Big Data and Analytics

1. Big Data

Definition: Big Data refers to massive and complex data sets that traditional database management systems cannot efficiently handle. This data is generated from various sources, such as social media, eCommerce, IoT, and more.

2. History of Big Data

The growth of Big Data is attributed to:

- **Proliferation of the Internet**
- **Rise of social media**
- **Increase in eCommerce** activities

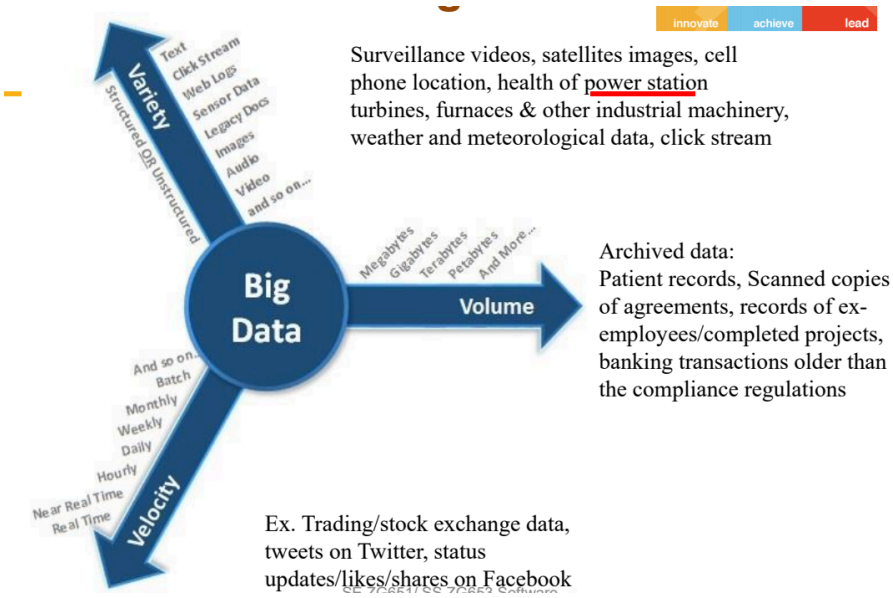
---

3. Big Data Characteristics

Big Data is characterized by:

- **Volume**: Large amounts of data.
- **Velocity**: Rapid data generation and processing.
- **Variety**: Diverse data types, including structured and unstructured data.

Diagram Placeholder: Characteristics of Big Data



---

4. Applications of Big Data

Examples:

- **Healthcare**: Recommending cancer treatment based on similar patient cases.
- **Weather Forecasting**: Alerts for farmers and fishermen.
- **Fraud Detection**: Monitoring transactions for suspicious patterns.

Big Data Technologies

1. Google BigTable

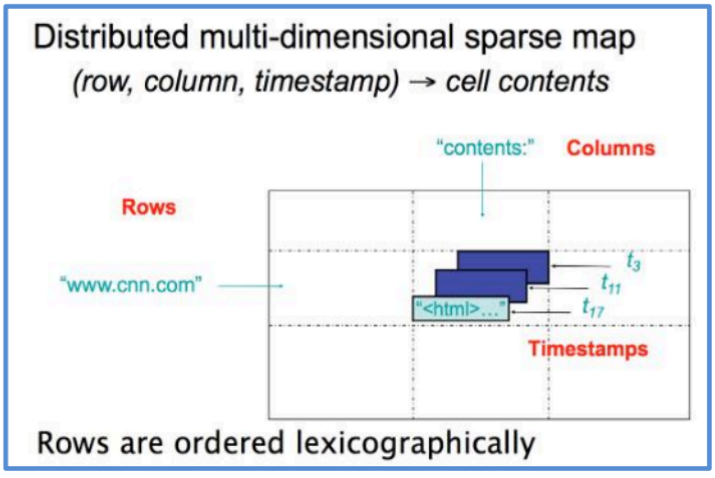
Google BigTable is a distributed storage system developed by Google to handle vast amounts of structured data. BigTable is optimized for scalability, allowing fast access across thousands of servers.

Key Features:

- **Versioning of data**
- **Data compression**
- **Distributed storage across servers**
- **Fault tolerance and fast access**
- **Dynamic addition of servers and load balancing**

Diagram Placeholder: Google BigTable

So Google invented a data storage structure called Big Table



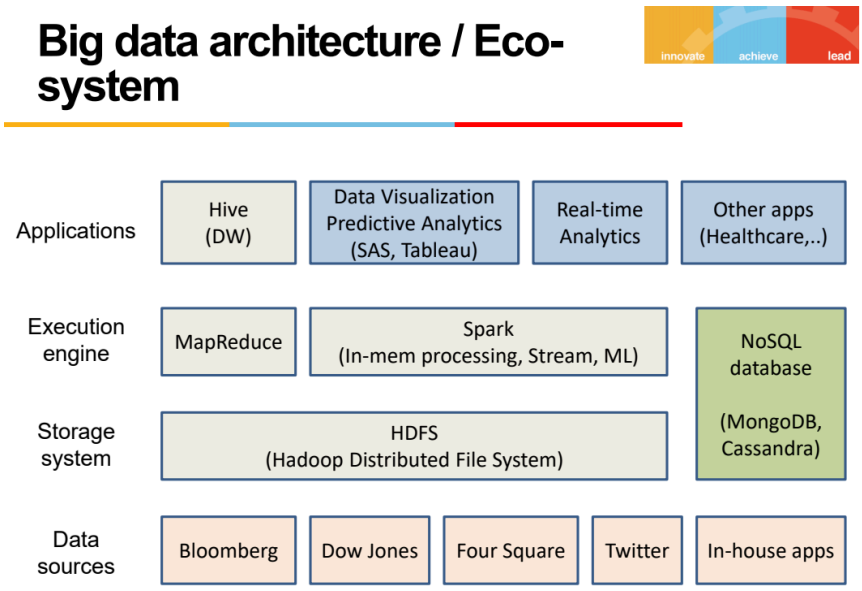
Uses of BigTable:

- Google services like **Google Maps** and **Google Analytics** rely on BigTable for storing and managing data.

Data Structure: BigTable stores data in a **wide table format** with various attributes, such as:

- **Content of the webpage**
- **Anchor text (text of hyperlinks)**
- **Websites referencing the page**
- **Timestamp for stored data**

Diagram Placeholder: Example of BigTable Structure



---

2. Hadoop, HDFS, and MapReduce

Hadoop: An open-source framework from Apache for distributed processing of large datasets.

- **HDFS (Hadoop Distributed File System)**: Distributes data across nodes, providing high availability and fault tolerance.
- **MapReduce**: A programming model for processing large datasets by distributing the workload across multiple servers.

Diagram Placeholder: Big Data Architecture

Hadoop - HDFS

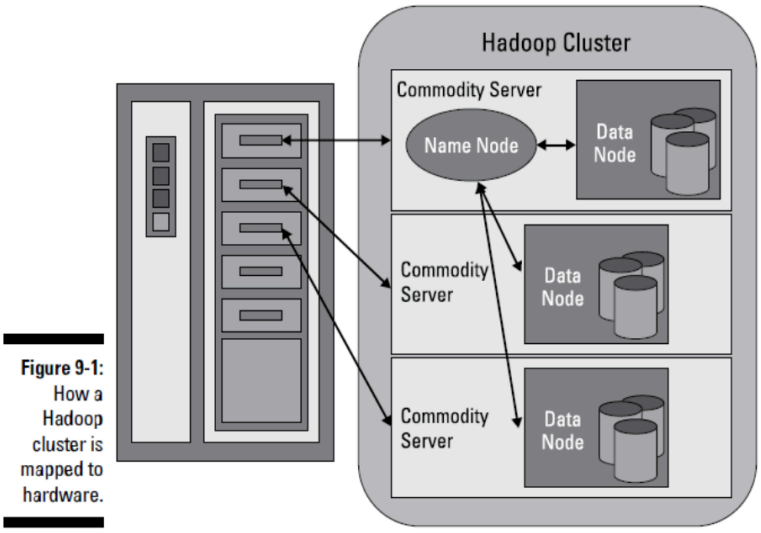


Figure 9-1: How a Hadoop cluster is mapped to hardware.

Ref: Big data for Dummies

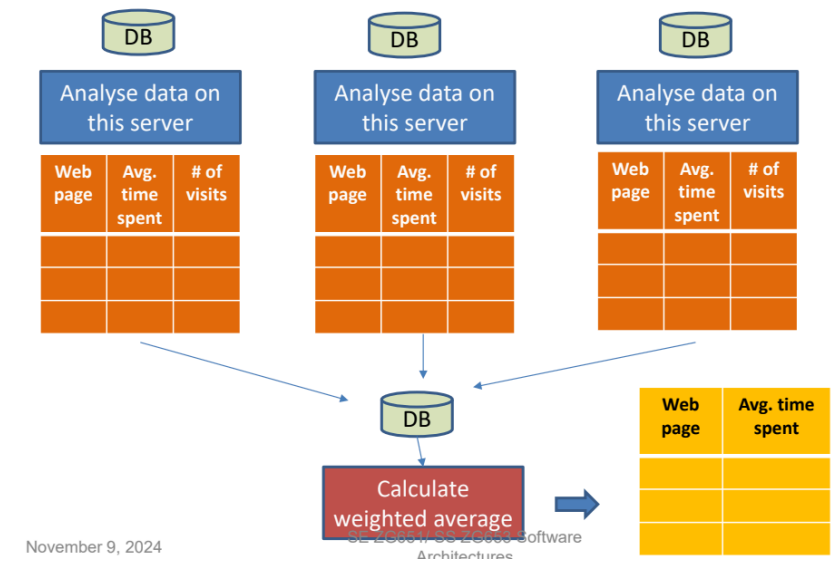
HDFS Features:

- Data split into blocks, distributed across nodes, and replicated for reliability.

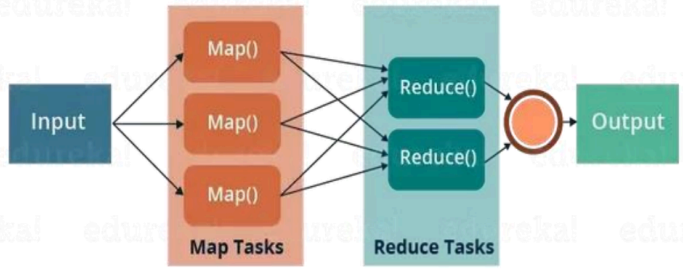
MapReduce Example: Calculates average time spent by users on each webpage by processing data in parallel.

Diagram Placeholder: MapReduce Example

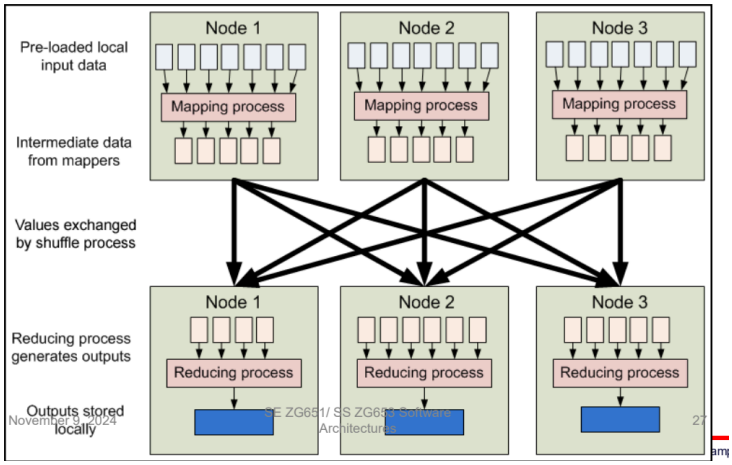
Map-Reduce pattern: Example



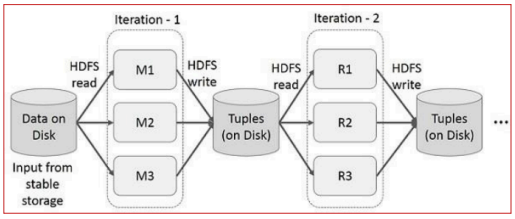
Map Reduce pattern



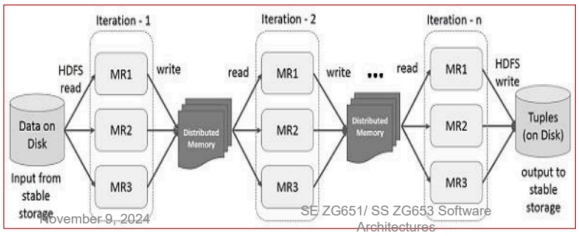
Example: Determine the average duration spent by users on different web pages



Difference between Hadoop & Spark



Iterative operation using Hadoop using disk



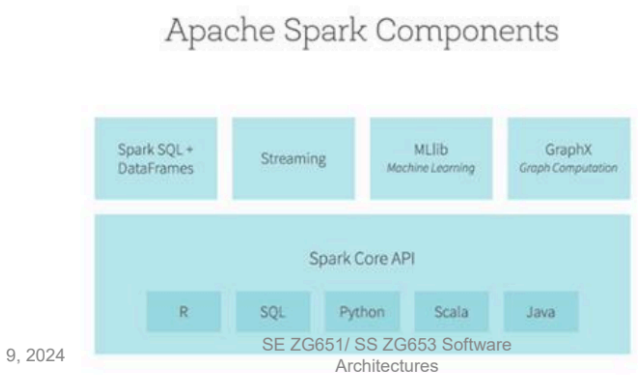
Iterative operation using Spark using memory

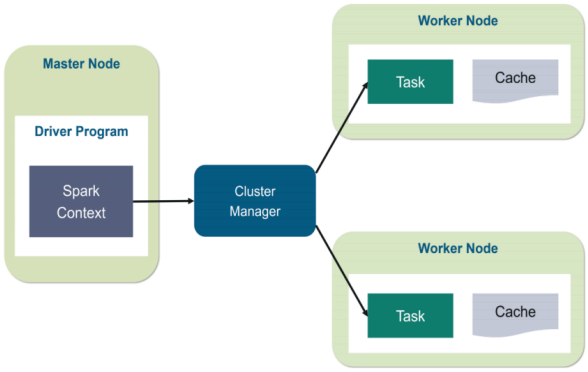
Source: https://www.tutorialspoint.com/apache_spark/apache_spark_rdd.htm

3. Apache Spark

Apache Spark: An open-source distributed computing engine for big data processing, enabling in-memory data processing, which is faster than disk-based processing in Hadoop.

Diagram Placeholder: Spark Architecture





---

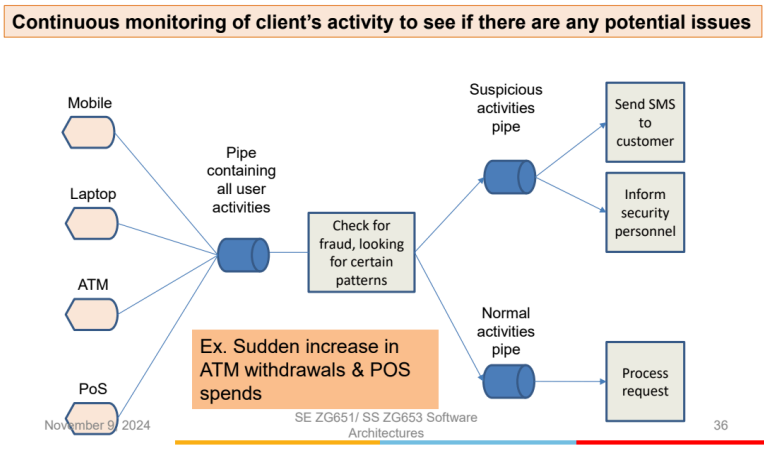
4. Real-Time Analytics

Real-Time Analytics allows users to gain insights from data as it arrives, enabling immediate actions.

Examples:

- **Real-Time Advertising**: Dynamic ad placements based on user behavior.
- **Fraud Detection**: Identifies suspicious activities in banking transactions.
- **Sensor Data Processing**: Predictive maintenance in industrial machines.

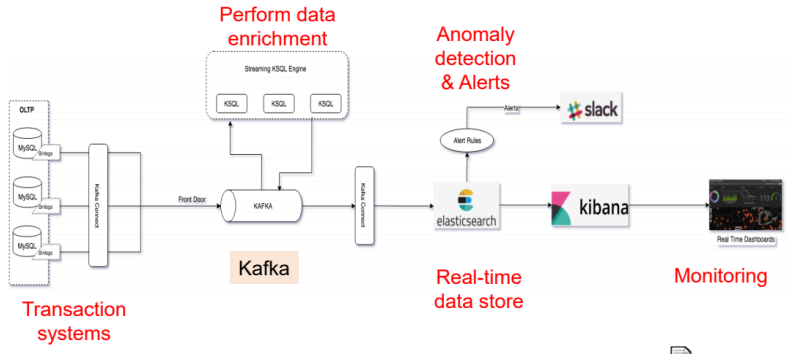
Diagram Placeholder: Real-Time Analytics - Fraud Detection



Case Study: Dream11

- Objectives: Monitor real-time contest participation, payment status, and detect unusual traffic.

Diagram Placeholder: Real-Time Analytics at Dream11



---

4. In-Memory Databases

In-Memory Databases: Store data in the main memory (RAM) instead of on disk, offering faster data access for applications that require real-time processing.

Examples:

- **SAP HANA**
- **IBM DB2 BLU**
- **Oracle In-Memory Database**

---

5. NoSQL Databases

NoSQL databases are designed for flexibility and scalability. Unlike SQL databases, they often lack support for transactions but are optimized for horizontal scaling.

Types of NoSQL Databases

1. **Document Databases**: Store data in document format (e.g., JSON), ideal for flexible and semi-structured data. Example: MongoDB.
2. **Key-Value Databases**: Store data as simple key-value pairs, which is ideal for caching. Example: Redis.
3. **Column-Oriented Databases**: Store data in columns, optimized for analytical tasks. Example: HBase.
4. **Graph Databases**: Represent data as nodes and edges, ideal for applications with interconnected data, such as social networks and fraud detection. Example: Neo4J.

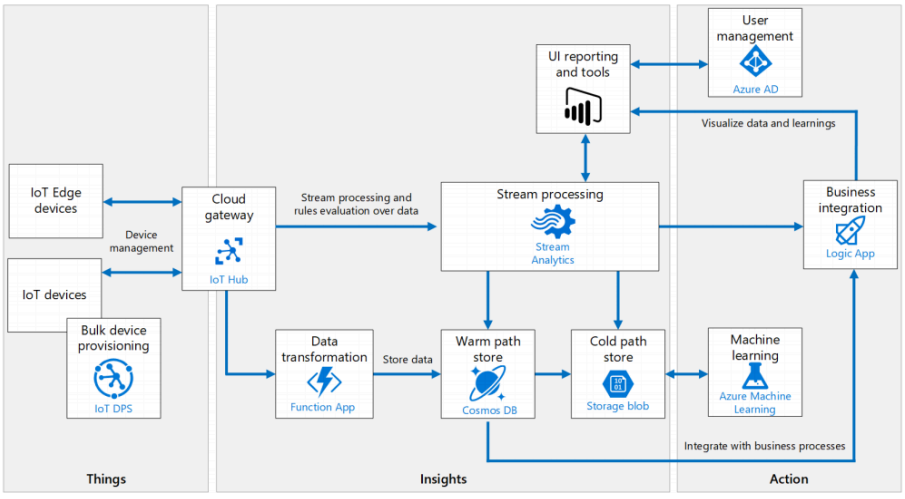
Internet of Things (IoT)

IoT: A network of interconnected devices that collect and share data with minimal human intervention, used in various applications.

Examples:

- **Smart Agriculture**: Soil moisture sensors control irrigation.
- **Supply Chain**: Real-time tracking of goods.

Diagram Placeholder: Azure IoT Reference Architecture



Examples of Big Data Use Cases

1. **Retail**:
    - **Fraud Detection**: Identifies suspicious transactions using patterns.
    - **Customer Targeting**: Customizes offers based on demographic and behavioral data.
2. **Banking and Financial Services**:
    - **Credit Risk**: Assess customer creditworthiness.
    - **Sentiment Analysis**: Analyzes customer feedback from social media.
3. **Industrial Equipment Monitoring**:
    - Monitors equipment for early fault detection.
4. **Weather Forecasting**:
    - Processes satellite data to predict weather patterns, including alerts for cyclones and heavy rainfall.

Summary

This document provides an overview of Big Data, Hadoop, Real-Time Analytics, In-Memory and NoSQL Databases, Google BigTable, and IoT applications. It introduces the characteristics, applications, and technologies enabling Big Data processing, and highlights various use cases across industries.