09-2024

code URJC

Sistemas Distribuidos de Procesamiento de Datos I

Tema 2: Arquitecturas de procesamiento de datos

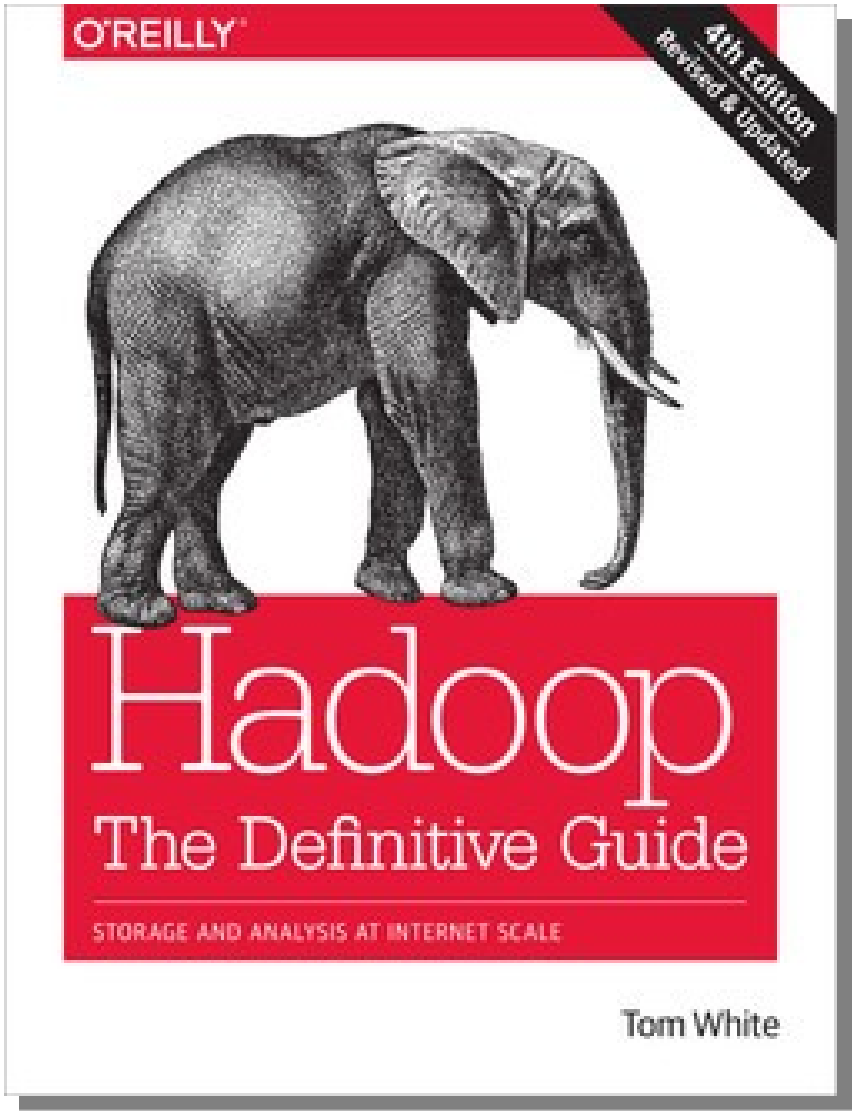
Tema 1.1 Introducción a Apache Hadoop

Universidad Rey Juan Carlos

Michel Maes
michel.maes@urjc.es

Bibliografía

- **Hadoop: The Definitive Guide, 4th Edition**
  - Tom White
  - 2015
  - Disponible gratis



https://www.oreilly.com/library/view/hadoop-the-definitive/9781491901687

Introducción a Apache Hadoop

code URJC

- Historia
- Distribución
- Arquitectura

Introducción a Apache Hadoop

code URJC

- **Historia**
- Distribución
- Arquitectura

Historia

code URJC

Precedentes

- Creado por Doug Cutting (Apache Lucene) con el combre original de Apache Nutch en 2002 como un *web search engine* open-source
- En 2003 Google publica un *paper* sobre su sistema de ficheros distribuido (*Google Distributed File System*)
- En 2004 Google publica un *paper* sobre el algoritmo de procesamiento MapReduce.
- Nutch toma estas ideas y y las aplica para crear Hadoop

Historia

code URJC

Evolución

- Abril 2008 → Hadoop procesa 1 TB de información en 3.5 minutos
- Noviembre de 2008 → La implementación de Google procesa 1TB de información en 68 segundos
- Abril de 2009 → Hadoop procesa 1TB de información en 62 segundos
- 2014→ Spark procesa 100TB de información en 23 minutos (4.27 TB/min)

Historia

Uso en la industria

- Utilizado por Yahoo, New York Times, Facebook ...
  - El New York Times convirtió en PDF todo su archivo en papel escaneado en menos de 24h utilizando un clúster Hadoop en 100 máquinas EC2 de Amazon Web Services (AWS)
- Sigue siendo un estándar en grandes compañias, a pesar de estar perdiendo terreno frente a Spark

Introducción a Apache Hadoop

code URJC

- Historia
- **Distribución**
- Arquitectura

Distribución

- Existen diferentes alternativas para usar Hadoop
  - On-premise
  - Proveedores especificos de Hadoop
  - Cloud público

Distribución

code URJC

On-premise

- Lo instalamos en nuestras propias máquinas
- Distintas opciones:
  - Single-node

https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html

  - Hadoop Cluster

https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/ClusterSetup.html

  - Docker

https://hub.docker.com/r/apache/hadoop

Distribución

Proveedores especificos de Hadoop

- Existen empresas específicas que ofrecen distribuciones concretas de Hadoop tanto para instalar como para utilizar en su nube
  - Cloudera (Hortonworks)
  - IBM
  - Microsoft
  - Oracle
  - …

Distribución

Cloud público

- La mayoría de clouds públicos ofrecen servicios de procesamiento de datos basados en Hadoop

amazon EMR

Cloud Dataproc

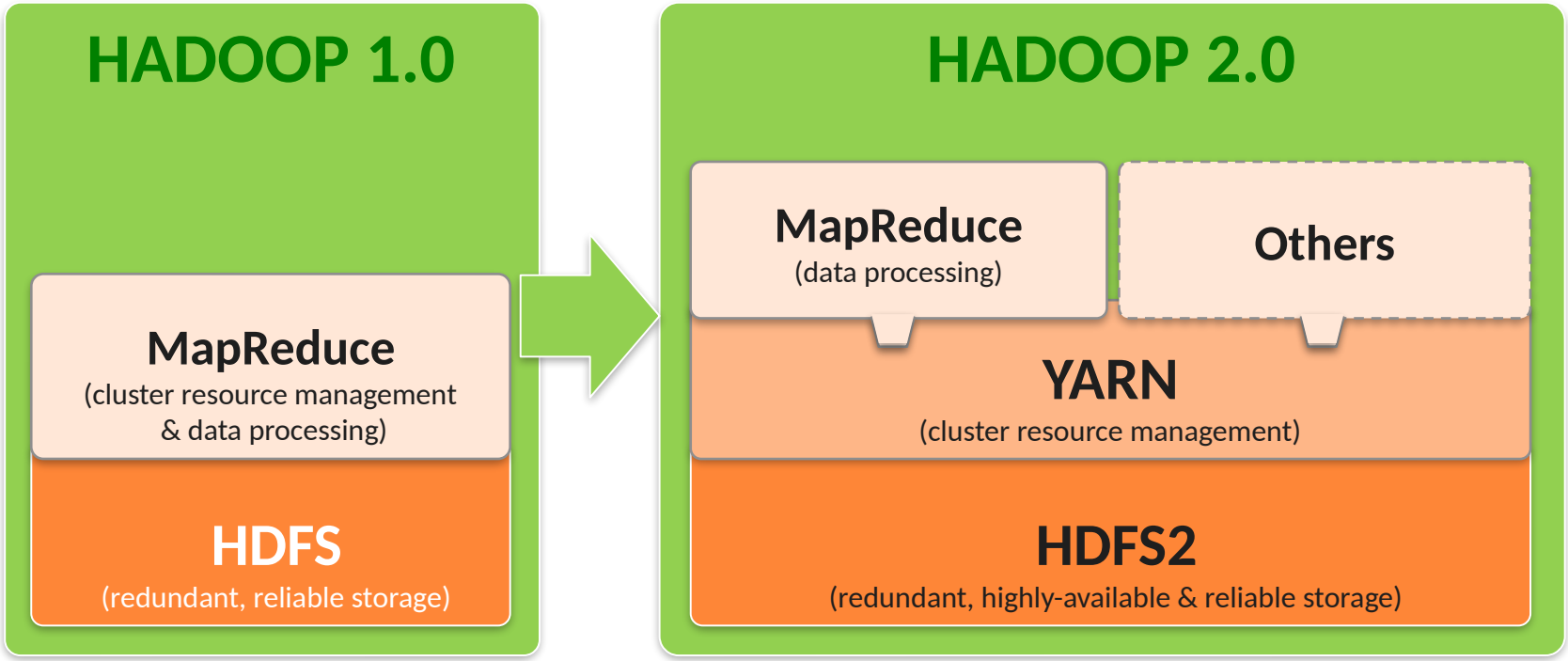
Introducción a Apache Hadoop

code URJC

- Historia
- Distribución
- **Arquitectura**

Arquitectura

Evolución de la arquitectura



Next steps

A lo largo del tema veremos …

- qué es Map Reduce
- cómo trabajar con HDFS
- cómo lanzar Hadoop utilizando distintas opciones
- cómo trabajar con YARN

Historia

code URJC

- **¿Cómo podemos acceder al objeto HttpClient?**
  - ¿Creamos el objeto con new?
    - Podríamos crear un objeto nuevo cada vez que nos haga falta
    - Esta opción **dificulta hacer tests automáticos** unitarios porque necesitaríamos que el servidor REST estuviese disponible y sería más difícil probar diferentes situaciones