
Universidad
Tecmilenio®

Universidad Tecmilenio

Infraestructura para Big Data

Introducción a Hadoop

Semana 9

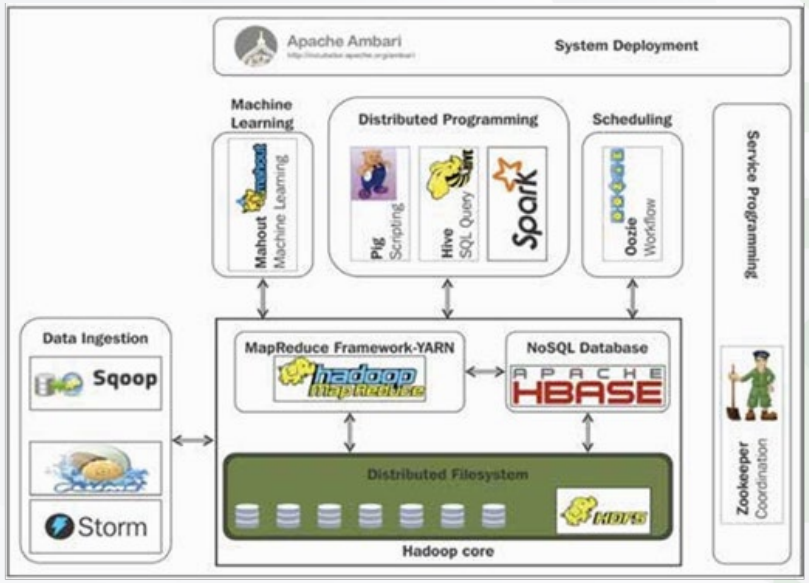
Introducción

Dentro de las variadas soluciones para el manejo de grandes volúmenes de datos, Hadoop supera a sus competidores al colocarse en uno de los principales lugares de popularidad, gracias a todo su ecosistema de código abierto, creado por la comunidad de Apache Software Foundation, que incluye diversas soluciones para preprocesamiento, análisis y visualización de datos, además de contar con un sistema propio de archivos distribuidos (HDFS).

Explicación

Hadoop es un *framework* (conjunto de tecnologías y funcionalidades estandarizadas, entrelazadas y empaquetadas para su uso inmediato dentro de otros aplicativos y desarrollos) de software abierto para almacenar, procesar y analizar grandes cantidades de datos en clústeres de equipos de cómputo, que invocamos a través de una librería para poder utilizarlo dentro del código (Marr, s.f.).

Núcleo de Hadoop



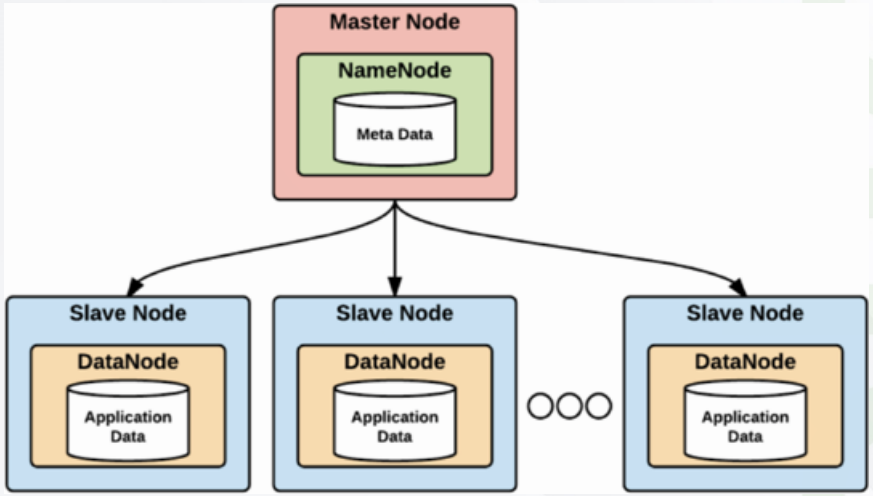
Fuente: Bahari, B. (2017). *Hadoop Fundamental.* Recuperado de https://medium.com/sarccom/hadoop-fundamental-5179099f5b2

Fue creado por Doug Cutting y Mike Cafarella. Su origen proviene del proyecto de un buscador web de nombre Apache Nutch, que es parte del proyecto Apache Lucene, una librería de búsqueda de texto. A partir del año 2008, Hadoop es el proyecto de mayor prioridad de Apache.

Explicación

Arquitectura

Hadoop trabaja en un ambiente de "maestro-esclavo", donde un equipo maestro puede estar formado de varios esclavos (computadoras), que pueden estar distribuidos en el mismo equipo, centro de datos o en cualquier parte del mundo.



Fuente: Tomcy, J., y Pankaj, M. (2017). *Data Lake for Enterprises*. Estados Unidos: Packt Publishing.

Las dos características más importantes de Hadoop son (Maklin, 2019):

Es un sistema distribuido de archivo que suministra una gran consistencia en los datos y provee mecanismos tolerantes a las fallas en cualquiera de los nodos donde se encuentra la información.

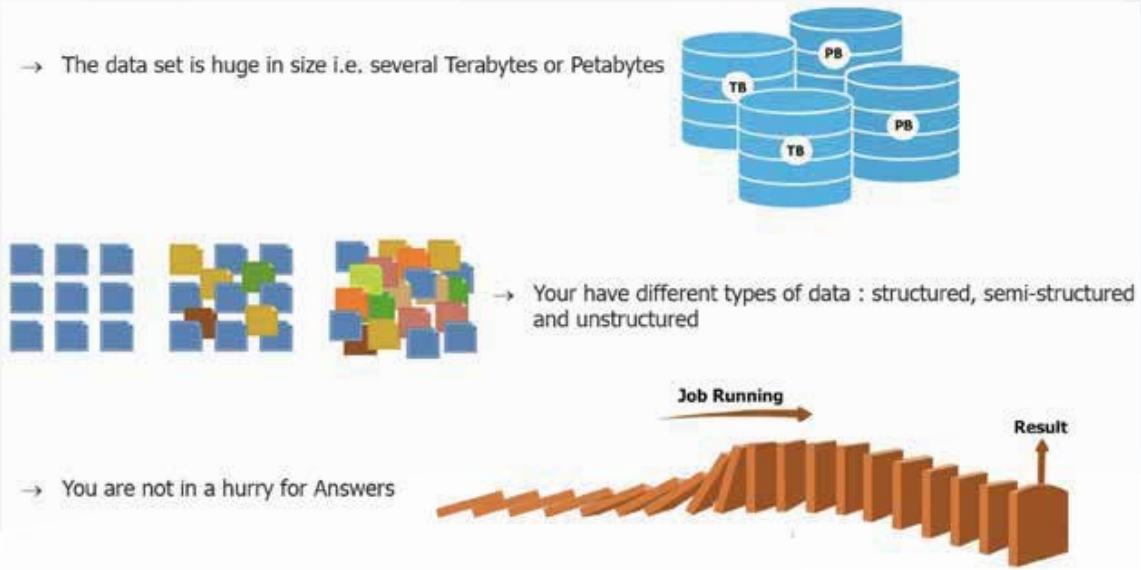
Ofrece un sistema de análisis que puede ejecutar tareas de cómputo en grandes conjuntos de información. Hadoop MapReduce es el responsable de ejecutar todas las tareas de procesamiento de datos, al dividirlas en múltiples tareas más pequeñas que son asignadas a los diferentes nodos; se encarga, además, de la coordinación y consolidación de los resultados.

Explicación

Puede utilizarse en gran variedad de escenarios, tales como (Leslie, 2018):

- **Análisis y retención de datos.**
- **Análisis de *logs* (archivos de registro) de información de equipos de cómputo.**
- **Análisis de texto, imágenes y/o video.**

Es importante recordar que los datos pueden provenir de diferentes fuentes (estructuradas, no estructuradas y semiestructuradas), además pueden ser de diferentes tipos. Si el proyecto o la empresa cumple con estas dos características, es seguro que Hadoop será la solución indicada.



Fuente: Chaturvedi, V. (2020). *5 Reasons When to and When not to use Hadoop.* Recuperado de https://www.edureka.co/blog/5-reasons-when-to-use-and-not-to-use-hadoop/

Explicación

Debes saber distinguir correctamente las necesidades. Aunque Hadoop es una excelente herramienta, implementarla equivocadamente te "cobrará" en desempeño y efectividad.

Se recomienda no usar Hadoop en los siguientes casos (Tomcy y Pankaj, 2017):

1. Cuando se requieren procesar grandes cantidades de datos en el menor tiempo posible.
2. Cuando se requiere un análisis en tiempo real.
3. Múltiples *datasets*, pero muy pequeños, por ejemplo, archivos de Excel con menos de un millón de registros.

Cierre

Gracias a las características, funcionalidades y beneficios de Hadoop, podemos procesar grandes volúmenes de información de forma distribuida a bajo costo y con gran confiablidad, permitiendo que muchos de los proyectos de Big Data en el mercado sean posibles. Sin olvidar que muchas herramientas también tienen retos y Hadoop no es la excepción (por ejemplo, la gestión y seguridad de los datos), por lo que podemos establecer que es la base de la mayoría de los proyectos de Big Data, los cuales abarcan una gran cantidad de sectores, como la mercadotecnia, las telecomunicaciones, el retail, los bancos, el sector público, entre otros.

Bibliografía

- Bahari, B. (2017). *Hadoop Fundamental.* Recuperado de https://medium.com/sarccom/hadoop-fundamental-5179099f5b21
  Chaturvedi, V. (2020). *5 Reasons When to and When not to use Hadoop.* Recuperado de https://www.edureka.co/blog/5-reasons-when-to-use-and-not-to-use-hadoop/
- Leslie, A. (2018). *What is Hadoop Good For? (Best Uses, Alternatives, & Tools).* Recuperado de https://www.hostingadvice.com/how-to/what-is-hadoop/
- Maklin, C. (2019). *Apache Hadoop — What Is YARN | HDFS | MapReduce.* Recuperado de https://towardsdatascience.com/big-data-what-is-apache-hadoop-3dafda16c98e
- Marr, B. (s.f.). *What Is Hadoop?* Recuperado de https://www.bernardmarr.com/default.asp?contentID=1080
- Tomcy, J., y Pankaj, M. (2017). *Data Lake for Enterprises.* Estados Unidos: Packt Publishing.

Universidad Tecmilenio

Infraestructura para Big Data

Arquitectura Hadoop

Semana 9

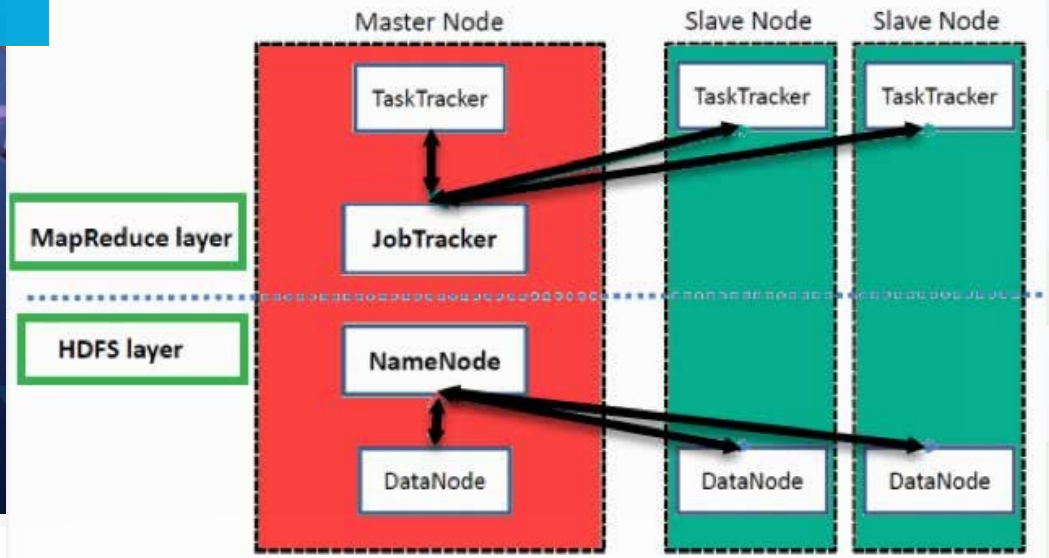
Introducción

Escalabilidad, alta disponibilidad y buen funcionamiento del sistema son las características esenciales de la arquitectura de Hadoop, convirtiéndola en un sistema de procesamiento y almacenamiento de una inmensa cantidad de datos, siendo muy flexible, rápida y robusta.

Al implementarse correctamente, la arquitectura Hadoop se acerca a los tres puntos ideales del tratamiento de datos del teorema CAP de Eric A. Brever: consistencia, disponibilidad y tolerancia a fallos (Mehra, 2019)

Explicación

Hadoop cuenta con una arquitectura basada en una topología maestro-esclavo (master-slave) para almacenar la información y procesarla de forma distribuida. Esta se conforma principalmente por dos elementos:
MapReduce y HDFS (Hadoop File System) (Kaplarevic, 2020).



Fuente: Guru99. (s.f.). *What is Hadoop? Introduction, Architecture, Ecosystem, Components.* Recuperado de https://www.guru99.com/learn-hadoop-in-10-minutes.html

Explicación

Existen otros elementos en la infraestructura de Hadoop que conforman el ecosistema completo de una solución de Big Data, tales como (Digital Guide IONOS by 1&1, 2019):

- ✓ **YARN:** Yet Another Resource Negotiator.
- ✓ **Spark:** procesamiento de datos en memoria.
- ✓ **PIG, HIVE:** procesamiento de datos basado en queries.
- ✓ **HBase:** base de datos NoSQL.
- ✓ **Mahout, Spark MLLib:** librerías de machine learning.
- ✓ **Solar, Lucene:** búsqueda e indexación.
- ✓ **Zookeeper:** administración del clúster.
- ✓ **Oozie:** administración de tareas.

Explicación

El *workflow* (flujo de trabajo) para la infraestructura de Hadoop es el siguiente (Sistla, 2021):

1. **Las aplicaciones del cliente envían solicitudes de tareas al JobTracker.**
2. **El JobTracker se comunica con el NameNode para determinar la localización de la información.**
3. **El JobTracker localiza el TaskTracker en los nodos con la información solicitada.**
4. **El JobTracker envía la tarea al TaskTracker seleccionado.**
5. **El TaskTracker es monitoreado para saber si se está recibiendo la información solicitada, de lo contrario, se busca un TaskTracker diferente.**
6. **El TaskTracker puede notificar al JobTracker en caso de falla y determina si buscar otro TaskTracker o enviar una notificación de falla.**
7. **Las aplicaciones del cliente envían solicitudes de tareas al JobTracker.**
8. **El JobTracker se comunica con el NameNode para determinar la localización de la información.**

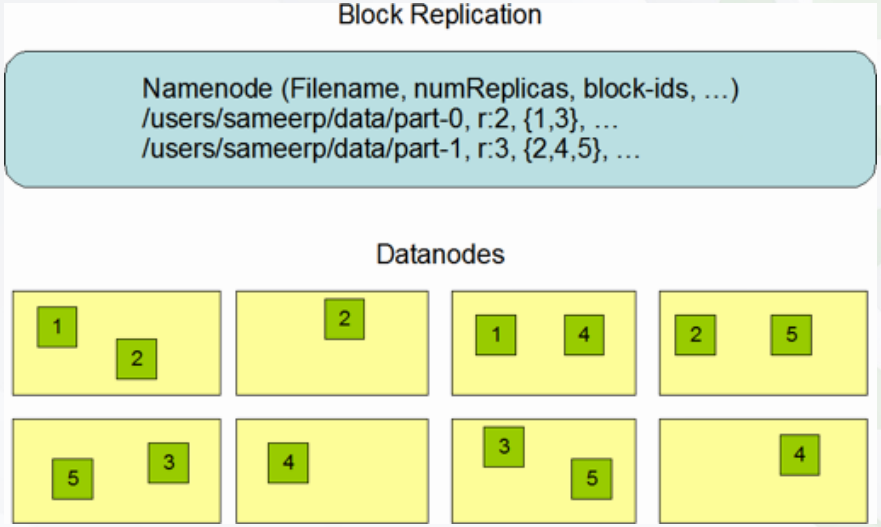
Explicación

HDFS (Hadoop File System)

Provee el almacenamiento de los datos de Hadoop. Divide los archivos en pequeñas unidades llamadas bloques y los almacena de forma distribuida (Apache Hadoop, 2020-a).

Los bloques en HDFS son la unidad de almacenamiento contiguo más pequeña en Hadoop.

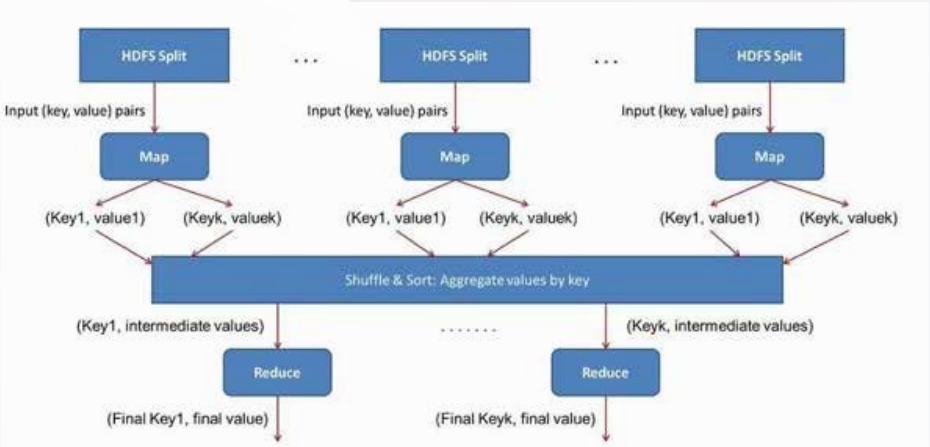
HDFS replica los bloques en diferentes DataNodes para proveer tolerancia a las fallas. Por defecto se hace la replicación en tres diferentes nodos, pero se puede configurar para que sea en un número mayor.



Fuente: ApacheHadoop. (2020-b). *HDFS Architecture*. Recuperado de https://hadoop.apache.org/docs/r1.2.1/images/hdfsdatanodes.gif

Explicación

MapReduce



Fuente: Big Data and Analytics-(BigAn). (2017). *Hadoop Reduce*. Recuperado de https://thebigan.wordpress.com/2017/08/28/hadoop-mapreduce/

En la primera parte del framework ocurre el Map, que contiene los siguientes pasos:

- ✓ **RecordReader:** transforma los datos divididos de entrada en registros (en pares), donde la llave es la posición y el valor del dato que compone al registro.
- ✓ **Map:** es una función definida por el usuario para procesar el par de datos generados.
- ✓ **Combiner:** agrupa los datos de la parte del Map (es opcional).
- ✓ **Partitioner:** funge como la condición al procesar los datos de entrada.

En la segunda parte del *framework* ocurre el Reduce, que contiene los siguientes pasos:

- ✓ **Shuffle and Sort:** se encarga de descargar la información del *partitioner* a la máquina que se va a encargar de hacer el Reduce.
- ✓ **Reduce:** se encarga de hacer la función de Reduce para cada par de datos.
- ✓ **Output format:** este es el paso final, toma el par de datos del *reducer* y escribe la información en el archivo de salida.

Es la capa que procesa la información en Hadoop; un framework que permite escribir aplicaciones para obtener grandes volúmenes de información.

Su función consiste usualmente en dividir la información en un conjunto de datos de entrada independientes que son procesados por la actividad de Map de forma paralela. El framework ordena los resultados de salida del Map y estos, a su vez, se convierten en los datos de entrada de la actividad Reduce.

Cierre

Hadoop, al ser un proyecto de código abierto, está en una constante mejora por parte de la comunidad internacional de desarrolladores, lo cual lo convierte en un ecosistema cada vez más rico, fortalecido y seguro.

No hay que olvidar que mientras más módulos se implementen, más complejo se volverá el ecosistema y debemos valorar el buen diseño de los repositorios, esquemas de archivos y rutinas para aprovechar al máximo el potencial de la arquitectura Hadoop.

Bibliografía

- Apache Hadoop. (2020-a). *Apache Hadoop 3.3.0*. Recuperado de https://hadoop.apache.org/docs/r3.3.0/index.html
- Apache Hadoop. (2020-b). *HDFS Architecture*. Recuperado de https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
- Big Data and Analytics-(BigAn). (2017). *Hadoop Reduce*. Recuperado de https://thebigan.wordpress.com/2017/08/28/hadoop-mapreduce/
- Digital Guide IONOS by 1&1. (2019). *Apache Hadoop: sistema de archivos distribuido*. Recuperado de https://www.ionos.mx/digitalguide/servidores/know-how/apache-hadoop-el-framework-para-big-data/
- Guru99. (s.f.). *What is Hadoop? Introduction, Architecture, Ecosystem, Components*. Recuperado de https://www.guru99.com/learn-hadoop-in-10-minutes.html
- Kaplarevic, V. (2020). *Apache Hadoop Architecture Explained (With Diagrams)*. Recuperado de https://phoenixnap.com/kb/apache-hadoop-architecture-explained
- Mehra, A. (2019). *Understanding the CAP Theorem*. Recuperado de https://dzone.com/articles/understanding-the-cap-theorem
- Sistla, D. (2021). *Hadoop Architecture Explained-What it is and why it matters*. Recuperado de https://www.dezyre.com/article/hadoop-architecture-explained-what-it-is-and-why-it-matters/317