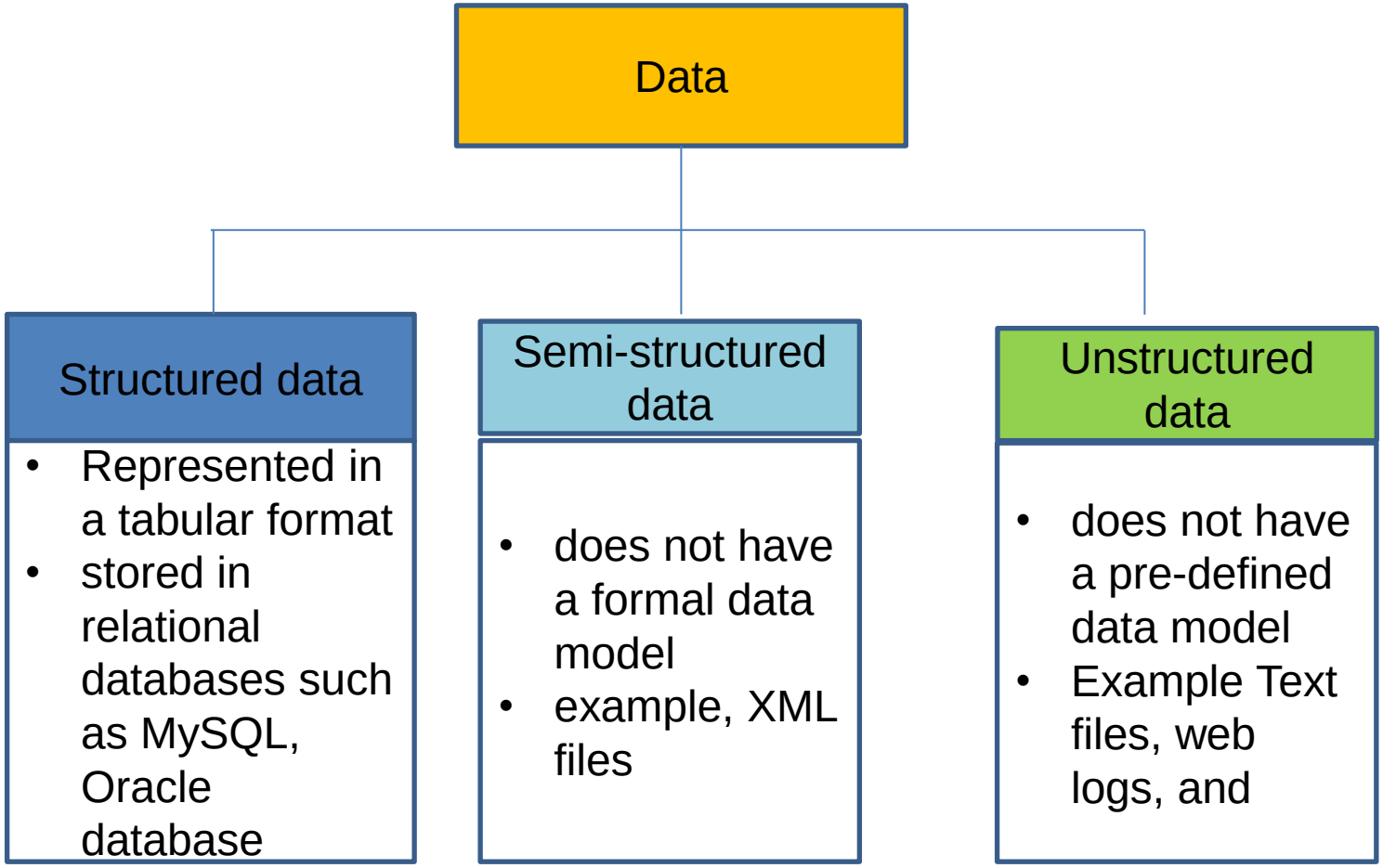
Introduction to Big Data and the Hadoop

Big Data and Hadoop Ecosystem

Evolution of Big Data

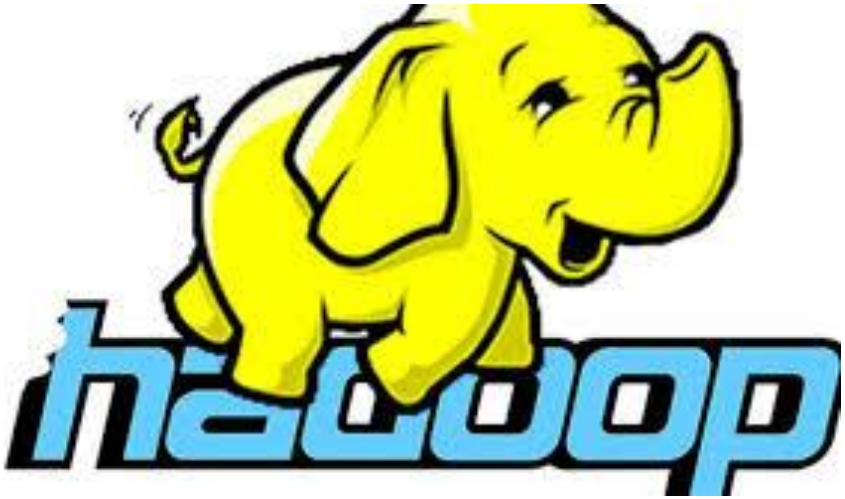
- Traditionally, computation has been Processor bound
- – Relatively small amounts of data
- – Lots of complex processing
- The early solution: bigger computers
- – Faster processor, more memory
- – But event this couldn’t keep up

Types of Data

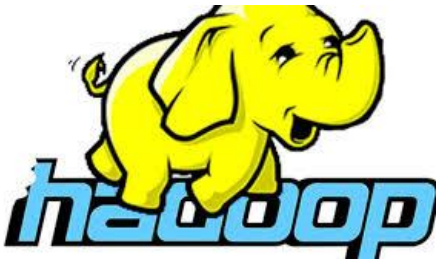


Analytics Vidhya
Learn everything about analytics

We need a new approach



Introducing Hadoop



- A radical new approach to distributed computing

– Distribute data when the data is stored
– Run computation where the data is

Google

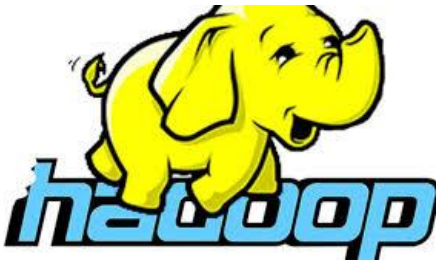
- Originally based on work done at Google

- Open-source project overseen by the Apache Software Foundation

- Doug Cutting discovered Hadoop and named it after his son's yellow-colored toy elephant.

Analytics Vidhya
Learn everything about analytics

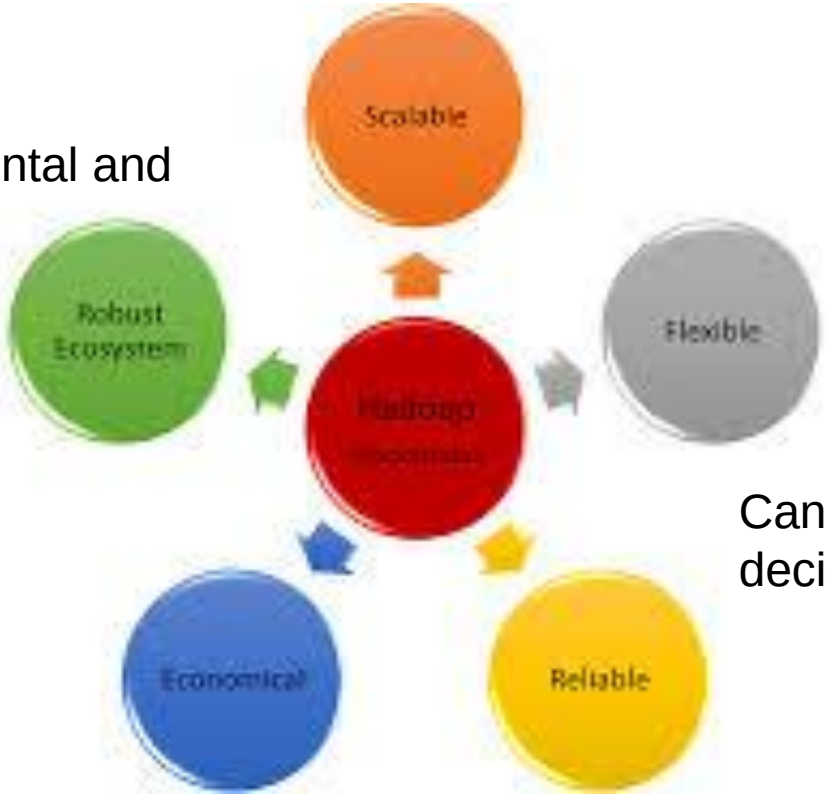
Introducing Hadoop



- Framework for solving data-intensive processes.
- Designed to scale massively
- Processes all the contents of a file (instead of attempting to read portions of a file)
- Hadoop is very fast for very large jobs
- Hadoop is not fast for small jobs
- It does not provide caching or indexing (tools like HBase can provide these features if needed)
- Designed for hardware and software failures

Hadoop Characteristics

Can follow both horizontal and vertical scaling



Can store a lot of data and decide to use it later

Ordinary computers can be used for data processing

Stores copies of the data on different machines and is resistant to hardware failure

Analytics Vidhya
Learn everything about analytics

Quiz

Do you see any common trend in the following use cases of Hadoop?

- Perception about the organization
- Risk Modelling in Banking domain

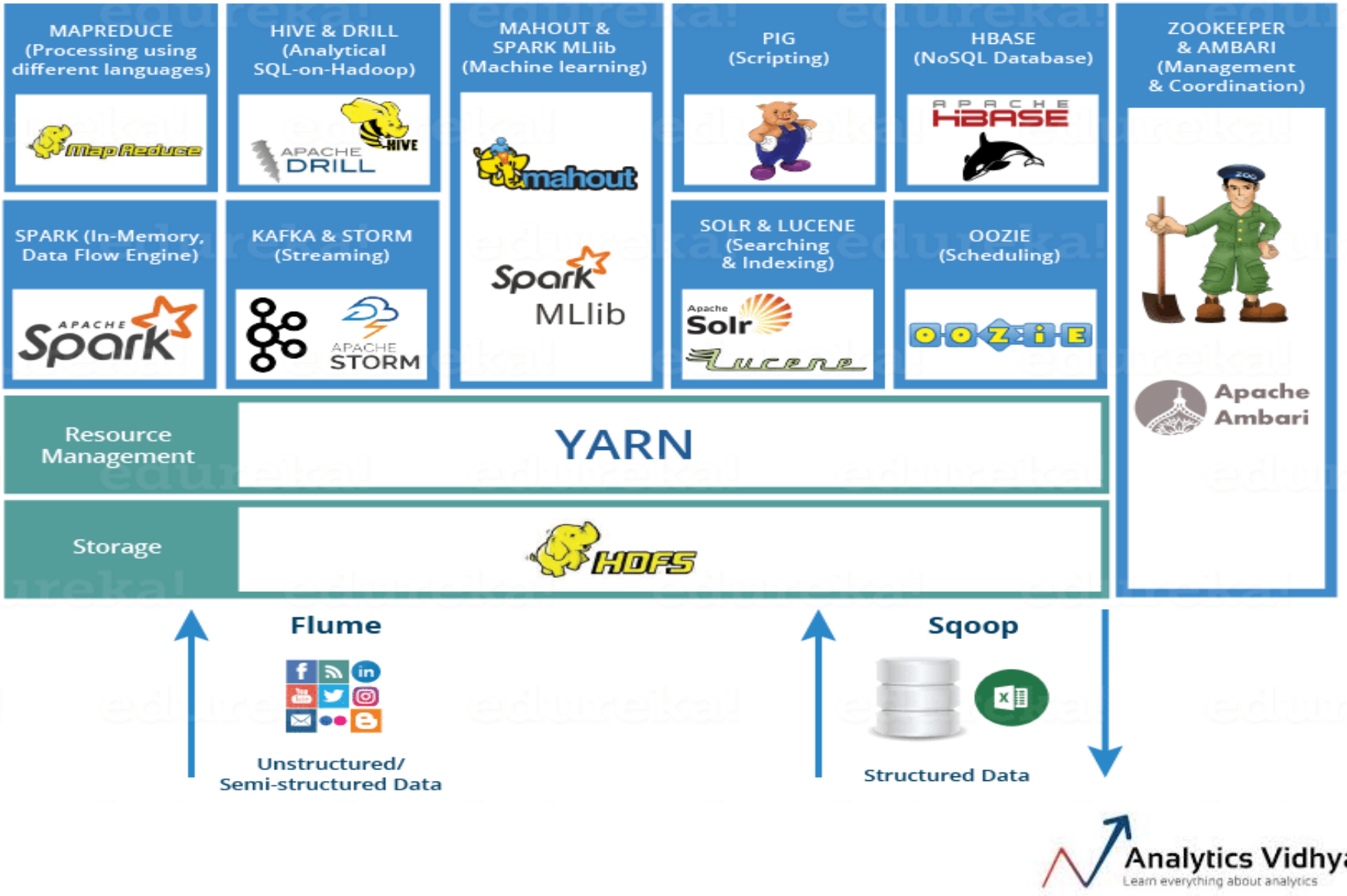
Hadoop-able Problems

- Text mining
- Index building
- Graph creation and analysis
- Pattern recognition
- Collaborative filtering
- Prediction models
- Sentiment analysis
- Risk assessment

Analytics Vidhya
Learn everything about analytics

Hadoop Components

Hadoop Ecosystem



Analytics Vidhya
Learn everything about analytics

Hadoop Ecosystem…

A storage layer for Hadoop

Hadoop provides a command line interface to interact with HDFS

Provides file permissions and authentication

Streaming access to file system data

Suitable for the distributed storage and processing

Hadoop Ecosystem…

Storesdata in HDFS

A NoSQLdatabase or non-relational database

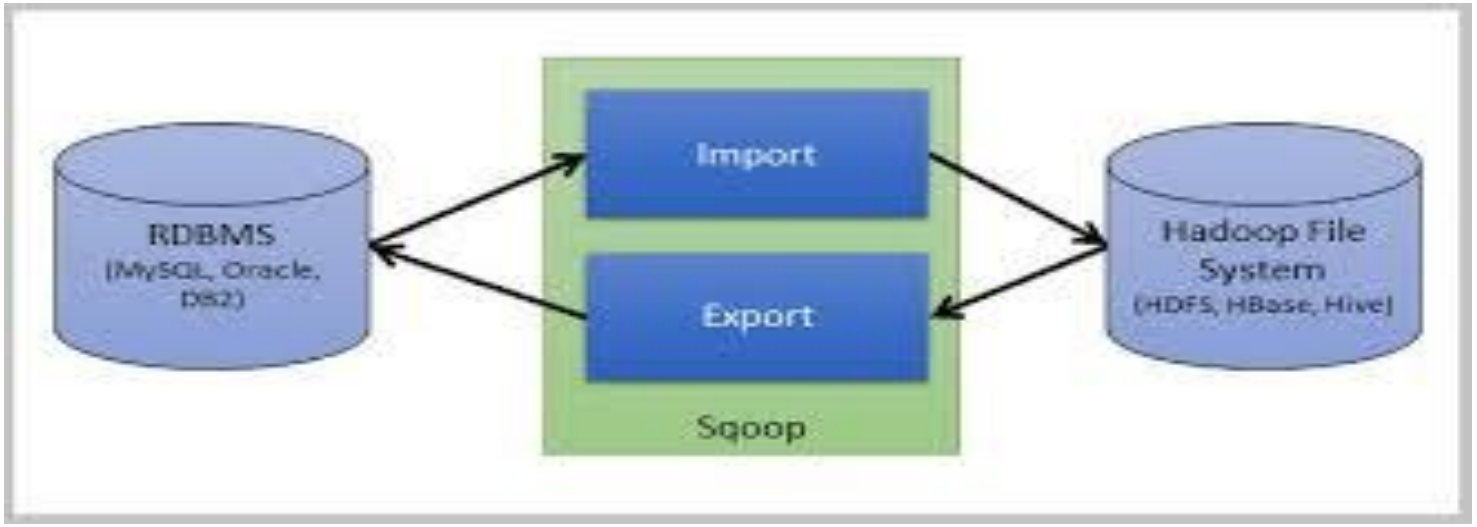
Provides support to high volume of data and high throughput

The table can have thousands of columns

Mainly used when you need random, real-time, read/write access to your Big Data

Analytics Vidhya
Learn everything about analytics

Hadoop Ecosystem…

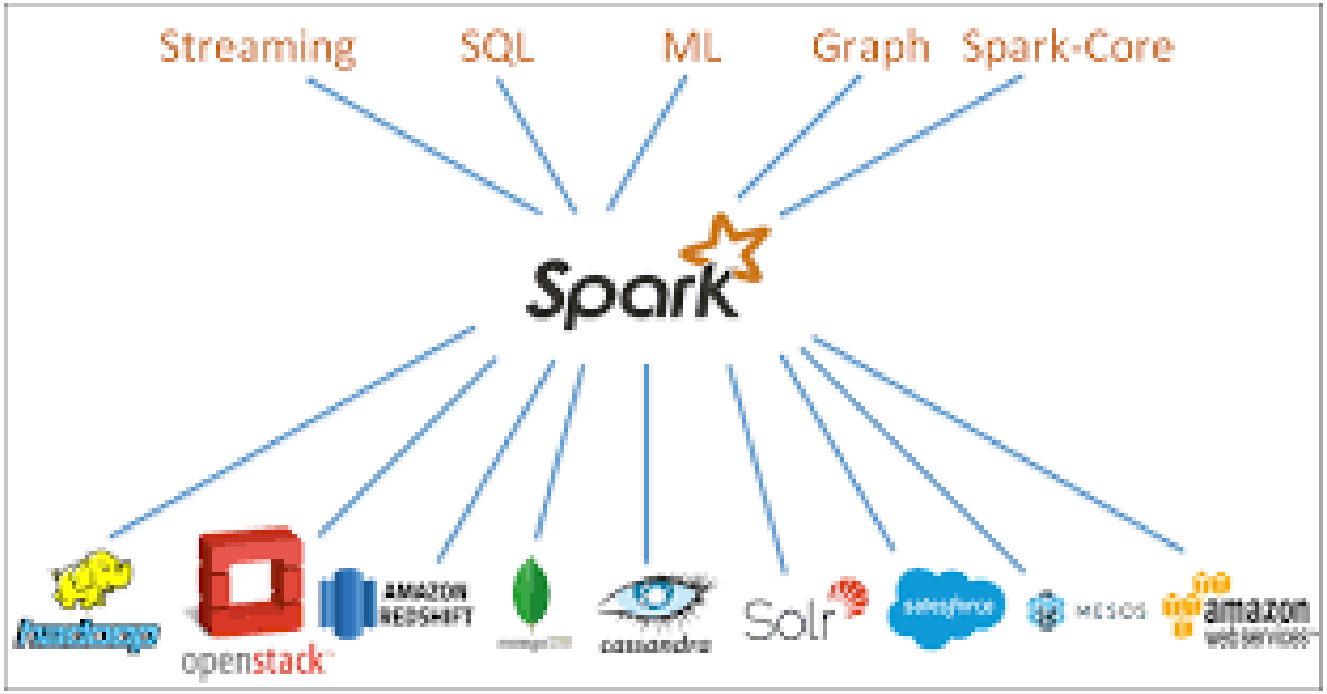


Sqoop is a tool designed to transfer data between Hadoop and relational database servers

Hadoop Ecosystem…

- A distributed service for ingesting streaming data
- Ideally suited for event data from multiple systems

Hadoop Ecosystem…

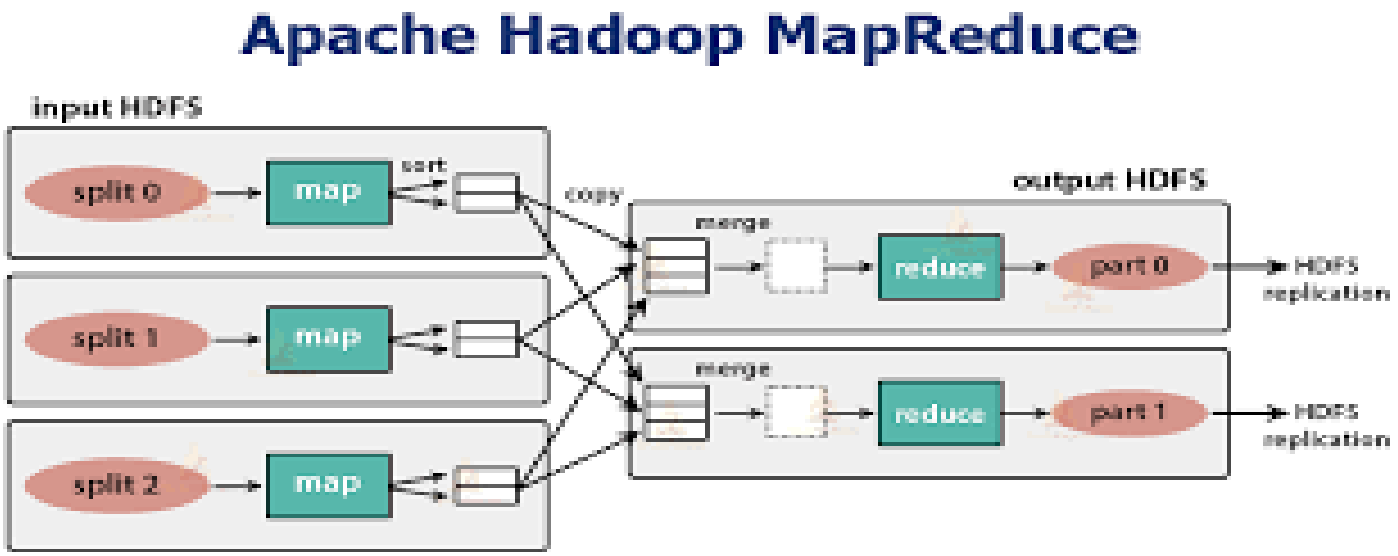


An open-source cluster computing framework

Provides 100 times faster performance than MapReduce

Analytics Vidhya
Learn everything about analytics

Hadoop Ecosystem…



Commonly used

The original Hadoop processing engine which is primarily Java based

An extensive and mature fault tolerance framework

Hadoop Ecosystem…

An open-source dataflow system

Converts pig script to Map-Reduce code

Apache Pig

An alternative to writing Map-Reduce code

Analytics Vidhya
Learn everything about analytics

Hadoop Ecosystem…

High performance SQL engine which runs on Hadoop cluster

Supports a dialect of SQL (Impala SQL)

Very low latency –measured in milliseconds

Hadoop Ecosystem…

Executes queries using MapReduce

Supports a dialect of SQL (HQL)

Similar to Impala

Hadoop Ecosystem…

Oozie is a workflow or coordination system used to manage the Hadoop jobs

Hadoop Ecosystem…

Hue is an acronym for Hadoop User Experience

Hue is an open source Web interface for analysing data with Hadoop

It provides SQL editors for Hive, Impala, MySQL, Oracle, PostgreSQL, Spark SQL, and SolrSQL

Hadoop Security

Kerberos Authentication	All services can be principals
HBase, Hive & HDFS Authorization	ACLs for each
Wire encryption	HDFS, Shuffle, JDBC
Basic Auditing	Logging of events
Knox: Hadoop REST Gateway	Perimeter security for Hadoop cluster
Centralized Security Administration	XA Secure
Centralized Audit Reporting	
Delegated Policy Administration	

RDBMS vs Hadoop

Relational		Hadoop
Required on write	**schema**	Required on read
Reads are fast	**speed**	Writes are fast
Standards and structured	**governance**	Loosely structured
Limited, no data processing	**processing**	Processing coupled with data
Structured	**data types**	Multi and unstructured
Interactive OLAP Analytics Complex ACID Transactions Operational Data Store	**best fit use**	Data Discovery Processing unstructured data Massive Storage/Processing

Apache Hadoop - Use cases

Product pricing

Stock levels

Sales

Stocking costs

Product costs & price

Advertisement pricing

Customer sentiment

Competitor prices

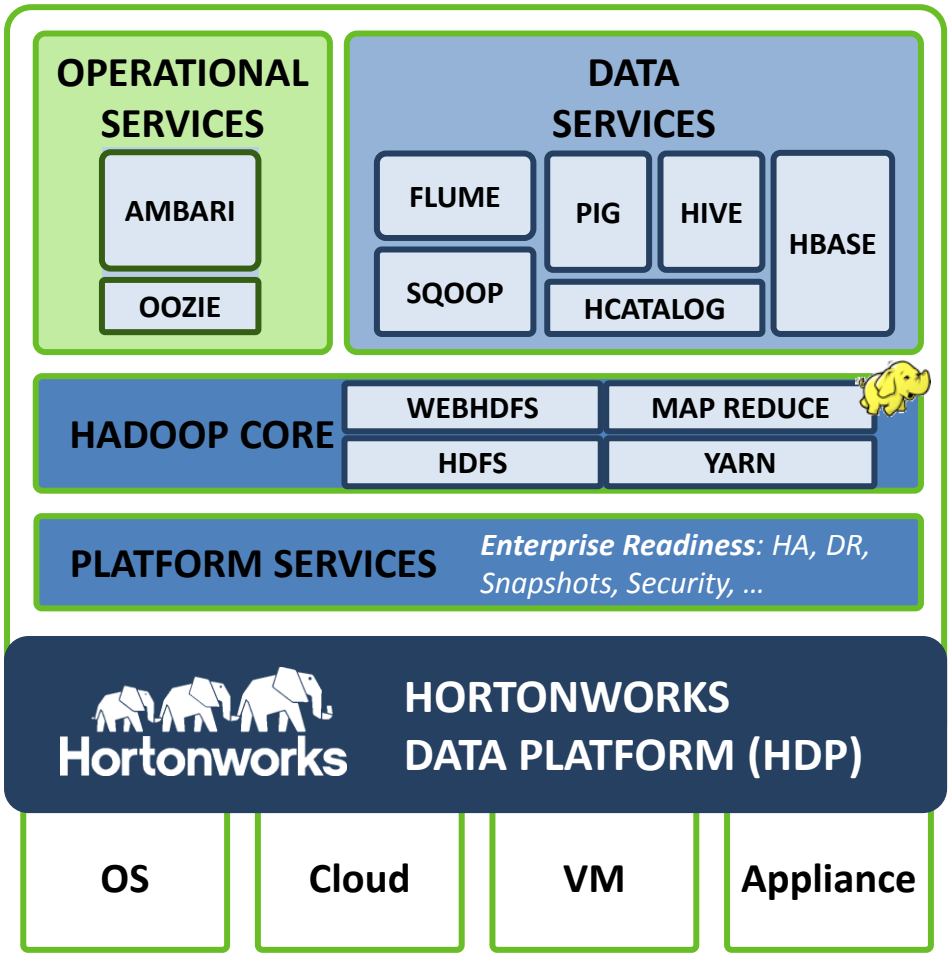
Analytics Vidhya
Learn everything about analytics

Cloudera Distribution Apache Hadoop

cloudera®

- Founded by leading experts on Hadoop from Facebook, Yahoo, Google and Oracle.
- Provides support, consulting, training and certification for Hadoop users.
- Staff includes committers to virtually all Hadoop projects.
- Many authors of industry standard books on Apache Hadoop projects
  - Tom White, Lars George, Kathleen Ting etc
- Customers includes many ke users of Hadoop
  - Altstate, AOL Advertising, Box, CBS Interactive, eBay, Experian, Groupon, National Cancer Institute, Orbitz, Social Security Administration, Trend Micro, Truila, US Army....

Hortonwork

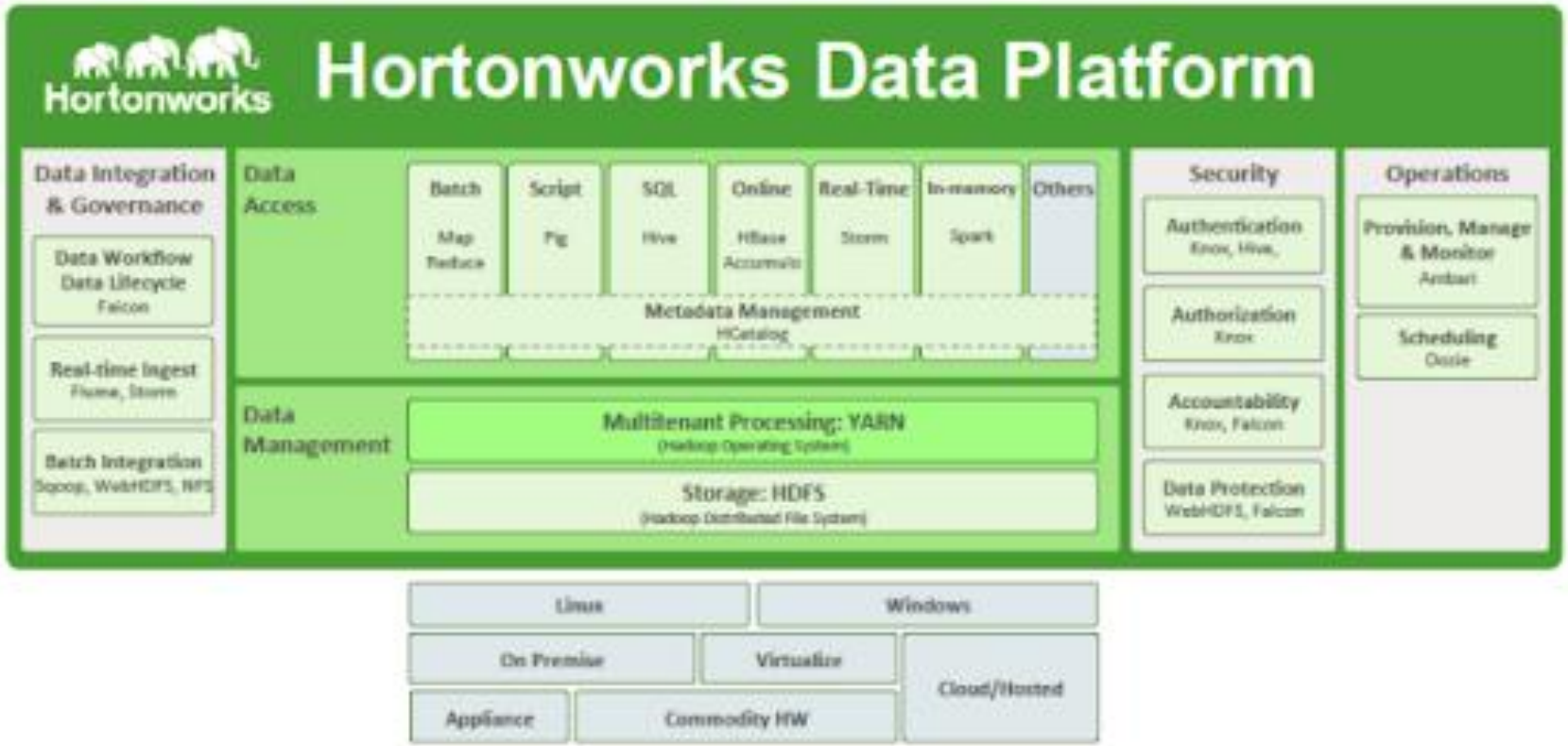


Enterprise Hadoop

- The ONLY 100% open source and complete distribution
- Enterprise grade, proven and tested at scale
- Ecosystem endorsed to ensure interoperability

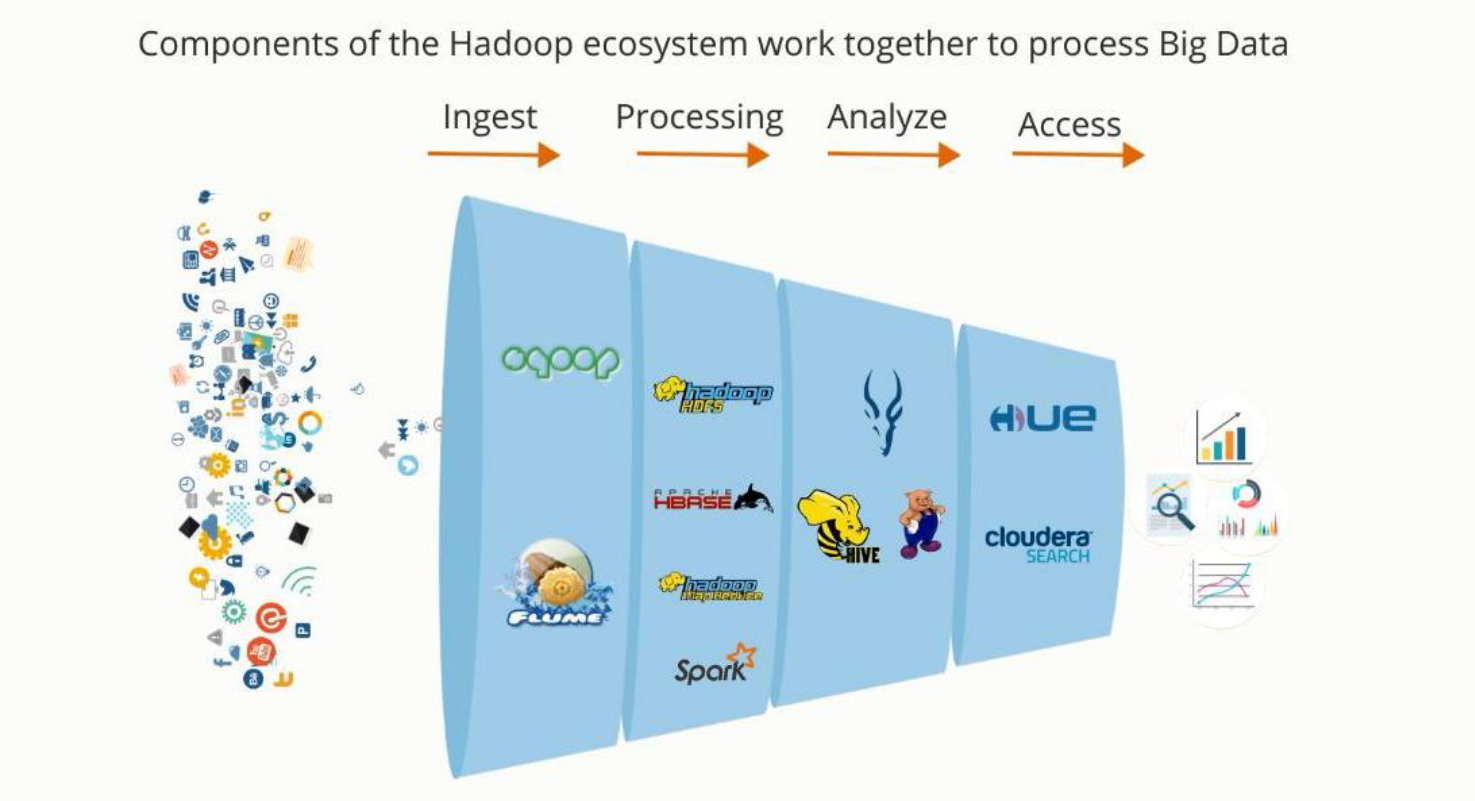
Hortonwork

Hortonwork



Big Data Processing

Four stages of Big Data

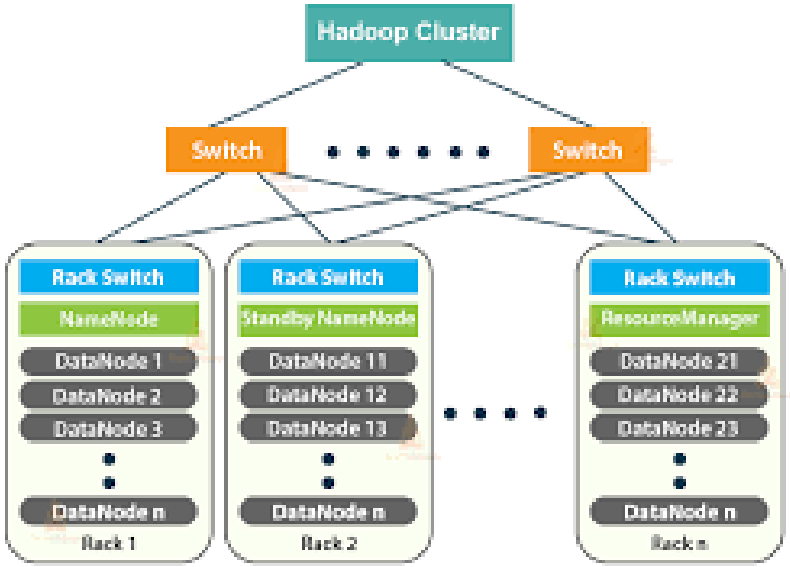


Analytics Vidhya
Learn everything about analytics

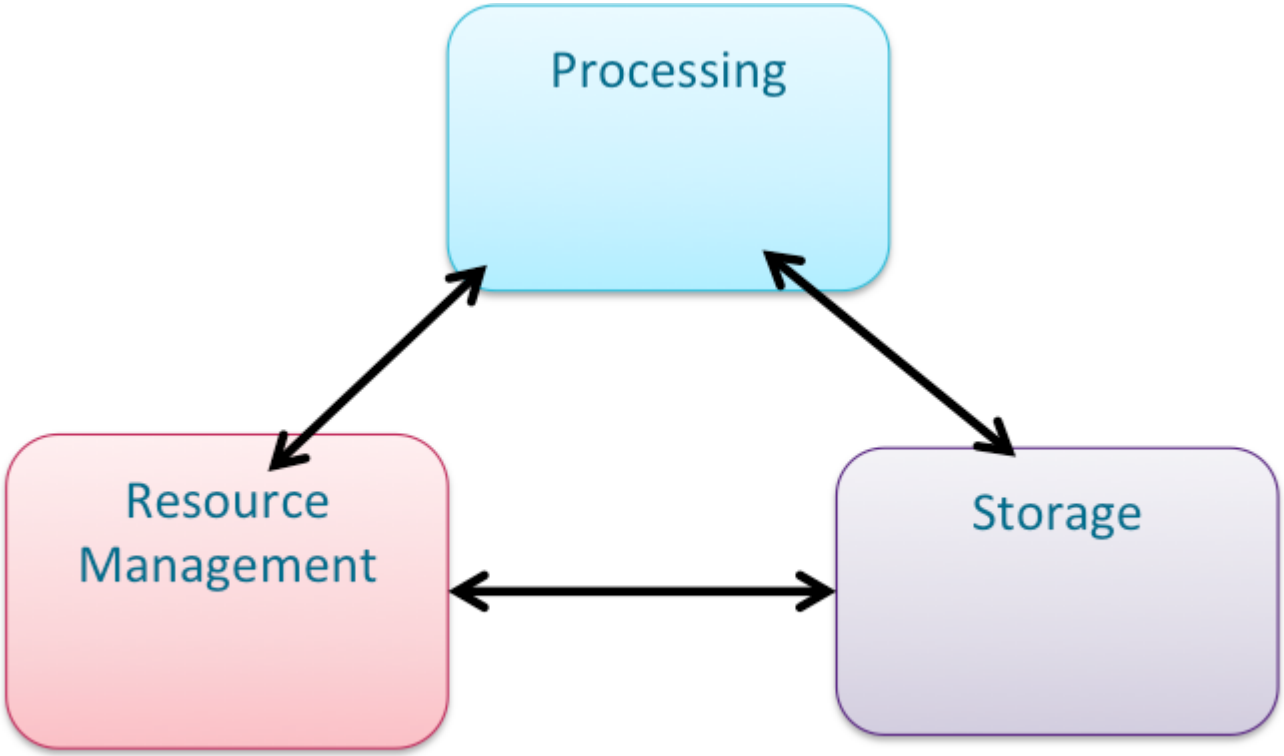
HADOOP DISTRIBUTED FILE SYSTEM

Hadoop Terminology

- A cluster is a group of computers working together
- A node is an individual computer in the cluster
- Daemons are long---running processes that typically start up when the machine starts and listens for requests from client processes.

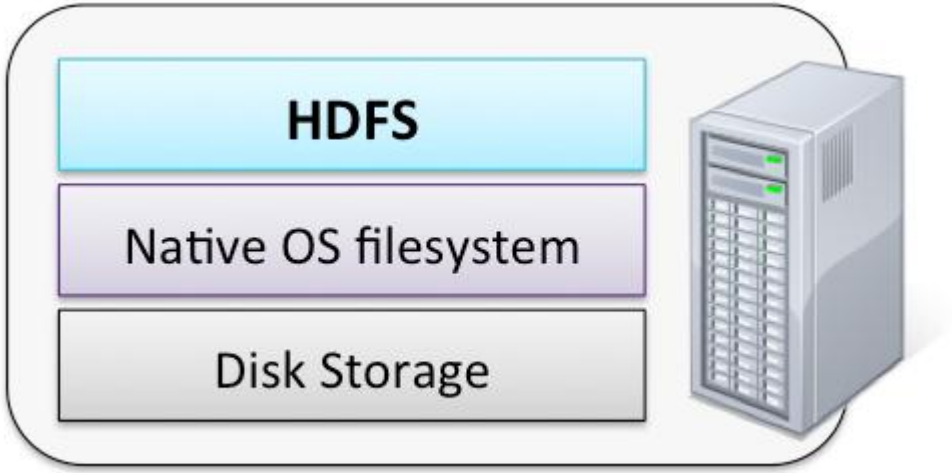


Cluster Components



What is HDFS

- HDFS is a distributed file system .
- It provides great access to data across Hadoop clusters.
- Based on GFS
- Sits on top of native file system
- Provides redundant storage for massive amount of data.
- Using readily available , industry standard computers.



What is HDFS

HDFS operates best with a ‘modest’ number of large files

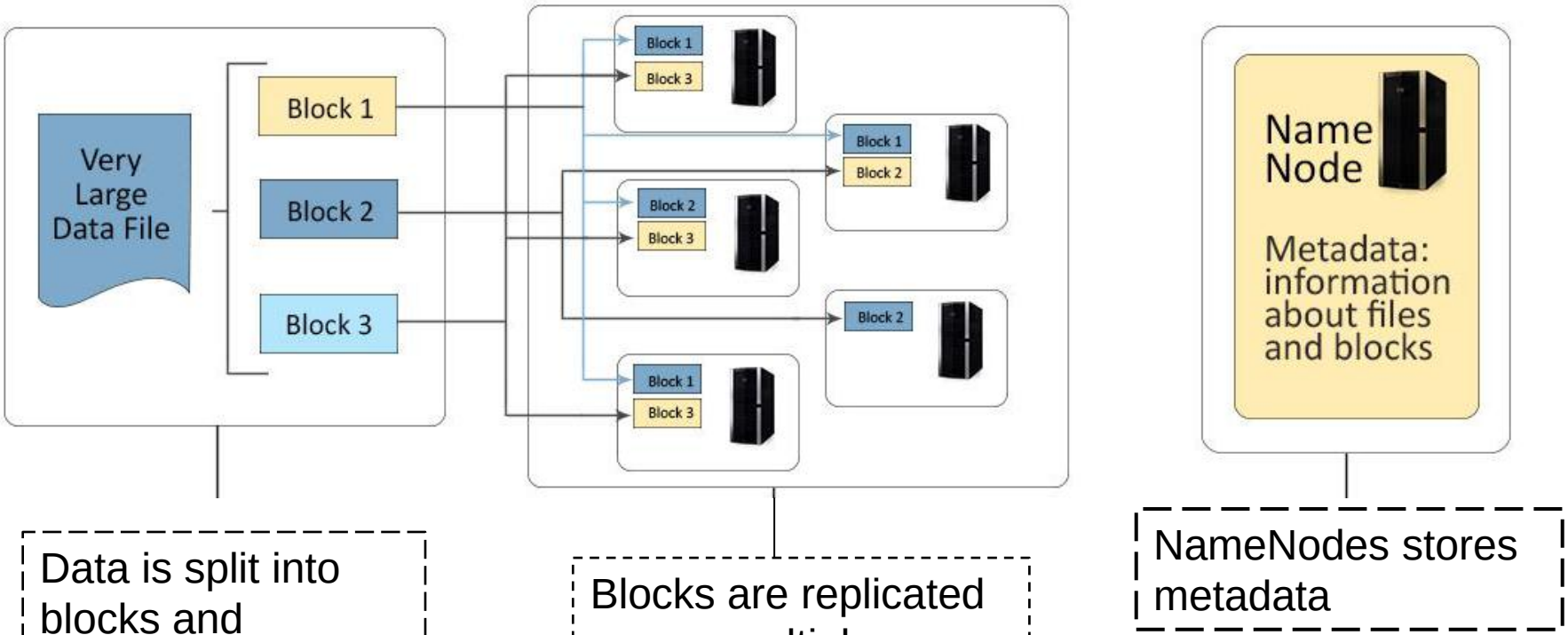
- Millions, rather than billions, of files
- Each file size typically100MB or more

Files in HDFS are ‘write once’

- No random writes to files are allowed

HDFS is optimized for large, streaming reads of files

How files are stored in HDFS



Data is split into blocks and distributed across multiple nodes in the cluster

Blocks are replicated across multiple machines, known as DataNodes

NameNodes stores metadata

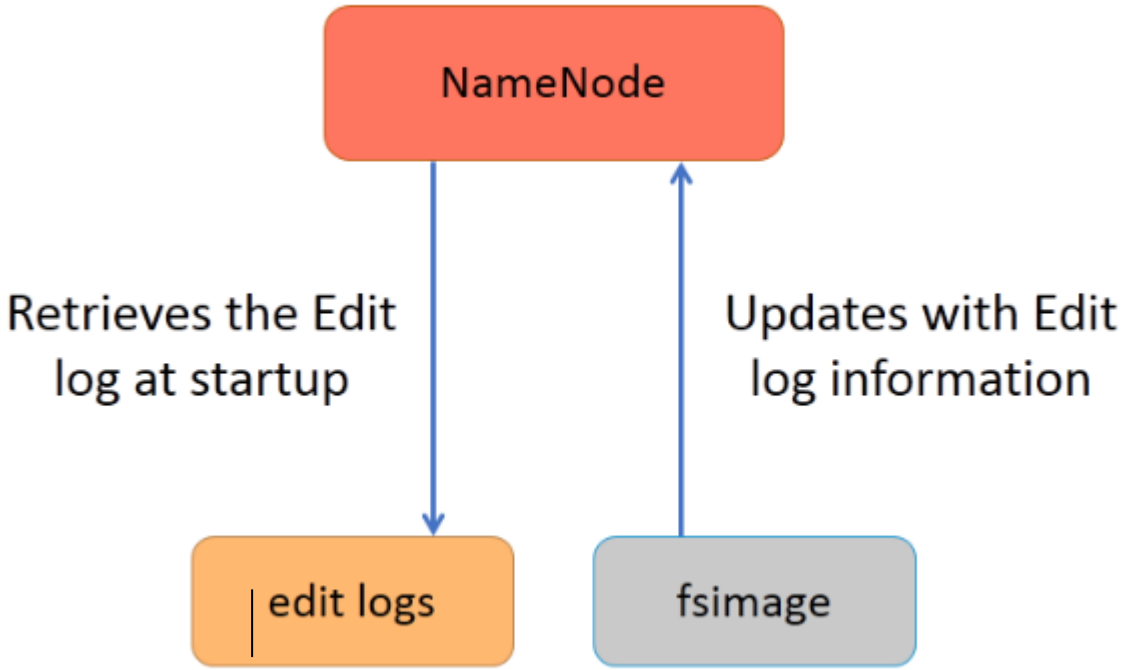
Analytics Vidhya
Learn everything about analytics

How files are stored in HDFS

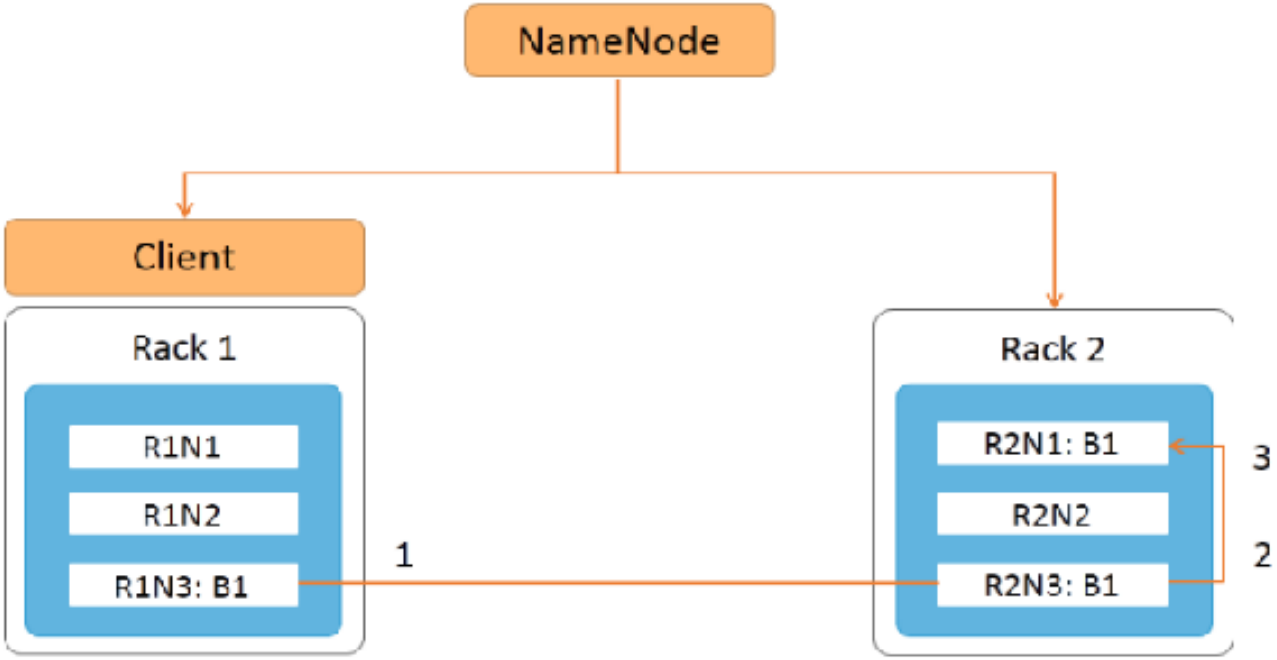
The NameNode maintains two persistent files

– a transaction log called an **Edit Log**

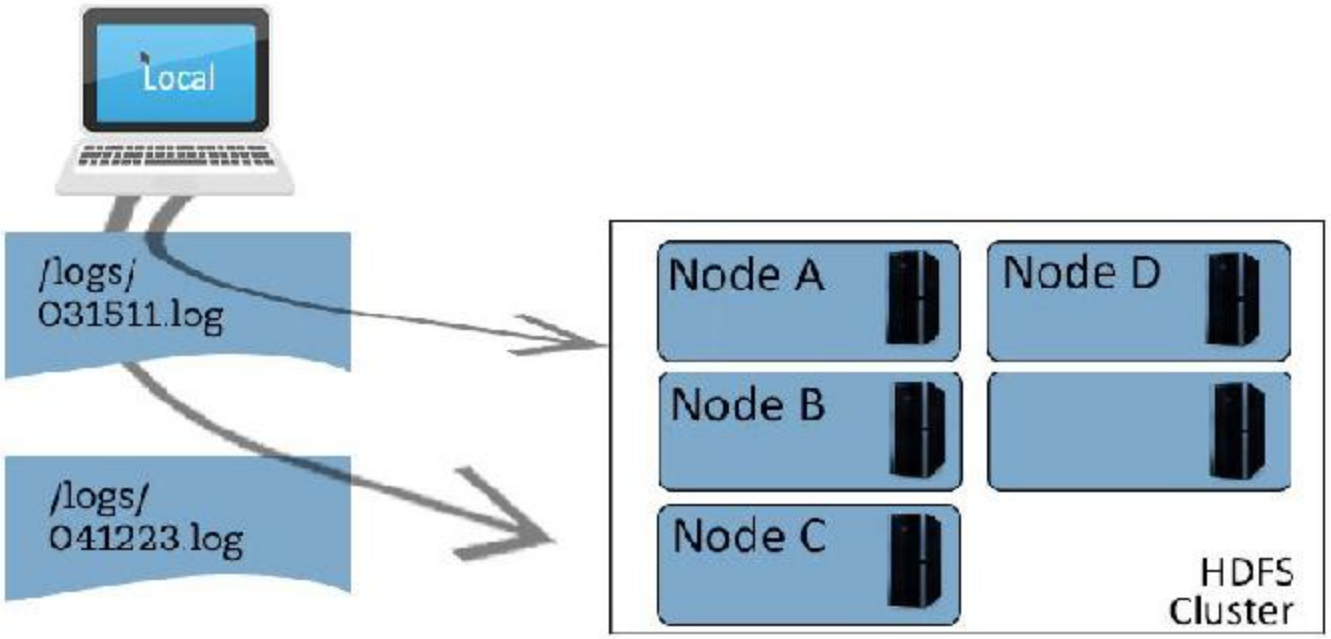
- a namespace image called an **FsImage**.



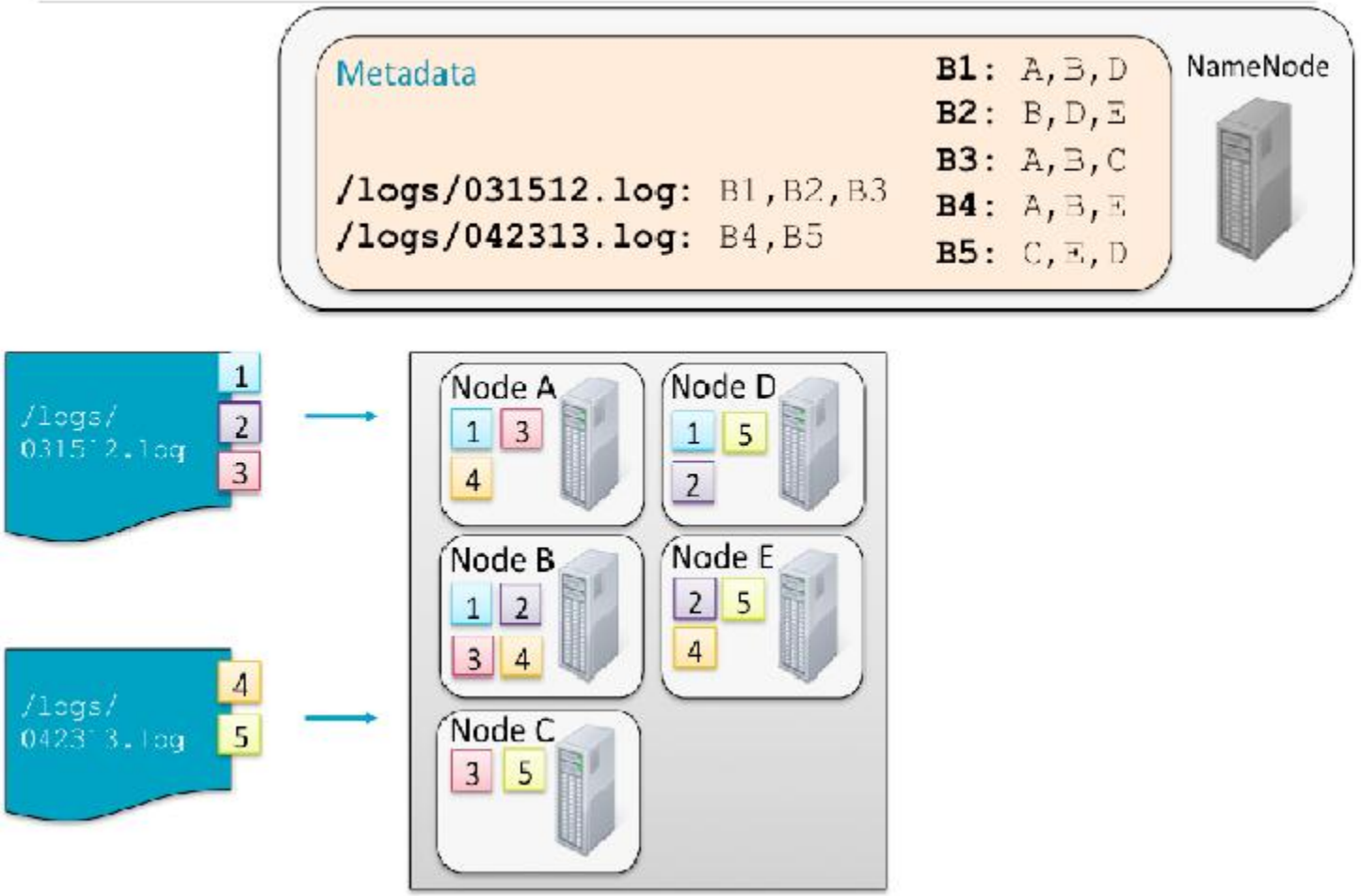
Data Replication Topology



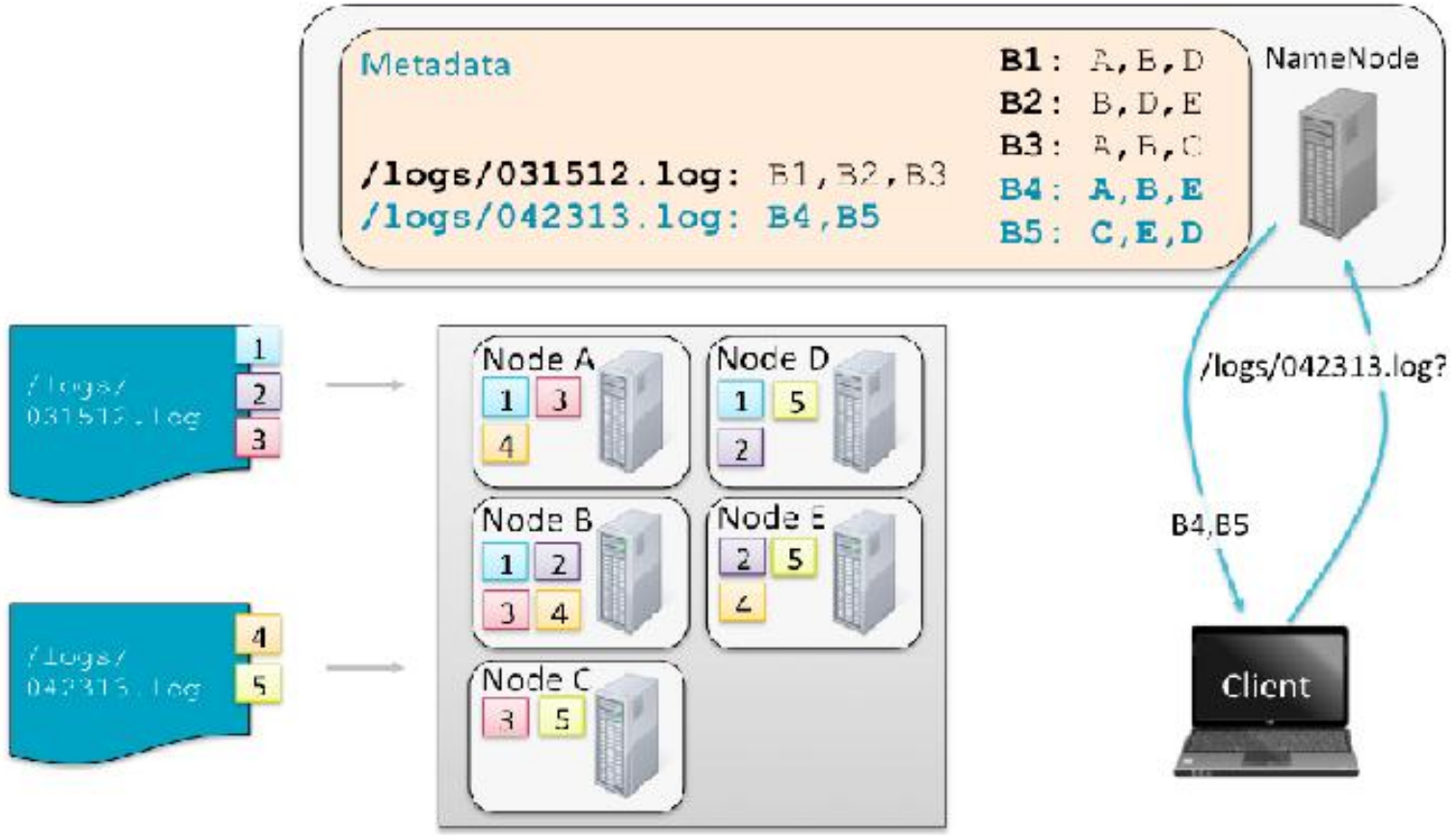
Storing and retrieving files

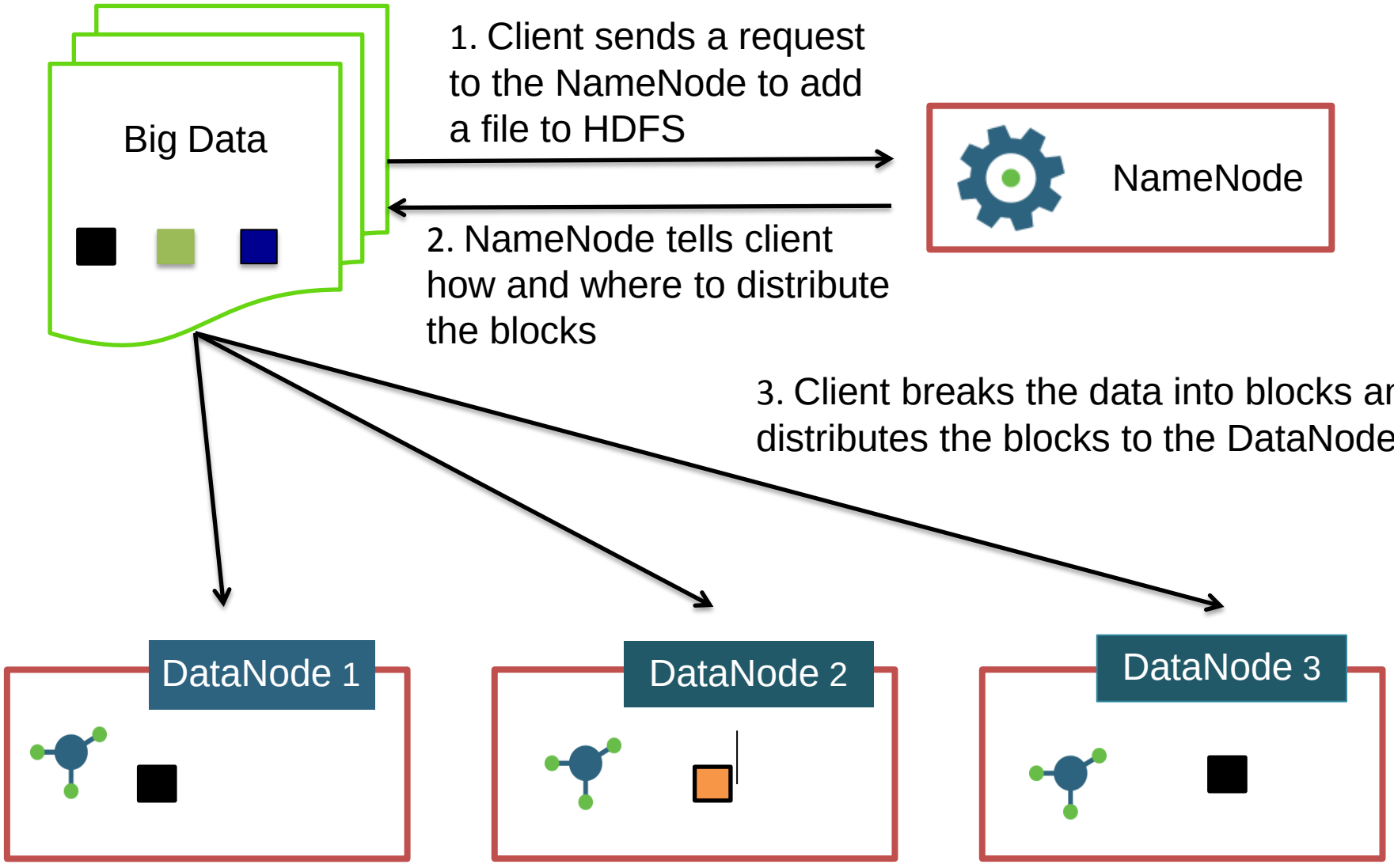


Storing and retrieving files



Storing and retrieving files





4. The DataNodes replicate the blocks (as instructed by the NameNode)

Accessing HDFS

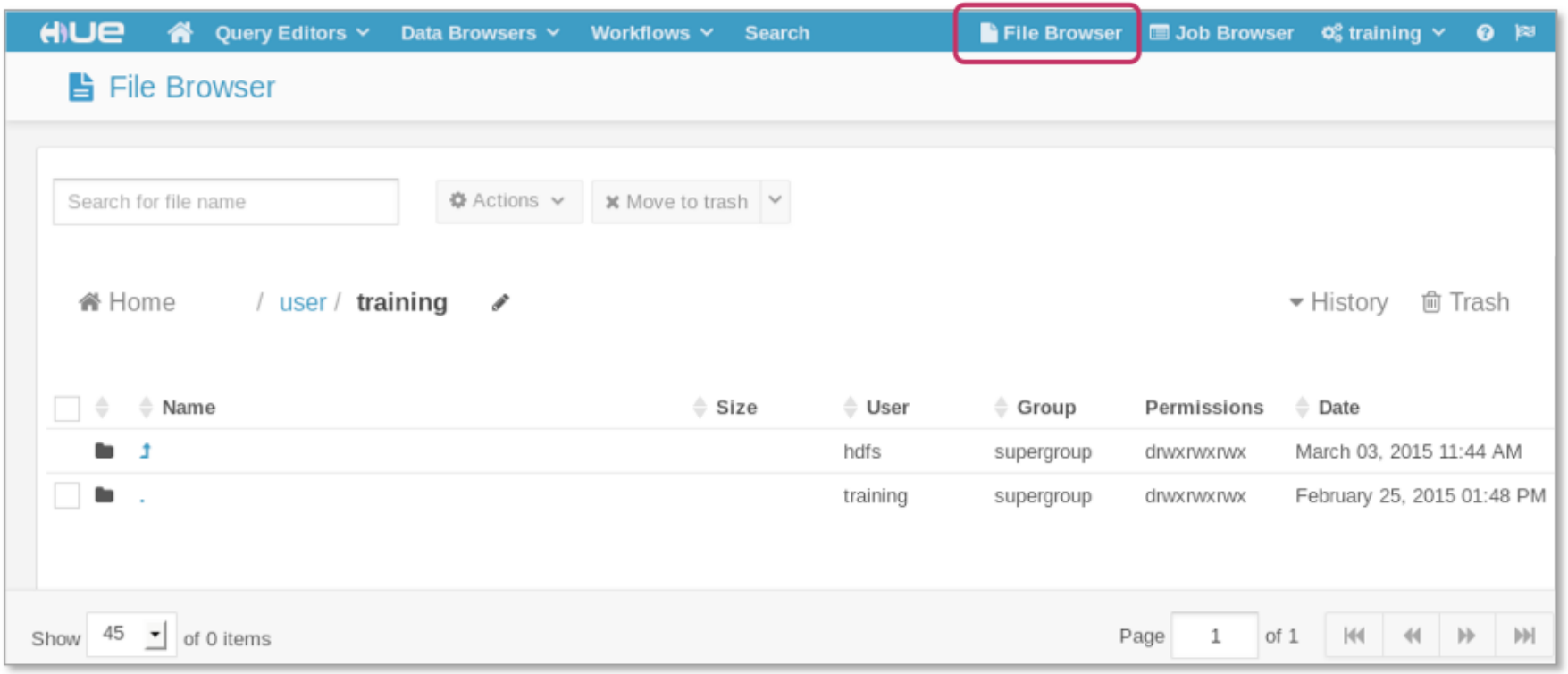
HDFS Commands

```
hdfs fs –command [args]
```

Here are a few (of the almost 30) HDFS commands:

- -cat: display file content (uncompressed)
- -text: just like cat but works on compressed files
- -chgrp,-chmod,-chown: changes file permissions
- -put,-get,-copyFromLocal,-copyToLocal: copies files from the local file system to the HDFS and vice-versa. Two versions.
- -ls, -ls -R: list files/directories
- -mv,-moveFromLocal,-moveToLocal: moves files
- -stat: statistical info for any given file (block size, number of blocks, file type, etc.)

HUE Browser



Analytics Vidhya
Learn everything about analytics

HDFS NameNode Availability

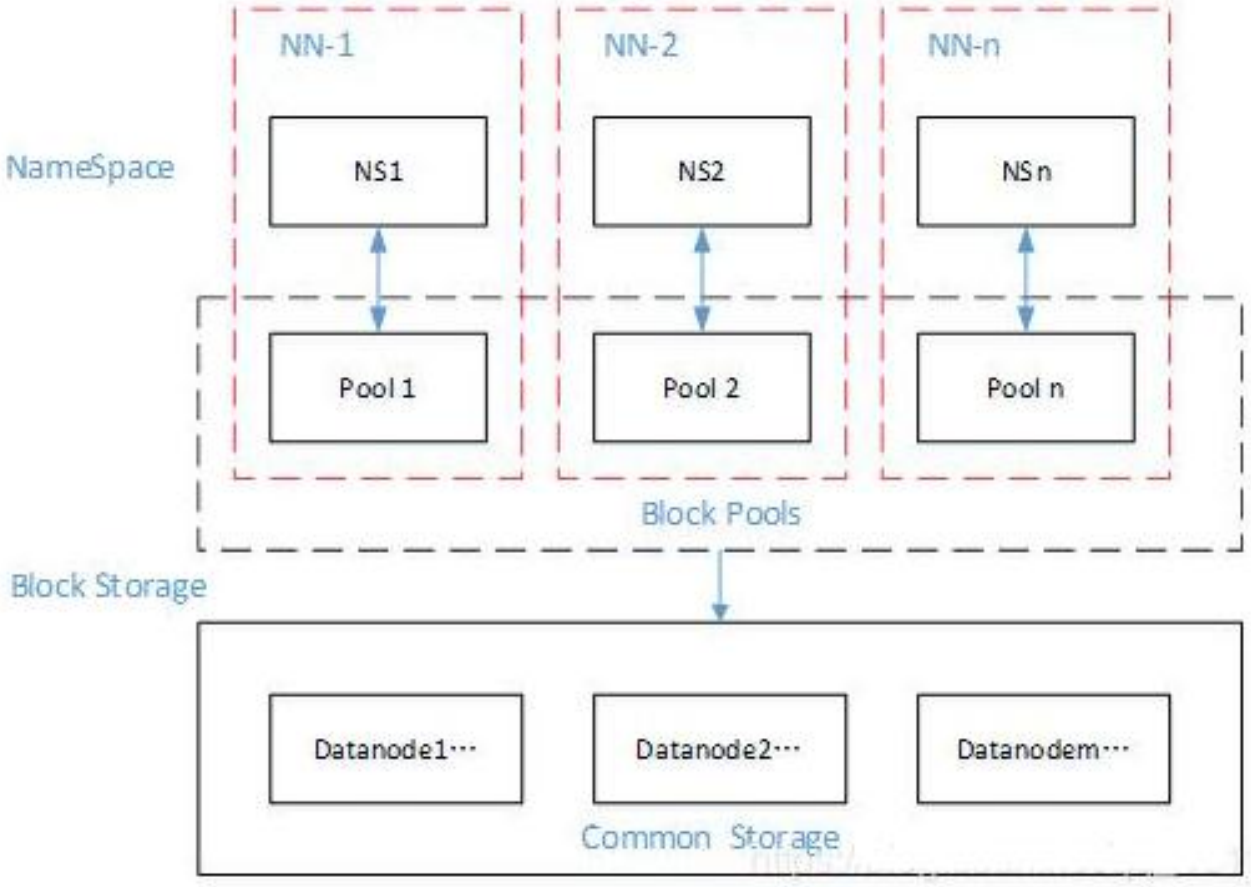
- Files and directories belong to a Namespace
- A NameNode manages a single Namespace Volume
- **The NameNode daemon must be running at all times**

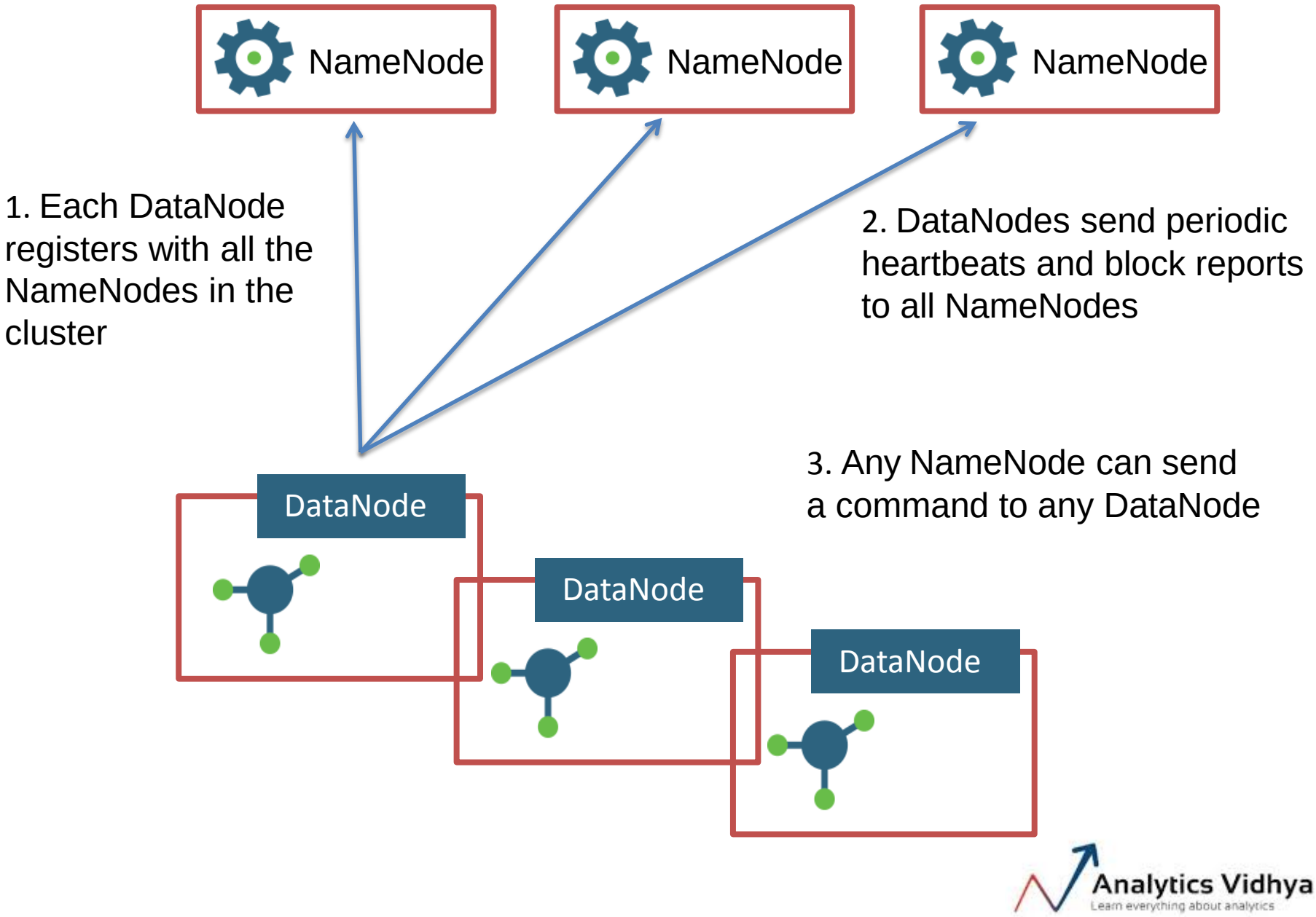
Two NameNodes:
- Active
- Standby Node

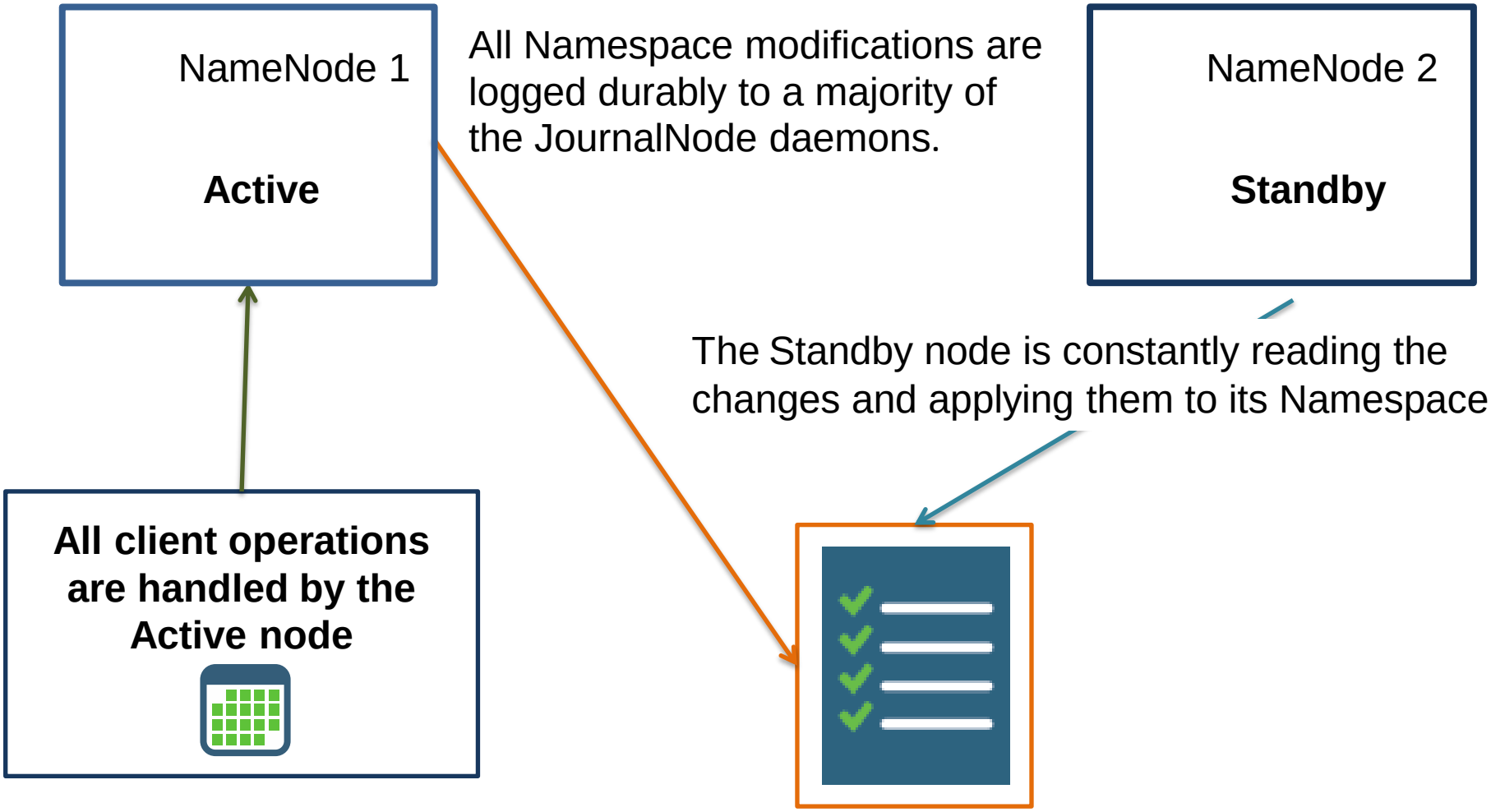
- **High Availability mode**

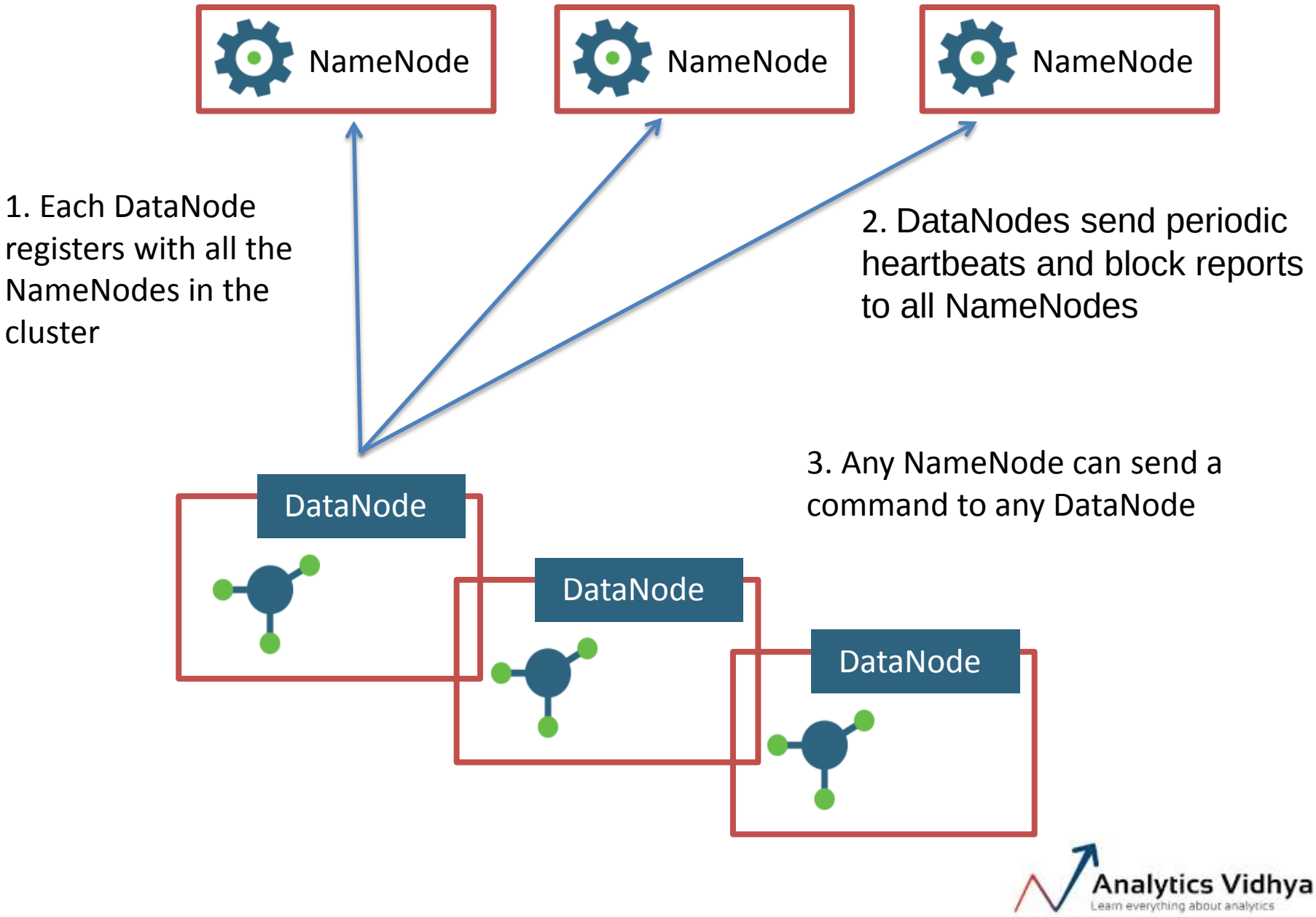
•One NameNode
•One “helper” node called Secondary NameNode for Bookkeeping, not backup.

HDFS Federation




NameNode
NameNode
NameNode
1. Each DataNode registers with all the NameNodes in the cluster
2. DataNodes send periodic heartbeats and block reports to all NameNodes
3. Any NameNode can send a command to any DataNode
DataNode
DataNode
DataNode
Analytics Vidhya
Learn everything about analytics




NameNode
NameNode
NameNode
1. Each DataNode registers with all the NameNodes in the cluster
2. DataNodes send periodic heartbeats and block reports to all NameNodes
3. Any NameNode can send a command to any DataNode
DataNode
DataNode
DataNode
Analytics Vidhya
Learn everything about analytics

HDFS Rebalance

A cluster is in a balanced status when,

- % of space used in each data node is within limits of Average % of space used on data nodes +/- Threshold size .
- Percentage space used on a data node should not be less than Average % of space used on data nodes – Threshold size.
- Percentage space used on a data node should not be greater than Average % of space used on data nodes + Threshold size.

Here Threshold size is configurable value which is 20 % of used spaced by default.

HDFS Rebalance

- Rebalancer is a administration tool in HDFS, to balance the distribution of blocks uniformly across all the data nodes in the cluster.
- Rebalancing will be done on demand only. It will not get triggered automatically.

```
hadoop1@ubuntu-1:~$ hdfs balancer
Time Stamp               Iteration#  Bytes Already Moved  Bytes Left To Move  Bytes Being Moved
The cluster is balanced. Exiting...
Balancing took 1.148 seconds
hadoop1@ubuntu-1:~$
```

HDFS Rebalance

If a Rebalancer is triggered, NameNode will scan entire data node list and when

- Under-utilized data node is found, it moves blocks from over-utilized data nodes or not-under-utilized data nodes to this current data node
- If Over-utilized data node is found, it moves blocks from this data node to other under-utilized or not-over-utilized data nodes.
- Spreads different replicas of a block across the racks so that cluster can survive loss of an entire rack.
- One of the replicas is placed on the same rack as the node writing to the file so that cross-rack network I/O is reduced.
- Spread HDFS data uniformly across the DataNodes in the cluster.

Yet Another Resource Negotiator

Analytics Vidhya
Learn everything about analytics

What is YARN

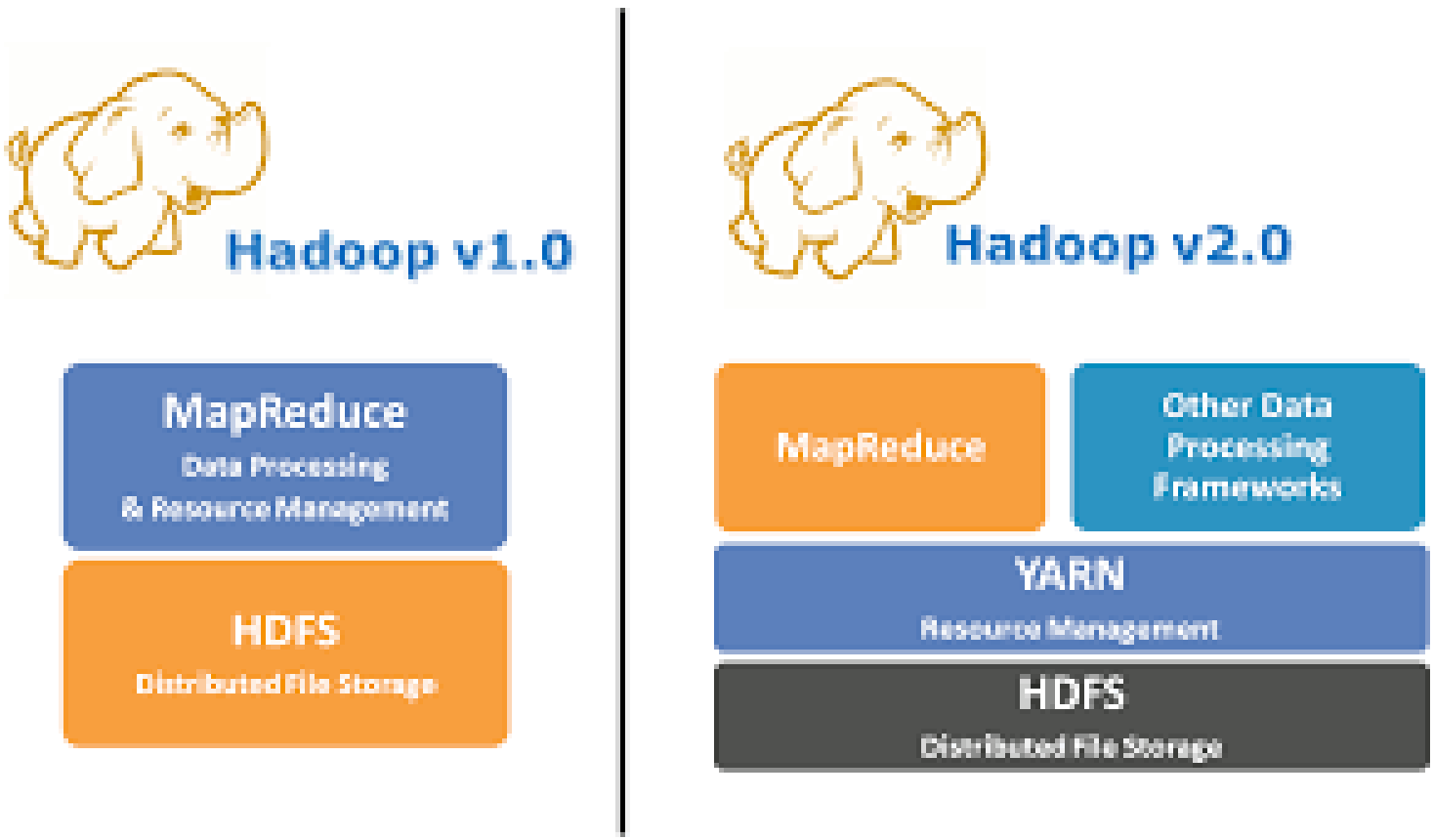
- **YARN** is a resource manager.
- It was created by separating the processing engine and the management function of MapReduce.
- It monitors and manages workloads, maintains a multi-tenant environment, manages the high availability features of Hadoop, and implements security controls
- YARN allow multiple processing engines

  Batch Processing

  Stream Processing

  Advanced Analytics

Before YARN



YARN Daemons

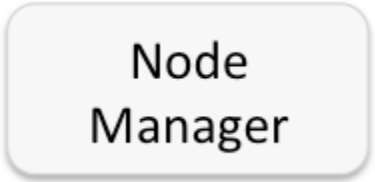
- **Resource Manager (RM**)

 - Runs as Master node

 - Global Resource Manager

 - Arbitrates system resources between competing applications

 - has a pluggable scheduler to support different algorithms



- **Node Manager (NM)**

 - Runs on slave nodes

 - communicates with RM

 - Node health

 - Reporting node and container status to the ResourceManager

Run an application on YARN

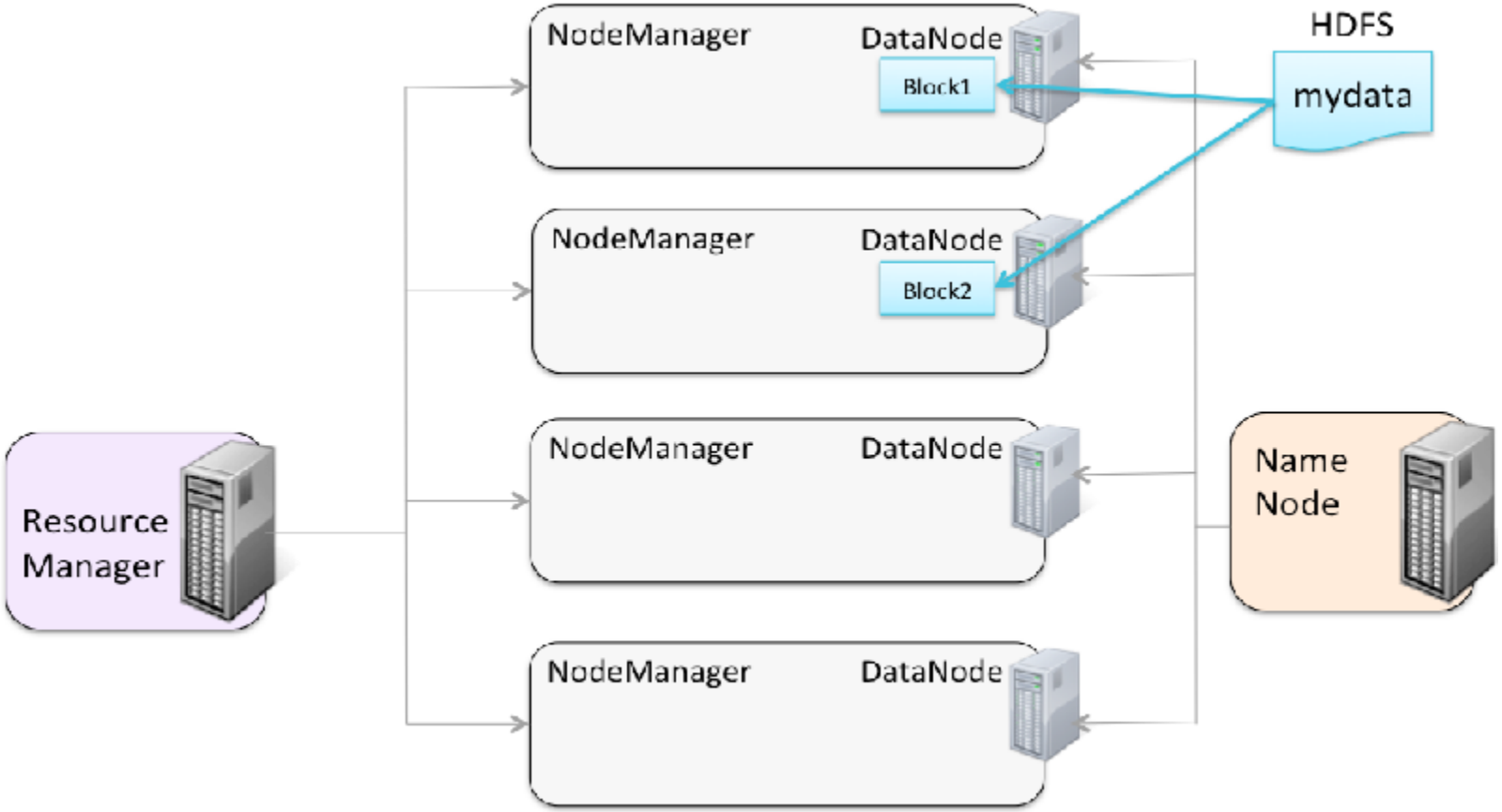
- **Application Master (AM)**

- One per application
- Framework/app specific
- Runs in a container
- Requests more containers to run application tasks
- manages the application lifecycle, dynamic adjustments to resource consumption, execution flow, faults, and provides status and metrics

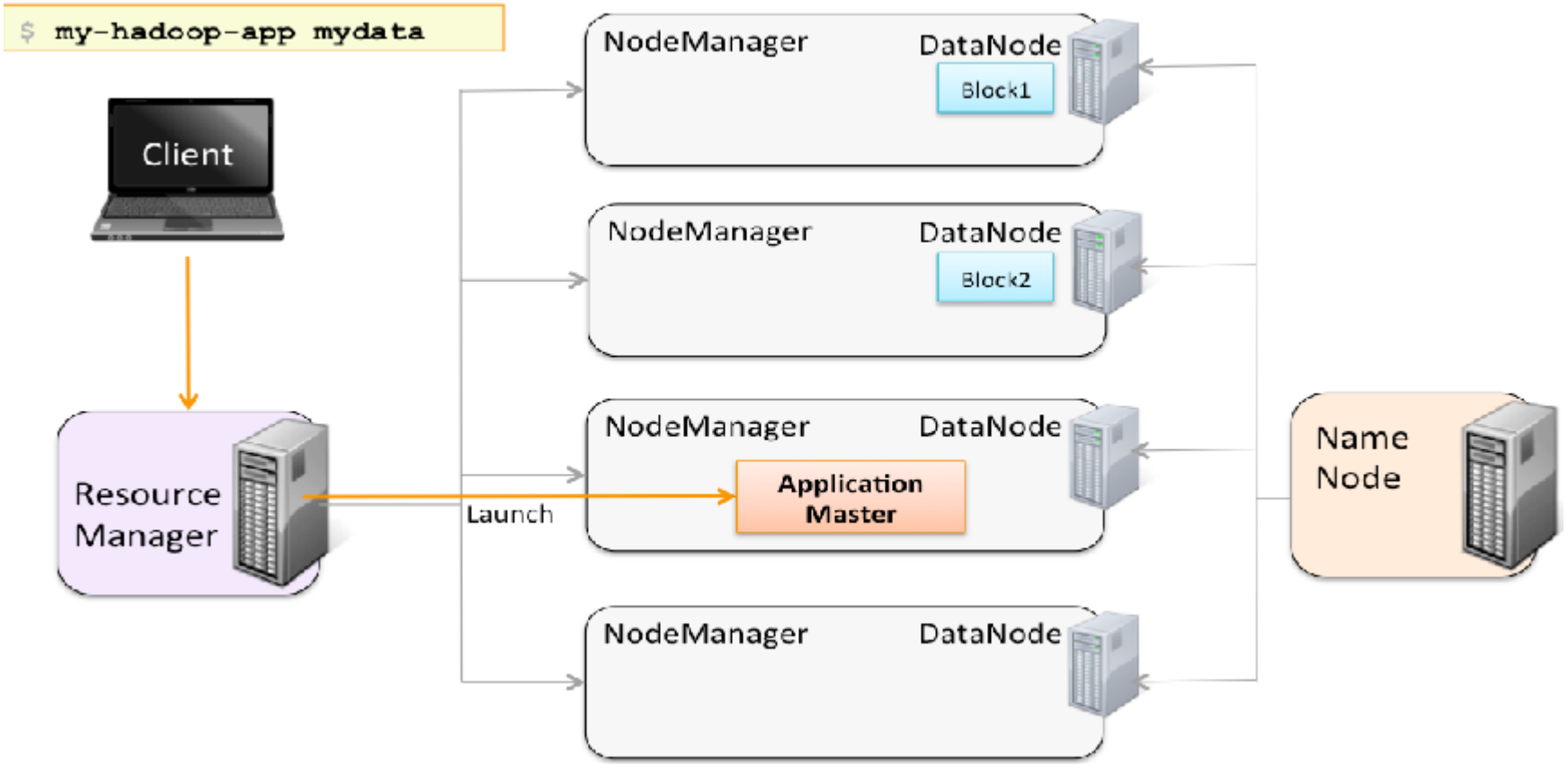
- **Containers**

- created by the RM on request
- Allocate a certain amount of resources on the slave nodes
- Applications run in one or more containers

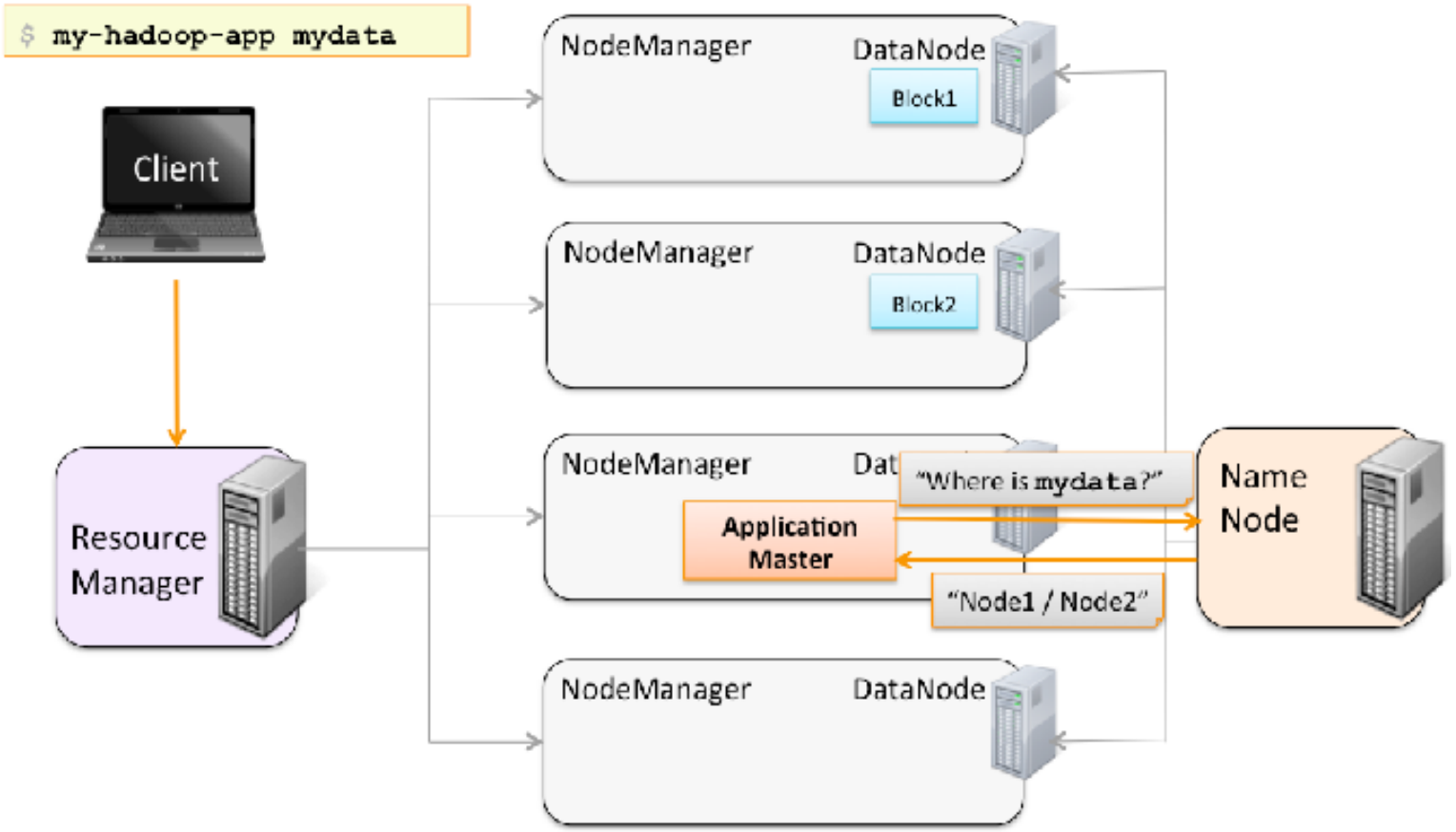
Run an application on YARN



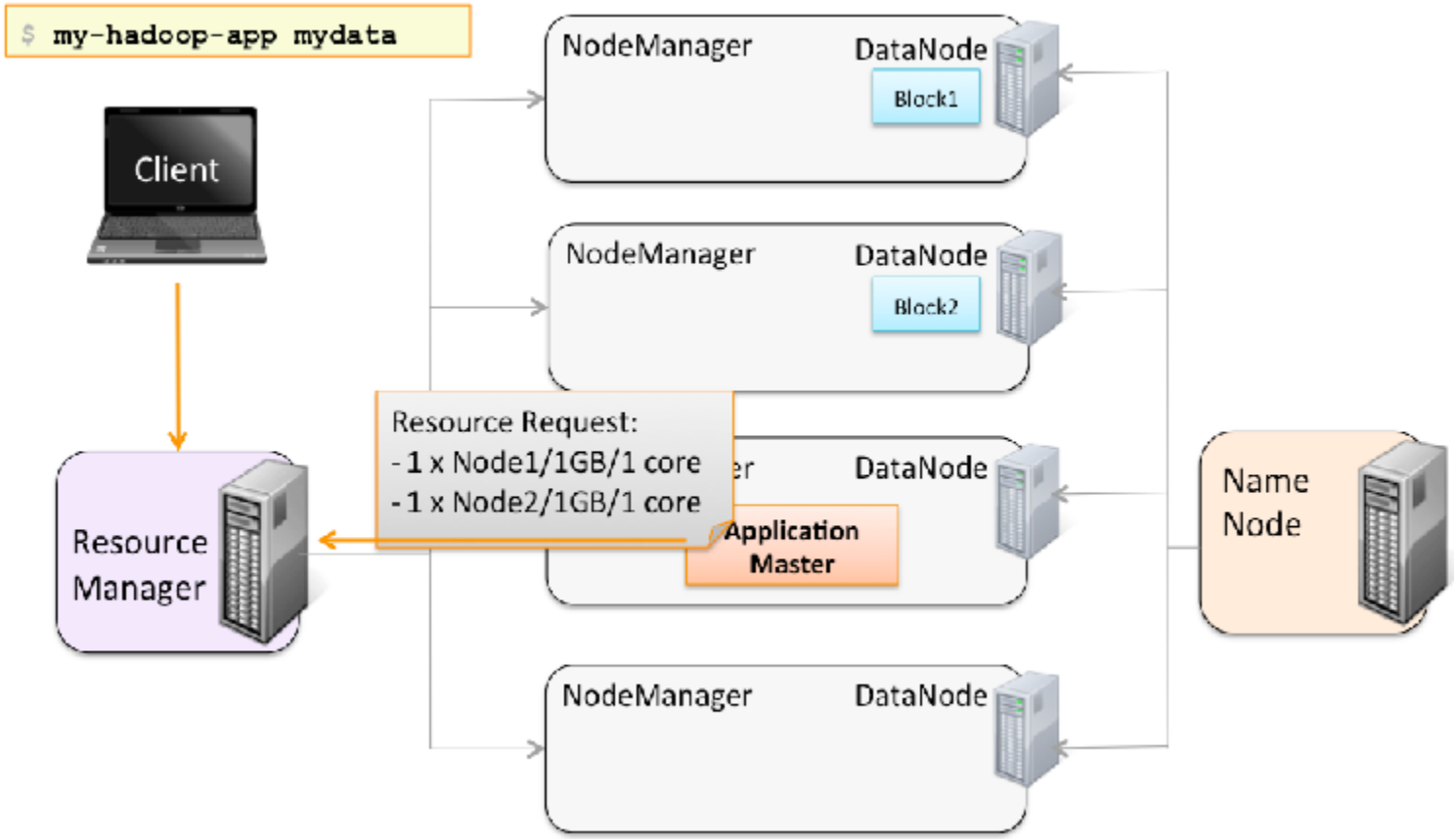
Run an application on YARN



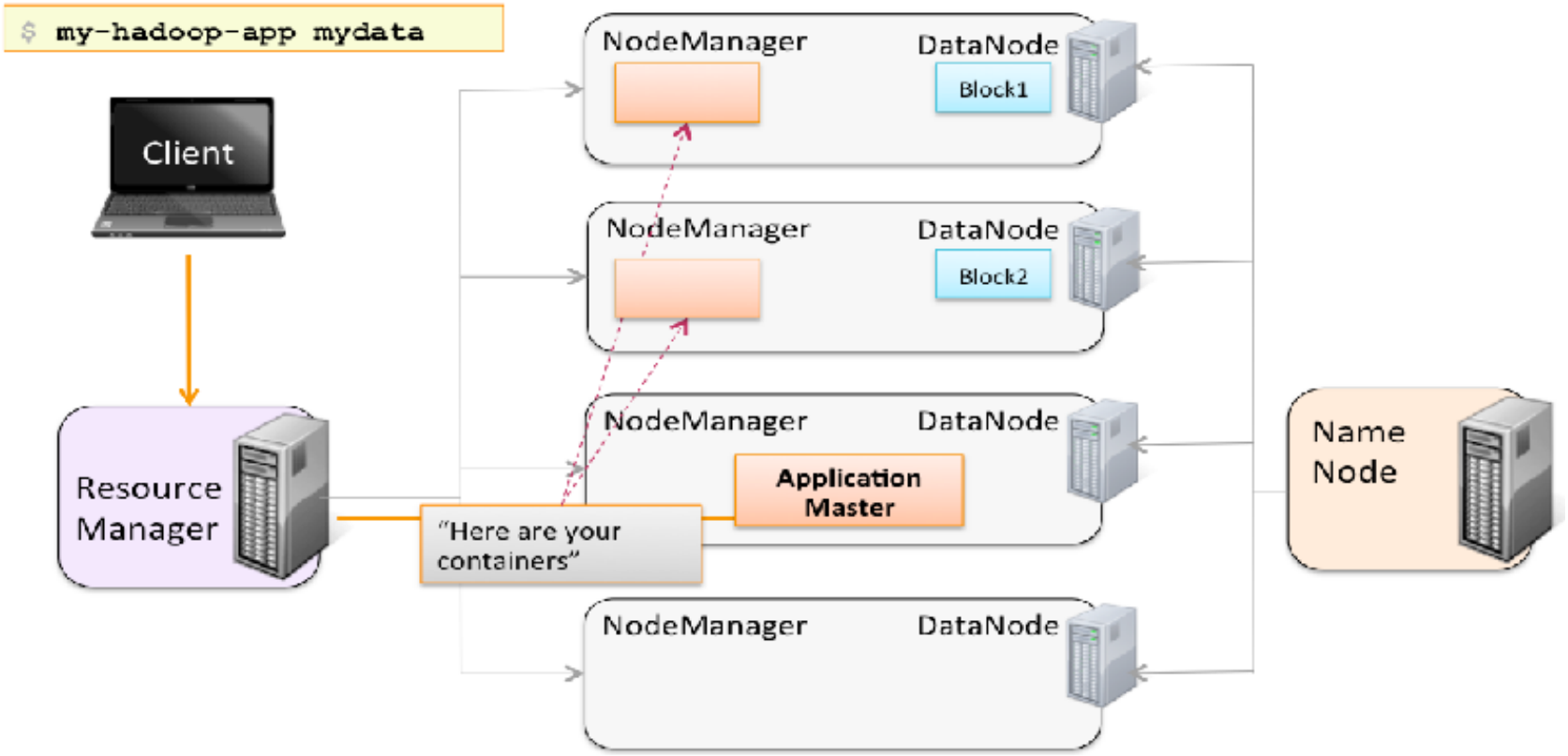
Run an application on YARN



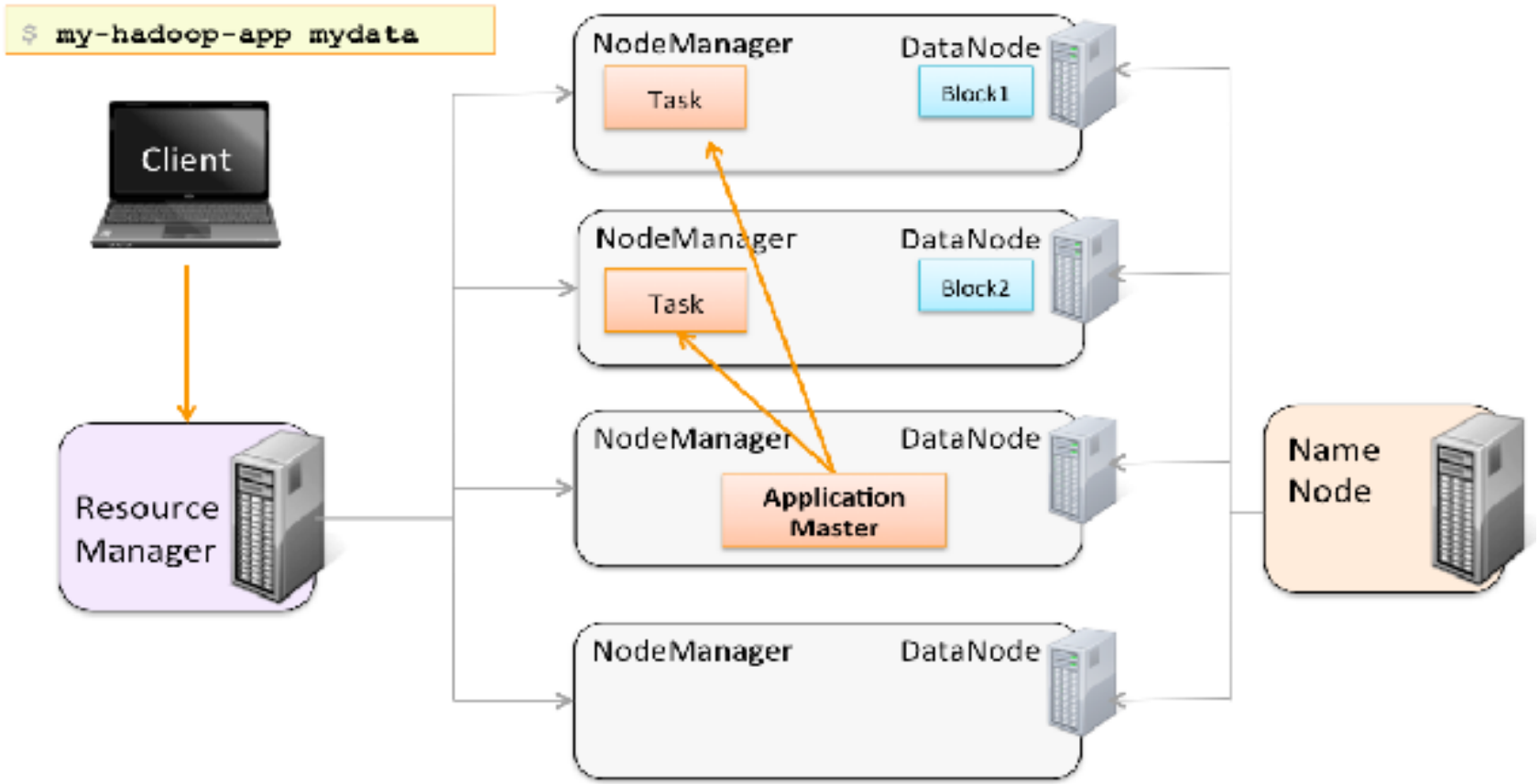
Run an application on YARN



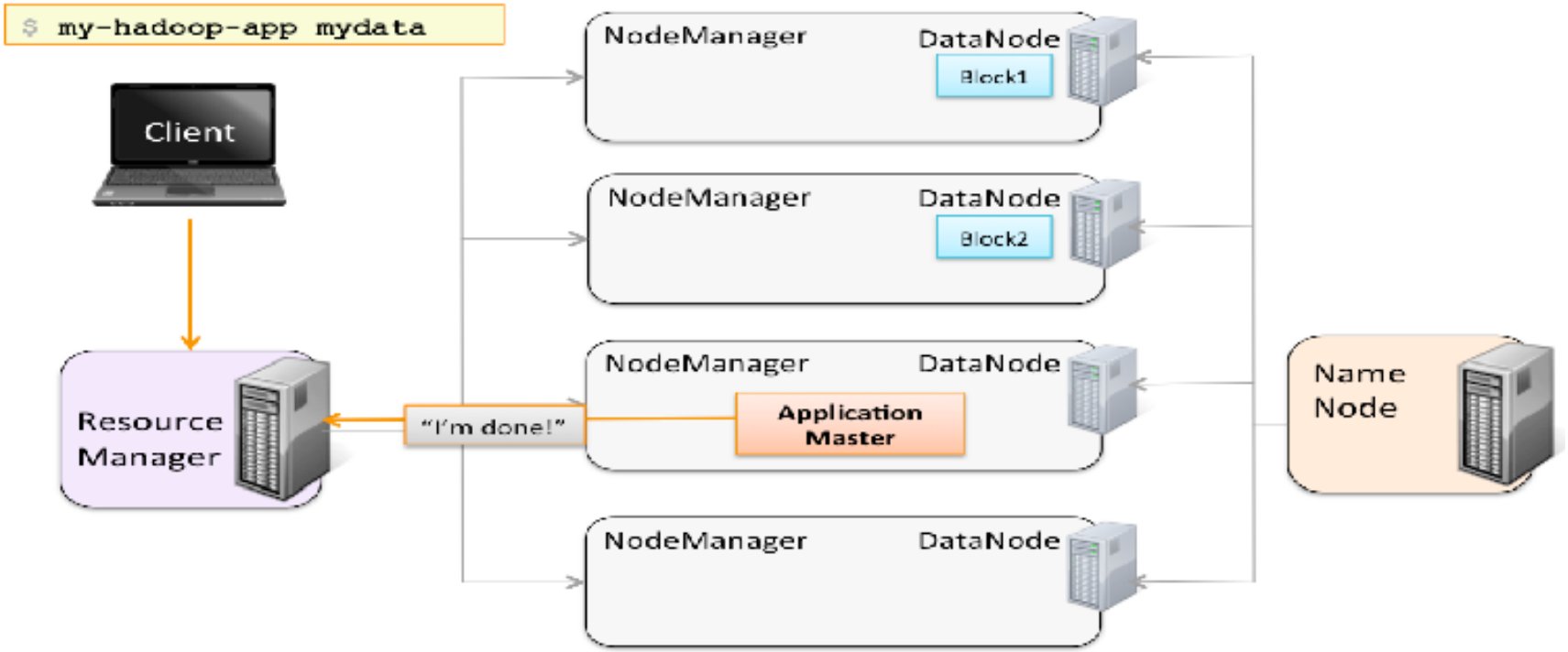
Run an application on YARN



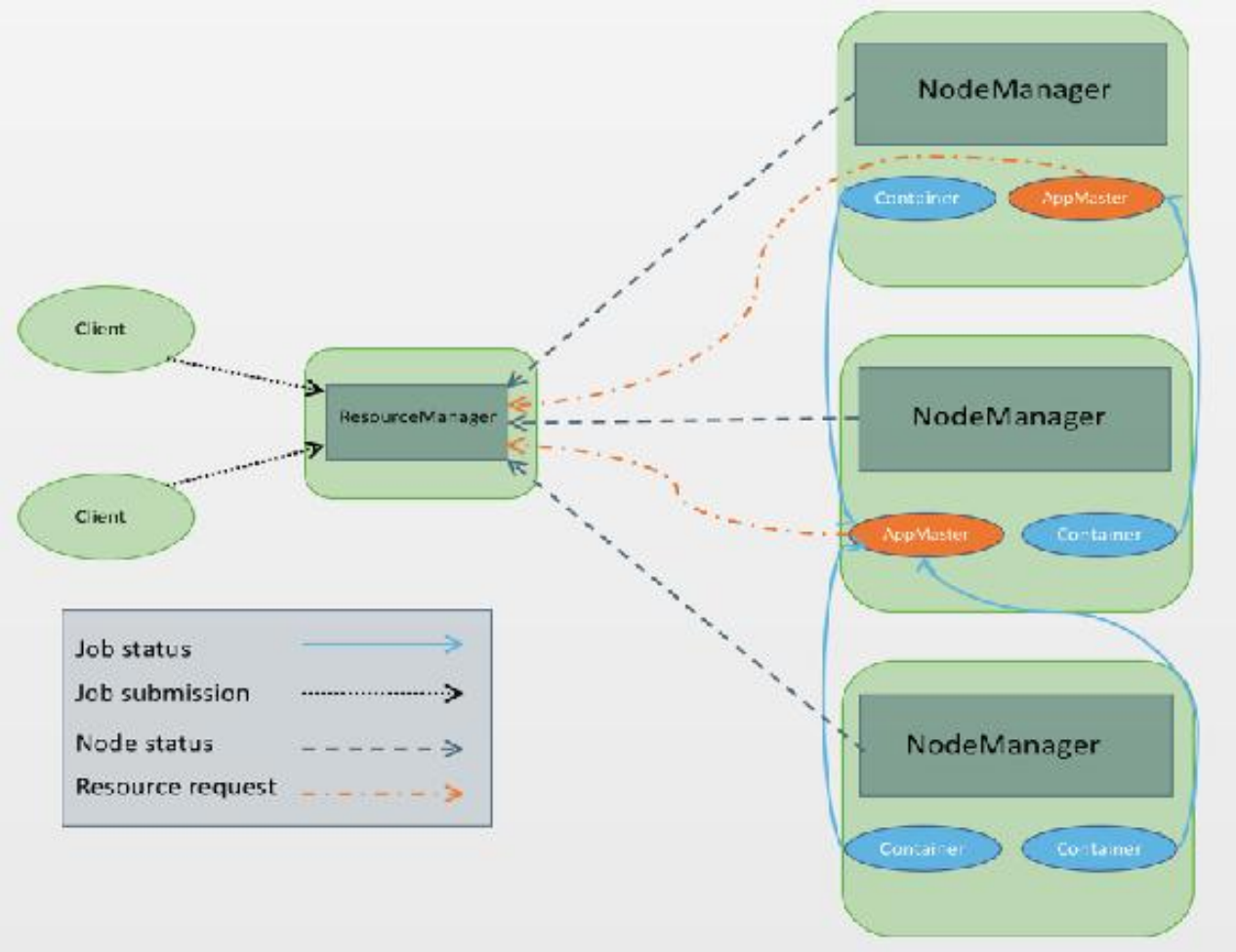
Run an application on YARN

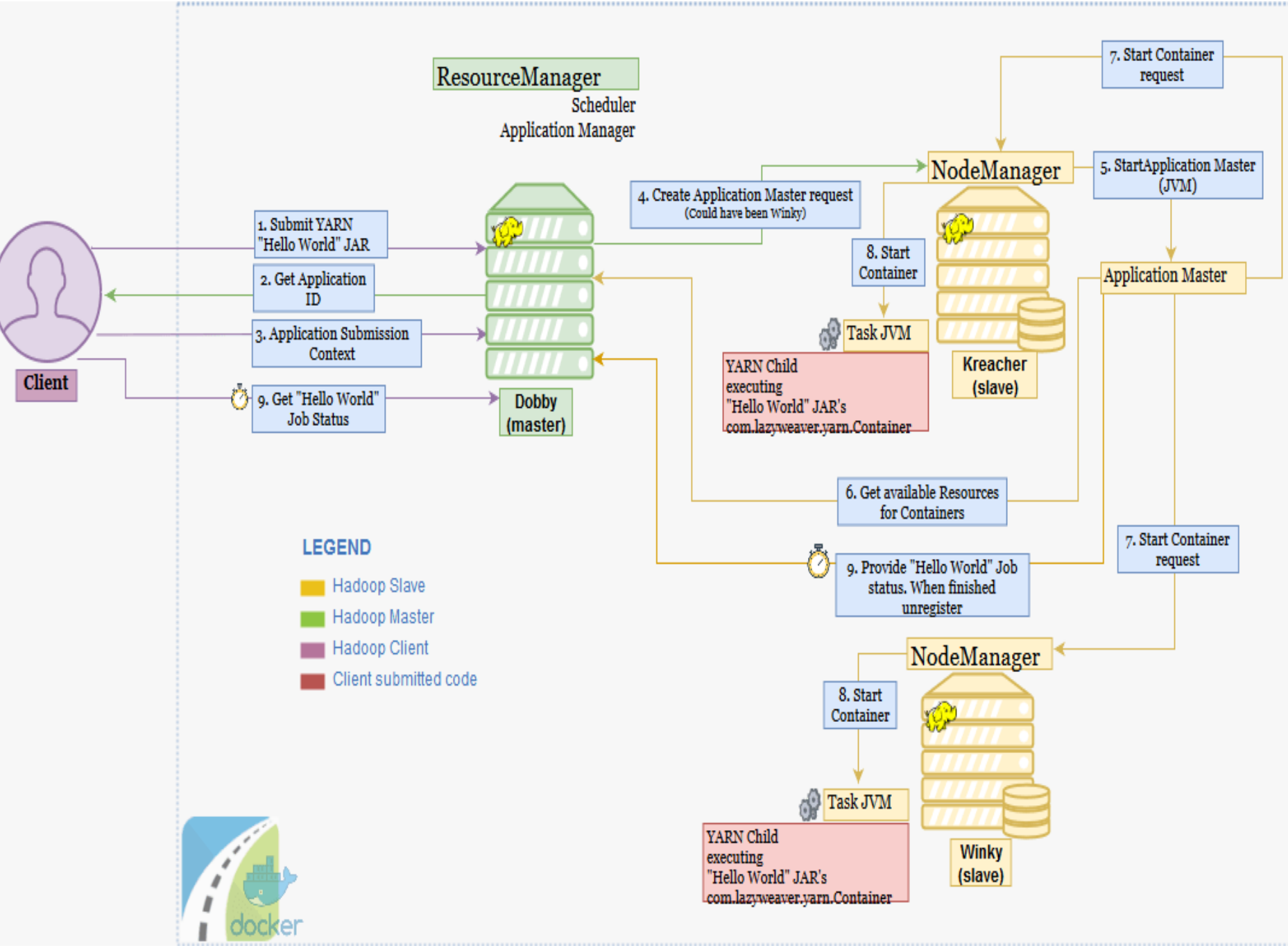


Run an application on YARN

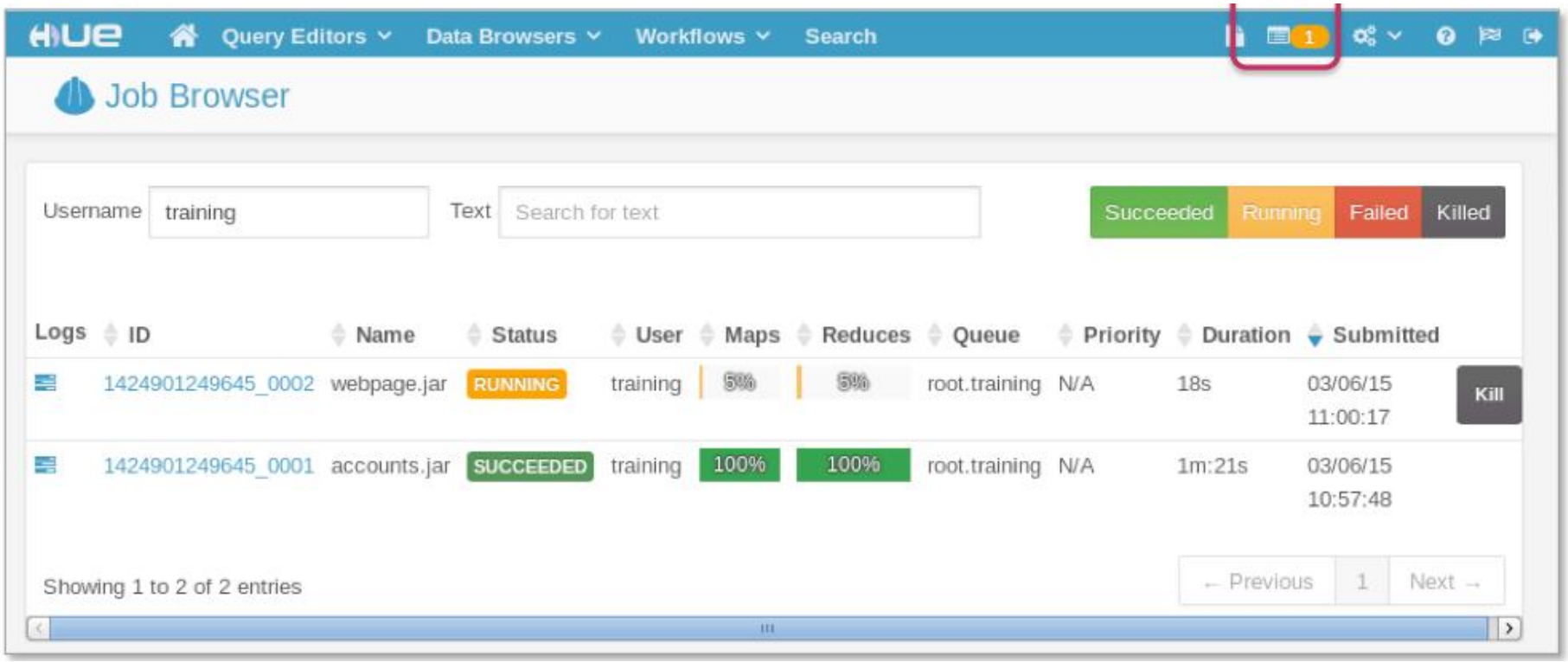


Cluster viewYARN



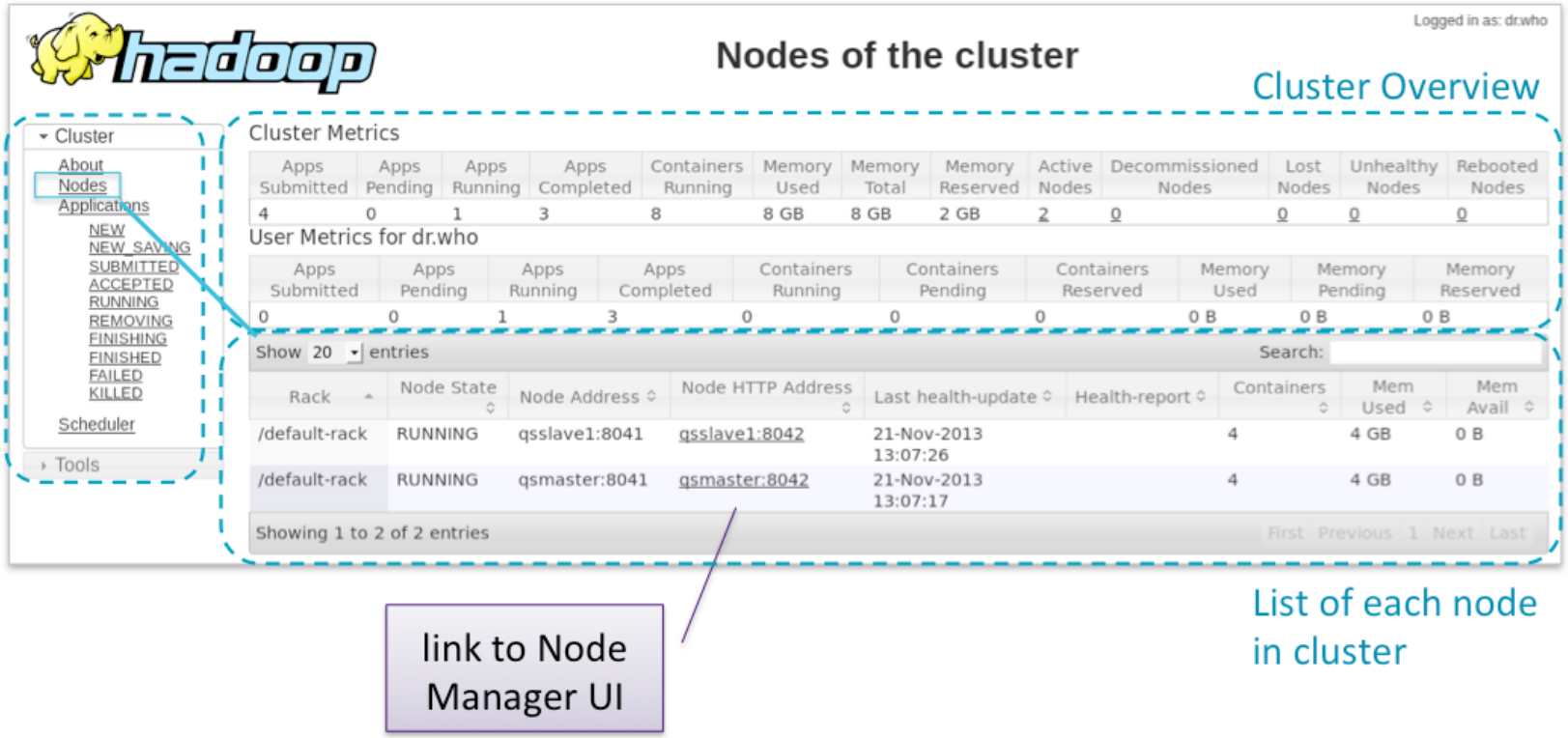

ResourceManager
Scheduler
Application Manager
7. Start Container request
NodeManager
5. StartApplication Master (JVM)
4. Create Application Master request
(Could have been Winky)
1. Submit YARN "Hello World" JAR
2. Get Application ID
3. Application Submission Context
9. Get "Hello World" Job Status
Client
Dobby (master)
8. Start Container
Application Master
Task JVM
YARN Child executing "Hello World" JAR's com.lazyweaver.yarn.Container
Kreacher (slave)
6. Get available Resources for Containers
9. Provide "Hello World" Job status. When finished unregister
7. Start Container request
LEGEND
Hadoop Slave
Hadoop Master
Hadoop Client
Client submitted code
NodeManager
8. Start Container
Task JVM
YARN Child executing "Hello World" JAR's com.lazyweaver.yarn.Container
Winky (slave)
docker

HUE Browser



Analytics Vidhya
Learn everything about analytics

YARN Manger Web UI



Analytics Vidhya
Learn everything about analytics

Job History Server

YARN does not keep track of job history