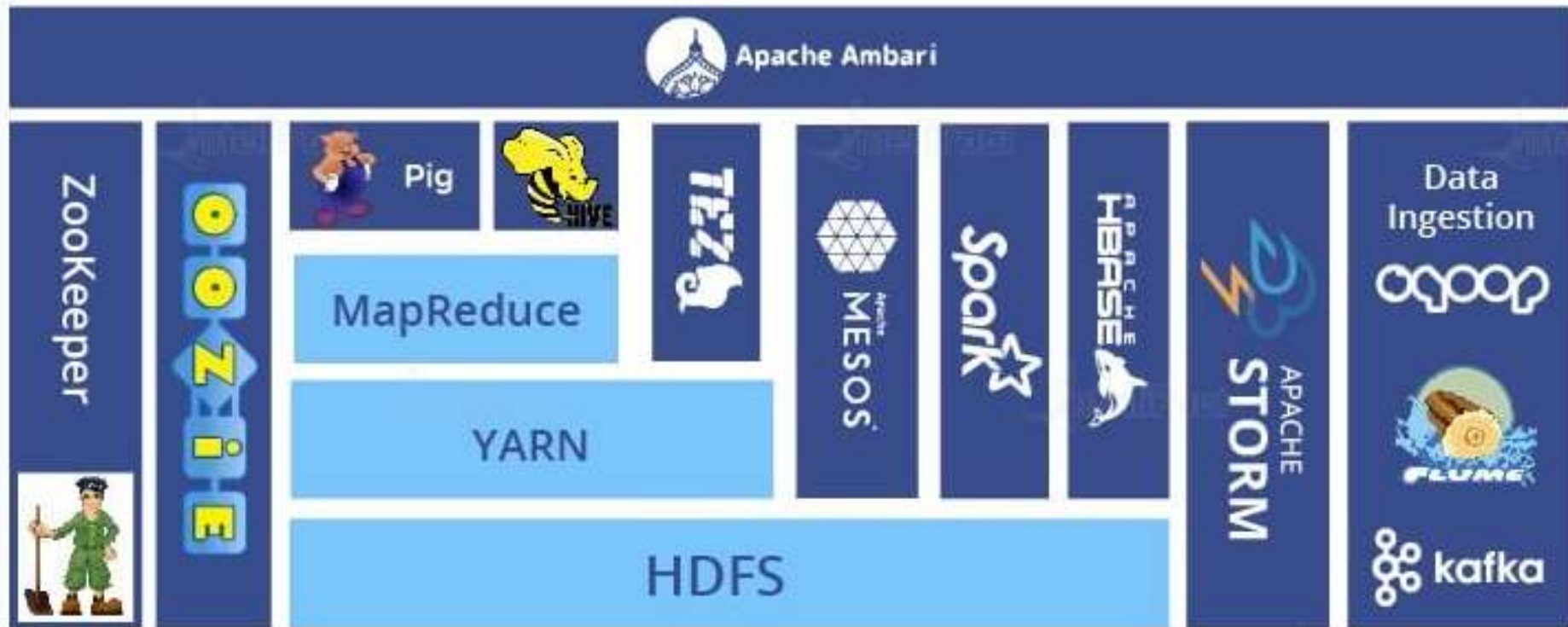


# Hadoop Ecosystem

# Hadoop Ecosystem



# HDFS (Hadoop Distributed File System)

---

Storage component of Hadoop

Stores data in the form of files

Each file is divided into blocks of 128MB (configurable)

Stores them on different machines in the cluster.

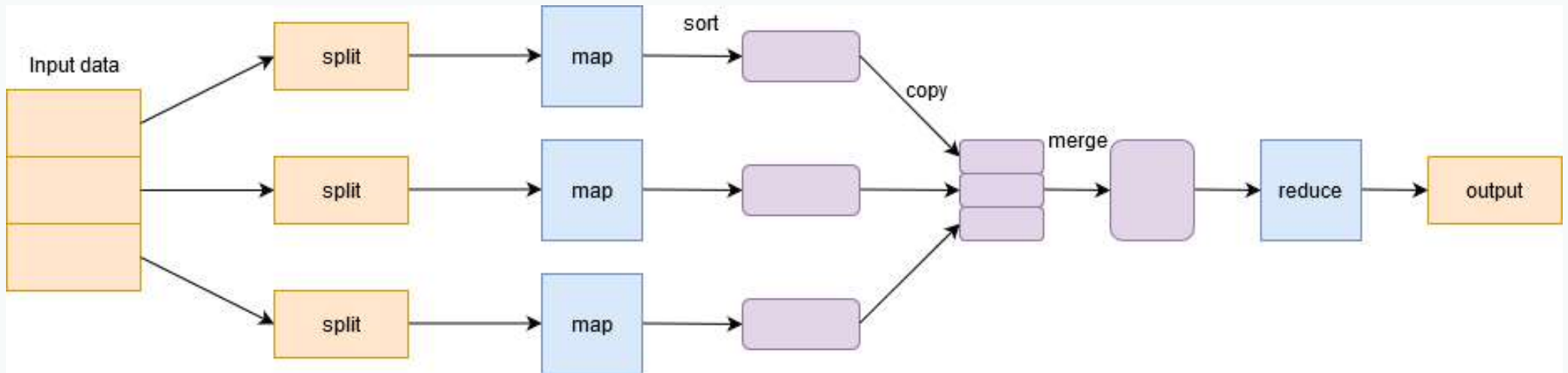
Has a master-slave architecture with two main components

- Name Node and Data Node

# MapReduce

---

- Divides a single task into multiple tasks and processes them on different machines.



- Map phase filters, groups, and sorts the data
- Reduce phase aggregates the data, summarises the result, and stores it on HDFS.

# YARN

---

- Manages resources in the cluster
- Allows data stored in HDFS to be processed and run by various data processing



# HBase

---

- Column-based NoSQL database
- Runs on top of HDFS
- Can handle any type of data



# Pig

---

- Analyses large datasets
- Overcomes difficulty to write map and reduce functions
- Consists of two components:
  - Pig Latin
  - Pig Engine
- Pig Latin
  - Scripting Language that is similar to SQL
- Pig Engine
  - Execution engine on which Pig Latin runs
- Internally, the code written in Pig is converted to MapReduce functions



# Hive

---

- Distributed data warehouse system
- Allows for easy reading, writing, and managing files on HDFS
- Has its own querying language
- Makes it easy for programmers to write MapReduce functions using simple HQL queries.

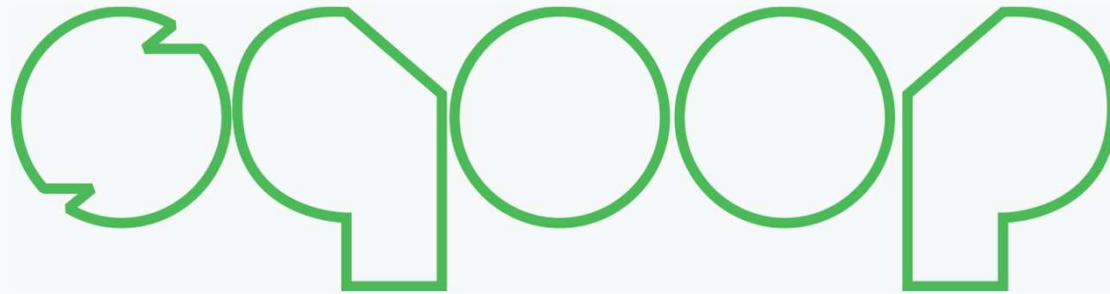




# Sqoop

---

- Bring data from Relational Databases into HDFS.
- The commands written in Sqoop internally converts into MapReduce tasks
- Can also be used to export data from HDFS to RDBMS



# Oozie

---

- A workflow scheduler system
- Allows to link jobs written on various platforms like MapReduce, Hive, Pig, etc
- Can create a pipeline of individual jobs to be executed to achieve a bigger task
- Can use Oozie to perform ETL operations



# Zookeeper

---

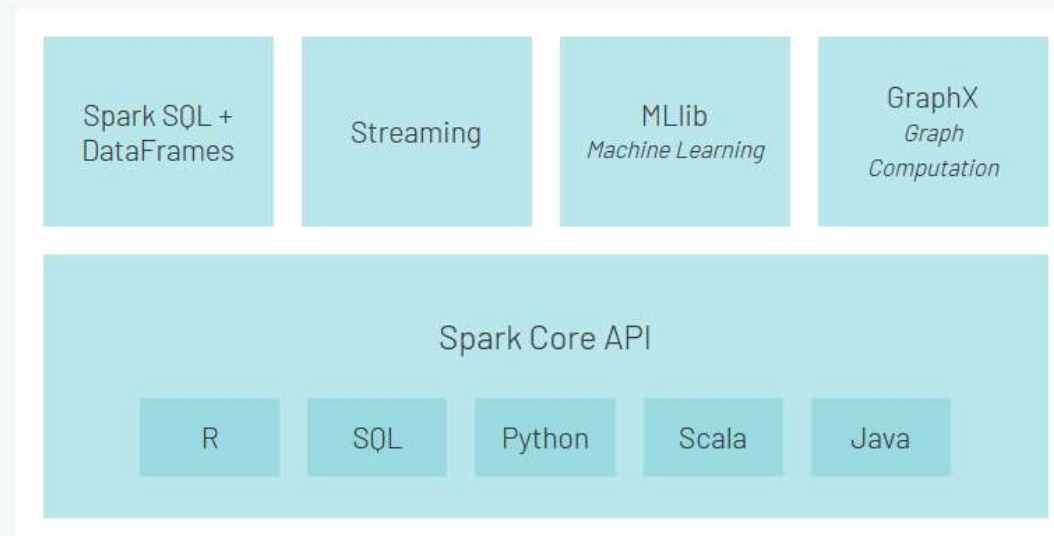
- Centralized service for
  - Maintaining configuration information
  - Naming
  - Providing distributed synchronization
  - Providing group services across the cluster.



# Spark



- Alternative framework to Hadoop
- Built on Scala
- Supports varied applications written in Java, Python, etc.
- It provides in-memory processing which accounts for faster processing



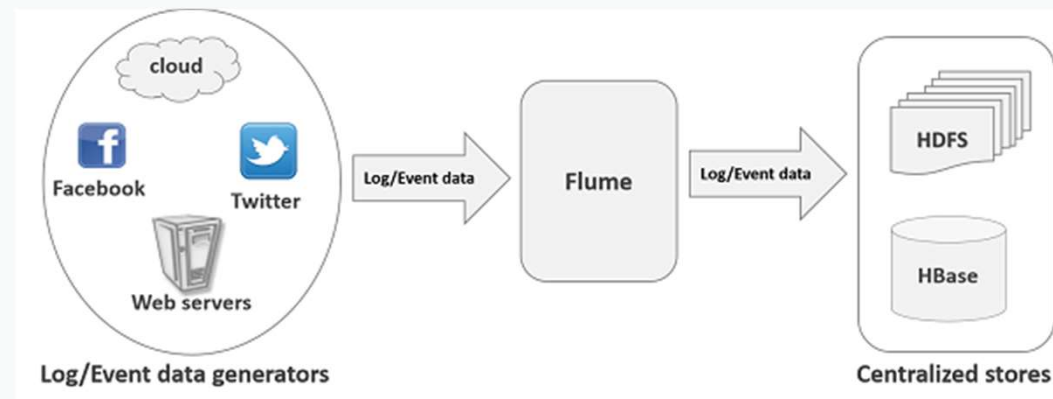


- Includes software for provisioning, managing and monitoring Apache Hadoop clusters.
- Provides:
- Hadoop cluster provisioning:
  - Gives us step by step process for installing Hadoop services across a number of hosts.
  - Also handles configuration of Hadoop services over a cluster.
- Hadoop cluster management:
  - Provides a central management service for starting, stopping and re-configuring Hadoop services across the cluster
- Hadoop cluster monitoring:
  - For monitoring health and status, Ambari provides us a dashboard

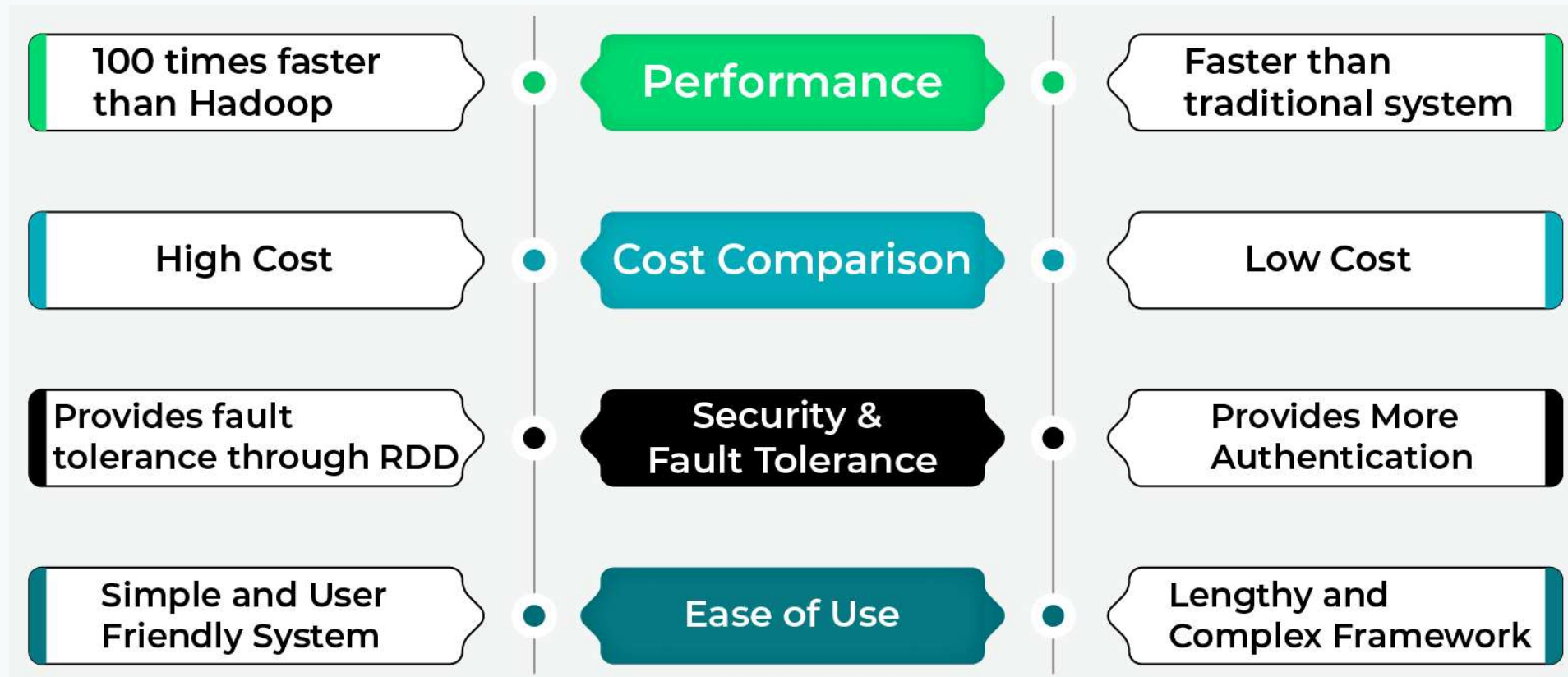
# Flume



- Assume an e-commerce web application wants to analyse the customer behaviour from a particular region
- To do so, they would need to move the available log data in to Hadoop for analysis
- Here, Apache Flume comes to our rescue.
- Flume is a system for collecting, aggregating and moving massive quantities of log data.

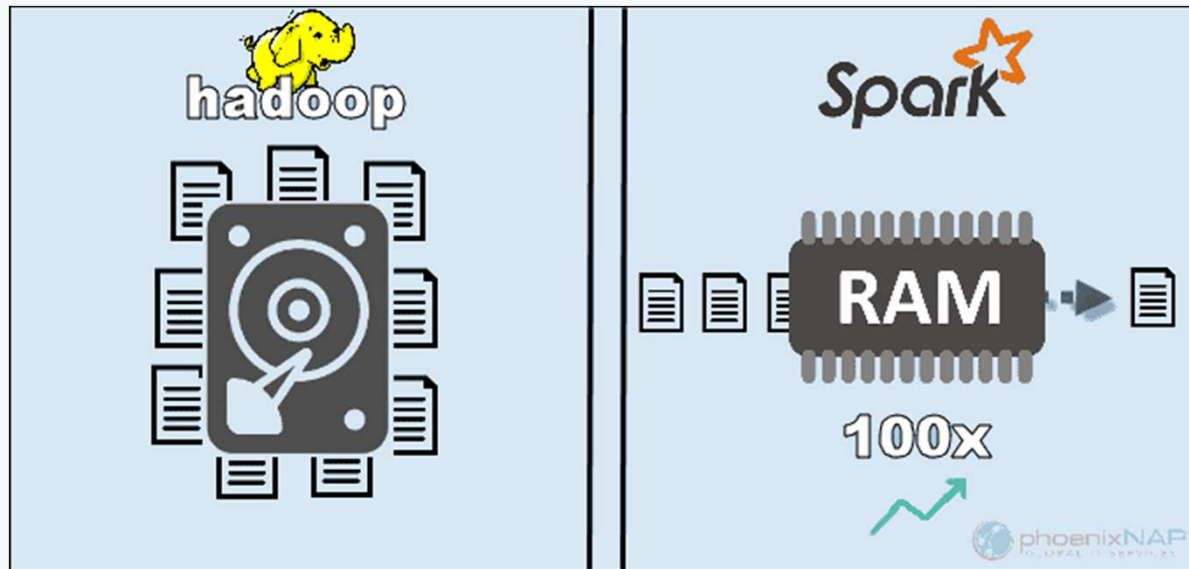


# Hadoop vs Spark



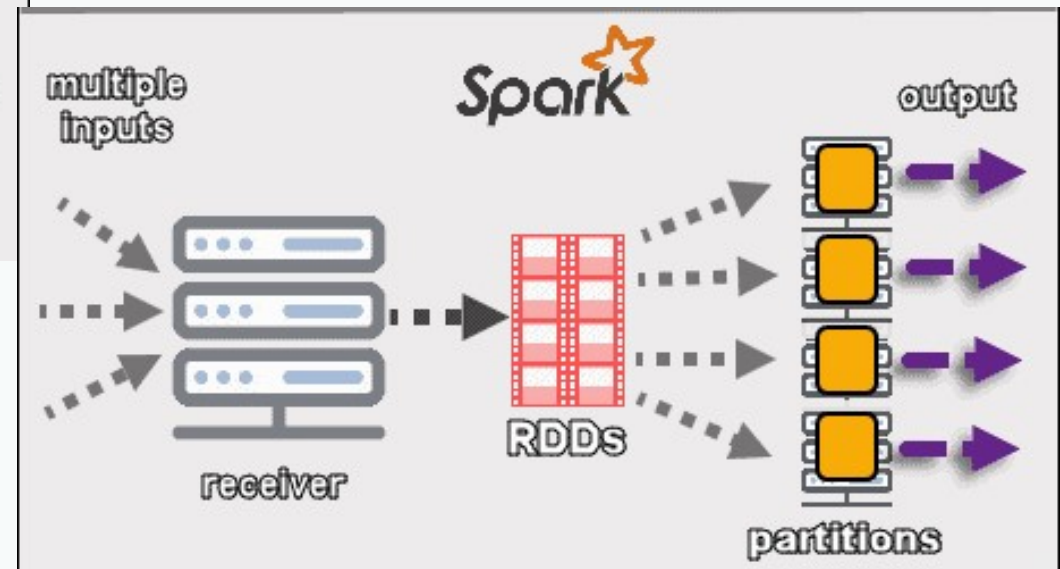
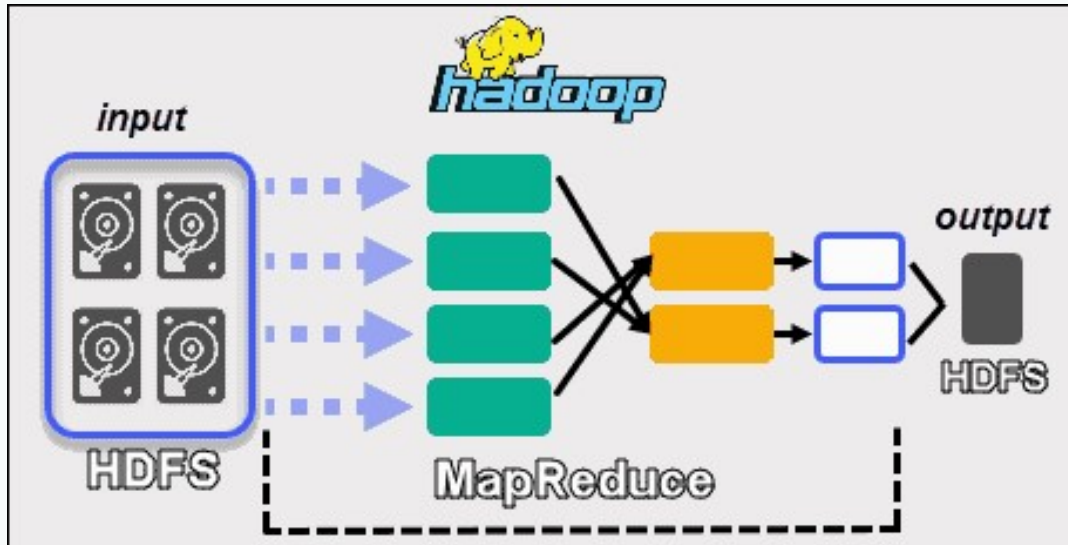
# Hadoop vs Spark

---





# Hadoop vs Spark



Thank You