Université Sultan Moulay Slimane

Département: ITG

Ecole Supérieure de Technologie de Béni Mellal

Module: Big Data & Bases de Données NoSQL (M236)

Filières: DUT IDIA & SIR - S3

Année universitaire: 2025-2026

Support de cours

Big Data & Bases de Données NoSQL

Chapitre I: Introduction au Big Data

Pr. Siham BAKKOURI

Email: s.bakkouri@usms.ma

Plan

- INTRODUCTION
- DÉFINITIONS ET CONTEXTE HISTORIQUE
- LES ENJEUX DU BIG DATA
- LES CARACTÉRISTIQUES : LES 5V DU BIG DATA
- DES APPLICATIONS CONCRÈTES DU BIG DATA
- LES LIMITES DES SYSTÈMES CLASSIQUES

Introduction

Le **Big Data** désigne l'ensemble des **technologies**, **méthodes** et **outils** permettant de collecter, stocker, traiter et analyser des volumes massifs de données.

→ Ces données proviennent de sources très diverses et arrivent souvent à grande vitesse, ce qui dépasse largement les capacités des systèmes traditionnels.

L'objectif principal du Big Data n'est pas seulement de stocker des données, mais de tirer de la valeur afin d'améliorer la prise de décision, la recherche scientifique ou l'innovation dans les entreprises et la société.

Introduction

Exemples

→ **Amazon** analyse chaque clic pour proposer des recommandations personnalisées.

→ **Netflix** adapte ses suggestions de films en temps réel grâce à l'analyse des habitudes de visionnage.

→ **Google Maps** utilise des flux de données massifs pour proposer les itinéraires les plus rapides.

Définitions et contexte historique

→ Définition du Big Data

Le **Big Data** concerne des **ensembles de données trop volumineux, trop rapides ou trop variés** pour être traités par les systèmes classiques.

❑ Les types de données dans le Big Data

- **Structurées :** Données bien organisées, généralement dans des tables relationnelles (ex. : informations clients, transactions bancaires).
- **Semi-structurées :** Données ayant une certaine organisation mais non totalement rigide (ex. : JSON, XML, fichiers logs).
- **Non structurées :** Données brutes et hétérogènes, sans modèle fixe (ex. : texte libre, vidéos, images, sons, capteurs IoT).

Définitions et contexte historique

→ Différences avec les systèmes classiques

❑ Bases relationnelles (SQL)

- Très efficaces pour les données structurées de taille moyenne (Go à quelques To).
- Peu adaptées aux volumes massifs (plusieurs Po) ou aux flux en temps réel.
- Pas conçues pour gérer directement de grandes quantités de données semi-structurées ou non structurées.

❑ Big Data (Hadoop, Spark, NoSQL, etc.)

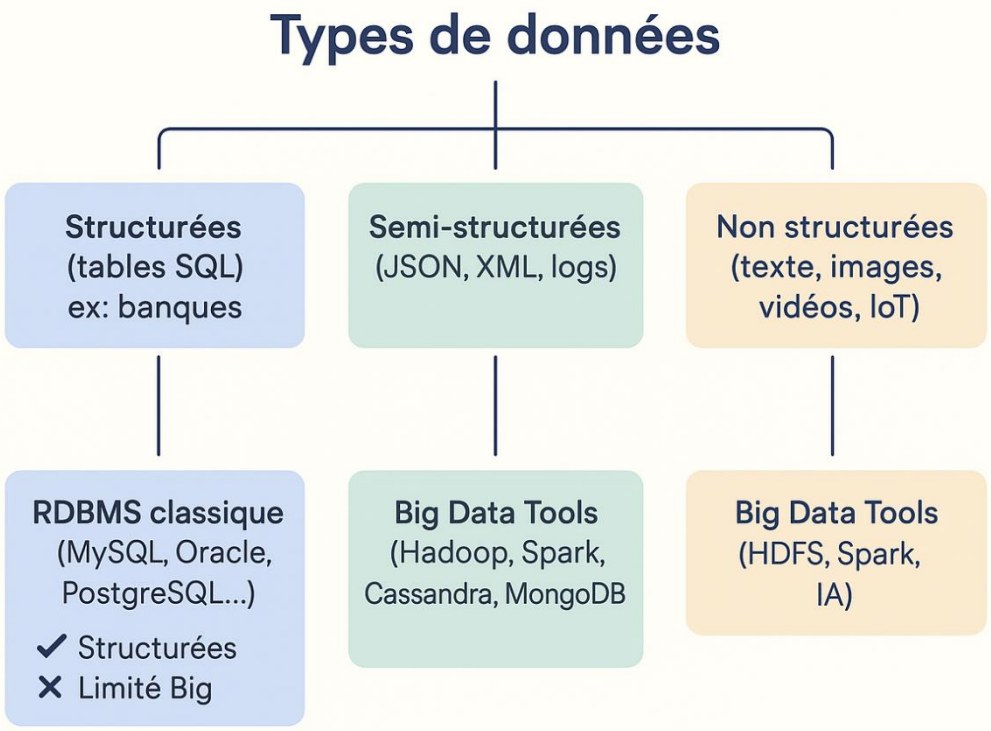
- Capable de gérer tous les types de données (structurées, semi-structurées et non structurées).
- S'appuie sur des architectures distribuées (données réparties sur plusieurs machines) et des traitements parallèles.
- Permet de traiter des volumes massifs et des flux continus en temps quasi réel.

Définitions et contexte historique

→ Différences avec les systèmes classiques

Le Big Data ne s'oppose donc pas aux bases relationnelles, mais il élargit le champ :

- Il inclut aussi les données structurées (mais à très grande échelle).
- Et il apporte surtout des solutions pour les données semi/non structurées, ainsi que pour la vitesse et la variété des flux.



Les limites des systèmes classiques

Les systèmes classiques de gestion des données, principalement les bases relationnelles et les entrepôts traditionnels, montrent rapidement leurs limites face aux besoins modernes. Ils peinent à gérer les volumes massifs de données, à intégrer des sources hétérogènes, à fournir des traitements en temps réel et à s'adapter à l'évolution rapide des besoins en scalabilité et en flexibilité.

- **Volume limité** : difficulté à stocker et traiter de très grandes quantités de données.
- **Variété réduite** : incapacité à gérer efficacement les données non structurées (textes, images, vidéos).
- **Faible vélocité** : manque de performance pour les traitements en temps réel.
- **Scalabilité limitée** : adaptation difficile à la croissance exponentielle des données.
- **Rigidité** : peu de flexibilité pour répondre aux nouveaux besoins.

Définitions et contexte historique

→ Historique

- **Années 1960–1970 : Premiers pas**
  - Développement des bases de données relationnelles (RDBMS) par Edgar F. Codd (1970).
  - L'ère des mainframes et du traitement par lots.
  - Données surtout structurées, stockées dans des bases centralisées.
- **Années 1980–1990 : Explosion des données**
  - Multiplication des systèmes de gestion de bases de données (Oracle, SQL Server, MySQL, PostgreSQL).
  - Début de l'Internet → génération massive de données textuelles et log files.
  - Les RDBMS deviennent la norme mais montrent des limites avec les très grands volumes.

Définitions et contexte historique

→ Historique

- **Début des années 2000 : Montée du Big Data**
  - Doug Laney (2001) formalise le concept des 3V du Big Data
  - Google publie des articles fondateurs sur MapReduce et Google File System (GFS).
  - Naissance de l'écosystème Hadoop (2005), inspiré de ces travaux.
- **Années 2010 : Industrialisation**
  - Émergence des NoSQL databases adaptées aux données non structurées.
  - Hadoop devient un standard pour le stockage distribué (**HDFS**) et le traitement batch.
  - Spark (2014) apporte un traitement in-memory, plus rapide que MapReduce.
  - Développement du Cloud computing (AWS, Azure, GCP), rendant le Big Data plus accessible.
  - Concept élargi aux 5V.

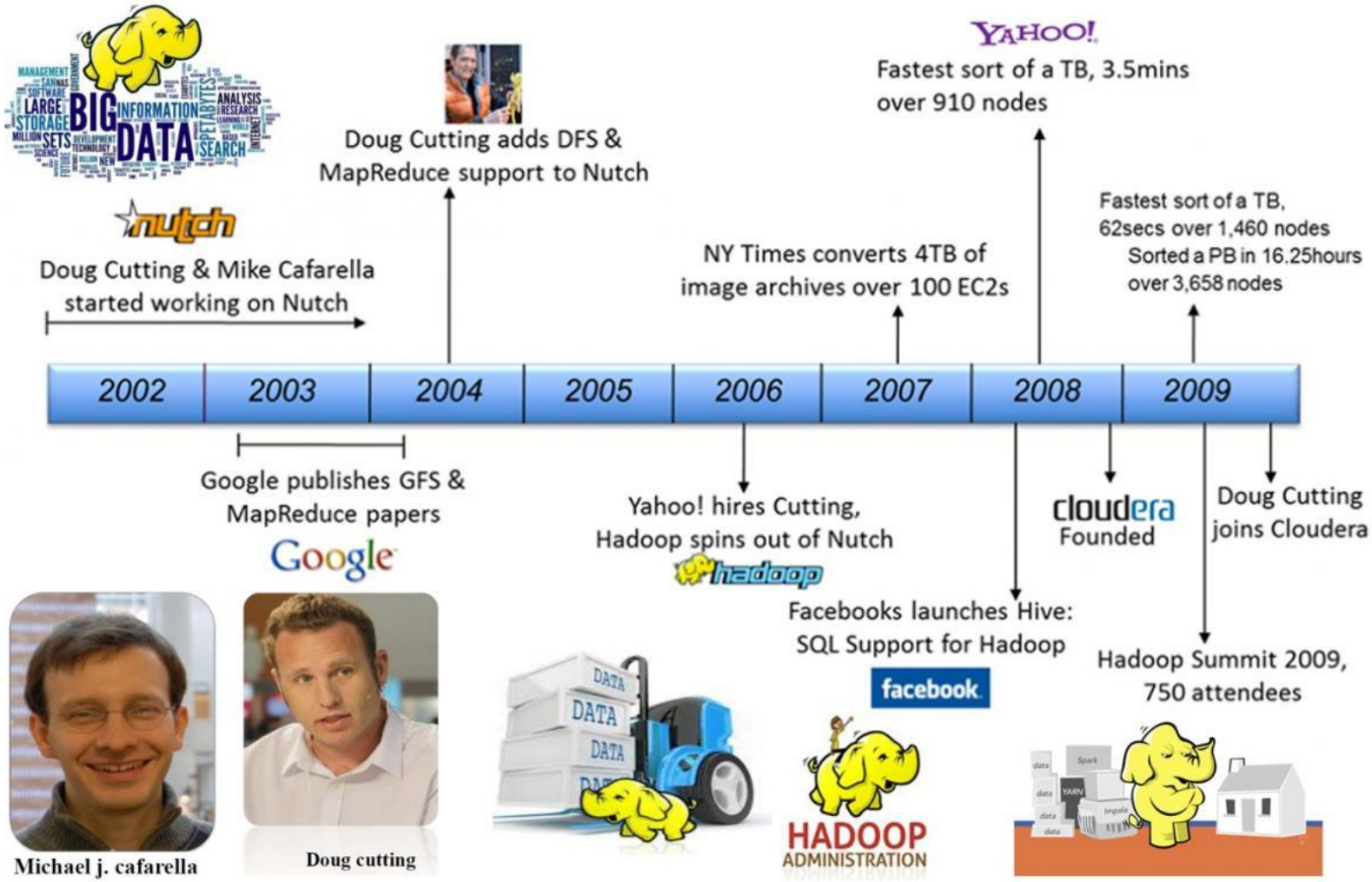
→ Historique

❑ Années 2020 – Aujourd'hui

- Data Lakes (Amazon S3, Azure Data Lake, Hadoop ecosystem) remplacent souvent les Data Warehouses pour le Big Data.
- Intelligence Artificielle et Machine Learning appliqués sur des volumes massifs (Deep Learning, NLP, Vision par ordinateur).
- Streaming temps réel (Kafka, Flink, Spark Streaming) pour traiter les données IoT, réseaux sociaux, transactions instantanées.
- Edge computing : traitement des données au plus près des capteurs pour réduire la latence.
- Cloud natif & Serverless Big Data : BigQuery (Google), Snowflake, Databricks.
- Apparition du terme Data Mesh et Data Fabric pour décrire de nouvelles architectures plus distribuées et orientées business.

Définitions et contexte historique

→ Historique



Les enjeux du Big Data

❑ Enjeux économiques

- Compréhension fine du client pour personnaliser l'offre.
- Optimisation de la chaîne logistique et réduction des coûts.
- Détection de fraude (banques, assurances).

Exemple **:** Amazon et Netflix analysent les clics et achats pour proposer des recommandations adaptées.

❑ Enjeux scientifiques

- Analyse du génome humain pour la médecine personnalisée.
- Simulation climatique et prévision météorologique.
- Recherche astrophysique ou physique des particules (CERN).

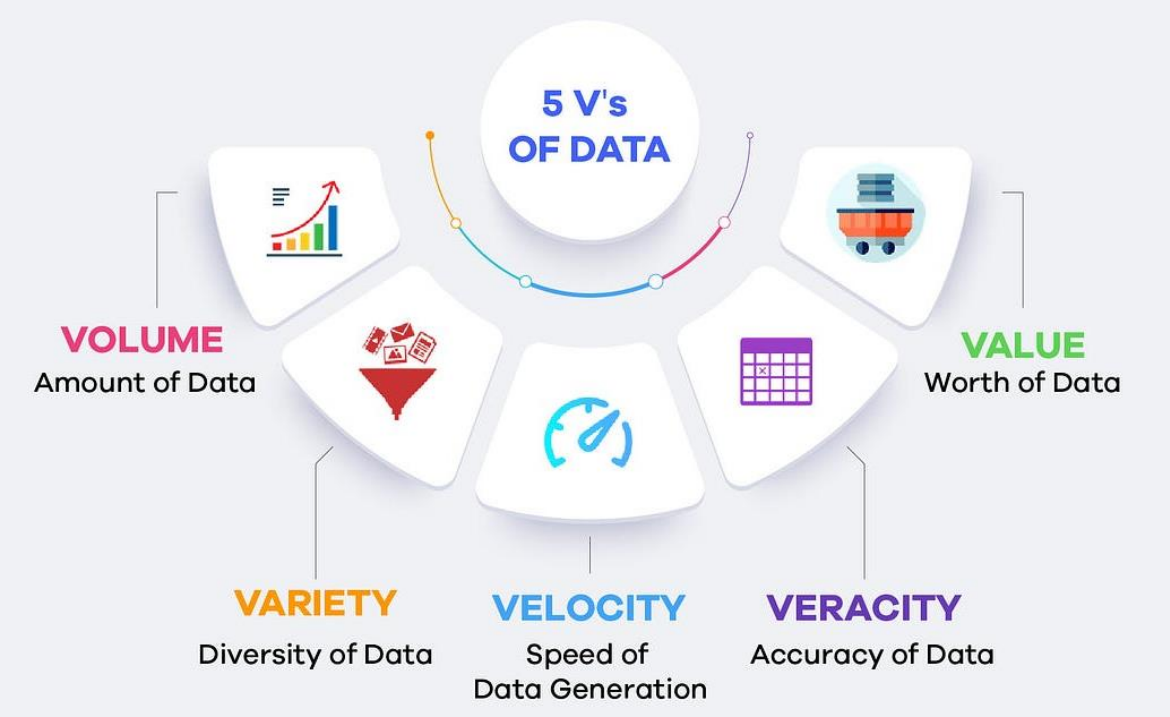
Les enjeux du Big Data

❑ Enjeux sociétaux

- ***Villes intelligentes*** : gestion du trafic et consommation énergétique optimisée.
- ***Santé publique*** : suivi épidémique et gestion de crises sanitaires.
- ***Sécurité*** : vidéosurveillance intelligente et détection d'anomalies.

Les caractéristiques : les 5V du Big Data

Le concept de Big Data se distingue des données traditionnelles grâce à certaines caractéristiques fondamentales.
À l'origine, ***Doug Laney*** avait proposé le modèle des **3V**, qui s'est ensuite élargi pour devenir les **5V**, afin de mieux représenter les défis et opportunités liés à la gestion et à l'analyse de grandes masses de données.



Les caractéristiques : les 5V du Big Data

❑ Les 5V du Big Data

1. **Volume:** Représente la quantité massive de données générées chaque jour (milliards de messages, vidéos, images, transactions…).

   → ***Défi*** : stockage et gestion.

2. **Vélocité :** Correspond à la vitesse à laquelle les données sont produites et doivent être traitées en temps réel (ex. transactions bancaires, capteurs IoT).

   → ***Défi*** : traitement rapide et sans latence.

3. **Variété :** Les données ne sont pas uniquement structurées (tableaux), mais aussi non structurées (textes, images, vidéos) ou semi-structurées.

   → ***Défi*** : intégrer et exploiter des formats hétérogènes.

Les caractéristiques : les 5V du Big Data

❑ Les 5V du Big Data

4. **Véracité :** Concerne la qualité et la fiabilité des données, qui peuvent être incomplètes, contradictoires ou bruitées.

   → ***Défi*** : assurer la confiance et la justesse des données.

5. **Valeur:** L'objectif final est d'extraire une valeur ajoutée des données : prise de décision, amélioration des services, création d'opportunités.

   → ***Défi*** : transformer les données en insights utiles.

❑ Les V supplémentaires

Avec le temps, de nouvelles caractéristiques se sont ajoutées :

- **Variabilité** : les données changent rapidement et peuvent être imprévisibles.
- **Visualisation** : nécessité de représenter les données de façon claire et compréhensible.
- **Validité** : s'assurer que les données sont pertinentes et adaptées au contexte.
- **Vulnérabilité** : protection des données et sécurité de l'information.

Des applications concrètes du Big Data

- **Santé** : diagnostic médical assisté, suivi des patients, recherche génétique.
- **Transport** : optimisation du trafic, maintenance prédictive, véhicules autonomes.
- **E-commerce** : recommandations personnalisées, analyse comportementale, détection de fraude.
- **Réseaux sociaux** : analyse de sentiments, ciblage publicitaire, détection de fake news.
- **Recherche scientifique** : CERN, observation spatiale, modélisation climatique.
- **Industrie** : maintenance prédictive et optimisation de la production via IoT.