Introduction to Hadoop

Open-Source Framework for Distributed Storage and Processing

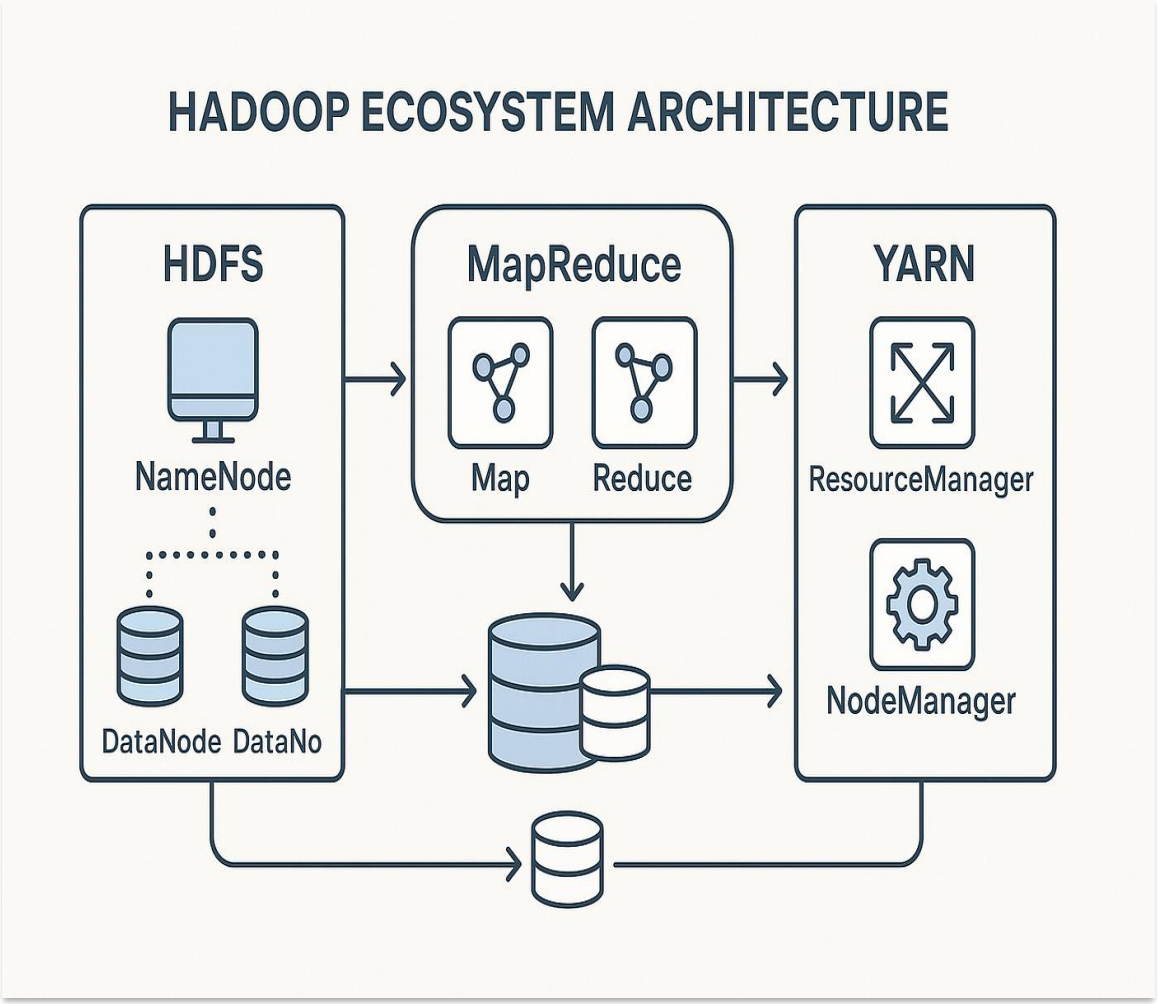
What is Hadoop?

Definition

Open-source framework for distributed storage and processing

Handles **large datasets** across clusters of computers

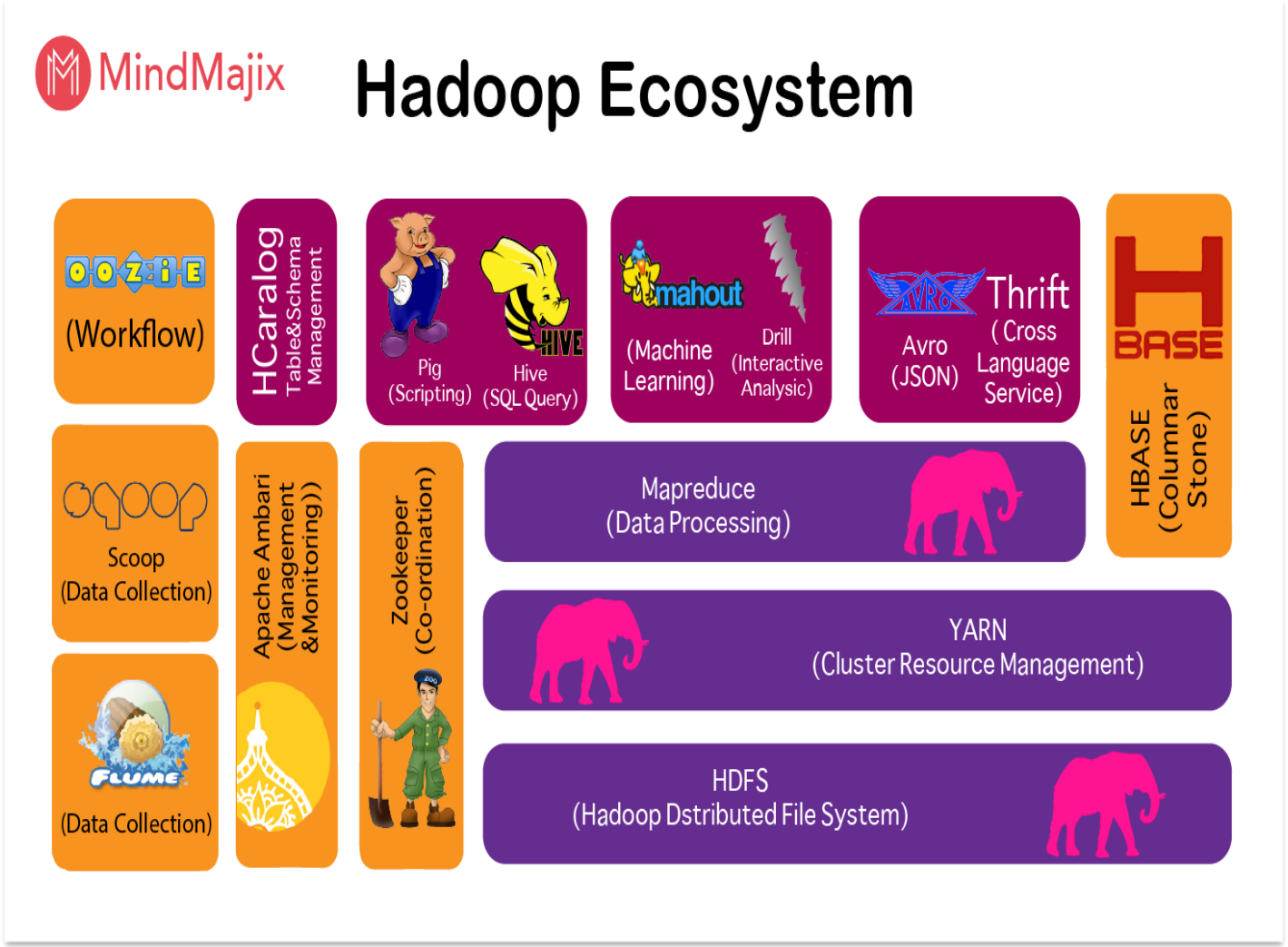
Designed to **scale up** from single servers to thousands



Hadoop Ecosystem

Key Components

- HDFS
- MapReduce
- YARN
- Hive
- Pig
- HBase
- Spark



Hadoop Core Components

HDFS

Hadoop Distributed File System

- Distributed **storage**
- High **throughput**
- Fault **tolerance**
- Scales to **petabytes**

MapReduce

Processing Framework

- Parallel **processing**
- Map + **Reduce** phases
- Large-scale **analytics**
- Batch **processing**

HDFS Architecture

Key Features

Distributed Storage

Files split into blocks

Block Replication

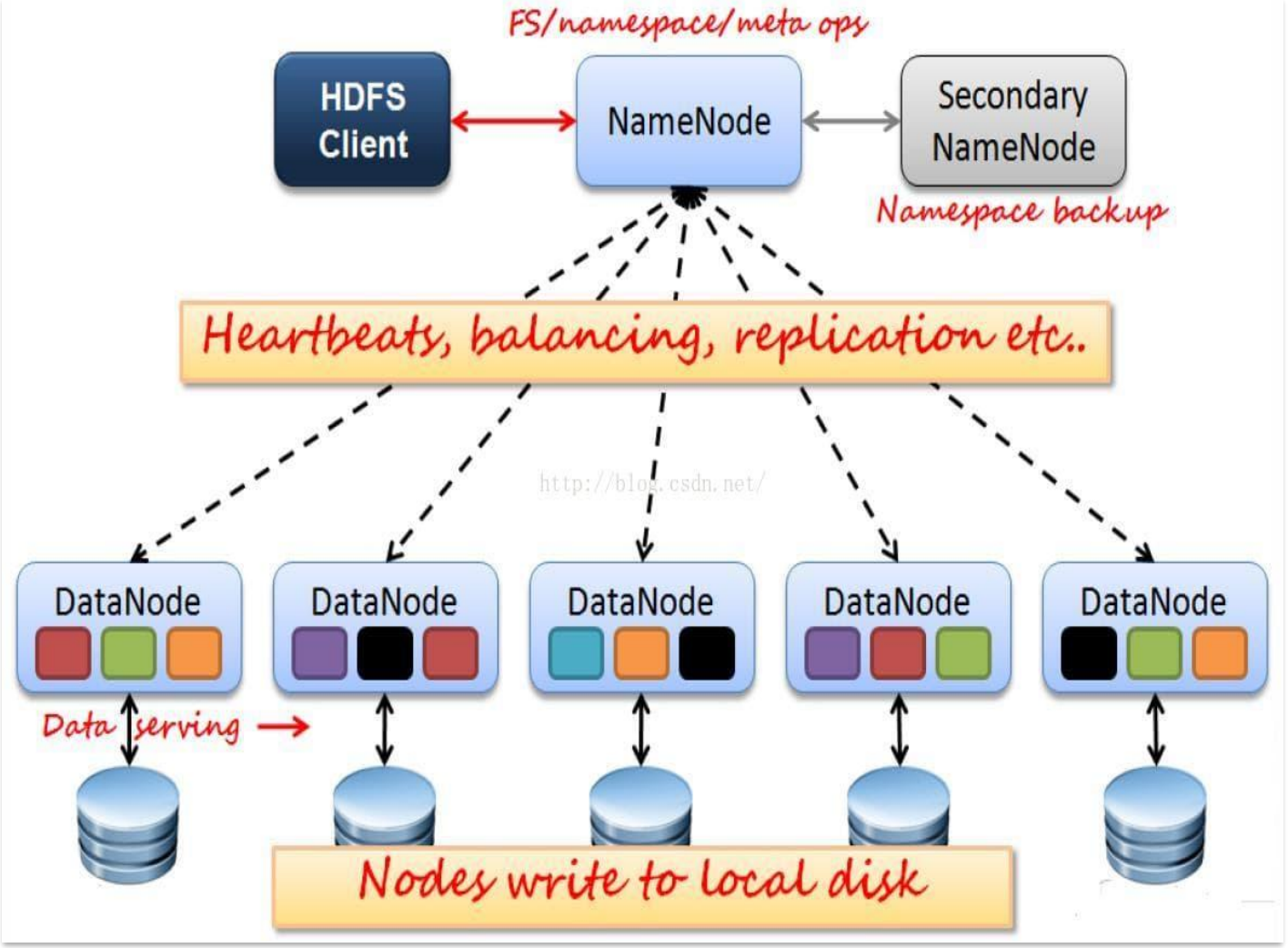
Default replication factor: 3

Fault Tolerance

Automatic recovery from failures

High Throughput

Optimized for batch processing



MapReduce Workflow

Processing Phases

Input Split

Data divided into splits

Map Phase

Process key-value pairs

Shuffle & Sort

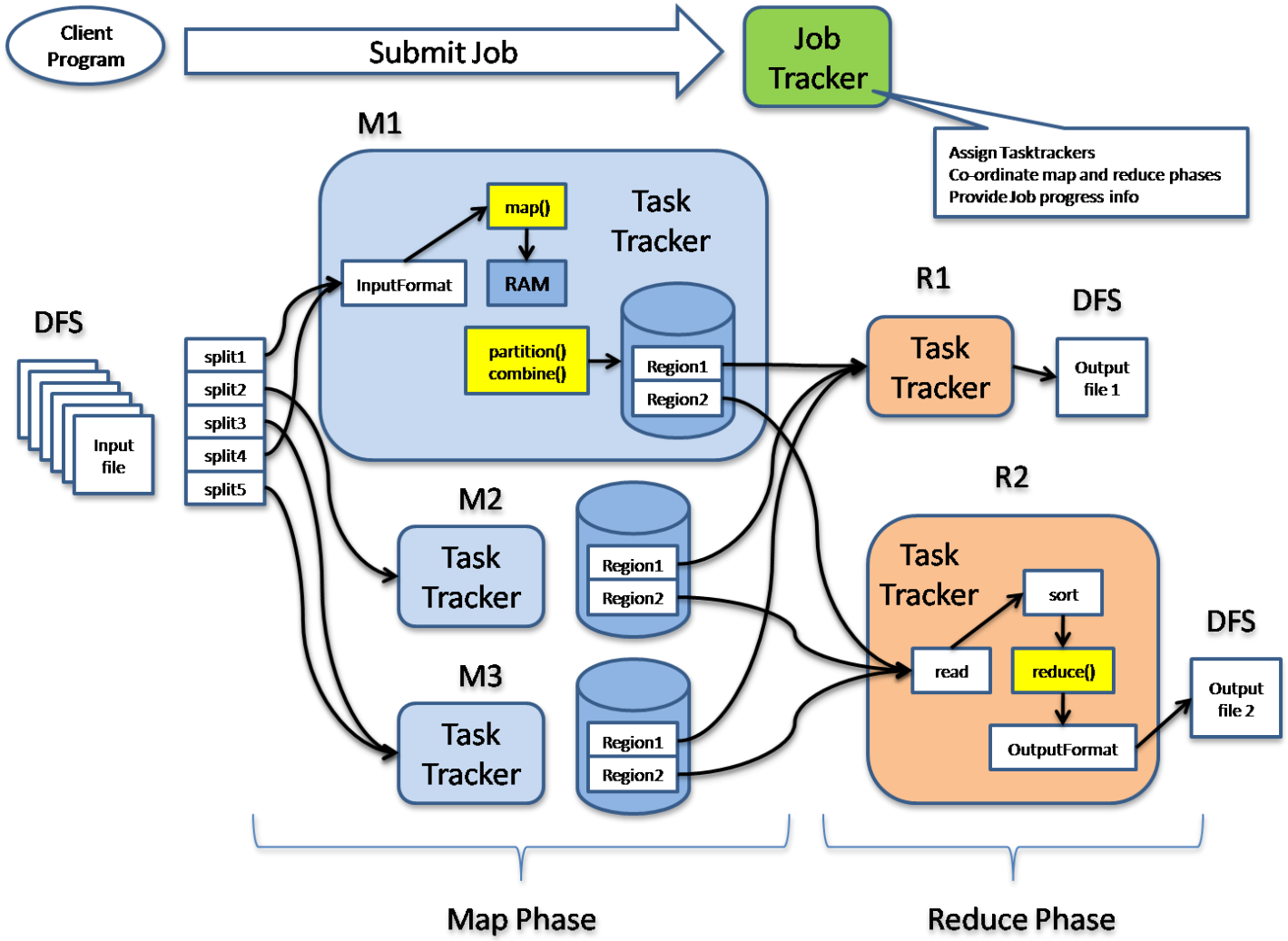
Intermediate data transfer

Reduce Phase

Aggregate results

Output

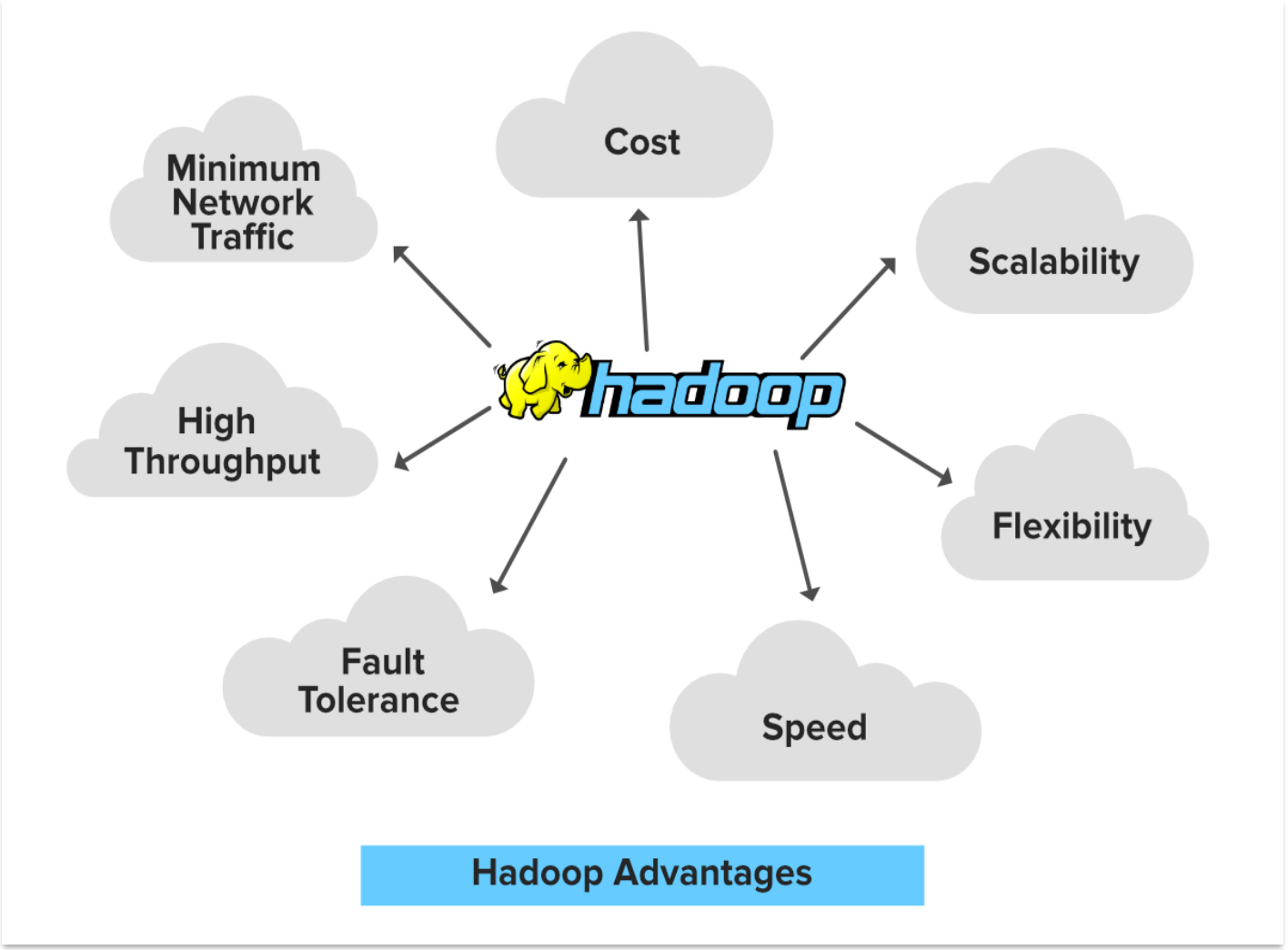
Final result stored



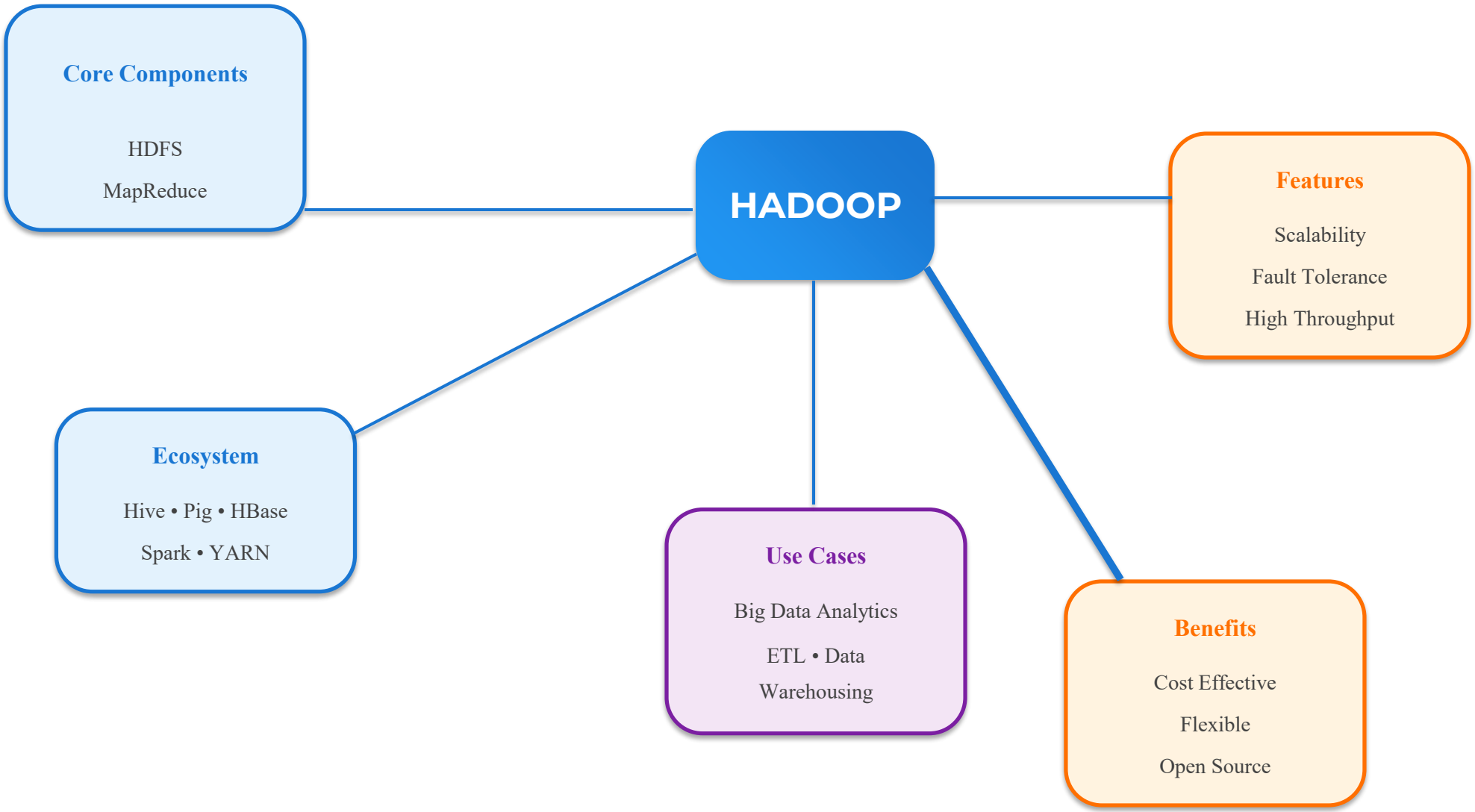
Hadoop Advantages

Key Benefits

- Cost-effective
- High scalability
- Flexible processing
- Fast processing
- Fault tolerance
- High throughput
- Minimized network traffic



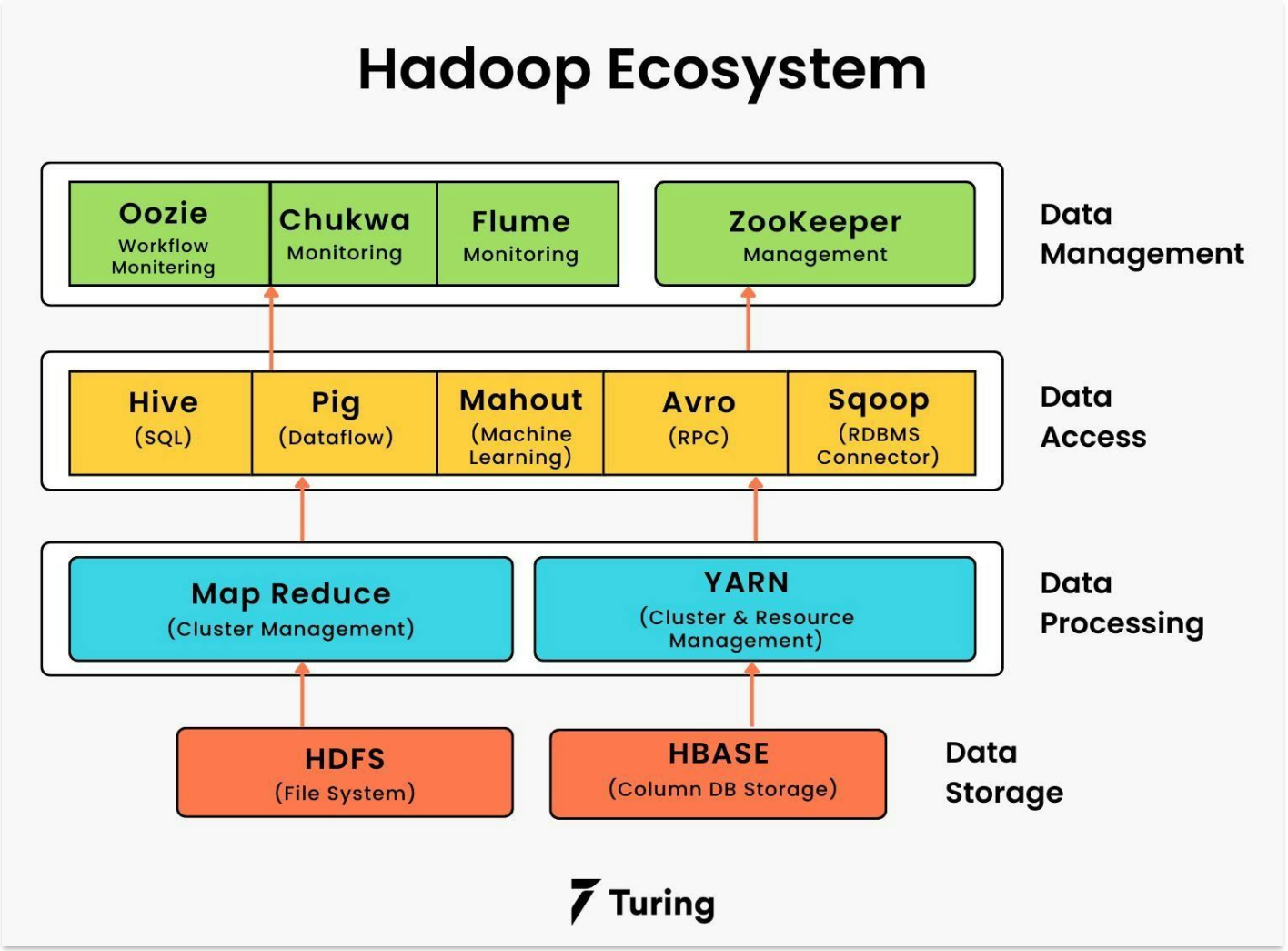
Hadoop Mind Map



Recap of Key Points

What We Learned

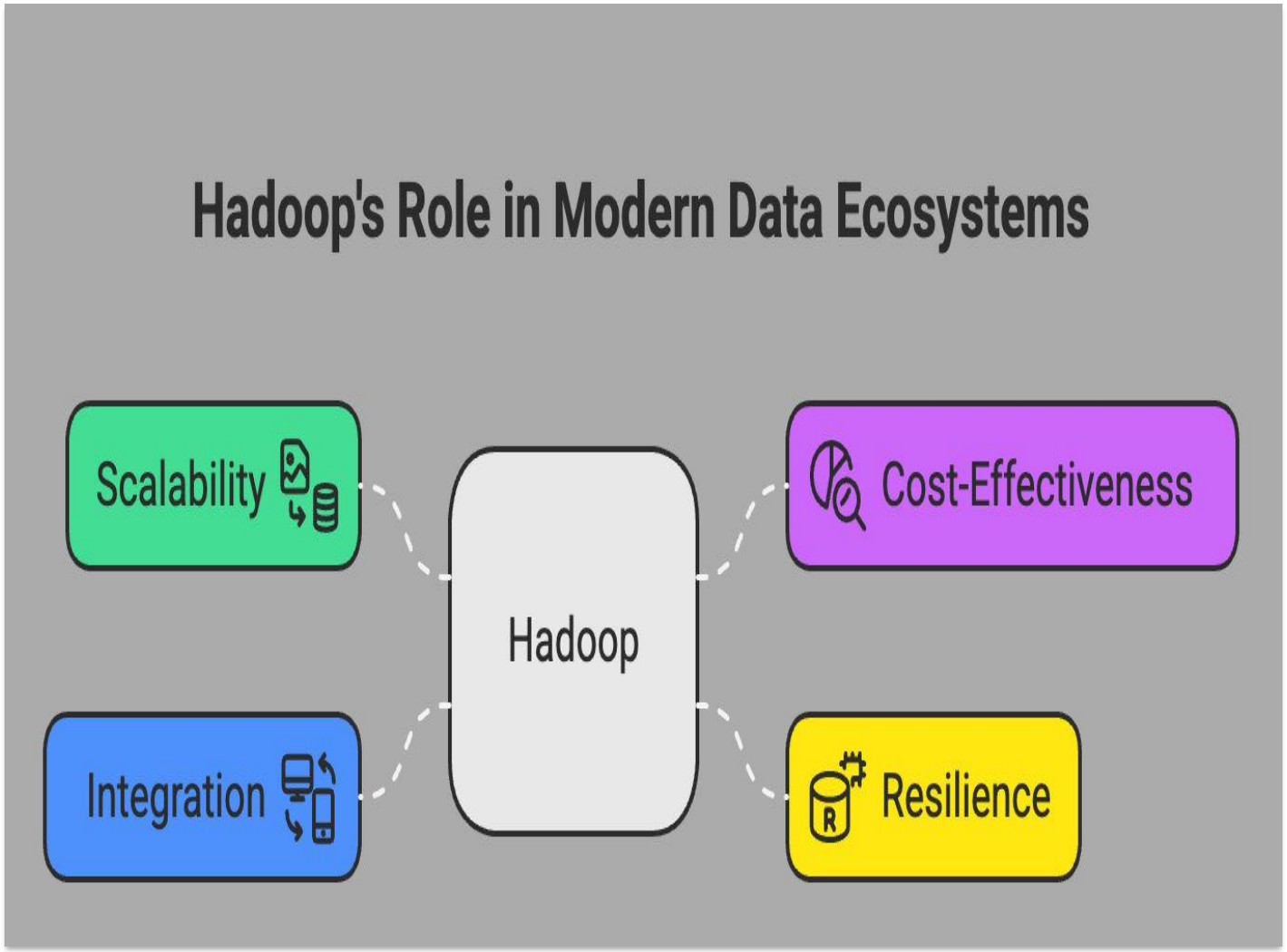
- What is Hadoop?
- Hadoop Ecosystem
- Core Components
- HDFS Architecture
- MapReduce Workflow
- Hadoop Advantages



Summary and References

Key Takeaways

- Scalable storage & processing
- Fault-tolerant architecture
- Cost-effective solution
- Distributed computing power



References

- Apache Hadoop Documentation
- O'Reilly Hadoop Guide
- Big Data Tutorials