UE BIG DATA ARCHITECTURE APPLIED TO ARTIFICIAL INTELLIGENCE @ ESIGELEC - 2025/2026

Abdel Dadouche - DJZ Consulting

adadouche@hotmail.com

linkedin.com/in/adadouche/

Modalités de Validation

- 50% : Examen Final
  - (QCM et question ouvertes)
- 50%: Contrôle continu
  - (TP, assiduité, implication et collaboration)

Qui suis-je?

Abdelhalim Dadouche

Solution Architect
Consultant Freelance

- Jobs:
  - Depuis Juin 2025: Databricks
  - 2020-2025: Amazon Web Services
  - 2019-2020: Freelance
  - 2016-2019: SAP Developer Relation
  - 2013-2016: SAP Predictive Go To Market and CoE
  - 2010-2013: Solution Architect @ KXEN (acquired by SAP)
- Langage:
  - Français & English
  - Java, SQL, “predictive” & “machine learning”
- Hobby:
  - Hackathon (InnoJam ESIGELEC)
  - Bricolage

Objectifs: Comprendre les concepts fondamentaux

- Les différences entres Data Lake, Data Warehouse, Data Mesh, …
- Le rôle du Data Engineering, Machine Learning Engineering, Data Scientist
- Les modèles de données en étoile, en flocon ...
- L'ingestion de données, les Pipelines, le Serverless, l'approche Medaillon
- Analytique et Smart Analytics
- Offres Cloud autour de ces sujets

Dans la pratique

- Batch Data Pipelines
  - Conception et construction de pipelines de traitement par lots
  - Outils et technologies (ex: Apache Hadoop, Apache Spark)
  - Optimisation et gestion des performances
- Systèmes d'Analyse de Streaming
  - Concepts et architecture des systèmes de streaming
  - Outils et technologies (ex: Apache Kafka, Apache Flink)
  - Développement de solutions résilientes pour le streaming de données
- Smart Analytics, Machine Learning et IA sur le Cloud
  - Introduction aux services cloud pour l'IA et l'analytique
  - Utilisation de plateformes cloud (ex: Google Cloud, AWS, Azure)
  - Déploiement et gestion de modèles de machine learning sur le cloud
- Traitement des Données Serverless : Fondations
  - Concepts de serverless computing
  - Services serverless (ex: AWS Lambda, Google Cloud Functions)
  - Avantages et défis des architectures serverless
- Traitement des Données Serverless : Développer des Pipelines
  - Conception de pipelines serverless
  - Intégration de services serverless pour le traitement des données
  - Bonnes pratiques et optimisation des performances
- Traitement des Données Serverless : Opérations
  - Surveillance et gestion des pipelines serverless
  - Sécurité et gouvernance des données dans un environnement serverless
  - Automatisation et scaling des opérationsAutomatisation et scaling des opérations
- Préparation des Données pour les ML APIs
  - Techniques de nettoyage et de transformation des données
  - Intégration des données pour les modèles de machine learning
  - Utilisation des API de machine learning pour les prédictions
- Construire un Data Warehouse
  - Conception et modélisation d'un data warehouse
  - Sélection des technologies et outils (ex: Snowflake, Amazon Redshift)
  - Stratégies de chargement et de transformation des données (ETL/ELT)
- Ingénierie des Données pour la Modélisation Prédictive
  - Processus de préparation des données pour la modélisation prédictive
  - Outils et techniques pour l'ingénierie des fonctionnalités
  - Évaluation et validation des modèles prédictifs
- Construire un Data Mesh
  - Concepts et principes de l'architecture de data mesh
  - Décentralisation et gouvernance des données
  - Cas d'utilisation et exemples d'implémentation

Petit sondage avant de (bien) commencer

- Qui a déjà utilisé du SQL?
- Qui a déjà suivi un cours ou travaillé avec une base de données? HADOOP? Spark?
- Qui a déjà suivi un cours ou travaillé avec des outils de BI? de Data Mining?
- Qui a déjà travaillé sur des « gros » volumes de données? (gros, c'est-à-dire?)
- Qui a déjà utilisé SQL? Java? Linux? Machines virtuelles? Cloud?
- Qui a déjà travaillé sur des « gros » volumes de données? (gros, c'est-à-dire?)
- Qui a déjà utilisé le Cloud (et je ne parle pas de Facebook, Instagram ou TikTok)?

Pourquoi cette UE?

- Depuis quelques années, de nouveaux types de compétences sont de plus en plus recherchées:
  - Business Intelligence
    - MS PowerBI, Qlik, Tableau, SAP Analytics Cloud
  - Data Scientist
    - SAS, R, Python, Neural Network, Python, Sci-Kit…
  - Data Engineer
    - ETL, Data Cleansing, Management & Governance, Performances…
  - Big Data Architect & Engineer
    - Hadoop & Distributed systems
- Ainsi que des rôles plus fonctionnels:
  - Master Data Manager
    - Est garant de la “single version of truth” des data dans l’organisation
  - Data Officer
    - Définit la stratégie « Data » de la companie
  - Data Protection Officer
    - Garantie que les lois et réglementations soient correctement implémentées et mise en place

Let’s ChatGPT a few terms: “Business Intelligence”, “Analytics”, “Predictive”…?

Let’s ChatGPT “Business Intelligence”

- **Business Intelligence (BI)** is a **technology-driven process** that gathers, integrates, analyzes, and presents business data to produce actionable insights for smarter decision-making.

- It turns both structured and unstructured data into easy-to-understand formats like reports, dashboards, and visualizations, enabling organizations to monitor performance and uncover trends.

- Ultimately, BI empowers businesses to make informed strategic and operational decisions through data-driven insights that help enhance efficiency and secure competitive advantage

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « Business Intelligence » ?

Let’s ChatGPT “Analytics”

- **Analytics** is the **systematic computational analysis of data or statistics** to discover, interpret, and communicate meaningful patterns, and to inform effective decision-making.

- It combines disciplines such as statistics, computer programming, and operations research to quantify performance and predict outcomes across diverse fields.

- In essence, analytics transforms raw data into actionable insights that support better-informed decisions.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « Analytics »?

Let’s ChatGPT “Predictive Analytics”

- Predictive analytics is an advanced form of analytics that uses historical and current data—often applying statistical modeling, machine learning, and data mining techniques—to anticipate future outcomes and behaviors.

- It enables organizations to uncover underlying patterns and assess the likelihood of upcoming events in order to inform proactive decision-making.

- By going beyond what happened and why, predictive analytics helps businesses forecast trends like customer behavior, risk exposure, or demand shifts to guide strategy and operations.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « Predictive Analytics »?

A lot of words to say more or less the same:

“Help you take better decisions based on data”

Let’s ChatGPT “Big Data”

- **Big Data** refers to extremely large and complex datasets—often comprising structured, semi-structured, and unstructured data—that are too voluminous or fast-moving for traditional data-processing tools.

- It’s commonly characterized by the “Three V’s”: **volume**, **velocity**, and **variety**, with some definitions also including additional V’s like **veracity** and **value** to emphasize data quality and usefulness.

- Organizations leverage Big Data to uncover patterns, gain insights, and drive decisions that are infeasible with smaller-scale data sets.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « Big Data »?

Let’s ChatGPT “Data Lake”

- A **Data Lake** is a centralized repository that allows you to store all your structured, semi-structured, and unstructured data at any scale in its raw format.

- Unlike traditional databases or data warehouses, it doesn’t require data to be cleaned or transformed before storage, making it highly flexible and scalable.

- Data Lakes are often used for advanced analytics, machine learning, and real-time data exploration due to their ability to handle diverse and large data sets.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « Data Lake »?

Let’s ChatGPT “Data Mesh”

- **Data Mesh** is a **decentralized, domain-oriented architecture** that empowers individual business domains to own, manage, and serve their own data as products—moving away from central IT-heavy, monolithic data platforms.

- It rests on four core principles: **domain ownership**, treating **data as a product**, providing a **self-serve data platform**, and practicing **federated computational governance** so that data is discoverable, trustworthy, and aligned with business context.

- By shifting accountability to those closest to the data and enabling autonomy with standard guardrails, Data Mesh addresses the bottlenecks of centralized data architectures—enhancing scalability, agility, and data quality.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « Data Mesh »?

Let’s ChatGPT “ETL”

- **ETL** is a data integration process that involves **extracting** data from various sources, **transforming** it into a consistent format, and then **loading** it into a data warehouse or other storage system.

- This process ensures that data is accurate, clean, and ready for analysis or reporting.

- ETL is essential for enabling business intelligence, as it consolidates data from multiple systems into a single, usable view.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « ETL »?

Thanks and see you at the exam

Just kidding!!!

The concept behind “Big Data”

The rise of Distributed Computing and Storage

A Quick Reminder

- "Big data" usually relates to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.
- These datasets can be unstructured, semi-structured or structured
- The "size" of "Big data" is constantly moving
- The term has been in use since the 1990s

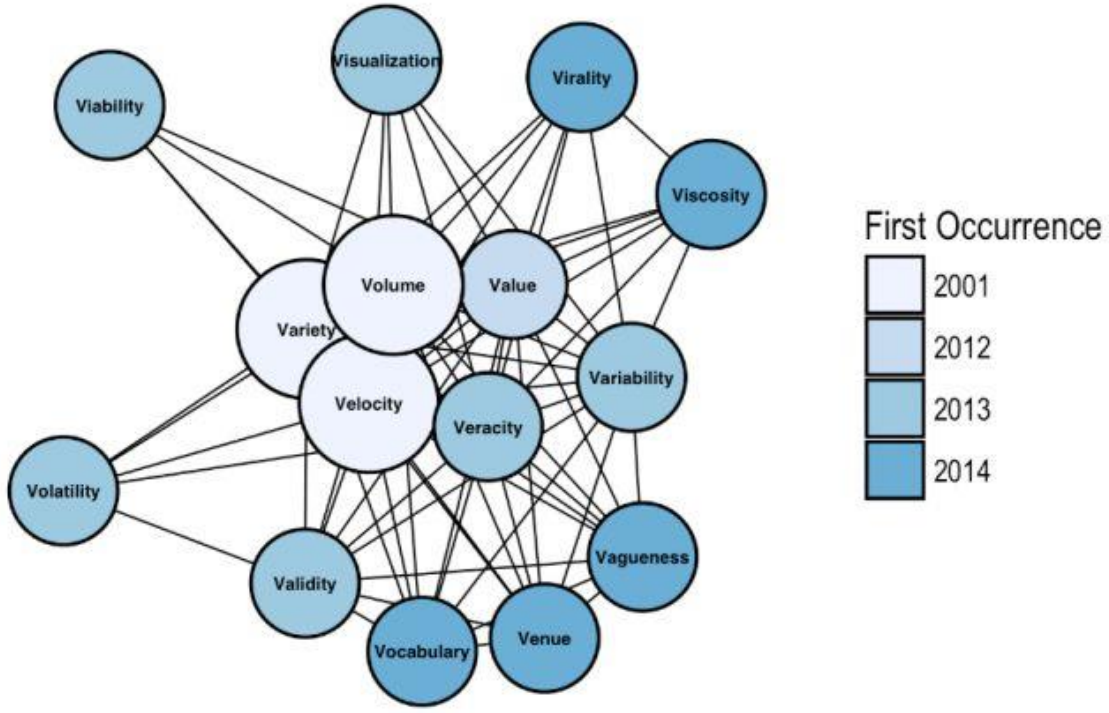
Source: https://en.wikipedia.org/wiki/Big_data

A History of V's

- The initial 3 V's of Big Data
  - **Volume**: The quantity of generated and stored data
  - **Velocity**: The speed at which the data is generated and processed
  - **Variety** : The type and nature of the data

- Then 7 V's: **Value, Veracity, Variability, Visualization**
- And some more: **Validity, Vulnerability, Volatility**

Now, up to the 42 V's of Big Data & Data Science by Tom Shafer, Elder Research, Inc.



Source: https://en.wikipedia.org/wiki/Big_data
https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html