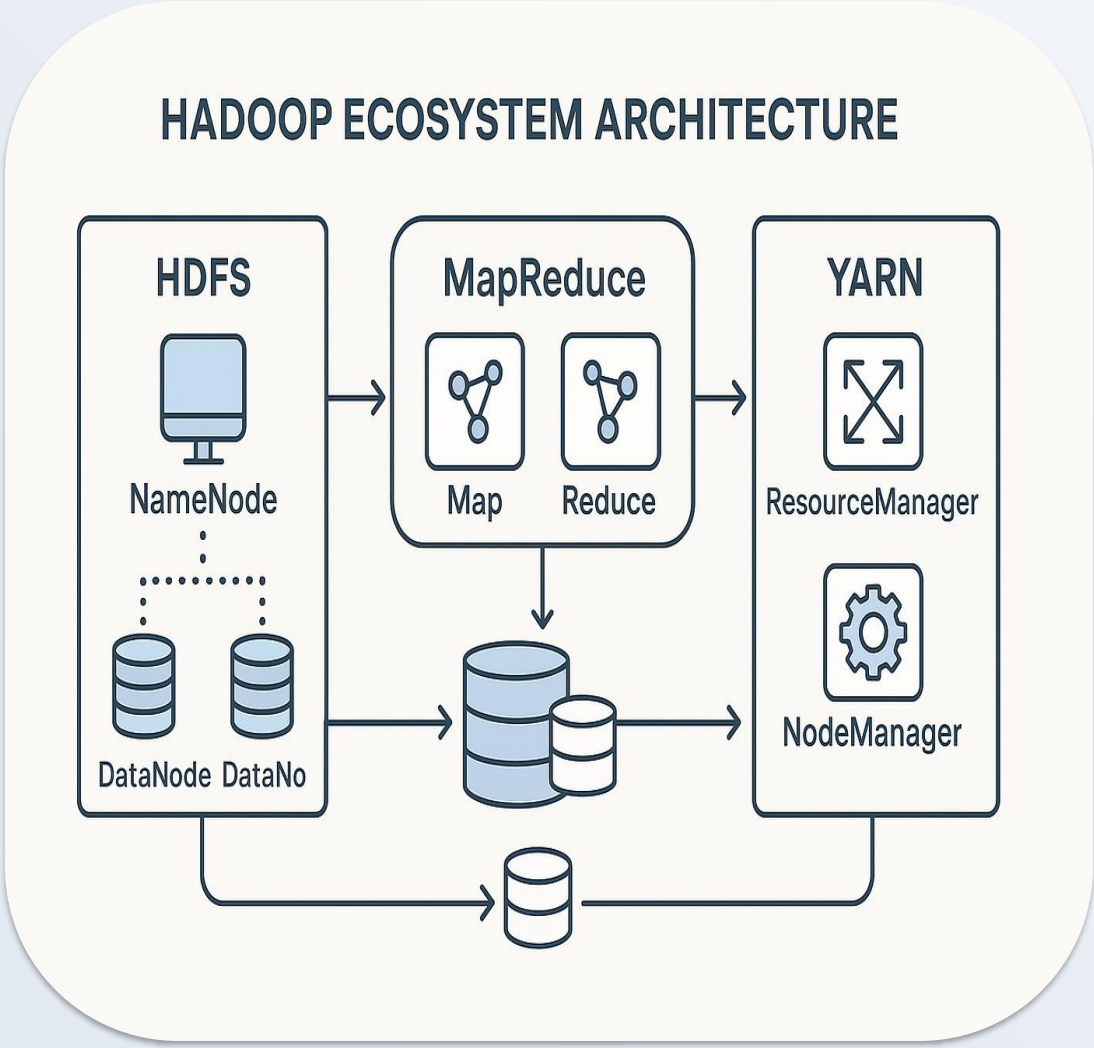
Managing Hadoop File-Automation Data in Hadoop

23MCT305 - Data Analytics in Automation System

Introduction to Hadoop



What is Hadoop?

Distributed framework for processing big data

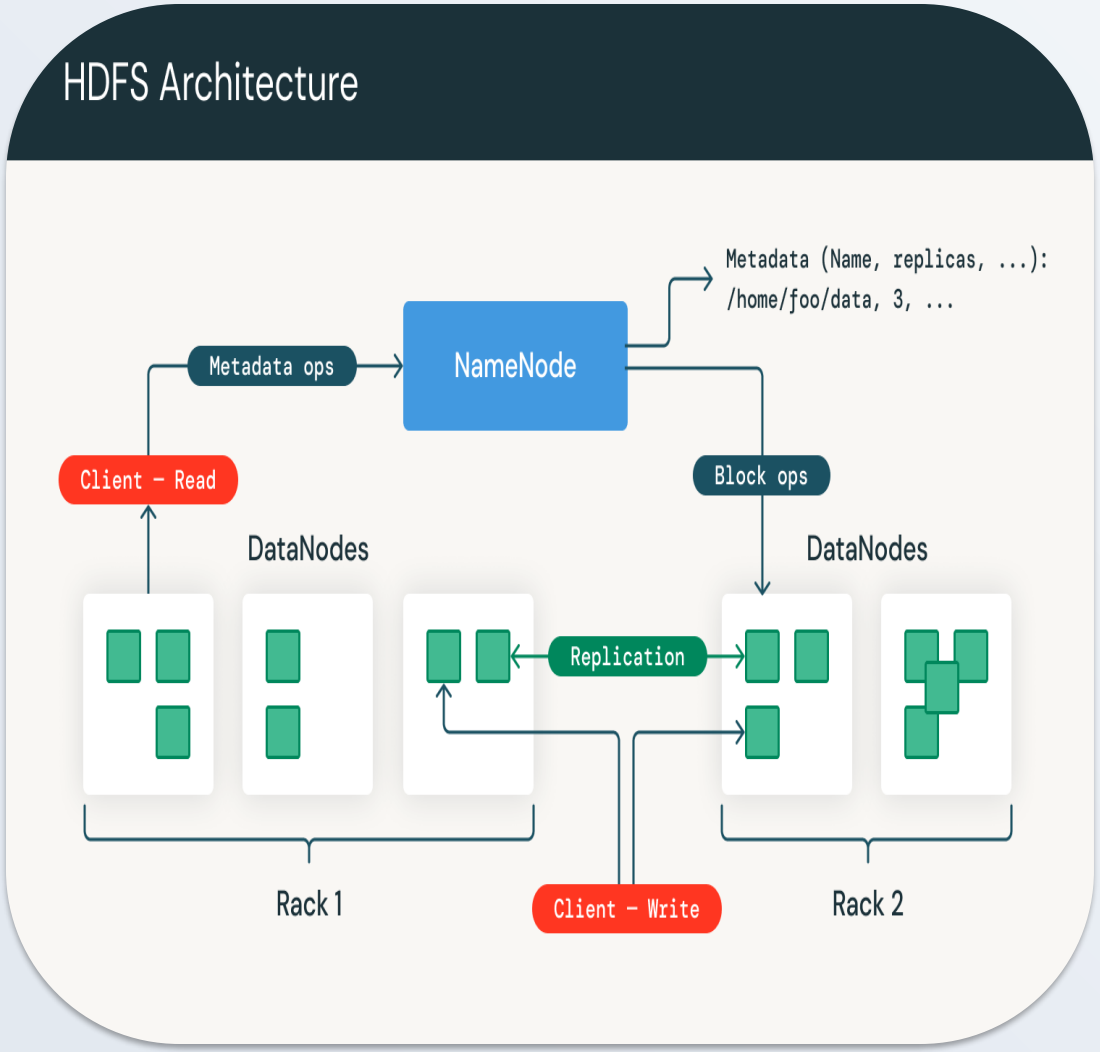
Core Components

- HDFS - Storage System
- MapReduce - Processing
- YARN - Resource Management

Key Features

- Scalable architecture
- Fault-tolerant design
- Cost-effective solution

HDFS Overview



HDFS Architecture

Master-Slave Distributed System

NameNode

Manages file system metadata & namespace

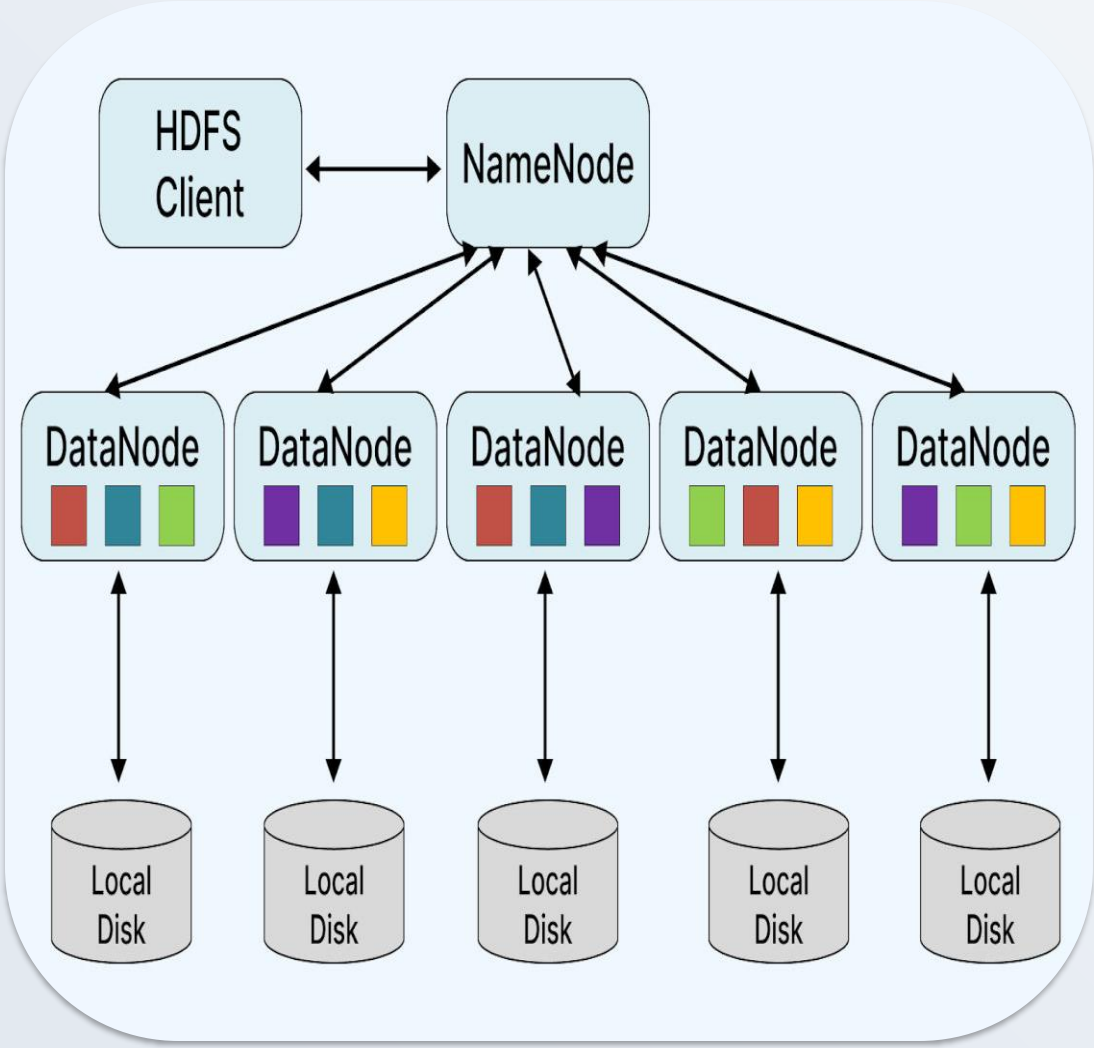
DataNode

Stores actual data blocks on local disk

Data Block Storage

Default block size: **128 MB**

Data Storage in Hadoop



Distributed Storage

Data split across multiple nodes

Data Replication

3 copies by default for fault tolerance

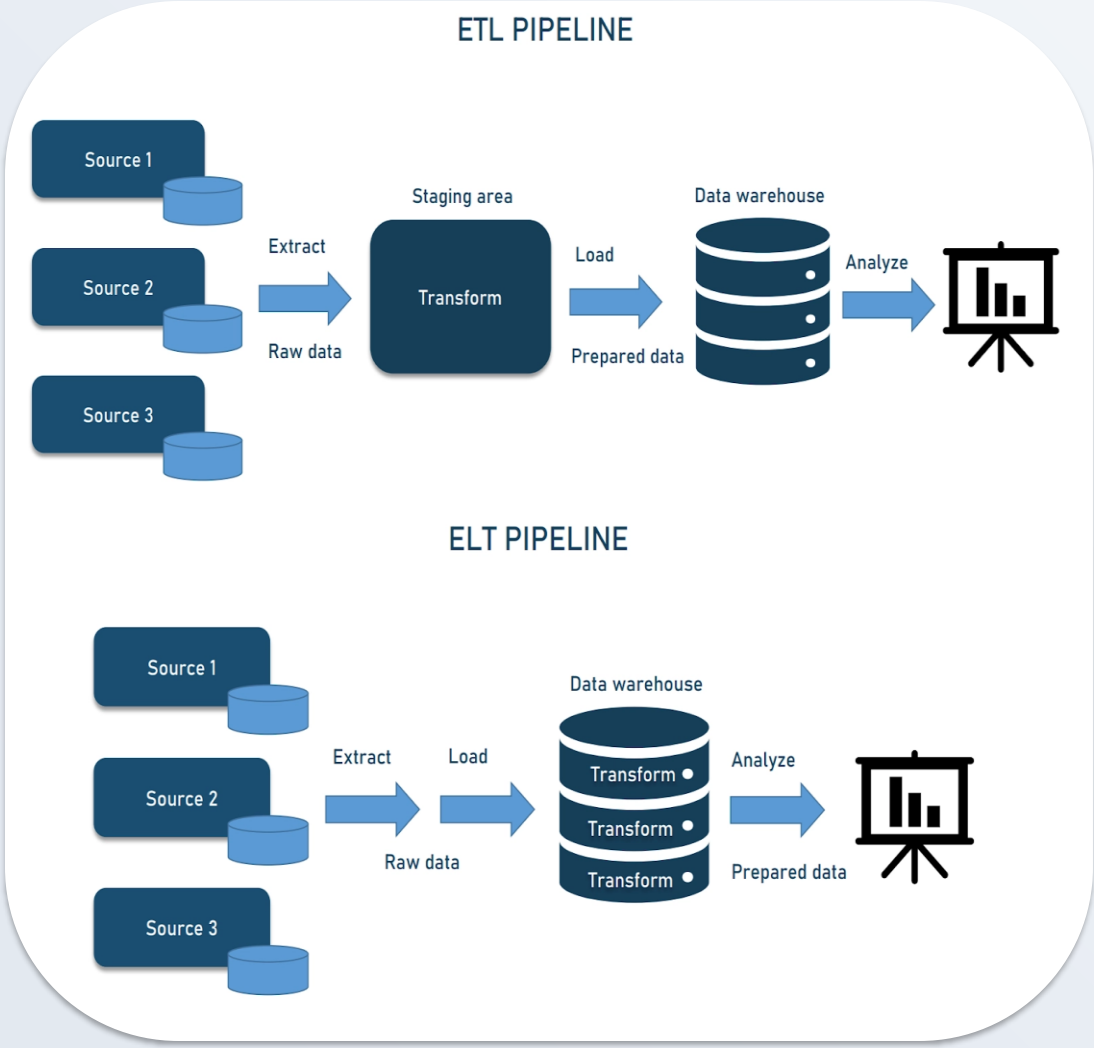
Fault Tolerance

Automatic recovery from node failures

Scalability

Add nodes horizontally as data grows

Automation in Hadoop Data Management



Data Ingestion

Automated collection from various sources

Automated Processing

MapReduce/YARN workflows

Data Transformation

ETL/ELT pipelines

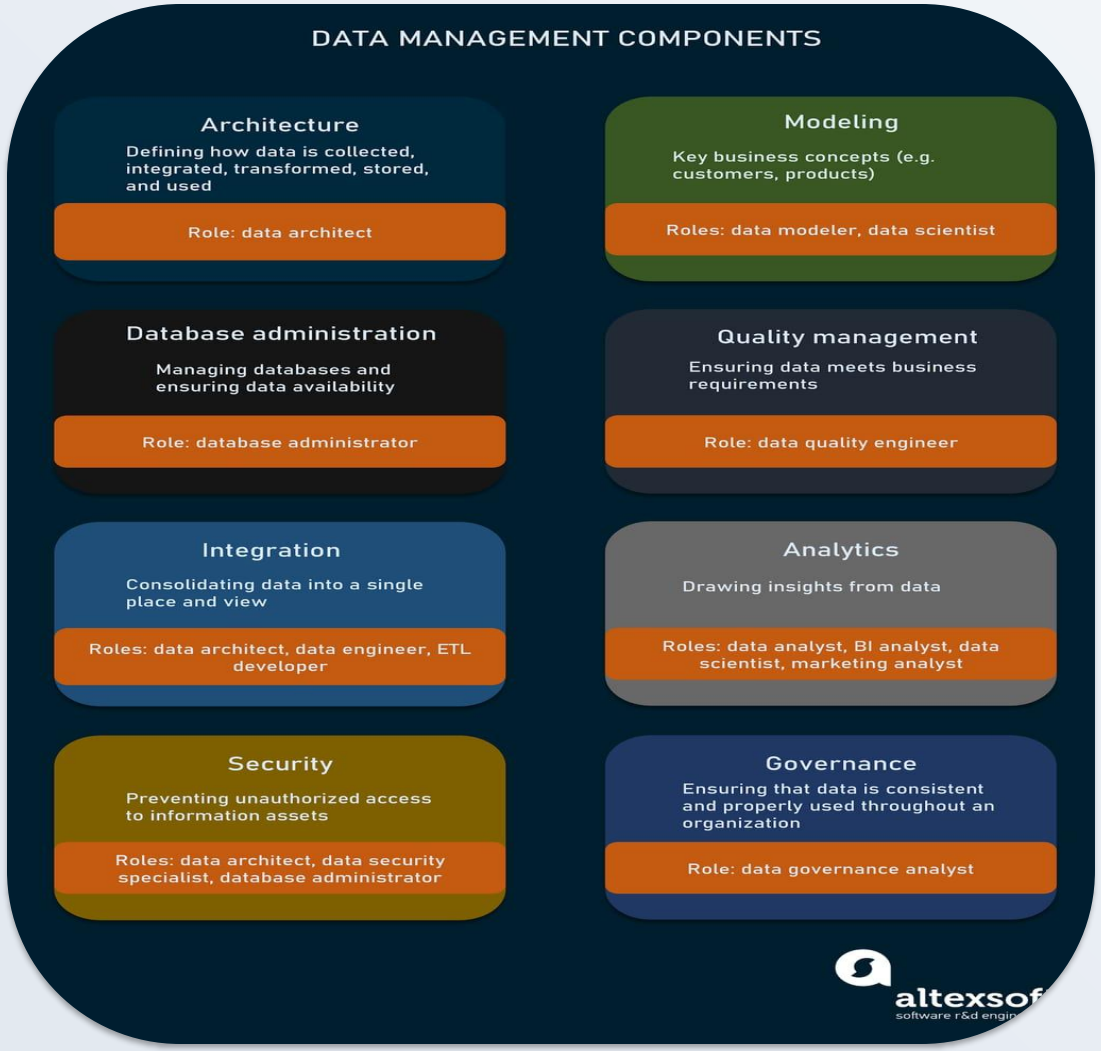
Workflow Orchestration

Oozie/Airflow scheduling

Mind Map - Key Concepts

- **HDFS Architecture**
- **Data Storage & Replication**
- **Automation & Workflows**
- **Hadoop Tools**
  - Hive, Pig, Sqoop
- **Best Practices**
- **Challenges & Solutions**

Best Practices for Managing Hadoop Data



Data Security

Implement authentication and authorization

Performance Optimization

Monitor cluster health and tune configurations

Scalability Planning

Design for future growth

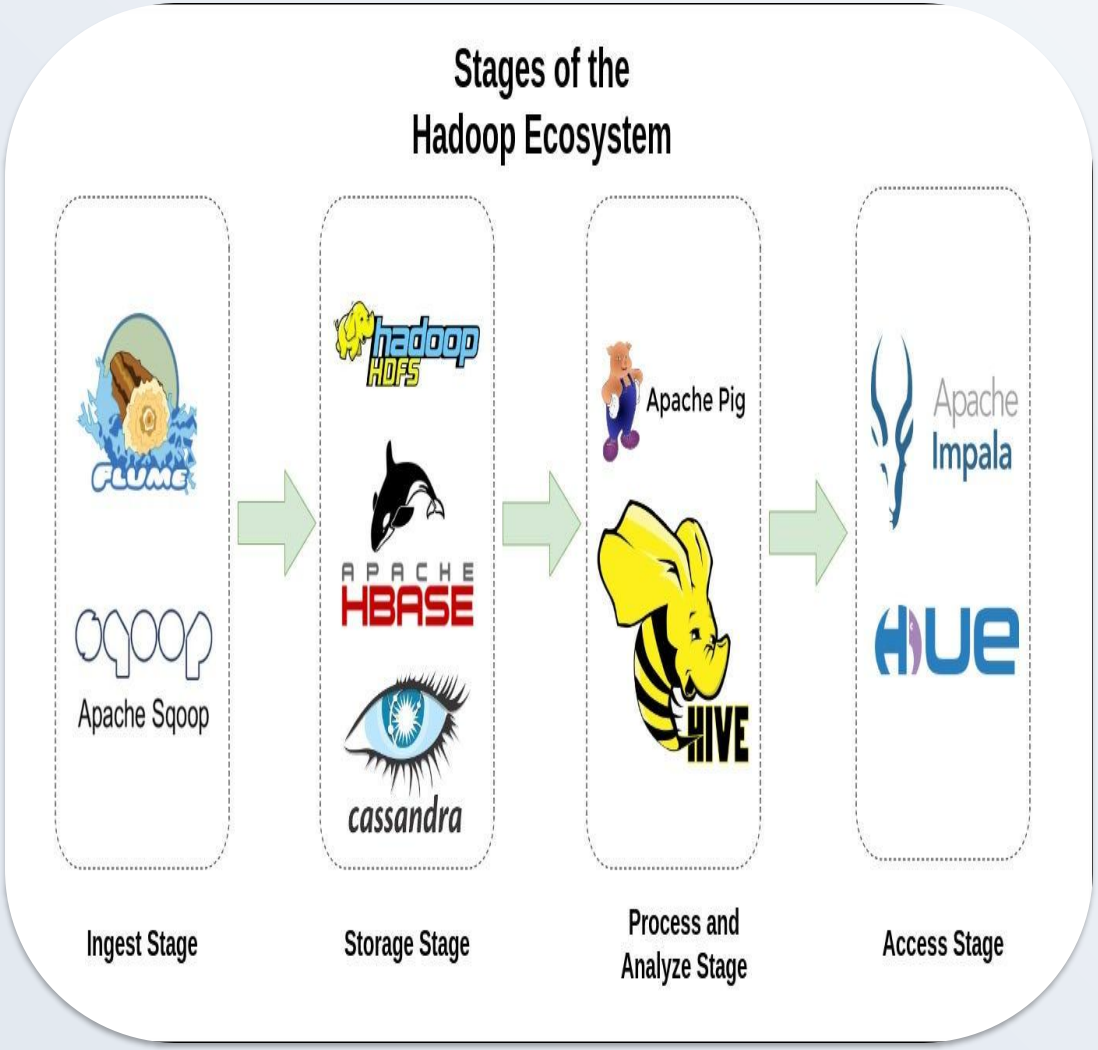
Regular Monitoring

Track metrics and alerts

Backup Strategies

Snapshot and disaster recovery planning

Tools for Hadoop Data Management



Hive

Data warehousing and SQL queries

Pig

Scripting and data transformation

Sqoop

Data transfer between Hadoop and RDBMS

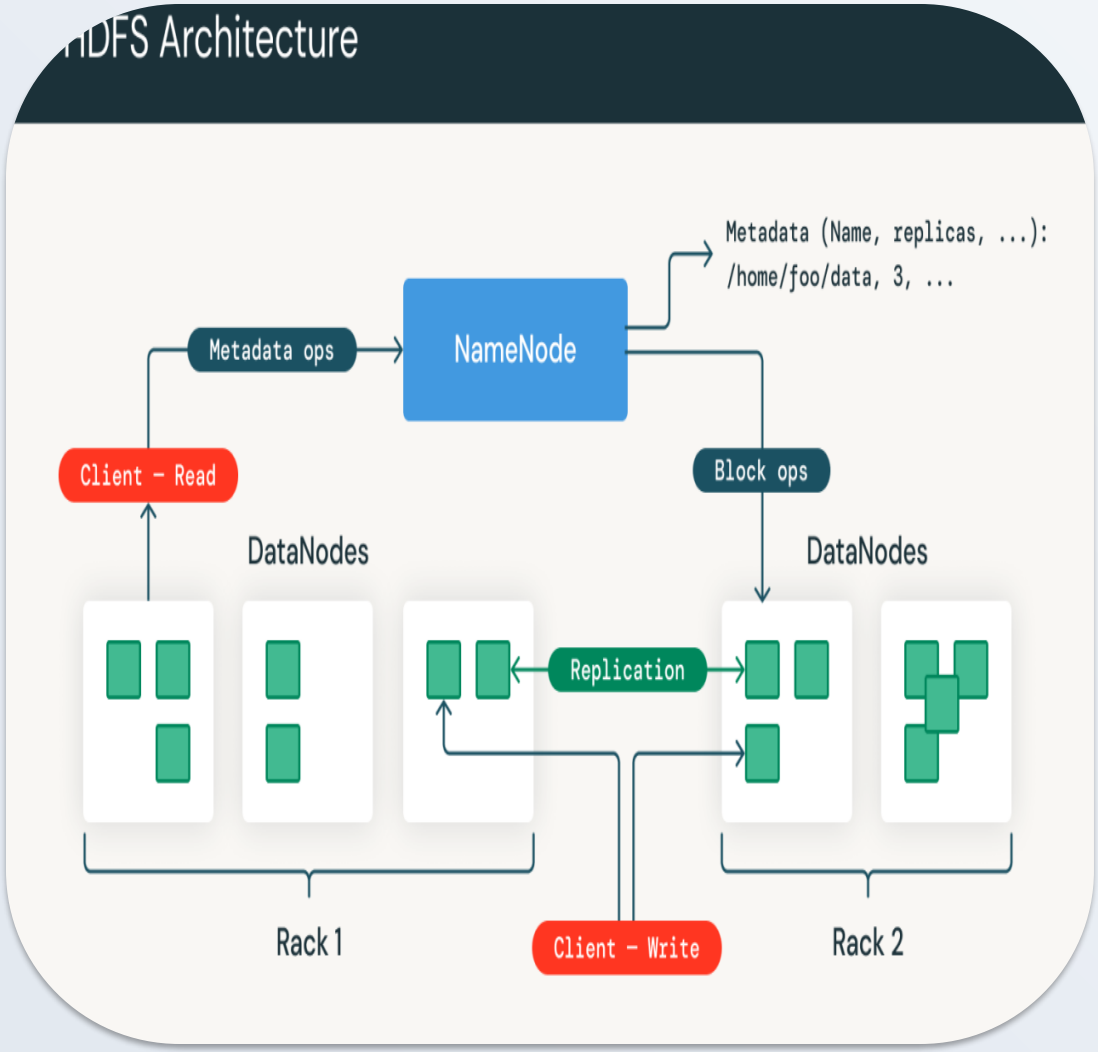
HBase

NoSQL database for real-time access

Oozie

Workflow scheduler and job coordination

Challenges and Solutions



Data Consistency

Use ZooKeeper coordination

Node Failures

Automatic replication recovery

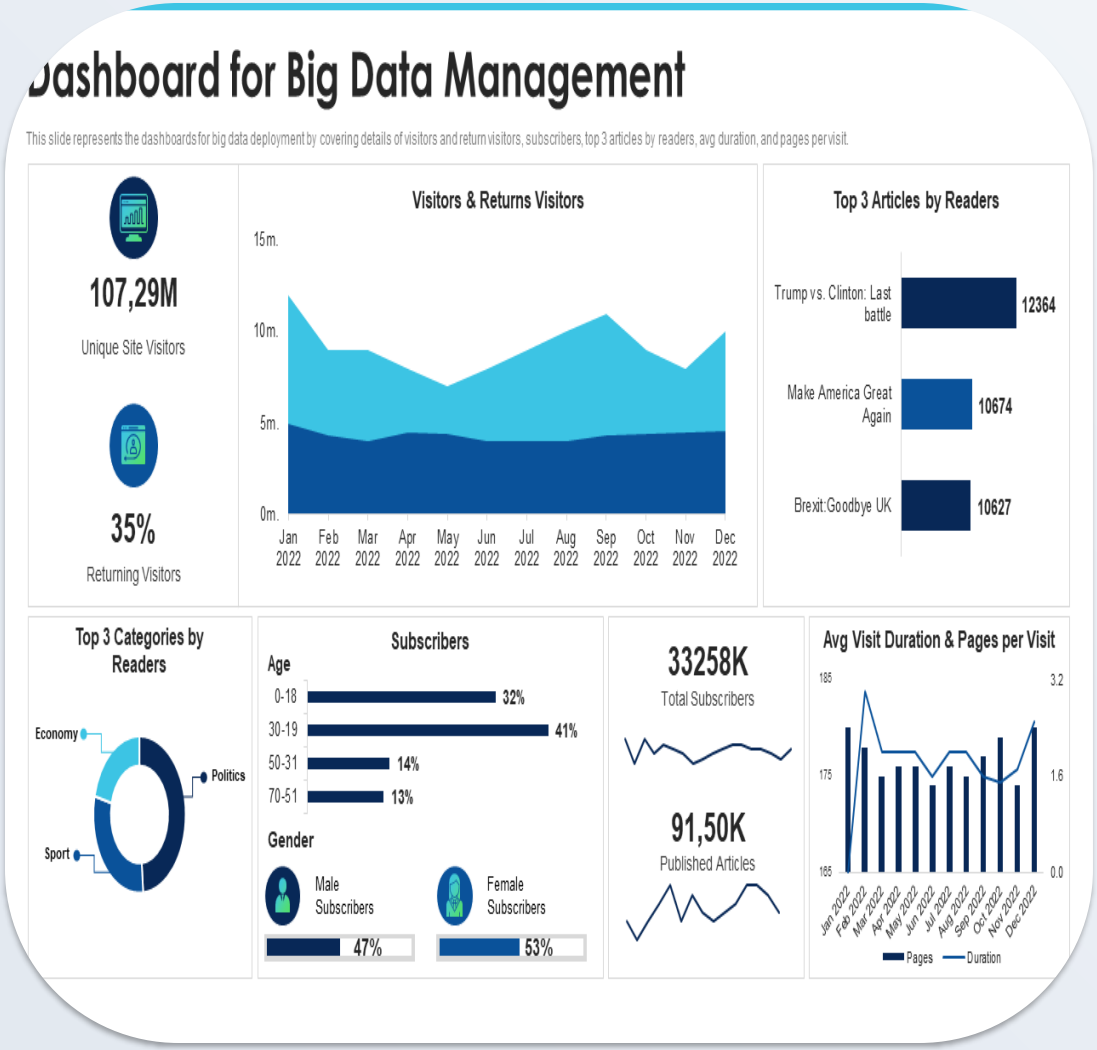
Network Latency

Optimize data locality

Resource Management

YARN scheduling

Recap and Summary



Key Takeaways

Hadoop enables efficient big data management

HDFS Importance

Distributed storage with fault tolerance

Automation Benefits

Streamlined data pipelines and workflows

Tools Overview

Hive, Pig, Sqoop, HBase for diverse data needs