Managing Hadoop File

Execution of Hadoop Jobs

23MCT305 - Data Analytics in Automation System

Faculty Name: N. KARTHI, AP/MCT

Introduction to Hadoop

Key Concepts

Distributed Computing

Processes data across multiple nodes

Scalable Storage

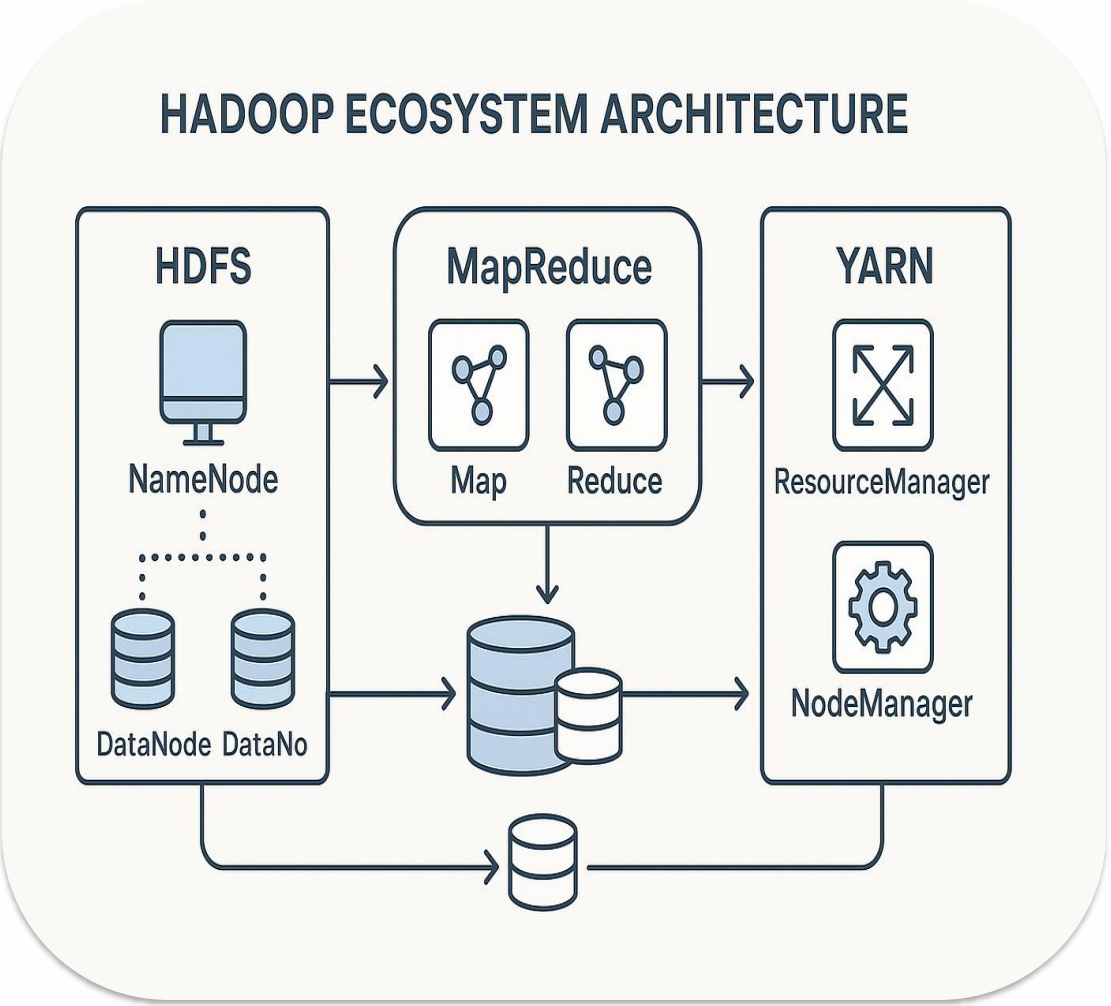
Handles petabytes of data efficiently

Fault Tolerance

Automatic data replication and recovery

Open Source

Free and community-driven development



Hadoop File System (HDFS) Architecture

Key Components

NameNode

Master server managing metadata

Master Node

DataNodes

Worker nodes storing data blocks

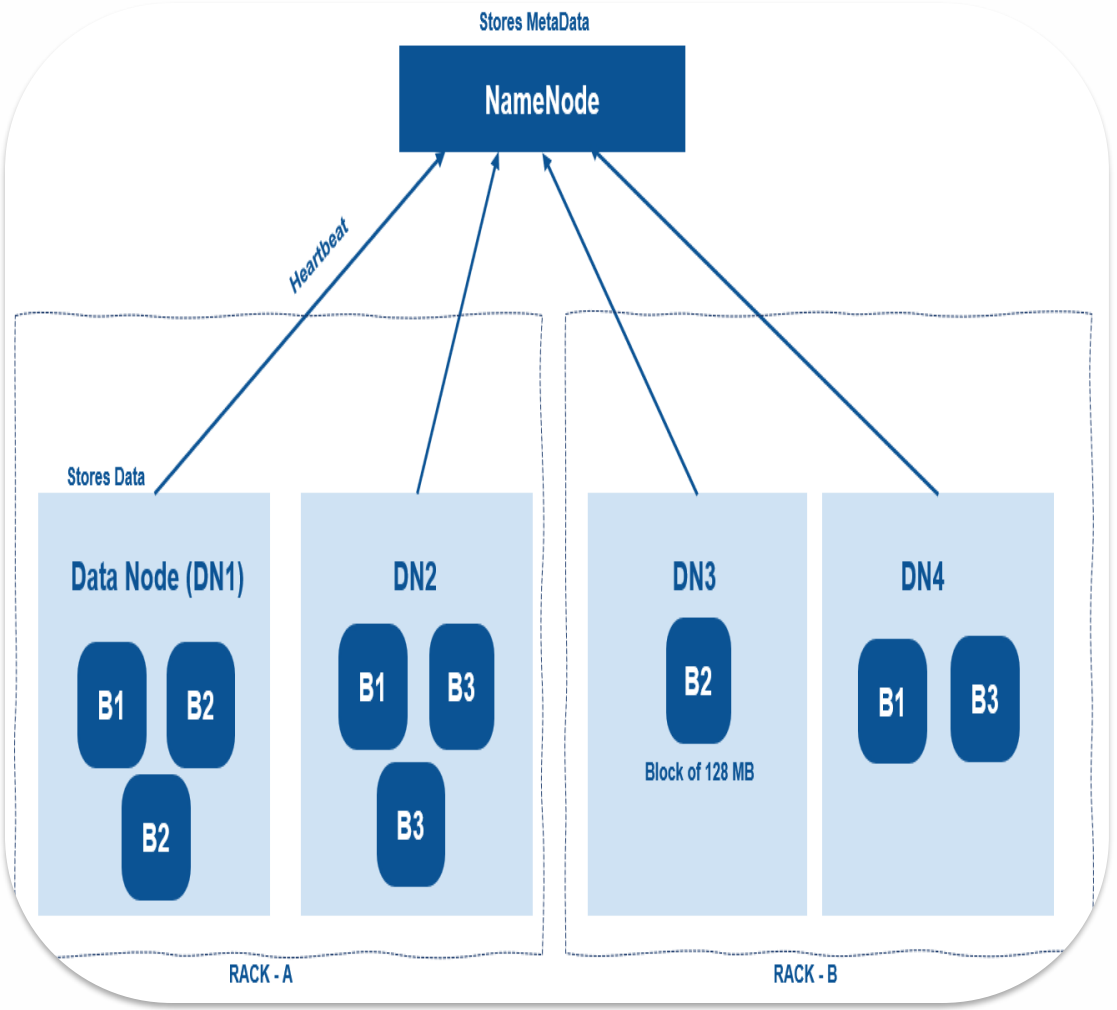
Worker Nodes

Secondary NameNode

Checkpoints and backup for NameNode

Blocks

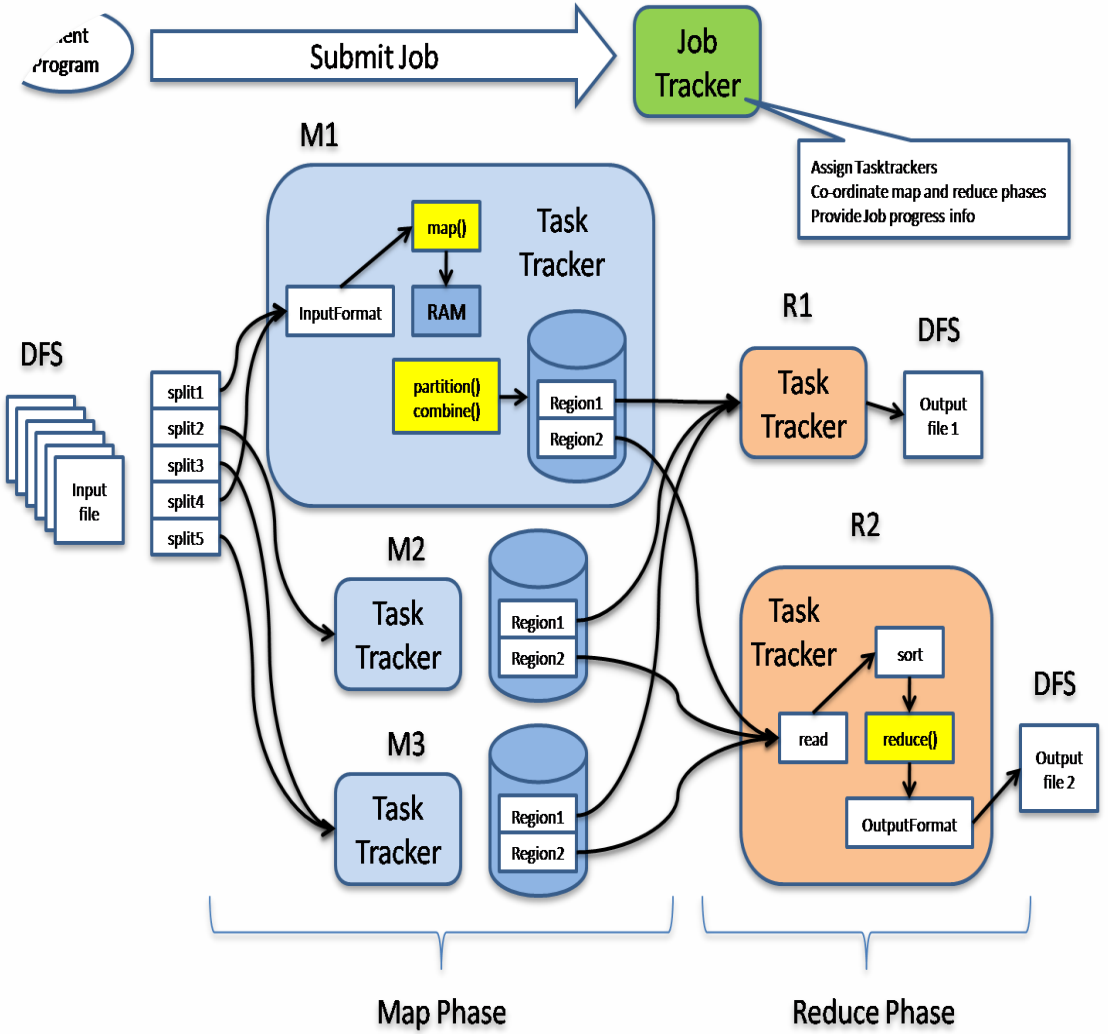
Default block size: 128 MB



Hadoop Job Execution Process

MapReduce Workflow

1. **Input Split**
   Divides input data
2. **Map**
   Processes key-value pairs
3. **Shuffle & Sort**
   Redistributes intermediate data
4. **Reduce**
   Aggregates and produces output
5. **Output**
   Final results stored



YARN Components

- ResourceManager
- NodeManager
- AppMaster

Managing Hadoop Files - Basic Operations

Essential Commands

hadoop fs -ls

List files and directories

hadoop fs -mkdir

Create new directory

hadoop fs -put

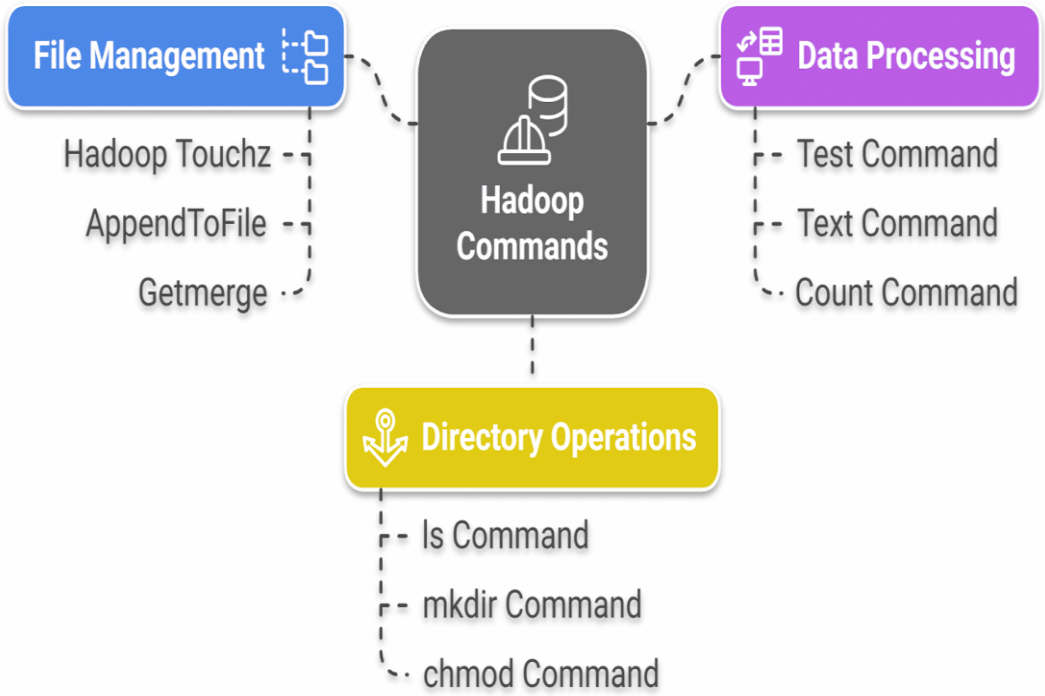
Upload files to HDFS

hadoop fs -get

Download files from HDFS

hadoop fs -rm

Delete files or directories



Managing Hadoop Files - Advanced Operations

Advanced Commands

hadoop fs -copyFromLocal

Copy local file to HDFS

hadoop fs -copyToLocal

Copy HDFS file to local system

hadoop fs -mv

Move or rename files/directories

hadoop fs -chmod

Change file/directory permissions

hadoop fs -chown

Change file/directory owner

hadoop fs -du

Display disk usage statistics

Command Usage Examples

Copy Local to HDFS

```
hadoop fs -copyFromLocal
  input.txt /user/hadoop/data/
```

Copy HDFS to Local

```
hadoop fs -copyToLocal
  /user/hadoop/output result.txt
```

Check Disk Usage

```
hadoop fs -du /user/hadoop/data
```

YARN Architecture

ResourceManager

Global resource scheduling

NodeManager

Per-node resource management

ApplicationMaster

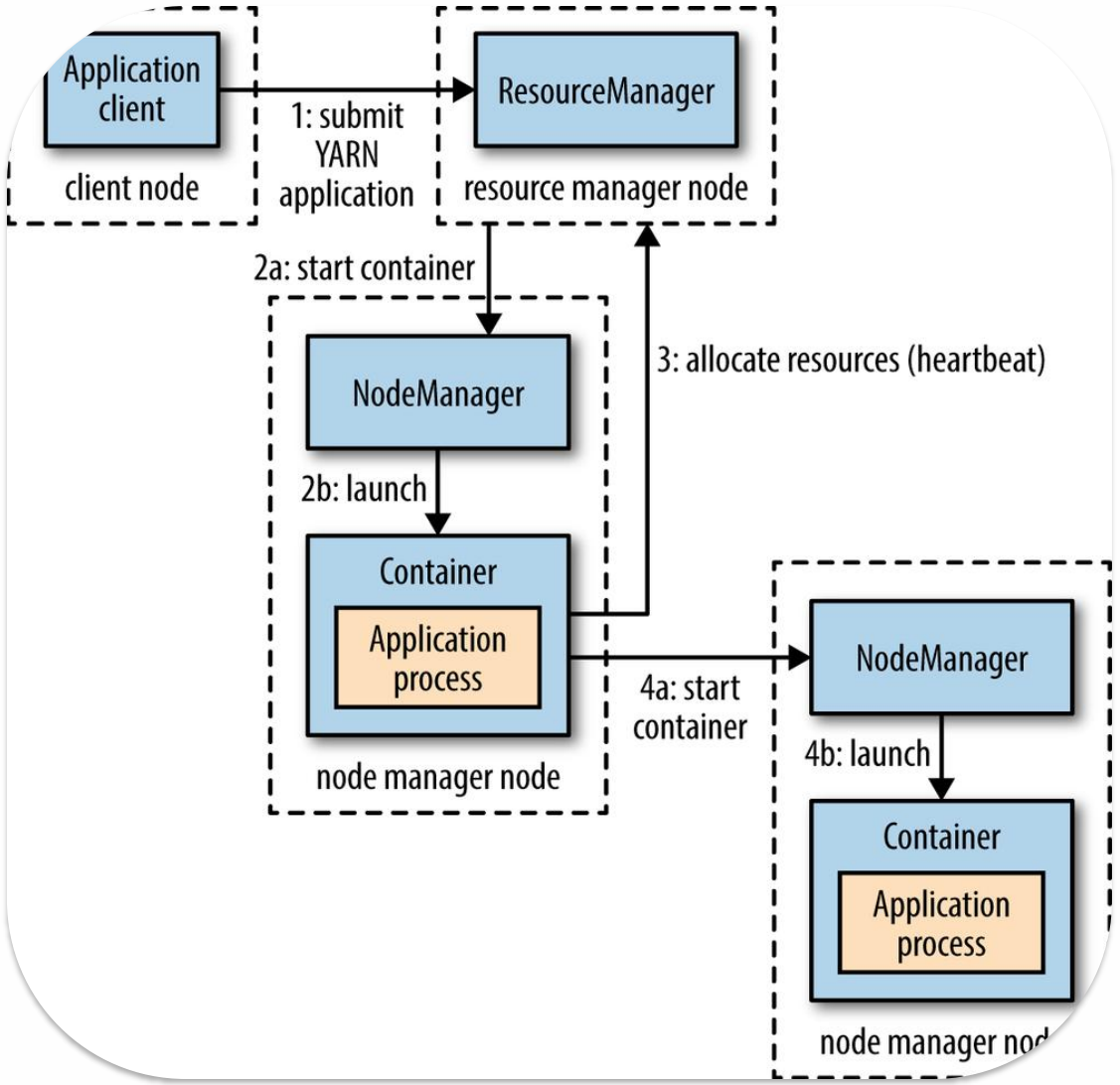
Application-specific scheduling

Containers

Resource allocation units

Scheduler Types

FIFO | Fair | Capacity



Hadoop Job Monitoring and Troubleshooting

Monitoring Tools

Web UI

JobTracker, NameNode dashboards

CLI Commands

job -list, job -status

Log Files

stdout, stderr, syslog

Common Issues & Solutions

Slow performance

Optimize cluster configuration

Failed tasks

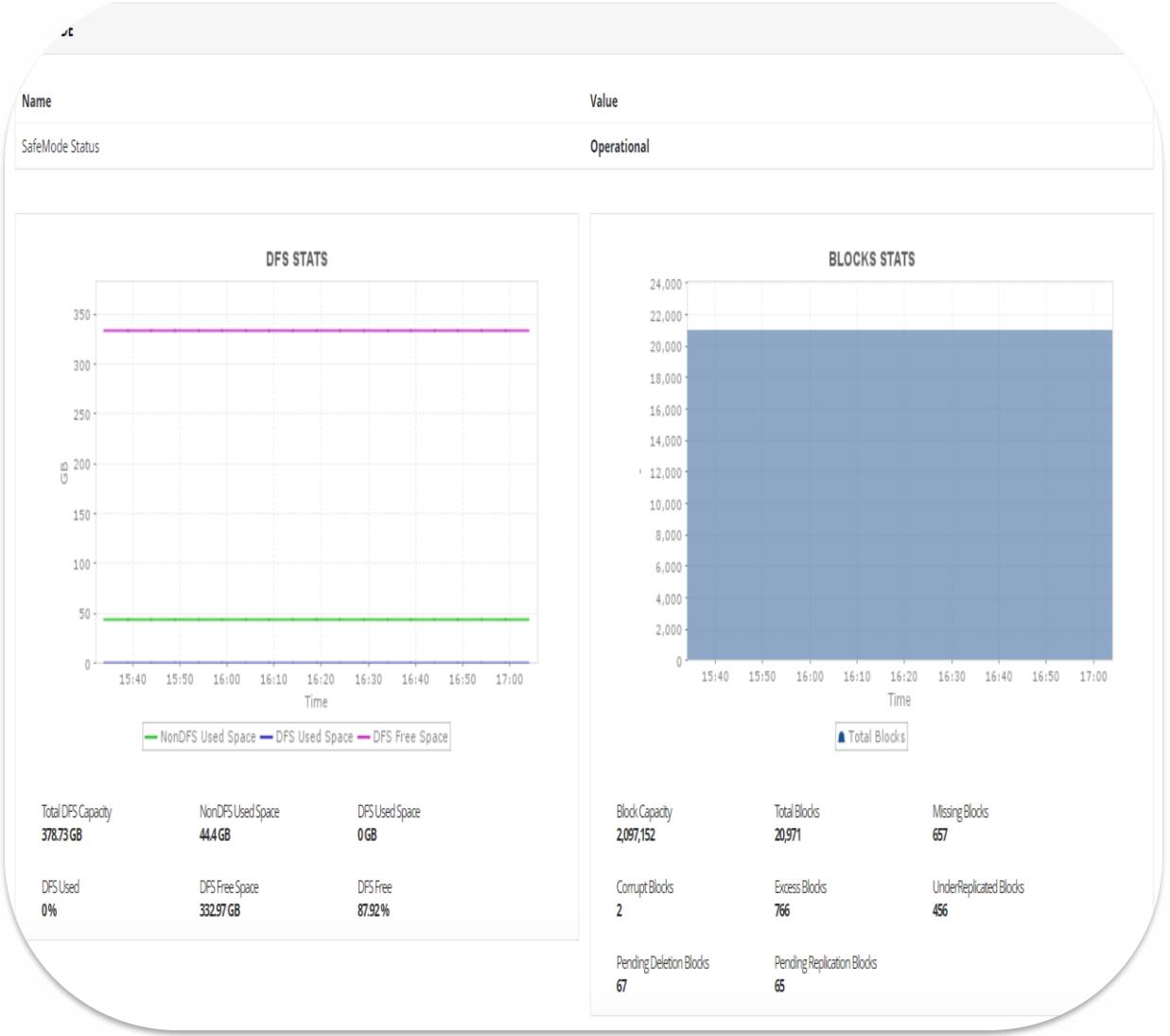
Check logs, tune parameters

Memory errors

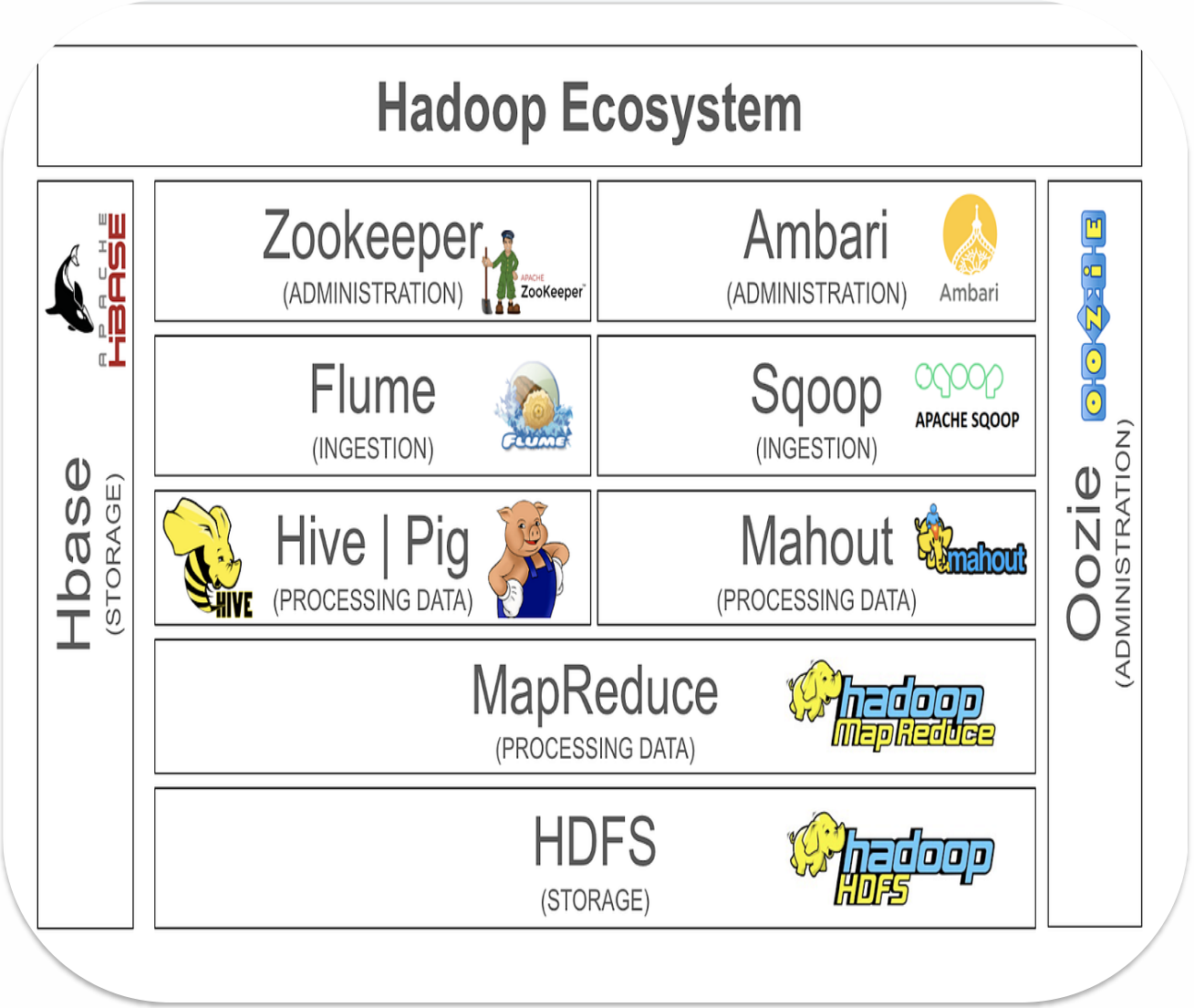
Increase heap size, optimize code

Network problems

Monitor connections, check timeouts



Mind Map of Hadoop File and Job Management



Core Components

HDFS Operations

Basic & Advanced File Management

Job Execution

MapReduce & YARN Framework

Monitoring

Tools, Metrics & Dashboards

Troubleshooting

Common Issues & Solutions

Best Practices

Optimization & Security

Recap and Summary

Key Points

- HDFS provides distributed storage with fault tolerance
- MapReduce processes large datasets in parallel
- YARN manages resources and job scheduling
- File operations use hadoop fs commands
- Monitoring ensures optimal performance
- Troubleshooting resolves common issues

Learning Outcomes

- **Understanding Hadoop architecture**
- **Managing HDFS files efficiently**
- **Executing jobs optimally**
- **Monitoring and debugging effectively**