# Managing Hadoop File-MapReduce Concept

23MCT305 - Data Analytics in Automation System

Faculty: **N. KARTHI**, AP/MCT

## What is Hadoop?

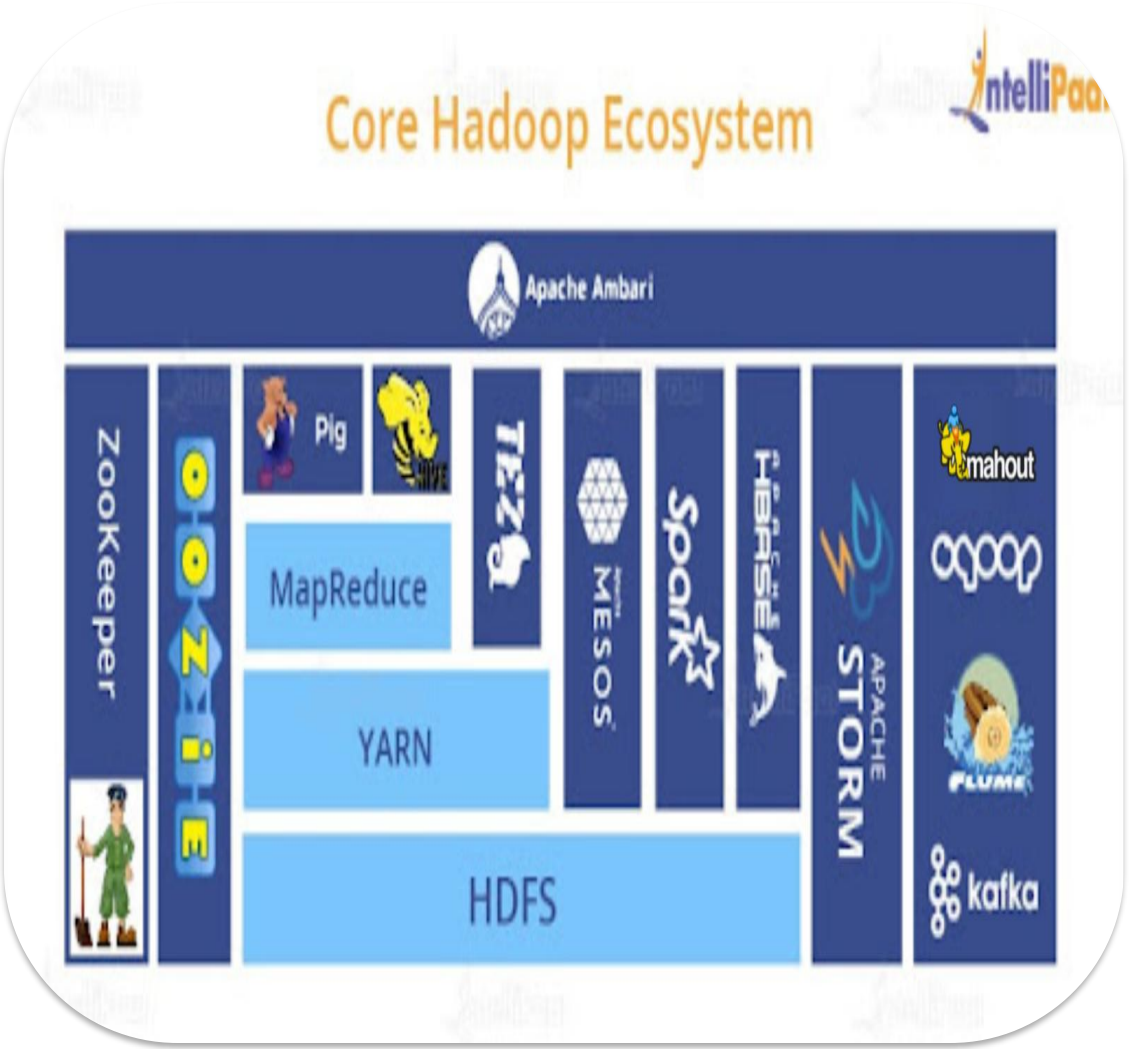
Distributed processing framework for big data

## Key Features

- Scalability
- Cost-effective
- Flexibility

## Hadoop Ecosystem

HDFS | MapReduce | YARN | Hive | Pig | HBase



# Hadoop Distributed File System (HDFS)

## Introduction to HDFS

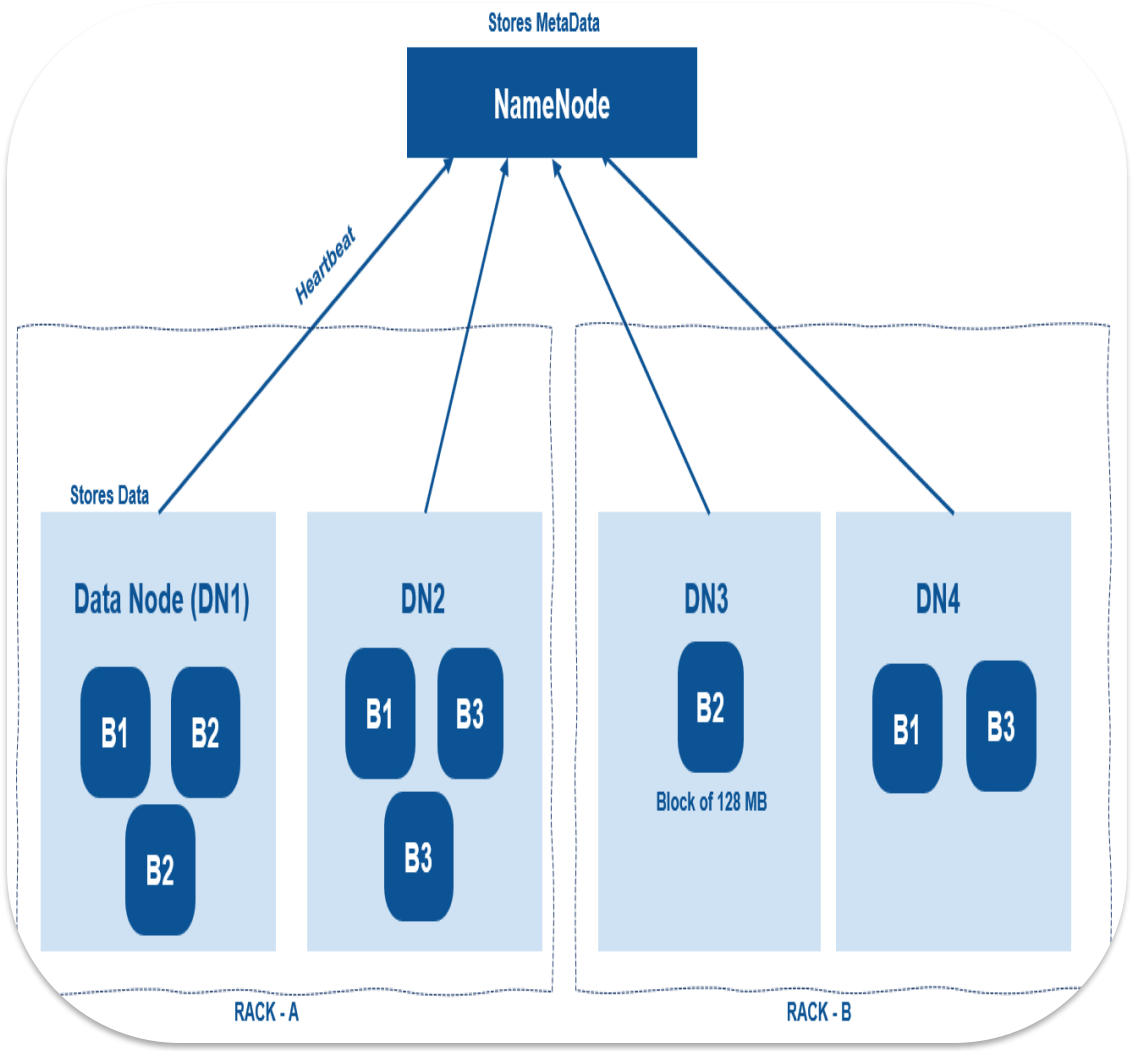
Distributed storage system designed for big data

## Key Characteristics

- Distributed
- Fault Tolerant
- High Throughput

## Benefits

- Scalability
- Reliability
- Data Locality



# HDFS Architecture Components

### NameNode

Master server • Manages metadata and namespace

### DataNode

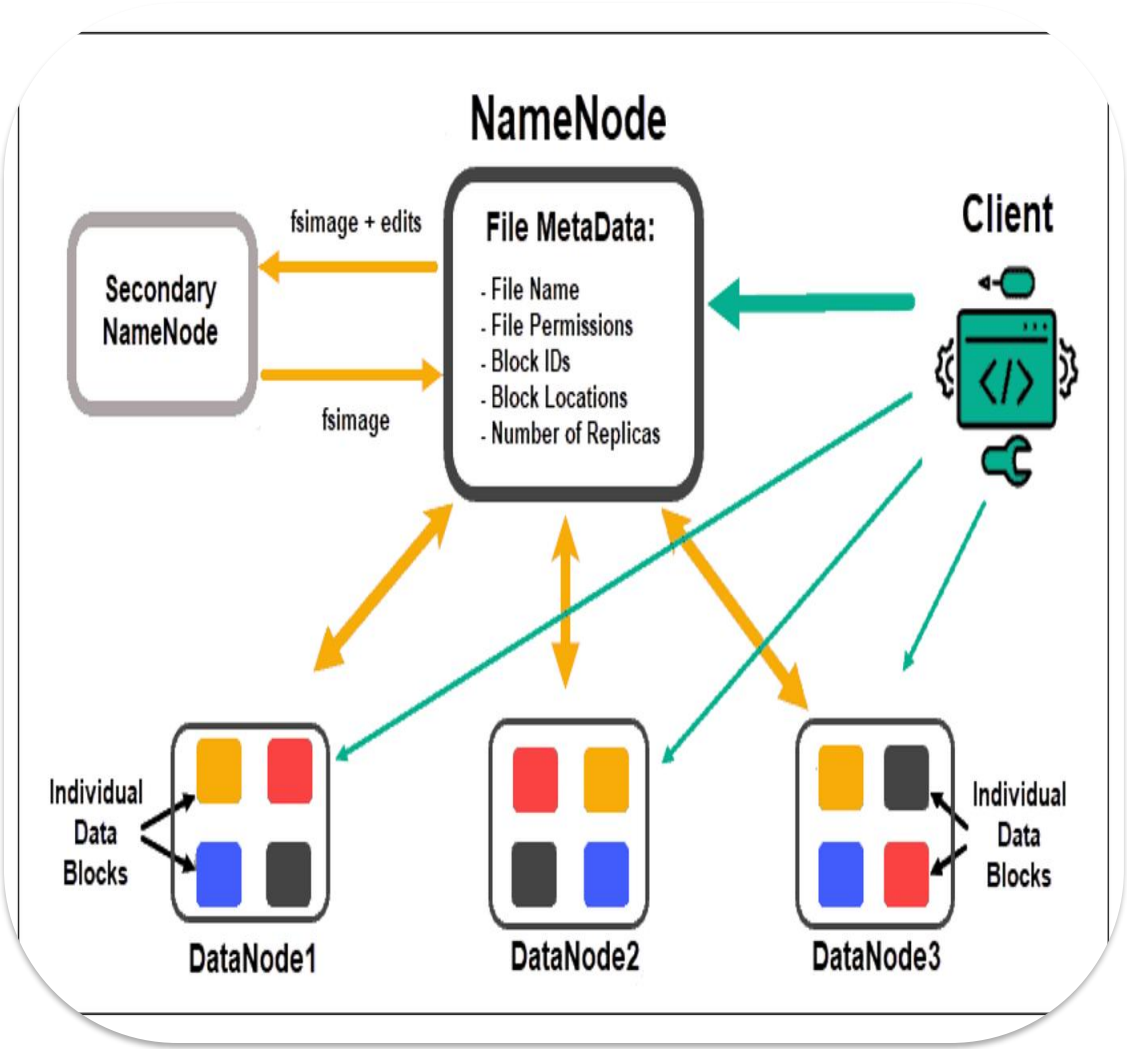
Slave servers • Store data blocks • Serve I/O requests

### Client

Interface for file access and operations

### Block Storage

Data split into blocks • Default 128MB • Replicated across DataNodes



# Introduction to MapReduce

## What is MapReduce?

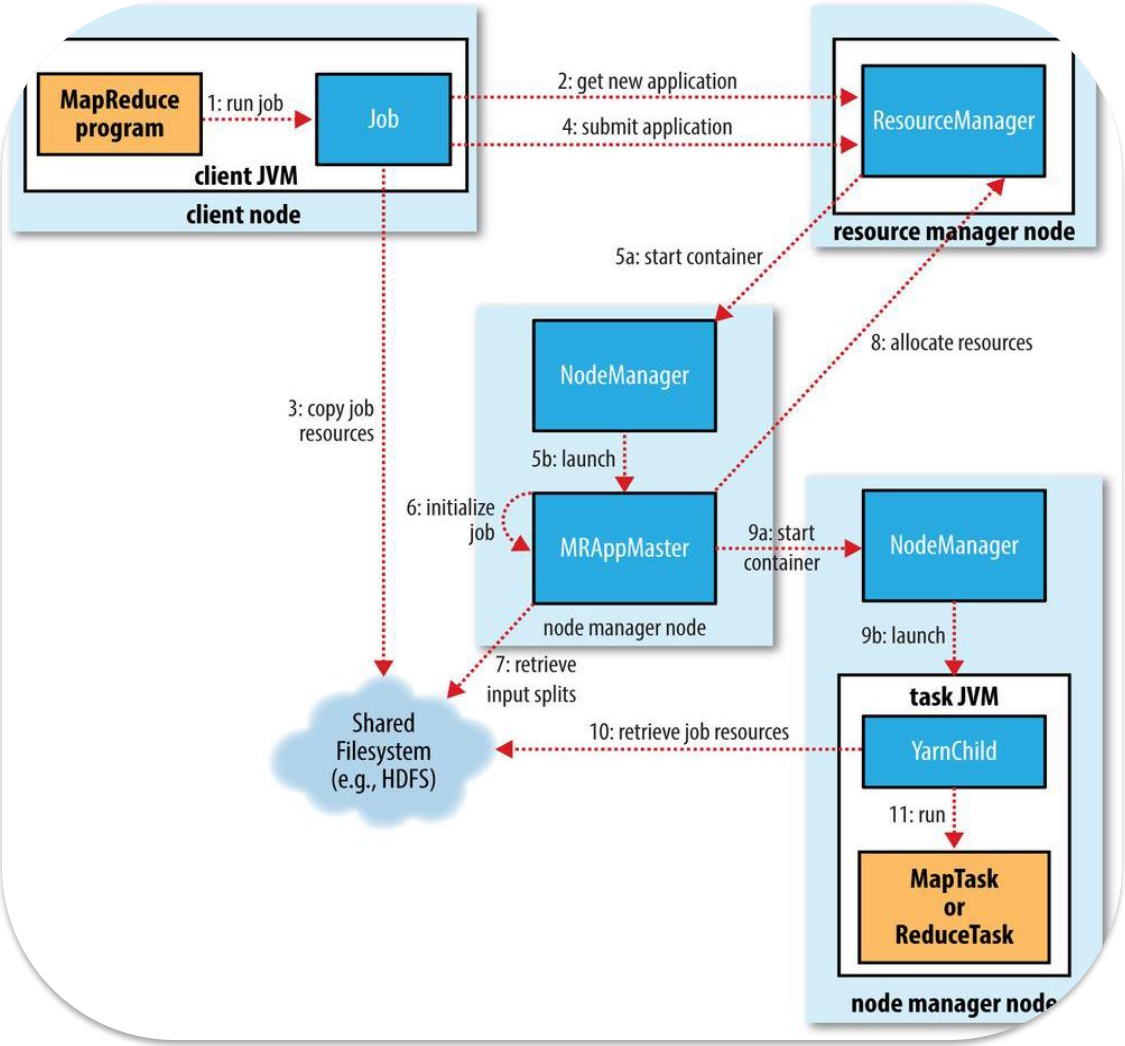
Distributed data processing framework for large datasets

## Programming Model

- Divide and conquer approach
- Parallel processing

## Key Advantages

- Parallel Processing
- Scalability
- Fault Tolerance
- High Throughput



# MapReduce Working Principle

## Map Phase

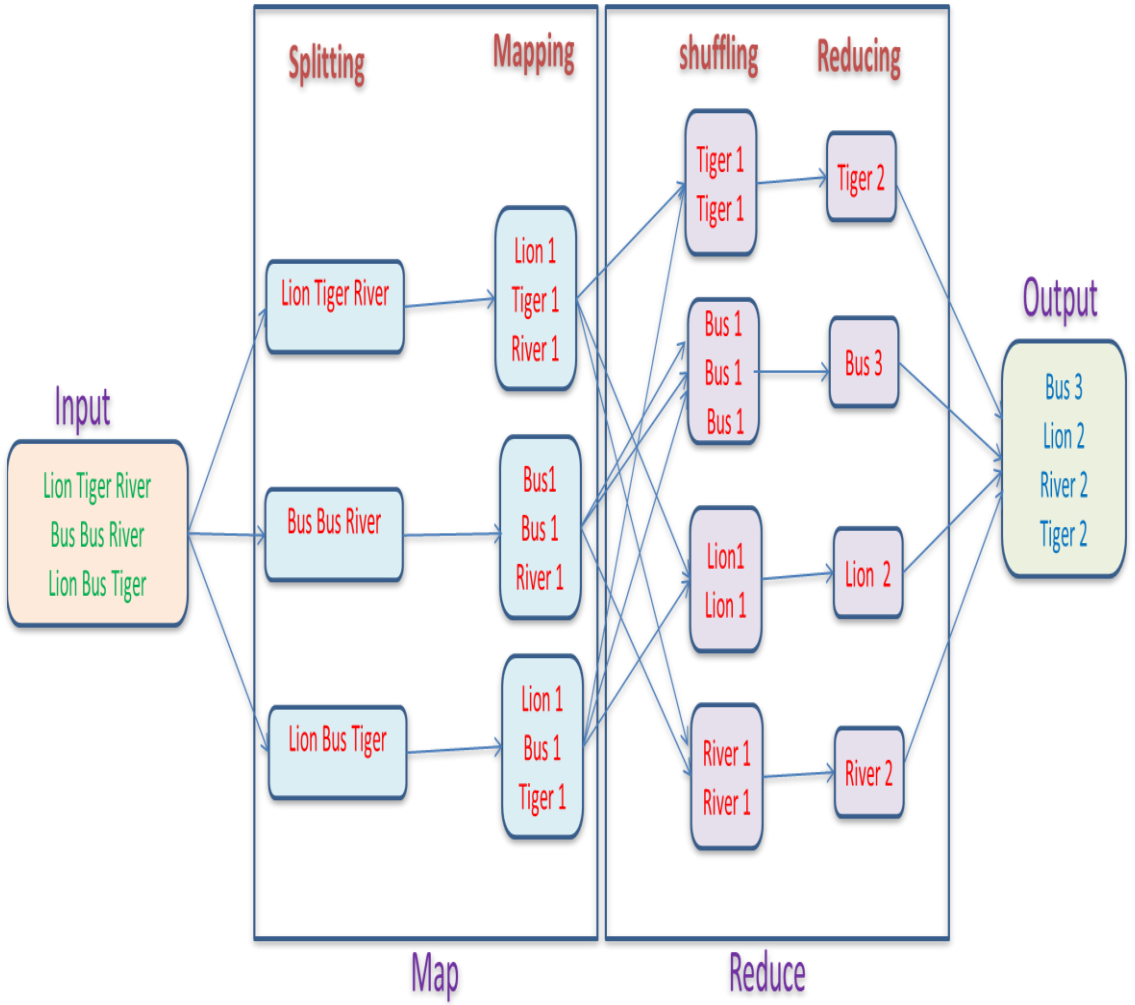
- Input splitting
- Mapping and processing
- Generate key-value pairs

## Shuffle and Sort Phase

- Transfer intermediate data
- Sort by keys

## Reduce Phase

- Aggregate intermediate data
- Generate final output



# MapReduce Job Execution Flow

## Job Submission

Client submits job to JobTracker/ResourceManager

## JobTracker/ResourceManager

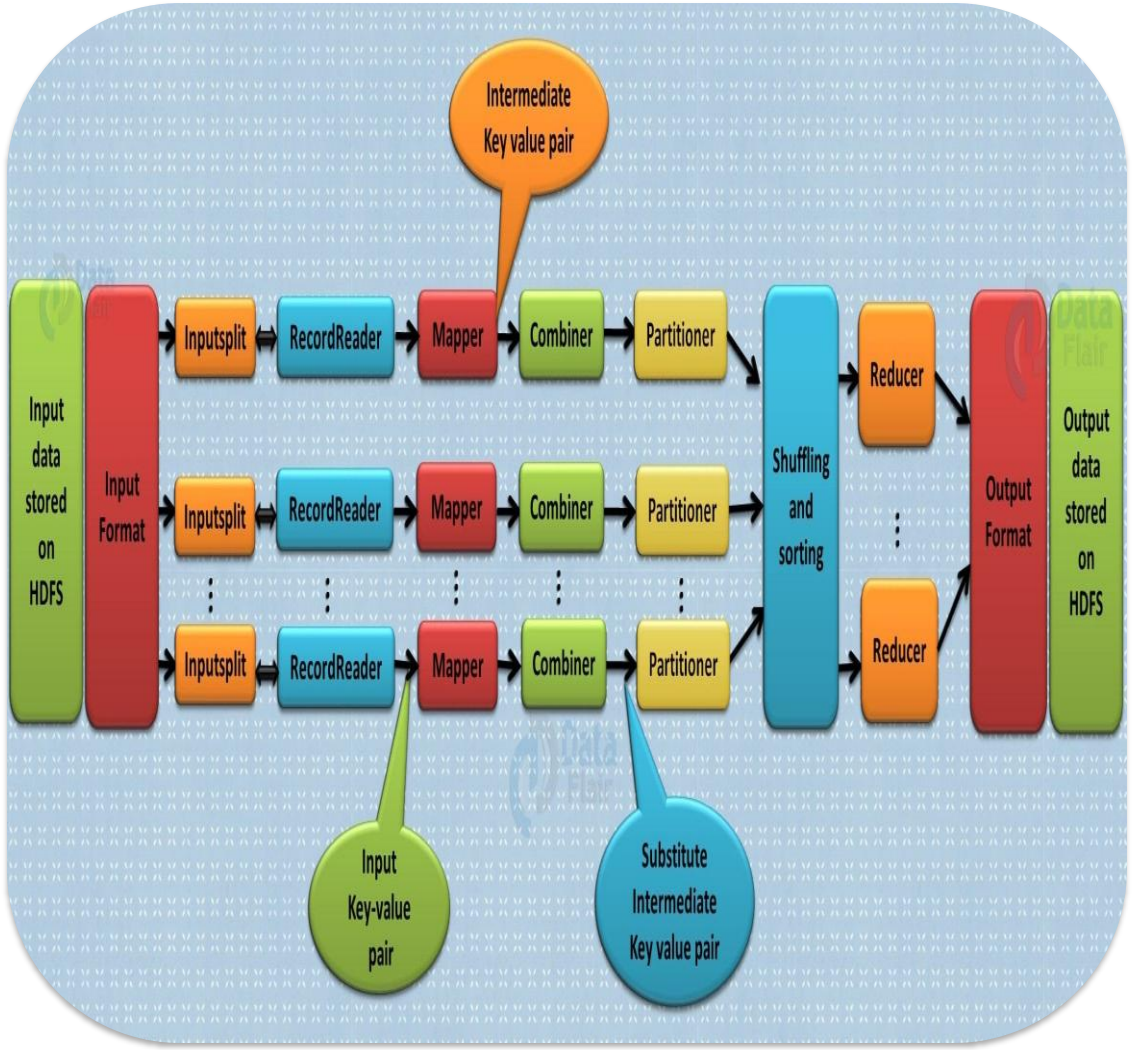
- Accepts job
- Splits job into tasks
- Schedules tasks on TaskTrackers

## TaskTracker/NodeManager

- Executes map and reduce tasks
- Reports status to JobTracker

## Task Monitoring

- Tracks task progress
- Handles failures
- Ensures job completion



# Managing Hadoop File-MapReduce

## File Management in HDFS

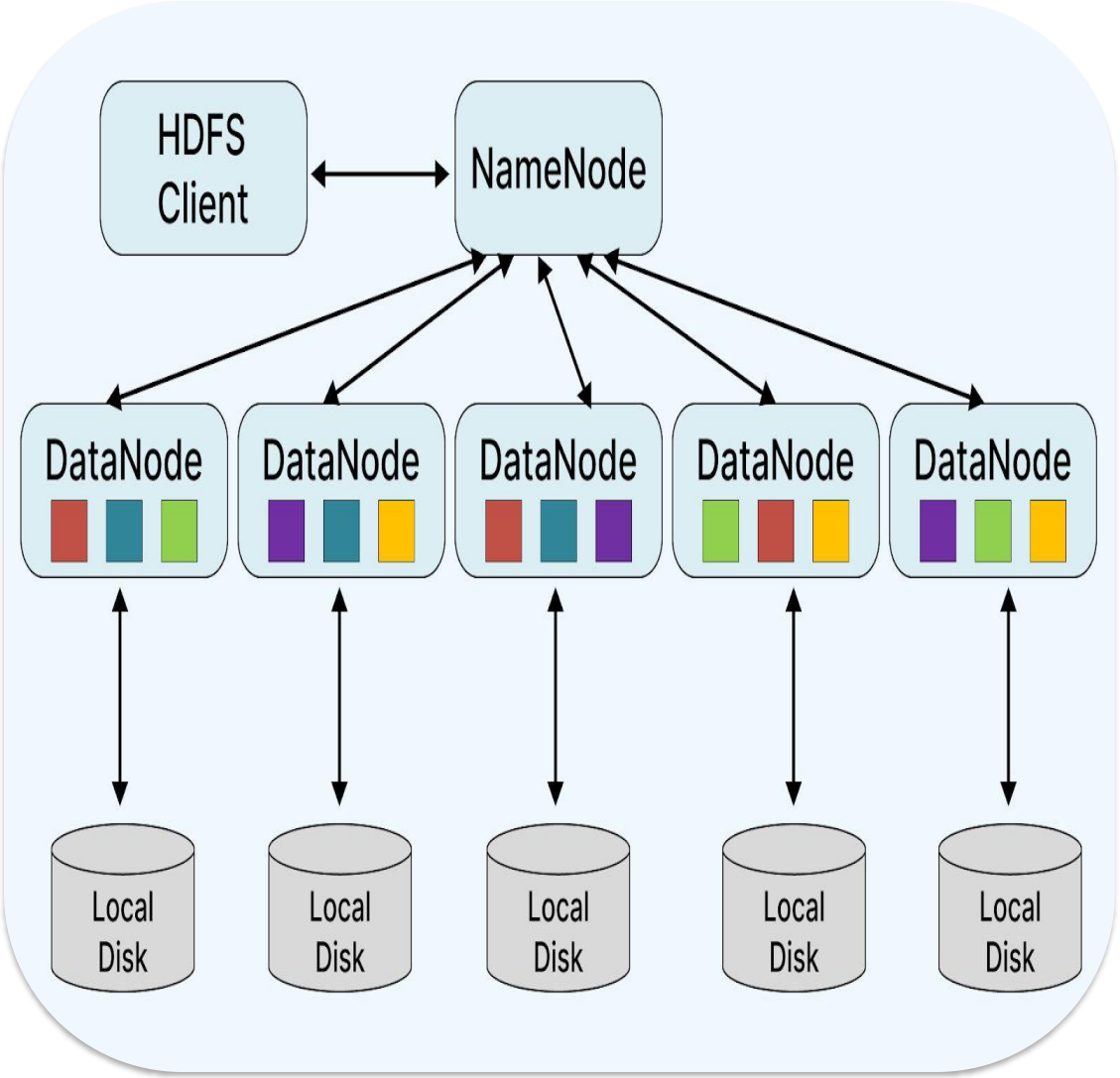
- Read/Write operations
- Data replication factor management
- Block management

## Job Management in MapReduce

- Job submission and scheduling
- Task monitoring and optimization
- Resource allocation

## Best Practices

- Proper data locality
- Optimize job configuration
- Monitor cluster health
- Handle failures gracefully



# Mind Map: Hadoop File-MapReduce Concept

## HDFS Components

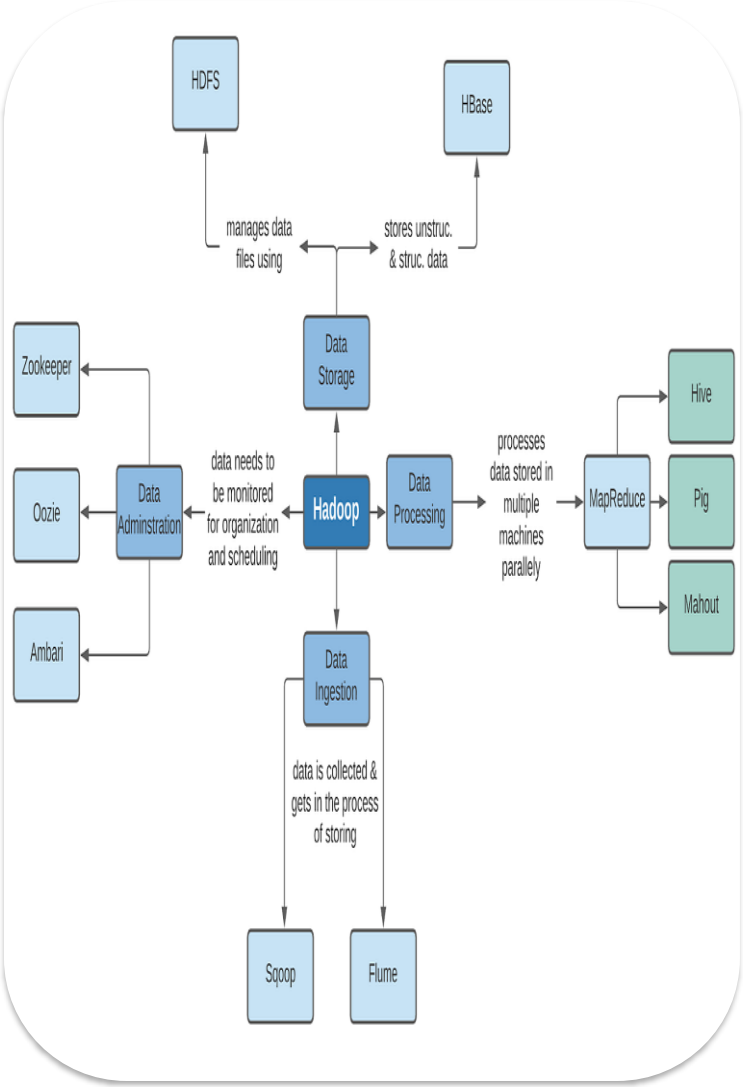
- NameNode
- DataNode
- Client

## MapReduce Phases

- Map Phase
- Shuffle & Sort
- Reduce Phase

## Management

- File Management
- Job Management
- Best Practices




## Architecture

- Distributed Storage
- Parallel Processing
- Fault Tolerance

## Features

- Scalability
- Fault Tolerance
- High Throughput

## Benefits

- Cost-effective
- Reliable
- Flexible

# Summary and Recap

## Key Points Covered

### HDFS Architecture

NameNode • DataNode • Client

### MapReduce Working Principle

Map • Shuffle/Sort • Reduce

### Job Execution Flow

Job submission • Scheduling • Monitoring

## Learning Objectives Achieved

- Understanding Hadoop components
- Mastering MapReduce phases
- Managing Hadoop file system

## Future Scope

- Apache Spark integration
- Advanced optimization techniques
- Real-world applications