Introduction To Big Data Analytics INSY 8413

ADVENTIST UNIVERSITY OF CENTRAL AFRICA

Instructor:

- Eric Maniraguha | eric.maniraguha@auca.ac.rw | LinkedIn Profile

6h00 pm – 8h50 pm

- Monday A-G104
- Tuesday E-G108
- Wednesday A-G104
- Thursday E-G108

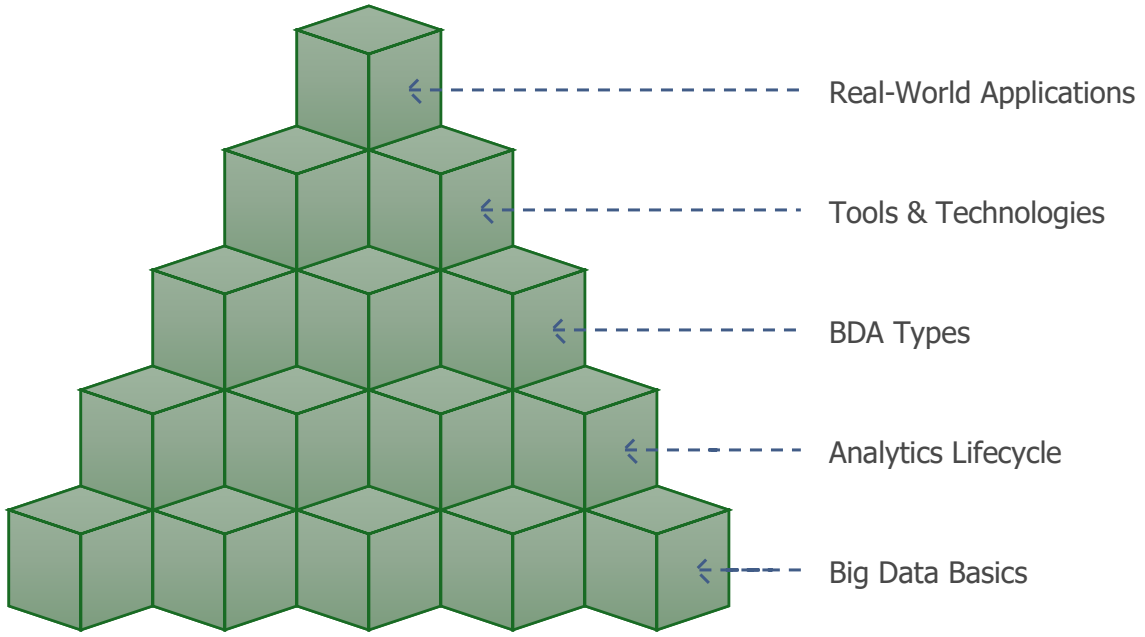
June 2025

Big Data, Explained: The 5 V’s of Big Data

Reference reading

- Big Data Analytics | What Is Big Data Analytics? | Big Data Analytics For Beginners | Simplilearn
- Amount of Data Created Daily (2025)
- Enterprise Data Warehouse: EDW Components, Key Concepts, and Architecture Types
- Big Data Analytics Tutorial

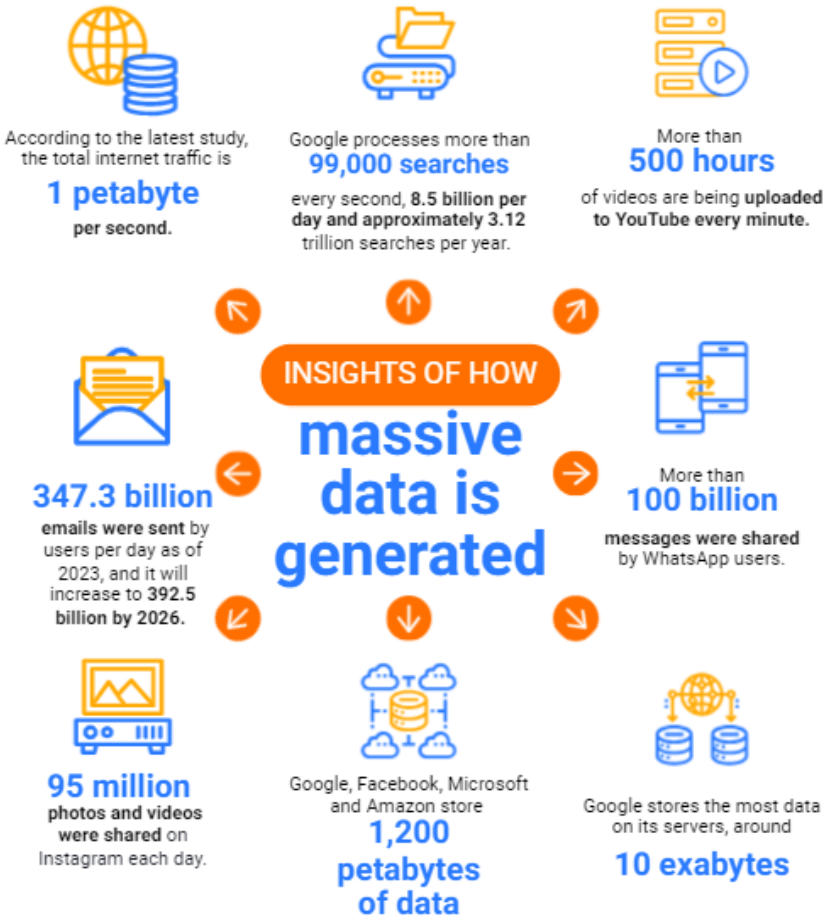
Big Data Analytics Learning Pyramid



Introduction to Big Data

What is Big Data?

Big Data refers to extremely large and complex **data sets generated continuously from various sources**, such as **social media platforms, sensors, devices, transactions**, and more. These data sets are so large and diverse that traditional data processing tools and methods cannot efficiently handle them.
Big Data is not just about volume — it's about **extracting meaningful insights from vast** and **varied information, helping organizations make smarter, data-driven decisions.**



Source Image: https://www.datasciencesociety.net/a-step-by-step-guide-to-data-visualization/

Big Data, Explained: The 5 V's of Big Data

To understand Big Data better, it helps to explore the **5 V's** — five key dimensions that describe its unique characteristics and challenges:

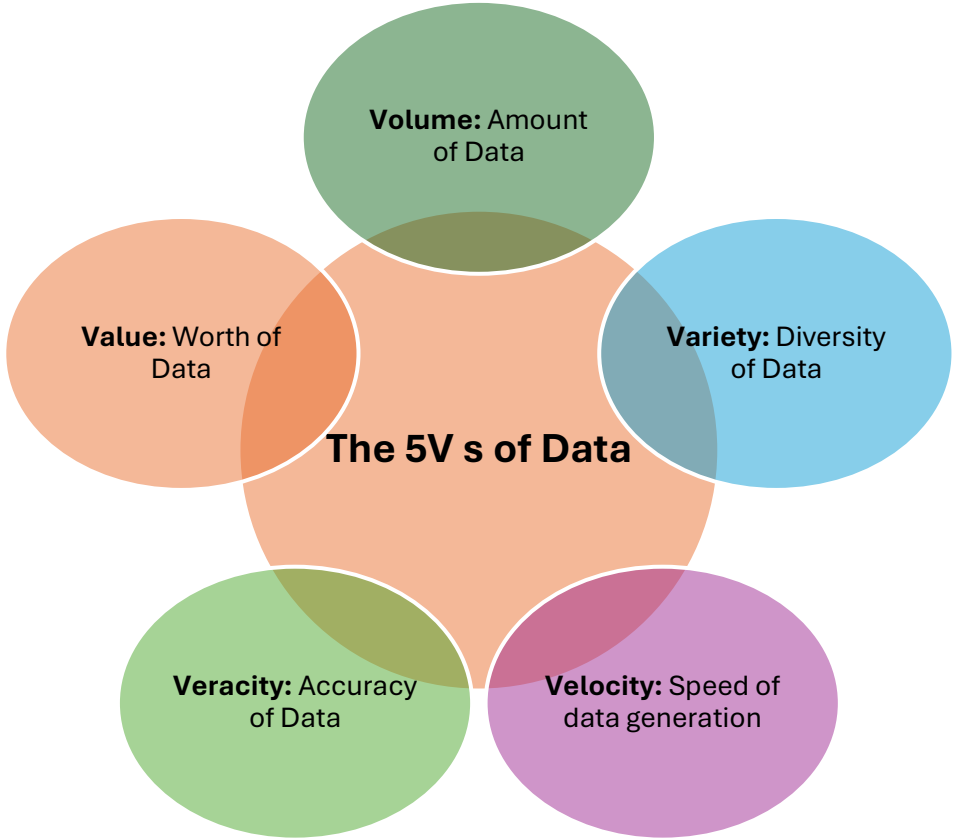
1. Volume
The enormous amount of data generated daily — ranging from social media posts, sensor outputs, financial transactions, and beyond. The scale is so vast that traditional storage and processing tools are inadequate.

2. Velocity
The speed at which new data is created and processed. For example, real-time social media updates, live financial trades, and IoT device signals demand rapid data ingestion and analysis to remain valuable.

3. Variety
Data today comes in many formats — structured (like databases), semi-structured (like JSON or XML), and unstructured (such as videos, images, emails, and text). This diversity requires flexible storage and processing solutions.

4. Veracity
The accuracy and trustworthiness of data. Big Data often contains noise, errors, or inconsistencies. Ensuring high-quality, reliable data is essential to make sound decisions.

5. Value
The ultimate goal of Big Data is to extract actionable insights that drive real business value — enabling better decisions, innovation, and competitive advantage.



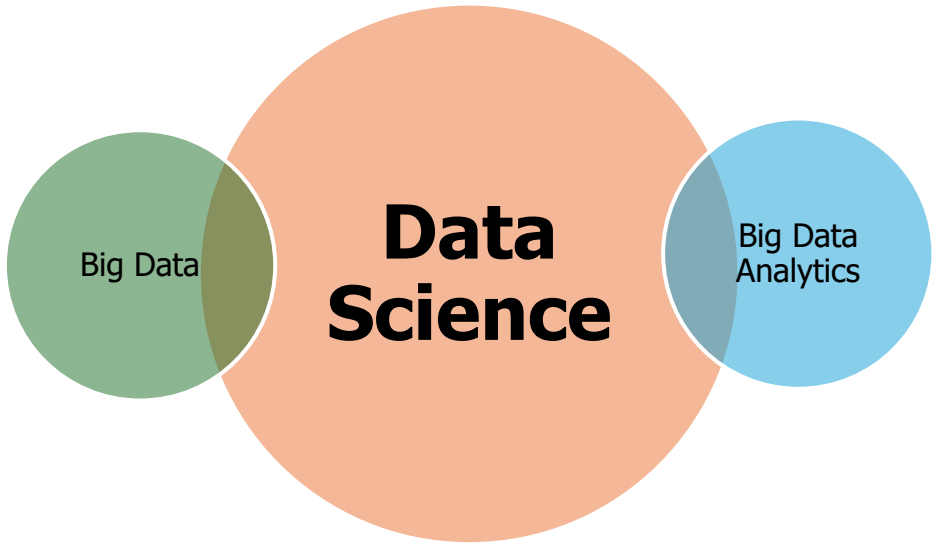
Why Big Data? (With Examples – Rwanda Context)

Make	**Make Data-Driven Decisions** **Example:** The City of Kigali uses traffic data and surveillance systems to plan infrastructure and manage congestion.
Gain	**Gain Deeper Insights** **Example:** Rwanda's Irembo platform collects and analyzes citizen interactions to improve e-government services.
Improve	**Improve Operational Efficiency** **Example:** Rwanda Energy Group (REG) leverages smart meters to monitor electricity usage and detect faults faster.
Enable	**Enable Innovation** **Example:** The Vision City housing project in Kigali integrates smart sensors for energy management, waste disposal, and water usage.
Support	**Support Predictive Analytics** **Example:** Rwanda National Police analyzes accident and crime data to deploy patrols more effectively and reduce incidents.
Respond in	**Respond in Real-Time** **Example**: Kigali's Smart Bus System uses GPS and passenger data to provide real-time updates and optimize routes.

Data related fields

Big Data vs. Big Data Analytics

Aspect	Big Data	Big Data Analytics
Definition	Massive volumes of structured, semi-structured, and unstructured data.	The process of examining big data to uncover insights.
Focus	Data characteristics (Volume, Velocity, Variety, etc.)	Extracting meaning and patterns from big data.
Purpose	To store and manage large datasets.	To analyze data for decision-making and predictions.
Technologies	HDFS, NoSQL databases, cloud storage, data lakes	Hadoop, Spark, Hive, ML algorithms, visualization tools
Output	Raw data	Actionable insights



Big Data Analytics Life Cycle (9 Phases)

Business Case / Problem Definition
Define the business problem, goals, and scope. Assess if it's truly a Big Data challenge (volume, velocity, variety).

Data Identification
Identify relevant data sources (internal or external) based on the business need.

Data Acquisition & Filtration
Collect and filter raw data, removing corrupt or irrelevant information.

Data Extraction
Transform incompatible data to a format aligned with analysis objectives.

Data Munging (Validation & Cleaning)
Clean and validate data by handling nulls, errors, and inconsistencies.

Data Aggregation & Representation
Join datasets into a unified form and store efficiently for analysis.

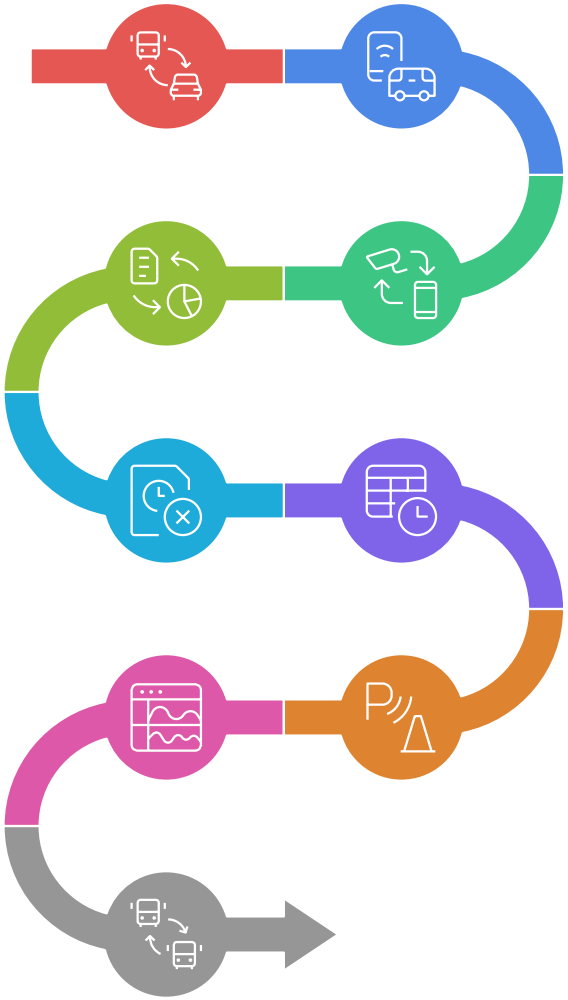
Exploratory Data Analysis (EDA)
Discover patterns and test hypotheses using statistical and ML techniques.

Data Visualization
Use charts and dashboards to present insights in an understandable way.

Utilization of Results
Apply findings to decision-making, optimization, or system improvements.

Scenario Big Data Analytics Life Cycle

Streamlining Kigali's Traffic with Big Data



Business Case / Problem Definition

Reduce traffic jams in Kigali

Data Identification

GPS data from buses, CCTV footage, road sensors

Data Extraction

Convert raw feeds into structured formats

Data Acquisition & Filtration

Collect video feeds and sensor logs

Data Munging (Validation & Cleaning)

Fix timestamp mismatches, remove duplicates

Data Aggregation & Representation

Combine data from all locations by timestamp

Data Visualization

Use heatmaps, time-series charts, and dashboards

Exploratory Data Analysis (EDA)

Identify areas with peak congestion

Utilization of Results

Redesign traffic light timings and reroute buses

Kigali Reduces Traffic Congestion Using Big Data Analytics

City experiences traffic jams during peak hours.

Analyze traffic data to optimize routes and timings.

Traffic flows smoothly, minimizing delays.

Impact:
Data-driven decisions lead to reduced commute times, improved air quality, and better citizen satisfaction.

Types of Data Analytics: A Guide to Informed Decision-Making

What is Data Analytics?
Data analytics involves collecting, cleaning, analyzing, and interpreting data to extract insights, support decision-making, and optimize operations.

Prescriptive

Recommend support strategies for students | Recommend support strategies for students

Predictive

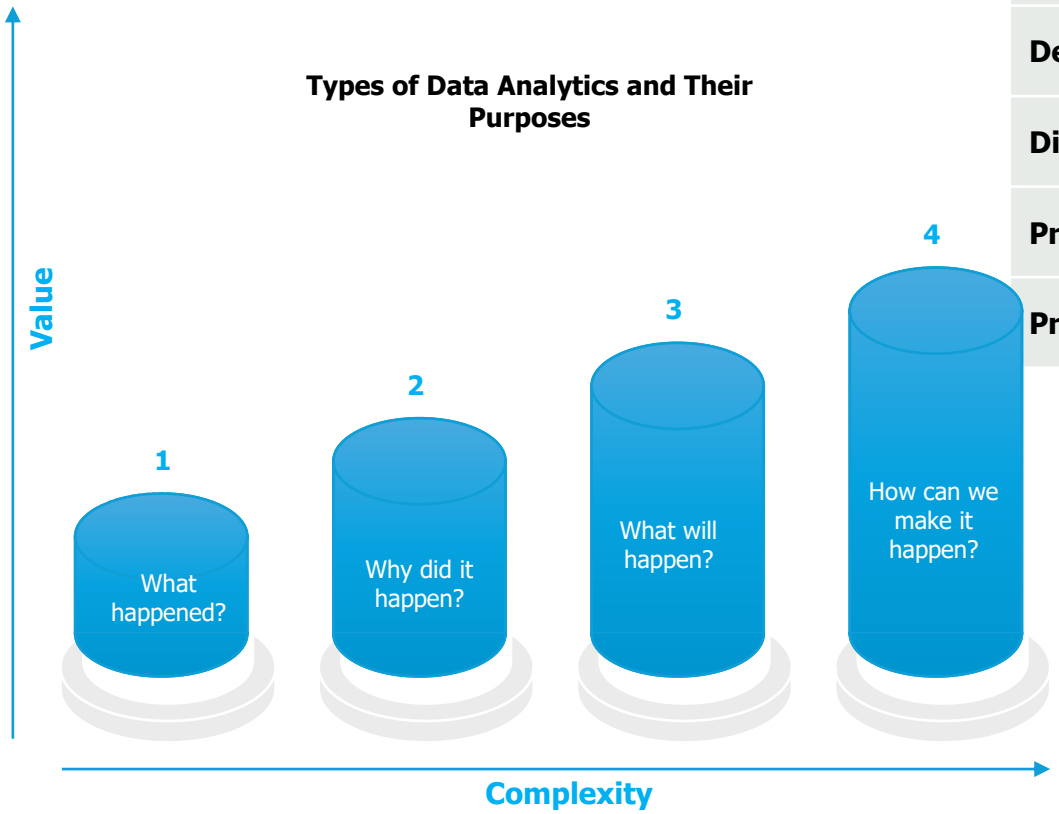
Forecast students who may struggle | Forecast student struggles with topics

Diagnostic

Investigate issues in LMS functionality | Investigate LMS functionality issues

Descriptive

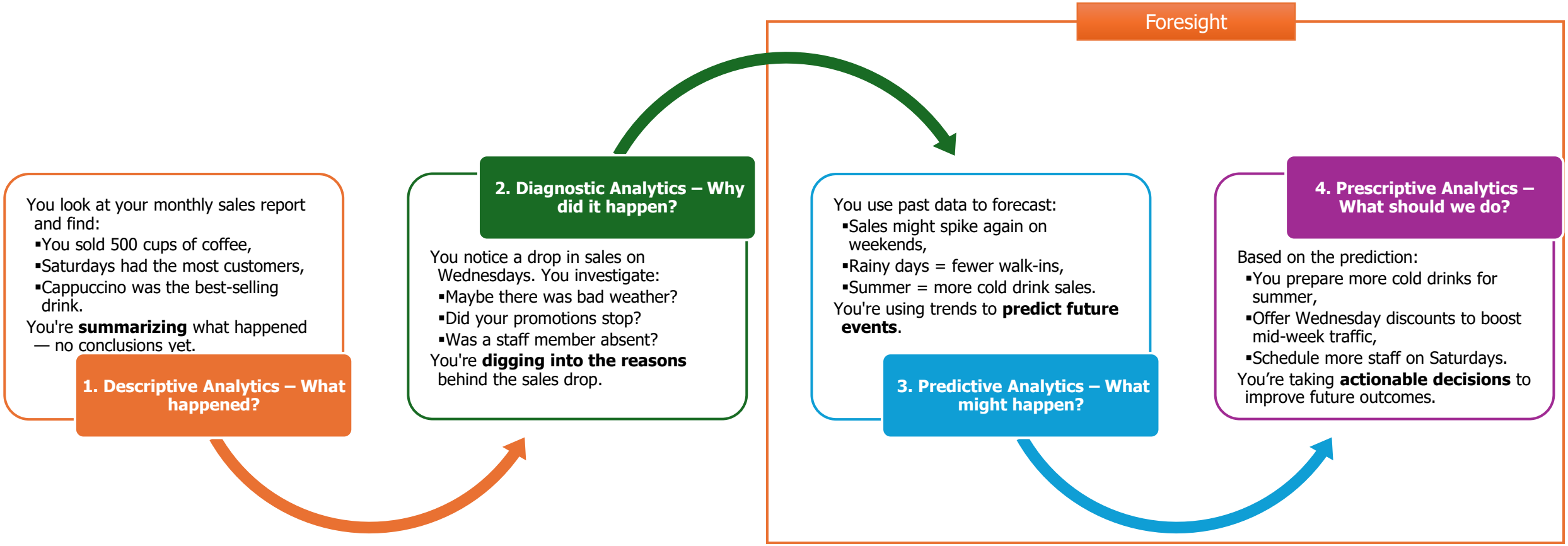
Summarize usage data in LMS | Summarize popular LMS features



Type	Purpose	Example
Descriptive	What happened?	Summarize usage data in an LMS to find popular features
Diagnostic	Why did it happen?	Investigate issues in LMS functionality based on user feedback
Predictive	What is likely to happen?	Forecast students who may struggle with certain topics
Prescriptive	What should be done?	Recommend support strategies based on predicted student performance

LMS stands for **Learning Management System**. It is a software platform used to **create, deliver, manage, and track educational courses or training programs**. LMSs are widely used in schools, universities, and corporate training environments.

Scenario: Running a Small Café



Wrap-Up:
"Just like managing a café, businesses use these 4 types of data analytics to understand the past, explain the present, predict the future, and take smart actions."

Types of Big Data

Big Data can be categorized into **three main types** based on its structure and source:

1. Structured Data

- **Definition**: Data that is organized in a predefined format (rows and columns).
- **Storage**: Stored in relational databases (RDBMS).
- **Examples**:
  - Transaction records
  - Sensor data in tables
  - Customer information (name, age, phone number)
- **Tools**: SQL, Oracle DB, MySQL

2. Unstructured Data

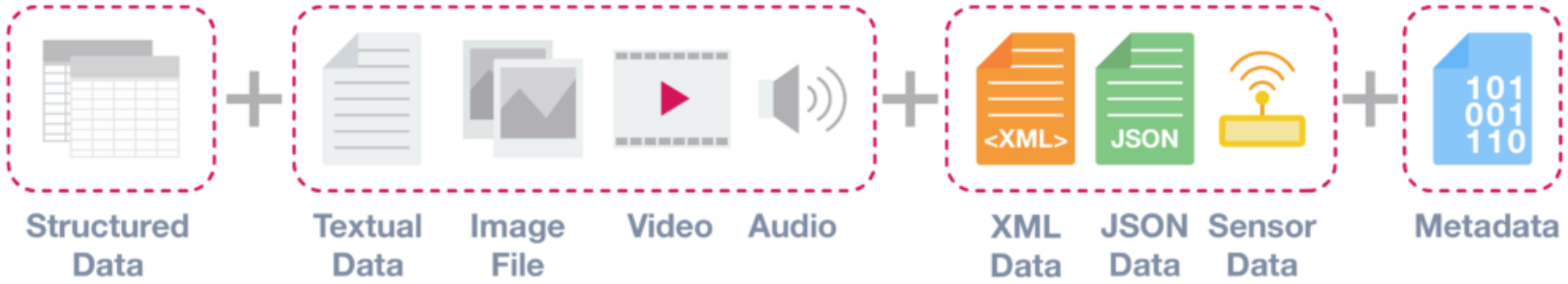
- **Definition**: Data that does not have a specific structure or format.
- **Storage**: Requires NoSQL databases or data lakes.
- **Examples**:
  - Social media posts
  - Emails
  - Videos, images, audio files
  - PDFs, Word documents
- **Tools**: Hadoop, Spark, MongoDB, Elasticsearch

3. Semi-Structured Data

- **Definition**: Data that does not reside in a relational database but still has some organizational properties (e.g., tags, hierarchies).
- **Examples**:
  - XML and JSON files
  - Web pages (HTML)
  - Sensor logs with labels
- **Tools**: NoSQL databases like Cassandra, MongoDB

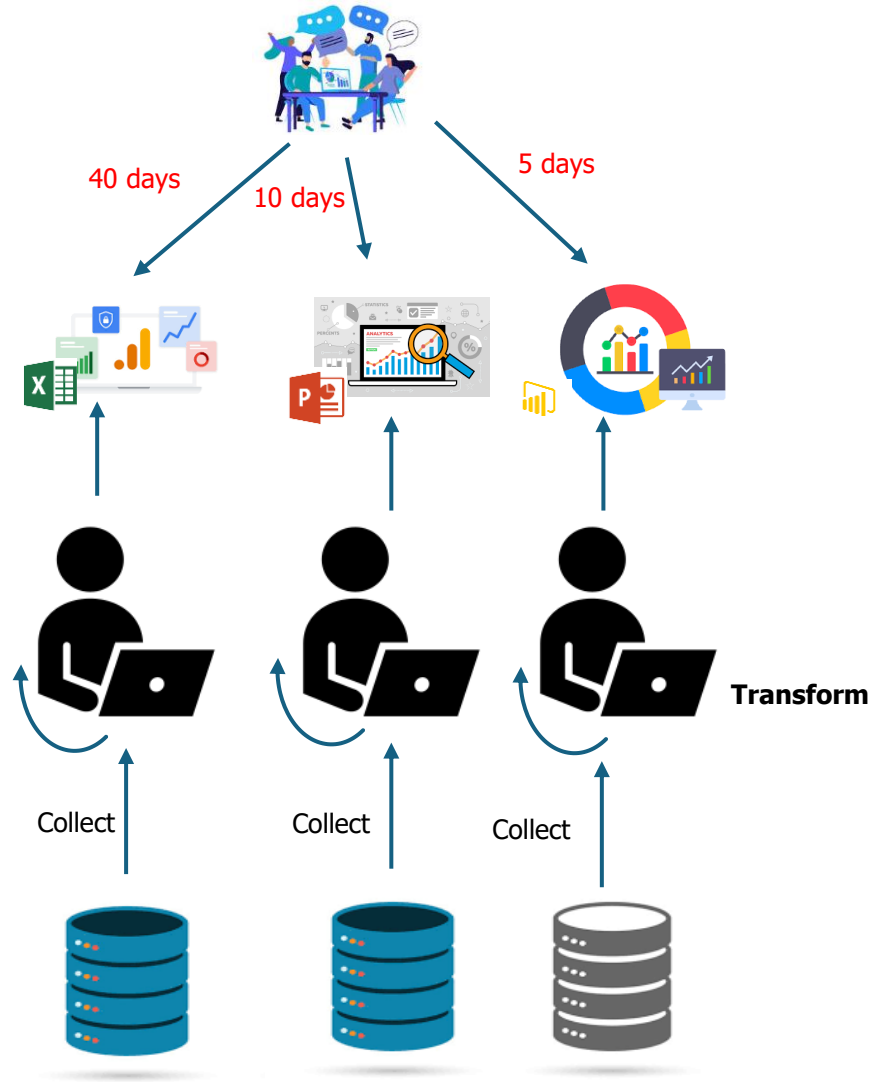
Type	Structure	Examples	Tools
Structured	Tabular	Sales records, SQL tables, spreadsheets	SQL, Oracle, MySQL
Unstructured	No clear format	Tweets, videos, documents	Hadoop, Spark, MongoDB
Semi-Structured	Partial structure	XML, JSON, sensor logs, zip file, audio	Cassandra, MongoDB, Hive

Data Types: Structured vs. Unstructured Data

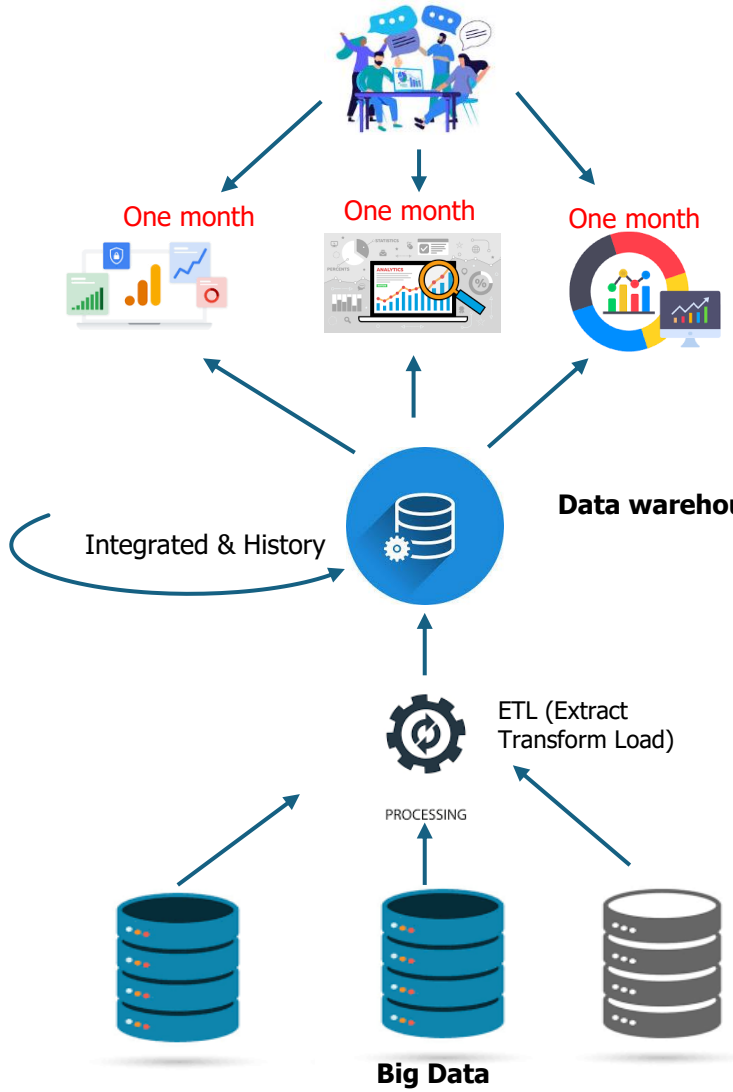


Metadata – Data About Data
Metadata is information that describes other data. While not a standalone data structure, it plays a key role in Big Data analysis by offering context—such as time and location—for datasets. For example, photo metadata may include the date and place the photo was taken. These details, often structured, support early stages of data analysis in Big Data systems.

Data Management Process - ETL



Automation



Data Analytics

Data Engineer

Key Terms in Big Data Analytics

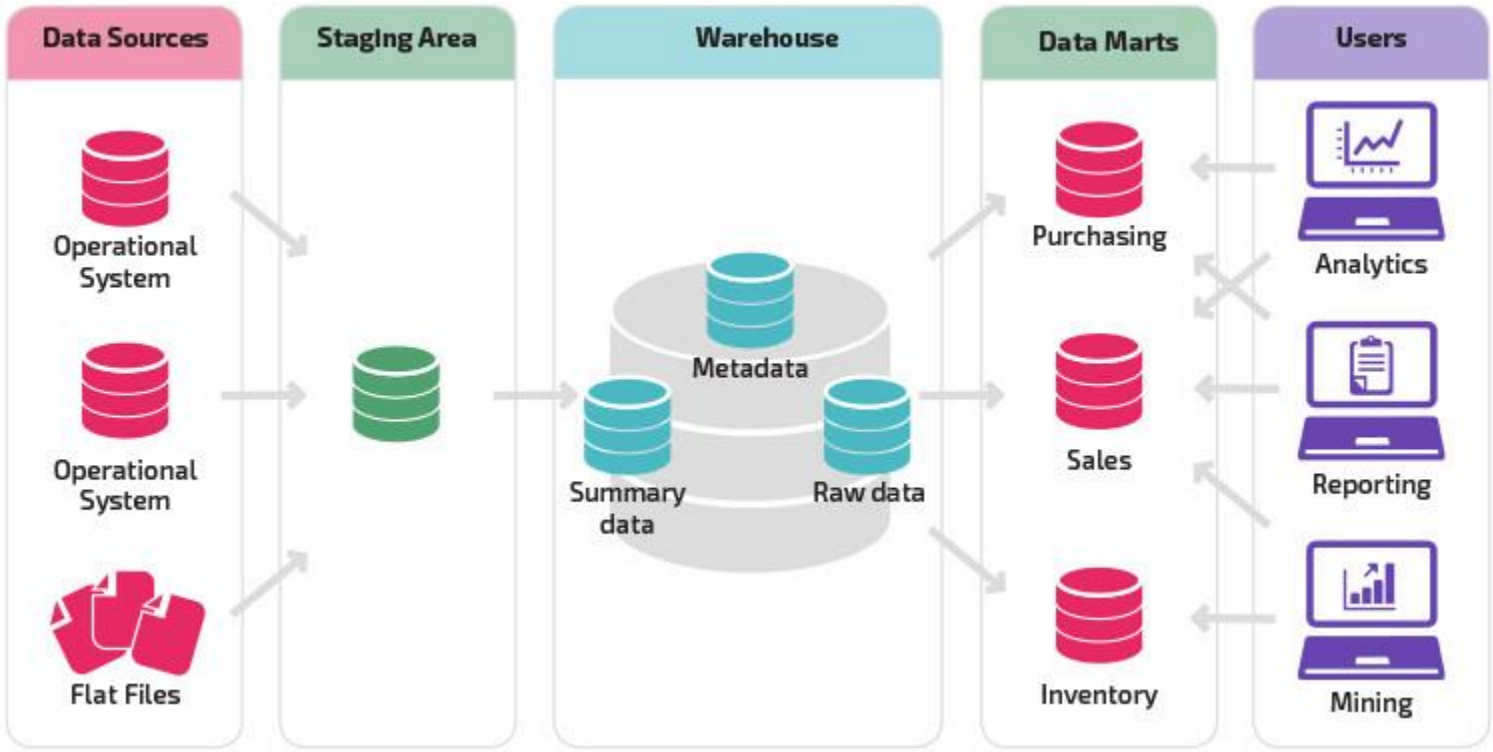
1. Big Data
Large, complex datasets defined by the 5 Vs:
Volume, **Velocity**, **Variety**, **Veracity**, **Value**

2. ETL (Extract, Transform, Load)
A process that:

- **Extracts** data from sources
- **Transforms** it into a usable format
- **Loads** it into a data storage system

3. Data Warehousing
Centralized storage for integrated, cleaned, and structured data—used for reporting and analytics.

4. Data Mining
Discovering hidden patterns, trends, and relationships in large datasets using statistical and machine learning methods.



Source Image: https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/

Data Lake Architecture

Source Image: https://www.qlik.com/us/data-lake

ETL stands for:

- **Extract** – Get data from different sources
- **Transform** – Clean and organize the data
- **Load** – Put the data into a system for reporting or analysis

Applications of Big Data

Big Data is transforming various industries by enabling deeper insights, automation, and smarter decision-making. Below are key application areas:

1. Healthcare

- **Applications**:
  - Predictive analytics for patient diagnosis
  - Tracking disease outbreaks
  - Personalized treatment plans
  - Genomic analysis
- **Example**: Using big data to detect early signs of cancer or manage pandemics (e.g., COVID-19 trends).

2. Finance and Banking

- **Applications**:
  - Fraud detection
  - Risk management and credit scoring
  - Algorithmic trading
  - Customer behavior analysis
- **Example**: Real-time monitoring of transactions for suspicious activities.

3. Retail and E-commerce

- **Applications**:
  - Customer segmentation and recommendation engines
  - Inventory and supply chain optimization
  - Dynamic pricing
- **Example**: Amazon using big data to recommend products based on purchase history.

4. Manufacturing

- **Applications**:
  - Predictive maintenance
  - Quality control using sensor data
  - Optimizing production processes
- **Example**: Monitoring equipment to prevent costly breakdowns.

5. Agriculture

- **Applications**:
  - Precision farming using satellite data
  - Crop yield prediction
  - Weather forecasting for planting
- **Example**: Using drones and sensors to optimize water and fertilizer usage.

Key Technologies in Big Data Analytics

Data Storage & Processing

- **Hadoop** – Distributed storage and batch processing framework.
- **Apache Spark** – Fast in-memory data processing engine.
- **Apache Flink** – Stream and batch processing.
- **Apache Storm** – Real-time data stream processing.

Data Warehousing

- **Hive** – Data warehouse on top of Hadoop, SQL-like querying.
- **Amazon Redshift** – Cloud data warehouse.
- **Google BigQuery** – Serverless, highly scalable cloud warehouse.

Data Ingestion

- **Apache Kafka** – Distributed event streaming platform.
- **Apache NiFi** – Data flow automation and management.
- **Flume** – Collecting and moving large log data.

Data Querying

- **Presto** – Distributed SQL query engine.
- **Impala** – Real-time queries on Hadoop.

Machine Learning & AI

- **TensorFlow** – Open-source ML framework.
- **PyTorch** – ML framework for deep learning.
- **H2O.ai** – Open-source ML platform.
- **Apache Mahout** – Scalable ML on Hadoop.

Data Visualization

- **Tableau** – Interactive data visualization.
- **Power BI** – Business analytics by Microsoft.
- **QlikView / Qlik Sense** – Associative data visualization.

Data Integration

- **Talend** – Open-source data integration.
- **Informatica** – Enterprise data integration suite.

Cloud Platforms

- **AWS (Amazon Web Services)** – Offers EMR, Redshift, etc.
- **Google Cloud Platform** – BigQuery, Dataproc, AI tools.
- **Microsoft Azure** – Azure Synapse, HDInsight, ML Studio.

Careers Involving Big Data Analytics

With the increasing reliance on data, demand for skilled professionals has surged:

- **Data Scientist:** Develop models to uncover insights and predict trends.
- **Data Analyst:** Transform data into actionable business insights.
- **Data Engineer:** Build and maintain the infrastructure for big data processing.
- **Machine Learning Engineer:** Create algorithms to learn from data and automate decisions.
- **Business Intelligence Analyst:** Deliver data visualizations and reports for stakeholders.
- **Data Visualization Specialist:** Design intuitive visuals to communicate data effectively.
- **Data Architect:** Design data systems to support big data initiatives.

Thank you!