



BIG DATA

INTRODUCCION A BIG DATA Y HADOOP

EDUARD LARA

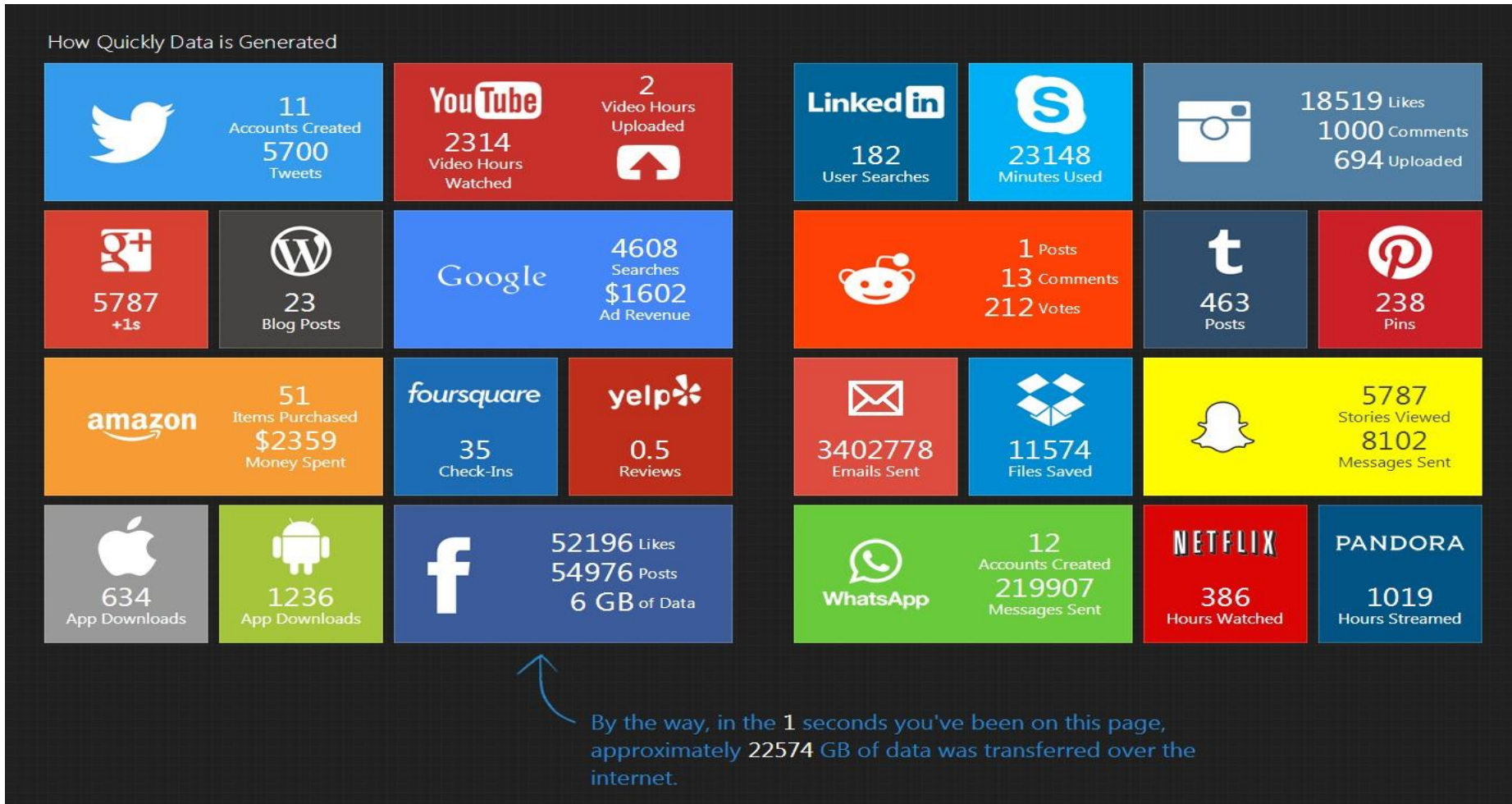
1. INTRODUCCION A BIG DATA

- ❖ **Big Data** es una filosofía o tipo de tecnología aplicable a productos tipo hadoop
- ❖ Big Data es la convergencia de enormes cantidades de datos tanto estructurados (datos en tablas-columnas, Oracle, mysql, muy relacionados) como no estructurados (fóruns, tweets, Facebook, blogs)
- ❖ **Petabytes de datos** creados diariamente
 - Redes sociales
 - Uso de Móviles,
 - Internet
 - Sensores físicos
 - Datos científicos,

Datos de los que extraer
algún tipo de información

1. INTRODUCCION A BIG DATA

❖ Lo que se genera en un 1 segundo en el mundo de Internet



1. INTRODUCCION A BIG DATA

- Hay una enorme cantidad de datos de los que queremos procesar y extraer información.



1. INTRODUCCION A BIG DATA

- Las tres Vs de BIG DATA

VOLUMEN

Terabytes
PetaBytes

VARIEDAD

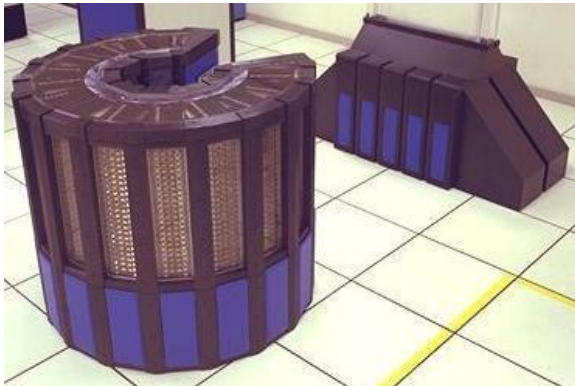
Estructurado
No estructurado
Semi-estructurado

VELOCIDAD

Batch
Tiempo real
Streams

1. INTRODUCCION A BIG DATA

- Dado que las tecnologías tradicionales no pueden hacer frente a esta cantidad de información es necesario utilizar nuevas estrategias.
- Informática distribuida: Distribución de datos y de procesos



Grandes servidores



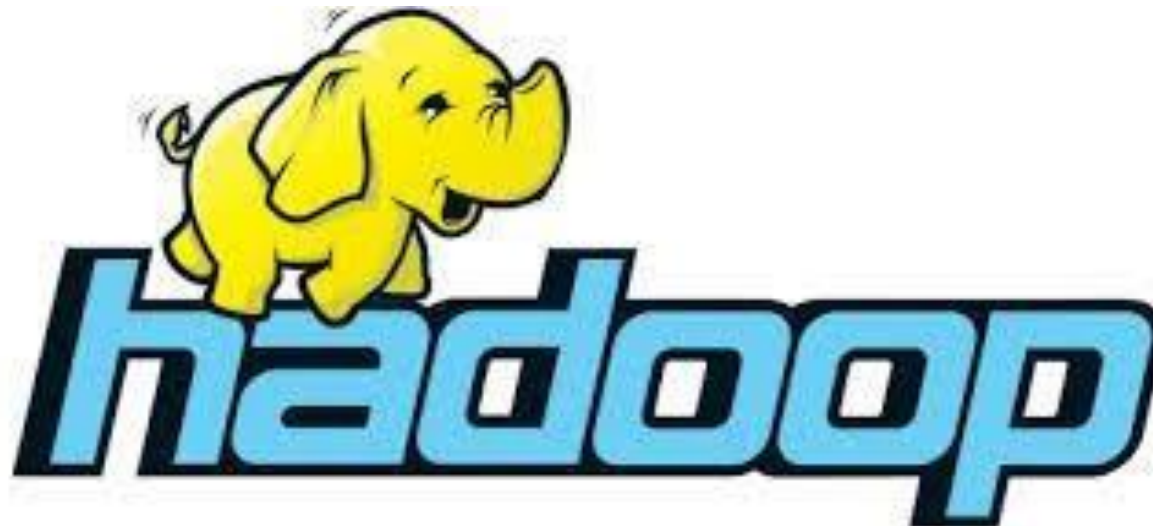
VERSUS



Entornos distribuidos

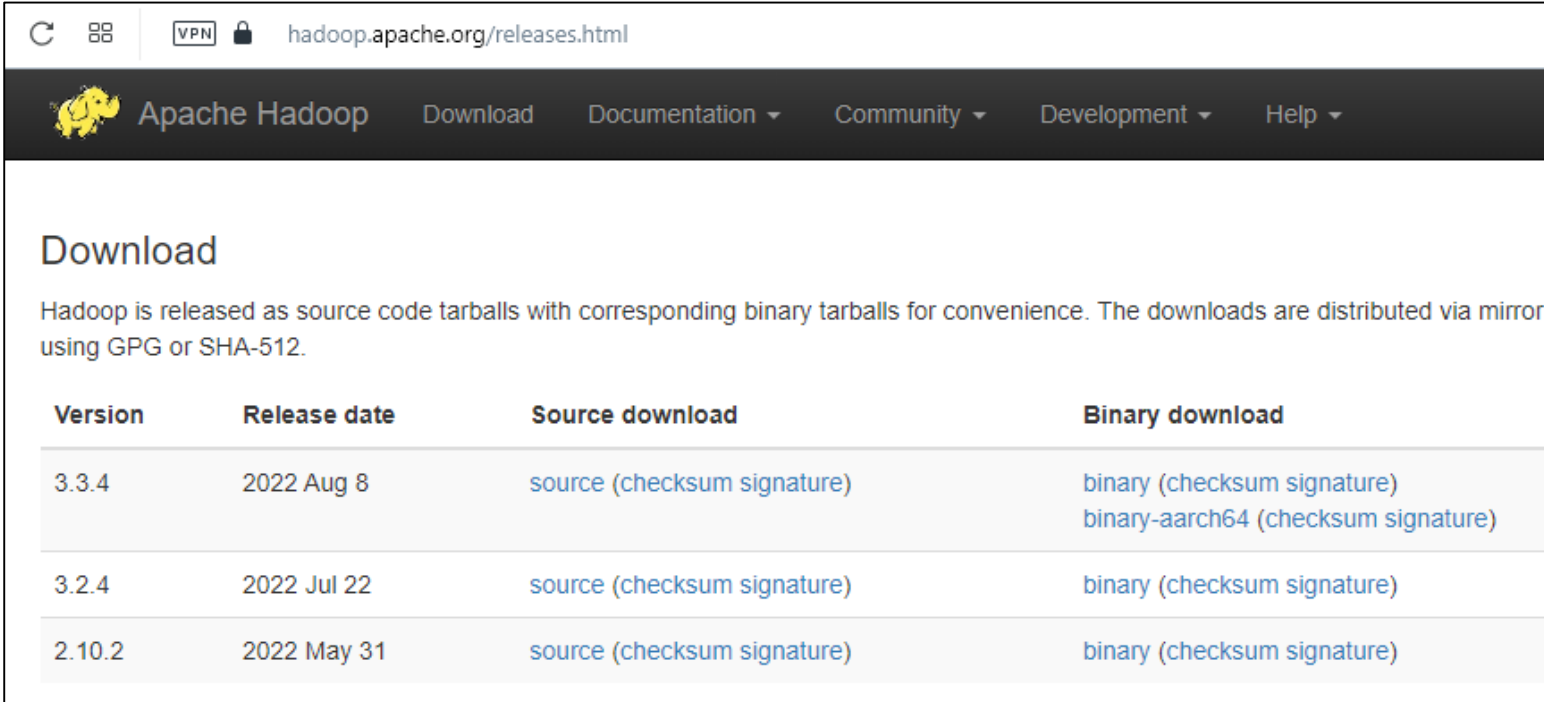
1. INTRODUCCION A BIG DATA

- ❖ Tecnología Big Data: Apache Hadoop
- ❖ Apache Hadoop es la traducción a nivel tecnológico de Big Data



2. INTRODUCCION A HADOOP

- ❖ Se puede hacer tanto con Versión 2 como con versión 3 de apache hadoop.
- ❖ La ultima versión 2 → 2.10.2 de Mayo 2022
- ❖ La ultima versión 3 → 3.3.4 de Agosto 2022



The screenshot shows the Apache Hadoop releases page. The browser address bar displays 'hadoop.apache.org/releases.html'. The page header includes the Apache Hadoop logo and navigation links: Download, Documentation, Community, Development, and Help. The main content area is titled 'Download' and contains a paragraph explaining that Hadoop is released as source code tarballs with corresponding binary tarballs for convenience, distributed via mirrors using GPG or SHA-512. Below this is a table with four columns: Version, Release date, Source download, and Binary download. The table lists three versions: 3.3.4 (released Aug 8, 2022), 3.2.4 (released Jul 22, 2022), and 2.10.2 (released May 31, 2022). Each version row provides links for source and binary downloads, with the 3.3.4 version also including a link for 'binary-aarch64'.

Version	Release date	Source download	Binary download
3.3.4	2022 Aug 8	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)
3.2.4	2022 Jul 22	source (checksum signature)	binary (checksum signature)
2.10.2	2022 May 31	source (checksum signature)	binary (checksum signature)

2. INTRODUCCION A HADOOP

- ❖ Hadoop es casi sinónimo del término "Big Data".
- ❖ Big Data es la filosofía o el tipo arquitectura y hadoop es la traducción física de esa filosofía
- ❖ 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"

2. INTRODUCCION A HADOOP

- ❖ Hadoop implementa procesamiento en paralelo a través de nodos de datos en un sistema de ficheros distribuidos. Si un proceso tarda 10 minutos, si tengo 10 maquinas y distribuyo el proceso entre esas 10^o maquinas me tardará 1 minuto
- ❖ Hadoop utiliza el principio divide y vencerás
- ❖ Infraestructura de un clúster hadoop: Nodos maestros que van a gobernar a los nodos esclavos que hacen procesamiento de la información



2. INTRODUCCION A HADOOP

- ❖ 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
- ❖ Presupone que siempre va a ver una máquina que se va a estropear. Como hay múltiples nodos, si uno falla no pasa nada, ese proceso se lo pasamos a otro.
- ❖ De hecho, en Hadoop, los fallos de hardware se tratan como una regla y no como una excepción.
- ❖ Se puede montar un cluster de servidores X86 a un precio razonable, comparando con grandes servidores
- ❖ Se monta la infraestructura de forma que trabajen de forma conjunta. Y si no, nos queda la nube

2. INTRODUCCION A HADOOP

- ❖ Hadoop en un entorno que suministra librerías open source para la computación distribuida usando varios componentes,
- ❖ Los principales componentes son:
 - ❖ Hadoop Common (librerías comunes)
 - ❖ MapReduce (Proceso)
 - ❖ Hadoop Distributed File System (HDFS). Donde se almacena los datos
- ❖ 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.

2. INTRODUCCION A HADOOP

- Versiones
 - Un poco enrevesadas al mantener varias líneas de trabajo

Download

Hadoop is released as source code tarballs with corresponding binary tarballs for conv using GPG or SHA-512.

Version	Release date	Source download
3.3.6	2023 Jun 23	source (checksum signature)
3.2.4	2022 Jul 22	source (checksum signature)
2.10.2	2022 May 31	source (checksum signature)

2. INTRODUCCION A HADOOP

- ❖ El “core” de Hadoop está formado por dos componentes básicos:
 - ❖ Distribución de datos
 - ❖ Procesamiento

DATOS

PROCESAMIENTO

2. INTRODUCCION A HADOOP

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.

2. INTRODUCCION A HADOOP

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



2. INTRODUCCION A HADOOP

Pagina descarga de hadoop

[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.

2. INTRODUCCION A HADOOP

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

3. PRODUCTOS BIG DATA

❖ Hive

- ❖ Permite acceder a HDFS como si fuera una Base de datos relacional
- ❖ Podemos ejecutar comandos muy parecidos a SQL para recuperar valores (HiveSQL)
- ❖ Esto simplifica enormemente el desarrollo y la gestión con Hadoop

<http://hive.apache.org/index.html>



3. PRODUCTOS BIG DATA

HBASE

- ❖ Es el sistema de almacenamiento no relacional por defecto para Hadoop. Hay otros como Cassandra, MongoDB. Es de tipo multi-columna
- ❖ Es una base de datos de código abierto, distribuida y escalable para el almacenamiento de Big Data.
- ❖ Está escrita en Java e implementa y proporciona capacidades similares sobre Hadoop y HDFS.
- ❖ El objetivo de este proyecto es el de trabajar con grandes tablas, miles de millones de filas de X millones de columnas, sobre un cluster Hadoop.

<http://hbase.apache.org/>

3. PRODUCTOS BIG DATA

PIG

- ❖ Pig es un lenguaje de alto de nivel para analizar grandes volúmenes de datos.
- ❖ Pig trabaja en paralelo lo que permite gestionar gran cantidad de información
- ❖ Es un compilador que genera comandos MapReduce.
- ❖ Es un lenguaje textual denominado Pig Latin.

<https://pig.apache.org/>



3. PRODUCTOS BIG DATA

Sqoop

- ❖ Permite transferir gran volumen de datos de manera eficiente entre Hadoop y gestores de datos estructurados, como Bases de datos relacionales
- ❖ Sqoop ofrece conectores para integrar Hadoop con otros sistemas, como por ejemplo Oracle o SqlServer

<http://sqoop.apache.org/>



3. PRODUCTOS BIG DATA

Flume

- ❖ Flume es un servicio distribuido y altamente eficiente para distribuir, agregar y recolectar grandes cantidades de información.
- ❖ Útil para cargar y mover en Hadoop información de tipo texto, como ficheros de logs, paquetes de twitter, etc.
- ❖ Tiene una arquitectura de tipo streaming con un flujo de datos muy potente y personalizables

<https://flume.apache.org/>



3. PRODUCTOS BIG DATA

Zookeeper

- ❖ ZooKeeper es un servicio para mantener la configuración, coordinación y aprovisionamiento de aplicaciones distribuidas
- ❖ No solo vale para Hadoop, pero es muy útil en esa arquitectura
- ❖ Elimina la complejidad de la gestión distribuido de la plataforma

<https://zookeeper.apache.org/>



3. PRODUCTOS BIG DATA

Spark

- ❖ Es un motor muy eficiente de procesamiento de datos a gran escala
- ❖ Implementa procesamiento en tiempo real al contrario que Map Reduce
- ❖ Es más rápida que MapReduce
- ❖ Trabaja de forma masiva en memoria
- ❖ Puede funcionar stand-alone

<http://spark.apache.org/>



3. PRODUCTOS BIG DATA

Otros

- ❖ **Avro:** Sistema para serialización de datos
- ❖ **Cassandra:** base de datos multi-master muy potente
- ❖ **Mohout:** machine learning y Data Mining
- ❖ **Tez, Chuwka.....**

4. DISTRIBUCIONES

Distribuciones Hadoop

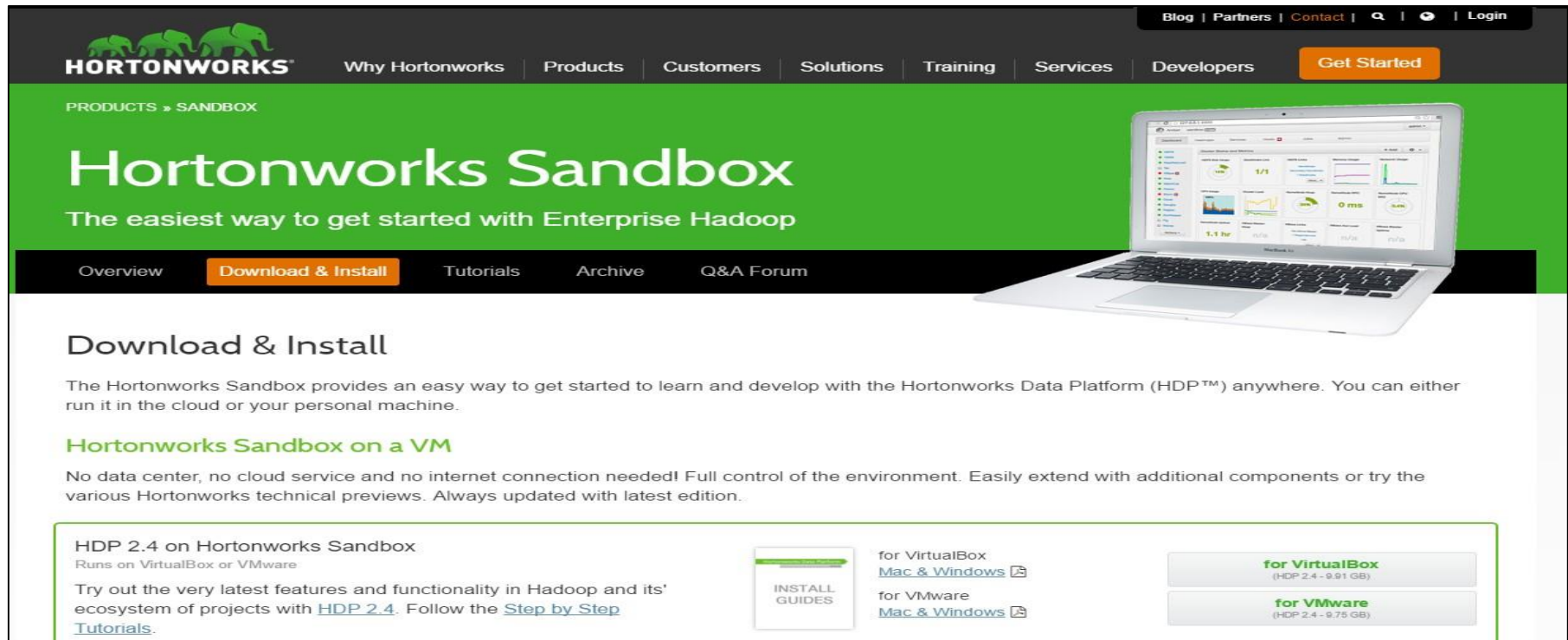
- ❖ Hay distintas empresas que ofrecen soluciones “empaquetadas” para Hadoop
- ❖ Podemos descargar máquinas virtuales pre-hechas o usar soluciones en la nube
- ❖ Entre las más importantes están:
 - **CLOUDERA**
 - **HORTONWORKS**
 - **IBM OPEN PLATFORM**
 -

4. DISTRIBUCIONES

- Distribuciones Hadoop

Nombre	Producto ofrecido
Amazon Web Service	Amazon Elastic MapReduce
Cloudera	Cloudera Enterprise
Hortonworks	Hortonworks Data Platform
Intel	Intel Distribution for Apache Hadoop
MapR Technologies	MapR M3 - MapR M7
Microsoft	Windows Azure HDInsight
Pivotal Software	Pivotal HD
Teradata	Teradata Open Distribution for Hadoop (TDH)

4. DISTRIBUCIONES



The screenshot shows the Hortonworks Sandbox website. The header includes the Hortonworks logo and navigation links: Why Hortonworks, Products, Customers, Solutions, Training, Services, Developers, and a Get Started button. Below the header, the main banner features the text 'Hortonworks Sandbox' and 'The easiest way to get started with Enterprise Hadoop'. A secondary navigation bar includes Overview, Download & Install, Tutorials, Archive, and Q&A Forum. The 'Download & Install' section is highlighted, showing a laptop displaying the Hortonworks Sandbox interface. Below this, the text describes the ease of use and availability of the sandbox. A section titled 'Hortonworks Sandbox on a VM' provides details on running HDP 2.4 on VirtualBox or VMware, including links to install guides and download buttons for both platforms.

Hortonworks

Why Hortonworks | Products | Customers | Solutions | Training | Services | Developers | [Get Started](#)

PRODUCTS » SANDBOX

Hortonworks Sandbox

The easiest way to get started with Enterprise Hadoop

[Overview](#) | [Download & Install](#) | [Tutorials](#) | [Archive](#) | [Q&A Forum](#)

Download & Install

The Hortonworks Sandbox provides an easy way to get started to learn and develop with the Hortonworks Data Platform (HDP™) anywhere. You can either run it in the cloud or your personal machine.

Hortonworks Sandbox on a VM

No data center, no cloud service and no internet connection needed! Full control of the environment. Easily extend with additional components or try the various Hortonworks technical previews. Always updated with latest edition.

HDP 2.4 on Hortonworks Sandbox
Runs on VirtualBox or VMware

Try out the very latest features and functionality in Hadoop and its' ecosystem of projects with [HDP 2.4](#). Follow the [Step by Step Tutorials](#).

INSTALL GUIDES

for VirtualBox
[Mac & Windows](#)

for VMware
[Mac & Windows](#)

for VirtualBox
(HDP 2.4 - 9.91 GB)

for VMware
(HDP 2.4 - 9.75 GB)

4. DISTRIBUCIONES



The screenshot shows the Cloudera website's 'Download Cloudera Enterprise' page. The header includes navigation links for Downloads, Training, Support Portal, Partners, Developers, and Community, along with Search, Sign In, and Language options. The main heading is 'Download Cloudera Enterprise' with the subtitle 'Local, On Premise, or Cloud-based Apache Hadoop Management'. Below this are three blue boxes representing different installation methods: QuickStarts (local machine), Cloudera Manager (enterprise data hub), and Cloudera Director (cloud). Each box contains a 'DOWNLOAD NOW' button and a 'Learn More' link. The footer mentions 'Cloudera Enterprise Extensions'.

Downloads Training Support Portal Partners Developers Community Search Sign In Language

cloudera Why Cloudera Products Services & Support Solutions Get Started

Download Cloudera Enterprise

Local, On Premise, or Cloud-based Apache Hadoop Management



QuickStarts

Get Started on your local machine using a QuickStart VM or Docker Image.

[DOWNLOAD NOW](#)

[Learn More](#)



Cloudera Manager

A unified interface to manage your enterprise data hub. Express and Enterprise editions available.

[DOWNLOAD NOW](#)

[Learn More](#)



Cloudera Director


Self-service, reliable experience for CDH and Cloudera Enterprise in the cloud

[DOWNLOAD NOW](#)

[Learn More](#)

Cloudera Enterprise Extensions

4. DISTRIBUCIONES



The screenshot displays the MapR Converged Data Platform website. The top navigation bar includes links for 'Why MapR', 'Products & Services', 'Partners', 'Training', 'Community', and 'Resources'. The main heading is 'MapR Converged Data Platform'. Below this, a navigation bar lists 'Overview', 'Platform Services', 'Open Source Engines', 'Commercial Engines', 'MapR Editions', and 'What's Included'. The 'Overview' section contains a paragraph describing the platform's integration of Hadoop and Spark with real-time database capabilities, global event streaming, and scalable enterprise storage. A 'Compare Editions' button is also present. Below the text, a diagram titled 'MapR Converged Data Platform' shows two categories: 'Open Source Engines and Tools' (including Hadoop, Spark, Apache Drill, and Search and Others) and 'Commercial Engines and Applications' (including Cloud and Managed Services, VERTICA, SAP, MySQL, and Custom Apps). The diagram is flanked by 'Processing' and 'Unified Management' labels.

MAPR

Why MapR Products & Services Partners Training Community Resources

MapR Converged Data Platform

Overview Platform Services Open Source Engines Commercial Engines MapR Editions What's Included

The MapR Converged Data Platform integrates Hadoop and Spark with real-time database capabilities, global event streaming, and scalable enterprise storage to power a new generation of big data applications. The MapR Platform delivers enterprise grade security, reliability, and real-time performance while dramatically lowering both hardware and operational costs of your most important applications and data.

[Compare Editions](#)

MapR Converged Data Platform

Open Source Engines and Tools

Processing


- Hadoop
- Spark
- APACHE DRILL
- Search and Others

Commercial Engines and Applications

- Cloud and Managed Services
- VERTICA
- SAP
- MySQL
- Custom Apps

Unified Management

4. DISTRIBUCIONES

 Analytics Cloud Commerce Infraestructura de TI MobileFirst Seguridad


Software de IBM > Productos > Bases de datos > Sistema Hadoop > IBM BigInsights >

IBM Open Platform with Apache Hadoop

Plataforma de código abierto 100% Apache Hadoop

No-charge download available

→ IBM Open Platform with Apache Hadoop



IBM Open Platform with Apache Hadoop crea la plataforma para proyectos de big data y proporciona el contenido de código abierto de Apache Hadoop más actual. IBM ofrece esta distribución Apache de código abierto como descarga gratuita, así como una oferta soportada para todas las cargas de trabajo Hadoop.

IBM Open Platform with Apache Hadoop

- Soporte nativo para actualizaciones continuas de los servicios Hadoop