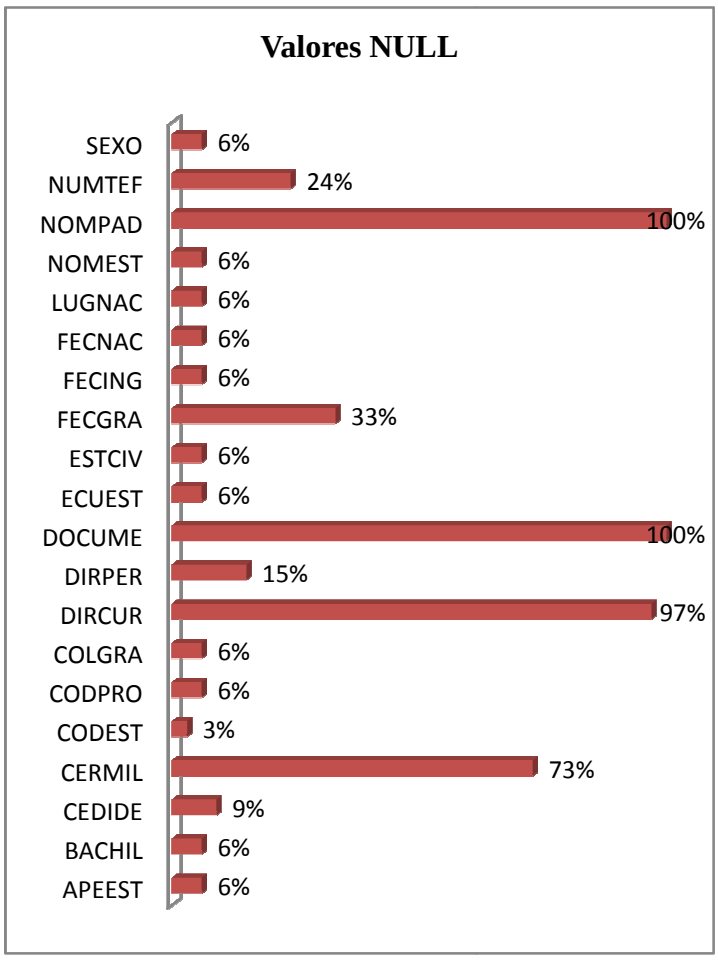


Figura V.27 Valores NULL FADE_FASE_1IC

- **Valores Vacíos**

Tabla V.18 Valores Vacíos FADE_FASE_1IC

	DIRCUR	COLGRA	ECUEST	FECNAC	NUMTEF	DOCUME	CODEST	FECGRA	APEEST	ESTCIV	BACHIL	CERMIL	DIRPER	CODPRO	NOMEST	SEXO	FECING	CEDIDE	NOMPAD	LUGNAC
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de longitud de columna**

Tabla V.19 Longitud de columna FADE_FASE_1IC

Columna	Longitud mínima	Longitud máxima
APEEST	8	18
BACHIL	11	25
CEDIDE	10	11
CERMIL	11	11
CODEST	2	6
CODPRO	3	3
COLGRA	5	24
DIRCUR	8	8
DIRPER	6	43
ECUEST	1	1
ESTCIV	1	1
LUGNAC	4	8
NOMEST	5	18
NUMTEF	6	6
SEXO	1	1

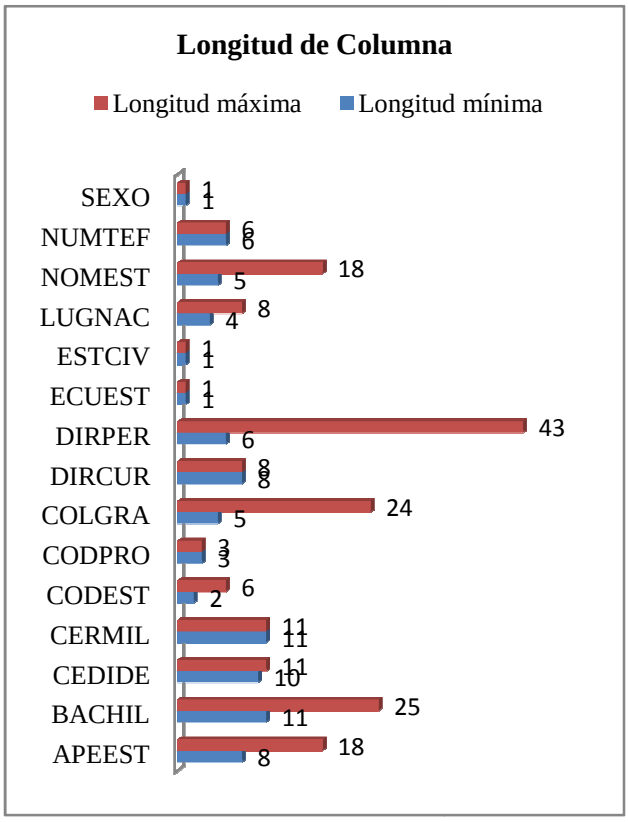


Figura V.28 Longitud de columna FADE_FASE_1IC

- **Distribución de valores de columna**

Tabla V.20 Distribución de valores de columna

FADE_FASE_1IC

Columna	Número de valores distintos
APEEST	30
BACHIL	11
CEDIDE	30
CERMIL	9
CODEST	32
CODPRO	3
COLGRA	21
DIRCUR	1
DIRPER	27
DOCUME	0
ECUEST	1
ESTCIV	3
FECGRA	21
FECING	24
FECNAC	31
LUGNAC	12
NOMEST	30
NOMPAD	0
NUMTEF	23
SEXO	2

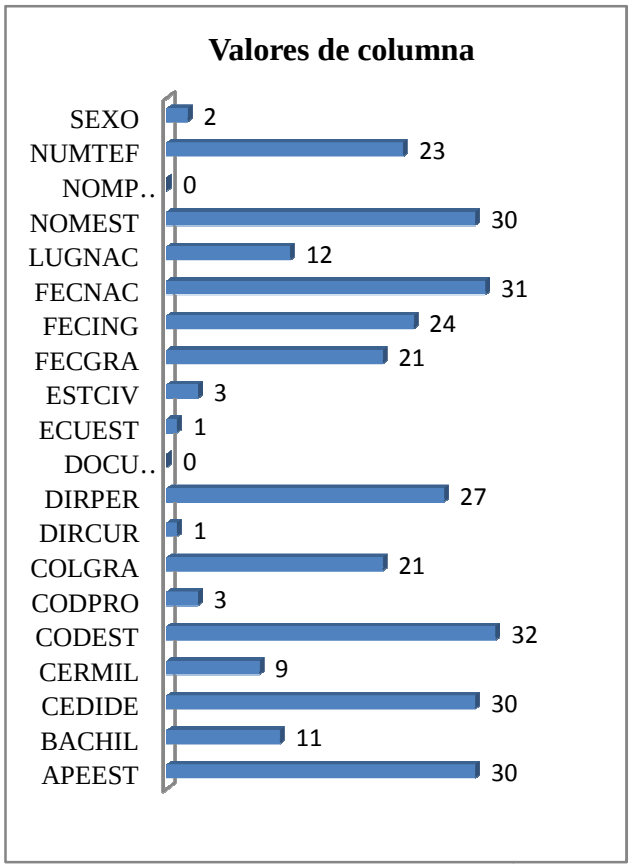


Figura V.29 Distribución de longitud de columnas

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.21 Caracteres Mayúsculas minúsculas FADE_FASE_1IC

	DIRCUR	COLGRA	ECUEST	NUMTEF	DOCUME	CODEST	APEEST	ESTCIV	BACHIL	CERMIL	DIRPER	CODPRO	NOMEST	SEXO	CEDIDE	NOMPAD	LUGNAC
Uppercase chars	100 %	88 %	100 %	0 %	0 %	0 %	100 %	100 %	93 %	0 %	73 %	100 %	100 %	100 %	0 %	0 %	100 %
Lowercase chars	0%	0 %	0%	0 %	0 %	0 %	0%	0%	0 %	0 %	0 %	0%	0%	0%	0 %	0 %	0%

- **Duplicación**

Tabla V.22 Duplicación FADE_FASE_1IC

Columnas de clave	Nivel de clave	Duplicación
APEEST	94%	6%
BACHIL	36%	64%
CEDIDE	94%	6%
CERMIL	30%	70%
CODEST	100%	0%
CODPRO	12%	88%
COLGRA	67%	33%
DIRCUR	6%	94%
DIRPER	85%	15%
ECUEST	6%	94%
ESTCIV	12%	88%
FECGRA	67%	33%
FECING	76%	24%
FECNAC	97%	3%
LUGNAC	39%	61%
NOMEST	94%	6%
NUMTEF	73%	27%
SEXO	9%	91%

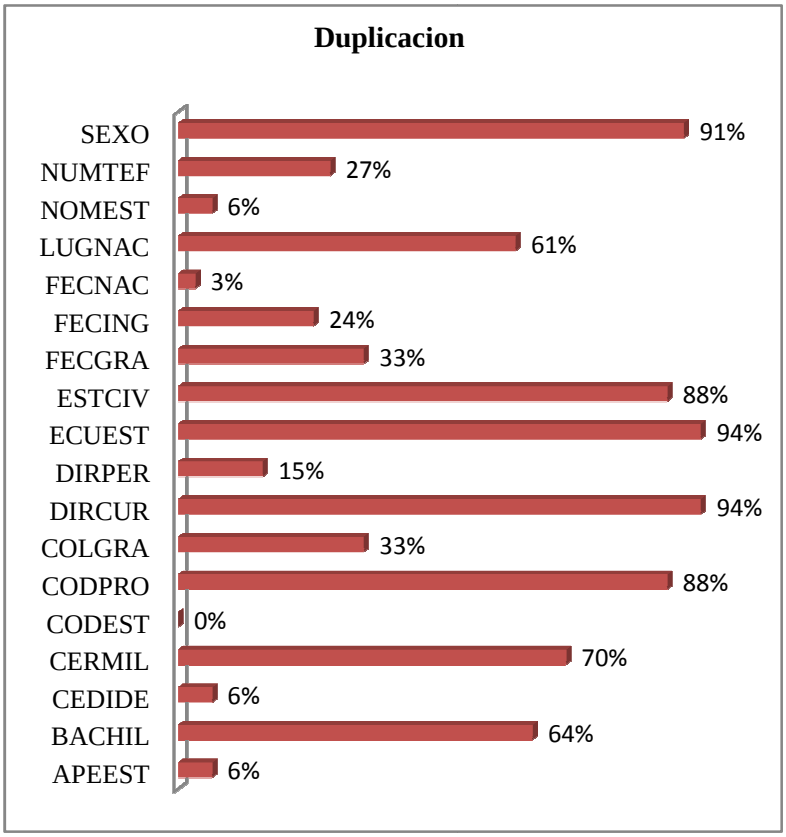


Figura V.30 Duplicación FADE_FASE_1IC

- **Tiempo**

Tabla V.23 Tiempo FADE_FASE_1IC

	FECNAC	FECGRA	FECING
Highest value	26/06/1987	08/07/2005	30/01/2007
Lowest value	20/07/1965	17/09/1982	20/07/2005

- **Patrones**

Tabla V.24 Patrones FADE_FASE_1IC

COLUMNA	PATRON	CANTIDAD
CEDIDE	9999999999	27
	999999999-9	3

Tabla CESTUD$-Base de Datos FADE_FASE_2IC

- **Valores NULL**

Tabla V.25 Valores NULL FADE_FASE_2IC

Columna	Recuento NULL	Porcentaje de NULL
APEEST	0	0%
BACHIL	0	0%
CEDIDE	0	0%
CERMIL	15	63%
CODEST	0	0%
CODPRO	0	0%
COLGRA	0	0%
DIRCUR	23	96%
DIRPER	1	0,04166667
DOCUME	24	100%
ECUEST	0	0%
ESTCIV	0	0%
FECGRA	0	0%
FECING	0	0%
FECNAC	0	0%
LUGNAC	0	0%
NOMEST	0	0%
NOMPAD	24	100%

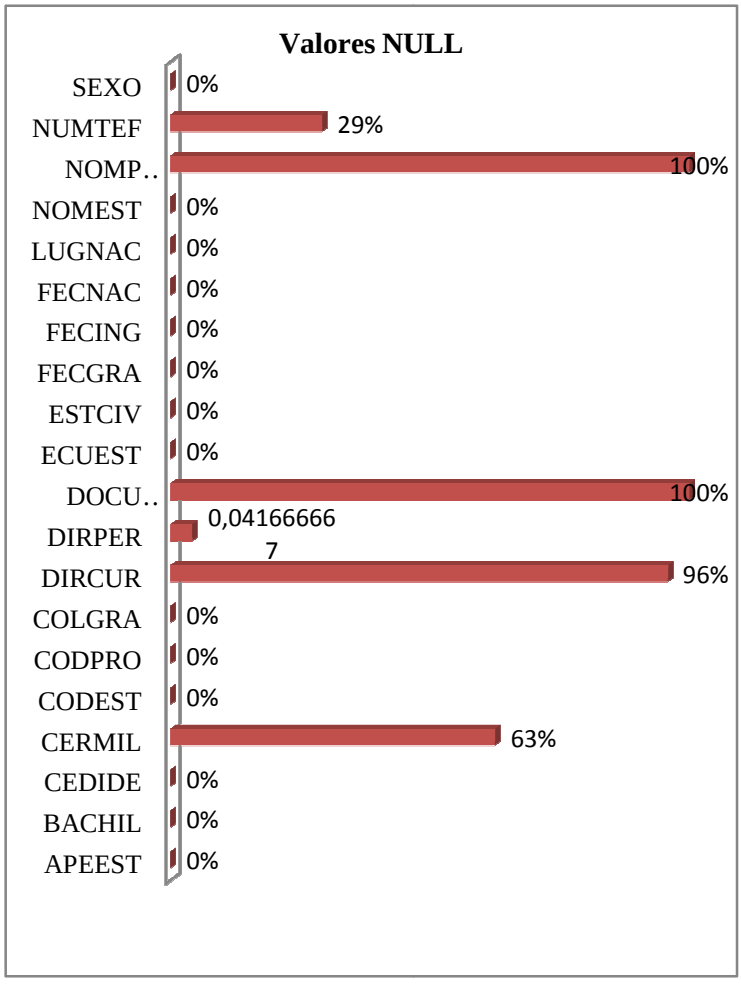


Figura V.31 Valores NULL FADE_FASE_2IC

NUMTEF	7	29%
SEXO	0	0%

- **Valores Vacíos**

	DIRCUR	COLGRA	ECUEST	FECNAC	NUMTEF	DOCUME	CODEST	FECGRA	APEEST	ESTCIV	BACHIL	CERMIL	DIRPER	CODPRO	NOMEST	SEXO	FECING	CEDIDE	NOMPAD	LUGNAC
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de longitud de columnas**

Tabla V.26 Longitud de columnas FADE_FASE_2IC

Columna	Longitud mínima	Longitud máxima
APEEST	10	19
BACHIL	8	25
CEDIDE	11	11
CERMIL	10	11
CODEST	6	6
CODPRO	3	3
COLGRA	14	25
DIRCUR	3	3
DIRPER	5	50
ECUEST	1	1
ESTCIV	1	1
LUGNAC	5	22
NOMEST	4	17
NUMTEF	5	6
SEXO	1	1

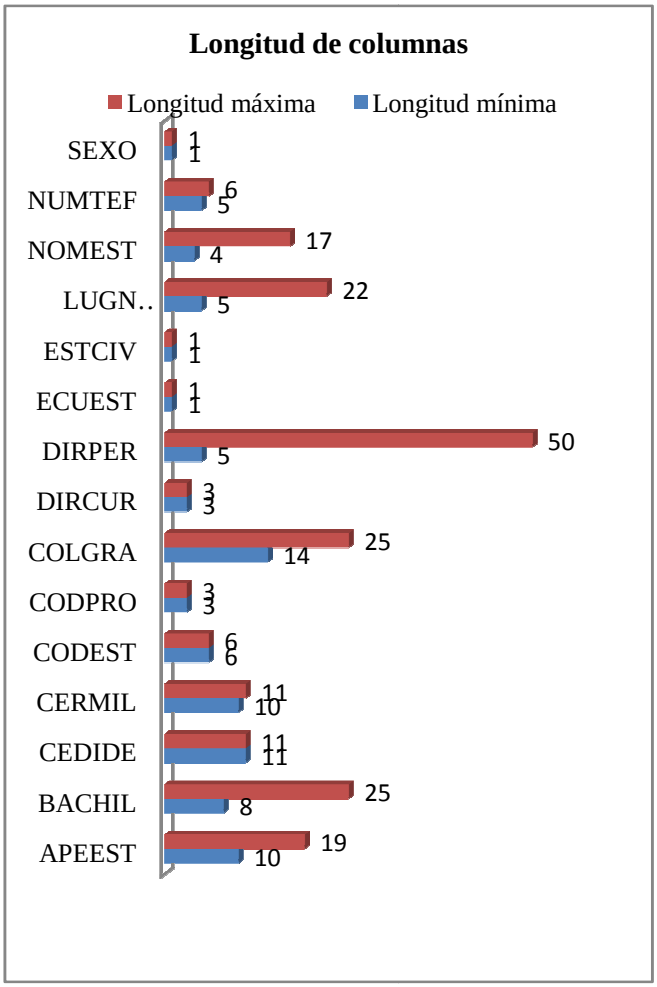


Figura V.32 Longitud de columnas FADE_FASE_2IC

- **Distribución de valores de columna**

Tabla V.27 Distribución de valores de columna

Columna	Número de valores distintos
APEEST	24
BACHIL	9
CEDIDE	24
CERMIL	9
CODEST	24
CODPRO	2
COLGRA	18
DIRCUR	1
DIRPER	21
DOCUME	0
ECUEST	1
ESTCIV	3
FECGRA	22
FECING	5
FECNAC	24
LUGNAC	13
NOMEST	24
NOMPAD	0
NUMTEF	16
SEXO	2

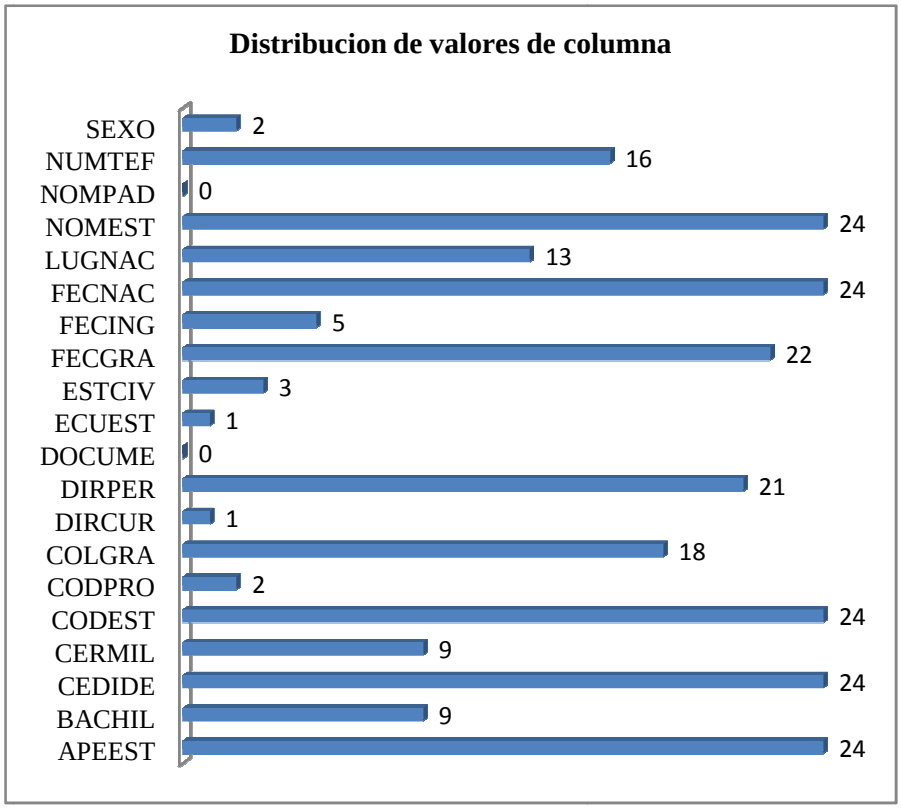


Figura V.33 Distribución de valores de columna

- **Caracteres Mayúsculas y Minúsculas**

TablaV.28 Caracteres Mayúsculas y minúsculas FADE_FASE_2IC

	DIRCUR	COLGRA	ECUEST	NUMTEF	DOCUME	CODEST	APEEST	ESTCIV	BACHIL	CERMIL	DIRPER	CODPRO	NOMEST	SEXO	CEDIDE	NOMPAD	LUGNAC
Uppercase chars	0%	88 %	100 %	4 %	0 %	0 %	100 %	100 %	94 %	0 %	58 %	100 %	100 %	100 %	0 %	0 %	94 %
Lowercase chars	100 %	0 %	0%	0 %	0 %	0 %	0%	0%	0 %	0 %	16 %	0%	0%	0%	0 %	0 %	0 %

- **Duplicación**

Tabla V.29 Duplicación FADE_FASE_2IC

Columnas de clave	Nivel de clave	Duplicación
APEEST	100%	0%
BACHIL	38%	63%
CEDIDE	100%	0%
CERMIL	42%	58%
CODEST	100%	0%
CODPRO	8%	92%
COLGRA	75%	25%
DIRCUR	8%	92%
DIRPER	92%	8%
ESTCIV	13%	88%
FECGRA	92%	8%
FECING	21%	79%
FECNAC	100%	0%
LUGNAC	54%	46%
NOMEST	100%	0%
NUMTEF	71%	29%
SEXO	8%	92%

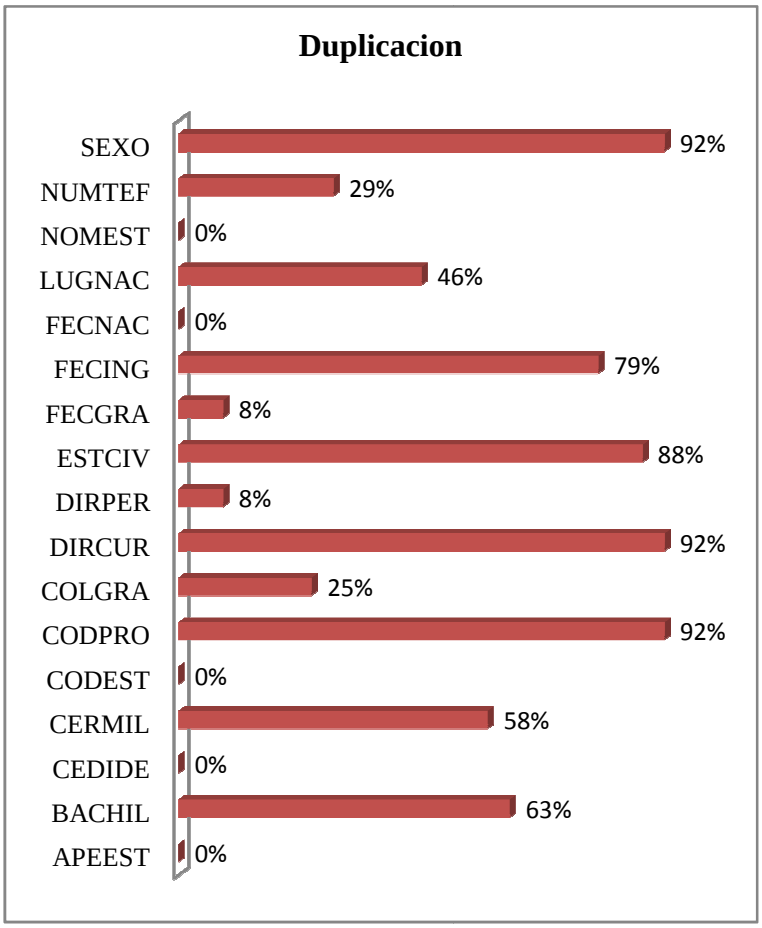


Figura V.35 Duplicacion FADE_FASE_2IC

- **Tiempo**

Tabla V.30 Tiempo FADE_FASE_2IC

	FECNAC	FECGRA	FECING
Highest value	03/07/1989	17/08/2007	13/12/2007
Lowest value	19/09/1960	22/01/1988	26/10/1907

- **Patrones**

Tabla V.0.31 Patrones FADE_FASE_2IC

Columna	Patrón	Cantidad
CEDIDE	999999999-9	24

Tabla CESTUD$ Base de Datos FADE_FASE_6

- **Valores NULL**

Tabla V.32 Valores NULL FADE_FASE_6

Columna	Recuento de NULL	Porcentaje de NULL
APEEST	1	0,2793%
BACHIL	14	4%
CEDIDE	5	1%
CERMIL	305	85%
CODEST	1	0,2793%
CODPRO	1	0,02%
COLGRA	14	4%
DIRCUR	320	89%
DIRPER	23	6%
DOCUME	355	99%
ECUEST	1	0,2793%
ESTCIV	1	0,02%
FECGRA	37	10%
FECING	4	1%
FECNAC	13	4%
LUGNAC	12	3%
NOMEST	1	0,2793%
NOMPAD	265	74%
NUMTEF	42	12%
SEXO	1	0,2793%

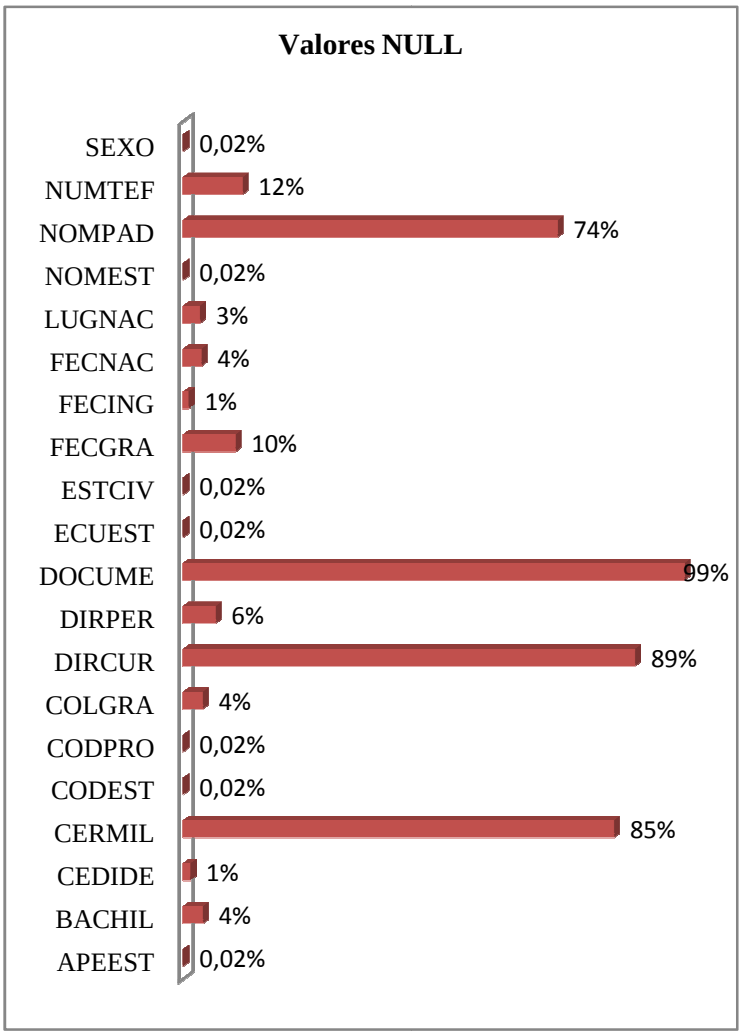


Figura V.36 Valores NULL FADE_FASE_6

- **Valores Vacíos**

Tabla V.33 Valores vacíos FADE_FASE_6

	APEEST	LUGNAC	DIRCUR	BACHIL	NUMTEF	DIRPER	FECING	CEDIDE	DOCUME	CODPRO	FECGRA	NOMEST	ESTCIV	ECUEST	CODEST	FECNAC	CERMIL	SEXO	NOMPAD	COLGRA
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de longitud de columna**

Tabla V.34 Distribución de longitud de columna FADE_FASE_6

Columna	Longitud mínima	Longitud máxima
APEEST	5	21
BACHIL	8	25
CEDIDE	7	11
CERMIL	9	11
CODEST	6	6
CODPRO	3	3
COLGRA	6	25
DIRCUR	3	42
DIRPER	5	50
DOCUME	1	1
ECUEST	1	1
ESTCIV	1	1
LUGNAC	4	25
NOMEST	4	24
NOMPAD	1	23
NUMTEF	4	9
SEXO	1	1

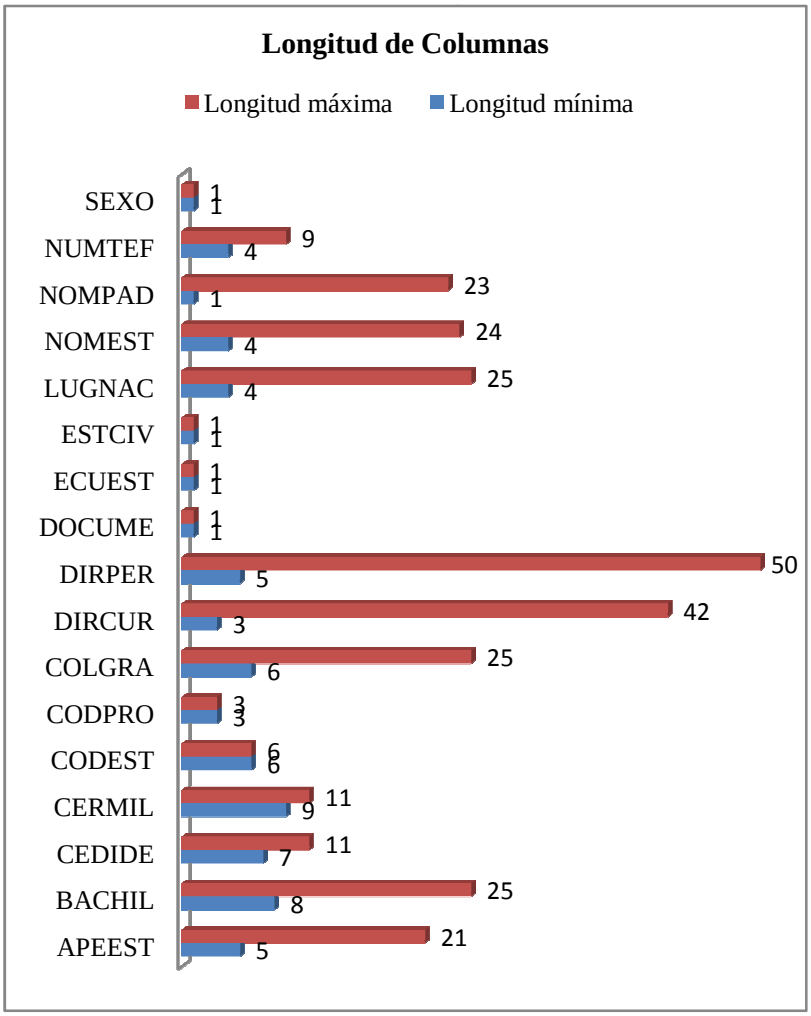


Figura V.37 Distribución de longitud FADE_FASE_6

- **Distribución de Valores de columna**

Tabla V.35 Distribución de longitud de columnas

Columna	Número de valores distintos
APEEST	339
BACHIL	55
CEDIDE	351
CERMIL	52
CODEST	357
CODPRO	3
COLGRA	90
DIRCUR	37
DIRPER	321
DOCUME	2
ECUEST	1
ESTCIV	4
FECGRA	208
FECING	66
FECNAC	339
LUGNAC	94
NOMEST	333
SEXO	2

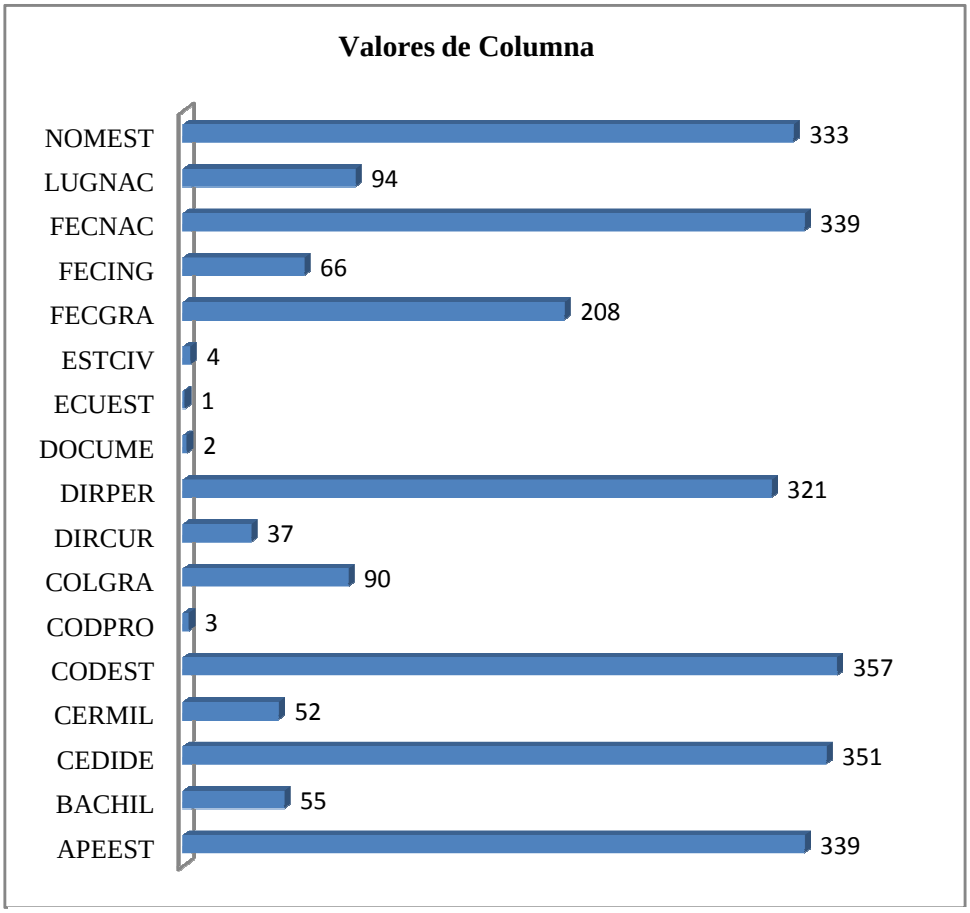
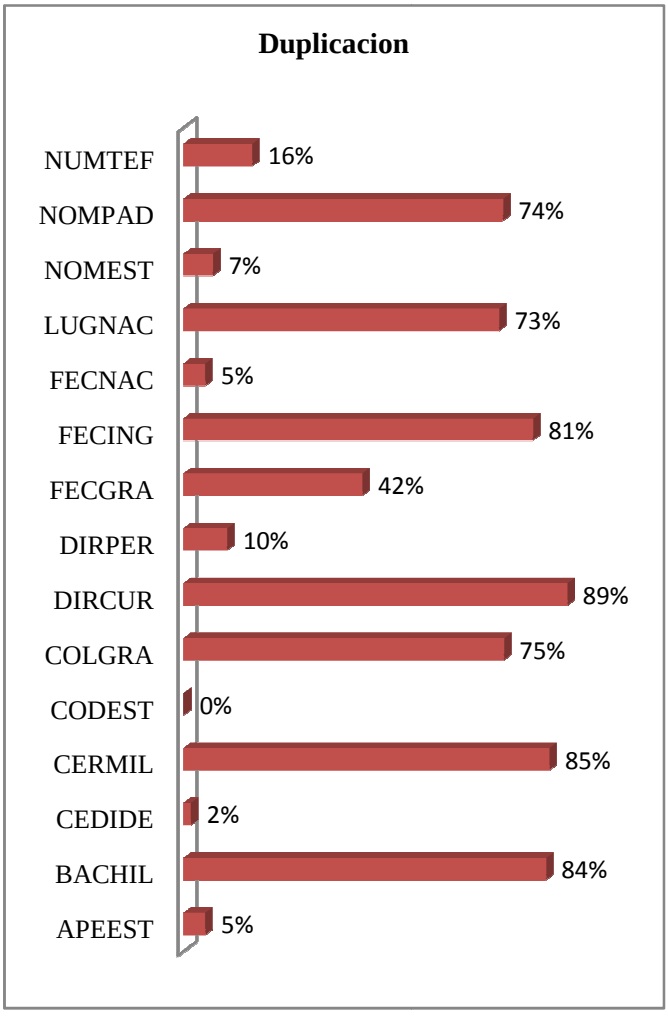


Figura V.0.38 Distribución de valores de columna

- **Duplicación**

Tabla V.36 Duplicación FADE_FASE_6

Columnas de clave	Nivel de clave	Duplicación
APEEST	95%	5%
BACHIL	16%	84%
CEDIDE	98%	2%
CERMIL	15%	85%
CODEST	100%	0%
COLGRA	25%	75%
DIRCUR	11%	89%
DIRPER	90%	10%
FECGRA	58%	42%
FECING	19%	81%
FECNAC	95%	5%
LUGNAC	27%	73%
NOMEST	93%	7%
NOMPAD	26%	74%
NUMTEF	84%	16%



FiguraV.40 Duplicacion FADE_FASE_6

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.37 Mayusculas y Minúsculas FADE_FASE_6

	APEEST	LUGNAC	DIRCUR	BACHIL	NUMTEF	DIRPER	CEDIDE	DOCUME	CODPRO	NOMEST	ESTCIV	ECUEST	CODEST	CERMIL	SEXO	NOMPAD	COLGRA
Uppercase chars	100 %	100 %	100 %	100 %	0 %	100 %	0 %	0 %	100 %	100 %	100 %	100 %	0 %	0 %	100 %	100 %	100 %
Lowercase chars	0%	0%	0%	0%	0 %	0%	0 %	0 %	0%	0%	0%	0%	0 %	0 %	0%	0%	0%

- **Tiempo**

Tabla V.38 Tiempo FADE_FASE_6

	FECING	FECGRA	FECNAC
Highest value	16/12/1999	09/11/1999	17/11/1993
Lowest value	12/02/1900	20/06/1900	06/03/1900

- **Patrones**

Tabla V.39 Patrones FADE_FASE_6

Columna	Patrón	Cantidad
CEDIDE	999999999-9	349
	99999999999	4

CESTUD -Base de Datos FADE_FASE_7

- **Valores NULL**

Tabla V.40 Valores NULL FADE_FASE_7

Columna	Recuento de NULL	Porcentaje de NULL
APEEST	1	0,0022
BACHIL	14	3%
CEDIDE	5	1%
CERMIL	375	84%
CODEST	1	0,002
CODPRO	1	0,002
COLGRA	14	3%
DIRCUR	409	91%
DIRPER	35	8%
DOCUME	444	99%
ECUEST	1	0,002
ESTCIV	1	0,002
FECGRA	48	11%
FECING	4	0,0089
FECNAC	13	3%
LUGNAC	12	3%
NOMEST	1	0%
NOMPAD	354	79%
NUMTEF	60	13%
SEXO	1	0,002

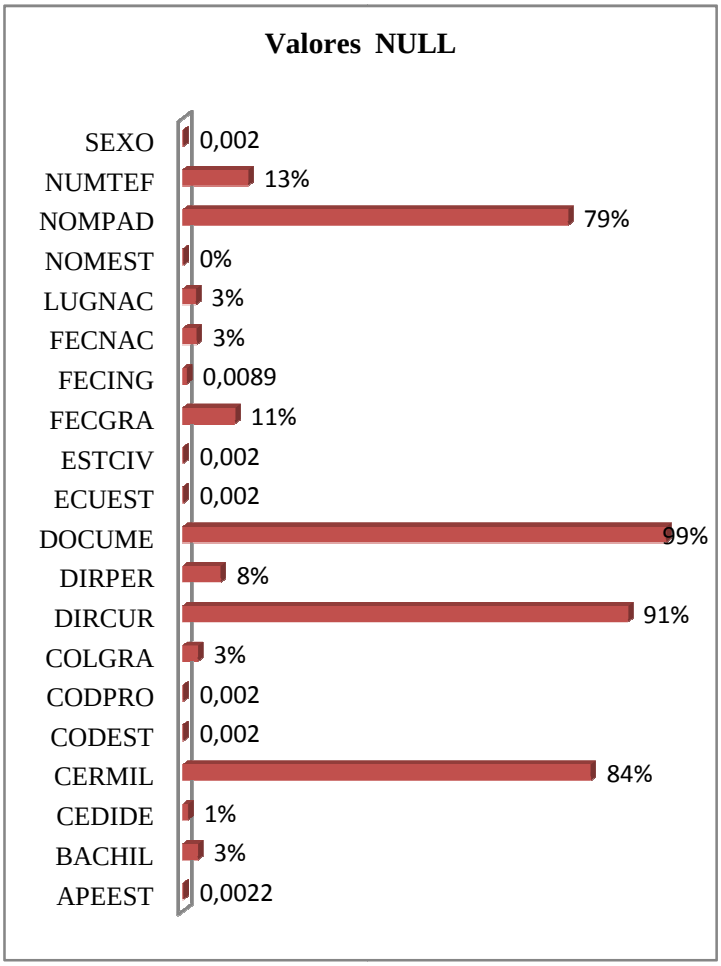


Figura V.42 Valores NULL FADE_FASE_7

- **Valores Vacíos**

Tabla V.41 Valores vacíos FADE_FASE_7

	NUMTEF	DIRCUR	CERMIL	FECING	SEXO	DOCUME	CODEST	BACHIL	ECUEST	NOMPAD	FECNAC	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	DIRPER	COLGRA	FECGRA
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de Longitud de Columnas**

Tabla V.42 Distribución de longitud de columnas FADE_FASE_7

Columna	Longitud mínima	Longitud máxima
APEEST	4	21
BACHIL	8	25
CEDIDE	7	11
CERMIL	9	11
CODEST	6	6
CODPRO	3	3
COLGRA	6	25
DIRCUR	3	42
DIRPER	5	50
DOCUME	1	1
ECUEST	1	1
ESTCIV	1	1
LUGNAC	4	25
NOMEST	4	24
NOMPAD	1	23
NUMTEF	4	10
SEXO	1	1

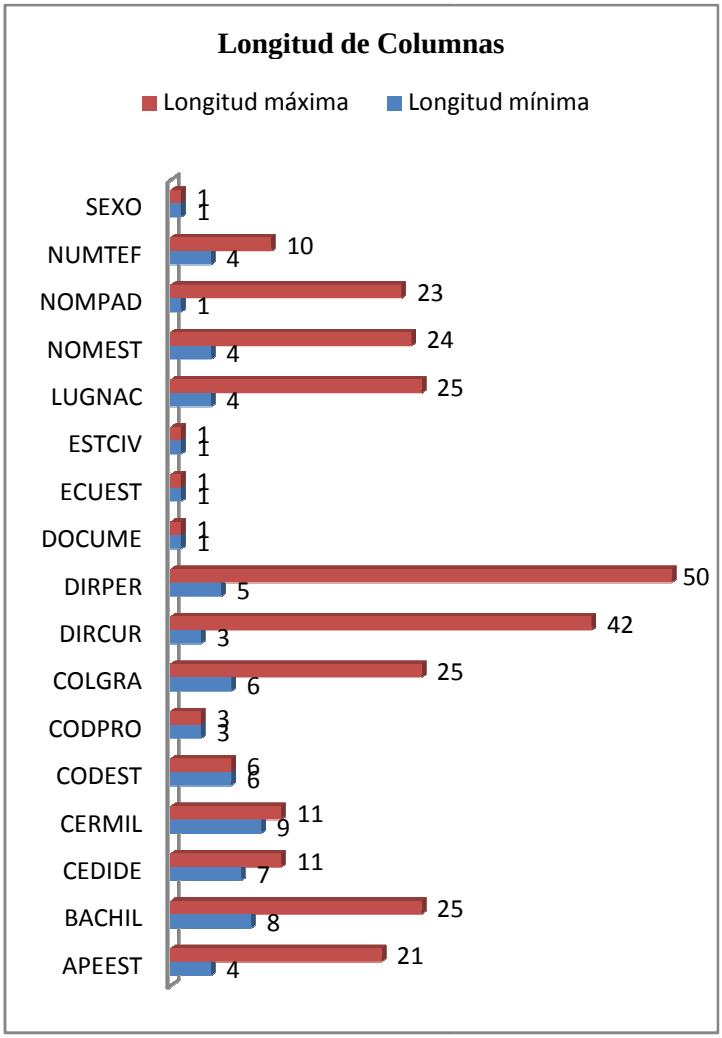


Figura V.43 Distribución de longitud FADE_FASE_6

- **Distribución de valores de columna**

Tabla V.43Distribución de valores de columna FADE_FASE_7

Columna	Número de valores distintos
APEEST	427
BACHIL	62
CEDIDE	440
CERMIL	71
CODEST	446
CODPRO	3
COLGRA	116
DIRCUR	37
DIRPER	392
DOCUME	2
ECUEST	1
ESTCIV	4
FECGRA	248
FECING	68
FECNAC	427
LUGNAC	107
NOMEST	416
NOMPAD	91
NUMTEF	367
SEXO	2

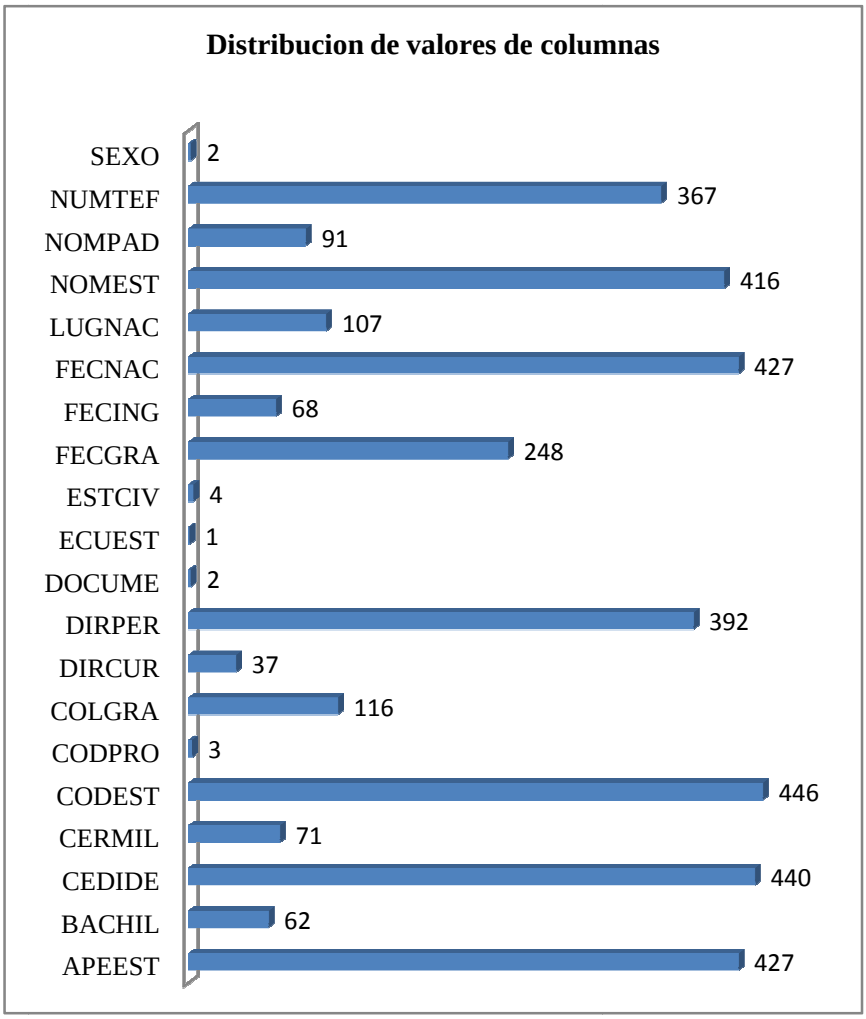


Figura V.44 Distribución de valores de columna FADE_FASE_7

- **Duplicación**

Tabla V.44 Duplicación FADE_FASE_7

Columnas de clave	Nivel de clave	Duplicación
APEEST	96%	4%
BACHIL	14%	86%
CEDIDE	99%	1%
CERMIL	16%	84%
CODEST	100%	0%
COLGRA	26%	74%
DIRCUR	9%	91%
DIRPER	88%	12%
FECGRA	56%	44%
FECING	15%	85%
FECNAC	96%	4%
LUGNAC	24%	76%
NOMEST	93%	7%
NOMPAD	21%	79%
NUMTEF	82%	18%

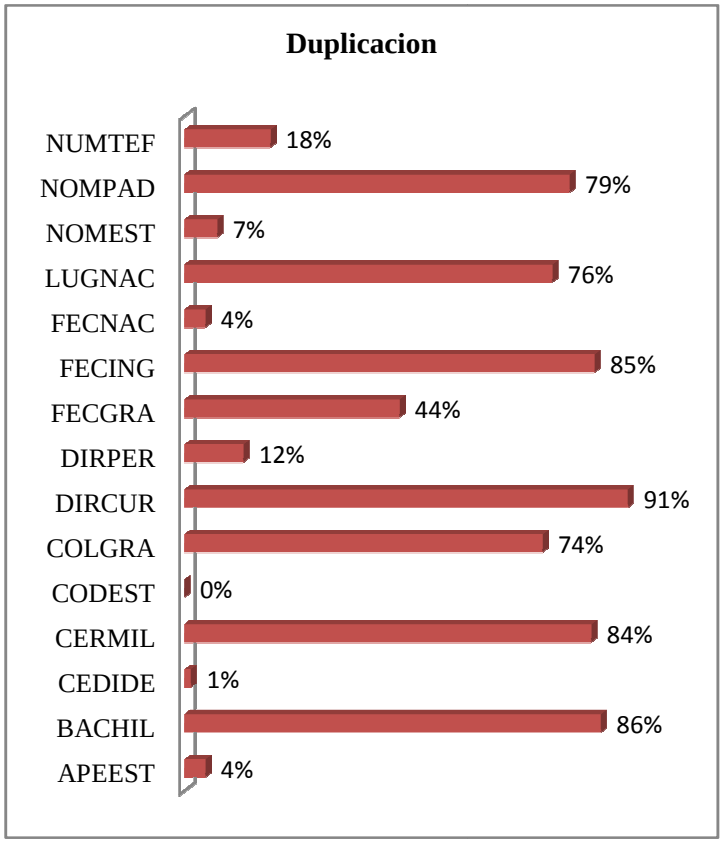


Figura V.45 Duplicación FADE_FASE_7

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.45 Mayúsculas y minúsculas FADE_FASE_7

	NUMTEF	DIRCUR	CERMIL	SEXO	DOCUME	CODEST	BACHIL	ECUEST	NOMPAD	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	DIRPER	COLGRA
Uppercase chars	0 %	76 %	0 %	100 %	0 %	0 %	94 %	100 %	92 %	0 %	100 %	100 %	100 %	96 %	100 %	35 %	88 %
Lowercase chars	0 %	0 %	0 %	0%	0 %	0 %	0 %	0%	0 %	0 %	0%	0%	0%	0 %	0%	35 %	0 %

- **Tiempo**

Tabla V.46 Tiempo FADE_FASE_7

	FECING	FECNAC	FECGRA
Highest value	16/12/1999 0:00	17/11/1993 0:00:00	23/11/1999 0:00:00
Lowest value	12/02/1900 0:00	06/03/1900 0:00:00	18/02/1900 0:00:00

- **Patrones**

Tabla V.47 Patrones FADE_FASE_7

Columna	Patrón	Cantidad
CEDIDE	999999999-9	438
	99999999999	4

Análisis Estadístico de la Tabla CESTUD Base de Datos FADE_FASE_8

- **Valores NULL**

Tabla V.48 Valores NULL FADE_FASE_8

Columna	Recuento de NULL	Porcentaje de NULL
APEEST	1	0,00223714
BACHIL	14	3%
CEDIDE	5	1%
CERMIL	375	84%
CODEST	1	0,00223714
CODPRO	1	0,00223714
COLGRA	14	3%
DIRCUR	409	91%
DIRPER	35	8%
DOCUME	444	99%
ECUEST	1	0,00223714
ESTCIV	1	0,00223714
FECGRA	48	11%
FECING	4	0,00894855
FECNAC	13	3%
LUGNAC	12	3%
NOMEST	1	0,00223714
NOMPAD	354	79%

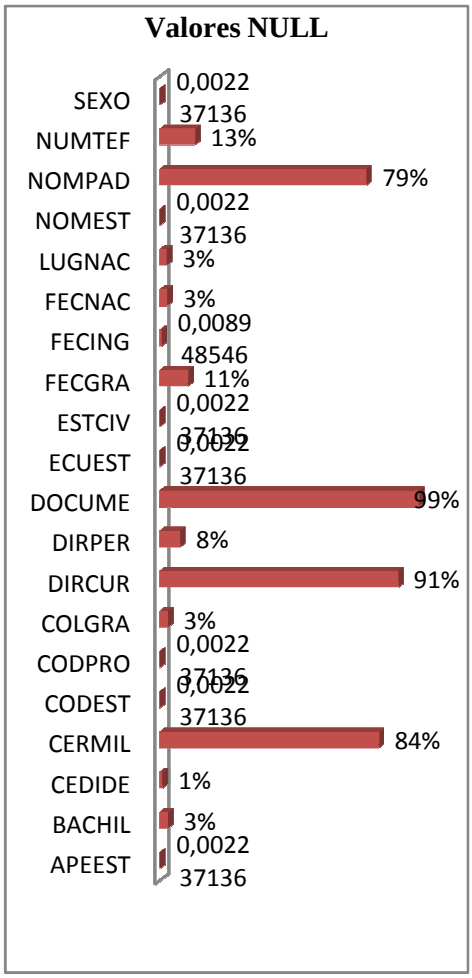


Figura V.46Valores NULL FADE_FASE_8

NUMTEF	60	13%
SEXO	1	0,00223714

- **Valores en Blanco**

Tabla V.49 Valores vacíos FADE_FASE_8

	NUMTEF	DIRCUR	CERMIL	FECING	SEXO	DOCUME	CODEST	BACHIL	ECUEST	NOMPAD	FECNAC	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	DIRPER	COLGRA	FECGRA
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de Longitud de Columnas**

Tabla V.50 Distribución de longitud de columnas FADE_FASE_8

Columna	Longitud mínima	Longitud máxima
APEEST	4	21
BACHIL	8	25
CEDIDE	7	11
CERMIL	9	11
CODEST	6	6
CODPRO	3	3
COLGRA	6	25
DIRCUR	3	42
DIRPER	5	50
DOCUME	1	1
ECUEST	1	1
ESTCIV	1	1
LUGNAC	4	25
NOMEST	4	24
NOMPAD	1	23
NUMTEF	4	10
SEXO	1	1

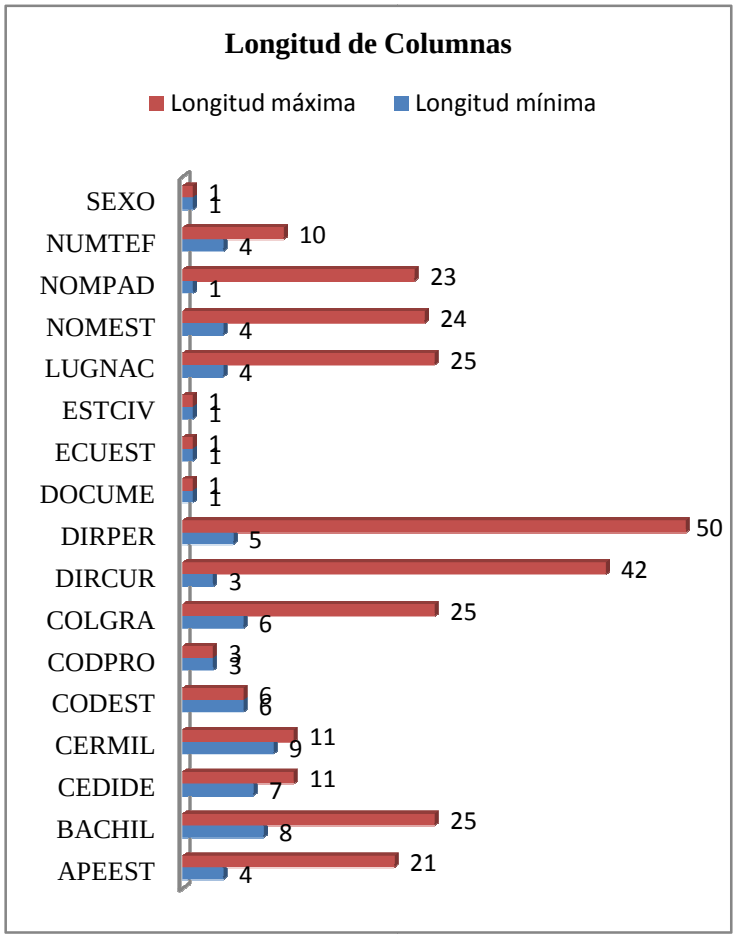


Figura V.47Distribucion de longitud de columnas FADE_FASE_8

- **Distribución de Valores de columnas**

Tabla V.51 Distribución de valores de columnas FADE_FASE_8

Columna	Longitud mínima	Longitud máxima
APEEST	4	21
BACHIL	8	25
CEDIDE	7	11
CERMIL	9	11
CODEST	6	6
CODPRO	3	3
COLGRA	6	25
DIRCUR	3	42
DIRPER	5	50
DOCUME	1	1
ECUEST	1	1
ESTCIV	1	1
LUGNAC	4	25
NOMEST	4	24
NOMPAD	1	23
NUMTEF	4	10
SEXO	1	1

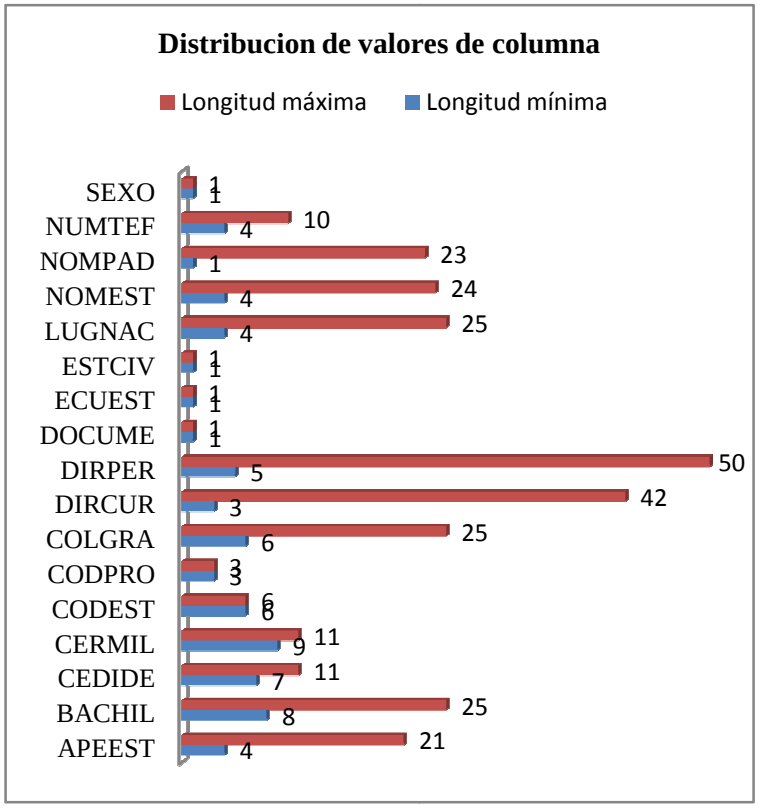
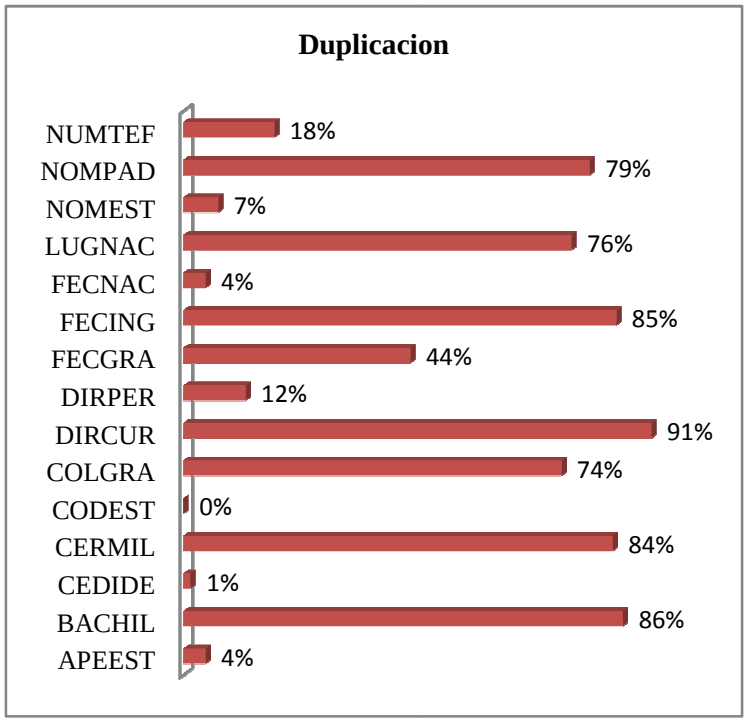


Figura V.48Distribucion de valores de columna FADE_FASE_8

- **Duplicación**

Tabla V.52 Duplicación FADE_FASE_8

Columnas de clave	Nivel de clave	Duplicación
APEEST	96%	4%
BACHIL	14%	86%
CEDIDE	99%	1%
CERMIL	16%	84%
CODEST	100%	0%
COLGRA	26%	74%
DIRCUR	9%	91%
DIRPER	88%	12%
FECGRA	56%	44%
FECING	15%	85%



FiguraV.49 Duplicacion FADE_FASE_8

FECNAC	96%	4%
LUGNAC	24%	76%
NOMEST	93%	7%
NOMPAD	21%	79%
NUMTEF	82%	18%

- **Caracteres Mayúsculas y minúsculas**

Tabla V.53 Mayúsculas y minúsculas FADE_FASE_8

	NOMPAD	**BACHILE**	**APEEST**	**ESTCIV**	**NOMEST**	**DIRPER**	**CODPRO**	**CERMIL**	**CEDIDE**	**DIRCUR**	**NUMTEF**	**LUGNAC**	**ECUEST**	**SEXO**	**CODEST**	**COLGRA**
Uppercase chars	92 %	94 %	0 %	100 %	100 %	100 %	35 %	100 %	0 %	0 %	76 %	0 %	96 %	100 %	100 %	0 %	88 %
Lowercase chars	0 %	0 %	0 %	0%	0%	0%	35 %	0%	0 %	0 %	0 %	0 %	0 %	0%	0%	0 %	0 %

- **Tiempo**

Tabla V.54 Tiempo FADE_FASE_8

	FECGRA	FECING	FECNAC
Highest value	03/10/1999	01/06/1909	09/11/1987 0:00:00
Lowest value	14/07/1900	01/01/1900	01/01/1900 0:00:00

- **Patrones**

Tabla V.55 Patrones FADE_FASE_8

Columna	**Patrón**	**Cantidad**
CEDIDE	999999999-9	46

Tabla CESTUD$ Base de Datos FADE_FASE_9

- **Valores NULL**

Tabla V.56 Valores NULL FADE_FASE_9

Columna	Recuento de NULL	Porcentaje de NULL
APEEST	1	2%
BACHIL	2	4%
CEDIDE	1	2%
CERMIL	46	94%
CODEST	1	2%
CODPRO	1	2%
COLGRA	2	4%
DIRCUR	17	35%
DIRPER	7	14%
DOCUME	49	100%
ECUEST	1	2%
ESTCIV	1	2%
FECGRA	8	16%
FECING	1	2%
FECNAC	2	4%
LUGNAC	2	4%
NOMEST	1	2%
NOMPAD	11	22%
NUMTEF	7	14%
SEXO	1	2%

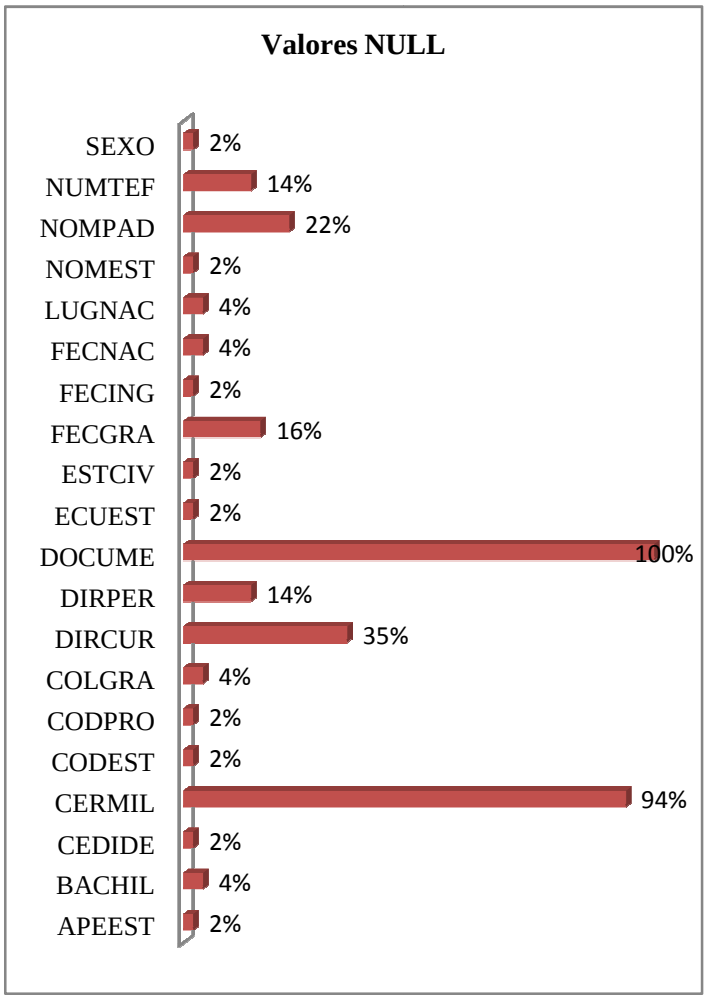


Figura V.50 Valores NULL FADE_FASE_9

- **Valores Vacíos**

Tabla V.57 Valores vacíos FADE_FASE_9

	NUMTEF	DIRCUR	CERMIL	FECING	SEXO	DOCUME	CODEST	BACHIL	ECUEST	NOMPAD	FECNAC	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	DIRPER	COLGRA	FECGRA
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de longitud de columnas**

Tabla V.58 Distribución de longitud de columnas FADE_FASE_9

Columna	Longitud mínima	Longitud máxima
APEEST	9	18
BACHIL	8	25
CEDIDE	11	11
CERMIL	10	10
CODEST	6	6
CODPRO	3	3
COLGRA	8	25
DIRCUR	7	39
DIRPER	7	50
ECUEST	1	1
ESTCIV	1	1
LUGNAC	4	15
NOMEST	5	19
NOMPAD	10	20
NUMTEF	7	9
SEXO	1	1

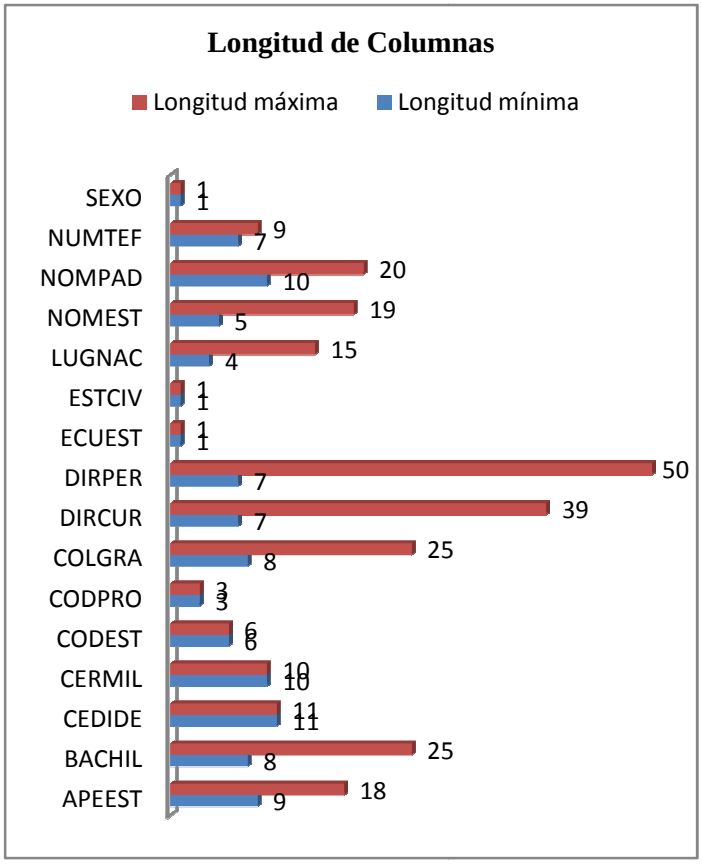


Figura V.51 Distribución de longitud de columnas FADE_FASE_9

- **Distribución de Valores de columna**

Tabla V.59 Distribución de valores de columna FADE_FASE_9

Columna	Número de valores distintos
APEEST	48
BACHIL	14
CEDIDE	47
CERMIL	3
CODEST	48
CODPRO	8
COLGRA	33
DIRCUR	6
DIRPER	40
DOCUME	0
ECUEST	1
ESTCIV	2
FECGRA	33
FECING	4
FECNAC	46
LUGNAC	21
NOMEST	47
NOMPAD	38
NUMTEF	40
SEXO	2

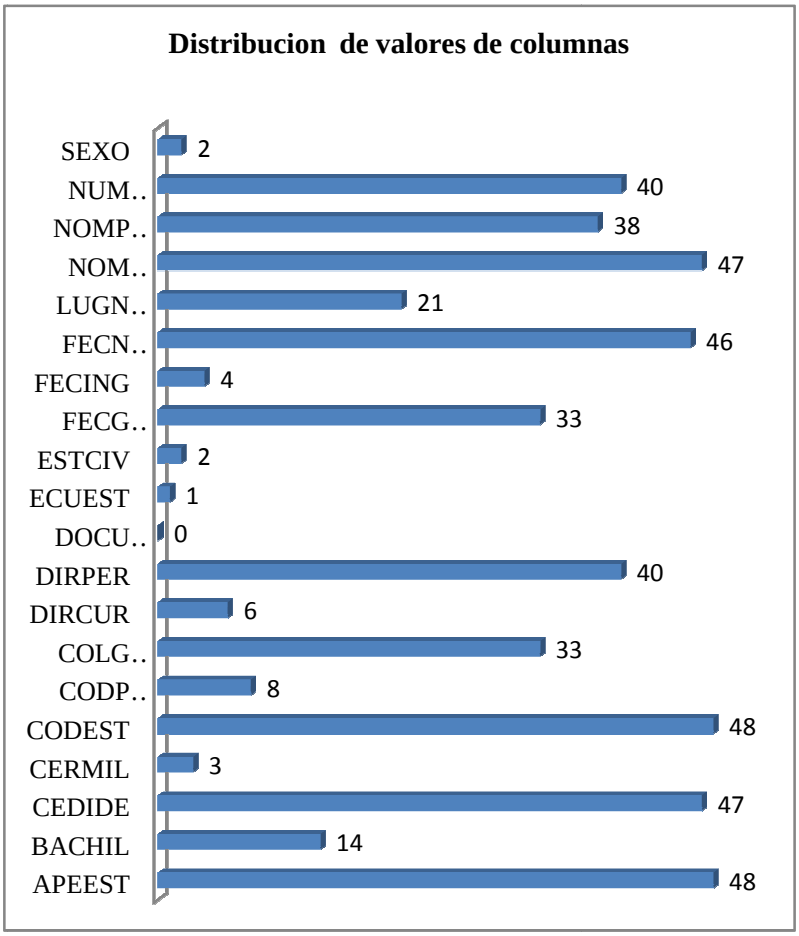


Figura V.52 Distribución de valores de columna FADE_FASE_9

- **Duplicación**

Tabla V.60 Duplicación FADE_FASE_9

Columnas de clave	Nivel de clave	Duplicación
APEEST	100%	0%
BACHIL	31%	69%
CEDIDE	98%	2%
CERMIL	8%	92%
CODEST	100%	0%
CODPRO	18%	82%
COLGRA	69%	31%
DIRCUR	14%	86%
DIRPER	84%	16%
ESTCIV	6%	94%
FECGRA	69%	31%
FECING	10%	90%
FECNAC	96%	4%
LUGNAC	45%	55%
NOMEST	98%	2%
NOMPAD	80%	20%
NUMTEF	84%	16%
SEXO	6%	94%

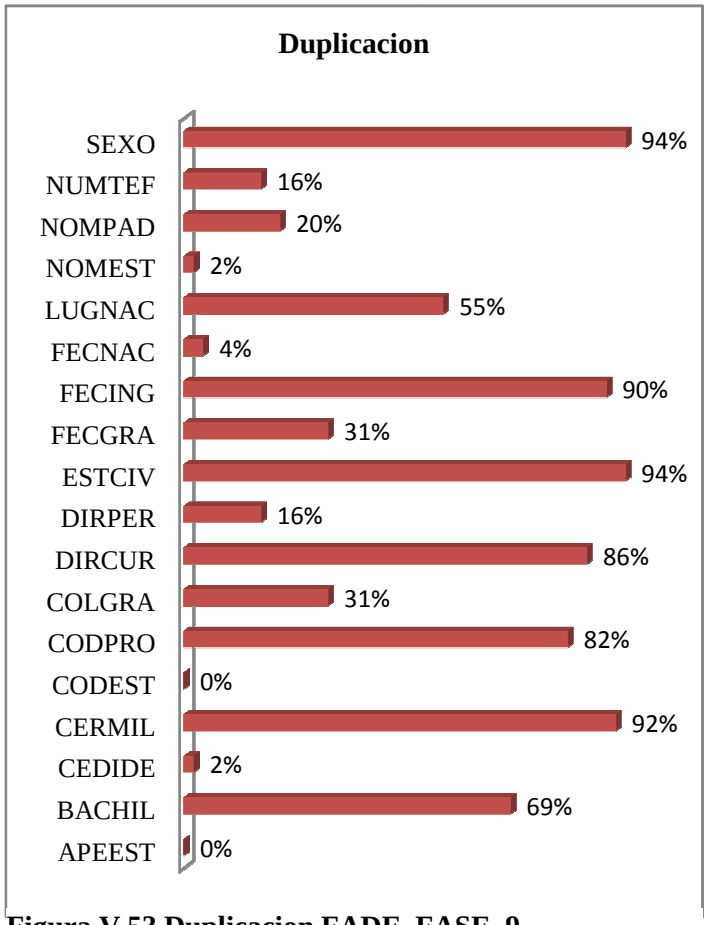


Figura V.53 Duplicacion FADE_FASE_9

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.61 Mayúsculas y minúsculas FADE_FASE_9

	NUMTEF	DIRCUR	CERMIL	SEXO	DOCUME	CODEST	BACHIL	ECUEST	NOMPAD	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	DIRPER	COLGRA
Uppercase chars	0 %	84 %	0 %	100 %	0 %	0 %	92 %	100 %	10 0	0 %	100 %	10 0	92 %	100 %	100 %	74 %	82 %
Lowercase chars	0 %	0%	0 %	0%	0 %	0 %	0%	0%	0 %	0 %	0%	0 %	0%	0%	0%	1%	0%

- **Tiempo**

Tabla V.62 Tiempo FADE_FASE_9

	FECING	FECNAC	FECGRA
Highest value	11/07/1908	09/02/1987	01/10/1999
Lowest value	23/03/1907	22/06/1954	31/08/1900

- **Patrones**

Tabla V.63 Patrones FADE_FASE_9

Columna	Patrón	Cantidad
CEDIDE	999999999-9	48

Tabla CESTUD -Base de Datos FADE_FASE_10

- **Valores NULL**

Tabla V.64 Valores NULL FADE_FASE_10

APEEST	0%	0
BACHIL	0,019607843	1
CEDIDE	0%	0
CERMIL	98%	50
CODEST	0%	0
CODPRO	0%	0
COLGRA	0%	0
DIRCUR	100%	51
DIRPER	0,019607843	1
DOCUME	100%	51
ECUEST	0%	0
ESTCIV	0%	0
FECGRA	14%	7
FECING	0%	0
FECNAC	0,019607843	1
LUGNAC	0%	0
NOMEST	0%	0
NOMPAD	100%	51

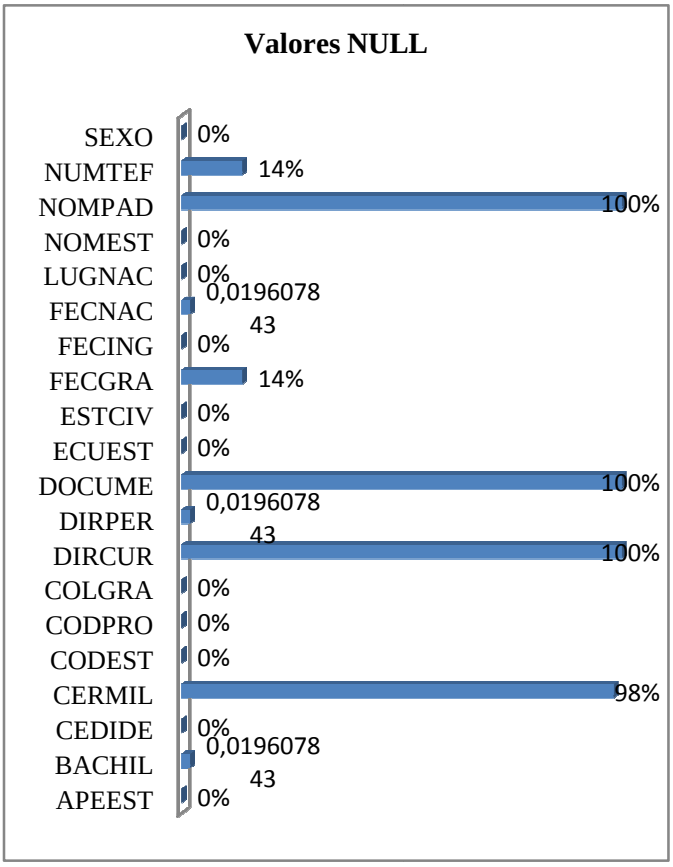


Figura V.54 Valores NULL FADE_FASE_10

NUMTEF	14%	7
SEXO	0%	0

- **Valores Vacíos**

Tabla V.65 Valores vacíos FADE_FASE_10

	DIRCUR	COLGRA	ECUEST	FECNAC	BACHIL	NUMTEF	DOCUME	CODEST	FECGRA	APEEST	ESTCIV	CERMIL	DIRPER	CODPRO	NOMEST	SEXO	FECING	CEDIDE	NOMPAD	LUGNAC
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de Longitud de Columnas**

Tabla V.66 Distribución de longitud de columnas FADE

Columna	Longitud mínima	Longitud máxima
APEEST	10	17
BACHIL	8	25
CEDIDE	11	11
CERMIL	11	11
CODEST	6	6
CODPRO	3	3
COLGRA	8	25
DIRPER	5	50
ECUEST	1	1
ESTCIV	1	1
LUGNAC	5	25
NOMEST	7	19
NUMTEF	7	10
SEXO	1	1

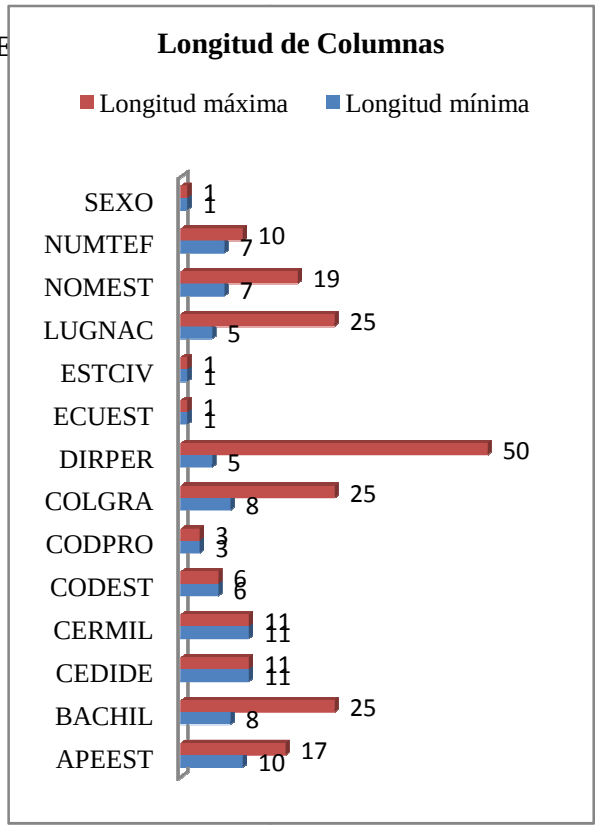


Figura V.55 Distribución de longitud de columnas

- **Distribución de valores de columna**

Tabla V.67 Distribución de valores d

Columna	Número de valores distintos
APEEST	48
BACHIL	12
CEDIDE	51
CERMIL	1
CODEST	51
CODPRO	1
COLGRA	24
DIRCUR	0
DIRPER	49
DOCUME	0
ECUEST	1
ESTCIV	3
FECGRA	39
FECING	10
FECNAC	50
LUGNAC	19
NOMEST	50
NOMPAD	0
NUMTEF	43
SEXO	2

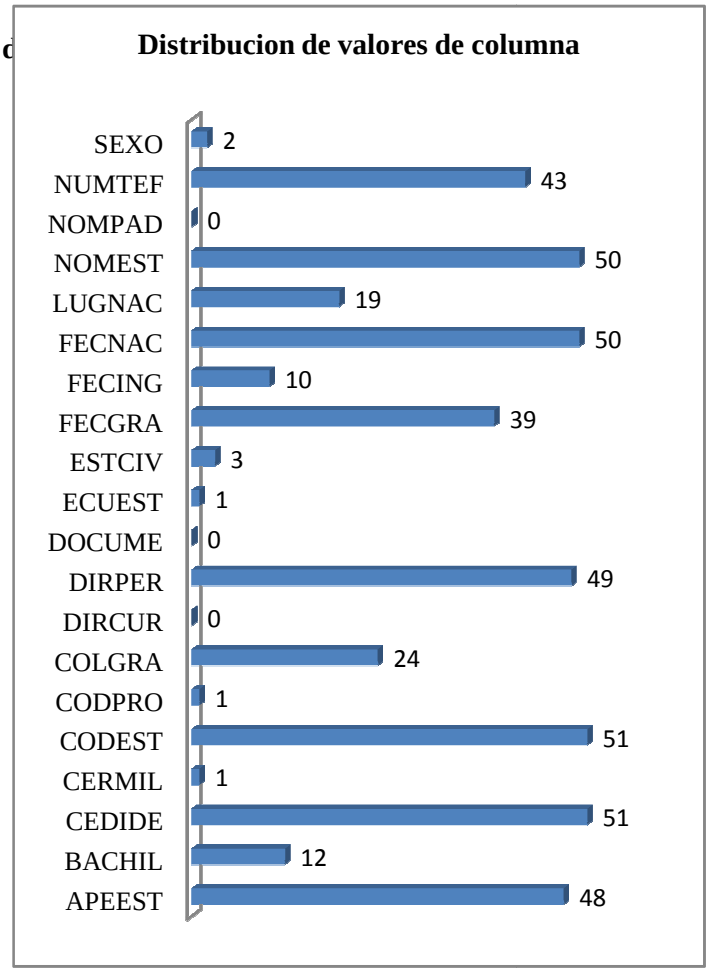


Figura V.56Distribucion de valores de columna FADE_FASE_10

- **Duplicación**

Tabla V.68 Duplicación FADE_FASE_10

Columnas de clave	Nivel de clave	Duplicación
APEEST	94%	6%
BACHIL	25%	75%
CEDIDE	100%	0%
CODEST	100%	0%
COLGRA	47%	53
DIRPER	98%	2%
ESTCIV	6%	94%
FECGRA	78%	22%
FECING	20%	80%
FECNAC	100%	0%
LUGNAC	37%	
NOMEST	98%	2%
NUMTEF	86%	14

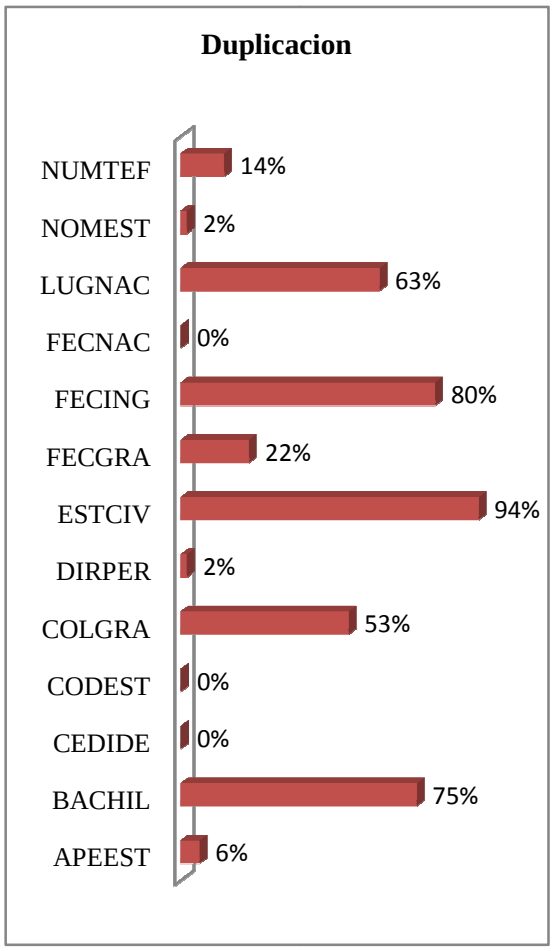


Figura V.57 Duplicación FADE_FASE_10

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.69 Mayúsculas y minúsculas FADE_FASE_10

	DIRCUR	COLGRA	ECUEST	BACHIL	NUMTEF	DOCUME	CODEST	APEEST	ESTCIV	CERMIL	DIRPER	CODPRO	NOMEST	SEXO	CEDIDE	NOMPAD	LUGNAC
Uppercase chars	0 %	87 %	100 %	94 %	0 %	0 %	0 %	100 %	100 %	0 %	67 %	100 %	100 %	100 %	0 %	0 %	97 %
Lowercase chars	0 %	0%	0%	0%	0 %	0 %	0 %	0%	0%	0 %	8%	0%	0%	0%	0 %	0 %	0%

- **Tiempo**

Tabla V.70 Tiempo FADE_FASE_10

	FECNAC	FECGRA	FECING
Highest value	18/09/1989	07/08/1999	26/10/1927
Lowest value	25/04/1964	22/09/1900	15/10/1904

- **Patrones**

Tabla V.71 Patrones FADE_FASE_10

COLUMNA	PATRON	CANTIDAD
CEDIDE	999999999-9	51

Tabla CESTUD- Base de Datos FADE_GGSBA

- **Valores NULL**

Tabla V.72 Valores NULL FADE_GGSBA

Columna	Recuento de NULL	Porcentaje de NULL
APEEST	1	3%
BACHIL	1	3%
CEDIDE	1	3%
CERMIL	12	32%
CODEST	**1**	**3%**
CODPRO	1	3%
COLGRA	1	3%
DIRCUR	33	89%
DIRPER	5	14%
DOCUME	37	100%
ECUEST	1	3%
ESTCIV	1	3%
FECGRA	3	8%
FECING	1	3%
FECNAC	1	3%
LUGNAC	1	3%
NOMEST	1	3%
NOMPAD	37	100%
NUMTEF	9	24%

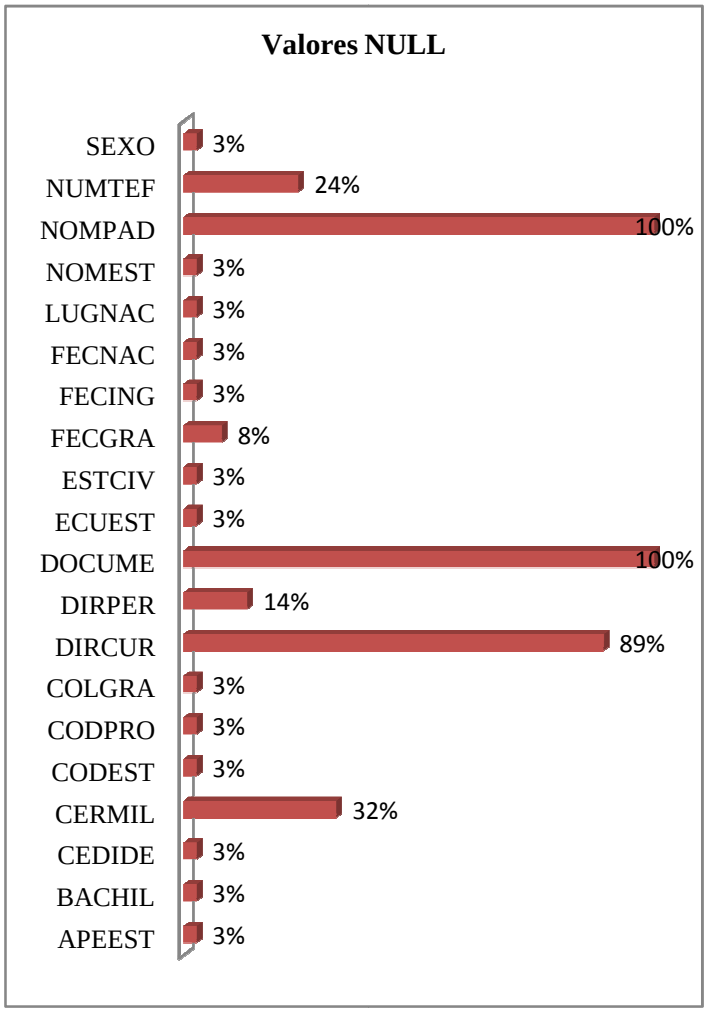


Figura V.58 Valores NULL FADE_FASE_GGSBA

SEXO	1	3%

- **Valores Vacíos**

Tabla V.73Valores vacíos FADE_FASE_GGSBA

	NUMTEF	DIRCUR	CERMIL	FECING	SEXO	DOCUME	CODEST	BACHIL	ECUEST	NOMPAD	FECNAC	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	DIRPER	COLGRA	FECGRA
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de longitud de columnas**

Tabla V.74 Distribucion de longitud de columnas FADE_FASE_GGSBA

Columna	Longitud mínima	Longitud máxima
APEEST	9	18
BACHIL	8	25
CEDIDE	10	11
CERMIL	10	11
CODEST	6	6
CODPRO	3	3
COLGRA	10	25
DIRCUR	16	26
DIRPER	7	43
ECUEST	1	1
ESTCIV	1	1
LUGNAC	5	18
NOMEST	4	16
NUMTEF	6	6
SEXO	1	1

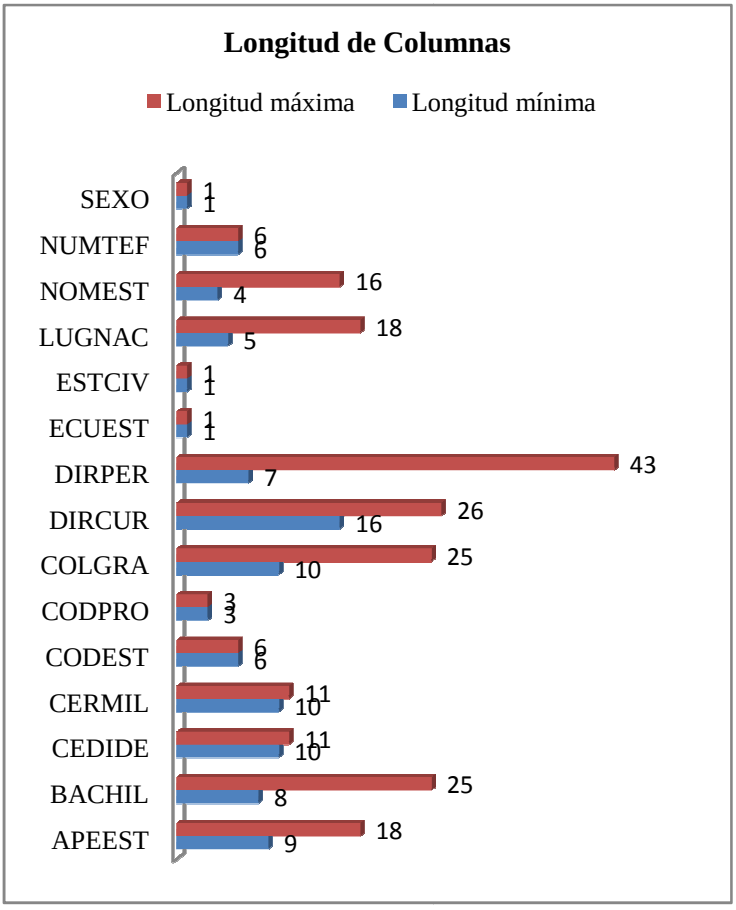


Figura V.59 Distribucion de longitud de columna FADE_FASE_GGSBA

- **Distribución de valores de columna**

Tabla V.75 Distribución de va

Columna	Número de valores distintos
APEEST	36
BACHIL	14
CEDIDE	36
CERMIL	25
CODEST	36
CODPRO	2
COLGRA	30
DIRCUR	4
DIRPER	32
DOCUME	0
ECUEST	1
ESTCIV	3
FECGRA	32
FECING	27
FECNAC	36
LUGNAC	12
NOMEST	34
NOMPAD	0
NUMTEF	28
SEXO	2

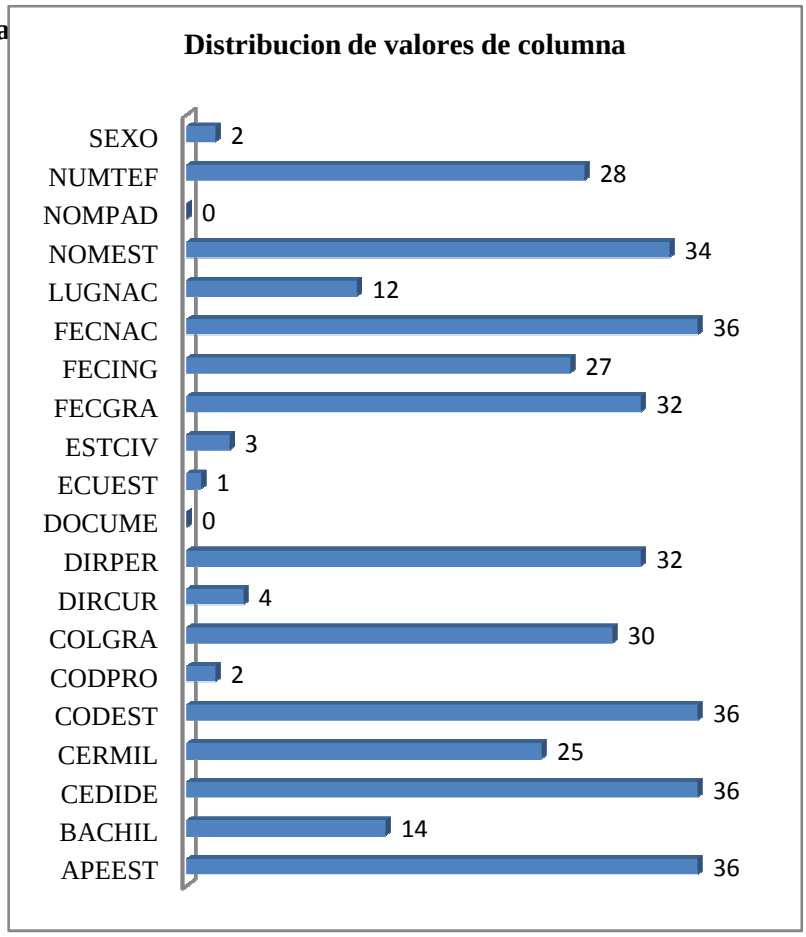


Figura V.60 Distribución de valores de columna

- **Duplicación**

Tabla V.76 Duplicación FADE_FASE_GGSBA

Columnas de clave	Nivel de clave	Duplicación
APEEST	100%	0%
BACHIL	41%	59%
CEDIDE	100%	0%
CERMIL	70%	30%
CODEST	100%	0%
CODPRO	8%	92%
COLGRA	84%	16%
DIRCUR	14%	86%
DIRPER	89%	11%
ECUEST	5%	95%
ESTCIV	11%	89%
FECGRA	89%	11%
FECING	76%	24%
FECNAC	100%	0%
LUGNAC	35%	65%
NOMEST	95%	5%
NUMTEF	78%	22%
SEXO	8%	92%

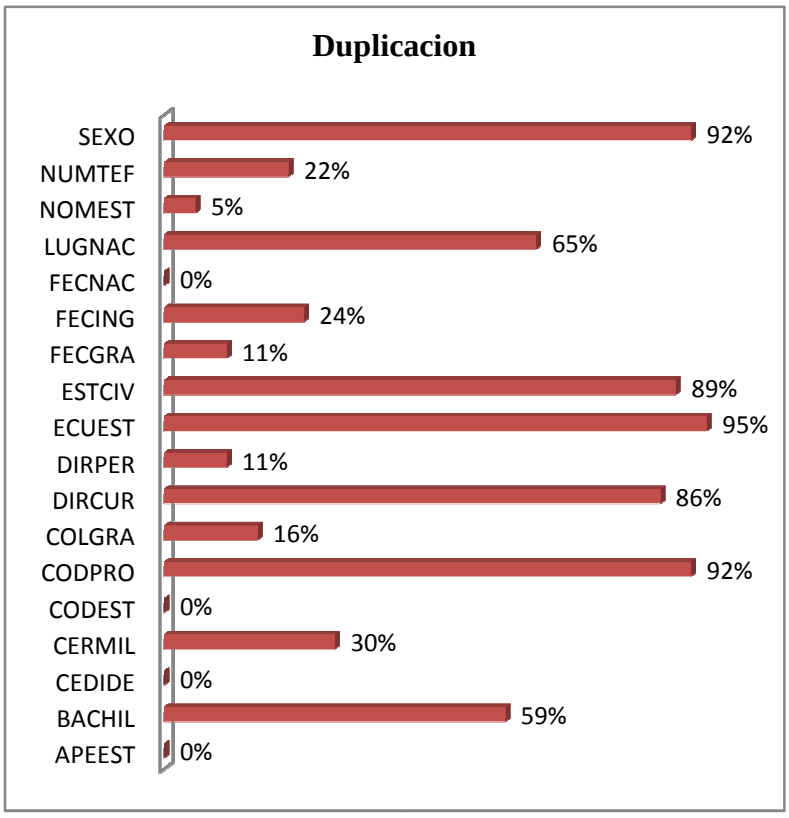


Figura V.61 Duplicacion FADE_FASE_GGSBA

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.77 Mayúsculas y minúsculas FADE_FASE_GGSBA

	NUMTEF	DIRCUR	CERMIL	SEXO	DOCUME	CODEST	BACHIL	ECUEST	NOMPAD	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	DIRPER	COLGRA
Uppercase chars	0 %	84 %	0 %	100 %	0 %	0 %	92 %	100 %	100 %	0 %	100 %	100 %	100 %	98 %	100 %	74 %	82 %
Lowercase chars	0 %	0 %	0 %	0%	0 %	0 %	0 %	0%	0%	0 %	0%	0%	0%	0 %	0%	1 %	0 %

- **Tiempo**

Tabla V.78 Tiempo FADE_FASE_GGSBA

	FECNAC	FECGRA	FECING
Highest value	16/07/2005	01/10/2005	02/08/2006
Lowest value	30/01/1957	21/09/1976	28/06/2005

- **Patrones**

Tabla V.79 Patrones FADE_FASE_GGSBA

Columna	Patrón	Cantidad
CEDIDE	9999999999	24
	999999999-9	12

Tabla CESTUD -Base de Datos FADE_GGSES

- **Valores NULL**

Tabla V.80 Valores NULL FADE_FASE_GGSES

Columna	Recuento de NULL	Porcentaje de NULL
APEEST	0	0%
BACHIL	1	5%
CEDIDE	0	0%
CERMIL	11	50%
CODEST	0	0%
CODPRO	0	0%
COLGRA	1	5%
DIRCUR	21	95%
DIRPER	3	14%
DOCUME	22	100%
ECUEST	0	0%
ESTCIV	0	0%
FECGRA	7	32%
FECING	0	0%
FECNAC	0	0%
LUGNAC	0	0%
NOMEST	0	0%
NOMPAD	22	100%
NUMTEF	7	32%
SEXO	0	0%

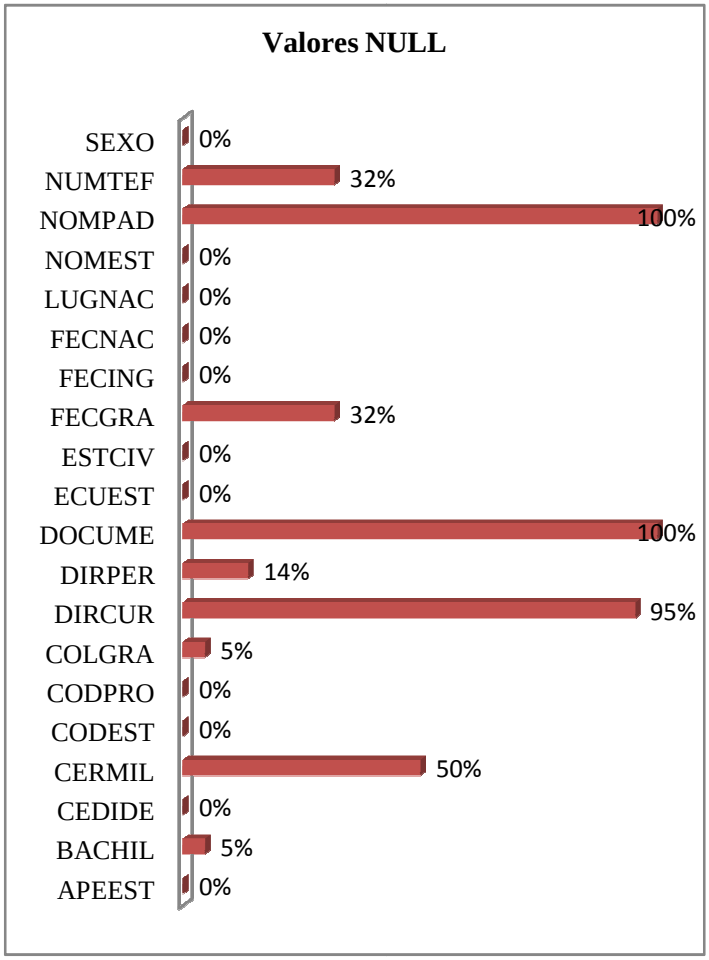


Figura V.62 Valores NULL FADE_FASE_ GGSES

- **Valores Vacíos**

Tabla V.81 Valores Vacíos FADE_FASE_GGSES

	NUMTEF	DIRCUR	CERMIL	FECING	SEXO	DOCUME	CODEST	ECUEST	NOMPAD	FECNAC	CEDIDE	FECGRA	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	BACHIL	DIRPER	COLGRA
Empty values	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de Longitud de Columnas**

Tabla V.82 Distribución de longitud de columnas FADE_FASE_GGSES

Columna	Longitud mínima	Longitud máxima
APEEST	11	19
BACHIL	8	25
CEDIDE	11	11
CERMIL	1	11
CODEST	6	6
CODPRO	3	3
COLGRA	13	25
DIRCUR	30	30
DIRPER	5	35
ECUEST	1	1
ESTCIV	1	1
LUGNAC	5	10
NOMEST	9	17
NUMTEF	6	6
SEXO	1	1

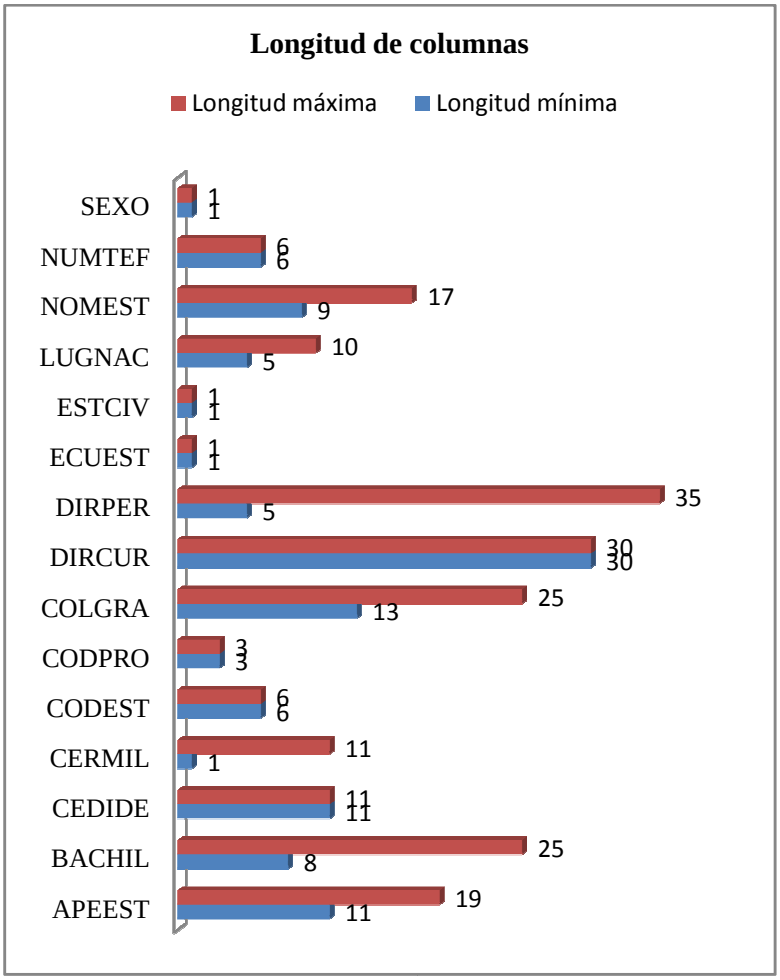


Figura V.63 Distribución de longitud de columnas GGSES

- **Distribución de Valores de columna**

Tabla V.83 Distribucion de valores de columna FADE_FASE_GGSES

Columna	Número de valores distintos
APEEST	22
BACHIL	10
CEDIDE	22
CERMIL	11
CODEST	22
CODPRO	3
COLGRA	21
DIRCUR	1
DIRPER	18
DOCUME	0
ECUEST	1
ESTCIV	2
FECGRA	15
FECING	3
FECNAC	22
LUGNAC	12
NOMEST	22
NOMPAD	0
NUMTEF	15
SEXO	2

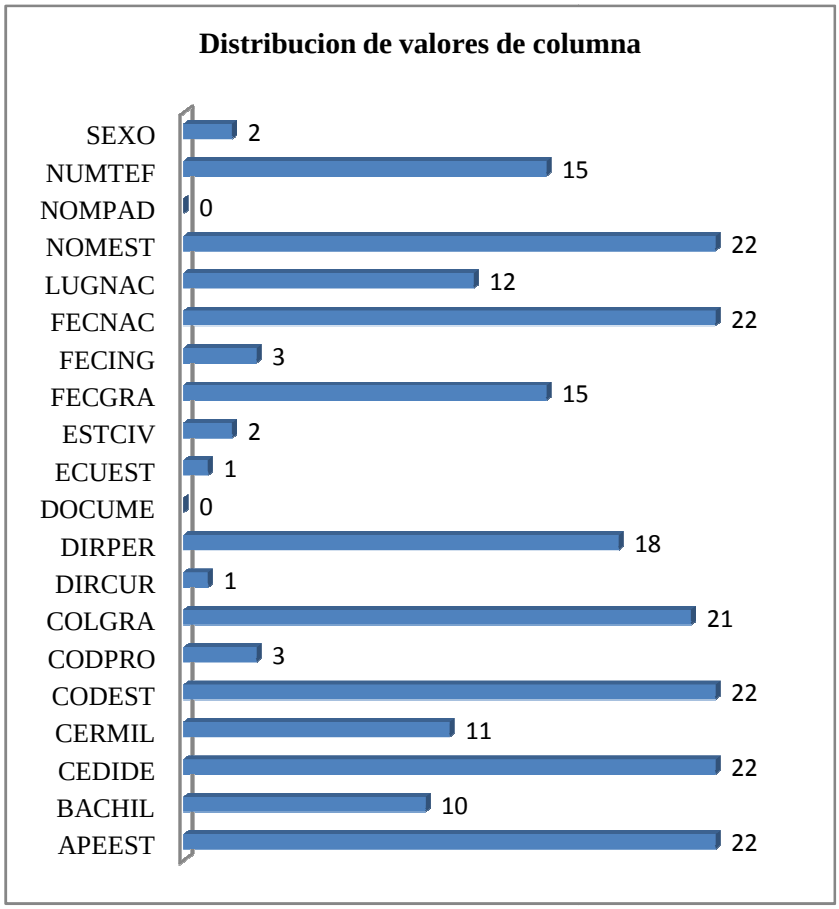
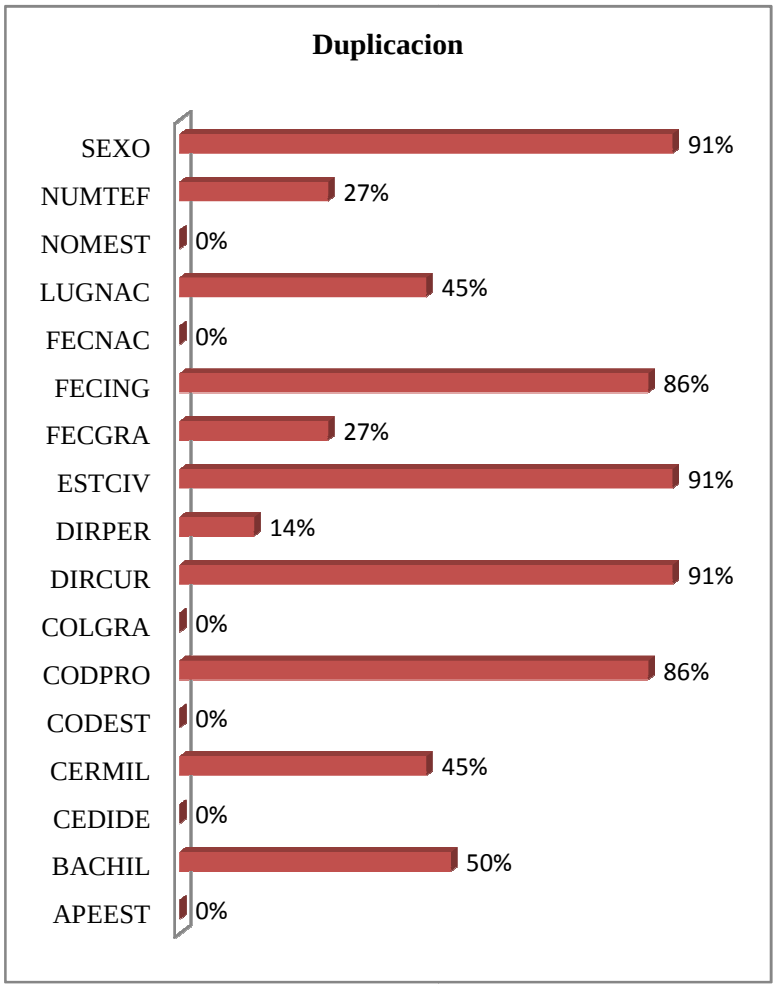


Figura V.64 Distribución de valores de columna FADE_FASE_GGSES

- **Duplicación**

Tabla V.84 Duplicación FADE_FASE_GGSES

Columnas de clave	Nivel de clave	Duplicación
APEEST	100%	0%
BACHIL	50%	50%
CEDIDE	100%	0%
CERMIL	55%	45%
CODEST	100%	0%
CODPRO	14%	86%
COLGRA	100%	0%
DIRCUR	9%	91%
DIRPER	86%	14%
ESTCIV	9%	91%
FECGRA	73%	27%
FECING	14%	86%
FECNAC	100%	0%
LUGNAC	55%	45%
NOMEST	100%	0%
NUMTEF	73%	27%
SEXO	9%	91%



FiguraV.65 Duplicación FADE_FASE_GGSES

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.85 Mayúsculas y minúsculas FADE_FASE_GGSES

	NUMTEF	DIRCUR	CERMIL	SEXO	E	CODEST	ECUEST	D	CEDIDE	CODPRO	APEEST	NOMEST	LUGNAC	ESTCIV	BACHIL	DIRPER	COLGRA
Uppercase chars	0 %	6 %	0 %	100 %	0 %	0 %	100 %	0 %	0 %	100 %	100 %	100 %	100 %	100 %	94 %	16 %	85 %
Lowercase chars	0 %	76 %	0 %	0%	0 %	0 %	0%	0 %	0 %	0%	0%	0%	0%	0%	0 %	66 %	0 %

- **Tiempo**

Tabla V.86 Tiempo FADE_FASE_GGSES

	FECING	FECNAC	FECGRA
Highest value	29/11/2007	12/10/1988	15/11/2006
Lowest value	01/10/2006	06/06/1954	29/07/1977

- **Patrones**

Tabla V.87 Patrones FADE_FASE_GGSES

Columna	Patrón	Cantidad
CEDIDE	999999999-9	22

Análisis de la Base de datos del Sistema Académico

Tabla Estudiantes Base de datos OAS_CicloFormativo_db

- **Valores NULL**

Tabla V.88 Valores NULL Ciclo Formativo

Columna	Recuento de NULL	Porcentaje de NULL
dtFechaIng	1	0,03%
dtFechaNac	0	0%
strApellidos	0	0%
strCedula	0	0%
strCedulaMil	735	23%
strCodigo	0	0%
strCodInt	978	30%
strCodSexo	0	0%
strCodTit	978	30%
strDocumentacion	3219	100%
strEmail	1072	33%
strFormaIns	0	0%
strNacionalidad	2	0,06%
strNombres	0	0%

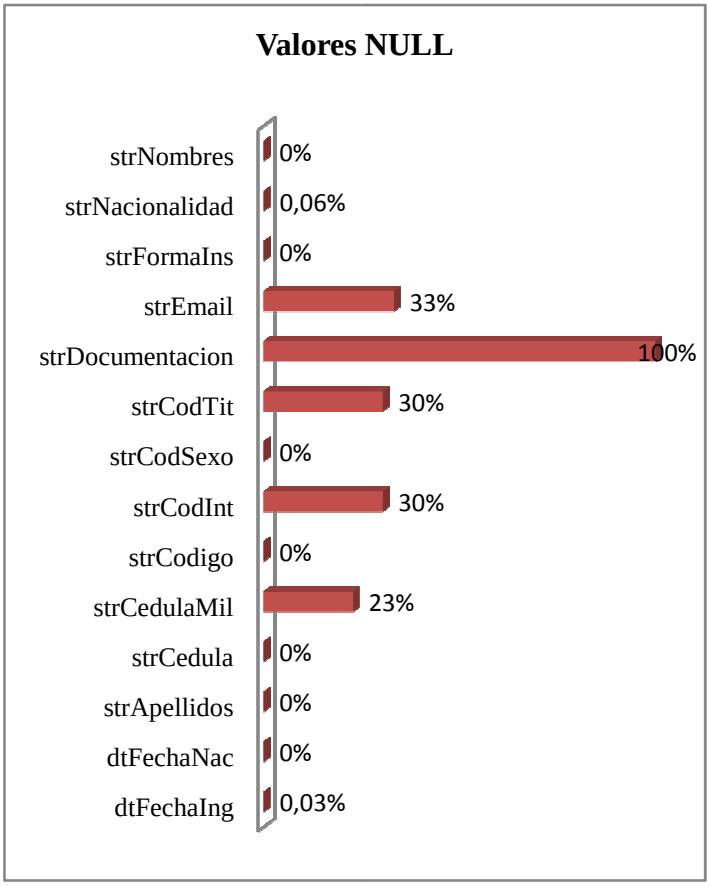


Figura V.66 Valores NULL Ciclo Formativo

- **Valores Vacíos**

Tabla V.89 Valores vacios Ciclo Formativo

	dtFechaNac	strCodTit	strCedulaMil	strDocumentacion	strNombres	trCodInt	strCodigo	strApellidos	dtFechaIng	strCodSexo	strFormaIns	strEmail	strCedula	strNacionalidad
Empty values	0	0	496	0	0	0	0	0	0	0	0	264	0	0

- **Distribución de longitud de columnas**

Tabla V.90 Distribución de longitud de columnas Ciclo Formativo

Columna	Longitud mínima	Longitud máxima
strApellidos	6	25
strCedula	11	11
strCedulaMil	0	13
strCodigo	5	5
strCodInt	2	10
strCodSexo	3	3
strCodTit	2	5
strEmail	0	38
strFormaIns	3	3
strNacionalidad	7	12
strNombres	4	25

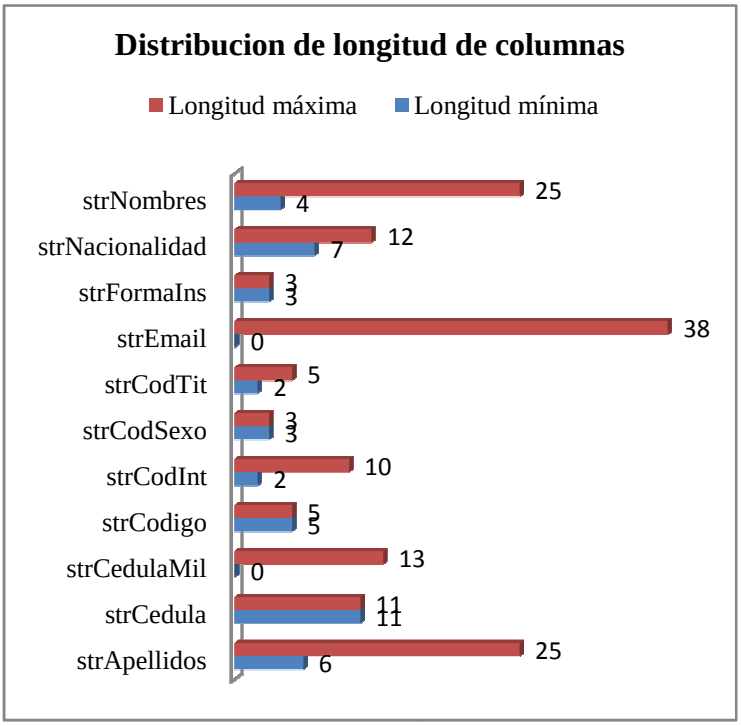


Figura V.67 Distribucion de longitud de columnas

Distribución de valores de columna

Tabla V.91 Distribución de valores de columna Ciclo Formativo

Columna	valores distintos
dtFechaIng	2375
dtFechaNac	2188
strApellidos	3011
strCedula	3219
strCedulaMil	994
strCodigo	3219
strCodInt	430
strCodSexo	2
strCodTit	48
strDocumentacion	0
strEmail	576
strFormaIns	3
strNacionalidad	19
strNombres	2564

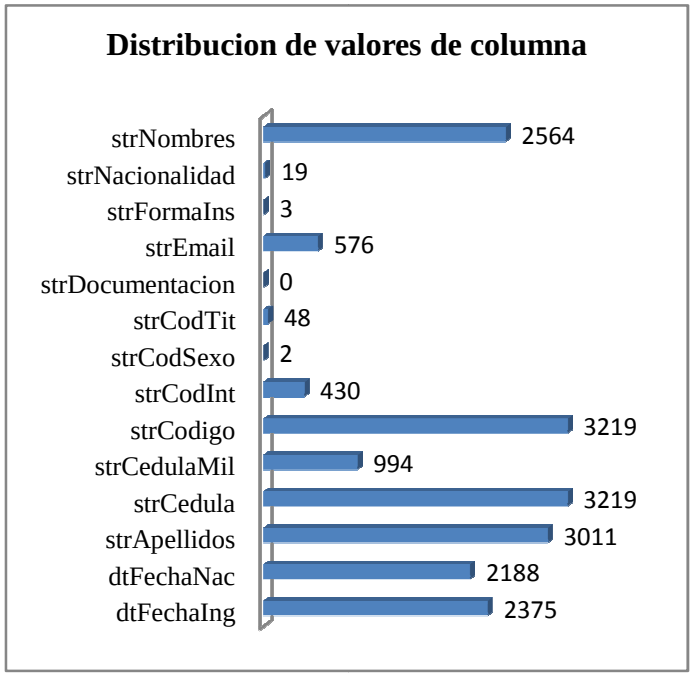


Figura V.68 Distribucion de valores de columna Ciclo Formativo

Duplicación

Tabla V.92 Duplicación Ciclo Formativo

Columnas de clave	Nivel de clave	Duplicación
dtFechaIng	74%	26%
dtFechaNac	68%	32%
strApellidos	94%	6%
strCedula	100%	0%
strCedulaMil	31%	69%
strCodigo	100%	0%
strCodInt	13%	87%
strEmail	18%	82%
strNombres	80%	20%

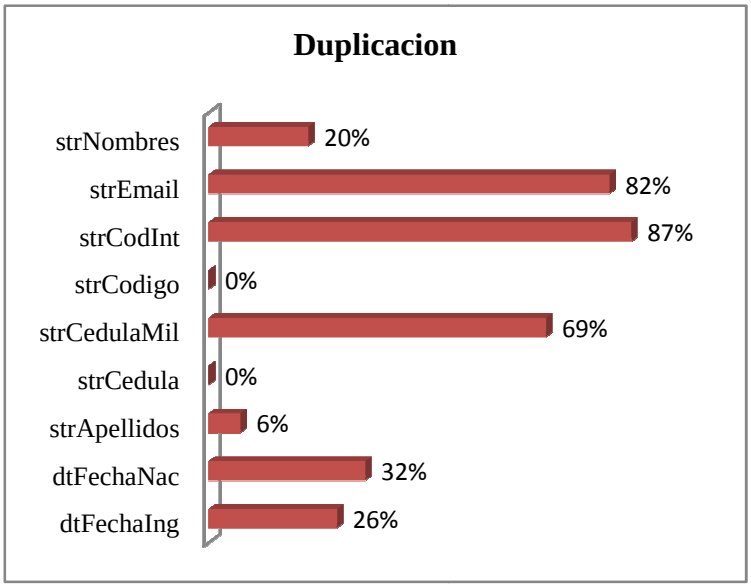


Figura V.69 Duplicacion Ciclo Formativo

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.93 Mayusculas y minusculas Ciclo Formativo

	strCodTit	strCedulaMil	strDocumentacion	strNombres	strCodInt	strCodigo	strApellidos	strCodSexo	strFormaIns	strEmail	strCedula	strNacionalidad
Uppercase chars	100%	0%	0%	97%	98%	0%	98%	100%	100%	1%	0%	97%
Lowercase chars	0%	0%	0%	3%	0%	0%	2%	0%	0%	78%	0%	1%

- **Tiempo**

Tabla V.94 Tiempo Ciclo Formativo

	dtFechaIng	dtFechaNac
Highest value	27/03/2009 15:34	13/03/2008
Lowest value	09/11/1988 0:00	1886-10-20 00:03:40

- **Patrones**

Tabla V.95 Patrones Ciclo Formativo

strCedula	
999999999-9	3219

Tabla Estudiantes Base de datos OAS_IngAgronomica

- **Valores NULL**

Tabla V.96 Valores NULL IngAgronomica

Columna	Recuento de NULL	Porcentaje de NULL
dtFechaIng	0	0%
dtFechaNac	3	0,43%
strApellidos	0	0%
strCedula	0	0%
strCedulaMil	99	14,08%
strCodigo	0	0%
strCodInt	409	58,18%
strCodSexo	0	0%
strCodTit	409	58,18%
strDocumentacion	703	100%
strEmail	251	35,70%
strFormaIns	0	0%
strNacionalidad	76	10,81%
strNombres	0	0%

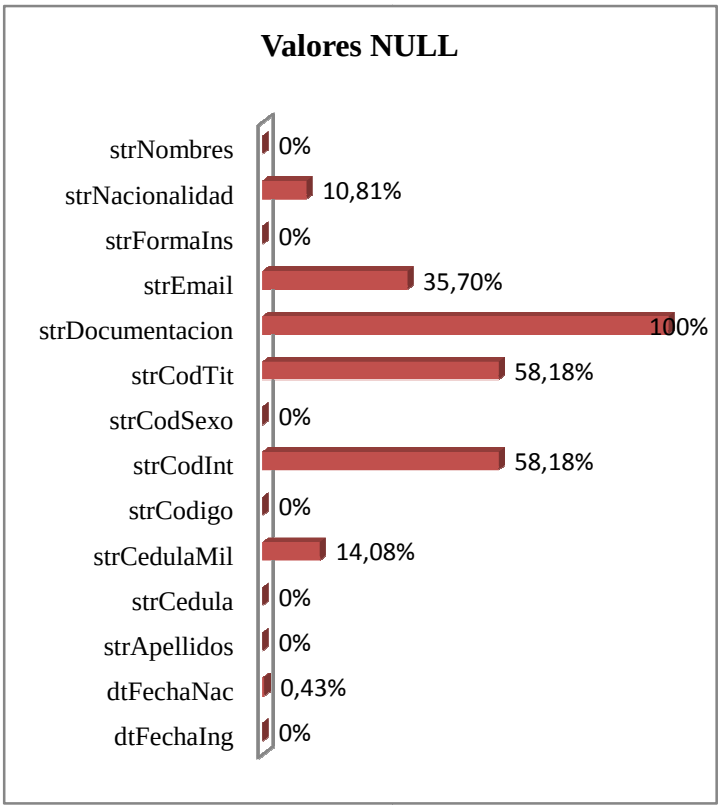


Figura V.70 Valores NULL IngAgronomica

- **Valores Vacíos**

Tabla V.97 Valores vacios IngAgronomica

	strCedulaMil	strEmail	strCedula	strCodInt	strFormaIns	strApellidos	dtFechaNac	strNacionalidad	strNombres	dtFechaIng	strCodTit	strCodigo	strDocumentacion	strCodSexo
Empty values	314	264	0	0	0	0	0	0	0	0	0	0	0	0

- **Distribución de longitud de columnas**

Tabla V.98 Distribución de longitud de columnas Ing Agronomica

Columna	Longitud mínima	Longitud máxima
strApellidos	5	21
strCedula	11	11
strCedulaMil	0	12
strCodigo	2	4
strCodInt	2	9
strCodSexo	3	3
strCodTit	2	5
strEmail	0	33
strFormaIns	3	3
strNacionalidad	10	12
strNombres	5	18

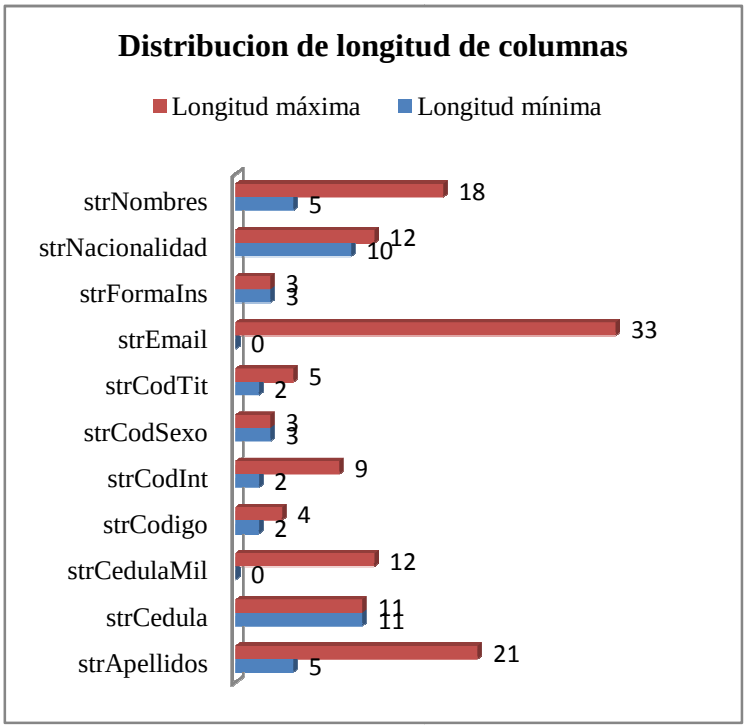


Figura V.71 Distribucion de longitud de columnas

- **Duplicación**

Tabla V.99 Duplicacion IngAgronomica

Columnas de clave	Nivel de clave	Duplicación
dtFechaIng	0,61024177	39%
dtFechaNac	0,93741113	6%
strApellidos	0,97866279	2%
strCedula	1	0%
strCedulaMil	0,41394031	59%
strCodigo	1	0%
strCodInt	0,2034139	80%
strEmail	0,27027029	73%
strNombres	0,88335699	12%

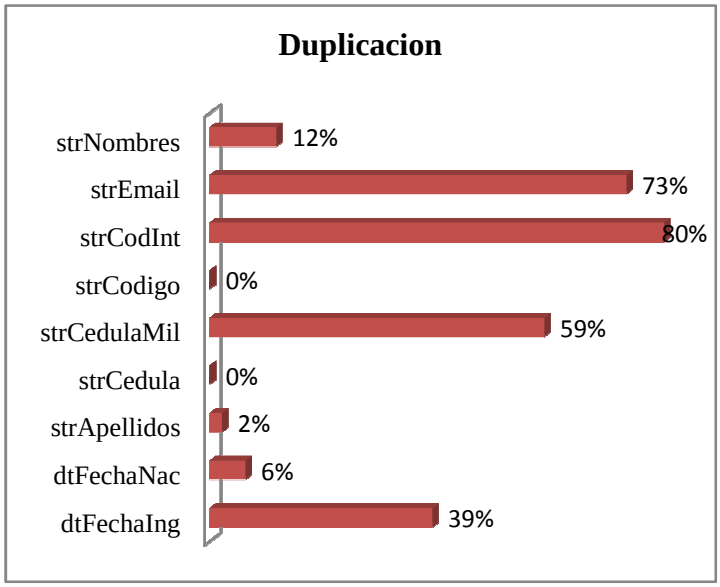


Figura V.72 Duplicacion IngAgronomica

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.100 Mayusculas y minusculas IngAgronomica

	strCedulaMil	strEmail	strCedula	strCodInt	strFormaIns	strApellidos	strNacionalidad	strNombres	strCodTit	strCodigo	strDocumentacion	strCodSexo
Uppercase chars	0%	2%	0%	99%	100%	99%	98%	99%	100%	0%	0%	100%
Lowercase chars	0%	78%	0%	0%	0%	1%	1%	1%	0%	0%	0%	0%

- **Tiempo**

Tabla V.101 Tiempo IngAgronomica

	dtFechaNac	dtFechaIng
Highest value	29/07/2006	22/09/2010 11:36:00
Lowest value	05/09/1903	20/11/1984 0:00:00

- **Patrones**

Tabla V.102 Patrones IngAgronomica

COLUMNA	PATRON	CANTIDAD
strCedula	999999999-9	314

Tabla Estudiantes Base de datos OAS_IngEmpresas_db

- **Valores NULL**

Tabla V.103 Valores NULL IngEmpresas

Columna	Recuento de NULL	Porcentaje de NULL
dtFechaIng	0	0
dtFechaNac	3	0,00214
strApellidos	0	0
strCedula	0	0
strCedulaMil	290	0,20685
strCodigo	0	0
strCodInt	595	0,42439
strCodSexo	0	0
strCodTit	595	0,42439
strDocumentacion	1402	1
strEmail	436	0,31098
strFormaIns	0	0

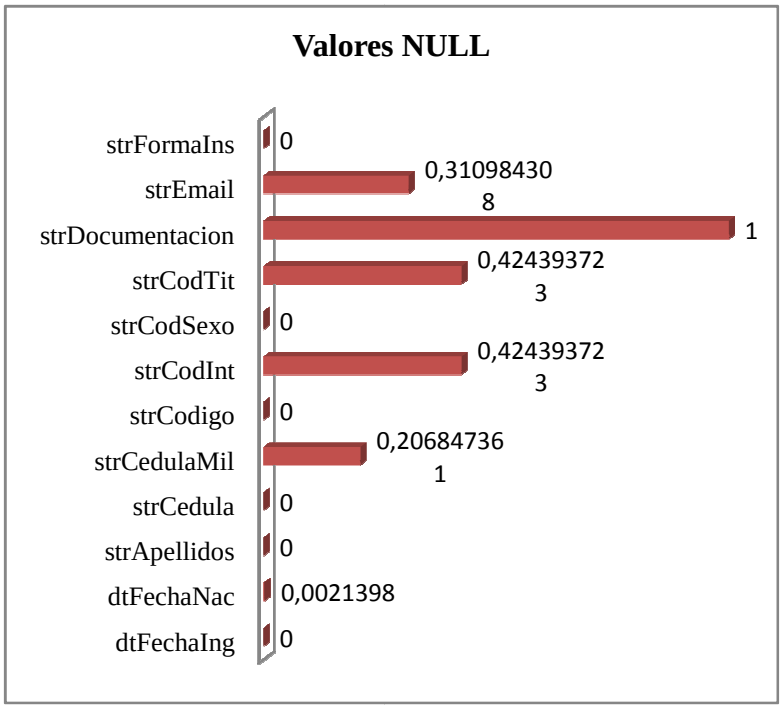


Figura V.73 Valores NULL IngEmpresas

- **Valores Vacíos**

Tabla V.104 Valores Vacios IngEmpresas

	strCodSexo	strEmail	strCedulaMil	strCodTit	dtFechaIng	dtFechaNac	strApellidos	strCodInt	strNombres	dad	acion	strFormaIns	strCodigo
Empty values	0	599	777	0	0	0	0	0	0	0	0	0	0

- **Distribución de valores de columna**

Tabla V.105 Distribución de valores de columna IngEmpresas

Columna	Número de valores distintos
dtFechaIng	699
dtFechaNac	1227
strApellidos	1349
strCedula	1402
strCedulaMil	335
strCodigo	1402
strCodInt	260
strCodSexo	2
strCodTit	39
strDocumentacion	0
strEmail	366
strFormaIns	3
strNacionalidad	11
strNombres	1227

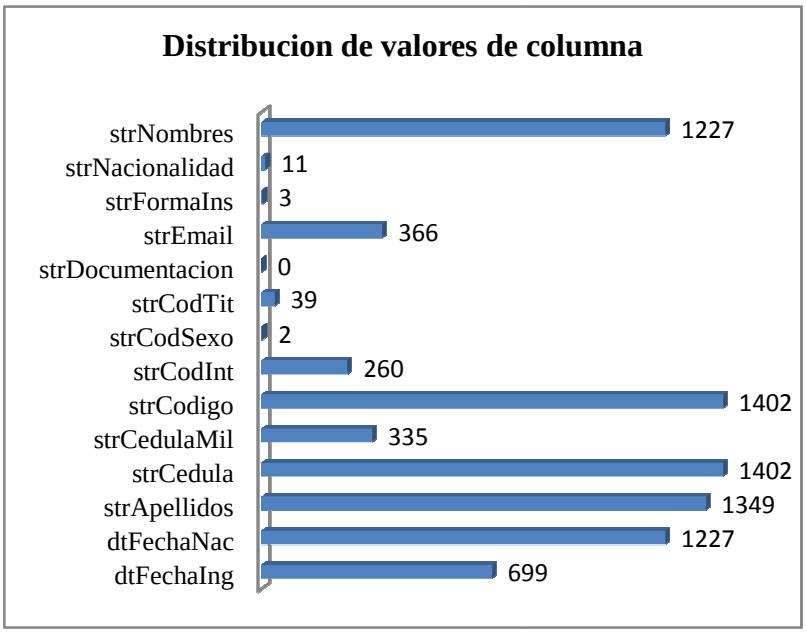


Figura V.74 Distribucion de valores de columna IngEmpresas

- **Distribución de longitud de columnas**

Tabla V.106Distribucion de longitud de columnas IngEmpresas

Columna	Longitud mínima	Longitud máxima
strApellidos	6	23
strCedula	11	11
strCedulaMil	0	13
strCodigo	5	5
strCodInt	3	10
strCodSexo	3	3
strCodTit	2	5
strEmail	0	38
strFormaIns	3	3
strNacionalidad	7	12
strNombres	4	25

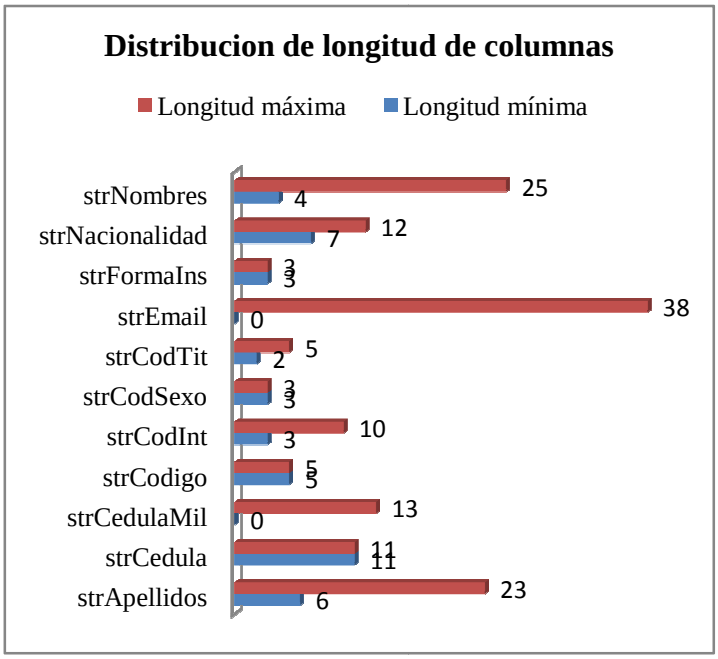


Figura V.75 Distribucion de longitud de columnas

- **Duplicación**

Tabla V.107 Duplicación IngEmpresas

Columnas de clave	Nivel de clave	Duplicación
dtFechaIng	50%	50%
dtFechaNac	88%	12%
strApellidos	96%	4%
strCedula	100%	0%
strCedulaMil	24%	76%
strCodigo	100%	0%
strCodInt	19%	81%
strEmail	26%	74%
strNombres	88%	12%

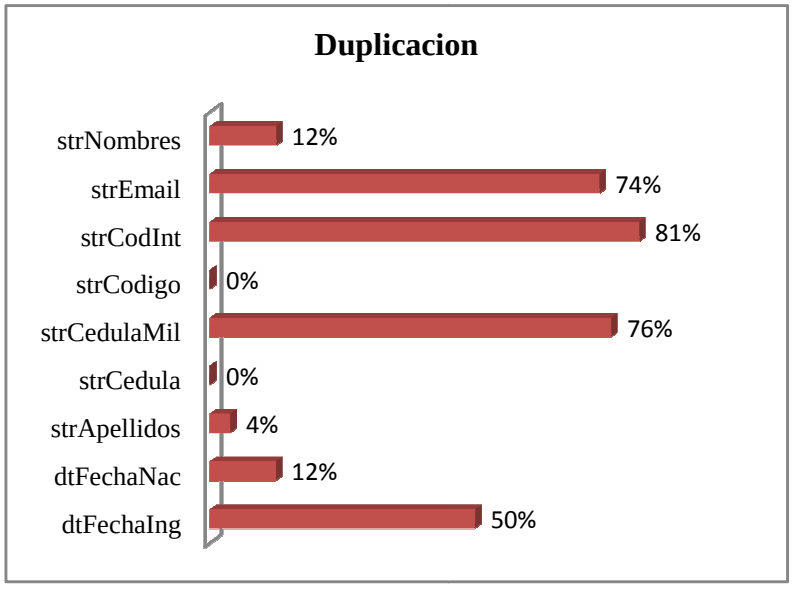


Figura V.76 Duplicacion IngEmpresas

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.108 Mayusculas y minusculas IngEmpresas

	strCodSexo	strEmail	strCedulaMil	strCodTit	strCedula	strApellidos	strNombres	strNacionalidad	strDocumentacion	strFormaIns	strCodigo
Uppercase chars	100%	1%	0%	99%	0%	89%	89%	97%	0%	100%	0%
Lowercase chars	0%	80%	0%	0%	0%	3%	3%	2%	0%	0%	0%

- **Tiempo**

Tabla V.109 Tiempo IngEmpresas

	dtFechaIng	dtFechaNac
Highest value	15/09/2010 17:21:51	13/03/2008
Lowest value	20/10/1988 0:00:00	16/05/1905

- **Patrones**

Tabla V.110 Patrones IngEmpresas

strCedula	
999999999-9	1402

Análisis Estadístico de la Tabla Estudiantes Base de datos OAS_NatPromSalud_db

- **Valores NULL**

Tabla V.111 Valores NULL NatPromSalud

Columna	Recuento de NULL	Porcentaje de NULL
dtFechaIng	0	0%
dtFechaNac	0	0%
strApellidos	0	0%
strCedula	0	0%
strCedulaMil	50	11%
strCodigo	0	0%
strCodInt	130	29%
strCodSexo	0	0%
strCodTit	130	29%
strDocumentacion	454	100%
strEmail	177	39%
strFormaIns	0	0%
strNacionalidad	112	25%
strNombres	0	0%

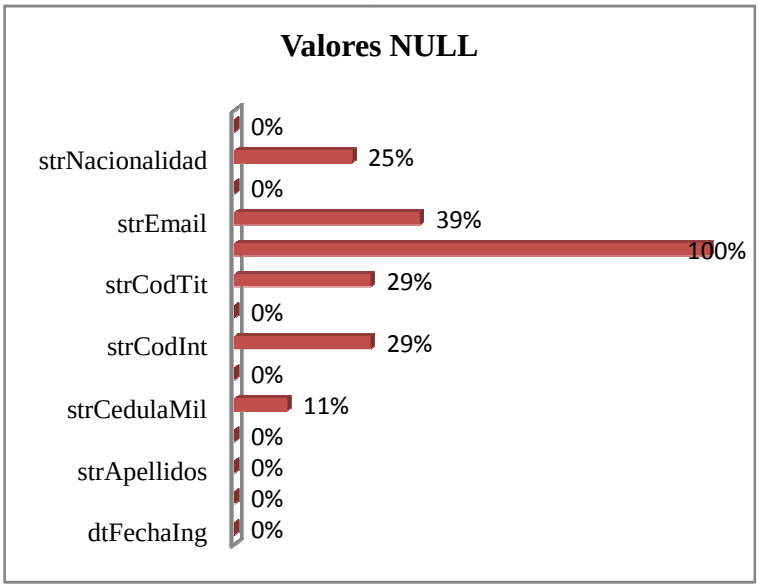


Figura V.77 Valores NULL NatPromSalud

- **Valores Vacíos**

Tabla V.112 Valores vacios NatPromSalud

	strCedula	acion	strCedulaMil	strNombres	strEmail	strCodSexo	strCodInt	ad	strApellidos	strCodTit	strFormaIns	dtFechaIng	dtFechaNac	strCodigo
Empty values	0	0	354	0	108	0	0	0	0	0	0	0	0	0

- **Distribución de valores de columna**

Tabla V.113 Distribución de valores de columna NatPromSalud

Columna	Número de valores distintos
dtFechaIng	450
dtFechaNac	419
strApellidos	448
strCedula	454
strCedulaMil	51
strCodigo	454
strCodInt	115
strCodSexo	2
strCodTit	20
strDocumentacion	0
strEmail	166
strFormaIns	3
strNacionalidad	6
strNombres	440

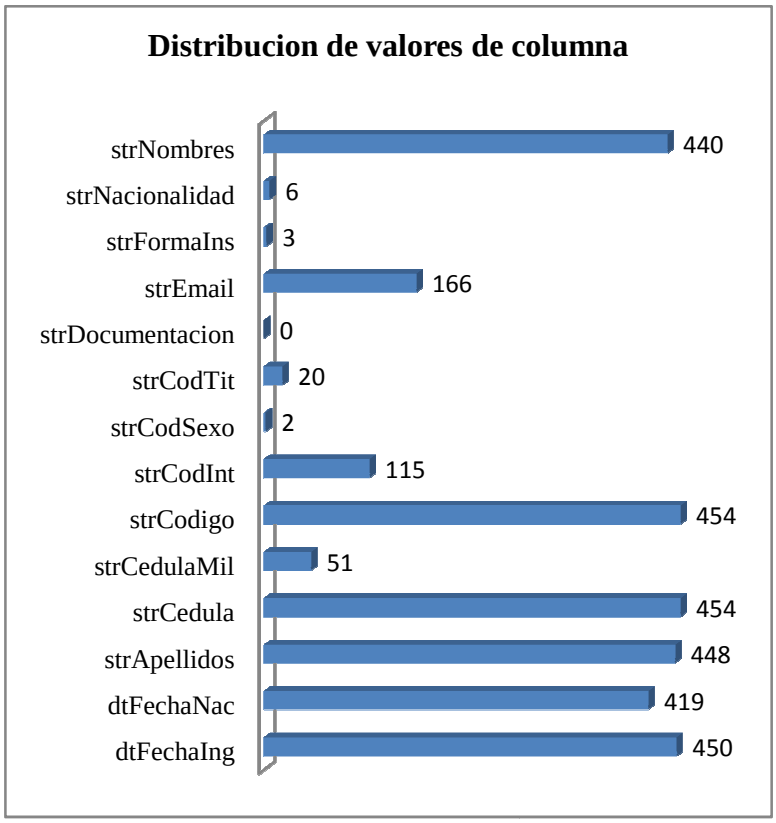


Figura V.78 Distribucion de valores de columna

- **Distribución de longitud de columnas**

Tabla V.114 Distribución de longitud de columnas NatPromSalud

Columna	Longitud mínima	Longitud máxima
strApellidos	7	23
strCedula	11	11
strCedulaMil	0	12
strCodigo	3	5
strCodInt	2	8
strCodSexo	3	3
strCodTit	2	5
strEmail	0	43
strFormaIns	3	3
strNacionalidad	7	11
strNombres	4	24

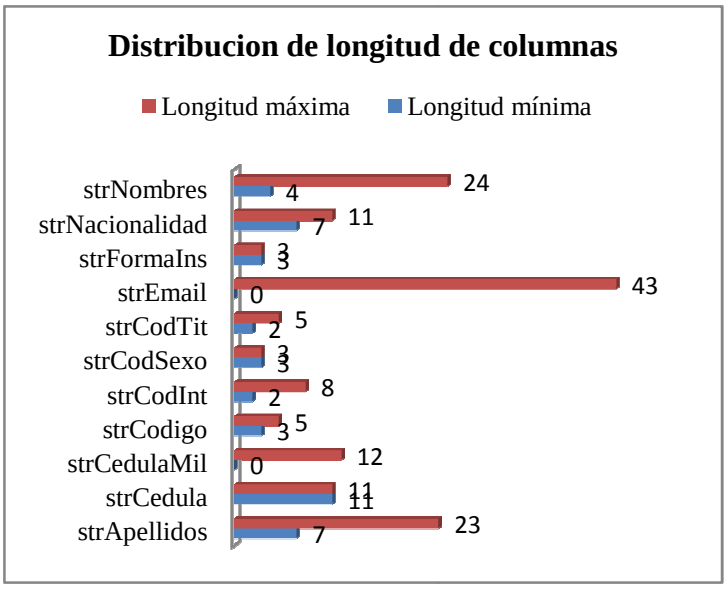


Figura V.79 Distribución de longitud de columnas NatPromSalud

- **Duplicación**

Tabla V.115 Duplicacion NatPromSalud

Columnas de clave	Nivel de clave	Duplicación
dtFechaIng	99%	1%
dtFechaNac	92%	8%
strApellidos	99%	1%
strCedula	100%	0%
strCedulaMil	11%	89%
strCodigo	100%	0%

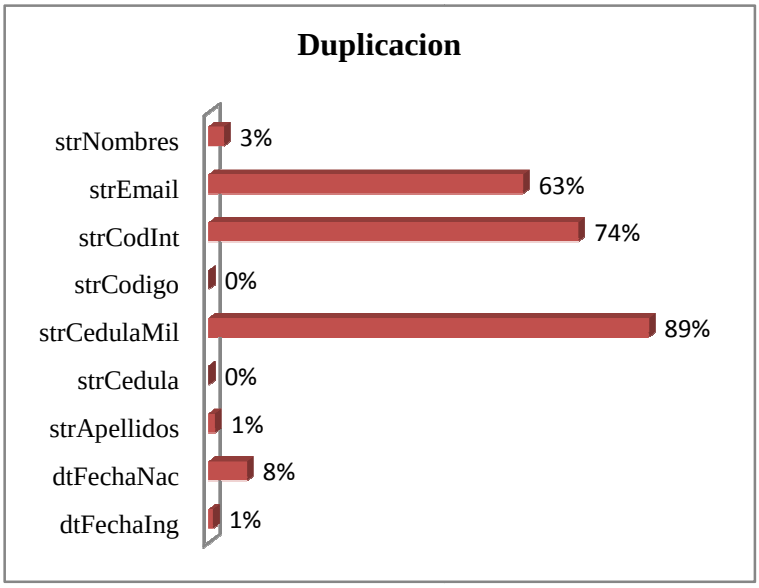


Figura V.80 Duplicacion NatPromSalud

strCodInt	26%	74%
strEmail	37%	63%
strNombres	97%	3%

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.116 Mayusculas y minusculas NatPromSalud

	strCedula	**[illegible]tacion**	**[illegible]il**	**strNombres**	**strEmail**	**strCodSexo**	**strCodInt**	**strNacionalidad**	**strApellidos**	**strCodTit**	**strFormaIngs**	**strCodigo**
Uppercase chars	0%	0%	0%	91%	3%	100%	99%	98%	91%	100%	100%	0%
Lowercase chars	0%	0%	0%	0%	77%	0%	0%	1%	0%	0%	0%	0%

- **Tiempo**

Tabla V.117 Tiempo NatPromSalud

	dtFechaIng	dtFechaNac
Highest value	17/09/2010 8:15:35	06/04/2010
Lowest value	20/12/2005 15:43:05	13/02/1964

- **Patrones**

Tabla V.118 Patrones NatPromSalud

	strCedula
999999999-9	454

Tabla Estudiantes Base de datos OAS_Nutricion_db

- **Valores NULL**

Tabla V.119 Valores NULL Nutrición

Columna	Recuento de NULL	Porcentaje de NULL
dtFechaIng	0	0%
dtFechaNac	13	2%
strApellidos	0	0%
strCedula	0	0%
strCedulaMil	166	22%
strCodigo	0	0%
strCodInt	248	33%
strCodSexo	0	0%
strCodTit	248	33%
strDocumentacion	758	100%
strEmail	285	38%
strFormaIns	0	0%
strNacionalidad	94	12%
strNombres	0	0%

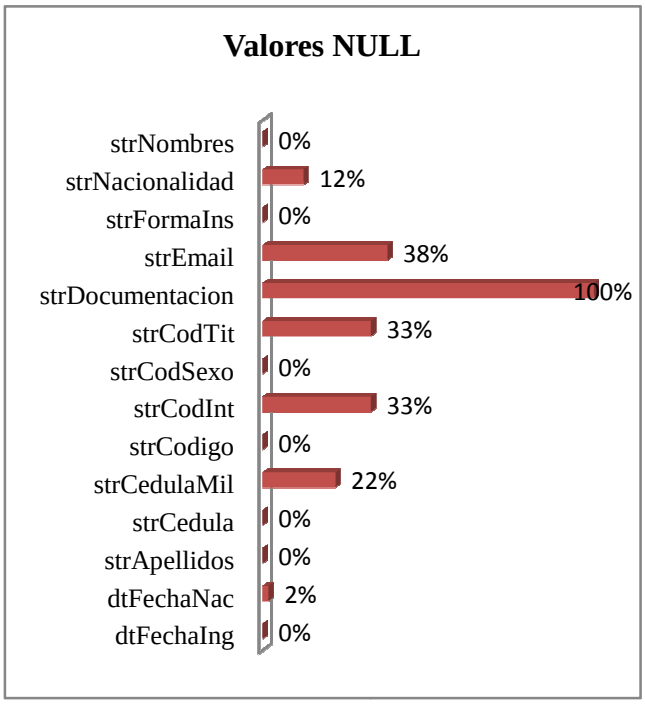


Figura V.81 Valores NULL Nutricion

- **Valores Vacíos**

Tabla V.120 Valores vacíos Nutrición

	strCodTit	strNacionalidad	strCedula	strEmail	strCodSexo	strNombre	strFormaIns	strApellidos	strCedulaMil	strCodInt	dtFechaNac	n	strCodigo	dtFechaIng
Empty values	0	0	0	599	0	0	0	0	777	0	0	0	0	0

- **Distribución de valores de columna**

Tabla V.121 Distribución de valores de columna Nutricion

Columna	Número de valores distintos
dtFechaIng	642
dtFechaNac	665
strApellidos	748
strCedula	758
strCedulaMil	97
strCodigo	758
strCodInt	196
strCodSexo	2
strCodTit	25
strDocumentacion	0
strEmail	208
strFormaIns	3
strNacionalidad	9
strNombres	696

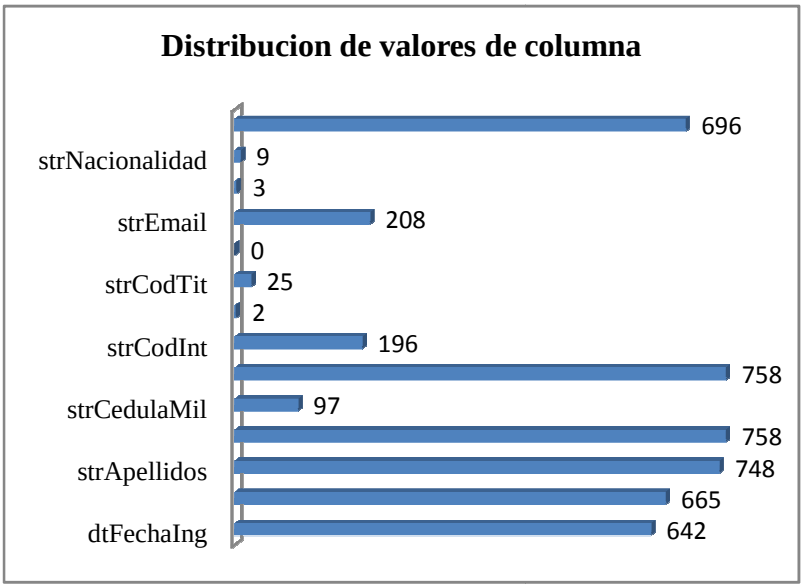


Figura V.82 Distribución de valores de columna Nutricion

- **Distribución de longitud de columnas**

Tabla V.122 Distribucion de longitud de co

Columna	Longitud mínima	Longitud máxima
strApellidos	9	22
strCedula	11	11
strCedulaMil	0	13
strCodigo	6	7
strCodInt	2	10
strCodSexo	3	3
strCodTit	2	5
strEmail	0	32
strFormaIns	3	3
strNacionalidad	2	22

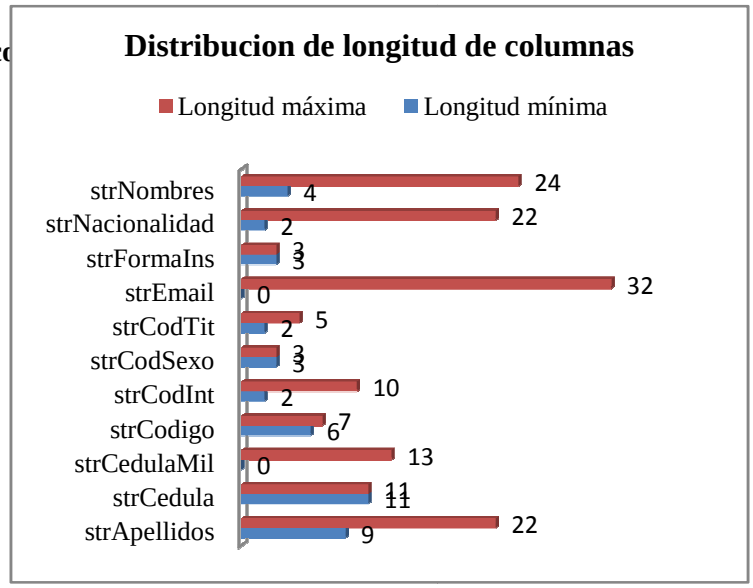


Figura V.83 Distribucion de longitud de columnas Nutricion

strNombres	4	24

- **Duplicación**

Tabla V.123 Duplicación Nutrición

Columnas de clave	Nivel de clave	Duplicación
dtFechaIng	85%	15%
dtFechaNac	88%	12%
strApellidos	99%	1%
strCedula	100%	0%
strCedulaMil	13%	87%
strCodigo	100%	0%
strCodInt	26%	74%
strEmail	28%	72%
strNombres	92%	8%

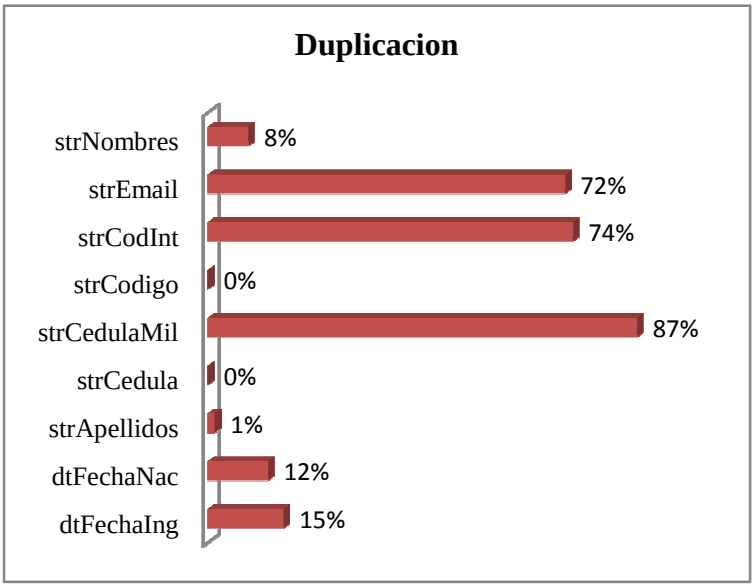


Figura V.84 Duplicacion Nutricion

- **Caracteres Mayúsculas y Minúsculas**

Tabla V.124 Mayúsculas y minúsculas Nutrición

	strCodTit	strNacionalidad	strCedula	strEmail	strCodSexo	strNombres	strFormaIns	strApellidos	strCedulaMil	strCodInt	[illegible]cion	strCodigo
Uppercase chars	99%	98%	0%	1%	100%	97%	100%	97%	0%	98%	0%	0%
Lowercase chars	0%	2%	0%	80%	0%	3%	0%	3%	0%	0%	0%	0%

- **Tiempo**

Tabla V.125 Tiempo Nutrición

	dtFechaNac	dtFechaIng
Highest value	13/03/2008 0:00:00	15/09/2010 17:21:51
Lowest value	16/05/1905 0:00:00	20/10/1988 0:00:00

- **Patrones**

Tabla V.126 Patrones Nutrición

strCedula	
999999999-9	1402

- **Dimensiones de calidad de datos Afectados**

Todos los problemas presentados afectan determinadas dimensiones de calidad , las cuales se muestras a continuación:

Tabla V.127 Dimensiones afectadas

No.	Problema	Dimensión de calidad Afectada
1	Datos NULL y blancos	Validez
2	Datos incompletos	Precisión
3	Datos duplicados	Duplicidad
4	Datos sin formatos necesarios	Conformidad
5	Datos sin un estándar especifico	Presentabilidad
6	Inconsistencias en los datos	Consistencia

- **Resultados de la Evaluación Inicial**

  La cantidad total para el calculo total será la cantidad total de datos de la tabla por el numero de problemas registrados:

FADE_FASE_1IC

Tabla V.128 Evaluación Inicial FADE_FASE_1IC

	CEDIDE	FECNAC	FECING	SEXO	NOMEST	APEEST	CODEST	ECUEST	Total	Comentarios
NULL	3	**2**	**2**	**2**	**2**	**2**	1	2	16	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	3								**3**	**Son únicamente NULL**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**27**								**27**	
ECUEST=S								**31**	**31**	
SEXO=F o M				**31**					**31**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**108**	

Resultados:

Total =264

Tabla V.129 Resultados Evaluación Inicial FADE_FASE_1IC

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_1IC	108	40,9%	59,1%

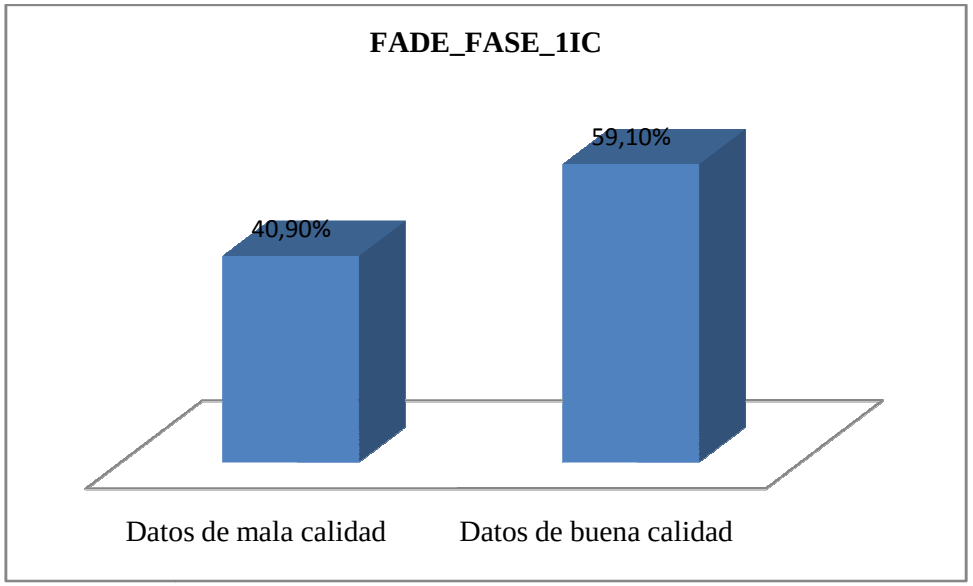


Figura V.85 Resultados Evaluación Inicial FADE_FASE_1IC

FADE_FASE_2IC

Tabla V.130 Evaluación Inicial FADE_FASE_2IC

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	**Son únicamente NULL**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**24**	**24**	
SEXO=F o M				**24**					**24**	
Inconsistencia de tiempo		**0**	**1**						**1**	**Se descarto la fecha 1907 de la fecha de ingreso**
								Total	**49**	

Resultados:

Total de datos=192

Tabla V.131 Resultados Evaluación Inicial FADE_FASE_2IC

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_2IC	49	25%	75%

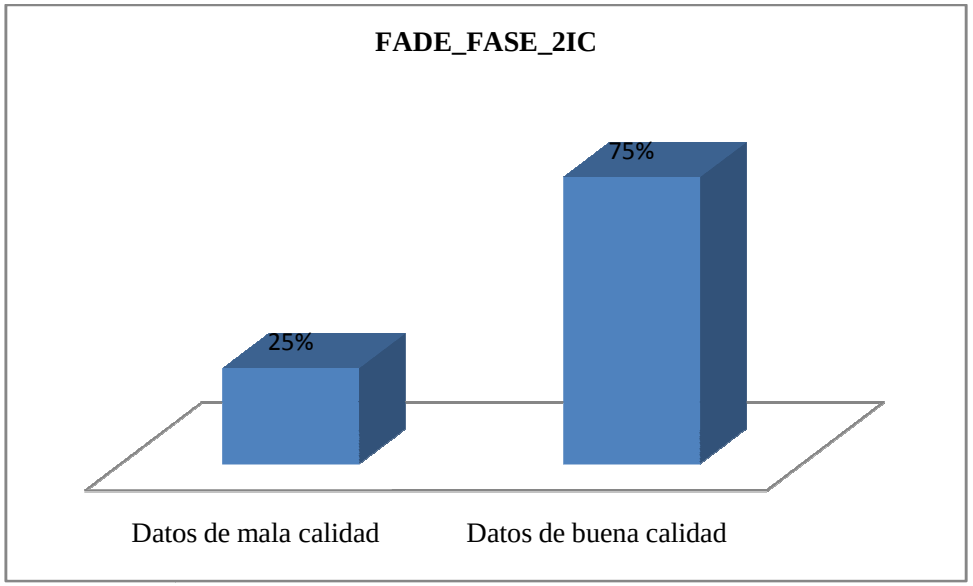


Figura V.86 Resultados Evaluación Inicial FADE_FASE_2IC

FADE_FASE_6

Tabla V.132Evaluación de datos FADE_FASE_6

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	5	**13**	**4**	**1**	**1**	**1**	1	1	27	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	9								**9**	**Existen: NULL=5 070227832-6=2 060287165-9=2**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**3**								**3**	
ECUEST=S								**357**	**357**	
SEXO=F o M				**357**					**357**	
Inconsistencia de tiempo		**1**	**192**						**193**	**Fechas Inconsistentes: FI=1900-1904 FN=1900**
								Total	**946**	

Total de datos=2864

Tabla V.133 Resultados Evaluación de Calidad FADE_FASE_6

BD	Total datos de mala calidad	Porcentaje de datos de mala calidad	Porcentaje de datos de buena calidad
FADE_FASE_6	946	33%	67%

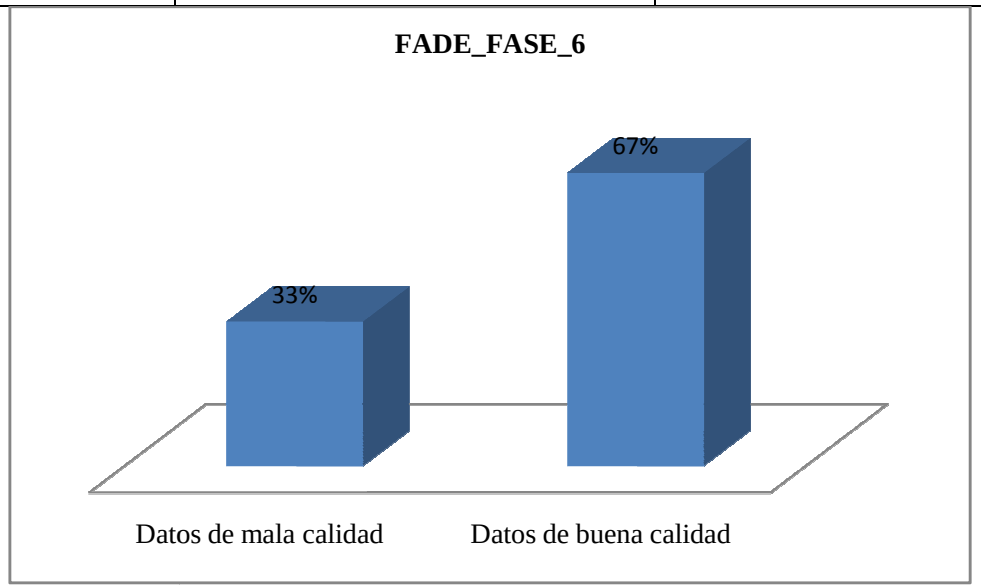


Figura V.87 Resultados Evaluación Inicial FADE_FASE_6

FADE_FASE_7

Tabla V.134 Evaluación Inicial FADE_FASE_7

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	5	**13**	**4**	**1**	**1**	**1**	1	1	27	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	9								**9**	**Existen: NULL=5 070227832-6=2 060287165-9=2**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**3**								**3**	
ECUEST=S								**446**	**446**	
SEXO=F o M				**446**					**446**	
Inconsistencia de tiempo		**1**	**281**						**281**	**Fechas Inconsistentes: FI=1900-1905 FN=1900**
								Total	**1212**	

Total de datos: 3576

Tabla V.135 Resultados de la Evaluación FADE_FASE_7

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_7	1212	34%	66%

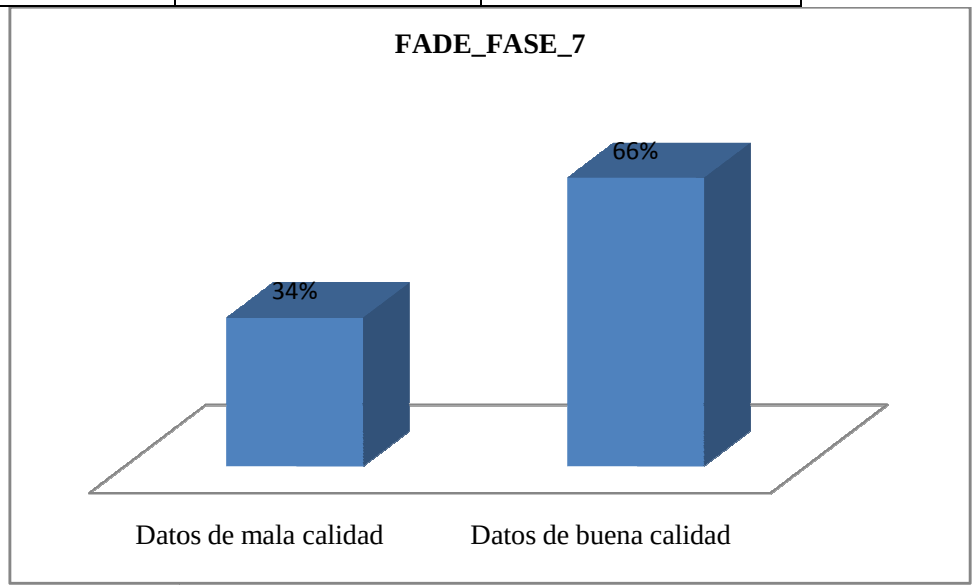


Figura V.88 Resultados Evaluación Inicial FADE_FASE_7

FADE_FASE_8

Tabla V.136 Evaluación Inicial FADE_FASE_8

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	5	**13**	**4**	**1**	**1**	**1**	1	1	27	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	9								**9**	**Existen: NULL=5 070227832-6=2 060287165-9=2**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**3**								**3**	
ECUEST=S								**45**	**45**	
SEXO=F o M				**45**					**45**	
Inconsistencia de tiempo		**0**	**45**						**45**	**Fechas Inconsistentes: FI=1903-1909**
								Total	**174**	

Resultados :

Total de datos: 368

Tabla V.137 Resultados Evaluación Inicial FADE_FASE_8

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_8	174	47%	53%

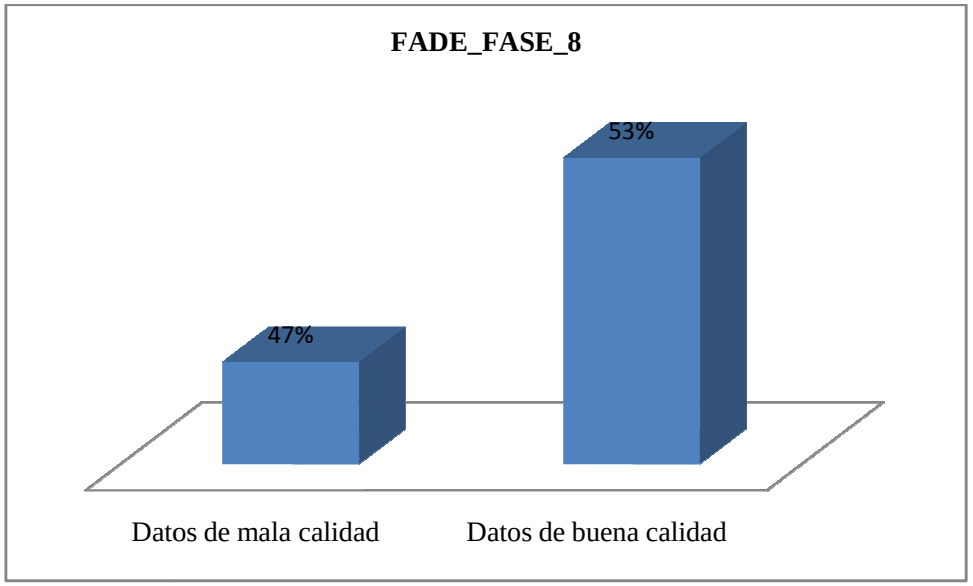


Figura V.89 Resultados Evaluación Inicial FADE_FASE_8

FADE_FASE_9

Tabla V.138 Evaluación Inicial FADE_FASE_9

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	1	**2**	**1**	**1**	**1**	**1**	1	1	9	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	2								**2**	**170883416-1=2**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**									
ECUEST=S								**48**	**48**	
SEXO=F o M				**48**					**48**	
Inconsistencia de tiempo		**0**	**48**						**48**	**Fechas Inconsistentes: FI=1907-1908**
								Total	**156**	

Resultados:

Total de datos: 392

Tabla V.139 Resultados Evaluación Inicial FADE_FASE_9

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_9	156	40%	60%

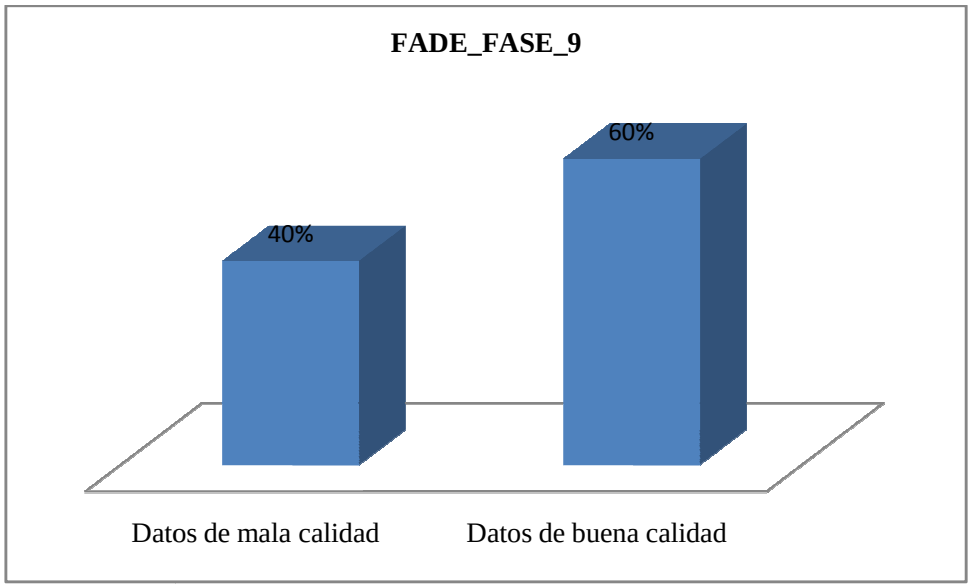


Figura V.90 Resultados Evaluación Inicial FADE_FASE_9

FADE_FASE_10

Tabla V.140 Evaluación Inicial FADE_FASE_10

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**1**	**0**	**0**	**0**	**0**	0	0	1	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**51**	**51**	
SEXO=F o M				**51**					**51**	
Inconsistencia de tiempo		**0**	**51**						**51**	**Fechas Inconsistentes: FI=1904-1927**
								Total	**155**	

Resultados:

Total de datos: 408

Tabla V.141 Resultados Evaluación Inicial FADE_FASE_10

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad

FADE_FASE_10	155	38%	62%

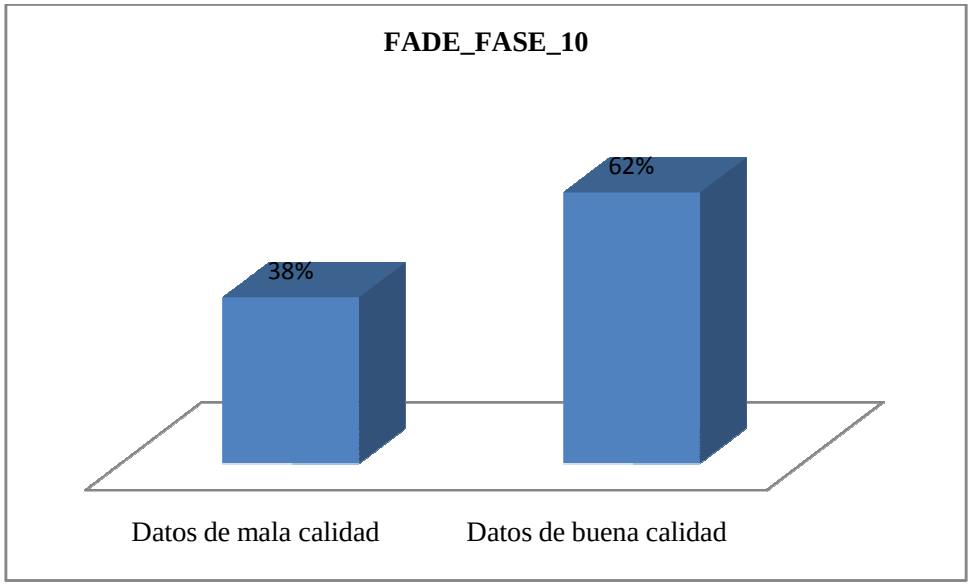


Figura V.91 Resultados Evaluación FADE_FASE_10

FADE_FASE_GGSBA

Tabla V.142 Evaluación Inicial FADE_FASE_GGSBA

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**1**	**1**	**1**	**1**	**1**	1	0	6	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0									
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**24**								**24**	
ECUEST=S								**36**	**36**	
SEXO=F o M				**36**					**36**	
Inconsistencia de tiempo		**0**	**0**						**0**	**Fechas Inconsistentes: FI=1904-1927**
								Total	**102**	

Resultados:

Total de datos: 296

Tabla V.143 Resultados Evaluación Inicial GGSBA

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_GGSBA	102	35%	65%

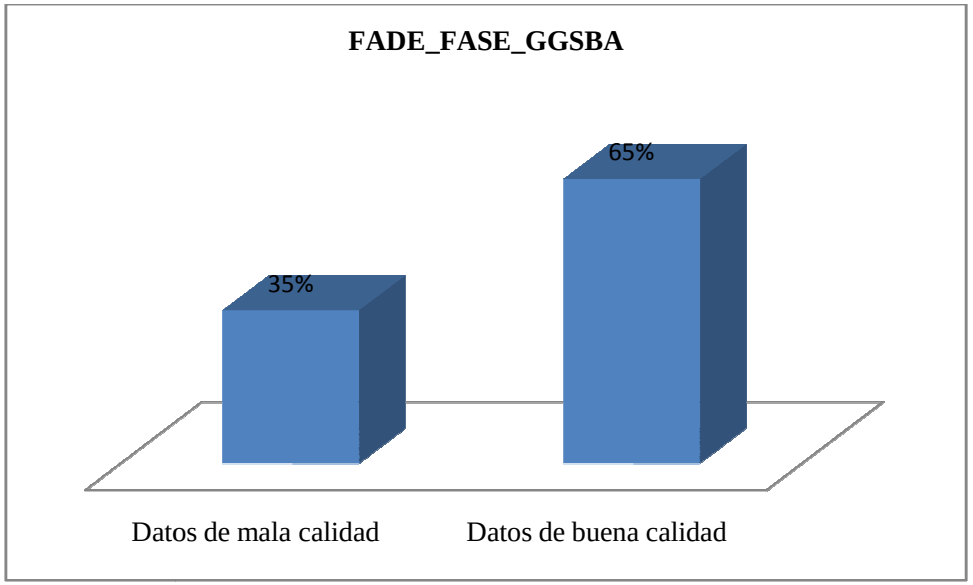


Figura V.92 Resultados Evaluación Inicial FADE_FASE_GGSBA

FADE_FASE_GGSES

Tabla V.144 Evaluación Inicial FADE_FASE_GGSES

	strCedul	dtFechaNa	dtFechaIn	strCodSex	strNombre	strApellido	strCodig	strNacionalida	Tota	Comentario

	a	c	g	o	s	s	o	d	l	s
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**22**	**22**	
SEXO=F o M				**22**					**22**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**44**	

Resultados:

Total de datos: 176

Tabla V.145 Resultados Evaluacion Inicial FADE_FASE_GGSES

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_GGSES	44	25%	75%

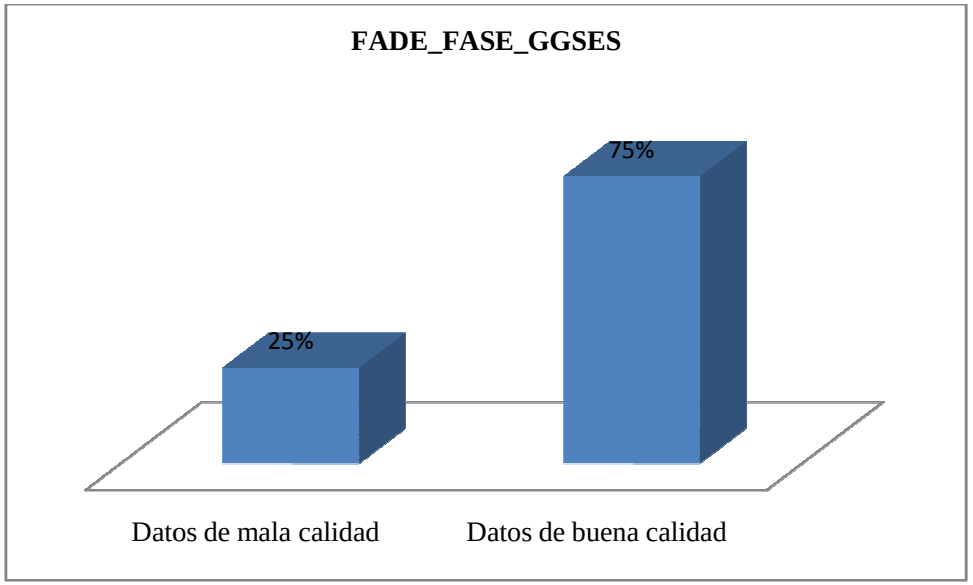


Figura V.93 Resultados Evaluación Inicial GGSES

Ciclo Formativo

Tabla V.146 Evaluación Inicial Ciclo Formativo

	strCedu	dtFechaN	dtFechaI	strCodSe	strNombr	strApellid	strCodi	strNacionali	Tot	Comentari

	la	ac	ng	xo	es	os	go	dad	al	os
NULL	0	**0**	**1**	**0**	**0**	**0**	0	2	3	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**97**	**64**		**0**	**161**	
Longitud <11	**0**									
ECUEST<>ECUATORIANO								**3099**	**3099**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**3263**	

Resultados:

Total de datos: 25752

Tabla V.147 Resultados Evaluación Inicial Ciclo Formativo

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
OAS_Ciclo Formativo	3263	13%	87%

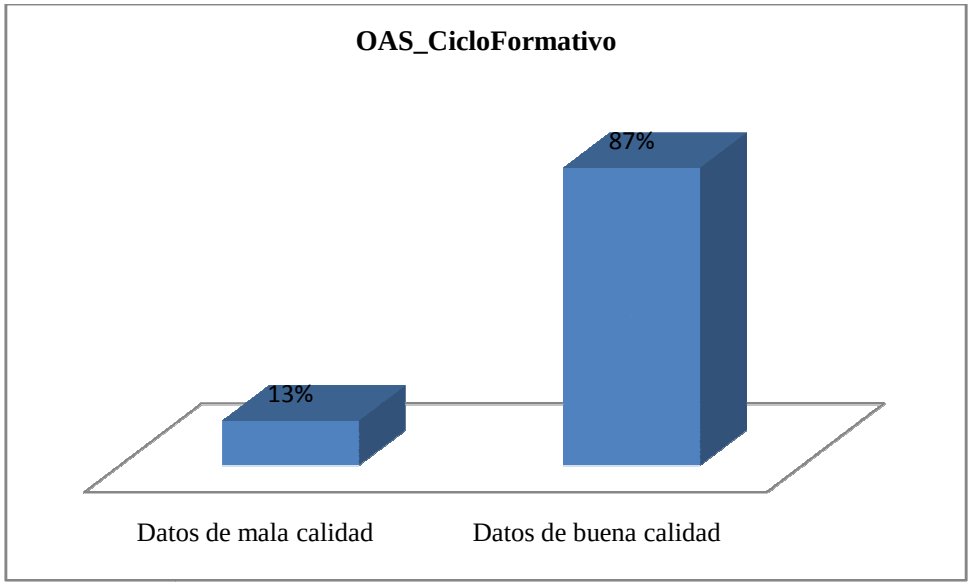


Figura V.94 Resultados Evaluación Inicial CicloFormativo

OAS_IngAgronomica

Tabla V.148 Evaluación Inicial OAS_IngAgronomica

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**3**	**0**	**0**	**0**	**0**	0	2	5	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**7**	**7**		**0**	**14**	
Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**577**	**577**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**0**						**0**	

	Total	**596**	

Resultados:

Total de datos: 5624

Tabla V.149Resultados Evaluación Inicial IngAgronomica

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
OAS_IngAgronómica	596	11%	89%

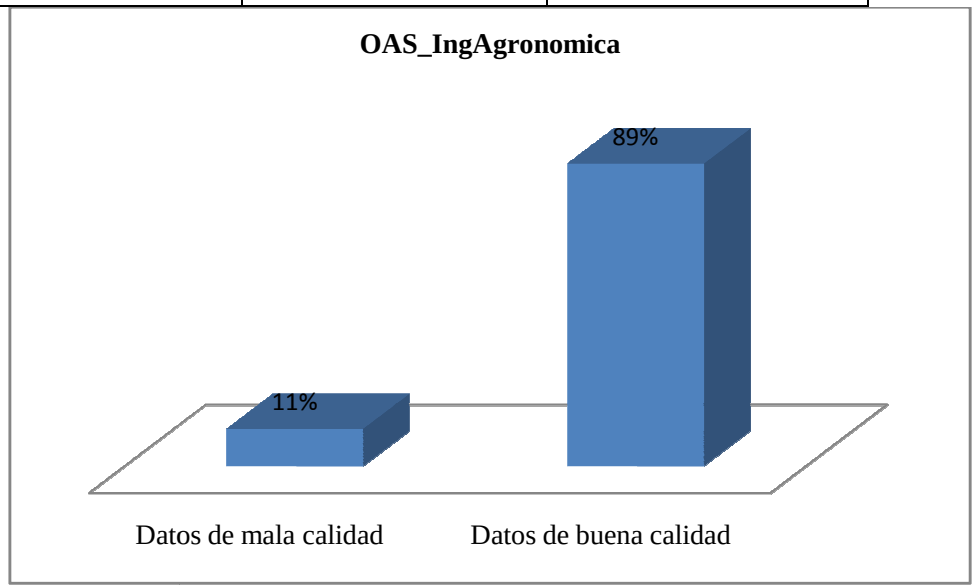


Figura V.95 Resultados Evaluación Inicial IngAgronomica

OAS_IngEmpresas

Tabla V.150 Evaluación Inicial OAS_IngEmpresas

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**3**	**0**	**0**	**0**	**0**	0	48	51	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**42**	**42**			**84**	
Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**1257**	**1257**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**4**	**0**						**4**	
								Total	**1396**	

Resultados:

Total de datos: 11216

Tabla V.151 Resultados Evaluación Inicial OAS_IngEmpresas

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
OAS_IngEmpresas	1396	13%	87%

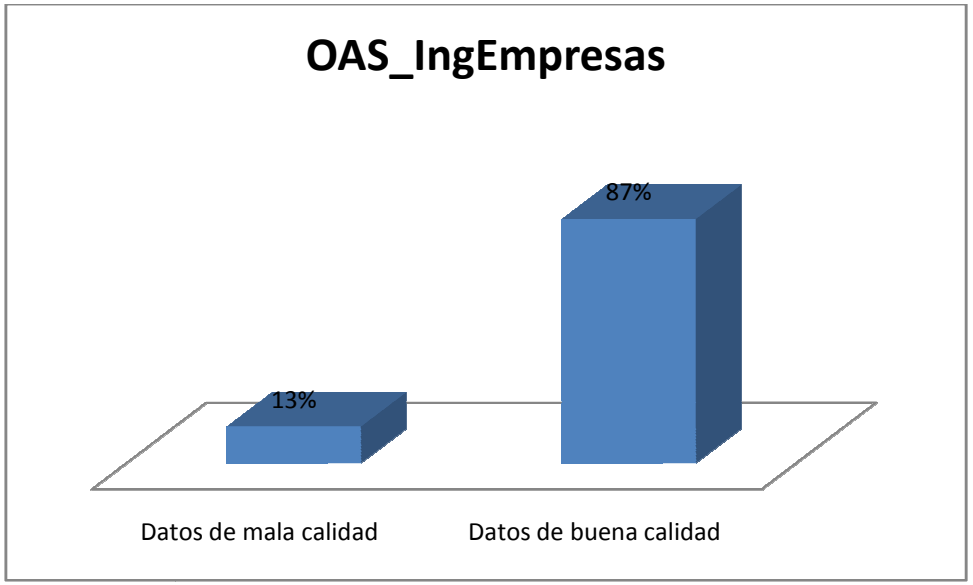


Figura V.96 Resultados Evaluación Inicial OAS_IngEmpresas

OAS_NatPromSalud

Tabla V.152 Evaluacion Inicial OAS_NatPromSalud

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	112	112	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**14**	**14**		**9**	**37**	
Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**305**	**305**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**454**	

Resultados:

Total de datos: 3632

Tabla V.153 Resultados Evaluación Inicial OAS_PromSalud

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
OAS_IngEmpresas	454	13%	87%

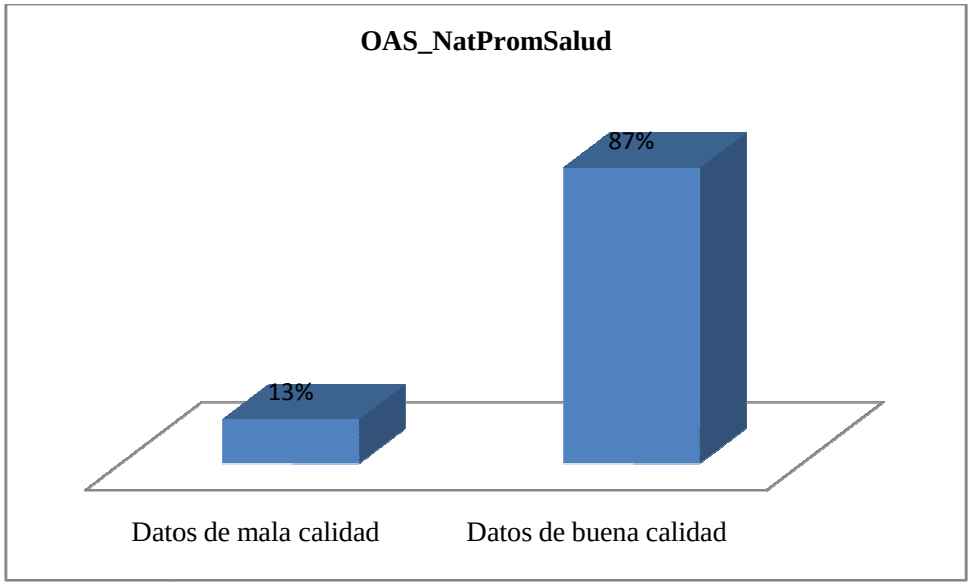


Figura V.97 Resultados OAS_NatPromSalud

Nutrición

Tabla V.154 Evaluación Inicial Nutrición

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**13**	**0**	**0**	**0**	**0**	0	94	107	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**14**	**14**		**9**	**37**	
Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**639**	**639**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**6**	**0**						**6**	
								Total	**789**	

Resultados:

Total de datos: 3632

Tabla V.155 Resultados Evaluación Inicial OAS_Nutricion

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
OAS_Nutricion	454	13%	87%

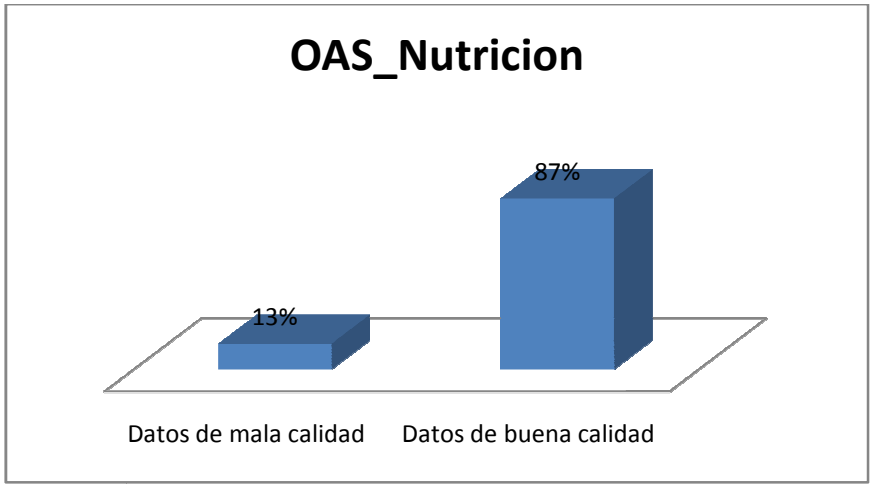


Figura V.98 Resultados Evaluación Inicial OAS_Nutricion

Base de Datos de la UED

Tabla V.156 Resultados Evaluación Inicial Base de Datos UED

Total datos	Buena Calidad	Mala Calidad
8536	65,49%	34,51%

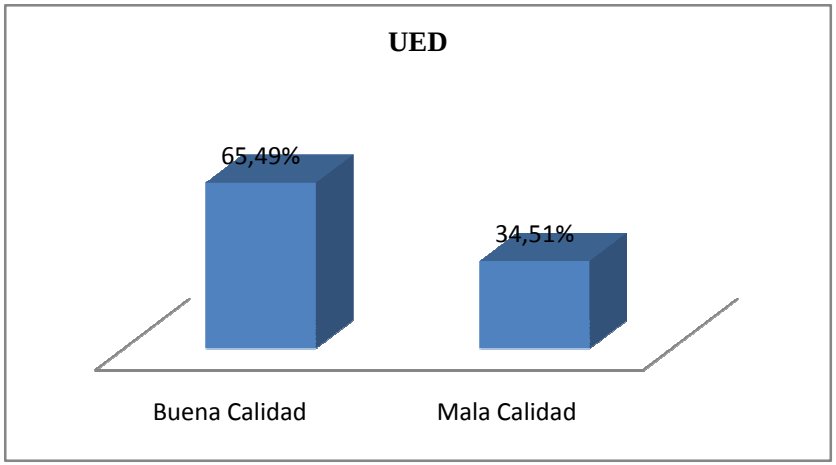


Figura V.99Resultado Evaluación Inicial UED

Base de Datos del Sistema Académico

Tabla V.157 Resultado Evaluacion Inicial Sistema Academico

Total datos	Buena Calidad	Mala Calidad
54288	87.58%	12.42%

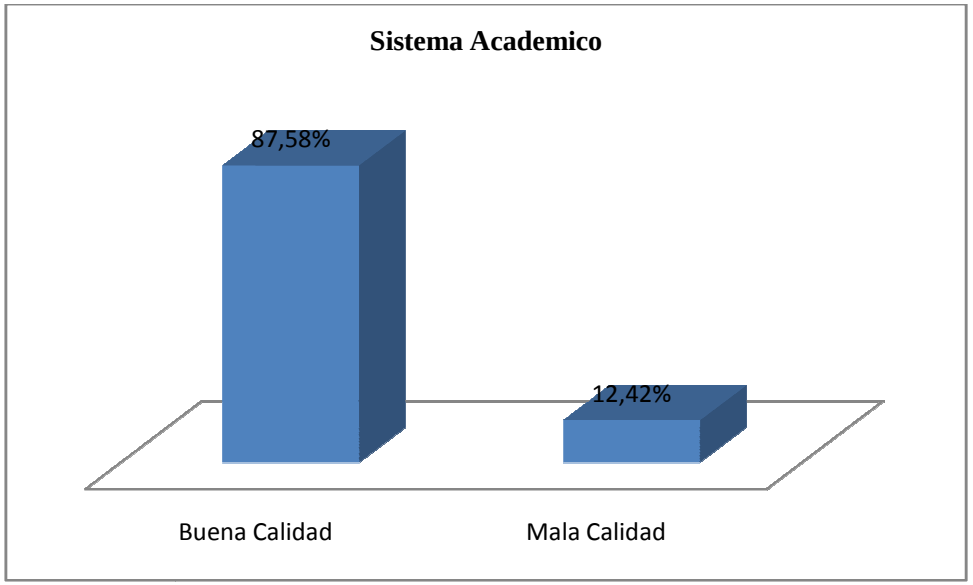


Figura V.100 Resultados evaluación Inicial Sistema Académico

5.2.3.4Etapa 3.4 Políticas internas de calidad

- **Políticas establecidas para la Base de Datos de las Escuelas a Distancia**

En la siguiente tabla se muestra las políticas establecidas para los datos de las bases de datos de la unidad de educación a distancia:

Tabla V.158 Políticas establecidas UED

No.	Problema	Tabla(s)	Columna(s)	Política(s)	Comentarios
1	Datos NULL y blancos	CESTUD	CEDIDE FECNAC	CEDIDE=000000000-0 FECNAC=1900-01-01	
2	Datos incompletos	CESTUD	CEDIDE	-Valor referencial en caso de cedula incompleta: 000000000-0	
3	Datos duplicados	CESTUD	CEDIDE	Eliminar	
4	Datos sin formatos necesarios	CESTUD	CEDIDE	Formato definido: Cedulas con guion	
5	Datos sin un estándar especifico	CESTUD	NOMEST APEEST	- Todos con mayúsculas	
			SEXO	-Estandarizar valor actual según corresponda a MAS y FEM	
			ECUEST	-Estandarizar valor actual según corresponda a ECUATORIANO	
6	Inconsistencias en los datos	CESTUD	FECING	-Valor referencial de inconsistencia: 1900-01-01	

- **Base de Datos de las Escuelas del Sistema Académico**

  En la siguiente tabla se muestra las políticas establecidas para los datos del Sistema Académico:

No.	Problema	Tabla(s)	Columna(s)	Política(s)	Comentarios
1	Datos NULL y blancos	Estudiantes	strCedula	CEDIDE=000000000-0 FECNAC=1900-01-01	
2	Datos incompletos	Estudiantes	strCedula	-Valor referencial en caso de cedula incompleta: 000000000-0	
3	Datos duplicados	Estudiantes	strCedula	Eliminar	
4	Datos sin formatos necesarios	Estudiantes	strCedula	Formato definido: Cedulas con guion	
5	Datos sin un estándar especifico	Estudiantes	strNombres strApellidos	- Todos con mayúsculas	
			strNacionalidad	-Estandarizar valor actual según corresponda a ECUATORIANO	
6	Inconsistencias en los datos	Estudiantes	dtFecIng dtFecNac	-Valor referencial de inconsistencia: 1900-01-01	

5.2.4 FASE IV. LIMPIEZA DE DATOS

Etapa 5.2.5 Limpieza

- **Ejecución de Limpieza**

Para la limpieza de datos se utilizo las siguientes herramientas:

- BayCastle Data Slave Map Editor
- Sql Power DQGuru

- **Sql Power DQGuru**

Requerimientos Software

Sql Power DQGuru requiere Java Runtime Environment (JRE) versión 5.0 o posterior.

Instalación

La instalación es sencilla únicamente se ejecuta el archivo SQL-Power-DQguru-Setup-Windows-0.9.7 y se procede a realizar la instalación mediante el botón next e indicándole donde instalar la herramienta, luego se finaliza

Figura V.101 Inicio de la instalación

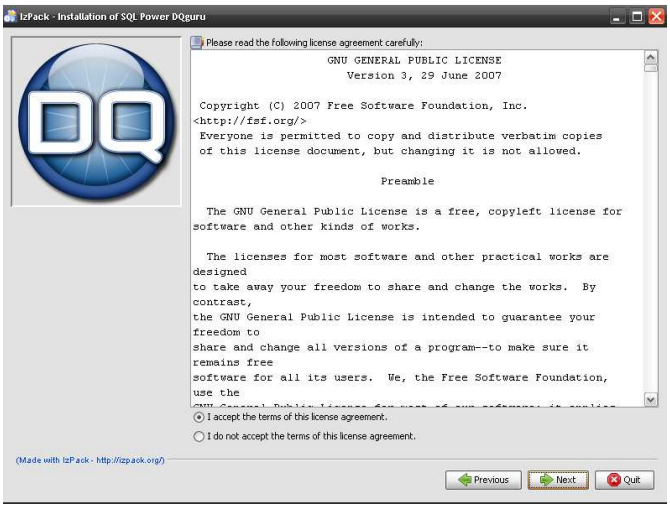


Figura V.102 Términos de GNU

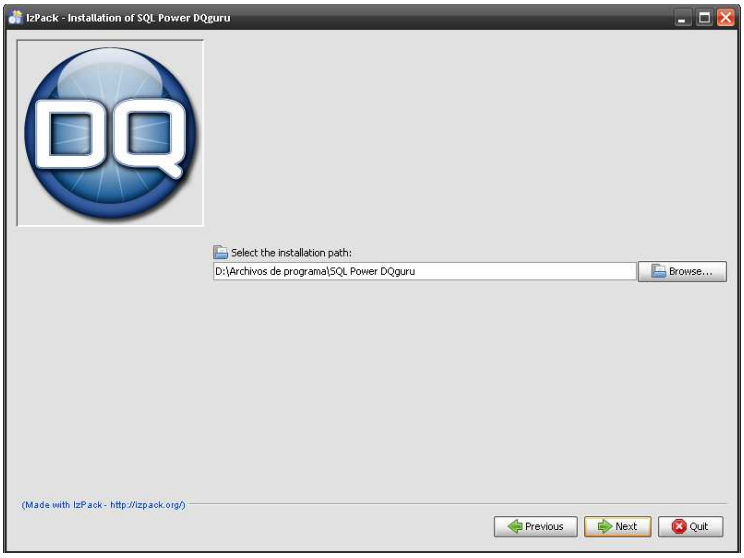


Figura V.103 Path de instalacion

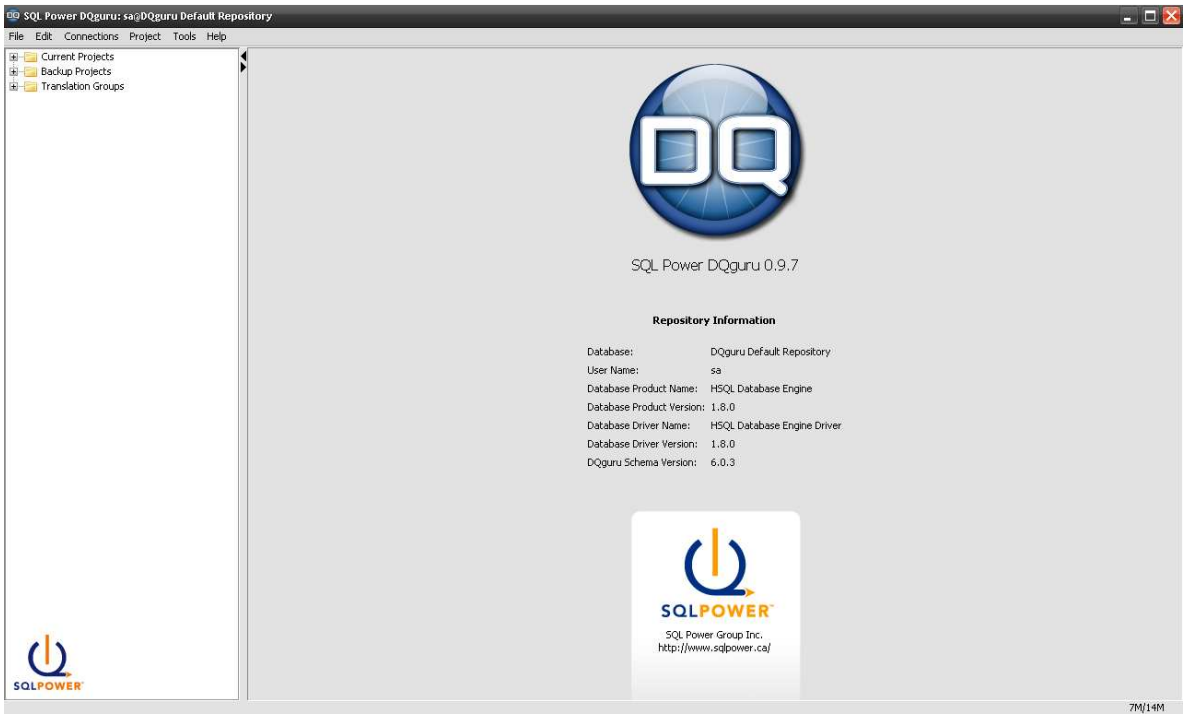


Figura V.104 Pantalla de inicio de SQL Power DQGuru

Configuración

Primero se realiza la configuración de las conexiones de las bases de datos con las que trabaja SQL Power DQGuru .Dado que esta herramienta esta desarrollada en java requiere de drivers para las respectivas conexiones con las bases de datos.

Se accede a **Conection** y se escoje**Manage Database Connections**:

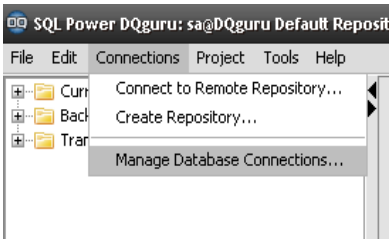


Figura V.105 Administrador de conexiones

Muestra las bases de datos que están disponibles para la conexión:

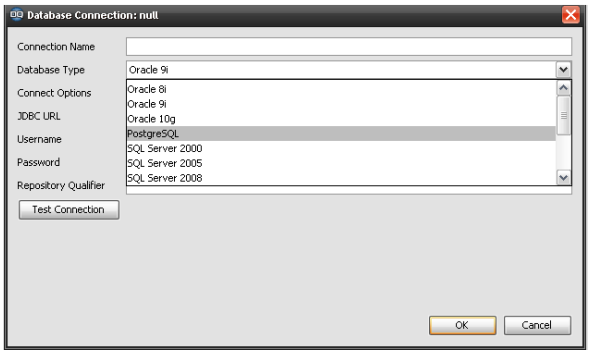


Figura V.106 Bases de datos disponibles para la conexion

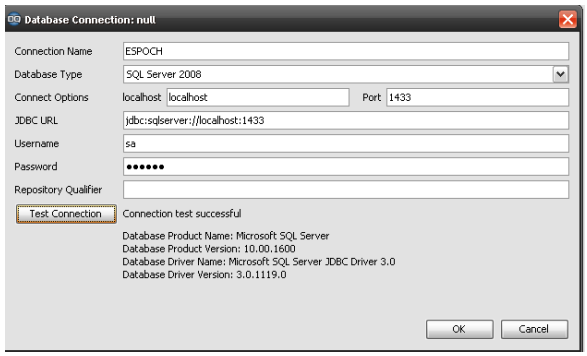


Figura V.107 Configuración de la conexión

En el caso de no estar instalado el driver únicamente se debe estar conectado a internet y se seleccionala base de datos que se necesita y automáticamente se descargara o si ya se descargado el driver en otra ubicación se selecciona Add JAR

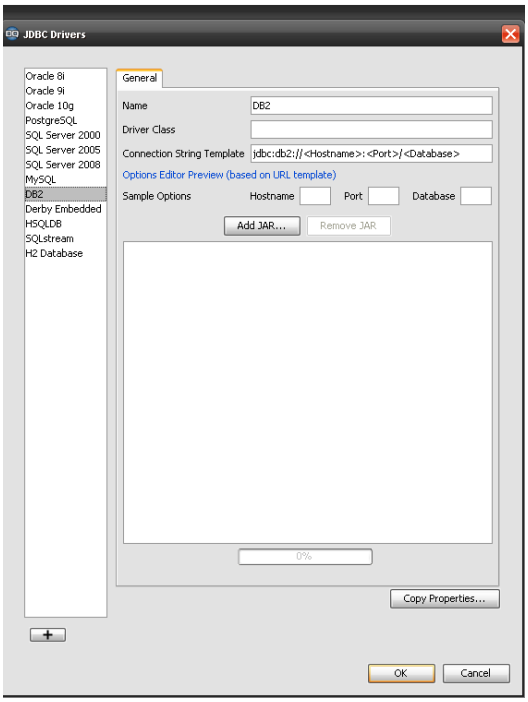


Figura V.108 Bases de datos

Ejecución

Problema: Datos Incompletos

Requerimiento de Calidad: Formato de Cedulas es de 11 dígitos separados por un guión.

Para la limpieza de cedulas definiéndolas un formato único se utilizara SQL Power DQGuru para lo cual se procede de la siguiente manera:

En **Current Projects** se agrega una nueva carpeta donde se almacenara las transformaciones que se realizara

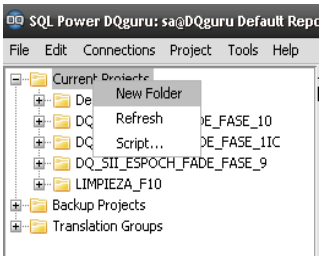


Figura V.109 Nueva carpeta para transformaciones

Se agrega un nuevo proyecto de limpieza

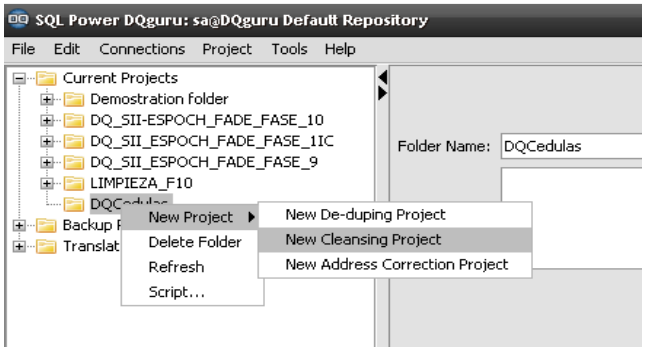


Figura V.110 Proyecto de Limpieza

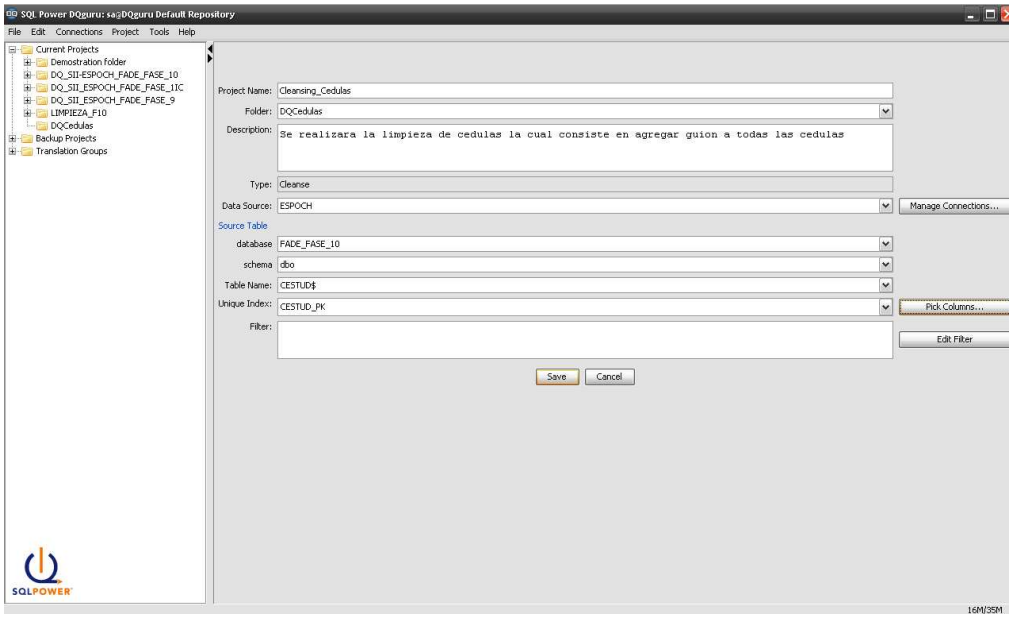


Figura V.111 Configuración de Proyecto

En la carpeta del proyecto clik derecho y se agrega una nueva transformación:

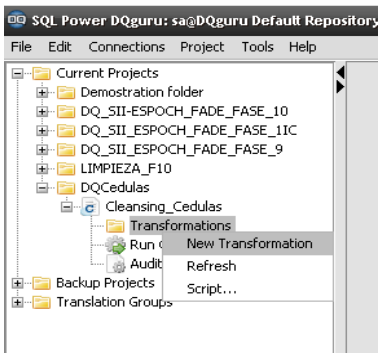


Figura V.112 Nueva transformación

Se muestra un origen y un destino de datos, es decir se obtiene el origen se realiza la transformación de datos (limpieza) y se lo envía al destino.

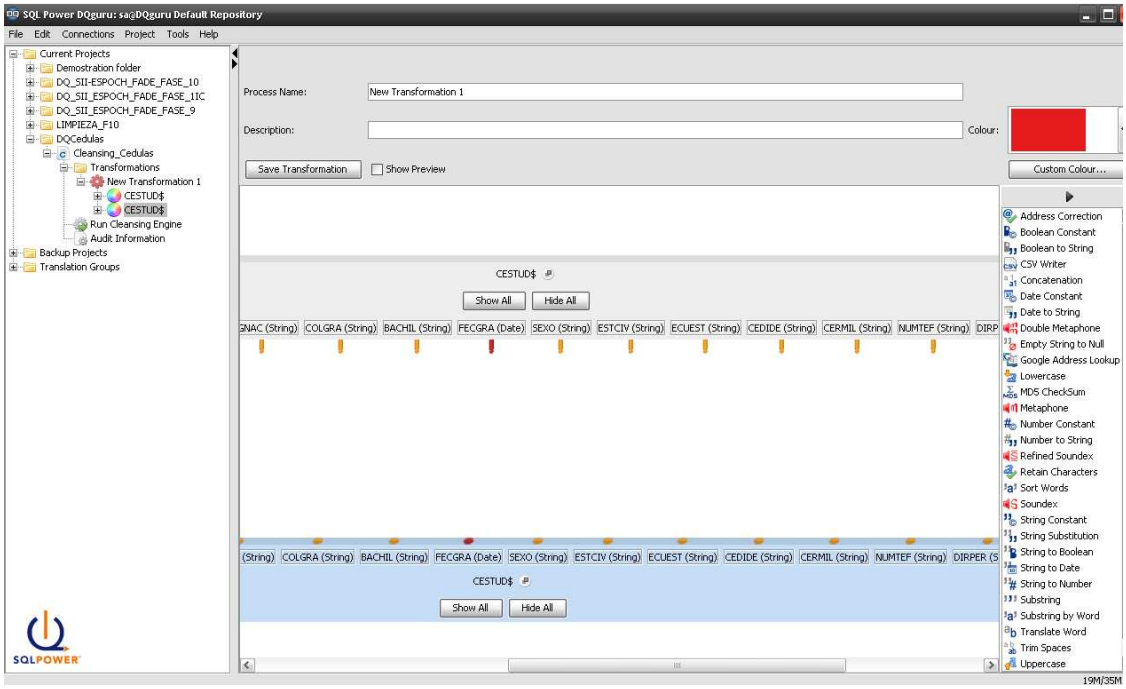


Figura V.113 Origen y Destino de datos

Para realizar la limpieza de las cedulas se obtiene como origen el campo de la cedula se agrega la tarea **Retain Chars** para obtener únicamente los dígitos del 0 al 9, luego se agrega dos tareas **Substring** el primero para obtener los 9 primeros dígitos y el segundo para obtener el ultimo digito, para finalizar se agrega una tarea **Concat** para concatenar los dos substrings mediante un delimitador que se lo ha definido con guion y se envia todo esto al destino.

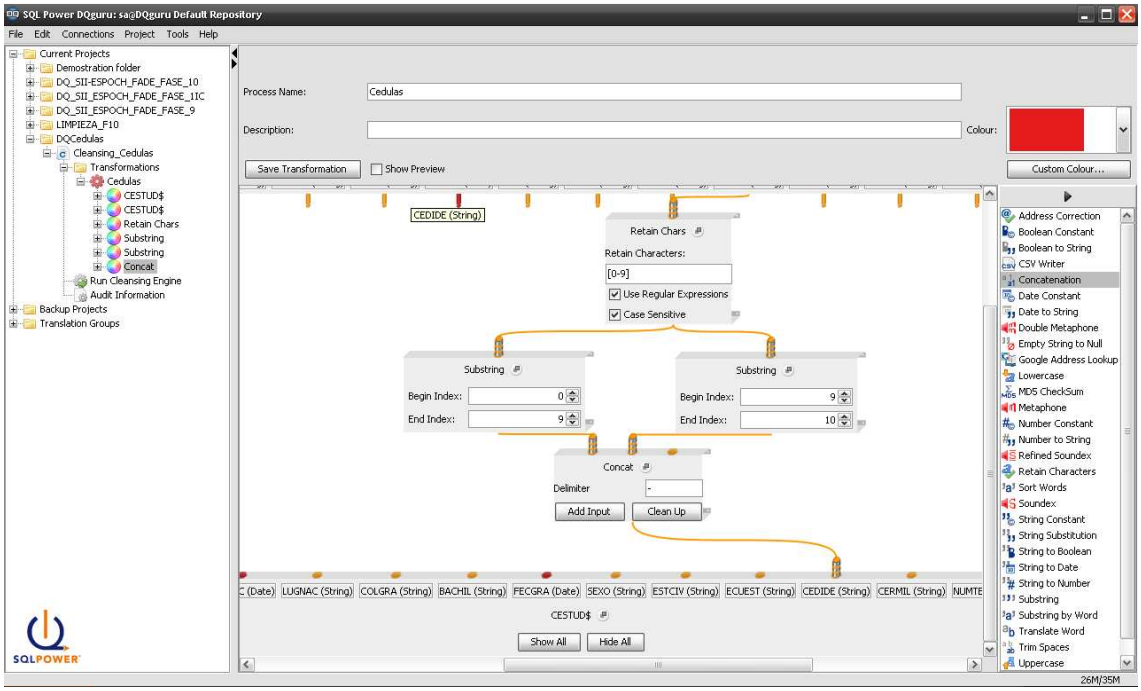


Figura V.114 Configuración de transformación

Para proceder a ejecutar la transformación, en el panel derecho **Run Cleansing Engine ,** se selecciona la transformación aceptamos y hacemos clik en **Run Engine**

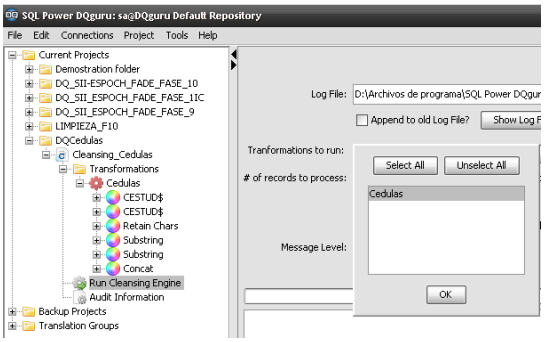


Figura V.115 Selección de Transformación

En el momento de ejecución se muestra una advertencia en la cual advierte que se realizara cambios en el origen de datos y que es necesario hacer un backup del mismo .Como ya se hizo el backup de las fuentes de datos seleccionamos si.

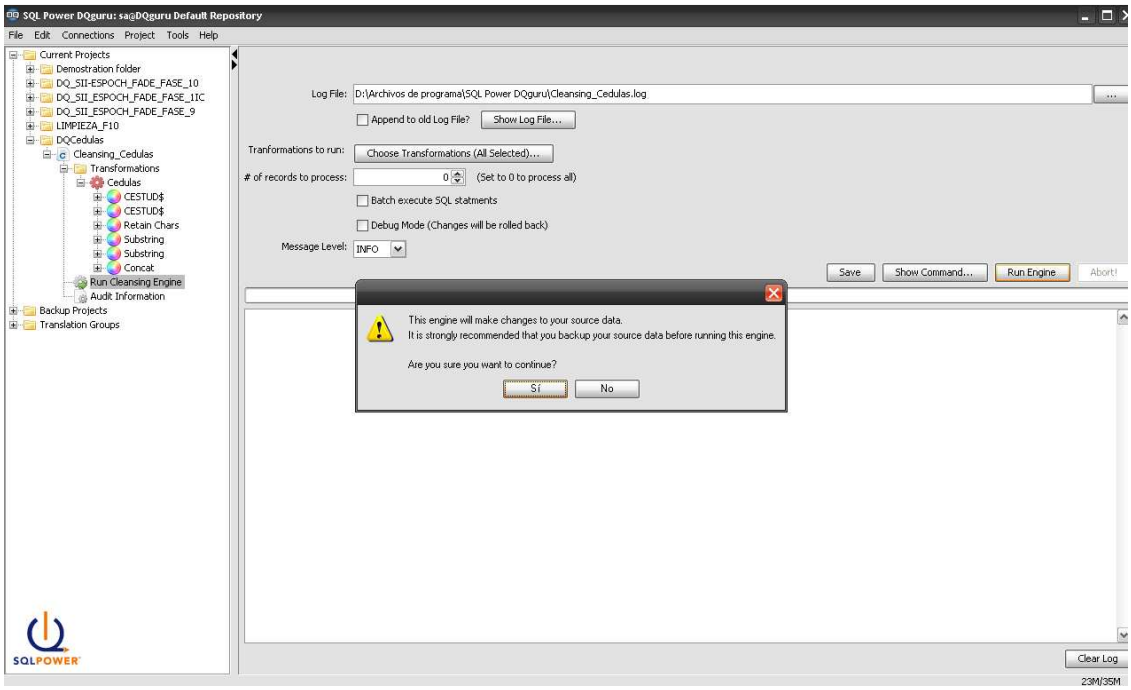


Figura V.116 Ejecución de la transformación

Luego de completar la ejecución de la transformación muestra un mensaje si todo se realizo correctamente:

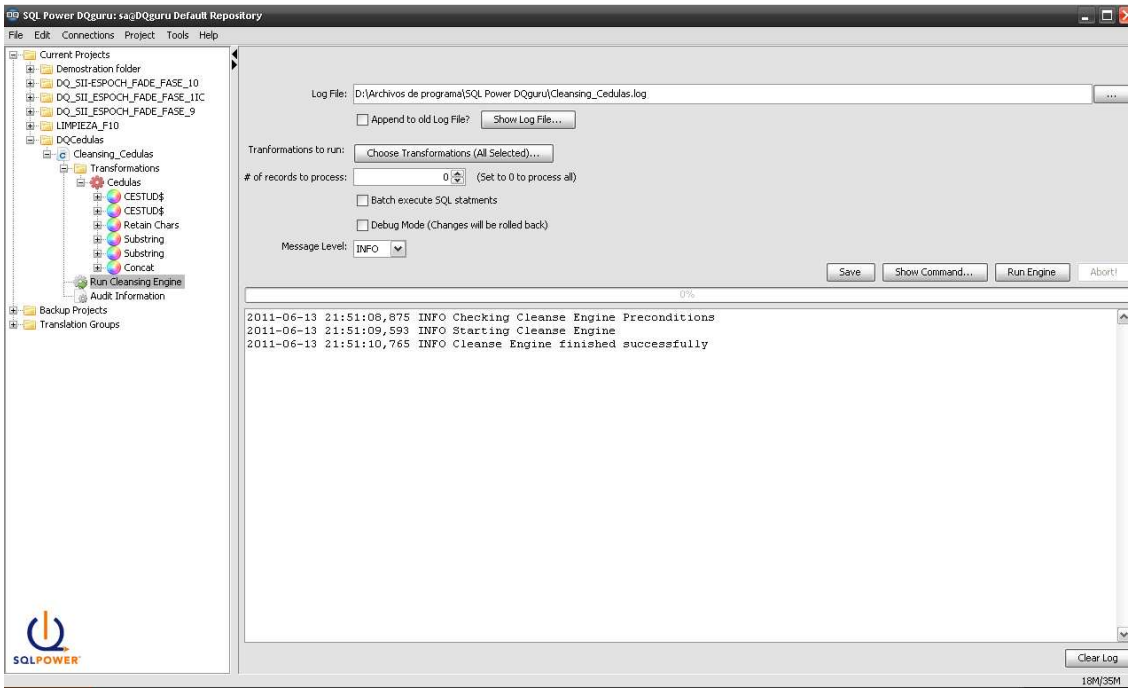


Figura V.117 Ejecución Finalizada

Para los demás requermientos de calidad se utilizara la siguiente herramienta:

- **BayCastle Data Slave MapEditor**

Configuración e Instalación

Requerimientos

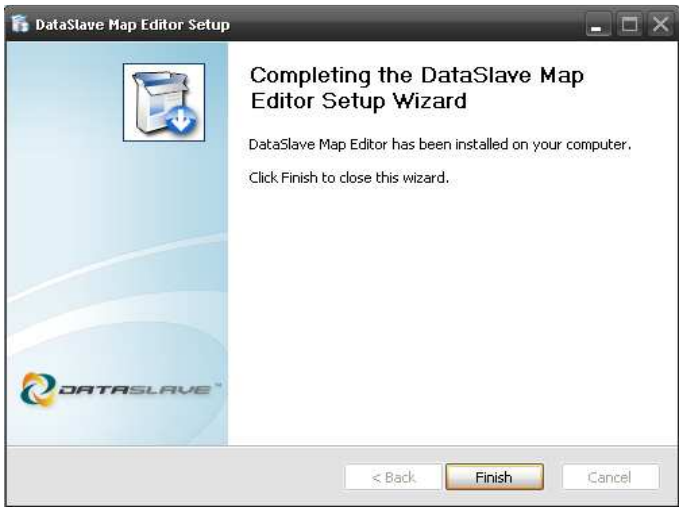
Los requerimientos para instalar es:

- Windows XP, Vista, 2003, 2008, 7
- Dual Core
- 1 GB RAM (2 GB recomendado)
- 40 MB de espacio libre en disco
- .NET Framework 3.51 SP1

Instalación

La instalación es sencilla únicamente se selecciona el botón Next hasta finalizar la instalación

Figura V.118 Instalador DataSlave



FiguraV.119 Finalización de instalación

Configuración

Antes de realizar la transformación de datos se debe realizar la conexión a la base de datos con la que se trabajará en este caso a una base de datos Sql Server por lo que se selecciona el componente **Read MS SQL Server** al área de transformación como se muestra en la siguiente figura:

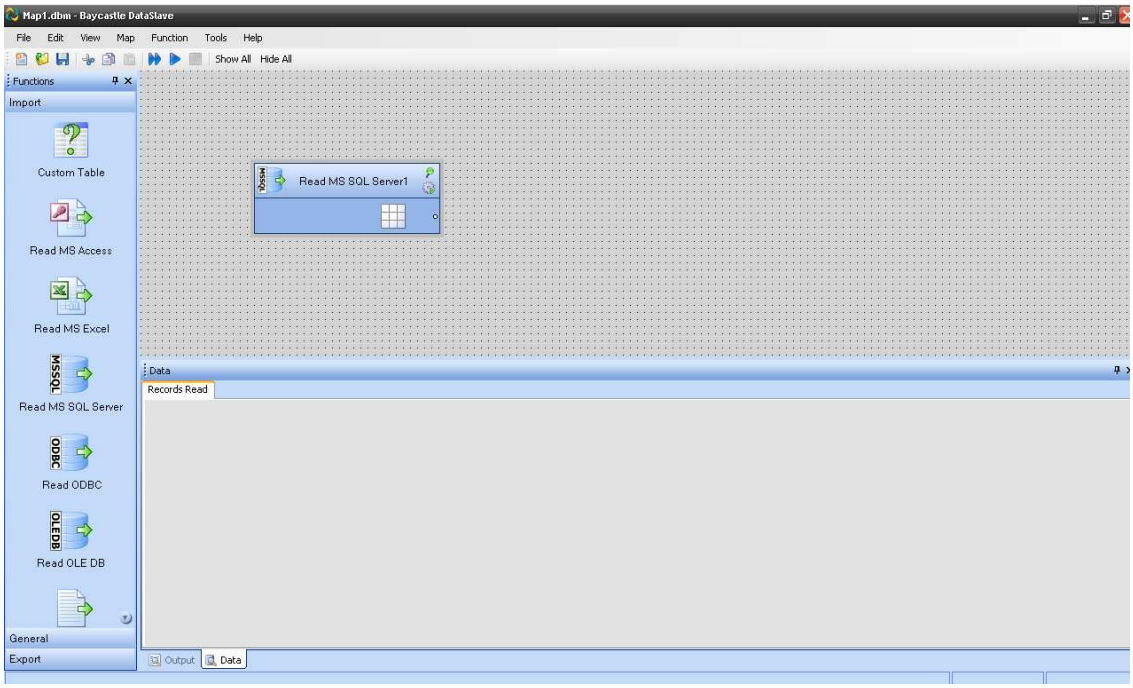


Figura V.120 Configuración de Data Slave

La conexión se realiza al servidor local y se utiliza la cuenta de acceso para este caso tenemos una forma de autenticación mixta así que se utilizara cualquiera de las dos:

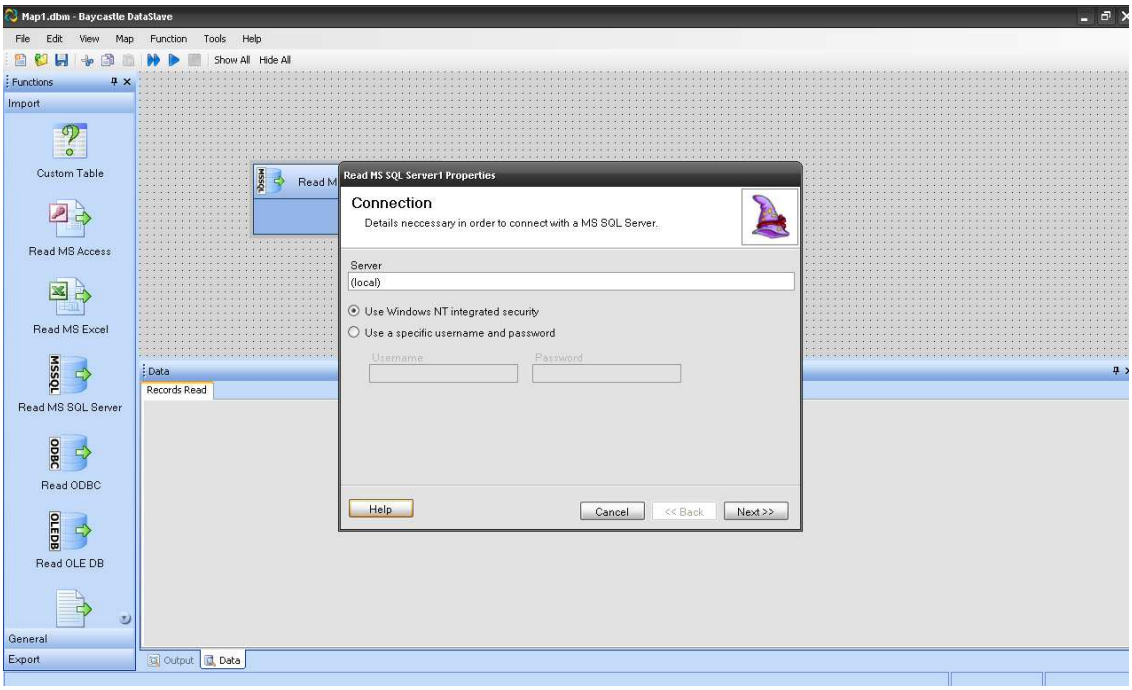


Figura V.121 Conexión al Servidor

Se selecciona Next y se selecciona la base de datos con la que se trabajara como se muestra a continuación:

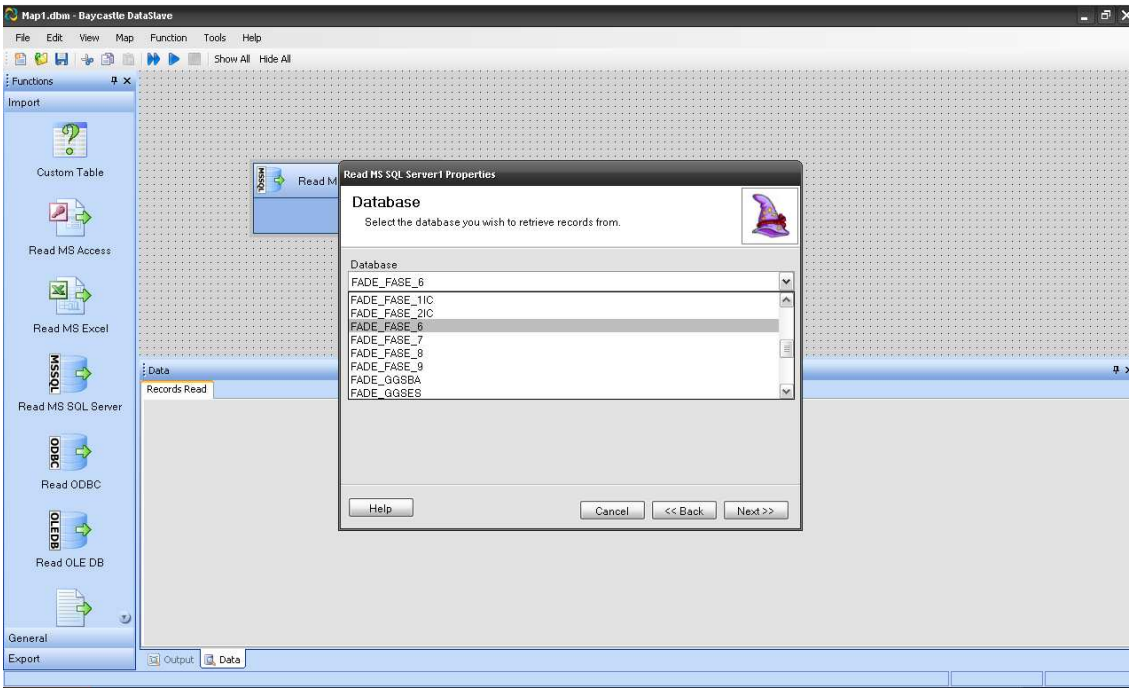


Figura V.122 Bases de datos disponibles

Se selecciona Next y se escoge la tabla que se utilizara para realizar la limpieza como se indica en la siguiente figura:

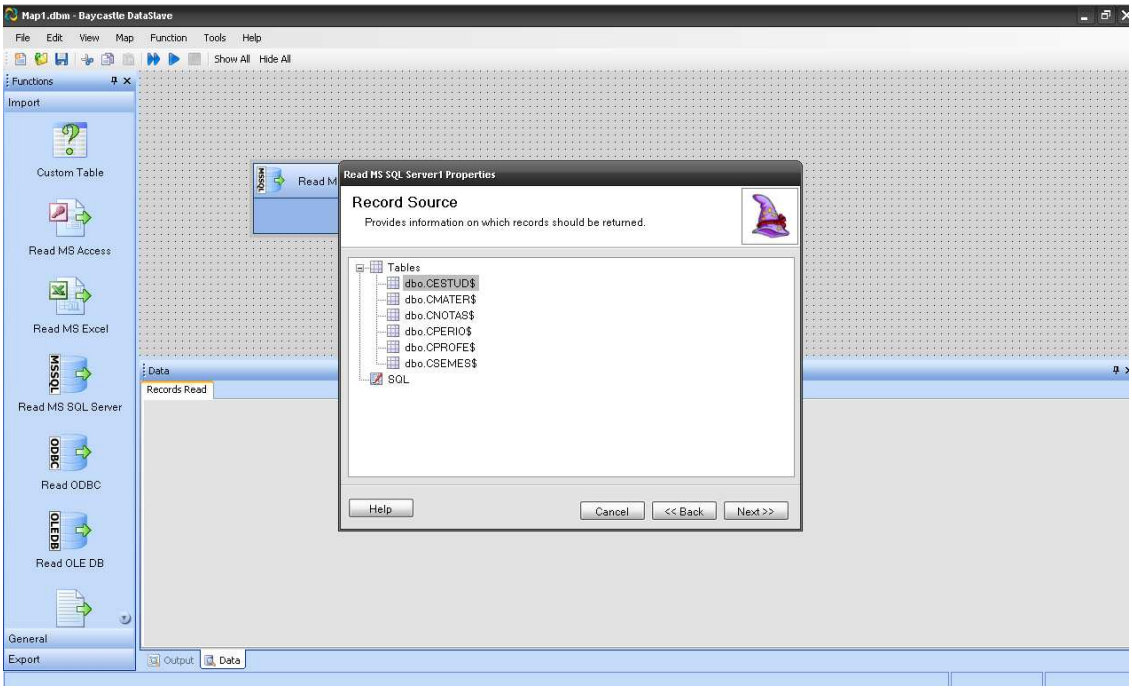


Figura V.123 Origen de registros

Luego se muestra una vista previa de los datos de la tabla seleccionada:

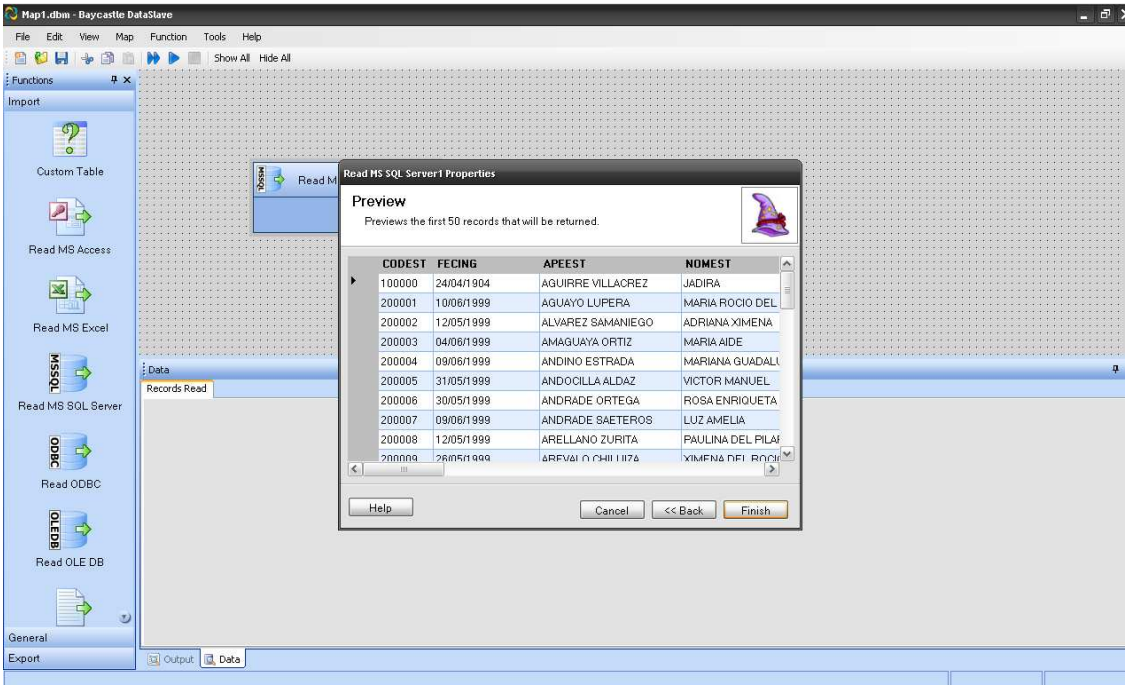


Figura V.124 Datos del Origen seleccionado

Finalizada la conexión se selecciona el recuadro blanco el cual muestra los datos leídos:

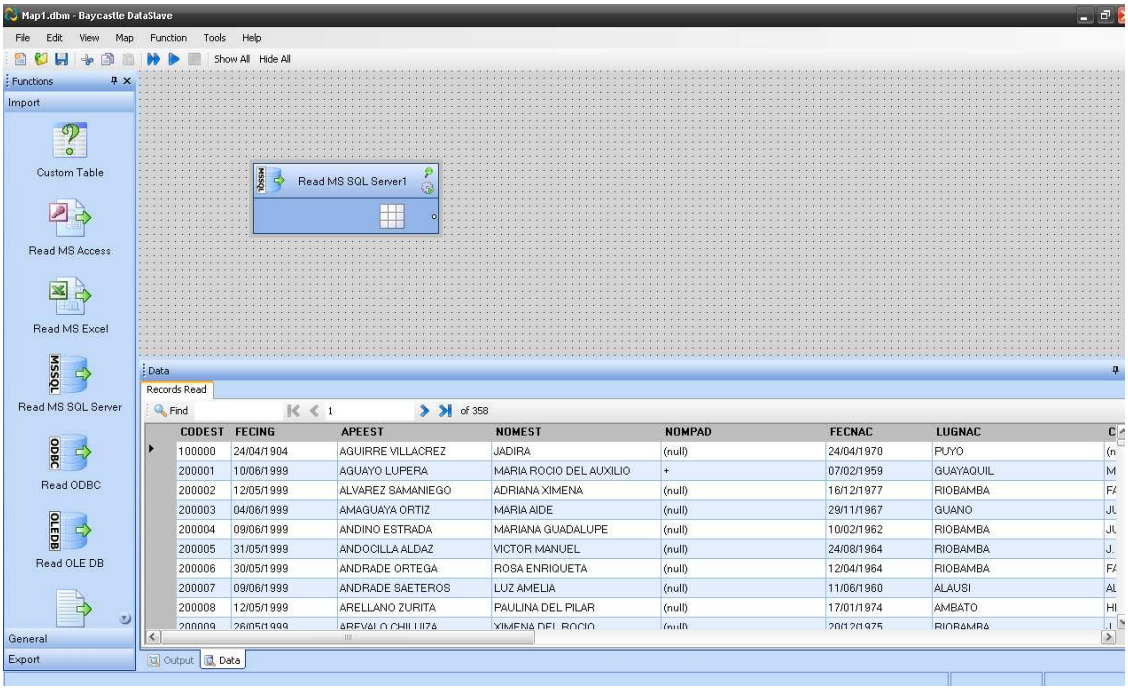


Figura V.125 Datos leídos desde el origen

Ejecución

BayCastle Map Editor se utiliza en los siguientes problemas de calidad encontrados en los datos:

Problema: Datos NULL y blancos en los campos CEDIDE y FECNAC

Para realizar el proceso de transformación de datos se debe utilizar el cuadro de funciones en la ficha General de la herramienta donde podemos utilizar lo siguiente:

Figura V.126 Elementos DataSlave

- **De-duplicación:** Filtra los registros duplicados de una tabla.
- **Merge:**Une los datos de dos tablas en una sola tabla
- **Sort:** Ordena los datos
- **Transform:** Contiene una serie de componentes para la transformación de los datos de entrada para luego producir un conjunto de datos transformados de salida.
- **Validate:**Valida los registros de datos y los ordena en dos tablas de salida aprobados y no aprobados.

Problema: Según el perfilado de datos tenemos Filas NULL completas en los datos

Acción a tomar: Validar únicamente las filas que contienen datos

En la siguiente figura se muestra como se realizo la validación de datos en este caso se toma la condición de que si el nombre y el apellido son NULL entonces no existe estudiante y por tanto la fila no es valida.

En la herramienta se agrega una tarea de validación se selcciona en  y se define estas condiciones arrastrando la función **Is Not NULL** para los campos APEEST y NOMEST para que únicamente filtre los datos que no sean NULL

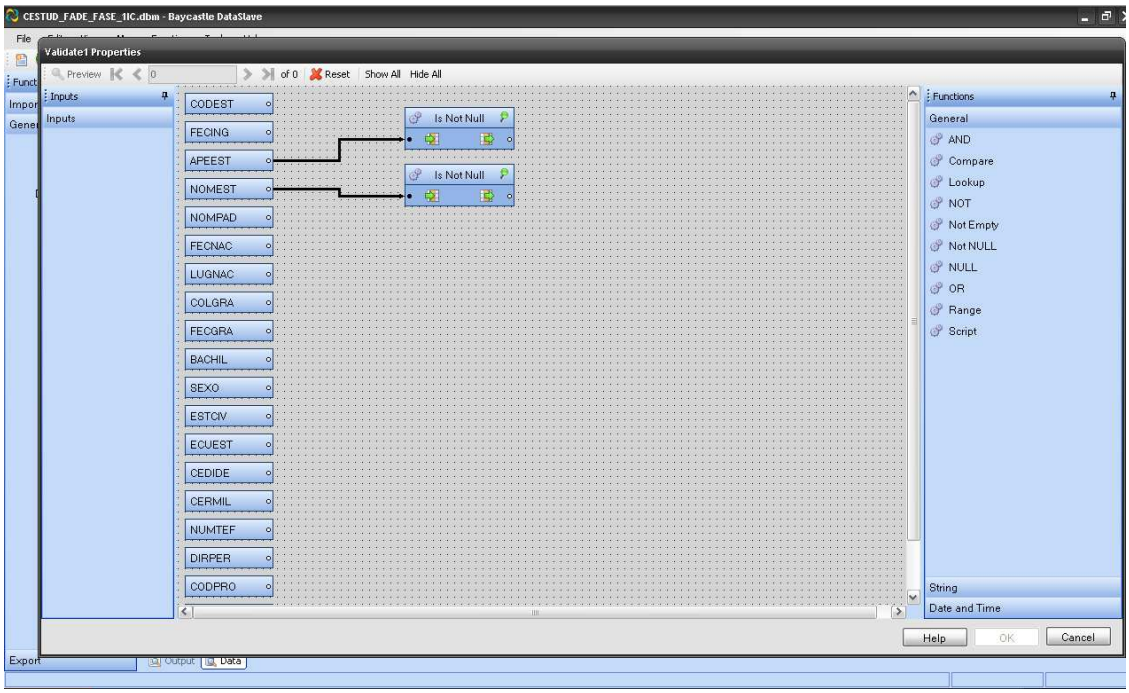


Figura V.127 Validación de registros NULL

Requerimiento de Calidad: El sexo del estudiante cambiar de M y F a MAS y FEM

Luego de realizar la validación correspondiente se procede al proceso de transformación en este caso para cambiar el sexo de M y F a MAS y FEM según corresponda para esto se arrastra la tarea de transformación y se une los registros de salida de la validación

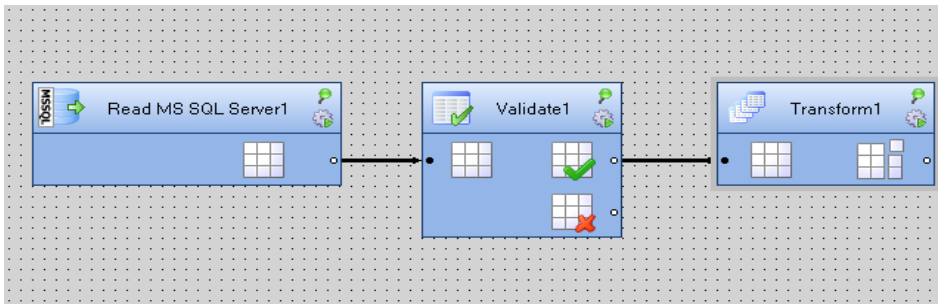


Figura V.128 Esquema de transformación

Se selecciona de la tarea de transformación y se procede a la configuración de la transformación que se quiere realizar.

Primero se realiza un auto-map Auto-Map para obtener las columnas origen y destino, luego en la columna SEXO se agrega una función **if** desde el cuadro de funciones, en la sección **General** seleccionamos y se coloca la siguiente condición:

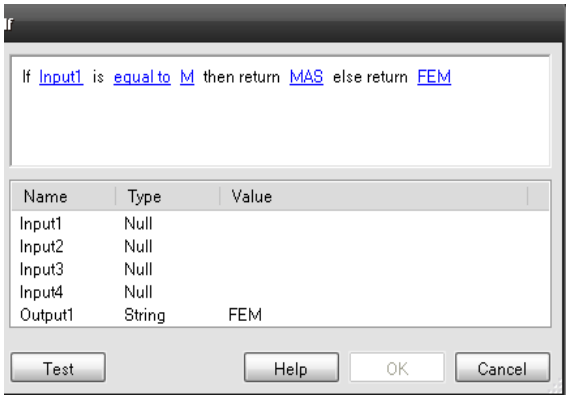


Figura V.129 Condiciones de transformación

Después de colocar la función se envia los resultados de la transformación al destino:

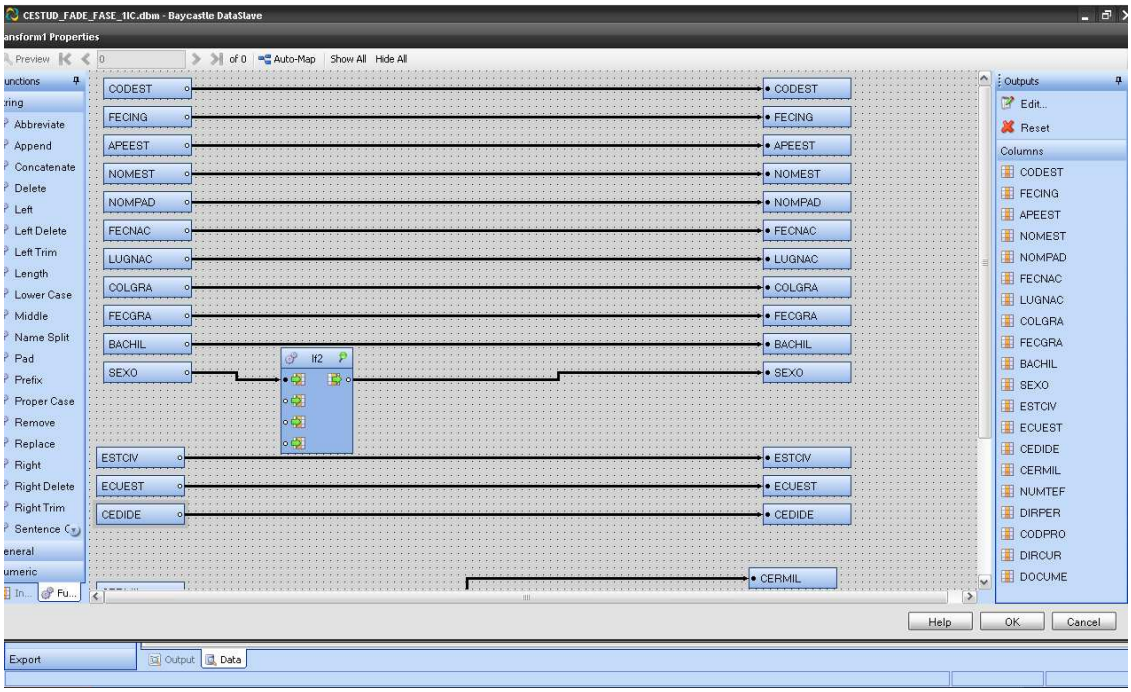


Figura V.130 Condición IF

Requerimiento de Calidad: Si en el campo ESCUEST que se refiere a la nacionalidad cambiar si esta S a ECUATORIANO

Se procede nuevamente en la tarea de transformación a agregar una nueva función llamada **replace** que se encuentra en el cuadro de funciones en la sección String y se agrega al campo ECUESTse selecciona y se coloca la siguiente condición:

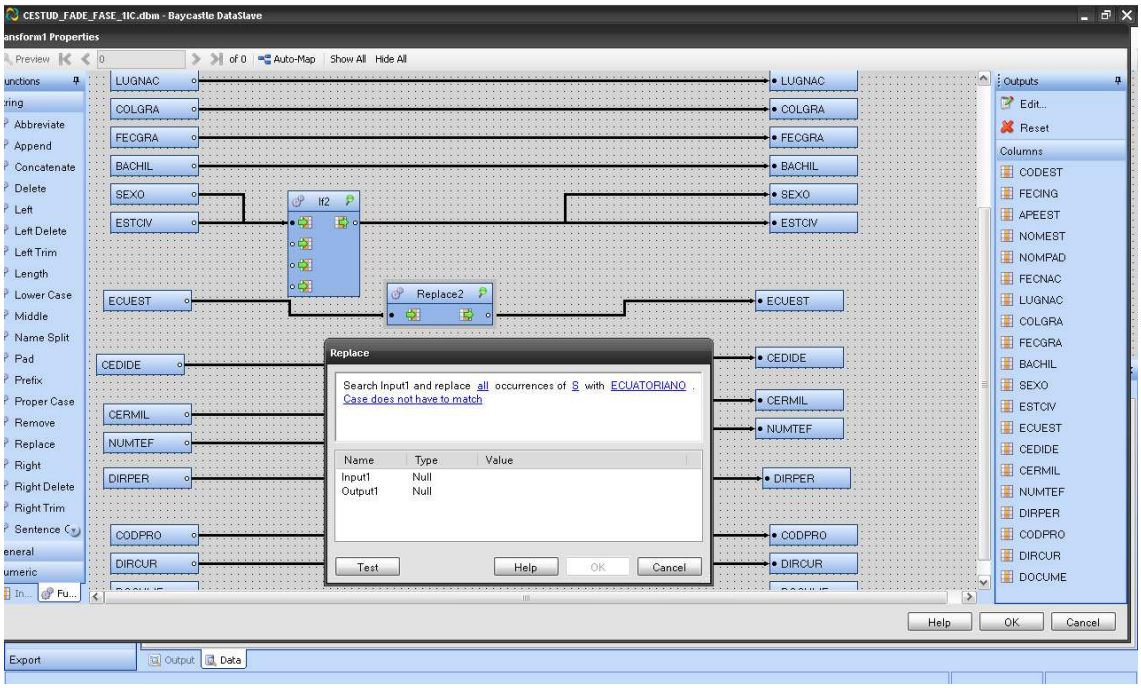


Figura V.131 Transformador Replace

Requerimiento de calidad: Si las fechas de ingreso y fecha de nacimiento del estudiante es NULL cambiarlos a la fecha 01-01-1900, esta fecha será la referencia de que el estudiante no tiene registrado una fecha de ingreso y/o fecha de nacimiento.

Se procede nuevamente en la tarea de transformación a agregar una nueva función **if** que se encuentra en el cuadro de funciones en la sección **General** y lo se lo agrega al campo FECING y al campo FECNAC ,se selecciona y se coloca la siguiente condición:

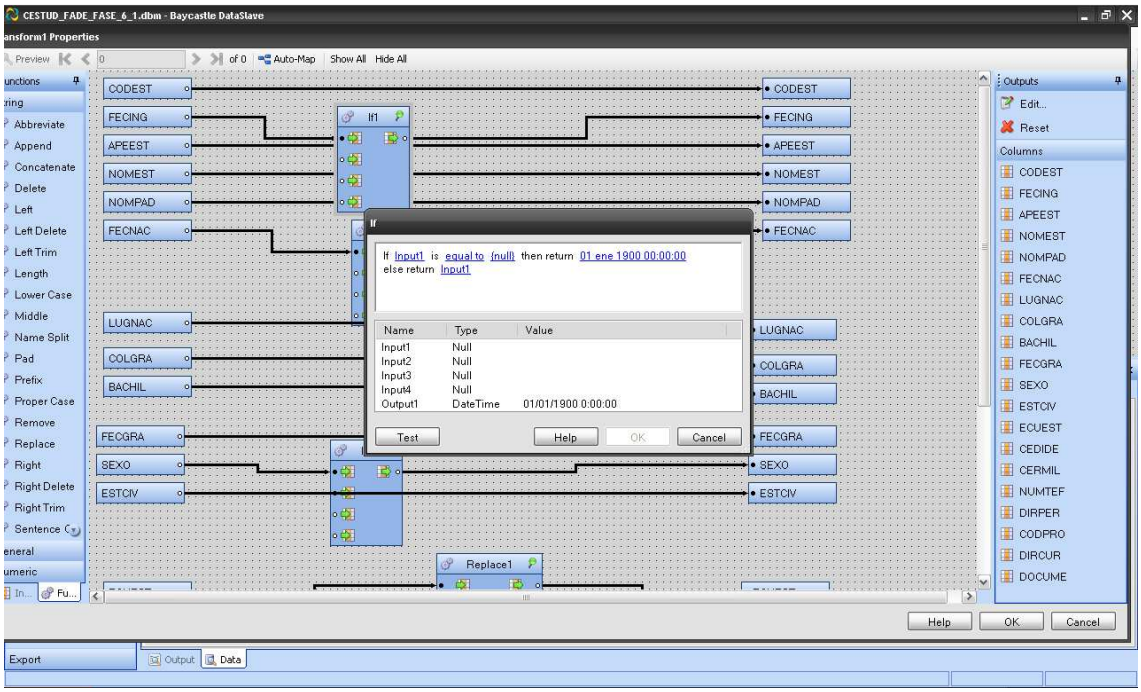


Figura V.132 Condición if para fecha

Requerimiento de calidad: Cambiar fechas inconsistentes a la fecha predeterminada 01-01-1900.

Para tratar las inconsistencias en fechas es de vital importancia utilizar el perfilado de datos de tiempo para poderse guiar que fechas son incorrectas y con esto conseguir un rango de fechas inconsistentes para de estar manera poder realizar lo siguiente:

Escogemos una tarea de validación y se filtra las fechas inconsistentes indicando el rango como se muestra a continuación:

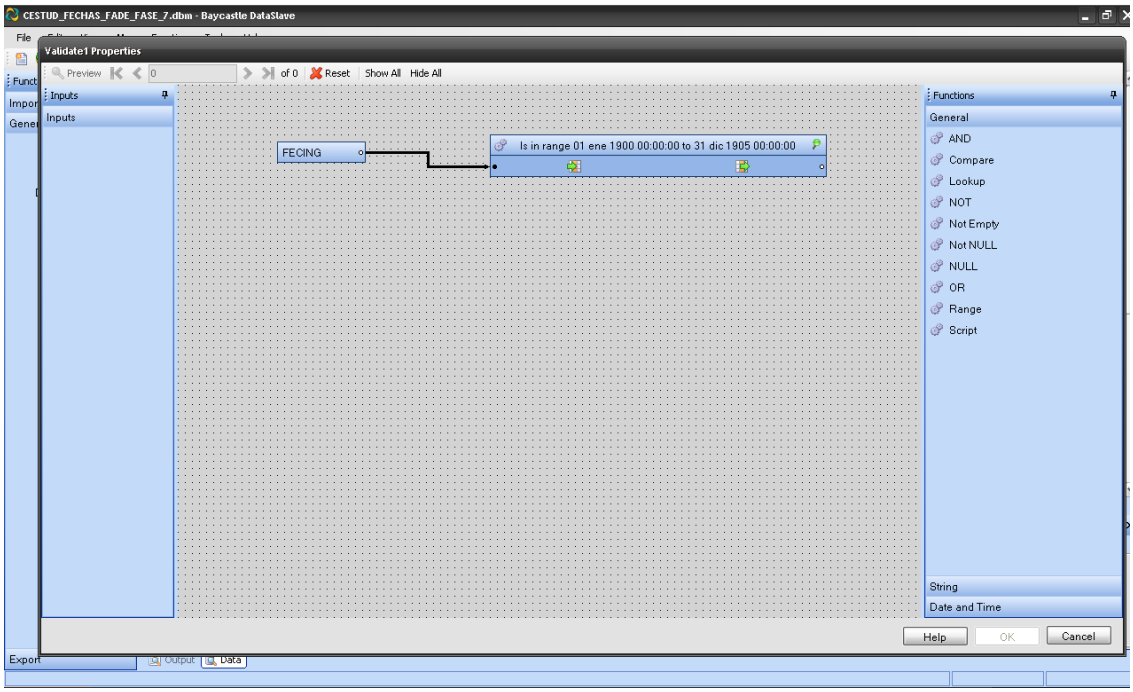


Figura V.133 Validación para fechas

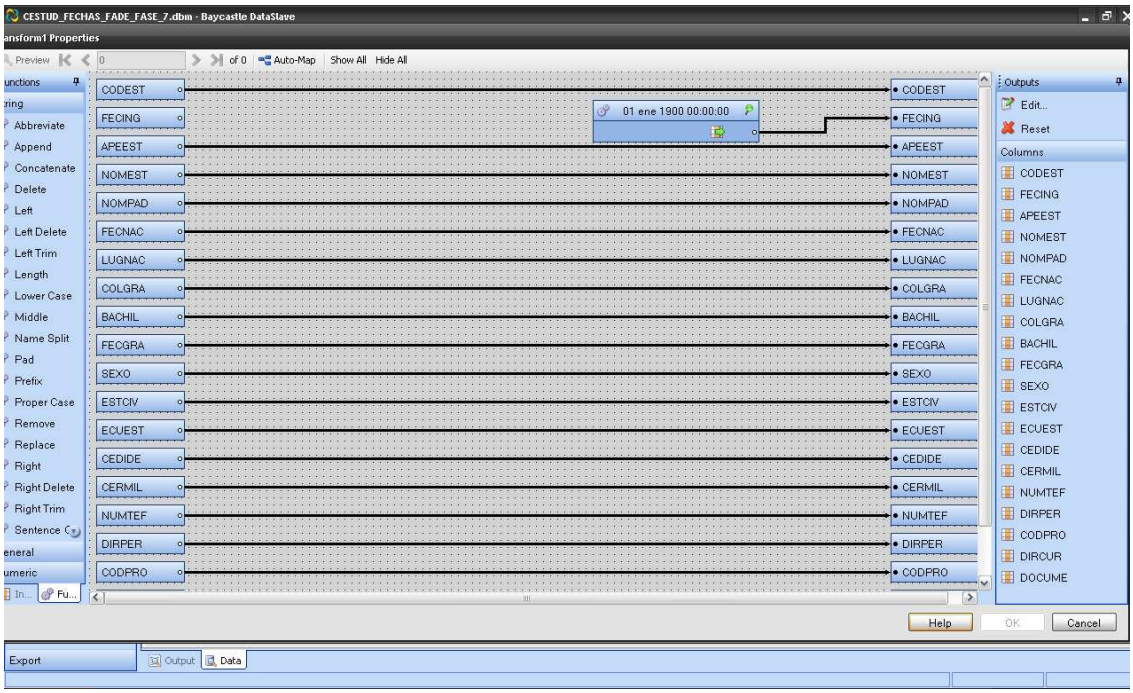


Figura V.134 Valor directo de fecha hacia el destino

Requerimiento de Calidad: No debe haber datos duplicados

Para no tener duplicación se realiza un proceso de matching en el caso de dos fuentes diferentes y datos duplicados numerosos, en este caso luego de la evaluación inicial se observa que la cantidad de duplicados es mínima por lo que se procede a realizar un matching tradicional mediante una sentencia sql delete ,utilizando la técnica de mejor registro.

En la siguiente figura se muestra uno de los casos de duplicación encontrados:

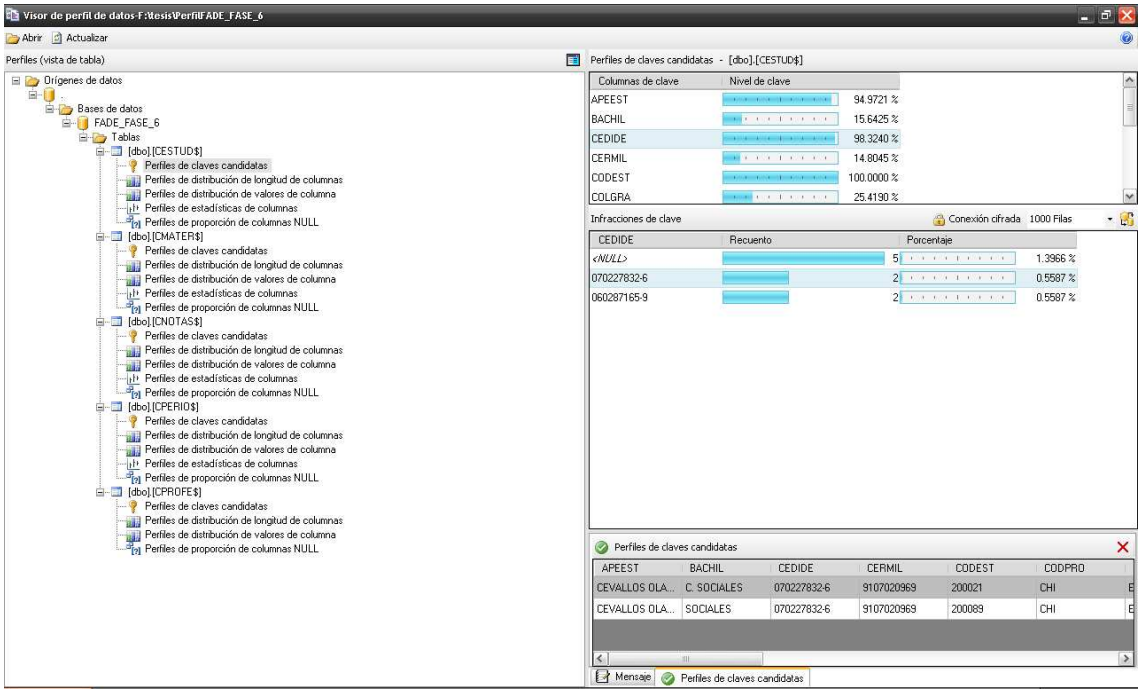


Figura V.138 Duplicados

5.2.5 FASE V. EVALUACIÓN Y ANÁLISIS FINAL DE LA CALIDAD DE DATOS

5.2.5.1 Etapa 5.1 Evaluación Final de los datos

- **Áreas del negocio en que se realizo la limpieza de datos**

Tabla V.159 Áreas del Negocio donde se realizó la limpieza de datos

Área	Fuente de datos	Total de Datos limpiados
Escuela de Unidad de Educación a distancia UED	FADE_FASE_1IC FADE_FASE_6 FADE_FASE_2IC FADE_FASE_7 FADE_FASE_8 FADE_FASE_9 FADE_FASE_10 FADE_GGSBA	1067
Escuela de Ingeniería en Empresas	OAS_IngEmpresas_db	1402
Escuela de Educación para la Salud	OAS_NatPromSalud_db	454
Escuela de Nutrición y Dietética	OAS_Nutricion_db	758
Escuela Ingeniería Agronómica	OAS_IngAgronomica	703
Escuela Ciclo Formativo	OAS_CicloFormativo_db	3219

- **Evaluación final de los Datos**

Para analizar los resultados se realizara un perfilado de los datos después de ejecutado la limpieza.

FADE_FASE_1IC

	CEDIDE	FECNAC	FECING	SEXO	NOMEST	APEEST	CODEST	ECUEST	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**0**	

Tabla V.160 Resultados evaluación Final FADE_FASE_1IC

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_1IC	0	0%	100%

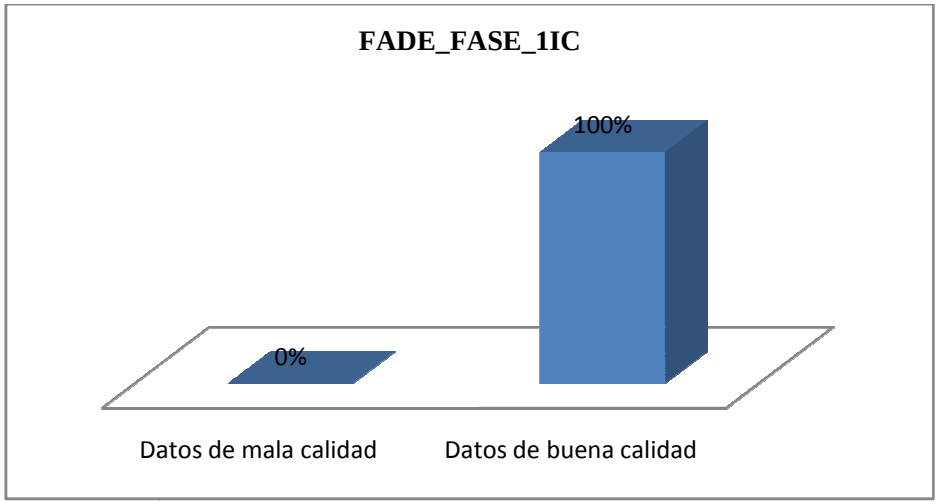


Figura V.135 Resultados evaluación Final FADE_FASE_1IC

FADE_FASE_2IC

Tabla V.161 Evaluación Final FADE_FASE_2IC

	CEDIDE	FECNAC	FECING	SEXO	NOMEST	APEEST	CODEST	ECUEST	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**1**						**1**	**Fechas inconsistentes 01/01/1900**
								Total	**1**	

Tabla V.162 Resultado Evaluacion Final FADE_FASE_2IC

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_2IC	1	0.52%	99.48%

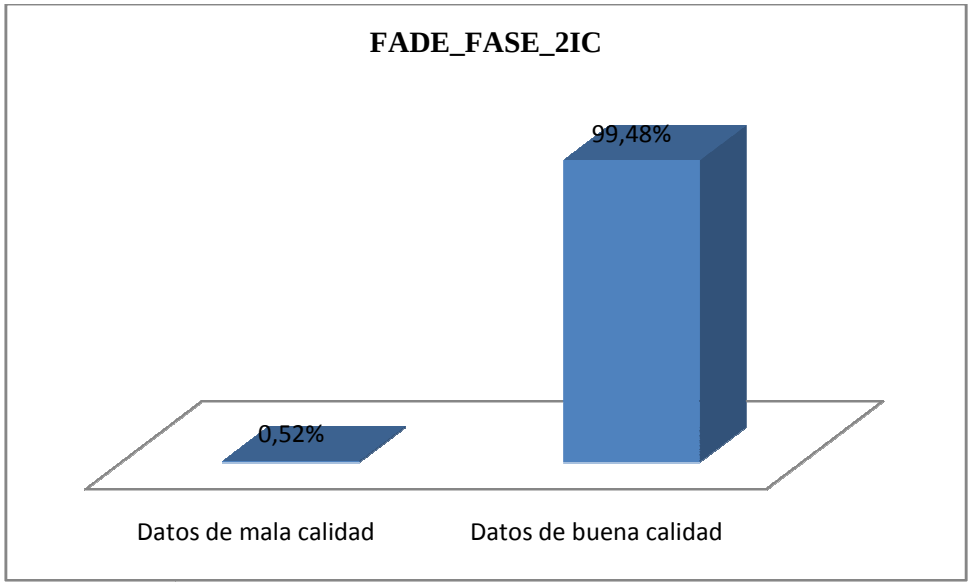


Figura V.136 Resultados Evaluacion Final FADE_FASE_2IC

FADE_FASE_6

Tabla V.163 Evaluacion Final FADE_FASE_6

	CEDIDE	FECNAC	FECING	SEXO	NOMEST	APEEST	CODEST	ECUEST	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	5								**5**	**Duplicación el valor por defecto: 0000000000-0**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**12**	**0**						**12**	**Valor por defecto=01/01/1900**
								Total	**17**	

Total=2864

Tabla V.164 Resultado Evaluacion Final FADE_FASE_2IC

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_6	17	0.6%	99.4%

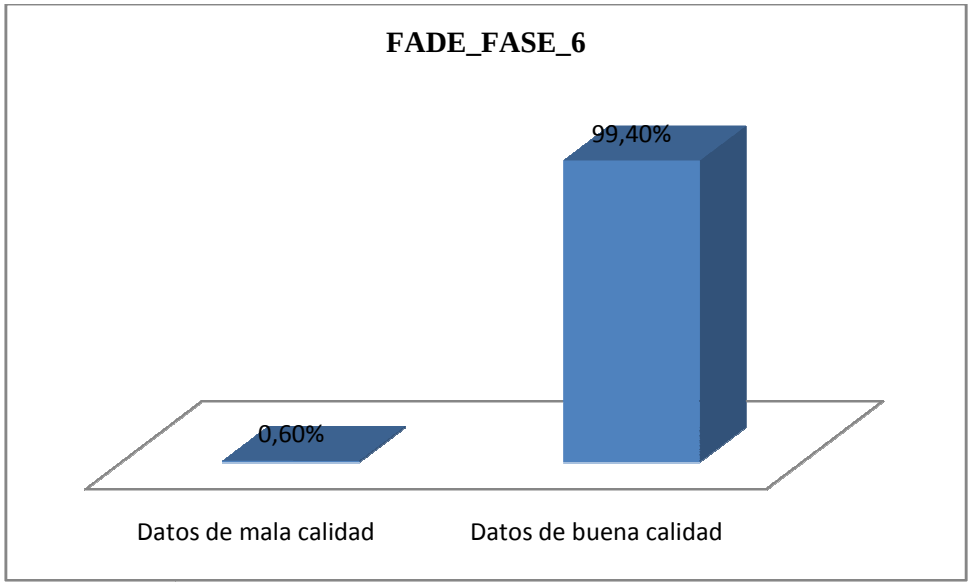


Figura V.137 Resultado Evaluacion Final FADE_FASE_6

FADE_FASE_7

Tabla V.165 Evaluación Final FADE_FASE_7

	CEDIDE	FECNAC	FECING	SEXO	NOMEST	APEEST	CODEST	ECUEST	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	5								**5**	**Valor por defecto =000000000-0**
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	

ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**13**	**282**						**295**	
								Total	**300**	

Resultados:

Total: 3576

Tabla V.166 Resultados Evaluacion Final FADE_FASE_7

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_7	300	8.3%	91.7%

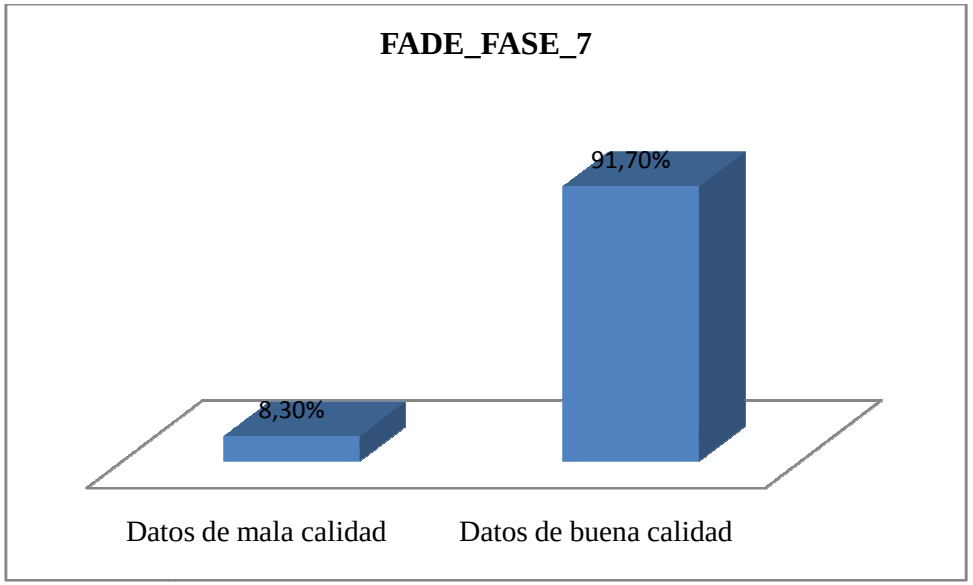


Figura V.138 Resultados Evaluación Final FADE_FASE_7

FADE_FASE_8

Tabla V.167 Evaluación Final FADE_FASE_8

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	

Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**1**	**1**						**2**	**Fechas Inconsistentes: 01/01/1900**
								Total	**2**	

Resultados:

Total: 368

TablaV.168 Resultados Evaluación Final FADE_FASE_8

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_8	2	0.5%	99.5%

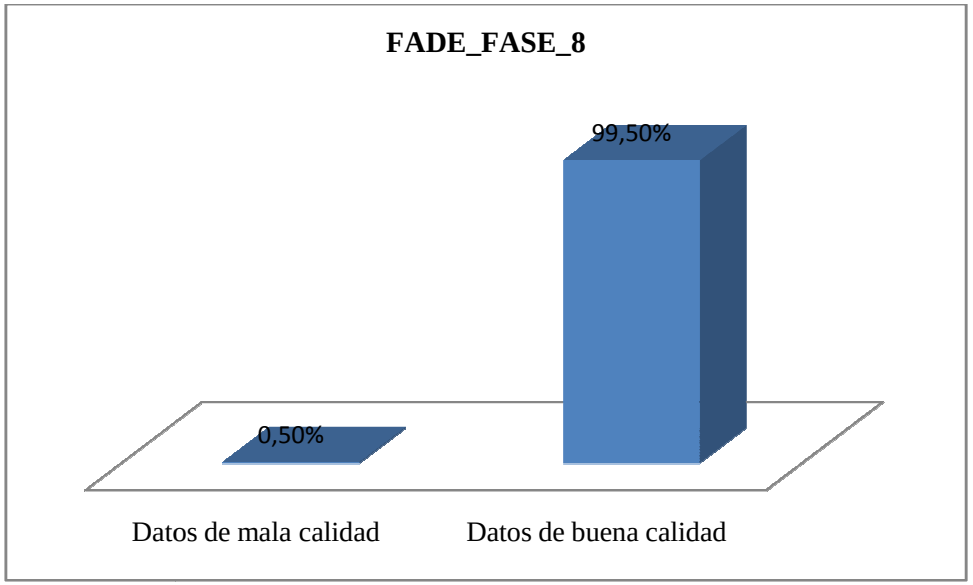


Figura V.139 Resultados Evaluación Final FADE_FASE_8

FADE_FASE_9

Tabla V.169 Evaluación Final FADE_FASE_9

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**									

ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**1**	**47**						**0**	
								Total	**48**	

Resultados:

Total :392

Tabla V.170 Resultados Evaluacion Final FADE_FASE_9

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_9	48	12.24%	87.76%

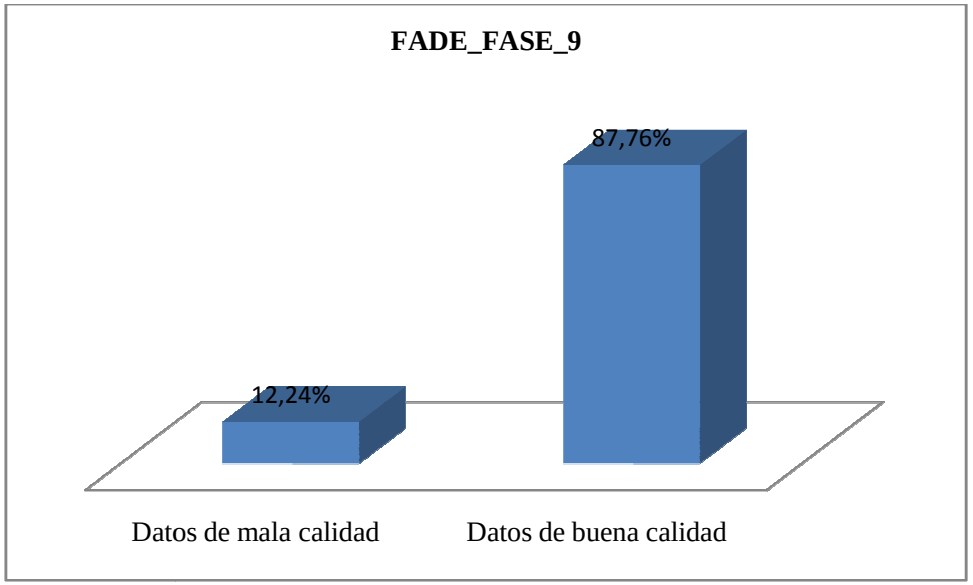


Figura V.140 Resultados Evaluacion Final FADE_FASE_9

FADE_FASE_10

Tabla V.171 Evaluación Final FADE_FASE_10

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	

ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**1**	**51**						**52**	
								Total		

Resultados:

Total :408

Tabla V.172 Resultados Evaluación Final FADE_FASE_10

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_10	52	12.7%	87.36%

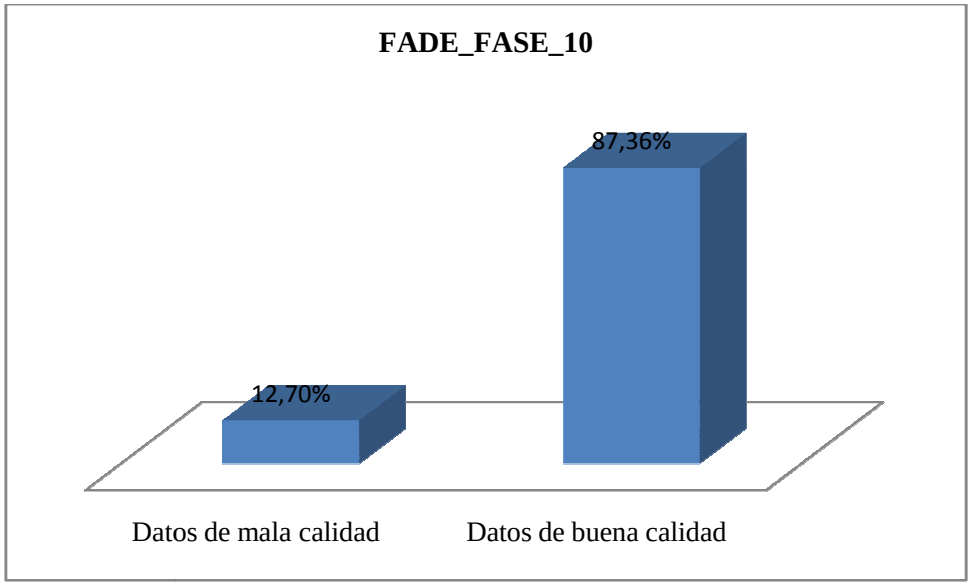


Figura V.141 Resultados Finales FADE_FASE_10

FADE_FASE_GGSBA

Tabla V.173 Evaluación Inicial FADE_FASE_GGSBA

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0									
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**0**	

Resultados:

Total: 296

Tabla V.174 Evaluación Final FADE_FASE_GGSBA

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_GGSBA	0	0%	100%

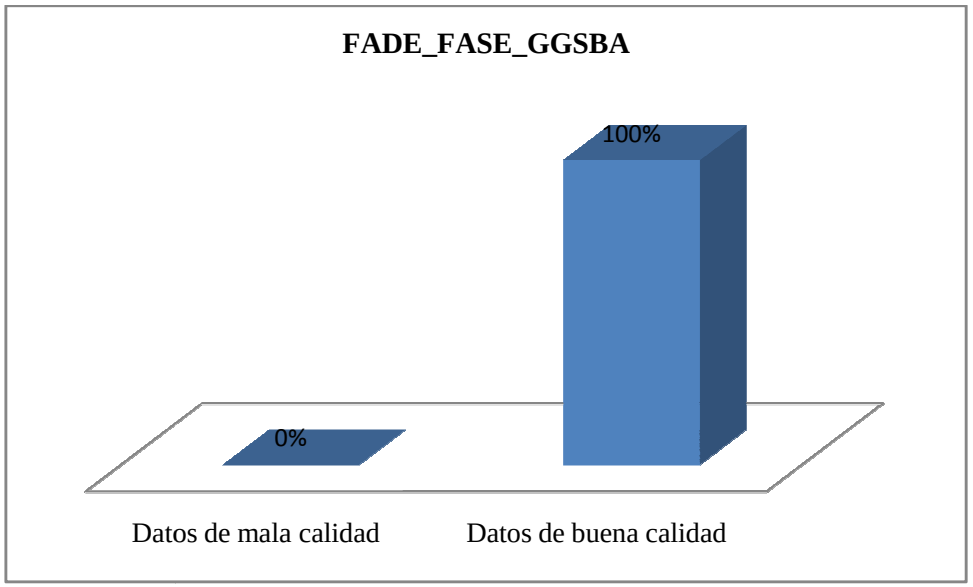


Figura V.142 Resultados Evaluacion Final FADE_FASE_GGSBA

FADE_FASE_GGSES

Tabla V.175 Evaluación Inicial FADE_FASE_GGSES

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST=S								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**44**	

Resultados:

Total: 176

Tabla V.176 Evaluación Final FADE_FASE_GGSES

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
FADE_FASE_GGSES	0	0%	100%

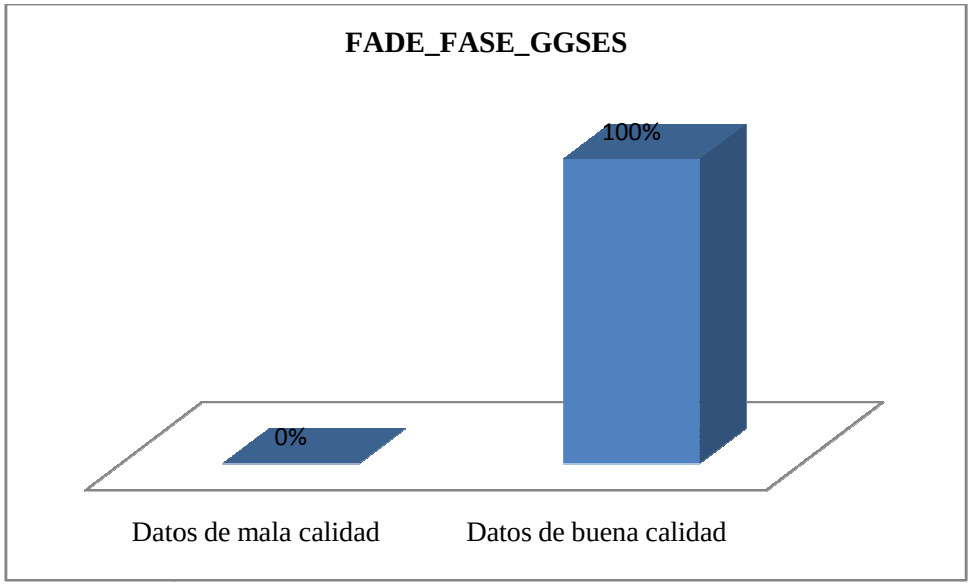


Figura V.143Evaluación Final FADE_FASE_GGSES

Ciclo Formativo

Tabla V.177 Evaluación Final Ciclo Formativo

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**1**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**									
ECUEST<>ECUATORIANO								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**3**	**0**						**3**	
								Total	**3**	

Resultados:

Total: 25752

Tabla V.178 Resultados Evaluación Final Ciclo Formativo

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
Ciclo Formativo	3	0.01%	99.99%

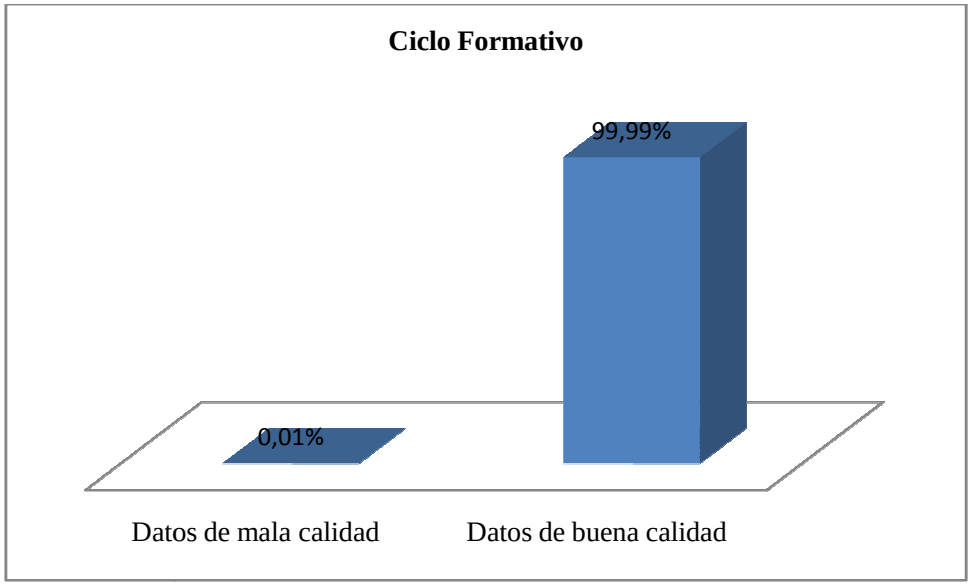


Figura V.144 Resultados Evaluación Inicial

OAS_IngAgronmica

Tabla V.179 Evaluación Final IngAgronomica

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	

Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**4**						**0**	
								Total	**4**	

Resultados:

Tabla V.180 Resultados IngAgronóica

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
IngAgronomica	3	0.07%	99.93%

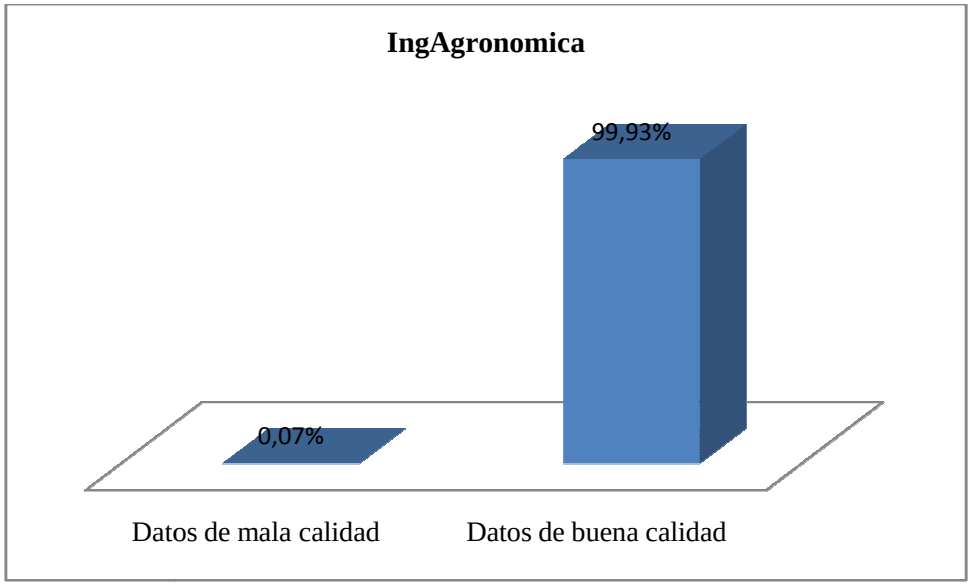


Figura V.145 Resultados Evaluación Final IngAgronomica

OAS_IngEmpresas

Tabla V.181 Evaluación Inicial IngEmpresas

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**			**0**	
Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**7**						**7**	
								Total	**7**	

Resultados:

Total:11216

Tabla V.182 Resultados Evaluación Final IngEmpresas

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad

IngEmpresas	7	0.06%	99.94%

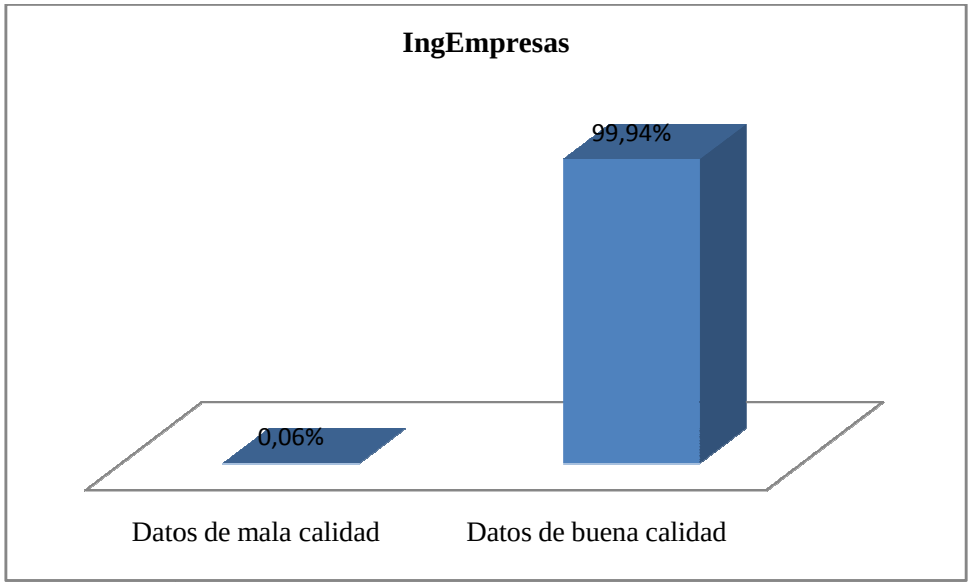


Figura V.146Resultados Evaluación Final IngEmpresas

OAS_NatPromSalud

Tabla V.183 Evaluación Final NatPromSalud

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**0**	**0**						**0**	
								Total	**0**	

Resultados:

Total:3632

Tabla V.184 Resultados Evaluación Final NatPromSalud

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
NatPromSalud	0	0%	100%

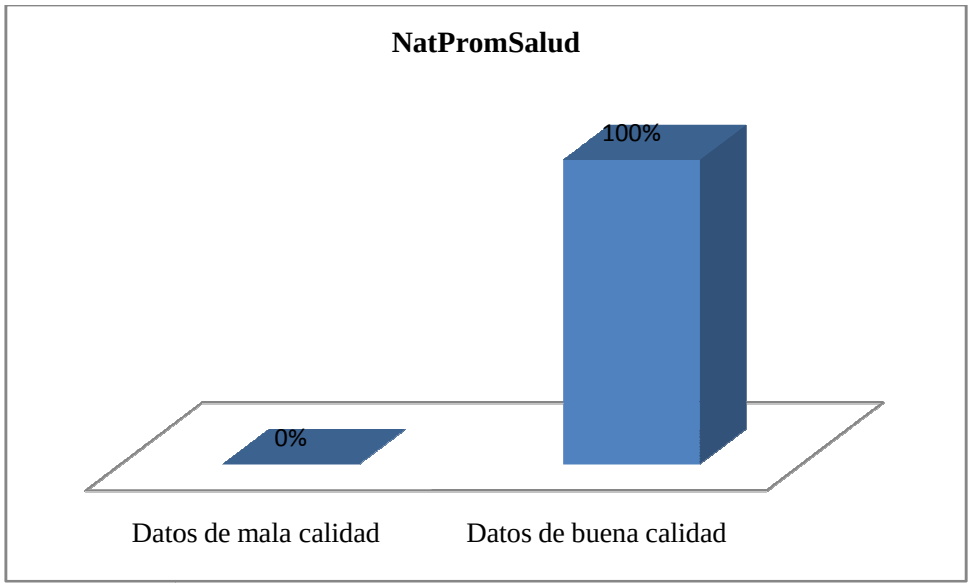


Figura V.147 Resultados Finales NatPromSalud

Nutrición

Tabla V.185 Evaluacion Final Nutricion

	strCedula	dtFechaNac	dtFechaIng	strCodSexo	strNombres	strApellidos	strCodigo	strNacionalidad	Total	Comentarios
NULL	0	**0**	**0**	**0**	**0**	**0**	0	0	0	
Vacíos	0	0	0	0	0	0	0	0	0	
Duplicación	0								**0**	
Minúsculas				**0**	**0**	**0**		**0**	**0**	
Longitud <11	**0**								**0**	
ECUEST<>ECUATORIANO								**0**	**0**	
SEXO=F o M				**0**					**0**	
Inconsistencia de tiempo		**19**	**0**						**19**	
								Total	**19**	

Resultados:

Total:6064

Tabla V.186 Resultados Evaluación Final Nutrición

BD	Total datos de mala calidad	Datos de mala calidad	Datos de buena calidad
Nutrición	19	2.5%	97.5%

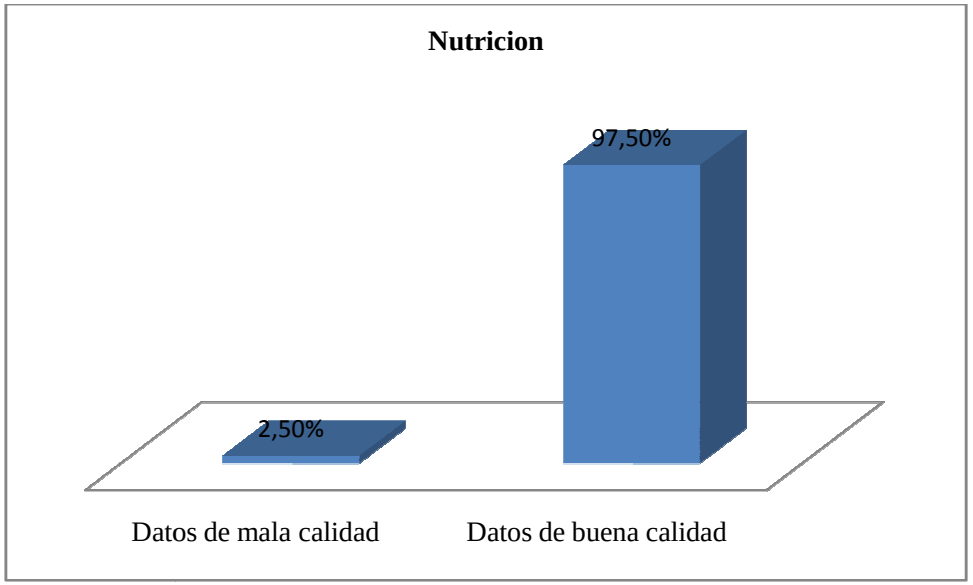


Figura V.148 Resultado Evaluación Final Nutrición

5.2.5.2 Etapa 5.2 Análisis de la Evaluaciónfinal

- **Resultado final de los datos**

Bases de datos de la Escuela de la Unidad de Educación a Distancia

Tabla V.187 Resultados Finales UED

Total	Buena Calidad	Mala Calidad
8536	95.1%	4.9%

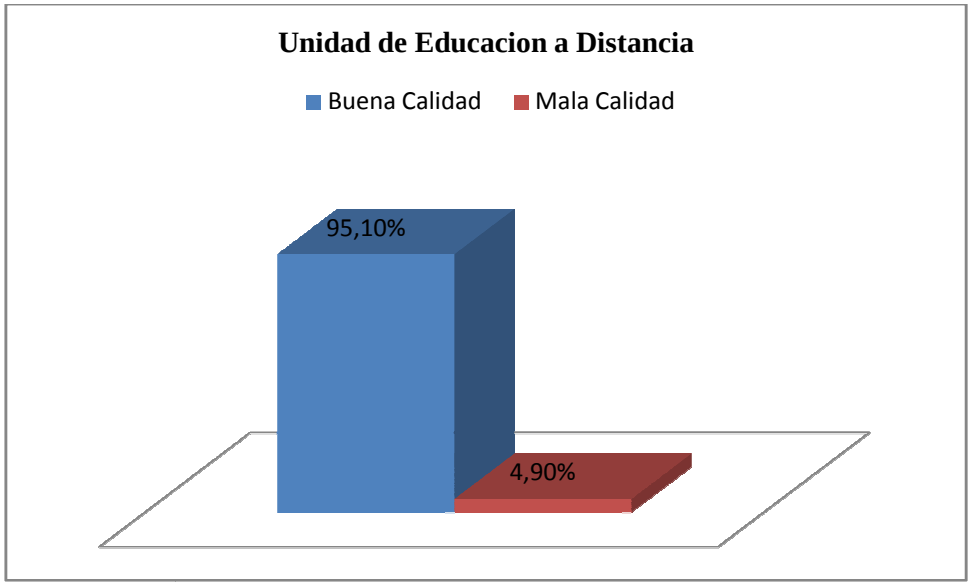
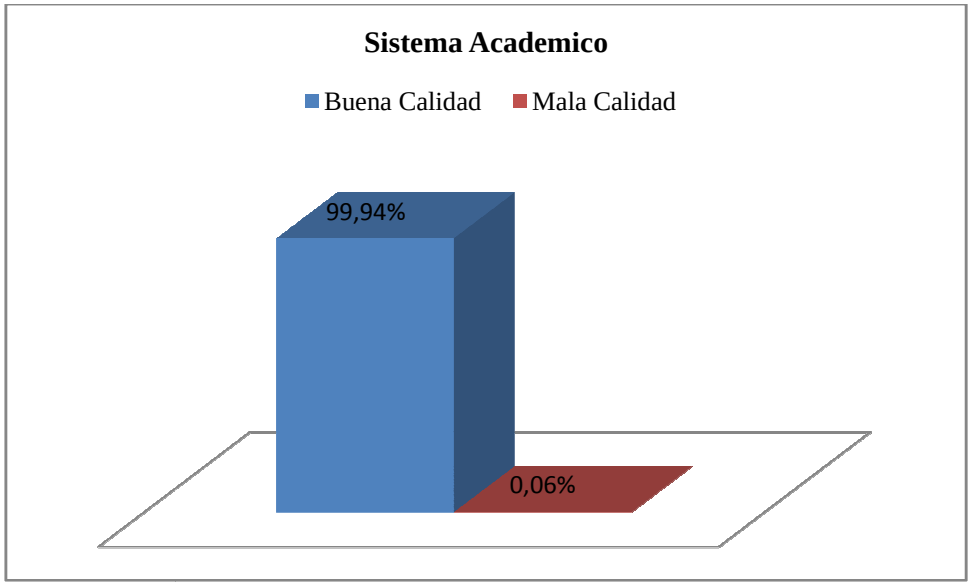


Figura V.149 Resultados Finales UED

Bases de Datos del Sistema Académico

Tabla V.188 Resultados Finales Sistema Académico

Total	Buena Calidad	Mala Calidad
52288	99.94%	0.06%



Análisis:

Al analizar la calidad de datos que se obtuvo al inicio y al final se observa que la calidad de datos en la UED tiene un incremento de 29.61% y del sistema Académico incremento en un 12.36%

5.2.6 FASE VI. MEJORAMIENTO Y PREVENCIÓN

5.2.6.1 Etapa 6.1 Analizar Causas de origen

- **Causas de origen**

Tabla V.189 Causas de Origen

Problema	Causa
Datos NULL y blancos	Falta de información por parte del estudiante
Datos incompletos	Falta de información por parte del estudiante
Datos duplicados	Falta de controles en la aplicación software para el ingreso de datos
Datos sin formatos necesarios	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos
Datos sin un estándar especifico	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos
Inconsistencias en los datos	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos Digitación errónea por parte del personal involucrado

- **Personal-tecnología involucrado con las causas de origen**

Tabla V.190 Personal-Tecnología Involucrado

Causa	Personal	Tecnología
Falta de información por parte del estudiante	Estudiante	
Falta de controles en la aplicación software para el ingreso de datos	Administrador de la base de datos Personal de desarrollo de software de la institución	
Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos	Administrador de base de datos Personal de desarrollo de software de la institución Promotores del desarrollo de software	

5.2.6.2Etapa 6.2 Diseñar plan de mejoramiento

- **Plan de Mejoramiento y prevención**

Tabla V.191 Plan de mejoramiento y prevención

Problema	Posibles acciones	Valoración de la viabilidad de cada acción	Importancia de cada acción para superar la debilidad	Responsable de verificación
Datos NULL y blancos	Desarrollar mejoras en cuanto al proceso de matriculación de estudiantes en los cuales se debe exigir la información completa de los mismos.	100%	Importante para la mejora de datos NULL y blancos encontrados en los datos de los estudiantes	Autoridades de cada escuela de la institución Equipo de desarrollo de software para la institucion Promotores de proyectos de gestión de calidad de datos
Datos incompletos	Desarrollar mejoras en cuanto al proceso de matriculación de estudiantes	100%	Importante para la mejora de datos incompletos encontrados en los datos de los estudiantes	Autoridades de cada escuela de la institución Equipo de desarrollo de software para la institucion

	en los cuales se debe exigir la información completa de los mismos			Promotores de proyectos de gestión de calidad de datos
Datos duplicados	Desarrollar mejoras en cuanto al desarrollo de software de la institución que son utilizados para la manipulación de datos	100%	Importante para la mejora en cuanto a duplicación encontrada en los datos de los estudiantes	Autoridades de cada escuela de la institución Equipo de desarrollo de software para la institucion Promotores de proyectos de gestión de calidad de datos
Datos sin formatos necesarios	Desarrollar mejoras en cuanto al desarrollo de software de la institución que son utilizados para la manipulación de datos	100%	Importante para la mejora de datos sin formatos encontrados en los datos de los estudiantes	Autoridades de cada escuela de la institución Equipo de desarrollo de software para la institucion Promotores de proyectos de gestión de calidad de datos

Datos sin un estándar especifico	Establecer políticas en cuanto a datos antes de realizar el desarrollo de algún software que permita la manipulación de datos	100%	Importante para la mejora de datos sin un estándar especifico encontrados en los datos de los estudiantes	Autoridades de cada escuela de la institución Equipo de desarrollo de software para la institucion Promotores de proyectos de gestión de calidad de datos
Inconsistencias en los datos	Establecer políticas en cuanto a datos antes de realizar el desarrollo de algún software que permita la manipulación de datos	100%	Importante para la mejora de datos inconsistentes encontrados en los datos de los estudiantes	Autoridades de cada escuela de la institución Equipo de desarrollo de software para la institución Promotores de proyectos de gestión de calidad de datos

5.2.7 FASE VII. SEGUIMIENTO Y CONTROL

5.2.7.1 Etapa 7.1 Diseñar Plan de seguimiento y control

- **Plan de seguimiento y control**

Tabla V.192 Plan de seguimiento y control

Problema	Causas de origen	Proceso de control sugerido	Responsable	Frecuencia
Datos NULL y blancos	Falta de información por parte del estudiante	Perfilamiento de datos Análisis de la calidad de datos	Equipo de calidad de datos conformado para el desarrollo de SII -ESPOCH	Semestral
Datos incompletos	Falta de información por parte del estudiante	Perfilamiento de datos Análisis de la calidad de datos	Equipo de calidad de datos conformado para el desarrollo de SII -ESPOCH	Semestral
Datos duplicados	Falta de controles en la aplicación software para el ingreso de datos	Perfilamiento de datos Análisis de la calidad de datos	Equipo de calidad de datos conformado para el desarrollo de SII -ESPOCH	Semestral
Datos sin formatos necesarios	Falta de políticas establecidas antes de la construcción del software para el ingreso de datos y del diseño de la base de datos	Perfilamiento de datos Análisis de la calidad de datos	Equipo de calidad de datos conformado para el desarrollo de SII -ESPOCH	Semestral
Datos sin un estándar especifico	Falta de políticas establecidas antes de la	Perfilamiento de datos Análisis de la	Equipo de calidad de datos conformado para el desarrollo de SII	

	construcción del software para el ingreso de datos y del diseño de la base de datos	calidad de datos	-ESPOCH	Semestral
Inconsistencias en los datos		Perfilamiento de datos Análisis de la calidad de datos	Equipo de calidad de datos conformado para el desarrollo de SII -ESPOCH	Semestral

5.3 RESULTADOS DE LA METODOLOGIA PROPUESTA

Luego del desarrollo y aplicación de la metodología para SII-ESPOCH, se presenta a continuación dos tablas de resumen de la calidad inicial y final

Tabla V.193 Resumen Evaluación Inicial

EVALUACION INICIAL			
Fuente de Datos	**Total Mala Calidad**	**% Total Mala Calidad**	**%Total Buena Calidad**
FADE_FASE_1IC	108	40.9%	59.1%
FADE_FASE_2IC	49	25%	75%
FADE_FASE_6	946	33%	67%
FADE_FASE_7	1212	34%	66%
FADE_FASE_8	174	47%	53%
FADE_FASE_9	156	40%	60%
FADE_FASE_10	155	38%	62%
GGSBA	102	35%	65%
GGSES	44	25%	75%
CICLO FORMATIVO	3263	13%	87%
INGAGRONOMICA	596	11%	89%
INGEMPRESAS	1396	13%	87%
NATPROMSALUD	454	13%	87%
NUTRICION	789	13%	87%

Tabla V.194 Resumen Evaluación Final

EVALUACION FINAL			
Fuente de Datos	**Total Mala Calidad**	**% Total Mala Calidad**	**%Total Buena Calidad**
FADE_FASE_1IC	0	0%	100%
FADE_FASE_2IC	1	0.52%	99.48%
FADE_FASE_6	17	0.6%	99.4%
FADE_FASE_7	300	8.3%	91.7%
FADE_FASE_8	2	0.5%	99.5%
FADE_FASE_9	48	12.24%	87.76%
FADE_FASE_10	52	12.7%	87.3%
GGSBA	0	0%	100%
GGSES	0	0%	100%
CICLO FORMATIVO	3	0.01%	99.9%
INGAGRONOMICA	4	0.07%	99.93%
INGEMPRESAS	7	0.06%	99.94%
NATPROMSALUD	0	0%	100%
NUTRICION	19	2.5%	97.5%

Resultados:

Para realizar el análisis correspondiente de los resultados obtenidos de la aplicación de la metodología propuesta se lo hará basándose en la FaseIII y Fase V que corresponde a la Evaluación y Análisis inicial y final de la calidad de datos en la cual se determina que la calidad de datos que se obtuvo al final de aplicada la metodología es la siguiente:

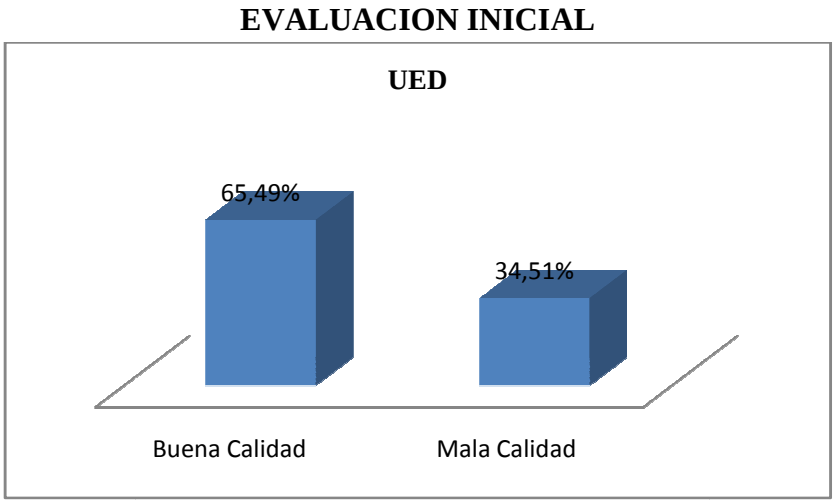


Figura V.150 Resultados UED Evaluación Inicial

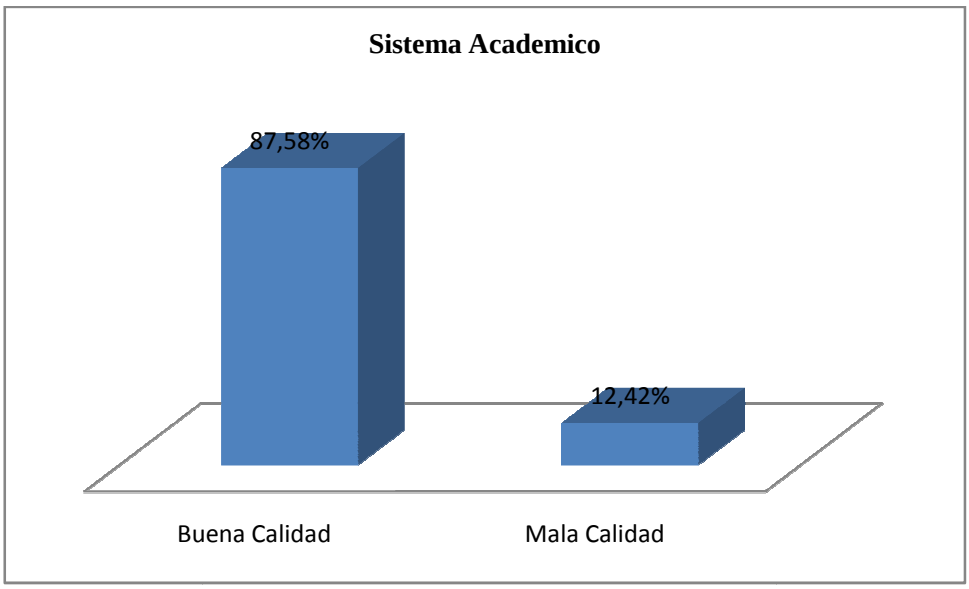


Figura V.151 Resultados Sistema Académico Inicial

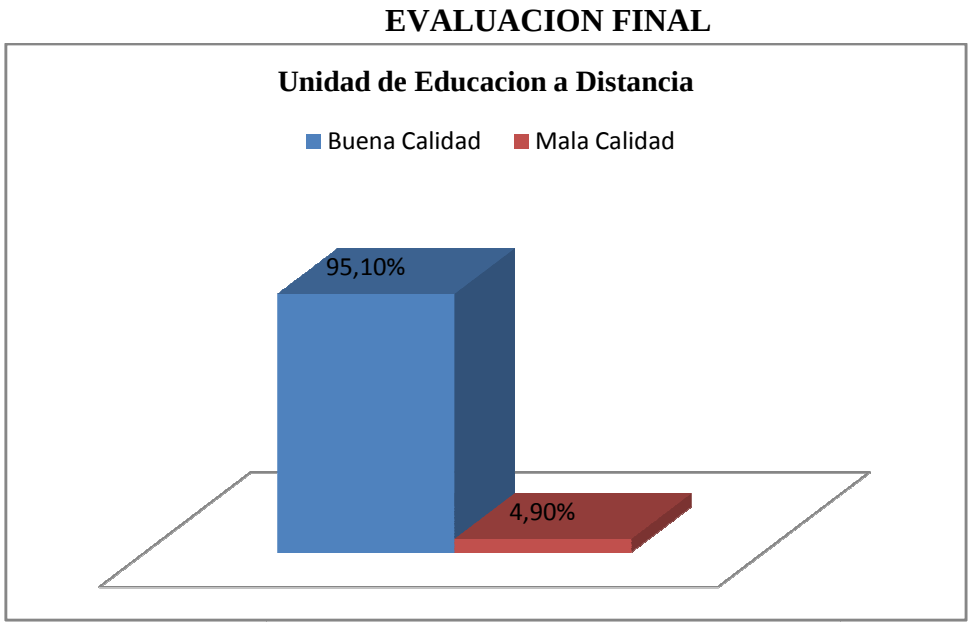


Figura V.152 Resultados Evaluación Final UED

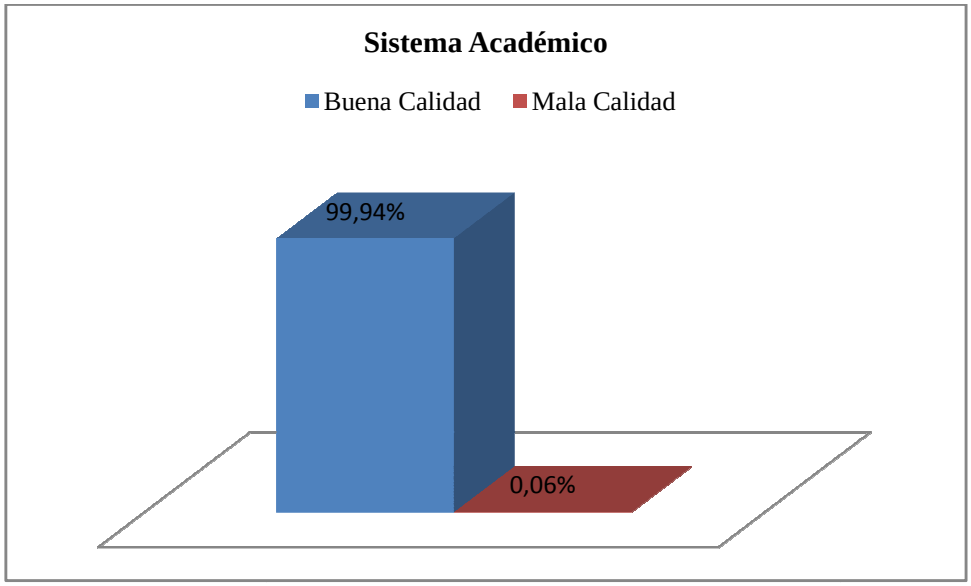


Figura V.153 Resultado Evaluación Final Sistema Académico

Por los resultados obtenidos se puede observar que en la unidad de educación a distancia la calidad de datos tiene un incremento de 29.61% teniendo como resultado final 95.10% de datos de buena calidad y 4.90% datos de mala calidad .En cuanto al sistema académico la calidad de datos incremento en un 29.61% teniendo como resultado final 99.94% datos de buena calidad y 0.06% datos de mala calidad. La cantidad de datos de mala calidad todavía existente se debe a las causas de origen analizadas en la FASE VI de Mejoramiento y prevención y no a la aplicación de la metodología. Por lo tanto la propuesta metodológica permite asegurar la calidad de datos para el proyecto de integración SII-ESPOCH.

CONCLUSIONES

- La influencia de la calidad de datos en proyectos de integración en la actualidad es de mucha importancia, ya que todo negocio empresa u organización ya sea pequeña mediana o grande ha visto la necesidad de contar con un proceso o metodología para la gestión de calidad de datos; la cual permite obtener información confiable para la correcta toma de decisiones.

- La propuesta metodológica para la gestión de calidad de datos se construyó en base a los procesos actuales para calidad de datos los cuales cuentan con pasos básicos para la gestión de los mismos; la metodología los consolida agregando una correcta estructuración y control de los mismos logrando asegurar la calidad de datos al final de su aplicación.

- La metodología propuesta para la gestión de calidad de datos se ha aplicado en las bases de datos del Sistema Académico Institucional "OASIS" y de la Unidad de Educación a Distancia de la ESPOCH, como parte del proyecto de integración SII-ESPOCH.

- Actualmente muchas empresas dedicadas a business intelligence se han enfocado en la parte de gestión de calidad de datos desarrollando procesos de gestión que conlleven su correcta actualización, además de desarrollar productos software tanto open source como propietarios los cuales todavía son pocos en el mercado teniendo algunos de estos algunas limitaciones.

- La metodología en cada una de sus fases sugiere técnicas y herramientas de software ,en el proyecto se aplicaron las siguientes :para la Fase 3 de Evaluación y Análisis Inicial de la Calidad de Datos se utilizo la herramienta Data Profile Task de Sql Server 2008 y la herramienta Data Cleaner con el propósito de realizar el perfilado de datos es decir medir la calidad, para la Fase 4 de

Limpieza de Datos se utilizo la herramienta BayCastle MapEditor y Sql Power DQGuru con el propósito de cumplir con la limpieza de datos establecidos en los requerimientos del negocio.

- Aplicada la metodología en el proyecto SII-ESPOCH en la Unidad de Educación a Distancia se obtuvo un incremento del 29.61% de calidad de datos teniendo como resultado final 95.10% de datos de buena calidad y 4.90% datos de mala calidad .En cuanto al sistema académico la calidad de datos se incremento en un 29.61% teniendo como resultado final 99.94% datos de buena calidad y 0.06% datos de mala calidad. Asegurando con esto la calidad de los datos para el proyecto de integración SII-ESPOCH

RECOMENDACIONES

- Contar con calidad de datos en un negocio empresa u organización no es únicamente utilizar una herramienta para calidad de datos por lo que es recomendable utilizar una metodología completa que involucre todo el negocio, para que tener calidad de datos no sea únicamente para un tiempo limitado.

- La metodología propuesta puede ser aplicada para cualquier proyecto de integración por lo que se recomienda trabajar en forma ordenada siguiendo todas sus fases, etapas y las diferentes actividades que se desarrolla en las mismas permitiendo esto obtener buenos resultados.

- Son muy pocos el software desarrollado para calidad de datos actualmente por lo que es recomendable ir buscando nuevas alternativas en cuanto a las herramientas ya que en los próximos años la cantidad de herramientas se ira incrementando, y de las herramientas con los que se cuenta en la actualidad es recomendable ir actualizando sus versiones.

- Se recomienda realizar una tesis que efectúe un estudio comparativo de herramientas de software para la calidad de datos, lo cual servirá para tener una instrumento base de utilización en la gestión de calidad de datos.

RESUMEN

Se desarrolló una propuesta metodológica para la gestión de calidad de datos en proyectos de integración para ser aplicado en el SII-ESPOCH (Sistema de Información Institucional de la Escuela Superior Politécnica de Chimborazo).

La propuesta consiste en cumplir siete fases, la primera para estudio y preparación del proyecto; la segunda para análisis de la información, la tercera consiste en evaluación y análisis de datos; la cuarta es limpieza de datos, a continuación en la quinta fase se realiza una evaluación y análisis final de la calidad de datos, continuando con un mejoramiento y prevención y el control de la calidad de datos que corresponde a la sexta y séptima fase.

Para el desarrollo de la investigación se utilizó las siguientes herramientas: Microsoft Sql Server 2008 Business Intelligence Development Studio, Data Cleaner, BayCastle Data Slave Map Editor, Sql Power DQGuru.

Aplicando esta metodología propuesta se logró que la unidad de educación a distancia tenga un incremento de 29.61% teniendo como resultado final 95.10% de datos de buena calidad y 4.90% datos de mala calidad .En cuanto al sistema académico, la calidad de datos se incrementó en un 29.61% y como resultado final 99.94% datos de buena calidad y un 0.06% datos de mala calidad

La propuesta metodológica asegura la calidad de datos en proyectos de integración, en este caso para el SII-ESPOCH. Es recomendable seguir en orden las fases de la misma, con esto se obtendrá información de calidad para la correcta toma de decisiones.

SUMMARY

It was developed a methodological approach for the management of data quality integration projects to be implemented in the IIS-ESPOCH (Institutional Information System of the Escuela Superior Politecnica de Chimborazo).

The proposal is to serve seven phases, the first to study and project preparation, the second for data analysis, the third is assessment and analysis, the four is data cleansing, then the fifth phase is done assessment and final analysis of data quality of data that corresponds to the sixth and seventh phase.

For the development of the research the following tools were used, Microsoft Sql Server 2008 Business Intelligence Development Studio, Data Cleaner, BayCastle Data Slave Map Editor, Sql Power DQGuru.

Applying this proposed methodology is obtained that the distance education unit has a 29.61% increase resulting in 95.10% of final good quality data and 4.9% poor quality data .As for the academic system ,data quality increased 29.61% and a final result of good quality data 99.94% and 0.06% and poor quality data.

The proposed methodology ensures the quality of data integration projects ,in this case for the IIS-ESPOCH .It is advisable to follow in order the phases of the same ,with this quality will provide information for right decision-making.

GLOSARIO

BUSINESS INTELLIGENCE

Se denomina inteligencia empresarial, inteligencia de negocios o BI (del inglés business intelligence) al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa.

CALIDAD

Una categoría tendiente siempre a la excelencia

GESTION

Realización de diligencias enfocadas a la obtención de algún beneficio

METADATA

La frase utilizada comúnmente para explicar los metadatos es los datos acerca de los datossimplemente significa la información descriptiva acerca de algo.

MIGRACION

Hablamos de migración de datos cuando nos referimos al traspaso de información entre bases de datos.

PERFILADO

Por perfilado de datos se entiende el análisis de los datos de los sistemas para entender su contenido, estructura, calidad y dependencias.

BIBLIOGRAFIA

LIBROS

[1] MCGILVRAY Danette. Executing Data Quality Projects. United States, Elsevier, 2008, pp. 163-165

BIBLIOGRAFIA DE INTERNET

[2] OÑATE, JUAN.Gestión de calidad de datos para business intelligence, 15 de Diciembre 2009
http://www.economiadehoy.com/modulos/mod_periodico/pub/mostrar_noticia.php?id=10466

(2010/12/11)

[3] CALIDAD DE DATOS:Beneficios de la calidad de datos ,12 de abril 2009

http://scanda.com.mx/scanda/pdf/QAR_CalidadDatos.pdf

(2010/11/20)

[4] CALIDAD DE DATOS FACTOR CRÍTICO:Impactos en el negocio, 08 de agosto2009

http://www.gestiopolis.com/canales5/emp/pymecommx/27.htm

[5] INFORMACION: Ciclo de vida de la Información y la calidad de datos, 04 diciembre 2009

http://www.bc.gov.cu/Anteriores/RevistaBCC/2010/Nro4_2010/datos.htm,

(2010/11/25)

[6] GESTION DE CALIDAD DE DATOS: Fuentes de errores en los datos, 09 de febrero 2009

www.aulesempresa.upc.edu/.../POWERDATA/.../3%20Gestion%20de%20Calidad%20de%20Datos.ppt

(2010/12/08)

[7] SOTO, DAVID:Integración y Calidad de Datos,22 de enero 2010

http://integracionycalidad.blogspot.com/2010_01_01_archive.html

(2010/12/10)

[8] CALIDAD DE DATOS EN LAS ORGANIZACIONES:Dimensiones de Calidad de Datos , 23 de junio 2009

http://www.wikilearning.com/monografia/calidad_de_datos-dimensiones_de_la_calidad_de_datos/14667-2

(2010/12/12)

[9] ERA DE LA INFORMACION: Categorías de Datos, 25 de febrero 2010

http://www.acis.org.co/fileadmin/Base_de_Conocimiento/XXVIII_Salon_de_Informatica/ConferenciaJorgeVillalobosAlvarado.pdf

(2010/12/20)

[10]INTEGRACION DE ASPECTOS DE CALIDAD DE DATOS:Personal de administración de calidad, 03 de marzo2009

http://www2.tdg-seville.info/cfp/zoco/zoco09/resources/Caballero.pdf

(2011/01/15)

[11] MANEJO DE INFORMACION: Niveles de madurez de calidad de datos

11,10 de octubre 2009

(2011/01/26)

[12] INTEGRACION DE DATOS:Proyectos de Integración, 11 de noviembre 2009

http://informationmanagement.wordpress.com/tag/integracion-de-datos/

(2011/01/27)

[13] POWER DATA: Gestión de calidad de datos,12 de diciembre 2010

http://www.infosysblogs.com/eim/2010/12/design_considerations_for_a_da.html

(2011/01/30)

[14] INFORMATICA: Estrategia de calidad de datos, 04 de agosto 2010

http://www.informatica.com/es/solutions/data_quality/Pages/data_quality_methodology.aspx

(2011/02/02)

[15] ADASTRA: Data Quality Strategy, 06 de julio 2009

http://ptgmedia.pearsoncmg.com/images/0321240995/samplechapter/Adelman_ch03.pdf

(2011/02/05)

[16]DATACTICS: Data Quality, 01 de septiembre 2009

http://www.datafactotum.com/2010/05/how-are-you-executing-your-data-quality.html

(2011/02/10)

[17] ORLI: Data Quality Methods, 20 de septiembre 2010

http://www.kismeta.com/cleand1.html

(2011/02/12)

[18] DMAIC: A Bifocal Strategy for Data Quality ,02 de octubre 2009

http://www.tdan.com/view-articles/8426

(2011/02/15)

DESCARGA DEL SOFTWARE

[19] SQL POWER DQGURU

http://www.sqlpower.ca/page/dqguru

[20] DATA CLEANER

http://datacleaner.eobjects.org/

[21] DATA SLAVE

http://www.baycastle.co.uk/V2/DataSlave/DataSlave.htm

[22] SQL SERVER 2008

http://www.microsoft.com/sqlserver/2008/en/us/default.aspx

ANEXOS

Perfilados de Datos

Los perfilados de datos que corresponden a la Fase II de la metodología propuesta se encuentran en el cd adjunto a este documento debido a su gran extensibilidad .