Anatomy of Hadoop

Understanding the Big Data Framework

HDFS

MapReduce

YARN

Introduction to Hadoop

Hadoop is an open-source framework for distributed storage and processing of large datasets

Scalability

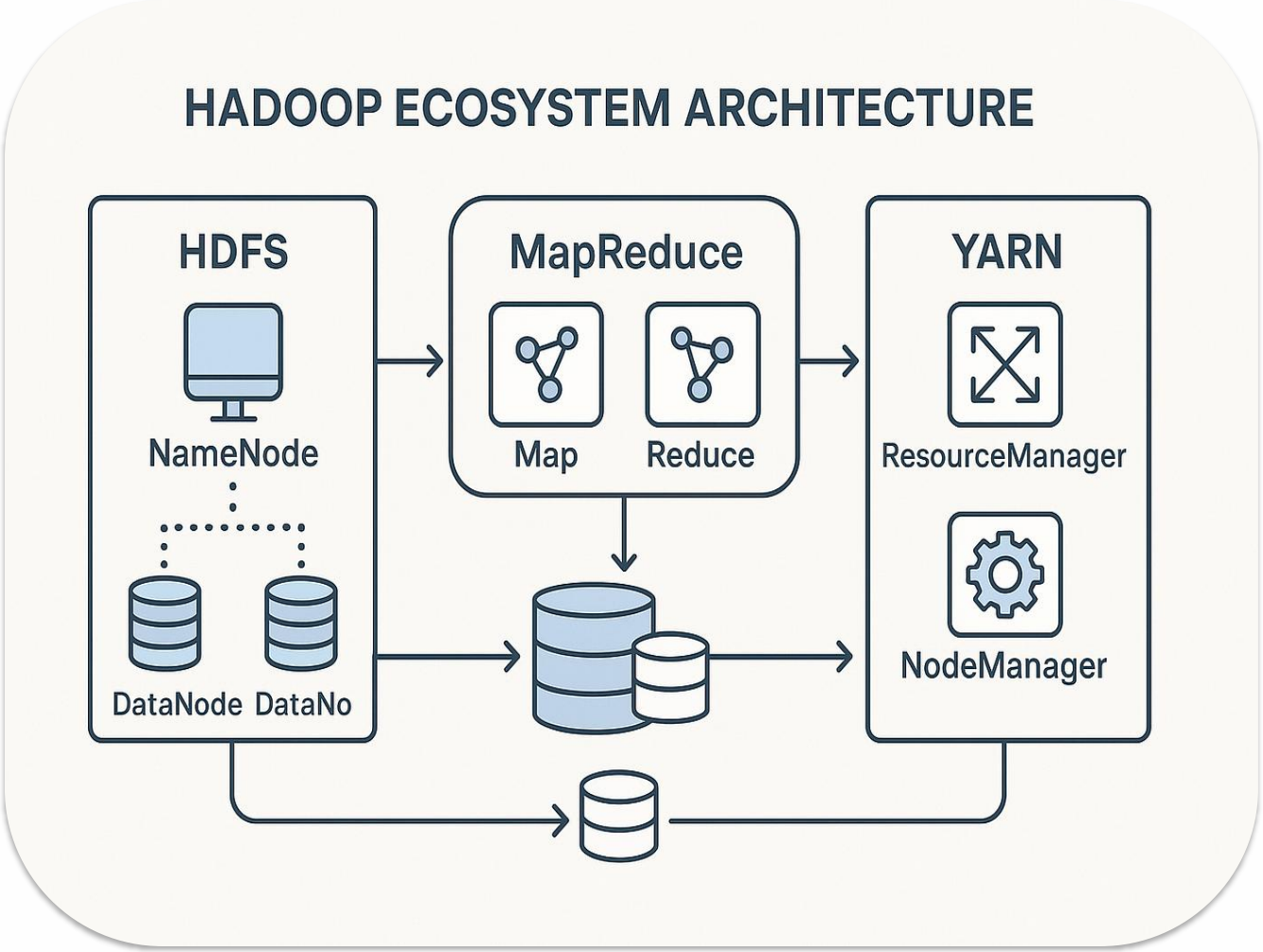
Handles massive data growth

Fault Tolerance

Data redundancy and recovery

Cost-Effectiveness

Runs on commodity hardware



Hadoop Core Components

HDFS

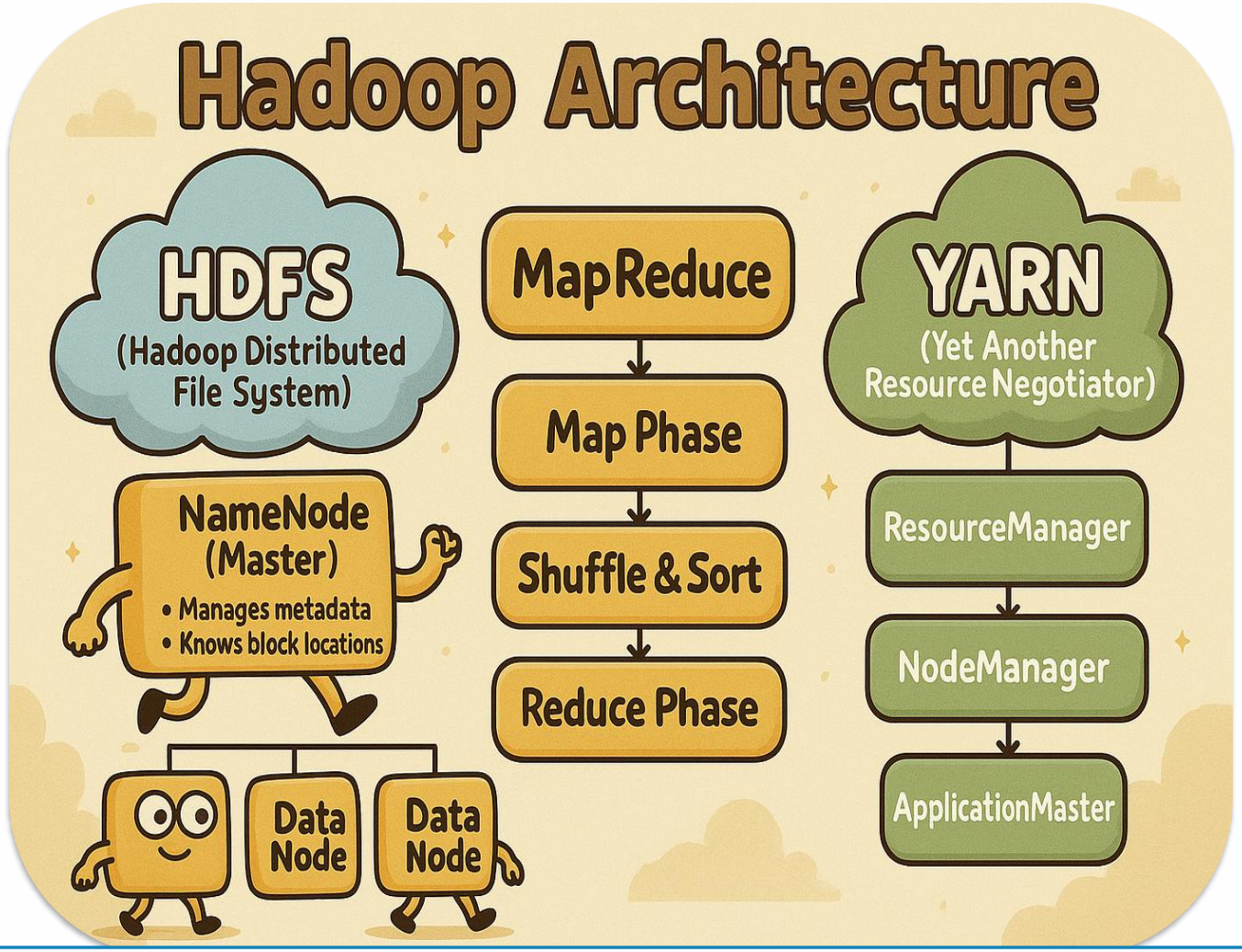
Distributed File System for reliable storage

MapReduce

Parallel processing framework for large datasets

YARN

Resource management and job scheduling



HDFS - Hadoop Distributed File System

NameNode

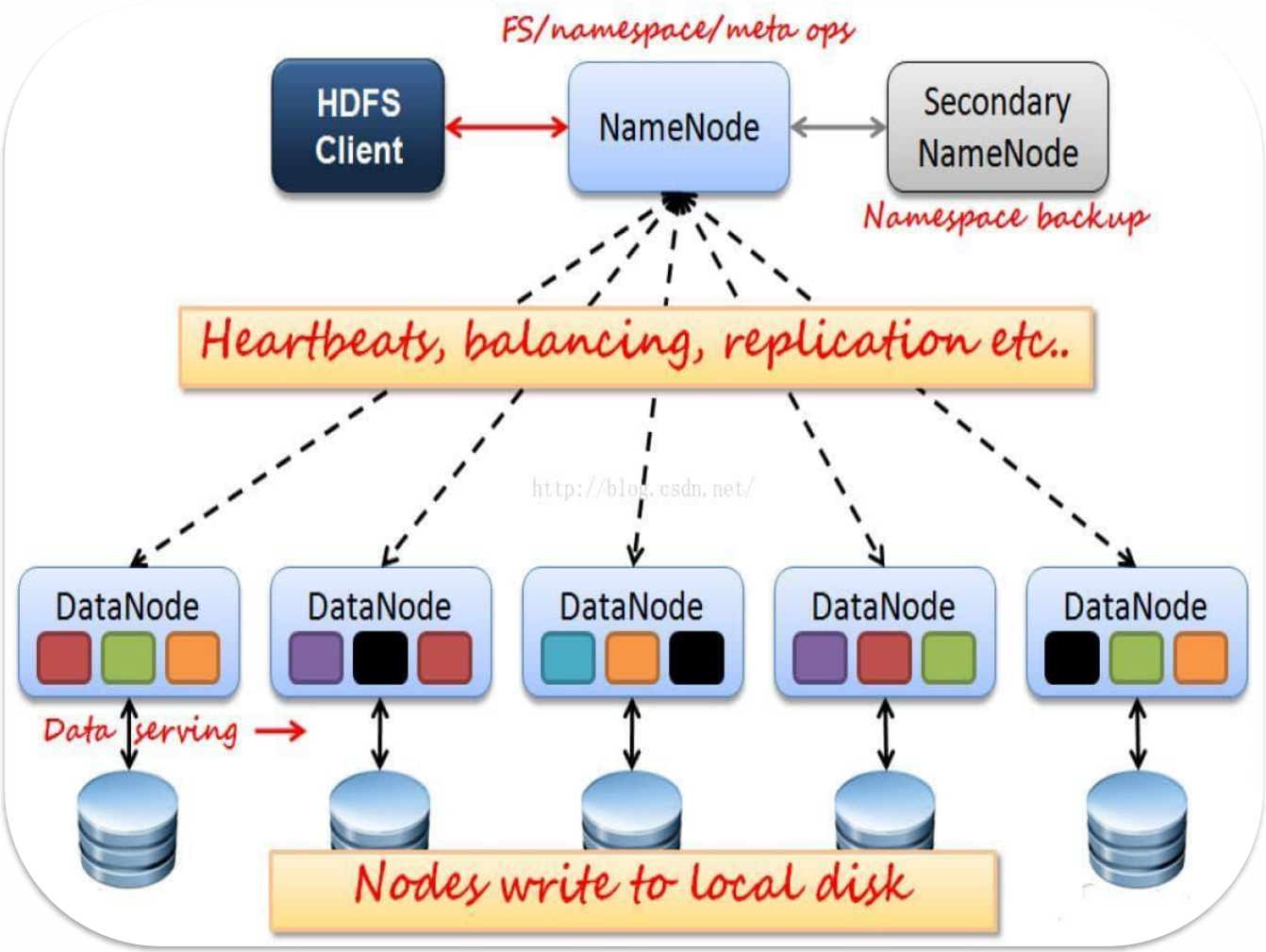
Manages metadata and filesystem namespace

DataNode

Stores data blocks and handles read/write requests

Secondary NameNode

Creates checkpoints and periodic snapshots



Key Features

- Block replication for fault tolerance
- Scalable to petabytes

MapReduce - Data Processing Framework

Map Phase

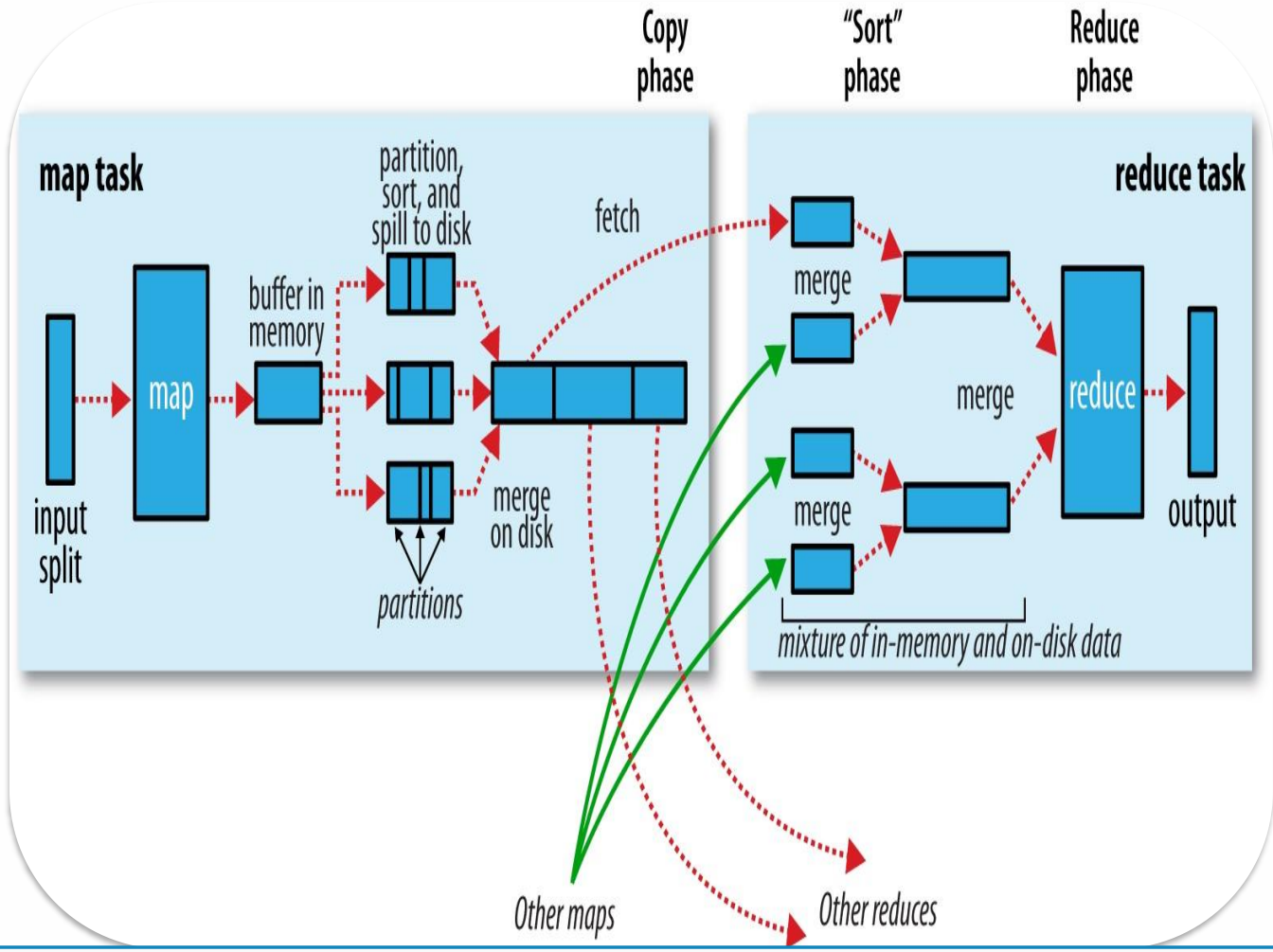
Processes and transforms input data into key-value pairs

Shuffle & Sort

Organizes and distributes key-value pairs to reducers

Reduce Phase

Aggregates and processes intermediate results to final output



Key Characteristics

- Parallel processing
- Fault tolerance

YARN - Yet Another Resource Negotiator

ResourceManager

Global resource scheduling and cluster management

NodeManager

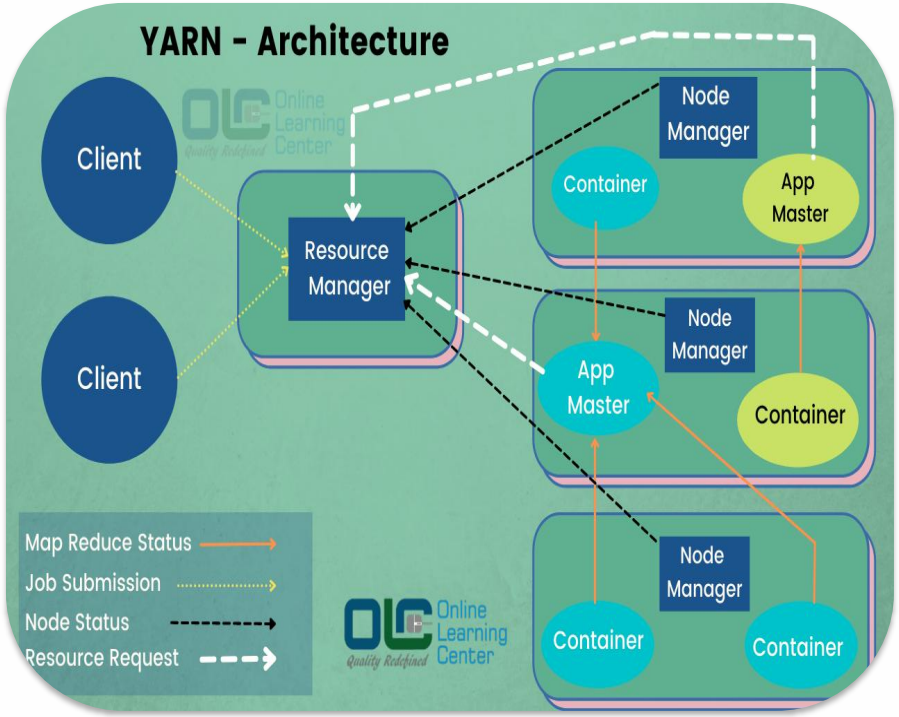
Per-node resource management and container execution

ApplicationMaster

Per-application resource negotiation and monitoring

Key Benefits

- Efficient resource utilization
- Multi-tenancy support
- Scalability



Hadoop Ecosystem

Hive

Data warehousing and SQL-like queries

Pig

Scripting platform for data analysis

HBase

NoSQL database for real-time access

Spark

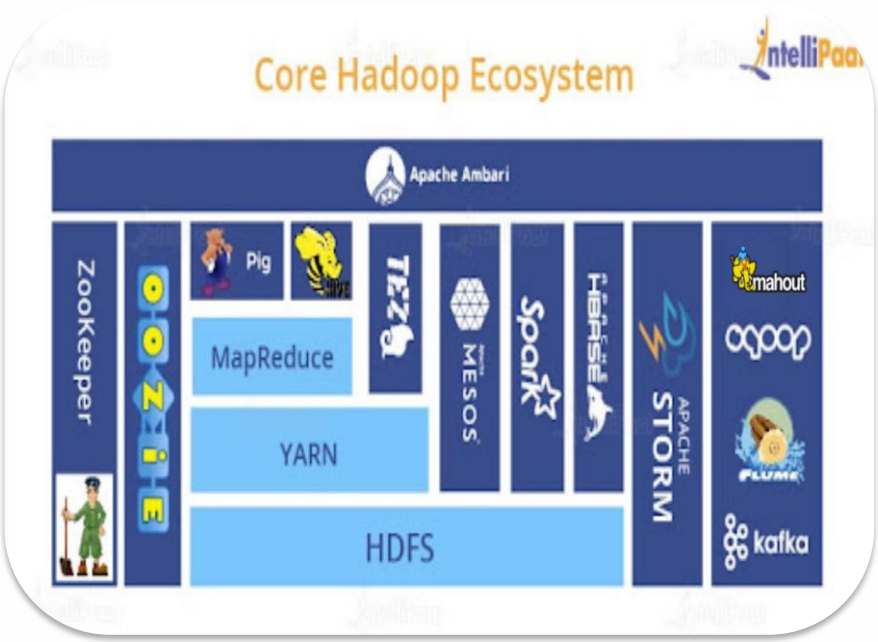
In-memory processing and analytics

ZooKeeper

Coordination and synchronization service

Integration Benefits

- Comprehensive big data platform
- Diverse data processing capabilities
- Industry-standard solutions



Hadoop Use Cases

Healthcare

Patient data analysis, disease research, clinical trials

Finance

Risk assessment, fraud detection, algorithmic trading

Social Media

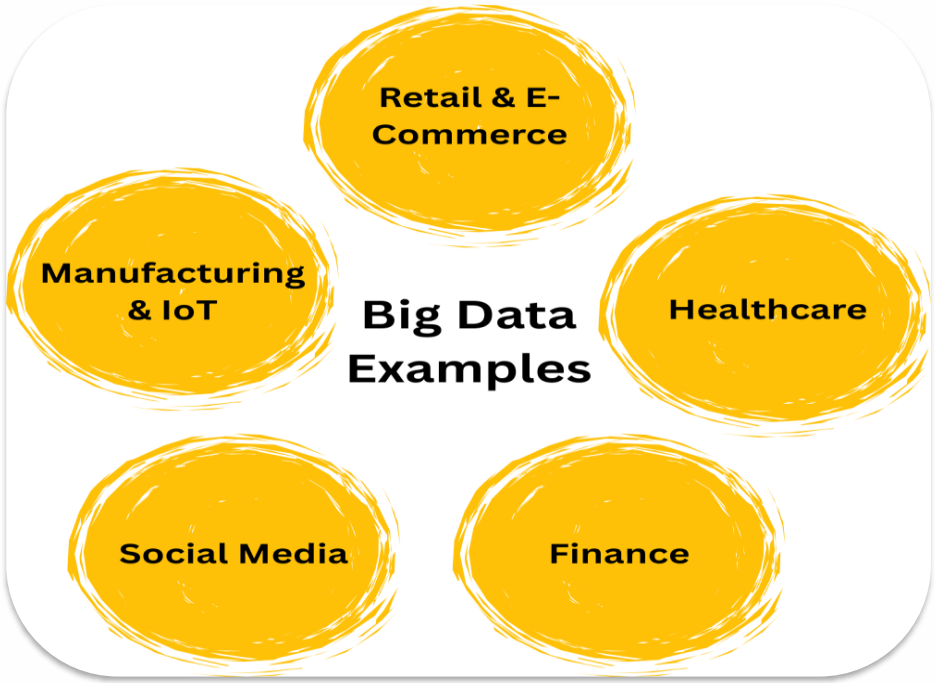
Sentiment analysis, trend detection, user engagement

Retail

Customer behavior analysis, inventory management, personalization

Real-World Impact

- Processing petabytes of data
- Real-time insights
- Data-driven decision making



Mind Map - Anatomy of Hadoop

Core Framework

Three pillars enabling distributed big data processing and storage

Ecosystem Integration

Hive, Pig, HBase, Spark, ZooKeeper, and more tools seamlessly integrate with core components

HDFS

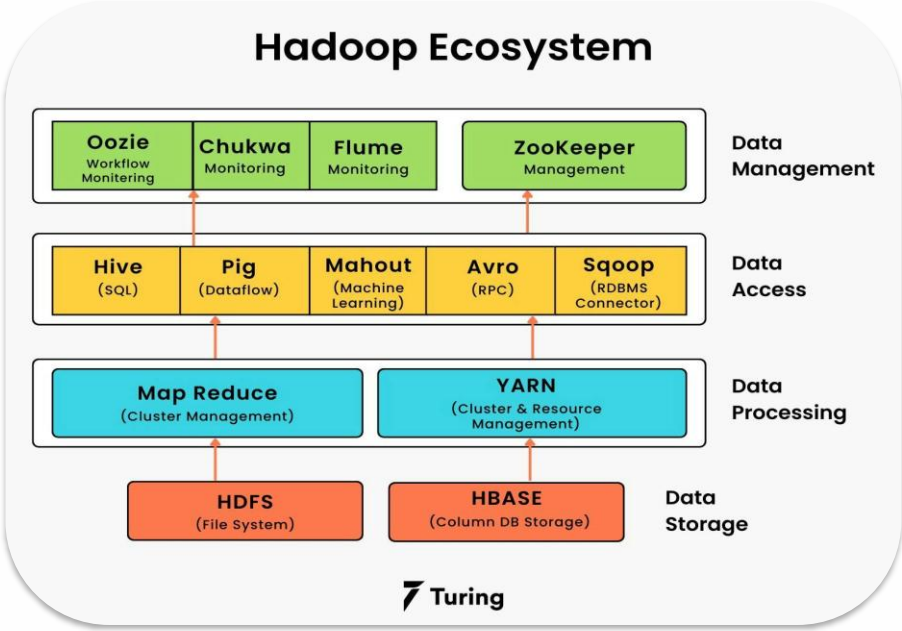
Storage layer • Metadata management • Replication

MapReduce

Processing engine • Parallel computation • Data transformation

YARN

Resource management • Job scheduling • Multi-tenancy



Summary and Recap

Core Takeaways

Understanding Hadoop's architecture and its role in big data processing

Distributed Framework

Provides scalable storage and processing capabilities

Three Core Components

HDFS, MapReduce, and YARN work together seamlessly

Rich Ecosystem

Extensive tools and frameworks for diverse applications