Hadoop Scalability

Data Analytics in Automation System

23MCT305

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

January 2026

Introduction to Hadoop and Scalability

What is Hadoop?

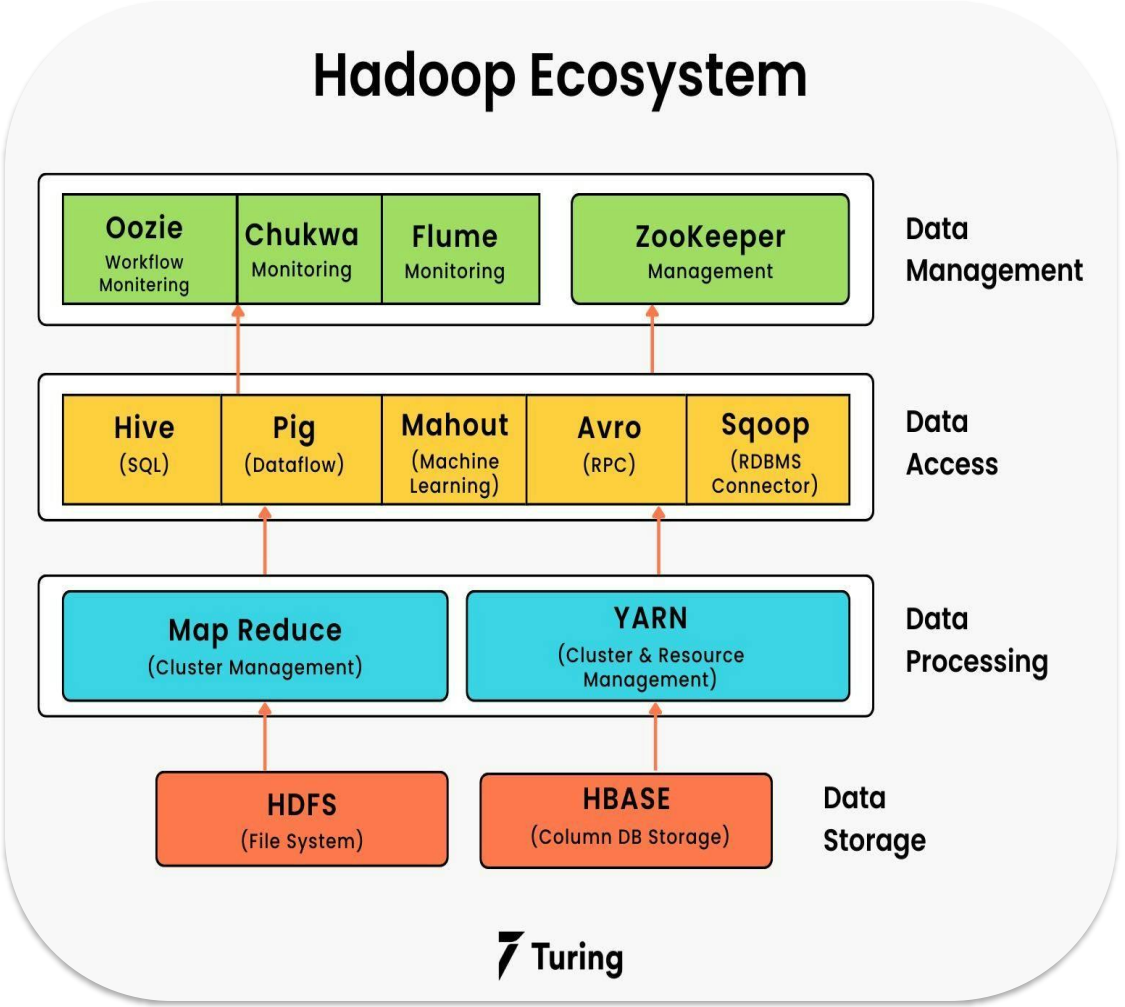
Open-source framework for distributed storage and processing of large datasets across clusters of computers

Scalability Defined

Ability of a system to handle growing workload by adding resources

Why Scalability Matters

- Handle petabytes of data
- Support thousands of concurrent users
- Ensure high availability
- Cost-effective growth



Hadoop Architecture Components

HDFS

Distributed File System

Storage component

MapReduce

Processing Framework

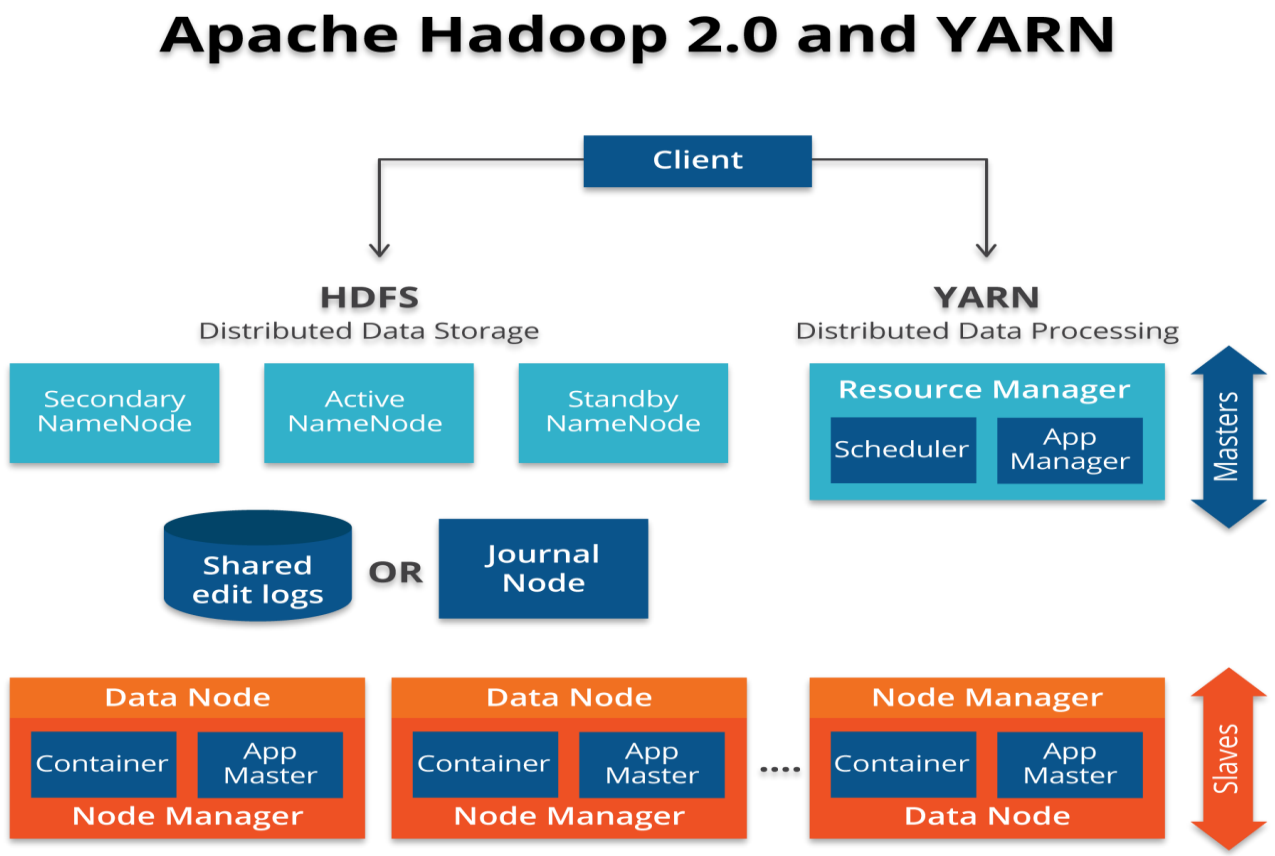
Computational engine

YARN

Yet Another Resource Negotiator

Resource management

Key Feature: Each component scales independently



HDFS Scalability Features

Block Storage

Default: 128MB blocks

Data Replication

3 copies by default

NameNode

Stores metadata

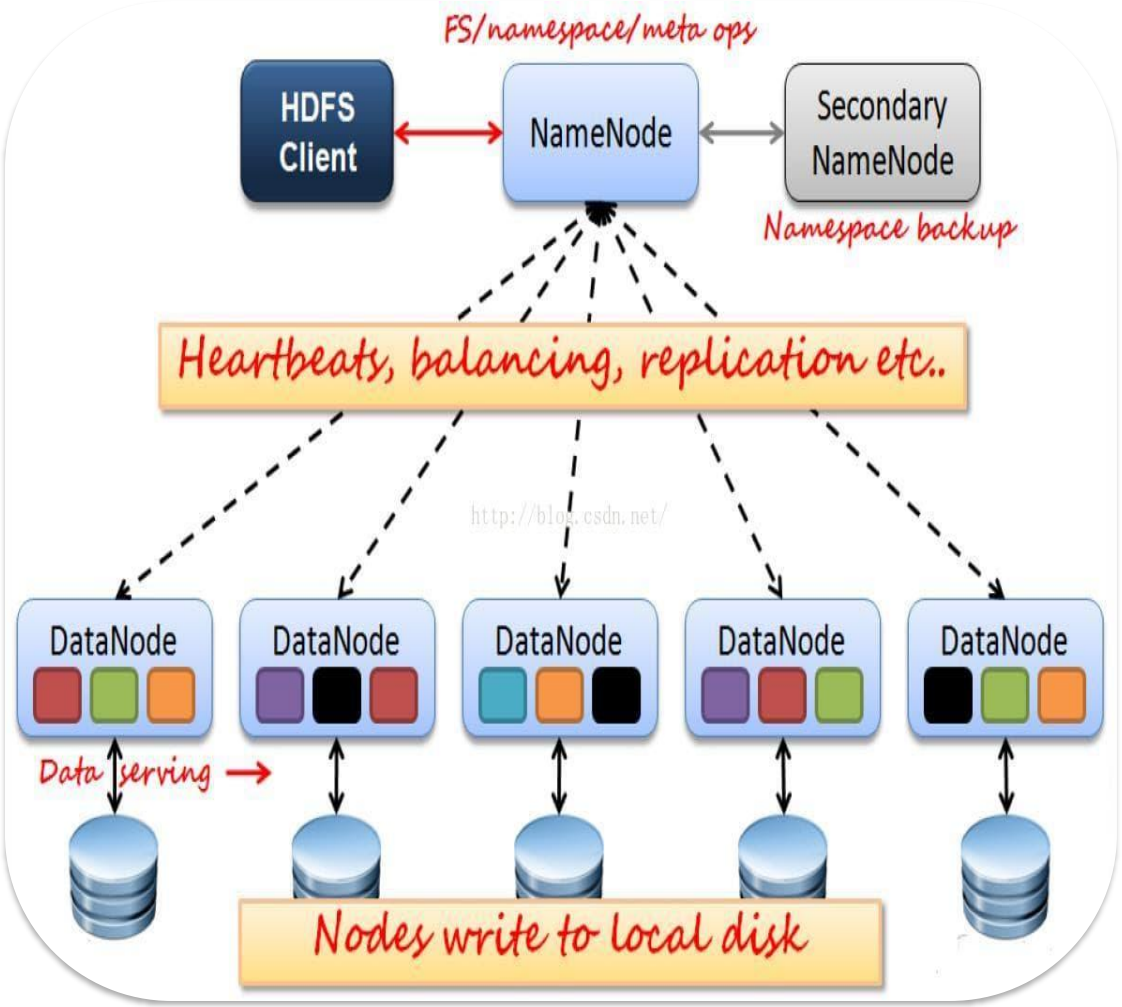
DataNodes

Store actual data blocks

Rack Awareness

Optimizes placement

Scalability: Add DataNodes to increase storage



MapReduce Scalability Mechanisms

Distributed Processing

Execute across cluster nodes

Map Phase

Split & process input data

Reduce Phase

Shuffle & aggregate results

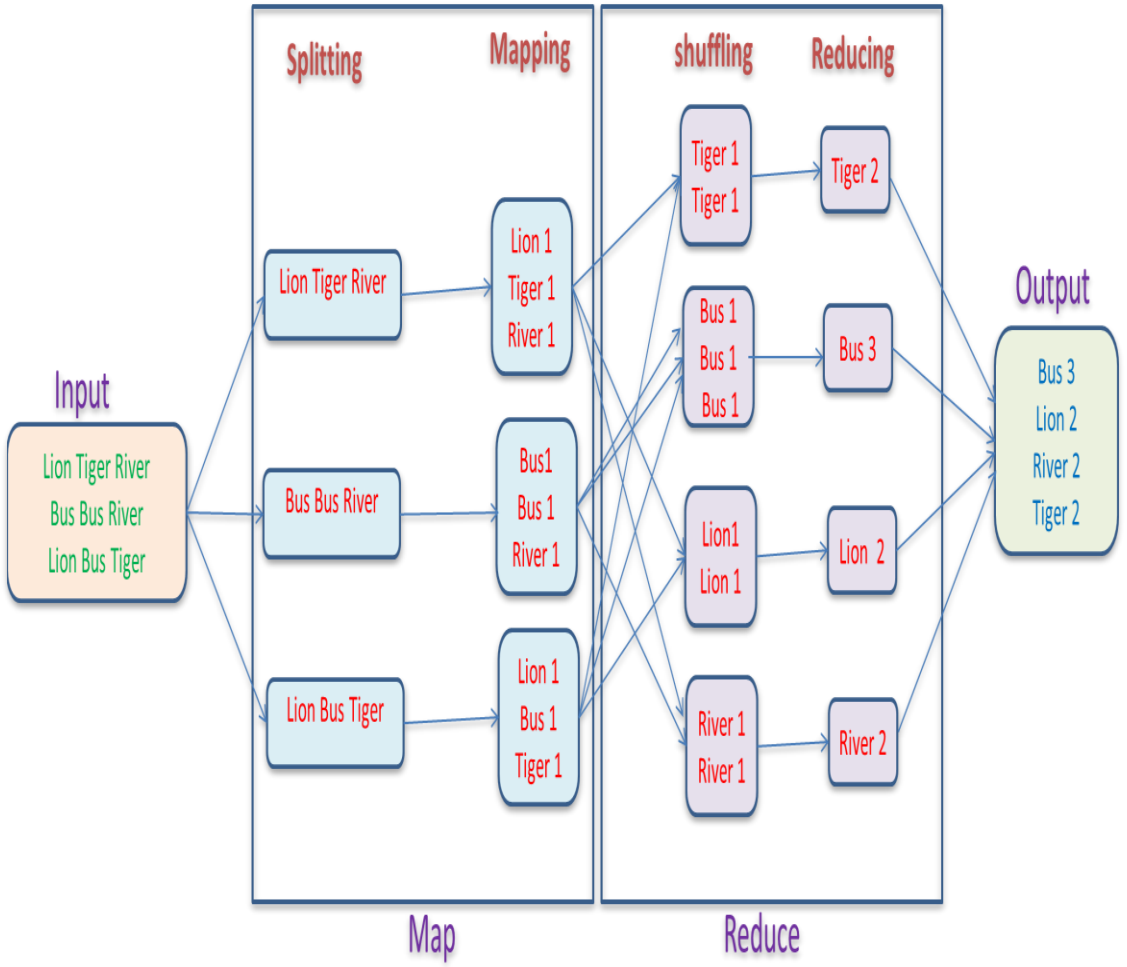
Job Scheduling

Optimized resource allocation

Fault Tolerance

Automatic task retry

Scalability: Parallel processing across nodes



YARN and Resource Management

ResourceManager

Global resource scheduler

NodeManager

Per-node agent

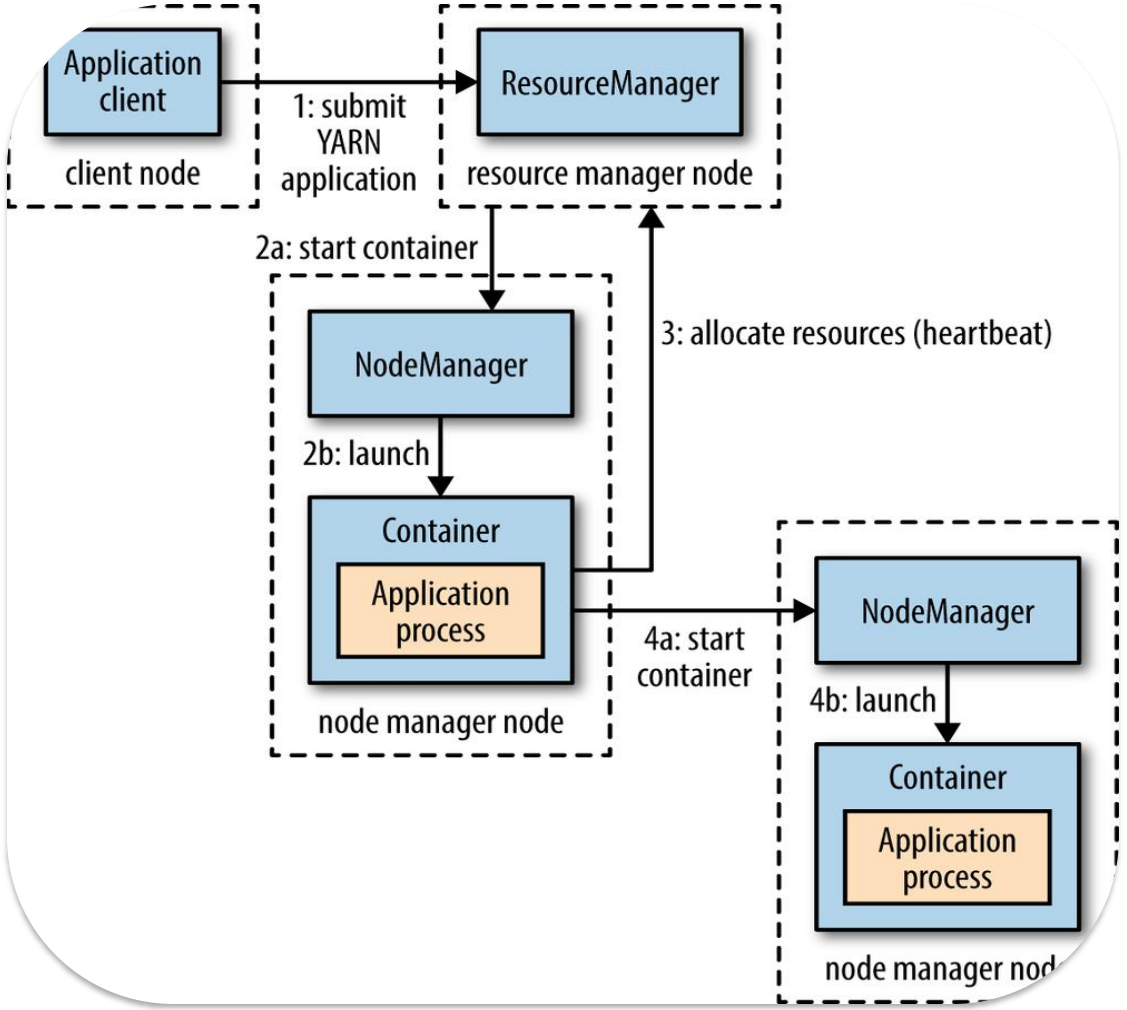
ApplicationMaster

Per-application scheduler

Container

Resource allocation unit

Scalability: Dynamic resource allocation



Horizontal vs Vertical Scaling

Horizontal Scaling

Scale Out - Add more nodes

✓ Cost-effective hardware

✓ Fault tolerant

✓ No single point failure

✓ Hadoop preferred approach

Vertical Scaling

Scale Up - Upgrade hardware

✗ Expensive hardware

✗ Single point of failure

✗ Limited scalability

✗ Not Hadoop recommended

Real-world Hadoop Scalability Examples

Enterprise Data Centers

Thousands of nodes

Petabytes of storage

Social Media Analytics

Facebook, Twitter, LinkedIn

Real-time user engagement

IoT Data Processing

Smart devices & sensors

Continuous data streams

E-commerce Platforms

Amazon, eBay

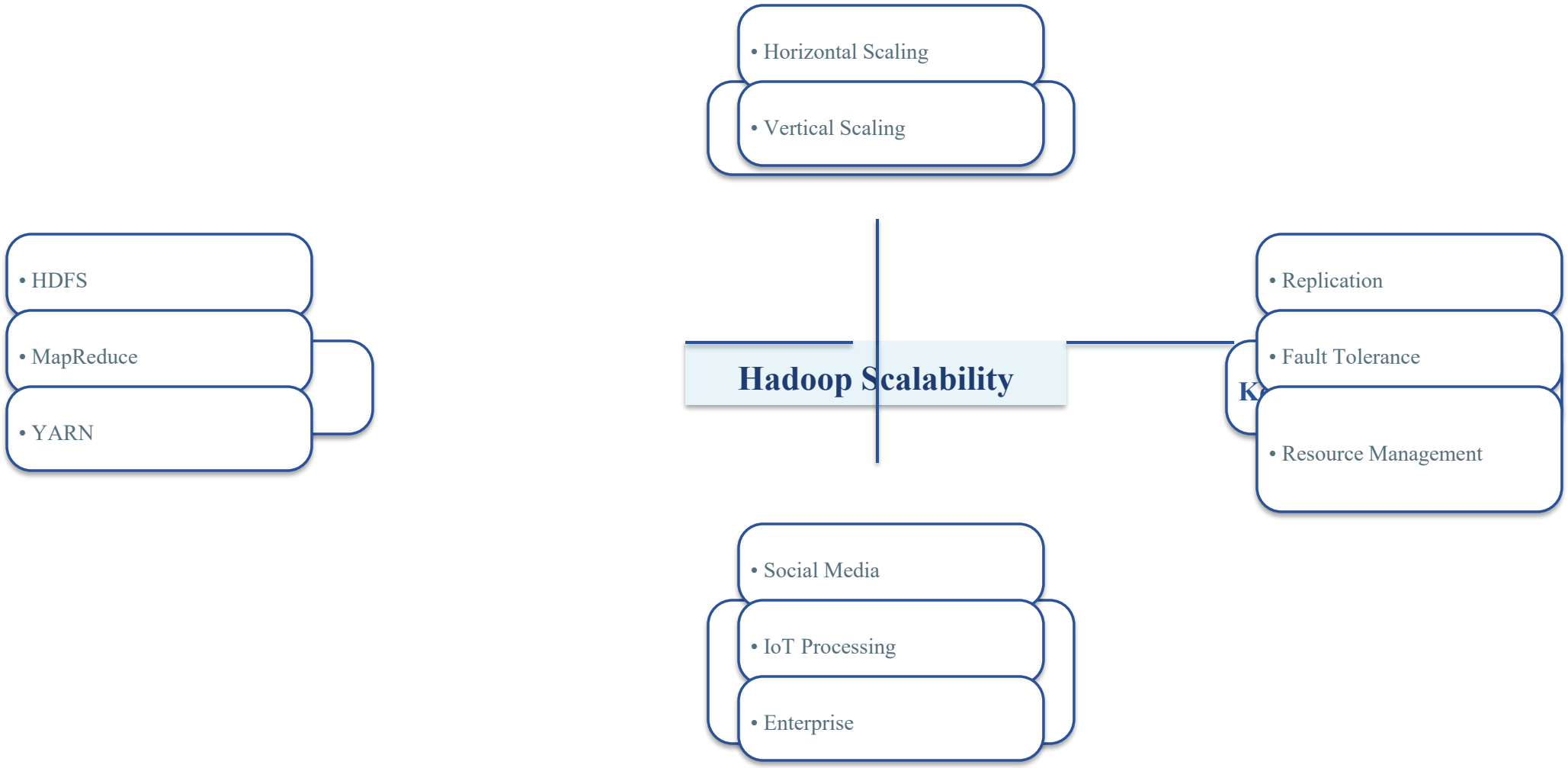
Customer behavior analysis

Financial Services

Fraud detection

Risk assessment

Mind Map of Hadoop Scalability Concepts



Summary and Recap

Key Points

- Hadoop scales horizontally
- HDFS provides distributed storage
- MapReduce enables parallel processing
- YARN manages resources efficiently
- Horizontal scaling preferred

Learning Objectives

✓ Understand scalability principles

✓ Identify Hadoop components

✓ Compare scaling approaches

✓ Recognize real-world applications

✓ Apply scalability best practices

Future Scope

- Cloud-based Hadoop
- Real-time processing
- AI/ML integration
- Hybrid cloud deployment
- Edge computing support