IMT Mines Alès
École Mines-Télécom

04 FÉVRIER 2026

IMT MINES ALÈS

3e année Développement Logiciel

Big Data et modélisation

Introduction au Big Data

FORMATEUR : DIALLO ALIMOU

Objectifs pédagogiques

- Concevoir une architecture Big Data
- Comprendre Spark, Hadoop, Data Lake
- Manipuler Spark en conditions quasi réelles
- Comprendre les problèmes de scalabilité
- Implémenter un pipeline de traitement distribué

Concepts, Enjeux & Architecture

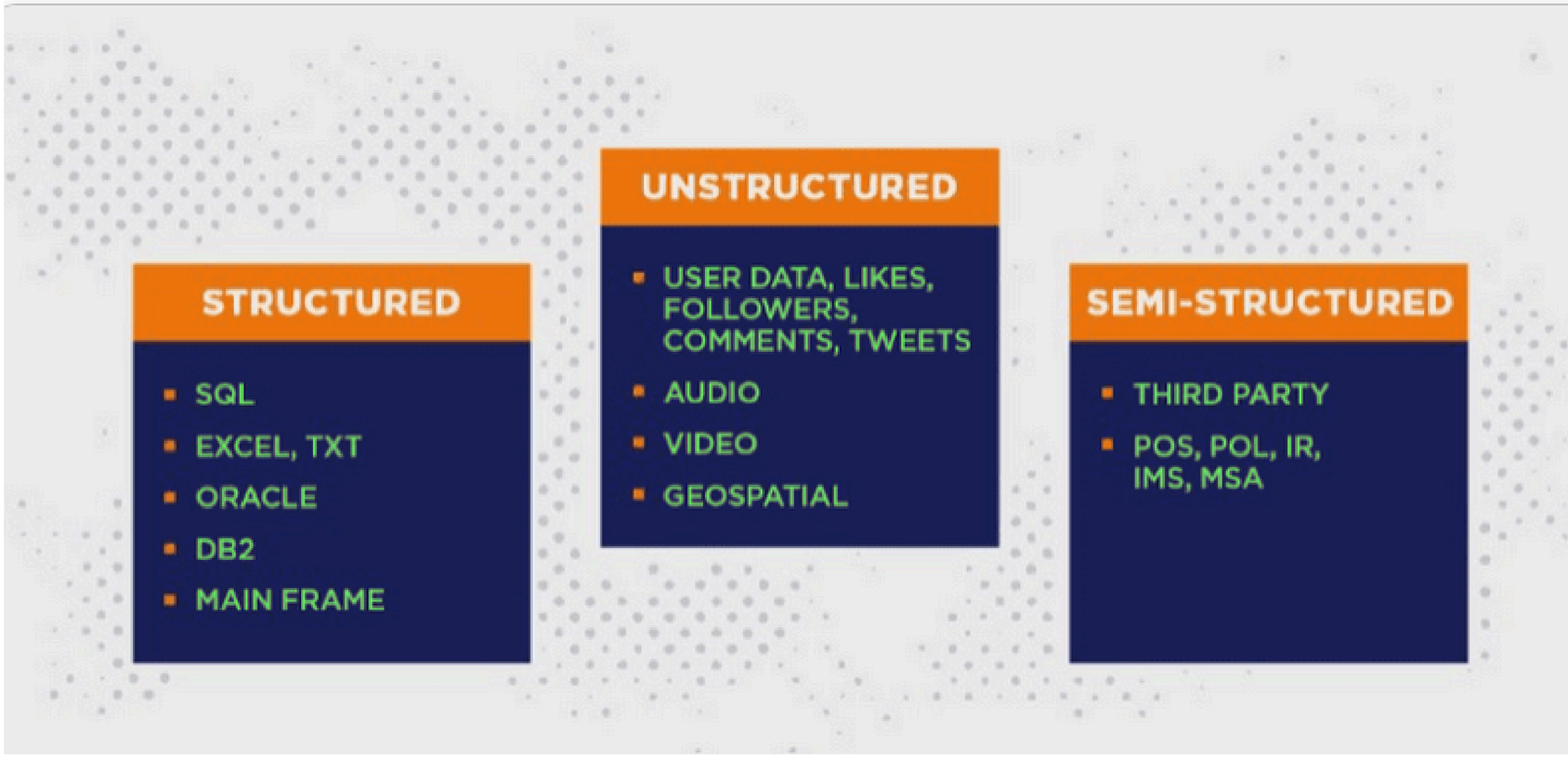
Dans la société d’aujourd’hui axée sur les données, les données imprègnent chaque partie de notre vie. Le volume de données sur Terre double tous les deux ans, mettant l'accent sur sa pertinence. « Big Data » fait référence à d’énormes quantités de données organisées et non structurées que les solutions standard de gestion des données ont du mal à gérer en raison de sa quantité et de sa complexité.

Qu’est-ce que le Big Data

Big Data est une **collection de données** qui est énorme en volume, mais en croissance exponentielle avec le temps. Ce sont des données avec une taille et une complexité si grandes qu'aucun des outils de gestion de données traditionnels ne peut les stocker ou les traiter efficacement. **Le big data** est aussi des données mais avec une taille énorme. Avec la croissance des technologies et des services, ces données importantes sont produites qui peuvent être **structuré, semi-structuré et non structuré de différentes sources.**

Types De Big Data

Les données structurées sont organisées selon **un schéma fixe** (ex : tables, bases de données). **Les données semi-structurées** possèdent une structure partielle **sans schéma rigide** (ex : XML, JSON). **Les données non structurées** n’ont pas de **structure prédéfinie** (ex : textes, images, vidéos).



Les 5V du Big Data

Volume : quantité massive de données à stocker et traiter (ex : plusieurs téraoctets de logs ou de données de capteurs).
Vélocité : vitesse à laquelle les données sont produites et doivent être traitées (ex : flux temps réel de transactions ou de capteurs IoT).
Variété : diversité des formats de données (ex : tables SQL, fichiers JSON, images, vidéos).
Véracité : fiabilité et qualité des données (ex : données bruitées, incomplètes ou erronées issues de capteurs).
Valeur : capacité à extraire de l'information utile à partir des données (ex : prédire la demande ou détecter une fraude).

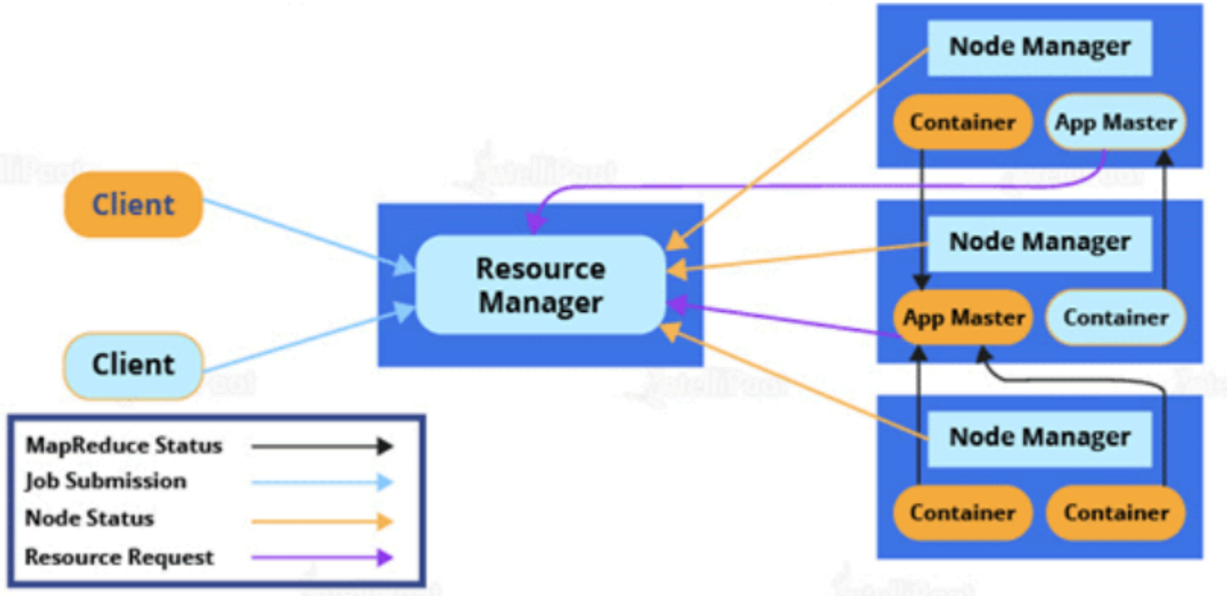
Projets Big Data

Un **projet Big Data** ne consiste pas seulement à analyser des données, mais à concevoir un **système complet capable d’ingérer, stocker, traiter et exploiter** des **données massives** efficacement.

1. Projet Hadoop / YARN

Objectif : gérer et optimiser l'exécution de traitements Big Data sur un cluster Hadoop.

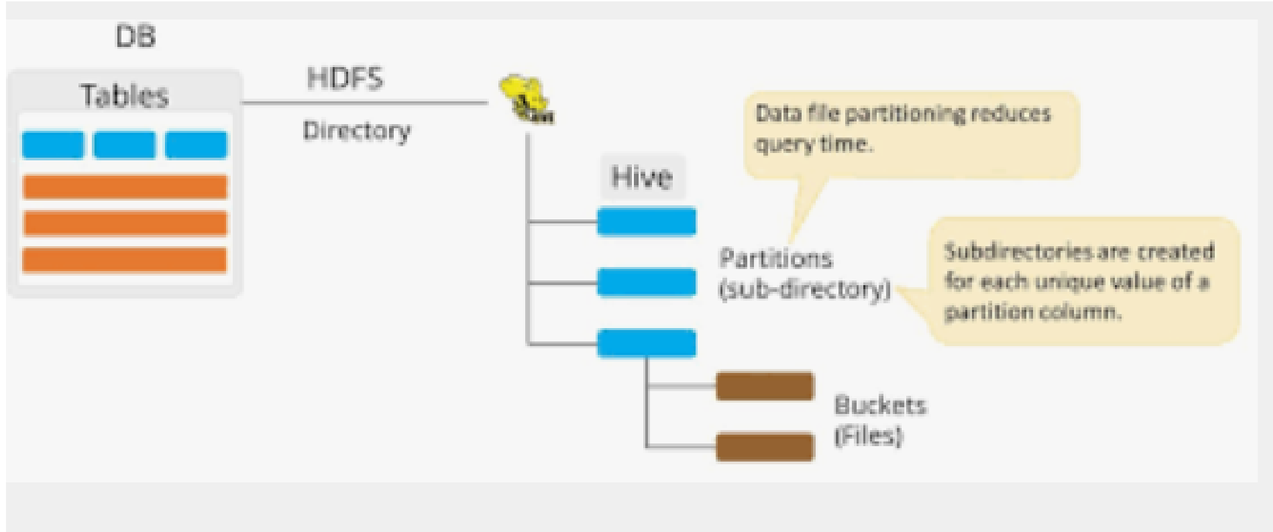
- Utilisation de **YARN** comme gestionnaire de ressources du cluster
- Ingestion de données massives (**via Sqoop vers HDFS**)
- Mise en place d'un pipeline de traitement de bout en bout (ingestion - stockage - traitement - résultats)
- Cas d'usage : traitement de logs, données clients, données IoT



2. Projet de partitionnement de tables Hive

Objectif : améliorer les performances de requêtes sur de très gros volumes de données.

- Stockage des données dans **Hive** sur **HDFS**
- Mise en place du partitionnement des tables pour accélérer les requêtes
- Deux types principaux :
  - Partitionnement statique (manuel)
  - Partitionnement dynamique
- **Bénéfices** : moins de données lues, requêtes plus rapides, meilleure scalabilité



Comment les technologies Big Data peuvent-elles améliorer l’efficacité opérationnelle ?

Le **Big Data** améliore les **opérations commerciales** en rationalisant les processus, des matières premières aux produits finaux. Il permet **une prise de décision** plus rapide, **améliore la qualité du service** et adapte les produits aux besoins des clients. En créant des zones de mise en scène pour organiser de nouvelles données et s’intégrer aux entrepôts de données, les entreprises priorisent les informations critiques, réduisent les coûts et optimisent les performances. Cette approche stratégique garantit l'efficacité et la prestation de services axés sur le client.

Architecture Big Data complète (vue système)

Pourquoi mettre en place une architecture Big Data ?

Les systèmes de données traditionnels ne permettent plus de gérer efficacement des volumes massifs, des vitesses élevées et une grande variété de données. Une architecture Big Data est donc nécessaire pour adapter l'écosystème informatique et exploiter pleinement ces données.

Elle permet notamment de :

- Traiter les données **en batch** et en **temps réel**
- Centraliser des données issues de sources et formats variés
- **Stocker**, **explorer** et **transformer** de grands volumes de données
- Réaliser des **analyses avancées**, **prédictives** et basées sur l'IA

Architecture Big Data complète (vue système)

Les deux principales architectures Big Data

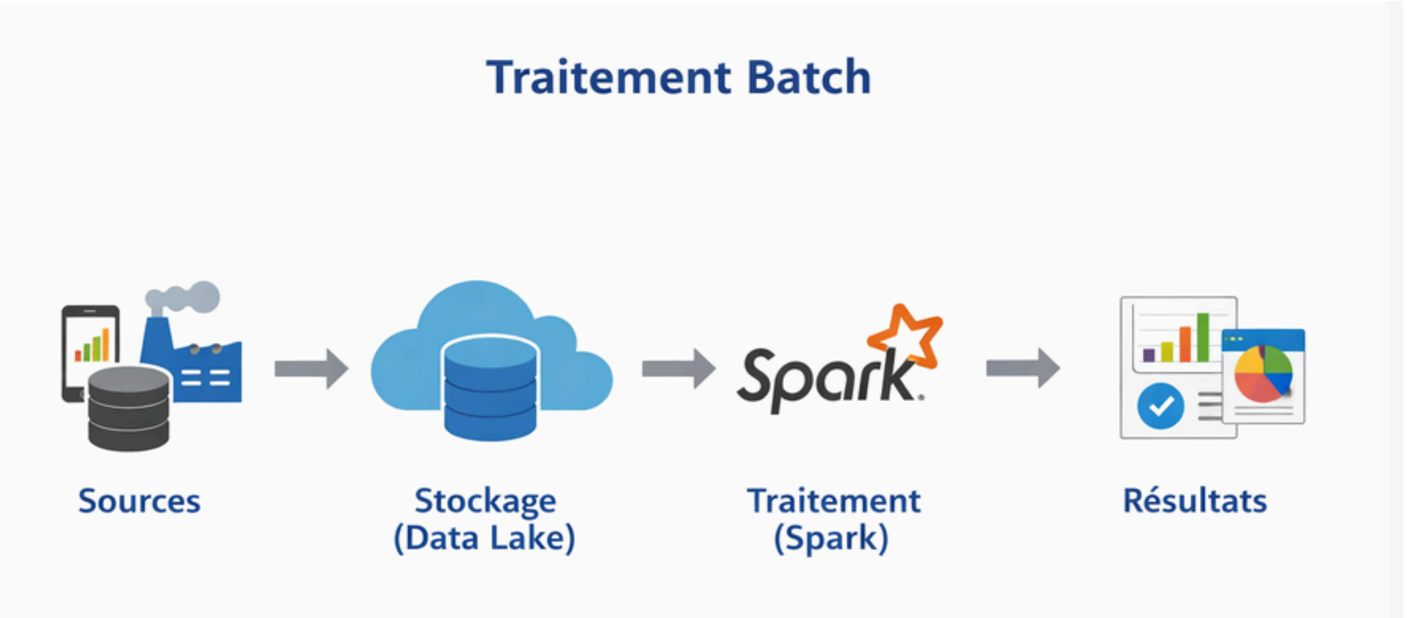
- **Architecture Lambda** : combine deux traitements séparés, un traitement batch (données historiques) et un traitement temps réel (streaming), afin de gérer simultanément données passées et données en flux.
- **Architecture Kappa** : repose uniquement sur le traitement streaming en fusionnant batch et temps réel dans un seul pipeline ; elle ne se concentre pas sur le stockage mais sur le traitement continu des données.

Architecture Big Data complète (vue système)

Traitement Batch

Objectif : traiter de grands volumes de données historiques en différé.

Principe



- **Les données sont :**
  - collectées
  - stockées
  - puis traitées par lots
- **Les traitements sont :**
  - périodiques (toutes les heures, tous les jours, etc.)
  - non temps réel

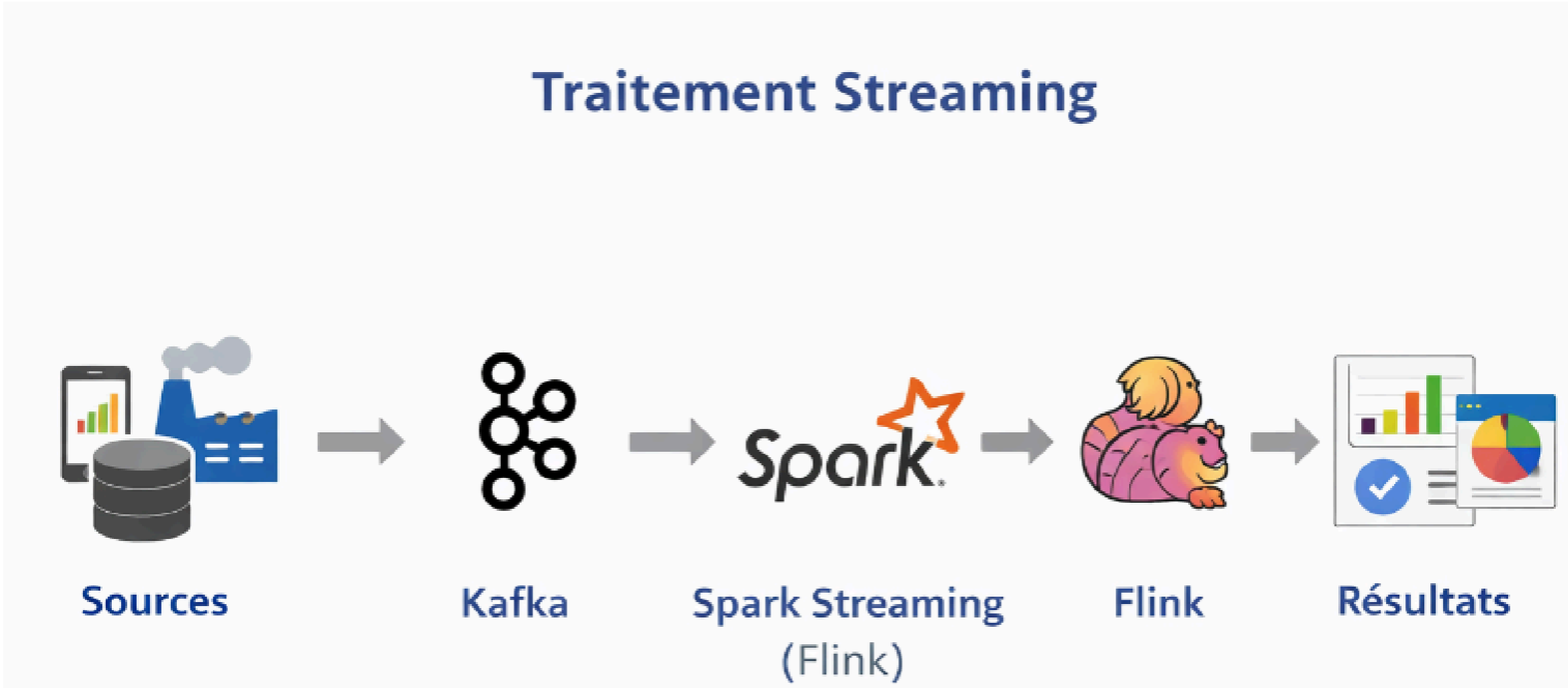
Architecture Big Data complète (vue système)

Traitement Streaming

Objectif : traiter les données en temps réel.

Principe

- **Les données sont traitées :**
  - dès leur arrivée
  - en continu



Ingestion des données

L’ingestion des données consiste à collecter et acheminer les données depuis différentes sources vers la plateforme Big Data.

Sources de données

- Applications
- Capteurs / IoT
- Logs systèmes
- Bases de données

Principaux modes et outils d’ingestion

- APIs : récupération de données depuis des services externes
- Kafka : ingestion de données en continu (streaming)
- Import batch : chargement périodique de gros volumes de données

L’ingestion est la porte d’entrée de toute architecture Big Data, en batch ou en temps réel.

Traitement des données Big Data

Le traitement consiste à analyser, transformer et agréger les données stockées ou en flux afin de produire de l'information utile.

Principaux moteurs de traitement

- **Apache Spark** : moteur de calcul distribué polyvalent pour le batch, le streaming et l'analyse de données à grande échelle.
- **Apache Flink** : moteur spécialisé dans le traitement temps réel, optimisé pour les flux continus à faible latence (ex : fraude, monitoring, alertes).

Stockage des données Big Data

Le stockage Big Data permet de conserver de très grands volumes de données de manière distribuée, fiable et scalable.

Principales solutions de stockage

- **Data Lake :**
  - Réservoir centralisé qui stocke toutes les données (brutes ou transformées), quel que soit leur format.
- **HDFS (Hadoop Distributed File System) :**
  - Système de fichiers distribué conçu pour stocker de très gros volumes de données sur un cluster de machines.
- **S3 (Cloud) :**
  - Stockage objet dans le cloud (ex : Amazon S3, Azure Blob, Google Cloud Storage), scalable et hautement disponible.

Le stockage Big Data est distribué, tolérant aux pannes et capable de passer à l'échelle horizontalement.

Énoncé du projet fil rouge — Big Data

Contexte

Vous êtes analyste de données / data engineer dans une entreprise de votre choix.

Votre entreprise dispose de grands volumes de données et souhaite les exploiter pour aider à la prise de décision.

On vous demande de concevoir et mettre en œuvre une solution Big Data permettant de répondre à une problématique métier précise.

Vous trouverez la description des différents blocs du projet ainsi que les consignes de dépôt des livrables sur mon outil de gestion et de suivi de projet, dont je vous partagerai le lien.