Introducción a la implementación de una Arquitectura Virtual para Big Data utilizando Cloudera basado en Hadoop

Krisel Salas Barrantes and Yirlania Torres Salas

Escuela de Ingeniería,
Universidad Latinoamericana de Ciencia y Tecnología,
ULACIT, Urbanización Tournón, 10235-1000
San José, Costa Rica
correo01@ulacit.ed.cr,correo02@ulacit.ed.cr
http://www.ulacit.ac.cr

Resumen El proyecto se basa básicamente en realizar una introducción para una correcta instarauración de Big Data con la utilización de la arquitectura Cloudera, la cual se basa en Hadoop. Esta determina quë componentes son necesarios para el buen funcionamiento de la plataforma, mediante la correcta revisión de diferentes artículos y videos técnicos, para ello, se obtuvó información que comprende una mejor tecnología y se adquiere prácticas para la creación de Hadoop. Se utilizaron diferentes herramientas sugeridas por el artículo, virtualización de servidores para implementar Hadoop, elaboración de una guía donde se pone en evidencia cada una de las pruebas realizadas en la plataforma. Se investigó sobre la implementación de Hadoop cuando se requiere más de un nodo, cuáles son los requisitos físicos de los equipos y los comandos con los que se puede llevar a cabo la creación de dichos nodos. También se logra observar el comportamient de las herramientas durante el proceso de investigación y ajuste. En el documento también se puede encontrar una pequeña comparación entre los Frameworks Hortonworks y Cloudera, al igual que se menciona Microsoft Azure el cual es una plataforma donde se puede trabajar con Cloudera y todos sus componentes ya viene previamente configurados.

Keywords: Big Data,Cloudera,Cloudera Impala,Cloudera Express,Cloudera Enterprise,Cloudera Director, Hadoop,MapReduce,HDFS,VMware,NoSql

¿Cuáles son los componentes para llevar a cabo la implementación de Cloudera basado en Hadoop?

Objetivo General:

- Desarrollar una guía que facilite la implementación de una arquitectura de Cloudera para el almacenamiento y procesamiento de datos.

Objetivos Específicos:

- Determinar los componentes y funciones de Hadoop que se requieren para diseñar e implementar una arquitectura de Big Data con Cloudera.
- Analizar y comparar los componentes de Cloudera para determinar cuáles son los más adecuados para implementar una infraestructura de Big Data.
- Comparar el funcionamiento de Cloudera con el de otros frameworks para Big Data.

1. Introducción

Big Data es un sistema de almacenamiento de grandes cantidades de datos en la nube y el cual con el pasar de los años, organizaciones buscan implementarlo en sus plataformas tecnológicas, Hadoop es unos sistemas de código abierto, que se caracteriza por su crecimiento y complejidad, y se constituye por una serie de subproyectos.Ver figura 1.

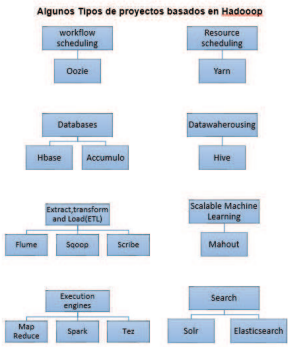

Figura 1. Proyectos Basados en Hadoop

En la imagen anterior(Figura 1.)podemos notar que cada uno de los compartimientos cumple un rol específico dentro el ecosistema, no obstante los más destacados son HDFS y MapReduce ya que se consideran los más importantes y descriptivos, ya que vienen a formar el "core" y definen el ecosistema.

Para entender un poco sobre Hadoop Distributed File System (HDFS), se refiere a un sistema de archivos distribuido, que tiene como principal objetivo el procesamiento de grandes cantidades de datos, principalmente aquellos que presentan problemas de escalabilidad, flexibilidad y rendimiento, HDFS acepta cualquier tipo de datos sin importar el esquema, el uso del ancho de banda y la escalabilidad de mayor a menor orden.

Características de HDFS:

- Arquitectura escalable: Es posible agregar más servidores para aumentar la capacidad.

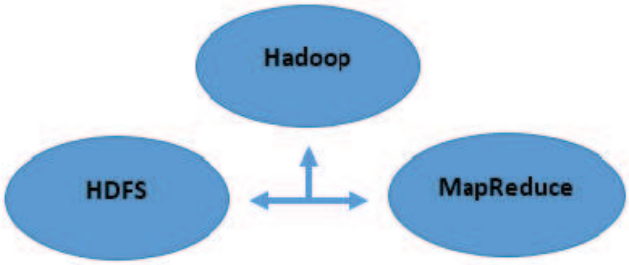

Figura 2. Hadoop HDFS MP

- Alta Disponibilidad: Cumple los objetivos de las aplicaciones y de aquellos flujos que son críticos.
- Tolerancia a fallas: Recuperación de fallas automáticas e inesperadas
- Acceso flexible: Prácticas de trabajo a la hora de ensamblar y asociar un sistema de archivos.
- Balance de Carga: Ubicación inteligente y flexible de los datos para obtener un mayor aprovechamiento y eficiencia.
- Replicación sintonizable: Copias múltiples de archivos para garantizar la protección de los datos y el rendimiento.
- Seguridad: permisos para el acceso de los grupos o usuarios a los diferentes archivos.

El HDFS se replica los diferentes nodos con el fin de obtener un mejor rendimiento y la protección de datos.Ver figura 3.

Por otra parte tenemos a MapReduce el cual se basa en primordialmente en realizar trabajos escalables, con mayor procesamiento y de una manera más ágil, al ser combinando con HDFS para llevar a cabo la ejecución en cada uno de los nodos para la realización de cálculos necesarios. El sistema a la hora de ser ejecutado con MapReduce y Hadoop se lleva acabo en la ubicación que están los datos, con esto se ahorra tiempo y no se requiere estar cambiando de ubicación, los datos almacenados y el sistema se localizan en un mismo grupo de servidores. Con esto podemos asegurar que MapReduce trata cientos de datos y estos no se ven interrumpidos por colisiones, ni por la cantidad de ancho de banda.

Características:

- Escalabilidad: Brinda la posibilidad de agregar más servidores con el fin de aumentar la productividad.
- Administra recursos: Trabaja buscando las mejores formas para gestionar los recursos y almacenamiento.

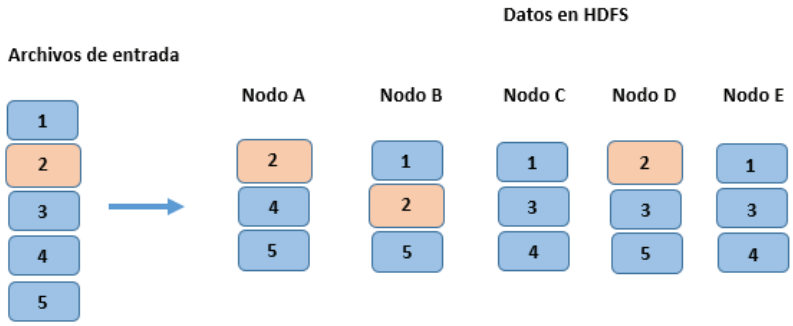

Figura 3. Datos en HDFS

- Optimiza la programación: Va ordenando cada trabajo según la prioridad de cada uno.
- Flexibilidad: Al trabajar los procesos de manera virtual es compatible con cualquier lenguaje.
- Resilencia y disponibilidad: Permite que varias tareas en caso de fallo sean independientes y se repongan automáticamente.Ver figura 4.

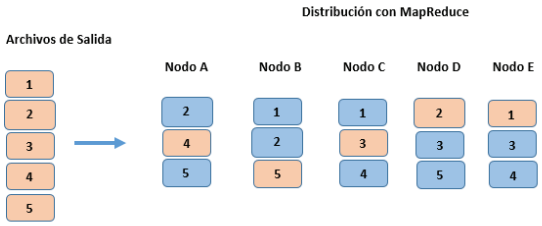

Figura 4. Distribucion con MapReduce

Estado del Arte

En los últimos años el crecimiento de los volúmenes de datos, en especial de tipo no estructurados que se quieren ser analizados en tiempo real, este análisis permite que los datos puedan ser organizados y gestionados con la mayor eficiencia posible, pero también debido a este crecimiento desmedido,se presentan

una serie de problemas para almacenar y gestionar los diferentes tipos de referencias,debido a las necesidades de software,hardware e infraestructura.Estos problemas requieren una solución inmediata como lo es la computación en la nube, la cual ofrece guardar y procesar inmediatamente los datos.Es por esta razón se requiere realizar mejoras en los diferentes problemas de seguridad que presenta HDFS, ya que almacena diferentes tipos de antecedentes, que al ser insertados pueden que se den robos de información,divulgación o bien accesos no deseados de los mismos, es por esto que la mayoría de organizaciones utilizan una plataforma para el almacenamiento de datos valiosos implementen, a la vez un firewall y un software de detección de intrusos, fuera de Hadoop o bien consideran la posibi- lidad de encriptación de los bloques y sistemas de archivos y utilizar métodos de seguridad como: Kerberos,Bull Eye Algorithm Approach,NameNode,los cuales son parte de seguridad.Podemos definir MapReduce como la resolución de algoritmos,sin embargo,no todos los procesos pueden ser tratados con MapReduce, pero para esta situación;por ejemplo, a la hora de trabajar con imágenes web,la función del Map se encarga de mapear,o sea calcular la distancia entre cada nota y la consulta realizada,luego de esto,entra el reduce a ordenar las distancias mencionadas, MapReduce ofrece el almacenamiento de grandes cantidades de datos, ya sean estructurados o no estructurados.También permite utilizar diferentes generadores de consultas como lo son: Hive, Sqoop, Pig e Impala,entre otros.Dentro de estos generadores mencionados,los cuales, se puede obtener por medio de consultas,por ejemplo,el nombre de cuál generador se utiza,al igual que una descripción de la consulta,quién es el propietario, qué tipo de consulta se hizo,el estado de dicha consulta y la última modificación.Y además dependiendo de los permisos asignados al administrador de la infraestructura,se podrá borrar,modificar,mover o restaurar tales notas Una vez dentro de la base de datos se puede obtener acceso a las tablas y a toda la información contenida.En la parte de flujos de trabajo se logra encontrar testimonio detallado sobre los cambios en tiempo real;además se puede enunciar de manera detallada,por ejemplo,si se utiliza en una red social o en una página web,siendo así un tipo de monitoreo,en tiempo real.Dentro de la parte de seguridad se trabaja mediante nodos, roles o grupos donde cada usuario tendrá permisos específicos,según la función que realice,teniendo acceso solo a la información necesaria.Con respecto al explorador de archivos cada carpeta posee un usuario determinado,la que se divide por subcarpetas,con usuarios,los cuales poseen diferentes permisos,como se mencionó antes y pueden pertenecer a diferentes grupos.También se pueden realizar múltiples búsquedas,en una sola consulta.

Mejores prácticas para la implementación de una plataforma virtual:

- Primeramente se requiere descargar un Vmware vSphere client para configurar el host y accesar desde el mismo a la máquina virtual.
- Antes de ejecutar el cliente y conectar el host al cable crossover se requiere cambiar la configuración de red de la pc que se va usar para conectarse al servidor, es por eso que se deben verificar la configuraciones de red para que todos los dispositivos estén dentro del mismo rango de direcciones ip.

- Una vez realizado el paso anterior se debe ejecutar el cliente y empezar su configuración.
- Se crea una VM basada en el HDD bajado de Cloudera. Se necesita VMWare WorkStation 8.x o superior.
- Se asigna el nombre y la configuración por defecto.

Ver figura 5.

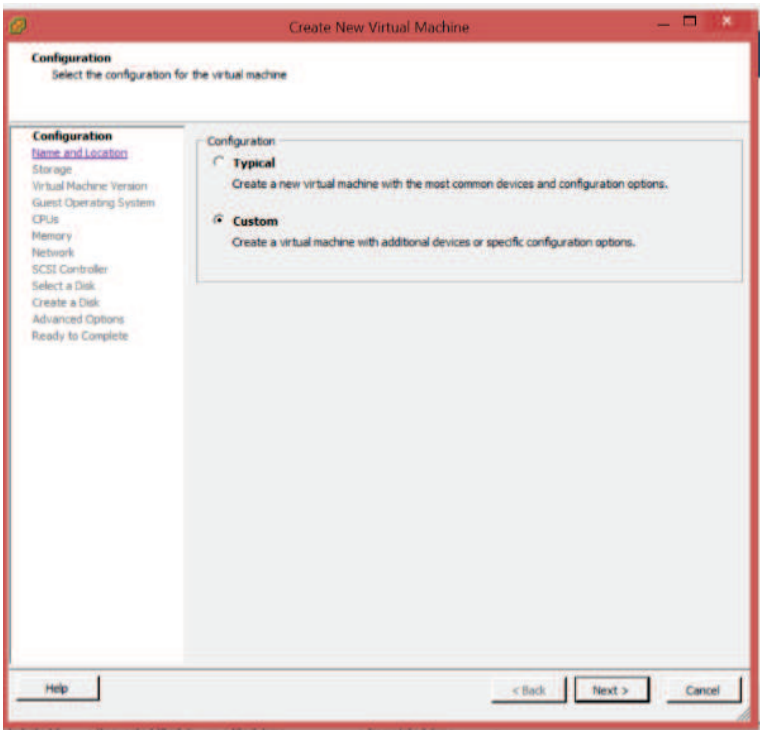

Figura 5. Configuracion de una maquina virtual

- Se define el destino de almacenamiento.
- La versión de la máquina virtual a crear.
- Se debe especificar el tipo de OS en este caso se utilizara Linux.
- Luego se procede a indicar el número de sockets requeridos en nuestro caso 2.
- La memoria debe ser según la versión del CDH ya sea el H por defecto con 4GB, Cloudera express con 8 GB ó Cloudera Enterprise e cual requiere de 10 GB.
- Referente a las conexiones de red con una Nick es sufriente, pero de igual manera esto varía según las necesidades.

- En la parte de SCSI controller se deje el que viene seleccionado por defecto: LSI Logic Paralel
- A la hora de seleccionar el disco nos ubicamos en la parte de crear un nuevo disco virtual.
- En el momento de que asignarle el disco duro, escogen el archivo con extensión .vmdk que descargaron del sitio web de Cloudera.
- Posterior a la creación del disco seleccionamos la capacidad de igual manera según la necesidad, y en la provisión del disco seleccionamos el thin provision y el la localización indicamos con la máquina virtual.
- En la parte de opciones avanzadas lo dejamos por defecto.
- Posterior a toda la configuración podemos volver a asegurarnos que lo realizado este en forma correcta al igual asegurarse que casilla de editar se encuentre habilitada, en caso que se requiera editar alguna configuración. Ver figura 6.

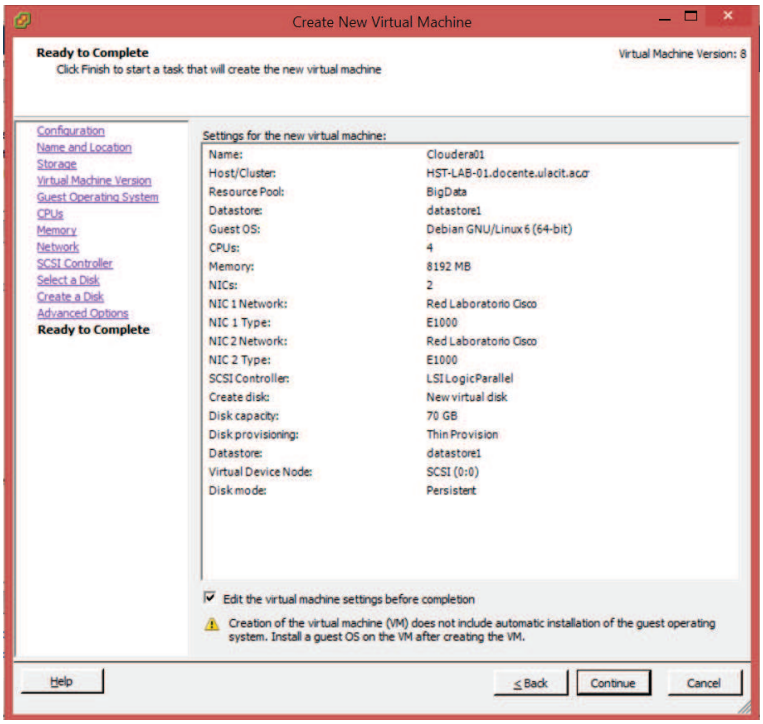

Figura 6. Resumen de la configuracion de la maquina virtual

- Se inicia la VM con 4 Gb de RAM como configuración básica de CDH 5 arranque.Ver figura 7.

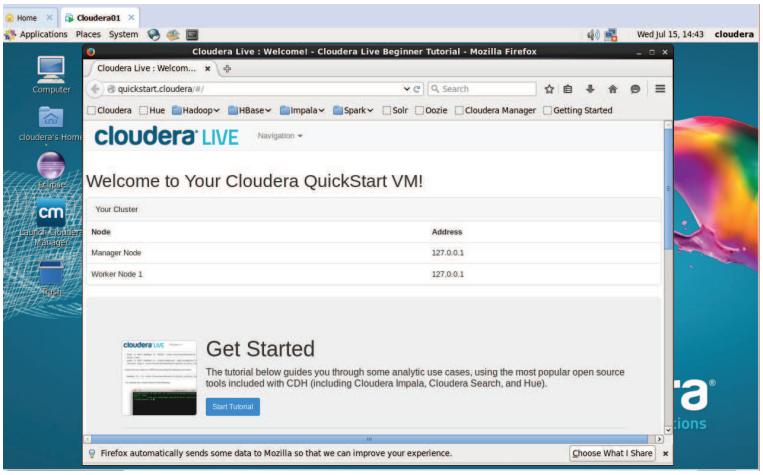

Figura 7. Cloudera Live

- Selecciona la opción Get Started.Ver figura 8.

Algunos inconvenientes que usted podría encontrar en la infraestructura física:

- Complicado acceso físico a los servidores ya que no se posee cables de tipo crossover los cuales son ideales para la conexión directa con el servidor.
- Dificultada de acceso al servidor uno (IP 10.10.1.1) por medio de puertos ya que no fue posible este tipo de conexión a pesar de que se hicieron los respectivos cambios de dirección IP al host que se requería que lograra la conexión al servidor, el objetivo de este cambio se realizó con el fin de que la computadora pertenezca al mismo rango de direcciones, una vez realizado el cambio, por medio de un ping al servidor atraves de la consola cmd se comprobó que no existía conexión entre la pc y el servidor, por lo que se probó la conexión de cada uno de los puertos y el resultado fue el mismo.

Los problemas mencionados anteriormente fueron solucionados de la siguiente manera:

- Se conectó el cable crossover aun segundo servidor el cual contenía la dirección IP 10.10.1.2 y la conexión fue exitosa.

Casos:

Como principal utilidad de una plataforma Cloudera es que se pueda obtener mayor beneficio a costos muy bajos,en un mismo sistema al utilizar diferentes tipos de análisis,lo que queda demostrado en estos casos es que usando CDH se pueden integrar y evaluar cualquier nueva infraestructura,sin cambiar ninguno de los pasos que normalmente se utilizan y de esta forma no se requiere eliminar informes o cargas de trabajo para los datos que se necesiten migar.Para llevar a

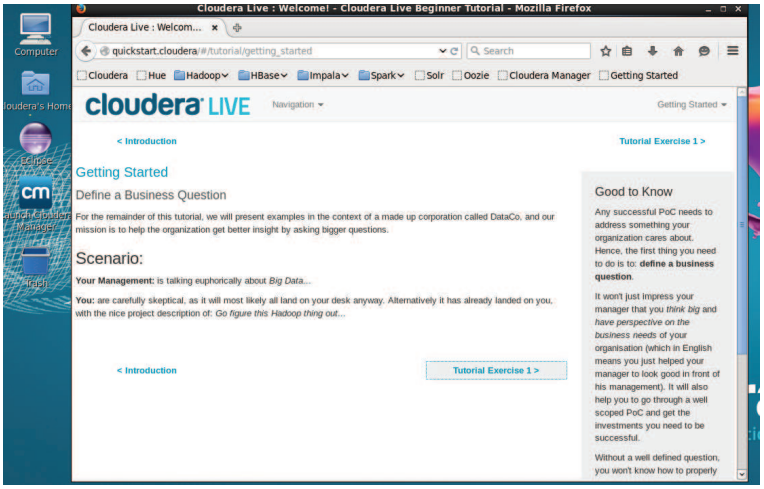

Figura 8. Get Started

cabo el análisis de datos que se trasladaran a la nueva plataforma, se necesita que estos se procesen en el sistema de archivos de Hadoop(HDFS),con una herramienta que facilite la trasferencia de datos estructurados desde un RDBMS para HDFS,utilizando la misma estructura la cual permite la consulta de elementos y sin interrumpir el trabajo que usualmente se realiza.Apache Sqoop forma parte de CDH(herramienta), en conjunto con Sqoop,los cuales permiten cargar automáticamente la información que se encuentre relacionada en MySQL hacia HDFS y posean el mismo tipo de lectura,en caso de utilizar parámetros de configuración adicional se pueden cargar datos relacionales para que puedan ser consultadas por impala que funciona con un motor de consultas de código abierto y en CDH.Si se requiere aprovechar la capacidad del formato de archivo de Apache Avro para futuras cargas de trabajo en el cluster usted debe tomar en cuenta una serie de medidas para que los datos sean cargados por impala utilizando el formato de archivo Avro, resultando altamente disponible tanto para impala como para otras cargas de trabajo.Se Debe abrir la terminal y comenzar su trabajo en Sqoop.Se deben ingresar los siguientes comandos:

- "sqoop import-all-tables m 1 –connect jdbc:mysql://quickstart:3306/retail_db –username=retail_dba –password=cloudera –compression-codec=snappy –as-avrodatafile –warehouse-dir=/user/hive/warehouse"Ver figura 9

Una vez que usted digita el comando anterior debe tomar en cuenta que su proceso puede tardar algunos minutos, pero es en este momento cuando se lograr poner en marcha los trabajo que realiza MapReduce para exporta los datos que se encuentren en nuestra base de datos y que los archivos de exportación estén un formato Avro en HDFS, también se crear un es quema que me permita cargar más fácilmente nuestros datos para futuros usos de impala.

Pasos de verificación:

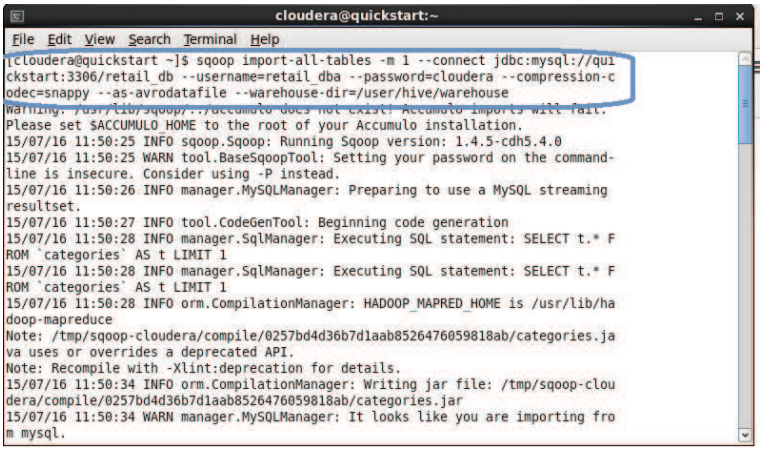

Figura 9. Comando para configurar Squoop

- Ingrese el siguiente comando y una vez completado verifique la existencia de un archivo Avro en HDFS.
- hadoop fs -ls/user/hive/warehouse
- Usted va observar una carpeta para para cada tabla. Ver figura 10.

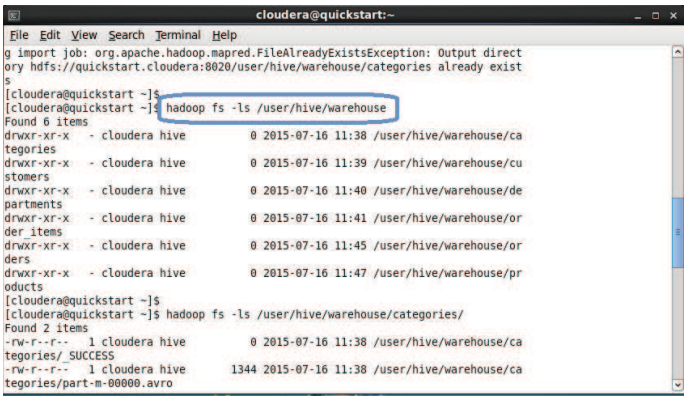

Figura 10. Comando para observar las carpetas de cada tabla

- hadoop fs -ls/user/hive/warehouse/categories/
- Usted puede observar los archivos que hay dentro de cada carpeta.Ver figura 11

Sqoop también debería haber creado archivos de esquema para estos datos en su directorio personal.

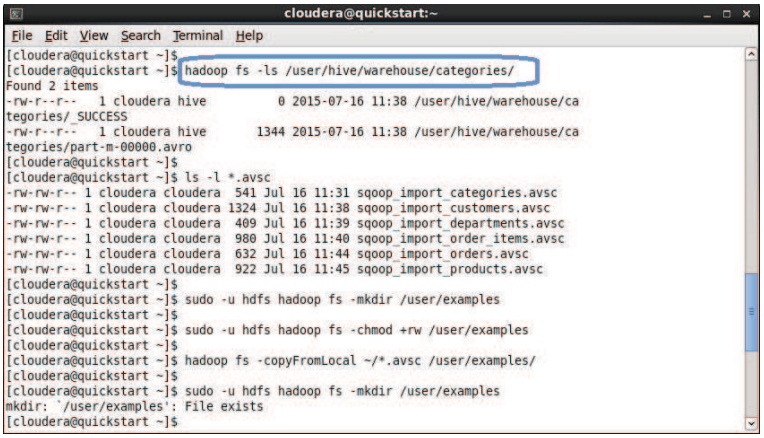

Figura 11. Comando para observar los archivos dentro de una carpeta

- ls -l *.avsc

Debe mostrar .avsc archivos para las seis tablas que estaban en nuestra retail_db. Ver figura 12.

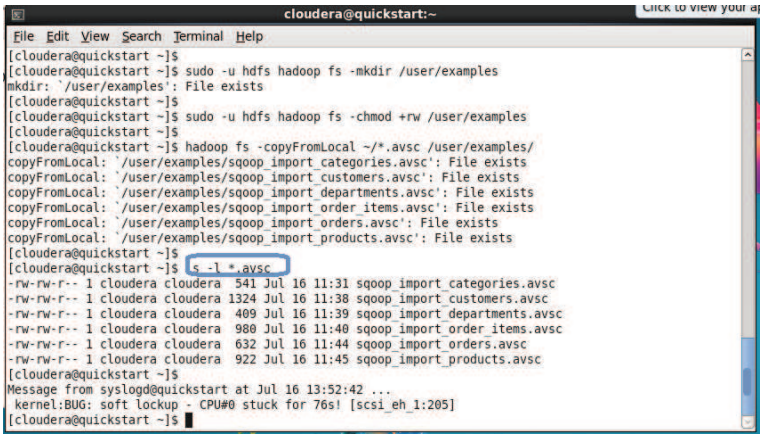

Figura 12. Comando para ver la informacion de las carpetas

Con la ejecución de estos comandos básicos para Sqoop con datos estructurados usando HDFS ya tenemos listos los archivos y podemos consultarlos en el momento necesario.

Agregación de Nodos

Con respecto a la implementación de un segundo nodo se realiza con el objetivo de se pueda librar de cargas al nodo principal, al instalar este tipo de

servicios se logra que sPara crear más de un nodo se requiere tener configurada una máquina virtual que cumpla las características y ajustes que le permitan establecerse como un nodo, es indispensable que se tome en cuenta la configuración de red, la primera máquina virtual que se crea es clonada las veces que se requiera según los nodos del clúster de hadoop, una vez que se finaliza este proceso se requiere modificar el nombre y la dirección IP del host para que el nodo sea funcional, el primer nodo llevara a cabo la mayor parte de servicios del clúster por lo que requiere se le asigne más memoria que a los demás nodos que se agreguen, cuando ya se poseen la cantidad de nodos que se requieren para continuar con la instalación se debe seleccionar la versión de licencia libre de cloudera e incluir todos los nodos que van a formar parte de la agrupación y se recomienda agregar cada uno de los nodos para mediante un método de automatización a los diferente nodos instalarle todos los paquetes y servicios, una vez que este proceso finaliza se pueden agregar otros servicios utilizando la configuración por defecto y de esta forma la información continuara y se completara utilizando el cluster hadoop el cual una vez que finaliza utiliza dos interfaces Cloudera Manager y Hue. Se guarden copias de los diferentes bloques en la memoria los cuales se modifican y actualizan en el nodo principal, sin embargo la creación de un segundo nodo no garantiza la disponibilidad ya que en el momento que el nodo principal sea afectado se cae el sistema de ficheros HDFS y MapReduce, con la creación de un nodo secundario se lograr tener los datos actualizados del fichero que contiene información sobre los bloques del clúster. Para crear más de un nodo se requiere tener configurada una máquina virtual que cumpla las características y ajustes que le permitan establecerse como un nodo, es indispensable que se tome en cuenta la configuración de red, la primera máquina virtual que se crea puede ser clonada las veces que se requiera según los nodos que deseamos agregar, una vez que se finaliza este proceso se requiere modificar el nombre y la dirección IP de cada host para que el nodo sea funcional, el primer nodo será el principal el cual llevara a cabo la mayor parte de servicios del clúster por lo que requiere se le asigne 8GB de memoria un poco más a diferencia de los demás los cuales podrían utilizar 2GB, por ejemplo si creamos 4 nodos en total seria 14GB de memoria, cabe destacar que si no se posee la suficiente memoria podrían darse fallas, en lo que se refiere a la capacidad del disco de almacenamiento debe ser de mínimo 40GB, una vez que ya se tenga la cantidad de nodos que se requieren para continuar con la instalación se debe seleccionar la versión de licencia libre de Cloudera e incluir todos los nodos que van a formar parte de la agrupación y se recomienda agregar cada uno de los nodos para mediante un método de automatización a los diferente nodos instalarle todos los paquetes y servicios, una vez que este proceso finaliza se pueden agregar otros servicios utilizando la configuración por defecto y de esta forma la información continuara y se completara utilizando el clúster hadoop el cual una vez que finaliza utiliza dos interfaces Cloudera Manager y Hue.
Lista de comandos a utilizar para la configuración:
Configuración de Red:
/etc /resolv.conf Para asignar el nombre y dirección del servidor.

/etc/sysconfig/network Para asignar el rango de Ip y el nombre del Host.
/etc/sysconfig /network-scripts /ifcfg-eth0 Para asignar la Ip a la interfaz de la red.
/etc/selinux/config Se debe desactiva la función del modo de seguridad selinux.
/etc/yum/pluginconf.d/fastestmirror.conf Luego de este comando se debe indicar enable=0.
Con estos comandos procedemos a reiniciar las configuraciones de red:
$ > chkconfig iptables off
$ > /etc/init.d/network restart.

Instalaciones para la VM:
Para actualizar paquetes requeridos por la máquina virtual $ > yum update $ > reboot

Para poder acceder a la ubicación del ISO $¿mkdir /media/VBGuest $¿mount -r /dev/cdrom /media/VBGuest

Definir y configurar los hosts: /etc /hosts Debe mostrar la lista de host configurados en red. Configuración y permisos de SSH: $¿yum -y install perl openssh-clients $¿ssh-keygen (type enter, enter, enter) $¿cd /.ssh $¿cp id_rsa.pub authorized_keys /etc/ssh/ssh_config StrictHostKeyChecking no Procedemos a apagar para poder clonar los nodos: Recordar asignar a uno de los nodos 8GB de memoria, luego para cada uno procedemos a ejecutar los siguientes comandos: /etc/sysconfig/network HOSTNAME=hadoop n Donde n es igual al numero de nodo 1,2,3 o 4. /etc/sysconfig/network-scripts/ifcfg-eth0 La ip de cada uno IPADDR=10.0.1.20[n]

Se reinicia la configuración de red

$¿/etc/init.d /network restart $¿init 6

Procedemos a ejecutar el comando para descargar el instalador de Cloudera en los nodos: $¿curl -O http://archive.cloudera.com/cm4/installer/latest/cloudera-manager-installer.bin $¿chmod +x cloudera-manager-installer.bin $¿./cloudera-manager-installer.bin

Cloudera Impala

Impala es una parte integral de cloudera, fue diseñado con el fin de que se puedan aprovechar la flexibilidad, escalabilidad y fortalezas de Hadoop , cloudera permite el aprovechamiento de los recursos unificados, metadatos, seguridad, así como la administración del sistema. Si antes de una solución como impala lo que posee es una base de datos de tipo relacional posiblemente su solución haya sido ampliar este sistema para mantener sus rendimiento, si utilizamos una herramienta como Hadoop para el análisis de los diferentes tipos de datos y se requiere que su rendimiento sea de tipo interactivo los datos tendrían que cumplir con una serie de condiciones que le limitarían en la toma de algunas decisiones, pero si se logara combinar impala con hadoop, se podrán obtener buenos resultados ya que impala como un componente del ecosistema de hadoop logra la combinación de todos los beneficios de este el cual permite flexibilidad, escalabilidad y rentabilidad, facilitando el uso y la funcionalidad de una base de datos a nivel empresarial.

Cloudera Express

Resulta una buena alternativa si la trabajamos con Hadoop, su descarga se puede hacer gratuita y su código es abierto, de la mano de hadoop y cloudera manager, se puede obtener cluster robustos, administración centralizada, monitoreo y diagnósticos, logrando así que la plataforma sea capaz de demostrar que tiene la tecnología, lo que se requiere es que con Hadoop se logre evaluar el procesamiento de los datos mejorando el rendimiento y realizando análisis de conjuntos de datos que otros escenarios no hubiera sido posible, una implementación de cloudera express con hadoop es considerada perfecta en la resolución de casos de primer uso.

Cloudera Enterprise

Implementado con hadoop ayuda a tener una mayor aprovechamiento del código abierto y con las capacidades que la organización requiera para que ApacheHadoop sea lo más exitoso posible para la organización, fue diseñado para entornos críticos, Enterprise incorpora CDH, esta plataforma es basado en hadoop y de código abierto además se considera una de las más reconocidas a nivel mundial, asi como las diferentes herramientas de gestión de sistema y gestión de datos, los expertos y desarrolladores de hadoop consideran a cloudera con su aliado en el desarrollo de los grandes datos.

Cloudera Director

Si este tipo de plataformas se implementa con Hadoop se puede lograr obtener una plataforma integral de gestión de datos logrando un mayor aprovechamiento del director de cloudera en los diferentes ambientes de producción, cuando se implemente el directorio como parte de cloudera Enterprise se logra habilitar la nube de líderes en el directorio de cloudera, logrando acceso al mejor soporte de la industria, seguridad de tipo integral y con la capacidad de tener un influencias para futuras versiones.

Windows Azure

Se tiene otra plataforma la cual mencionaremos brevemente para llevar a cabo la implementación de big data basada en cloudera la cual es Windows Azure, esta ofrece la posibilidad de que las organizaciones puedan implementar big data y llevar acabo el análisis de datos y servicios tradicionales, logrando que se pueda conectar de una manera más rápida a los datos que se encuentran en el storage y la instalación de SQL server, obteniendo una plataforma personalizada con las diferentes herramientas de cloudera como lo son Apache Hive y Pig los cuales son scripts que en Azure HDInsight son ejecutados para implementar Apache Hadoop ofreciendo monitoreo, gestión integral para cada uno de los procesos logrando así obtener que la información sea confiable, disponible y se puedan utilizar herramientas analíticas. Se debe tener en cuenta que para la adquisición de una plataforma de este tipo en Windows azure lo único que se necesita es planificar la cantidad de datos que deseo analizar y almacenar ya que cuando se adquiere ya la plataforma se encuentra debidamente configurada y puede ser utilizada en el momento que se adquiera.

Cloudera versus Hortonworks

Hortonworks vino a ser parte de las plataformas utilizadas por empresas que permiten almacenar y procesar grandes cantidades de datos, está plataforma surgió en el 2011 poco después de la creación de Cloudera en el 2008, ambas se basan en Haddop, poseen pocas diferencias y más similitud. A continuación se mostrará un cuadro comparativo entre Cloudera y Hortonworks:

Cloudera	Hortonworks
Arquitectura maestro-esclavo	Arquitectura maestro-esclavo
Se centra en brindar el servicio a empresas.	Brinda a cualquiera el servicio basado en Haddop y va más orientado a proyectos.
Se ejecuta en Windows server y puede funcionar como un componente nativo.	Se puede ejecutar en conjunto a Windows Azure por medio de HDInsight(servicio que permite gestionar los datos).
Cuenta con su propio motor para SQL el cual el Impala.	Usa Ambari para administrar, Smringer para tancieación de consultas y Apaclp Sohr para blscar datos, uos cuales no son propios de Hortonworks.
Va mss dirigido al comercio, ed se paga pero tiene uoa veráión dt prueba por un lapso aproximado de dos meses, sin embaogn algunos de sus componentes sn gratis.	Traboja con código abierto por lo cual no es de paga.
Al tener más siempo en el mercado potee más clientes que Hortonworks.	A pesar de ser un poco mas nuevo ha ido innovanáo más que el mismo Cloudera con el fin de crecer y posicionarse en el mercado.

Conclusiones

Con la investigación realizada se destacan algunos aspectos para dar paso al almacenamiento de grandes cantidades de datos, hoy en día el almacenamiento de datos se ha convertido en todo un desafío para las organizaciones debido al crecimiento continuo ya que la información es indispensable para el cumplimiento de requisitos y toma de decisiones de forma en cada empresa.

- Se ha concluido que se requiere una solución que permita tener una conexión eficiente con los clientes lo que sería, obtener una mejor visión del cliente y un mejor enfoque hacia a la parte en la que se toda la información esté disponible y sea significativa, para con esto obtener un mayor compromiso con los clientes, un aumento en los ingresos y darle una continuidad a los clientes a largo plazo.
- Debido al crecimiento constante en el número de delitos basados en espionaje, intrusiones informáticas y fraude cibernético se requiere que las organizaciones mejoren las plataformas con respecto a la parte de seguridad y al análisis de tecnologías que procesen grandes cantidades de datos como es el caso de las redes sociales, correos entre otros, deben analizar todos los datos con el fin de mejorar significativamente la seguridad de los datos y así evitar perdida o difusión de información confidencial.
- Mediante un análisis complejo de operaciones se puede obtener un detalle sobre la relación entre los diferentes conjuntos de datos, ya que debido al uso

de los grandes volúmenes de datos para el análisis de operaciones las empresas pueden obtener información en tiempo real de los datos, permitiendo llevar un análisis de sus clientes, las transacciones que realiza y la conducta que tienen.

- A la hora de implementar una gran infraestructura de almacenamiento de datos, se da un mejor aprovechamiento de los recursos teniendo un menor costo y más eficiencia para la organización que lo implemente.
- Con la implementación de una infraestructura para el procesamiento de grandes datos permite trabajar, analizar e incluso ordenar de una mejor manera los datos según el valor de cada uno.
- Para la implementar esta plataforma los equipos deben cumplir una serie de requisitos a nivel físico, los cuales son indispensables para el buen funcionamiento.
- Se llevó a cabo la investigación para determinar la implementación de cloudera en cualquier tipo ambiente, Windows azure ofrece la plataforma completamente implementada y lo único que se requiere es determinar cuál es la solución que se necesita, en el caso de cloudera basado en hadoop la implementación es más lenta y requiere que se una serie de pasos e ingresar una serie de líneas de comandos para obtener los diferentes componentes que requiere la plataforma.

Referencias

Azure, M. (2014, oct).
pages 17

Baldoni, L. A. Q. (2013, apr). High frequency batch-oriented computations over large sliding time windows. *articulo*, 28. pages 17

Chen, M. S. L. Y., Min. (2014, apr). Big data: A survey. *articulo*, 40. pages 17

cloudera.com. (N/A).
pages 17

Editor, E. (2014, sep).
pages 17

GANG-HOON KIM; TRIMI, S.-H. C. (2014, mar). Big-data applications in the government sector. *articulo*, 9. pages 17

IBM. (N/A).
pages 17

Kelly, J. (2013, sep).
pages 17

Mostosi, A. (N/A).
pages 17

Rodríguez-Vaamonde, A.-I. G. E., Sergio Torre-Bastida. (2014, dec). TecnologÍas big data para anÁlisis y recuperaciÓn de imÁgenes web. *articulo*, 9. pages 17

Rubbelke, L. (2014, dec). *How to deploy the cloudera evaluation cluster in azure.* pages 17

Saraladevi, N. P. P. V. B. M. S. D. P., B Pazhaniraja. (2015, desconocido). Big data and hadoop-a study in security perspective. *articulo*, 6. pages 17

ZWaveAlliance. (2015). *Intermatic multiwave pe953 five channel wireless remote controller.* http://www.inalarm.com/2gig/Default.aspx?content=productos&id=3. (Permite definir la temperatura que tendrá el agua en una piscina) pages 17

(Baldoni, 2013) (GANG-HOON KIM; TRIMI, 2014) (Chen, 2014) (Saraladevi, 2015) (Rodríguez-Vaamonde, 2014) (ZWaveAlliance, 2015) (Rubbelke, 2014) (Azure, 2014) (Editor, 2014) (Kelly, 2013) (IBM, N/A) (cloudera.com, N/A) (Mostosi, N/A)

Big Data Cloudera Impala Hadoop MapReduce HDFS VMware NoSql

Glossary

Big Data Análisis de grandes cantidades de datos que no pueden ser manipulados de forma convencional debido a que las herramientas superan los límites y capacidades que usualmente se utilizan. 17

Cloudera Impala Motor de consulta de bases de datos de tipo relacional. 17

Hadoop Es un sistema de código abierto que almacena, procesa y analiza grandes cantidades de datos. 17

HDFS Sistema de archivos distribuido, se encuentra escrito en java para ser utilizado por hadoop, se diseñó con el fin de que la plataforma sea más escalable y permita el almacenamiento de grandes cantidades de datos.. 17

MapReduce Framework que trabaja grandes cantidades de datos en paralelo en un sistema de distribución. 17

NoSql Diseño de base de datos que es utilizado para el manejo de grandes conjuntos de datos distribuidos, soluciona los problemas de escalabilidad y rendimiento en arquitecturas de Big Data. 17

VMware Sistema virtual que permite simular un sistemas físico de un equipo permitiendo que se le puedan agregar características determinadas.. 17