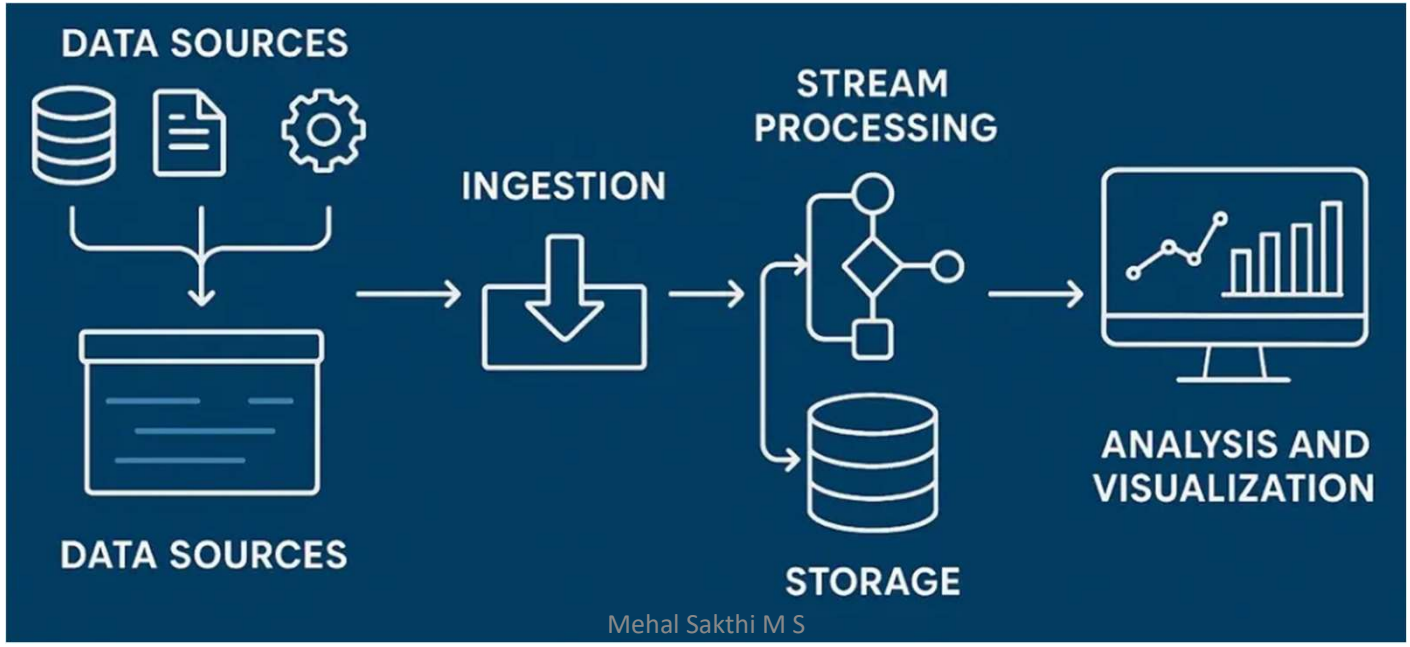
Big Data Architecture & Ecosystem Overview

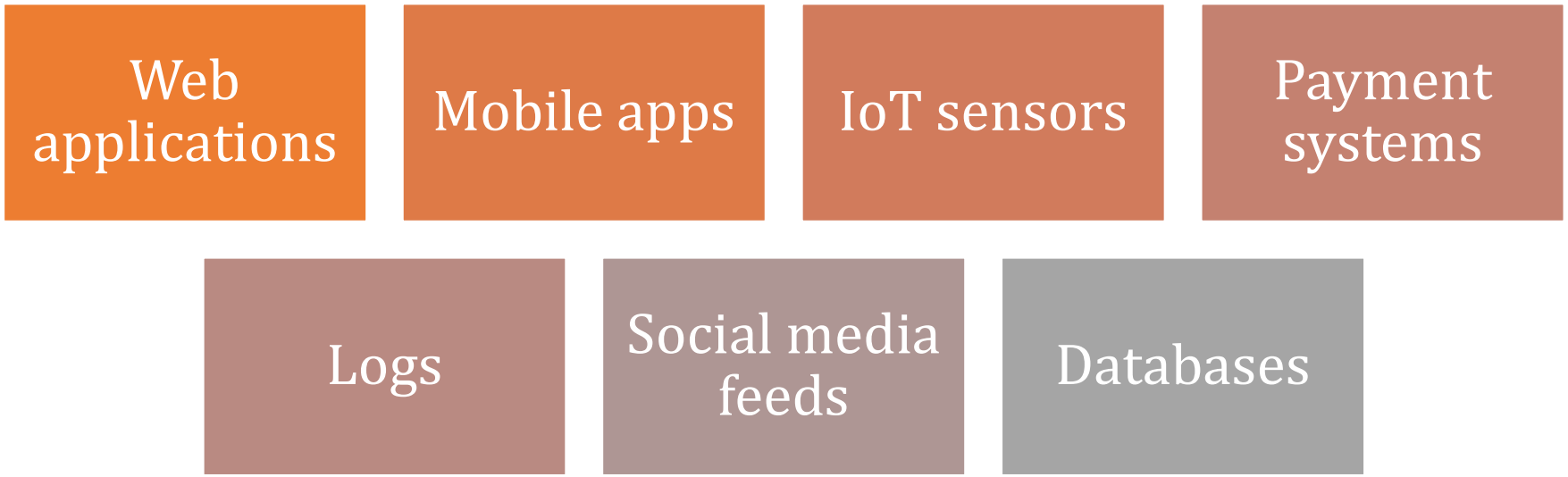
Mehal Sakthi M S
Analyst SDE

What is Big Data Architecture

Where is data stored?	How does it move?	How is it processed?	How does it become insights?



Data Sources Layer



Structured

Semi-structured

Unstructured data

Data Ingestion Layer

The process of collecting and moving data into storage systems.

Batch Ingestion

Real-time or Streaming Ingestion

Tools Used

Apache Kafka

Apache Flume

APIs

Storage Layer

Where is Big Data Stored?

Data Lakes	Stores raw data in original format
Data Warehouses	Stores structured, processed data.

Distributed File Systems

A system that stores data across multiple machines (nodes) but appears as a single storage system.

HDFS (Hadoop Distributed File System)

- Splits large files into blocks
- Stores blocks across different machines
- Replicates data for fault tolerance
- Automatically handles failures

Instead of moving data to computation
Move computation to where data exists.

Storage Layer

Data Lake	Data Warehouse
Raw data	Processed data
All formats	Structured
Flexible	Optimized
Schema-on-read	Schema-on-write

Data Lakehouse

Data Lake flexibility
+
Data Warehouse performance

- **No duplication of data**
- **Unified analytics + AI**
- **Better governance**
- **ACID transactions on data lakes**

Processing Layer

Batch Processing	Stream Processing
Processes data in bulk	Processes data continuously
Scheduled jobs (hourly/daily)	Real-time / near real-time
High latency (minutes–hours)	Low latency (milliseconds–seconds)
Suitable for historical analysis	Suitable for instant decision-making
Simpler architecture	More complex distributed systems
Example: Monthly sales report	Example: Fraud detection alert

Processing Layer

Apache Spark

An open-source distributed processing engine designed for large-scale data processing.

Key Characteristics

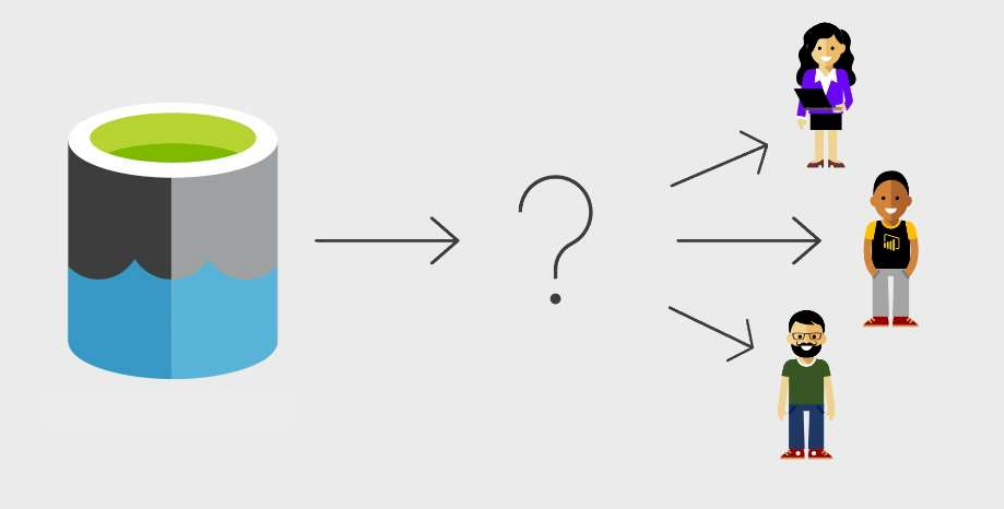
- Distributed Processing Engine
- In-Memory Processing
- Fault Tolerant
- Lazy Evaluation
- Unified Analytics Engine

Spark Ecosystem Components

- Spark Core
- Spark SQL
- Spark Streaming
- Spark MLlib
- GraphX

Serving Layer

This is where we make data usable



- Dashboards
- BI tools
- APIs
- Machine learning models

Analytics & AI Layer

- Reports are created
- Insights are generated
- ML models are trained
- Predictions are made

Use Case Discussion

Thank You