
BIG DATA

Introducción a Hadoop



Introducción a Big Data

- Hadoop es casi sinónimo del término “Big Data”.
- Es un entorno distribuido de:
 - Datos
 - Procesos
- Hadoop es un entorno de tipo High Performance Super Computer que se puede escalar horizontalmente con hardware relativamente barato “commodity hardware”



Introducción a Big Data

- Hadoop implementa procesamiento en paralelo a través de nodos de datos en un sistema de ficheros distribuidos.

NODOS MAESTROS



NODOS
ESCLAVOS



Introducción a Big Data

- Uno de los puntos fuertes de Hadoop es que está diseñado para ejecutarse en servidores de bajo coste y que dispone de una gran tolerancia a fallos
 - De hecho, en Hadoop, los fallos de hardware se tratan como una regla y no como una excepción.
 - Por tanto se puede montar un cluster de servidores X86 a un precio muy razonable si lo comparamos con grandes servidores
 - Y si no, nos queda la nube
-



Introducción a Big Data

- Hadoop en un entorno que suministra librerías open source para la computación distribuida usando varios componentes, Los principales son:
 - Hadoop Common
 - MapReduce
 - Hadoop Distributed File System (HDFS).
- Está diseñado para escalar desde unos pocos nodos a miles de máquinas, cada uno de ellas ofreciendo la lógica de negocio y el almacenamiento a nivel local.



Introducción a Big Data

• Versiones

- Un poco enrevesadas al mantener varias líneas de trabajo
- En el momento de escribir este manual
 - **1.2.X** - current stable version, 1.2 release
 - **2.7.1** - stable 2.x version
 - **0.23.X** - similar to 2.X.X but missing NN HA.



Introducción a Big Data

- El “core” de Hadoop está formado por dos componentes básicos:



DATOS

PROCESAMIENTO



Introducción a Big Data

- **HDFS**

- HDFS es un sistema de almacenamiento tolerante a fallos que puede almacenar gran cantidad de datos, escalar de forma incremental y sobrevivir a fallos de hardware sin perder datos
- Si un nodo falla, el cluster puede continuar trabajando sin perder datos o interrumpir el trabajo.
- Sencillamente redistribuye el trabajo entre los nodos restantes del cluster.




Introducción a Big Data

- **Procesos**
 - En la actualidad existen dos formas de procesamiento distintos
 - Map Reduce V1
 - Map Reduce V2- YARN
 - De forma general son algoritmos de procesamientos de datos que implementan procesos en paralelo
 - Es decir distribuye las tareas a través de los nodos de un cluster
-



Introducción a Big Data

[Apache](#) > [Hadoop](#) >



[Top](#) [Wiki](#)

Last Published: 12/18/2015 15:22:33

About

Welcome

What Is Apache Hadoop...

Getting Started ...

Download Hadoop

Who Uses Hadoop?...

News

Releases

Mailing Lists

Issue Tracking

Who We Are?

Who Uses Hadoop?

Buy Stuff

Sponsorship

Thanks

Privacy Policy

Bylaws

License

Documentation

Related Projects

built with Apache Forrest

Welcome to Apache™ Hadoop®!

What Is Apache Hadoop?

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects at Apache include:

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A Scalable machine learning and data mining library.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.
- **Spark™**: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez™**: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
- **ZooKeeper™**: A high-performance coordination service for distributed applications.



Introducción a Big Data

- Otros proyectos implicados en Hadoop
 - HBase : Una base de datos orientada a valores/claves que se ejecuta sobre HDFS
 - Hive : sistema de funciones que soportan agregación de datos y consultas ad hoc sobre MapReduce
 - Pig: Lenguaje de alto nivel para gestionar flujos de datos y ejecución de aplicaciones sobre Hadoop
 - Mahout: entorno de aprendizaje de máquinas implementado en hadoop
 - Zookeeper : servicio centralizado para mantener información de configuración, gestión de nombre, y para facilitar la sincronización de servicios
 - Sqoop : Herramienta diseñada para transferir datos masivos desde Hadoop a otros entornos como Bases de datos relacionales

