

Pr. Mohamed  
GHAZOUANI

# BIG DATA ET HADOOP

Partie 2

1

18 novembre 2024

## Plan du cours

1	• C'est quoi Hadoop ?
2	• Domaines d'application d'Hadoop
3	• Modes de fonctionnement de Hadoop
4	• L'eco-système Hadoop

Pr. M. Ghazouani

2

18 novembre 2024

Excel

Registre

Pr. M. Ghazouani

3

18 novembre 2024

Excel-S/W

Excel-S/W

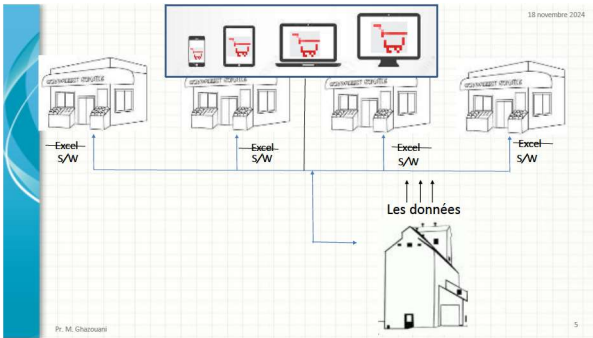
Excel-S/W

Excel-S/W

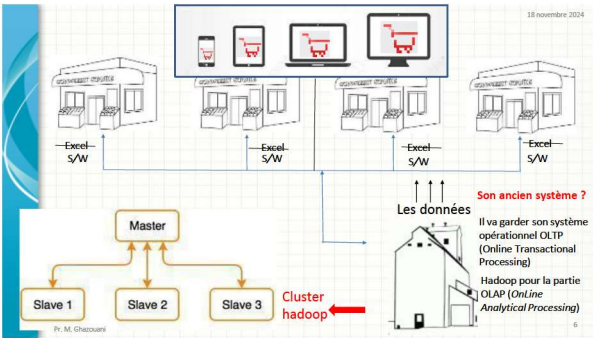
Les données

Pr. M. Ghazouani

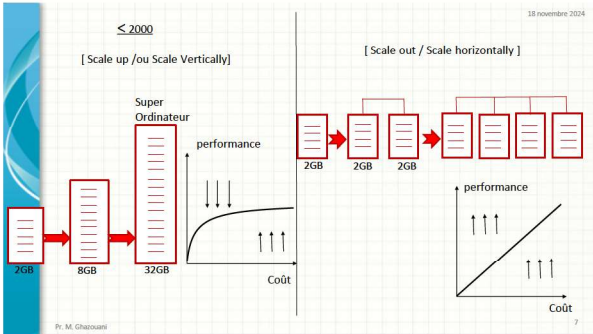
4



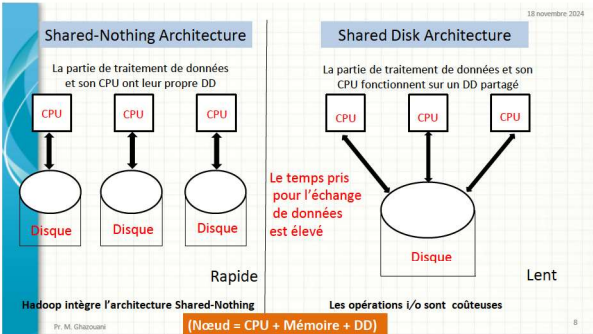
5



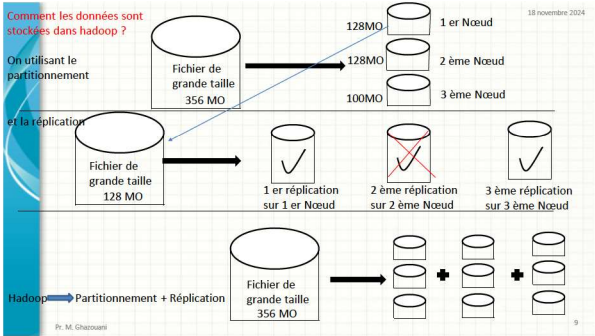
6



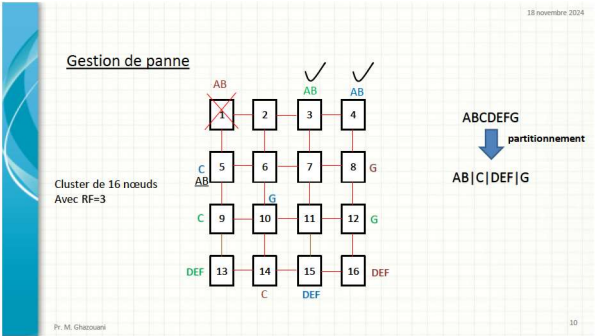
7



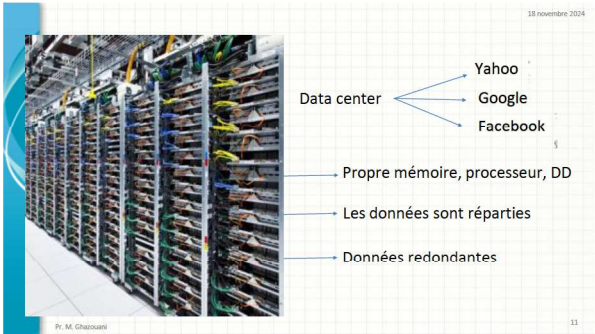
8



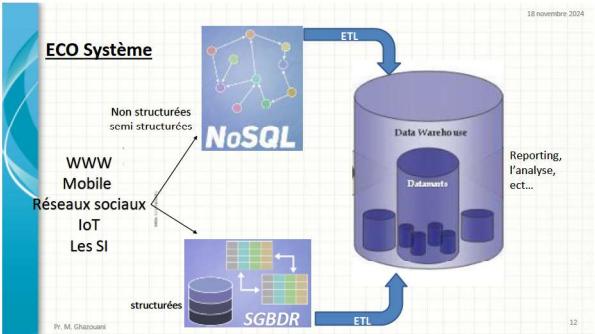
9



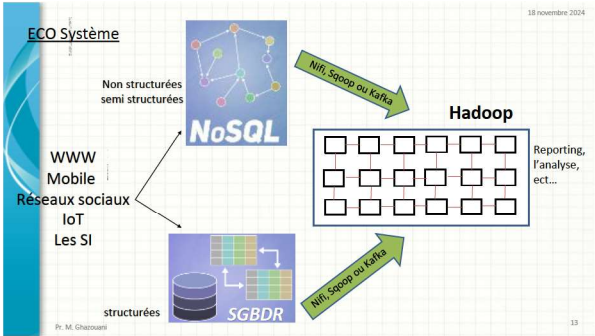
10



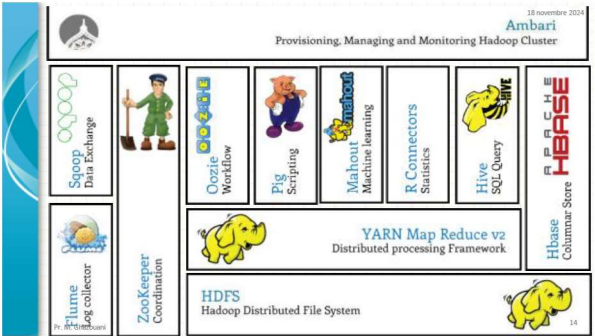
11



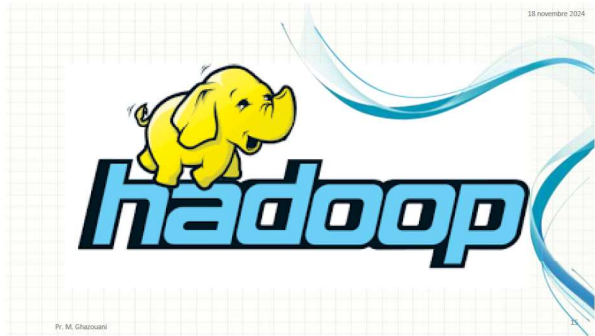
12



13




14



15

### Hadoop?

- **Hadoop** est un framework libre et open source écrit en Java destiné à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données.
- Tous les modules de Hadoop sont conçus dans l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par le framework.
- Le nom "Hadoop" était initialement celui d'un éléphant en peluche, jouet préféré du fils de Doug Cutting.



2004 : conçu par Doug Cutting.

16

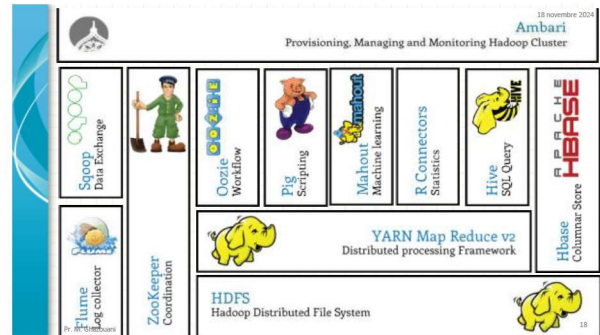
## Domaines d'application d'Hadoop

Hadoop est utilisé par des entreprises ayant de très fortes volumétries de données à traiter. Parmi elles, on trouve notamment des géants du web comme Facebook, Twitter, LinkedIn, ou encore les géants de l'e-commerce à l'instar de eBay et Amazon.



- Facebook
- Google
- Yahoo
- Amazon
- Microsoft
- Twitter
- LinkedIn
- IBM
- Airbnb
- Uber
- Netflix
- eBay
- Airbnb
- Spotify
- Adobe
- Intel
- NASA
- Goldman Sachs
- Walmart
- General Electric

17



18



## Apache Hadoop Ecosystem

**HDFS**  
Hadoop Distributed File System



19

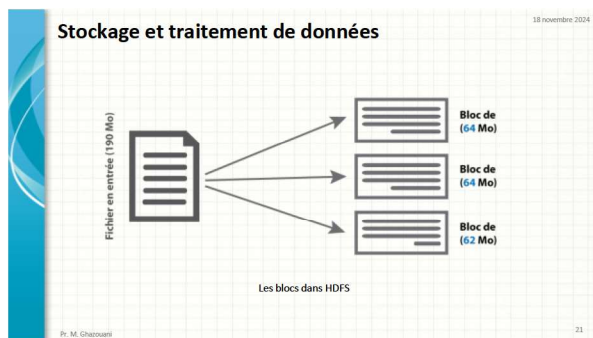
## Hadoop Distributed File System (HDFS)

HDFS est un système de fichiers distribué qui donne un accès haute-performance aux données réparties dans des clusters Hadoop.

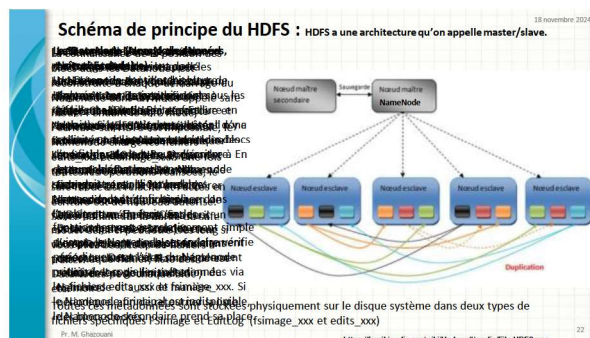
- **Système de stockage** : HDFS utilise des tailles de blocs largement supérieures à ceux des systèmes classiques. Par défaut, la taille est fixée à 64 Mo. Il est toutefois possible de monter à 128 Mo, 256 Mo, 512 Mo voire 1 Go. Alors que sur des systèmes classiques, la taille est généralement de 4 Ko. L'intérêt de fournir des tailles plus grandes permet de réduire le temps d'accès à un bloc.
- **Possibilité de stocker des pétaoctets de données**
- **Traitement parallèle et distribué** : le système segmente l'information en plusieurs briques et les distribue sur plusieurs nœuds du cluster, ce qui permet alors le traitement en parallèle.
- **Tolérance aux pannes avec la réplication des données** : Pour la phase de lecture, si un bloc est indisponible sur un nœud, des copies de ce bloc seront disponibles sur d'autres nœuds.

20

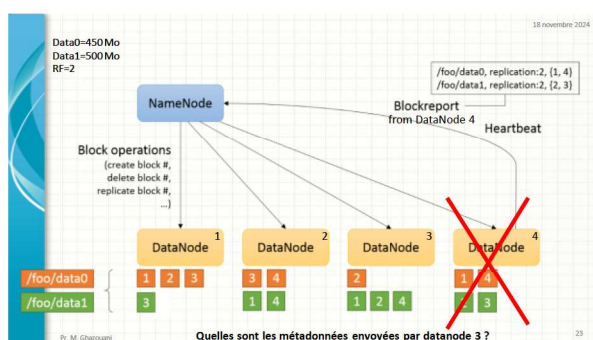




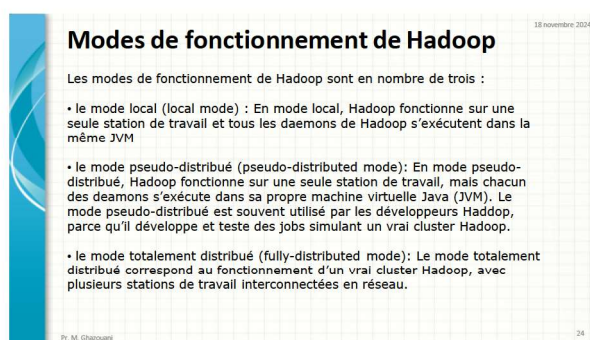
21



22



23



24

## Installation - Hadoop-v2.7.3

Cette partie de cours s'intéresse à l'installation et la configuration d'un cluster Hadoop simple en utilisant la version proposée par la fondation Apache. Hadoop n'a d'intérêt que s'il est utilisé dans un cluster composé de plusieurs machines.

Pour télécharger Hadoop deux solutions sont disponibles :

1. Utiliser la version proposée par la fondation Apache.
2. Utiliser les distributions fournies par des entreprises qui font du service autour de Hadoop.
  - Cloudera - <https://www.cloudera.com/>
  - Hortonworks - <https://hortonworks.com/>
  - MapR Technologies - <https://mapr.com/>



Pr. M. Ghannassi

25

« Ce que j'entends je l'oublie, ce que je vois je m'en souviens, et ce que je fais je le comprends ».



Pr. M. Ghannassi

26

## Utilisation de Hadoop

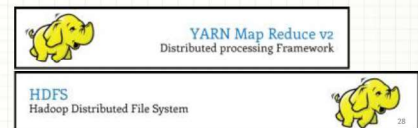
### Manipulation de HDFS

27



## Apache Hadoop Ecosystem

- Map reduce est le composant de base de traitement : Il fournit la logique de traitement. Est un Framework logiciel qui aide à écrire des applications qui traitent de grands ensembles de données en utilisant des algorithmes distribués et parallèles dans l'environnement Hadoop.
- YARN est l'un des principaux composants de Apache Hadoop. Il permet de gérer les ressources du système et de planifier les tâches. On peut dire que Yarn est le cerveau de l'écosystème de Hadoop.



Pr. M. Ghannassi

28

