ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE INFORMÁTICA Y ELECTRÓNICA

ESCUELA DE INGENIERÍA EN SISTEMAS INFORMÁTICOS

"EVALUACIÓN DEL RENDIMIENTO EN LA INTEGRACIÓN DE DATOS CON HERRAMIENTAS DE SOFTWARE LIBRE, EN AMBIENTES CUYAS FUENTES DE DATOS SEAN BIG DATA"

Trabajo de titulación presentado para optar el grado académico de:

INGENIERO EN SISTEMAS INFORMÁTICOS

AUTOR: **LÓPEZ ESPINOZA GUIDO EFRAÍN**

TUTOR: **IVÁN MENES CAMEJO**

Riobamba – Ecuador

2015

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE INFORMÁTICA Y ELECTRÓNICA

ESCUELA DE INGENIERÍA EN SISTEMAS INFORMÁTICOS

EL Tribunal de Tesis certifica que: El trabajo de investigación: "EVALUACIÓN DEL RENDIMIENTO EN LA INTEGRACIÓN DE DATOS CON HERRAMIENTAS DE SOFTWARE LIBRE, EN AMBIENTES CUYAS FUENTES DE DATOS SEAN BIG DATA", de responsabilidad del señor Guido Efraín López Espinoza, ha sido revisado minuciosamente por los miembros del Tribunal de Tesis, quedando autorizada su presentación:

ING. GONZALO SAMANIEGO **DECANO DE LA FACULTAD DE** **INFORMÁTICA Y ELECTRÓNICA**	_____	_____
DR. JULIO SANTILLÁN **DIRECTOR DE LA ESCUELA** **INGENIERÍA EN SISTEMAS**	_____	_____
ING. IVÁN MENES **DIRECTOR DE TESIS**	_____	_____
DR. ALONSO ÁLVAREZ **MIEMBRO DE TESIS**	_____	_____
DOCUMENTALISTA **SISBIB ESPOCH**	_____	_____

Yo, Guido Efraín López Espinoza soy responsable de las ideas, doctrinas y resultados expuestos en esta Tesis y el patrimonio intelectual de la Tesis de Grado pertenece a la Escuela Superior Politécnica de Chimborazo.

_____

GUIDO EFRAÍN LÓPEZ ESPINOZA

DEDICATORIA

Este trabajo va dedicado a mi querida esposa y a mi hija; por su apoyo incondicional; a mis padres y hermanos que han sido un pilar fundamental para obtener esta distinción; a mis maestros por darme todo su apoyo para llegar a un feliz término y en general a todos mis familiares por su constante apoyo para la obtención de esta meta tan importante en la vida de un estudiante.

AGRADECIMIENTO

Agradezco a los miembros de esta Tesis; Ing. Iván Menes y Dr. Alonso Álvarez, por su apoyo en este trabajo de investigación; al Ing. Hugo Vera por su incondicional apoyo para el desarrollo de este trabajo; finalmente agradezco a todas las personas que me han apoyado en este camino tan importante para el desarrollo de mi vida profesional.

Mi profundo agradecimiento a mi querida familia por su apoyo incondicional para poder desarrollar este trabajo y obtener una meta muy importante en mi vida.

Guido

CONTENIDO

ANEXOS

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

RESUMEN

El análisis comparativo de la Evaluación del Rendimiento en la Integración de Datos con Herramientas de Software Libre, en ambientes cuyas fuentes de datos sean Big Data, se lo realizó con el propósito de determinar cuál de las dos herramientas de integración de datos de software libre establecidas brindan un mejor rendimiento para el desarrollo del prototipo de un observatorio de indicadores educativos para la Facultad de Informática y Electrónica (FIE), de la Escuela Superior Politécnica de Chimborazo (ESPOCH). Se utilizó la metodología de construcción de prototipos de comparación, para realizar con las dos herramientas que se escogieron para el desarrollo del ambiente de comparación y analizar el comportamiento en el análisis de datos. La construcción de prototipos de comparación, nos sirvió para determinar a través de sus parámetros y sub parámetros, la herramienta de integración que mejor rendimiento presenta en el desarrollo de los procesos de integración de datos de los indicadores educativos de la FIE de la ESPOCH. El resultado obtenido determinó que la herramienta de integración de datos Pentaho Data Integrator posee un mejor rendimiento**,** con un valor de 95/100; dando paso a la construcción del prototipo del observatorio de indicadores académicos en la Facultad de Informática y Electrónica. El observatorio de indicadores se realizó con las herramientas Microsoft SQL Server, Pentaho Data Integrator **y** la herramienta de visualización de datos Tableau**,** mediante el cual se puede analizar el comportamiento de los indicadores entre las escuelas de la Facultad. Al final se concluyó que la Herramienta de Integración Pentaho Data Integrator brindó un mejor rendimiento en el desarrollo de prototipos de integración de datos. Se recomienda el uso de herramientas fuente con actualización de datos en tiempo real, para desarrollar ambientes actualizados y así presentar los resultados más reales.

PALABRAS CLAVES: <INTEGRACIÓN DE DATOS>, < [PENTAHO DATA INTEGRATION] HERRAMIENTA DE INTEGRACIÓN DE DATOS>, <TALEND OPEN STUDIO HERRAMIENTA DE INTEGRACIÓN DE DATOS>, <OBSERVATORIO DE INDICADORES>, <PROTOTIPO>, <SOFTWARE LIBRE>, < [BIG DATA] GRAN CANTIDAD DE DATOS>

SUMMARY

This investigation was carried out to make a comparative analysis about the Evaluation of Performance Data Integration with free software tools, environments whose data sources are Big Data, it made in order to determine which of the two data integration tools established free software to develop the prototype of an observatory of educational indicators for computer and electronics (CEF) from Escuela Superior Politecnica de Chimborazo (ESPOCH). Construction methodology prototypes comparison was used to make tools with both were chosen for comparison development environment and analyze the behavior data analysis. This helped to determinate through them parameters and sub-parameters, the integration tool that presents better performance in the development of integration process data indicators educational from this institution. The result found that data integration tool Pentaho Data Integrator has a better performance with 95/100; leading to the construction of the prototype observatory of the Faculty academic indicators. The observatory of indicators was done with Microsoft SQL Server, Pentaho Data Integrator and Tableau visualization tool data by which to analyze the behavior of the indicators between the schools at Faculty tools. Finally, it was concluded that integration tool Pentaho Data Integrator gave better performance in prototype development data integration. It is recommended to use of power tools to update data in real time, developing current environments and thus present the actual results.

KEYWORDS: <DATA INTEGRATION>, < [PENTAHO DATA INTEGRATION] DATA INTEGRATION TOOL>, <TALEND DATA INTEGRATION TOOLS>, <MONITORING INDICATORS>, <PROTOTYPE>, <FREE SOFTWARE>, < [BIG DATA] LARGE AMOUNT OF DATA>

INTRODUCCIÓN

En toda organización o empresa surge la necesidad de innovar su área informática, que se sustenta en los avances tecnológicos que fluyen día a día a su alrededor, constituyendo así en un punto de competición con sectores afines a él.
Sin embargo, cada uno de estos cambios va de la mano con una situación económica y metas realizables a futuro que desea alcanzar la organización o empresa. No obstante todo sacrificio que se realice permitirá posesionar a la empresa en un punto de máxima calidad en lo que hace.
Debido a esta razón es importante seleccionar adecuadamente las herramientas de Integración de Datos que se analizan en esta tesis, siendo estas herramientas las que nos proporcionen una mayor fiabilidad en la calidad de datos para que los usuarios posean información coherente para la toma de decisiones dentro de un sistema informacional.

Mediante el desarrollo del presente trabajo de investigación, se busca determinar la Herramienta de Integración de Datos más adecuada para integrar información de distintas fuentes que sean de ámbito Big Data, y otorguen información veraz y adecuada para el personal que tome las decisiones en la empresa.
El presente trabajo de investigación se enfoca en analizar las diferentes herramientas de integración de software libre y su desempeño en la fase de integración para lograr optimizar tanto los tiempos de respuesta, compatibilidad en los datos, etc.

Para la obtención de los resultados se realizó un estudio de las herramientas usando diferentes parámetros de comparación realizando un cuadro comparativo que permitirá demostrar cuál es la mejor opción para la parte aplicativa de esta tesis, para los escenarios de prueba se establecieron orígenes y destinos de datos de diferentes tamaños de registros con los siguientes ambientes de prueba, SQL Server 2008, Archivos en Excel, Archivos Planos.

Antecedentes

Dentro de una empresa grande, mediana o pequeña, el manejo que se le dé a la información es muy importante, por tal razón no es recomendable hacerlo manualmente, sin embargo existen compañías que lo hacen, presentándose perdida de información y lentitud en los procesos. Por esto han surgido cada vez más en el mercado, soluciones informáticas conformadas por diferentes módulos.
Debido a la creación de herramientas libres para el desarrollo de sistemas de información, gran parte de las empresas están optando por utilizarlas, de esta forma optimizan los procesos a menor

precio. Sin embargo, entre los analistas y diseñadores de software, algunas no son muy conocidas, por tal razón tienen que estar documentándose y a la vanguardia de la tecnología.
Para la creación y administración de una red de datos, existe gran variedad en el mercado de software y dispositivos de hardware para cumplir esta tarea, de tal manera que se debe tener en cuenta el tamaño de la empresa en cuanto a equipos activos, servidores, impresoras y clientes.
El problema que queremos evaluar radica en, que una Base de Datos concentradora implica grandes volúmenes de datos, estos datos generalmente suelen ser de fuentes transaccionales de menor tamaño y volumen de datos.

Estas fuentes provienen de fuentes de los sistemas transaccionales de una empresa; que contienen datos, los mismos que se necesitan ser integrados para realizar su análisis.
La integración de los datos se caracteriza por la conexión entre los datos y la información digital. Además la utilización de herramientas Open Source para la Integración proporcionan una serie de ventajas gracias a sus magníficas cualidades como estabilidad, seguridad, confiabilidad, multiplataforma, optimización de recursos, gratuidad entre otros. Se puede acceder al código y aprender de él, se puede modificar adaptándole para realizar áreas específicas, adaptación tecnológica Open Source con tecnología propietaria entre otras más.

Justificación

Justificación Teórica

Las herramientas de integración de datos, están destinadas a facilitar la realización de las tareas ETL, para así poder lograr una mejor integración de los datos provenientes de fuentes Big Data. Los procesos ETL, permiten a las organizaciones mover datos de unas o varias fuentes, reformatearlos, limpiarlos y cargarlos en una Base de Datos centralizadora para poder realizar un análisis de los datos y emitir decisiones que ayuden a mejorar los negocios.
Para analizar y determinar que herramienta es la más eficiente y de mayor rendimiento, nos basamos en factores y parámetros como puede ser: tiempo de acceso, integridad con los datos, interactividad con las fuentes, complejidad con las sentencias de integración; también se realizan pruebas reales mediante la creación de prototipos de cada una de las herramientas de integración seleccionadas.

Para el estudio de las herramientas de integración lo realizaremos mediante los parámetros propuestos anteriormente en prototipos y ambientes de prueba, para de esta manera determinar cuál de las herramientas seleccionadas es la de mejor rendimiento al momento de realizar la integración de los datos.

Justificación Metodológica

Para la elaboración de la propuesta metodológica nos basaremos en la recopilación de información hallada a través de papers, blogs, libros, foros de Inteligencia Artificial, comentarios, sugerencias, guías, recomendaciones, información compartida de empresas dedicadas al análisis de datos.
Para la realización de este estudio se utilizará la técnica de recopilación de información en fuentes secundarias, publicaciones que se hayan realizado en base a este tema, el mismo que se encuentra en auge en estos días.

Justificación Práctica

Este trabajo de investigación que se pretende realizar tiene un enfoque puramente investigativo; por lo que no podemos analizar la justificación práctica se lo realizará mediante unos ambientes de prueba en los cuales podremos determinar el rendimiento que ofrece esta integración de datos, con fuentes Big Data.

Objetivos

Objetivo General

- ✓ Evaluar el rendimiento en la integración de datos con herramientas de Software Libre, en ambientes cuyas fuentes de datos sean Big Data.

Objetivos Específicos

- ✓ Realizar un estudio de las herramientas de integración de datos con herramientas de software libre, compatibles con Big Data.
- ✓ Seleccionar los parámetros y criterios de evaluación para medir el rendimiento.
- ✓ Construir un prototipo para integración de datos para pruebas y análisis de resultados con las herramientas seleccionadas.
- ✓ Desarrollar un prototipo para la construcción de un observatorio de indicadores en la Facultad de Informática y Electrónica, basado en tecnología de integración de datos con la factibilidad de fuentes Big Data.

Hipótesis

La herramienta de integración de datos Pentaho Data Integrator posee un mejor rendimiento en ambientes cuyas fuentes de datos sean Big Data.

Métodos y Técnicas

Métodos

Para la comprobación de la hipótesis será aplicado un método científico que permitirá establecer una secuencia ordenada de actividades que nos llevará a establecer nuestras conclusiones sobre la investigación realizada.
También se utilizará como complemento del presente trabajo al método, por cuanto, este establece el procedimiento necesario para la recopilación y análisis de comparación para la realización de un observatorio de indicadores en la Facultad de Informática y Electrónica, basado en la tecnología de integración de datos con fuentes Big Data.

Técnicas

En cuanto a fuentes de información se utilizará principalmente fuentes que se refieren al tema de investigación como páginas web, también se empleará la observación y experimentación por parte de los investigadores.

Técnicas:

- ✓ Observación
- ✓ Revisión de Documentos
- ✓ Técnicas Estadísticas para comprobar la Hipótesis

Fuentes:

- Internet

CAPITULO I

1. MARCO TEÓRICO REFERENCIAL

1.1 Introducción

Durante el desarrollo del estudio de la Herramientas de Integración de Datos en fuentes Big Data, es importante analizar todos los componentes que forman parte de esta investigación así como las herramientas que van a ser objeto de estudio de nuestro trabajo, así como características, ventajas y desventajas de cada una de ellas; por tal motivo este capítulo es el conjunto introductorio al desarrollo de la Tesis.

Una de los antecedentes del Big Data es que para representar fácil y rápidamente el rendimiento de una supercomputadora los expertos recurren a su particular notación científica, los FLOPS ("floating point operations per second"), es decir, la cantidad de operaciones que procesa por segundo, por lo que hablamos de teras y petas, es decir, respectivamente, de al menos un billón y mil billones de operaciones por segundo. El siguiente paso, el grail santo de la supercomputación actual, al decir de Clay Dillow, sería una máquina con capacidad exaflop, toda esta cantidad de información se tiene que almacenar.

Más de 900 millones de usuarios de Facebook registrados generan más de 1500 actualizaciones de estado cada segundo de sus intereses y su paradero. En 2011, la plataforma de comercio electrónico eBay, recolectó datos sobre más de 100 millones de usuarios activos, incluyendo los 6 millones de nuevos bienes que se ofrecen todos los días. Cuando el 14 de febrero del 2013, cerró sus puertas tras tres años de operación para una etapa de mantenimiento y renovación de equipos, el Large Hadron Collider (LHC), que hizo posible el descubrimiento de la Partícula de Higgs, entre la frontera de Suiza y Francia, había logrado acumular 100 petaflops de datos, dos veces una biblioteca colectiva que incluiría cada palabra escrita de todas las lenguas, más o menos el equivalente a 700 años de películas HD de plena calidad o mil veces todo el texto disponible en la Biblioteca del Congreso de los Estados Unidos.

La información es un activo fundamental por su capacidad para impulsar los negocios. Conocer más sobre la forma de comportamiento de los consumidores, saber a qué herramientas recurren a la hora de informarse sobre productos y servicios, identificar a los líderes de opinión en un determinado mercado, detectar amenazas y actividades fraudulentas antes de que lleguen a concretarse o identificar las posibles fuentes de problemas para predecir fallos en las redes son factores cruciales que pueden hacer que un negocio incremente su rentabilidad enormemente.

Mucha de esa información está al alcance de las empresas, prácticamente delante de sus ojos, esperando sólo que alguien se detenga en ella. El nuevo paradigma de la Big Data implica que las fuentes a partir de las cuales obtener una visión profunda del mercado y las operaciones se han multiplicado. Los datos ya no sólo provienen de las bases estructuradas tradicionales, si no de interfaces de usuarios, redes sociales, foros, mensajes de texto y entornos diversos en los que los consumidores interactúan día a día entre sí y con otros actores del mercado. En el entorno de la Big Data, las organizaciones se encuentran ante el desafío de incorporar información en crudo, sin procesar, que se actualiza en tiempo real y que presenta una enorme complejidad.

1.2 Big Data

Existen muchos conceptos que definen el termino BIG DATA, a continuación se va a citar uno de las explicaciones que más se asemejan a su significación, para comprender el amplio campo que abarca.

"Big data" son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones. (*http://www.cnis.es/images/informes/Articulo%20Big%20Data%200.0.pdf, p.p 1-3, 9-10)*

Big Data es el término que se emplea hoy en día para describir el conjunto de procesos, tecnologías y modelos de negocio que están basados en datos y en capturar el valor que los propios datos encierran. (MORO, Esteban & LUENGO-OROZ, Miguel & DE LA TORRE, Javier; 2013; p.p 5-10, 12-15). En 2001, en un informe de investigación que se fundamentaba en congresos y presentaciones relacionadas, el analista Doug Laney del META Group (ahora Gartner) definía el crecimiento constante de datos como una oportunidad y un reto para investigar en el volumen, la velocidad y la variedad. Gartner continúa usando Big Data como referencia de este. (RINDLER Andreas, *2011)*

De acuerdo a los conceptos citados, podemos concluir que BIG DATA, en un conjunto de información que posee un tamaño absolutamente grande, que hace referencia a la cantidad de datos existentes, los cuales necesitan un proceso de minería de datos para poder obtener información realmente necesaria para una organización para poder lograr que se tomen las decisiones más acertadas, para mejorar los procesos que desarrolla la institución.

Big Data se genera por muchos aspectos, entre los cuales surgen aspectos como los que a continuación se citan; para lograr entender de donde proviene este término y su importancia en el mundo de la tecnología y de los datos en el campo informático.

Todo lo que hacemos en internet queda registrado en diferentes servicios, incluso lo que hacemos en nuestra vida por el simple hecho de llevar un dispositivo móvil con nosotros conectado a una línea de teléfono y/o 3G. Todos esos datos se almacenan (dónde has estado, a quién has llamado, qué has comprado con tu tarjeta, qué hábitos tienes, qué páginas visitas) en enormes bases de datos, que son luego procesadas por auténticas factorías de información. (IGLESIAS Pablo; 2012)

Big Data está aquí para optimizar el beneficio de las empresas, y evitar el spam en los consumidores ofreciéndoles información que en verdad sea lo que buscan.
Una mayor cantidad de fuentes de datos y el desarrollo de fuentes digitales que permiten su recolección en tiempo real, como instrumentos, sensores, transacciones de internet, entre otras muchas, impulsan al sector del Big Data. De acuerdo a cifras de IBM, el 90% de los datos en el mundo se ha creado en los últimos dos años y hoy, todos los días, se crean 2.5 quintillones de bytes de datos. (FARAH CALDERÓN, Walter, 2013, p.p 1-2)

Las empresas en la actualidad generan mucha información diariamente, la misma que se almacena en los dispositivos de almacenamiento secundario, lo que hace genera determinar un alto presupuesto para el almacenamiento para la información, lo que hace necesario la utilización de herramientas de gestión de datos para realizar una integración adecuada para preservar la información que se necesita y es importante para la empresa y de esta manera reducir el presupuesto para el almacenamiento y tener un almacén de datos que sea con información muy útil.

Las redes permiten al usuario hablar de tú a tú con la marca y con sus clientes, el cliente pasa de ser target a ser tratado individualmente expresando sus preferencias y opiniones respecto a sus productos y servicios ejerciendo un poder de influencia que como compañía, no podemos obviar. Diversos estudios sectoriales demuestran que los volúmenes de información se duplican cada 18 meses, además de mencionar que tan solo el 20% de la información disponible son datos estructurados y de fácil acceso. De ahí que uno de los mayores retos para las organizaciones sea la capacidad para gestionar e interpretar todos estos datos obteniendo una ventaja competitiva sólida y diferenciada. (*EPSILONTEC, p.p 1)*

Estos estudios que se analizan de la fuente obtenida nos indica que se debe tener una política rigurosa en el aspecto de integración de datos, ya que la mayoría de los mismos no constituyen una información realmente importante para la empresa, sino que se debe realizar un proceso de minería de datos para poder obtener lo más importante y que sea de beneficio para nuestra empresa.

1.3 Integración de datos.

La integración de datos se centra principalmente en las bases de datos. Una base de datos es una colección organizada de datos. Podemos decir que es algo similar a un sistema de archivos, el cual es un grupo estructurado de archivos para que puedan ser encontrados, accedidos y manipulados fácilmente. (http://www.latinobi-ven.com/, 2012)
Las empresas en ocasiones es probable que ya tengan los datos que necesita para ejecutar aplicaciones, comprender a los clientes y tomar decisiones de negocios importantes. Pero para la mayoría de estas, esta información es almacenada en sistemas dispares, por medio de formatos incoherentes.

La integración de datos puede ofrecer beneficios considerables para su organización, eficiencia y producción, al aprovechar los datos empresariales y la información incluida en los registros de los clientes. (www.ordenadores-y-portatiles.com/integracion-de-datos.html. 2014)
De acuerdo con los conceptos citados anteriormente se puede mencionar que este proceso de integración beneficia en muchos aspectos a la empresa para obtener información que se encuentra almacenada en diferentes dispositivos y que requieren ser agrupados, para realizar un mejor análisis y obtener decisiones más acertadas para una mejor toma de decisiones que beneficien a la empresa.

La integración de datos la podemos definir como el proceso de combinar datos que residen en diferentes fuentes y permitirle al usuario final tener una vista unificada de todos sus datos. La habilidad de transformar datos inter-departamentales de fuentes heterogéneas en un plan de acción que se convertido en un reto y en una ventaja competitiva para compañías que requieran la integración de datos.
La integración de datos es un elemento fundamental y crítico en la variedad de tecnologías incluyendo Data Warehouse, aplicaciones de inteligencia de negocio, arquitecturas orientada a servicio, aplicaciones MDM y arquitecturas data-centric. (**RIOS, Angel.** 2009)

1.3.1 *Generalidades*

El desarrollo produce cada vez mayor poder de procesamiento y sofisticación de las herramientas y técnicas analíticas, esto ha dado como resultado la creación de los almacenes de datos.
Proporcionan almacenamiento, funcionalidad y receptividad a las consultas que van más allá de las posibilidades de las bases de datos destinadas a transacciones.
Los almacenes de datos proporcionan acceso a datos para análisis complejos, revelación de conocimientos y toma de decisiones. (MORALES Yessica, 2012, p.p 2-209.

Las herramientas de integración de datos son aquellas destinadas a facilitar la realización de las tareas de integración, que permitirán estructurar los datos de forma sencilla, pensando en el rendimiento, rapidez, facilidad de uso y garantizar la calidad de la información.
Estas pueden disminuir los mayores tiempos de carga, mayores costes de almacenamiento, peores rendimientos, mayores costes, publicidad, etc. todo un despilfarro realmente. (BETANCUR, Daniel. p: p 16-38)

1.3.2 Almacenamiento de datos

Durante la última década, las organizaciones han desplegado gradualmente aplicaciones de inteligencia de negocio (BI) para equipar a los responsables de la toma de decisiones y a los analistas con la mejor y más fiable información. La recompensa ha sido enorme, y mientras otros sectores de TI han sufrido, la inversión de las compañías en BI se ha mantenido de manera consistente.

Es por esta razón que se requiere realizar un estudio de las herramientas que permitan realizar una integración de los datos de una manera tal que permita a los miembros de los departamentos tomar decisiones que sean muy acertadas para el mejor funcionamiento de la empresa y proyectarla hacia un mejor porvenir, por lo que se hace muy necesario contar con herramientas que faciliten este proceso y brinden la información necesaria y coherente para realizar los procesos de interpretación y análisis de la información.

Las funciones clave para llegar al buen aprovechamiento de los sistemas tienen que ver con un rendimiento y escalabilidad mejorados que permitan manejar grandes volúmenes de datos y procesarlos en tiempo real. Además, se ha de trabajar en pro de una limpieza y transformación especializadas, así como en base a la generación de informes claros y concretos a raíz de dichos datos y un linaje completo desde el informe hasta la fuente, que permita lograr una pista de auditoría integral, de manera que el negocio confíe y comprenda la manera en que se hace entrega de todos los datos.

1.3.3 *Herramientas para realizar integración de datos*

Uno de los mejores en calificar a las herramientas de integración de datos es Gartner; en su reciente estudio ha reconocido como Líderes a IBM, Informática, SAP, Oracle y Sas; en el área de Challengers a Microsoft; en el cuadrante de Niche Players se encuentra Syncsort; y en la categoría de Visionarios destacan iWay Software, Talend y Pervasive Software, nombra a las herramientas que más sobresalen en este campo de estudio, estadísticas que se muestra en la figura

1-1. (*Cuadrante Mágico de Gartner sobre integración de datos 2014 – Informática es uno de los líderes.* 2014)

Figura 1-1 Cuadrante de Gartner sobre Herramientas de Integración de Datos.
Fuente: Gartner, 2014**.**

1.3.4 *Pasos para realizar la integración de datos*

Para la realización del proceso de integración de datos debemos tener en cuenta los siguientes aspectos:

- La situación previa a la implementación.-
  - Se analiza las diferentes fuentes de información que poseen los sistemas informativos u operacionales que se encuentran ejecutando.
  - Se analiza el procesamiento de los datos así como la validación e incorporación de los datos en un Almacén de Datos.
  - Explotación de los datos a través de las herramientas seleccionadas con este objetivo.
  - Automatizar los procesos que forman parte del ciclo de vida de los datos en la ESPOCH, reduciendo el tiempo que se ocupa para las tareas de carga de los datos.
  - Crear repositorios de procesos documentados para que nos reduzca la pendiente de la curva de aprendizaje, en el caso de incorporación o sustitución de personal.
  - Aumentar la calidad y el control en los procesos ETL (Extracción, Transformación y Carga), para disminuir el número de errores así como el tiempo de demora en la corrección.
- Elección de la Solución.-

Para realizar el proceso de integración de datos con fuentes de datos Big Data, debemos analizar las herramientas disponibles para este efecto.

  - Talend Open Studio.-

Es una Suite de herramientas potentes y flexibles de software libre que nos ayudan a la realización de las tareas de integración de datos, además que posee la funcionalidad de trabajar con fuentes de Datos BigData.

- Pentaho Data Integration.- Pentaho se define a sí mismo como una plataforma de BI "orientada a la solución" y "centrada en procesos" que incluye todos los principales componentes requeridos para implementar soluciones basados en procesos y ha sido concebido desde el principio para estar basada en procesos. (*http://latam.pbinsight.com/*, 2012)

1.3.5 *Herramientas de integración en el mercado compatible con Big Data*

En el mercado existen muchas herramientas de integración de datos compatibles con fuentes Big Data, tanto herramientas de software libre como software propietario, las mismas que se pueden visualizar en la siguiente tabla. (Press Gil, *2014)*

Tabla 1-1 Herramientas de Integración de Datos compatibles con Big Data

Herramientas de integración de Software Propietario, compatible con Big Data	Herramientas de integración de Software Libre, compatible con Big Data
Microsoft	Talend Open Studio
IBM	Pentaho Data Integration
Oracle	Scriptella ETL
Datameer Big Data Analytics	Hadoop
Kapow Software	Platfora
ZettaSet	YarcData
Space Time Insight	clearstorydata

Fuente: «Top 10 Big Data Pure-Plays 2014».
Realizado por: Guido López

1.4 Fuentes de información Big Data

En los últimos años, debido al avance de las Tecnologías de la Información, estamos siendo testigos de una verdadera explosión en la cantidad de datos disponibles, listos para ser analizados y así convertirse en información importante para la inteligencia de negocio.

Este nuevo escenario se refiere no sólo al volumen de datos, sino también a la velocidad, complejidad y variedad de los tipos de información disponible, como acontece con los datos de las redes sociales, logs de acceso a Internet o datos generados por máquinas. (Breva Systems Administrator**,** 2012)

Los Big Data o Grandes Datos son datos ricos y extremadamente útiles para análisis, pero que no se encuentra disponibles al menos inicialmente de una manera estructurada, ya sea por la alta velocidad con que son producidos o por los mecanismos a través de los cuales son generados. Siendo así, más allá de la gran cantidad de información disponible hoy, los Big Data se relacionan directamente a la capacidad de manipular y analizar datos multi-estructurados no relacionados, que requieren de una interacción rápida y adaptable. (DEL PINO, Manuel, 2014, p.p 53)

En base a la tecnología existente, hoy en día se cuenta con sistemas que proporcionan datos en una cantidad muy elevada para ser manejada por los sistemas tradicionales.
Entre estos tipos de fuentes de datos podemos tener:

- Bases de Datos.- Estas bases cuentan con información que son almacenada en millones de registros, lo cual hace muy difícil su manipulación y extracción de datos confiables que brinden información adecuada.
- Archivos Planos.- Este tipo de fuente es muy común, contiene información no estructurada en la cual debemos utilizar técnicas y herramientas que faciliten la integración de estos datos y poder convertirlos en información.
- Sistemas ERP.- Estos sistemas proveen un alto contenido de información, lo que los convierten en fuentes Big Data; para lo cual es necesario analizar y usar herramientas de integración de estos datos para su análisis.
- Redes Sociales.- En esta era las redes sociales son la mayor fuente de datos Big Data; ya que millones de personas diariamente utilizan estas redes, ya sea para subir fotografías, música, videos; etc.

Las fuentes de información Big Data se pueden administrar mediante la utilización del framework HADOOP, también con gestores de bases de datos: SQL Server y otros, el mismo que posee herramientas necesarias para el tratamiento de grandes cantidades de información.

1.5 Hadoop

Hadoop es un sistema de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos; cientos de terabytes, petabytes o incluso más.
Hadoop surgió como iniciativa open source (software libre) a raíz de la publicación de varios papers de Google sobre sus sistemas de archivo, su herramienta de mapas y el sistema BigTable Reduce.

Como resultado nació un conjunto de soluciones en el entorno Apache: HDFS Apache, Apache MapReduce y Apache HBase; que se conocen como Hadoop, con herramientas como Sqoop (para

importar datos estructurados en Hadoop cluster) o NoSQL (para realizar el análisis de los datos no estructurados) entre otros. (LURIE, Marty. 2013).
De este concepto se concluye que HADOOP es una tecnología de código abierto que facilita la manipulación de datos, mediante una serie de herramientas en conjunto; las cuales realizan tareas específicas para la integración y manipulación de los datos de una organización.

1.5.1 *Antecedentes*

El origen de Hadoop se remonta a 2004, cuando el ingeniero de software Doug Cutting, que por aquel entonces trabajaba en Google, describe en un documento técnicas para manejar grandes volúmenes de datos, desgranándolos en problemas cada vez más pequeños para hacerlos abordables. Poco después se marchó a Yahoo y allí siguió investigando hasta completar el desarrollo de la plataforma en 2008. El propio buscador utilizaría la tecnología para su negocio, así como otras grandes compañías de Internet, como Facebook, Twitter o eBay.

La procedencia del nombre es mucho menos técnica de lo que se podía esperar. El hijo de tres años de Cutting llamaba a su peluche Hadoop y así bautizó su inventor a la plataforma, que también tomaría de ahí su logo, un elefante amarillo. (LEO-REVILLA, Ángel, 2013)

1.5.2 *Aplicaciones que manejan Hadoop*

1.5.2.1 *Cloudera Hadoop (Cdh)*

Cloudera Hadoop *(*CDH) es la distribución más completa, probada y popular del mundo de Hadoop y proyectos relacionados. CDH es 100% Apache-licencia de código abierto y es la solución Hadoop sólo para ofrecer unificado de procesamiento por lotes, SQL interactivo, y de búsqueda interactiva, y los controles de acceso basados en roles. Más empresas han descargado CDH que todos los otros combinados tales distribuciones.

CDH incluye los elementos básicos de Hadoop además de varios proyectos de código abierto clave adicionales que, cuando se combina con la atención al cliente, la gestión y la gobernanza a través de una suscripción de Cloudera Empresa, pueden ofrecer un centro de datos de la empresa; en la figura 1-2 podemos observar lo expuesto anteriormente. (PATTERSON Clarke., *2014)*

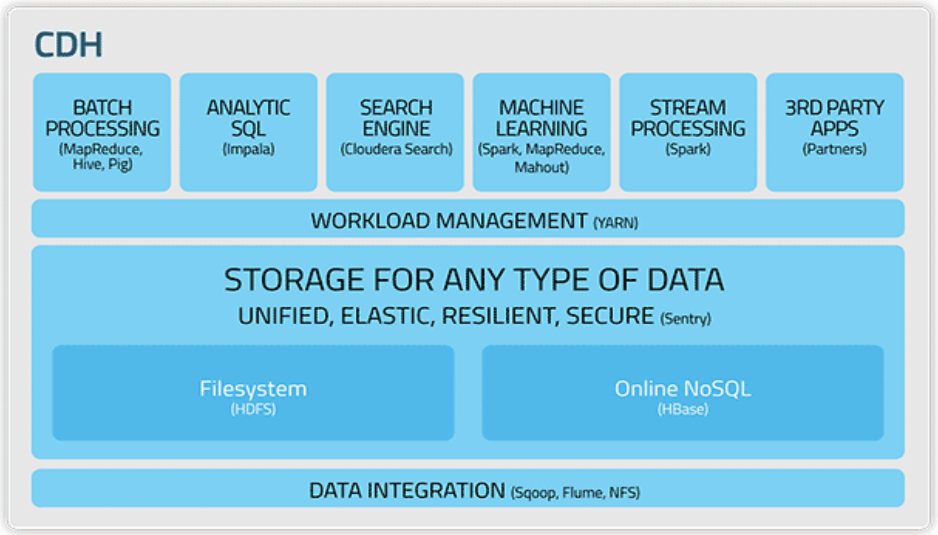


Figura 2-1 Arquitectura de Cloudera Hadoop
Fuente: «CDH». : http://www.cloudera.com/content/cloudera/en/products-and-services/cdh.html.

Del concepto anteriormente citado se concluye que esta distribución de HADOOP posee los elementos necesarios para la realización del proceso de integración de los datos de una forma eficaz y rápida, lo que facilita a los técnicos para obtener información realmente importante para sus intereses y tomar decisiones muy acertadas para el desarrollo de la organización.

CDH es:

- Flexible - Almacenar cualquier tipo de datos y procesar con una gran variedad de diferentes marcos de computación, incluyendo el procesamiento por lotes, SQL interactivo, búsqueda de texto libre, aprendizaje automático y cómputo estadístico.
- Integrado - Empiece a trabajar rápidamente en un empaquetado, la plataforma completa, Hadoop.
- Seguro - Proceso y control de datos sensibles y facilitar multiempresa.
- Escalable y extensible - Habilitar una amplia gama de aplicaciones y escalar con su negocio.
- De alta disponibilidad - Ejecutar las cargas de trabajo de misión crítica con confianza.
- Abrir - Beneficiarse de la rápida innovación patentada sin el encadenamiento con proveedores.
- Compatible - Ampliar y aprovechar las inversiones existentes en TI.

Esta herramienta brinda las facilidades necesarias para poder realizar los procesos de integración de datos, mediante el uso de los instrumentos que facilitan este proceso de una manera gráfica y amigable para el usuario que utiliza esta opción.

1.5.2.2 *Hortonworks*

Fue fundada en 2011, ha surgido rápidamente como uno de los proveedores líderes de Hadoop. La distribución proporciona la plataforma de código abierto basado en Apache Hadoop para el análisis, el almacenamiento y la gestión de grandes volúmenes de datos. Hortonworks es el único proveedor comercial para distribuir código abierto completa Hadoop sin software propietario adicional. HDP2.0 distribución Hortonworks se puede descargar directamente desde su página web de forma gratuita y es fácil de instalar. Los ingenieros de Hortonworks están detrás de la mayoría de las innovaciones recientes de Hadoop incluidos los hilados, que es mejor que MapReduce en el sentido de que permitirá inclusión de más marcos de procesamiento de datos. (Experfy Editor., 2014)

Hortonworks al igual que Cloudera está desarrollado bajo el mismo núcleo de Apache Hadoop por lo que posee algunas características similares y muy pocas diferencias con Cloudera.
Del concepto citado se puede concluir que esta herramienta es muy similar a Cloudera Hadoop, ya que se ha desarrollado bajo la misma arquitectura y presenta muchas facilidades para el manejo de las opciones que proporciona este conjunto de iniciativas para la administración de grandes volúmenes de datos.

1.5.2.3 *Mapr*

MapR cumple con la promesa de Hadoop con una plataforma de nivel empresarial que es compatible con una amplia gama de usos de producción de misión crítica y en tiempo real. MapR aporta fiabilidad sin precedentes, facilidad de uso y velocidad récord mundial para Hadoop, NoSQL, base de datos y aplicaciones de streaming en una distribución unificada para Hadoop. MapR es utilizado por más de 500 clientes en todo los servicios financieros, gobierno, salud, manufactura, medios de comunicación, el comercio minorista y las telecomunicaciones, así como por los líderes del índice Global 2000 y Web 2.0 las empresas. Amazon, Cisco, Google y HP son parte del ecosistema amplio de socios de MapR (www.mapr.com, 2014)

Mapr de acuerdo a la cita anterior, se dice que es una tecnología de manejo de Hadoop muy versátil en el mercado en el momento, posee varias utilidades muy importantes que benefician a los usuarios que deseen trabajar con grandes cantidades de datos y procesarlos para obtener una información veraz y útil.

1.6 Funcionamiento de Hadoop

La plataforma de código abierto dispone de un sistema para almacenar información en el que ésta se replica en varias máquinas, distribuyéndose de tal manera que si una máquina se cae no se pierdan los datos. Si es necesario añadir más información se añaden más servidores sin que haya problemas de compatibilidad o reorganización de los datos.
Al igual que ocurre con Linux, cualquiera puede tomar Hadoop, empaquetarlo y ofrecerlo como una distribución de la plataforma. Son varias las compañías que comercializan este tipo de solución y uno de sus principales atractivos es el algoritmo de procesamiento y búsquedas: MapReduce.

Esta herramienta permite hacer consultas a una base de datos inmensa y obtener respuestas rápidas. Es capaz de enviar una orden a cada máquina para que busque en su disco duro, recolectar todas las contestaciones y ordenarlas para resolver la consulta. (**LÁZARO, Miguel.** Como usar Hadoop y sobrevivir a la experiencia.). En las definiciones anteriores se puede observar que Hadoop es una distribución gratuita que se puede obtener, empaquetarlo y adaptarlo a nuestras necesidades debido a que es de código abierto, y da apertura para poder analizar su código y adecuarlo a lo que necesitemos; esta es una gran ventaja en relación a las herramientas propietarias que no admiten adecuación alguna en su código; para los usuarios y administradores de las organizaciones les ayuda mucho para analizar de acuerdo a su perspectiva los datos e información que pueden obtener.

MapReduce puede resolver con éxito cargas de trabajo de gran complejidad, como el procesamiento del lenguaje humano o el aprendizaje de las máquinas. Pero no es el único algoritmo que se puede utilizar. Recientemente ha aparecido la versión 2.0 de Hadoop, que permite construir otros algoritmos y utilizar otros lenguajes, lo que es un estímulo para los desarrolladores. (GRACIA Luis Miguel, 2012)

Existen plataformas que compiten con Hadoop en el escenario de Big data, aunque el elefante amarillo de momento ha tomado la delantera a todas ellas. El proyecto Spark, también de código abierto, avanza a marchas forzadas con el apoyo de Yahoo, quien estuvo involucrado en el desarrollo de su rival. Las soluciones de HPCC Systems y Pervasive Software son otras de las propuestas que flotan en el mercado.
Hadoop se ha consolidado en el ámbito Big Data, debido a su gran versatilidad y situación en el mercado, debido a que brinda las facilidades de poseer herramientas muy potentes para el tratamiento de los datos y la información que se encuentra procesando.

1.7 Componentes de Hadoop

Hadoop posee un grupo importante de herramientas para su trabajo; los componentes básicos de Hadoop son los siguientes:

1.7.1 ***Hdfs***.-

Consiste en un sistema de archivo distribuido, que permite que el fichero de datos no se guarde en una única máquina sino que sea capaz de distribuir la información a distintos dispositivos.
HDFS es el sistema de almacenamiento, fue creado a partir del Google File System (GFS). HDFS se encuentra optimizado para grandes flujos y trabajar con ficheros grandes en sus lecturas y escrituras. Su diseño reduce la E/S en la red. La escalabilidad y disponibilidad son otras de sus claves, gracias a la replicación de los datos y tolerancia a los fallos.

HDFS es un sistema de ficheros pensado para el almacenamiento de ficheros "grandes" (por encima de 100 MB) y en la que el acceso a esa información está orientado hacia procesamiento en batch o lectura de tipo "write once"-"read-many-times" (ideal para temas de MapReduce, pero no para necesidades de baja latencia) y cuyo diseño está pensado para ser ejecutado en máquinas "baratas". (TILVES Mónica, 2012)
Como todo sistema de ficheros, los ficheros son almacenados en bloques de un mismo tamaño, que en un sistema de ficheros tradicional suele ser de unos 4KB. En HDFS también existe este concepto, pero el tamaño del bloque es mucho mayor (64MB por defecto) para minimizar el coste de acceso a los bloques, y lógicamente, los bloques de un mismo fichero no tienen por qué residir en el mismo nodo. (DOMÍNGUEZ, Enrique**.** . 2007)

De acuerdo con las citas anteriores se puede mencionar que HDFS es una de las herramientas muy importantes dentro de Hadoop, se utiliza para dividir los datos que tengamos para la integración en bloques de un tamaño mayor a los tradicionales, para de esta manera agilizar el proceso de integración y minimizar los cotos que esto genera; aspecto muy importante al momento de utilizar la tecnología.
Además esta división de bloques facilita tener un almacenamiento distribuido para no abarcar todos los datos en el mismo nodo, por lo que el nodo puede funcionar tranquilamente sin tener una sobrecarga de datos y saturar el proceso de integración.

Los elementos importantes del clúster:

- NameNode: Sólo hay uno en el clúster. Regula el acceso a los ficheros por parte de los clientes. Mantiene en memoria la metadata del sistema de ficheros y control de los bloques de fichero que tiene cada DataNode.
- DataNode: Son los responsables de leer y escribir las peticiones de los clientes. Los ficheros están formados por bloques, estos se encuentran replicados en diferentes nodos.

A continuación en la ilustración N° 1-3. Se puede observar la arquitectura que presenta HDFS con sus respectivos componentes:

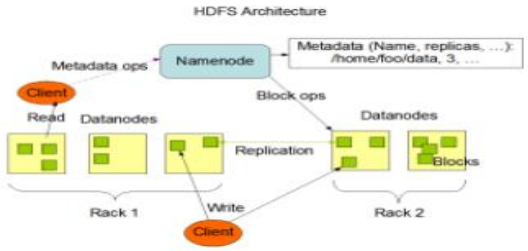


Figura 2-1 Arquitectura de HDFS
Fuente: http://java4developers.com/2013/introduccion-a-big-data-y-hadoop/

Las principales características de HDFS son:

- Tolerancia a fallos.
- Acceso a datos en streaming.
- Facilidad para grandes volúmenes
- Modelo sencillo de coherencia.
- Portabilidad entre hardware heterogéneo. (**FERNÁNDEZ, Iván**, 2013)

Estas características citadas hacen que HDFS sea lo más utilizado dentro de Hadoop para la integración de datos, con fuentes Big Data; ya que brinda muchas facilidades para la realización de estas tareas.

1.7.2 *Mapreduce*

Se trata de un framework de trabajo que hace posible aislar al programador de todas las tareas propias de la programación en paralelo. Es decir, permite que un programa que ha sido escrito en los lenguajes de programación más comunes, se pueda ejecutar en un cluster de Hadoop.
La gran ventaja es que hace posible escoger y utilizar el lenguaje y las herramientas más adecuadas para la tarea concreta que se va a realizar.

MapReduce es un proceso batch, creado para el proceso distribuido de los datos. Permite de una forma simple, paralelizar trabajo sobre los grandes volúmenes de datos, como combinar web logs

con los datos relacionales de una base de datos OLTP, de esta forma ver como los usuarios interactúan con el website.

El modelo de MapReduce simplifica el procesamiento en paralelo, abstrayéndonos de la complejidad que hay en los sistemas distribuidos. Básicamente las funciones Map transforman un conjunto de datos a un número de pares key/value. Cada uno de estos elementos se encontrará ordenado por su clave, y la función reduce es usada para combinar los valores (con la misma clave) en un mismo resultado.

Un programa en MapReduce, se suele conocer como Job, la ejecución de un Job empieza cuando el cliente manda la configuración de Job al JobTracker, esta configuración especifica las funciones Map, Combine (shuttle) y Reduce, además de la entrada y salida de los datos. (*http://www.pragsis.com, 2012)*

Este framework es basado en Java y es muy útil para el rastreo de los datos, mediante las función MAP y REDUCE, esta funciones están desarrolladas en su totalidad bajo el lenguaje de código abierto Java; las cuales sirven para mapear los datos y reducirlos de manera que la integración que se va a realizar sea de la mejor manera y de un menor tiempo posible.

Este funcionamiento podemos observar en al siguiente figura 1-4.

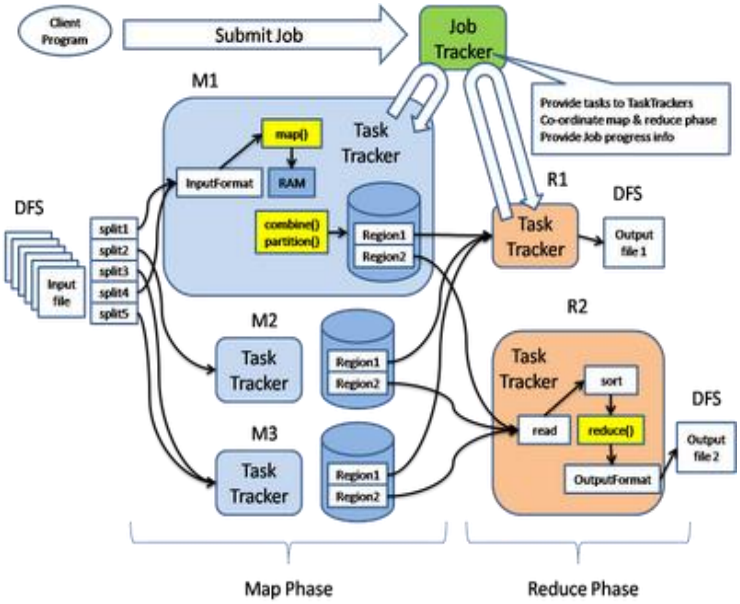


Figura 3-1 Arquitectura MAPREDUCE

Fuente:http://www.pragsis.com/sites/default/files/pdf/Hadoop,%20MapReduce,%20Bidoop.pdf**.**

1.7.3 *Chukwa*.

"Es un sistema de recopilación de datos de código abierto para el seguimiento de grandes sistemas distribuidos. Se construye en la parte superior del sistema de archivos distribuido Hadoop (HDFS) y Map/Reduce. Chukwa también incluye un conjunto de herramientas flexibles para la

visualización, seguimiento y análisis de resultados de los datos recogidos". (www.chukwa.apache.org, 2013,)

En la ilustración nos muestra la interacción de Chukwa con las otras tecnologías y su función que tiene dentro de los procesos de datos.

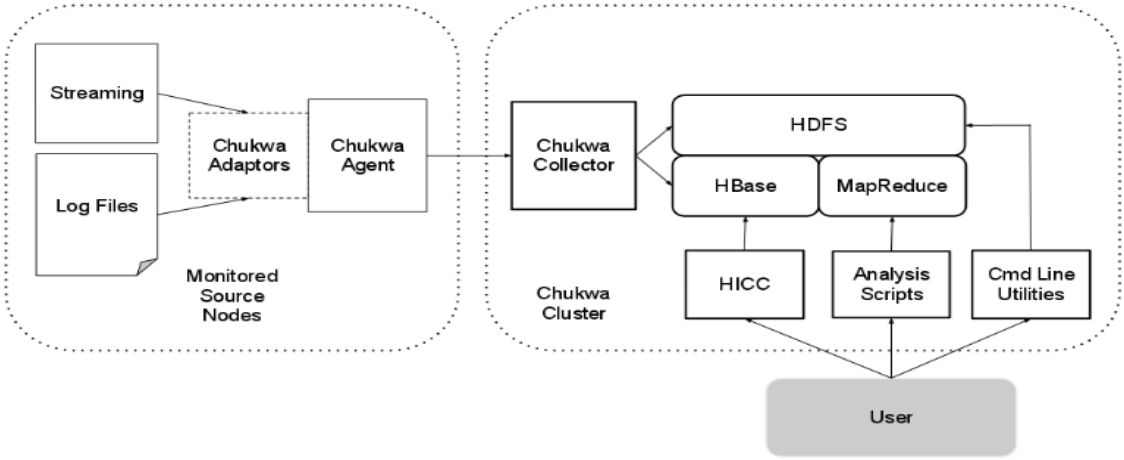


Figura 4-1 Arquitectura Chukwa
Fuente: Arquitectura de Chukwa [En línea]. Disponible en: http://chukwa.apache.org/docs/r0.5.0/admin.html.

1.7.4 *Sqoop*

"Es una herramienta diseñada para transferir datos entre Hadoop y bases de datos relacionales. Sqoop importa los datos de un sistema de gestión de bases de datos relacionales (RDBMS) como MySQL u Oracle al sistema de archivos distribuido Hadoop (HDFS), donde transforma los datos y luego los exporta de nuevo a un RDBMS". (http://sqoop.apache.org/docs, 2012)

"Sqoop automatiza la mayor parte de este proceso, basándose en la base de datos para describir el esquema de importación de los datos. Sqoop utiliza MapReduce para importar y exportar los datos, lo que proporciona el funcionamiento en paralelo, así como tolerancia a fallos".

Es una herramienta muy útil ya que permite la interacción entre Hadoop y las bases de datos relacionales; esto sirve para el almacenamiento de los datos luego del proceso de integración de los datos que se han obtenido de las fuentes de datos Big Data: Sistemas Operaciones, Transaccionales, Redes Sociales, entre otros.

Algunas de sus características son:

- Permite importar tablas individuales o bases de datos enteras a HDFS.
- Genera clases Java que permiten interactuar con los datos importados.
- Además, permite importar de las bases de datos SQL a Hive.

En la Figura 1-6 podemos visualizar la arquitectura de SQOOP para el proceso de integración de los datos,

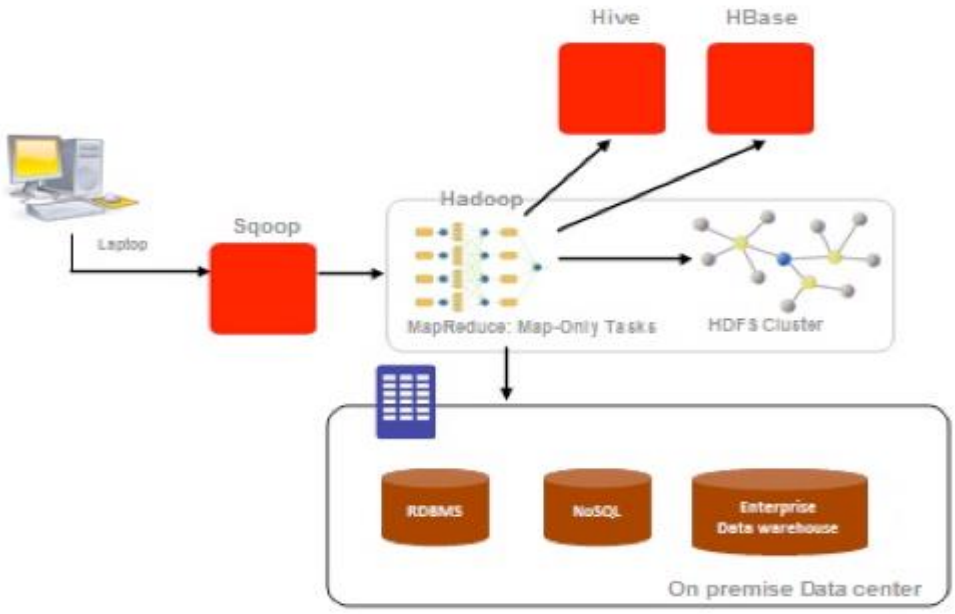


Figura 5-1 Arquitectura de Sqoop
Fuente:http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/.

1.7.5 *Pig*

"Es una plataforma para el análisis de grandes conjuntos de información que consiste en un lenguaje de alto nivel para la expresión de los programas de análisis de datos, junto con la infraestructura necesaria para la evaluación de estos programas. La propiedad más importante es que su estructura es susceptible de paralelismo, que a su vez les permite manejar grandes conjuntos de datos". (http://pig.apache.org/, 2013)

"En la actualidad, la capa de infraestructura de Pig se compone de un compilador que produce secuencias de programas de Mapa-Reduce, para el que ya existen implementaciones paralelas a gran escala (por ejemplo, el Hadoop subproyectos)".

Pig posee las siguientes características:

- Facilidad de programación. Es trivial para lograr la ejecución paralela de tareas de análisis simples. Las tareas complejas compuestas de múltiples transformaciones de datos relacionados entre sí, están codificados explícitamente como secuencias de flujo de datos, lo que hace que sean fáciles de escribir, entender y mantener.
- Oportunidades de optimización. La forma en que se codifican las tareas permite que el sistema pueda optimizar su ejecución de forma automática, lo que permite al usuario centrarse en la semántica en lugar de la eficiencia.
- Extensibilidad: los usuarios pueden crear sus propias funciones para realizar el procesamiento de propósito especial.

Esta herramienta es muy similar a MAPREDUCE, produce secuencias de esta función para poder realizar tareas de análisis de los datos mucho más simples y no tener un proceso de integración muy complejo, de esta manera se reduce el tiempo que toma el proceso de integración y otros beneficios adicionales para el equipo que trabaja con los datos.

1.7.6 H*ive*

Hive es un sistema de DataWarehouse para Hadoop que facilita el uso de la agregación de los datos, ad-hoc queries, y el análisis de grandes Datasets almacenados en Hadoop. Hive proporciona métodos de consulta de los datos usando un lenguaje parecido al SQL, llamado HiveQL. Además permite de usar los tradicionales Map/Reduce cuando el rendimiento no es el correcto. Tiene interfaces JDBC/ODBC, por lo que empieza a funcionar su integración con herramientas de BI y otras herramientas de integración de datos. (http://hive.apache.org/ . 2011 – 2014).

En la figura 1-7 podemos visualizar gráficamente la arquitectura básica del funcionamiento de Hive.

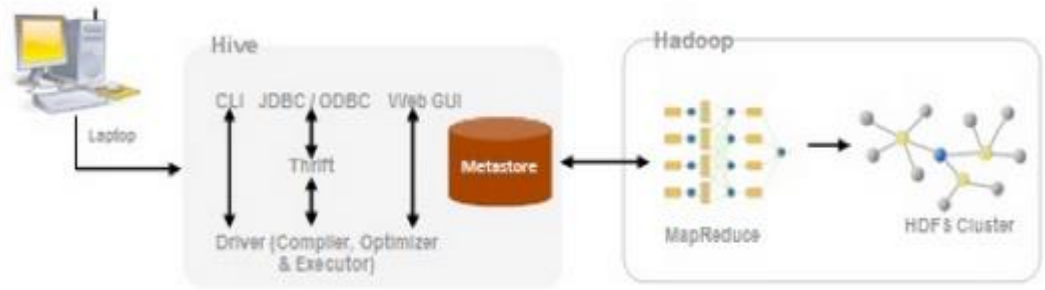


Figura 6-1 Arquitectura de Hive

Fuente: http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/.

1.7.7 *Hbase*

HBase, se trata de la base de datos de Hadoop. HBase es el componente de Hadoop a usar, cuando se requiere escrituras/lecturas en tiempo real y acceso aleatorio para grandes conjuntos de datos. Es una base de datos orientada a la columna, eso quiere decir que no sigue el esquema relacional no admite SQL. (**WPlook***; 2013***)**

Esta herramienta proporciona un acceso en tiempo real y aleatorio para grandes volúmenes de datos, esto es muy importante debido a que no se pierde tiempo en la búsqueda de los datos en las fuentes de datos que estamos utilizando en nuestro proceso.

En la figura 1-8 que se indica a continuación podemos observar la arquitectura básica de HBase de manera gráfica.

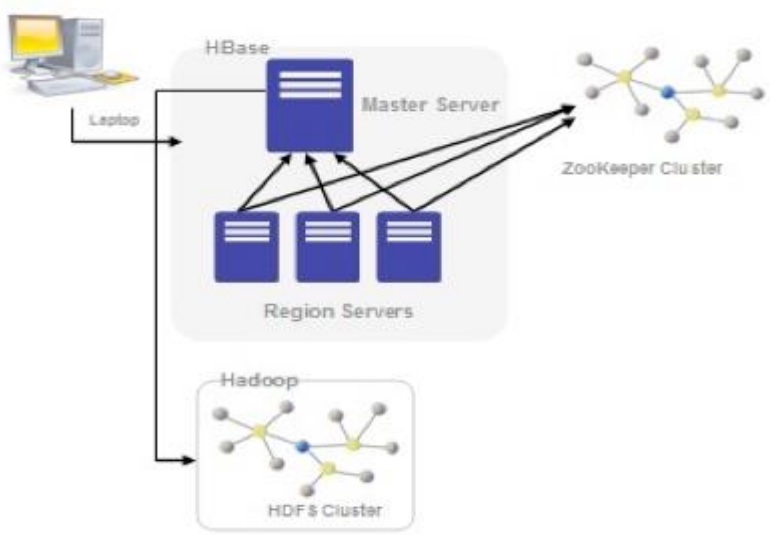


Figura 8-1 Arquitectura Básica de HBase
Fuente: http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/.

1.8 Herramientas de integración compatibles con Big Data

1.8.1 *Pentaho Data Integration (Pdi)*

La herramienta PDI (Pentaho Data Integration) cuenta con la interfaz gráfica llamada Spoon la misma que permitirá realizar la integración de la información de distintos orígenes y destinos. (PÉREZ Arian. , 2011)

Pentaho Data Integration proporciona la Extracción de gran alcance, Transformación y Carga (ETL) utilizando un enfoque innovador, orientado a los metadatos y a los grandes volúmenes de datos BigData. Con una interfaz gráfica de arrastre, y una probada arquitectura escalable y basada en estándares; en las organizaciones se están inclinando por Pentaho Data Integration para la integración de datos.

Pentaho Data Integration unifica el ETL, modelado y visualización de procesos en un entorno único e integrado que permite a los desarrolladores y usuarios finales a trabajar juntos sin problemas. (*www.cognus.biz,* 2011**,)**

PENTAHO es un proyecto iniciado por una comunidad OpenSource, provee una alternativa de soluciones de BI en distintas áreas como en la Arquitectura, Soporte, Funcionalidad e Implantación. Estas soluciones al igual que su ambiente de implantación están basados en JAVA, haciéndolo flexible en cubrir amplias necesidades empresariales.

A través de la integración funcional de diversos proyectos de OpenSource permite ofrecer soluciones en áreas como: Análisis de información, Reportes, Tableros de mando conocido como "DashBoards", Flujos de Trabajo y Minería de Datos. (*http://www.dataprix.com, 2011*)

En base a los conceptos anteriormente citados se concluye que Pentaho Data Integrator es una herramienta de integración de gran funcionalidad y compatibilidad con lo que es Big Data, es de código abierto y ayuda mucho para los procesos de manejos de datos, posee una interfaz muy amigable y fácil de utilizar.

1.8.1.1. *Antecedentes*

Pentaho, creada en el 2004 es el actual líder en cuanto a soluciones de Business Intelligence Open Source. Ofrece, con soluciones propias, todo el espectro de recursos para desarrollar, mantener y explotar un proyecto de B.I. Desde las ETL con Data Integration hasta los cuadros de mando con el Dashboard Designer o el Comunity Dashboard Framework. (PUENAYÁN, Adriana Del Rocío, & AYNAGUANO, Diana; 2012)

La forma como Pentaho ha construido su solución de B.I. es integrando diferentes proyectos ya existentes y de solvencia reconocida. Data Integration anteriormente era Kettle, de hecho sigue conservando su antiguo nombre como nombre coloquial. Mondrian es el otro componente de Pentaho que sigue manteniendo entidad propia. (SALINAS, Alexandro, 2008)

1.8.1.2 *Requisitos previos a la instalación de PDI*

Requisitos mínimos de hardware

- Procesador: Celeron, 2.0GHz
- Memoria RAM: 128Mb o superior.
- Espacio libre en disco duro: 200Mb

Requisitos mínimos de software

- Máquina virtual de Java (JRE) versión 1.5 o superior.
- MySQL versión 5 o posteriores. (PEREZ Arian., 2011)

En base a estos requerimientos citados se observa que no es ninguna complicación su instalación, además funciona en máquinas relativamente potentes, lo que haces que esta herramienta sea muy utilizada en el mercado en la actualidad para este proceso que es muy requerido en el ámbito de la tecnología y la información.

1.8.1.3 *Características de Pentaho Data Integrator*

a) Pentaho Data Integration incluye un conjunto de componentes para realizar ETL (Extracción, transformación y carga de datos). Uno de sus objetivos es que el proyecto ETL sea fácil de generar, mantener y desplegar.
b) Exploración del repositorio de archivos (tablas, vistas) y metadatos.
c) Asistente para la creación de conexiones a base de datos.
d) Soporta muchos tipos de transformaciones básicas: mapeo de campos y valores, filtrado de filas, ordenación, secuencias, partición de campos, agrupación, adicción de constantes, normalización/desnormalización de filas, uniones (join) de filas, fusión de filas y algunas operaciones matemáticas.
e) Ejecución de procedimientos almacenados y SQLs.
f) Las transformaciones pueden ser llamadas los trabajos y los trabajos pueden ser llamados por otros trabajos. Por ello hay mecanismos para pasar la información entre transformaciones y trabajos.
g) Ejecución de shellscript y comprobación de la existencia de ficheros y tablas.
h) Definición del intervalo de ejecución en el planificador de trabajos.
i) Entorno gráfico de desarrollo.
j) Código: Aplicación 100% Java con transformaciones. Diseño orientado a metadatos.
k) Conectividad: Soporta Oracle, DB2, SQL Server y Sybase. Compatible también con MySQL, PostGres, Hypersonic, FireBird SQL e Ingres. Soporta conectividad con SAP R/3 si se paga la licencia.
l) El diseño de la interfaz puede resultar un poco pobre, y no hay una interfaz unificada en todos los componentes, siendo en ocasiones la interfaz de los componentes confusa.
m) Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).
n) Evolución mucho más lenta de la herramienta e incierta, pues Pentaho Data Integration tiende a abandonar la parte Software Libre.
o) Incluye cuatro herramientas, complementarias para hacer uso de los ETLs realizados en PDI:

1. Spoon: para diseñar transformaciones ETL usando el entorno gráfico.
2. PAN: para ejecutar transformaciones diseñadas con spoon.
3. CHEF: para crear trabajos.
4. Kitchen: para ejecutar trabajos.

p) Los principales transformadores con los que cuenta la herramienta es:

1. Diseño de dimensiones.
2. Diseño de tablas FAC. (www.datprix.com, 2011)

1.8.2 *Talend Open Studio (Tos)*

Talend Open Studio es una herramienta Open Source de integración y gestión de datos, así como integración de aplicaciones empresariales: en palabras simples una herramienta ETL.
Talend está basado en Java, requiere específicamente JDK 6 y por tanto puede ser ejecutado en Windows y Linux sin mayor dificultad, solo basta con descomprimir su 'instalador". (AVILÉS Marco, 2011)

1.8.2.1 N*iveles de Talend Open Studio*

Talend basa su diseño en 3 niveles:

- Business Models (Modelos de Negocios): es nivel diseñado para modelar de manera teórica la aplicación, para lo cual se realizan diagramas de flujo básicos con actores de los procesos.
- Job Designs (Diseño de Trabajos): el nivel más interesante, en el cual se diseña el trabajo en sí, el código que será ejecutado.
- Contexts (Contextos): él es nivel que contiene los contextos, los cuales pueden ser definidos como variables globales de ejecución del programa, como la carpeta donde se ejecutará la aplicación final o variables iniciales de entrada. (AVILÉS Marco, 2011)

De acuerdo a las definiciones citadas se puede mencionar que Talend Open Studio posee requerimientos accesibles para su instalación, así como también una arquitectura de trabajo muy adecuados para los usuarios, además posee una interfaz gráfica muy amigable para su uso, lo cual es muy importante en las herramientas de integración de datos.

1.8.2.2 *Antecedentes*

Talend Open Studio se basa en programación por componentes (para algunos en cajitas) por lo que el desarrollo de cualquier script/programa varía bastante respecto a la programación habitual. Esta forma de programar consiste en ir uniendo diferentes componentes con funcionalidades diversas mediante sus flujos de entrada y flujos de salida para realizar una tarea más compleja. Como ya os habréis dado cuenta, esto cumpliría con el método algorítmico de Divide y vencerás (DYV) que consiste en resolver un problema complejo dividiéndolo en partes más simples tantas veces como sea necesario, hasta que la resolución de las partes sencillas se torne obvia. Con lo que la solución del problema principal se construye a partir de las soluciones de los problemas más simples. (MADRID, Víctor, 2010)

1.8.2.3 Requisitos previos de instalación

Requisitos mínimos de hardware

- Procesador de arquitectura Pentium de 2.0 GHZ.
- 3 GB de memoria RAM.
- Disco Duro con al menos 3 GB libres.

Requisitos de software

- Java Home Enviroment 5 o posteriores.
- MySQL versión 5 o posteriores. (*https://help.talend.com)*

1.8.2.4 Características de Talend Open Studio

- Sincronización o replicación de bases de datos
- Intercambios de datos en el momento correcto o por lotes entre los sistemas de la infraestructura de TI
- Migración de datos
- Transformación y carga de datos complejas
- Modelo de proceso orientado al negocio.
- Repositorio centralizado: Información de todos los proyectos.
- Consistencia de datos y reutilización de componentes.
- Rápido desarrollo.
- Fácil de mantener.
- Desarrollo gráfico.
- Aumenta la productividad.
- Combina vistas gráficas con técnicas.
- Arrastrar y soltar componentes en la ventana de diseño.
- Amplia gama de componentes y conectores.
- Ejecución robusta y escalable.
- Proceso distribuido en red. Aprovecha al máximo el hardware. Indicado para todo tipo se servidores (gama alta y baja), ya que maximiza la tasa de utilización de los recursos.
- Genera código estándar. Utiliza motores optimizados (JAVA o PERL). (MADRID, Víctor; 2010)

Las características que posee Talend Open Studio lo sitúan en una posición muy privilegiada de las herramientas de integración de datos, debido a su gran funcionalidad y a su desempeño con los datos.

1.8.3 *Scriptella ETL*

1.8.3.1 *Antecedentes*

Las herramientas de análisis e integración de datos han sufrido una transformación asombrosa; con el avance del tiempo, por lo cual se hace necesario el estudio de estas herramientas; para tener conocimiento sobre su funcionamiento y uso para integrar datos desde las diferentes fuentes de información, que en la mayoría son fuentes Big Data.

1.8.3.2 *Definición*

Scriptella es una herramienta de integración basada en Java y herramientas de ejecución de scripts. El lenguaje de scripts primario es un lenguaje SQL normal ejecutado por un conector JDBC. Al mismo tiempo se puede mesclar con otros lenguajes de ejecución de scripts, para lograr un mejor desempeño y evitar perdida de datos.

1.8.3.3 *Requisitos previos de instalación*

Requisitos mínimos de hardware

- Procesador de arquitectura Pentium de 2.0 GHZ.
- 512 MB a 1 GB de memoria RAM.
- Disco Duro con al menos 3 GB libres.

Requisitos de software

- JDK 5 o posteriores.
- JRE 5 o posterior. (*http://mscerts.programming4.us*, 2012 – 2014.)

1.8.3.4 Características Scriptella ETL

- Sintaxis simple de XML para las escrituras. Agregar la dinámica a sus escrituras existentes del SQL creando un archivo fino de la envoltura XML:
- Ayuda para los datasources múltiples (o las conexiones múltiples a una sola base de datos) en un archivo de ETL.
- Ayuda para muchas características útiles de JDBC, e.g. parámetros en el SQL incluyendo referencias a las gotas del archivo y al escape de JDBC.
- Funcionamiento. El uso del funcionamiento y de la memoria baja es una de nuestras metas fundamentales.

- Ayuda para las expresiones y las características evaluadas (sintaxis de JEXL)
- La ayuda para la cruz-base de datos ETL scripts usando elementos.
- Ejecución transaccional.
- Tratamiento de errores vía elementos.
- Escrituras/ejecución condicionales de las preguntas (similar a las cualidades de la hormiga if/unless pero más de gran alcance).
- Fácil de utilizar como una herramienta o tarea independiente de la hormiga. Ningún despliegue/instalación requerida.
- Fácil-A-Funcionar los archivos de ETL directo del código de Java.
- Adaptadores incorporados para las bases de datos populares para una integración apretada. Ayuda para cualquier base de datos con el conductor obediente de JDBC/ODBC.
- Interfaz de Service Provider (SPI) para la interoperabilidad con DataSources del non-JDBC e idiomas scripting. Fuera de la ayuda de la caja para los idiomas compatibles de JSR 223 (Scripting para la plataforma de Java).
- Abastecedores incorporados de CSV, de XML, de LDAP, de Lucene, de la velocidad, de Excel, del texto, de JEXL y de Janino.
- Integración con Java EE, el marco del resorte, JMX, JNDI y JavaMail para las escrituras listas de la empresa. (*http://mscerts.programming4.us*, 2012 – 2014.)

CAPITULO II

2. MARCO METODOLÓGICO.-

2.1 ESTUDIO COMPARATIVO DE LAS HERRAMIENTAS DE INTEGRACIÓN DE DATOS

2.1.1 Elección de las herramientas a utilizar

Las herramientas que se va a utilizar para el desarrollo de los prototipos de prueba son Talend Open Studio y Pentaho Data Integrator; se han seleccionado de acuerdo a los puntos analizados en el marco teórico; tomando en cuenta la disponibilidad en el mercado, posición frente a otras herramientas, ambiente de uso, licenciamiento, entre otros aspectos; además se eligió la tecnología de manejo de Big Data a la herramienta SQLServer, basándose en las características que presentan estas tecnologías que se han analizado en el Marco Teórico.

2.1.2 Determinación de los escenarios de comparación

Para la realización de nuestro estudio investigativo se procede a realizar los siguientes escenarios de prueba para realizar los prototipos y verificar de esta manera el rendimiento que posee cada una de las herramientas al trabajar con grandes volúmenes de datos (Big Data).

2.1.2.1 *Escenario Pentaho Data Integrator*

Para este escenario utilizaremos la Herramienta de Integración de Datos PENTAHO DATA INTEGRATOR.

- Fuentes Big Data (Bases de Datos de las escuelas de la Facultad de Informática y Electrónica)
- Pentaho Data Integrator
- SQL Server

2.1.2.2 Escenario Talend Open Studio

Para este escenario se utilizará la herramienta de Integración Talend Open Studio:

- Fuentes Big Data (Bases de Datos de las escuelas de la Facultad de Informática y Electrónica)

- Talend Open Studio
- SQL Server

2.2 Determinación del ámbito de los parámetros de comparación

Para el trabajo de investigación a desarrollar se tomará en cuenta los siguientes parámetros y Sub parámetros de comparación, los mismos que se encuentran reflejados en Tabla 2-1.

Tabla 1-2 Parámetros de Comparación

ÁMBITO DE LOS PARÁMETROS	PARÁMETROS	SUB PARÁMETROS
ATRIBUTOS PROPIOS DE LA HERRAMIENTA	Conectividad	Soporte de conexión a las fuentes Big Data.
		Seguridad de acceso a datos.
		Control de errores
	Compatibilidad	Compatibilidad de tipos de datos
		Soporte de tipos de datos.
	Funcionalidad	Carga de datos desde las fuentes Big Data.
		Soporte de sentencias SQL
		Manejo de Integración de datos.
		Soporte Técnico
ATRIBUTOS DE USABILIDAD	Interfaz	Interfaz en el proceso de integración de datos.

Fuente: Autor
Realizado por: Guido López

2.3 Descripción de los sub parámetros de comparación

2.3.1 A*tributos propios de la herramienta*

- Conectividad

La conectividad hace referencia hacia la disponibilidad de las fuentes que posee cada una de las herramientas que se van a utilizar en el desarrollo del prototipo del observatorio de indicadores de la Facultad de Informática y Electrónica de la ESPOCH.

- Soporte de conexión a las fuentes Big Data.- Se analiza si la herramienta de integración de datos requiere de un ODBC para su conexión o se conecta a las fuentes de forma directa.
- Seguridad de acceso a datos.- Se refiere al nivel de seguridad que posee cada una de las fuentes Big Data; así como las fuentes de destino.

- Control de errores.- La herramienta debe permitir reconocer y emitir mensajes de error, para de esta manera irlos corrigiendo en el momento.

- Compatibilidad.

  La compatibilidad es la condición que hace que una base de datos logre compactarse correctamente tanto directamente o indirectamente con las diferentes plataformas de tecnologías y bases de datos destino.

- Compatibilidad de tipos de datos.- Esto se mide en base a que en el momento de realizar la carga de datos no exista conflicto con las fuentes de datos así como los destinos de la herramienta de manejo de Big Data.
- Soporte de tipos de datos.- La herramienta deberá contar con una adecuación de todos los tipos de datos que posee la herramienta Hadoop, para lograr una integración de forma correcta y no produzca errores al momento de cargar los datos.

- Funcionalidad

Son varias las características funcionales que posee una herramienta de manejo de Big Data y las Bases de datos de destino, con las cuales se puede establecer una solución adecuada para el manejo de datos dentro de una organización.

- Carga de datos desde las fuentes Big Data.- La herramienta debe permitir medir la cantidad de datos obtenidos y el tiempo que ocupa desde la fuente y desde las herramientas de la tecnología Hadoop, para efectuar el proceso de integración.
- Rendimiento en el proceso de integración.- Efectuar una medición del rendimiento, esto es uso de Memoria, CPU y el empleo de la red que posee en efectuar el proceso de integración de datos.
- Soporte de sentencias SQL.- Las herramientas deben permitir el uso de sentencias SQL, para realizar el proceso de integración de los datos; la unión tablas y otras opciones de selección y carga de datos.
- Manejo de integración de datos.- Se medirá si las herramientas de integración de datos manejas claven primaria o secundarias al momento de realizar el proceso de integración de los datos desde las fuentes Big Data,
- Soporte Técnico.- Hace referencia a los tipos de soporte técnico que se puede dar a las herramientas de integración; en ocasiones se lo realiza mediante la web o también de forma personal acudiendo a las instalaciones de las empresas que brindan el software de integración de datos.

2.3.2 *Atributos de usabilidad.*

- Interfaz

Se refiere a la interfaz de trabajo que cada herramienta de integración de datos dispone, de la facilidad de hacer un proceso carga de datos desde las fuentes Big Data, y el almacenamiento en los repositorios de cada una de las herramientas.

- Interfaz en el proceso de integración de datos.- Las herramientas pueden tener dos formas de realizar un proceso integración de datos

La primera mediante una interfaz gráfica la cual hace que el usuario interactúe y establezca un contacto más fácil e intuitivo con la herramienta.

La segunda es de forma manual realizando todos los pasos median la línea de comandos escribiendo código fuente para que el ordenador interprete que debe realizar alguna acción.

2.4 Definición de los pesos de ponderación.

Para realizar la comparación entre las herramientas y determinar su rendimiento en el proceso de integración de datos con fuentes Big Data se les asigna a cada uno de los parámetros, los correspondientes pesos para poder determinar un mejor rendimiento en el desarrollo de las actividades que involucra este proyecto.

Los pesos asignados a cada uno de los parámetros y sub-parámetros que se van a utilizar se los muestra en la Tabla 3 que se indica a continuación.

Tabla 2-2 Determinación de los pesos de ponderación para los atributos.

PARÁMETROS	SUB PARÁMETROS	PESO
Conectividad	Soporte de conexión a las fuentes Big Data.	5
	Seguridad de acceso a datos.	10
	Control de errores	5
TOTAL CONECTIVIDAD		**20**
Compatibilidad	Compatibilidad de tipos de datos	10
	Soporte de tipos de datos.	10
TOTAL COMPATIBILIDAD		**20**
Funcionalidad	Carga de datos desde las fuentes Big Data.	30
	Soporte de sentencias SQL	5
	Manejo de Integración de datos.	5
	Soporte Técnico	10
TOTAL FUNCIONALIDAD		**50**
Interfaz	Interfaz en el proceso de integración de datos.	10
TOTAL INTERFAZ		**10**
TOTAL		**100**

Fuente: Autor
Realizado por: Guido López

2.5 Determinación de las condiciones para la asignación de los pesos de los parámetros de comparación atributos propios de la herramienta

- Conectividad

Este parámetro posee una importancia moderada para nuestro trabajo investigativo por lo que posee un valor del 20%, el mismo que se ha repartido entre los sub-parámetros de la siguiente manera.

- Soporte de conexión a las fuentes Big Data.- El valor máximo que puede alcanzar este indicador es 5, en caso de no cumplir con la conexión a los datos el valor es 0.
- Seguridad de acceso a datos.- El valor de máximo de este indicador es 10 si cumple con todos los requisitos de seguridad; en el caso de contar con la seguridad en un bajo nivel tendrá un valor de 5; y si no posee ninguna seguridad su valor es 0**.**
- Control de errores.- El valor para este indicador si presenta una interfaz fácil al mostrar un error y su explicación es 5, en caso de presentar solo el error y no su explicación, su puntuación es 2 y en caso de no poseer ninguna de las dos alternativas su valor será 0.

- Compatibilidad.

Este parámetro de comparación posee una relevancia del 20%, para lo que se ha dividido en sus sub-parámetros para el establecimiento de valores de medición.

- Compatibilidad de tipos de datos.- Este indicador posee un valor máximo de 10 en caso de ser totalmente compatible con los tipos de datos de las fuentes, caso contrario se evaluará cada tipo con un valor de 1.
- Soporte de tipos de datos.- Este indicador tendrá un valor global de 10, en caso contrario se valorará cada uno de los tipos de datos con un valor de 1.

- Funcionalidad

Este parámetro está considerado como fundamental para el desarrollo del presente estudio comparativo, por lo que posee un valor de 50%, para el análisis y determinación de resultados.

- Carga de datos desde las fuentes Big Data.- Para este parámetro se debe tener en cuenta, los siguientes indicadores a utilizar para poder determinar los tiempos de carga de los datos desde las fuentes, los procesos de integración y los resultados en el destino.
- Soporte de sentencias SQL.- El valor máximo que posee este indicador es de 5, si permite un alto nivel de joins y condiciones entre las tablas; en caso contrario tendrá un valor de 3.
- Manejo de integración de datos.- El valor máximo de este indicador es 5, si manejas las claves primarias y foráneas; en caso de no manejarlas se puntuará con un valor de 2,5 a cada una de ellas.

- Soporte Técnico.- Se calificará con un valor de 10 si posee las dos formas de soporte técnico; en caso contrario se puntuará con un valor de 5 a cada una de ellas.

2.6 Determinación de las condiciones para la asignación de los pesos de los parámetros de comparación atributos de usabilidad.

- Interfaz
- Interfaz en el proceso de integración de datos.- A este indicador se le puntuará con un valor máximo de 10 si posee una interfaz interactiva y fácil de manejar; además mediante líneas de comandos; en caso de tener una de las dos, se lo dará un puntaje de 5; en caso de no contar con ninguna se lo dará un valor de 0.

2.7 Desarrollo de las pruebas de integración de los datos

Para el desarrollo del presente trabajo investigativo se va a desarrollar los prototipos de trabajo para determinar el rendimiento que poseen las herramientas de integración citadas en el Marco Teórico.

2.7.1 Desarrollo del prototipo con la herramienta Talend Open Studio (tos)

Para el desarrollo del prototipo que se va a utilizar emplearemos las siguientes herramientas: Microsoft SQL Server, Talend Open Studio, Gestores de Bases de Datos, Fuentes de Información Big Data.

Para este prototipo se utiliza las siguientes herramientas:

- La fuente de datos para el prototipo se va a utilizar Bases de Datos de la Facultad de Informática y Electrónica; con sus escuelas de Ingeniería en Sistemas, Ingeniería en Diseño Gráfico, Ingeniería en Redes Y Telecomunicaciones, Ingeniería en Redes y Control Industrial, e Ingeniería Electrónica. Las mismas que utilizan el Gestor de Base de Datos Microsoft SQL Server, las mismas que se pueden visualizar en el Anexo A.
- La Herramienta de Integración de Software Libre Talend Open Studio, disponible en versiones de manera libre; como se puede observar en el Anexo B.
- Un repositorio para la realización de los Jobs y Transformaciones necesarias para el trabajo, el que está realizado en Microsoft SQL Server; como se puede visualizar en el Anexo C.

2.7.1.1 *Desarrollo del Prototipo*

- Se instala la versión de SQL Server para almacenar las bases de datos que sirven de Fuentes para el prototipo.
- Se realiza la restauración de las Bases de Datos de la Facultad de Informática y Electrónica de la ESPOCH, como se puede visualizar en el anexo D.
- Se ejecuta la herramienta de Integración de Datos Talend Open Studio para la realización de los trabajos de integración de datos que se utilizaran en el presente trabajo de investigación; como se puede visualizar en el Anexo E.
- Luego se realiza las conexiones correspondientes con los orígenes de datos, mediantes la herramienta Talend Open Studio y Microsoft SQL Server.

- Para realizar la conexión se sigue lo siguiente.
- Se selecciona la opción de Metadata, luego seleccionamos la opción Db Conections, luego clic derecho y optamos por la opción Crear Conexión, seleccionamos un nombre para la conexión, y damos siguiente; seleccionamos el tipo de Base de Datos a la que queremos conectarnos, en nuestro caso SQL Server, ingresamos los datos de conexión y damos clic en Finalizar; cómo podemos observar en el anexo F.
- na vez realizada la conexión hacia todas fuentes de datos, de nuestra Facultad de Informática y Electrónica; para obtener todos los indicadores académicos necesarios.

Para la determinación de los parámetros necesarios para la comparación entre las herramientas, realizamos los procesos necesarios para obtener la información necesaria.
Se desarrolla un índice académico como ejemplo; los demás indicadores se encuentran en los anexos correspondientes.

Malla curricular (por pensum de estudios):

a. Títulos que se ofertan

Para la realización de este indicador se realiza lo siguiente:
Se selecciona el origen de los datos para nuestro Job, como se muestra en la imagen.

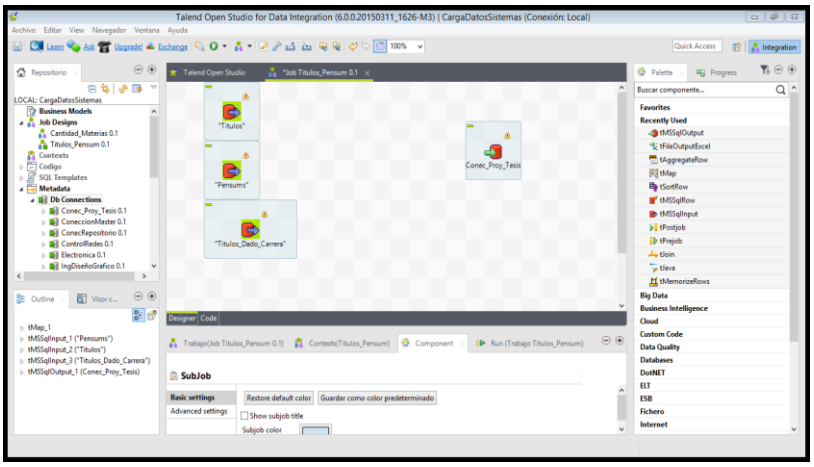


Figura 1-2 Selección de los orígenes y destino de los datos
Realizado por: Guido López

Luego seleccionamos la opción tMap y arrastramos hacia el ambiente de trabajo.

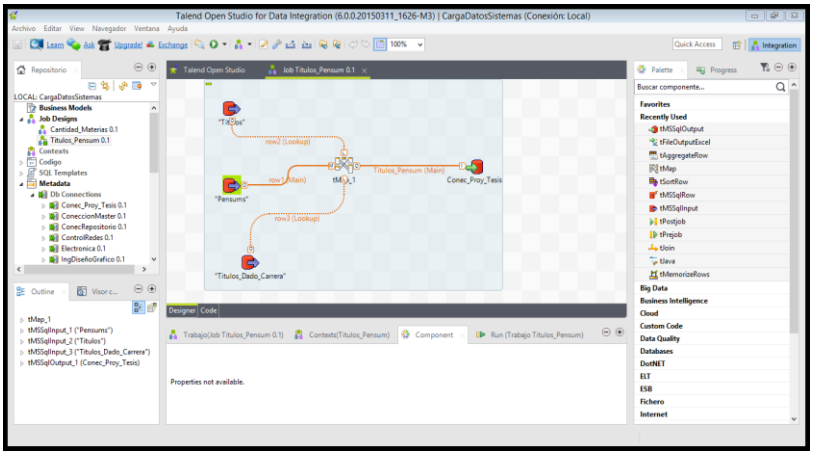


Figura 2-2 Selección de la opción para seleccionar los campos a utilizar.
Realizado por: Guido López

En la opción Tmap, se asigna los campos a utilizar en el trabajo de integración y los resultados a mostrar, teniendo una imagen como la siguiente.

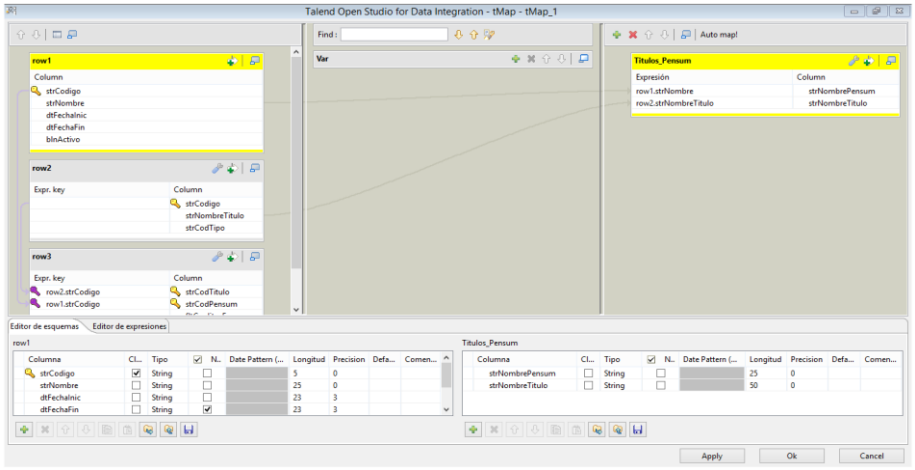


Figura 3-2 Selección de los campos a utilizar en los destinos.
Realizado por: Guido López

Para ejecutar el Job, seleccionamos la opción Run; luego se cargan los datos necesarios y operaciones correspondientes; además se muestra los tiempos empleados en la realización de estas tareas.

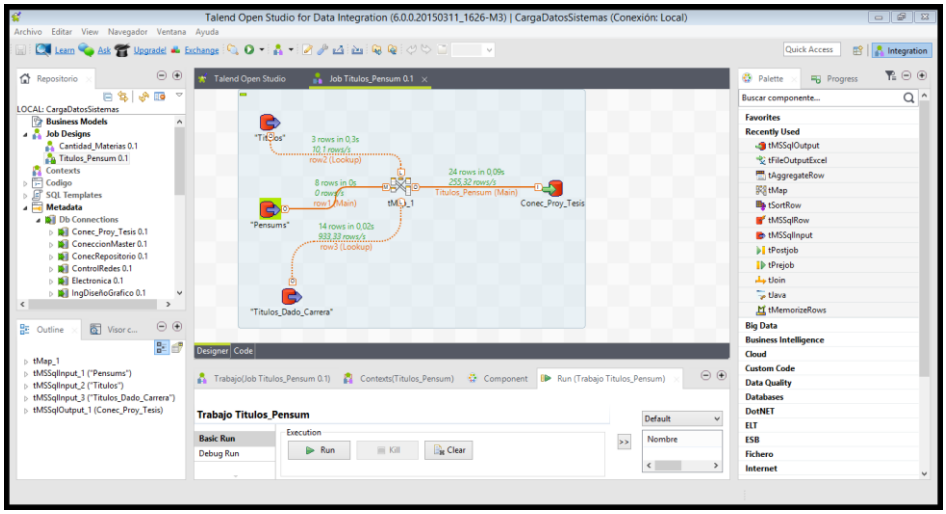


Figura 4-2 Selección de los campos a utilizar en los destinos.
Realizado por: Guido López

2.7.1.2 Desarrollo del prototipo con la herramienta Pentaho Data Integrator (PDI)

Para la realización de este prototipo de análisis de la integración de datos con fuentes Big Data se va a utilizar las siguientes herramientas: Fuentes de Información Big Data, Pentaho Data Integrator, Gestores de Bases de Datos.

Este prototipo se lo realiza de acuerdo a la utilización de las siguientes herramientas.

- Su utilizaran las mismas fuentes de datos, para de esta manera poder obtener los resultados más exactos posibles, así poder determinar cuál de las herramientas citadas posee un mejor rendimiento en la integración de datos.
- La herramienta de integración Pentaho Data Integrator, disponible en la versión Comunity para poderla utilizar de manera gratuita todos los módulos para los trabajos de integración de datos; cómo podemos observar en el anexo G.
- Un repositorio de datos en el Gestor de Datos Microsoft SQL Server, que permitirá utilizar las opciones para realizar las transformaciones y Jobs necesarios para la realización del prototipo de prueba; como se puede visualizar en el anexo H.

Para este caso se inicia la Herramienta de Integración Pentaho Data Integrator, para realizar las tareas de integración mediante el uso de Trasformaciones y Jobs; en base a los indicadores necesarios para el estudio comparativo de las herramientas.

Esta operación que se realizarán para la determinación de los tiempos de carga de los datos, se desarrolla y se encuentran en los anexos; para mostrar un ejemplo se desarrolla un indicador correspondiente con la herramienta Pentaho Data Integrator.

Malla curricular (por pensum de estudios):

a. Títulos que se ofertan

Se selecciona la entrada de datos, un paso para ordenar filas y la salida de datos; para obtener el indicador solicitado, como se puede observar en la imagen a continuación.

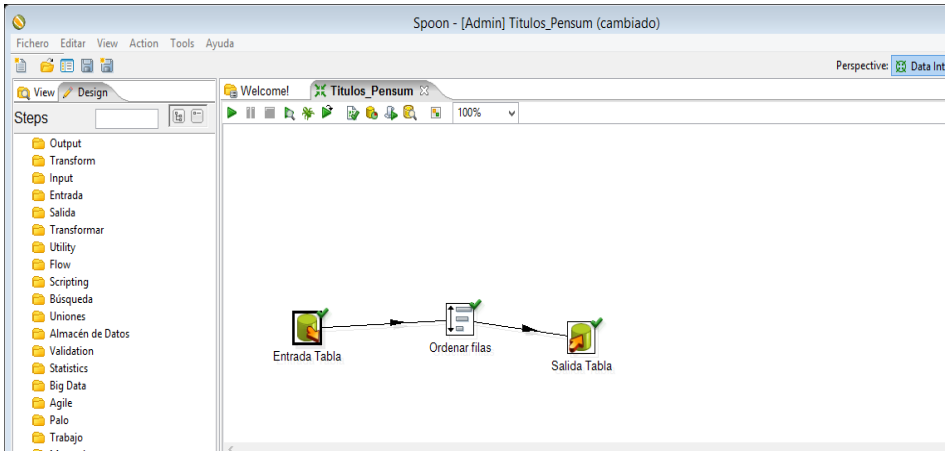


Figura 5-2 Selección de la entrada de datos.
Realizado por: Guido López

Luego damos doble clic en la Entrada de Datos, y colocamos la sentencia SQL para obtener dicho indicador, como podemos visualizar en la pantalla que se muestra.

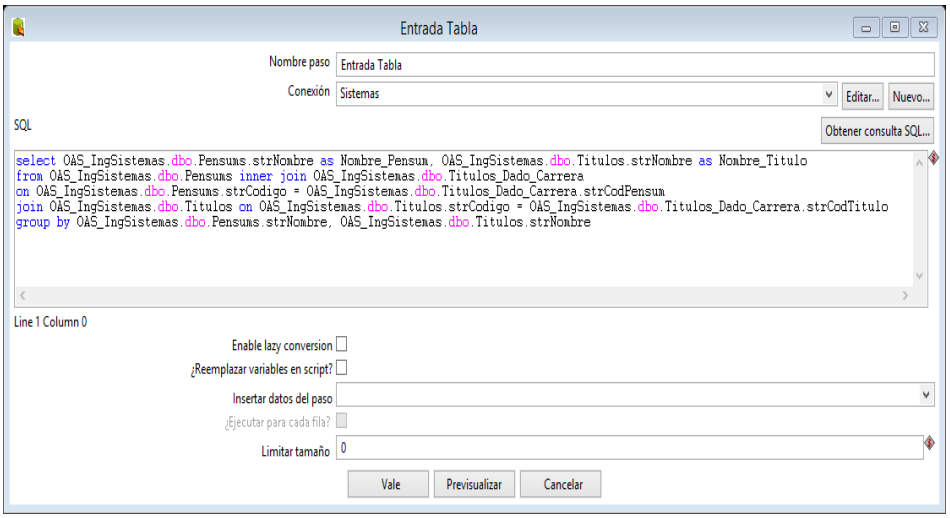


Figura 6-2 Sentencia SQL de la entrada de datos.
Realizado por: Guido López

Luego debemos dar doble clic en el salto de ordenar filas e ingresamos los campos para ordenar los datos, como se muestra en la figura que se indica.

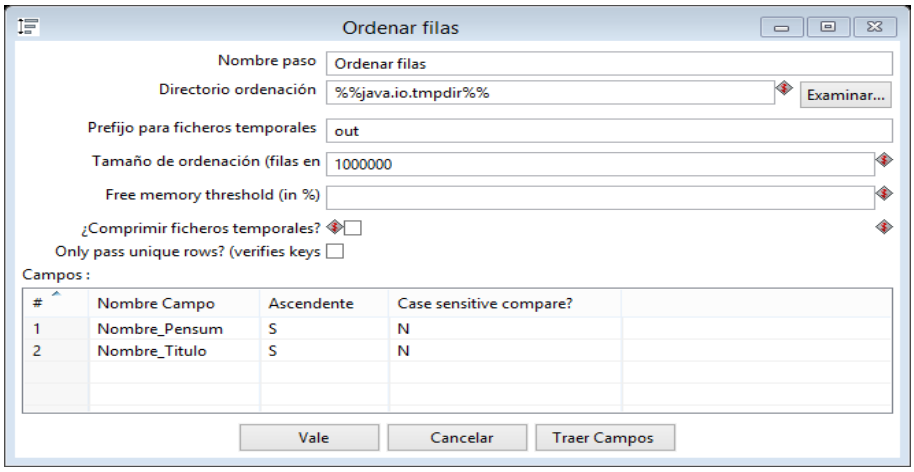


Figura 7-2 Selección de los campos para el ordenamiento de los registros
Realizado por: Guido López

Luego seleccionamos con doble clic, sobre el paso de salida de datos; en la cual debemos ingresar algunos datos, la conexión hacia el destino de los datos, así como el esquema y la tabla donde se va a guardar la información; como se puede visualizar en la siguiente imagen.

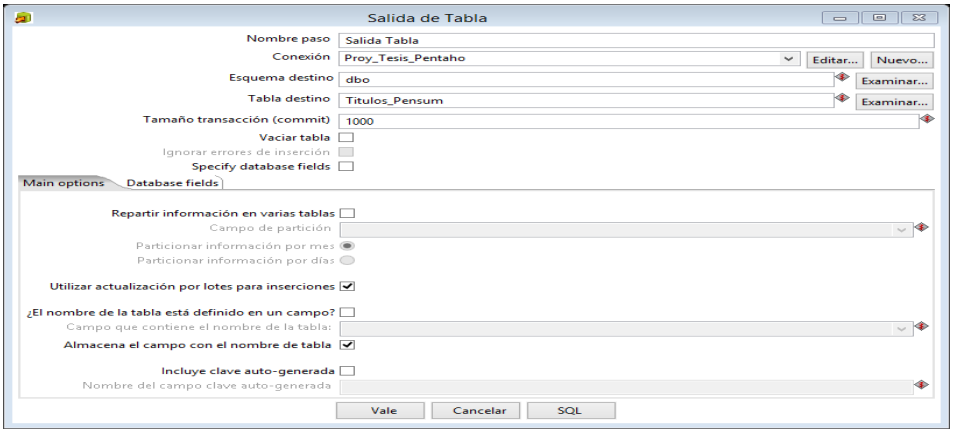


Figura 8-2 Selección de la tabla destino de los datos que se seleccionan
Realizado por: Guido López

Luego se selecciona la opción SQL de la parte inferior para crear la tabla donde se va a almacenar los datos del proceso de integración.

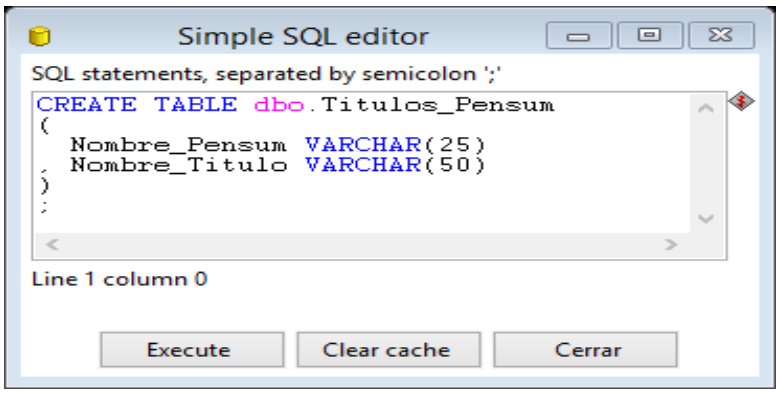


Figura 9-2 Ejecución de las sentencias de creación de los campos
Realizado por: Guido López

Luego, se selecciona la opción de Execute, después la opción Cerrar.

Finalmente se ejecuta la transformación que se está realizando, para obtener el resultado deseado.

Para los demás trabajos de integración los resultados se encuentran en los anexos correspondientes.

CAPITULO III

3. MARCO DE RESULTADOS, DISCUSIÓN Y ANÁLISIS DE RESULTADOS

3.1- Análisis de los resultados

En base a los parámetros de valoración anteriormente expuestos se puede verificar los siguientes resultados:

3.1.1 *Conectividad*

- Soporte de conexión a las fuentes Big Data.- Se analizará la conexión hacia las fuentes de datos en las dos herramientas.

- Talend Open Studio.- En este indicador esta herramienta posee un valor de 5 debido a que facilita la conexión hacia las fuentes de datos Big Data; cómo podemos observar en la imagen siguiente.

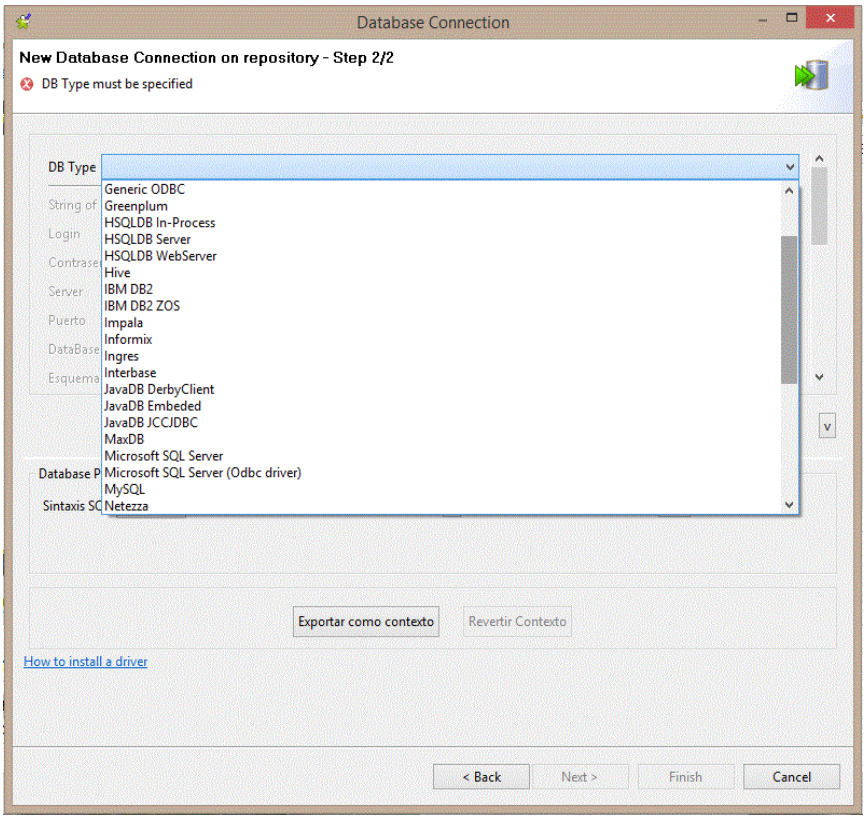


Figura 1-3 Disponibilidad de Fuentes de Datos con Talend Open Studio
Realizado por: Guido López

- Pentaho Data Integration.- Esta herramienta posee un valor de 5 en este indicador, debido a que posee una muy buena conexión hacia las fuentes de datos Big Data; para los trabajos necesarios.

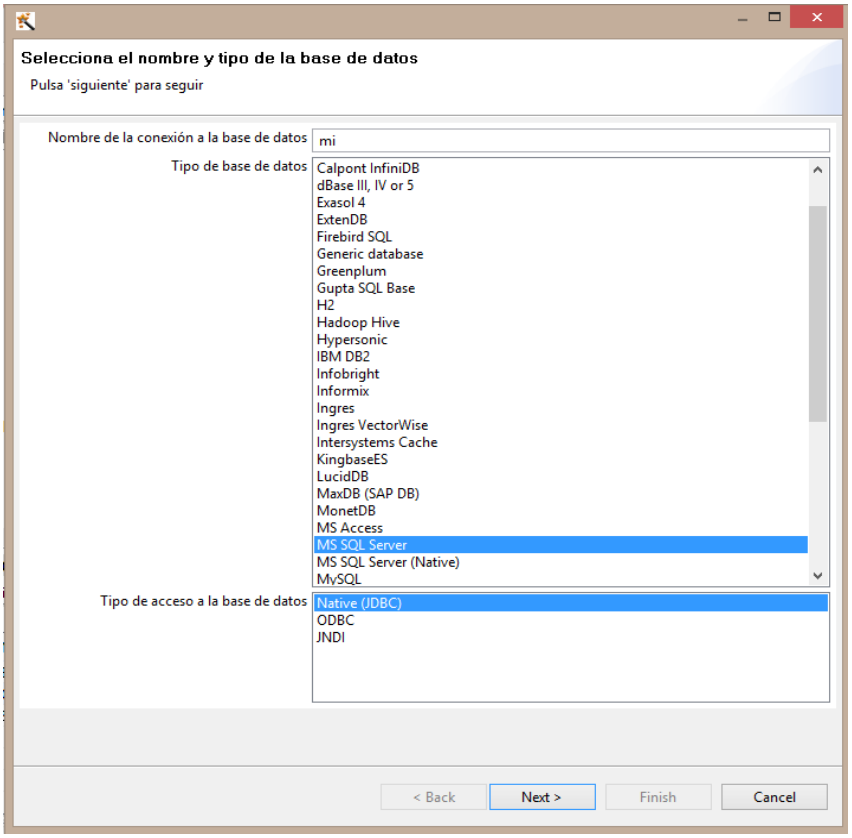


Figura 2-3 Disponibilidad de Fuentes de Datos con Pentaho Data Integration
Realizado por: Guido López

- Seguridad de acceso a datos.- En este parámetro se va a medir la seguridad con la que se realiza la conexión hacia las fuentes de datos que se van a utilizar, para el trabajo de investigación.

- Talend Open Studio.- En este indicador esta herramienta posee un valor de 10 debido a que facilita la conexión hacia las fuentes de datos Big Data; cómo podemos observar en la imagen siguiente.

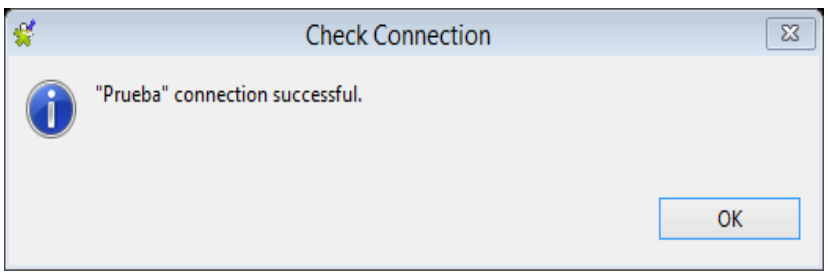


Figura 3-3 Aseguramiento de éxito en la conexión hacia las fuentes de datos con Talend Open Studio
Realizado por: Guido López

- **Pentaho Data Integration.-** Esta herramienta posee un valor de 10 en este indicador, debido a que posee una muy buena conexión hacia las fuentes de datos Big Data; para los trabajos necesarios.

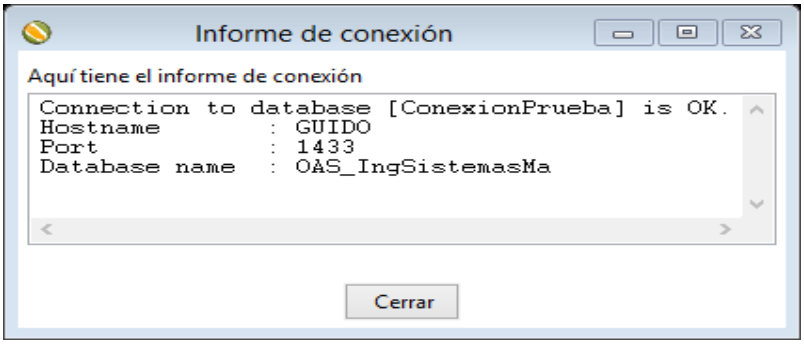


Figura 4-3 Aseguramiento de éxito en la conexión hacia las fuentes de datos con Pentaho Data Integration
Realizado por: Guido López

- Control de errores.- En este indicador se mide la forma como controla los errores que se producen durante la conexión o realización de los trabajos de integración de los datos.
- Talend Open Studio**.-** Esta herramienta gestiona de manera gráfica y muy versátil los errores que pueden surgir durante el proceso de integración de los datos, por lo que posee un valor de 5.

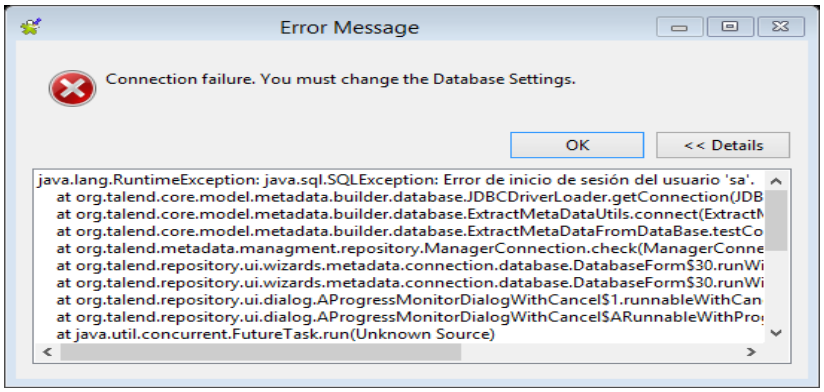


Figura 5-3 Gestión de Errores con Talend Open Studio
Realizado por: Guido López

- Pentaho Data Integration**.-** Esta herramienta posee un valor de 5 en este indicador, debido a que posee una muy buena gestión de los errores que se produjesen, debido a que nos muestra los errores y su correspondiente explicación.

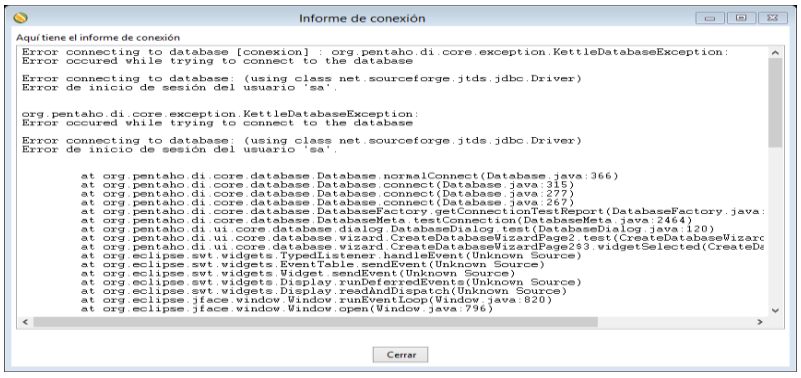


Figura 6-3 Gestión de Errores con Pentaho Data Integrator
Realizado por: Guido López

3.1.2 *Compatibilidad*.

- Compatibilidad de tipos de datos.- Se analiza la compatibilidad existente entre los tipos de datos desde la fuente hacia el destino del proceso de integración.

- Talend Open Studio.- Esta herramienta posee la mayoría de tipos de datos compatibles, para ejecutar trabajos de integración y evitar los errores por incompatibilidad; poseyendo un valor de 10 en este indicador.

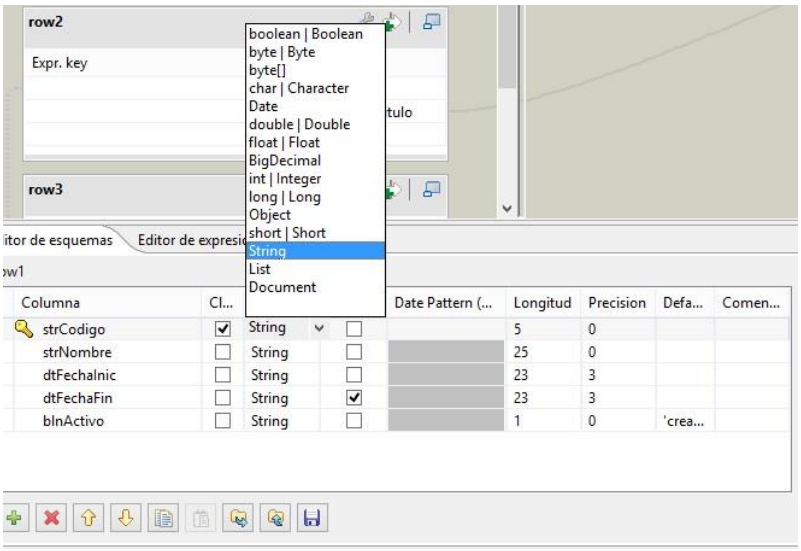


Figura 7-3 Tipos de datos compatibles con Talend Open Studio
Realizado por: Guido López

- Pentaho Data Integration**.-** Esta herramienta tiene un alto grado de compatibilidad en los tipos de datos a usar en los trabajos de integración; por lo que posee un valor de 10 en este indicador.

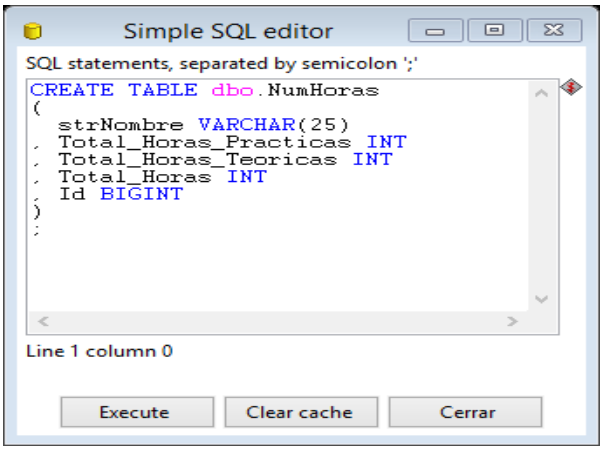


Figura 8-3 Tipos de datos compatibles con Pentaho Data Integrator
Realizado por: Guido López

- Soporte de tipos de datos.- Se verifica los tipos de datos soportados por las herramientas de integración, y administrar los procesos de integración de datos.

- Talend Open Studio.- Esta herramienta soporta todos los tipos de datos utilizados en los procesos de integración desde las fuentes hacia el destino; poseyendo un valor de 10 en este indicador.

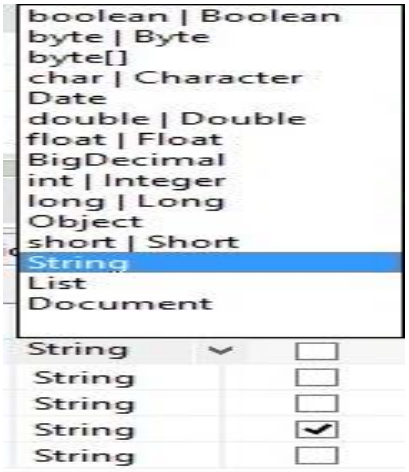


Figura 9-3 Tipos de datos soportados por Talend Open Studio
Realizado por: Guido López

- Pentaho Data Integration.- Esta herramienta tiene un alto grado de compatibilidad en los tipos de datos a usar en los trabajos de integración; por lo que posee un valor de 10 en este indicador.

3.1.3 *Funcionalidad*

- Carga de datos desde las fuentes Big Data.- Este indicador se puntuará en base a los tiempos de carga de los datos del correspondiente indicador a desarrollar.

Una vez analizados los tiempos de carga de cada uno de los indicadores propuesto se determinará el valor de este atributo para determinar cuál de las herramientas tiene un valor más alto.

Solo se van a indicar los tiempos de ejecución de los Jobs para la integración de los datos, la parte demostrativa se encuentran en los anexos correspondientes.

Se realizan los indicadores correspondientes para la escuela de Ingeniería en Sistemas Informáticos; para las demás escuelas se realizará en la parte aplicativa.

Tabla 1-3 Tiempos de carga de los datos, para los indicadores

Indicador	**Tiempo TOS**	**Tiempo PDI**
Malla curricular (por pensum de estudios):		
Títulos que se ofertan	0.41 s Ver Anexo I	0.1 s Ver Anexo I
Número de asignaturas por pensum de estudios	0.48 s Ver Anexo J	0.1 s Ver Anexo J
Número de asignaturas por tipo por pensum de estudios.	0.53 s Ver Anexo K	0.1 s Ver Anexo K
Número de horas, teóricas, prácticas, en total	0.45 s Ver Anexo L	0.1 s Ver Anexo L
Número de créditos por pensum de estudios.	0.56 s Ver Anexo M	0.4 s Ver Anexo M
Total Tiempo de Carga	**2.43 s**	**0.8 s**
Información del personal académico agrupado por período académico:		
Titularidad: N°. docentes con nombramiento por periodo	1.05 s Ver Anexo N	0.1 s 0.2 Ver Anexo N
Titularidad: N°. docentes con contrato por periodo	0.76 s Ver Anexo O	0.1 s Ver Anexo O
Titularidad: N°. docentes empleados por periodo	0.72 s Ver Anexo P	0.1 s Ver Anexo P
Dedicación N°. Docentes de tiempo completo por periodo.	1.99 s Ver Anexo Q	0.1 s Ver Anexo Q
Dedicación N°. Docentes de medio tiempo por periodo.	1.41 s Ver Anexo R	0.1 s Ver Anexo R
Dedicación N°. Docentes de tiempo parcial por periodo.	0.86 s Ver Anexo S	0.5 s Ver Anexo S
Categoría N°. Docentes principales por periodo.	1.52 s Ver Anexo T	0.4 s Ver Anexo T
Categoría N°. Docentes auxiliares por periodo.	1.69 s Ver Anexo U	0.3 s Ver Anexo U
Categoría N°. Docentes agregados por periodo.	1.76 Ver Anexo V	0.1 s Ver Anexo V
Número de horas clases presenciales del profesor en la carrera.	1.14 s Ver Anexo W	0.5 s Ver Anexo W
Total Información docentes	**12.9 s**	**2.30 s**
Información de los estudiantes (por período académico):		
Estudiantes y su zona de Procedencia.	17.15 s Ver Anexo X	5.7 s Ver Anexo X
Estudiantes y su colegio de donde proviene.	4.03 s Ver Anexo Y	7.03 s Ver Anexo Y
Numero de materias matriculadas con primera matricula	4.72 s Ver Anexo Z	3.7 s Ver Anexo Z

Numero de materias matriculadas con segunda matricula	4.93 s Ver Anexo A1	1 s Ver Anexo A1
Numero de materias matriculadas con tercera matricula	3.89 s Ver Anexo B1	0.5 s Ver Anexo B1
Total Información estudiantes	**34.72 s**	**17,93 s**
Información General de los estudiantes (por periodos académicos)		
Número de estudiantes matriculados por nivel.	1.58 s Ver Anexo C1	0.1 s Ver Anexo C1
Número de estudiantes de sexo masculino por periodo	1.56 s Ver Anexo D1	0.4 s Ver Anexo D1
Número de estudiantes de sexo femenino por periodo	1.97 s Ver Anexo E1	0.1 s Ver Anexo E1
Número de estudiantes agrupados por ciudad de procedencia.	4.42 s Ver Anexo F1	0.4 s Ver Anexo F1
Número de estudiantes agrupados por colegio de procedencia.	1.79 s Ver Anexo G1	0.4 s Ver Anexo G1
Número de estudiantes que tienen primera matricula en las materias por periodos.	4.08 s Ver Anexo H1	0.4 s Ver Anexo H1
Número de estudiantes que tienen segunda matricula en las materias por periodo.	2.53 s Ver Anexo I1	0.1 s Ver Anexo I1
Número de estudiantes que tienen tercera matricula en las materias por periodo	2.74 s Ver Anexo J1	0.1 s Ver Anexo J1
Total Información General Estudiantes	**20.67 s**	**2 s**

Fuente: Tipos de datos compatibles con Talend Open Studio
Realizado por: Guido López

- Soporte de sentencias SQL.- Se verifica el soporte a sentencias de lenguaje SQL, para la realización de cálculos y tareas de agrupación y ordenamiento de los datos.

- Talend Open Studio.- Esta herramienta posee un entorno mayormente grafico para el análisis de los datos de entrada para obtener un resultado, por lo que se asignará un valor de 3 en este indicador.

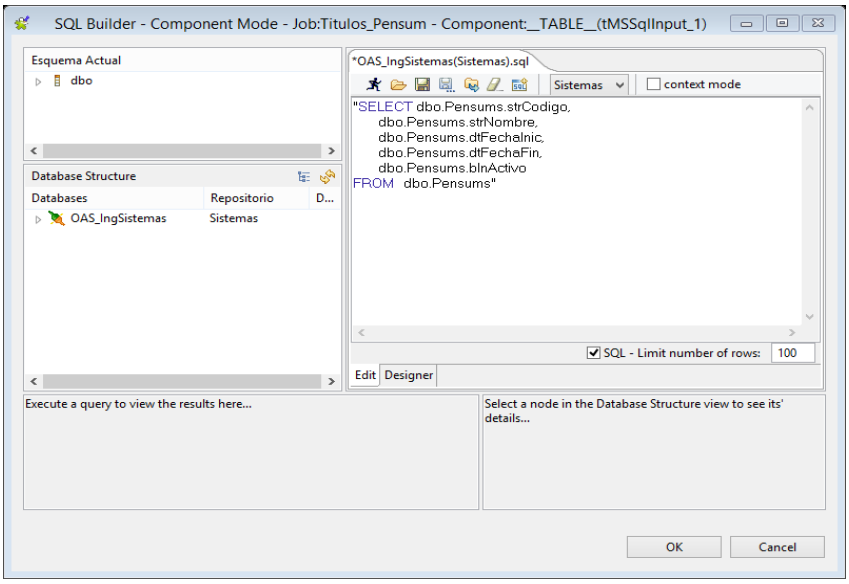


Figura 10-3 Ejecución de sentencias SQL en Talend Open Studio
Realizado por: Guido López

- Pentaho Data Integration.- Esta herramienta basa sus trabajos de integración de datos en sentencias de lenguaje SQL, por lo que posee un indicador alto, con un valor de 5.

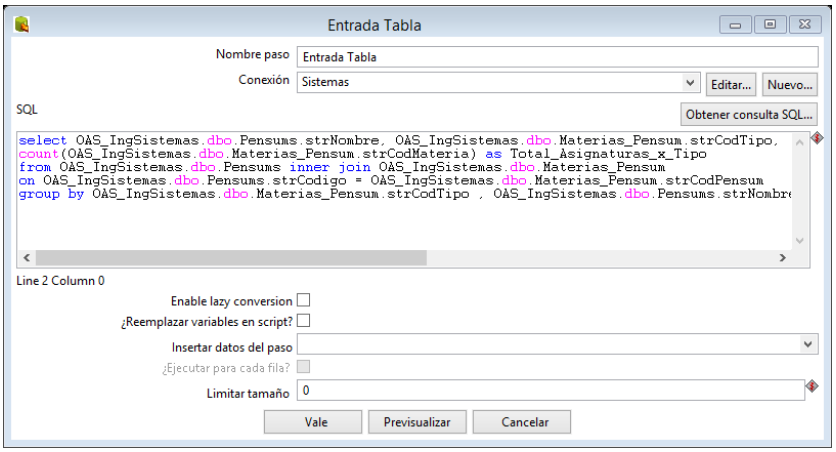


Figura 11-3 Ejecución de sentencias SQL en Talend Open Studio
Realizado por: Guido López

- Manejo de integración de datos.- Este indicador hace referencia al manejo de claves primarias y otros elementos que aseguran la integridad de los datos de salida del proceso de integración de los datos.
- Talend Open Studio.- Esta herramienta permite el control de claves primarias con cierta facilidad por lo que permite mucha integridad en los datos de salida del proceso de integración; por lo que posee un valor de 5.

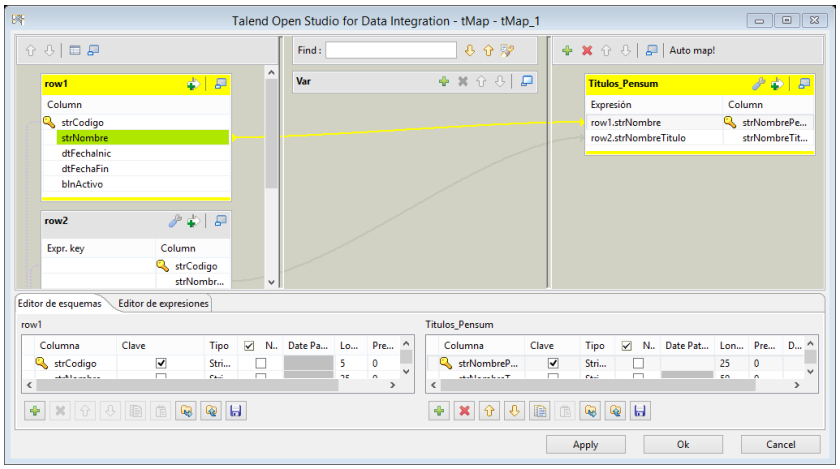


Figura 12-3 Generación de claves primarias para la salida de datos TOS
Realizado por: Guido López

- Pentaho Data Integration.- Esta herramienta tiene un grado elevado de control en base a claves autogeneradas, para garantizar el grado de integridad de los datos de las fuentes así como los destinos; teniendo un valor de 5 en este indicador.

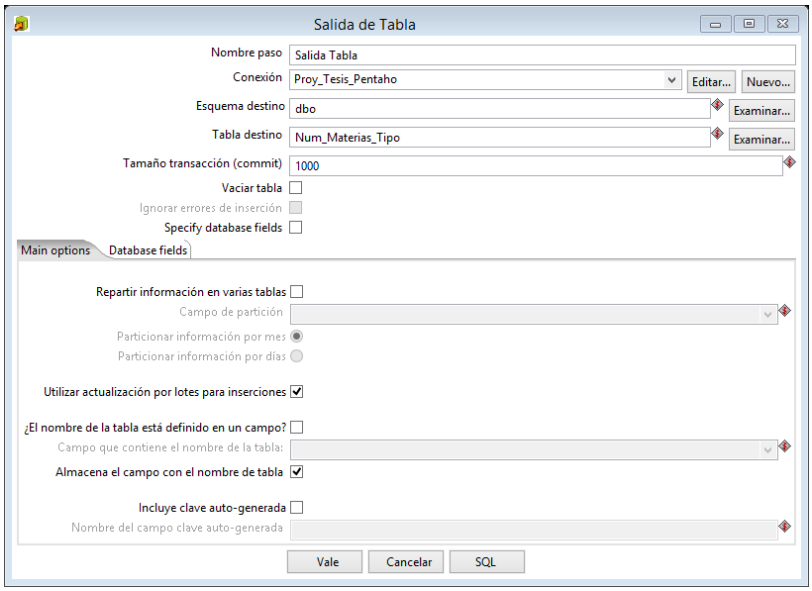


Figura 13-3 Generación de claves primarias para la salida de datos PDI
Realizado por: Guido López

- Soporte Técnico.- Se analiza las formas de acceder al soporte de los errores que se producen durante los procesos de integración, siendo este de manera gráfica o de forma de comandos.

- Talend Open Studio.- Esta herramienta proporciona el soporte de Manera gráfica y nos proporciona la información necesaria del error y así poder analizar y determinar la solución al problema que se presenta; teniendo un valor de 10 en este indicador.

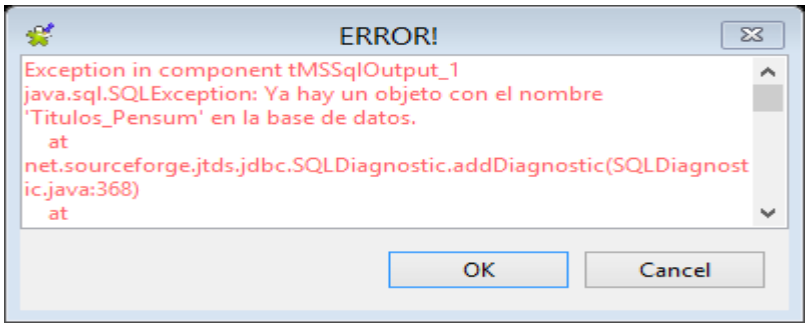


Figura 14-3 Soporte ante los errores surgidos Talend Open Studio
Realizado por: Guido López

- Pentaho Data Integration.- Esta herramienta posee una descripción gráfica y con líneas de código para determinar el problema y dar soporte necesario para dar solución a un problema; poseyendo un valor de 10 en este indicador.

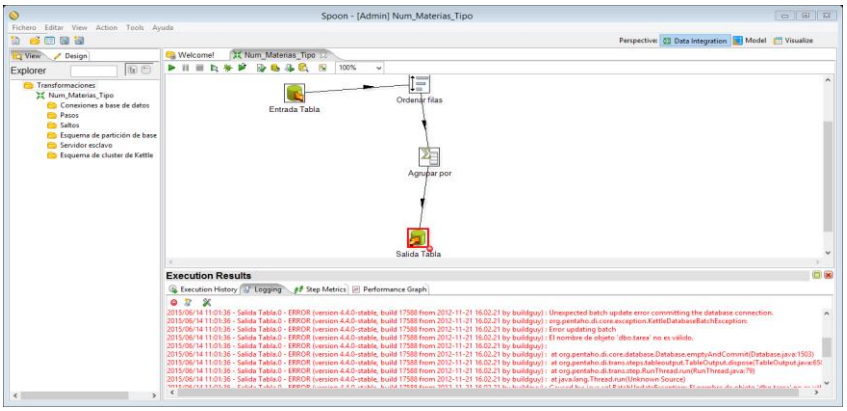


Figura 15-3 Soporte por errores surgidos Pentaho Data Integrator
Realizado por: Guido López

3.1.4 *Interfaz*

- Interfaz en el proceso de integración de datos.- Se analiza la forma de realización de los trabajos de integración, siendo estos de manera gráfica o mediante sentencias de lenguaje SQL en caso de poseer las dos formas tendrán el puntaje máximo.

- Talend Open Studio**.-** Esta herramienta posee las dos maneras de realización de los trabajos de integración, siendo la manera gráfica la más predominante y la parte de comandos en menor importancia, dándole un valor de 8 en este indicador.

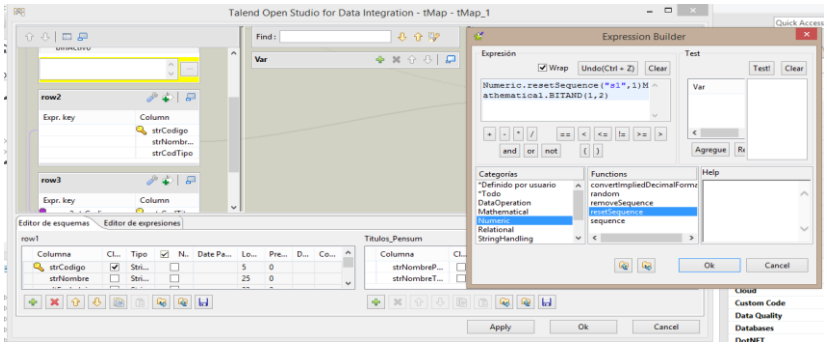


Figura 16-3 Interfaz gráfica de Talend Open Studio
Realizado por: Guido López

- Pentaho Data Integration.- Esta herramienta posee un valor de 7 en este indicador, debido a que posee las dos formas de realización de los trabajos de integración de datos, siendo la parte de sentencias SQL la más predominante y la parte gráfica en menor proporción.

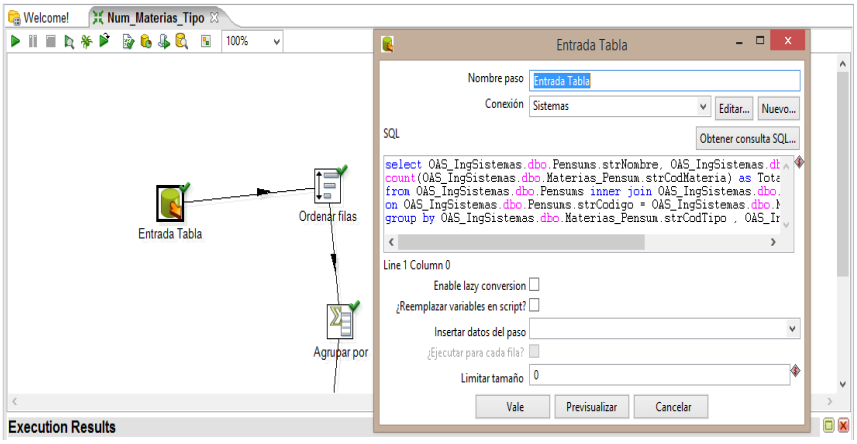


Figura 17-3 Interfaz gráfica de Pentaho Data Integrator
Realizado por: Guido López

3.2 Resultados totales

Para la determinación de los resultados parciales, tomando en cuenta los atributos seleccionados anteriormente para determinar el rendimiento en el proceso de integración de los datos, con la factibilidad de fuentes de datos Big Data.

Para determinar el desempeño de las herramientas seleccionadas se enfocará en el atributo de tiempo de carga de los datos, resolviendo los indicadores propuestos.

Para determinar el valor del parámetro de carga de datos desde las fuentes Big Data; se desarrollará en base a una tabla donde se seleccionará la opción que menor tiempo toma en realizar las cargas de los datos y los procesos de integración.

Para dar un valor al parámetro de la carga de datos, se tiene que evaluar el resultado obtenido en el proceso de carga de datos de los indicadores propuestos anteriormente.

Tabla para establecer el valor del parámetro de carga de datos desde las fuentes Big Data, hacia el destino.

Tabla 2-3 Rango de Tiempos y la asignación de valores

Atributo	Rango (s)	Valor
Tiempo de carga de datos	0.1 – 0.5	30
	0.5 – 1	28
	1 – 1.5	26
	1.5 – 2	24
	2 – 2.5	22
	2.5 – 3	20
	3 – 3.5	18
	3.5 – 4	16
	4 – 4.5	14
	4.5 – 5	12
	5 en adelante	10

Fuente: Autor
Realizado por: Guido López

Luego se determina el valor promedio del tiempo de carga de datos, para los indicadores; para determinar el valor del parámetro que se va a usar para evaluar cuál de las dos herramientas ofrece un mejor rendimiento en los procesos integración.

Tabla 3-3 Tiempos de carga de los datos con las dos herramientas

Parámetro	TOS	PDI
Tiempo de carga de datos para los indicadores.	0,41	0,1
	0,48	0,1
	0,53	0,1
	0,45	0,1
	0,56	0,4
	1,05	0,1
	0,76	0,1
	0,72	0,1
	1,99	0,1
	1,41	0,1
	0,86	0,5
	1,52	0,4
	1,69	0,3
	1,76	0,1
	1,14	0,5
	17,15	5,7
	4,03	7,03
	4,72	3,7
	4,93	1
	3,89	0,5
	1,58	0,1
	1,56	0,4
	1,97	0,1
	4,42	0,4
	1,79	0,4
	4,08	0,4
	2,53	0,1
	2,74	0,1
Promedio	2,53	0,85

Fuente: Autor
Realizado por: Guido López

Una vez determinado el tiempo promedio de ejecución de los indicadores necesarios, se establece el peso y el valor correspondiente para analizar los tiempos.

Tabla 4-3 Asignación de los pesos de cada uno de los parámetros seleccionados

PARÁMETROS	SUB PARÁMETROS	TOS	PDI
Conectividad	Soporte de conexión a las fuentes Big Data.	5	5
	Seguridad de acceso a datos.	10	10
	Control de errores	5	5
TOTAL CONECTIVIDAD		20	20
Compatibilidad	Compatibilidad de tipos de datos	10	10
	Soporte de tipos de datos.	10	10
TOTAL COMPATIBILIDAD		20	20
Funcionalidad	Carga de datos desde las fuentes Big Data.	20	28
	Soporte de sentencias SQL	3	5
	Manejo de Integración de datos.	5	5
	Soporte Técnico	10	10
TOTAL FUNCIONALIDAD		38	48
Interfaz	Interfaz en el proceso de integración de datos.	8	7
TOTAL INTERFAZ		8	7
TOTAL		86 / 100	95/100

Fuente: Autor
Realizado por: Guido López

3.3 Interpretación de los resultados

Se analizará los resultados en base a lo analizado en el capítulo anterior, para determinar de manera visual como es el comportamiento de las herramientas de integración de datos seleccionadas.

Tabla 5-3 Tabla de comparación de Parámetros entre las dos herramientas de integración de datos.

Parámetros de comparación	TOS	PDI
Soporte de conexión de fuentes Big Data	5	5
Seguridad de acceso a datos	10	10
Control de errores	5	5
Compatibilidad de tipos de datos	10	10
Soporte de tipos de datos	10	10
Carga de datos desde las fuentes Big Data	20	28
Soporte de sentencias SQL	3	5
Manejo de integración de datos	5	5
Soporte Técnico	10	10
Interfaz en el proceso de integración de datos	8	7
Total / 100	86	95

Fuente: Autor
Realizado por: Guido López

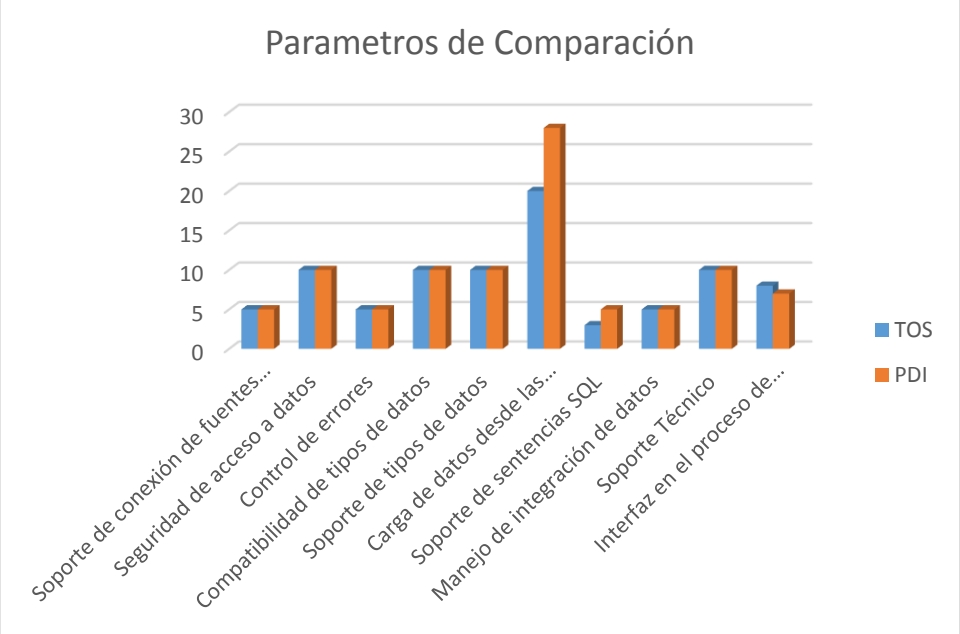


Figura 18-3 Cuadro estadístico de los valores de los parámetros de comparación
Realizado por: Guido López

Según el gráfico estadístico expuesto podemos decir que lo herramienta de integración Pentaho Data Integrator (PDI) posee valores más altos en los parámetros de comparación expuestos para determinar el rendimiento en los procesos de integración de datos.

3.4 Comprobación de la hipótesis

Para la comprobación de la hipótesis, se basará en el atributo **Carga de Datos desde las fuentes,** para lo cual se realizará mediante procesos estadísticos conocidos como **la Media, y la Varianza;** para determinar cuál de las herramientas, ofrece un mejor rendimiento en los procesos de integración de datos con fuentes Big Data.

Tabla 6-3 Tabla de tiempos de carga de los indicadores entre las dos herramientas de integración de datos.

Parámetro	TOS	PDI
Tiempo de carga de datos para los indicadores.	0,41	0,1
	0,48	0,1
	0,53	0,1
	0,45	0,1
	0,56	0,4
	1,05	0,1
	0,76	0,1
	0,72	0,1
	1,99	0,1
	1,41	0,1
	0,86	0,5
	1,52	0,4
	1,69	0,3
	1,76	0,1
	1,14	0,5
	17,15	5,7
	4,03	7,03
	4,72	3,7
	4,93	1
	3,89	0,5
	1,58	0,1
	1,56	0,4
	1,97	0,1
	4,42	0,4
	1,79	0,4
	4,08	0,4
	2,53	0,1
	2,74	0,1

Fuente: Autor
Realizado por: Guido López

Fórmula correspondiente para hallar la media aritmética y la desviación estándar de los datos obtenidos de los procesos de integración de los datos para la obtención de los indicadores académicos propuestos.

Calculo de la media aritmética de la herramienta Talend Open Studio.

$$\bar{X} = \frac{x1 + x2 + x3 + \cdots \ldots\ldots\ldots\ldots\ldots\ldots + xn}{n}$$

$$\bar{X} = \frac{70{,}72}{28}$$

$$\overline{X1} = 2{,}53$$

Calculo de la media aritmética de la herramienta Pentaho Data Integrator.

$$\bar{X} = \frac{x1 + x2 + x3 + \cdots \ldots\ldots\ldots\ldots\ldots\ldots + xn}{n}$$

$$\bar{X} = \frac{22{,}13}{28}$$

$$\overline{X2} = 0{,}85$$

Calculo de la varianza de la herramienta Talend Open Studio

$$V = \sum \frac{(xi - \bar{x})^2}{n}$$

$$V = \frac{275{,}2126}{27}$$

$$V1 = 10{,}1931$$

Calculo de la desviación estándar

$$\sigma = \sqrt{V}$$

$$\sigma = \sqrt{10}, 1931$$

$$\sigma 1 = 3{,}19$$

Calculo de la varianza de la herramienta Pentaho Data Integrator

$$V = \sum \frac{(xi - \bar{x})^2}{n}$$

$$V = \frac{79{,}5989}{27}$$

$$V2 = 2{,}9481$$

Calculo de la desviación estándar

$$\sigma = \sqrt{V}$$

$$\sigma = \sqrt{2}, 9481$$

$$\sigma 2 = 1{,}7170$$

Teniendo en cuenta los valores obtenidos mediante las fórmulas de la media aritmética y la varianza resultantes podemos mencionar que la herramienta de integración de datos **Pentaho**

Data Integrator posee un mejor rendimiento en cuanto a la realización de procesos de integración de datos.

Tabla 7-3 Valores de la Media Aritmética y la Varianza

	TOS	PDI
Media Aritmética	2,53	0,82
Varianza	10,1931	2,9481

Fuente: Autor
Realizado por: Guido López

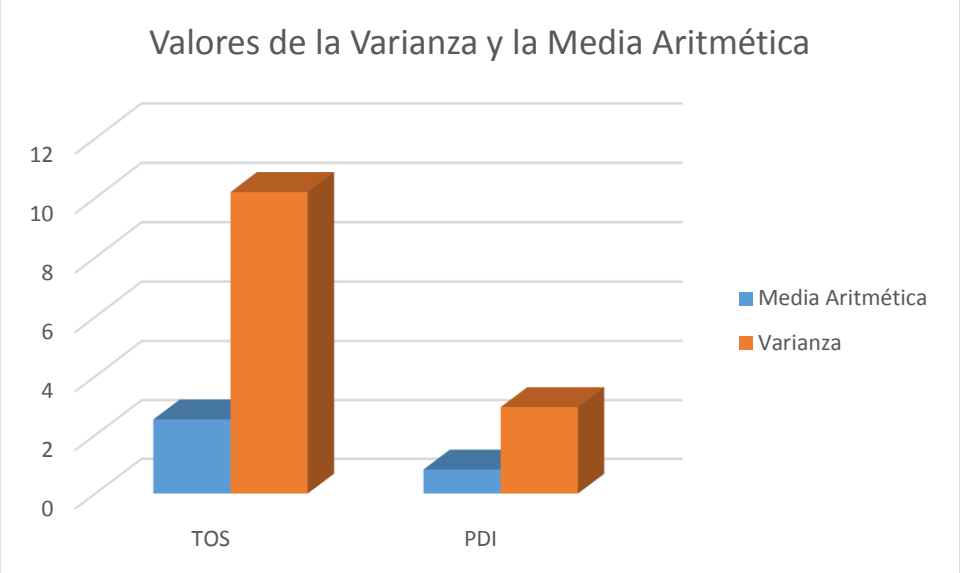


Figura 19-3 Cuadro estadístico de los valores de la Media Aritmética y Varianza
Realizado por: Guido López

Una vez determinados los valores de la Media y Varianza respectivamente; se procede a la comprobación de la Hipótesis, utilizando la técnica de t-student para dos muestras de datos.

3.4.1 Técnica t-student

Cuando se tiene una muestra de datos menor de 30, se hace indispensable la utilización de esta técnica.

Esta Técnica se posee varios usos entre los cuales citamos:

- Para determinar el intervalo de confianza dentro del cual se puede estimar la media de una población a partir de muestras pequeña (n < 30).
- Para probar hipótesis cuando una investigación se basa en muestreo pequeño.
- Para probar si dos muestras provienen de una misma población.

Para utilizar esta técnica estadística, se debe plantear las hipótesis **Ho** y **H1**; para esto se plantea de la siguiente manera:

Ho= $\overline{X1} = \overline{X2}$

Ho= La herramienta de integración de datos Pentaho Data Integrator posee igual rendimiento en ambientes cuyas fuentes de datos sean Big Data.

H1= $\overline{X2} < \overline{X1}$

H1= La herramienta de integración de datos Pentaho Data Integrator posee un mejor rendimiento en ambientes cuyas fuentes de datos sean Big Data.

Una vez determinadas las hipótesis Nula e Hipótesis Alternativa del trabajo de investigación se procede a determinar el valor de significancia conocido como **α**, para trabajos de investigación este valor se asigna el 5%, por lo que **α=0.05.**

Una vez determinado el valor de significancia; se procede a la obtención de los valores de la Media y de la Desviación Estándar a partir de la muestra de los datos; estos valores ya tenemos anteriormente.

$\overline{X1} = 2{,}53$

$\overline{X2} = 0{,}85$

$V1 = 10{,}1931$

$V2 = 2{,}9481$

Luego se aplica la fórmula de la distribución estadística t-student para dos grupos de datos:

$$t = \frac{\overline{X1} - \overline{X2}}{\sqrt{\frac{V1}{n} + \frac{V2}{n}}}$$

$$t = \frac{2{,}53 - 0{,}85}{\sqrt{\frac{10{,}1931}{28} + \frac{2{,}9481}{28}}}$$

$$t = \frac{1{,}68}{\sqrt{0{,}3640 + 0{,}1053}}$$

$$t = \frac{1{,}68}{\sqrt{0{,}3640 + 0{,}1053}}$$

$$t = \frac{1{,}68}{\sqrt{0{,}3640 + 0{,}1053}}$$

$$t = \frac{1{,}68}{0{,}6851}$$

$$t = 2{,}4522$$

Seguidamente se procede a calcular los grados de libertad correspondiente a en trabajo de investigación propuesto, para lo cual se lo realiza de la siguiente manera.

$$gl = \frac{\left(\frac{V1}{n} + \frac{V2}{n}\right)^2}{\frac{\left(\frac{V1}{n-1}\right)^2}{n} + \frac{\left(\frac{V2}{n-1}\right)^2}{n}}$$

$$gl = \frac{\left(\frac{10,1931}{28} + \frac{2,9481}{28}\right)^2}{\frac{\left(\frac{10,1931}{27}\right)^2}{28} + \frac{\left(\frac{2,9481}{27}\right)^2}{28}}$$

$$gl = \frac{(0,4693)^2}{0,005514}$$

$$gl = \frac{0,2203}{0,005516}$$

$$gl = 39,94$$

Finalmente se obtiene el valor correspondiente en la tabla de t-student con los grados de libertad y el valor de t obtenidos anteriormente. Basándose en la tabla de t-student se determina que el valor correspondiente a t de la tabla.

Luego se realiza la correspondiente comparación ente el valor de la función de la distribución t-student encontrado con el valor de t correspondiente a la tabla.

T-enc=2.4532

t-tabla=1.6839

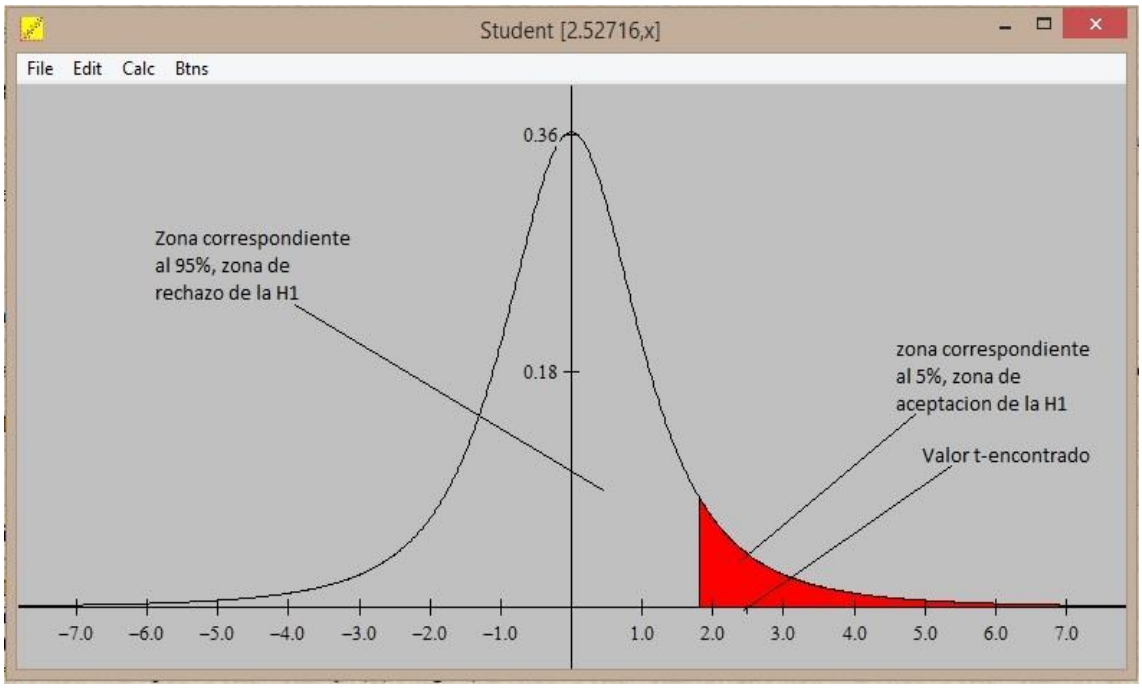


Figura 20-3 Figura de la distribución t-student correspondiente.
Realizado por: Guido López

De acuerdo a los resultados obtenidos y debido a que tabla<tencontrado; entonces se acepta la Hipótesis Alternativa o H1.

H1= La herramienta de integración de datos Pentaho Data Integrator posee un mejor rendimiento en ambientes cuyas fuentes de datos sean Big Data.

3.5 Propuesta a realizar

Para la determinación de la Propuesta, se basara en el objetivo específico; que se mencionó anteriormente, que dice: Desarrollar un prototipo para la construcción de un observatorio de indicadores en la Facultad de Informática y Electrónica, basado en tecnología de integración de datos con la factibilidad de fuentes Big Data; para lo cual se utilizan las Bases de Datos de la Facultad de Informática y Electrónica de la Escuela Superior Politécnica de Chimborazo; correspondientes a las escuelas de: Ingeniería en Sistemas, Ingeniería Electrónica, Ingeniería Electrónica en Control y Redes Industriales, Ingeniería Electrónica en Redes y Telecomunicaciones, Ingeniería en Diseña Grafico.

3.6 Desarrollo de la propuesta

Para el desarrollo de la propuesta se requiere la instalación de las siguientes herramientas y desarrollo del Prototipo para analizar los indicadores mencionados:
Se realiza la instalación de la herramienta Microsoft SQL Server 2012, la utilización de la herramienta que resulto la mejor en el rendimiento con fuentes Big Data.
También se selecciona una herramienta para la visualización de datos.

- Se instala el software de Manejo de Base de Datos Microsoft SQL Server, se utiliza la herramienta de Integración Pentaho Data Integrator y la herramienta de visualización de datos Tableau.
- Se realiza los procesos de carga de los datos correspondientes a cada uno de los indicadores académicos mencionados anteriormente para cada una de las Escuelas de la Facultad de Informática y Electrónica.
- Luego se carga los datos de los indicadores en la herramienta de visualización de datos, en la misma que se puede realizar análisis de datos como: Obtener promedios, sacar totales y comparar entre las Escuelas de la Facultad y Periodos de la Escuela y otros análisis de los datos.
- Se muestra a continuación las tablas relacionadas a cada uno de los indicadores propuestos anteriormente.

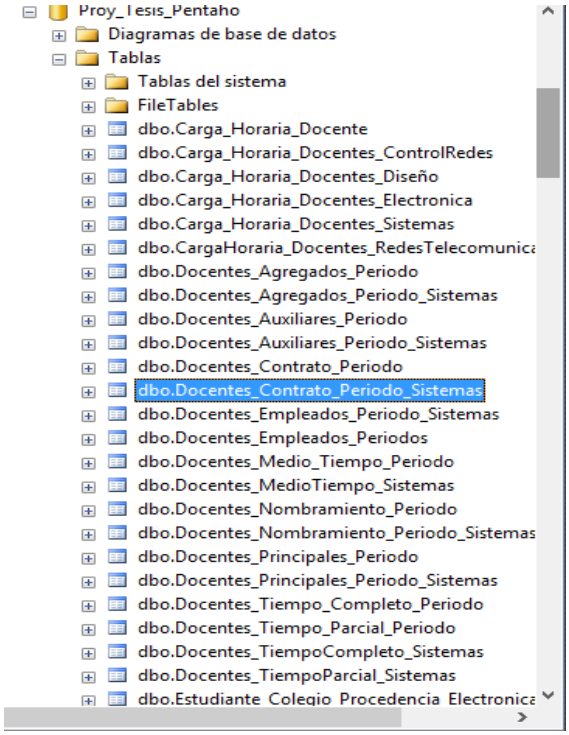


Figura 21-3 Figura de las tablas correspondientes a los indicadores.
Realizado por: Guido López

- Luego utilizamos la herramienta de integración **Pentaho Data Integrator**, para desarrollar los procesos de integración de los datos correspondientes a los indicadores educativos, los mismos que se almacenaran en la herramienta de Gestión de Base de Datos **Microsoft SQL Server.**

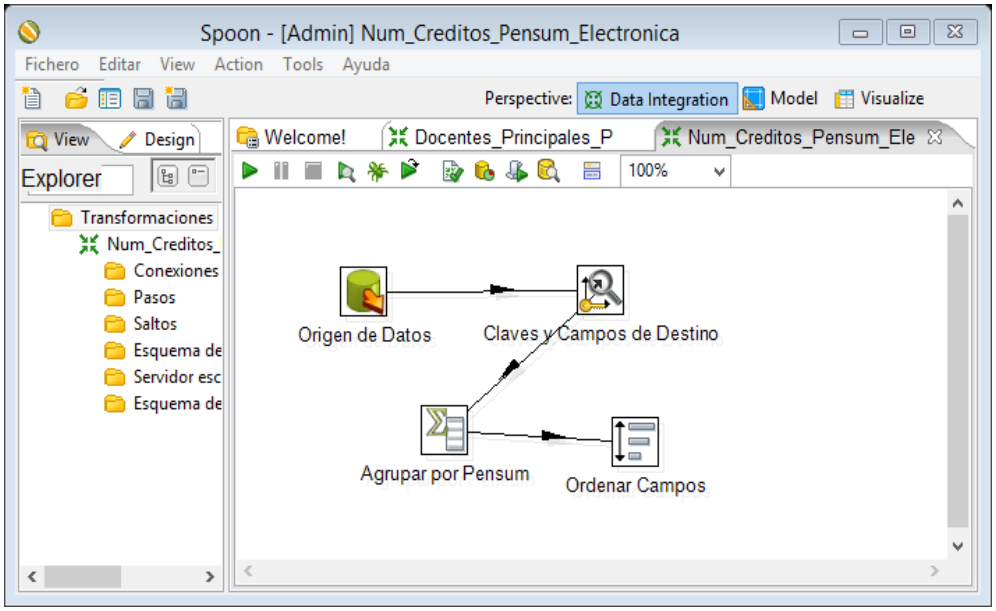


Figura 22-3 Figura del proceso de integración de datos con Pentaho Data Integrator
Realizado por: Guido López

- Finalmente se desarrolla el prototipo del observatorio de indicadores educativos para las escuelas de la Facultad de Informática y Electrónica.

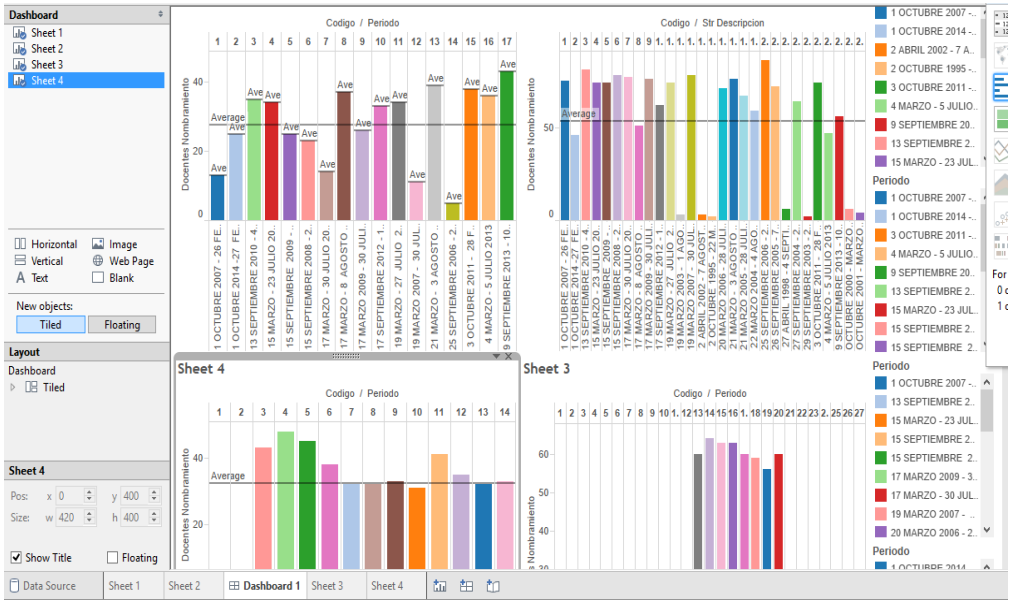


Figura 23-3 Observatorio de indicadores de cada una de las escuelas.
Realizado por: Guido López

Este observatorio consta con los indicadores necesarios que se analizaron anteriormente, para determinar varios factores que afectan al correcto funcionamiento de las escuelas de la Facultad de Informática y Electrónica.

CONCLUSIONES

- El mundo del Big Data se desarrolla muy rápidamente por lo que se debe estar en constante actualización sobre los métodos y herramientas para la integración de los datos, para obtener una información veraz y oportuna para de esta manera tomar decisiones muy acertadas.

- Los criterios de evaluación del rendimiento en la integración de datos, se seleccionan en base a criterios específicos como: Conectividad, Compatibilidad, Funcionalidad e Interfaz.

- El prototipo con la herramienta de integración Pentaho Data Integrator, se desempeña con un mejor rendimiento en la ejecución de los procesos de integración, que el prototipo con la herramienta Talend Open Studio.

- El observatorio de indicadores es muy importante para poder analizar cómo se relacionan estos indicadores de cada una de las escuelas, determinando promedio y totales.

RECOMENDACIONES

- Obtener fuentes confiables de información, para poder obtener un resultado muy confiable en el trabajo de investigación que se está desarrollando.

- Seleccionar adecuadamente los parámetros o criterios de evaluación del rendimiento, dentro del campo de integración de datos; para poder obtener valores adecuados para poder realizar comparaciones que satisfagan una evaluación eficaz.

- Obtener suficientes fuentes sobre las herramientas de integración más posicionadas en el mercado; poder tomar una decisión adecuada sobre cual herramienta se debe escoger, para la realización de los prototipos de estudio.

- Desarrollar un prototipo con el mayor número de indicadores, necesarios para obtener información necesaria; para poder tomar decisiones acertadas para el mejor funcionamiento de la organización e institución educativa.

- Hacer uso de observatorios de indicadores para el monitoreo de las escuelas de las facultades de la Escuela Superior Politécnica de Chimborazo.

GLOSARIO

DataWarehouse: proceso mediante el cual una organización o empresa particular almacena todos aquellos datos e información necesarios para el propio desempeño de la misma.

DashBoards: Es una representación gráfica de los principales indicadores (KPI) que intervienen en la consecución de los objetivos de negocio, y que está orientada a la toma de decisiones para optimizar la estrategia de la empresa.

Framework: es una estructura conceptual y tecnológica de soporte definido, normalmente con artefactos o módulos de software concretos, que puede servir de base para la organización y desarrollo de software.

Hadoop: Sistema de administración de grandes cantidades de información.

JDBC: es una API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java, independientemente del sistema operativo donde se ejecute.

Join: Función que permite unir dos o más tablas de una base de datos.

Media: Valor promedio de un conjunto de datos, obtenidos de una Fuente.

My SQL: Gestor de Base de Datos de software libre.

Open Source: Lenguaje de código abierto, manejable y actualizable por los usuarios que lo utilicen.

Prototipo: Primer ejemplar de alguna cosa que se toma como modelo para crear otros de la misma clase.

Shell script: Es un programa usualmente simple, que por lo regular se almacena en un archivo de texto plano.

Sistemas ERP: Sistemas de planificación de recursos empresariales.

Streaming: Se utiliza para optimizar la descarga y reproducción de archivos de audio y video que suelen tener un cierto peso.

Varianza: Función estadística que permite obtener valores para comprobar la Hipótesis

BIBLIOGRAFÍA

About MapR. [En línea]. San José, CA, About MapR Technologies, 2014, [Accedido: 02-sep-2014**].** Disponible en: **https://www.mapr.com/company.**

Apache Hive TM. [En línea]. 2011 – 2014 [Accedido: 17-sep-2014]**.** Disponible en: **http://hive.apache.org/.**

Articulo Big Data 0.0.pdf. [En línea]. España, [Accedido: 11-nov-2013] Disponible en: **http://www.cnis.es/images/informes/Articulo%20Big%20Data%200.0.pdf .**

AVILÉS, Marco; *Talend Open Studio: Introducción a ETL y posibilidades con Job Designs | El Mundo es Open Source.* [En línea]. Marco Avilés, 1 Noviembre 2011, Niveles de Talend, [Accedido: 30-mar-2014]. Disponible en**:** **http://blogs.antartec.com/opensource/2011/11/primeros-pasos-en-talend/.**

AVILÉS Marco*; Talend Open Studio: Introducción a ETL y posibilidades con Job Designs | El Mundo es Open Source.* [En línea]. Marco Avilés, 1 Noviembre 2011, Job Desing, [Accedido: 31-mar-2014]. Disponible en**:** **http://blogs.antartec.com/opensource/2011/11/primeros-pasos-en-talend/.**

BETANCUR CALDERÓN, Daniel. Modelo basado en agentes para las etapas de recopilación e integración de datos en el proceso de kdd (Tesis de Maestría) [En línea], Universidad Nacional de Colombia – Medellín-Colombia, p: p 16-38**,** [Accedido: 24-ene-2014]. Disponible en: **http://www.bdigital.unal.edu.co/2034/1/10366054583.2010.pdf**

Breva Systems Administrator, *Big data y análisis de la información,* [Blog]. México**.** 19/10/2012, 2012. [Accedido: 21-mar-2014]. Disponible en: **http://www.breva.biz/big-data-y-analisis-de-la-informacion/#!prettyPhoto.**

Características de Pentaho | Dataprix TI. [En línea]. Pentaho Data Integration, [Accedido: 30-mar-2014]. Disponible en: **http://www.dataprix.com/723-caracter-sticas-pentaho.**

chukwa.apache.org/docs/r0.5.0/admin.html, [En línea]. 2013-06-30, 2013, Chukwa Administration Guide, [Accedido: 14-sep-2014]. Disponible en: **http://chukwa.apache.org/docs/r0.5.0/admin.html.**

Cuadrante Mágico de Gartner sobre integración de datos 2014 – Informática es uno de los líderes. [En línea]. 2014**,** Informatica se posiciona como líder 8 años consecutivos**,** [Accedido: 01-sep-2014]. Disponible en: **http://www.informatica.com/es/data-integration-magic-quadrant/.**

DEL PINO, Manuel. Cuando el Big Data y el Internet de las Cosas Colisionan**,** *SG Software Guru,* [En línea]. #46**,** 2014**,** p.p 53 [Accedido: 21-mar-2014]. Disponible en: http://sg.com.mx/revista/46/cuando-el-big-data-y-el-internet-las-cosas-colisionan#.VcI48_l_Oko

DOMÍNGUEZ, Enrique. HDFS: Hadoop Distributed File System : Enrique Domínguez. [Blog]. 25 Aug, 2007, 2007, [Accedido: 26-ago-2014]. Disponible en**:** **http://www.enriquedominguez.com/hdfs-hadoop-distributed-file-system/.**

EPSILONTEC*; Gestión de los grandes volúmenes de información: Big Data| EPSILONTEC* [En línea]. *EPSILONTEC***,** Big Data, [Accedido: 09-ene-2014]. Disponible en: http://www.epsilontec.com/gestion-de-los-grandes-volumenes-de-informacion-big-data/

Experfy Editor. Cloudera vs Hortonworks: Comparing Hadoop Distributions - Experfy Insights. [Blog]. September 5 2014, 2014, [Accedido: 20-sep-2014]. Disponible en: **http://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/.**

FARAH CALDERÓN, **Walter. "**Introducción al Big Data" [En línea]**.** 2013,p.p 1-2; [Accedido: 09-ene-2014]**.** Disponible en**:** **http://www.academia.edu/3011182/Introduccion_al_Big_Data.**

FERNÁNDEZ PEREA, Iván: Introducción a Big Data y Hadoop, 19/08/2013, 2013, [Accedido: 29-mar-2014]. Disponible en: http://java4developers.com/2013/introduccion-a-big-data-y-hadoop/

GRACIA, Luis Miguel: Sqoop: Integrando Hadoop con nuestra base de datos. [En línea]. Luis Miguel Gracia, 3 diciembre 2012, Sqoop: Integrando Hadoop [Accedido: 21-jul-2014]. Disponible en: **http://unpocodejava.wordpress.com/2012/12/03/sqoop-integrando-hadoop-con-nuestra-base-de-datos/.**

IGLESIAS Pablo, *Qué es y qué significa para el internauta Big Data? | PabloYglesias | internet + móvil + actualidad* [En línea]. Pablo Iglesias, 2012, Qué es y qué significa para el internauta

Big Data **[**Accedido: 09-ene-2014**].** Disponible en: **http://www.pabloyglesias.com/que-es-y-que-significa-para-el-internauta-big-data/.**

Integración de datos. [En línea]. ¿Cómo funciona la integración de datos? 2014, [Accedido: 08-ene-2014]. Disponible en: **http://www.ordenadores-y-portatiles.com/integracion-de-datos.html.**

Integración de Datos|LatinoBI, CA **[**En línea]. Caracas, Venezuela: Productos y Servicios, [Accedido: 09-mar-2014]. Disponible en: **http://www.latinobi-ven.com/productos-y-servicios/**

Latest Pentaho Data Integration (aka Kettle) Documentation, [En línea]. Pentaho Data Integration (Kettle) Tutorial [Accedido: 24-mar-2014]. Disponible en: **http://openbi.ning.com/group/pentahodataintegration?groupUrl=pentahodataintegration&xg_source=activity&id=2400100%3AGroup%3A8163&xg_pw=&page=9%3E**

LÁZARO, Miguel. Como usar Hadoop y sobrevivir a la experiencia. [En línea]. [Accedido: 26-ago-2014]. Disponible en: **http://www.tsc.uc3m.es/~miguel/MLG/adjuntos/Hadoop.pdf.**

LEO-REVILLA, Ángel, ¿Qué es Hadoop? [Blog]. May 16 2013, 2013, [Accedido: 11-jun-2014]. Disponible en: **http://momentotic.wordpress.com/2013/05/16/que-es-hadoop/.**

LURIE, Marty. Big data de código abierto para el impaciente, Parte 1: Tutorial Hadoop: Hello World con Java, Pig, Hive, Flume, Fuse, Oozie, y Sqoop con Informix, DB2, y MySQL. [Blog]. USA**,** 20-05-2013**,** 2013, [Accedido: 21-jul-2014]. Disponible en: **http://www.ibm.com/developerworks/ssa/data/library/techarticle/dm-1209hadoopbigdata/.**

MADRID, Víctor Javier. Talend Open Studio (TOS) 4.0**.** [Blog**].** España, Mayo 6 2010, 2010 [Accedido: 30-mar-2014]. Disponible en: **http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=TOS4.0.**

MADRID, Víctor Javier. Talend Open Studio (TOS) 4.0**.** [Blog**].** España, Mayo 6 2010, 2010 [Accedido: 20-mar-2014]. Disponible en: **http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=TOS4.0.**

MORALES Yessica, "Integración de Datos (ETL) y Almacenes de Datos" [En línea]. Maestría en Ingeniería de Software, 2012, p.p 2-20 [Accedido: 09-ene-2014]. Disponible en:

http://basesdatoscms.files.wordpress.com/2012/09/interacion-de-datos-y-almacenes-de-datos.pdf.

MORO, Esteban & **LUENGO-OROZ, Miguel** & **DE LA TORRE**, **Javier**; Big Data ¿En qué punto estamos? [En línea]**.** 6, 2013**,** Una definición; p.p 5-10; 12-15; [Accedido: 16-abr-2014]. Disponible en: http://www.centrodeinnovacionbbva.com/sites/default/files/bigdata_spanish.pdf

PATTERSON Clarke, *CDH.* [En línea]. Palo Alto: California, 2014, What's Inside? [Accedido: 19-sep-2014]. Disponible en: **http://www.cloudera.com/content/cloudera/en/products-and-services/cdh.html.**

*Pentaho BI Suite|Dataprix TI***,** [En línea]. Pentaho, [Accedido: 29-mar-2014]. Disponible en: **http://www.dataprix.com/empresa/productos/pentaho-bi-suite.**

PÉREZ Arian.; *Pentaho Data Integration - EcuRed.* **[**En línea]. Arian.Perez jc, 22 jun 2011, 2011, Requisitos mínimos para su funcionamiento [Accedido: 30-mar-2014]. Disponible en: **http://www.ecured.cu/index.php/Pentaho_Data_Integration.**

Press, Gil, *Top 10 Big Data Pure-Plays 2014.* [Blog]. [Accedido: 24-mar-2015]. Disponible en: **http://www.forbes.com/sites/gilpress/2014/02/11/top-10-big-data-pure-plays-2014/.**

Productos. [En línea]. Chile, Pentaho**,** [Accedido: 29-mar-2014]. Disponible en: **http://www.cognus.biz/productos/.**

PUENAYÁN CHAPI, Adriana Del Rocío, & AYNAGUANO SALGUERO, Diana Verónica Estudio Comparativo de ETLS Propietario vs Software Libre para la Implementación de una Solución Business Intelligence. (Tesis de Pregrado) [En linea]. Escuela Superior Politécnica de Chimborazo, Facultad de Informática y Electrónica, Escuela Ingeniería en Sistemas, Riobamba, Ecuador; 2012, p.p 24-26/ 52-62, Disponible en: http://dspace.espoch.edu.ec/bitstream/123456789/1519/1/18T00463.pdf

Que es Hadoop. [En línea]. Madrid, [Accedido: 12-jun-2014]. Disponible en: **http://www.pragsis.com/sites/default/files/pdf/Hadoop,%20MapReduce,%20Bidoop.pdf**

RIOS, Angel. "Oracle Data Integrator", [En línea]. U.S.A. 2009, ¿Que es Integración de Datos? [Accedido: 24-ene-2014]. Disponible en: **http://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/317498_esa.pdf.**

RINDLER Andreas. *BIG DATA DEFINITION - MIKE2.0, the open source methodology for Information Development* [En línea]. Big Data Definition [Accedido: 09-ene-2014]. Disponible en: **http://mike2.openmethodology.org/wiki/Big_Data_Definition.**

SALINAS, Alexandro, *Introducción a PENTAHO (parte 1 de 2)* [Blog]. 2008-03-12, 2008, [Accedido: 29-mar-2014]. Disponible en: **http://gravitar.biz/bi/introduccion-pentaho-parte-1/**

Scriptella 1.0 RC [En línea]. [Accedido: 20-mar-2014]. Disponible en: **http://mscerts.programming4.us/es/239746.aspx.**

Scriptella ETL Reference Documentation [En línea]. 2012 – 2014. [Accedido: 20-mar-2014]. Disponible en: **http://mscerts.programming4.us/es/239746.aspx.**

*Sqoop User Guide (v1.4.2)***,** [En línea]. Failed Exports [Accedido: 15-sep-2014**].** Disponible en**:** **http://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html#_introduction.**

Soluciones: Integración de datos para mejorar la eficiencia operativa | Pitney Bowes Software». [En línea]. Miami**:** USA**,** 2012, [Accedido: 08-ene-2014]. Disponible en: **http://latam.pbinsight.com/soluciones/por-oportunidades-de-negocio/mejorar-la-eficiencia-operativa/integracion-de-datos/.**

TILVES, Mónica; Hadoop: El gran aliado open source para afrontar el reto de 'Big Data', **[**En línea]. Mónica Tilves, 11 de mayo de 2012, [Accedido: 26-ago-2014]. Disponible **en: http://www.siliconweek.es/knowledge-center/hadoop-el-gran-aliado-open-source-para-afrontar-el-reto-de-big-data-22867.**

Welcome to Apache Pig! [En línea]. 06/06/2013, 2013 [Accedido: 17-sep-2014]. Disponible en**:** **http://pig.apache.org/.**

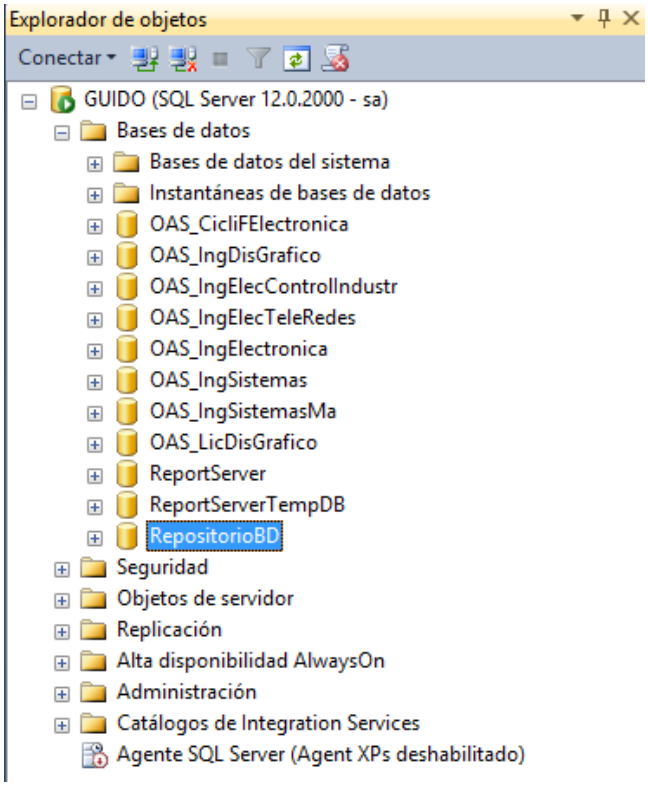
WPlook*; Introducción a Hadoop y su ecosistema***,** [En línea].Que es el Big Data, 2013**,**[Accedido: 18-sep-2014]. Disponible en: **http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/.**

1.1 Installation requirements - Talend Open Studio for Data Integration v5.4.2 - Installation and Upgrade Guide. [En línea]. Installation requirements, [Accedido: 20-mar-2014]. Disponible en: **https://help.talend.com/display/TalendOpenStudioforDataIntegrationInstallationandUpgradeGuide54EN/1.1+Installation+requirements.**

ANEXOS

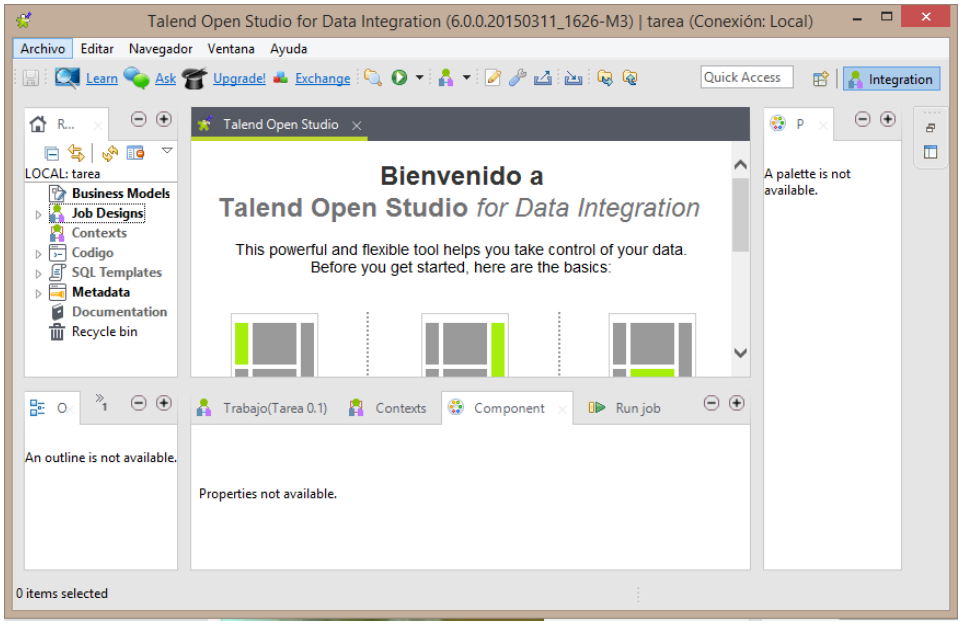
ANEXO A

- **Fuentes de datos a utilizar para realizar las cargas de datos en las herramientas que se van a utilizar.**



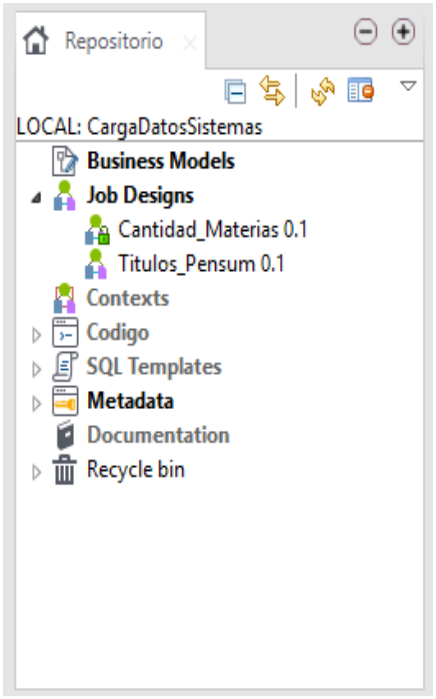
ANEXO B

- **Herramienta de Integración de Datos Talend Open Studio; ambiente para realizar los trabajos de Integración de Datos.**



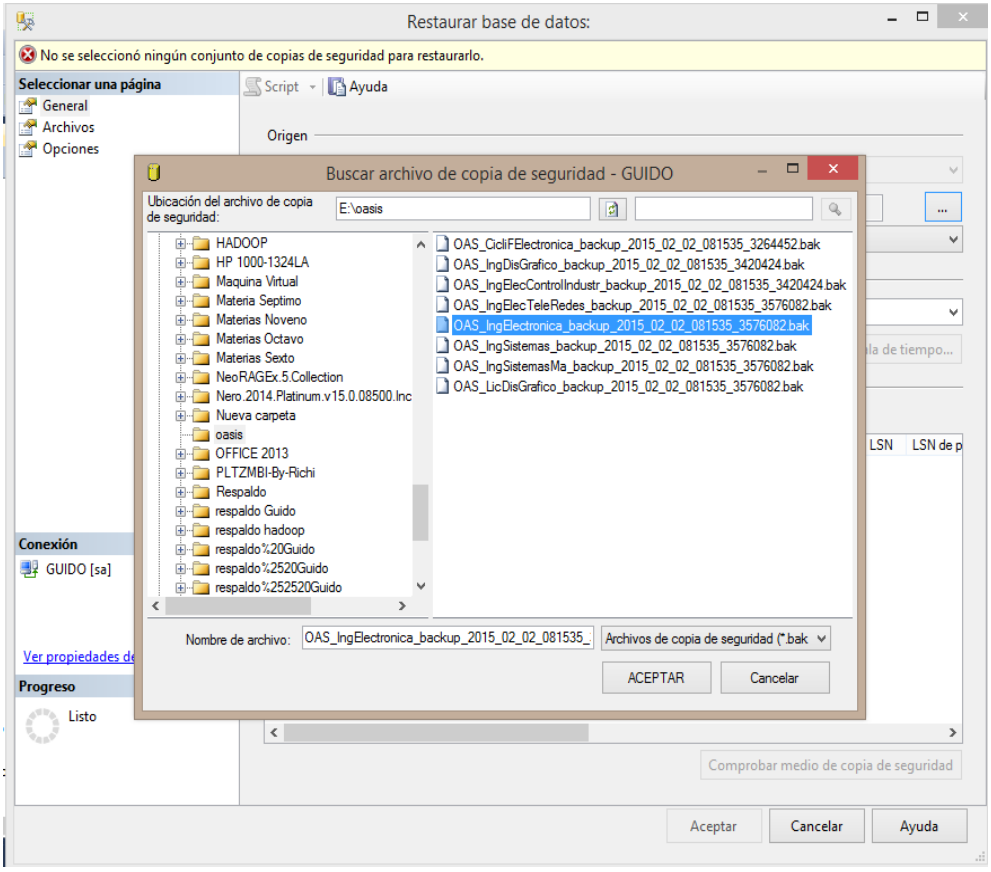
ANEXO C

- **Repositorio para la creación de los Jobs para nuestro estudio comparativo.**



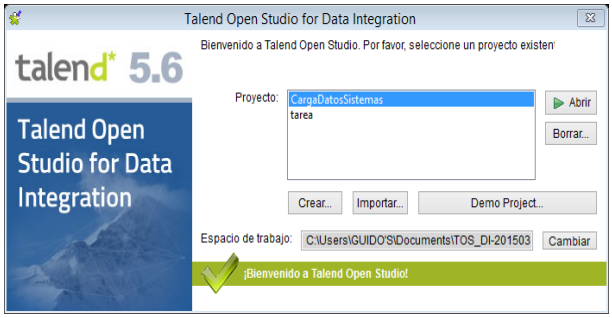
ANEXO D

- **Realización de la Restauración de las bases de datos que nos servirán como fuentes de datos Big Data.**



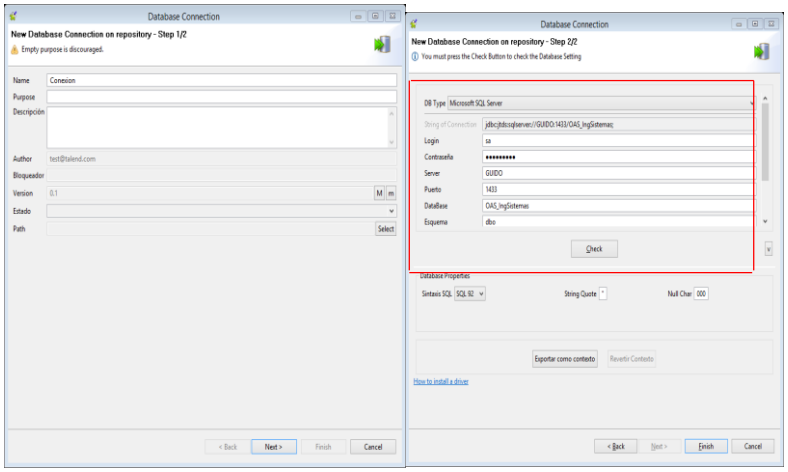
ANEXO E

- **Ejecución de la herramienta de integración Talend Open Studio, para realizar la realización de los Jobs.**



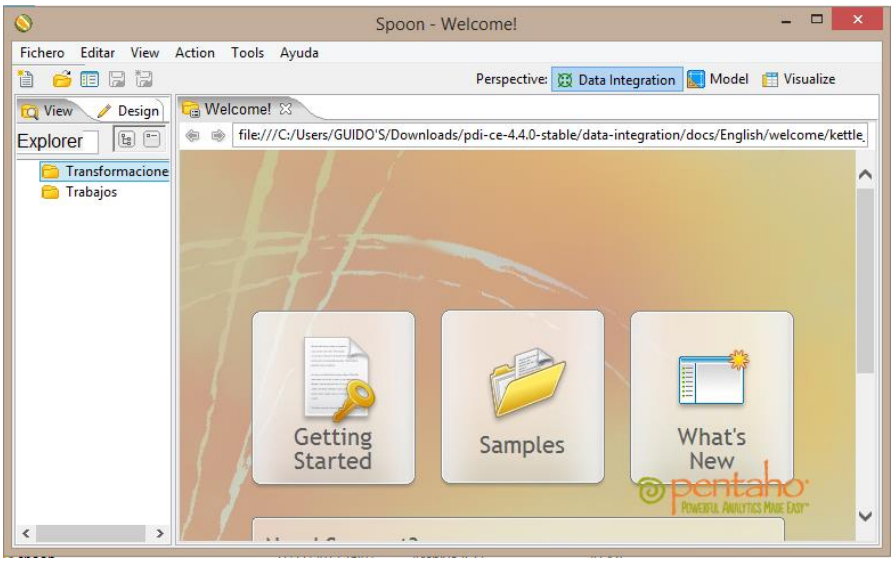
ANEXO F

- **Creación de la conexión hacia nuestra fuente de datos Big Data; que vamos a utilizar.**



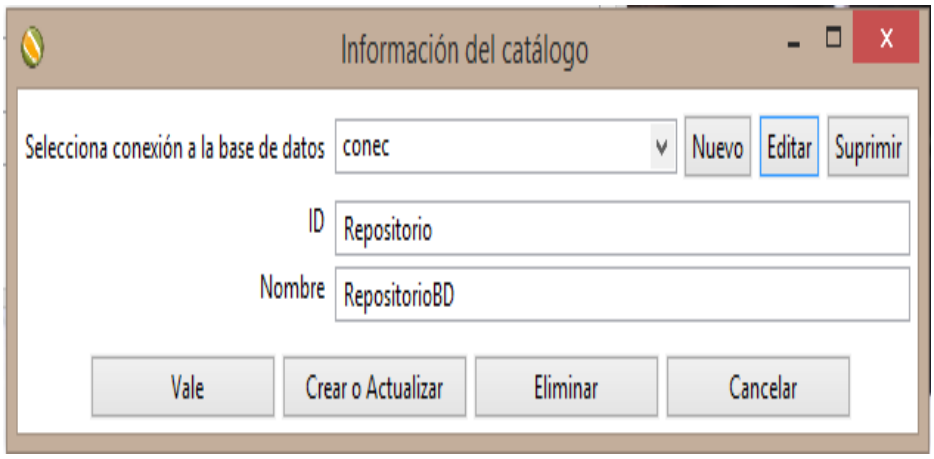
ANEXO G

- **Herramienta de Integración Pentaho Data Integrator, disponible en versión gratuita para la realización de tareas de integración de datos, con diferentes fuentes, incluidas Fuentes Big Data.**



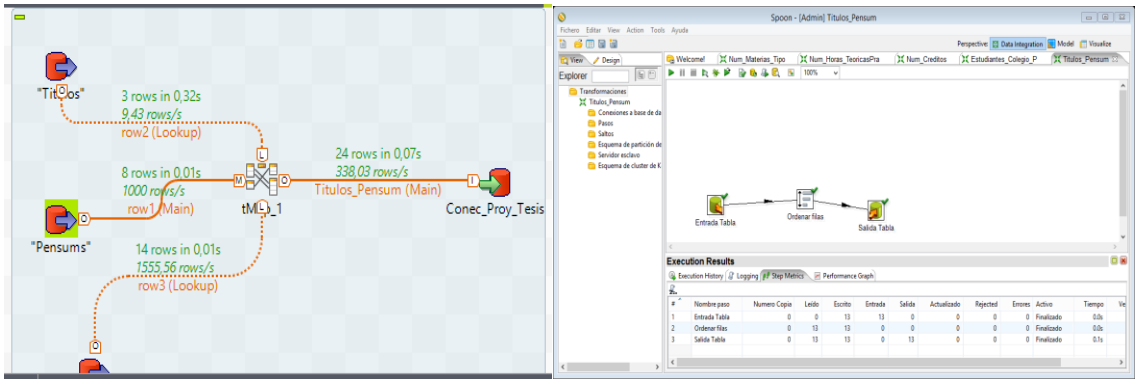
ANEXO H

- **Repositorio de Datos para los trabajos y transformaciones de Datos que se van a desarrollar para las pruebas.**



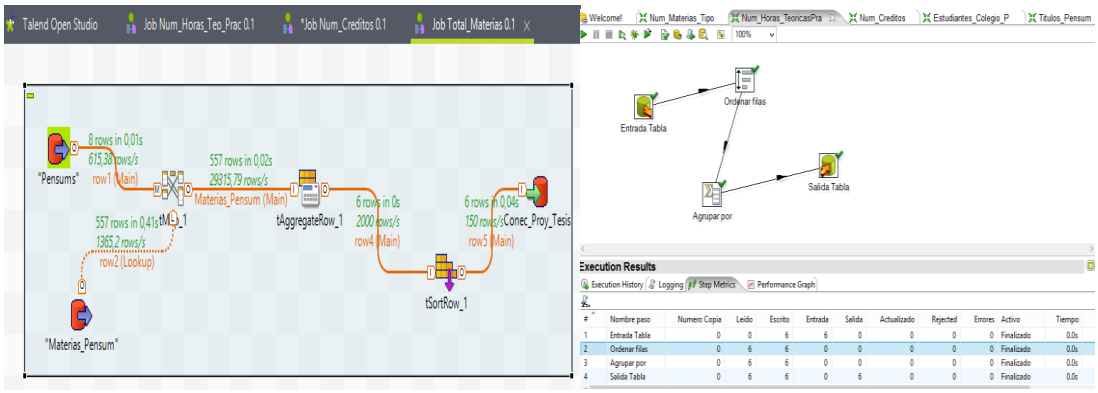
ANEXO I

- **Ejecución del Job para determinar los títulos que proporcionan por pensum con la herramienta Talend Open Studio y Pentaho Data Integrator.**



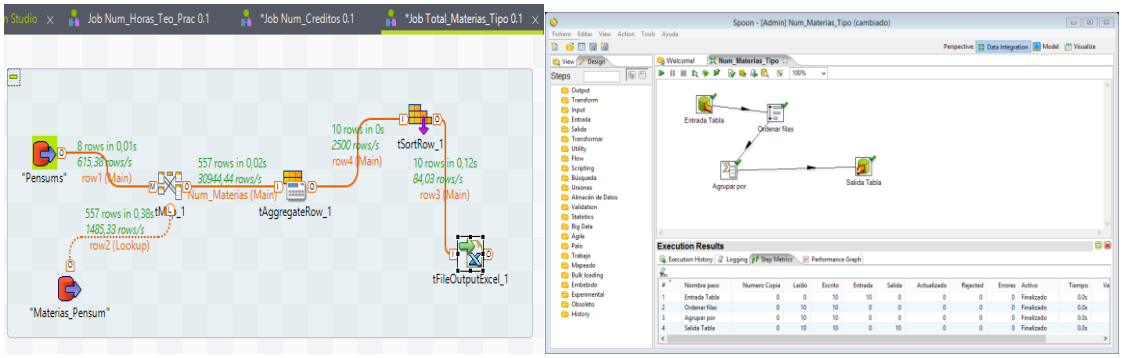
ANEXO J

- **Ejecución del Job para determinar el número de materias por pensum con la herramienta Talend Open Studio y Pentaho Data Integrator.**



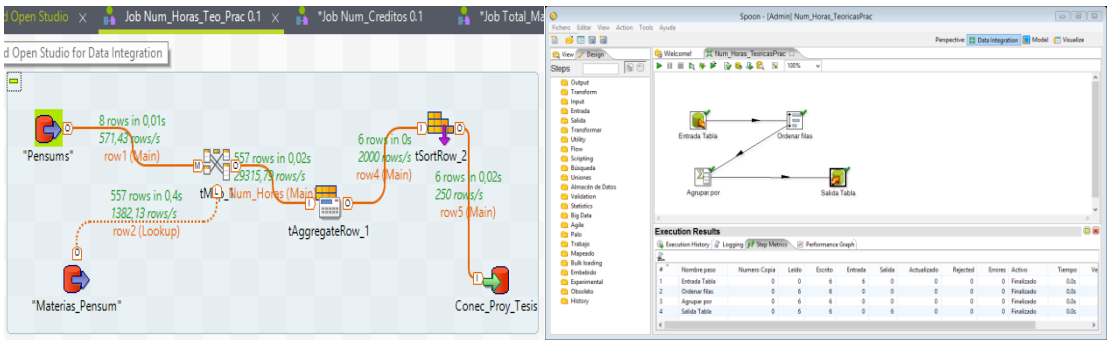
ANEXO K

- **Ejecución del Job para determinar el número de materias por tipo por pensum con la herramienta Talend Open Studio y Pentaho Data Integrator.**



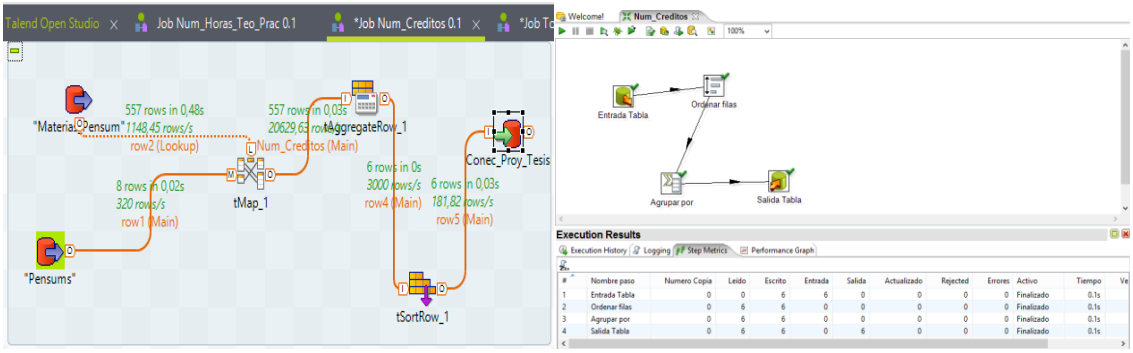
ANEXO L

- **Ejecución del Job para determinar el número de horas teóricas y prácticas por pensum con la herramienta Talend Open Studio y Pentaho Data Integrator.**



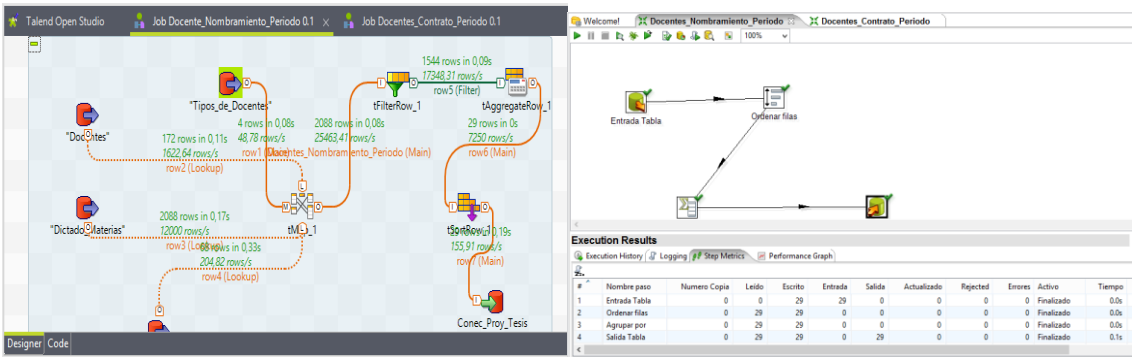
ANEXO M

- **Ejecución del Job para determinar el número de créditos por pensum con la herramienta Talend Open Studio**



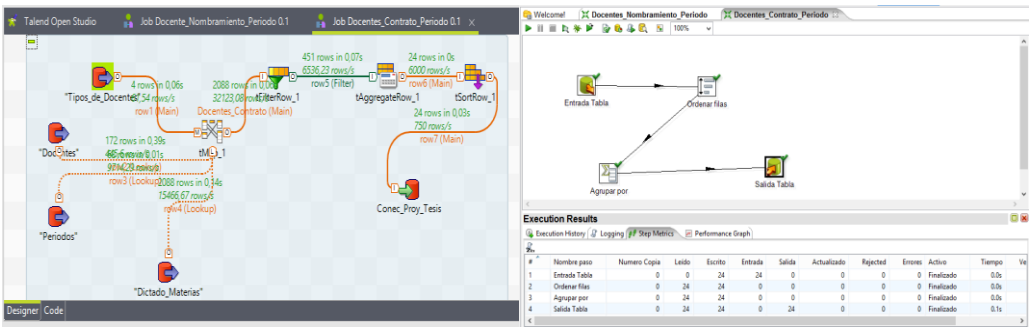
ANEXO N

- **Ejecución del Job: N°. docentes con nombramiento por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator**



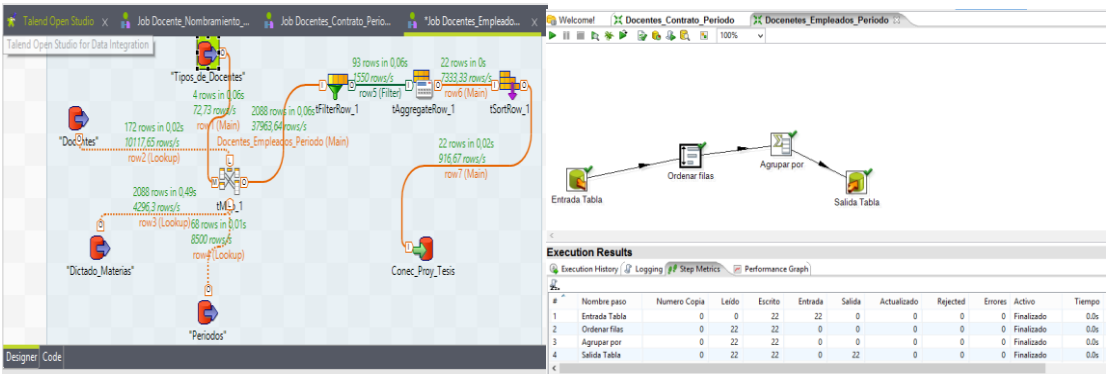
ANEXO O

- **Ejecución del Job: N°. docentes con contrato por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



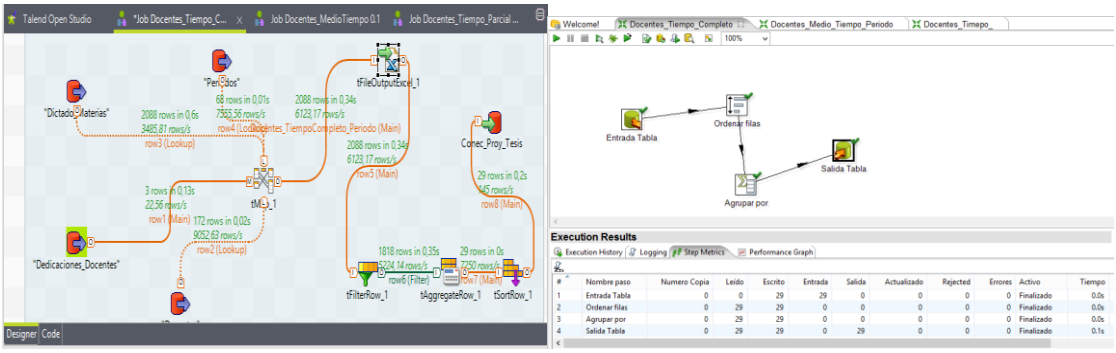
ANEXO P

- **Ejecución del Job: N°. docentes empleados por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



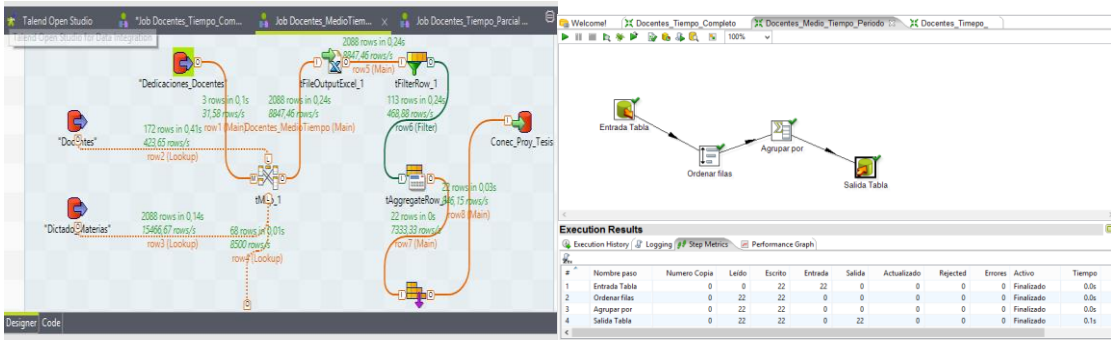
ANEXO Q

- **Ejecución del Job: N°. docentes por tiempo completo por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



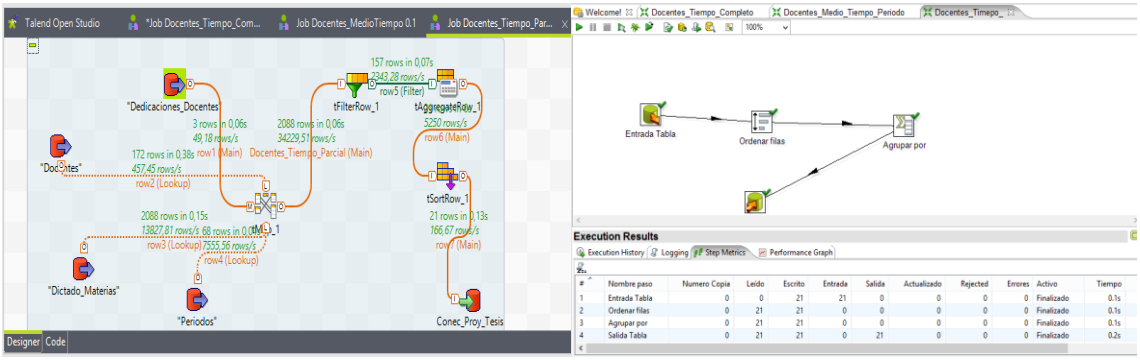
ANEXO R

- **Ejecución del Job: N°. docentes por medio tiempo por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



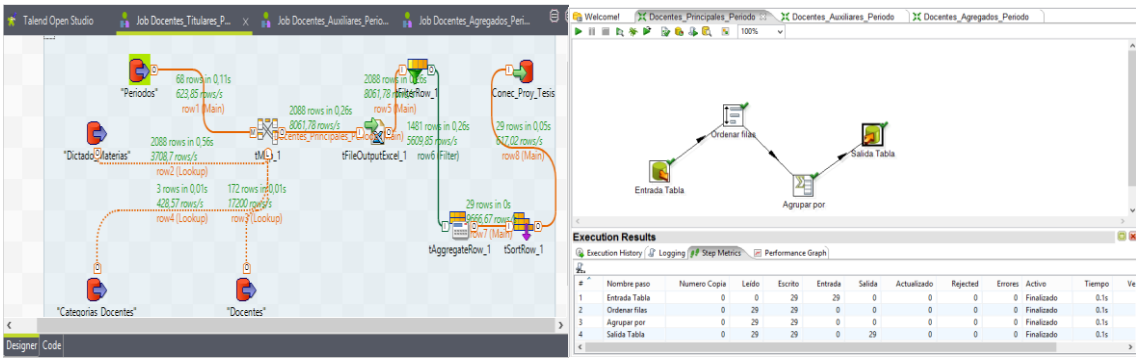
ANEXO S

- **Ejecución del Job: N°. docentes por tiempo parcial por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



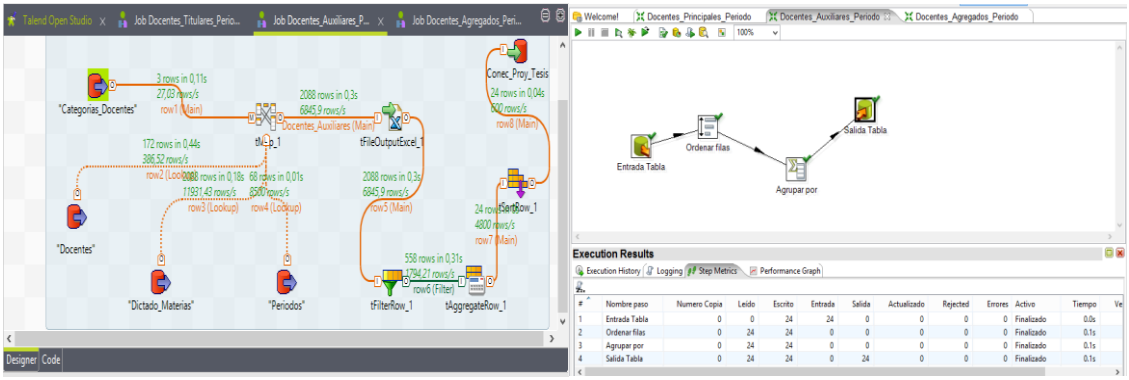
ANEXO T

- **Ejecución del Job: N°. Docentes principales por periodo, con la herramienta Talend Open Studio y Pentaho Data Integrator.**



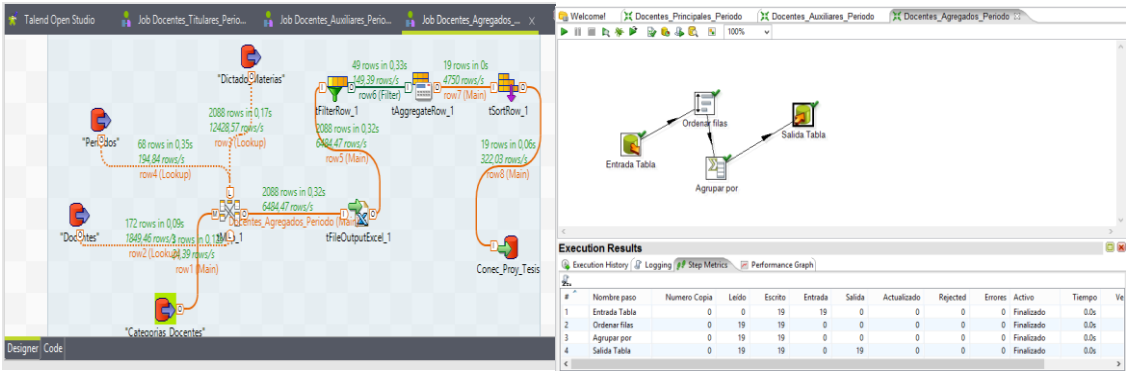
ANEXO U

- **Ejecución del Job: N°. Docentes auxiliares por periodo, con la herramienta Talend Open Studio y Pentaho Data Integrator.**



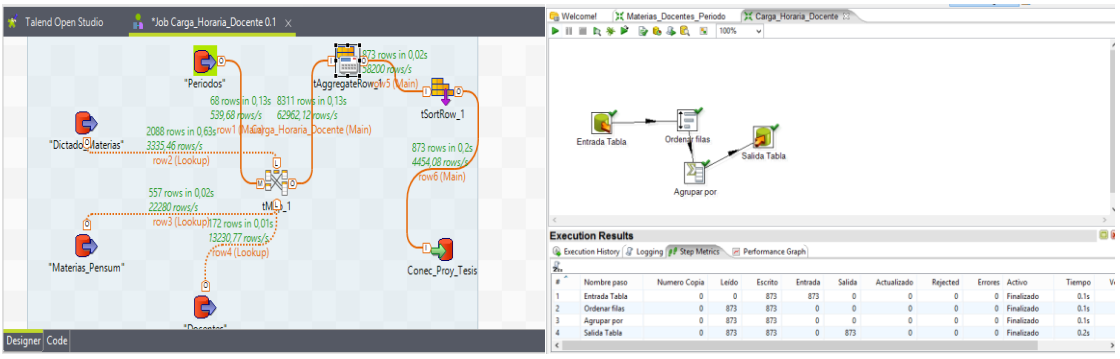
ANEXO V

- **Ejecución del Job: N°. Docentes agregados por periodo, con la herramienta Talend Open Studio y Pentaho Data Integrator.**



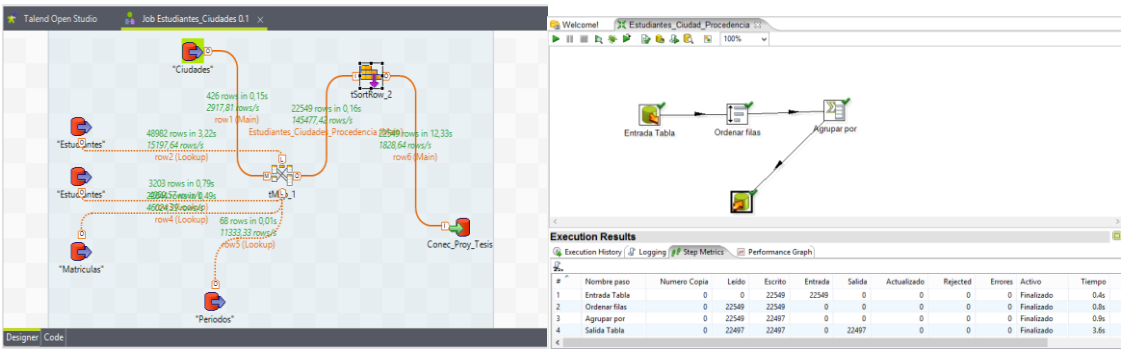
ANEXO W

- **Ejecución del Job. número de horas clases presenciales del profesor en la carrera con la herramienta Talend Open Studio y Pentaho Data Integrator.**



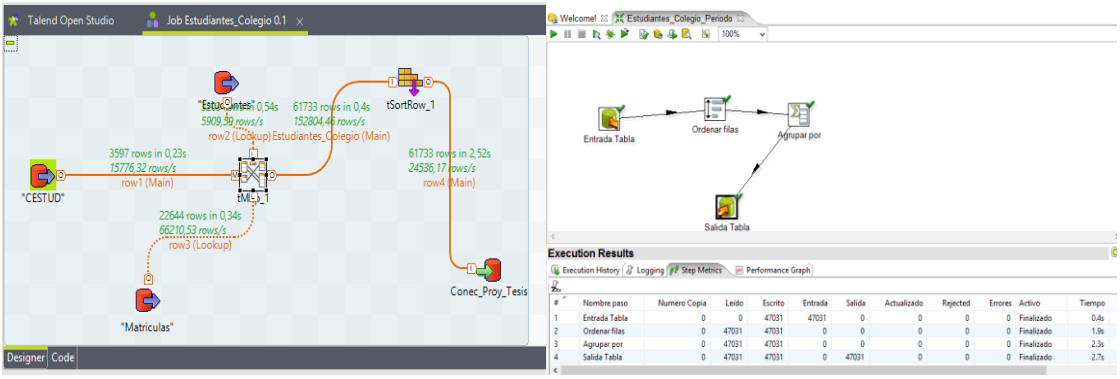
ANEXO X

- **Ejecución del Job. Estudiantes y la ciudad de procedencia por periodo con la herramienta Talend Open Integrator y Pentaho Data Integrator.**



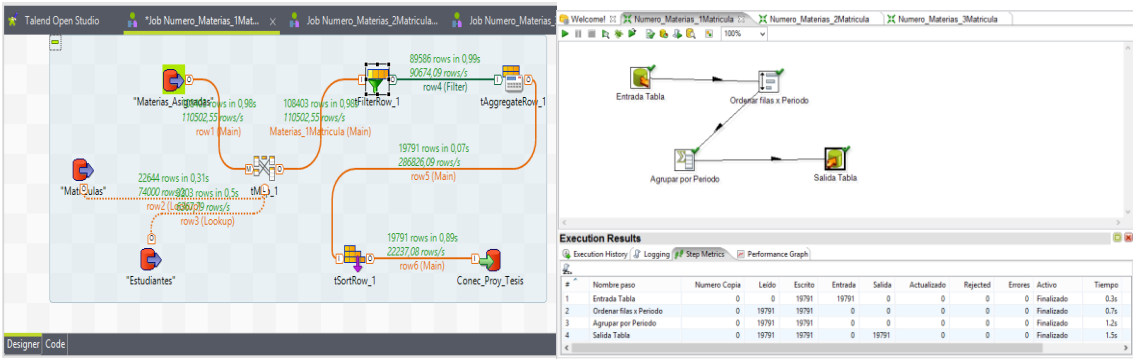
ANEXO Y

- **Ejecución del Job. Estudiantes y el colegio de procedencia por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



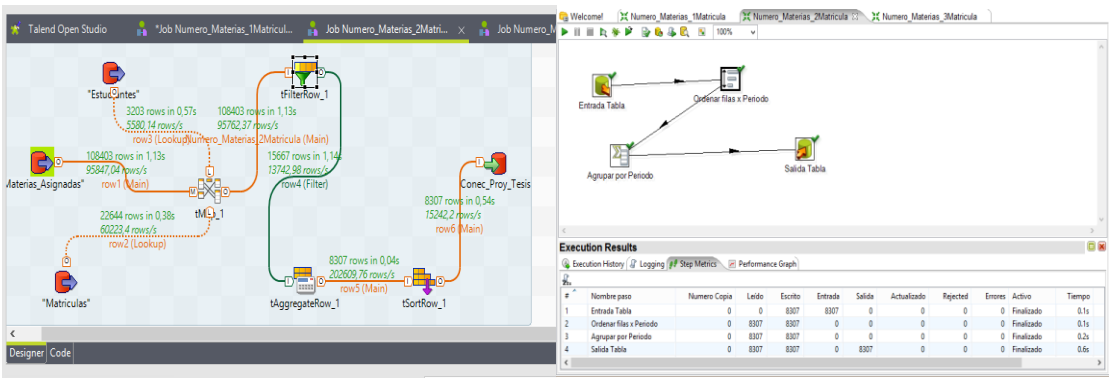
ANEXO Z

- **Ejecución del Job. N°. de materias matriculadas con primera matricula agrupado por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



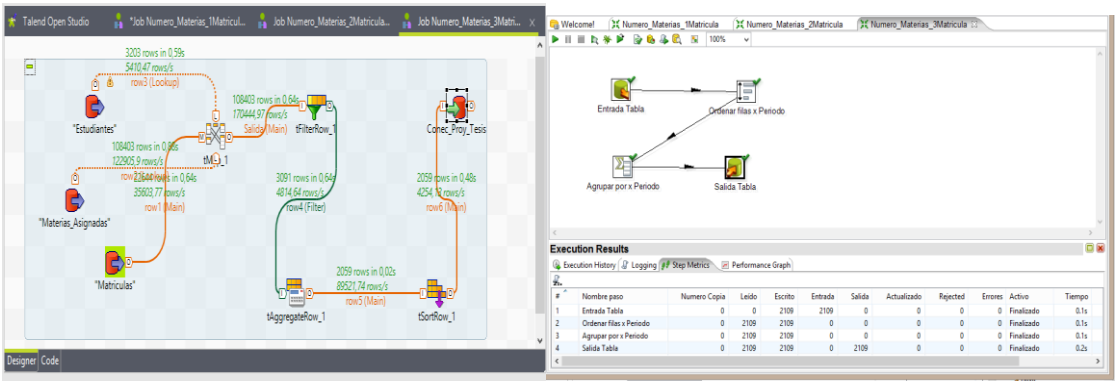
ANEXO A1

- **Ejecución del Job. N°. de materias matriculadas con segunda matricula agrupado por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



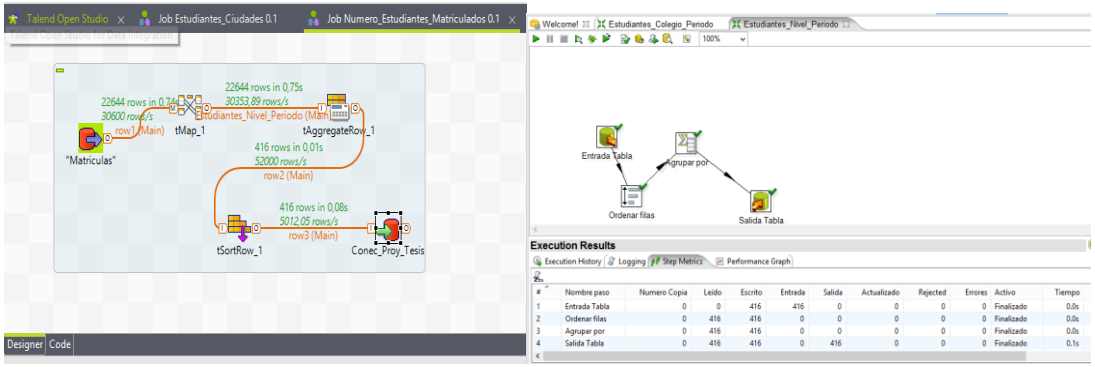
ANEXO B1

- **Ejecución del Job. N°. de materias matriculadas con tercera matricula agrupado por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



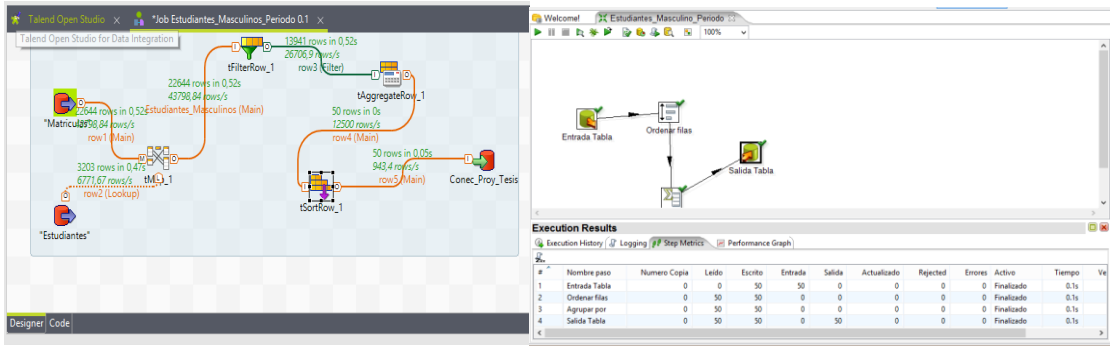
ANEXO C1

- **Ejecución del Job. N°. Estudiantes matriculados por nivel agrupado por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



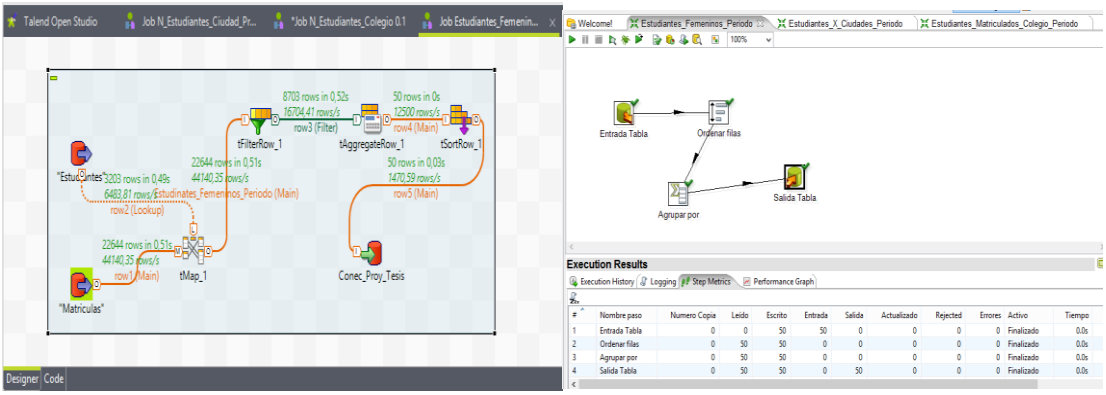
ANEXO D1

- **Ejecución del Job. N°. Estudiantes de sexo Masculino agrupado por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



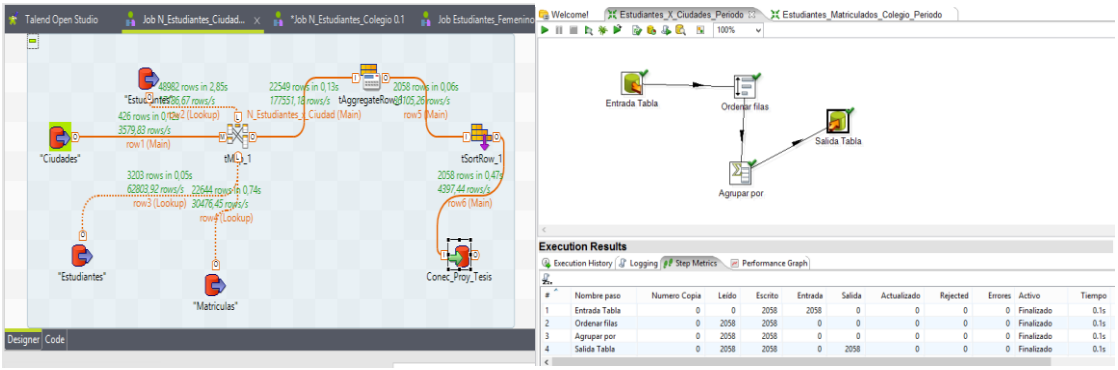
ANEXO E1

- **Ejecución del Job. N°. Estudiantes de sexo femenino agrupado por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



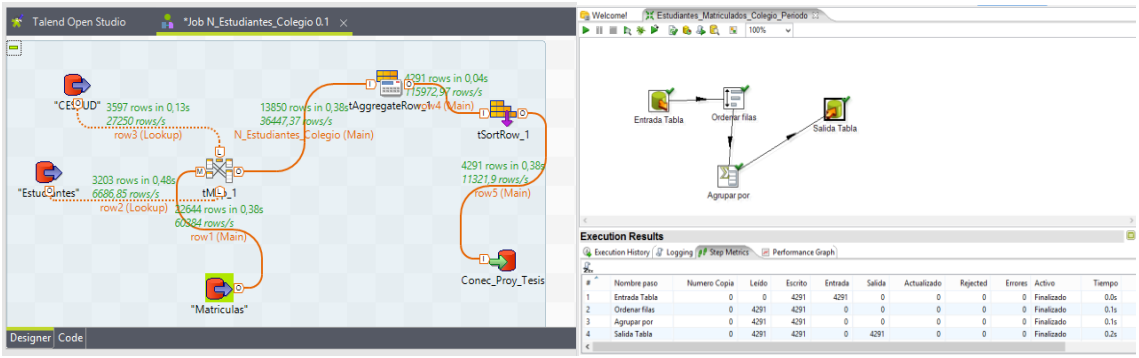
ANEXO F1

- **Ejecución del Job. N°. Estudiantes agrupados por ciudad de procedencia por periodo con la herramienta Talend Open Studio.**



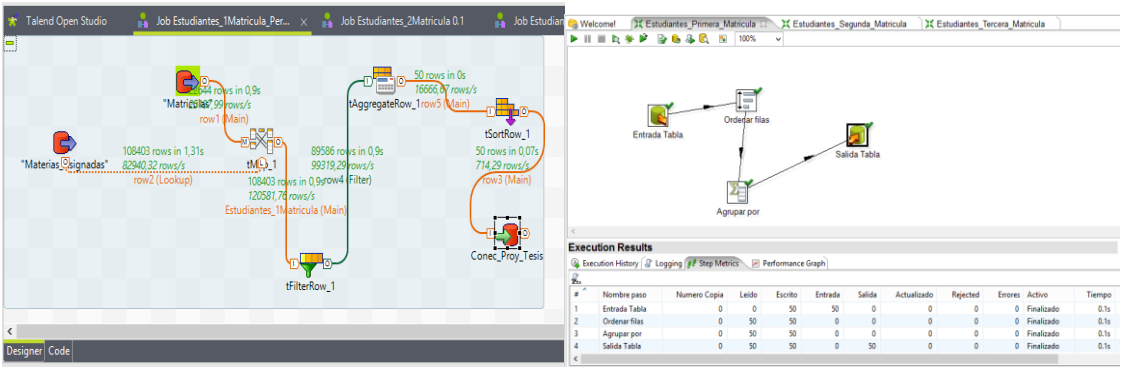
ANEXO G1

- **Ejecución del Job. N°. Estudiantes agrupados por colegio de procedencia por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



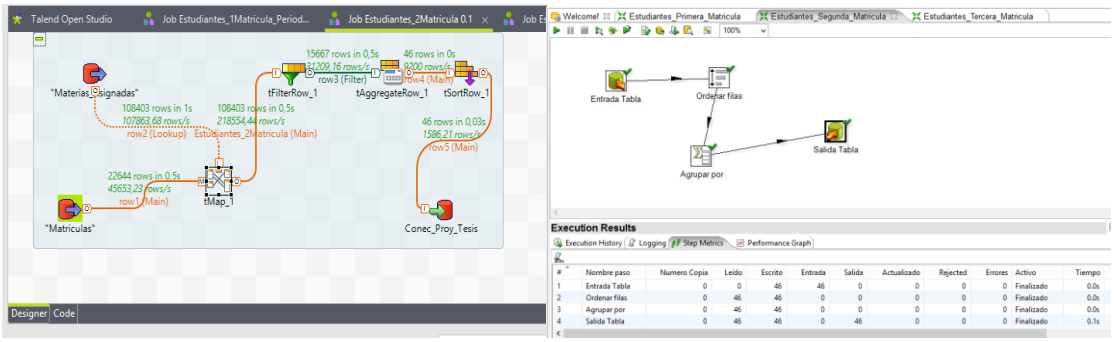
ANEXO H1

- **Ejecución del Job. N°. Estudiantes matriculados con primera matricula por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



ANEXO I1

- **Ejecución del Job. N°. Estudiantes matriculados con segunda matricula por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**



ANEXO J1

- **Ejecución del Job. N°. Estudiantes matriculados con tercera matricula por periodo con la herramienta Talend Open Studio y Pentaho Data Integrator.**