Universidad Técnica Nacional

Investigación Grupal

Tema: Hadoop

Curso:

ISW-412

Integrantes:

Luis Eduardo Jiménez Mata

Warren Carvajal Hernández

Rudy Castro Marín

Profesor:

Efrén Jiménez Delgado

Fecha de entrega:

Viernes 24 de febrero de 2017

Tabla de contenido

Introducción

Se dice por ahí que la mejor arma es el conocimiento y no es para menos pues gracias al conocimiento, gracias a un simple dato recolectado, se han ganado guerras en el pasado, gracias a un simple dato se han podido frenar catástrofes, gracias a un simple dato se han podido curar enfermedades. No es para menos entonces que el ser humano haya dado tanta importancia al conocimiento. Ahora bien, si nos movemos en el tiempo hacia la actualidad podemos ver que vivimos en una época diferente, una época donde la información paso de ser algo transmitido por las palabras y el papel a unos cuantos a usar medios de comunicación virtuales para hacer dicha información pública al mundo. Y es ahí donde nos hemos dado cuenta de la importancia de saber manipular esa información de manera segura, correcta y ágilmente.

De esta necesidad es de donde nacen tecnologías como la que veremos a continuación.

Resumen Ejecutivo

Producto de la transmisión masiva de información es que cada día es más y más necesario procesar esa información de manera eficaz. No solo el procesamiento es necesario, la seguridad, el almacenamiento y la flexibilidad son indispensables.

De ahí que Apache Hadoop sea una herramienta tan eficaz. El mismo es un framework que permite el procesamiento de grandes volúmenes de datos a través de clusters, usando un modelo simple de programación. Además su diseño permite pasar de pocos nodos a miles de nodos de forma ágil. Hadoop es un sistema distribuido usando una arquitectura Master-Slave, usando para almacenar su Hadoop Distributed File System (HDFS) y algoritmos de MapReduce para hacer cálculos. Es una Tecnología altamente escalable, posee un almacenamiento a bajo costo con una flexibilidad y velocidad excepcionales.

Historia

Hadoop fue creado por Doug Cutting, que lo nombró así por el elefante de juguete de su hijo.

Fue desarrollado originalmente para apoyar la distribución del proyecto de motor de búsqueda, denominado Nutch.

Caracteristicas

- Es adecuado para el almacenamiento y procesamiento distribuido.
- Hadoop proporciona una interfaz de comandos para interactuar con HDFS.
- Los servidores de namenode datanode y ayudar a los usuarios a comprobar fácilmente el estado del clúster.
- Streaming el acceso a los datos del sistema de ficheros.
- HDFS proporciona permisos de archivo y la autenticación.
- Hadoop distribuye los datos y los procesa en paralelo en los nodos donde los datos se encuentran localizados.
- Es capaz de mantener múltiples copias de los datos y automáticamente hacer un redespliegue de las tareas. El aspecto clave de Hadoop es que en lugar de mover los datos hacia donde se hace el procesamiento, Hadoop mueve el procesamiento (Tasks) a donde están los datos.

Componentes

El proyecto Hadoop consta de los siguientes módulos:

- **Hadoop Common**: Son las utilidades comunes que soportan a los demás módulos Hadoop.
- **Hadoop Distribuited File System (HDFS):** Es un sistema distribuido de archivos que provee un acceso de alto rendimiento a los datos de la aplicación.

- **Hadoop YARN**: Es un framework para programar tareas y gestionar los recursos del cluster.

- **Hadoop MapReduce**: Es un sistema basado en YARN para procesamiento en paralelo de grandes conjuntos de datos.

Además, otros componentes de los que se sirve hadoop son:

- **Ambari:** Es una herramienta web para aprovisionar, gestionar y monitorear los clusters Apache Hadoop.

- **Avro**: Es un sistema de serialización de datos.

- **Cassandra**: Es una base de datos multi-maestro en configuración non-single-failure.

- **Chukwa**: Es un sistema de recolección de datos para gestionar grandes sistemas distribuidos.

- **HBase:** Es una base de datos escalable y distribuida, que soporte almacenamiento de datos estructurados para tablas grandes.

- **Hive:** Es una infraestructura data-warehouse que provee resumen de datos y consultas ad-hoc.

- **Mohout:** Es una librería escalable de aprendizaje de máquina y de minería de datos.

- **Pig:** Es un framework para procesamiento en paralelo, con un lenguaje de alto nivel de flujo de datos.

- **Spark:** Es un motor de computo rápido y general para los datos Hadoop. Provee un modelo de programación simpre y expresivo que soporta un amplio rango de aplicaciones, incluyendo ETL, aprendizaje de máquina, flujos de procesos y computación gráfica.

- **Tez:** Es un framework de programación de flujo de datos, construido sobre Hadoop YARN.

- **ZooKeeper:** Es un serivicio de coordinación de alto rendimiento para aplicaciones distrubuidas.

Apache Hadoop es una herramienta muy potente y al alcance de cualquiera para iniciarse en el mundo del manejo de big data, y el procesamiento en paralelo de grandes cantidades de información.

Ventajas

1. **Tecnología altamente escalable:** Un clúster de Hadoop puede crecer simplemente añadiendo nuevos nodos. No es necesario hacer ajustes que modifiquen la estructura inicial. Por lo tanto, nos permite un crecimiento fácil, sin estar atados a las características iniciales del diseño. Gracias al procesamiento distribuido de MapReduce, los archivos se dividen en bloques de forma sencilla.
2. **Almacenamiento a bajo costo:** La información no se almacena de forma predefinida, en filas y columnas, como ocurre con las bases de datos tradicionales, sino que Hadoop asigna datos categorizados a través de miles de computadoras baratas, y ello supone un gran ahorro.
3. **Flexibilidad:** Al incrementar el número de nodos del sistema también ganamos en capacidad de almacenamiento y procesamiento. A su vez, es posible agregar o acceder a nuevas y diferentes fuentes de datos (estructurados, semiestructurados y no estructurados), al tiempo que existe la posibilidad de adaptar herramientas accesorias que funcionan en el entorno Hadoop y ayudan en el diseño de procesos, la integración o mejorar otros aspectos.
4. **Velocidad:** Permite ejecutar procesamientos y realizar análisis muy rápidos.
5. **Tolerante a fallos:** Hadoop es una tecnología que facilita almacenar grandes volúmenes de información, lo que a su vez permite recuperar datos de forma segura. Si un equipo se cae, siempre hay otra copia disponible, con lo que es posible la recuperación de datos en caso de producirse fallos.

Desventajas

En lo que respecta al HDFS:

1. Latencia para el acceso a datos: HDFS está orientado a procesos batch y operaciones en streaming. Por lo tanto, la latencia de cualquier operación IO no ha sido optimizada y sistemas de archivos tradicionales (como ext4, XFS...) suelen ser más rápidos en estos aspectos.

2. Cantidades grandes de ficheros pequeños: El límite del número de ficheros en este sistema está limitado por la memoria del NameNode, que es en su RAM donde se encuentran los metadata.
3. Escribe una vez, lee varias: En HDFS los ficheros solo se pueden escribir una vez.
4. No se puede acceder con los comandos tradicionales de Linux (ls, cat, vim...). Y aunque exista "HDFS fuse" para montar HDFS como cualquier otro sistema de archivo Linux, esta solución no ofrece un buen rendimiento.

En lo que respecta a MapReduce:

1. Es muy difícil de depurar: Al procesarse el programa en los nodos donde se encuentran los bloques de datos, no es fácil encontrar los fallos de código.
2. No todos los algoritmos se pueden escribir con el paradigma MapReduce.
3. Latencia: cualquier job MapReduce suele tardar por lo menos 10 segundos.

Uso en el mercado

Hadoop se puede utilizar en teoría para casi cualquier tipo de trabajo batch, mejor que para trabajos en tiempo real, ya que son más fáciles de dividir y ejecutar en paralelo. Entre los campos actuales a aplicación se encuentran:

- Análisis de logs
- Análisis de mercado
- Data mining
- Procesamiento de imágenes
- Procesamiento de mensajes XML
- Web crawling
- Indexación de textos

Algunas empresas que lo usan: Yahoo, Facebook, Amazon A9, The New York Times.

Para que funciona

Apache Hadoop es un “framework” (marco de trabajo) que permite procesamiento distribuido de grandes conjuntos de datos, a través de “clusters” (grupos) de computadores, usando modelos simples de programación. Está diseñado para escalar desde pocos servidores a miles de servidores, cada uno ofreciendo su propio almacenamiento y procesamiento local.

Tipo de licencia comercial

La licencia Apache (Apache License o Apache Software License para versiones anteriores a 2.0) es una licencia de software libre permisiva creada por la Apache Software Foundation (ASF).6 La licencia Apache (con versiones 1.0, 1.1 y 2.0) requiere la conservación del aviso de derecho de autor y el descargo de responsabilidad, pero no es una licencia copyleft, ya que no requiere la redistribución del código fuente cuando se distribuyen versiones modificadas.

Al igual que otras licencias de software libre, todo el software producido por la ASF o cualquiera de sus proyectos está desarrollado bajo los términos de esta licencia, es decir, la licencia permite al usuario del software de la libertad de usar el software para cualquier propósito, para distribuirlo, modificarlo y distribuir versiones modificadas del software, bajo los términos de la licencia, sin preocuparse de las regalías.

Diferentes versiones

3.0.0-alpha2 25 January, 2017

3.0.0-alpha1 03 September, 2016

2.7.3 25 August, 2016

2.6.5 08 October, 2016

2.5.2 19 Nov, 2014

Futuro de la herramienta

La gran tendencia Big Data analytics en los últimos meses está siendo focalizada en la inteligencia artificial (IA), en sus varias formas y sabores. La tendencia lógica es, una vez cubierta la parte de captura e infraestructura, buscar ayuda a la hora de analizar conjuntos de datos masivos y predecir insights.

Gran parte del motivo de la reciente resurrección de la IA es debido al Big Data. Los algoritmos detrás del deep learning (área de IA que más atención recibe últimamente) fueron creados en su gran mayoría hace décadas, pero hasta ahora no habían podido aplicarse a grandes volúmenes de datos de manera suficientemente rápida y a unos costes asumibles. Al fin han podido alcanzar su potencial. La relación entre la IA y el Big Data es más cercana que nunca y están destinados a ir de la mano en el futuro.

Conclusión

Hadoop como tecnología de manejo de información como vimos mantiene un perfil de excelencia siendo este confiable a la vez barato. Cabe rescatar que aunque posee ciertas desventajas en procesamiento de grandes combos de información sus demás características la vuelven una herramienta realmente útil. Ante este mundo de la información es importante que herramientas como estas sigan avanzando y evolucionando para conformar así la base de futuras bases de datos.

Recomendaciones

1. Como se vio existen diferentes componentes para Hadoop, por lo que se recomienda una investigación más afondo de cada uno de ellos para la utilización del mismo.
2. El futuro de esta herramienta es incierto y podrían existir otras posibilidades, en este documento se cita solo una de ellas.

Bibliografía

http://hadoop.apache.org/releases.html

http://searchdatacenter.techtarget.com/es/cronica/Explorando-distribuciones-Hadoop-para-gestionar-big-data

https://unpocodejava.wordpress.com/2012/08/29/un-poco-de-hadoop/

https://www.tutorialspoint.com/es/hadoop/hadoop_hdfs_overview.htm

https://unpocodejava.wordpress.com/2012/08/29/un-poco-de-hadoop/

http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/387622/Las-ventajas-definitivas-de-utilizar-Hadoop

https://indizen.com/5-ventajas-que-obtendra-tu-organizacion-al-incorporar-hadoop/

http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/402826/5-ventajas-de-la-arquitectura-de-Hadoop

https://www-01.ibm.com/software/cl/data/infosphere/hadoop/que-es.html

https://es.wikipedia.org/wiki/Hadoop#Historia

https://indizen.com/5-ventajas-que-obtendra-tu-organizacion-al-incorporar-hadoop/

http://fireosoft.com.co/blogs/que-es-apache-hadoop/