Data Mining et bases de données NoSQL

Wilfried Y. H. Adoni

Université Internationale de Casablanca

Informations pratiques

Déroulé du cours

- 10,5 heures de cours
- 15 heures TD/TP
- Evaluation : Examen +TP+ Mini projet

Objectifs pédagogiques :

- Comprendre les fondamentaux du big data et le paradigmes NoSQL
- Choisir une solution NoSQL adaptée aux besoins
- Déployer, administrer et utiliser un cluster Hadoop, Spark et Neo4J
- Effectuer des requêtages sur une base de données distribuée

Compétences acquises :

- Capacité à manipuler des données massives
- Capacité à déployer, administrer et utiliser une base de donnée distribué
- Capacité à migrer vers des calculs de haute performance

Université Internationale de Casablanca

Prérequis

Bases de données

- Compréhension des bases de données relationnelles
- Langage de requêtes SQL
- Notions d'optimisation de bases de données : indexation, hachage, plans d’exécution

Ingénierie informatique

- Bonne connaissance de l'environnement UNIX
- Connaissances réseau élémentaires

Université Internationale de Casablanca

Déroulé du cours

- **Partie I: Technologie NoSQL**
  - Cours 1: Data Mining et Technologie Big Data
  - Cours 2: Les familles NoSQL
  - Cours 3: Théorème de CAP
- **Partie II: Les bases de données NoSQL**
  - Cours 4: Hadoop
  - Cours 5: Spark
  - Cours 5: Neo4j
  - Cours 6: MongoDB

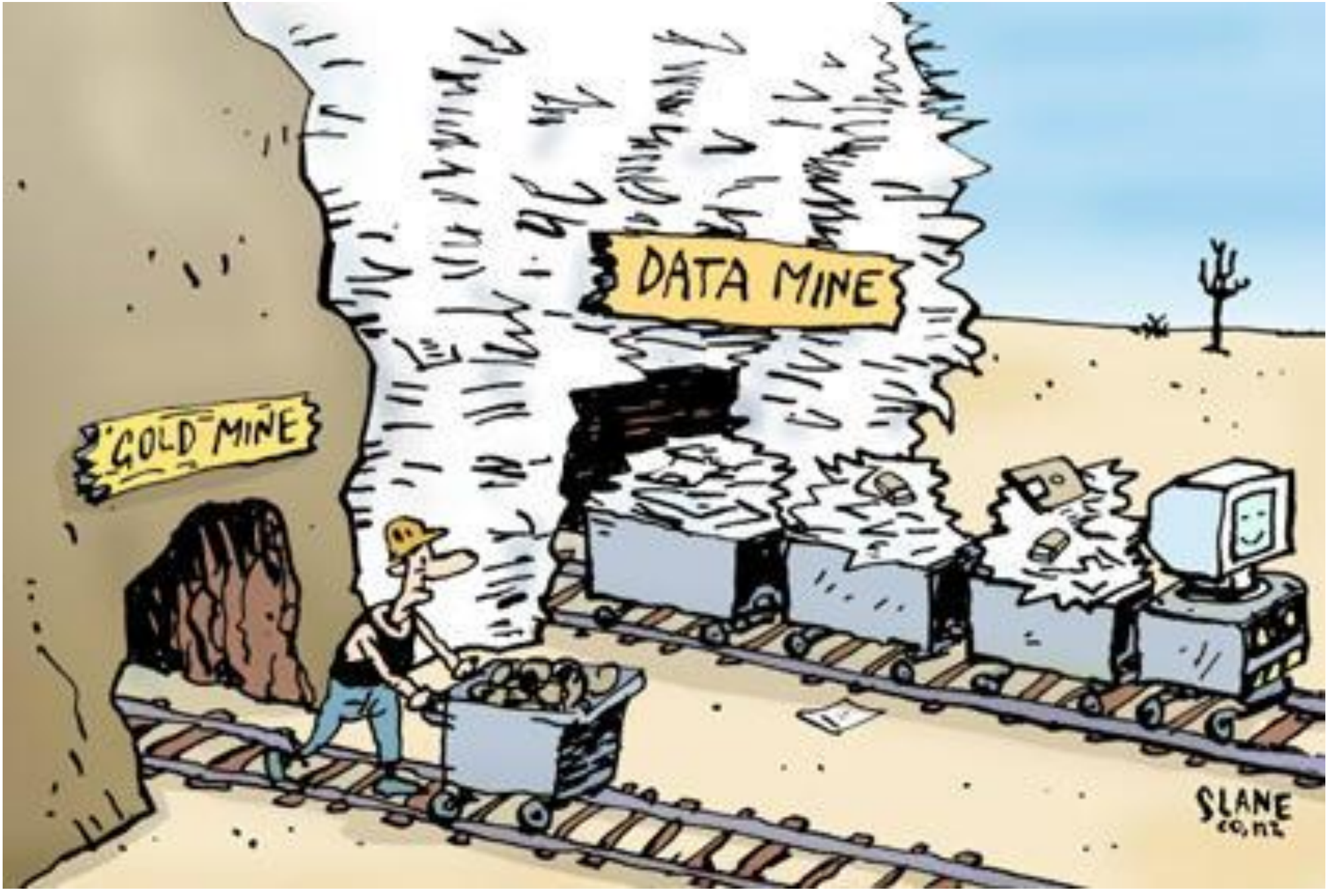
Data Mining et Technologie Big Data

Cours 1/7

Data Mining, ETL, Big Data, 5V, NoSQL

Université Internationale de Casablanca

Data Mining ?



Université Internationale de Casablanca

Procédé vers l'exploration des données

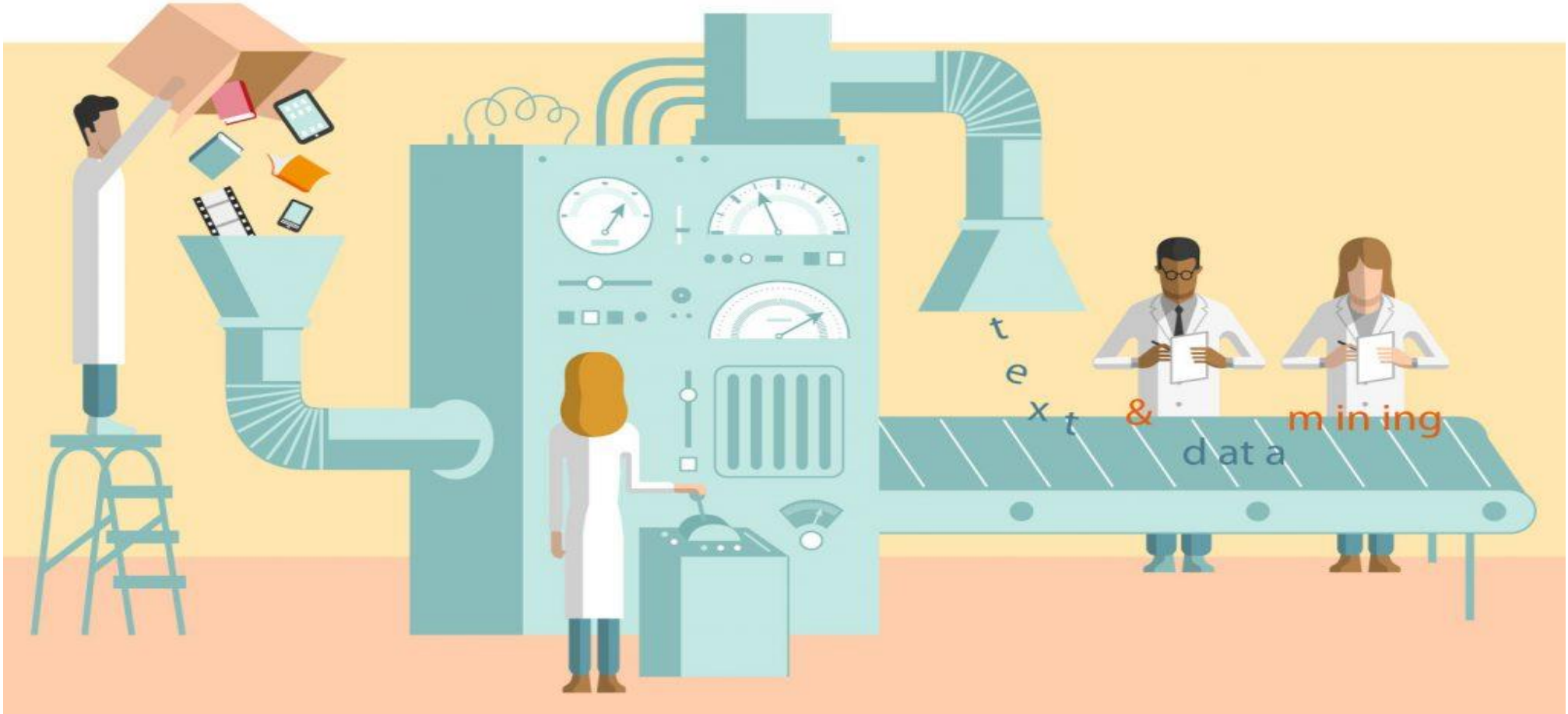
- **Les opérations ETL**
  - Extract
  - Transform
  - Load

Opérations ETL

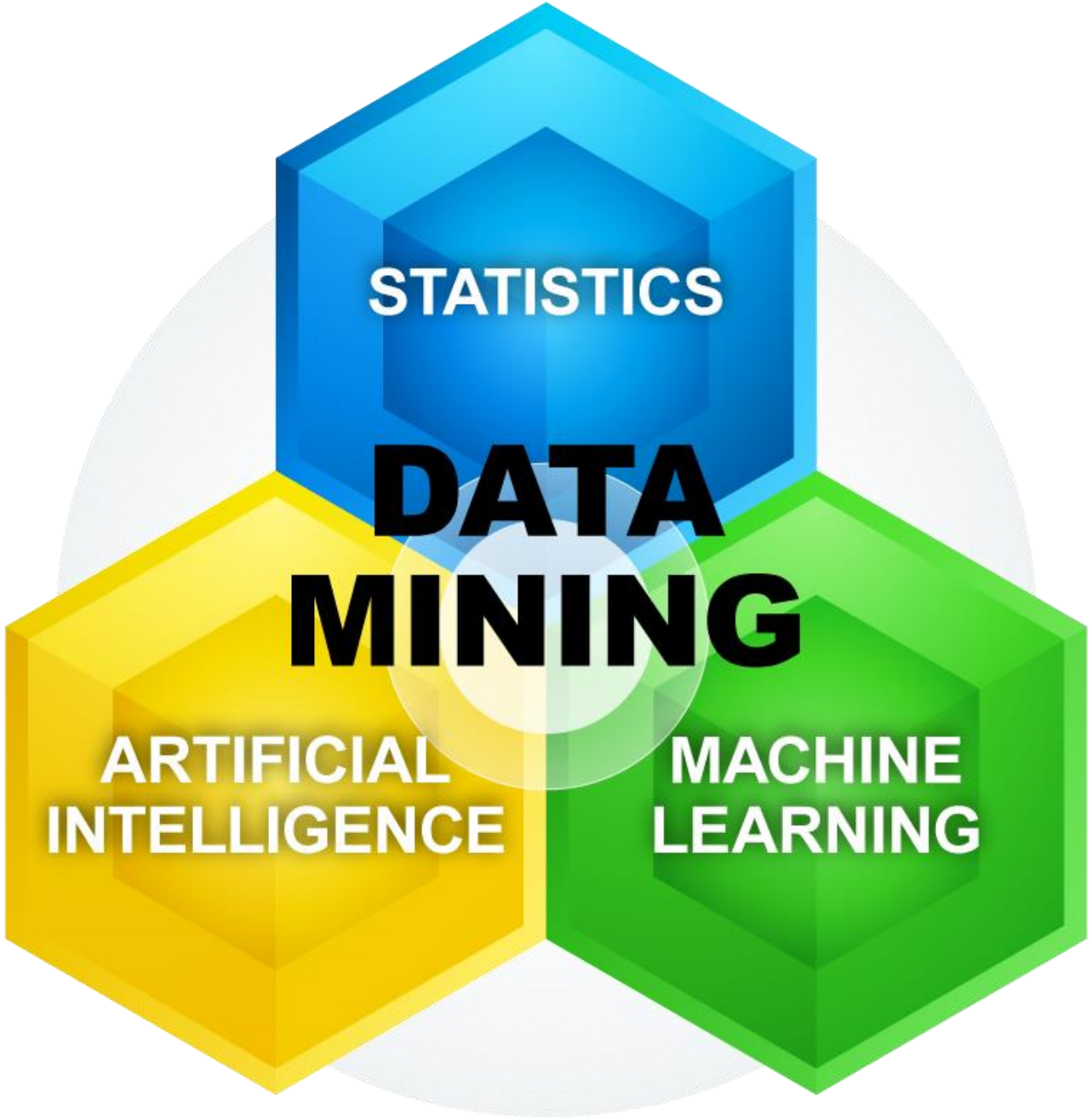
Université Internationale de Casablanca

Data Mining ?

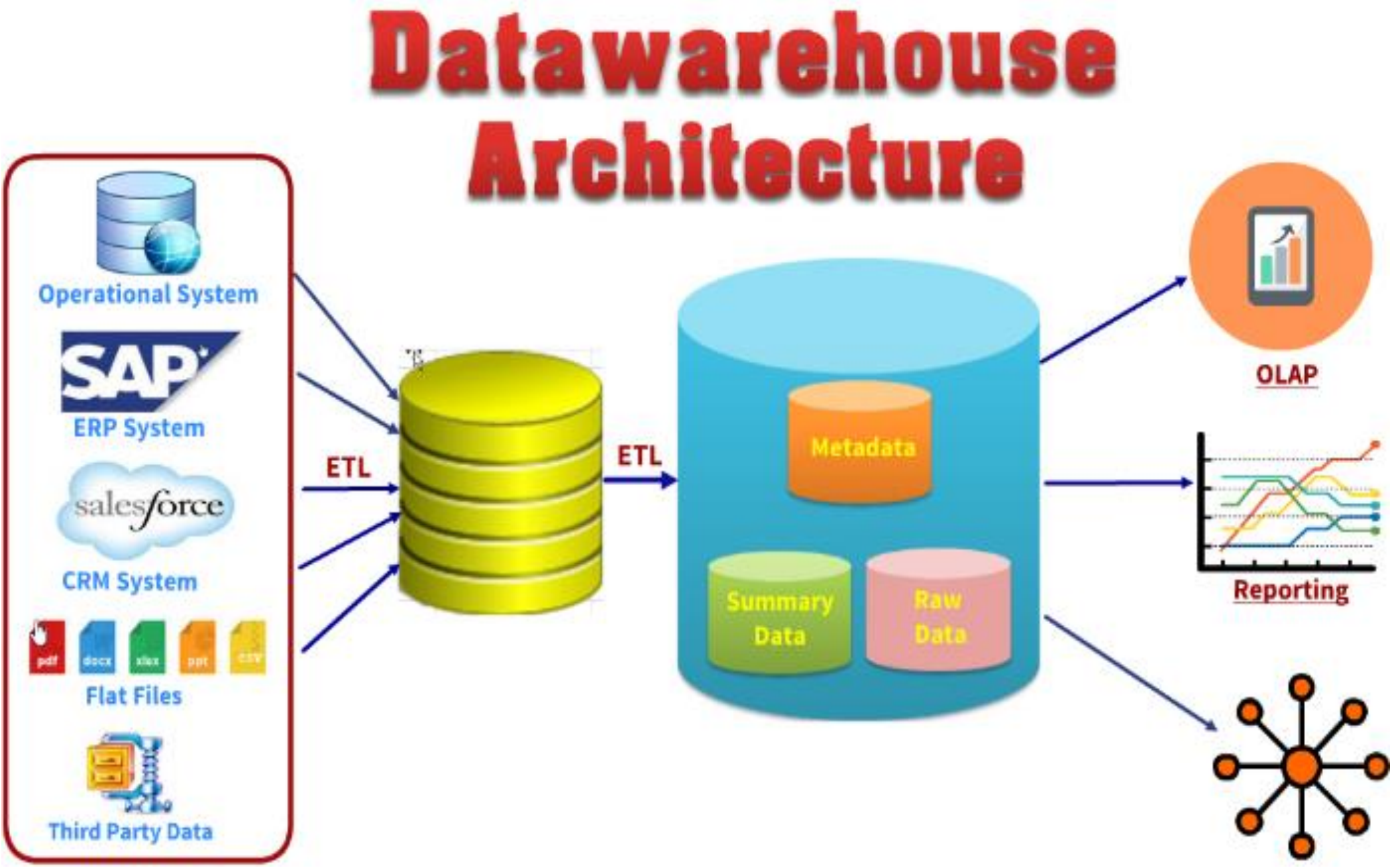
- **Forage de données, explorations de données ou fouilles de données**, ce sont les traductions possibles du data mining.
- En règle générale, le terme Data Mining désigne l'analyse de données depuis différentes perspectives et le fait de **transformer** ces données en informations utiles, en établissant des relations entre les données.



Cas d'application du data mining

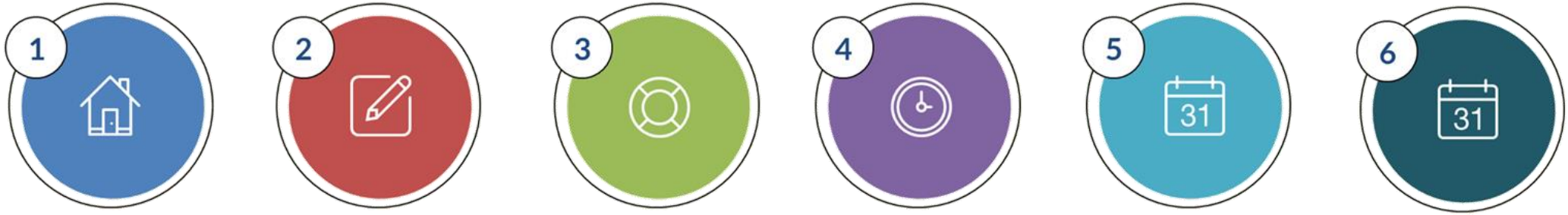


Qu'est ce qu'un Data Warehouse



Université Internationale de Casablanca

Data Mining les différentes phases



Define the Problem

Identify business goals

Identify data mining goals

Identify Required Data

Assess needed data

Collect and understand data

Prepare and Pre-process

Select required data

Cleanse/format data as necessary

Model the Data

Select algorithms

Build predictive models

Train and Test

Train the model with sample data sets

Test and iterate

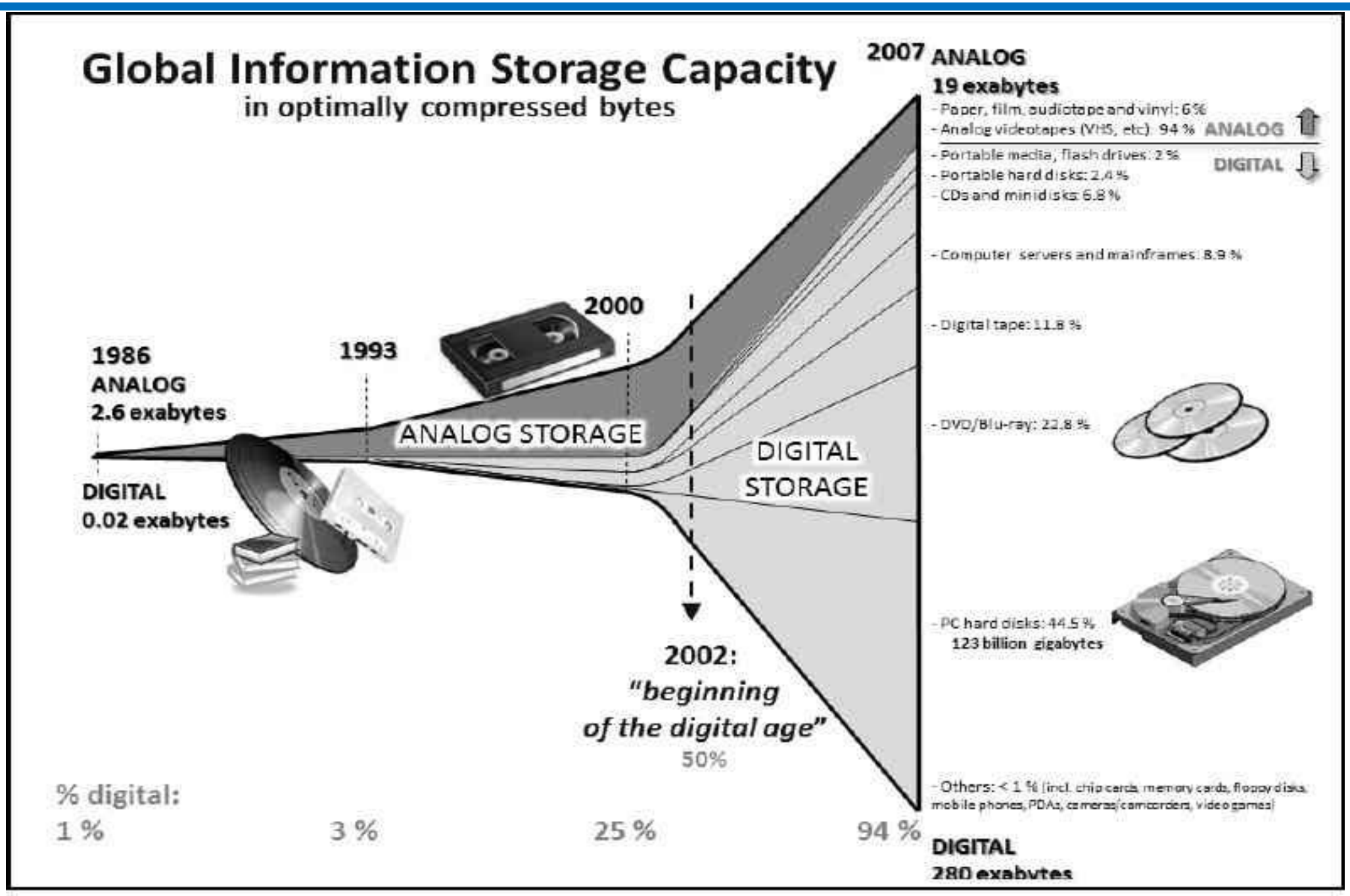
Verify and Deploy

Verify final model

Prepare visualizations and deploy

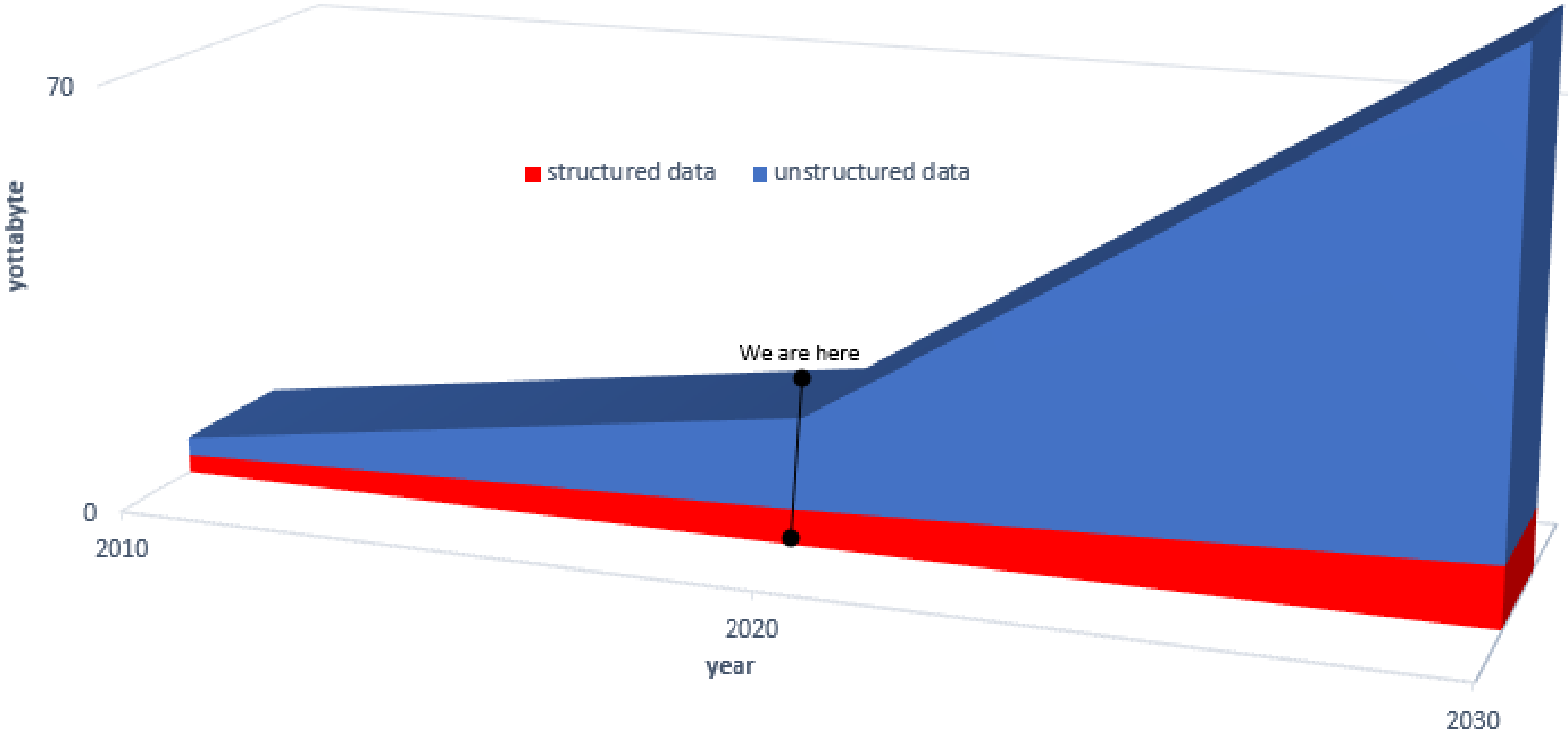
Université Internationale de Casablanca

Big data : plongée dans un océan de données



Université Internationale de Casablanca

Big data : plongée dans un océan de données



Université Internationale de Casablanca

Big data : plongée dans un océan de données

- Le **Big Data** peut se définir avant tout par un volume de plus en plus important de données difficiles à gérer par les bases de données traditionnelles.
- 90% des données existantes aujourd'hui ont été créées au cours des deux dernières années.

90%
des données du monde a été créé au cours des trois dernières années

80%
des données sont non structurées.

20%
des données peuvent être gérées par les bases de données traditionnelles

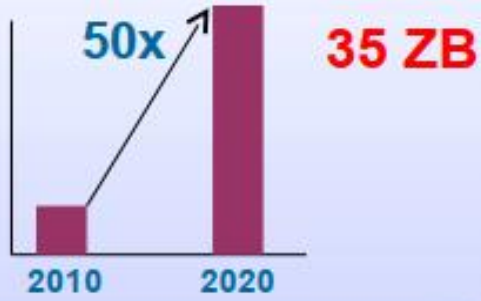
Université Internationale de Casablanca

Caractéristiques du big data

- **4V:** **V**olume **V**élocité **V**ariété **V**éracité

Volume

Capacité à stocker de très volume de données



Vélocité

Temps de réponse ultra-rapide

30 milliards de capteurs RFID et objets connectés

Variété

Analyser diverses variétés de donnés

80% des données mondiale sont non-structurées

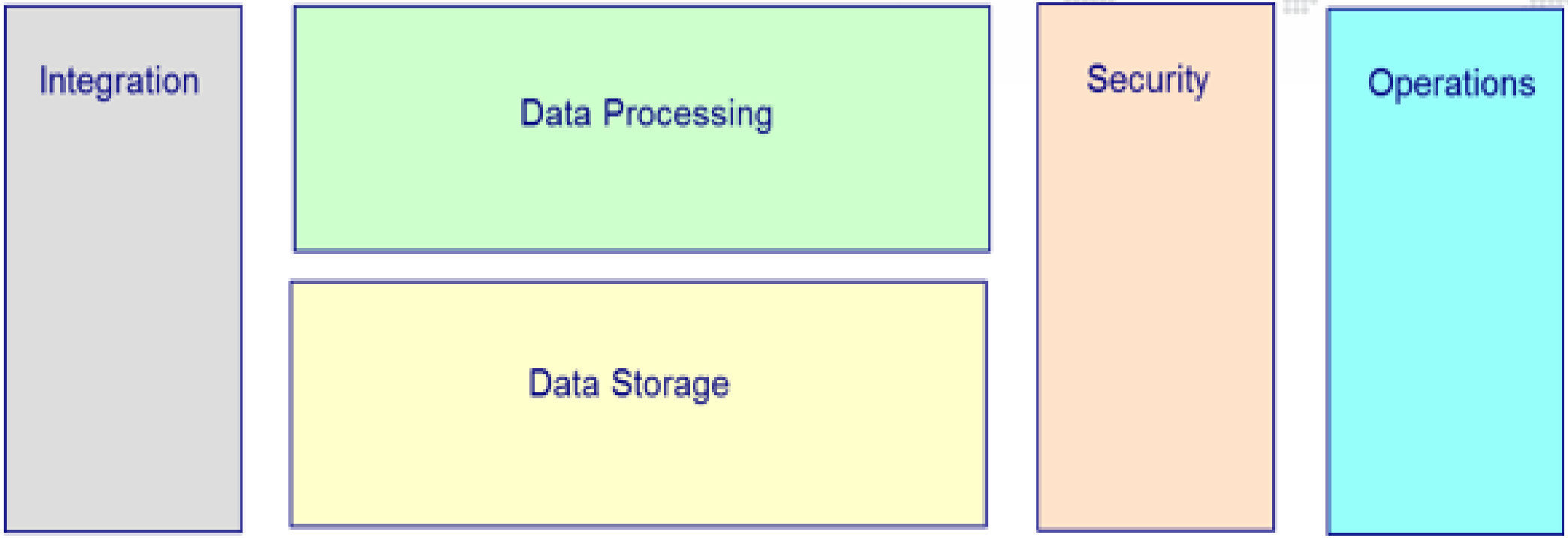
Véracité

Assurer la fiabilité des sources de données.

1 sur 3 business leaders ne fait pas confiance aux informations qu'il utilise pour ces décisions

Université Internationale de Casablanca

Composants principaux d'une architecture Big Data



Université Internationale de Casablanca

Exigences d'une architecture Big Data

❖ **Data Storage**

- Stocker le volume et la variété des données de manière rentable

❖ **Data Processing**

- Répondre à une grande variété d'exigences de traitement

  ❑ Batch
  ❑ Ad hoc querying
  ❑ Real-time stream processing
  ❑ Search
  ❑ Machine learning, discovery

Université Internationale de Casablanca

Exigences d'une architecture Big Data

- **Intégration**
  - Ingérer des données de diverses sources
- **Sécurité**
  - Autorisation d'authentification
  - Gestion des comptes utilisateurs
  - protection des données
- **Operations**
  - Fournir, gérer, surveiller et planifier des ressources

Université Internationale de Casablanca

Améliorations hardwares au fil des ans ...

- **CPU Speeds:**
  - 1990 – 44 MIPS at 40 MHz
  - 2010 – 147,600 MIPS at 3.3 GHz
- **RAM Memory**
  - 1990 – 640K conventional memory (256K extended memory recommended)
  - 2010 – 8-32GB (and more)
- **Disk Capacity**
  - 1990 – 20MB
  - 2010 – 1TB
- **Disk Latency (speed of reads and writes) – not much improvement in last 7-10 years, currently around 70 – 80MB / sec**

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

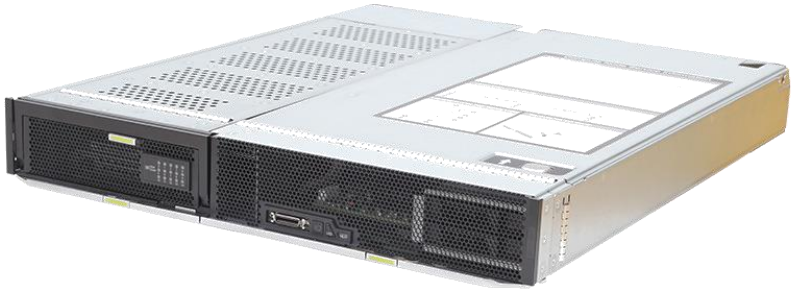
1TB (at 80Mb / sec):

1 disk - 3.4 hours
10 disks - 20 min
100 disks - 2 min
1000 disks - 12 sec

Université Internationale de Casablanca

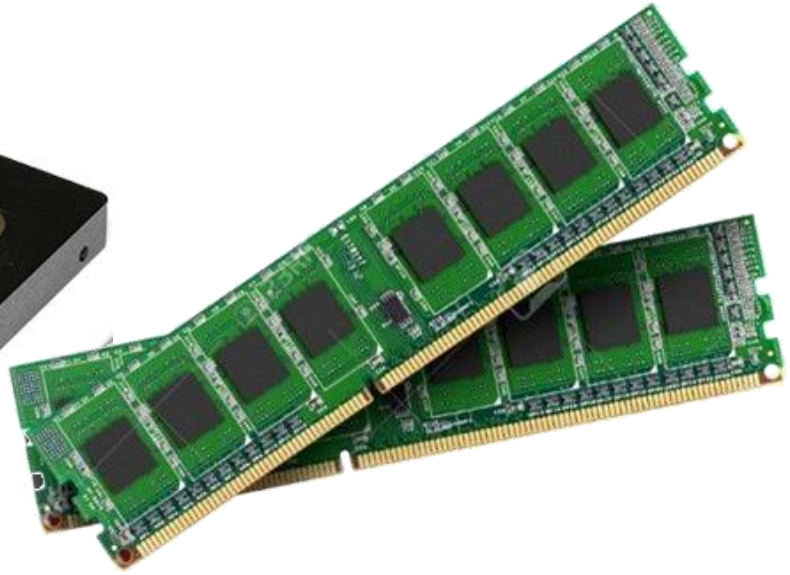
Composant d'un nœud de calcul/stockage

Noeud de calcul / stockage

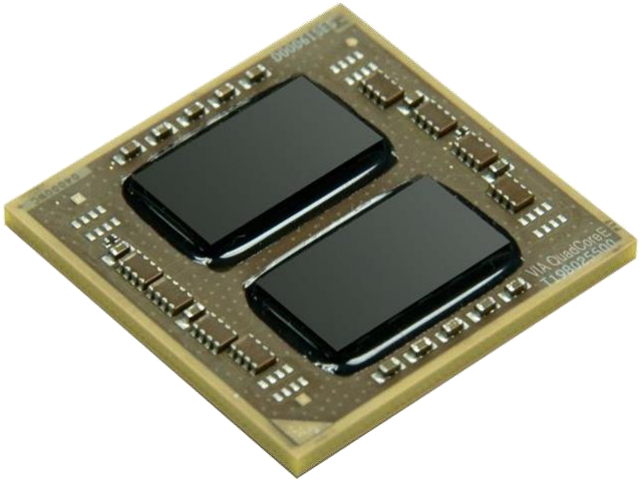


=

Disque dur



Mémoire Ram



Processeur

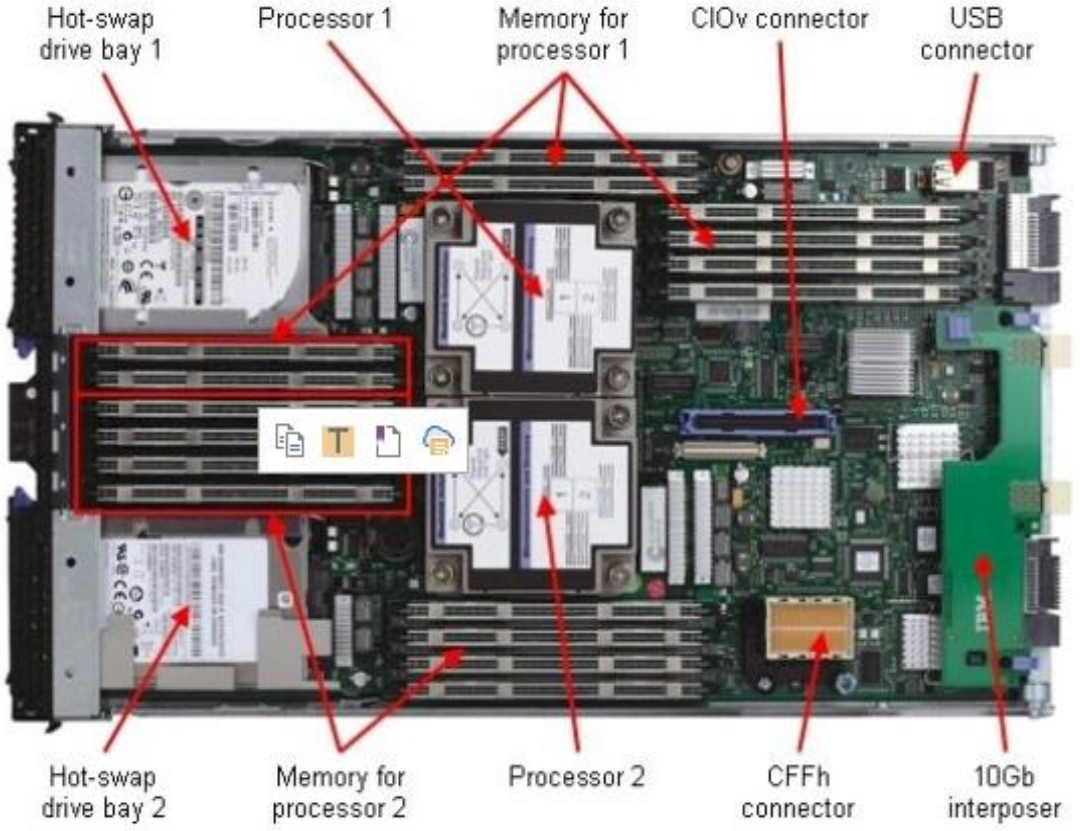
Big data favorise l'usage des serveurs low-cost

Commoditisation du matériel permettant l'analyses des données

- Resource hardware à faible coût
- Les softwares big data sont conçus pour des traitements sur du matériel de base.
- **Prérequis hardware**

Switch Ethernet

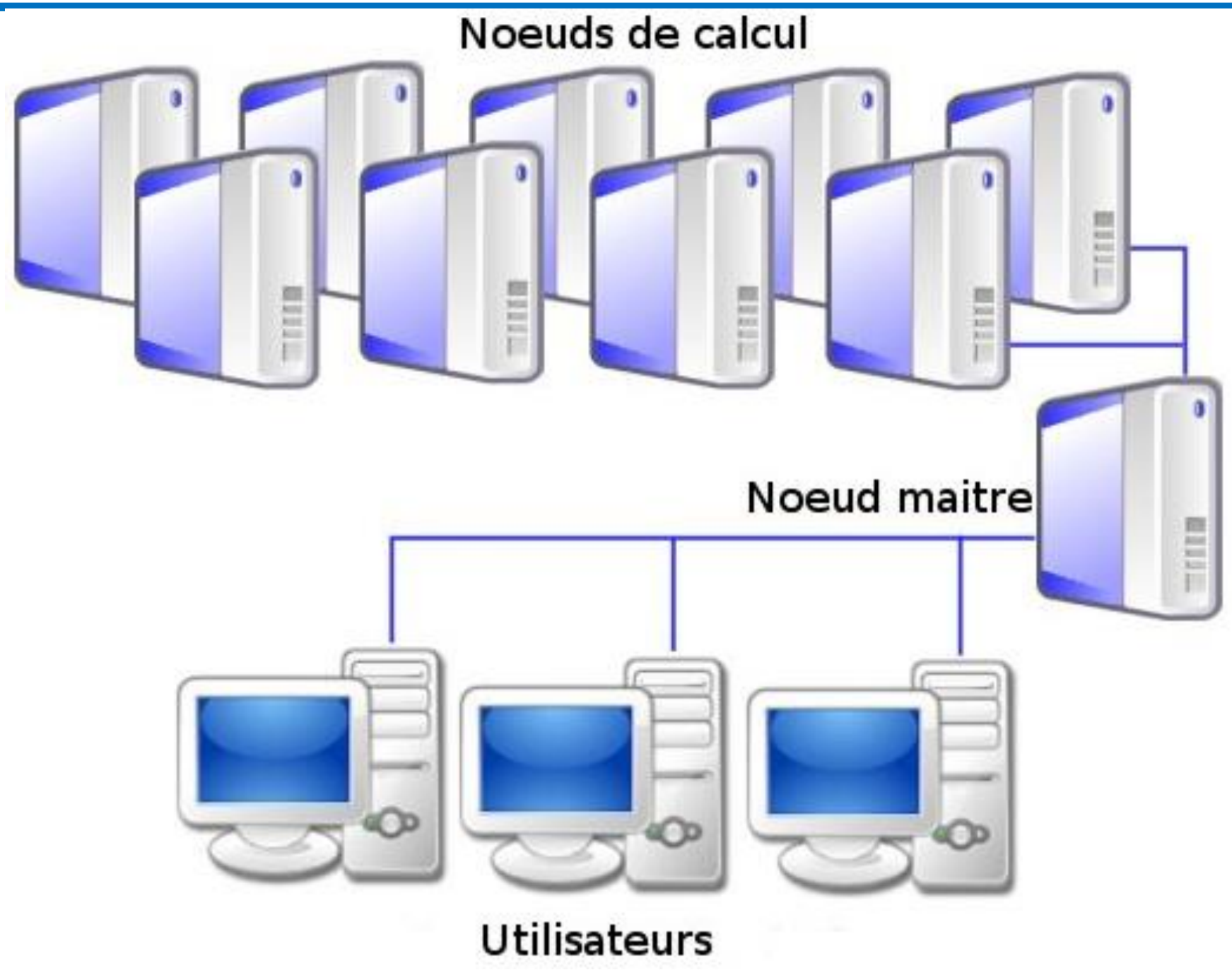
Câble RG 45



Nœud de stockage/ calcul

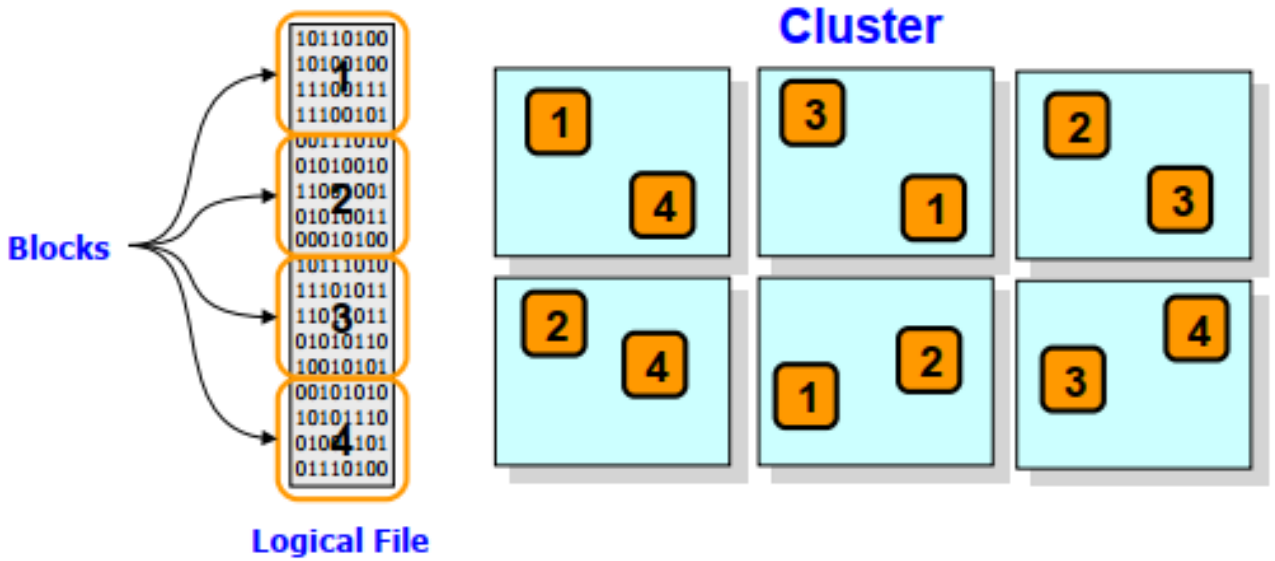
Université Internationale de Casablanca

Architecture distribuée master-slaves

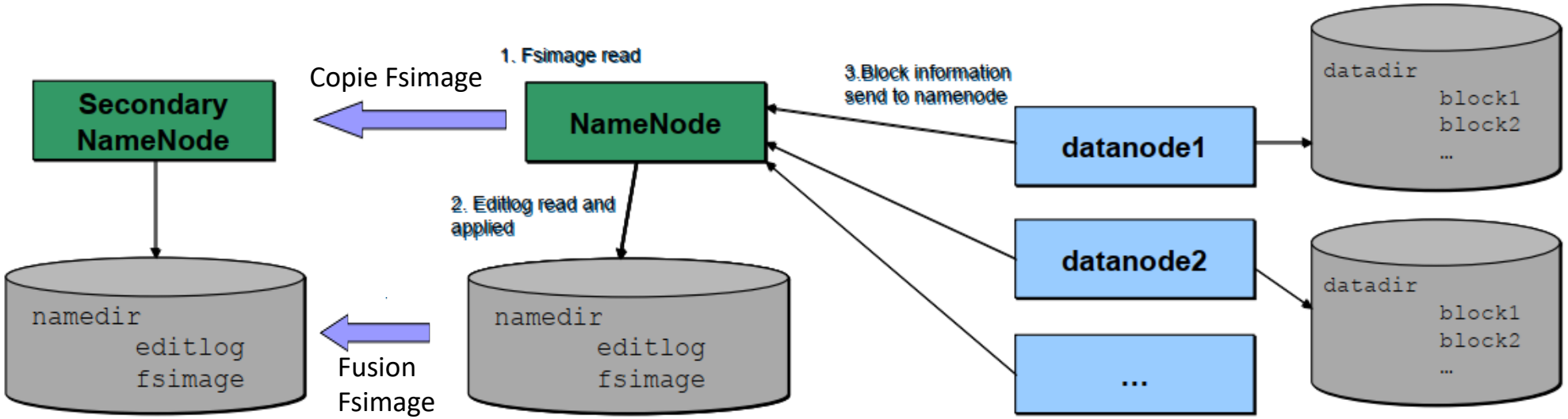


Stockage distribué et architecture distribuée

Les données sont physiquement ou logiquement découpées en blocs de fichier



- **Big Data repose sur une architecture master-slaves**



Université Internationale de Casablanca

Qui utilise l'écosystème Big Data

Université Internationale de Casablanca

Qu'est ce que le NoSQL ?

- Le **NoSQL**, pour « **Not only SQL**", désigne les bases de données qui ne sont pas fondées sur l'architecture classique des bases de données relationnelles.
- Développé à l'origine pour gérer du big data, l'utilisation de base de données NoSQL a explosée depuis quelques années.
- Le NoSQL permet de surpasser les limites des bases de données relationnelles SGBDR.
- Il est préférable d'avoir un langage de haut niveau pour interroger les données plutôt que tout exprimer en Map/Reduce.

- Le **NewSQL** est un terme inventé par Matt Aslett, analyste du groupe 451, pour décrire un nouveau groupe de bases de données partageant la plupart des fonctionnalités des bases de données relationnelles SQL traditionnelles, tout en offrant certains des avantages du NoSQL.

Université Internationale de Casablanca

Différence entre SQL et NoSQL (1)

Type

- Les bases de données SQL sont principalement appelées bases de données centralisées ou relationnelles (SGBDR);
- Les bases de données NoSQL sont principalement appelées bases de données non relationnelles ou distribuées (décentralisée).

Schéma des données

- SQL requiert que vous utilisiez des schémas prédéfinis pour déterminer la structure de vos données avant de les utiliser. De plus, toutes vos données doivent suivre la même structure.
- Une base de données NoSQL possède un schéma dynamique pour les données non structurées (orientées document, colonne, graphe ou organisées en pair de clé-valeur.
- Cette flexibilité signifie que les données peuvent être traitées sans avoir au préalable une structure définie.

Université Internationale de Casablanca

Différence entre SQL et NoSQL (2)

Evolutivité

- Les bases de données SQL sont évolutives verticalement.
- Par contre, les bases de données NoSQL sont évolutives horizontalement.

Structure

- Les bases de données SQL sont basées sur des tables.
- Les bases de données NoSQL sont des paires clé-valeur, des documents, graphes, ou des colonnes.

Différence entre SQL et NoSQL (3)

Propriété

- Les bases de données SQL suivent les propriétés ACID (**A**tomicity, **C**onsistency, **I**solation et **D**urability), tandis que la base de données NoSQL suit les propriétés CAP (**C**ohérence, **D**isponibilité **et P**artition).

Support

- PostgreSQL, MySQL, Oracle et Microsoft SQL Server sont des exemples de bases de données SQL.
- Les exemples de base de données NoSQL incluent Hive, RavenDB Cassandra, MongoDB, BigTable, HBase, Neo4j et CouchDB.

Points forts de SQL vs. NoSQL (3)

SQL	NoSQL
Système de base de données relationnelles et centralisé.	Système de base de données non relationnel ou distribué.
Ces bases de données ont un schéma fixe ou statique ou prédéfini.	Ces bases de données ont un schéma ont un schéma dynamique.
Ces bases de données ne conviennent pas au stockage de données hiérarchique.	Ces bases de données conviennent le mieux au stockage de données non structurées.
Ces bases de données conviennent mieux aux requêtes complexes.	Ces bases de données ne conviennent pas aux requêtes complexes.
Evolutif verticalement	Evolutif horizontalement

Université Internationale de Casablanca

Ressources

- Documentation officielle :
  - https://www.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/c0057605.html
  - https://insidebigdata.com/category/whitepapers/
  - https://spark.apache.org/
  - https://hadoop.apache.org/
  - https://hive.apache.org/

- Livre :
  - "*Les bases de données NoSQL et le Big Data: Comprendre et mettre en oeuvre*" par Rudi Bruchez.
  - *"Big Data white paper" par* Arzu Barske