ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

PROYECTO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS

TEMA: "CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO DE LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM."

AUTOR: TORRES FLORES, ROBERT ANDRES

SALAZAR SÀNCHEZ, DANIEL FABRICIO

DIRECTOR: ING. RON, MARIO B.

SANGOLQUÍ

JUNIO, 2015

UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE
CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

CERTIFICADO

Ing. Mario Ron (DIRECTOR DE TESIS)

CERTIFICA

Que el presente trabajo titulado "CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM." fue realizado en su totalidad por el Sr. Robert Andrés Torres Flores y el Sr. Daniel Fabricio Salazar Sánchez como requerimiento parcial a la obtención del título de INGENIERO EN SISTEMAS E INFORMÁTICA

UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE
CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

DECLARACIÓN DE RESPONSABILIDAD

Nosotros, Robert Andrés Torres Flores y Daniel Fabricio Salazar Sánchez

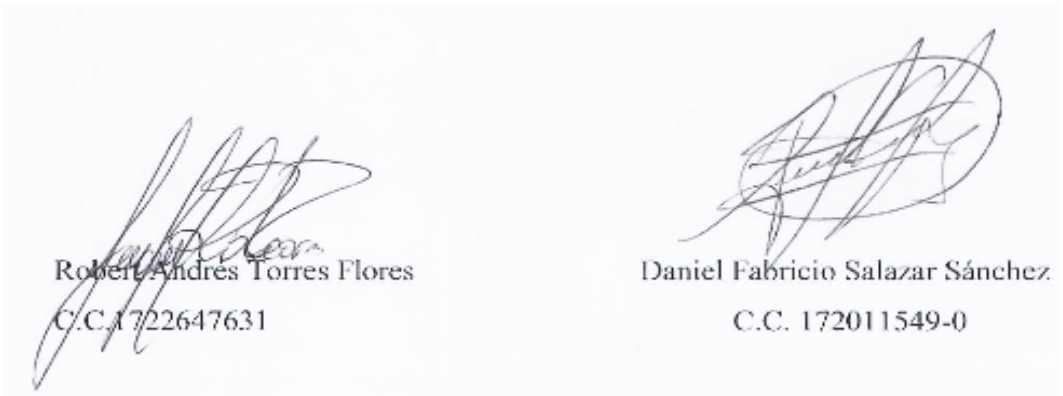
DECLARAMOS QUE:

El proyecto de grado denominado "CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM", ha sido desarrollado con base a una investigación exhaustiva, respetando derechos intelectuales de terceros, conforme las citas, cuyas fuentes se incorporan en la bibliografía.

Consecuentemente este trabajo es de nuestra autoría.

En virtud de esta declaración, nos responsabilizamos del contenido veracidad y alcance científico del proyecto de grado en mención.

Sangolquí, Junio de 2015

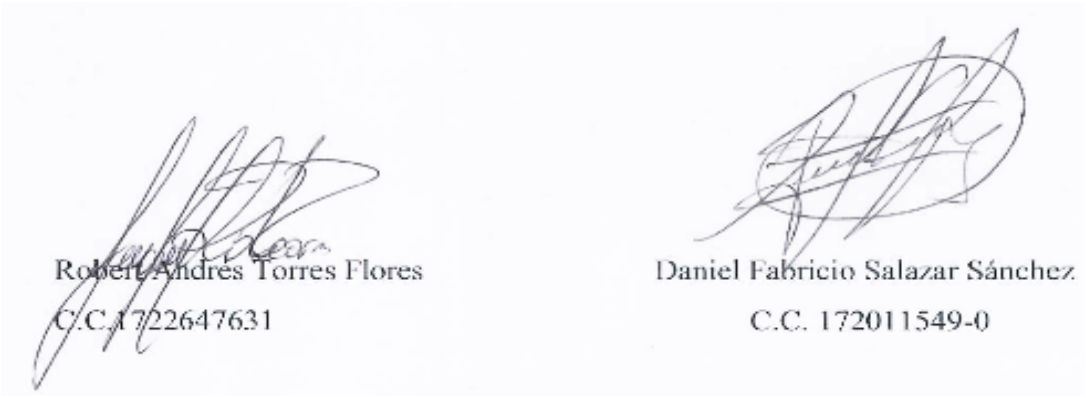


UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE
CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

AUTORIZACIÓN DE PUBLICACIÓN

Nosotros, Robert Andrés Torres Flores y Daniel Fabricio Salazar Sánchez, autorizamos a la UNIVERSIDAD DE LAS FUERZAS ARMADAS – ESPE, la publicación, en la biblioteca virtual de la Institución del proyecto de tesis "CONSTRUCCIÓN DE UN REPOSITORIO POR MEDIO DE LA EXTRACCIÓN Y PROCESAMIENTO LA INFORMACIÓN PROVENIENTE DE CDRS GENERADOS POR TELCOS PARA ANALIZAR EL CONSUMO DE LOS DATOS OBTENIDOS, DENTRO DE LA EMPRESA SOFTCONSULTING S.A UTILIZANDO HERRAMIENTAS DE BIG DATA DE IBM", cuyo contenido, ideas y criterios son de nuestra exclusiva responsabilidad y autoría.

Sangolquí, Junio de 2015



DEDICATORIA

El presente proyecto va dedicado de manera muy especial a mi madre que ha sido mi roca este último año tan difícil en la vida de los dos y gracias a ella he podido lograr tantas cosas buenas en este año, ya que siempre ha estado a mi lado brindándome su mano amiga, dándome a cada instante una palabra de aliento para llegar a culminar mi profesión.

A mí amada esposa Mishell, por su apoyo y ánimo que me brinda día con día para alcanzar nuevas metas tanto profesionales y personales.

A fin, A ti madre y mi amada Mishell, ya que sin ustedes no me hubiese levantado tan rápido de este último año, en cada una de ellas se ve en gran refrán: "detrás de todo gran hombre hay una gran mujer"

A mí adorada hija Valentina y a mi nuevo bebé Robertito, las personitas que me alientan día a día a no desmayar en mi crecimiento personal, profesional y como padre.

A mi padre que está en el cielo, a mi hermano Xavier y a Luis Felipe, por sus palabras de aliento.

Robert Andrés Torres Flores

El presente proyecto va dedicado con todo mi corazón a mi madre que me cuida desde el cielo, que siempre me motivo siempre a seguir adelante a pesar de las adversidades, me enseño valores los que me han servido para salir

victorioso en mi carrera la cual estoy culminándola y a quien le prometí llegar hasta aquí.

A mis abuelitos que fueros mis padres por muchos años quienes me enseñaron el valor de las cosas y por quien ahora estoy culminando esta etapa de mi vida.

A mis hermanas que son mi motor para seguir como profesional y brindarles un buen ejemplo y a todas aquellas personas que confiaron en mí.

Daniel Fabricio Salazar Sánchez

AGRADECIMIENTO

Agradezco de todo corazón a mi familia por su apoyo en todo mi desarrollo como profesional.

A mi esposa por su constancia en el desarrollo del proyecto de tesis.

A la empresa SoftConsulting S.A por el apoyo técnico y profesional en el desarrollo de la misma, por brindar su infraestructura y conocimientos de arquitectura de software empresarial.

Al Ing. Mario Ron por el seguimiento constante en el desarrollo de la tesis y empeño en este gran tema de tesis.

Robert Andrés Torres Flores

Agradezco en primer lugar a Dios por darme la fuerza y la paz para enfrentar momentos duros de mi vida, darme, a mis abuelitos que desde el principio me brindaron su apoyo incondicional y su entera confianza en mi meta de llegar a ser ingeniero.

Agradezco a nuestro director de Tesis Ing. Mario Ron por el gran afecto, la confianza, la paciencia y el apoyo brindado para culminar el presente proyecto, finalmente a mi mejor amiga y novia que siempre me dio ánimos y me motivo cada día para lograr mis objetivos

Daniel Fabricio Salazar Sánchez

TABLA DE CONTENIDO

Índice de Figuras

Índice de Tablas

RESUMEN

El propósito de este trabajo fue crear un repositorio de Big Data dentro de la empresa SoftConsulting S.A para almacenar grandes cantidades de información, acceder con mayor rapidez a realizar consultas correspondientes a la información de los CDRS que generan las TELCOS, evitar o manipulación de información y brinde grandes facilidades de escalamiento en horizontal y vertical. En la realización del proyecto utilizó el formato de CDR de Asterisk con el fin de utilizar como referencia de n estándar, además también se utilizó las herramientas de IBM para extracción, procesamiento y almacenamiento de grandes cantidades de información como se hace referencia a Big Data. Los resultados de la creación del repositorio en el clúster de Big Data y las pruebas de rendimiento y estrés muestran la curva de tendencia a estabilización del tiempo de respuesta independiente a la cantidad de datos retornados utilizando IBM InfoSphere BigInsights V 2.1.1.1Standard Edition por su gran ventaja en el mercado de IBM BigSQL V 3.0 comparado con otras herramientas similares es de 10 a 1.

Palabras Clave:
CDRS
TELCOS
ASTERISK
INFOSPHERE BIGINSIGHTS
BIGSQL

ABSTRACT

The purpose of this work was to create a Big Data repository within the company SoftConsulting S.A to store large amounts of information, faster access to conduct appropriate queries to the CDRs information that generate the TELCOS, avoiding or manipulating information and offer great facilities of vertical and horizontal scaling. In the project he used the CDR format Asterisk in order to be used as reference standard, plus IBM tools for extraction, processing and storage of large amounts of information are also used as a reference to Big Data. The repository's creation results in the Big Data cluster and the performance and stress test show the trend curve stabilization time independent response to the amount of data returned using IBM InfoSphere BigInsights V 2.1.1.1Standard Edition for its great market advantage of IBM BigSQL V 3.0 compared with other similar tools is 10 to 1.

KeyWords:
CDRS
TELCOS
ASTERISK
INFOSPHERE BIGINSIGHTS
IBM BIGSQL

CAPÌTULO 1

1.1 INTRODUCCION

El primer cuestionamiento que posiblemente llegue a su mente en este momento es ¿Qué es Big Data y porqué se ha vuelto tan importante? pues bien, en términos generales se puede referir como a la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de peta bytes y exabytes de datos. Entonces ¿Cuánto es demasiada información de manera que sea elegible para ser procesada y analizada utilizando Big Data? Analizar primeramente en términos de bytes:

Gigabyte = 109 = 1, 000, 000,000

Terabyte = 1012 = 1, 000, 000, 000,000

Petabyte = 1015 = 1, 000, 000, 000, 000,000

Exabyte = 1018 = 1, 000, 000, 000, 000, 000,000

Además del gran volumen de información, esta existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso.

Estas son las características principales de una oportunidad para Big Data.

¿De dónde proviene esta información?

Los seres humanos están creando y almacenando información constantemente y cada vez más en cantidades astronómicas. Se podría decir que si todos los bits y bytes de datos del último año fueran guardados en CD, se generaría una gran torre desde la Tierra hasta la Luna y de regreso.

Esta contribución a la acumulación masiva de datos la pueden encontrar en diversas industrias, las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. En muchos países se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, etc., y si a todo esto se añaden transacciones financieras realizadas en línea o por dispositivos móviles, análisis de redes sociales (en Twitter son cerca de 12 Terabytes de tweets creados diariamente y Facebook almacena alrededor de 100 Petabytes de fotos y videos), ubicación geográfica mediante coordenadas GPS, en otras palabras, todas aquellas actividades que la mayoría de las personas realizan varias veces al día con sus "smartphones", Se estaría hablando de que se generan alrededor de 2.5 quintillones de bytes diariamente en el mundo.

1 quintillón = 10 30 = 1,000,000,000,000,000,000,000,000,000,000

De acuerdo con un estudio realizado por Cisco, entre el 2011 y el 2016 la cantidad de tráfico de datos móviles crecerá a una tasa anual de 78%, así como el número de dispositivos móviles conectados a Internet excederá el número de habitantes en el planeta. Las naciones unidas proyectan que la población mundial alcanzará los 7.5 billones para el 2016 de tal modo que habrá cerca de 18.9 billones de dispositivos conectados a la red a escala mundial, esto conllevaría a que el tráfico global de datos móviles alcance 10.8 Exabytes mensuales o 130 Exabytes anuales. Este volumen de tráfico

previsto para 2016 equivale a 33 billones de Dvd anuales o 813 cuatrillones de mensajes de texto.

Pero no solamente los seres humanos son quienes contribuyen a este crecimiento enorme de información, existe también la comunicación denominada máquina a máquina (M2M machine-to-machine) cuyo valor en la creación de grandes cantidades de datos también es muy importante. Sensores digitales instalados en contenedores para determinar la ruta generada durante una entrega de algún paquete y que esta información sea enviada a las compañías de transportación, sensores en medidores eléctricos para determinar el consumo de energía a intervalos regulares para que sea enviada esta información a las compañías del sector energético. Se estima que hay más de 30 millones de sensores interconectados en distintos sectores como automotriz, transportación, industrial, servicios, comercial, etc. y se espera que este número crezca en un 30% anualmente.

¿Qué tipos de datos debo explorar?

Muchas organizaciones se enfrentan a la pregunta sobre ¿qué información es la que se debe analizar?, sin embargo, el cuestionamiento debería estar enfocado hacia ¿qué problema es el que se está tratando de resolver?

Si bien se sabe que existe una amplia variedad de tipos de datos a analizar, una buena clasificación nos ayudaría a entender mejor su representación, aunque es muy probable que estas categorías puedan extenderse con el avance tecnológico.

- Web y Medios Sociales: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, blogs, entre otros.
- Maquina--Maquina (M2M): M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de

redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

- Transacciones Grandes de datos: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.
- Biométricos: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.
- Generación de Información Humana: Las personas generan diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

1.2 ANTECEDENTES

SOFTCONSULTING S.A es una empresa ecuatoriana legalmente establecida, con residencia en la ciudad de Quito, república del Ecuador. La cual se encarga de brindar servicios de asesoría, consultoría de data Warehousing y soluciones innovadoras de tecnología para empresas medianas y grandes utilizando en gran parte software e infraestructura IBM, teniendo principalmente en su cartera de clientes varias TELCOS donde se realiza tareas de consultoría, arquitectura y desarrollo de ETLs e innovaciones tecnológicas para la automatización de procesos internos de la compañía.

Desde el punto de vista judicial y de control de fraude por parte de las autoridades del estado Ecuatoriano hacia las operadoras telefónicas que utilizan el espectro radioeléctrico, en el cual se requerirá a su tiempo, que provean la información de los CDRs (Call Detail Record) por medio de un único repositorio disponible para las autoridades judiciales por lo que

SOFTCONSULTING S.A como principal proveedor de innovación tecnológica se ve en la necesidad de desarrollar un repositorio por medio de la extracción, procesamiento y consumo de la información de los CDRs generados día a día del tráfico de voz de los usuarios que utilizan el servicio de telefonía celular de las operadoras denominadas TELCO, para poseer los procedimientos necesarios para manejar esta información y generar el repositorio sin afectaciones a la misma en toda la construcción del repositorio si se requiriese por parte de su cliente TELCOS cualquier otra TELCO existente en el país o fuera del mismo, a un bajo impacto en su infraestructura y costo.

1.3 PLANTEAMIENTO DEL PROBLEMA

El almacenamiento y procesamiento de CDRs es un requerimiento que se exige en la mayoría de los países con entidades regulatorias de telecomunicaciones, el CDR es un archivo plano en el que la operadora registra los detalles de llamadas, como tipo de tecnología (2G y 3G), tiempo, duración, origen y destino, entre otros. Esta información llega a ser bastante extensa dependiendo del número de usuarios de la red celular asignada a la operadora por lo cual utilizar un motor de BD sea este: Teradata, Oracle, MS SQL, DB2 u otro, implica costos extremadamente grandes tanto en infraestructura como en licenciamiento, por lo que en la actualidad la tendencia es utilizar tecnología relevante a Big Data en donde gracias a las características de gran capacidad de almacenamiento y procesamiento puede solventar las exigencias de las entidades regulatorias, brindando la posibilidad de crear un repositorio de información de los CDRs con la información relevante y brindando adicionalmente la capacidad de consultar dicha información en menor tiempo y con diversidad de tipos de datos que se requiera estructurar.

1.4 OBJETIVOS DEL PROYECTO

1.4.1 OBJETIVOS GENERAL

Construir un repositorio extrayendo y procesando la información proveniente de CDRs generados por TELCOS para analizar el consumo de los datos obtenidos, utilizando herramientas de Big Data de IBM, dentro de la empresa SOFTCONSULTING S.A.

1.4.2 OBJETIVOS ESPECÍFICOS

Configurar un ambiente de instalación de la solución IBM InfoSphere BigInsights Standard Version.

Instalar un clúster de IBM InfoSphere BigInsights Standard Edition con 5 nodos, 2 de administración y 3 de datos.

Configurar IBM InfoSphere Information Server – Data Stage para extraer el CDR, procesarlo y cargarlo a IBM InfoSphere BigInsights Standard Edition.

Analizar la estructura de la información de los CDRs generados por el MEDIADOR de la TELCO por medio de una muestra proporcionada a SOFTCONSULTING S.A por parte de las TELCOS.

Seleccionar la información concerniente y pertinente a los usuarios como: fechas, y horas de llamadas, números de origen y destino de las llamadas, celdas de comunicación, tecnología usada y centrales de comunicación.

Realizar procesos de calidad de datos a bajo nivel en el proceso de carga de los CDRs utilizando ETLs (Jobs Paralelos) en los números de origen y destino de llamadas, celdas, imsi, imei, fechas y horas de llamadas utilizando la herramienta de IBM InfoSphere Information Management DataStage y almacenado la información en IBM InfoSphere BigInsights Standard Edition y cuantificar los datos ingresados si existiesen registros

rezagados para determinar la tendencia de llamadas en un determinado periodo utilizando un motor de BD Teradata .

Reestructurar la información estableciendo tipos de datos de mayor impacto en reducción de búsqueda y almacenamiento en tablas temporales y tablas RCFiles (Registro de archivos en columnas) utilizando HIVE IBM BIGSQL.

Realizar Jobs Secuenciales a través de IBM InfoSphere Information Management – DataStage para generar un proceso cíclico y automatizado que garantice una autonomía del sistema para evitar la manipulación o alteración del flujo o información de los CDRs

Realizar Shell Scripts de carga de Hadoop a HIVE para serializar la información y garantizar la seguridad de la información.

1.5 JUSTIFICACIÓN DEL PROYECTO

Actualmente las TELCOS proveen esta información con cierto retraso de 48 horas laborables o más dependiendo de la solicitud de las autoridades judiciales del Ecuador. Las TELCOS usan sistemas de bases de datos relacionales, causando afectaciones a sus procesos de negocio afectando así la carga de procesamiento en disco, RAM y CPU a sus servidores de producción, por lo que se genera la necesidad de utilizar un paradigma diferente al de bases de datos al momento que se necesite un mayor nivel de disponibilidad de la información provista en los CDRs Por lo cual SOFTCONSULTING S.A analiza el escenario del cliente y decide realizar la metodología de extracción procesamiento y consumo de información provenientes de los MEDIADORES, que generan las CDRs utilizando herramientas IBM existentes en las TELCOS para el paradigma del manejo de información de gran volumen, velocidad y variedad de datos denominada Big Data.

Con el desarrollo de la metodología basada en tecnología IBM en SOFTCONSULTING S.A se quiere evitar los siguientes aspectos:

JUDICIAL

- Sanciones por la manipulación errónea y divulgación de la información concerniente a las llamadas de los usuarios.
- Sanciones por retraso en la entrega de la información solicitada por las autoridades.

EMPRESA

- Dependencia de una o varias plazas de trabajo para manipular la información.
- Impacto en los procesos de negocio: facturación, marketing, entre otros.
- Incremento en el costo de adquirir hardware o nuevo licenciamiento para bases de datos relacionales
- Incremento en el almacenamiento de CDRs en sistemas alternos que difieren de su objetivo de funcionamiento por ejemplo: servidores de Stagging del DWH (Data Warehousing), servidores SFTP (SSH File Transfer Protocol), de transición de procesos batch, entre otros.

SOFTCONSULTING S.A con el desarrollo de una metodología basada en Big Data podrá garantizar a la TELCO el acceso a la información de los CDRs sin importar del tráfico generado por los usuarios, con un tiempo mínimo, empleando seguridades en el manejo de la información desde el origen del CDR hasta su destino (repositorio).

La implementación de esta metodología requiere de procesos en la extracción, manipulación y consumo de la información de los CDRs, para los cuales se utilizará software IBM mientras que el hardware queda a libre disposición del cliente.

EL justificante de la elección de software IBM nace de la necesidad de obtener soporte inmediato tras cualquier incidente en todo el proceso, esto se debe a que IBM es una de las marcas que se encuentra más de cuatro años seguidos en el cuadrante de Gartner como una de las soluciones que satisface las necesidades de los clientes.

1.6 ALCANCE DEL PROYECTO

En el citado desarrollo se construirá un repositorio por medio de la extracción y procesamiento de la información proveniente de CDRs generados por TELCOS para analizar el consumo de los datos obtenidos, dentro de la empresa SOFTCONSULTING S.A utilizando herramientas de Big Data de IBM, que se encuentra en la ciudad de Quito, Edificio Puerto de Palos, Av. 12 de octubre N3-25 Y Cordero, la metodología tiene como objetivo principal el servir de base procedimental para aplicar el procesamiento de CDRs generados por la TELCO para proveer rapidez de acceso, integridad y seguridad de la información cuando las autoridades judiciales del Ecuador lo soliciten .

La construcción del repositorio permitirá determinar factores críticos de éxito por medio del análisis del tiempo de:

- Disponibilidad del CDR
- Tiempo en la Extracción, procesamiento y carga a IBM InfoSphere BigInsights.
- Tiempo Carga a HIVE y serialización de la información
- Tiempo de respuesta por medio de consultas realizadas desde IBM BIGSQL por parámetros de: número celular, celda, imei, imsi y fecha.
- Cantidad de meses de almacenamiento de los CDRs en IBM InfoSphereBigInsights.

1.7 SOFTCONSULTING

Es una compañía dedicada al desarrollo de proyectos de Gestión de Información por medio de soluciones IBM, incluyendo: Motores de Bases de Datos, Data Warehousing, Gobernabilidad de Datos, Integración de Información, Calidad de Datos y Portales Corporativos.

Sofconsulting S.A cuenta con experiencia de más de 10 años en el área y más de 200 proyectos exitosos en la región, lo que ha permitido que IBM le otorgue varios reconocimientos.

Ofrece asesoría y consultoría para permitir que sus negocios crezcan de la mano de su infraestructura tecnológica.

Cuenta con ingenieros certificados y con la suficiente experiencia en administración y optimización de motores de bases de datos IBM INFORMIX y DB2, Bodegas de datos en ambientes IBM INFORMIX y DB2 con herramientas de COGNOS. Soluciones de Integración de Información con DATASTAGE y QUALITY STAGE, soluciones de gobernabilidad de datos a través de OPTIM y desarrollos de WEBSPHERE y RATIONAL para permitirle a una compañía centrarse en su negocio y a la vez con conocimientos bastos en las nuevas tecnologías de Big Data como IBM INFOSPHERE BIGINSIGHTS, BIM STREAMS Y IBM WATSON EXPLORER.

CAPÌTULO 2

2.1 BIG DATA

Para poder esquematizar el concepto, uso y aplicaciones de Big Data deben centrarse inicialmente en la pregunta **¿Qué es Big Data?**
“Todos forman parte de ese gran crecimiento de datos” (IBM Corporation)
Debido al gran avance que existe día con día en las tecnologías de información, las organizaciones se han tenido que enfrentar a nuevos desafíos que les permitan analizar, descubrir y entender más allá de lo que sus herramientas tradicionales reportan sobre su información, al mismo tiempo que durante los últimos años el gran crecimiento de las aplicaciones disponibles en internet (geo-referenciamiento, redes sociales, entre otras.) han sido parte importante en las decisiones de negocio de las empresas. El presente artículo tiene como propósito introducir al lector en el concepto de Big Data y describir algunas características de los componentes principales que constituyen una solución de este tipo.

2.2 Componentes de una plataforma Big Data

La extensa cantidad de información que debe ser procesada ha generado un costo potencial. Desde luego, el ángulo correcto que actualmente tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto Hadoop.

2.2.1 Hadoop

Hadoop *está inspirado en el proyecto de Google File System (GFS) y en el paradigma de programación* MapReduce*, el cual consiste en dividir en dos tareas (mapper – reducer) para manipular los datos distribuidos a nodos de un clúster logrando un alto paralelismo en el procesamiento. [5] Hadoop está*

compuesto de tres piezas: Hadoop Distributed File System *(HDFS),* Hadoop MapReduce *y* Hadoop Common.

2.2.1.1 *Hadoop Distributed File System (HDFS)*

Los datos en el clúster de Hadoop son divididos en pequeñas piezas llamadas *bloques* y distribuidas a través del clúster; de esta manera, las funciones map y reduce pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes.

La siguiente figura ejemplifica como los bloques de datos son escritos hacia HDFS. Observe que cada bloque es almacenado tres veces y al menos un bloque se almacena en un diferente rack para lograr redundancia.

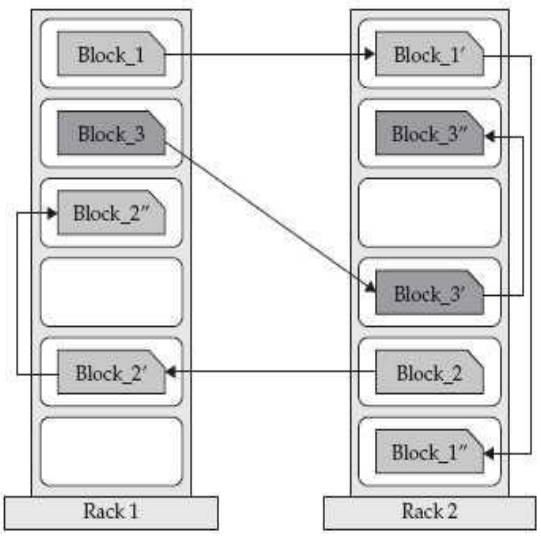


Figura 1: Ejemplo de HDFS(Franco, 2012)

2.2.1.2 *Hadoop MapReduce*

MapReduce es el núcleo de Hadoop. El término MapReduce en realidad se refiere a dos procesos separados que Hadoop ejecuta. El primer proceso map, el cual toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en tuplas (pares de llave/valor). El proceso reduce obtiene la salida de map como datos de entrada y combina las tuplas en un conjunto más pequeño de las mismas.

Una fase intermedia es la denominada Shuffle la cual obtiene las tuplas del proceso map y determina que nodo procesará estos datos dirigiendo la salida a una tarea reduce en específico.

La siguiente figura ejemplifica un flujo de datos en un proceso sencillo de MapReduce.

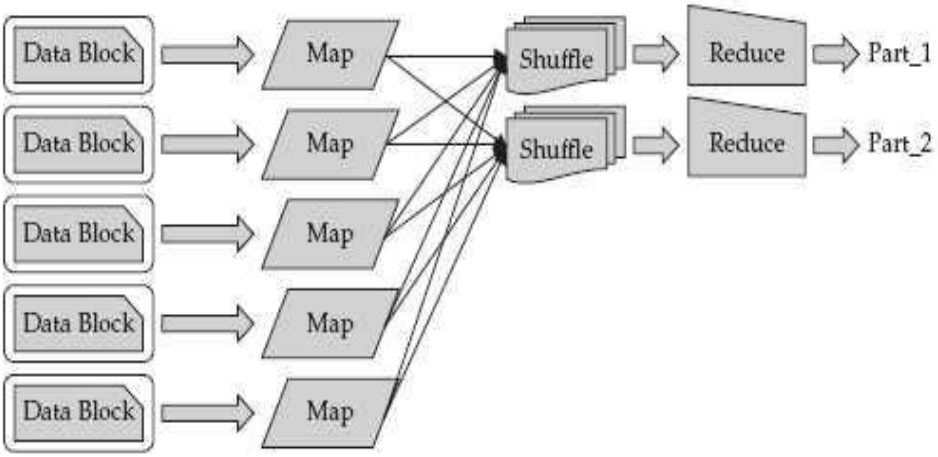


Figura 2 Ejemplo de MapReduce (Franco, 2012)

2.2.1.3 *Hadoop Common Components*

Son un conjunto de librerías que soportan varios subproyectos de Hadoop.

Además de estos tres componentes principales de Hadoop, existen otros proyectos relacionados los cuales son definidos a continuación:

Avro

Avro es un proyecto de Apache que provee servicios de serialización. Cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema está definido dentro del archivo.

Cassandra

Cassandra es una base de datos no relacional distribuida y basada en un modelo de almacenamiento de <clave-valor>, desarrollada en Java.

Permite grandes volúmenes de datos en forma distribuida. Twitter es una de las empresas que utiliza Cassandra dentro de su plataforma.

Chukwa

Diseñado para la colección y análisis a gran escala de "logs". Incluye un toolkit para desplegar los resultados del análisis y monitoreo.

Flume

Dirige los datos de una fuente hacia alguna otra localidad, en este caso hacia el ambiente de Hadoop. Existen tres entidades principales: sources, decorators y sinks. Un source es básicamente cualquier fuente de datos, sink es el destino de una operación en específico y undecorator es una operación dentro del flujo de datos que transforma esa información de alguna manera, como por ejemplo comprimir o descomprimir los datos o alguna otra operación en particular sobre los mismos.

Hbase

Es una base de datos columnar (column-oriented database) que se ejecuta en HDFS. HBase no es una base de datos relacional. Cada tabla contiene filas y columnas como una base de datos relacional. HBase permite que muchos atributos sean agrupados llamándolos *familias de columnas,* de tal manera que los elementos de una familia de columnas son almacenados en un solo conjunto. Eso es distinto a las bases de datos relacionales orientadas a filas, donde todas las columnas de una fila dada son almacenadas en conjunto. Facebook utiliza HBase en su plataforma desde Noviembre del 2010**.**

Hive

Es una infraestructura de datawarehouse que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar a SQL llamado Hive Query Language (HQL), estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el clúster de Hadoop.

El siguiente es un ejemplo en HQL para crear una tabla, cargar datos y obtener información de la tabla utilizando Hive:

```
CREATE TABLE Tweets (from_user STRING, userid BIGINT, tweettext STRING, retweets INT)
COMMENT 'This is the Twitter feed table'
STORED AS SEQUENCEFILE;
LOAD DATA INPATH 'hdfs://node/tweetdata' INTO TABLE TWEETS;
```

```
SELECT from_user, SUM (retweets)
FROM TWEETS
GROUP BY from_user;
```

Jaql

Fue donado por IBM a la comunidad de software libre. Query Language for Javascript Object Notation (JSON) es un lenguaje funcional y declarativo que permite la explotación de datos en formato JSON diseñado para procesar grandes volúmenes de información. Para explotar el paralelismo, Jaql reescribe los queries de alto nivel (cuando es necesario) en queries de "bajo nivel" para distribuirlos como procesos **MapReduce.**

Internamente el motor de Jaql transforma el Query en procesos map y reduce para reducir el tiempo de desarrollo asociado en analizar los datos en Hadoop. Jaql posee de una infraestructura flexible para administrar y analizar datos semiestructurados como XML, archivos CSV, archivos planos, datos relacionales, etc.

Lucene

Es un proyecto de Apache bastante popular para realizar búsquedas sobre textos. Lucene provee de librerías para indexación y búsqueda de texto. Ha sido principalmente utilizado en la implementación de motores de búsqueda (aunque hay que considerar que no tiene funciones de "crawling" ni análisis de documentos HTML ya incorporadas). El concepto a nivel de arquitectura de Lucene es simple, básicamente los documentos (*document*) son divididos en campos de texto (*fields*) y se genera un índice sobre estos campos de texto. La indexación es el componente clave de Lucene, lo que le

permite realizar búsquedas rápidamente independientemente del formato del archivo, ya sean PDF, documentos HTML, etc.

Oozie

Oozie es un proyecto de código abierto que simplifica los flujos de trabajo y la coordinación entre cada uno de los procesos. Permite que el usuario pueda definir acciones y las dependencias entre dichas acciones.

Un flujo de trabajo en Oozie es definido mediante un grafo acíclico llamado *Directed Acyclical Graph (DAG),* y es acíclico puesto que no permite ciclos en el grafo; es decir, solo hay un punto de entrada y de salida y todas las tareas y dependencias parten del punto inicial al punto final sin puntos de retorno. Un ejemplo de un flujo de trabajo en Oozie se representa de la siguiente manera:

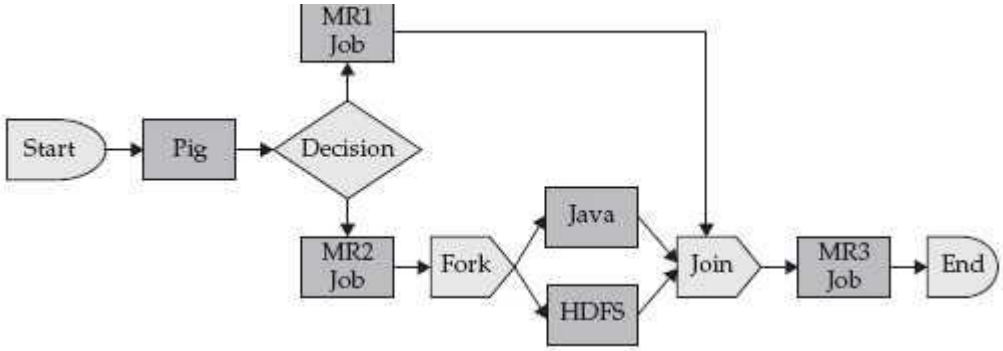


Figura 3 Flujo de Trabajo en Oozie(Franco, 2012)

Pig

Inicialmente desarrollado por Yahoo para permitir a los usuarios de Hadoop enfocarse más en analizar todos los conjuntos de datos y dedicar menos tiempo en construir los programas MapReduce. Tal como su nombre lo indica al igual que cualquier cerdo que come cualquier cosa, el lenguaje *PigLatin* fue diseñado para manejar cualquier tipo de dato y *Pig* es el ambiente de ejecución donde estos programas son ejecutados, de manera muy similar a la relación entre la máquina virtual de Java (JVM) y una aplicación Java.

ZooKeeper

ZooKeeper es otro proyecto de código abierto de Apache que provee de una infraestructura centralizada y de servicios que pueden ser utilizados por aplicaciones para asegurarse de que los procesos a través de un clúster sean serializados o sincronizados.
Internamente en ZooKeeper una aplicación puede crear un archivo que se persiste en memoria en los servidores ZooKeeper llamado *znode.* Este archivo *znode* puede ser actualizado por cualquier nodo en el clúster, y cualquier nodo puede registrar que sea informado de los cambios ocurridos en ese *znode*; es decir, un servidor puede ser configurado para "vigilar" un *znode* en particular. De este modo, las aplicaciones pueden sincronizar sus procesos a través de un clúster distribuido actualizando su estatus en cada *znode*, el cual informará al resto del clúster sobre el estatus correspondiente de algún nodo en específico.

Como podrá observar, más allá de Hadoop, una plataforma de Big Data consiste de todo un ecosistema de proyectos que en conjunto permiten simplificar, administrar, coordinar y analizar grandes volúmenes de información.

2.3 CDR

Un registro detallado de llamadas (CDR) es un registro de datos producida por una central telefónica equipo u otras telecomunicaciones que documenta los detalles de una llamada telefónica que pasa a través de la instalación o el dispositivo. El registro contiene varios atributos de la llamada, como el tiempo, la duración, el estado de finalización, número de la fuente, y el número de destino. Es el automatizado equivalente de las tarjetas de peaje de papel que se han escrito y con fecha por operadores para llamadas de larga distancia en una central telefónica Manual.

2.3.1 Contenido del CDR

Un registro detallado de llamadas contiene metadatos - es decir, datos sobre los datos - que contiene los campos de datos que describen una instancia específica de una transacción de telecomunicaciones, pero no incluye el contenido de esa transacción. A modo de ejemplo, un registro detallado de llamadas describiendo una llamada de teléfono en particular podría incluir los números de teléfono tanto de la vocación y las partes que reciben, la hora de inicio y la duración de la llamada. En la práctica moderna actual, registros detallados de llamadas son mucho más detallados, y contienen atributos como:

- El número de teléfono del abonado que origina la llamada (el que llama, de la parte A).
- El número de teléfono que recibe la llamada (la parte llamada, la parte B).
- La hora de inicio de la llamada (fecha y hora).
- La duración de la llamada.
- El número de teléfono de facturación que se cobra por la llamada.
- La identificación de la central telefónica o equipo escribiendo el registro.
- Un único número de secuencia que identifica el registro.
- Cifras adicionales sobre el número llamado utilizan para direccionar o cobran la llamada.
- La disposición o los resultados de la convocatoria, indicando, por ejemplo, si se conectó la llamada.
- La vía por la que la llamada entró en el intercambio.
- La ruta por la cual la llamada abandonó el intercambio.
- Tipo de llamada (voz, SMS, etc.).
- Cualquier condición de fallo encontró

Cada fabricante de cambio decide qué información se emite en los tickets y la forma en que está formateada. Ejemplos:

- Enviar la marca de tiempo de la final de la llamada en lugar de duración.
- Máquinas-Voice sólo no pueden enviar tipo de llamada.
- Algunos pequeños PBX no envía la persona que llama.

2.4 IBM INFOSPHERE BIGINSIGHTS

InfoSphere BigInsights es una plataforma de software diseñada para ayudar a las empresas a descubrir y analizar ideas de negocios ocultos en grandes volúmenes de una amplia gama de datos de datos que a menudo se ignora o se desecha porque es muy poco práctico o difícil de procesar por los medios tradicionales. Ejemplos de tales datos incluyen registros de registro, haga clic en los arroyos, los datos de medios sociales, feeds de noticias, salida de sensor electrónico, e incluso algunos datos transaccionales. Para ayudar a las empresas obtener valor de dichos datos de manera eficiente, BigInsights incorpora varios proyectos de código abierto (incluyendo Apache Hadoop) y una serie de tecnologías desarrolladas por IBM.

La Figura 4, muestra la plataforma de datos de IBM, que incluye software para el procesamiento de datos en streaming y datos persistentes. BigInsights apoya esta última, mientras que InfoSphere Streams apoya la primera. Los dos se pueden implementar en conjunto para apoyar en tiempo real y análisis de lotes de varias formas de datos en bruto, o que se pueden implementar individualmente para cumplir objetivos específicos de la aplicación. El resto de este artículo se centra en BigInsights.

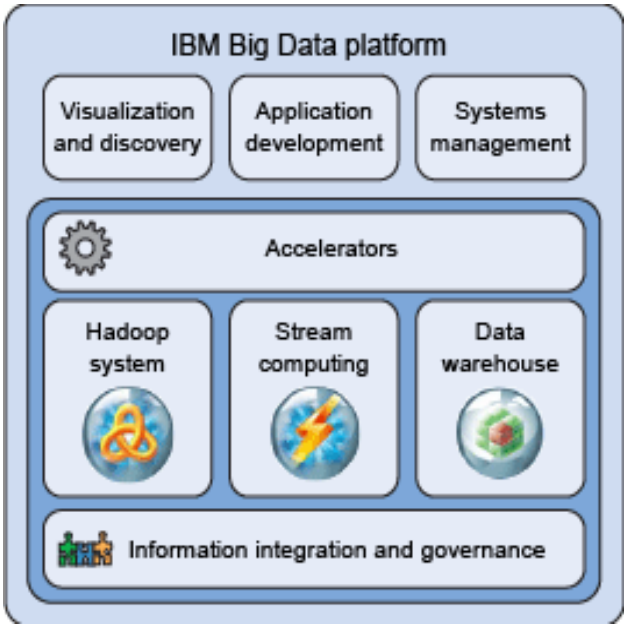


Figura 4 Plataforma de Big Data de IBM (Seeling, Resende, Saraco, & Linder, 2013)

IBM desarrolló BigInsights para ayudar a las empresas a procesar y analizar el aumento del volumen, variedad y velocidad de datos de interés. Considerar la posibilidad de que las industrias esperan que la cantidad de datos digitales aumenten rápidamente en los próximos años. De hecho, una empresa, International Data Corporation, espera que los volúmenes crezcan hasta 44 veces en 2020, en comparación con los niveles de 2009, y la mayoría de los datos estará en formatos no estructurados o semi-estructurados. Como resultado, muchos profesionales de TI anticipan los nuevos desafíos de procesamiento de datos; a menudo utilizan el término "Big Data" (o "big data") para referirse a este tema.

Sin embargo, muchas empresas reconocen que el análisis de grandes volúmenes de datos puede revelar patrones y perspectivas importantes para sus organizaciones. Las áreas de aplicación abarcan muchos campos, incluyendo la retención de clientes, servicio al cliente, inteligencia de mercado, planificación de negocios y operaciones, la investigación científica, la seguridad, y otras áreas. Tales aplicaciones pueden requerir la aplicación de analizar, sistema o datos de registro del sensor; consumidor o el

sentimiento público expresan a través de varios lugares electrónicos; datos de texto rico, incluyendo documentos, correos electrónicos y mensajes; y varias otras fuentes de datos. Por desgracia, el gran esfuerzo que supone para recopilar, procesar, analizar y gestionar estos datos puede parecer desalentador.

BigInsights ofrece un plus en tecnologías analíticas y exploits racimos hardware nada compartida. Se distribuye de forma transparente los datos almacenados en archivos a través de discos conectados a varios nodos en un clúster, dirigir subtareas de aplicaciones para los procesadores que están cerca de los subconjuntos de destino de sus datos. Este enfoque minimiza el tráfico de red y mejora el rendimiento en tiempo de ejecución. Por la tolerancia a fallos, BigInsights replica automáticamente cada porción de sus datos en varios discos en base a parámetros especificados por el administrador. Tal replicación permite BigInsights recuperar automáticamente desde un disco o nodo falla al redirigir el trabajo en otros lugares.

BigInsights no sustituye a un sistema relacional de gestión de base de datos (DBMS) o un almacén de datos tradicional. No está optimizado para consultas interactivas sobre estructuras de datos tabulares, procesamiento analítico en línea (OLAP), o de procesamiento de transacciones en línea (OLTP) aplicaciones. Más bien, es una plataforma que puede aumentar su infraestructura analítica existente, lo que le permite filtrar a través de grandes volúmenes de datos en bruto y combinar los resultados con los datos estructurados almacenados en su DBMS o almacén, si se desea escenarios de integración potenciales serán tratados más adelante.

2.4.1 Ediciones básicas y Enterprise

La plataforma BigInsights componen de una serie de tecnologías de código abierto de IBM y son parte de BigInsights, que está disponible en dos ediciones: Básico y Enterprise. Como se muestra en la Ilustración 6, ambas

ediciones incluyen Apache Hadoop y otro software de código abierto, que se explican con más detalle más adelante.

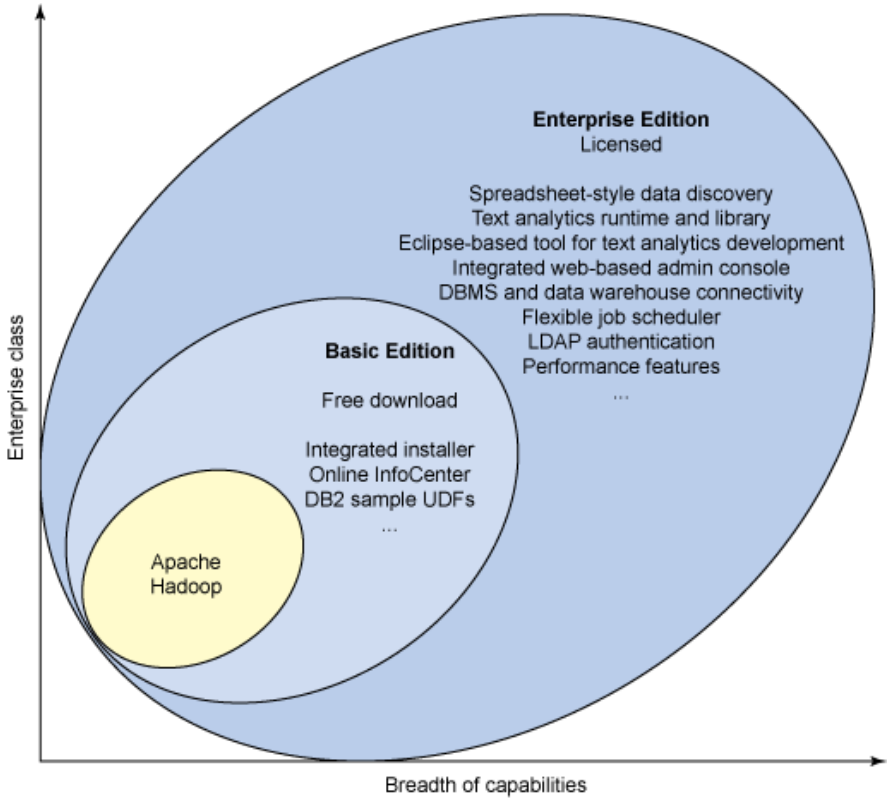


Figura 5 IBM InfoSphere BigInsights 1.2(Seeling, Resende, Saraco, & Linder, 2013)

Basic Edition está disponible para su descarga gratuita y puede gestionar hasta 10 TB de datos. Como tal, es conveniente para los proyectos piloto y trabajos exploratorios. La Enterprise Edition es una oferta basada en honorarios, sin restricciones de licencia sobre la cantidad de datos que se pueden gestionar. Incluye todas las características de la edición básica y ofrece analítica adicional, administrativa, y las capacidades de integración de software. Como tal, Enterprise Edition es adecuado para aplicaciones de producción.

2.4.2 Tecnologías de código abierto

Los proyectos de código abierto incluidos con BigInsights 1,2 y 3 ediciones básicas y Enterprise son:

- Hadoop
- Pig
- Jaql
- Hive
- Tablas HBase
- Canal de flujo
- Lucene
- Avro
- ZooKeeper
- Oozie

Estos proyectos están bien documentados en sitios web de acceso público.

2.4.3 Tecnologías de IBM

Además de software de código abierto, BigInsights incluye una serie de tecnologías desarrolladas por IBM para ayudarle a ser productivo rápidamente. Los ejemplos incluyen un motor de texto de análisis y apoyo de herramientas de desarrollo, una herramienta de exploración de datos para que los analistas de negocio, integración de software empresarial, y varias mejoras de la plataforma para simplificar la administración y ayudar a mejorar el rendimiento en tiempo de ejecución. Echa un vistazo más de cerca.

2.4.3.1 Text Analitics (Análisis y herramientas basados en texto)

Como se mencionó anteriormente, BigInsights está diseñado para ayudar a las empresas a analizar una amplia gama de datos, incluidos los datos que está poco estructurada o no estructurada en gran medida. Varios

tipos de datos de texto se incluyen en esta categoría. De hecho, los documentos financieros, documentos legales, material de marketing, correos electrónicos, blogs, noticias, comunicados de prensa y sitios web de medios sociales contienen datos de texto que las empresas pueden querer procesar y evaluar.

Por esta razón, BigInsights Enterprise Edition incluye un motor de procesamiento de texto y de la biblioteca de anotadores que permiten a los desarrolladores para consultar e identificar temas de interés en los documentos y mensajes. Ejemplos de entidades empresariales que BigInsights pueden extraer de los datos basados en texto incluyen personas, direcciones de correo electrónico, direcciones postales, números de teléfono, direcciones URL, joint ventures, alianzas y otros.

Además, los programadores pueden utilizar el plug-in basado en Eclipse para construir su propia biblioteca de funciones analíticas de texto para BigInsights. Se muestra en la Ilustración 7, el plug-in incluye un generador de expresiones, la tecnología patrón de descubrimiento, un entorno de prueba, y un explorador de resultados para promover el prototipado rápido y el perfeccionamiento de complejas funciones analíticas de textos adaptados a las necesidades específicas de la aplicación.

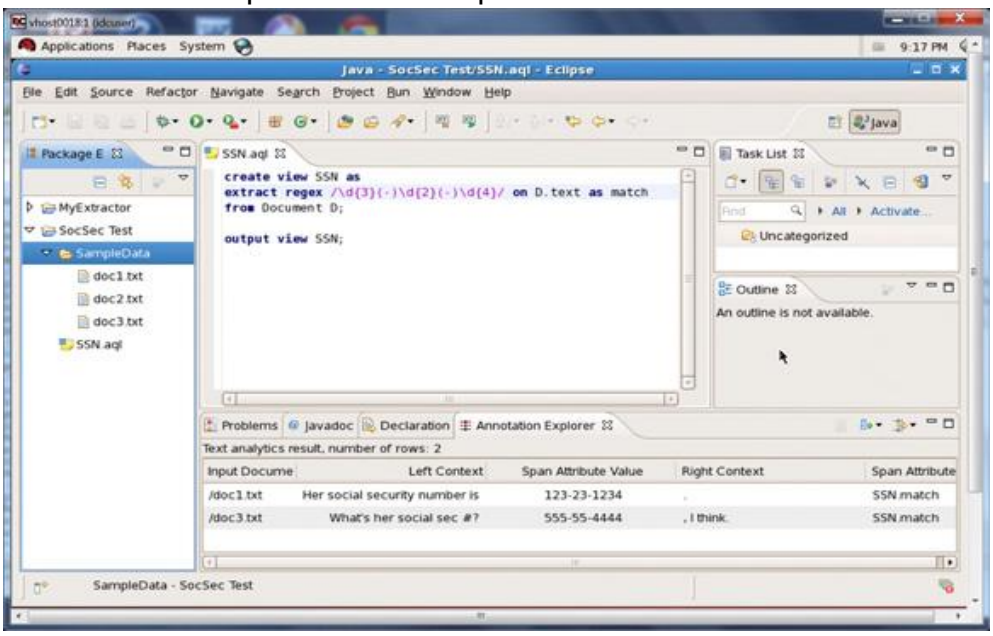


Figura 6 BigInsights - plug-in para Eclipse IDE

En esta Figura, la caja central muestra la sintaxis para crear un anotador "SSN" para identificar los números de seguridad social de Estados Unidos. La definición del anotador pide 3 números de un solo dígito, un guión, 2 números de un solo dígito, un guión y 4 números de un solo dígito. Aunque este ejemplo es muy simple, se ilustra el concepto básico de cómo definir un anotador. Por supuesto, los programadores pueden utilizar esta herramienta para construir, anotadores complejos de calidad de producción, incluidos los anotadores que se basan en los anotadores creados previamente.

El panel inferior ilustra el resultado de una prueba de funcionamiento el uso de estos documentos como entrada. Los resultados muestran las entidades "SSN" que el anotador identificadas, el contexto en el que cada SSN apareció en el documento (es decir, el texto inmediatamente a la izquierda ya la derecha del número de seguro social identificados), y el nombre del documento de origen. Una vez que se completa el anotador (o un conjunto de anotadores), un desarrollador puede exportar el código resultante en la forma de un Operador Gráfico.

2.4.3.2 BigSheets

Para ayudar a los analistas de negocio y los no programadores trabajan con "grandes datos", BigInsights Enterprise Edition proporciona una herramienta de análisis de datos de hoja de cálculo. Lanzado a través de un navegador Web, BigSheets permite a los analistas de negocio para crear *colecciones* de datos para explorar. Para crear una colección, analista especifica la fuente deseada de datos (s), que podría incluir la BigInsights distribuidas sistema de archivos, un sistema de archivos local, o la salida de un rastreo Web. BigSheets Incorpora soporte para formatos de datos populares, como los datos JSON, valores separados por comas (CSV), valores separados por tabuladores (TSV), de datos de carácter delimitado, y otros. Si se desea, los programadores pueden crear plug-ins para procesar formatos de datos adicionales y ejecutar funciones personalizadas.

Cuando un analista ejecuta (o carreras) la definición de la colección, BigSheets genera empleos MapReduce detrás de escena para recuperar y procesar los datos necesarios. Los analistas también pueden revisar y manipular los datos de la colección utilizando funciones y macros incorporadas. Este tipo de trabajo se realiza a través de una interfaz de hoja de cálculo tradicional, como se muestra en la Ilustración 8, que representa una simple fórmula definida por el usuario para poblar una nueva columna con valores derivados de otras columnas de la colección.

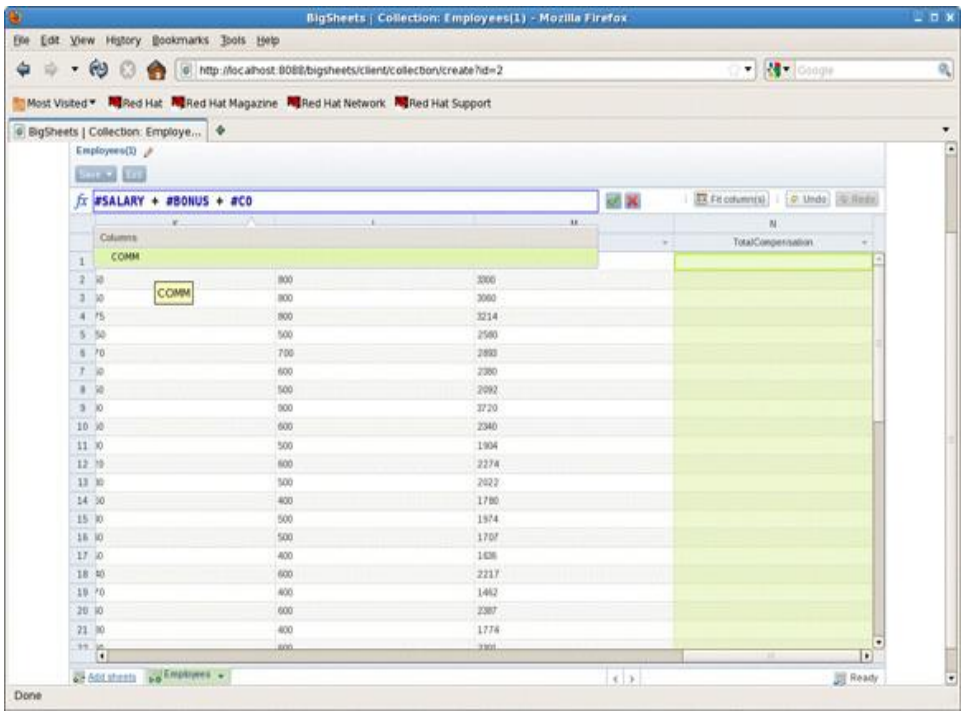


Figura 7 BigSheets

Por último, los analistas pueden utilizar las instalaciones de gráficos en BigSheets visualizar algunos o todos los contenidos de su colección, si se desea. Además, se pueden exportar los datos de recogida en uno de varios formatos populares para su uso por otras aplicaciones. HTML, CSV, y JSON son algunos de los formatos de exportación admitidos.

2.4.3.3 *Instalación integrada y herramientas de administración*

Para ayudar a las empresas a tener un inicio rápido, ediciones BigInsights básicos y Enterprise proporcionan una herramienta basada en

web que instala y configura todo el software de IBM y no IBM apoyado seleccionado por un administrador. Los detalles sobre el progreso de una instalación BigInsights se informan en tiempo real, y un "chequeo" mecanismo integrado verifica e informa automáticamente sobre el éxito de la instalación.

Por el contrario, los que trabajan con las ofertas de código abierto individuo necesitaría para descargar de forma iterativa, configurar y probar cada proyecto de software que querían utilizar. Además, tendría que ser sensible a cualquier software pre-requisitos e incompatibilidades que puedan existir entre los proyectos deseados.

Una vez BigInsights está instalado, los administradores de la empresa pueden trabajar con una consola de administración basada en Web para inspeccionar el estado de su entorno BigInsights en cualquier momento. A través de esta consola, pueden iniciar y detener nodos, investigar el estado de los trabajos de MapReduce, revisar los archivos de registro, evaluar la salud general del sistema, iniciar y detener componentes opcionales, navegar por el sistema de archivos distribuido, y más. La Figura 8 muestra una parte del panel principal de la consola Web.

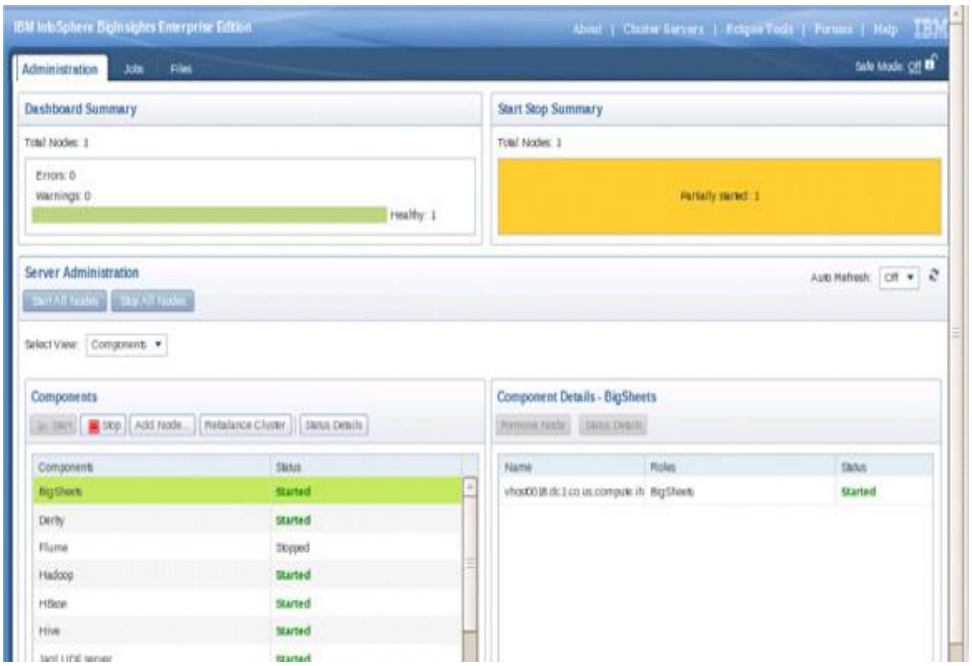


Figura 8 Consola Web BigInsights 1

2.4.3.4 La integración de software para empresas

Muchas organizaciones están preocupadas por la introducción de una nueva plataforma de gestión de la información en su infraestructura de TI existente. Muy comúnmente, los arquitectos de TI se preocupan por la integración de la información manejada por el nuevo sistema con otros datos importantes que ya tienen en su empresa.

Para solucionar este problema, BigInsights Enterprise Edition proporciona a los desarrolladores Jaql con conectividad JDBC para Netezza y DB2 para que puedan transferir datos hacia y desde estas fuentes de manera que explota las capacidades de procesamiento paralelo nativos de esas plataformas. Este apoyo es útil para los desarrolladores BigInsights que quieran unirse a los datos de referencia almacenados en un DBMS relacional con datos gestionados por BigInsights. Para acceder a otras fuentes de datos relacionales, BigInsights proporciona un conector JDBC genérico.

Además, ambos BigInsights básicos y ediciones Enterprise proporcionan funciones definidas por el usuario de ejemplo de DB2 (UDF) que permiten a los programadores de DB2 para iniciar consultas Jaql en BigInsights, se unen a la salida con los datos de DB2, y presentar los resultados a los usuarios de DB2 y aplicaciones. Estas UDF se pueden registrar con un servidor DB2 9.7 para plataformas Linux, UNIX y Windows.

La siguiente figura muestra la conectividad DBMS y datos almacén proporcionada a través BigInsights.

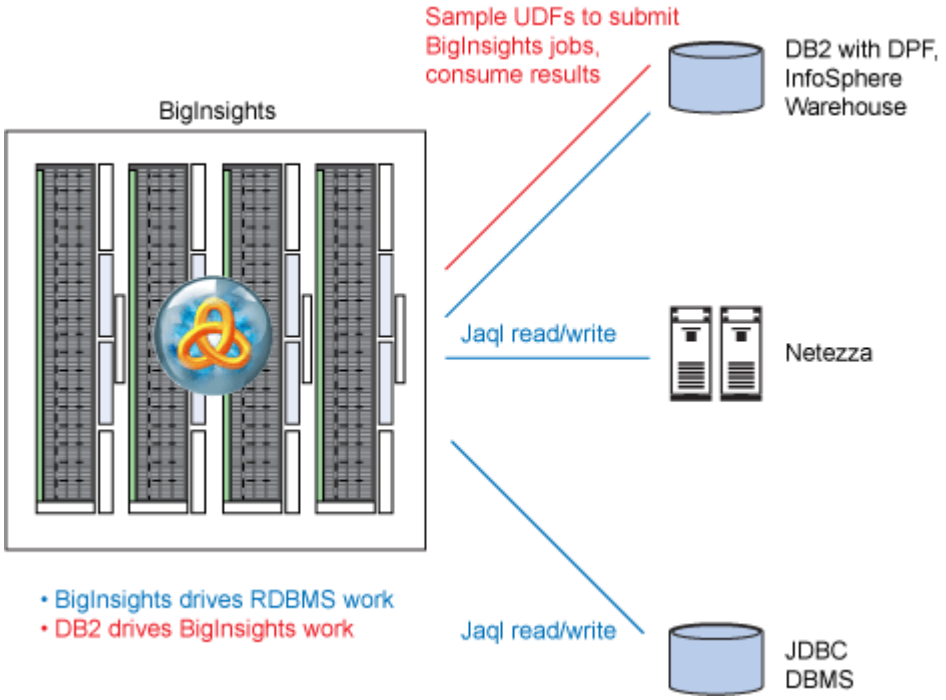


Figura 9 DBMS y conectividad de almacenamiento de datos para BigInsights (Seeling, Resende, Saraco, & Linder, 2013)

2.4.3.5 Mejoras de la plataforma y las características de rendimiento

Mientras BigInsights utiliza tecnologías de código abierto que ofrecen buenos resultados en tiempo de ejecución y los altos niveles de escalabilidad, la Enterprise Edition también emplea el software de IBM específica para mejorar aún más la administración y el rendimiento.

Por ejemplo, BigInsights ofrece un mecanismo de planificación de trabajos opcional para la asignación de recursos sintonía fina entre los puestos de trabajo de larga duración y de corta corriente. Los administradores pueden utilizar una configuración de propiedad para asignar el máximo de recursos para pequeños trabajos para ayudar a asegurar que completen rápidamente. Esta opción de programación de trabajo está disponible, además de Hadoop de primero en entrar / primero en salir (FIFO) y enfoques de programación "justos".

Además, BigInsights proporciona seguridad mejorada mediante el apoyo a la autenticación LDAP para su consola Web. Administradores LDAP y soporte de proxy inverso de ayuda restringen el acceso a los usuarios con la debida autorización.

Mejoras en el rendimiento incluyen el procesamiento eficiente de los datos comprimidos basados en texto a través del uso de la tecnología de compresión basada en LZO IBM. BigInsights también incluye técnicas de ejecución de adaptación para los puestos de trabajo Jaql que pueden ayudar a mejorar el rendimiento en tiempo de ejecución de aplicaciones de destino. Cuando la tecnología MapReduce adaptativa de IBM está encendido (a través de una configuración de propiedad o una opción de Jaql), tareas Mapa comunican a través ZooKeeper para entender el estado global del trabajo. Cuando tiene sentido hacerlo, tareas de mapa pueden tomar de forma dinámica en el trabajo adicional, que puede conducir a un mejor desempeño en tiempo de ejecución para el trabajo en general.

2.5 BIGINSIGHTS EN ARQUITECTURA DE DATOS EMPRESARIALES

Trabajar con grandes datos se está convirtiendo en una parte integral de la estrategia de datos empresariales en muchas empresas. De hecho, un número de organizaciones están buscando implementar una plataforma de software como BigInsights para que puedan gestionar grandes volúmenes de datos desde el momento en que entra en su empresa. Después de almacenar los datos en bruto en BigInsights, las empresas pueden manipular, analizar y resumir los datos a obtener nuevos conocimientos, así como sistemas de alimentación aguas abajo. De esta manera, tanto los datos originales (RAW) y formas modificadas son accesibles para su posterior procesamiento.

Un enfoque implica el uso de despliegue potencial BigInsights como una fuente para un almacén de datos. BigInsights pueden tamizar a través de

grandes volúmenes de datos no estructurados o semi-estructurados, captura de información relevante que puede aumentar los datos corporativos existentes en un almacén. La ilustración 11 muestra un escenario de este tipo, que ofrece a las empresas la posibilidad de ampliar su cobertura analítica sin crear una carga excesiva para sus sistemas existentes.

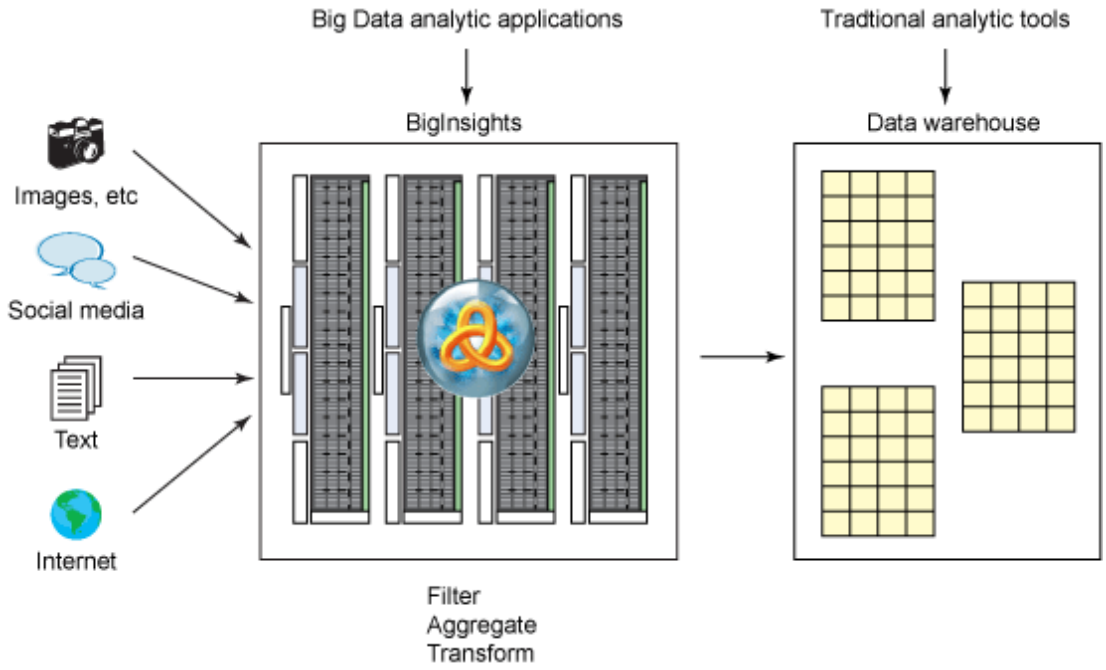


Figura 10 BigInsights: Filtra y resume grandes datos para el DWH(Seeling, Resende, Saraco, & Linder, 2013)

Otro enfoque implica el uso de BigInsights como un archivo Query listo para un almacén de datos. Con este enfoque, los datos de acceso frecuente se pueden mantener en el almacén mientras que la información "frío" o no actualizados se pueda descargar a BigInsights. Esto permite a las empresas gestionar el tamaño de sus plataformas de gestión de datos existentes, mientras que el servicio de las necesidades bien establecidas de sus aplicaciones existentes. Offloading raramente consulta de datos a BigInsights permite que los datos permanezcan accesibles para aplicaciones que pueden tener una necesidad ocasional o impredecible de trabajar con él.

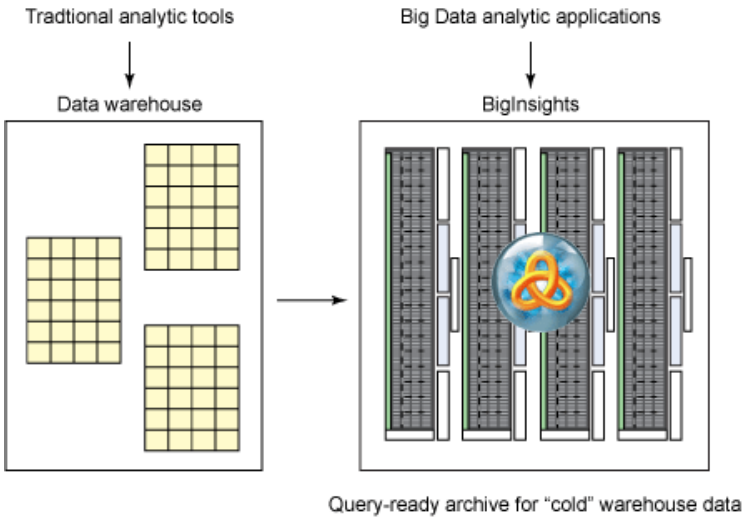


Figura 11 BigInsights actúa como un archivo Query listo para un DWH(Seeling, Resende, Saraco, & Linder, 2013)

2.6 IBM INFOSPHERE INFORMATION MANAGEMENT – DATASTAGE

IBM InfoSphere DataStage integra datos de múltiples sistemas utilizando una estructura paralela de elevado rendimiento y da soporte a la gestión ampliada de metadatos y la conectividad de la empresa. Esta plataforma escalable proporciona una integración más flexible de todos los tipos de datos, incluidos big data inactivos (basados en Hadoop) o en ejecución (basados en secuencias), en plataformas de mainframe y distribuidas.

InfoSphere DataStage incluye estas características y beneficios:

- La plataforma ETL potente y escalable: admite la recopilación, integración y transformación de grandes volúmenes de datos con estructuras de datos tanto simples como complejos.
- El soporte para big data y Hadoop permite el acceso directo a big data en un sistema de archivos distribuido.
- Integración de datos prácticamente en tiempo real y conectividad entre aplicaciones y orígenes de datos.

- La gestión de cargas de trabajo y reglas de negocio, optimiza la utilización de hardware y prioriza las tareas más importantes.
- La facilidad de uso, amplía la velocidad, la flexibilidad y la efectividad para crear, desplegar, actualizar y gestionar la infraestructura de integración de datos.

CAPÌTULO 3

3.1 CREACIÓN, CONFIGURACIÓN E INSTALACIÓN DEL CLÚSTER.

3.1.1 Especificación de los servidores del clúster

Tabla 1 - Hardware del Clúster de Big Data

O	NOMBRE	IP	CPU	RAM	DISCO	LAN	ROL	SISTEMA OPERATIVO
1	Softnamenode	192.168.10.166	8 Cores,4 Sockets per core	64Gb	1,1 TB – Raid 1+0	1 Gigabyte	NameNode	Red Hat Enterprise Linux 6.5
2	Softsecnamenode	192.168.10.167	8 Cores,4 Sockets per core	64Gb	1,1 TB – Raid 1+0	1 Gigabyte	Secondary NameNode	Red Hat Enterprise Linux 6.5
3	Softdatanode1	192.168.10.163	8 Cores,4 Sockets per core	64Gb	4 TB – Raid 1+0	1 Gigabyte	Data Node	Red Hat Enterprise Linux 6.5
4	Softdatanode2	192.168.10.164	8 Cores,4 Sockets per core	64Gb	4 TB – Raid 1+0	1 Gigabyte	Data Node	Red Hat Enterprise Linux 6.5
5	Softdatanode3	192.168.10.165	8 Cores,4 Sockets per core	64Gb	4 TB – Raid 1+0	1 Gigabyte	Data Node	Red Hat Enterprise Linux 6.5

3.1.2 Creación del ambiente del clúster

- Revise la lista de discos disponibles en el clúster. Se utilizan los nombres de las particiones de disco cuando se especifica el caché y directorios de datos para el sistema de archivos distribuido.

  Df -h

- Asegúrese de que existe suficiente espacio en disco para los siguientes directorios necesarios.

Tabla 2 - Especificaciones de Particiones

Directory	Available disk space
/	10 GB
/tmp	50 GB
/$BIGINSIGHTS_HOME The default directory for this variable is /opt/ibm.	15 GB
/$BIGINSIGHTS_VAR The default directory for this variable is /var/ibm.	5 GB
/home/$USER_HOME The default directory for this variable is /home/biadmin.	5 GB

- Compruebe que todos los dispositivos tienen un identificador único universal (UUID) y que los dispositivos se asignan al punto de montaje.
- Visualice los UUID asignados actualmente para todos los dispositivos en el clúster.

```
sudo blkid
```

Figura 12 Aplicación de comando SUDO

La salida muestra todos los dispositivos y su UUID. En el siguiente ejemplo, tres discos están listados: /dev/sda3, /dev/sda1 y /dev/sda2.

```
/dev/sda3: UUID="1632fdf8-2283-4771-9fdd-664964ee7fcf" TYPE="ext3"
/dev/sda1: UUID="8ed83d7a-4e5f-44a1-8448-533da7109312" TYPE="ext3"
/dev/sda2: UUID="59f180e3-931f-4b50-aa94-4b3cb0ab2c0a" TYPE="swap"
```

Figura 13 Listado de Dispositivos

- Actualizar ***/etc/fstab*** para codificarlas referencias cartográficas de dispositivos al punto de montaje. Estas referencias aseguran que la asignación de dispositivos no cambia si un dispositivo deja de estar disponible o deja de funcionar. El punto de montaje debe existir antes de crear referencias cartográficas.

  Importante: Antes de editar/etc /fstab, guarde una copia del archivo original.

```
#UUID=<device-UID> </path/to/mount/point> <file-system-type> <options> <dump> <pass>

#/dev/sda3
UUID=1632fdf8-2283-4771-9fdd-664964ee7fcf /      ext3  defaults  1 1
#/dev/sda1
UUID=8ed83d7a-4e5f-44a1-8448-533da7109312 /boot  ext3  defaults  1 2
#/dev/sda2
UUID=59f180e3-931f-4b50-aa94-4b3cb0ab2c0a swap   swap  defaults  0 0
```

Figura 14 Listado de discos con fstab

- En el ejemplo anterior, tanto/dev/sda3, que es el sistema de ficheros raíz, y /dev/sda1se incluyen en el vertedero de copia de seguridad, como se indica por el número entero que aparece primero. El segundo número entero determina el orden en el que se comprueban los sistemas de archivos. En el ejemplo anterior, /dev/sda3se comprueba primero, /dev/sda1se comprueba en segundo lugar, y /dev/sda2no está marcada.
- Cree el usuario biadmin y su grupo
- En cada nodo del clúster, como usuario root, cree el grupo Biadmin y luego agregar el usuario Biadmin a ella.
- Agregue el grupo Biadmin.
  Groupadd –g biadmin
- Agregue el usuario Biadmin al grupo Biadmin
  Useradd –g biadmin –u 123 biadmin

  Un usuario predeterminado se crea como el propietario de cada componente

  Nota: Si los usuarios se gestionan de forma centralizada a través de LDAP, cada usuario del servicio tiene que ser creado en LDAP antes de la instalación.
- Establezca la contraseña para el usuario Biadmin.
  Passwd biadmin
- En el nodo maestro, agregue el usuario Biadmin al grupo sudoers.
- Editar el archivo sudoers
  Sudo visudo –f /etc/sudoers
- Quitar el comentario de la siguiente línea
  Default requiretty
- Localice la siguiente línea
  # %wheel ALL= (ALL) NOPASSWD: ALL

- reemplazar esa línea con una de las líneas siguientes, dependiendo de qué tipo de acceso se requiere.

```
##Permits users in the biadmin group to run all commands without
##supplying a password
biadmin ALL=(ALL) NOPASSWD:ALL
```

Figura 15 Reemplazo Defaults requieretty

- Crear directorios para los archivos de datos y archivos de caché del sistema de archivos distribuido. Estos directorios deben ser propiedad de Biadmin: Biadmin.
  Mkdir /disk_name/directory

  Por ejemplo, existen los siguientes directorios en un servidor con puntos de montajeDISK1travésdisk10.
- Configura la red
- En el directorio/etc, editar el archivo hosts para que incluya la dirección IP, el nombre de dominio completo, y nombre corto de cada host del clúster, separados por espacios. Debe editar este archivo en cada servidor del clúster.
  El formato es: IP_address domain_name short_name. Por ejemplo
  127.0.0.1 Localhost.localdomain localhost
  123.123.123.123 server_name.server_domain.com server_name
- Si el clúster incluye nodos que sólo utilizan las redes privadas, a continuación, debe configurar una puerta de enlace predeterminada a un host que tenga acceso al nodo de administración, que debe residir en una red pública.

  Importante: siempre debe utilizar un nombre de host público para el nodo de administración.
- En todos los nodos privadas en el clúster, edite el archivo /etc/sysconfig/network-scripts/ifcfg-eth0y añada la dirección IP privada del nodo de administración. Este archivo contiene la configuración de red privada para el controlador de interfaz de red (NIC).

GATEWAY=magnament_node_IP

management_node_IP es la IP privada del nodo de administración, tales como192.0.2.21.

- Guarde los cambios y reinicie su red.
  Service network restart

- Revise las tablas de enrutamiento para asegurar que su puerta de entrada está activada.
  Sudo route -n

  Usted debe ver la puerta de entrada que agregó que la última línea de la tabla de enrutamiento IP del núcleo. Instala InfoSphereBigInsights utilizando el nombre de host público para el nodo de administración, y luego utiliza los nombres de host públicos o privados para otros nodos en el clúster.

- Configurar SSH sin contraseña entre cada nodo y el nodo maestro, entre el nodo maestro y en sí, y tanto para el usuario y Biadmin usuario root. Por otra parte, el programa de instalación puede configurar SSH sin contraseña si no desea realizar esta tarea manualmente.

  Importante: Debe configurar SSH sin contraseña para el usuario root cuando se utiliza GPFS™como sistema de archivos distribuido. Si no configura SSH sin contraseña para el usuario root puede hacer GPFS inoperable.
- En cada nodo del clúster, ejecute el comando siguiente ya que tanto el usuario Biadmin y el usuario root. Seleccione la ubicación de archivos por defecto y dejar la contraseña en blanco.

  Ssh-keygen -t rs

- En el nodo maestro, ejecute el comando siguiente ya que tanto el usuario Biadmin y el usuario root para cada nodo, y después de cada nodo de nuevo al maestro.

  Ssh-copy-id -i ~/.ssh/id_rsa.pub user@server_name

- Asegúrese de que puede iniciar sesión en el servidor remoto sin contraseña.

  Ssh biadmin@server_name.com

- Ejecute los siguientes comandos en la sucesión a desactivar los cortafuegos en todos los nodos del clúster.

  Importante: Asegúrese de que vuelva a activar el cortafuegos en todos los nodos del clúster después de instalar InfoSphereBigInsights.

  Service iptables save
  Service iptables stop
  Chkconfig iptables off

- En todos los servidores del clúster, deshabilitar IPv6.
- Desde la línea de comandos, escriba ifconfig para comprobar siIPv6está funcionando. En la salida, una entrada parainet6indicaque IPv6se está ejecutando.
- Ejecute el comando de la siguiente tabla que se aplica a su sistema operativo paradeshabilitarIPv6.

Tabla 3: Comandos para deshabilitar IPV6.

Sistema Operativo	**Comando**
Red Hat Enterprise Linux 5.6, 6.2, 6.3, y 6.4	Añada al archivo modprobe.d. vi /etc/modprobe.d/disable-ipv6.conf install ipv6 /bin/true Edite /etc/sysconfig/network y añada las siguientes líneas. vi /etc/sysconfig/network NETWORKING=yes NETWORKING_IPV6=no Escriba los cambios en este archivo /etc/sysctl.conf. echo "net.ipv6.conf.all.disable_ ipv6 = 1" >> /etc/sysctl.conf Reinicié su máquina. reboot Verifique que IPV6 está deshabilitado. lsmod | grep ipv6 IPV6 está deshabilitado no retornará nada.

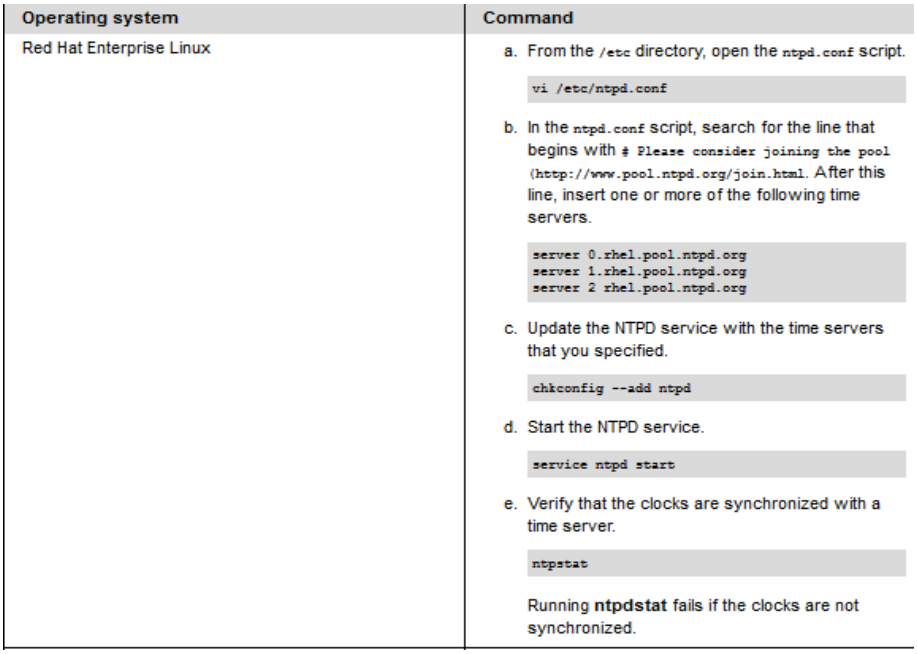
- Asegurarse que las propiedades de ulimit para su sistema operativo se configuran.
- En el directorio/etc / security, abra el archivo limits.conf.
- Asegurarse de que el nofile y propiedad nproc contienen los siguientes valores o mayor. El parámetro no file establece el número máximo de archivos que pueden estar abiertos, y la propiedad nproc

establece el número máximo de procesos que se pueden ejecutar. Los siguientes valores son los valores mínimos que se requieren.

Nofile-16384

Nproc–10240

- Sincroniza los relojes de todos los servidores del clúster mediante una fuente interna o externa Network Time Protocol (NTP / NTPD). El programa de instalación de InfoSphereBigInsights sincroniza los demás relojes de los servidores con el servidor principal durante la instalación. Debe habilitar el servicioNTP /NTPD en el nodo de gestión y permitir a los clientes la sincronización con el nodo maestro.



Operating system	Command

Red Hat Enterprise Linux	a. From the `/etc` directory, open the `ntpd.conf` script.
	`vi /etc/ntpd.conf`
	b. In the `ntpd.conf` script, search for the line that begins with `# Please consider joining the pool (http://www.pool.ntpd.org/join.html.` After this line, insert one or more of the following time servers.
	`server 0.rhel.pool.ntpd.org`
	`server 1.rhel.pool.ntpd.org`
	`server 2 rhel.pool.ntpd.org`
	c. Update the NTPD service with the time servers that you specified.
	`chkconfig --add ntpd`
	d. Start the NTPD service.
	`service ntpd start`
	e. Verify that the clocks are synchronized with a time server.
	`ntpstat`
	Running **ntpdstat** fails if the clocks are not synchronized.

Figura 16 Sincronización de Relojes (knowledgecenter, IBM;, 2012)

- Asegúrese de que el intérprete de comandos de la ID de usuario del administrador es bash.
- Diríjase al directorio */etc*
- Corra el siguiente comando para mostrar el intérprete por defecto para el usuario de administración

  Grep biadmin /etc/passwd

La información mostrada por este comando debe ser similar al siguiente ejemplo.

Biadmin: x: 10539:10539:: /Home/biadmin:/bin/bash

- Si el valor para el ID del usuario administrador (por defecto *biadmin*) no es */bin/bash*, abra el archivo *passwd* en el directorio */etc* y cambien el valor.
- Comprobar la disponibilidad del puerto y resolver nombres de host.
- Asegurarse que todos los puertos necesarios estén disponibles. El siguiente comando muestra el estado de todos los puertos en el sistema, su estado actual y el ID del proceso que se ejecuta encada puerto.

  Número de puerto8300debe estar disponible para que el programa de instalación se ejecute. Para obtener una lista de los componentes necesarios y sus puertos

  Netstat –ap more
- Asegurarse de que los nombres de host para todos los nodos del clúster se resuelven. Los nombres de host deben estar configurados para las mismas direcciones IP que los servidores reales, porque InfoSphereBigInsights no soporta direcciones IP dinámicas. Puede resolverlos nombres de host mediante el uso deservidores DNS, o asegurándose de que los nombres de host se asignan correctamente en los archivos / etc/ hosts en todos los nodos del clúster.
- En el archivo */etc /hosts*, asegúrese de que *local host* se asigna a la dirección *127.0.0.1 loopback*, como se muestra en el siguiente ejemplo.
- Actualizar el firmware para sus controladores de disco. Si los discos no se usan durante un período prolongado de tiempo, pueden entrar en el modo de suspensión. Este comportamiento podría ser percibido como un retraso en el programa de instalación de InfoSphereBigInsights o procesos relacionados tratan de acceder a los discos. Los discos pueden requerir un tiempo de respuesta más largo debido a que los discos comienzan sólo cuando accede.

- En el nodo donde va a ejecutar el programa de instalación, verificar o instalar el Linux Expect package.
- Verify that the Linux Expect package is installed.
  Rpm –qa grep expect
- Si el paquete no está instalado, a continuación, ejecute el comando siguiente para instalarlo.
  Yum install expect

3.1.3 Configuración del ambiente en los nodos del clúster

3.1.3.1 *Verificación de Pre Requisitos*

Revisión del paquete EXPECT

Ejecutando el comando *rpm –qa | grep expect* revisar si el paquete *expect* está instalado

Rpm –qa grep expect

Revisión del paquete NC

Ejecutando el comando *rpm –qa | grep nc* revisar si el paquete *nc* está instalado

Rpm –qa grep cn-1

Verificación del sistema SE Linux deshabilitado

Ejecutando el comando *vim /etc/selinux/config* revisar si **SELINUX** está en *disabled*

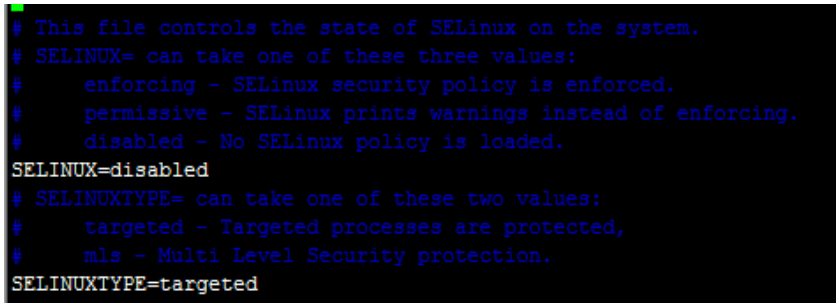


Figura 17 Verificación de Pre-requisitos

Sincronización con el protocolo NTP

Ejecutando el comando *ntpstat* revisar que esté en estado "*synchronised*"

ntpstat

Propiedades de ulimit

Ejecutando el comando *vim /etc/security/limits.conf, verificar* que los parámetros nofile, nproc, hard y soft estén con los usuarios y los valores configurados

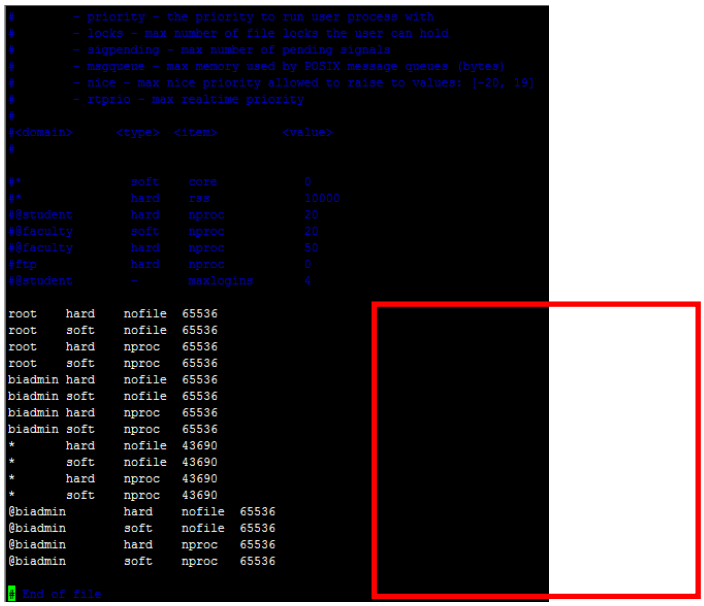


Figura 18 Propiedades de Ulimit (Salazar, Torres)

IPV6 deshabilitada

- Se modifica el archivo:*/etc/modprobe.conf*
- Se añade la siguiente línea:

  Install ipv6 /bin/true
- Se modifica el archivo con el siguiente comando:

  vim /etc/sysconfig/network
- Se añade las siguientes líneas:

  NETWORKING=yes

  NETWORKING_IPV6=no
- Se reinicia el equipo.

- Se verifica que IPv6 está desactivado con:

  lsmod | grep ipv6

Firewall deshabilitado

- Se ejecuta los siguientes comandos:

  Service iptables save

  Service iptables stop

  Chkconfig iptables off

Configuración del archivo /etc/hosts

Verificar en el archivo */etc/hosts* que se encuentren las ips y los nombres largos y alias de los nodos del clúster incluido si mismo

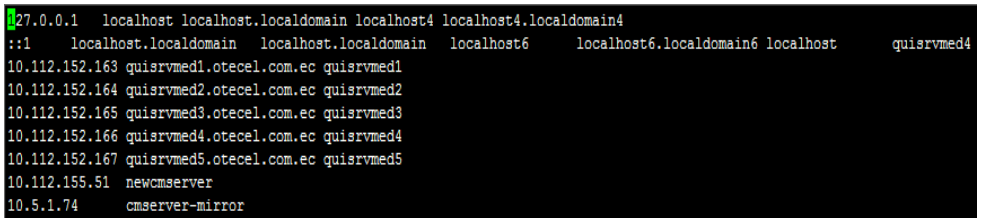


Figura 19 Configuración de Archivo etc/host

Creación del usuario de administración biadmin

Verificar si el usuario biadmin está creado, caso contrario utilizar los siguientes comandos para crear el grupo, el usuario y establecer un password

groupadd -g 168 biadmin

useradd -g biadmin -u 168 biadmin

Passwd biadmin

Permisos de ejecución como root en "visudo"

Verificar con el comando *visudo* la línea: *%biadmin ALL= (ALL) NOPASSWD: ALL,* esta línea debe estar situada debajo de la siguiente línea:

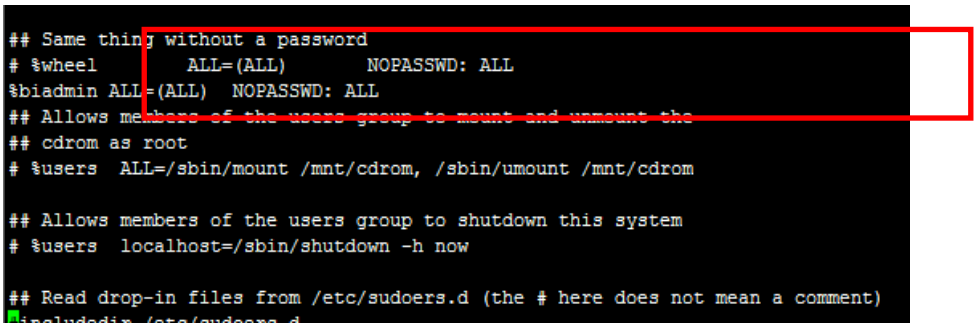


Figura 20 Permisos de Ejecución como ROOT

3.1.4 Instalación de IBM InfoSphere BigInsights Standard Edition

3.1.4.1 *Copia del instalador del producto IBM InfoSphere BigInsights Standard Edition 3.0.0.1*

Ejecutar la Shell de instalación “start.sh” en el directorio donde se descomprimió el instalador de IBM BigInsights Standard Edition V 3.0.0.1, ejecutar la instalación con el usuario “biadmin” caso contrario no funcionará con ningún otro usuario inclusive con el usuario “root”

Inicio de instalación

En este paso se muestra el asistente de instalación de IBM InfoSphere BigInsights Standard Edition V 2.1.1, presionar el botón “Next”

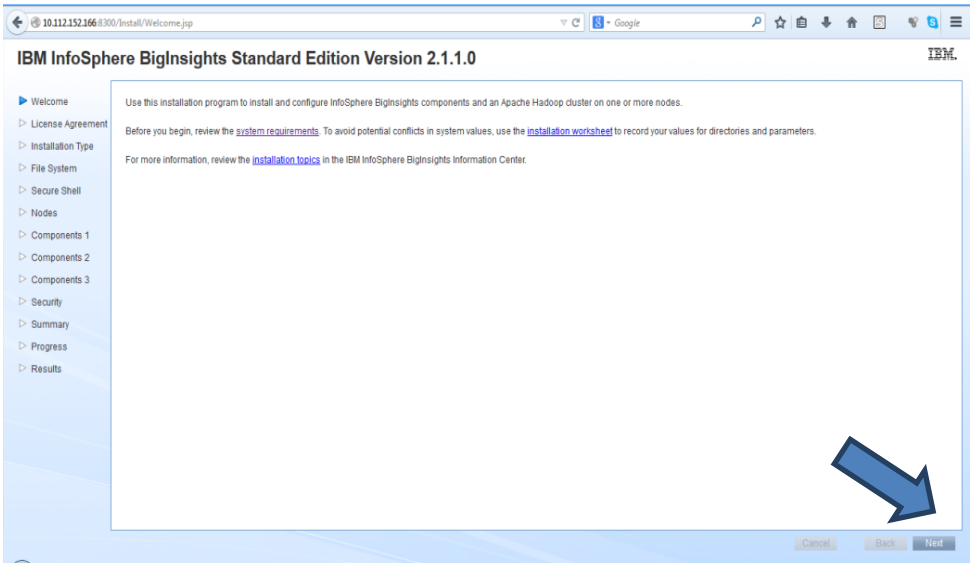


Figura 21 Asistente de instalación de IBM InfoSphere BigInsights Standard Edition

Aceptación del acuerdo de licenciamiento

En este paso se debe leer y realizar clic en “I accept the terms in the license agreement”, de ahí presionar en “Next”.

Selección del tipo de instalación

Para este ambiente seleccionar:

“Clúster Installation”, para poder crear el clúster de Big Data - Hadoop,

Dar clic en “Create a response file without completing an installation”, esto generará un archive de log de instalación para reiniciar en pasos anteriores a la instalación final si existiese necesidad.

Presionar en “Next”.

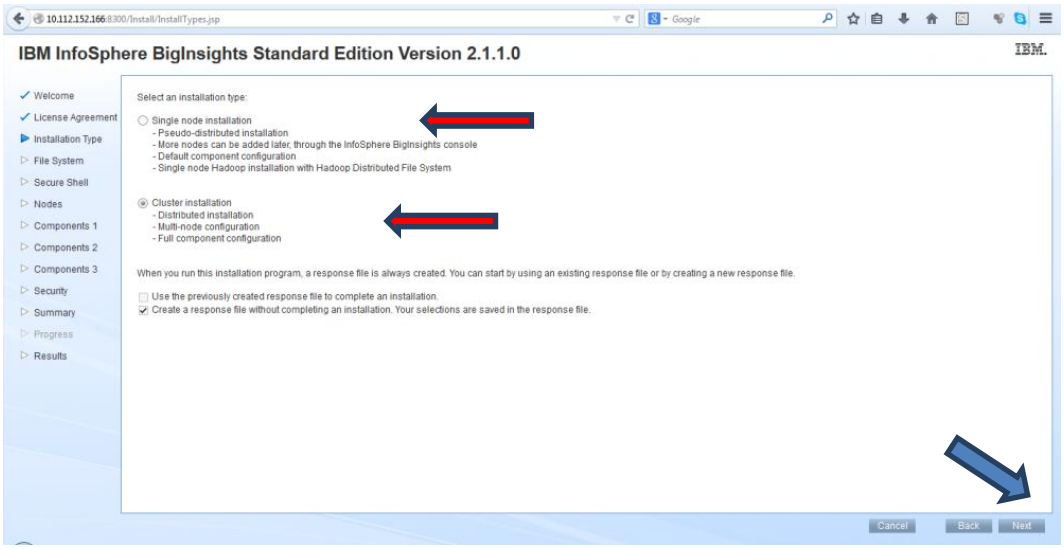


Figura 22 Selección Tipo de Instalación

Selección del tipo de clúster para el sistema

- Para el clúster seleccionar “Install Hadoop Distributed File System (HDFS)
- Dar click en “Overwrite existing files and directories”.

  Presionar en la pestaña “MapReduce general settings”
- Verificar que estén llenos los campos

  Presionar en “Next”.

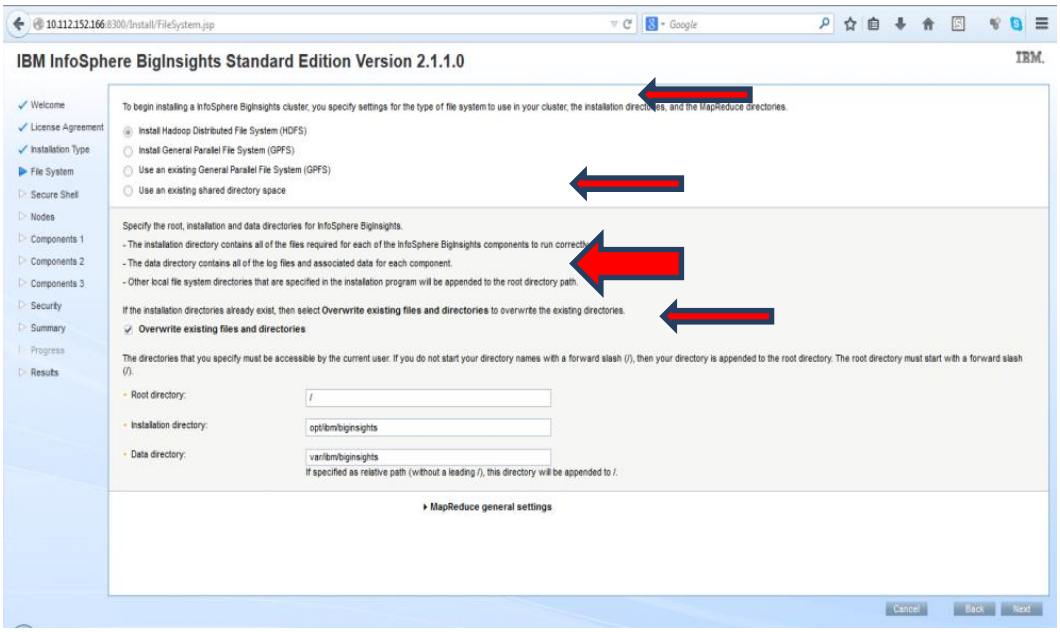


Figura 23 Selección Tipo de Clúster

3.1.5 Configuración SSH Passwordless (Relaciones de Confianza)

Los nodos del clúster deben tener acceso directo entre ellos tanto por el usuario “root” y para el usuario “biadmin”

- Seleccionar “Use the root user”. Select this option if the following statement is true.
- Especificar el password del usuario root.
- Especificar el usuario de administración de BigInsights “biadmin”.
- Ingresar la contraseña y confirmar la misma.
- Ingresar el grupo de administración del grupo de BigInsights “168”.
- Presionar en “Next”.

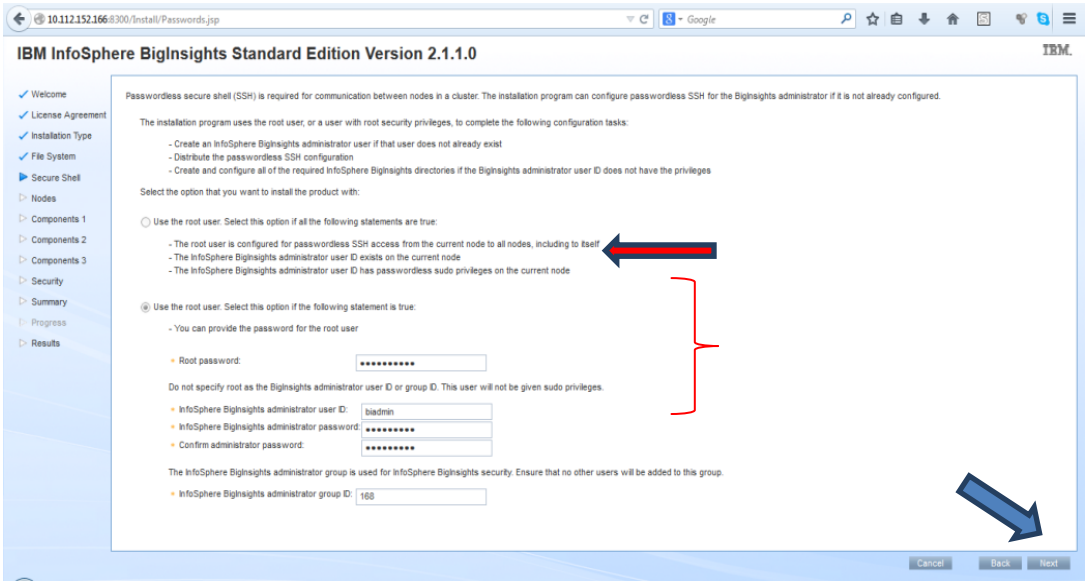


Figura 24 Configuración de SSH

Ingreso de Nodos (Servidores) al clúster

Ingresar los servidores del clúster siguiendo los siguientes pasos:

- Dar clic en Add Node por cada servidor que desean ingresar al clúster:
- Ingresar el nombre del servidor incluido dominio.
- Especificar el password del usuario root del servidor.
- Presionar OK.
- Presionar en “Next”.

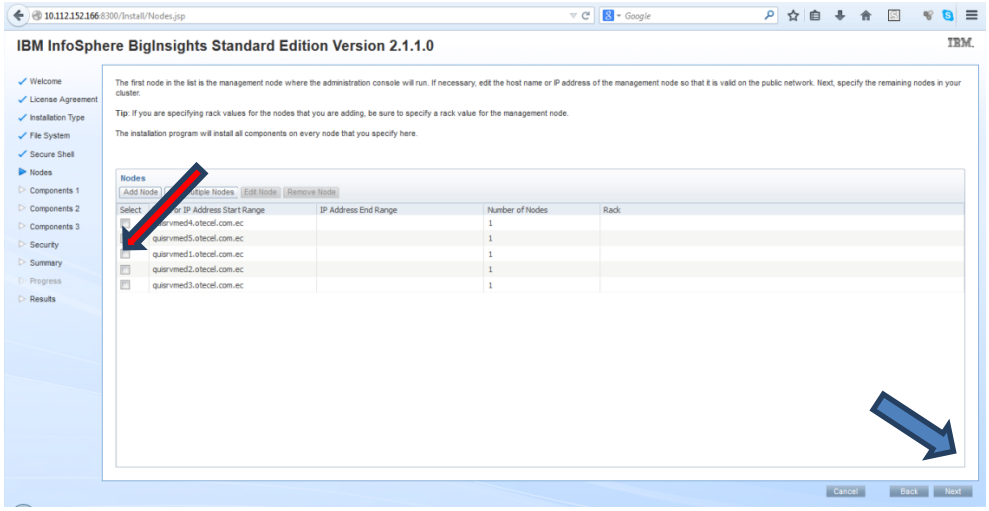


Figura 25 Ingreso de Nodos al Clúster

3.1.6 Configuración de los nodos y puertos

Revisar las configuraciones que se encuentren llenadas, todos los campos que contengan la etiqueta Nodo deben de estar llenado con el servidor “Name Node” principal en este caso “quisrvmed4.TELCOS.com.ec” Presionar en “Next”.

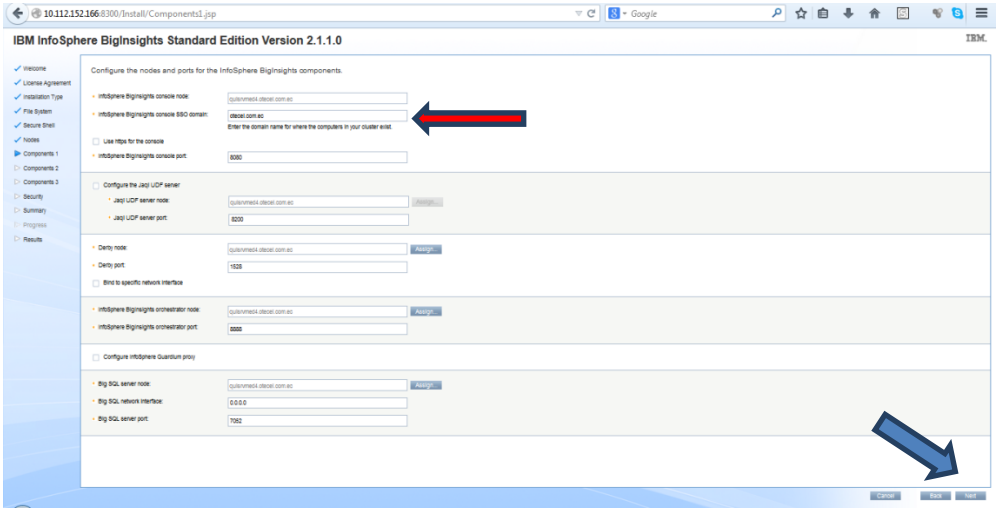


Figura 26 Configuración de Nodos

3.1.7 Especificación del Secondary Name Node, DataNode y TaskTracker

Pasos para definir el Secondary Name Node

- Presionar Assign
- Ingresar el nombre del servidor asignado para Name Node, en este caso “quisrvmed5.TELCOS.com.ec”

Pasos pare definir DataNode y TaskTracker

- Dar click en la opción “Use all nodes except the Name Node, Job Tracker and Secondary NameNode nodes”.
- Presionar en “Next”.

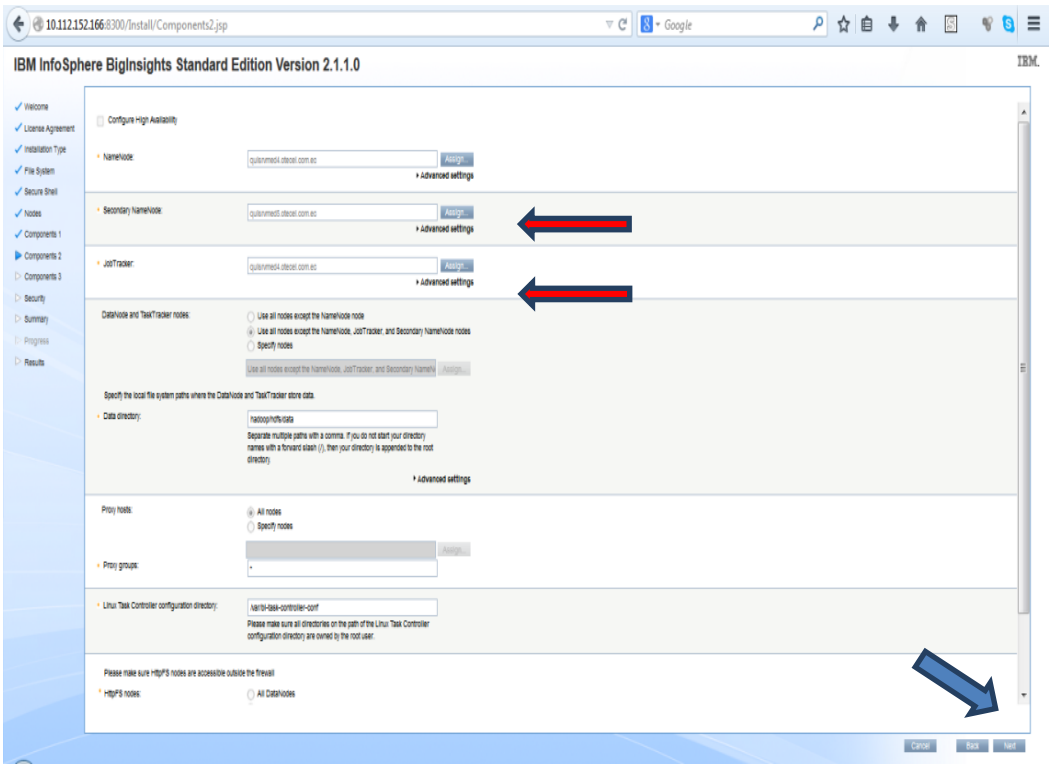


Figura 27 Especificación de Nodos

3.1.8 Configuración de HBase y ZooKeeper

Revisar todos los campos que estén llenos y presionar en “Next”

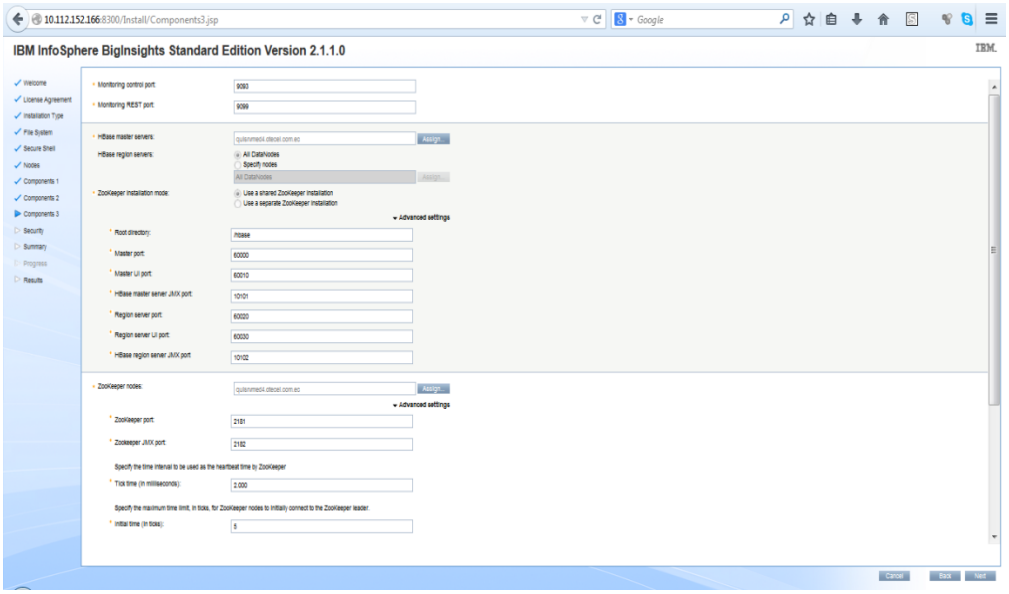


Figura 28 Configuración de HBase y ZooKeeper

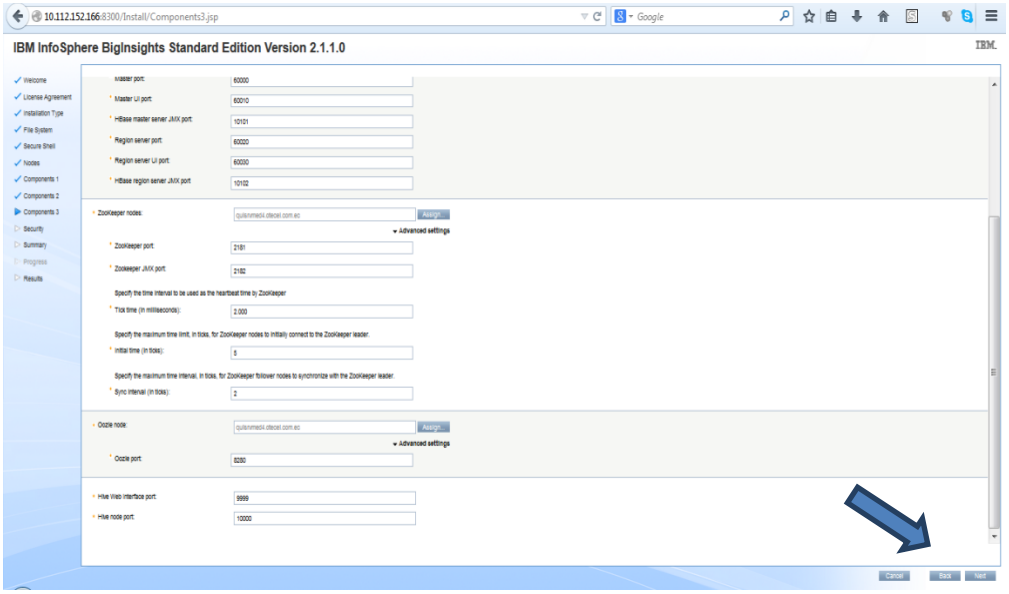


Figura 29 Configuración de HBase y ZooKeeper

Especificación del password para el usuario "biadmin"

En el directorio donde fue descomprimido el instalador, en este caso "/home/biadmin/biginsights-standard-linux64_b20140123_1205" dirigirse al directorio "/home/biadmin/biginsights-standard linux64_b20140123_1205/artifacts/security/flatfile".

- Editar el archivo "biginsights_users.xml"
- En la etiqueta user name="biadmin" cambiar el password al password que especifico anteriormente.
- Cerrar el archivo guardando los cambios.
- Regresar a la consola y presionar "Next".

Figura 30 Especificación de PSW para Biadmin

Revisión de las configuraciones y creación del archivo "Response File"

Revisar las configuraciones tanto del ambiente en general, nodos y Roles y presionar "Create Response File"

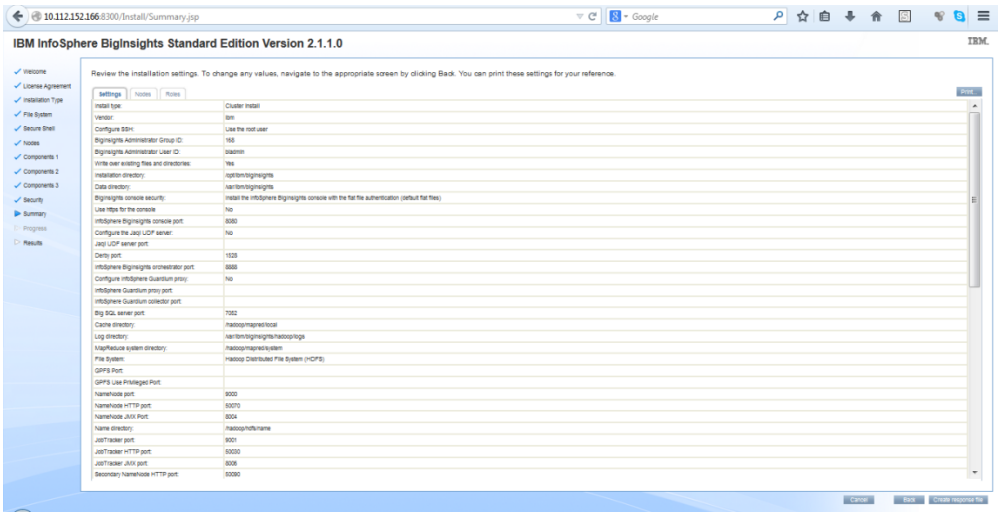


Figura 31 Verificación de las Configuraciones

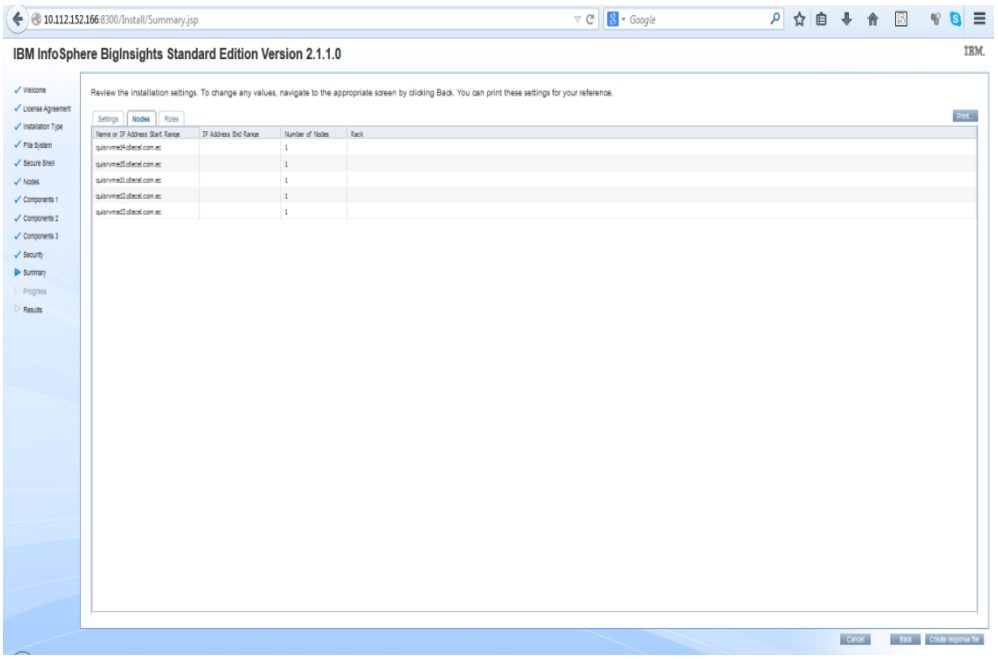


Figura 32 Creación del Archivo Response File

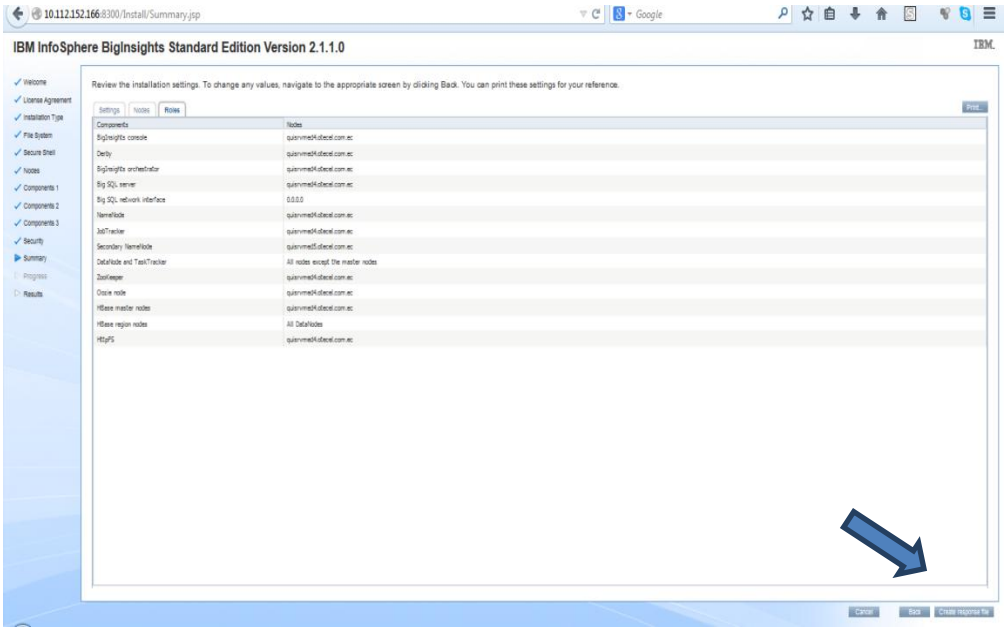


Figura 33 Creación de Archivo Response File

Presionar Restart Now para reiniciar el proceso de instalación en base al "Response File", despues, cuando reinicie, deseleccionar la opción "Create Response File" y seleccionar "Use the previously created response file to complete the installation" y presionar "Next" hasta el final e iniciará el proceso de instalación.

Finalización de la instalación 1

Si se ha especificado todos los parámetros conforme a este documento la instalación resultará exitosa

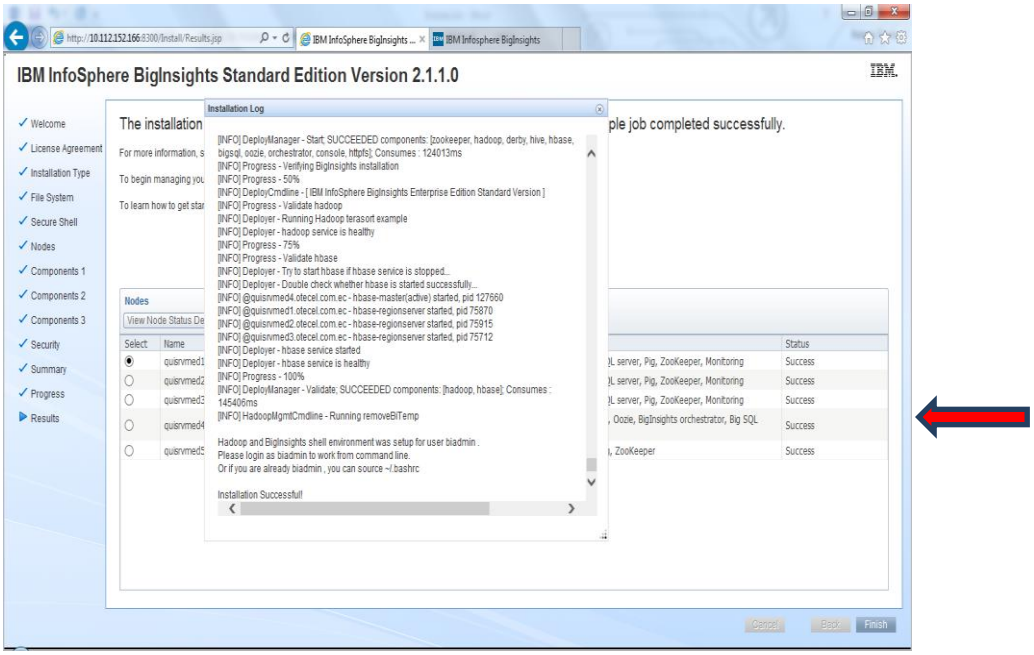


Figura 34 Finalización Instalación

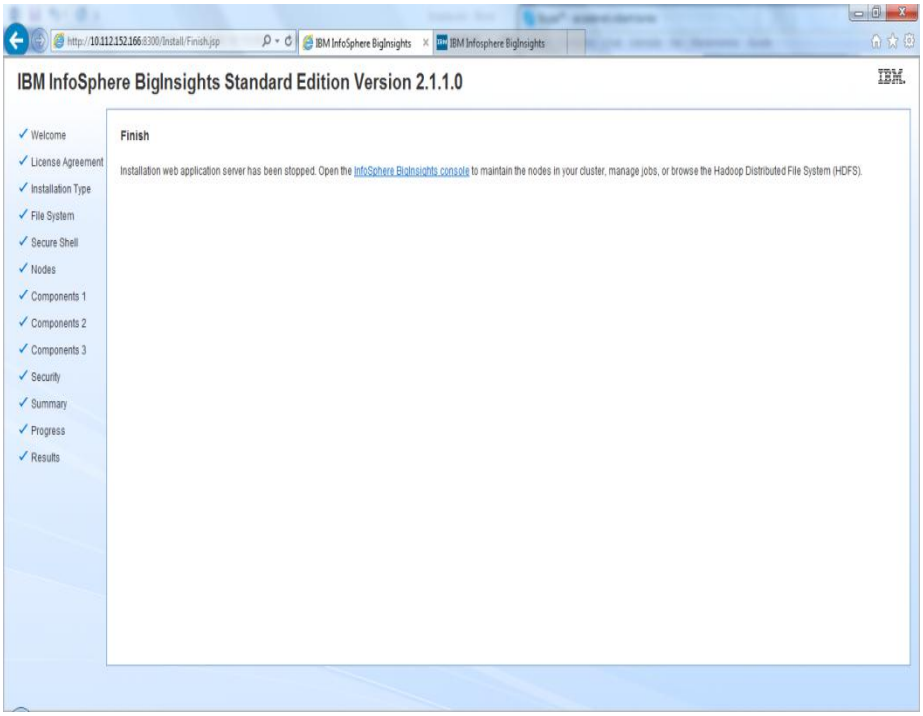


Figura 35 Finalización Instalación 1

Finalización de la instalación 2

Para finalizar el proceso de instalación, en el directorio de instalación ejecutar el siguiente comando descrito en la gráfica y el proceso de instalación terminará

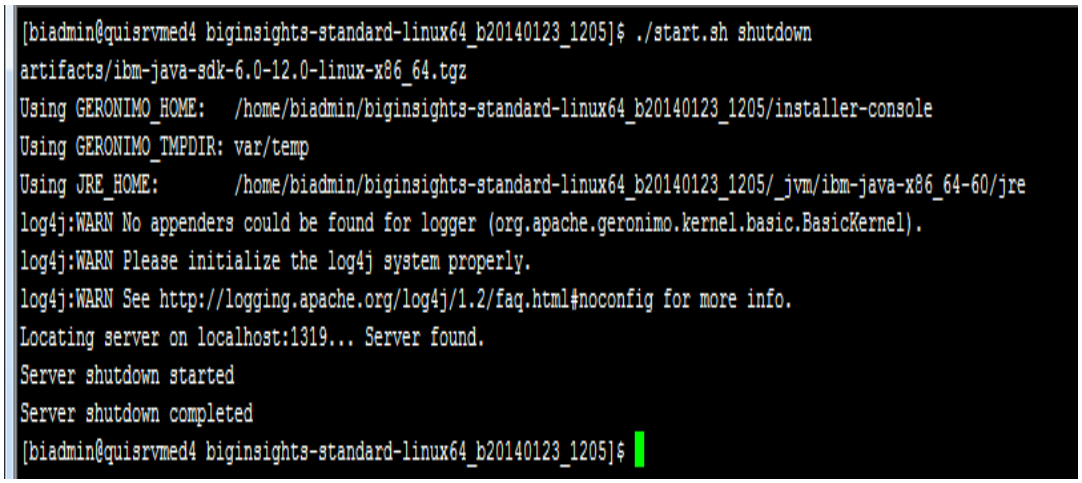


Figura 36 Finalización instalación

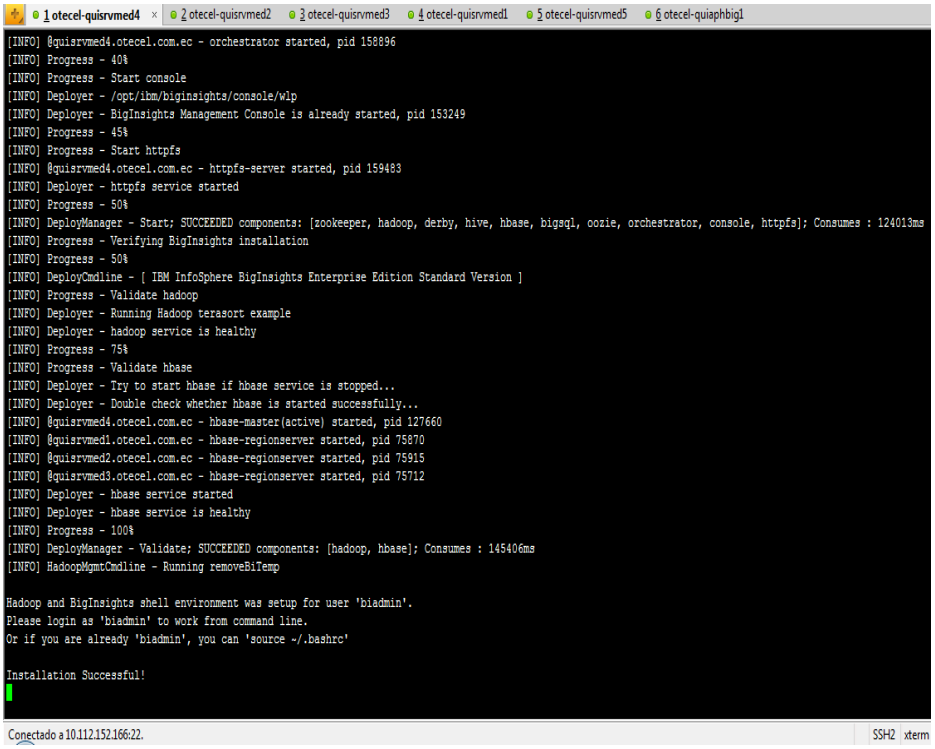


Figura 37 Archivo de log generado en el directorio de instalación

3.2 ANÁLISIS, PROCESAMIENTO Y CARGA DEL CDR

3.2.1 Búsqueda de información relevante y concerniente a los usuarios

La información de un archivo CDR consta mínimo 100 campos con información muy importante para las TELCOS.

Para ello se utilizará los campos por defecto del sistema Asterisk

Tabla 4: Análisis de Estructura de la información

FIELD_NAME	EXAMPLE	DESCRIPTION
Accountcode	12345	An account ID. This field is user-defined and is empty by default.
Src	12565551212	The calling party's caller ID number. It is set automatically and is read-only.
Dst	102	The destination extension for the call. This field is set automatically and is read-only.
Dcontext	PublicExtensions	The destination context for the call. This field is set automatically and is read-only.
Clid	"Big Bird" <12565551212>	The full caller ID, including the name, of the calling party. This field is set automatically and is read-only.
Channel	SIP/0004F2040808-a1bc23ef	The calling party's channel. This field is set automatically and is read-only.
Dstchannel	SIP/0004F2046969-9786b0b0	The called party's channel. This field is set automatically and is read-only.
lastapp	Dial	The last dialplan application that was executed. This field is set automatically and is read-only.
Lastdata	SIP/0004F2046969,30,tT	The arguments passed to the lastapp. This field is set automatically and is read-only.

CONTINUA

start	2010-10-26 12:00:00	The start time of the call. This field is set automatically and is read-only.
answer	2010-10-26 12:00:15	The answered time of the call. This field is set automatically and is read-only.
End	2010-10-26 12:03:15	The end time of the call. This field is set automatically and is read-only.
Duration	195	The number of seconds between the start and end times for the call. This field is set automatically and is read-only.
billsec	180	The number of seconds between the answer and end times for the call. This field is set automatically and is read-only.
Disposition	ANSWERED	An indication of what happened to the call. This may be NO ANSWER, FAILED, BUSY, ANSWERED, or UNKNOWN.
Amaflags	DOCUMENTATION	The Automatic Message Accounting (AMA) flag associated with this call. This may be one of the following: OMIT, BILLING, DOCUMENTATION, or Unknown.
Userfield	PerMinuteCharge:0.02	A general-purpose user field. This field is empty by default and can be set to a user-defined string.[a]
Uniqueid	1288112400.1	The unique ID for the src channel. This field is set automatically and is read-only.

3.2.2 Análisis de la estructura de la información relevada.

Para el análisis de la información se utilizará la información de la tabla de con los campos por defecto del sistema Asterisk, en este proceso se detalla en un mapeo los campos, sus tipos de datos y descripciones desde los campos destino a los campos origen en un formato más comprensible y apto para el almacenamiento en el repositorio de Big Data.

Tabla 5: Análisis de Estructura de la información

NO	CAMPO DESTINO	TIPO DATO	CAMPO ORIGEN	TIPO DATO	DESCRIPCIÓN	TRANSFORMACIÓN
1	CALL_START_TIME	TimeStamp	start	TimeStamp	Fecha de llamada	Mapeo directo
2	A_DIRECTION_NUMBER	VarChar(25)	src	VarChar(50)	Origen de llamada	Limpieza de caracteres especiales
3	B_DIRECTION_NUMBER	VarChar(25)	dst	VarChar(50)	Destino de llamada	Limpieza de caracteres especiales
4	FORWARDED_TO_NUMBER	VarChar(25)	dcontext	VarChar(50)	Llamada re-direccionada	Limpieza de caracteres especiales
5	A_FIRST_CELL	Integer	srcclid	VarChar(50)	Celda 2G Origen	Casting a entero
6	A_CELL	Integer	srcchannel	VarChar(50)	Celda 3G Origen	Casting a entero
7	A_IMEI	Bigint	srclastapp	VarChar(50)	IMEI Origen	Casting a bigint
8	A_IMSI	VarChar(20)	srclastdata	VarChar(50)	IMSI Origen	Limpieza de caracteres especiales
9	B_IMEI	Bigint	dstlastapp	VarChar(50)	IMEI Destino	Casting a bigint
10	B_IMSI	VarChar(20)	dstlastdata	VarChar(50)	IMSI Destino	Limpieza de caracteres especiales
11	B_FIRST_CELL	Integer	dstclid	VarChar(50)	Celda 2G Destino	Casting a entero
12	B_CELL	Integer	dstchannel	VarChar(50)	Celda 3G Destino	Casting a entero
13	DX_CAUSE	Integer	disposition	VarChar(50)	Estado llamada	Casting a entero
14	B_ANSWERED_TIME	TimeStamp	answer	TimeStamp	Fecha de Contestación	Mapeo directo
15	CHARGINING_END_TIME	TimeStamp	end	TimeStamp	Fecha de Finalización	Mapeo directo
16	GLOBAL_CALL_REFERENCE	VarChar(50)	amaflags	VarChar(50)	Identificador de llamada	Casting directo
17	NOMBRE_ARCHIVO	VarChar(50)	uniqueid	VarChar(50)	Archivo Origen	Casting directo
18	DURACION	Integer	duration	Integer	Duración llamada	Mapeo directo

3.2.3 Patrones de procesamiento de la información con calidad de datos

Para la especificación de los patrones de calidad de datos en la información se plantean las siguientes consideraciones:

Consideraciones:

- **Para los campos de información de números celulares**:
  - Limpieza de caracteres especiales: +,- y...
  - Limpieza del número de ubicación regional nacional: 593
- **Para los campos de información de IMEI:**
  - Limpieza de caracteres especiales: +,- y...
- **Para los campos de información de IMSI:**
  - Limpieza de caracteres especiales: +,- y...
- **Para los campos de información de Celdas:**
  - Limpieza de caracteres especiales: +,- y...
- **Para los campos de información de Fechas y Horas:**
  - Limpieza de caracteres especiales: - y:
- **Para el campo de información del CDR:**
  - Reasignación de códigos para reducción de tamaño del campo 3 campos: NOMBRE_ARCHIVO en: COD_CENTRAL, COD_CENTRAL_BASE, COD_CENTRAL_RTT

3.2.4 Procedimientos de carga de la información

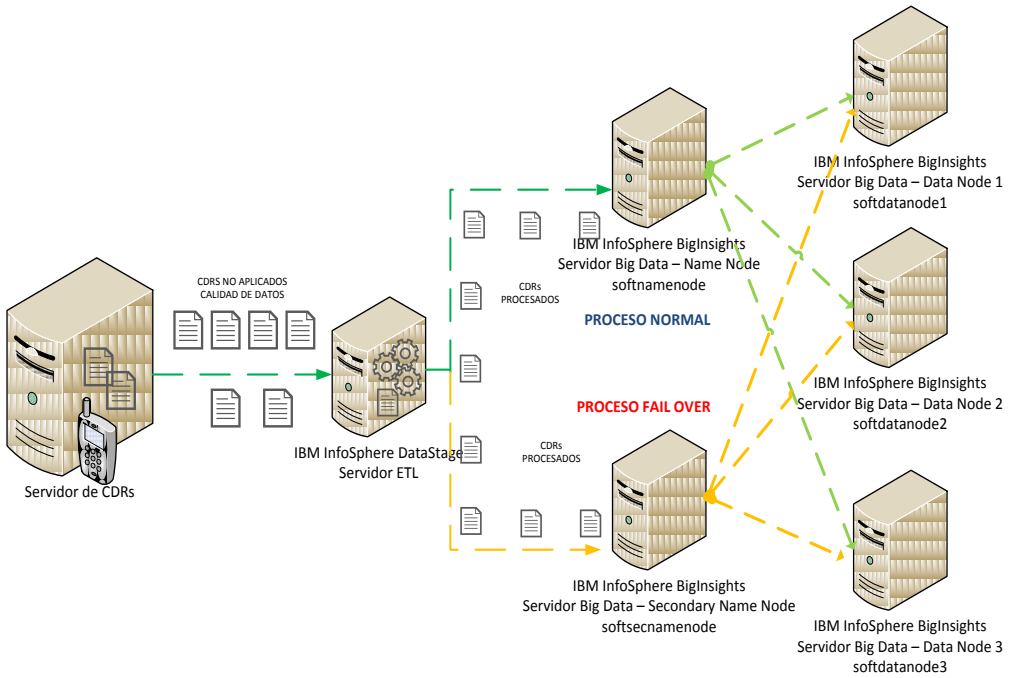


Figura 38 Procedimiento Carga de Información

El diagrama de arquitectura detallada el proceso general de Extracción, Transformación y Carga del CDR desde el servidor de CDRs hasta IBM InfoSphere BigInsights V2.1.1.1

3.3 DISEÑO DE ESTRUCTURA DE DATOS EN HIVE

Diagrama de datos – FUENTE – DESTINO

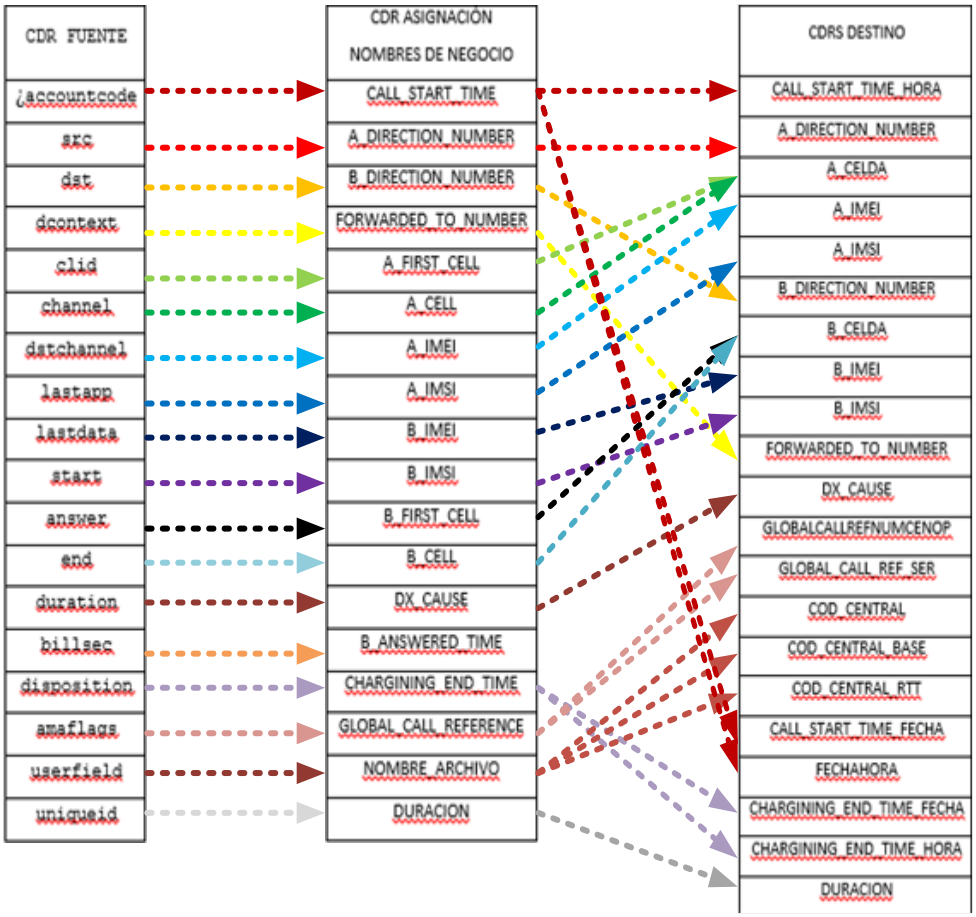


Figura 39 Diagrama Datos- Fuente -Destino

3.3.1 Diseño de estructuras de datos en HIVE

Para las estructuras de datos para almacenar en HIVE, se toma las consideraciones importantes de seguridad e integridad de la información, los datos en Big Data se consideran por lo general archivos planos, lo que causa afectaciones de seguridad en la confidencialidad de la información, por lo que la información deberá ser pasada a un formato binario.

El formato de archivo RCFiles es muy usada para cifrar la información contenida en Big Data y al ser de tipo columnar nos permite procesar la información más rápidamente, para ello, se necesita primero albergar la información en formato Text File y transformarla a RCFiles con ciertas características importantes.

Las características son:

Partitioning

El partitioning (particionamiento) ayudará en la distribución organizada a nivel de datos en porciones especificas basadas en por lo menos un campo de la tabla en Text File, para ello en vista al análisis de la información, se toma en consideración al campo CALL_START_TIME_FECHA, el cual al ser de tipo entero brinda2 beneficios directos:

- Tipo de Dato: Al ser de tipo de dato entero, la información es procesada más rápido y en menor tiempo.
- Información Relevante: Al ser un dato referente a la fecha, contiene información importante y referente a un día en específico lo que ayuda en la administración de la información referente al CDR

Bucketing

El Bucketing ayudará en la organización de la información ordenándola de manera ascendente o descendente, para ello en vista al análisis de la información, se toma en consideración a los campos: A_DIRECTION_NUMBER y B_DIRECTION_NUMBER, el cual al ser de tipo VarChar brinda el beneficio de búsquedas organizadas lo que acelera el retorno de la información.

3.3.2 Creación de estructuras de datos en HIVE

En la creación de las estructuras de datos en HIVE, se utiliza el entorno integrado de desarrollo, IBM DATA STUDIO 4.1.0, el cual posee configuraciones específicas para conectarse a IBM INFOSPHERE BIGINSIGHTS V 2.1.1.1.

La creación de las estructuras a la que se denomina tablas o tuplas poseen estructura similar a SQL de cualquier BD del mercado, con ciertas características especiales.

Las tablas a crearse son:

1. **default.telco_cdrstxt**

La tabla contendrá la información del CDR transformado en formato texto, se realizará un alias a un nombre de referencia: default.otc_t_cdrs_txt

2. **defautl.telco_cdrs**

La tabla contendrá la información del CDR transformado en formato RCFile, se realizará un alias a un nombre de referencia: default.otc_t_cdrsf2

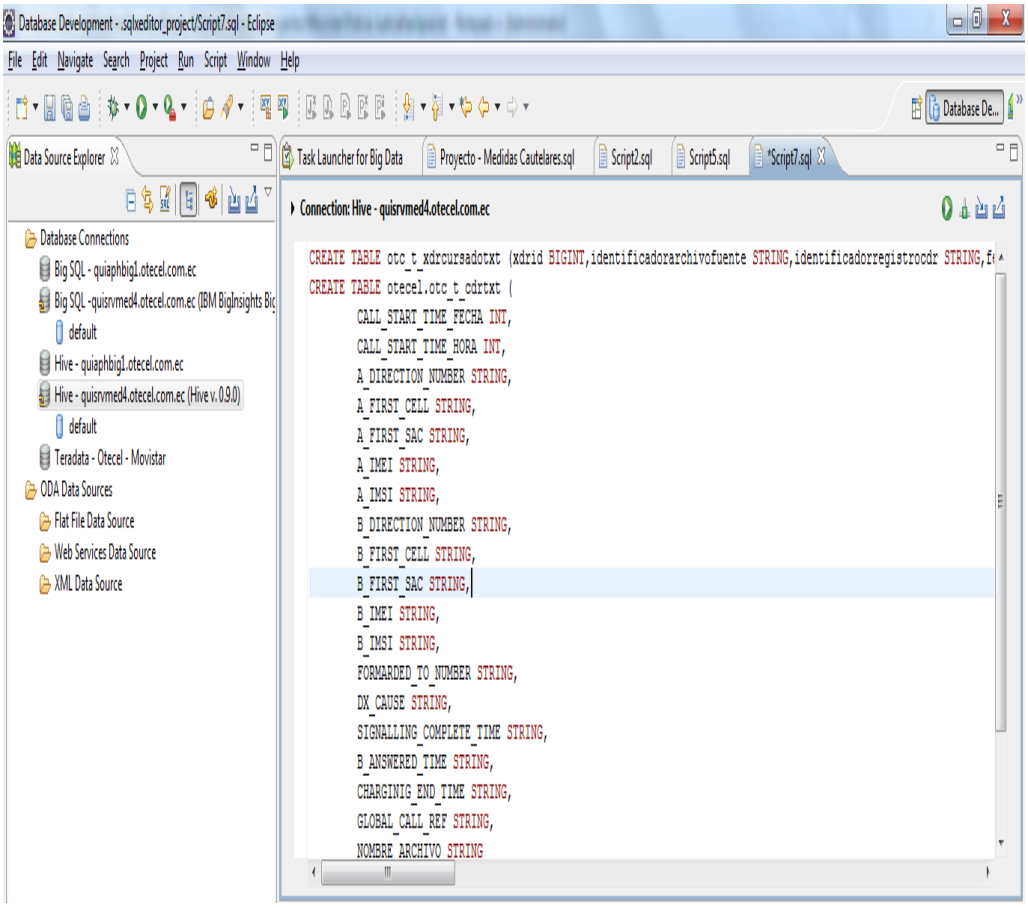


Figura 40 Entorno de IBM Data Studio

3.3.3 Diseño de la Shell de procesamiento y carga

La información procesada desde el servidor de CDRs hasta el clúster de IBM INFOSPHERE BIGINSIGHTS por medio del servidor IBM INFOSPHERE DATASTAGE solo deposita la información en HADOOP y no en HIVE, lo que no brinda la óptima explotación de la información almacenada.

Para ello se necesita de un proceso extra que se debe ejecutar en el name node del clúster, el proceso se detalla en una Shell llamada CDRHadoopHIve.sh donde recibe como parámetro una fecha en formato entero yyyymmdd que representa la fecha de ejecución y de carga del CDR.

El procesamiento y carga se detalla en 3 procesos:

1. Carga la información en HIVE en formato Text File.
2. Lee la información de las fechas contenidas en el CDR.
3. Inserta la información del archivo Text File en base a las fechas leídas y las carga en la tabla defautl.telco_cdrs en formato RCFile.

Trunca la tabla en formato Text File.DISEÑO DE ETLS DE EXTRACCIÓN, PROCESAMIENTO Y CARGA

Para el diseño del Job Paralelo se toma las consideraciones y nombres claves para el desarrollo del mismo de cada Stage y su función:

3.3.4 Diseño del Job Paralelo de Extracción, procesamiento y carga

ETL_PARALLEL_JOB_BIG DATA_TELCOCDRS

Origen:

FF_CDRS_Trafica

Lee los registros del CDR especificado por el parámetro File**"/archivos/CDR_CURSADO/VOZ/INPUT_CDR_VOZ_#FechaCorteInicial#.out"**, donde "#FechaCorteInicial#" es una parámetro de ejecución en formato integer.

Este archivo está en el servidor de desarrollo o producción generado previamente por otro proceso Shell: **"TRAF_CUR_VOZ_PREV.sh"** el cual genera el archivo desde el servidor de ***TRAFFICA.***

Procesamiento:

TRN_FECHAS_NUMEROS

Transforma los números de los campos A_DIRECTION_NUMBER y B_DIRECTION_NUMBER a un formato estándar regido por las TELCOS.

AGG_FECHASCDR

Agrupa por cada fecha proveniente en el capo CALL_START_TIME_TOT y los cuenta para analizar que fechas vienen en cada CDR debido a que los CDRs tienen información retrasada de 1 o variosdías anteriores al procesado.

FLT_ZERO_NON_ZERO

Filtra los registros en los cuales no tienen fecha, envía a BDFS_HADOOOP_CDR_FECHA si tienen fecha, caso contrario a TRN_FECHACORTE

TRN_FECHACORTE

Quema en el registro la fecha en la cual el CDR fue procesado en DataStage.

CP_AGG_FECHASCDR

Crea copias de los registros para ser procesados en TRN_FIELDS y FLT_GREATER_THAN_0.

FLT_GREATER_THAN_0

Filtra las fechas que sean mayor a 0 (debido a que pueden venir son fecha uno o varios registros del CDR), para que sirva de guía para procesos Hadoop (Ver documento Procesos Hadoop).

Destinos de Información:

BDFS_FECHAS_CDR

Guarda la información de las fechas únicas mayores a 0 en el filesystem de Hadoop en el directorio: “/user/biadmin/Archivos/Trafica/CDRs”.

BDFS_HADOOP_CDR_FECHASCDR

Guarda los registros de los CDRs con fecha mayor a 0 en el filesystem de Hadoop en el directorio: “/user/biadmin/Archivos/Trafica/CDRs/FechasXcdr”.

BDFS_HADOOP_CDR_FECHAS_00

Guarda los registros que tiene fecha 0 en el filesystem de Hadoop en el directorio: “/user/biadmin/Archivos/Trafica/FechasCero”.

3.3.5 Diseño de Jobs Secuenciales de orquestación de procesos

ETL_SECUENCE_JOB_BIG DATA_TELCO

EXC_TRAFFICA_CDR_SRV_DATASTAGE

Se ejecuta la Shell script /archivo/CDR_CURSADO/VOZ/TRAF_CUR_VOZ_PREV.sh con la sobrecarga de la fecha en formato integer “yyyymmdd”.

JAC_OTL_LOAD_CDR_TRAFFICA_BDFS

Se Ejecuta el Job Paralelo que recibe como parámetro principal FECHAPROCESO en la Variable FechaCorte.

EXC_RM_TRAFFICA

Ejecuta el comando “rm –rf” con los parámetros “/archivos/CDR_CURSADO/VOZ/INPUT_CDR_VOZ_#FECHAPROCESO.out” donde #FECHOPROCESO es un parámetro sobrecargado en formato Integer “yyyymmdd”, el cual elimina el CDR de esa fecha específica.

EXC_RM_RF_RESPALDOS

Ejecuta el comando “rm –rf” con los parámetros “/archivos/CDR_CURSADO/VOZ/respaldo/TEMSS*”, el cual elimina los backups del CDR de esa fecha específica.

CAPÌTULO 4

4.1 SCRIPTS DE LA ESTRUCTURAS DE DATOS

SCRIPT DE LA TABLA default.telco_cdrstxt

```
CREATE HADOOP TABLE"DEFAULT"."TELCO_CDRSTXT" (
        "CALL_START_TIME_HORA"INTEGER,
        "A_DIRECTION_NUMBER"VARCHAR (50),
        "A_CELDA"INTEGER,
        "A_IMEI"BIGINT,
        "A_IMSI"VARCHAR (50),
        "B_DIRECTION_NUMBER"VARCHAR (50),
        "B_CELDA"INTEGER,
        "B_IMEI"BIGINT,
        "B_IMSI"VARCHAR (50),
        "FORWARDED_TO_NUMBER"VARCHAR (50),
        "DX_CAUSE"SMALLINT,
        "GLOBALCALLREFNUMCENOP"BIGINT,
        "GLOBAL_CALL_REF_SER"VARCHAR (11),
        "COD_CENTRAL"INTEGER,
        "COD_CENTRAL_BASE"INTEGER,
        "COD_CENTRAL_RTT"INTEGER,
        "CALL_START_TIME_FECHA"INTEGER,
        "FECHAHORA"BIGINT,
         "CHARGINIG_END_TIME_FECHA"INTEGER,
    "CHARGINIG_END_TIME_HORA"INTEGER,
    "DURACION"INTEGER

    )    ROWFORMATDELIMITEDFIELDSTERMINATEDBY','
    LINES TERMINATEDBY'\n'
    STOREDASTEXTFILE ALIAS HIVE
METADATA"DEFAULT"."OTC_T_CDRSTXT_F2";
```

SCRIPT DE LA TABLA defautl.telco_cdrs

```
CREATE HADOOP TABLE"DEFAULT"."TELCO_CDRS" (
        "CALL_START_TIME_HORA"INTEGER,
        "A_DIRECTION_NUMBER"VARCHAR (50),
        "A_CELDA"INTEGER,
        "A_IMEI"BIGINT,
        "A_IMSI"VARCHAR (50),
        "B_DIRECTION_NUMBER"VARCHAR (50),
        "B_CELDA"INTEGER,
        "B_IMEI"BIGINT,
        "B_IMSI"VARCHAR (50),
        "FORWARDED_TO_NUMBER"VARCHAR (50),
        "DX_CAUSE"SMALLINT,
        "GLOBALCALLREFNUMCENOP"BIGINT,
        "GLOBAL_CALL_REF_SER"VARCHAR (11),
        "COD_CENTRAL"INTEGER,
        "COD_CENTRAL_BASE"INTEGER,
        "COD_CENTRAL_RTT"INTEGER,
        "FECHAHORA"BIGINT,
    "CHARGINIG_END_TIME_FECHA"INTEGER,
      "CHARGINIG_END_TIME_HORA"INTEGER,
      "DURACION"INTEGER

      ) partitionedby (CALL_START_TIME_FECHA INT)
    Clusteredby (A_DIRECTION_NUMBER, B_DIRECTION_NUMBER)
    into 48 bucketsstoredasrcfile ALIAS HIVE
METADATA"DEFAULT"."OTC_T_CDRSF2";
```

4.2 CAPTURAS DEL DESARROLLO DE LOS ETLS

4.2.1 JOBS PARALELOS

- **ETL_PARALLEL_JOB_BIG DATA_TELCOCDRS**

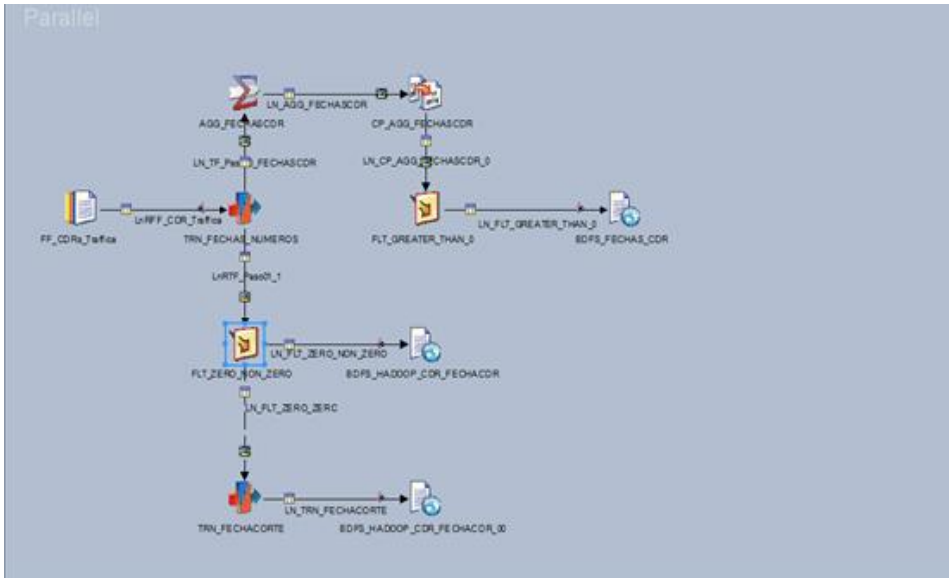


Figura 41 ETL_PARALLEL_JOB_BIG DATA_TELCOCDRS

4.2.2 JOBS SECUENCIALES

- **ETL_SECUENCE_JOB_BIG DATA_TELCO**

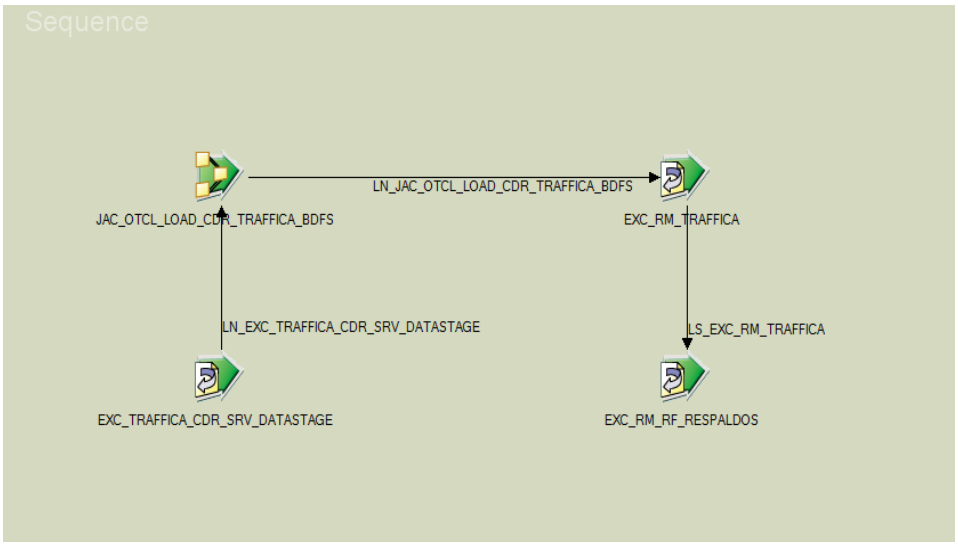


Figura 42 ETL Secuencial

4.3 SEGURIDADES

- **Nivel de datos**
    - Información protegida al estar en formato RCFile que se encuentra en binario.
- **Nivel de aplicativo**
    - Restricciones de acceso al sistema por consola Web.
    - Restricción de modificar la estructura o integridad del archivo si no se posee la credencial adecuada.

4.4 PRUEBAS

- Tiempo de Carga de datos

  Para la toma de tiempo de carga de datos, se tomó en consideración 2 semanas como tiempo de muestreo, ya que en una TELCO, existe una tendencia casi fija de datos como se muestra en la tabla siguiente de la cantidad de registros generados en el CDR por llamadas celular.

Tabla 6: Pruebas

DÍAS	NUM REGISTROS
LUNES 1	38674459
MARTES 2	33107746
MIERCOLES 3	27953239
JUEVES 4	22095231
VIERNES 5	33131211
SABADO 6	22811957
DOMINGO 7	17787172
LUNES 8	41964157
MARTES 9	35390863
MIERCOLES 10	29586755
JUEVES 11	22537049
VIERNES 12	33038385
SABADO 13	21608576
DOMINGO 14	16533952

En base a esta recopilación de datos, se muestra la siguiente Figura:

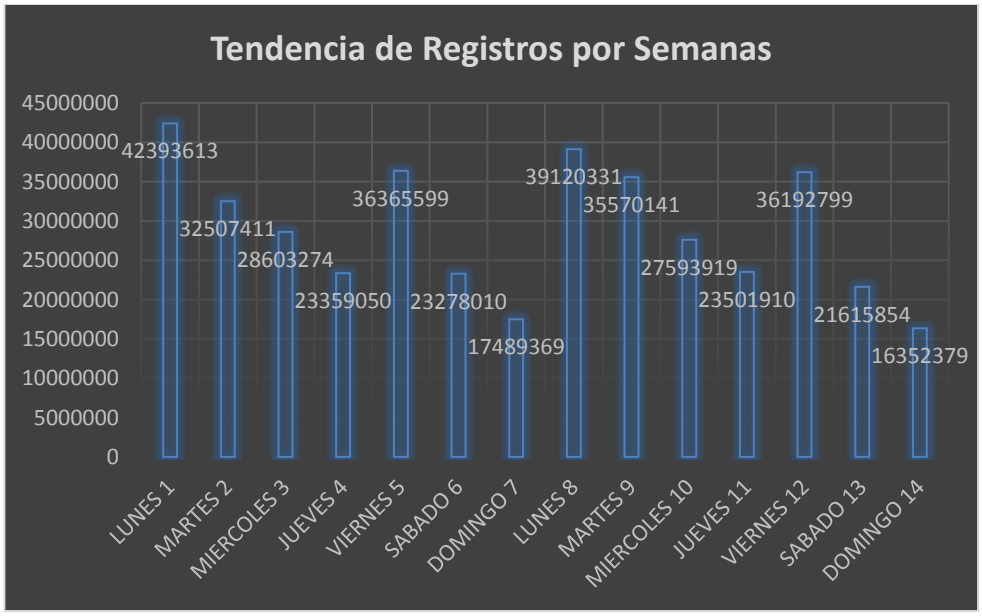


Figura 43 Tendencias de Registros Por Semana

- Tiempo de Respuesta de Consultas

  Para la toma del tiempo de respuesta tras ejecutar consultas, se tomó la siguiente consulta como base para la ejecución de las pruebas, se tomará muestras en consideración al número de consultas simultáneas y la cantidad de registros devueltos, los campos a retornar son: call_start_time_fecha, call_start_time_hora, a_direction_number, b_direction_number, forwarded_to_number a_celda, b_celda, chargining_end_time_fecha, chargining_end_time_hora y duración.

Tabla 7: Tiempo de Respuesta

O	I	NUM CONSULTAS	NUM TELEFÓNICOS	TIEMPO DE CONSULTA (días)	TIEMPO RESPUESTA (seg)	PROMEDIO RESPUESTA	CANTIDAD DE REGISTROS	PROMEDIO REGISTROS	FEC INI	FEC FIN
	1	1	0999927666		14	**14**	27	**27**	20150601	20150601
	2	**2**	0999927635		22	**24,5**	0	**9,5**	20150601	20150601
	3		0986940633		27		19			
	4	**3**	0999927797	1	22	**17**	12	**19,66666667**	20150601	20150601
	5		0999927638		14		19			
	6		0999927710		15		28			
	7	**4**	0999927892		31	**24**	22	**22,25**	20150601	20150601
	8		0999927608		32		24			
	9		0999927603		21		31			
0	0		0999927772		12		12			
1	1	**5**	0999927607		66	**67,8**	32	**34,8**	20150601	20150605
2	2		0999927918		69		30			
3	3		0999927835		64		52			
4	4		0999927646		72		8			
5	5		0999927889		68		52			
6	6	**6**	0999927961	**5**	73	**68,83333333**	53	**35,16666667**	20150601	20150605
7	7		0999927864		72		6			
8	8		0999927720		73		54			
9	9		0999927710		60		34			
0	0		0999927858		69		55			
1	1		0999927808		66		9			
2	2	**7**	0999927787		71	**68,571428 57**	8	**30,571428 57**	20150601	20150605
3	3		0999927666		72		42			
4	4		0999927769		73		52			
5	5		0999927706		66		28			
6	6		0984004704		64		39			

CONTINUA

7		0999927908		66		14			
8		0999927868		68		31			
9	8	0999927762	5	67	66,375	11	27,625	20150601	20150605
0		0999927880		74		52			
1		0999005763		60		33			
2		0998521848		62		28			
3		0999927683		61		12			
4		0999927936		67		31			
5		0999927777		71		24			
6		0999927617		69		30			
7	9	0999927608	10	71	110,6666667	16	46,11111111	20150601	20150610
8		0999927807		108		47			
9		0999927635		133		61			
0		0986940633		64		36			
1		0999927916		169		67			
2		0999927848		134		36			
3		0999927769		137		72			
4		0999927719		112		21			
5		0999927659		68		59			
6	10	0999927865		108	97,5	16	40,3	20150601	20150610
7		0999927888		107		21			
8		0999927843		107		77			
9		0999927760		144		66			
0		0999927753		103		83			
1		0999927650		67		21			
2		0999927921		66		19			
3		0999927665		65		17			
4		0999927912		85		53			

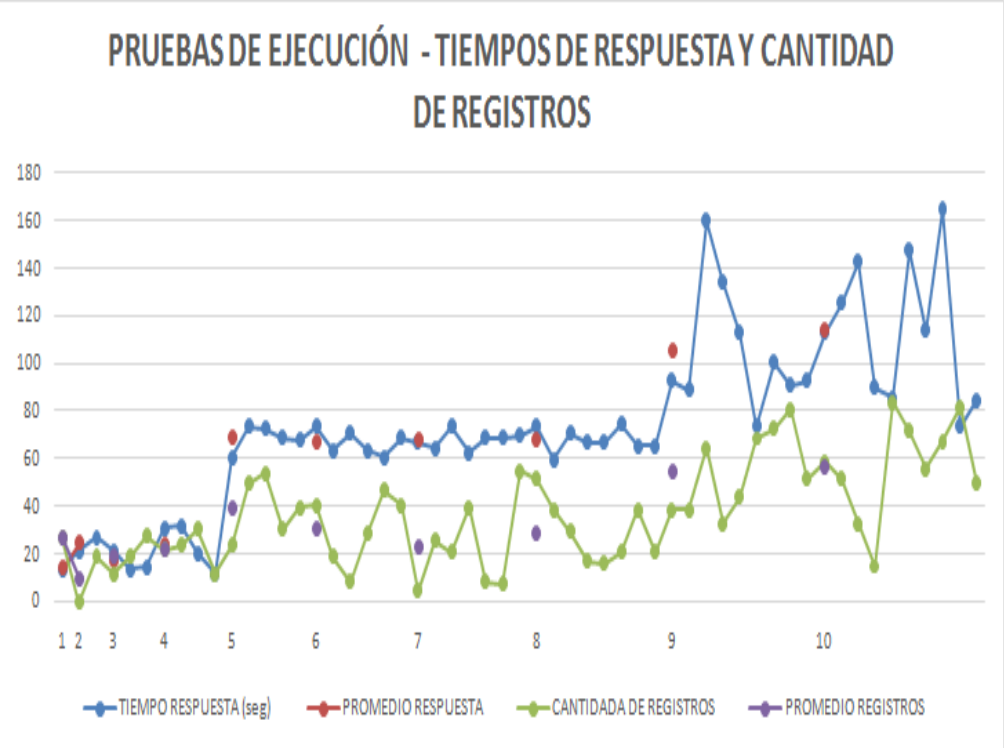


Figura 44 Pruebas Tiempos de Respuesta y Cantidad de Registros

CAPÌTULO 5

CONCLUSIONES Y RECOMEDACIONES

5.1 CONCLUSIONES

- La tendencia de la curva del tiempo de respuesta, se estabiliza cuando se incrementa la cantidad de consultas y el número de registros retornados, esto se debe a la naturaleza de la estructura de la tabla que se encuentra particionada y con Bucketing.
- En Big Data es importante tener una arquitectura heterogénea de componentes, considerando que son difíciles de acoplar hacia otros sistemas que no sean empresariales o que no tengan conexiones nativas por el "Paralelismo".
- En arquitecturas de Big Data se puede notar como ventaja, que al tener réplica de datos en 3 en los datanodes, no se tiene pérdida de servicio, considerando que la probabilidad de fallo es de 1 a 10000 y que se diseña para que los nodos de datos tengan un balanceo de carga y colas de procesos tras falla de uno o más nodos de datos, además, envía alertas tras cada fallo de uno o más componentes.
- Utilizando las herramientas de IBM se comprueba la alta compatibilidad, comunicación y manejo de información entre productos de IBM y de otros sistemas ya que éstas se integran con la mayoría de fuentes de datos, sistemas y recursos tecnológicos, sean open source o de carácter privativo.

5.2 RECOMENDACIONES

- Se recomienda que las personas que implementen una solución de Big Data, tengan conocimientos de Networking, sistemas operativos, bases de datos e infraestructura para crear una solución adecuada.
- Para optimizar más aún la información en Big Data, se necesita tener conocimientos de minería de datos y optimización de datos, especialmente en consultas masivas de gran retorno. Si es mal estructurada la información y no tiene un adecuado manejo, no se evidencia ningún beneficio de poseer una arquitectura de esta naturaleza.
- No es necesario invertir en Hardware especializado para procesamiento y almacenamiento de grandes volúmenes de datos, debido a que se diseña para alto desempeño y escalabilidad a todo nivel tanto horizontal como vertical.
- Big Data no solo puede albergar información en texto plano, también se podría utilizar para analizar las llamadas efectuadas y extraer la información de voz a texto para realizar análisis de patrones delictivos, sean de índole criminalística o judicial y de esta manera brindar un servicio de protección, monitoreo y acción ante eventos malignos.

BIBLIOGRAFIA

bjelvert, P. (junio de 2013). *http://www.slideshare.net.* Obtenido de Big Data using Data Stage: http://www.slideshare.net/ibmsverige/big-data-for-dummies-using-data-stage-live-tool-demo

Franco, R. (18 de 06 de 2012). *Que es Big Data*. Obtenido de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

knowledgecenter, IBM;. (2012). IBM InfoSphere BigInsights Version 3.0. EUU.

Mike, E., De Souza, R., Lima, M., Nobles, M., & Waters, B. (junio de 2013). *http://www.redbooks.ibm.com/.* Obtenido de http://www.redbooks.ibm.com/: http://www.redbooks.ibm.com/

Robinson, Gary;IBM. (7 de noviembre de 2013). *http://www.slideshare.net.* Obtenido de Text Analytics: http://www.slideshare.net/NicolasJMorales/text-analytics-30065459

Seeling, C., Resende, L., Saraco, C., & Linder, S. (23 de 04 de 2013). *Desarrollo de una Aplicación de Big Data para Explorar y Descubrir Datos*. Obtenido de http://www.ibm.com/developerworks/ssa/data/library/app_bigdata/872864.html

Yazid, H. (2012). *Protocolos de Transporte y Aplicación.*

ELABORADO POR:

SALAZAR SÁNCHEZ DANIEL FABRICIO

ELABORADO POR:

TORRES FLORES ROBERT ANDRÉS

DIRECTOR DE LA CARRERA

ING. MAURICIO CAMPAÑA

Sangolquí, Junio del 2015