ITAB
Academy

Présentation de

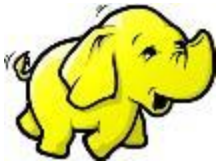
BIG DATA

ITAB ACADEMY

Hadoop

- Hadoop
- Before Hadoop
- What is Hadoop?
- Hadoop Timeline
- Who uses Hadoop
- Hadoop ecosytem
- Hadoop Vendors
- HDFS
- Common HDFS Commands

ITAB ACADEMY

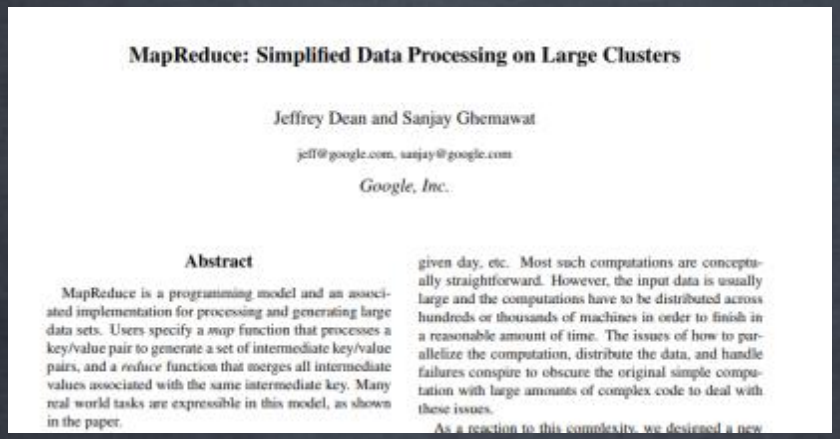


ITAB Academy

Before Hadoop?

Google

- **Before Hadoop?**
- En 2004, **Dean et Ghemawat** de **Google** ont publié un article sur MapReduce qui a donné naissance du BIG DATA



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

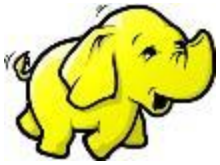
Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new



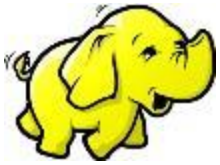
ITAB Academy

What is Hadoop?

- **What is Hadoop?**

- **Doug Cutting, Mike Cafarella et son équipe** ont pris la solution fournie par Google et ont lancé un projet Open Source appelé HADOOP en 2005.

- Doug l'a nommé après l'éléphant de jouet de son fils.



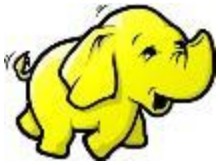
ITAB Academy

What is Hadoop?

- **What is Hadoop?**
- Apache Hadoop est un **framework open-source** écrit en JAVA utilisé pour:
  - le stockage distribué
  - et le traitement de l'ensemble de données de Big Data.
- Il utilise le modèle de programmation **MapReduce** en parallèle sur différents noeuds CPU.
- Hadoop is composed of both: HDFS et MapReduce.

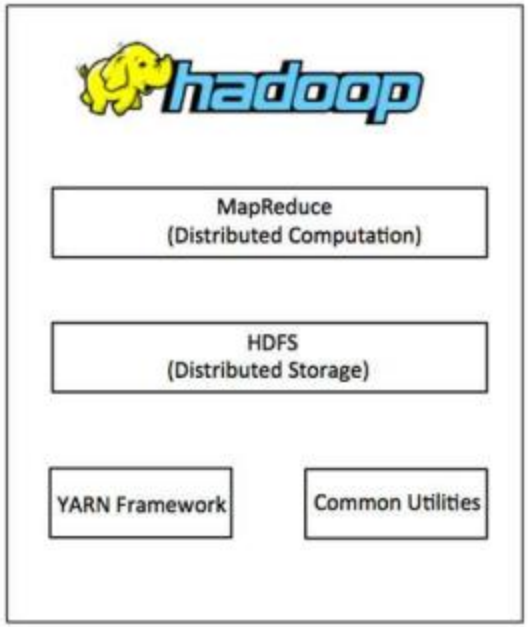
What is Hadoop?

- **What is Hadoop?**
- Hadoop fonctionne sur un cluster de machines.
- Apache Hadoop est aussi un **framework:**
  - **Scalable** (Evolutif)
  - **Fault tolerant** (Tolérance de panne)
  - **Distributed**
  - **Reliable** (Fiable)
  - **Highly available** (Hautement disponible)
  - **Economic** (On peut l'installer sur des machine pas chère)
  - **Easy to use**



ITAB Academy

What is Hadoop?



- **Noyau de Hadoop**

- **Hadoop Common (anciennement Hadoop Core)**
  - Ce sont des bibliothèques Java et des utilitaires requis par d'autres modules Hadoop et des fichiers Java et des scripts requis pour démarrer Hadoop.

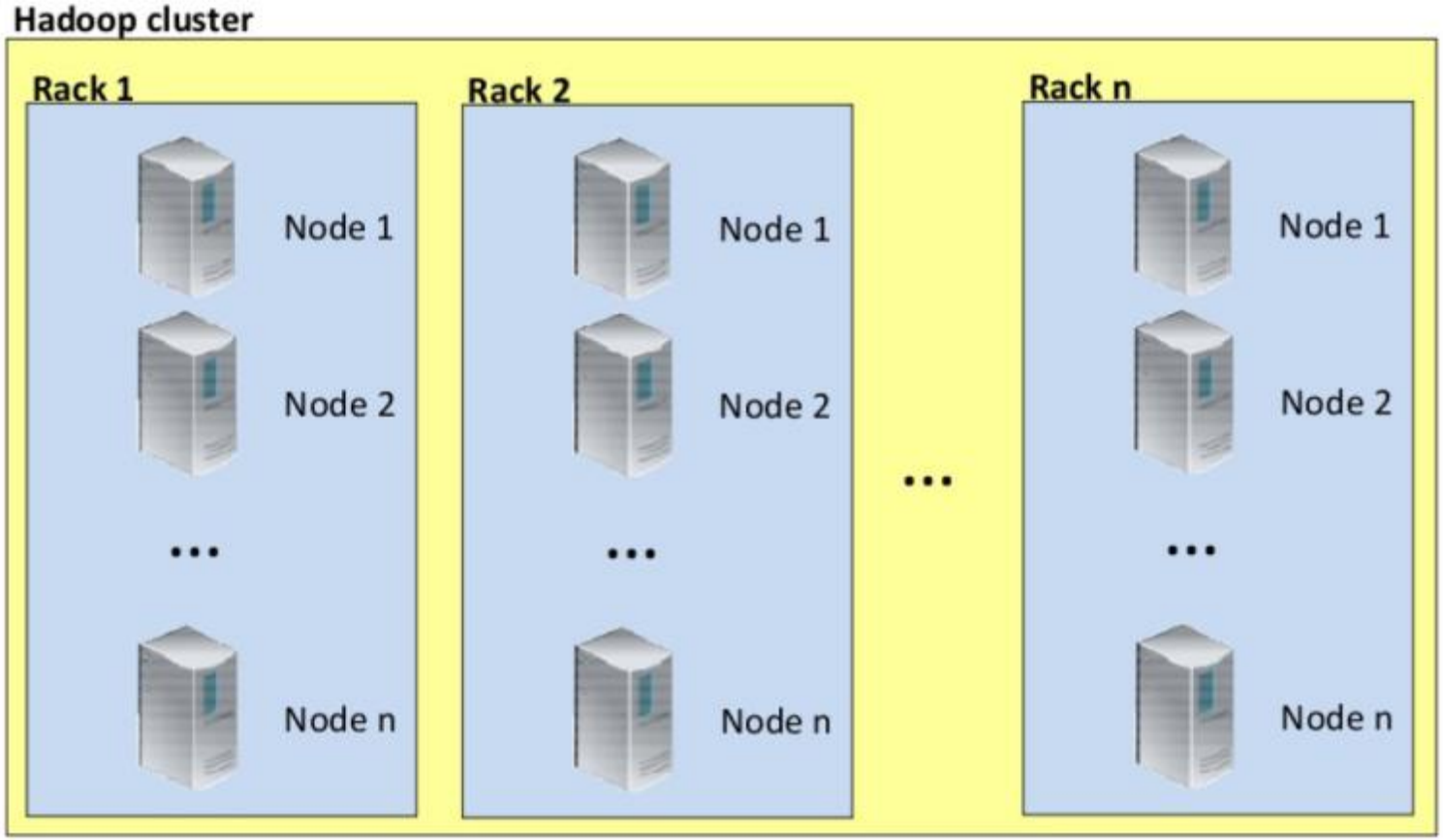
- **Hadoop MapReduce**
  - Système pour le traitement parallèle de grands ensembles de données.

- **Hadoop YARN (MapReduce 2.0)**
  - Responsable de la planification des tâches et la gestion des ressources de cluster.

- **Système de fichiers distribué Hadoop (HDFS)**

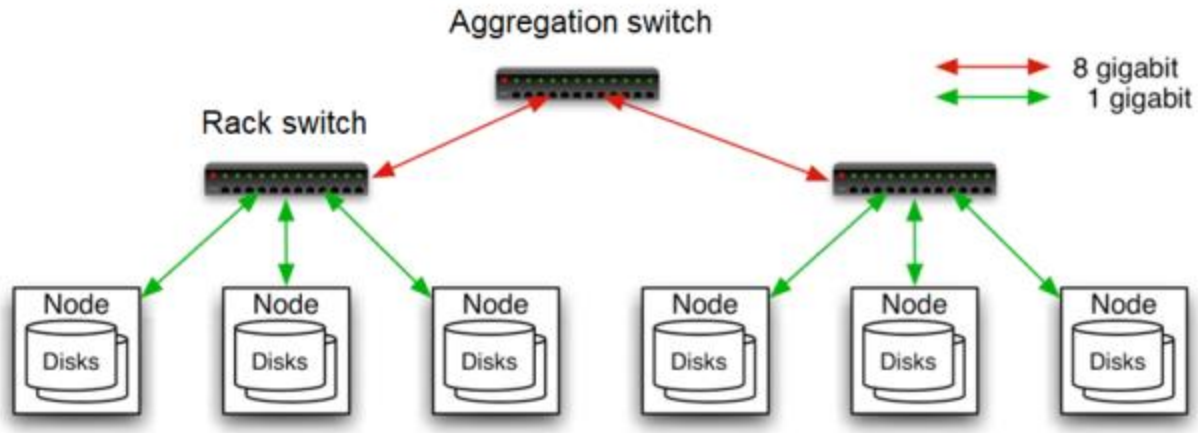
Hadoop

- Hadoop Cluster



Hadoop

- Hadoop Cluster



- Typically in 2 level architecture
  - Nodes are commodity PCs
  - 30-40 nodes/rack
  - Uplink from rack is 3-4 gigabit
  - Rack-internal is 1 gigabit

- Hadoop Cluster

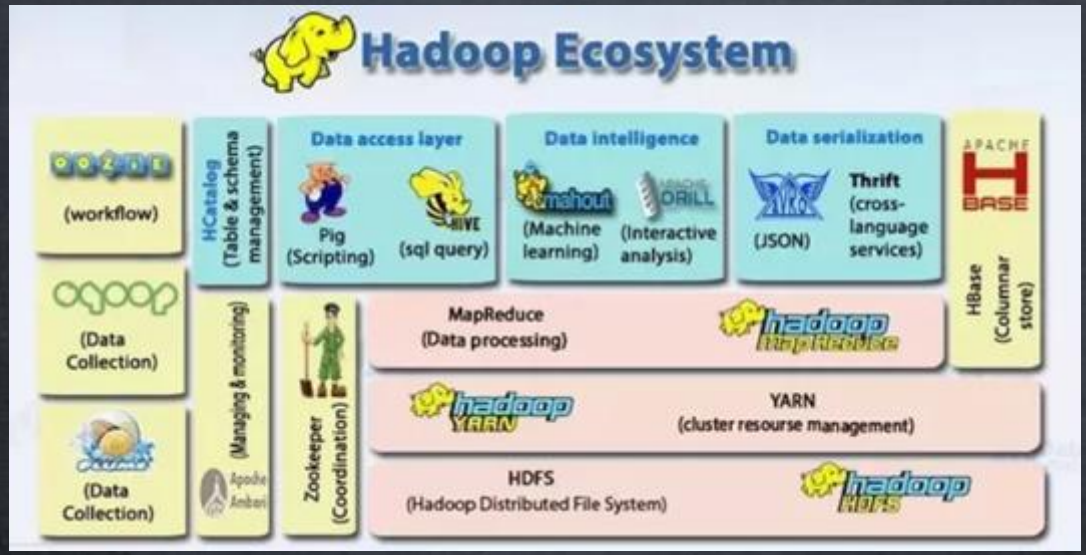
- **Un Data Center**
  - Imaginez 5000 ordinateurs connectés entre eux ; c'est un cluster :

Quelle sont les outils de le BIG DATA?

Big DATA means not only enormous DATA but it is a complete subject with a large set of tools.

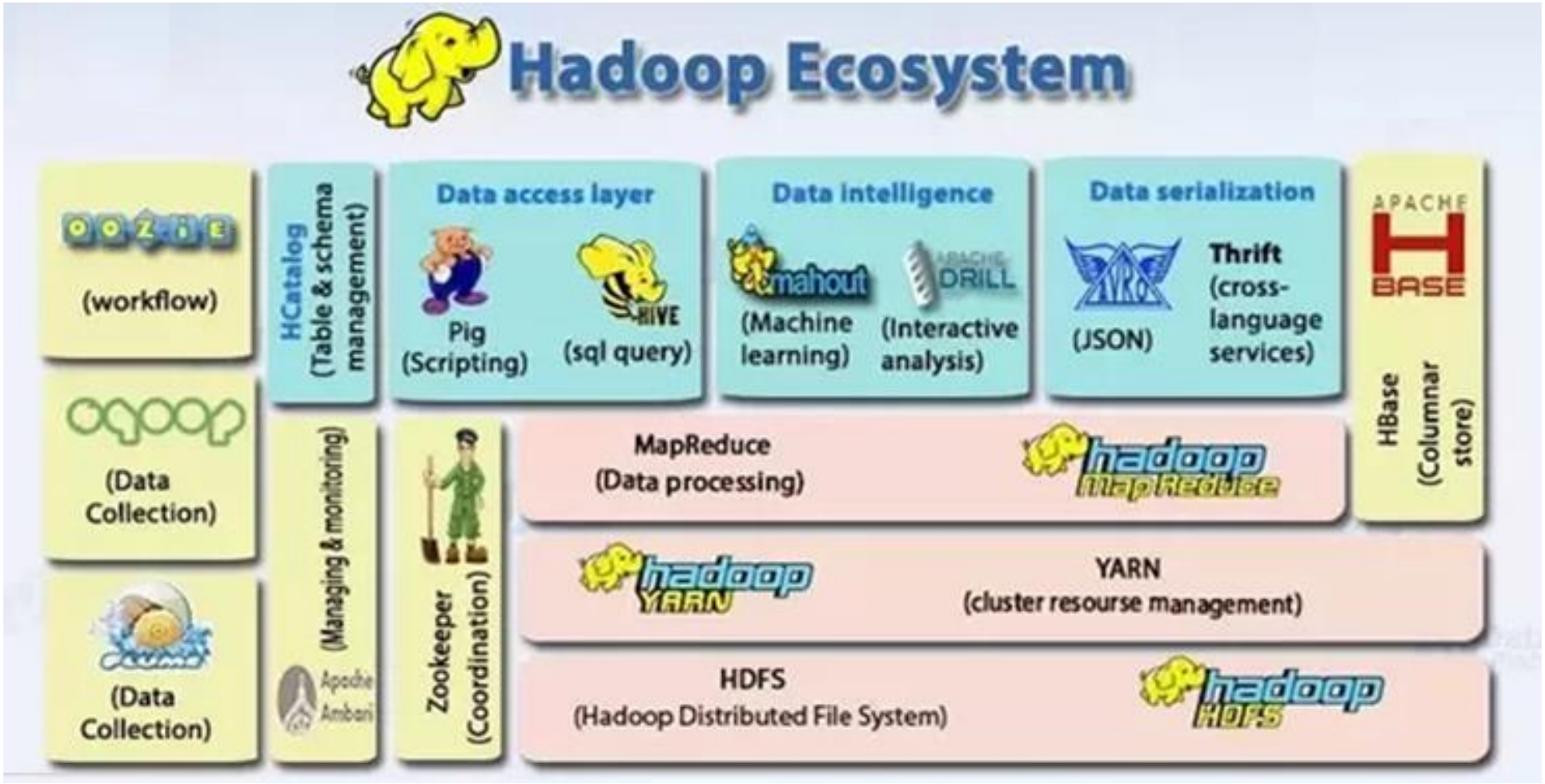
- **Hadoop Ecosystème**

- Le terme **"Hadoop"** fait souvent référence non seulement aux modules de base déjà discutés, mais aussi à la **collection d'outils supplémentaires** qui peuvent être installés sur ou à côté de Hadoop.



Quelle sont les outils de le BIG DATA?

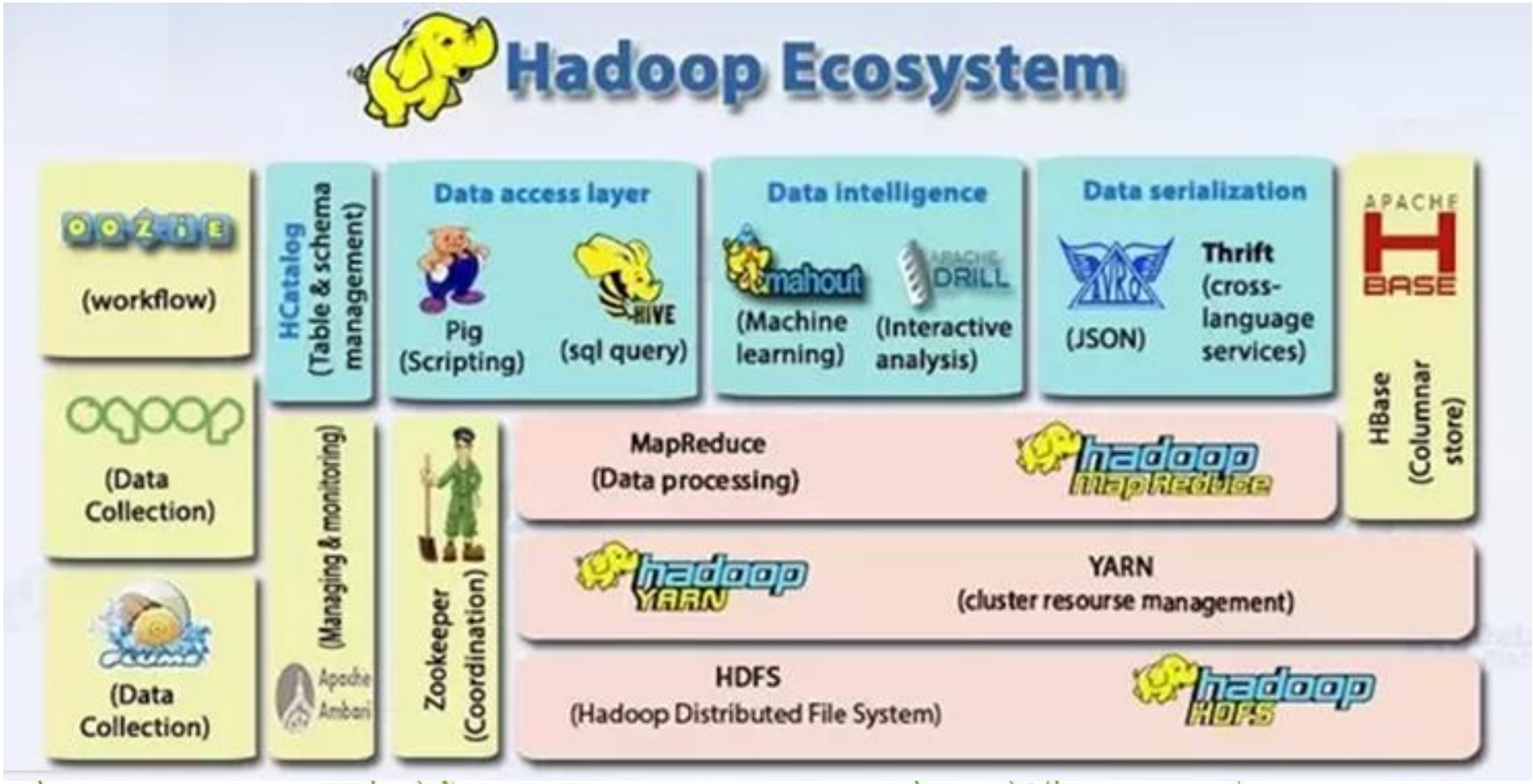
Hadoop



Quelle sont les outils de le BIG DATA?

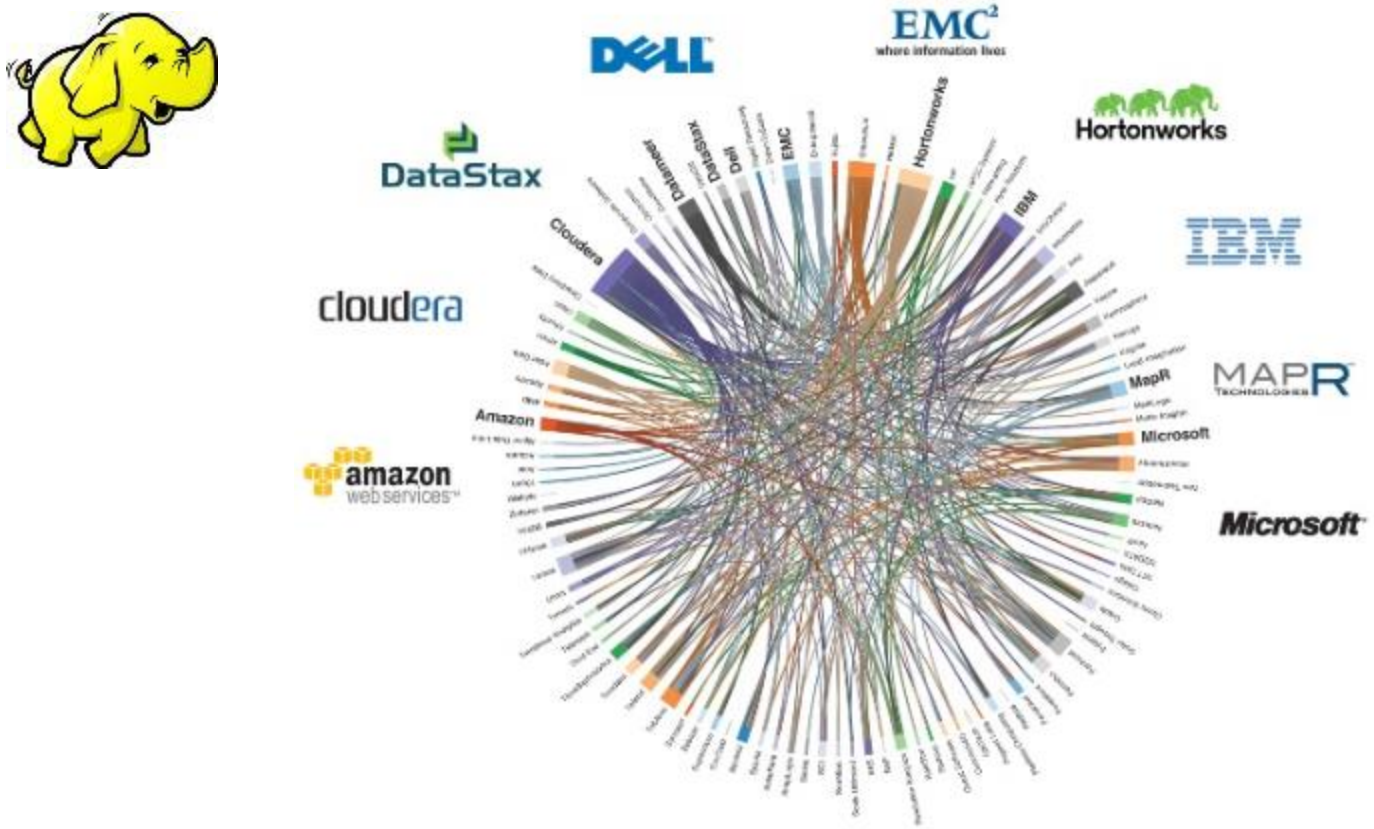
Hadoop

What is a BIG DATA distribution?



Hadoop

- Hadoop vendors



Hadoop

- Who uses Hadoop
  - Amazon
  - Facebook
  - Google
  - New York Times
  - Yahoo!
  - …. many more