Big Data desde cero

Contexto, tecnologías y aplicación

HADOOP: CORE TECNOLÓGICO BIG DATA

- Apache Hadoop: core tecnológico de Big Data
- Historia de Apache Hadoop
- HDFS y YARN: Almacenamiento y procesamiento como punto central
- Ecosistema Hadoop: diferentes productos para diferentes finalidades
- Ubicar los componentes en sus capas correspondientes
- Hadoop es al Big Data lo que Linux a los Sistemas Operativos

APACHE HADOOP: CORE TECNOLÓGICO DE BIG DATA

- **Apache Hadoop:** la tecnología que propició cumplir las 3 V´s del Big Data.

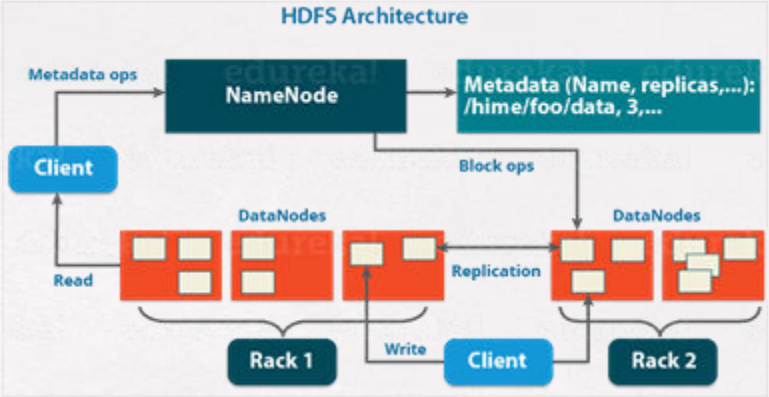
- Pertenece al **proyecto Apache** y tiene una de las mayores comunidades activas.

HISTORIA DE APACHE HADOOP

- 1997: **Lucene** → Doug Cutting (D.C.) crea el motor de indexación Lucene.
- 2002: **Nutch 1-Mach. Nutch 4-M.** → D. C. crea un buscador distribuido, pero sólo con 4 máquinas.
- 2003: **GFS y Map/Reduce** → Google publica cómo almacena y procesa internet (GFS y Map/Reduce)
- 2005: **Nutch + Hadoop** → D.C. basado en los whitepapers de Google crea Hadoop.
- 2006 - 2008: **Yahoo y Yahoo!!** → Es contratado por Yahoo para desarrollar Hadoop (para adelantar a Google).
- 2009: **Cloudera y MapR** → Nacen empresas que dan soporte sobre Hadoop.
- 2010: **Hortonworks** → Spin-off de Yahoo de la parte de soporte a Hadoop.
- 2011: **Hadoop 2.0 (YARN)** → Evolución de Hadoop separando el procesamiento de la gestión de recursos.
- 2012: **Spark** → Revolución en el mundo Big Data, el framework de procesamiento estrella hoy en día.

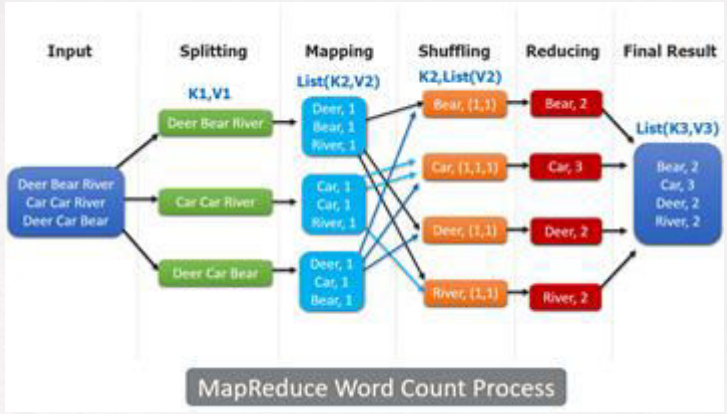
HDFS Y YARN: ALMACENAMIENTO Y PROCESAMIENTO

- **HDFS: Almacenamiento**
  - Hadoop Distributed File System.
  - Sistema de almacenamiento de Hadoop.
  - Apariencia de un único sistema de ficheros, similar a Linux.
  - Es altamente escalable y tolerante a fallos.



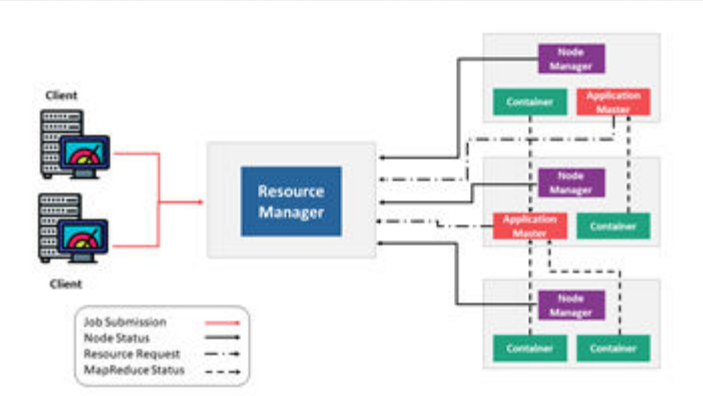
HDFS Y YARN: ALMACENAMIENTO Y PROCESAMIENTO

- **MapReduce: Procesamiento**
  - MapReduce permite el procesamiento distribuido sobre HDFS.
  - Permite obtener Data Locallity (revolución tecnológica de Big Data).
  - Es es el primer paradigma de programación de Hadoop.
    - Con una fase de Mapeo (Map), y otra de Agrupación (Reduce).
  - Su funcionamiento lo ilustra un "Contar palabras".



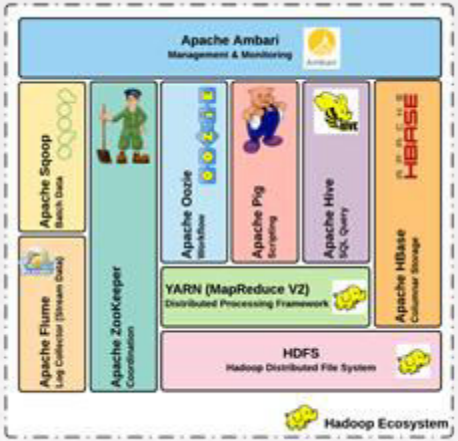
HDFS Y YARN: ALMACENAMIENTO Y PROCESAMIENTO

- **YARN: Gestión de recursos**
  - YARN gestiona los recursos (CPU y RAM) del cluster para procesar
  - Es tolerante a fallos y permite multitud de ejecuciones en paralelo.
  - Gestiona cuotas de ejecución según prioridad, necesidad, disponibilidad, ...
  - Apareció con MapReduce, pero puede trabajar con otros otros frameworks de procesamiento, como Spark.

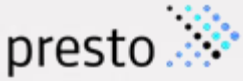


ECOSISTEMA HADOOP

- Hay una gran cantidad de componentes que trabajan de manera nativa con HDFS y YARN.
  - Algunos también del Proyecto Apache:



  - Otros no...



- Pero al conjunto de todos se conoce como **"Ecosistema Hadoop"**

UBICAR LOS COMPONENTES EN SUS CAPAS

- Hay un gran cantidad de productos en el mundo Big Data (dentro y fuera del ecosistema Hadoop)

http://mattturck.com/wp-content/uploads/2019/07/2019_Matt_Turck_Big_Data_Landscape_Final_Fullsize.png

- Importante ubicar el producto en las Capas de la Plataforma Big Data:

UBICAR LOS COMPONENTES EN SUS CAPAS

- Distribución de los principales componentes Big Data en sus capas correspondientes:

HADOOP ES AL BIG DATA LO QUE LINUX A LOS S.O.

- Apache Hadoop es **software libre**, pero sin un soporte oficial por parte de Apache - como Linux.

- Hay empresas que ofrecen **soporte comercial sobre software libre**: tanto en Linux como en Hadoop.

  - **Hadoop** sería comparable a **Linux**.
  - **Cloudera y Hortonworks** son en Hadoop como **Red Hat y Ubuntu** en Linux.