Hadoop Cloud

Abirami M - 22z204

Adish Kumar S - 22z206

Aravinth Cheran K S - 22z212

Kabhinyasri S V - 22z229

Madhisha S - 22z235

Karthikeyan S - 22z256

Introduction to Hadoop & Cloud Computing

ADISH KUMAR S
22Z206

BIG DATA:

To handle extremely large & complex datasets that are difficult to processing. It is characterized by 5V's,

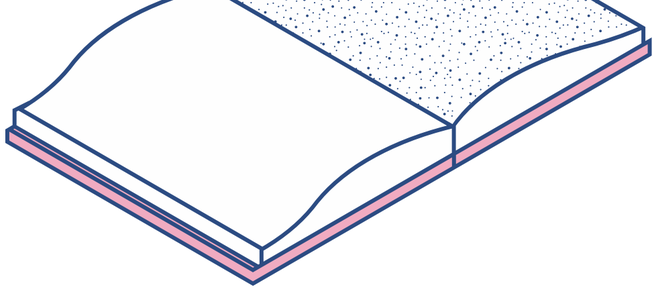
- Volume
- Velocity
- Variety
- Veracity
- Value

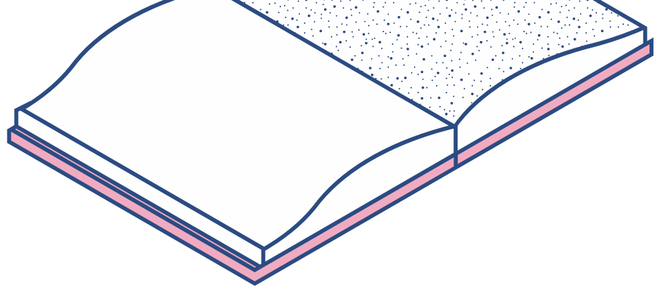
HADOOP:

Open-source framework for storing and processing Big Data in a distributed environment.

Key Features:

- **Scalable**
- **Fault - tolerant**
- **Cost-effective**





Why to Hadoop in cloud?

- Scalability: Easily scales up or down
- Cost-effectiveness: Pay-as-you-go model
- High availability and fault tolerance
- Managed services simplify deployment

Cloud providers that supports Hadoop:

- AWS EMR (Elastic MapReduce)
- Google Cloud Dataproc

Hadoop Architecture and Components

Madhisha S
(22z235)

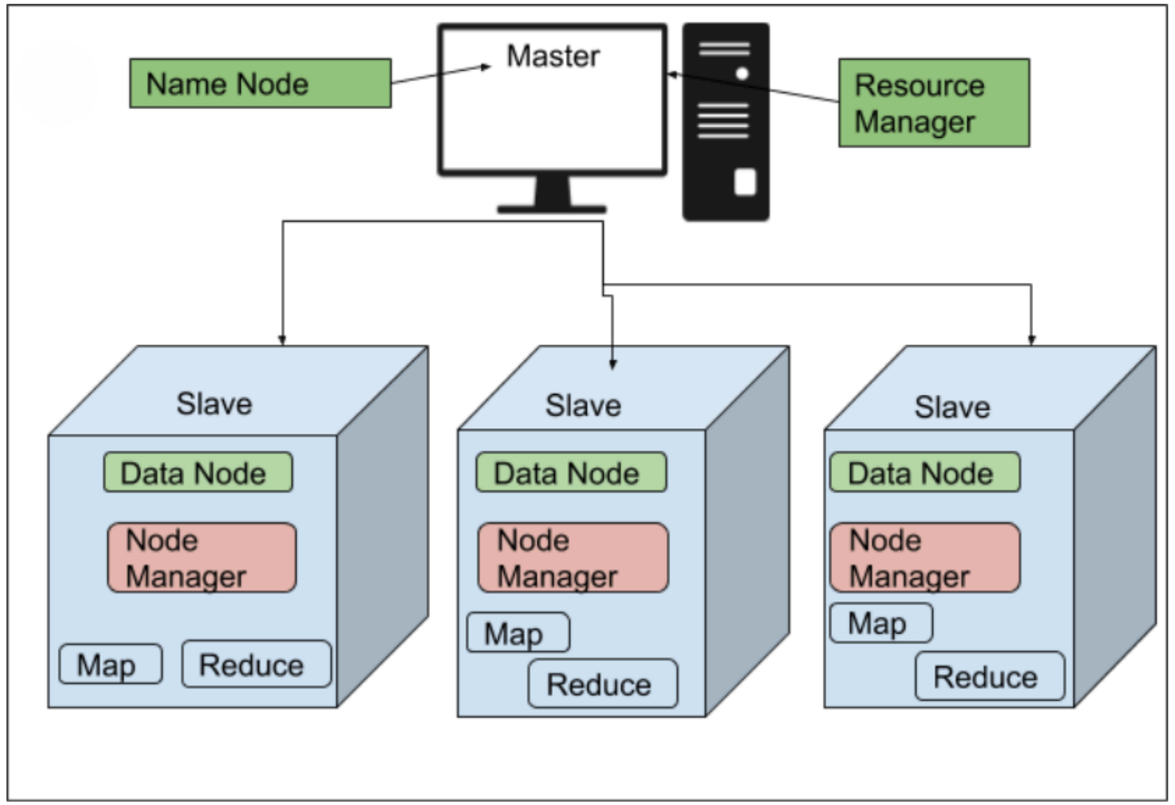
Hadoop Architecture

Hadoop follows a **distributed computing** model designed for handling **large-scale** data processing efficiently. It has a **master-slave** architecture and consists of several key components that enable reliable storage and parallel processing of big data.

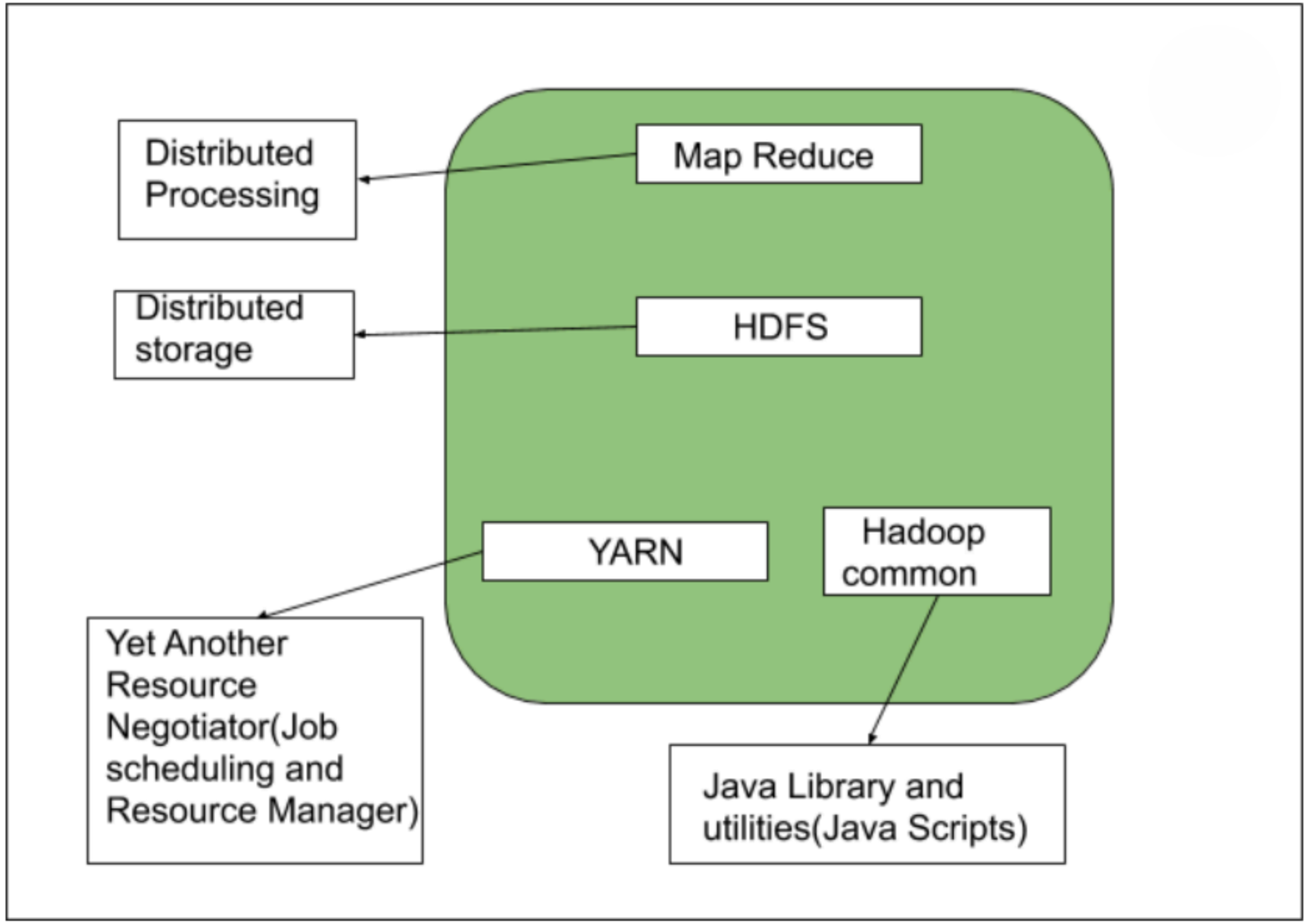
Hadoop Architecture

- **NameNode** (Master Node) – Manages the metadata (file system namespace, location of blocks).

- **DataNodes** (Slave Nodes) – Store actual data in the form of blocks and respond to read/write requests.

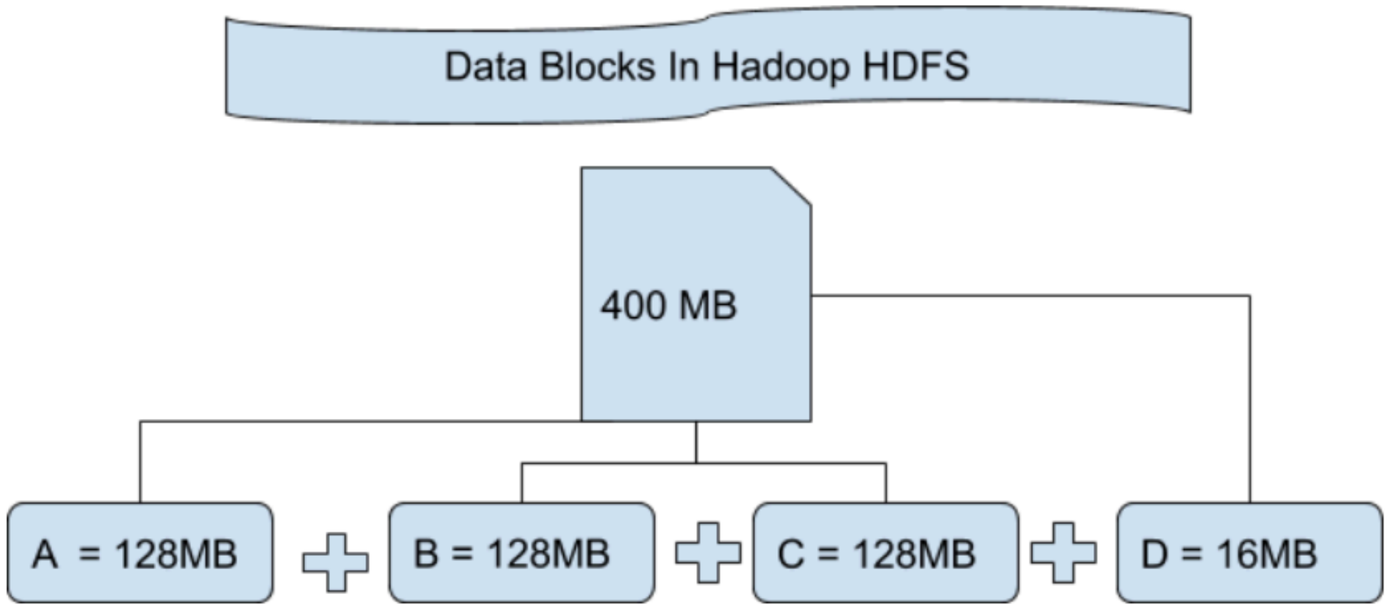


Four main components of Hadoop



1. HDFS (Hadoop Distributed File System)

HDFS is the **storage layer** of Hadoop. It is designed to store large datasets across multiple machines in a **distributed manner.** It contains the **NameNode** and **DataNode.**



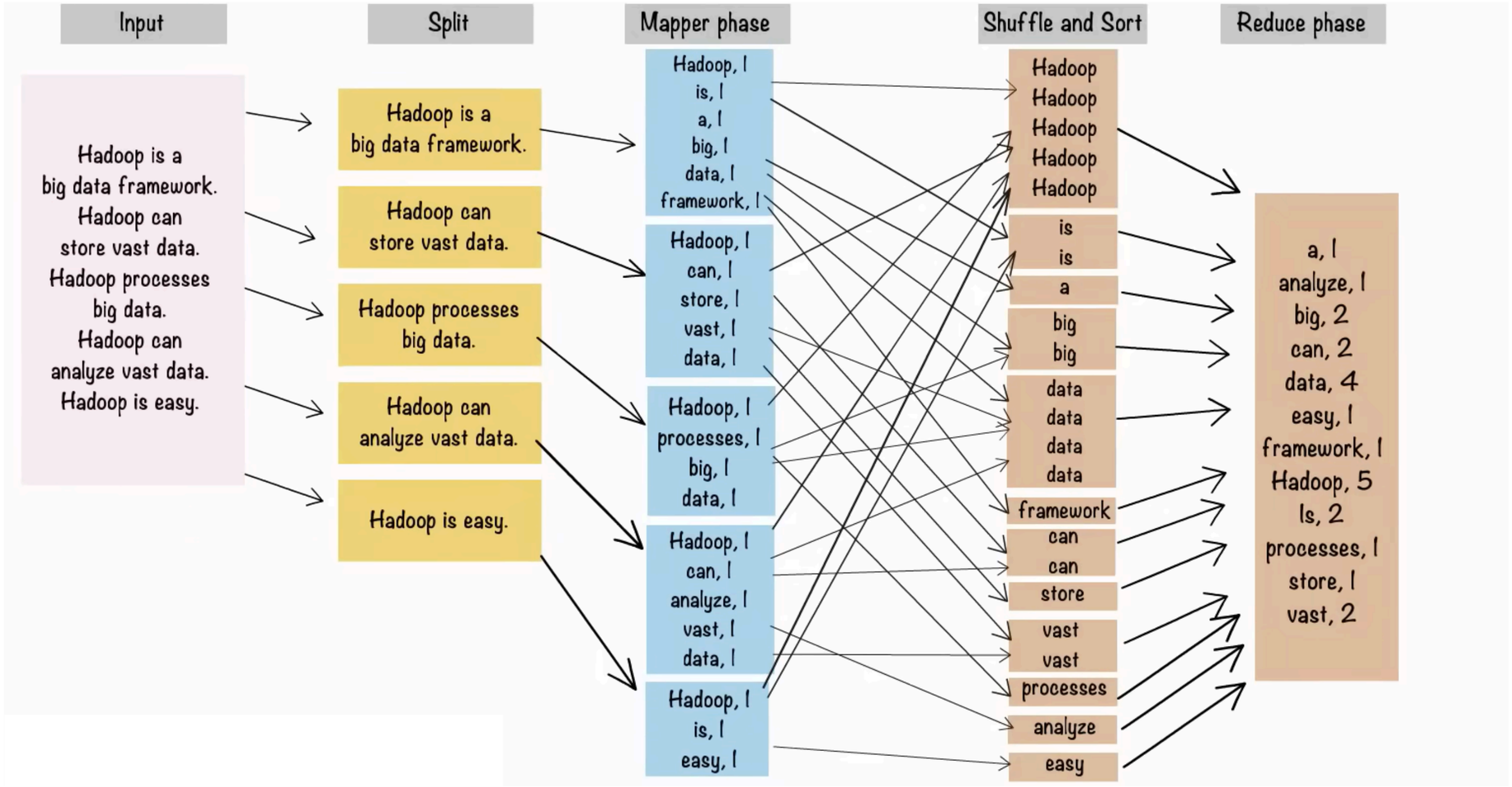
2. MapReduce (Processing Layer)

MapReduce is a **framework** conducting distributed and parallel processing of large volumes of data. It is nothing but just like an Algorithm or a data structure.

How MapReduce Works:

- Map Phase – Splits input data into small chunks and processes them in parallel.
- Shuffle & Sort Phase – Groups and organizes the intermediate data for processing.
- Reduce Phase – Aggregates the processed data and produces the final output.

2. MapReduce (Processing Layer)



3. YARN (Yet Another Resource Negotiator)

YARN is the **resource management layer** of Hadoop. YARN performs 2 operations that are Job scheduling and Resource Management. It helps in efficiently **allocating resources** for running multiple jobs.

YARN Components:

- ResourceManager (Master Node) – Allocates resources for running applications.
- NodeManagers (Slave Nodes) – Monitors resource usage and manages containers on each node.
- ApplicationMaster – Handles execution of individual applications on YARN.
- Container - Container houses a collection of resources like RAM, CPU, and network badnwidth.

4. Common Utilities

Hadoop Common is a crucial part of the Hadoop ecosystem, providing essential utilities and libraries that support the core components of Hadoop. It includes Java libraries and files necessary for the functioning of HDFS, YARN, and MapReduce.

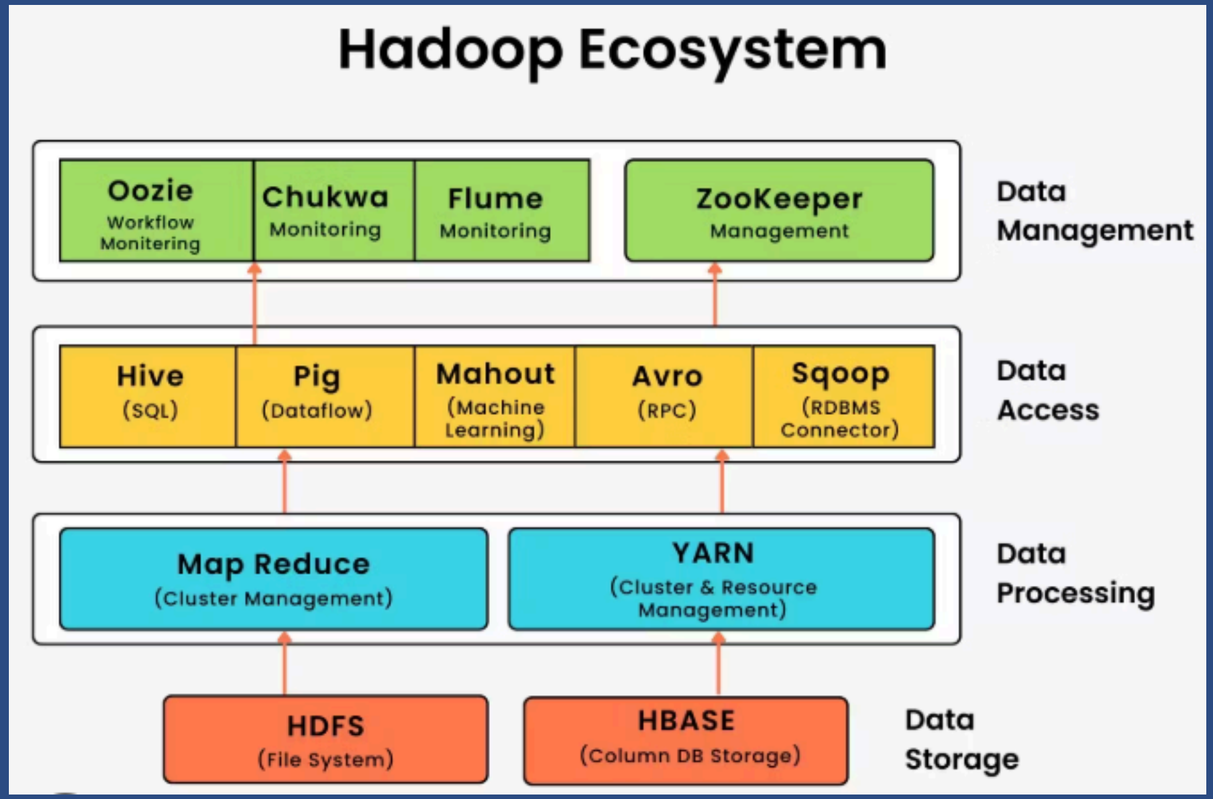
Hadoop Ecosystem
&
Tools

KABHINYASRI S V
22Z229

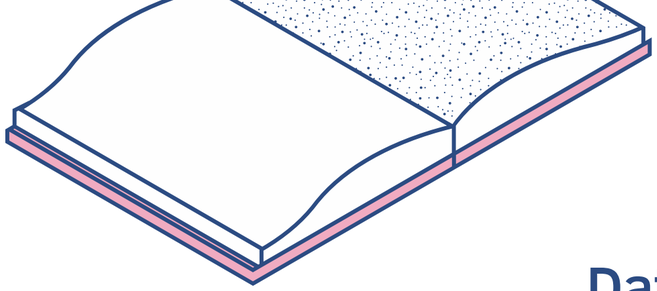
HADOOP ECOSYSTEM

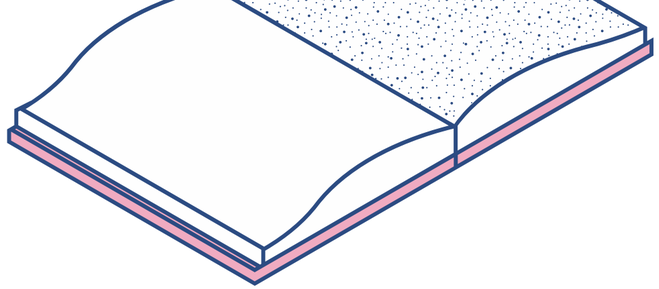
Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems.

- data storage
- processing
- querying
- security
- machine learning
- visualization
- ingestion
- workflow automation



The four major elements of Hadoop : HDFS, MapReduce, YARN, and Hadoop Common Utilities.





Data Storage:

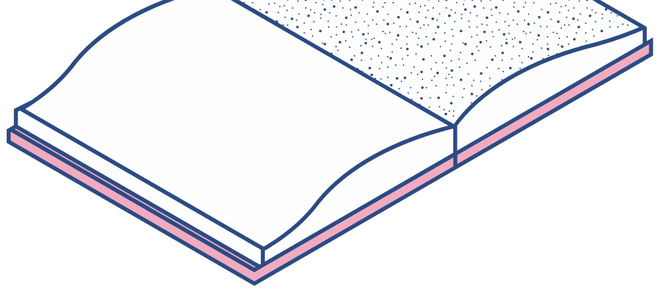
- HDFS (Hadoop Distributed File System)
- HBase (Hadoop dataBase)

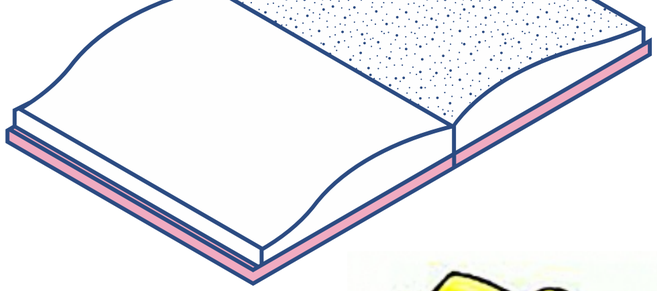
  A NoSQL database for real-time read/write operations.

  (Facebook)

Data Processing:

- MapReduce
- YARN (Yet Another Resource Negotiator)





Data Querying & Access :

- Hive: SQL-based querying (HiveQL) for Hadoop.(Facebook)
- Pig: High-level scripting language (Pig Latin) for data transformation. (Twitter)

Data Ingestion & Integration:

- Sqoop: Transfers data between Hadoop and relational databases.
- Flume: Collects, aggregates, and moves large-scale streaming data (logs, IoT data) into HDFS. (LinkedIn)

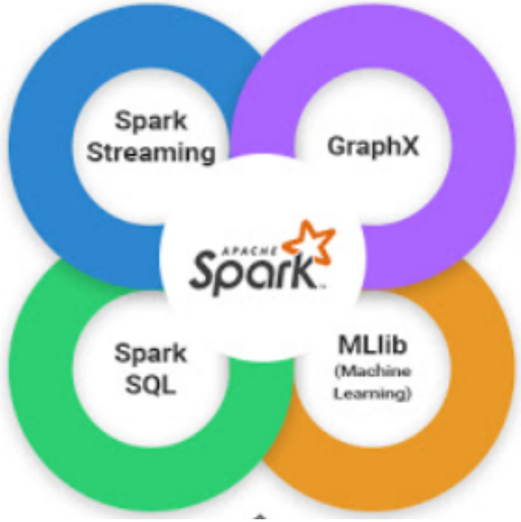
Data Management & Monitoring:

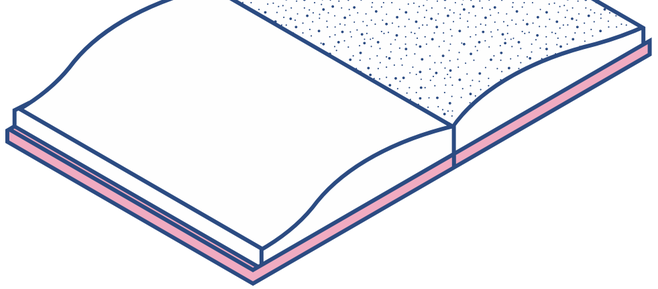
- Oozie: Workflow scheduler for automating Hadoop jobs.
- ZooKeeper: Ensures synchronization and coordination in distributed systems.(Yahoo,LinkedIn)

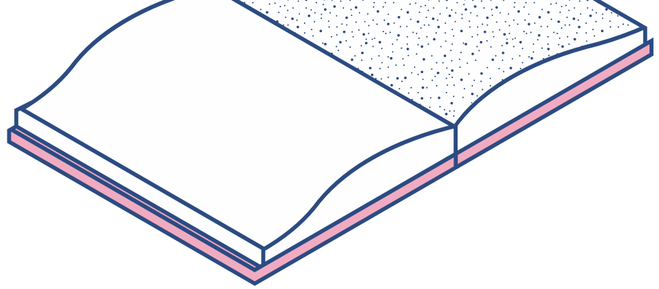
Apache Spark

- Faster alternative to MapReduce using in-memory computing.
- Supports batch processing, real-time analytics, and machine learning.

(Netflix, Uber, and Airbnb)







Why Use the Hadoop Ecosystem?

Advantages:

- Scalability
- Cost-Effective
- Fault Tolerance
- Flexibility
- Integration with ML & AI

Use Cases:

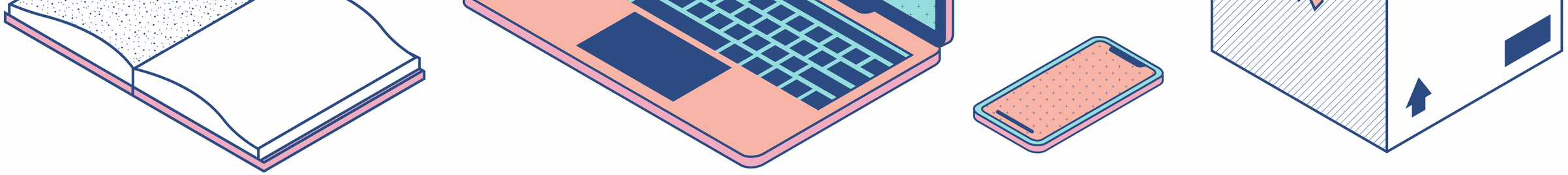
- Social Media (Facebook, Twitter) - user data
- E-Commerce (Amazon, Flipkart) - recommendation
- Banking & Finance - fraud detection
- Healthcare & Genomics - medical records
- IoT & Smart Cities - energy management

Hadoop Deployment in Cloud

Why and How Hadoop is Deployed in the Cloud?

ARAVINTH CHERAN K S

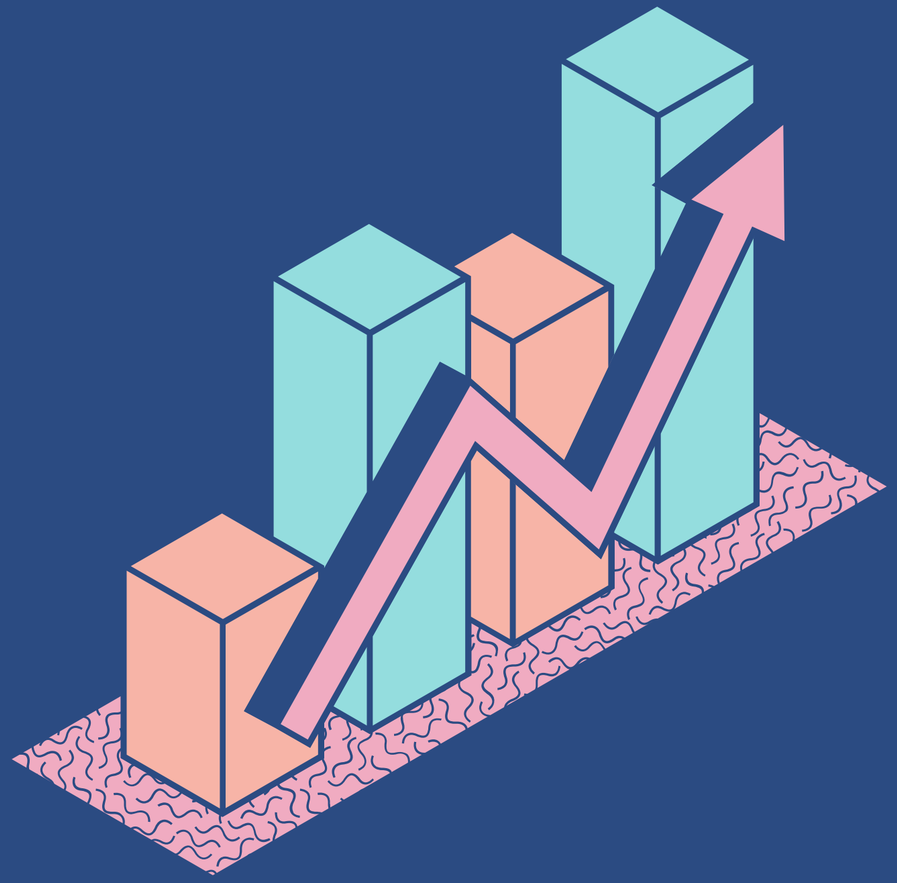
22Z212



Why Deploy Hadoop in the Cloud?

- Scalability: Cloud allows dynamic resource allocation based on demand.
- Cost-Effectiveness: Pay-as-you-go pricing avoids high infrastructure costs.
- Flexibility: No need for on-premise hardware; easily integrates with cloud services
- Disaster Recovery: Built-in backup and replication for high availability.

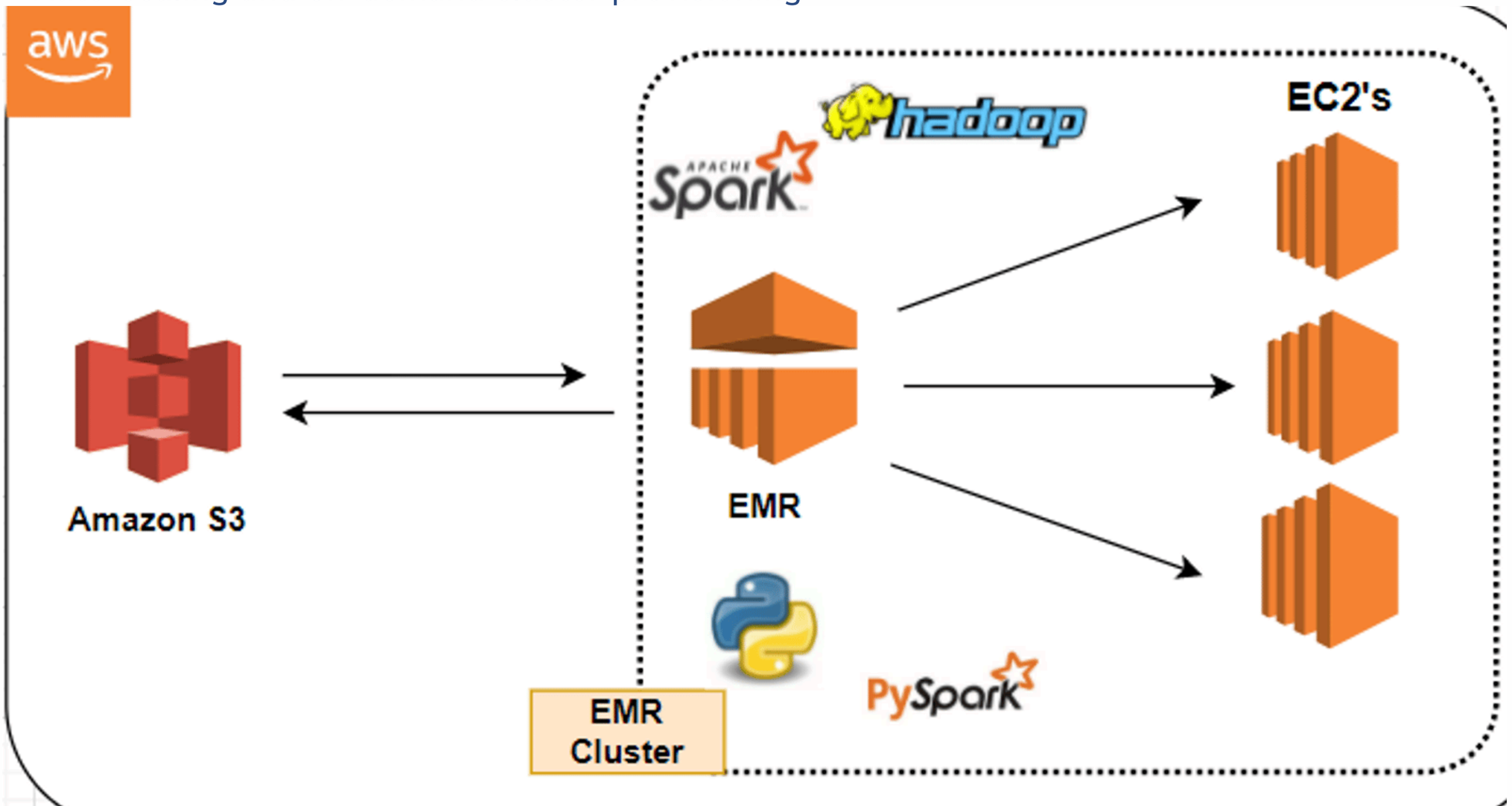
Cloud Service Providers for Hadoop



AWS EMR (Elastic MapReduce)

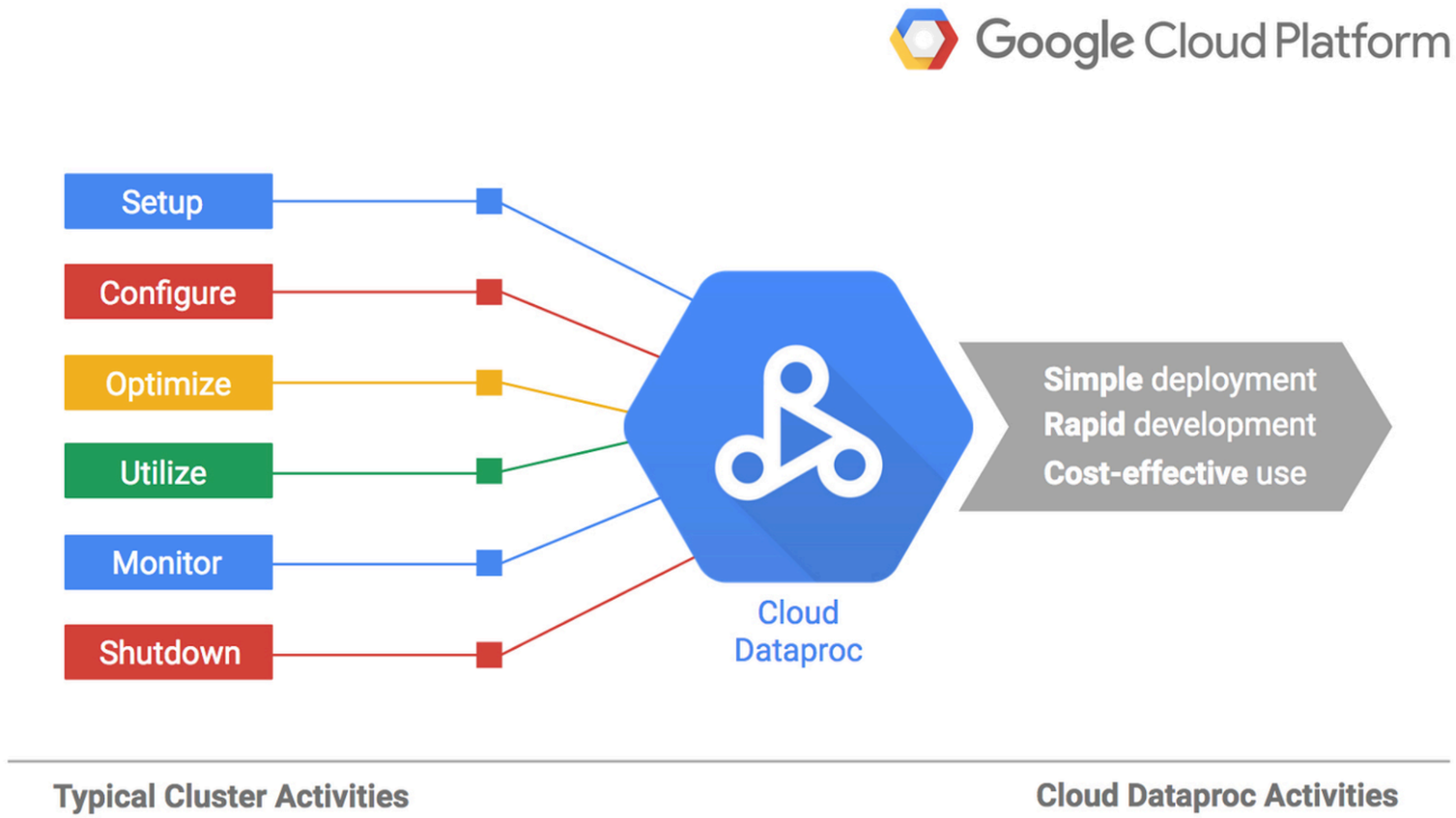
- Fully managed Hadoop service

- Auto-scaling and on-demand cluster provisioning



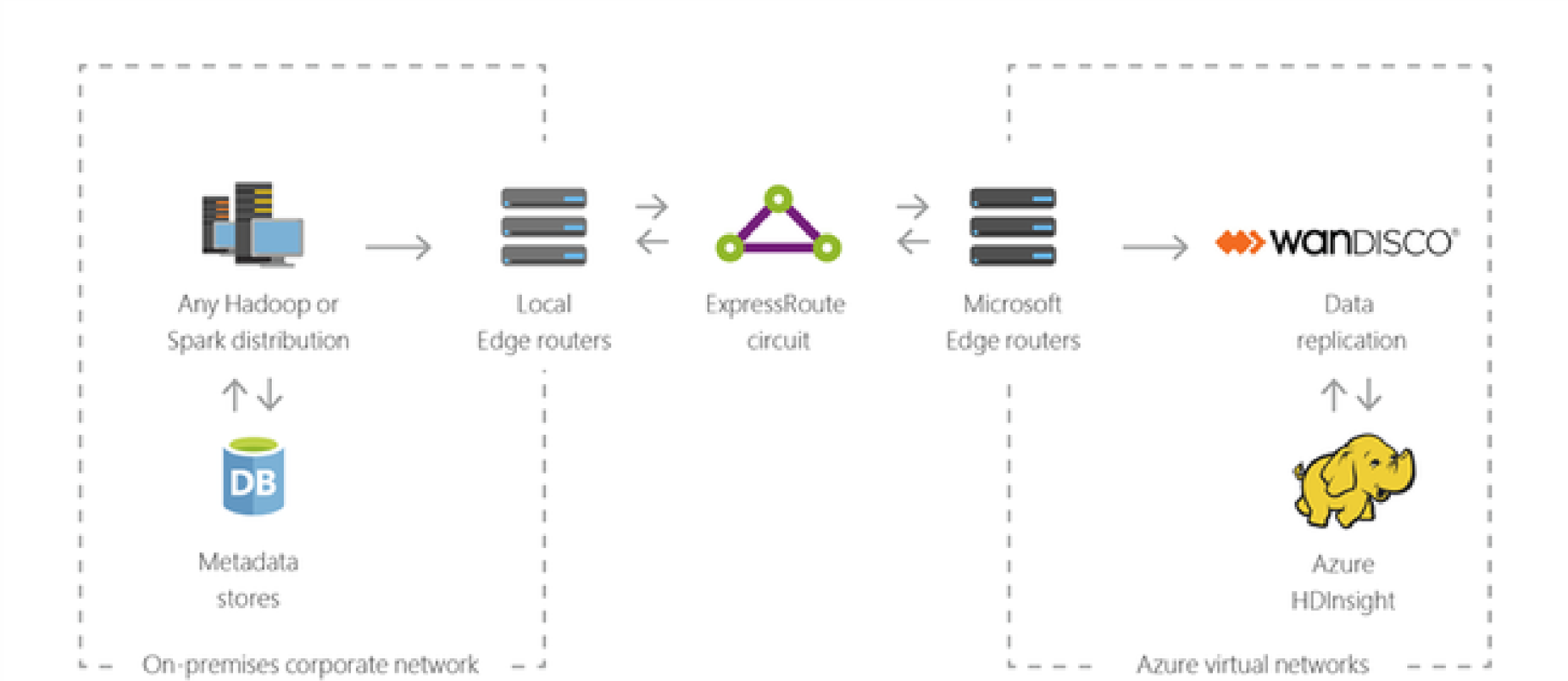
Google Cloud Dataproc

- Optimized for fast job execution
- Integration with Google Cloud Storage and BigQuery



Azure HDInsight

- Supports Hadoop, Spark, and Kafka
- Enterprise-grade security and compliance



How Hadoop is Deployed in the Cloud?

Provisioning Cloud Resources

Setting up virtual machines, storage, and networking

Configuring Hadoop

Installing Hadoop ecosystem tools (HDFS, YARN, Hive, etc.).

Uploading & Processing Data

Data is stored in cloud storage (S3, GCS, or Azure Blob) and processed using MapReduce/Spark

Monitoring & Optimization

Auto-scaling and performance tuning for efficiency

Applications and Real World Use Cases

Karthikeyan Sivarasu

22z256

Hadoop In Cloud

- Hadoop is an open-source framework for processing large datasets.
- When used in the cloud, it provides scalability, flexibility, and cost efficiency.
- Cloud-based Hadoop eliminates the need for expensive hardware.
- Hadoop helps in managing and analyzing vast amounts of data. Running Hadoop on the cloud makes it easier and cheaper because you don’t need to buy and maintain servers.

Key Benefits of Hadoop in the Cloud

✔ Scalability – Easily add more storage and computing power.
✔ Flexibility – Works with structured & unstructured data.
✔ Cost-Effective – Pay only for what you use.
✔ Performance – Fast processing of huge datasets.

- Hadoop in the cloud allows companies to grow without worrying about storage limits. They can process all types of data, save money by paying only for used resources, and analyze data quickly.

Industries Using Hadoop in the Cloud

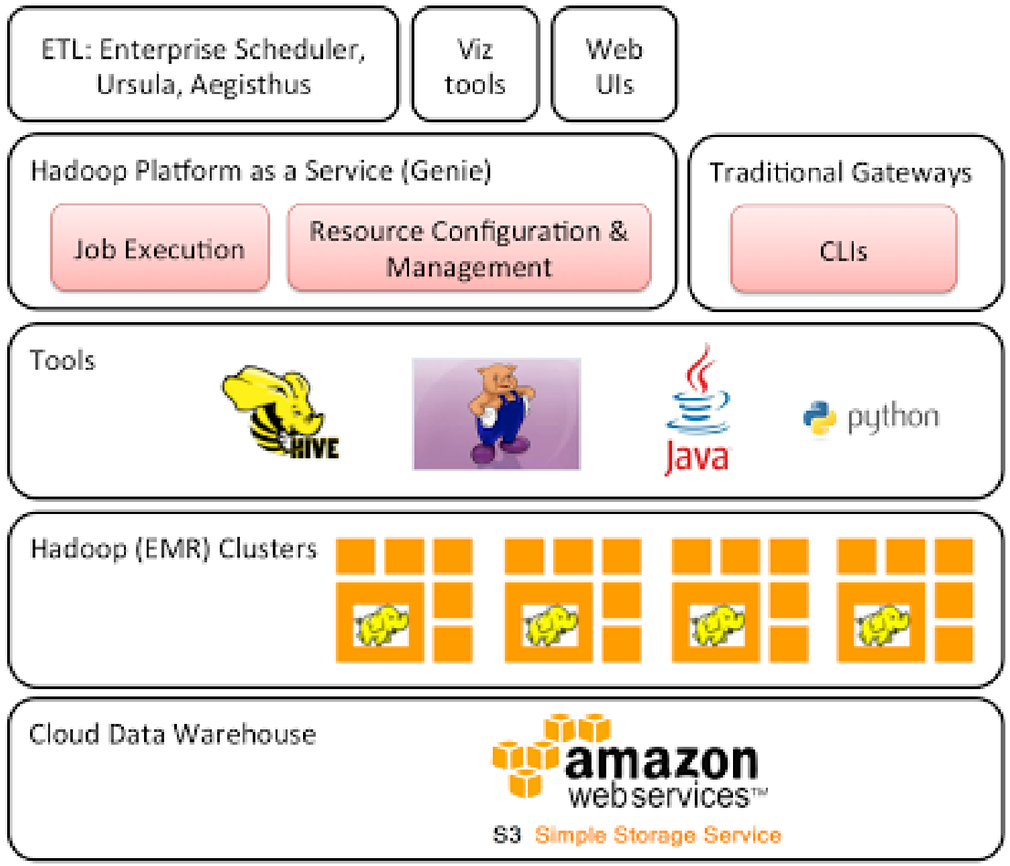
💰 E-commerce – Customer recommendations & fraud detection.
🏥 Healthcare – Patient data analysis & disease prediction.
🏦 Finance – Risk management & transaction monitoring.
🚗 Automobile – Connected car analytics & traffic prediction.
📺 Entertainment – Streaming services (Netflix, Spotify) for content suggestions.

- Many industries use Hadoop in the cloud. For example, e-commerce companies suggest products based on customer behavior, while hospitals use it to analyze patient data. Banks monitor transactions for fraud, and streaming platforms recommend movies based on your interests.

Real-World Case Study - Netflix

- Challenge: Storing & analyzing massive streaming data.
- Solution: Netflix uses Hadoop on AWS to process user preferences.
- Result: Personalized recommendations & seamless streaming.

- Netflix has millions of users streaming videos. It collects data on what people watch and suggests movies based on this data. With Hadoop in the cloud, Netflix can process this huge amount of information quickly.



Cloud-Based Deployment

- Hadoop on AWS (Amazon Web Services)
- Amazon S3 replaces HDFS for cloud storage

Key Hadoop Components Used

- Apache Hive – Querying structured data
- Apache Spark – Fast in-memory processing
- Apache Pig – Large-scale data transformations

Uses Apache Kafka + Hadoop for real-time data streaming analysis.

Cloud-Based Deployment

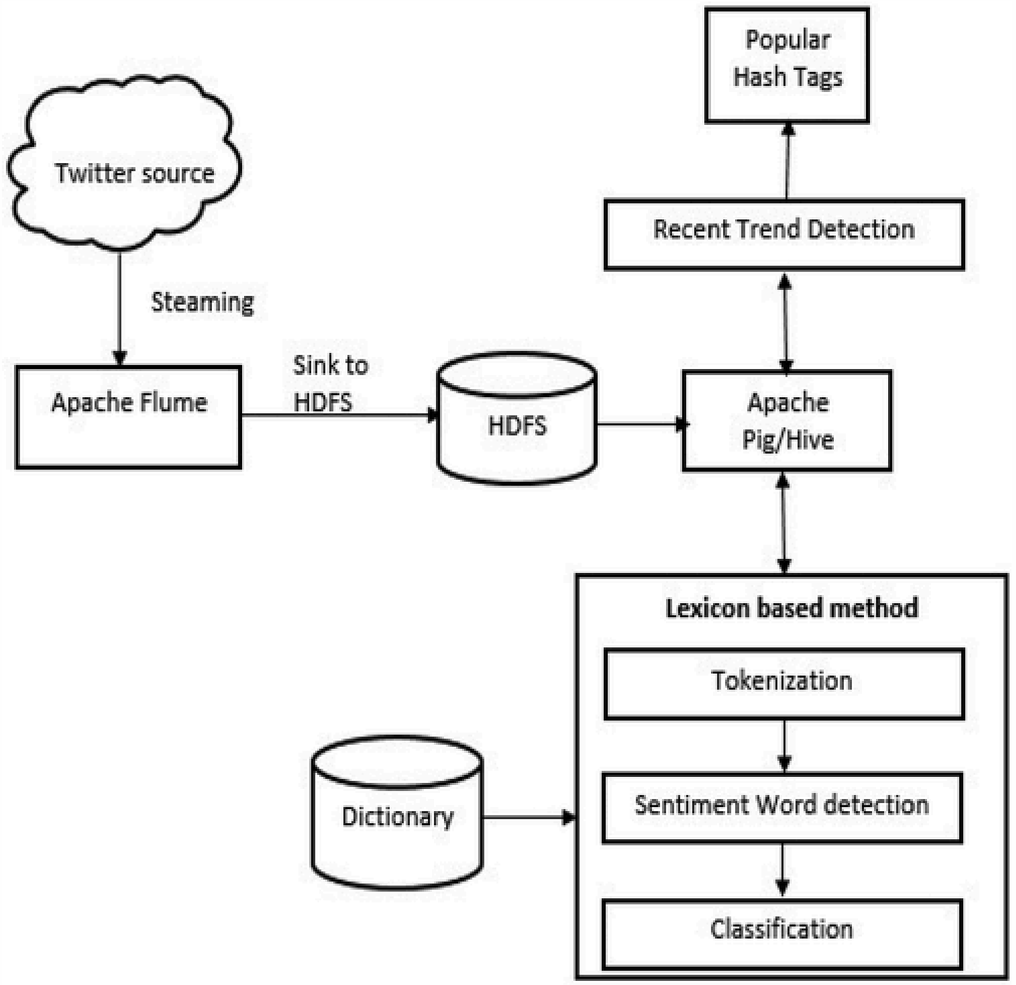
- Hadoop on AWS & On-Premises Hybrid Model
- Amazon S3 + HDFS for storage

Key Hadoop Components Used

- Apache Storm – Real-time tweet processing
- Apache HBase – Stores tweet metadata for fast retrieval
- Apache Pig – Processes large-scale data transformations
- Apache Spark – Analyzes and recommends trending top

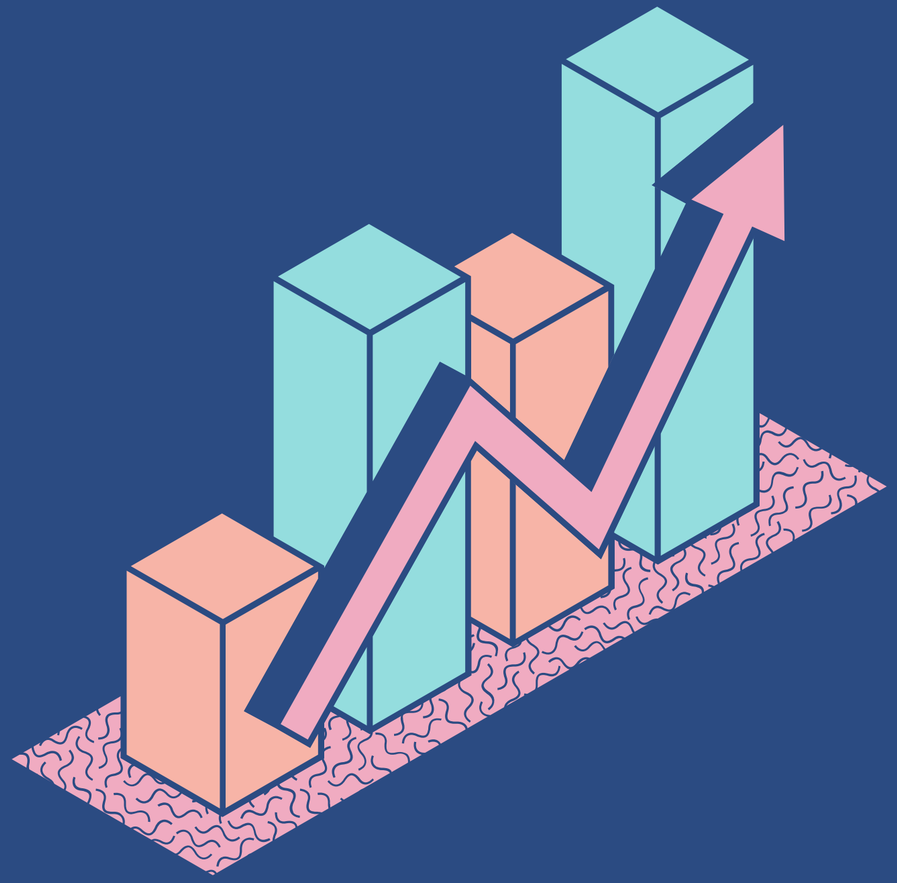
- Implemented Hadoop to process and store vast amounts of tweet data.
- Utilized Hadoop's ecosystem tools for text analysis and data mining.

- Enhanced ability to monitor trends and user sentiments in real time.
- Improved targeting for advertisements and content recommendations.

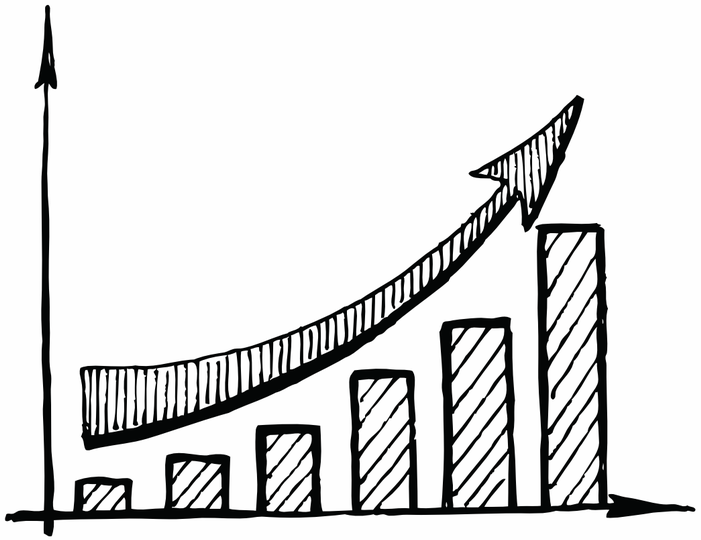


Challenges and Future trends

-22z204



Scalability

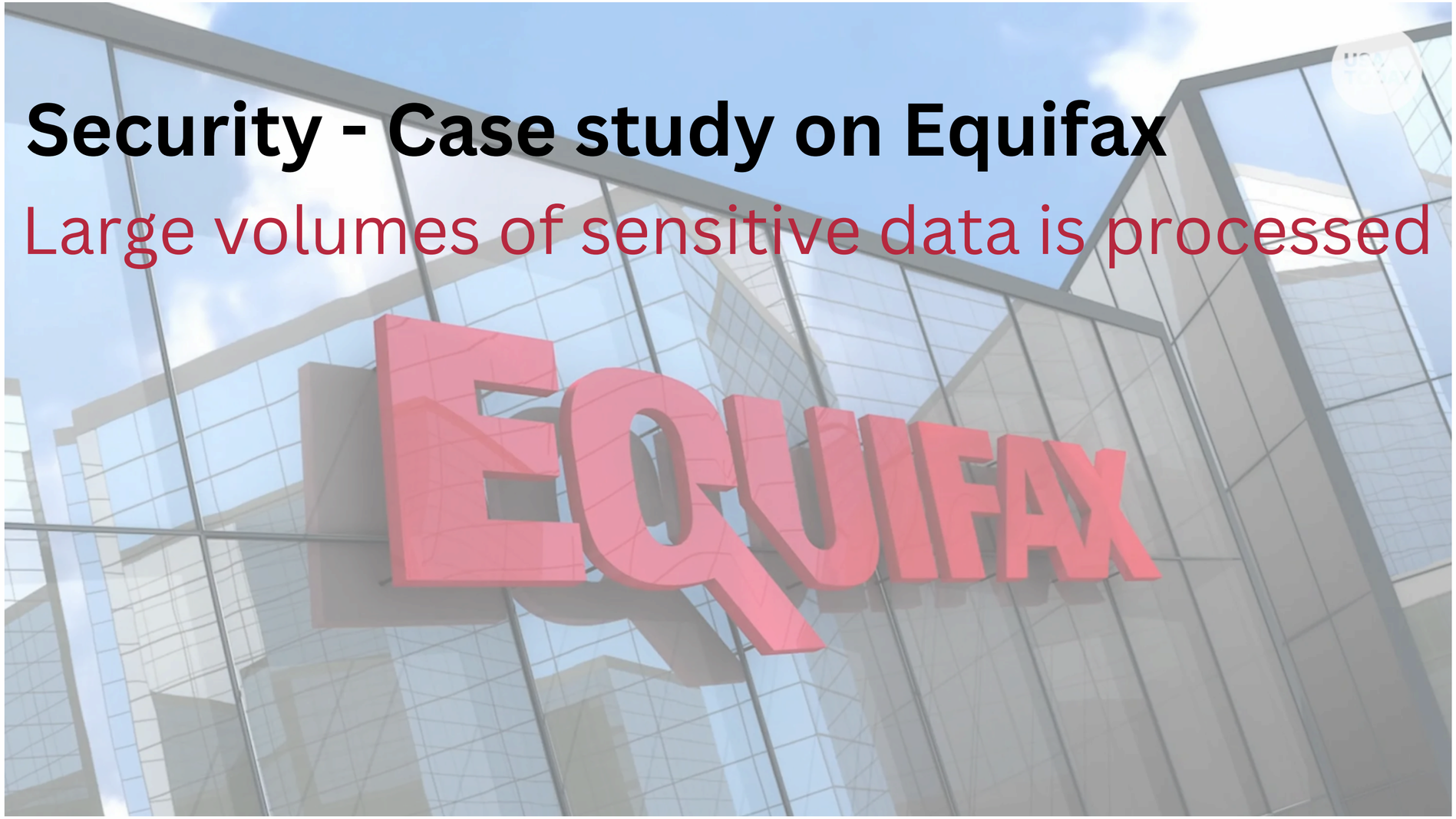


Challenges in

- Autoscaling
- Resource Allocation
- Performance Bottlenecks

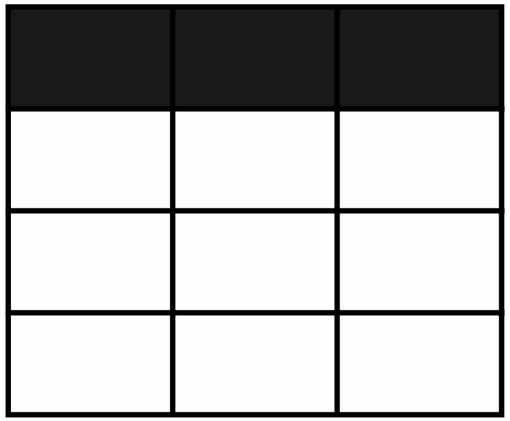
Security - Case study on Equifax

Large volumes of sensitive data is processed



EX-CEO says that the data was not encrypted

Data Integration



- Handling Structured, Unstructured and Semi structured data together is difficult
- Solution: Use Polyglot Persistence

Hybrid Architectures

- Hybrid Architecture is combination of Hadoop with other technologies like data warehouses

- Challenges:
  Data movement costs
  Security

Future Trends

Hybrid Data mesh
Multi-Cloud Strategy

Emerging field : Federated Learning on Hadoop

- Hadoop’s HDFS (Hadoop Distributed File System) stores data across multiple nodes, making it ideal for federated learning since data can remain in its original location while model training happens in a distributed fashion.

- Banks have customer transaction data stored across multiple branches and locations.

- Instead of moving sensitive customer data, only model weight updates are shared and aggregated to improve fraud detection accuracy globally.

Thankyou