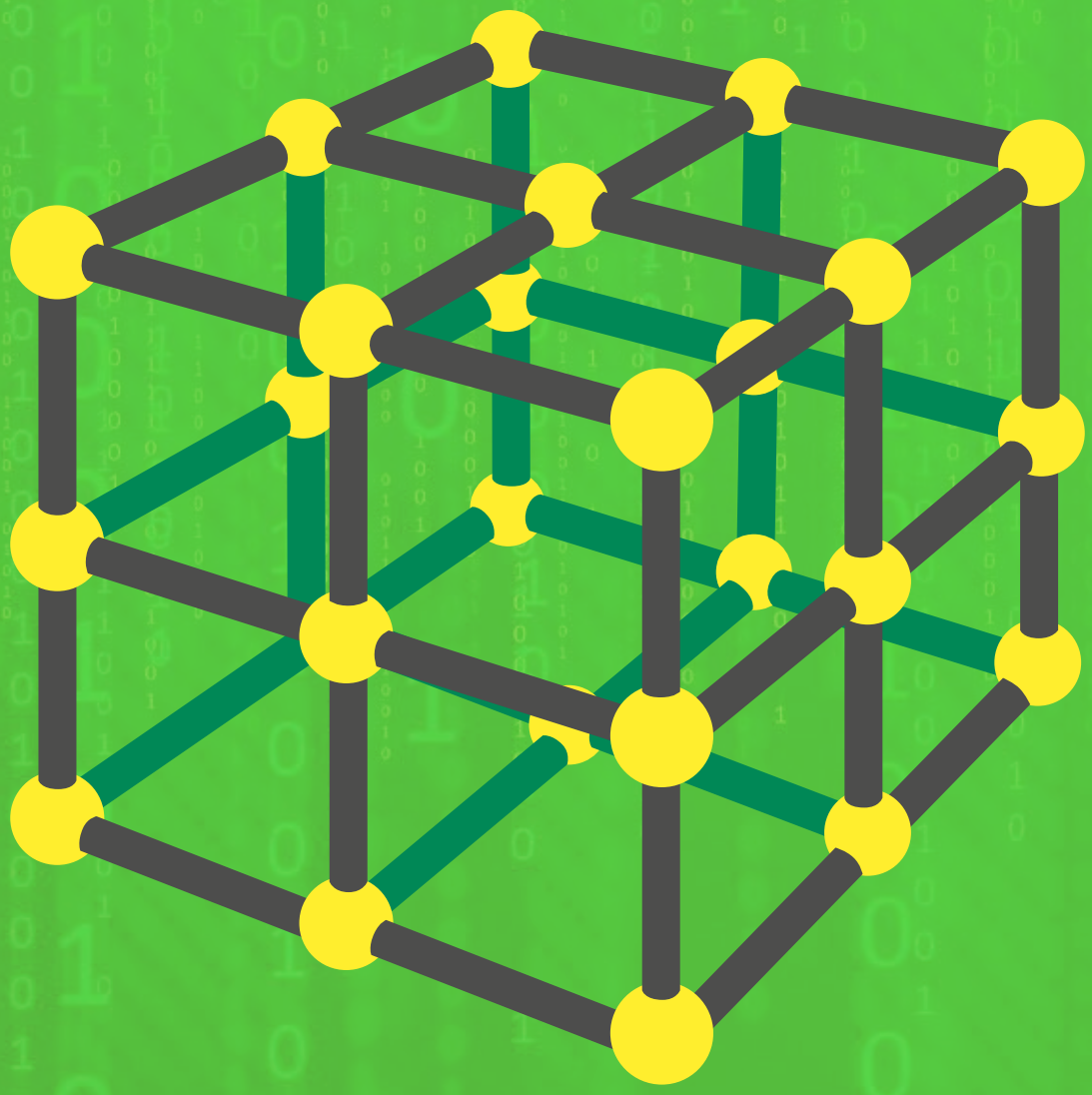
PowerData

Especialistas en Gestión de Datos

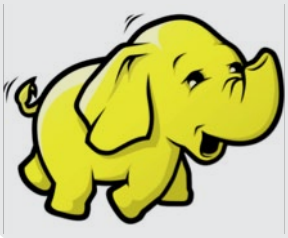
¿Qué significa Hadoop en el mundo del Big Data?

Un contenido para perfiles técnicos



ÍNDICE

¿Qué significa Hadoop en el Universo Big Data?



¿Qué sucede cuando las técnicas de análisis tradicionales se encuentran con sus límites? ¿Cuándo llega el momento en que la minería de datos no aporta las soluciones esperadas? ¿Cómo se enfrentan los clusters de Hadoop al desafío de los grandes datos y su expresión más desestructurada? ¿Es Hadoop una buena opción para mi negocio?

El planteamiento: big data y data science

El concepto big data hace referencia a los sistemas que manipulan grandes conjuntos de datos, también conocidos como data sets. Entre sus principales cualidades se encuentran la **heterogeneidad** y la **volatilidad**, como datos que son. Sin embargo, son su **volumen** y su **velocidad** de generación las que plantean mayores dificultades a la hora de trabajar a un nivel **big data** en el entorno empresarial. Los retos tienen que ver con:

- Captura de datos.
- Almacenamiento de tales volúmenes de información.
- Capacidad de realizar búsquedas eficientes.
- Compartición.
- Posibilidad de llevar a cabo análisis efectivos.
- Visualización de los datos.

© Globe On Binary Background" by digitalart

Además de los desafíos, big data supone un nuevo mundo de oportunidades para las empresas. Aquí es donde entra en juego el **data science**. Este concepto es algo más genérico y hace referencia a las técnicas necesarias para manipular y tratar la información desde un punto de vista estadístico/matemático.

Data Science está basado en algoritmos, aplicados al problema de big data, entre otros. Incorporar la figura del data scientist en la organización implica dejar atrás las conocidas limitaciones del data mining. Esta evolución traspasa las fronteras de una simple query, hallando correlaciones, aplicando algoritmos más complejos y proporcionando unos niveles de visibilidad que transforman por completo el contacto de la empresa con su entorno, otorgándole la capacidad, por primera vez, de descubrir y estudiar sus oportunidades.

Los desafíos que big data lanza y la solución que propone Hadoop

> La mayoría de las empresas estiman que sólo analizan el 12% de los datos que tienen, dejando 88% de ellos en la sala de montaje
>
> Encuesta de Forrester Software Q4, 2013.

El software de Hadoop surge para resolver parte de los problemas asociados a big data y a la aparición del data science. Entre sus puntos clave se encuentran su capacidad de almacenamiento y procesamiento local. Partiendo de ellos:

- Consigue escalar desde unos pocos servidores hasta miles de máquinas, todas ellas ofreciendo idéntica calidad de servicio.
- Permite el procesamiento distribuido de grandes conjuntos de datos en clusters de computadoras utilizando modelos sencillos de programación.

Se trata, en definitiva, de un proyecto de desarrollo de software orientado hacia la computación distribuida, donde la escalabilidad y la fiabilidad son los dos atributos más importantes. En otras palabras, Hadoop completa el círculo, erigiéndose en complemento perfecto de big data porque:

- **Simplifica la interacción** con su aportación informativa.
- **Economiza los procesos**.
- **Palia las carencias** que big data puede presentar de cara al usuario.

	TRADITIONAL RDBMS	MAPREDUCE
Data Size	Gigabytes *(Terabytes)*	Petabytes *(Hexabytes)*
Access	Interactive and Batch	Batch
Updates	Read / Write many times	Write once, Read many times
Structure	Static Schema	Dynamic Schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear
DBA Ratio	*1:40*	*1:3000*

Reference: Tom White's Hadoop: The Definitive Guide

Pero, ¿cómo lo hace? ¿Cómo consigue dejar atrás los límites y superar las dificultades?

Los dos conceptos en los que se apoya Hadoop son, por un lado, la técnica de MapReduce y, por otro, el sistema distribuido de archivos HDFS. Una primera toma de contacto con estos términos revelará que:

- **HDFS** (Hadoop Distributed File System): es un sistema de archivos distribuido, escalable y portátil típicamente escrito en JAVA.
- **MapReduce**: es el modelo de programación utilizado por Google para dar soporte a la computación paralela. Trabaja sobre grandes colecciones de datos en grupos de computadoras y sobre commodity hardware.

Sin embargo, para entender cómo funciona Hadoop hace falta adentrarse en estos dos conceptos y profundizar en las implicaciones que cada uno de ellos tiene en la interacción con big data.

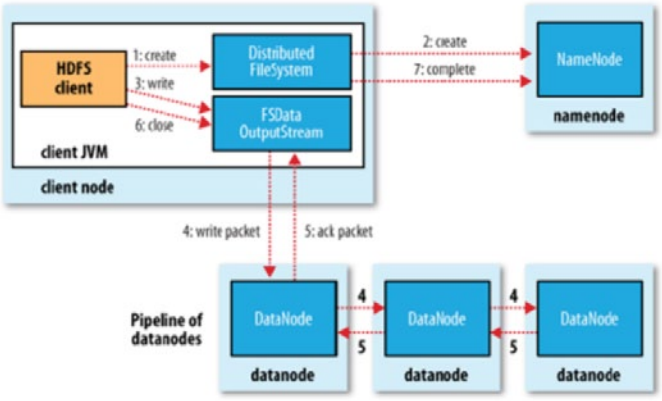
HDFS y MapReduce: estructura, características y fases

¿Quiénes conocen y usan Hadoop? Conócelos todos, de Amazon a Zvents. La respuesta aquí.

Arquitectura básica de HDFS

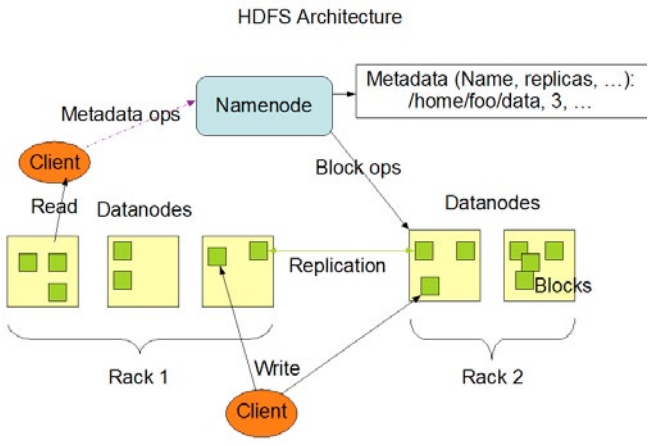
Este sistema de archivos sobre el que se estructura Hadoop cuenta con tres pilares básicos:

- **Namenode**: se ocupa del control de acceso y tiene la información sobre la distribución de datos en el resto de nodos.
- **Datanodes**: son los encargados de ejecutar el cómputo, es decir, las funciones Map y Reduce, sobre los datos almacenados de manera local en cada uno de dichos nodos.
- **Jobtracker**: este nodo se encarga de las tareas y ejerce el control sobre la ejecución del proceso de MapReduce.



© megaupl0ad.com

Principales características de HDFS



© Cloudera

Las cualidades de HDFS son la clave que aporta a Hadoop la versatilidad, segurida d y eficiencia que se echaba de menos en el trabajo con big data. Ello se pone de manifiesto en sus 5 principales características:

- **Tolerancia a fallos**: que consigue que no se pierda la información ni se generen retrasos. De hecho, incluso aunque se produzca la caída de algunos datanodes, el cluster sigue funcionando.
- **Acceso a datos en streaming**: los datos son facilitados a medida que se consumen, por lo que no hace falta descargarlos.
- **Facilidad para el trabajo con grandes volúmenes de datos**: los clusters de Hadoop están preparados para almacenar grandes ficheros de todo tipo.
- **Modelo sencillo de coherencia**: ya que no implementa la regla POSIX en un 100% para poder aumentar los ratios de transferencia de datos.
- **Portabilidad de convivencia entre hardware heterogéneo**: Hadoop puede correr en máquinas de distintos fabricantes.

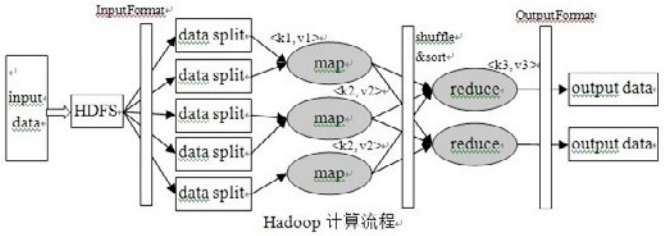
Fases de MapReduce

Su ejecución consta de dos fases principales, Map y Reduce, ambas programadas por el desarrollador. Se completan con una etapa interna, llamada "Shuffle and sort", que permite vincular las dos fases anteriores. Cada una de ellas tiene una función:

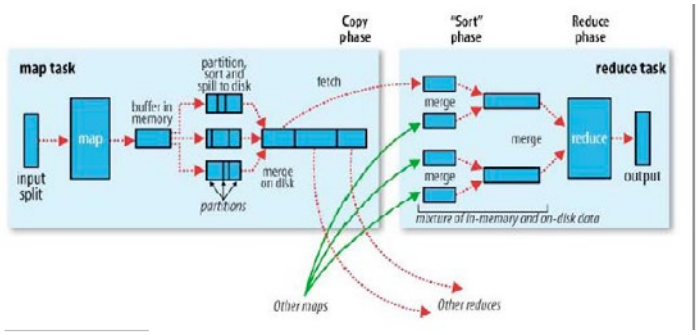
- **Map**: se aplica en paralelo para cada ítem en la entrada de datos. Gracias a este mapeo, a cada llamada se asignará una lista de pares (key/value, también conocidos como clave/valor). El resultado es la creación de un grupo por cada clave generada. El framework de mapreduce agrupará todos los pares con la misma clave de todas las listas.

- **Shuffle and sort**: tiene dos misiones. Por una parte, se encarga de ordenar por clave (key) todos los resultados emitidos por el mapper; y, por otra, de recoger todos los valores intermedios pertenecientes a una clave para combinarlos en una lista asociada a ella.

- **Reduce**: esta función se aplica en paralelo para cada grupo asociado a una clave. El resultado es la producción de una colección de valores para cada dominio, aunque también puede darse el caso de generar una llamada vacía. El resultado es la obtención de una lista de valores.



© csdn.net

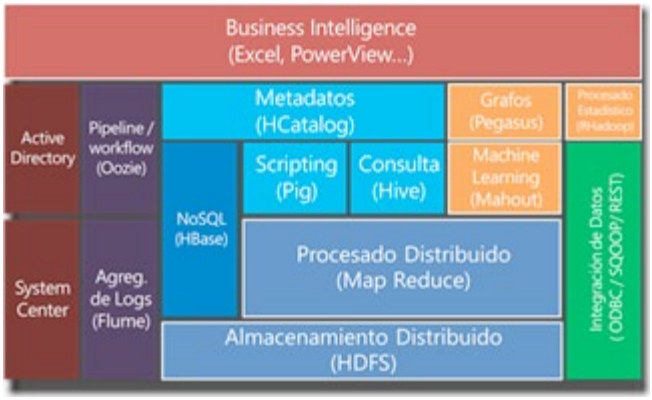


© nosqlfan.com

Principales características de MapReduce

- **Distribución y paralelización (automáticas).**
- **Tolerancia a fallos y a redundancias.**
- **Transparencia: su funcionamiento interno y su mantenimiento son transparentes para los desarrolladores. Es decir, que sólo tienen que programar la lógica de negocio del algoritmo, en vez de necesitar invertir tiempo gestionando errores o parámetros de la computación distribuida.**
- **Escalabilidad horizontal: permite que, si se necesita más potencia de computación, baste con añadir más nodos en el clúster.**
- **Localización de los datos: se desplaza el algoritmo a los datos y no al contrario, como suele suceder en sistemas distribuidos tradicionales.**
- **Dispone de herramientas de monitorización.**

Fases de Big Data y sus soluciones dentro del ecosistema Hadoop



© msdn.com

1. Descubrimiento de grandes datos

El procedimiento a seguir se resume en cuatro pasos:

- **Definir cuáles son los datos de interés.**
- **Encontrar sus fuentes (históricos o Social Media, entre otros)**
- **Grabar los datos en el sistema.**
- **Determinar cómo serán procesados.**

Dentro de Hadoop, pueden emplearse:

- **Flume y Chukwa** framework para datos no estructurados: se ocupan de los ficheros de logs.
- **Sqoop**: si los datos provienen de una base de datos relacional.

2. Extracción y limpieza de los grandes volúmenes de datos

Para llevar a cabo la extracción y pre-procesamiento de los datos es necesario:

- **Extraer los datos de la fuente de origen datos.**
- **Perfilar y limpiar los datos.**
- **Adecuarlos a las necesidades de la empresa, de acuerdo a las reglas de negocio.**
- **Aplicar los estándares de calidad de datos.**

Las dos aplicaciones mencionadas en la fase anterior también son de utilidad en esta etapa, ya que suelen contar con opciones de filtrado previo que proporcionan, a su vez, la estructura conveniente para servir de entrada a Hadoop.

3. Estructuración y análisis de big data

La integración es crucial y aquí es donde intervienen las tres siguientes acciones:

- **Dotar de estructura lógica a los conjuntos de datos tratados.**
- **Almacenar los datos en el repositorio elegido (puede ser una base de datos o u sistema)**
- **Analizar los datos disponibles para hallar relaciones.**

En el ecosistema de Hadoop pueden encontrarse dos alternativas para solucionar los problemas de estructuración:

- **HDFS**: antes de que Hadoop pueda proceder al tratamiento de la información, los datos pre-procesados han de almacenarse en un sistema distribuido de ficheros. Este es el rol que cumple HDFS como componente core dentro de Hadoop.
- **Avro**: se trata de un sistema que hace posible serializar datos para codificar los que va a manejar Hadoop. Permite también definir interfaces a la hora de "parsear" información.

4. Modelado de datos

Esta etapa se orienta al procesamiento de datos apoyándose en el modelado y para ello requiere de:

- **Aplicar algoritmos a los datos.**
- **Aplicar procesos estadísticos.**
- **Resolver las peticiones lanzadas mediante el modelado de datos en base a técnicas de minería.**

Hay muchas maneras de llevar a cabo estos cometidos. Las más eficientes son:

- **Bases Datos NoSQL**: no tienen esquema de datos fijo, por lo que no es necesario preocuparse de comprobar el esquema cada vez que se realiza la inscripción de un registro en la base de datos. Ello supone una ventaja considerable cuando se trabaja con cifras que alcanzan los millones de registros. Algunas de las más conocidas son MongoDB o Impala, aunque existen muchas más opciones.
- **Frameworks de consultas**:
  - **HIVE:** es un framework que permite crear tablas, insertar datos y realizar consultas con un lenguaje similar al que podría llevarse a cabo utilizando queries SQL (el lenguaje no es SQL, sino HQL).
  - **PIG:** permite manejar datos mediante un lenguaje textual conocido como Pig Latin.

5. Interpretación de grandes datos

El fin de todo trabajo con big data pasa por:

- **Interpretar las distintas soluciones.**
- **Aportar un resultado final.**

Las mejores opciones para levar a cabo esta fase de big data son:

- **Mahout y R**: librería de minería de datos que permite realizar clustering, algoritmos de regresión e implementación de modelos estadísticos sobre los datos de salida ya procesados.

PowerData
Especialistas en Gestión de Datos

ESPAÑA

MADRID

C/ Miguel Yuste, 17, 4º, C
28037 Madrid
Tel:(+34) 91 129 72 97
marketing@powerdata.es
www.powerdata.es

BARCELONA

C/ Pau Claris, 95
08009 Barcelona
Tel: (+34) 934 45 60 01
marketing@powerdata.es
www.powerdata.es

VALENCIA

Edificio Europa - 5º I Avda. Aragón, 30
46021 Valencia
Tel: (+34) 960916025
marketing@powerdata.es
www.powerdata.es

LATINOAMÉRICA

ARGENTINA

Avenida Leandro N Alem 530, Piso 4
CD C100 1AAN Ciudad Autónoma de Buenos Aires
Tel: (+54) 11 4314 1370
marketing@powerdataam.com
www.powerdataam.com

CHILE

Av. Presidente Errázuriz Nº 2999 - Oficina 202
Las Condes, Santiago CP 7550357
Tel: (+56) 2 29363-100
marketing@powerdataam.com
www.powerdataam.com

COLOMBIA

Calle 100 No. 8A-55 Torre C. Of. 718
Bogotá
Tel: (+57 1) 6167796
marketing@powerdataam.com
www.powerdataam.com

MÉXICO

Homero 906, Colonia Polanco, Miguel Hidalgo
C.P. 11550, México, D.F.
Tel: +52 (55) 6552-7039
marketing@powerdataam.com
www.powerdataam.com

PERÚ

Calle Los Zorzales Nº 160, piso 9
San Isidro, Lima 27
Tel: (+51) 1 6344900
marketing@powerdataam.com
www.powerdataam.com

Síguenos en: