

# Conceptos y Aplicaciones de Big Data

---

ECOSISTEMA HADOOP

Prof. Waldo Hasperué  
[whasperue@lidi.info.unlp.edu.ar](mailto:whasperue@lidi.info.unlp.edu.ar)

# Temario

---

## Ecosistema Hadoop

- HDFS
- Ejecución de una aplicación en el entorno Hadoop

## Introducción al paradigma MapReduce

## Etapas de un trabajo en MapReduce

- Map
- Shuffle
- Sort
- Reduce

# Historia

---

Para procesar grandes conjuntos de datos, en 2003 Google creó el framework Hadoop capaz de poder procesar grandes volúmenes de datos.

En 2006, Yahoo continúa con el desarrollo del proyecto Hadoop. Aparece Hadoop MapReduce.

Actualmente pertenece a Apache

- Apache Hadoop ([hadoop.apache.org](http://hadoop.apache.org))

# Hadoop

---

Es un framework que soporta procesamiento de grandes bases de datos en un ambiente distribuido.

Ejecuta aplicaciones para el tratamiento de grandes volúmenes de datos.

Incluye un sistema de archivos distribuidos (HDFS).

Tolerante a fallas.

# Hadoop

---

- ✓ Diseñado para el procesamiento off-line de los datos (procesamiento en batch)
- ✓ Funciona con la idea de "escriba una sola vez y lea muchas"
- ✗ No permite lectura aleatoria
- ✗ No permite el procesamiento on-line

Se ejecuta en el "lugar" donde se encuentran los datos

# Componentes Hadoop

---

Common (I/O, serialización, RPC)

HDFS (file system distribuido)

Zookeeper (servicio de coordinación de procesos)

MapReduce (modelo de procesamiento de datos)

Pig (lenguaje de scripting sobre MapReduce)

Cascading (framework que simplifica el uso de MapReduce)

Hive (lenguaje basado en SQL)



# Apache Hadoop Ecosystem



**Ambari**

Provisioning, Managing and Monitoring Hadoop Clusters



**Scoop**  
Data Exchange



**Zookeeper**  
Coordination



**Oozie**  
Workflow



**Pig**  
Scripting



**Mahout**  
Machine Learning

**R Connectors**  
Statistics



**Hive**  
SQL Query



**Hbase**  
Columnar Store



**Flume**  
Log Collector



**YARN Map Reduce v2**

Distributed Processing Framework

**HDFS**

Hadoop Distributed File System



# Componentes Hadoop

---

## Almacenamiento: Distributed File System (DFS)

- Los archivos están distribuidos
- Ofrece transparencia al usuario permitiendo operar con todos los archivos del cluster a través del file system distribuído.
- Un mismo archivo podría estar almacenado en varias computadoras.
- Hadoop tiene su propio filesystem distribuido: el HDFS (Hadoop Distributed FileSystem)



# DFS

---

Hay varios sistemas de archivos distribuidos

- HDFS
- HFTP
- HSFTP
- HAR
- FTP
- S3

# HDFS

---

Todos los archivos se dividen en bloques del mismo tamaño (64MB por defecto, aunque es configurable).

Los bloques pueden estar físicamente en cualquier computadora.

Permite la réplica de bloques para optimización y recupero de fallas.

# Procesos del HDFS

---

## Namenode

- Maneja el árbol del filesystem y los metadatos de cada archivo y carpeta.
- Conoce para cada bloque del FS que datanode lo maneja.
- Vínculo con el filesystem del SO

## Datanode

- Son lo que llevan a cabo la lectura y escritura de los bloques en el filesystem del SO.
- Lleva a cabo la creación, borrado y replicado de los bloques.

Secondary namenode: realiza tareas auxiliares al namenode.

# Componentes Hadoop

---

En Hadoop la administración de los procesos que se ejecutan en el cluster la lleva a cabo un framework llamado Yarn MapReduce.

Básicamente Yarn realiza los trabajos usando dos procesos diferentes:

- Job tracker: maneja todos los trabajos a ser procesados. Tiene en cuenta el mapa del cluster al momento de crear los procesos Task
- Task tracker: son los encargados de realizar el procesamiento de los datos