

Cours Gestion des Données Massives

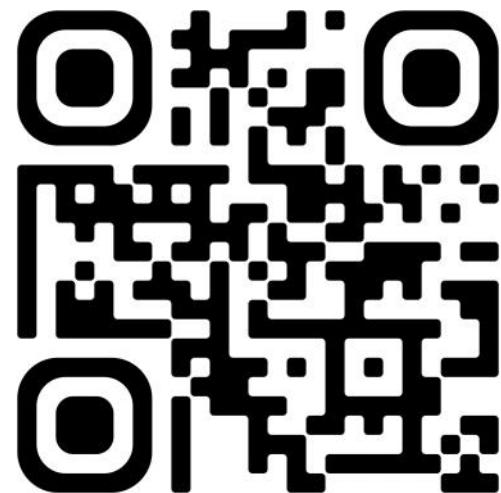
DSI3

SEHIMI SABEUR

Sommaire

1. Introduction au Big Data
2. Présentation de Hadoop
3. Hadoop Distributed File System (HDFS)
4. MapReduce
5. BD NoSQL

Join Google Classroom



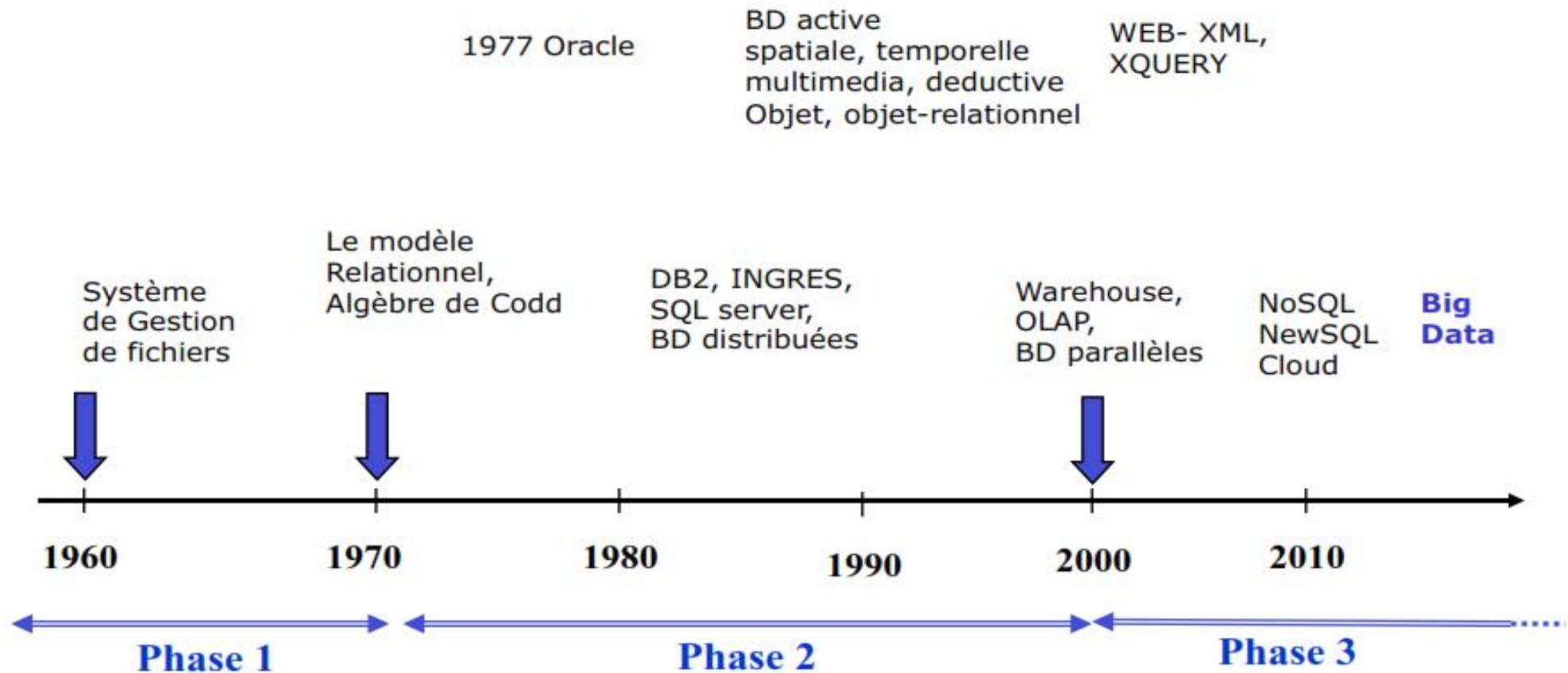
Scan Me

Introduction au Big Data

Big Data – Définition 1

Les big data, (grosses données, ou mégadonnées, données massives), désignent des ensembles de données qui deviennent tellement **volumineux** qu'ils en deviennent **difficiles** à travailler avec des **outils classiques** de gestion de base de données ou de gestion de l'information.

Stockage Data

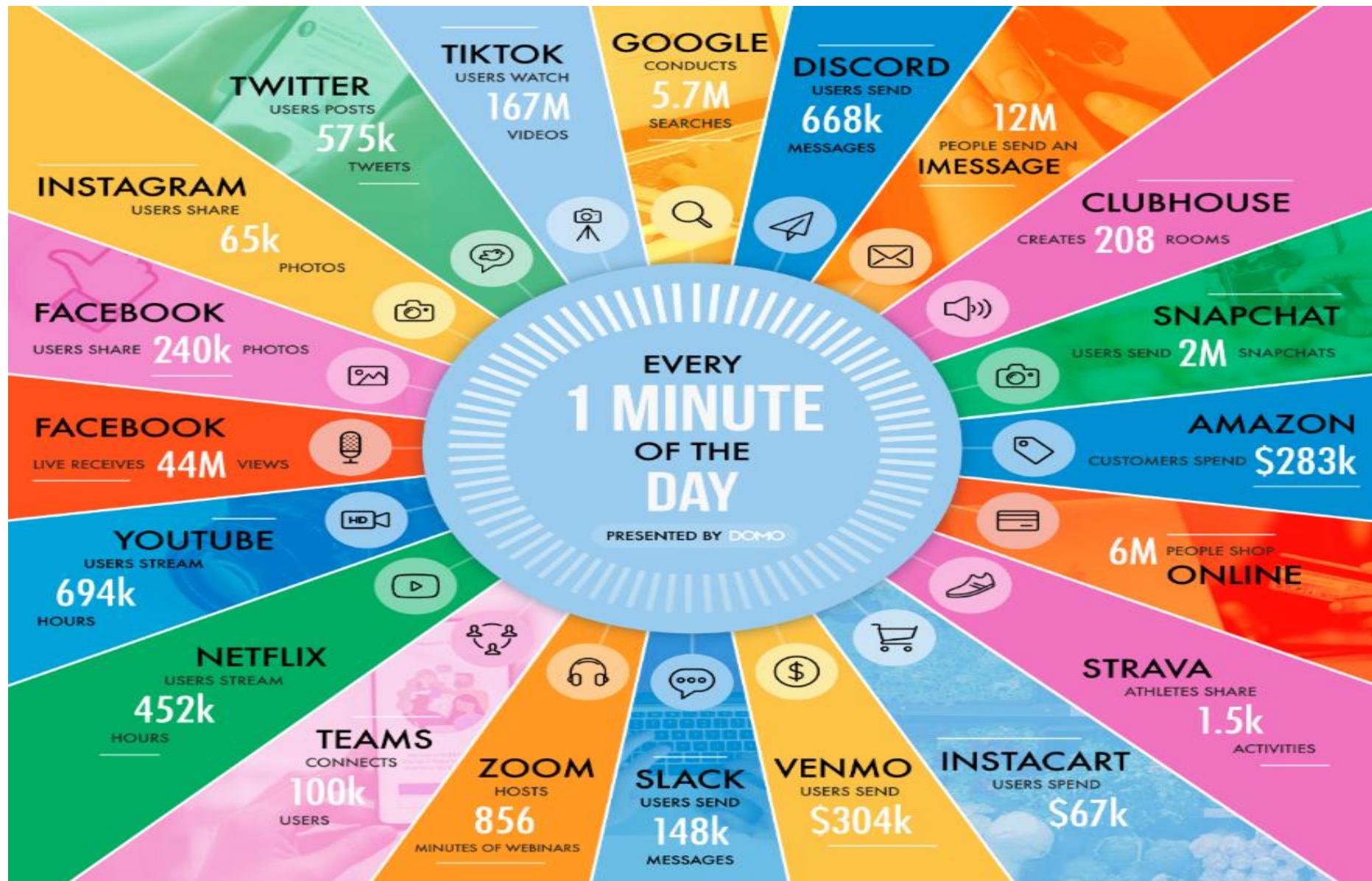


Big Data – Données Massives

- ▶ En avril 2022, l'internet atteint 63 % de la population mondiale, soit environ 5 milliards.
- ▶ Sur ce total, 4,65 milliards, soit plus de 93 %, étaient des utilisateurs des médias sociaux.
- ▶ La quantité totale de données qui devraient être créées, capturées, copiées et consommées dans le monde en 2022 est de 97 zettaoctets,
- ▶ La taille devrait atteindre 181 zettaoctets d'ici 2025.

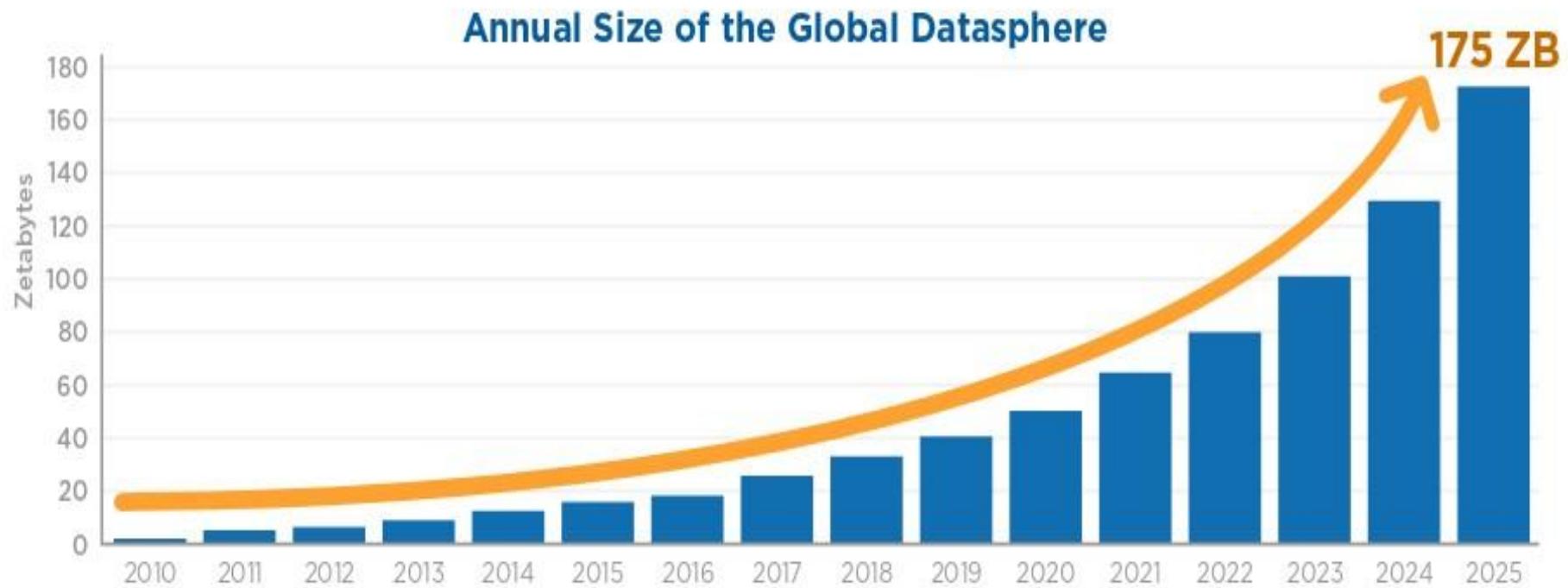
Big Data – volume de données

10000000000000000000000000000000	10^{24}	yotta	Y	septillion
1000000000000000000000000000000	10^{21}	zetta	Z	sextrillion
1000000000000000000000000	10^{18}	exa	E	quintillion
1000000000000000	10^{15}	peta	P	quadrillion
1000000000000	10^{12}	tera	T	trillion
1000000000	10^9	giga	G	billion
1000000	10^6	mega	M	million
1000	10^3	kilo	k	thousand



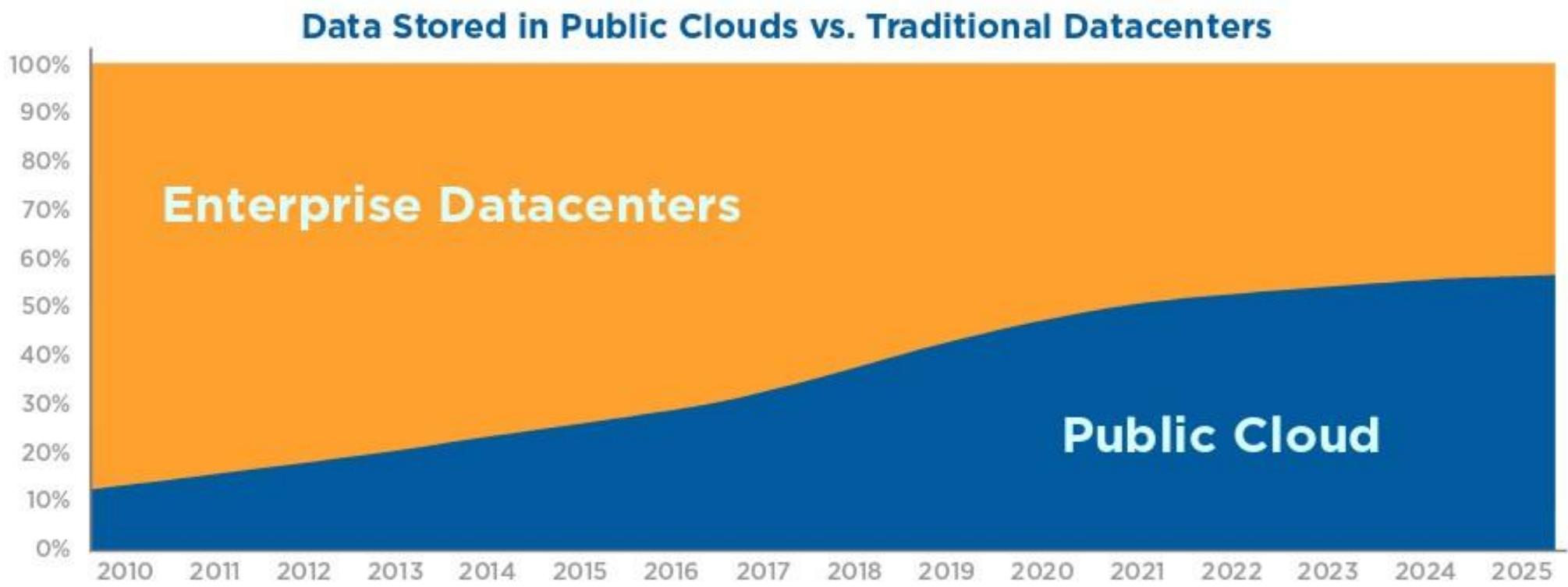
9

Big Data – Données Massives



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Big Data – Sources de données



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Big Data – Données Massives



Big Data – Domaines

- Quelques exemples d'utilisation de la Big Data

- Science
- Astronomy
- Atmospheric science
- Genomics
- Biogeochemical
- Biological
- Social
- Social networks
- Social data
 - * Twitter
 - * Facebook
 - * LinkedIn
- Commercial
- Web / event / database logs
- Sensor networks
- Internet text and documents
- Medical records
- Photographic archives
- Video / audio archives
- Government
- Military and homeland security surveillance

Big Data – 4V

On utilise souvent le précepte des **4 V du Big Data** établis par des experts du domaine:

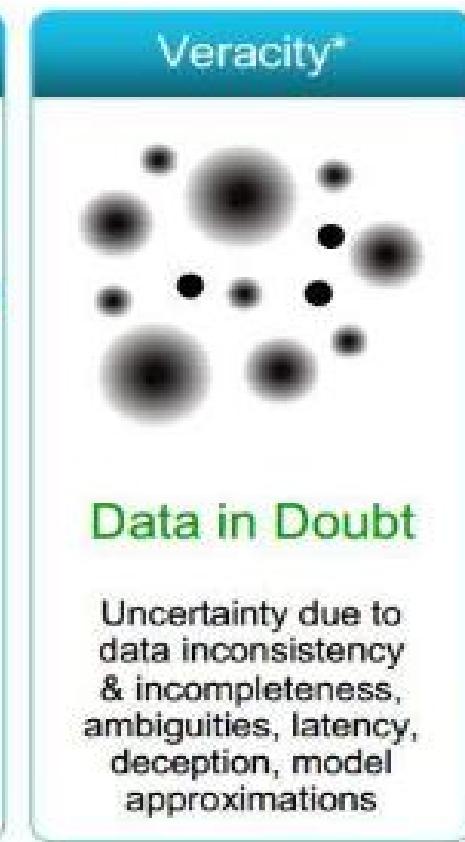
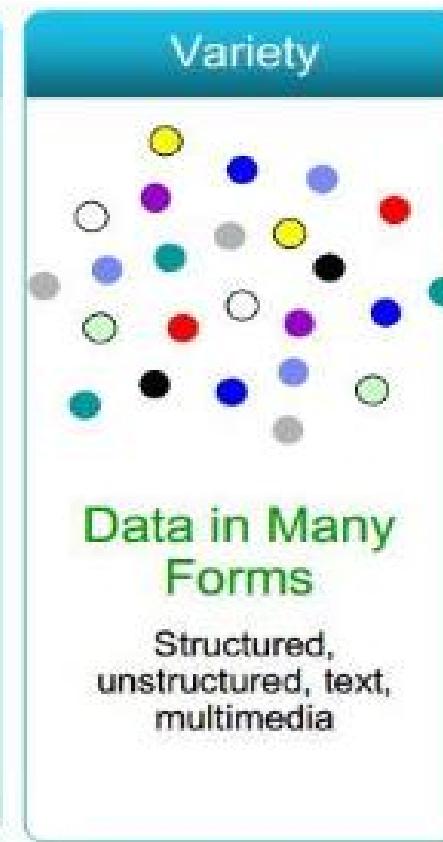
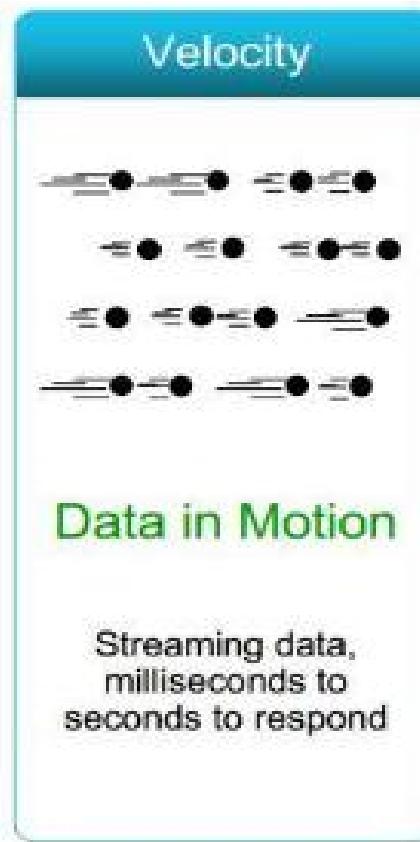
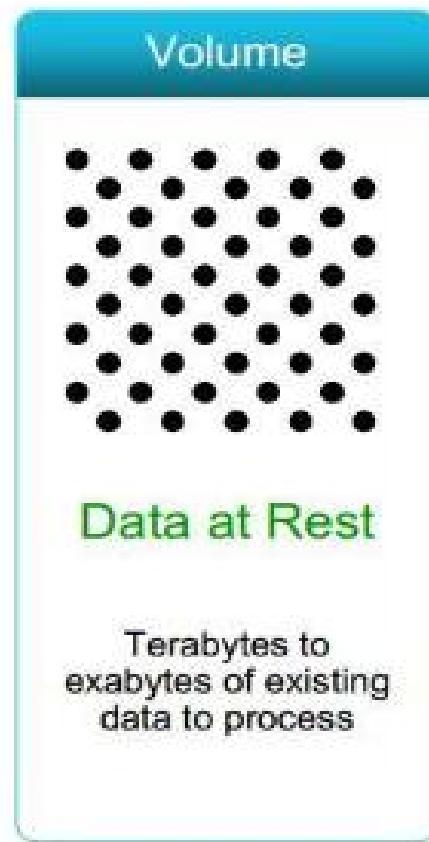
- ▶ **Volume** : développer un plan pour **gérer la quantité de données** qui seront en jeu et où et comment elles seront hébergées
- ▶ **Variété** : identifier toutes les différentes **sources de données** dans l'écosystème numérique et s'équiper des bons outils pour l'ingestion.
- ▶ **Vitesse** : rechercher et déployer les bonnes technologies pour s'assurer que les **données volumineuses** sont traitées de manière à être utilisées **quasiment en temps réel**.
- ▶ **Véracité** : **nettoyer les datas** et faire en sorte que les données collectées soient exactes et prêtes à l'emploi.

Big Data – 7V

On peut ajouter le paramètre suivant:

- ▶ **Valeur** : créer un **environnement Big Data** qui met en évidence la BI de manière exploitable et priorise les informations importantes pour chaque équipe du personnel, les données auront donc une **utilisation bénéfique** pour le tiers qui investi dans le Big Data.
- ▶ **Visualisation**: Les données générées doivent être accessibles et lisibles de la part des prestataires de service.
- ▶ **Variabilité**: il est important d'avoir à disposition des outils permettant d'identifier, de traiter et de filtrer les données de qualité variable pour en optimiser l'utilisabilité. Elle dépend directement de l'hétérogénéité des plateforme et infrastructures générant les données.

Big Data – Données Massives



Big Data – Définition 2

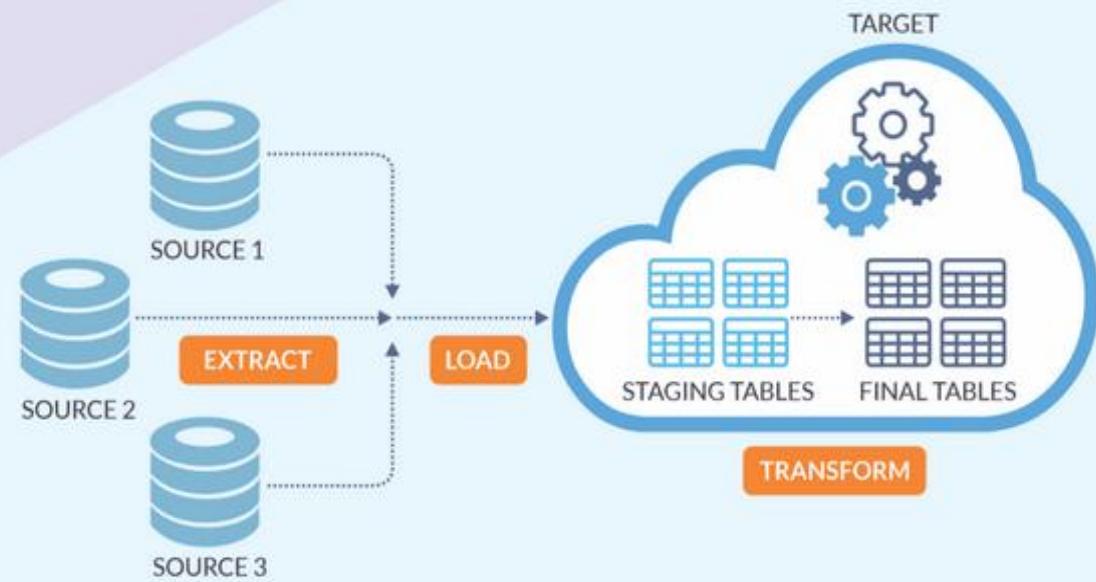
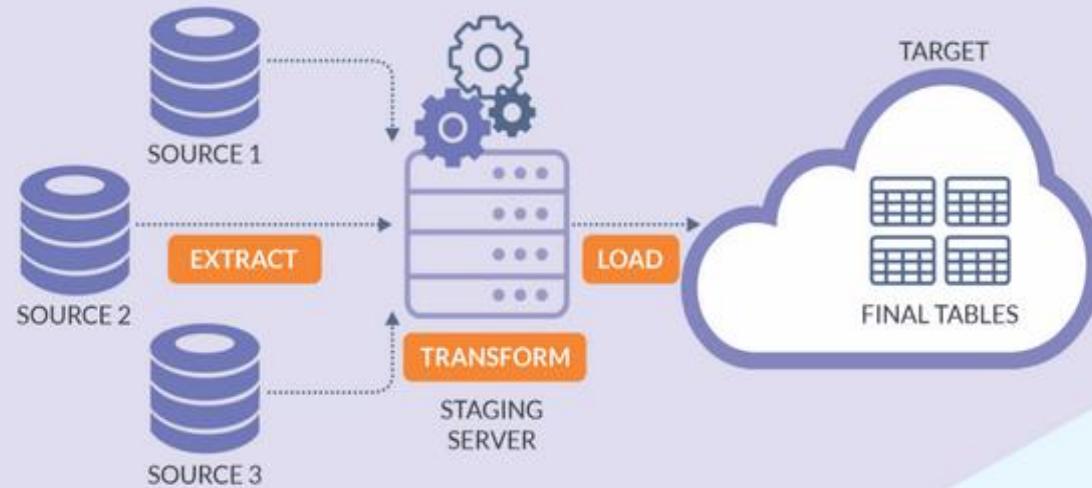
Autre définition:

“Le Big Data (ou mégadonnées) représente les collections de données caractérisées par un **volume**, une **vélocité** et une **variété** si grands que leur transformation en **valeur utilisable** requiert l'utilisation de technologies et de **méthodes analytiques spécifiques**.”

Big Data et Business Intelligence

- ▶ Business Intelligence conçue pour décrire
 - ▶ l'ingestion
 - ▶ l'analyse
 - ▶ l'application d'ensembles de données au profit d'une stratégie d'entreprise
- ▶ la Business Intelligence emploie le **Big Data** au service du produit
- ▶ cartographier et en prédire l'activité et les points clés qui constituent des challenges à relever pour la **prise de décision**.

ETL VS ELT



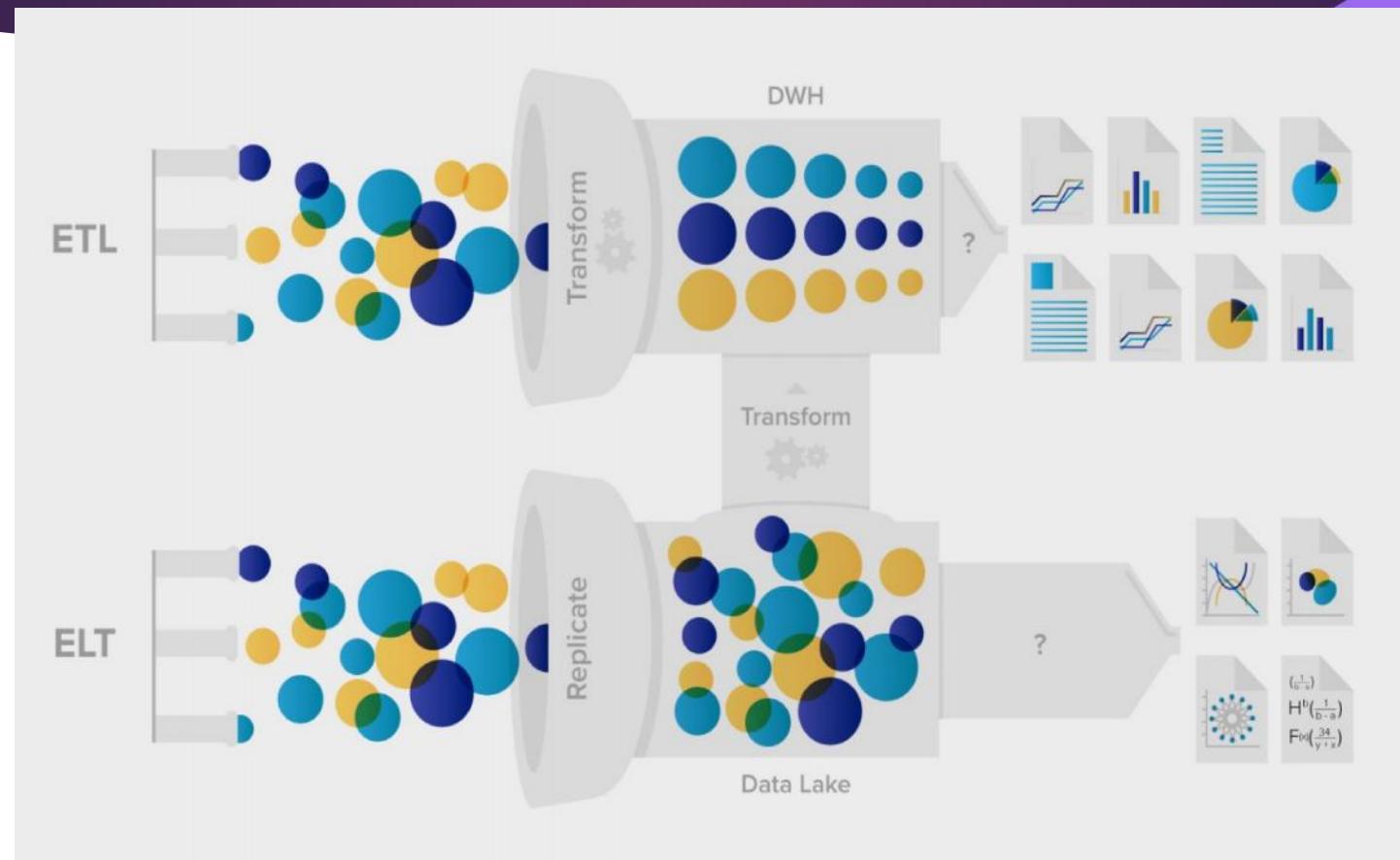
Les entrepôts de données et data lakes

- ▶ L'**analyse Big Data** consiste à examiner de très grands ensembles de données granulaires pour découvrir des **modèles** et **corrélations** cachées ainsi que des tendances et de **nouvelles informations** aux niveaux commercial et marketing (entre autres)
- ▶ Le modèle classique de stockage **entrepôt de données (data warehouse)**
- ▶ **Un data warehouse est un vaste gisement de données qui facilite la prise de décision dans l'entreprise**
- ▶ Peut être géré :
 - ▶ Localement (stockage privé)
 - ▶ cloud

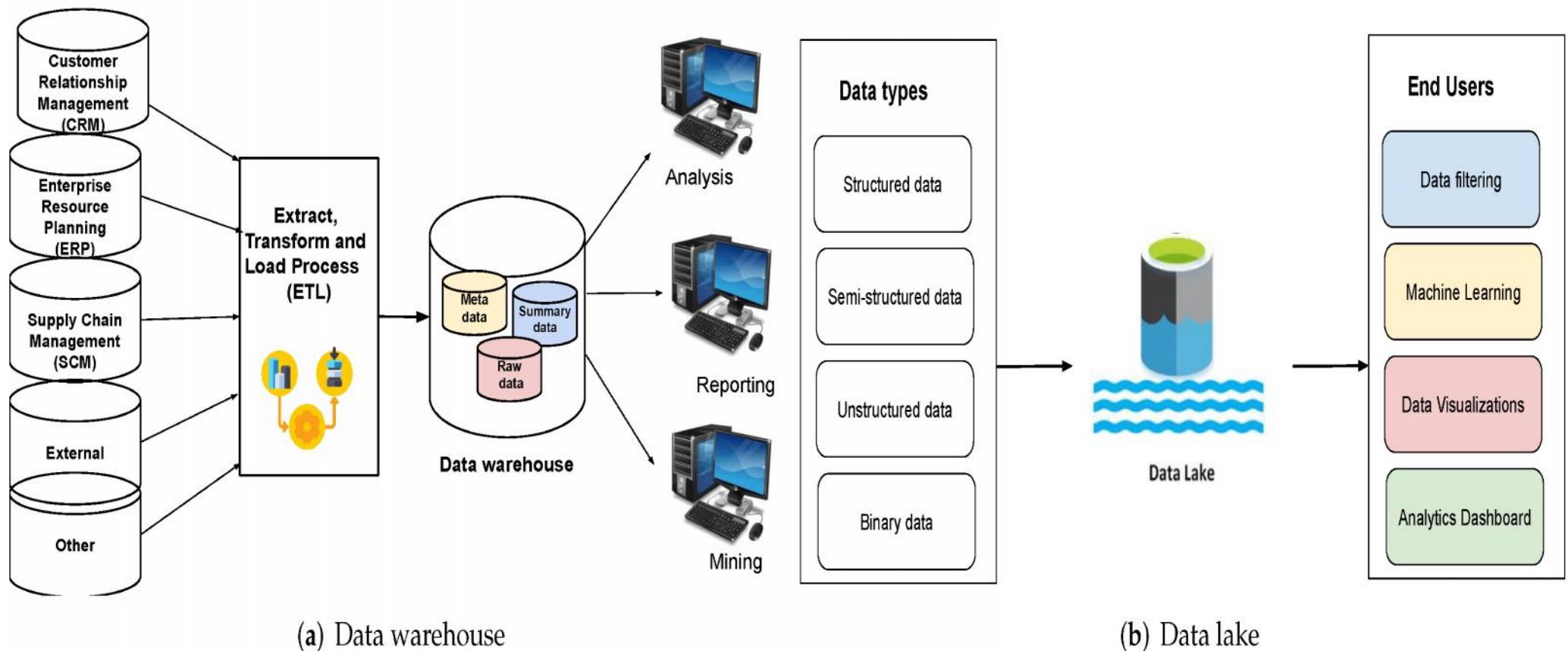
Les entrepôts de données et data lakes

- ▶ Les **data lakes** sont un référentiel de stockage central qui contient les **données volumineuses provenant de nombreuses sources** différentes et dans un format brut et granulaire...
- ▶ Le **data lake** peut stocker des données structurées, semi-structurées ou non structurées.
- ▶ Elles peuvent être conservées dans un format quelconque pour une utilisation flexible et un traitement **futur**.

Les entrepôts de données et data lakes



Les entrepôts de données et data lakes



Présentation de Hadoop

Evolution Hardware

- Vitesse CPU
 - 1990 – 44 MIPS at 40 MHz
 - 2020 – 147,600 MIPS at 4GHz
- RAM Memory
 - 1990 – 640K conventional memory (256K extended memory recommended)
 - 2020 – 32GB (and more)
- Disk Capacity
 - 1990 – 20MB
 - 2020 – 1TB (and more)

Evolution Hardware (débit de transfert de données)

- Disk Latency (vitesse de lecture et d'écriture) - pas beaucoup d'améliorations dans les 10 dernières années , actuellement environ 70 - 80MB / sec

How long will it take to read 1TB of data?

1TB (at 80Mb / sec):

1 disk - 3.4 hours

10 disks - 20 min

100 disks - 2 min

1000 disks - 12 sec

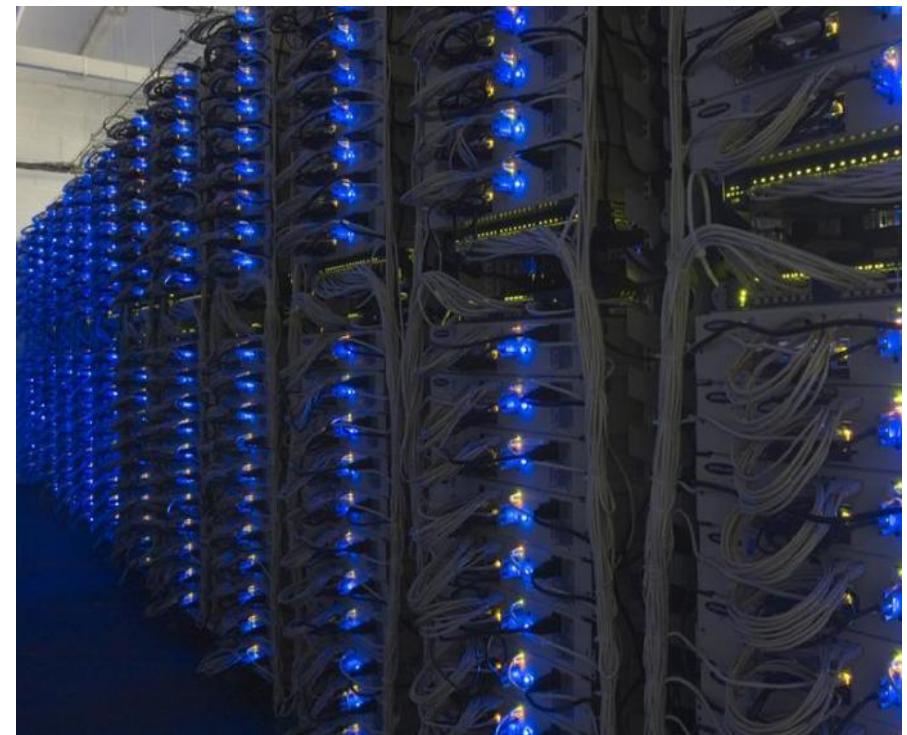
Système Distribué

Nous avons travaillé longtemps avec:

- GRID computing - propage la charge de traitement –
- Distributed workload - difficile pour gérer les applications , les frais sont lourds.
- Parallel databases - DB2 DPF , Teradata , Netezza , etc (distribue les données)

Système Distribué

- Distributed computing : plusieurs ordinateurs apparaissent comme un super-ordinateur pour communiquer les uns avec les autres par envoi de messages , fonctionnent ensemble pour atteindre un but commun



Système Distribué

- Les technologies traditionnelles ne permettent pas de traiter efficacement de grandes quantités de données.
- Le calcul distribué permet de travailler avec le Big Data en utilisant des quantités raisonnables de temps et de ressources.



Système Distribué

- Challenges du Distributed computing :
- Hétérogénéité
- Extensibilité
- Sécurité
- Evolutivité
- Concurrence
- La tolérance aux pannes



Hadoop

- Apache framework est un environnement open source fiable, évolutif et distribué pour le calcul de quantité massive de données
 - Masque les détails et la complexité du système sous-jacents à l'utilisateur
 - Développé en Java



Hadoop

- Consiste en 3 Sous projets:
 - Hadoop Common
 - MapReduce
 - Hadoop Distributed File System (HDFS)
- Conçu pour du matériel de base hétérogène
- Nouvelle façon de stocker et traiter les données
- + *Apporte le traitement aux données*



Hadoop

- Optimisé pour gérer :
 - Quantités massives de données à travers le parallélisme
 - **Variété** de données (structurées , non structurées , semi-structurées)
 - Utilisation de matériel de base peu coûteux
- Fiabilité fournie par la **réPLICATION**

Hadoop

- **Flexible,**
- **Supporte un large processing de données**
 - Inspiré par la technologie Google (MapReduce, GFS, BigTable, ...)
 - Initié par Yahoo
 - Well-suited to batch-oriented, read-intensive applications
 - Supporte une large variété de données

Hadoop

Permet aux applications de travailler avec des milliers de nœuds et des pétaoctets parallèlement et avec coût modéré.

- CPU + disks = “node”
- “Nodes” peuvent être combinés en clusters
- de nouveaux “nodes” peuvent être ajoutés selon besoin sans changement de:
 - Data formats
 - comment les données sont chargées
 - comment les jobs sont écrites

Hadoop

- **MapReduce framework**
 - Comment Hadoop comprend et affecte le travail aux nodes (machines)
- **Hadoop Distributed File System = HDFS**
 - Où Hadoop stocke les données
 - Un système de fichiers qui couvre tous les «nodes» d'un cluster Hadoop
 - Il relie les systèmes de fichiers sur de nombreux «nodes » locaux pour en faire un grand système de fichiers



Apache Hadoop Ecosystem



Ambari

Provisioning, Managing and Monitoring Hadoop Cluster



Sqoop
Data Exchange



ZooKeeper
Coordination



Oozie
Workflow



Pig
Scripting



Mahout
Machine learning

R Connectors
Statistics



Hive
SQL Query



Hbase
Columnar Store



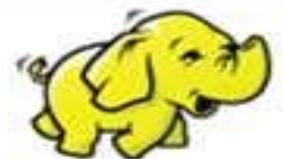
Flume
Log collector



YARN Map Reduce v2
Distributed processing Framework

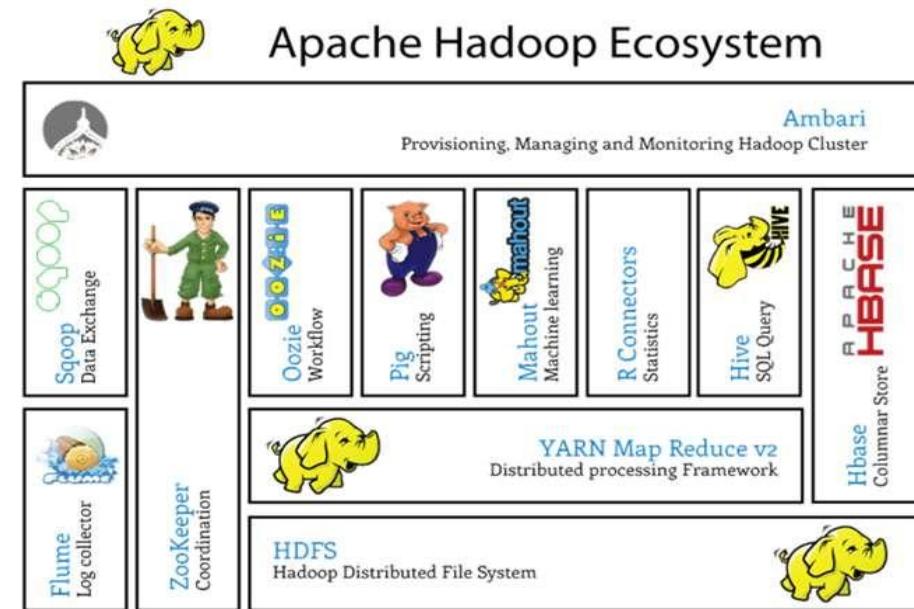


HDFS
Hadoop Distributed File System



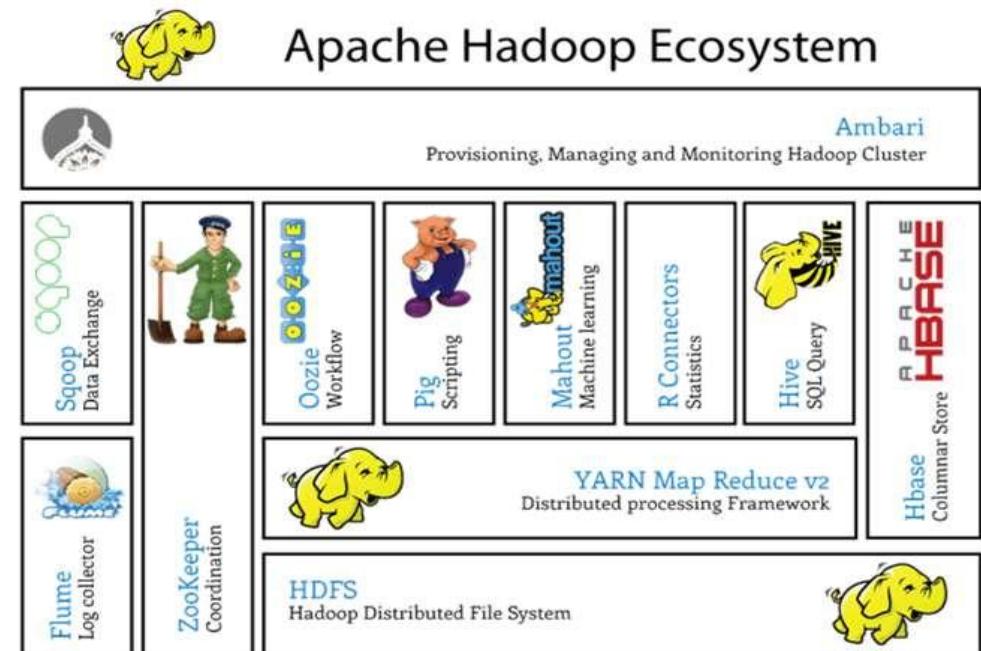
Hadoop ecosystem

- ▶ **Pig:** Plateforme haut niveau pour le traitement de données, basée sur un langage de script Pig Latin
- ▶ **Hive:** Environnement de haut niveau pour le traitement de données, basé sur un langage proche de SQL (Hive QL)
- ▶ **R Connectors:** permet l'accès à HDFS et l'exécution de requêtes Map/Reduce à partir du langage R
- ▶ **Mahout:** bibliothèque de machine learning et mathématiques
- ▶ **Oozie:** permet d'ordonnancer les jobs Map Reduce (Java, Python, Pig, Hive, ...), en définissant des workflows



Hadoop ecosystem

- ▶ **Hbase** : Base de données NoSQL orientée colonnes.
- ▶ **Sqoop**: Lecture et écriture des données à partir de BD externes
- ▶ **Flume**: Collecte de logs et stockage dans HDFS
- ▶ **Ambari**: outil pour la gestion et monitoring des clusters
- ▶ **Zookeeper**: fournit un service centralisé pour maintenir les information de configuration, de nommage et de synchronisation distribuée



Hadoop en pratique

