KEMENTERIAN SUMBER MANUSIA

MALAYSIA MADANI

KELAS BIMBINGAN PEMBANGUNAN KERANGKA DATA RAYA

10 – 12 DIS 2024 (SELASA – KHAMIS)
8:30 PAGI HINGGA 4:30 PETANG
OrenG ACADEMY

Day #1

Introduction to Big Data and Types of Platforms

Definition

BIG DATA

refers to extremely **large and complex datasets**

that are challenging to process, store, and analyze using traditional methods

due to their size, speed of generation, and variety of formats.

Importance of Big Data

Cost Saving

- Optimize resources and reduce inefficiencies, lowering operational costs

Time Saving

- Automated data processing and real-time analytics speed up tasks that would take hours or days manually

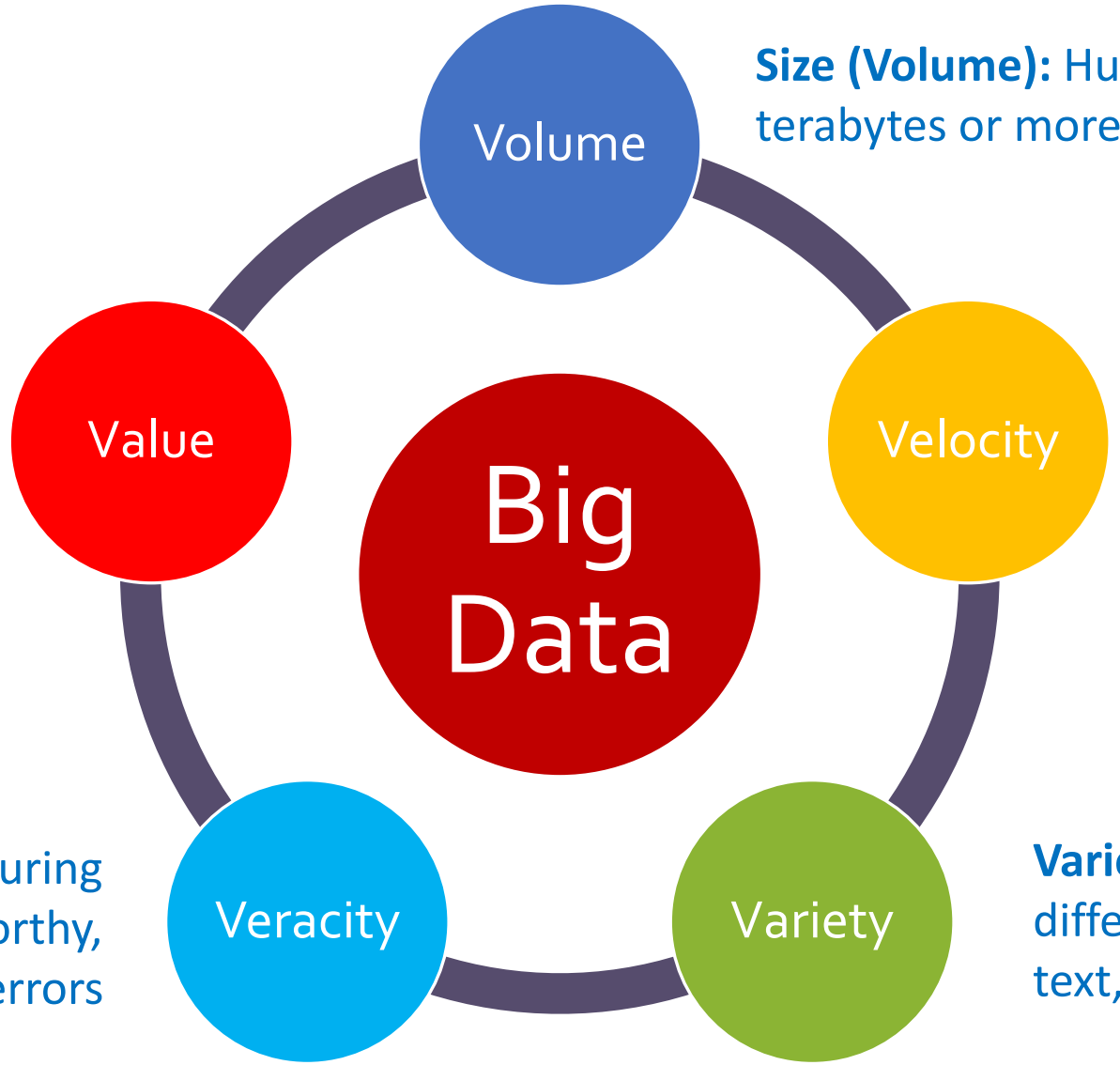
Faster & Better Decision Making

- Data-driven insights enable more accurate and timely decisions, improving outcomes

Data-Driven Policy Making

- In sectors like government, Big Data provides empirical evidence for crafting effective policies

Characteristics of Big Data



Size (Volume): Huge amounts of data, often in terabytes or more

Speed (Velocity): Data is generated and processed very quickly, sometimes in real-time

Variety: Data comes in different formats – numbers, text, images, videos, etc

Accuracy (Veracity): Ensuring the data is clean, trustworthy, and free of errors

Value: The insights and benefits you can get from analyzing the data

Big Data Platform

Understanding Data

What Data (Represent)?

How Data is Used?

SIMPLIFIED EXPLANATION

How Data is Stored?

Data Life Cycle/Stage

Understanding Data

What Data / What Data Represents?

Data is **information** that can be collected, stored, and analyzed

Understanding Data

How Data is Stored?

Structured Data:
Highly organized and easily searchable within relational databases

Unstructured Data:
Data that lacks a predefined model or format

Semi-structured Data:
Data that has some organizational properties but does not fit neatly into a structured database.

Understanding Data

How Data is Used?

Operational Data:
Used to run day-to-day tasks

Analytical Data:
Used to find trends and make decisions

Real-Time Data:
Used for live updates

Historical Data:
Stored for later use

Understanding Data

Data Life Cycle/Stage

Raw Data:

Like ingredients in the kitchen—collected but unprocessed

Cleaned Data:

Prepped and ready to use—errors removed

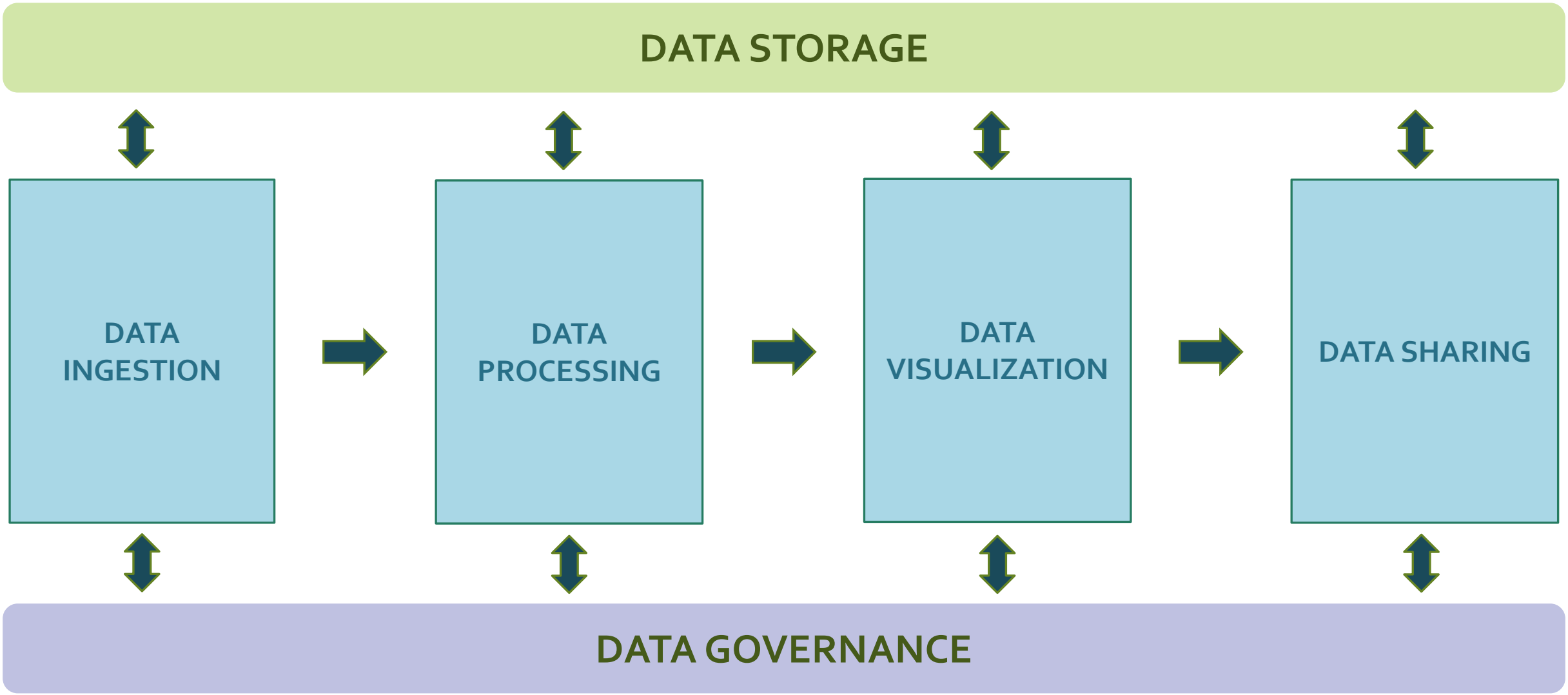
Analyzed Data:

Insights drawn—helps make decisions

Archived Data:

Stored away for future reference

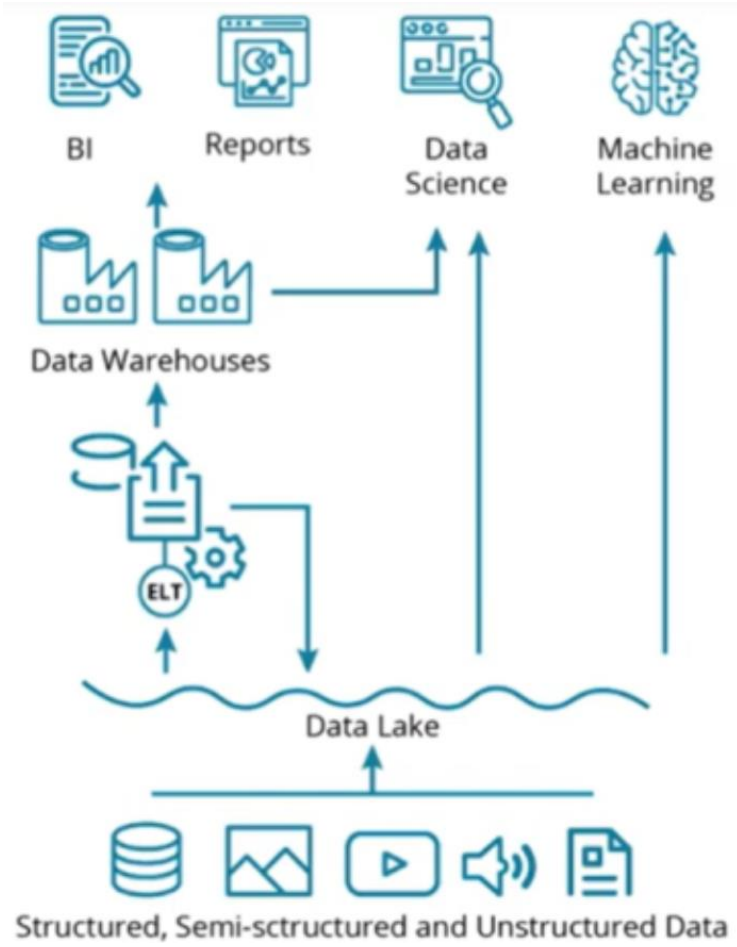
Foundation/Components of Big Data



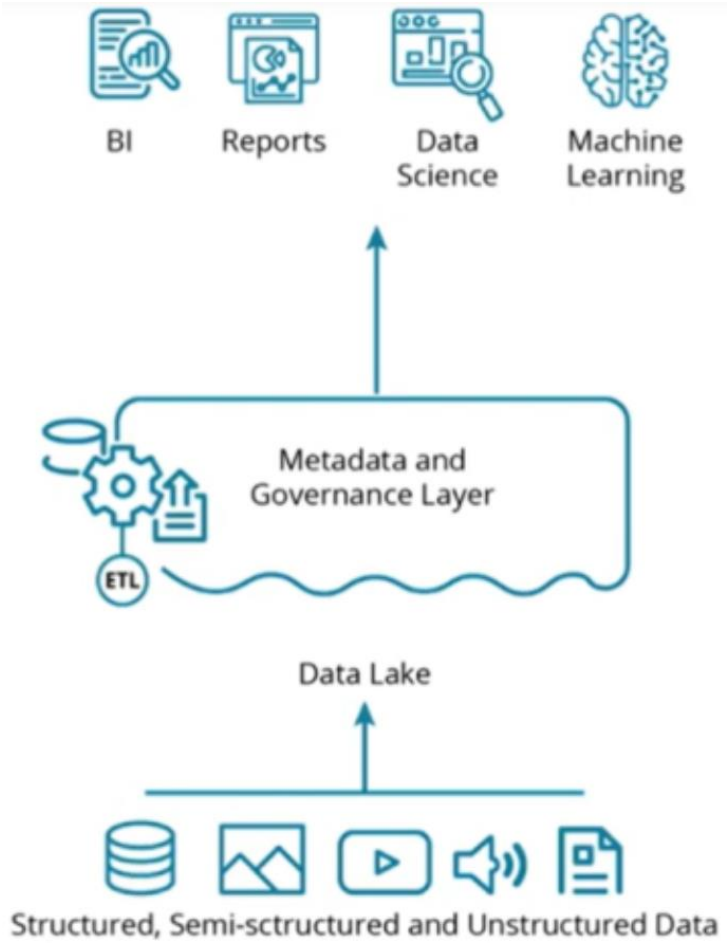
Types of Big Data Platforms

Data Warehouse

Data Lake



Data Lakehouse



Comparison of Data Management Systems

Feature	Data Warehouse	Data Lake	Data Lakehouse
Data Type	Structured	Structured, Semi-structured, Unstructured	Structured, Semi-structured, Unstructured
Use Case	Analytical, Reporting	Storage, Exploration, Analytics	Hybrid Storage and Analytics
Schema	Defined	No strict rules; you can dump data in any format	A hybrid approach, combining both strict and flexible organization
Query Complexity	Complex	Varied	Complex
ACID Compliance	Limited	Typically None	Provided by technologies like Delta Lake
Data Processing Tools	Business Intelligence Tools (e.g., Tableau)	Big Data Tools (e.g., Spark, Hadoop)	Hybrid Approach with Big Data Tools
Scalability	Scalable for analytics	Highly Scalable	Scalable, but may require a robust architecture
Example Tools	Snowflake, Redshift, BigQuery	Amazon S3, Azure Data Lake Storage, Hadoop	Delta Lake, Databricks, AWS Glue
Data Integration	ETL and ELT (Extract, Load, Transform) processes	Often uses ETL/ELT, supports data from various sources	ETL and ELT for structured and raw data
Common Use Cases	Historical Sales Analysis, Reporting	Raw data storage, Sensor Data, User-generated Content	Healthcare Data, IoT Data, Financial Data
Storage Efficiency	Optimized for query performance	Low-cost storage for diverse data types	Storage efficiency can vary based on architecture

** ACID is principles make sure your data processes are reliable, accurate, and protected, even when working with massive amounts of information or during unexpected problems

Comparison Simple Analogy

Database	Data Warehouse	Data Lake	Data Lakehouse
Like a well-organized filing cabinet	Like a library where books (data) are categorized for easy access	Like a giant storage room where you toss everything without sorting	Like a library that also has a storage room for unsorted items

Relationship Between Big Data And Data Analysis

Definition

BIG DATA

refers to extremely **large and complex datasets**
that are challenging to process, store, and analyze using traditional methods
due to their size, speed of generation, and variety of formats.

DATA ANALYTICS

is the **process** of examining, interpreting, and analyzing Big Data
to uncover patterns, trends, correlations, and insights.

KEMENTERIAN SUMBER MANUSIA

Simplified Definition

MALAYSIA MADANI

Aspect	Big Data	Data Analytics
What it is	Large amounts of data	Process of analyzing that data
Focus	It emphasizes the storage, processing, and management of large, complex datasets	It focuses on extracting value and actionable insights from Big Data
Purpose	Serves as the raw material. It provides the information but doesn't derive meaning from it	Serves as the tool or technique used to interpret and make use of the raw data
Example	Collecting data from millions of users about their online shopping habits	Analyzing shopping habit data to recommend products to users or improve sales strategies

Key Elements in Data Analysis

Data Sources	Data Processing	Data Cleaning	Analysis & Modelling	Visualization & Presentation
Objective: Define the problem or question and determine the required data	**Objective:** Gather raw data from identified sources	**Objective:** Ensure the data is accurate, consistent, and usable	**Objective:** Apply analytical techniques to derive insights	**Objective:** Present insights visually for clarity and impact
Actions: • Identify internal and external sources • Classify data as structured or unstructured	• **Actions:** • Web scraping • Surveys • Database queries • Upload Module • API • IoT sensors	**Actions:** • Handle missing data (impute, remove, or flag) • Remove duplicates data • Standardize data formats	**Actions:** • Summarize data (mean, median, standard deviation). • Perform statistical analysis (e.g., hypothesis testing, regression)	**Actions:** • Create dashboards • Create Charts • Create Reports

Types of Data Analytics

Descriptive Analytics

- Summarizes past data to understand what has happened

Diagnostic Analytics

- Investigates data to determine why something happened

Predictive Analytics

- Uses statistical models and machine learning to forecast future trends

Prescriptive Analytics

- Suggests actions based on predictive insights to achieve desired outcomes

Challenges of Data Analytics

Data Quality Issues	Volume and Variety	Skill Gap	Privacy and Security
• Missing, incomplete, or unreliable data	• Handling large and diverse datasets	• Need for expertise in programming, statistics, and domain knowledge	• Ethical considerations and legal compliance in data handling

Popular Open Source Technology in Big Data

Apache NiFi

Kafka

Airbyte

ClickHouse

Apache Hadoop

PostgreSQL

Open for Innovation KNIME

Python

R

Apache Spark

Airflow

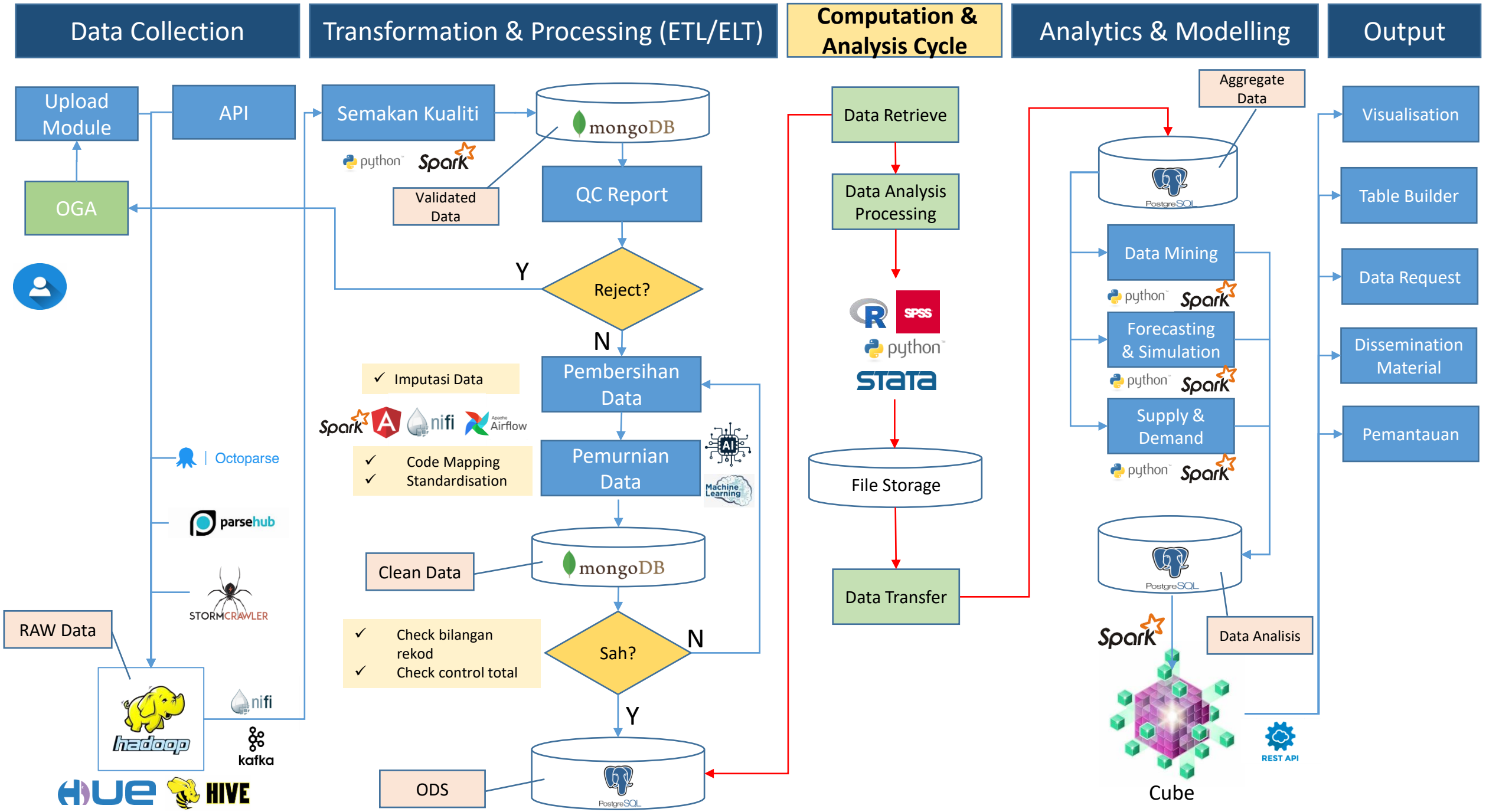
FastAPI

Python

Flask

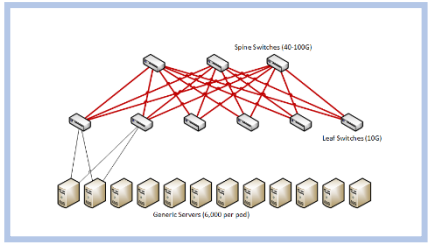
Apache Superset

Metabase


Data Collection
Transformation & Processing (ETL/ELT)
Computation & Analysis Cycle
Analytics & Modelling
Output
Upload Module
API
OGA
Octoparse
parsehub
STORMCRAWLER
RAW Data
hadoop
nifi
kafka
HUE
HIVE
Semakan Kualiti
mongoDB
python
Spark
Validated Data
QC Report
Reject?
Y
N
Pembersihan Data
✓ Imputasi Data
Spark
nifi
Airflow
Pemurnian Data
✓ Code Mapping
✓ Standardisation
AI
Machine Learning
Clean Data
mongoDB
✓ Check bilangan rekod
✓ Check control total
Sah?
N
Y
ODS
PostgreSQL
Data Retrieve
Data Analysis Processing
R
SPSS
python
STATA
File Storage
Data Transfer
Aggregate Data
PostgreSQL
Data Mining
python
Spark
Forecasting & Simulation
Supply & Demand
PostgreSQL
Spark
Data Analisis
REST API
Cube
Visualisation
Table Builder
Data Request
Dissemination Material
Pemantauan

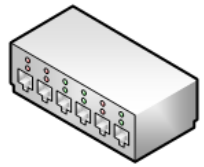
Massive Scale Data Analytic Platform

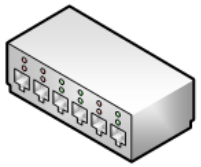
Spine-Leaf network architecture with all software running in containers



To Existing Data Center Infrastructure
1GB/10Gb/25Gb/100Gb connectivity

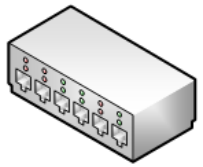
2 x Spine Switch
32 x 100Gb

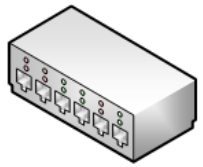


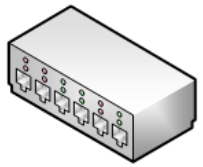


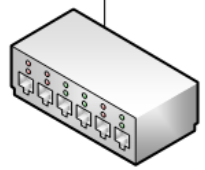
8 x 100Gb (800GB) Connectivity per Switch

4 x Leaf Switch
48 x 25Gb + 8 x 100Gb









4 x 25Gb (100GB) Connectivity per Node

4 x 25Gb (100GB) Connectivity per Node

Typical Cluster Size

TOTAL CPU 1,040 Core
TOTAL MEMORY 27.8Tb
TOTAL STORAGE 1,704Tb
TOTAL APPLIANCE 24 Nodes

796,480 Nodes Max

Horizontally scale up to 760 Appliance per Pod, 1,048 pods per Cluster, for maximum of 796,480 Nodes of Data Analytic Nodes per Data Center.
Providing long-term, low cost, on-premise, private, hyperscale technology for large scale data analytic

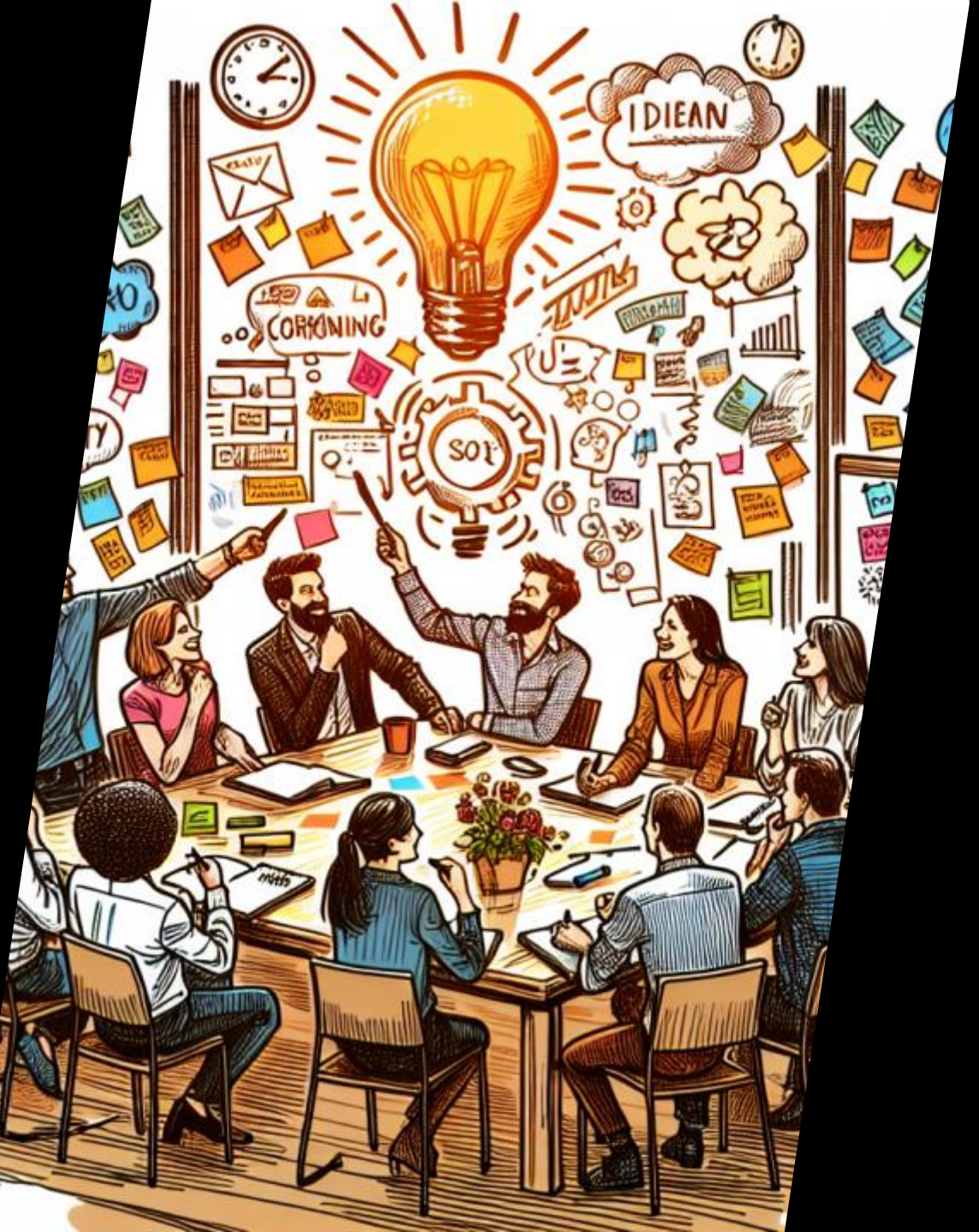
4 x Storage Appliance
2x8c CPU, 512Gb Memory, 192/250Tb Storage
TOTAL CPU 64 Core
TOTAL MEMORY 2Tb
TOTAL STORAGE 1Pb

5 x Management Appliance
2x24c CPU, 1Tb Memory, 24Tb Storage
TOTAL CPU 240 Core
TOTAL MEMORY 5Tb
TOTAL STORAGE 120 Tb

5 x Data Stream Appliance
2x24c CPU, 1Tb Memory, 64Tb Storage
TOTAL CPU 240 Core
TOTAL MEMORY 5Tb
TOTAL STORAGE 320 Tb

5 x Compute Appliance
2x24c CPU, 1Tb Memory, 24Tb Storage
TOTAL CPU 240 Core
TOTAL MEMORY 5Tb
TOTAL STORAGE 120 Tb

5 x Query Appliance
2x24c CPU, 2Tb Memory, 24Tb Storage
TOTAL CPU 240 Core
TOTAL MEMORY 10Tb
TOTAL STORAGE 120 Tb



Brainstorming Activity

KEMENTERIAN SUMBER MANUSIA

List of Datasets does KSM/Jabatan Generate or Use

MALAYSIA MADANI