Big Data and Cloud Computing

Dr. Walid Miloud Dahmane khemis miliana University SMI Department Email : w.miloud-dahmane@univ-dbkm.dz

1.0 March 2025

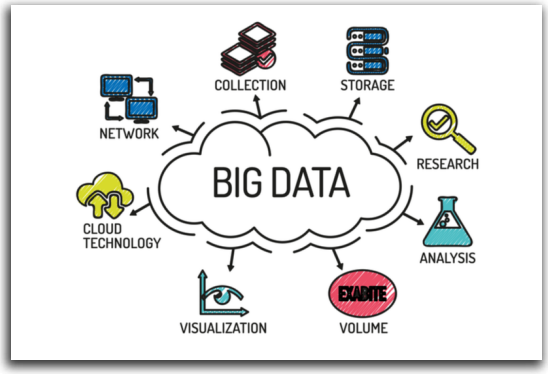
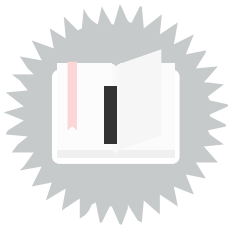


Table of contents

Chapter objectives



By the end of this chapter, the student will be able to:

- Understand the meaning of big data.
- Classify types of data.
- Understand the process of handling big data from its source to the analysis stage.

Chapter 1: Introduction to Big Data

1. What is a Big Data ?

Definition

Big Data refers to extremely large datasets that are complex and difficult to manage, process, or analyze using traditional data management tools.

Big Data Elements

Applications of Big Data

Extra

- **Healthcare:** Analyzing patient records to improve treatment plans.
- **Retail:** Predicting customer preferences for personalized shopping experiences.
- **Finance:** Detecting fraud and managing risk.
- **IoT:** Monitoring real-time data from connected devices.
- **Transportation:** Optimizing routes and reducing delivery times.

2. Characteristics of big data

Definition

Characteristics of BD

1. The "***volume***" in Big Data refers to the **enormous amount of data** generated every second. It's one of the key characteristics of Big Data and highlights the sheer size of the data being collected and stored.

   The **size** can range from terabytes (TB) to petabytes (PB) and even more.

2. The "***velocity***" in Big Data refers to the **speed at which data is generated, processed, and analyzed**. It highlights how quickly data flows into systems and the need to handle this fast-moving data in real time or near real time.

3. The "***variety***" in Big Data refers to the **different types and formats of data** that are collected from multiple sources. It highlights how Big Data includes both structured and unstructured information.

   Key Types of Data:

   - **Structured Data**
   - **Unstructured Data**
   - **Semi-Structured Data**

4. The "***veracity***" in big data refers to the quality, accuracy, and trustworthiness of the data being analyzed. It highlights the challenges posed by **uncertainty, inconsistencies**, and **biases** in data.

   Big data often comes from multiple sources, like social media, sensors, and logs, which may contain **errors, noise, or conflicting information.**

   How It's Managed :

   - **Data Cleaning**
   - **Validation**
   - **Algorithms**

5. The "***value***" in Big Data refers to the usefulness or insights that can be extracted from data. It emphasizes that data is only meaningful if it can provide actionable benefits to businesses, organizations, or individuals.

Not All Data is Useful: Just collecting large amounts of data is not enough. The value comes from analyzing and using it effectively.

Business Impact: Big Data helps businesses make better decisions, improve customer experiences, and reduce costs.

3. Differents Types of Data

- **Structured Data:** Organized data that follows a predefined format, stored in rows and columns, **e.g.,** Data in relational databases like SQL tables, and excel files (CSV).
- **Unstructured Data:** Data without a predefined structure, often qualitative and harder to process, **e.g.,** Social media posts, images, and videos.
- **Semi-Structured Data:** Data with some organizational structure but not fully organized into tables or databases, **e.g.,** JSON and XML, etc.

4. Big Data Structuring

Definition

Big Data structuring refers to the process of organizing and categorizing vast amounts of data into formats that can be efficiently stored, processed, and analyzed. It ensures that data can be accessed, interpreted, and used effectively for decision-making and insights.

Why is Structuring Important?

Note

- **Improved Data Management:** Well-structured data is easier to store, retrieve, and process.
- **Enhanced Analysis:** Proper structuring allows for meaningful insights by enabling better use of analytical tools.
- **Scalability:** Structured data grows easily.

How is Data Structured?

Method

- **Data Integration:** Combining data from multiple sources into one centralized system.
- **Data Cleansing:** Removing duplicates, errors, and irrelevant information.
- **Indexing and Metadata:** Adding labels or tags to make data easily searchable.
- **Partitioning:** Dividing data into smaller, manageable chunks for efficient processing.

Storage Solutions:

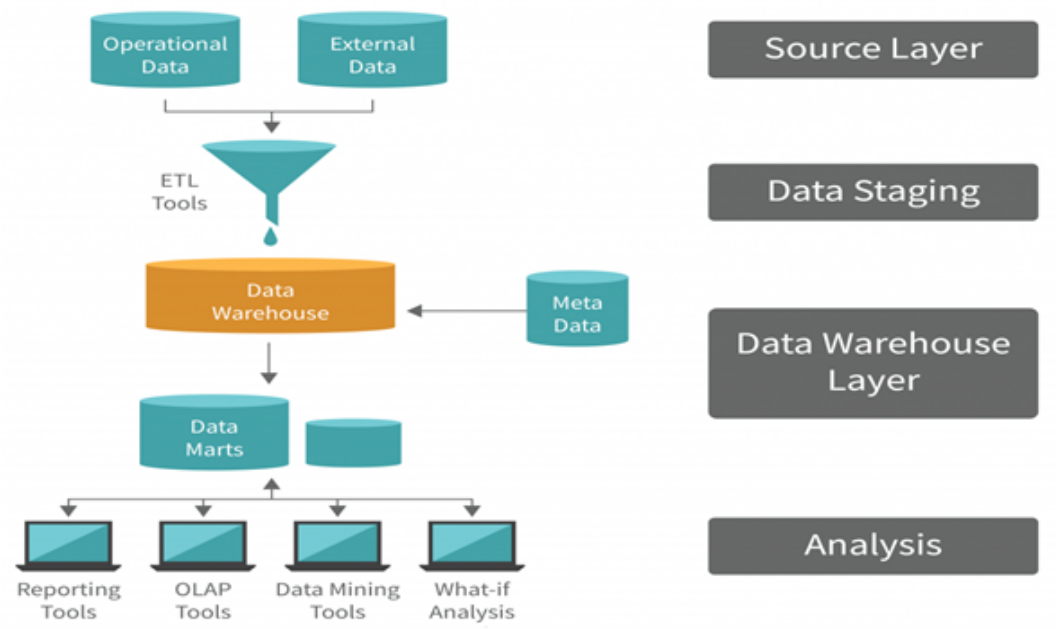
Note

- **Data Warehouses:** for structured data.
- **Data Lakes:** all types of data.

How a Data Warehouse Works

Extra

A **data warehouse** is a centralized repository designed specifically for storing and managing **structured data**. It is optimized for querying and analyzing large datasets, making it essential for business intelligence (BI) and decision-making processes.



Data Warehouse

1. **Source Layer:** This layer includes the **operational data** (such as transactional databases) and **external data** (e.g., third-party data, web data, etc.).
    - Acts as the input point for the data warehouse.
    - Collects raw data from multiple sources, which may have different formats, structures, or storage systems.
2. **Data Staging:** Data from the source layer is **E**xtracted, **T**ransformed, and **L**oaded (**ETL** process) into this staging area.
    - Cleans, integrates, and formats raw data to ensure consistency.
    - Handles data transformation, such as converting data types, handling missing values, or merging datasets.
3. **Data Warehouse Layer:** This is the central repository that stores the processed data in a structured format.
    - Stores historical and integrated data to support analysis and reporting.
    - Allows querying large datasets efficiently.

    Metadata: Provides information about the data, such as schema, relationships, and lineage.

    Data Marts: Smaller subsets of the data warehouse focused on specific business domains, like sales or finance.
4. **Analysis Layer:** This layer is where users interact with the data warehouse to extract insights.

    Provides tools and applications for analysis, reporting, and visualization.

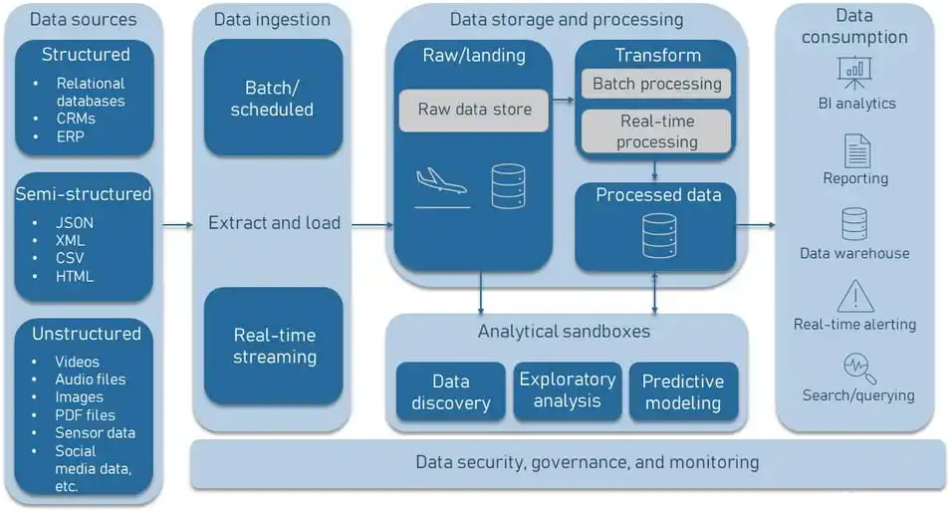
Benefits of Data Warehouses *Extra*

- **Improved Decision-Making:** Provides a single source of truth for organizational data.
- **Faster Query Performance:** Optimized for complex queries compared to operational databases.
- **Scalability:** Handles growing data volumes effectively.
- **Historical Analysis:** Stores time-variant data for trend and pattern detection.

How a Data Lake Works

Extra

A **data lake** is a centralized repository that stores large volumes of raw, unprocessed data in its native format, whether structured, semi-structured, or unstructured. It enables organizations to collect, manage, and process diverse datasets at scale, supporting a wide variety of use cases such as analytics, machine learning, and real-time processing.



Data Lake

1. Data Ingestion:
    - Data is gathered from multiple sources, including databases, streaming platforms, IoT devices, and APIs.
    - Tools like Apache Kafka, Flume, or AWS Glue facilitate ingestion.
2. Data Storage:
    - Raw data is stored in its native format (e.g., CSV, JSON, images, or videos).
    - Common storage solutions include Amazon S3, Azure Data Lake, or Hadoop Distributed File System (HDFS).
3. Data Processing:
    - Tools like Apache Spark or MapReduce process raw data for specific use cases.
    - Processing can be batch-oriented or real-time depending on the requirements.
4. Data Analytics and Machine Learning:
    - Analysts and data scientists use tools like TensorFlow, PyTorch, or BI tools (Power BI, Tableau) to analyze data or build predictive models.

Benefits of Data Lakes

Extra

- **Flexibility:**
    - Accommodates all data types (structured, semi-structured, and unstructured).
    - Allows experimentation with data without predefined schemas.
- **Scalable Storage:**
    - Handles petabytes or even exabytes of data efficiently.
- **Support for Advanced Analytics:**
    - Facilitates machine learning and predictive analytics by retaining raw data.

- **Unified Repository:**
  - Acts as a single source for organizational data.
- **Cost Savings:**
  - Storing raw data in inexpensive object storage is cost-efficient compared to traditional databases.

5. Governance and performance requirements

Governance in Big Data: *Definition*

Governance in Big Data refers to the **policies, procedures, and standards** that ensure data is managed, protected, and used effectively and ethically. It covers various aspects of data handling and management to maintain its quality, security, privacy, and compliance with regulations.

Why Governance Matters? *Note*

Effective governance is essential to ensure that Big Data can be **trusted, properly analyzed,** and used for **decision-making**, while meeting legal and regulatory standards.

Components of Governance: *Extra*

- **Data Quality:** Ensuring data is accurate, consistent, and up-to-date.
- **Data Security:** Protecting data from unauthorized access or breaches.
- **Data Privacy:** Ensuring that personal data is handled in compliance with privacy laws.
- **Compliance:** Following legal and industry regulations regarding data storage and processing.
- **Data Ownership:** Defining who is responsible for the data and who can access or use it.
- **Data Provenance:** Tracking the origins and history of data to ensure transparency and reliability.

Performance Requirements in Big Data: *Definition*

Performance requirements in Big Data refer to the **expectations for how well Big Data systems should perform** in terms of speed, scalability, reliability, and efficiency.

Why Performance Matters? *Note*

To derive valuable insights from Big Data, systems must meet these performance requirements. Poor performance can lead to slow data processing, delays in decision-making, and missed opportunities.

6. The job of data scientist

Who is Data Scientist? *Definition*

A **data scientist** is a professional who analyzes and interprets complex data to help organizations make informed decisions. They combine expertise in **statistics, programming, and domain knowledge** to extract actionable insights from structured and unstructured data.

Skills Required to Be a Data Scientist *Extra*

1. **Technical Skills**
    - **Programming:** (Object-Oriented , Declarative , Procedural, Logic, etc) programming languages.
    - **Data Manipulation:** Expertise in libraries like Pandas, NumPy, or databases like MySQL, PostgreSQL.
    - **Machine Learning:** Knowledge of algorithms, supervised/unsupervised learning, and deep learning.
    - **Big Data Tools:** Familiarity with Spark, Hadoop, or similar technologies.
    - **Cloud Platforms:** Experience with AWS, Azure, or Google Cloud for data storage and processing.
    - **Visualization:** Skilled in Tableau, Power BI, or matplotlib for creating visual insights.
2. **Soft Skills**
    - **Critical Thinking:** Ability to ask the right questions and find innovative solutions.
    - **Communication:** Translate technical results into business language.
    - **Problem-Solving:** Approach challenges systematically to find effective solutions.
    - **Collaboration:** Work with teams across different domains.

Challenges Faced by Data Scientists *Warning*

- **Data Quality Issues:** Ability to handle incomplete, inconsistent, or noisy data.
- **Scalability:** Processing large datasets efficiently.
- **Stakeholder Expectations:** Bridging the gap between technical possibilities and business needs.
- **Keeping Up-to-Date:** Constantly evolving tools, techniques, and algorithms.
- **Ethical Issues:** Ensuring privacy and integrity in data analysis.

7. Quiz: Practical Work

Exercise 1:

1. Define Big Data in your own words.
2. What are the three key characteristics (the "5 Vs") of Big Data? Explain each briefly.
3. Give two real-world examples of Big Data applications.

Exercise 2:

1. How does Volume differ from Velocity in Big Data? Provide examples.
2. Why is Variety a challenge in Big Data processing? Name three data formats it includes.
3. True or False: Big Data only refers to structured data. Justify your answer.

Exercise 3:

1. What is the purpose of structuring Big Data? Name two tools used for this.
2. Compare data lakes and data warehouses. When would you use each?

8. Submission Answer

Reply on the pratical work at PDF.

Send your answer through to me: **w.miloud-dahmane@univ-dbkm.dz**

Pupular Accounts:

- Gmail[1]
- Outlook/Hotmail[2]
- Yahoo[3]
- iCloud[4]

1. https://mail.google.com
2. https://outlook.live.com
3. https://mail.yahoo.com
4. https://www.icloud.com/mail