

# **BIGDATA**

**AND HADOOP ADMINISTRATOR**

A stylized, blue-outlined illustration of an elephant's head and trunk, rendered in a folded paper or origami style. The elephant is facing right and is positioned to the right of the text, partially overlapping the word 'BIGDATA'.

# What's in it for you?

---

Use case of Hadoop

Demo on HDFS, MapReduce  
and YARN

Why Hadoop?

Hadoop YARN

What is Hadoop?

Hadoop MapReduce

Hadoop HDFS



# Why Hadoop?



In a town far away..



Tim sells food grains in his shop



The customers were happy as Tim was very quick with the orders



Tim sensed a good demand for other products, so he thought of expanding his business



He started selling fruits, vegetables, meat, and dairy products in addition to food grains





But it wasn't as easy as he expected it to be. The number of customers increased, and he was not able to cater to their needs on time



He had to look into assisting his customers with each of their orders and billing. It was too difficult for him to manage alone



To start delivering orders on time and to manage the customers' demands, Tim hired 3 more people to work with him



Matt took care of the fruits and vegetable section.  
Luke handled the dairy and meat section. Ann was  
appointed as the cashier



Tim



Matt



Luke



Ann



However, this was still not a solution to Tim's problem as there was not enough space in the shop for all the items





The storage was a bottleneck since storing and accessing became more and more difficult with increased supply and demand



Tim came up with an idea to overcome this issue. He decided to expand the storage area and distribute each category of product on different floors




Now, customers were happy, and after picking up their products from the respective sections, it was then billed



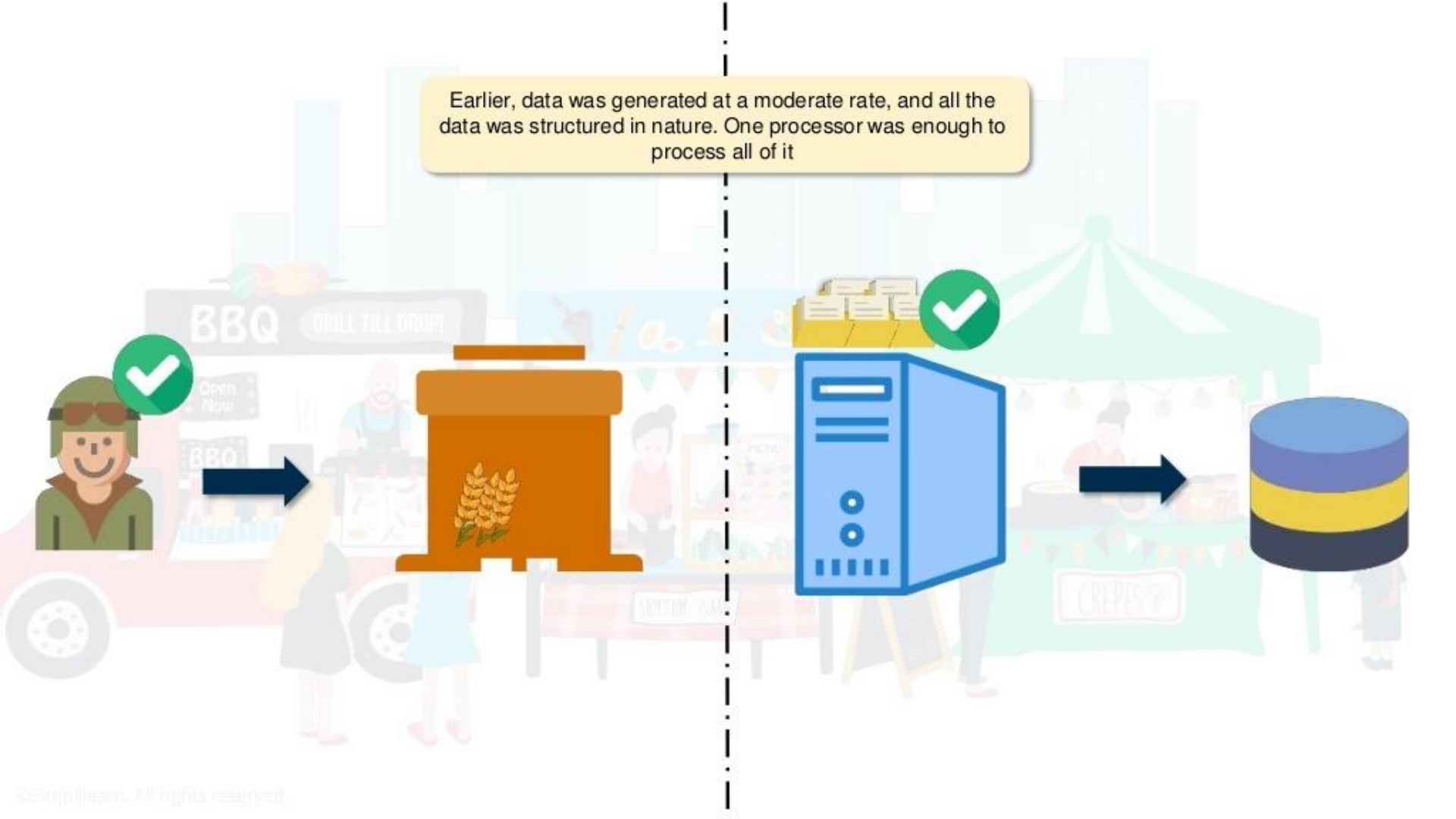


Now, customers were happy, and after picking up their products from the respective sections, it was then billed

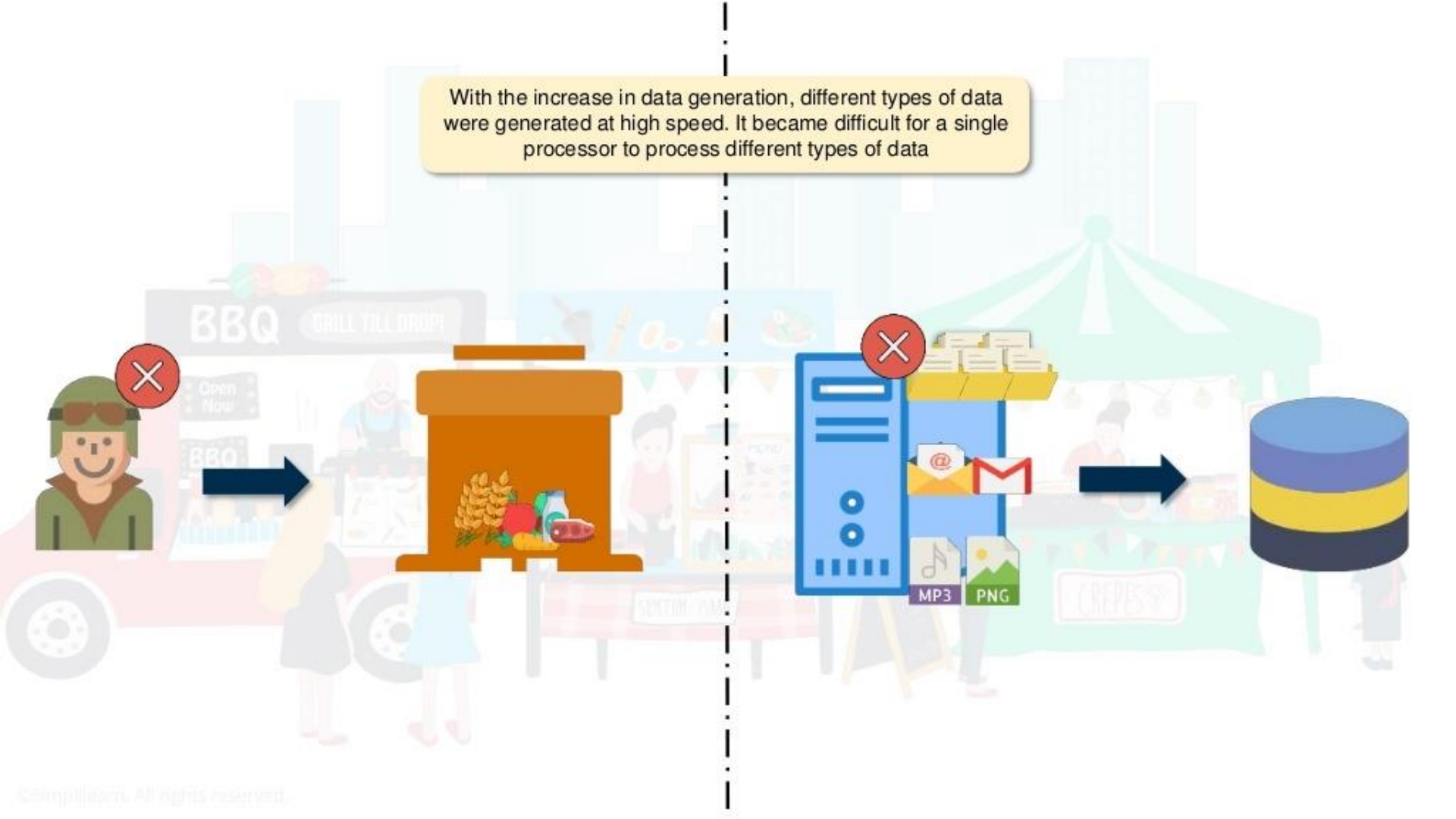


Now, let us compare this story to big data

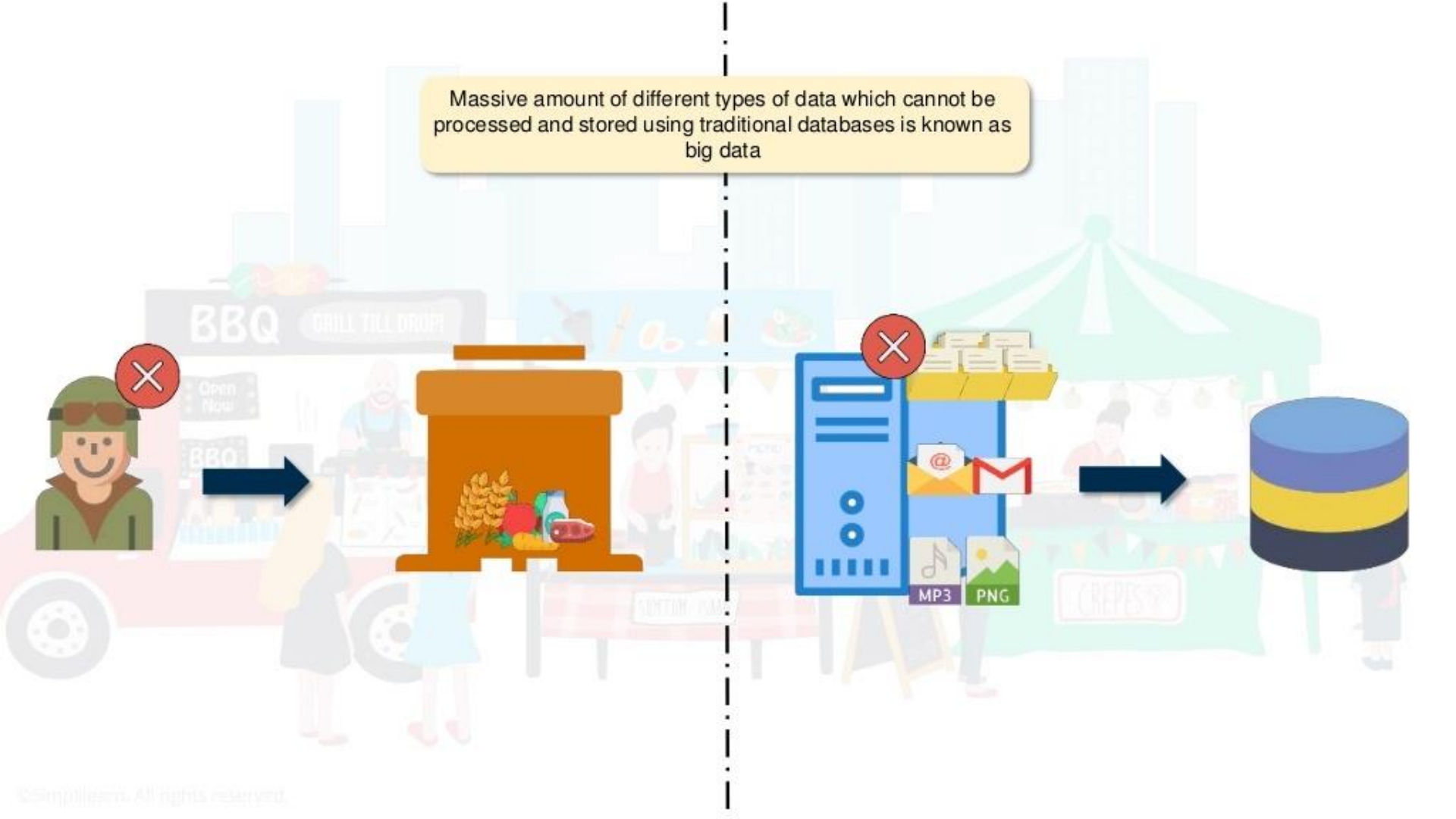
Earlier, data was generated at a moderate rate, and all the data was structured in nature. One processor was enough to process all of it



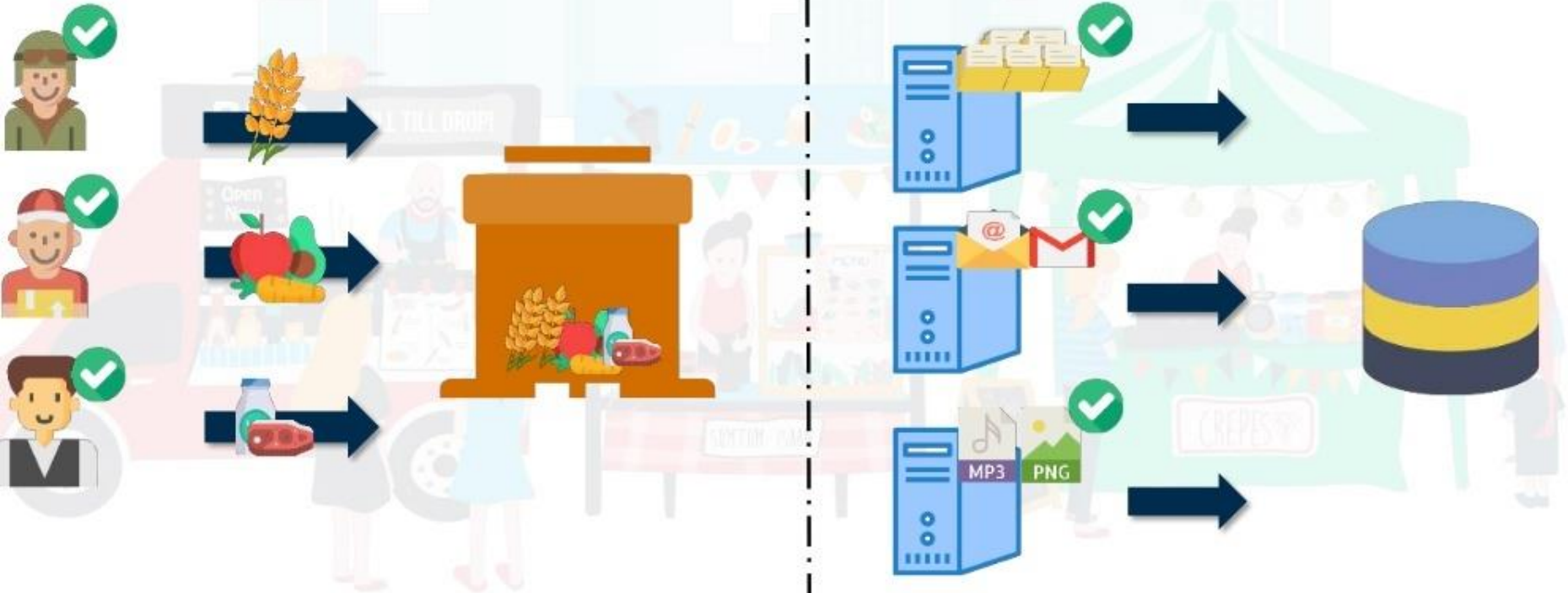
With the increase in data generation, different types of data were generated at high speed. It became difficult for a single processor to process different types of data



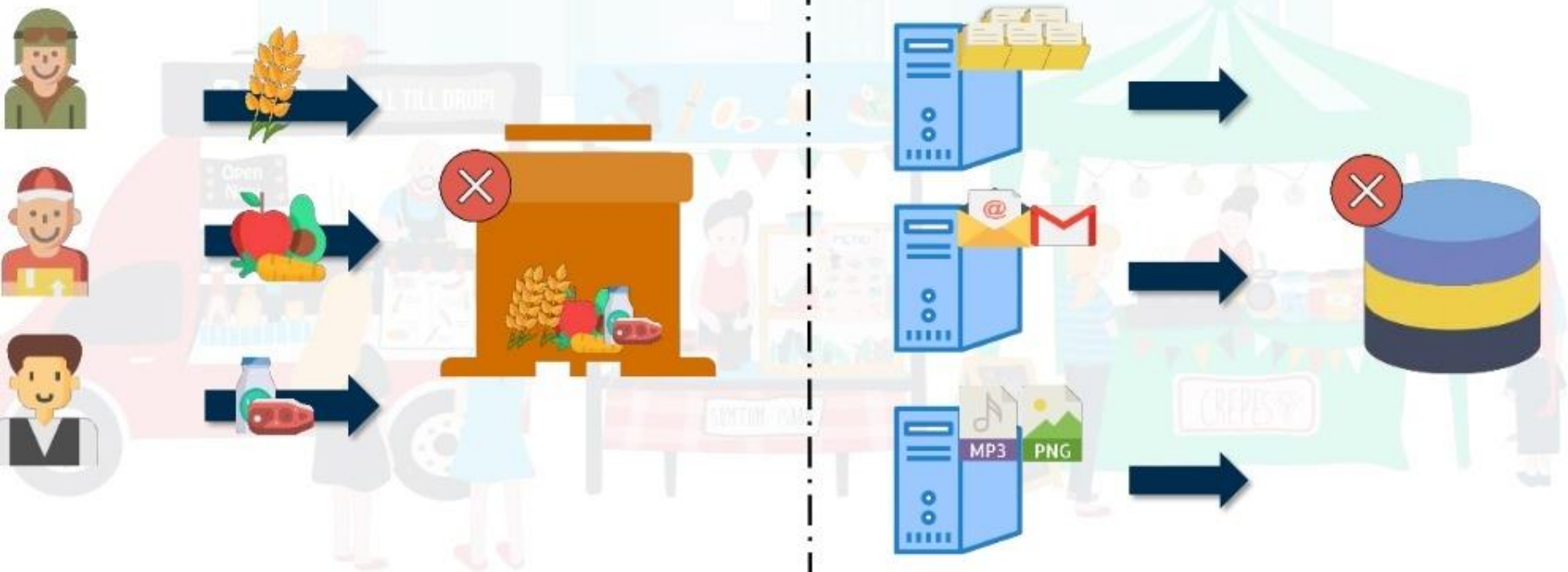
Massive amount of different types of data which cannot be processed and stored using traditional databases is known as big data



To overcome this issue, multiple processors were used to process each type of data

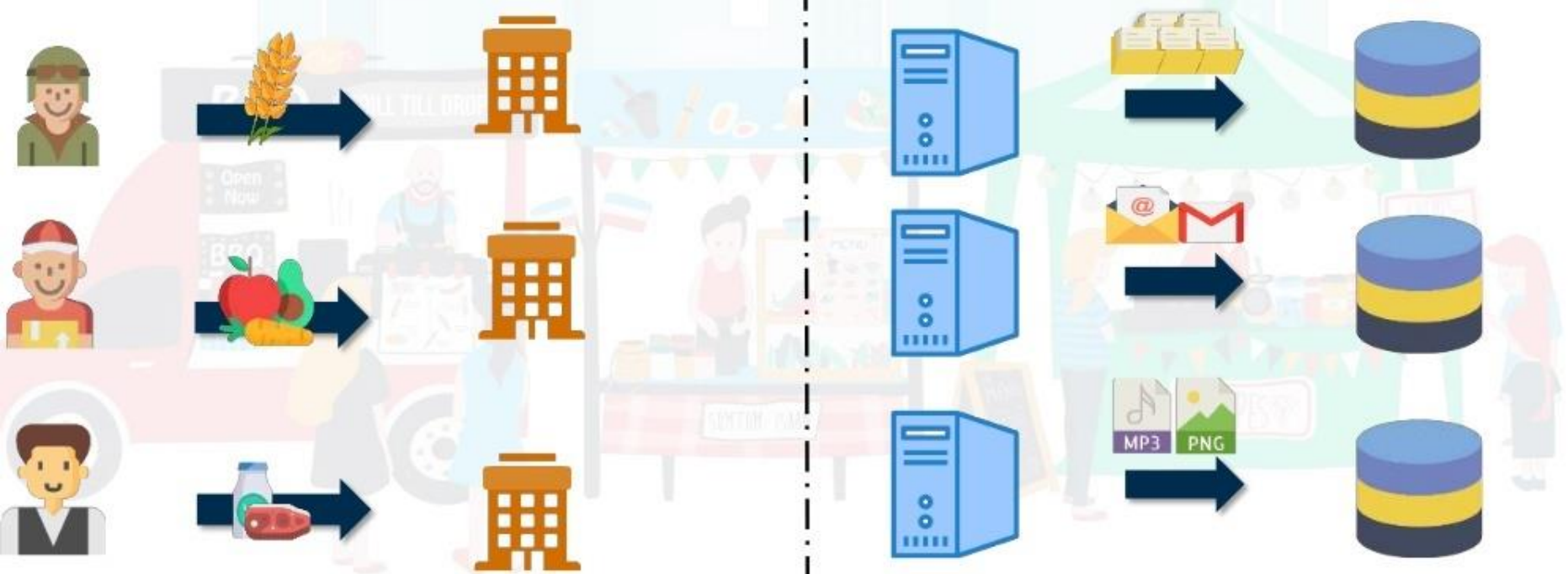


But now the problem was that one storage system was accessed by all the processors and the storage became the bottleneck





Just like how Tim adopted the distributed approach, the storage system was also distributed and by doing so, the data was stored in individual databases



Just like how Tim adopted the distributed approach, the storage system was also distributed and by doing so, the data was stored in individual databases

Through this story, we see the two approaches that are used by Hadoop that is HDFS and MapReduce





HDFS refers to the distributed storage space just like how Tim distributed the storage space amongst the various sections



Each person took care of a separate section and at the end the customers went to the cashier for the final billing, this sorted the process and made it easier. This is how Hadoop MapReduce works



This was a rough story of big data generation and why Hadoop is required. I will now explain in detail as to what Hadoop is





This sounds interesting. I would like to know more about Hadoop

# What is Hadoop?



# What is Hadoop?

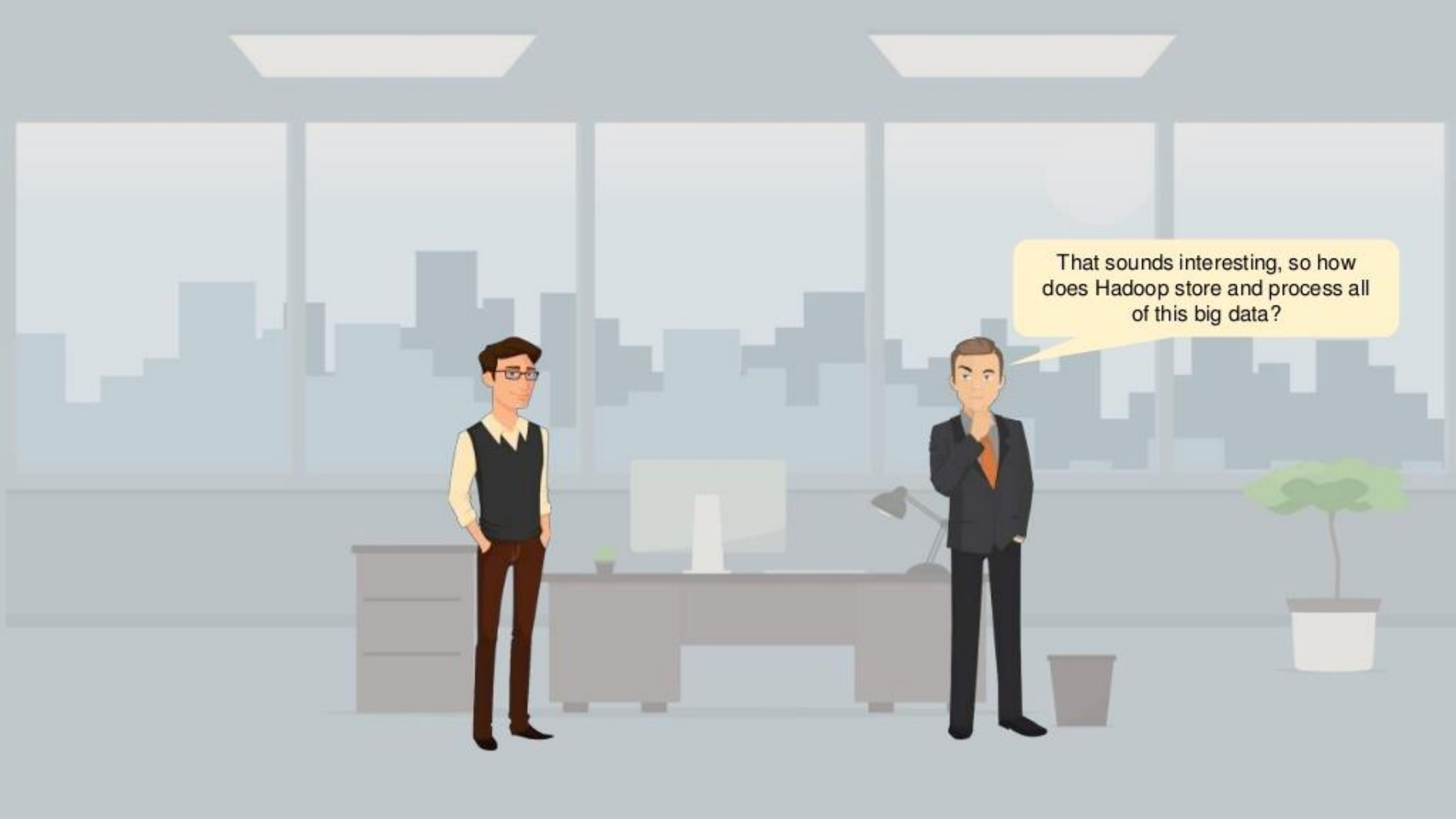
---

Hadoop is a framework which stores and processes big data in a distributed and parallel fashion

# What is Hadoop?

Hadoop is a framework which stores and processes big data in a distributed and parallel fashion





That sounds interesting, so how does Hadoop store and process all of this big data?



Hadoop has individual components, which are used for storing and processing big data



Components of Hadoop

HDFS

The storage unit of Hadoop



## Components of Hadoop



HDFS

The storage unit of Hadoop

MapReduce

The processing unit of Hadoop

Components of Hadoop

HDFS

The storage unit of Hadoop

MapReduce

The processing unit of Hadoop

YARN

The resource management unit of Hadoop

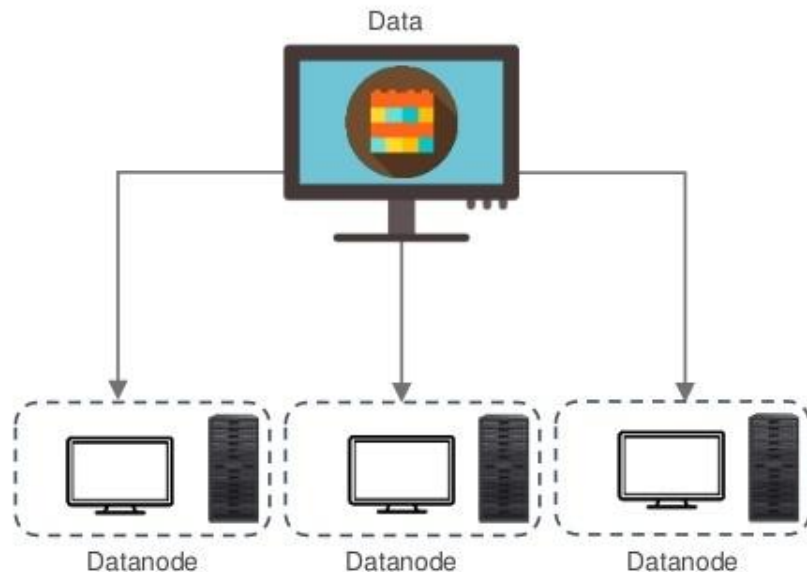


# Hadoop HDFS



# What is HDFS?

Hadoop Distributed File System (HDFS) is known for its distributed storage method. It distributes the data amongst many computers. In addition to this, replication of data is also done to avoid loss of data



Each block of data is stored on multiple systems and by default has 128 MB of data

## What is HDFS?

---

Let us now see how 500 MB of data is stored in the traditional method

# What is HDFS?

---

Let us now see how 500 MB of data is stored in the traditional method

500 MB data





# What is HDFS?

Let us now see how 500 MB of data is stored in the traditional method

500 MB data



Here, the entire set of data is stored in one database. This overloads the database, and if it crashes, we lose all our data

# What is HDFS?

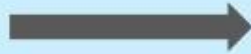
---

Using Hadoop HDFS, this problem is taken care of as data is distributed amongst many systems

# What is HDFS?

Using Hadoop HDFS, this problem is taken care of as data is distributed amongst many systems

500 MB data



By doing so, a single database is not overloaded

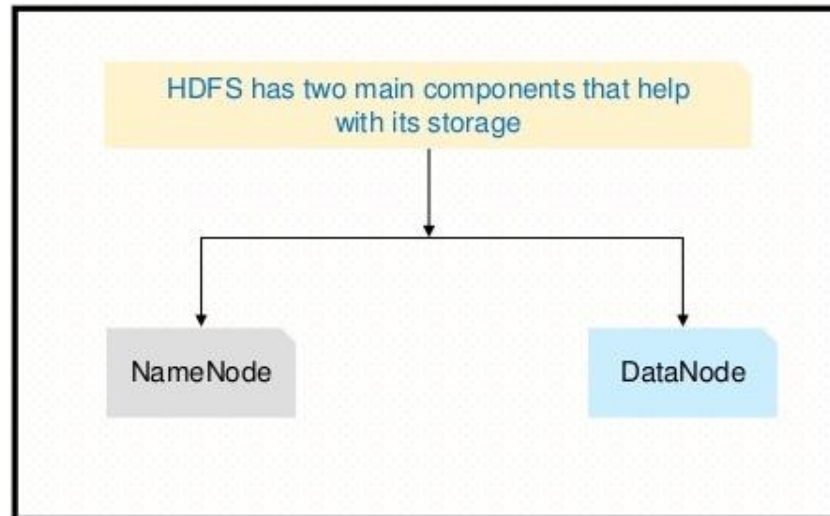
# What is HDFS?

---

Hadoop Distributed File System (HDFS) is specially designed for storing massive datasets in commodity hardware

# What is HDFS?

Hadoop Distributed File System (HDFS) is specially designed for storing massive datasets in commodity hardware



# What is HDFS?

---



NameNode

- NameNode is the master of the system
- It stores all the metadata



DataNode



DataNode



DataNode



DataNode



# What is HDFS?

---

- DataNode is known as the slave node. There are multiple DataNodes
- It performs the read/write operations and stores the actual data



NameNode



DataNode



DataNode



DataNode

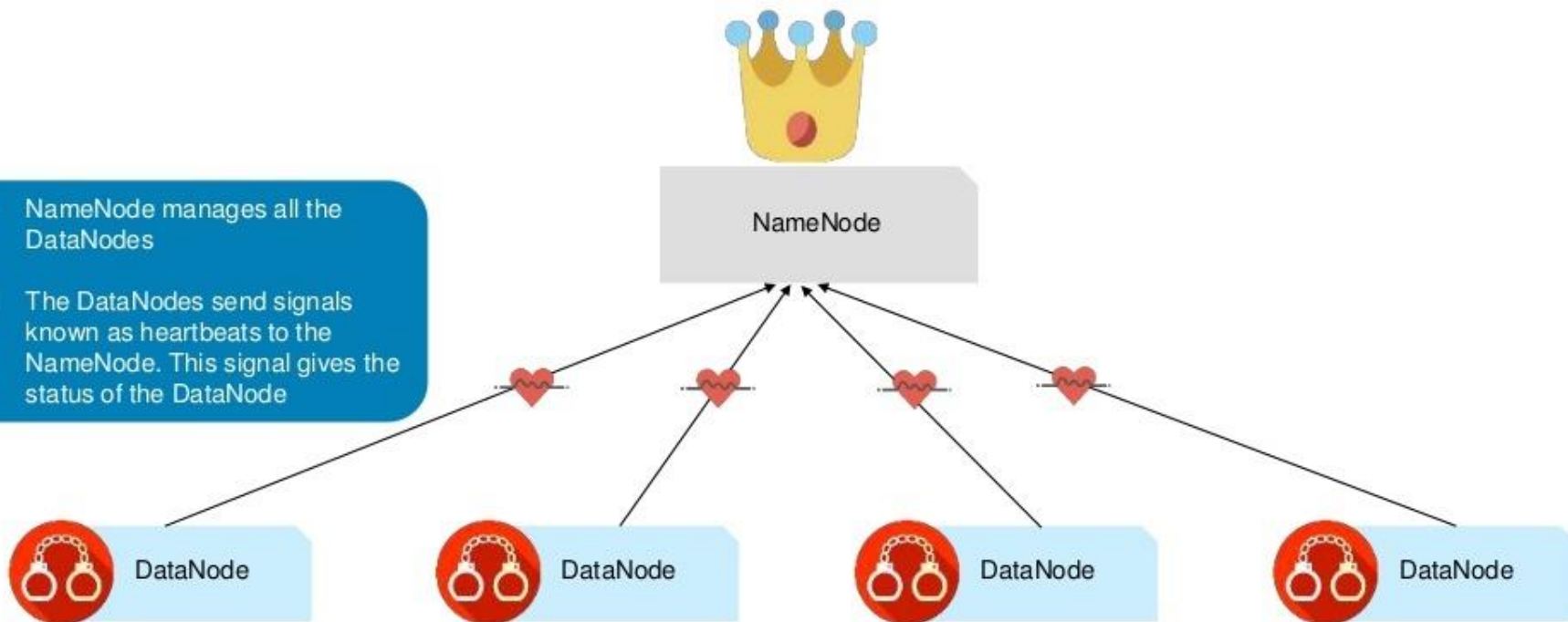


DataNode



# What is HDFS?

- NameNode manages all the DataNodes
- The DataNodes send signals known as heartbeats to the NameNode. This signal gives the status of the DataNode





## What is HDFS?

---

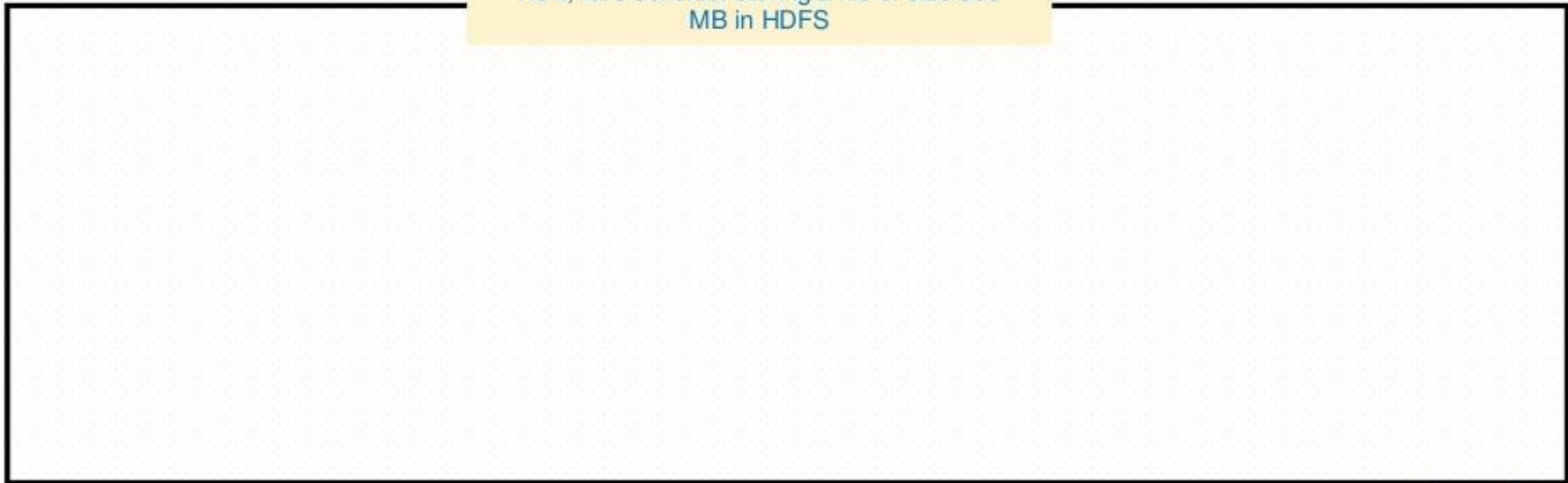
As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

# What is HDFS?

---

As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

Now, let's consider storing a file of size 530 MB in HDFS



# What is HDFS?

---

As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

Now, let's consider storing a file of size 530 MB in HDFS

File.txt  
530 MB

# What is HDFS?

As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

Now, let's consider storing a file of size 530 MB in HDFS

File.txt  
530 MB

Block A

128 MB

Block B

128 MB

Block C

128 MB

Block D

128 MB

# What is HDFS?

As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

Now, let's consider storing a file of size 530 MB in HDFS

File.txt  
530 MB

Block A

128 MB

Block B

128 MB

Block C

128 MB

Block D

128 MB

Block E

18 MB

# What is HDFS?

As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

Now, let's consider storing a file of size 530 MB in HDFS

File.txt  
530 MB

Block A

128 MB

Block B

128 MB

Block C

128 MB

Block D

128 MB

Block E

18 MB

The final block uses only the remaining space for storage

# What is HDFS?

As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

Now, let's consider storing a file of size 530 MB in HDFS

File.txt  
530 MB

DataNode 1



DataNode 2



DataNode 3



DataNode 4



DataNode 5



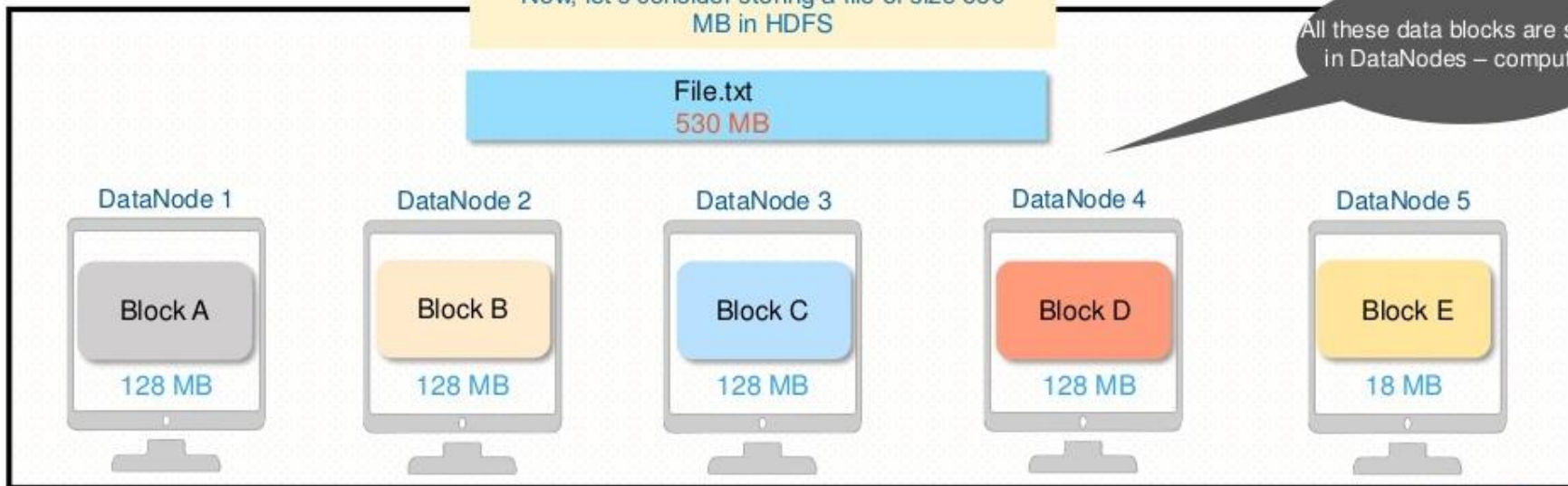
# What is HDFS?

As mentioned earlier, the actual data is stored in DataNodes. Data is stored in the form of blocks here. The default size of each block is 128 MB

Now, let's consider storing a file of size 530 MB in HDFS

File.txt  
530 MB

All these data blocks are stored in DataNodes – computers







What happens if the computer that contains block A crashes? Do we lose the data in block A?

No, we don't. That's the beauty of Hadoop HDFS. It uses replication to prevent the loss of data



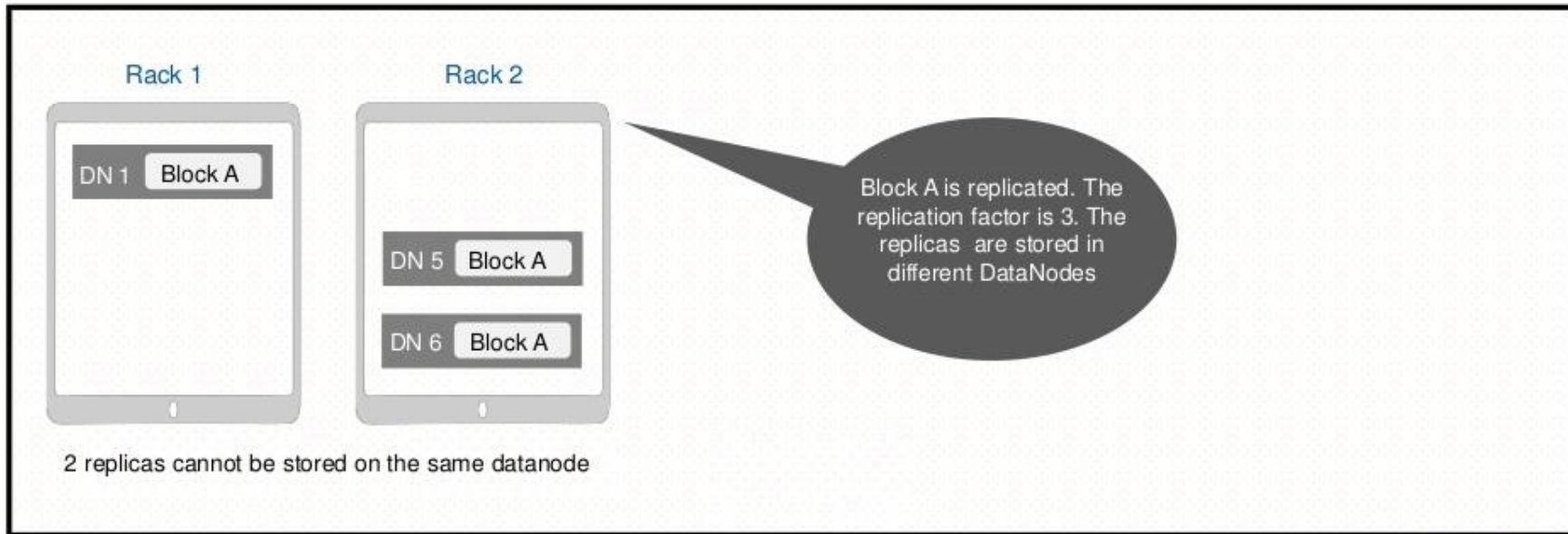
# Replication in HDFS

HDFS overcomes the issue of DataNode failure by creating copies of the data; this is known as the replication method



# Replication in HDFS

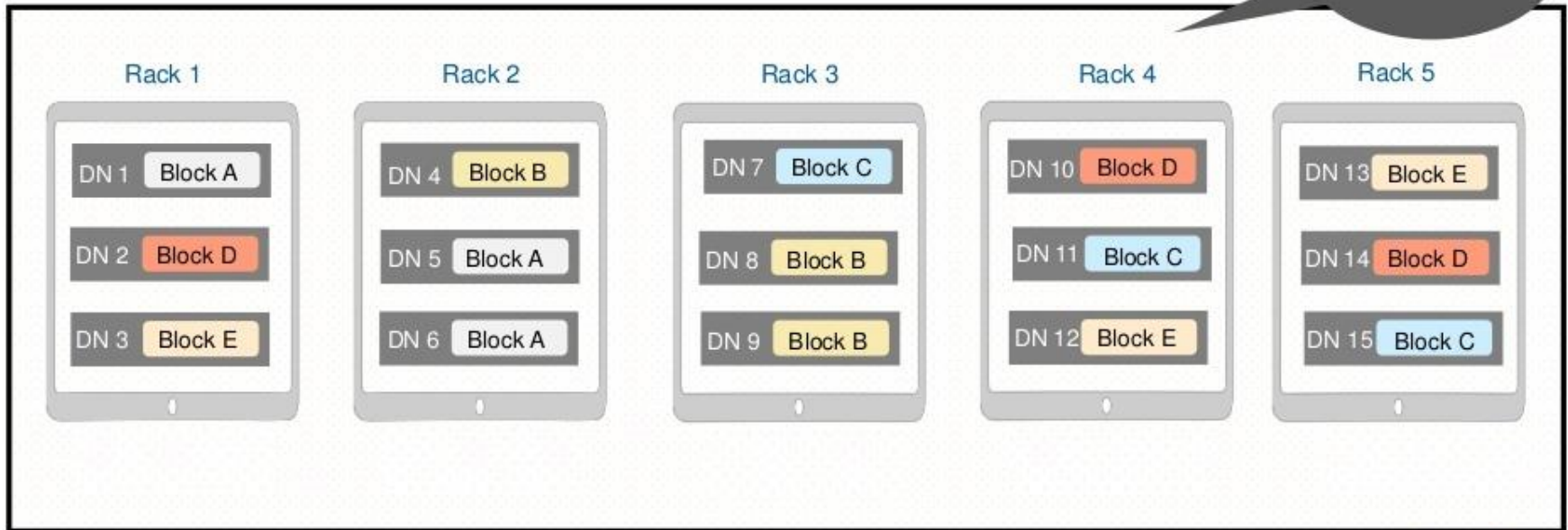
HDFS overcomes the issue of DataNode failure by creating copies of the data; this is known as the replication method



# Replication in HDFS

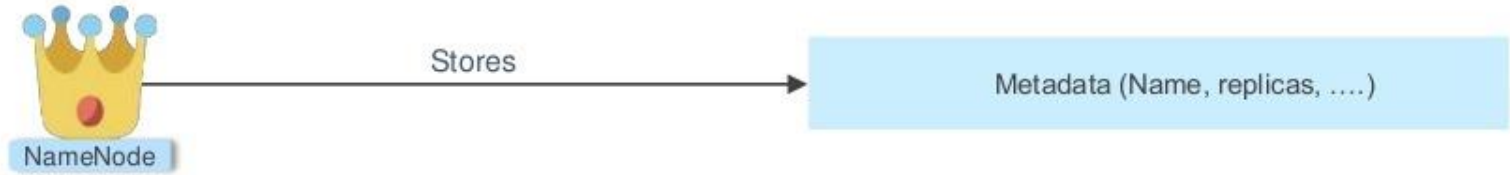
HDFS overcomes the issue of DataNode failure by creating copies of the data; this is known as the replication method

Similarly, every other block is replicated

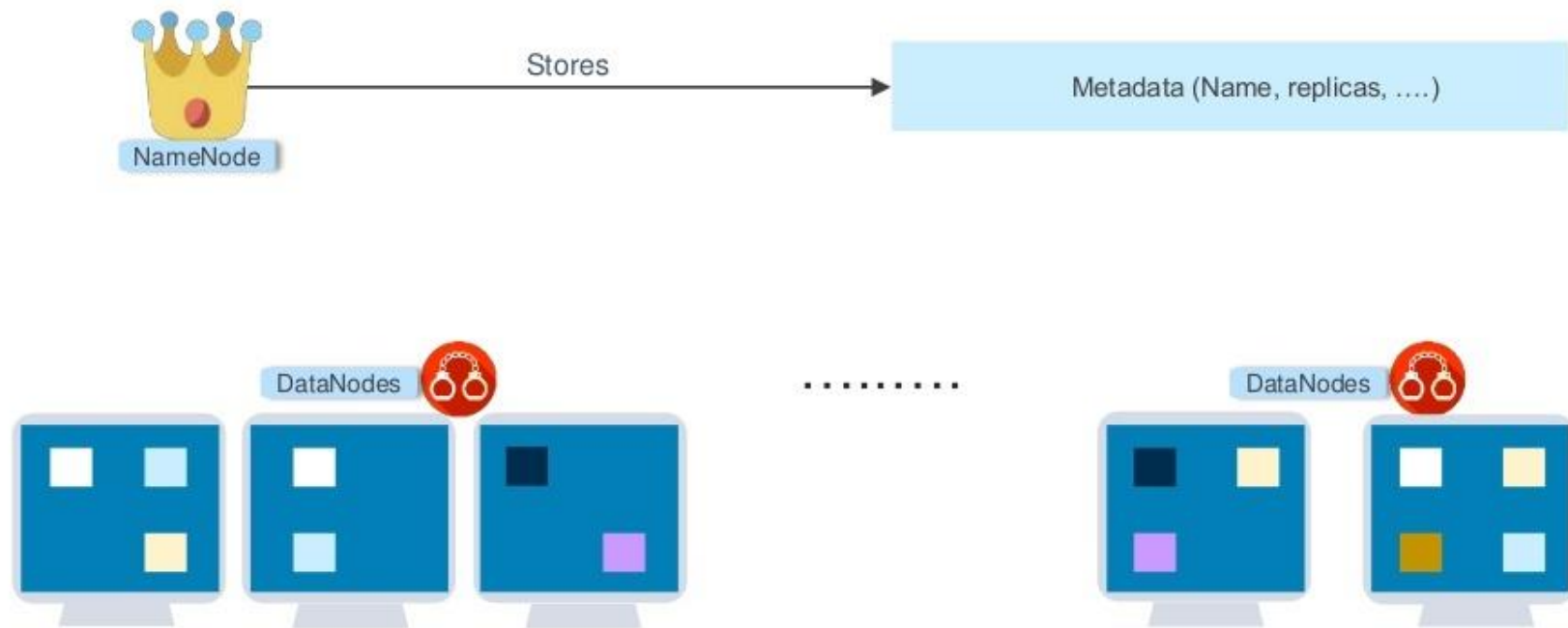


# Architecture of HDFS

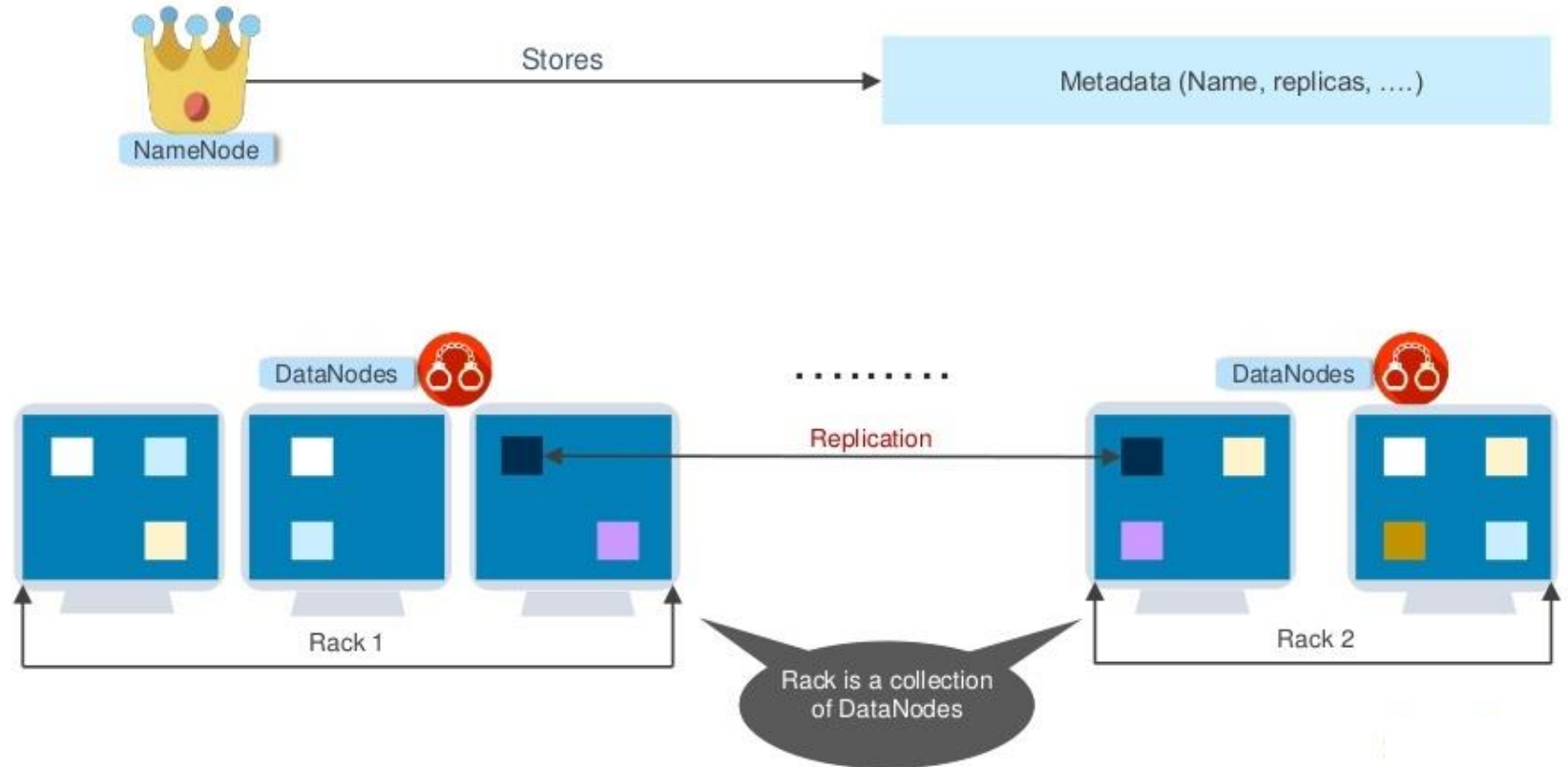
---



# Architecture of HDFS

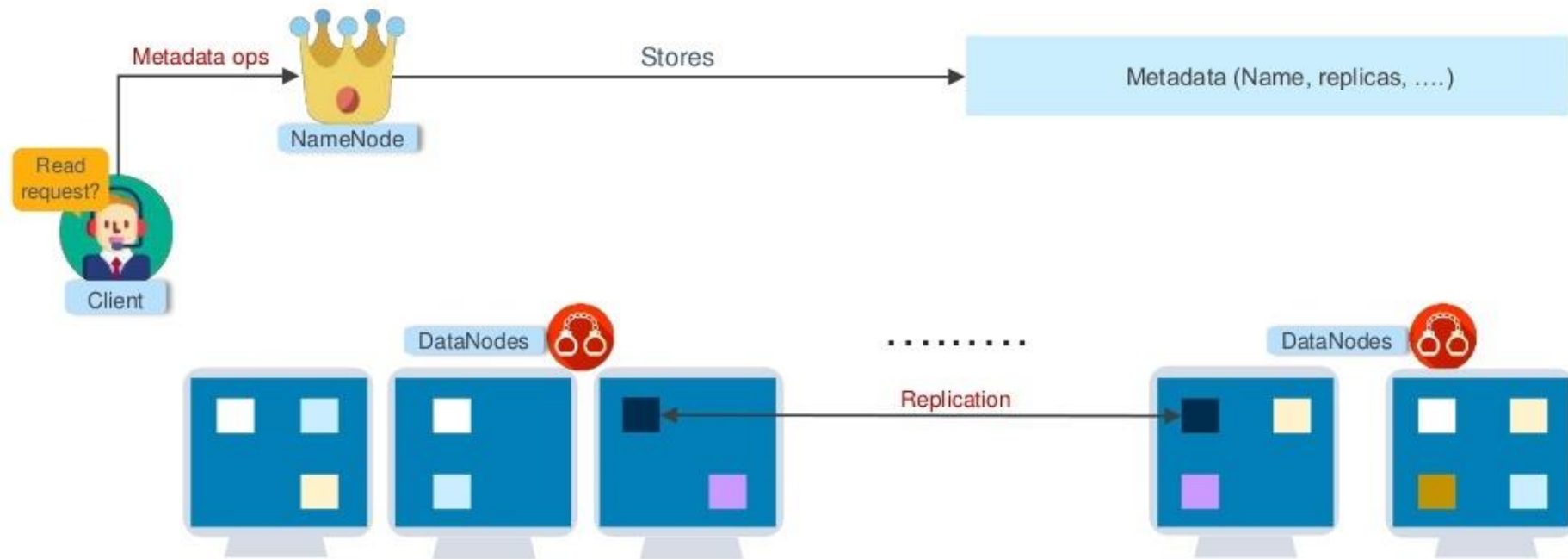


# Architecture of HDFS

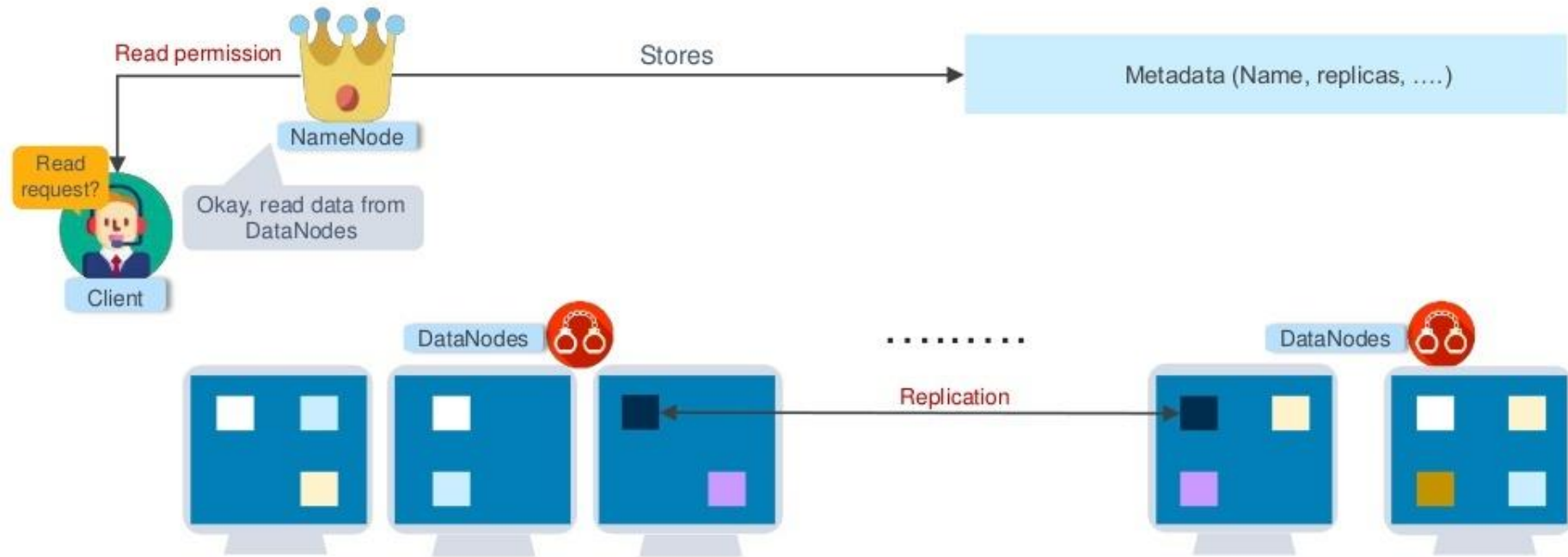




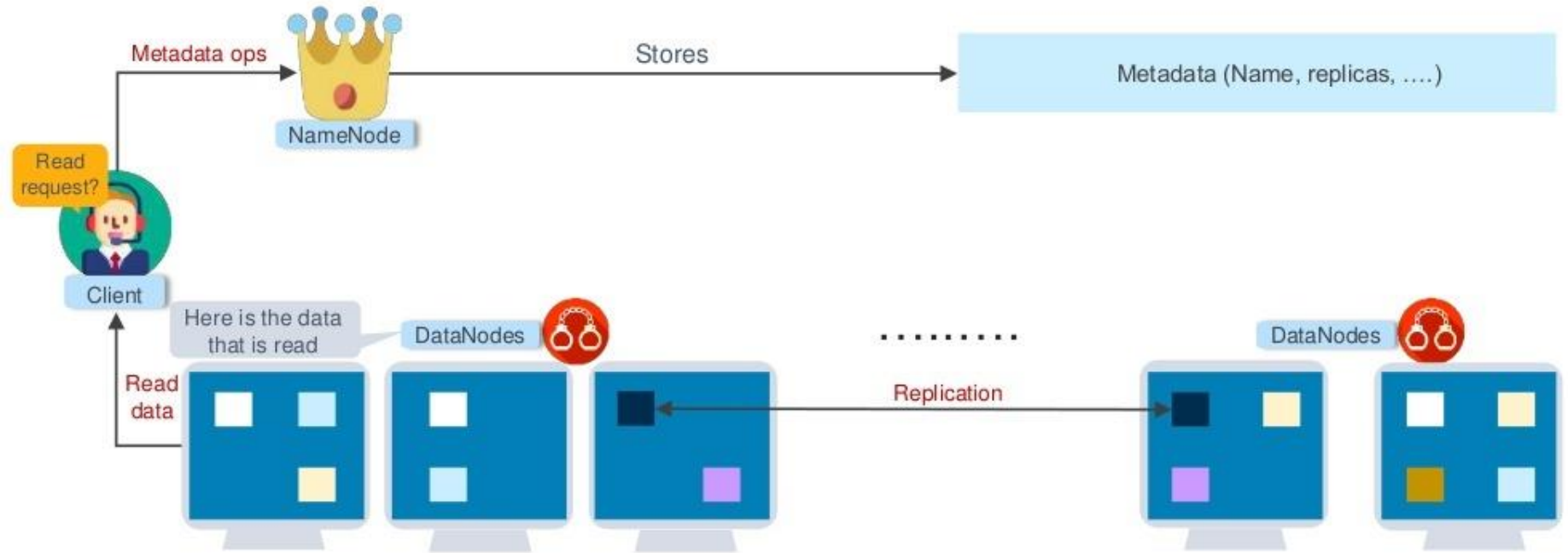
# Architecture of HDFS



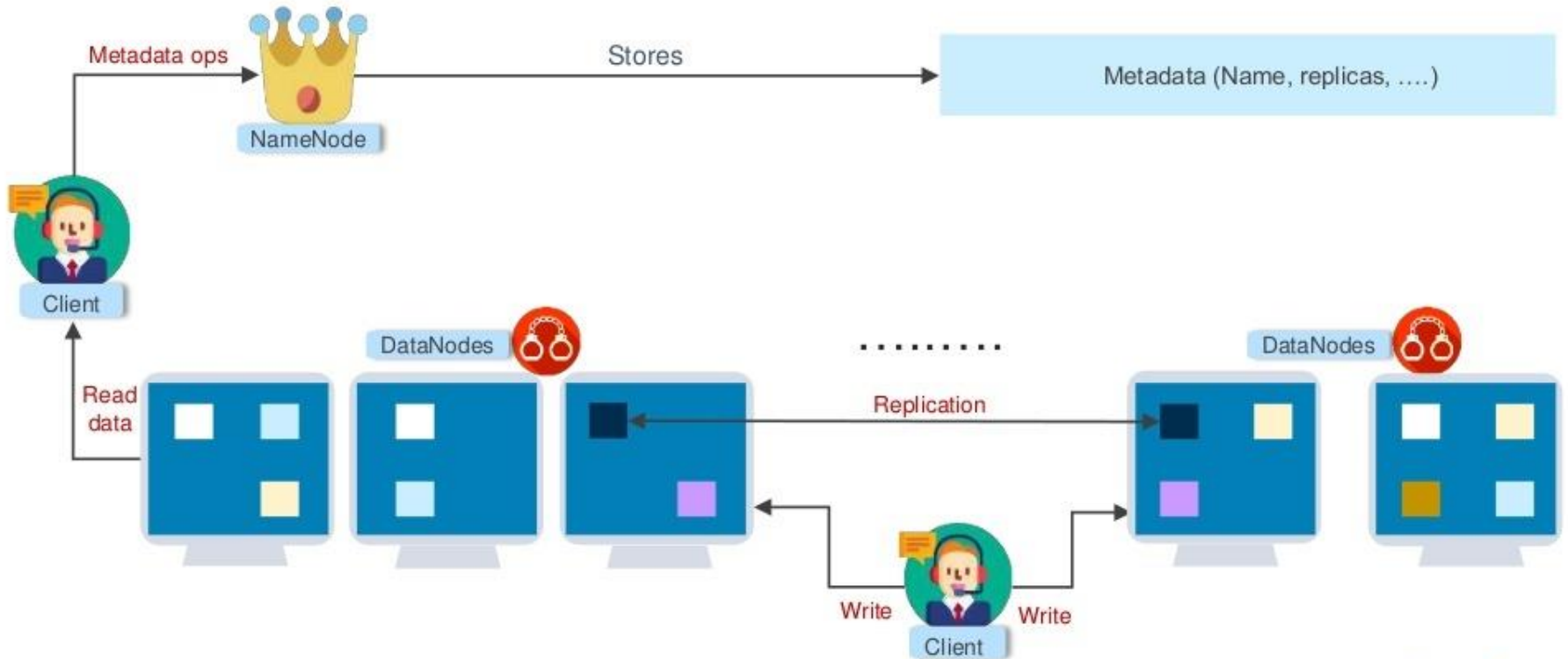
# Architecture of HDFS



# Architecture of HDFS



# Architecture of HDFS



# Features of HDFS

Fault tolerant



HDFS is fault tolerant as multiple copies of data are made

Data security



Scalability



Flexibility



# Features of HDFS

Fault tolerant



Data security



Provides end-to-end encryption that protects data

Scalability



Flexibility



# Features of HDFS

Fault tolerant



Data security



Scalability



Multiple nodes can be added to the cluster depending on the requirement

Flexibility



# Features of HDFS

Fault tolerant



Data security



Scalability

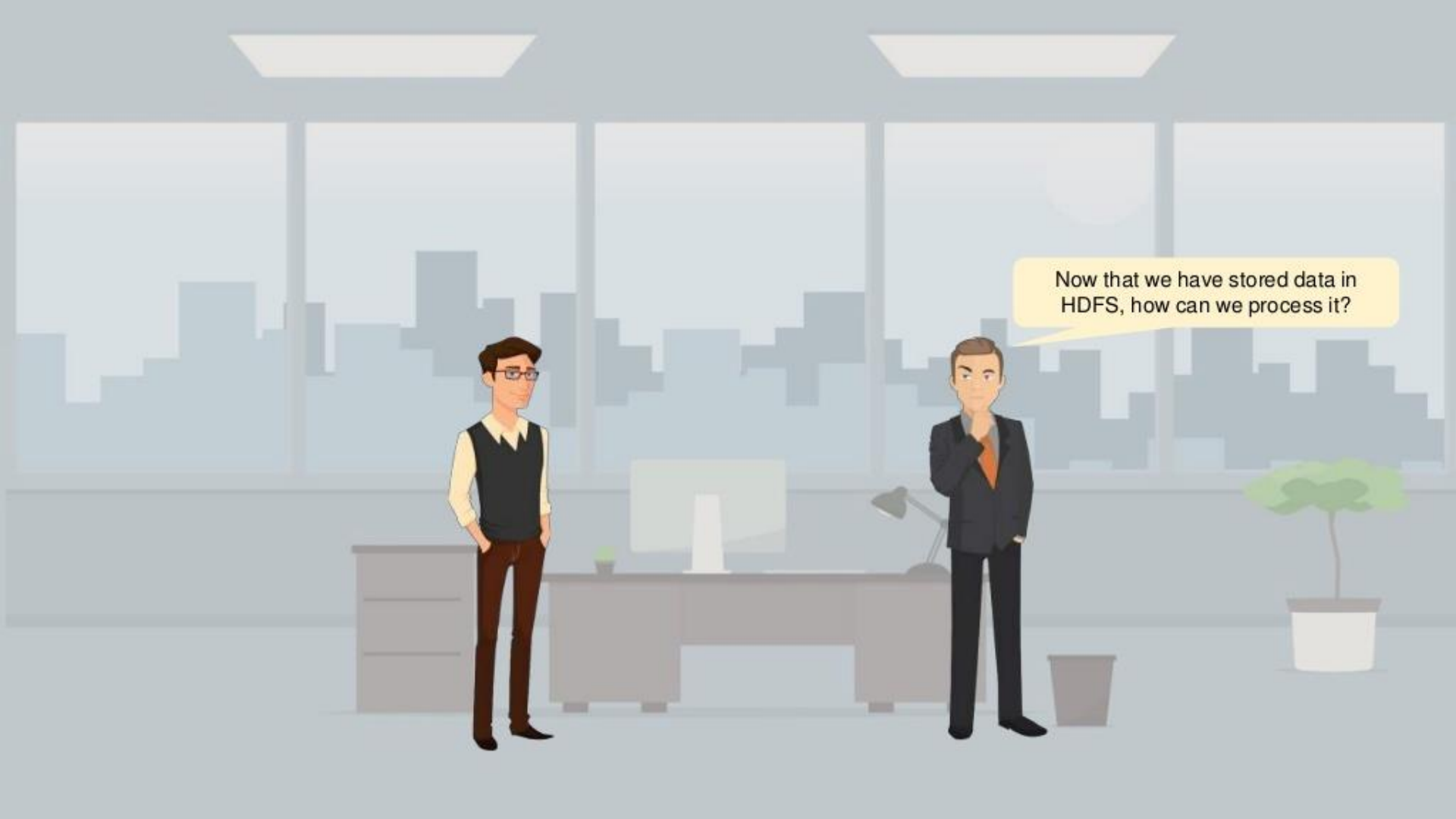


Flexibility



Hadoop is flexible in storing any type of data, like structured, semi structured or unstructured data





Now that we have stored data in HDFS, how can we process it?

For processing data, Hadoop has a unit known as MapReduce



# Why MapReduce?

In the traditional approach, big data was processed at the master node



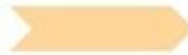
big data

# Why MapReduce?

In the traditional approach, big data was processed at the master node



big data

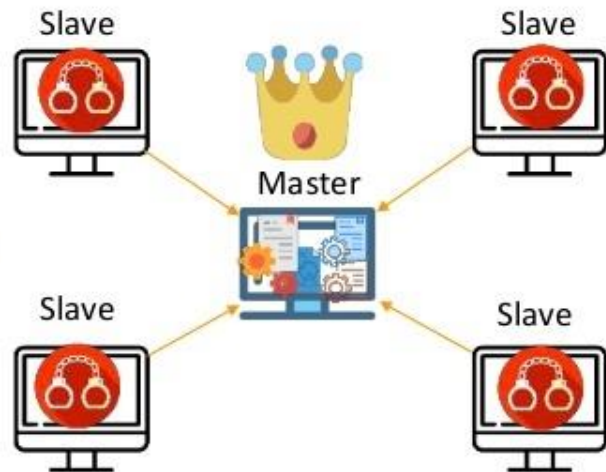


# Why MapReduce?

This was a disadvantage as it consumed more time to process various types of data



big data

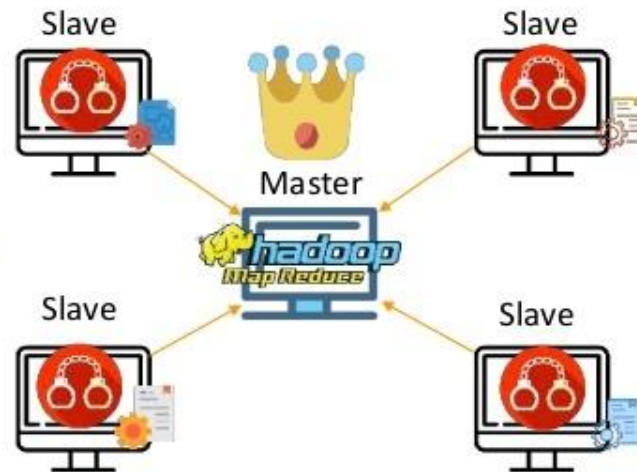


# Why MapReduce?

To overcome this issue, data was processed at each slave node. This approach is known as MapReduce



big data



# Hadoop MapReduce



# What is MapReduce?

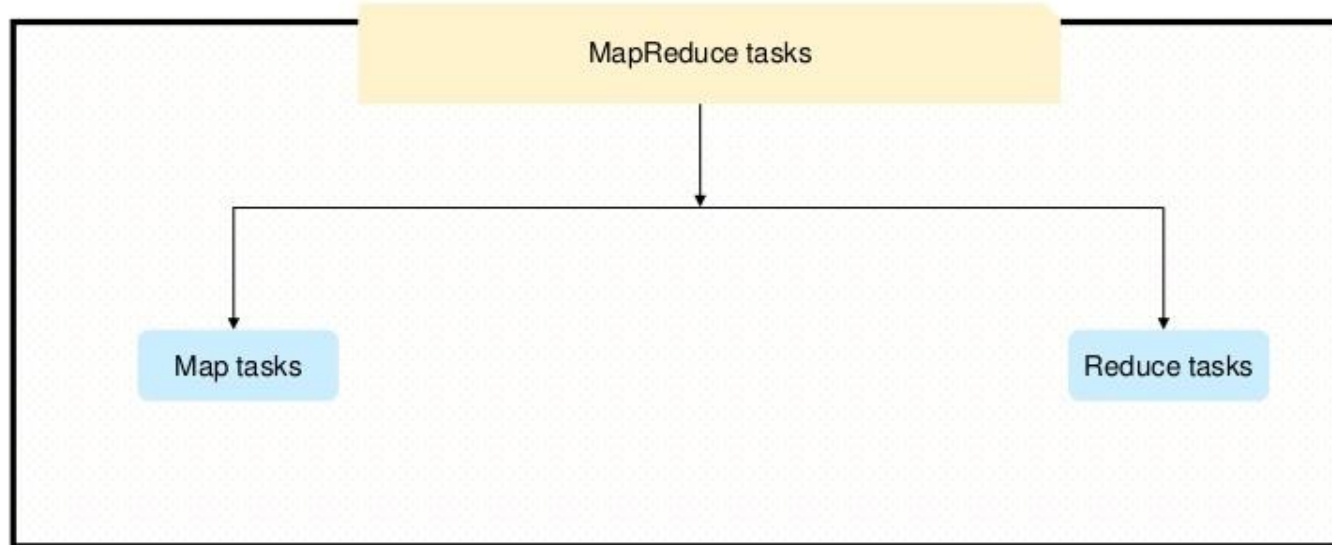
---

Programming technique where huge data is processed in a parallel and distributed fashion is known as Hadoop MapReduce

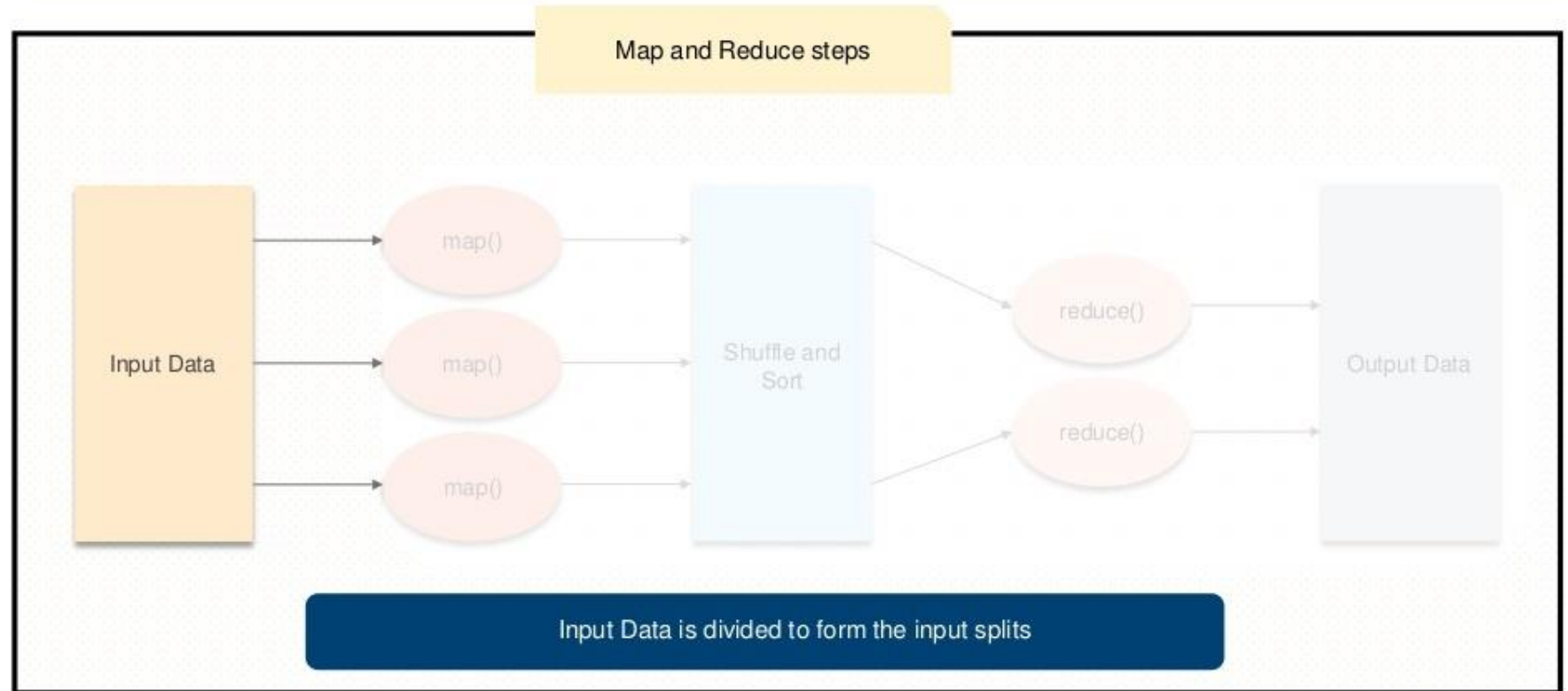


# What is MapReduce?

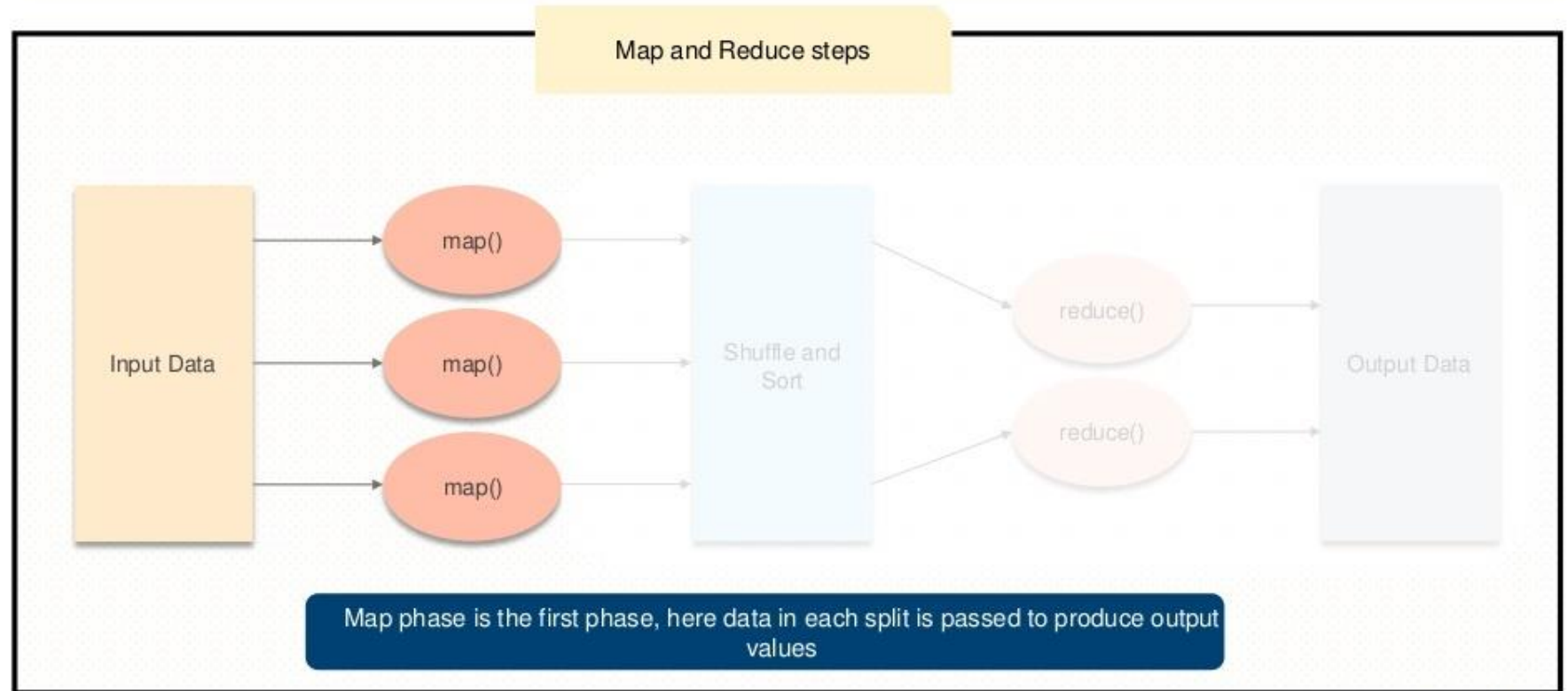
Programming technique where huge data is processed in a parallel and distributed fashion is known as Hadoop MapReduce



# What is MapReduce?

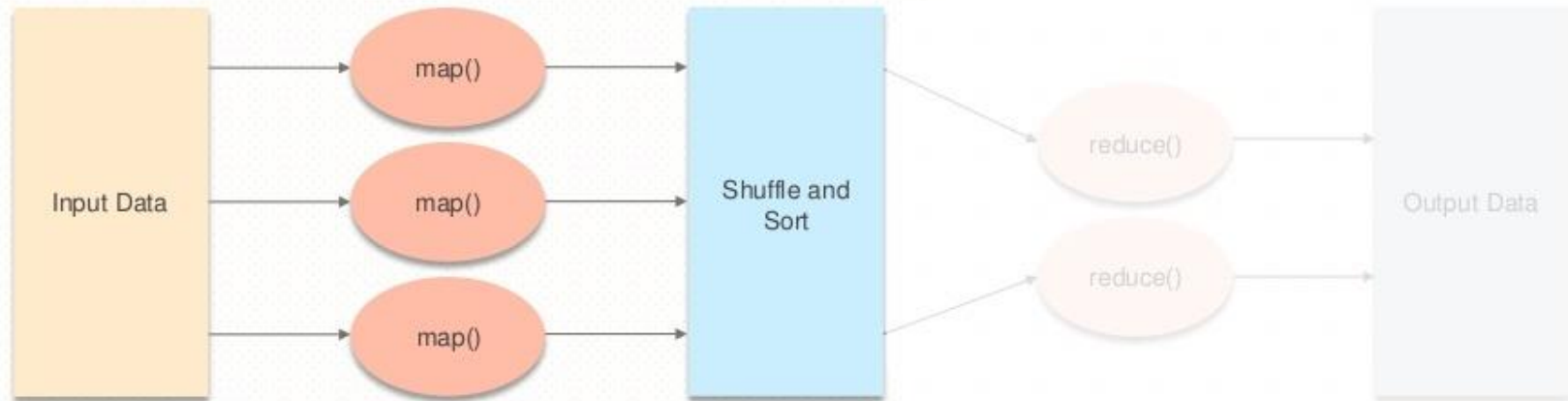


# What is MapReduce?



# What is MapReduce?

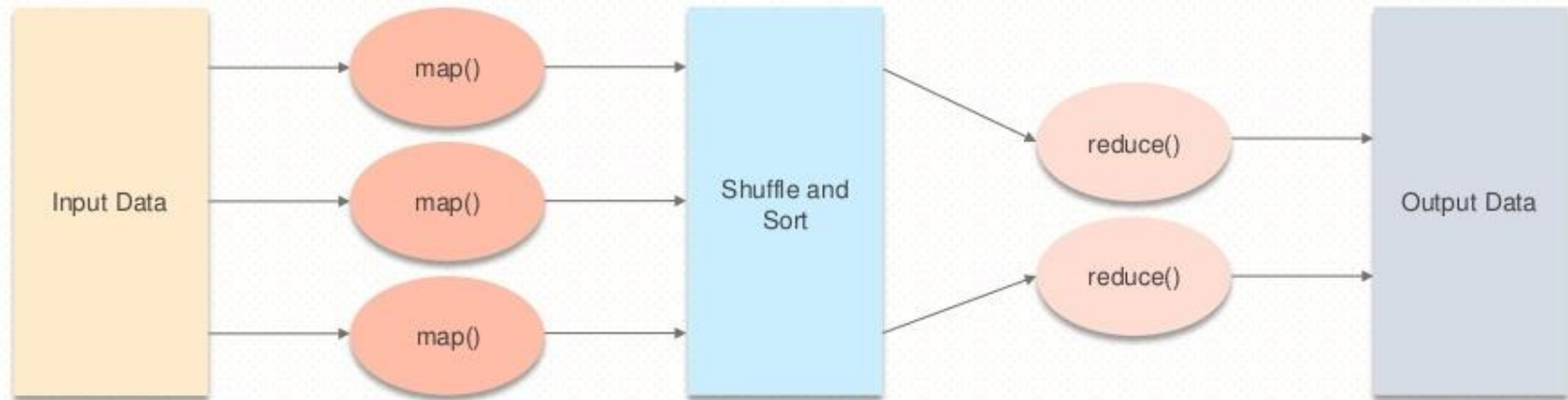
Map and Reduce steps



In the shuffle and sort phase, output of mapping phase is taken and similar data is grouped

# What is MapReduce?

Map and Reduce steps



Here, the output values from the shuffling phase are aggregated. It then returns a single output value

# What is MapReduce?

---

Let us now see how MapReduce works with an example

# What is MapReduce?

---

Let us now see how MapReduce works with an example

Input data

Welcome to Hadoop  
Hadoop is interesting  
Hadoop is easy

# What is MapReduce?

Let us now see how MapReduce works with an example

Input data

Input Splits

Welcome to Hadoop  
Hadoop is interesting  
Hadoop is easy

Welcome to Hadoop

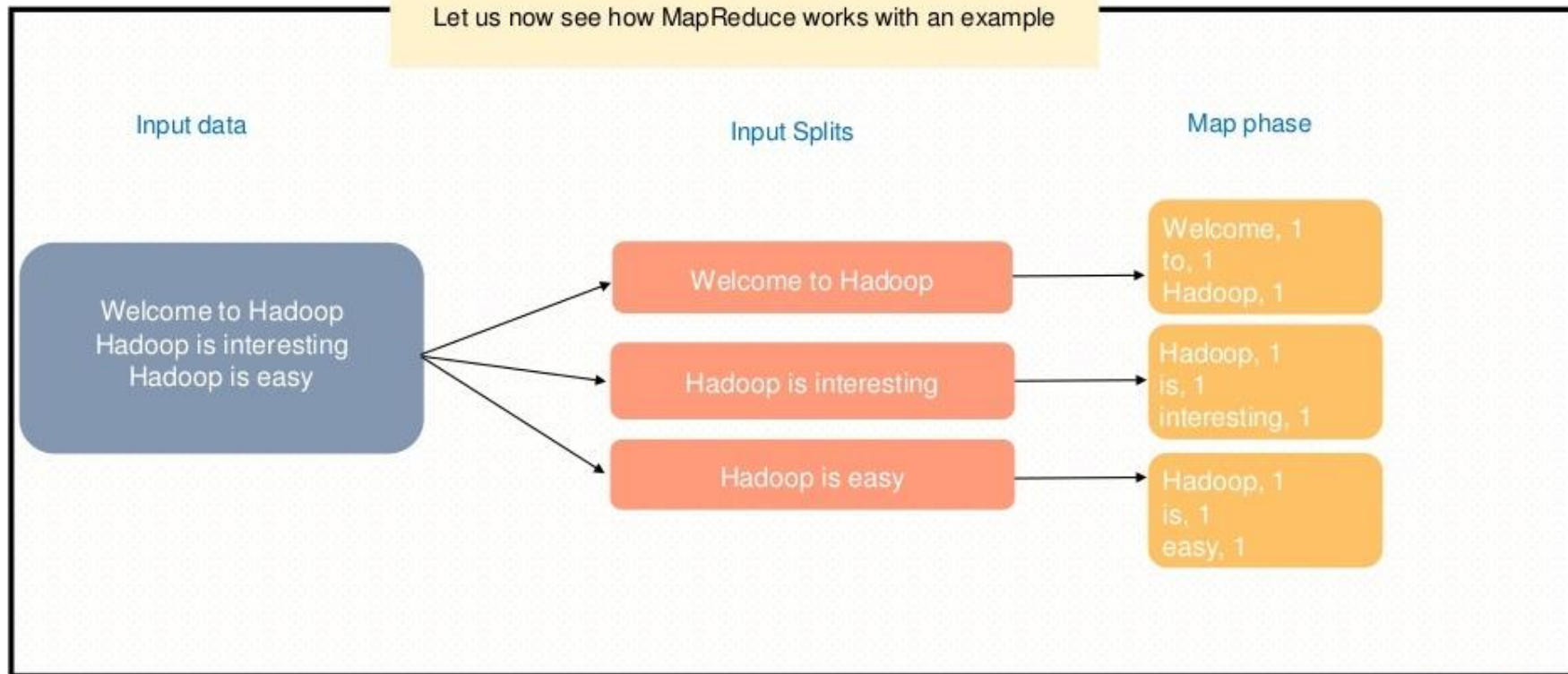
Hadoop is interesting

Hadoop is easy



# What is MapReduce?

Let us now see how MapReduce works with an example



# What is MapReduce?

Let us now see how MapReduce works with an example

Map phase

Welcome, 1  
to, 1  
Hadoop, 1

Hadoop, 1  
is, 1  
interesting, 1

Hadoop, 1  
is, 1  
easy, 1

Shuffle and Sort phase

easy, 1

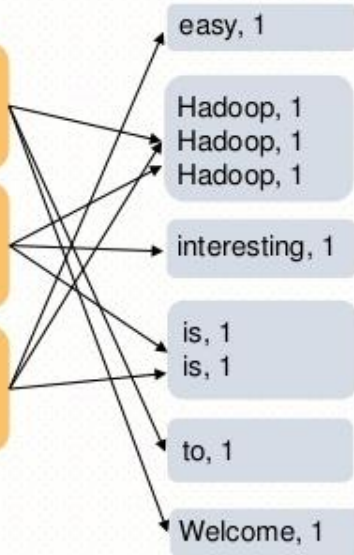
Hadoop, 1  
Hadoop, 1  
Hadoop, 1

interesting, 1

is, 1  
is, 1

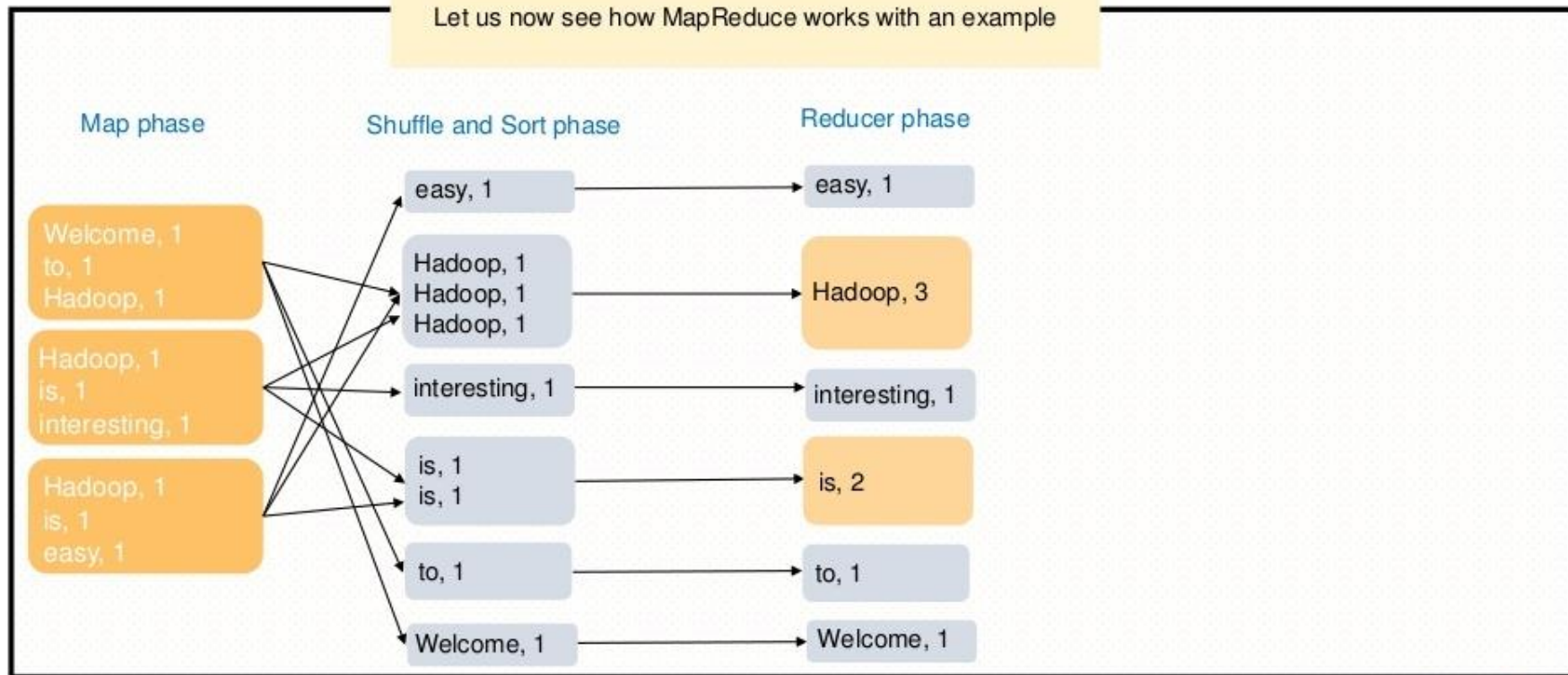
to, 1

Welcome, 1



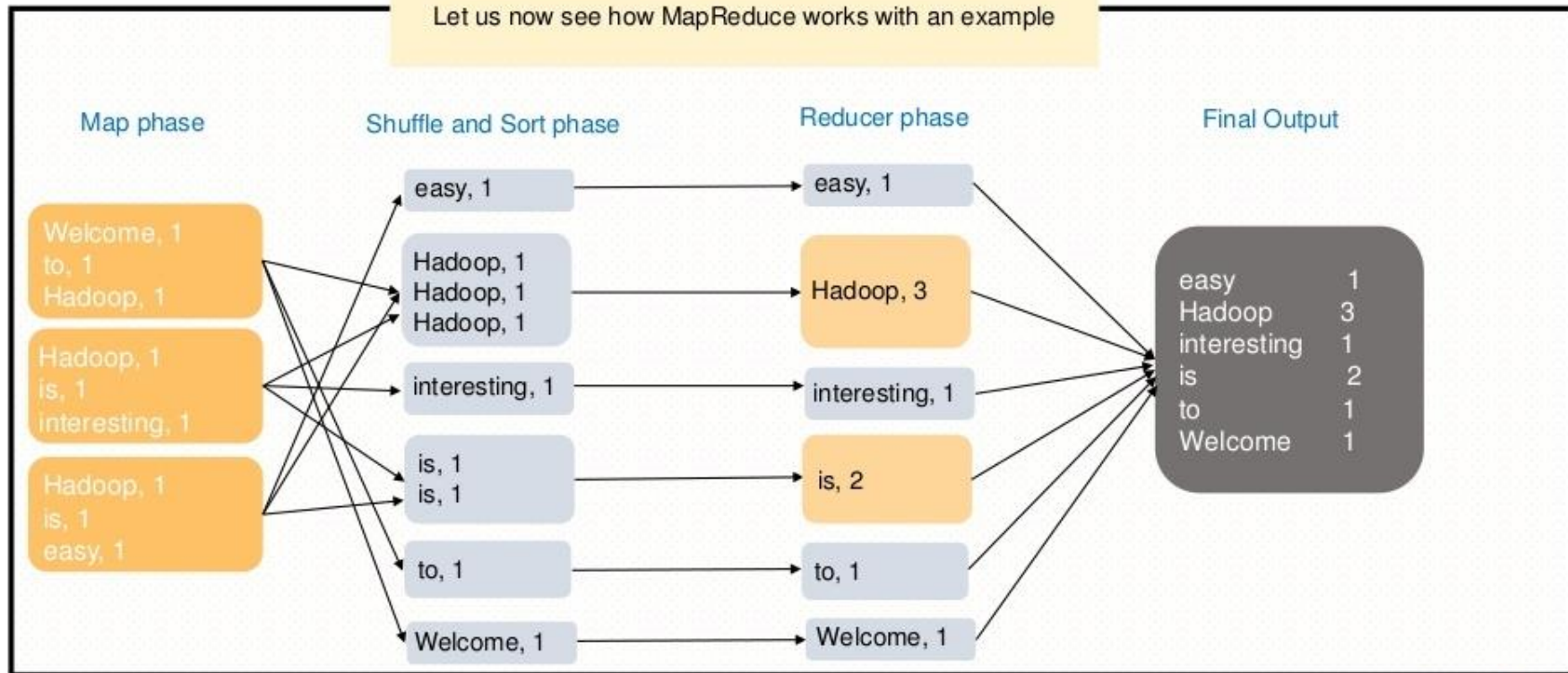
# What is MapReduce?

Let us now see how MapReduce works with an example

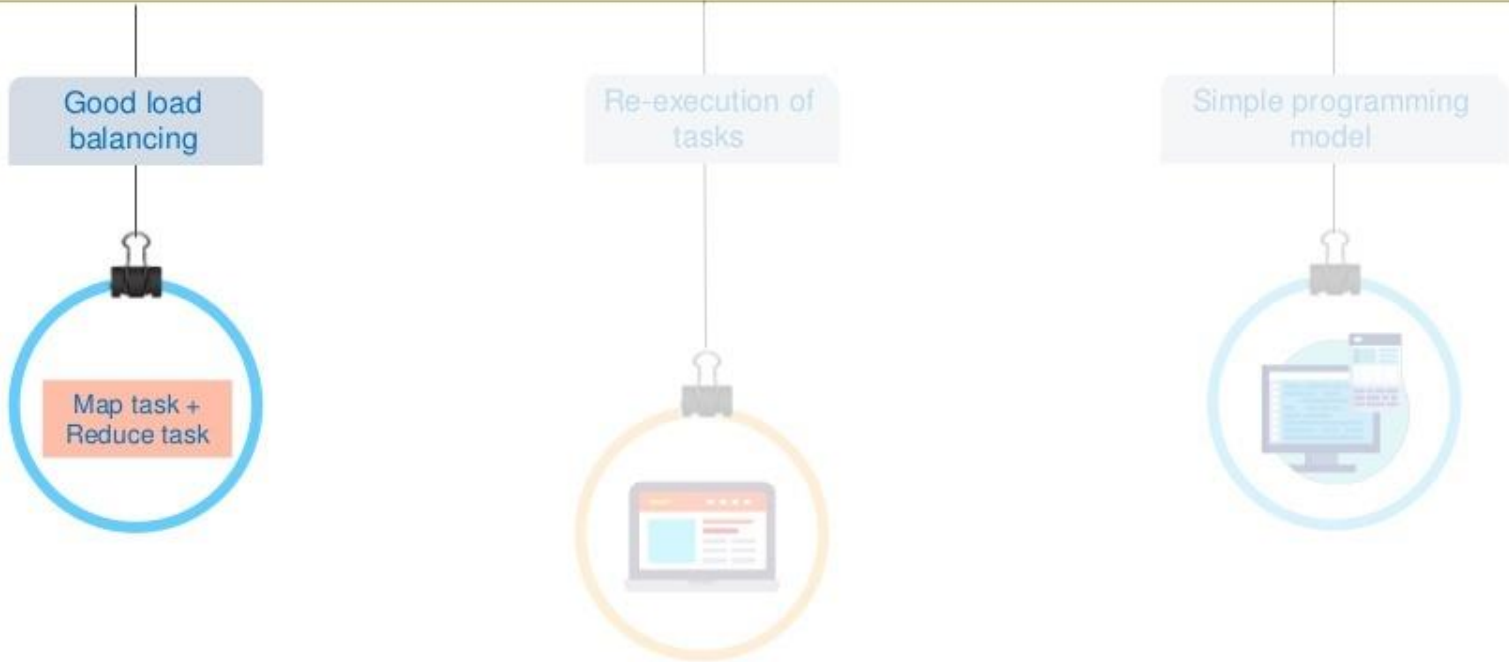


# What is MapReduce?

Let us now see how MapReduce works with an example



# Features of MapReduce



Splitting the stages into Map and Reduce tasks improves the load balancing

# Features of MapReduce

---

Good load  
balancing



Re-execution of  
tasks



Simple programming  
model



# Features of MapReduce

Good load  
balancing



Re-execution of  
tasks



Simple programming  
model



MapReduce has one of the simplest programming model which is based on Java. Java is a very common programming language

HDFS and MapReduce were the two units  
of Hadoop 1.0





Hadoop 1.0 was also known as  
MapReduce Version 1



The disadvantage with this version was that the Job tracker did both the processing of data and resource allocation



As a result, Job tracker was overburdened due to handling job scheduling, and resource management



To overcome this issue, Hadoop 2 introduced YARN as the processing layer that supported many frameworks



# Hadoop YARN



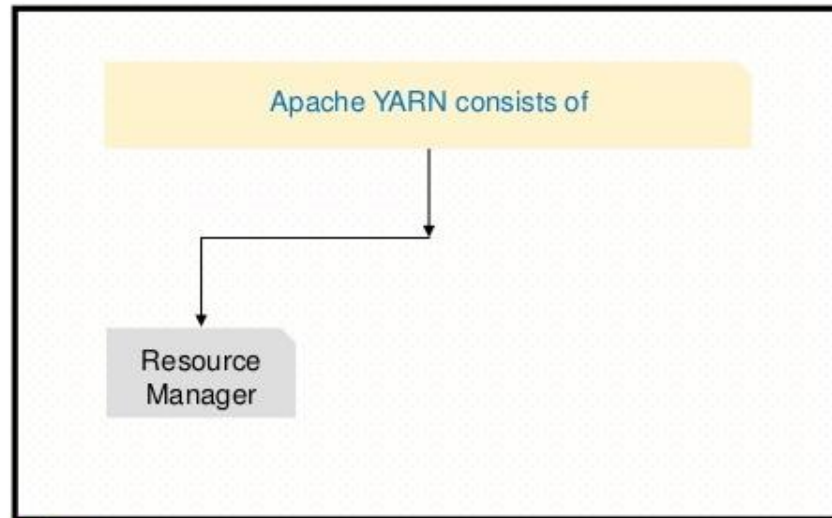
# What is YARN?

---

Yet Another Resource Negotiator (YARN) acts as the resource management unit of Hadoop

# What is YARN?

Yet Another Resource Negotiator (YARN) acts as the resource management unit of Hadoop

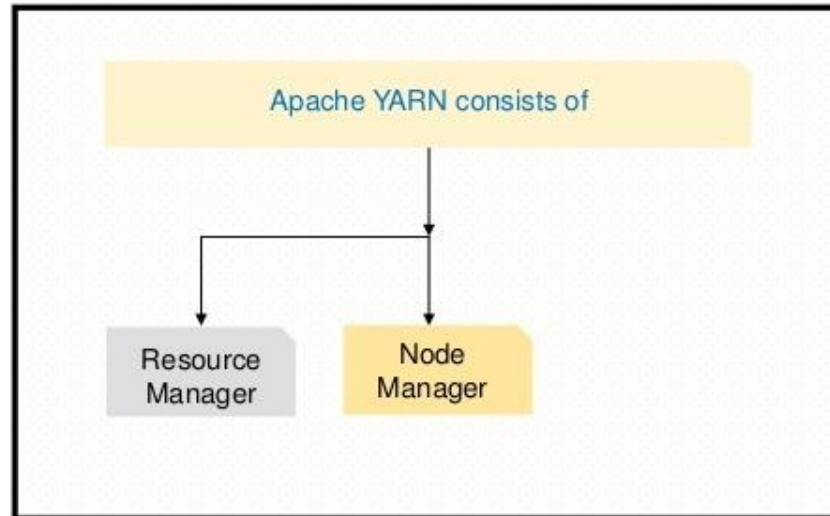


It is the master daemon. Manages the assignment of resources such as CPU, memory



# What is YARN?

Yet Another Resource Negotiator (YARN) acts as the resource management unit of Hadoop



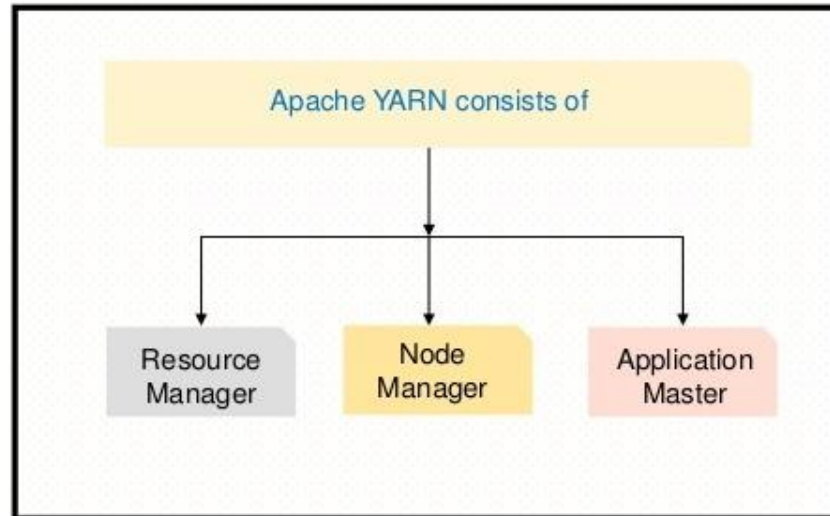
It is the slave daemon. It reports the resource usage to the Resource Manager





# What is YARN?

Yet Another Resource Negotiator (YARN) acts as the resource management unit of Hadoop



Works with the negotiation of resources from resource manager and works with node manager



# What is YARN?

---



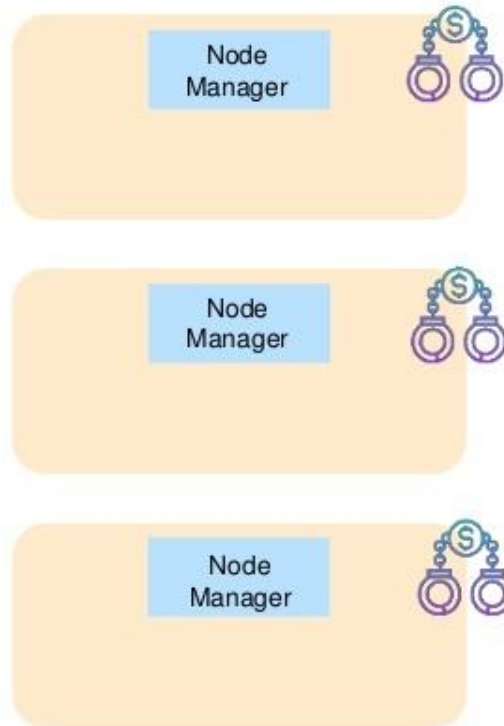
# What is YARN?

---

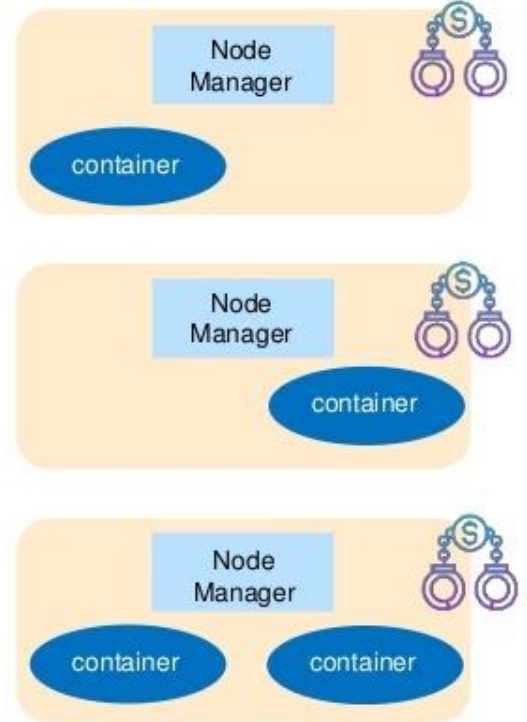


# What is YARN?

---

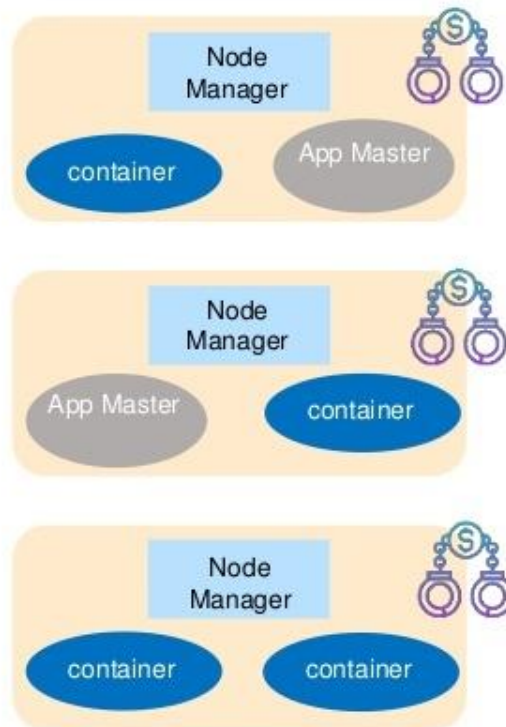


# What is YARN?



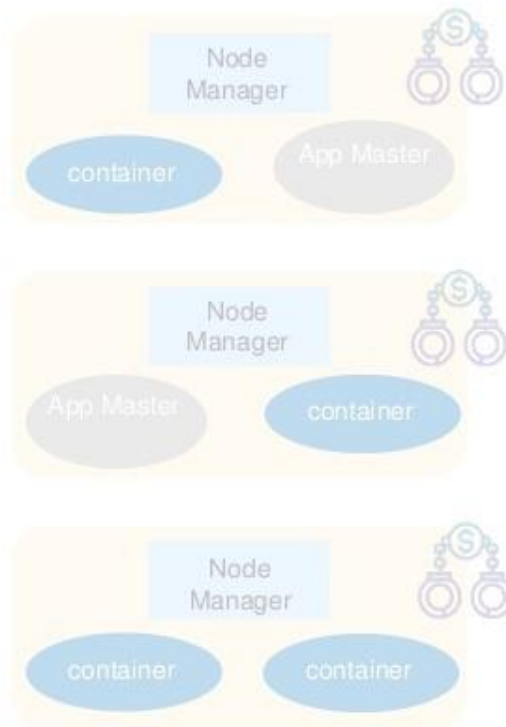
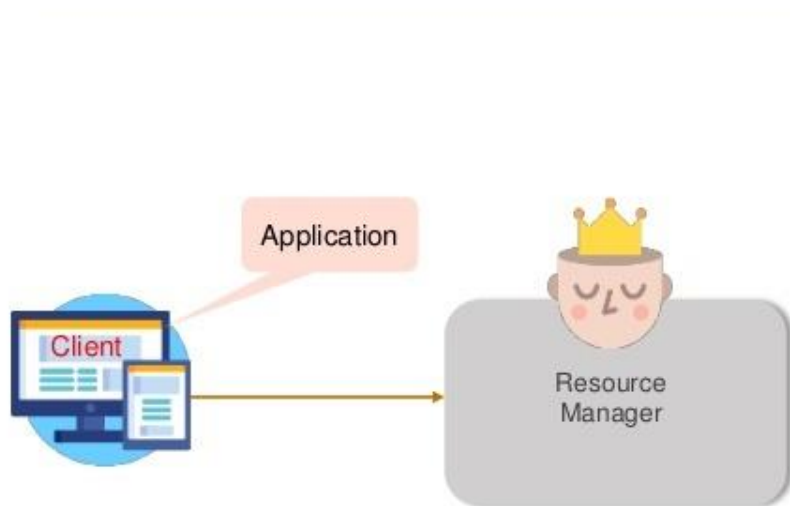
Container is a collection of physical resources such as CPU, RAM

# What is YARN?



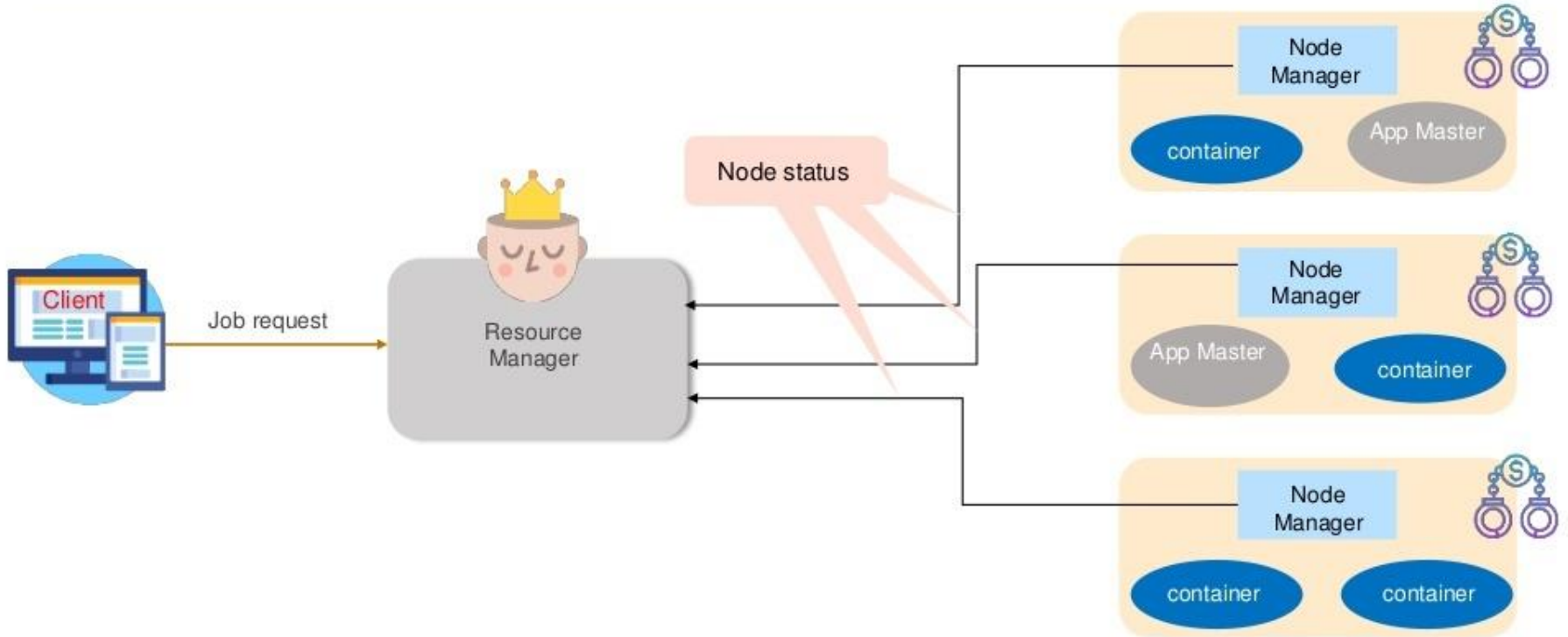
App Master requests container to Resource Manager. It uses container allocated by Node Manager

# What is YARN?



Client program sends application request to the resource manager

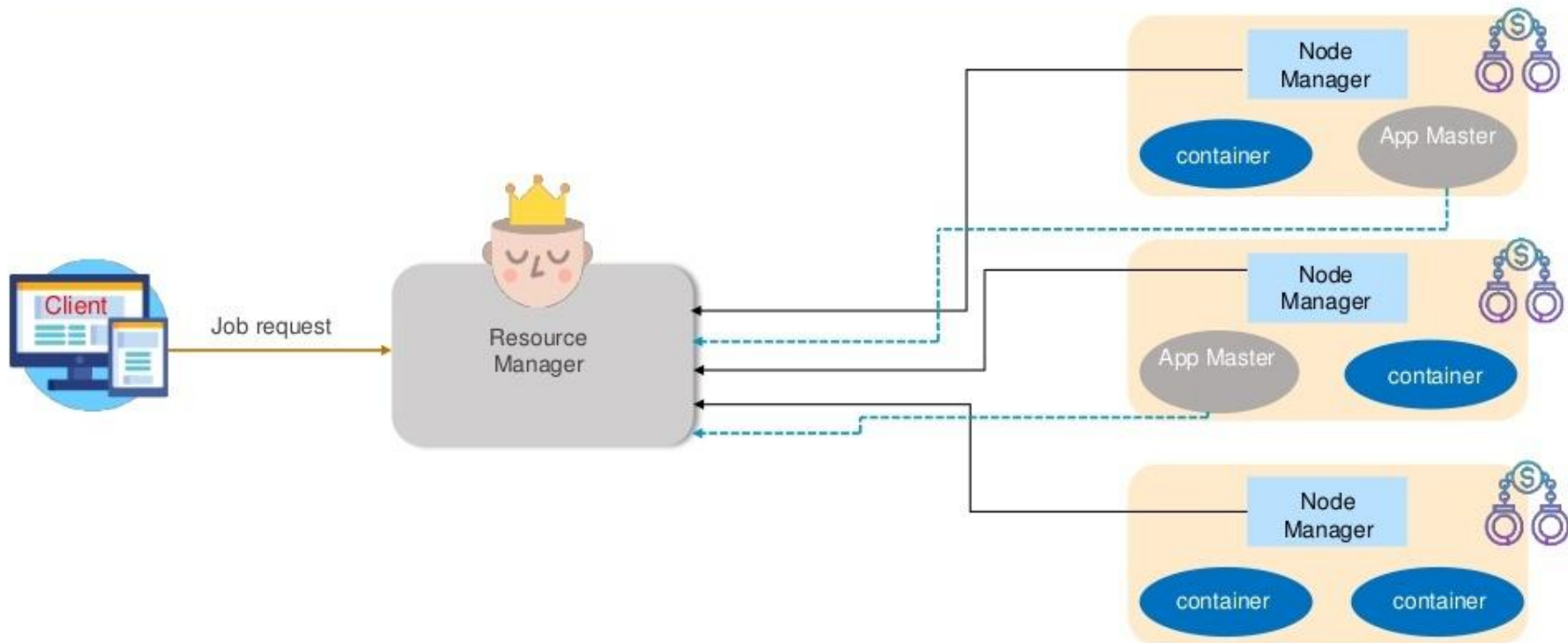
# What is YARN?



Node manager updates the status of the nodes to the resource manager

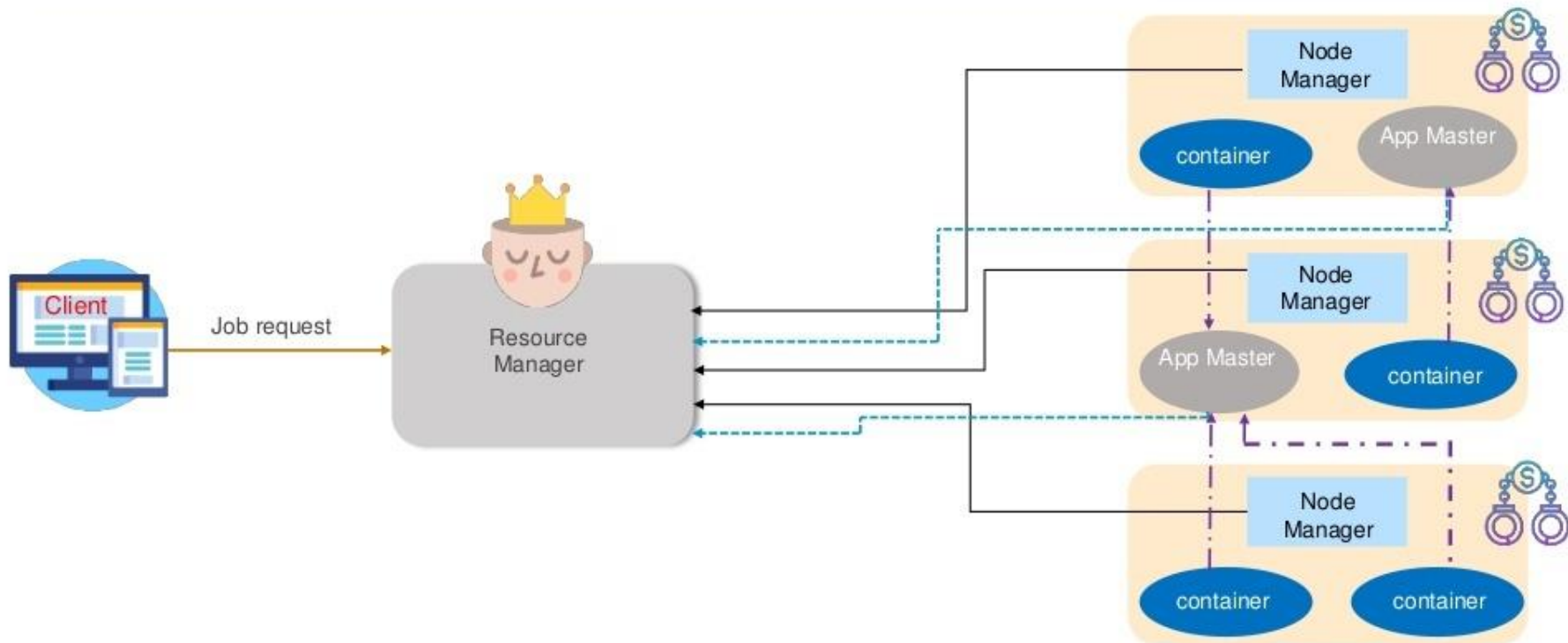


# What is YARN?



Resource Manager contacts the Node Manager requesting for resources(containers). The Node Manager grants the request

# What is YARN?



App Master contacts the Node Manager to use the container and runs in one of the container allocated on one of the nodes

# Features of YARN

Job scheduling



Multitenancy



Scalability



YARN is responsible to process job requests and allocate resources

# Features of YARN

Job scheduling



Multitenancy



Scalability



Different versions of MapReduce can run on YARN. This makes upgrading of MapReduce manageable

# Features of YARN

Job scheduling



Multitenancy



Scalability



Depending on the requirement, the number of nodes can be increased

Many companies use Hadoop for storing and processing data. Now, let me tell you about one such company



# Use Case



You would have probably heard of the popular image sharing website Pinterest







All Pins



Home

Following



2050

Concept

Design

Illustration

House

Gadgets

Aesthetic

Military

Sci Fi

Fashion

Inventions

Cities



...



These slick, geeky playing cards are all about the...

...

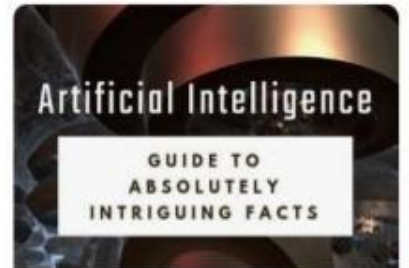


...



What Is Blockchain Technology And How Will It...

...





All Pins

Home

Following



2050

Concept

Design

Illustration

House

Budgets

Aesthetic

Military

Sci Fi

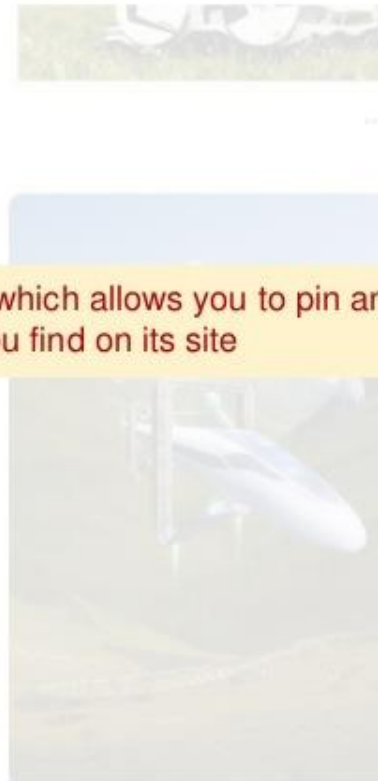
Fashion

Inventions

Circle



These slick, geeky playing cards are all about the...



What is Blockchain Technology And How Will It...



Pinterest is a social media platform which allows you to pin any interesting information you find on its site



All Pins



Home

Following



2050

Concept

Design

Illustration

House

Budgets

Aesthetic

Military

Sci Fi

Fashion

Inventions

Circle

Pinterest has more than 250 million users and nearly 30 billion pins. All these account to big data concerning Pinterest



These slick, geeky playing cards are all about the...



# How Will it Change the World

What is Blockchain Technology  
And How Will It...



## Artificial Intelligence

GUIDE TO  
ABSOLUTELY  
INTRIGUING FACTS





All Pins

Home

Following



2050

Concept

Design

Illustration

House

Budgets

Aesthetic

Military

Sci Fi

Fashion

Inventions

Circuits

Pinterest has more than 250 million users and nearly 30 billion pins. All these account to big data concerning Pinterest



These slick, geeky playing cards are all about the...



Problem

Pinterest faced a challenge in processing tremendous amount of data

What is Blockchain Technology  
How Will It...





All Pins

Home

Following

2050

Concept

Design

Illustration

House

Budgets

Aesthetic

Military

Sci Fi

Fashion

Inventions

Class

Pinterest has more than 250 million users and nearly 30 billion pins. All these account to big data concerning Pinterest



These slick, geeky playing cards are all about the...

### Problem

Pinterest faced a challenge in processing tremendous amount of data

There was a difficulty in analyzing which data needs to be displayed in a user's personalized discovery engine



What is Blockchain Technology  
How Will It...





All Pins

Home

Following

2050

Concept

Design

Illustration

House

Budgets

Aesthetic

Military

Sci Fi

Fashion

Inventions

Class

Pinterest has more than 250 million users and nearly 30 billion pins. All these account to big data concerning Pinterest



These slick, geeky playing cards are all about the...



How Will it  
Change the  
World

What is Blockchain Technology  
How Will It...

Solution



Artificial Intelligence

GUIDE TO  
ABSOLUTELY  
INTRIGUING FACTS



All Pins

Home

Following



2050

Concept

Design

Illustration

House

Budgets

Aesthetic

Military

Sci Fi

Fashion

Inventions

Cin...

Pinterest has more than 250 million users and nearly 30 billion pins. All these account to big data concerning Pinterest



These slick, geeky playing cards are all about the...



How Will it  
Change the  
World

What is Blockchain Technology  
How Will It...



Solution

Pinterest uses Hadoop to process and analyze big data in a way that it helps the company to show the most relevant content to its users

Artificial Intelligence

GUIDE TO  
ABSOLUTELY  
INTRIGUING FACTS





All Pins

Home

Following

2050

Concept

Design

Illustration

House

Budgets

Aesthetic

Military

Sci Fi

Fashion

Inventions

Class

Pinterest has more than 250 million users and nearly 30 billion pins. All these account to big data concerning Pinterest



### Solution

Pinterest uses Hadoop to process and analyze big data in a way that it helps the company to show the most relevant content to its users

Through continuous analysis of the data, Pinterest can provide its users with features such as related pins, guided search and so on



These slick, geeky playing cards are all about the...



How Will it  
Change the  
World

What is Blockchain Technology  
How Will It...

Artificial Intelligence

GUIDE TO  
ABSOLUTELY  
INTRIGUING FACTS



This is how Pinterest benefited from Hadoop. Let's also start using Hadoop to put an end to the big data challenges we are facing



# Key Takeaways

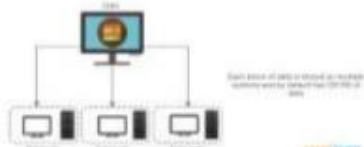
## What is Hadoop?

Hadoop is a framework which stores and processes big data in a distributed and parallel manner.

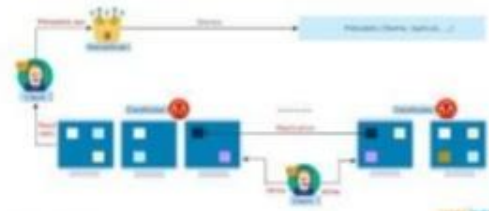


## What is HDFS?

Hadoop Distributed File System (HDFS) is used for distributed storage method. It splits files into data blocks and replicates it across in two replication of those data is done to be saved on or off.



## Architecture of HDFS

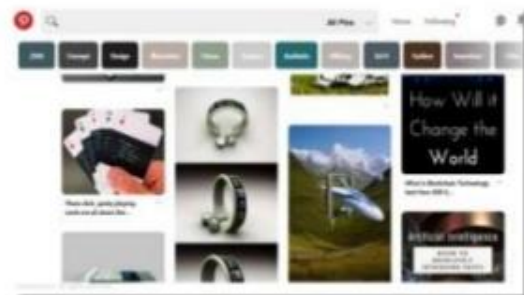
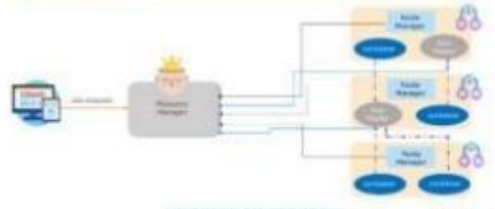


## What is MapReduce?

All to know and how MapReduce works with an example.



## What is YARN?



# Let's do some hands-on

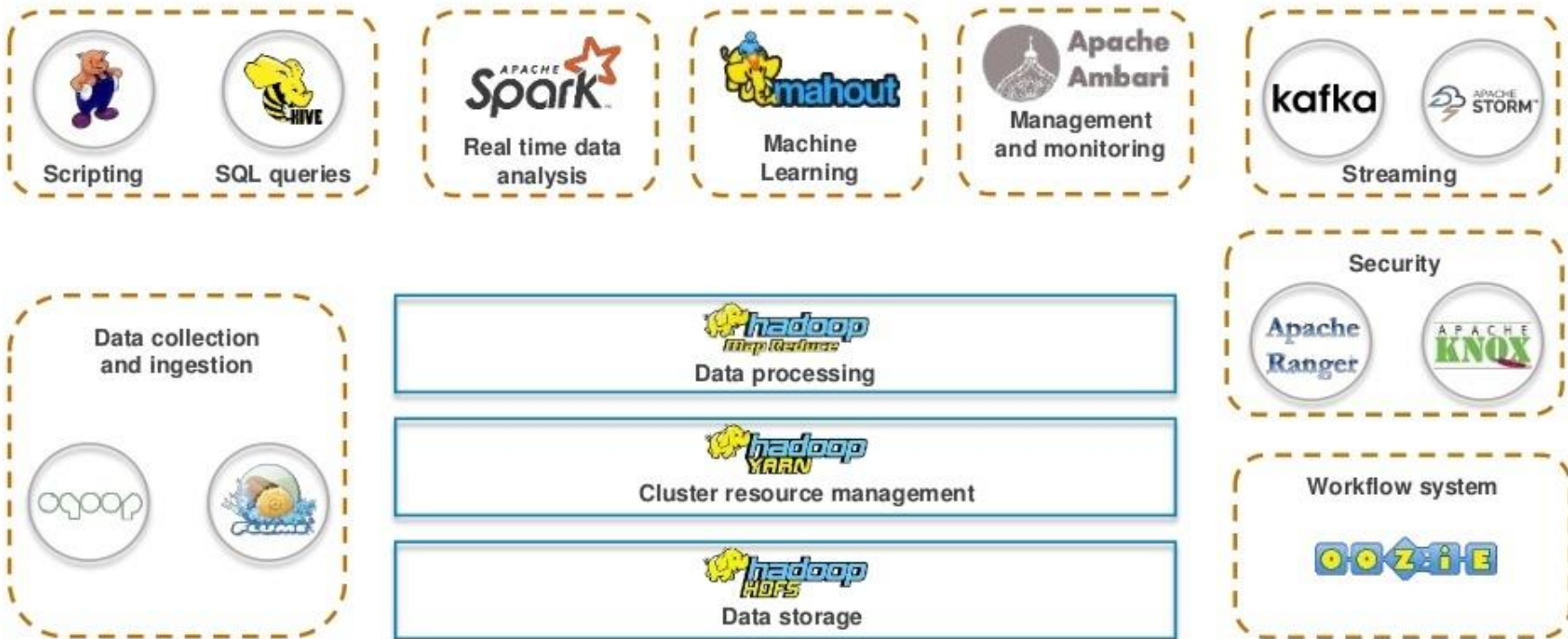
Installation / Working with HDFS / Working with MapReduce



# Hadoop Ecosystem



# Hadoop Ecosystem

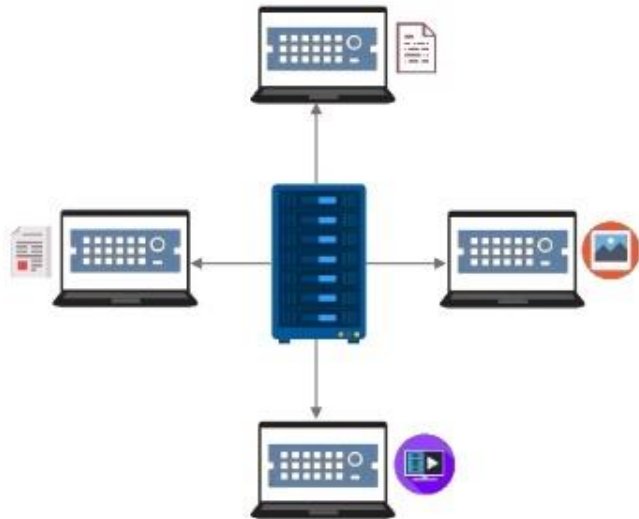


# Hadoop Ecosystem

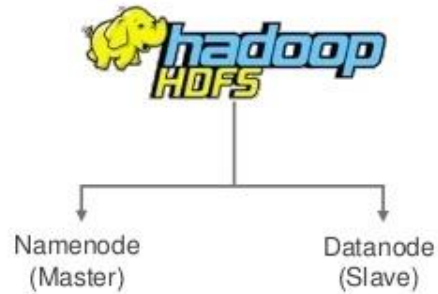


# HDFS

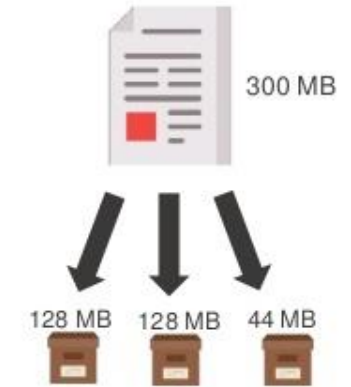
HDFS stands for Hadoop Distributed File System



Stores different formats of data on various machines



2 major components



Splits the data into multiple blocks (128 MB by default)

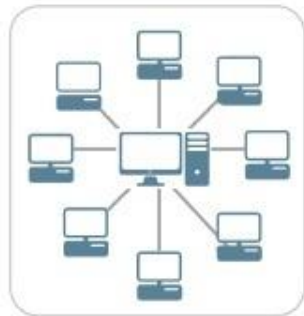
# Hadoop Ecosystem



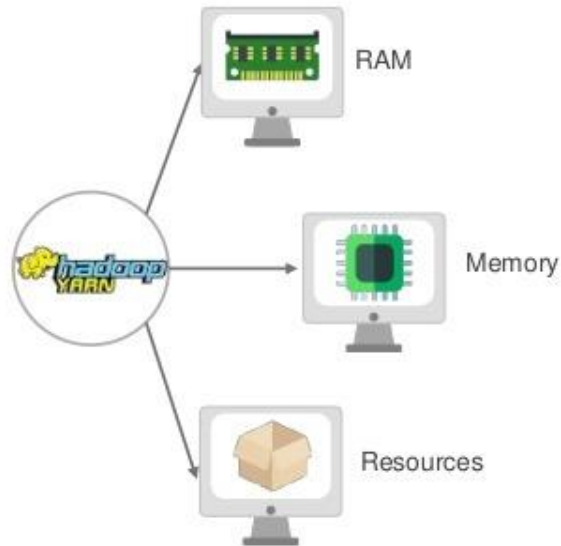


# YARN

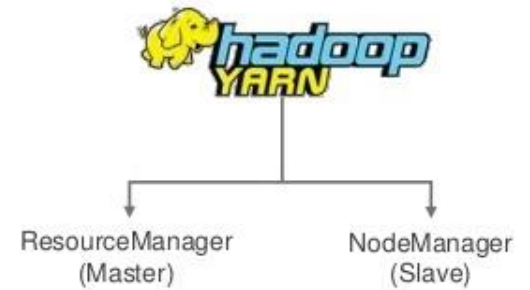
YARN stands for Yet Another Resource Negotiator



Handles the cluster of nodes



Allocates RAM, memory and other resources to different applications



2 major components



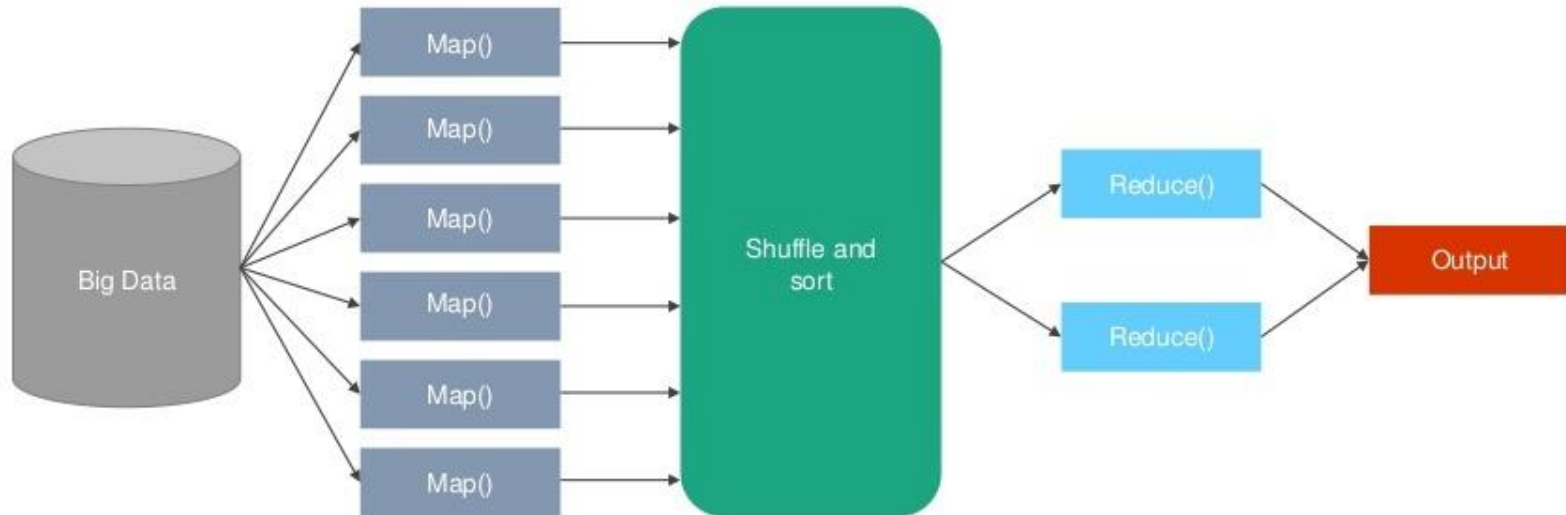
# Hadoop Ecosystem



# MapReduce



MapReduce processes large volumes of data in a parallelly distributed manner



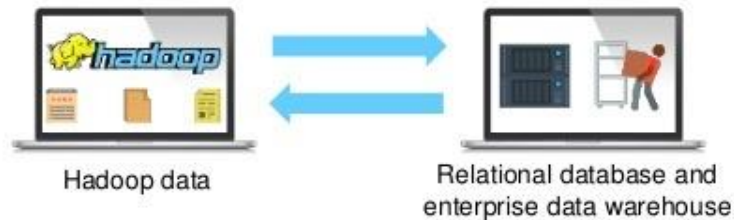
# Hadoop Ecosystem



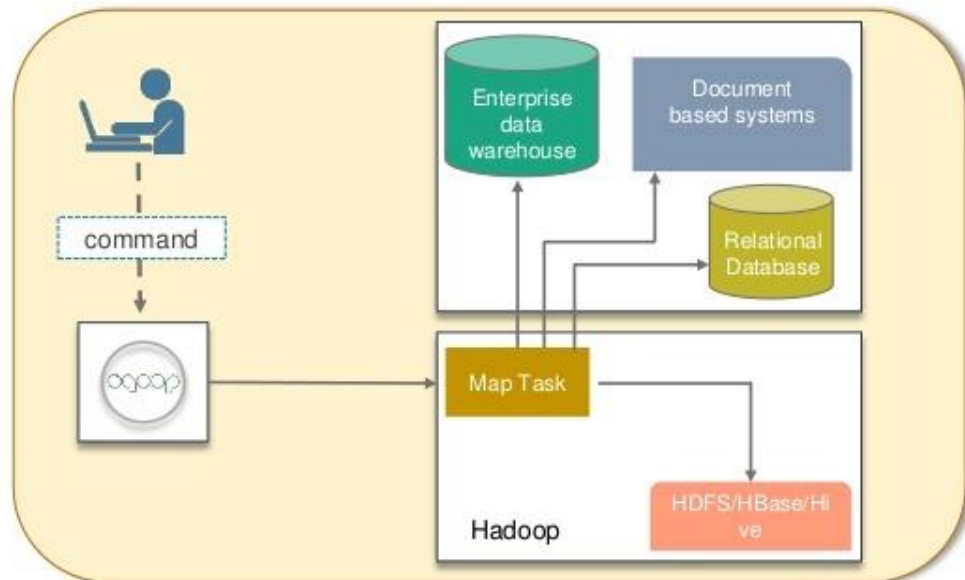
# Sqoop



Sqoop is used to transfer data between Hadoop and external datastores such as relational databases and enterprise data warehouses



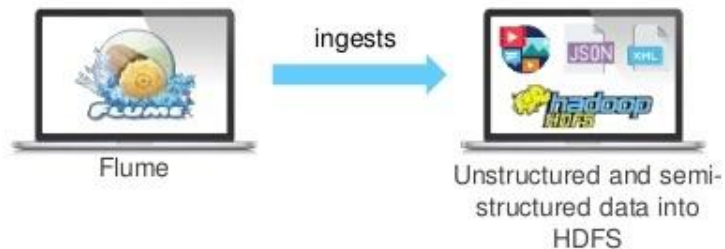
It imports data from external datastores into HDFS, Hive and HBase



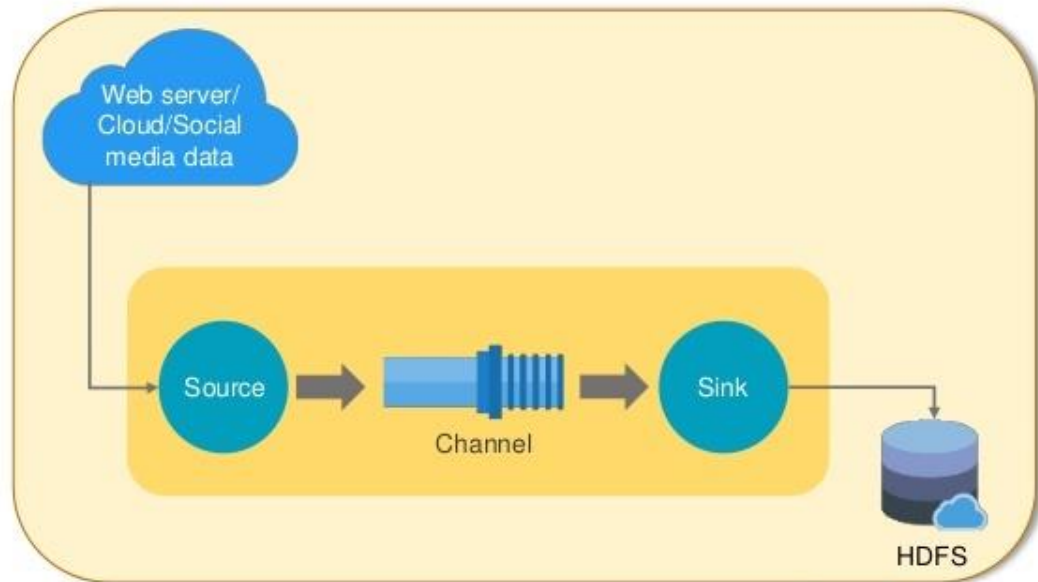
# Flume



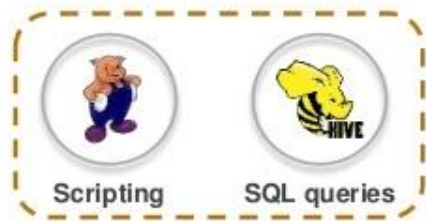
Flume is distributed service for collecting, aggregating and moving large amounts of log data



Ingests online streaming data from social media, log files, web server into HDFS



# Hadoop Ecosystem



# Pig



Pig is used to analyze data in Hadoop. It provides a high level data processing language to perform numerous operations on the data



Pig Latin

Language for scripting

Pig Latin Compiler

Converts Pig Latin code to executable code

ETL

Provides a platform for building data flow for ETL



10 lines of Pig Latin script is around 200 lines of MapReduce job

Pig Latin Scripts

Grunt Shell

Pig Server

Parser

Optimizer

Compiler

Execution Engine

Apache Pig

MapReduce

HDFS





# Hive



Hive facilitates reading, writing and managing large datasets residing in the distributed storage using SQL (Hive Query Language)

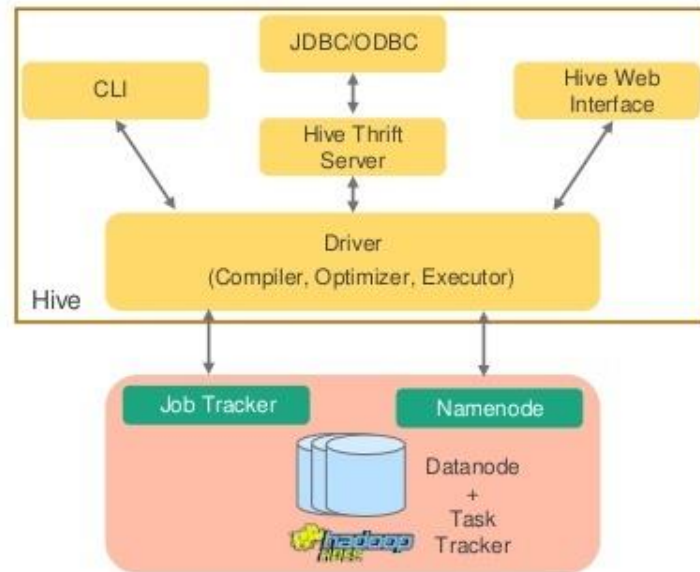


Hive Command Line

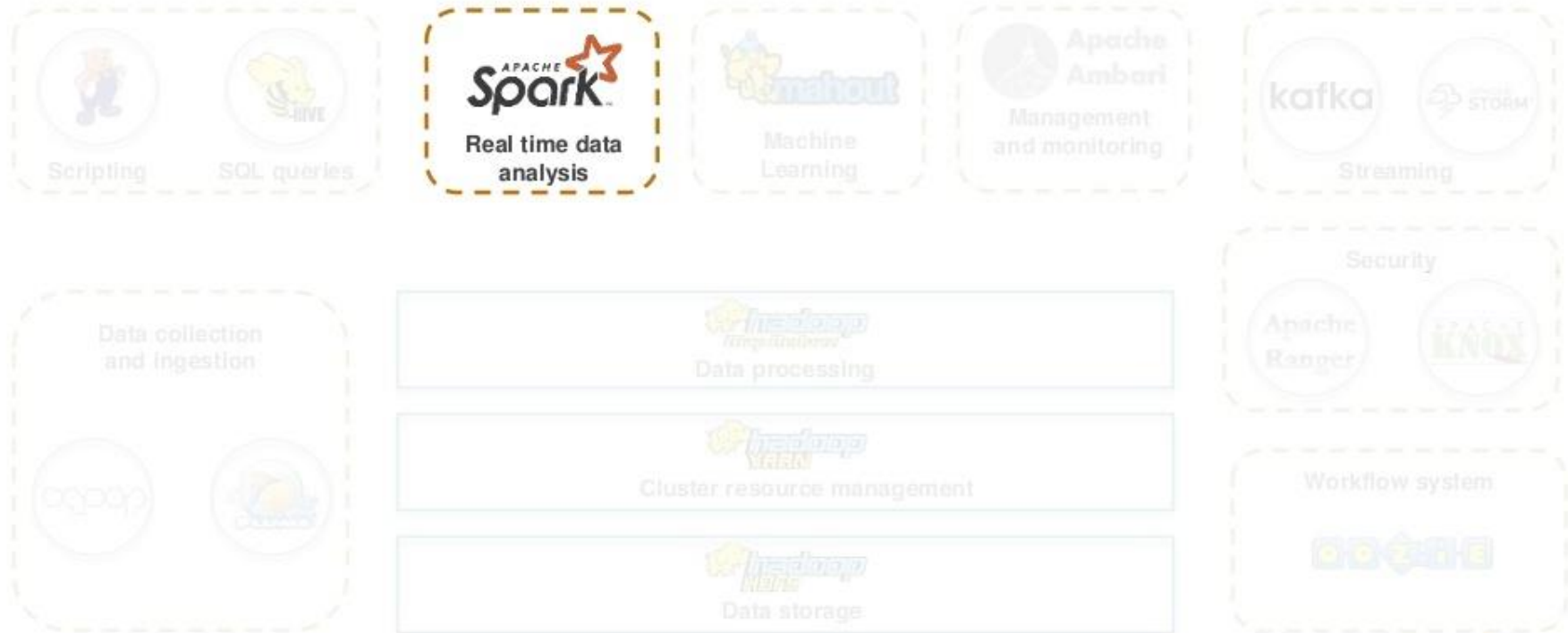
JDBC/ODBC driver

2 major components

Provides User Defined Functions (UDF) for data mining, document indexing, log processing, etc.



# Hadoop Ecosystem



# Spark



Spark is an open-source distributed computing engine for processing and analyzing huge volumes of real time data

Written in  
**Scala**



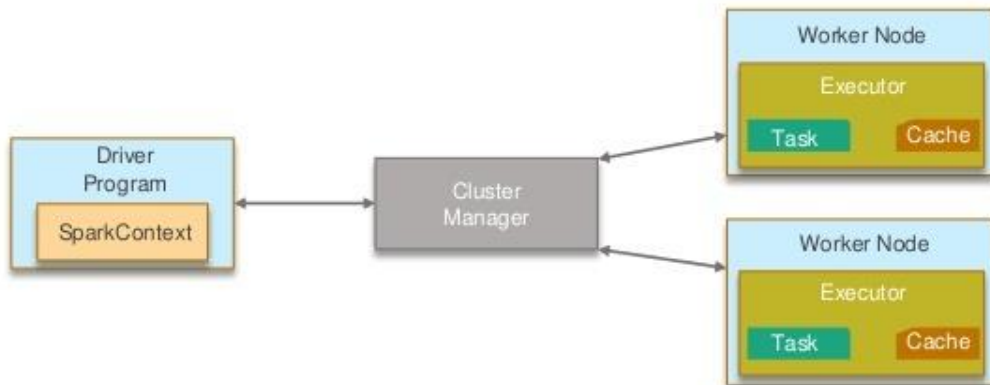
Runs 100x times faster than MapReduce



Provides in-memory computation of data



Used to process and analyze real time streaming data such as stock market and banking data



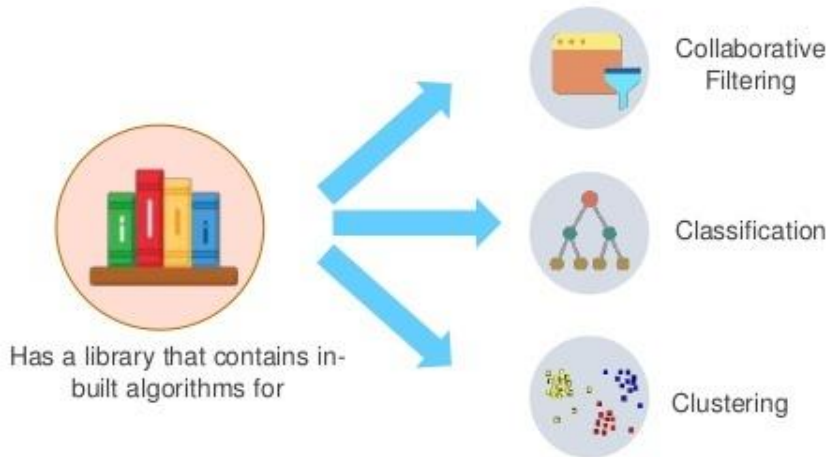
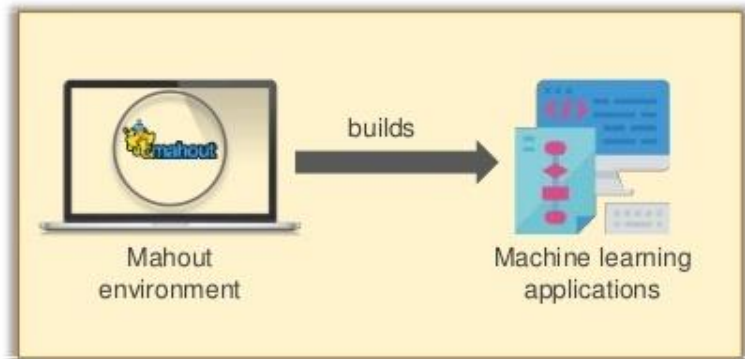
# Hadoop Ecosystem



# Mahout



Mahout is used to create scalable and distributed machine learning algorithms



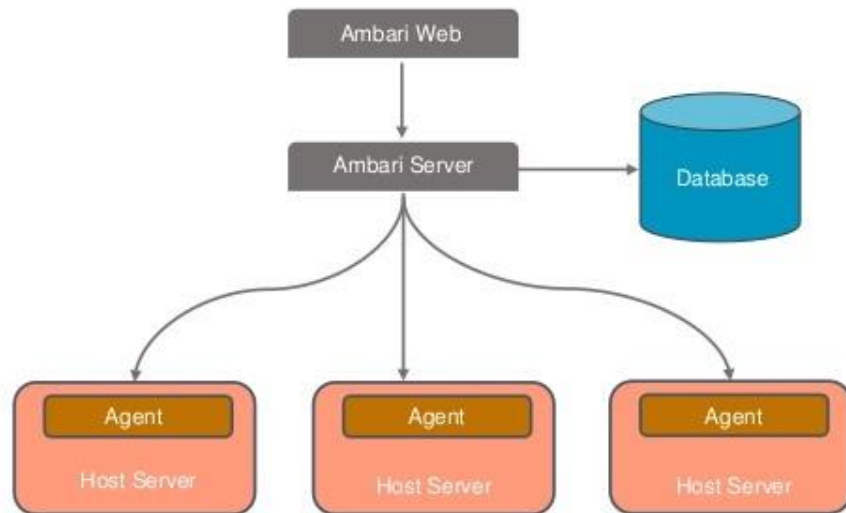
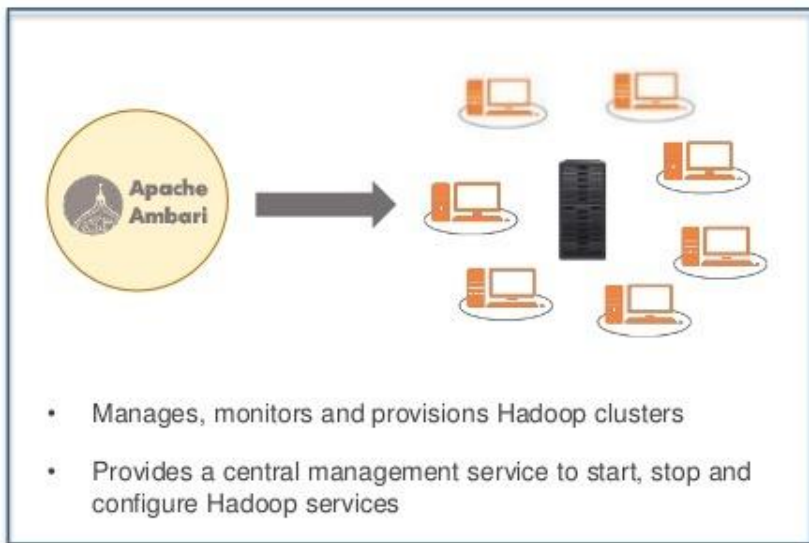
# Hadoop Ecosystem



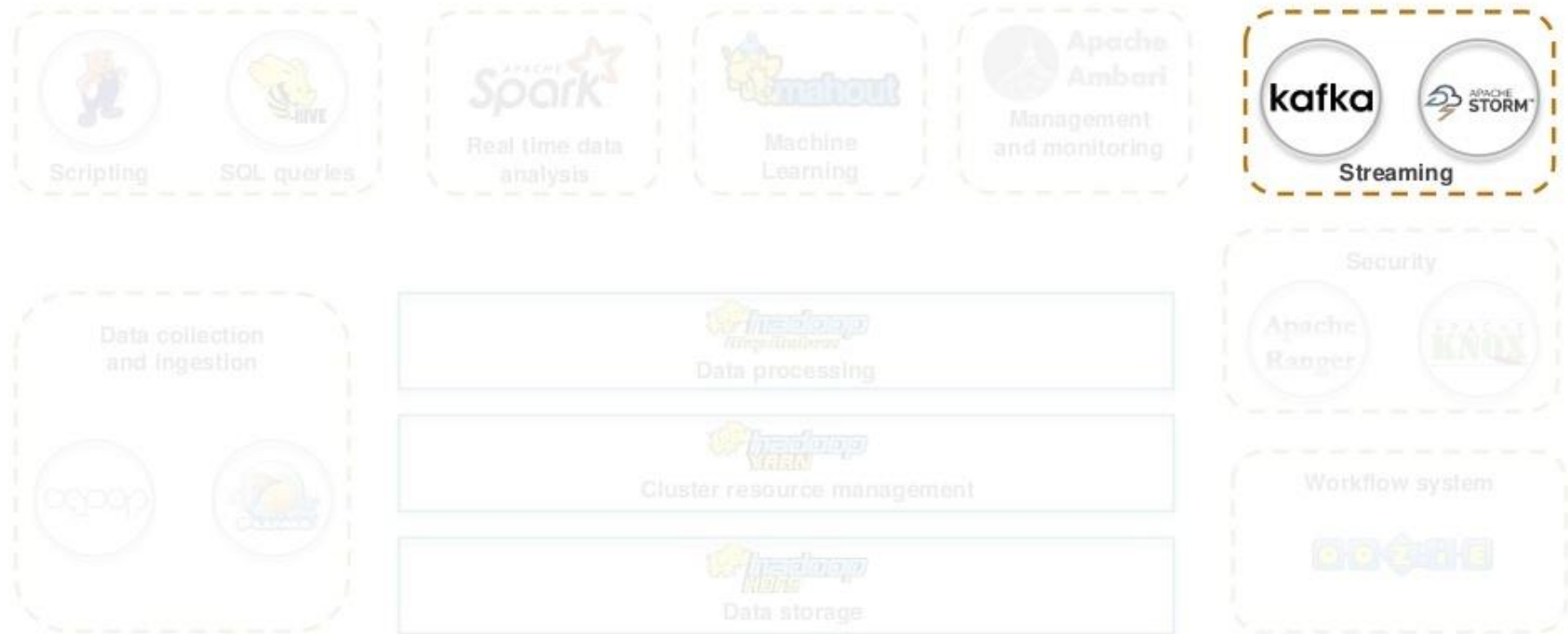
# Ambari



Ambari is an open-source tool responsible for keeping track of running applications and their statuses



# Hadoop Ecosystem



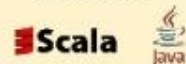


# Kafka



Kafka is a distributed streaming platform to store and process streams of records

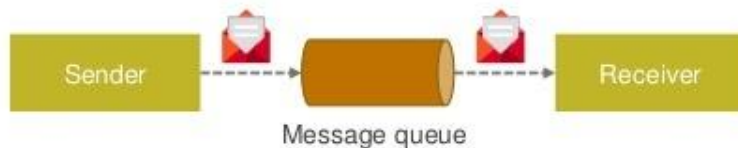
Written in



Builds real-time streaming data pipelines that reliably get data between applications

Builds real-time streaming applications that transforms data into streams

Kafka uses a messaging system for transferring data from one application to another



# Storm



Storm is a processing engine that processes real-time streaming data at a very high speed

Written in

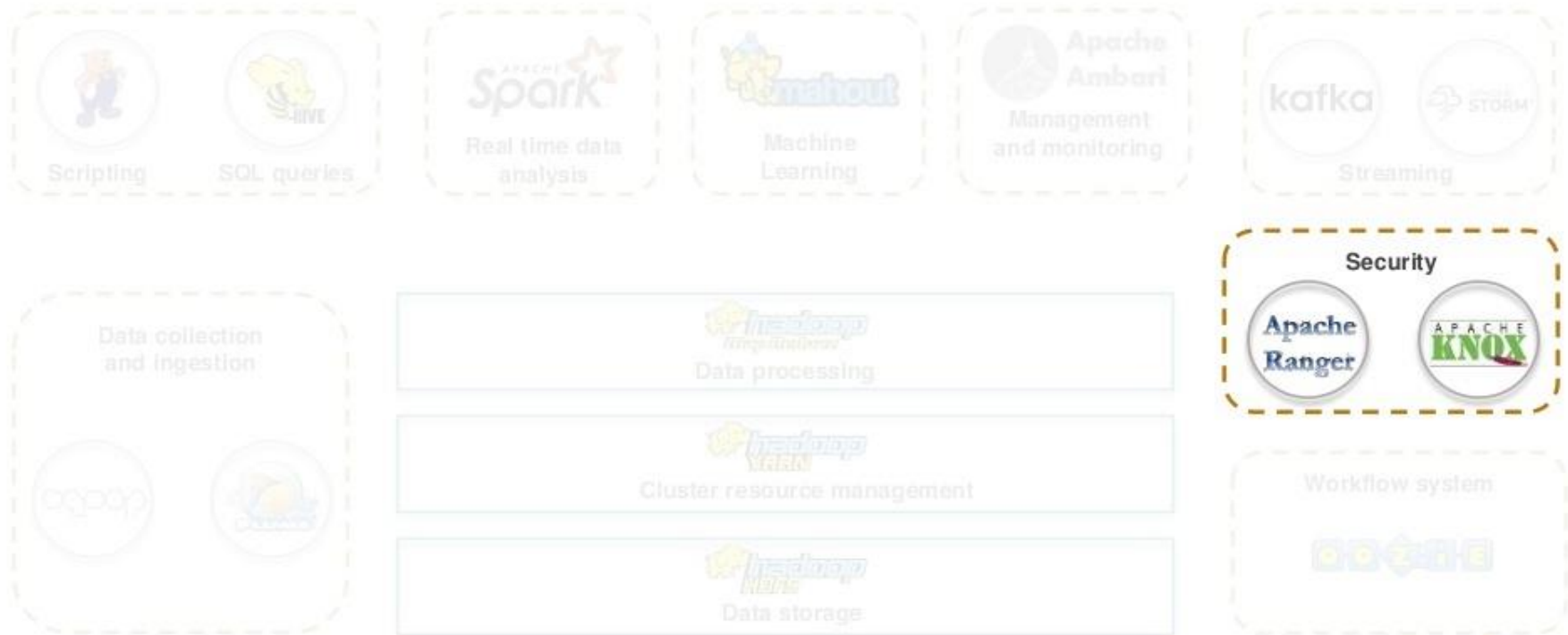


Ability to process over a million jobs in a fraction of seconds on a node



It is integrated with Hadoop to harness higher throughputs

# Hadoop Ecosystem



# Ranger



Ranger is a framework to enable, monitor and manage data securities across the Hadoop platform

1

Provides centralized security administration to manage all security related tasks



2

Standardize authorization across all Hadoop components



3

Enhanced support for different authorization methods – Role based access control, attribute based access control, etc.



# Knox



Knox is an application gateway for interacting with the REST APIs and UIs of Hadoop deployments

Knox delivers 3 groups of user facing services:

1

Proxying Services

Provides access to Hadoop via proxying the HTTP request

http://

2

Authentication Services

Authentication for REST API access and WebSSO flow for user interfaces

{ REST }

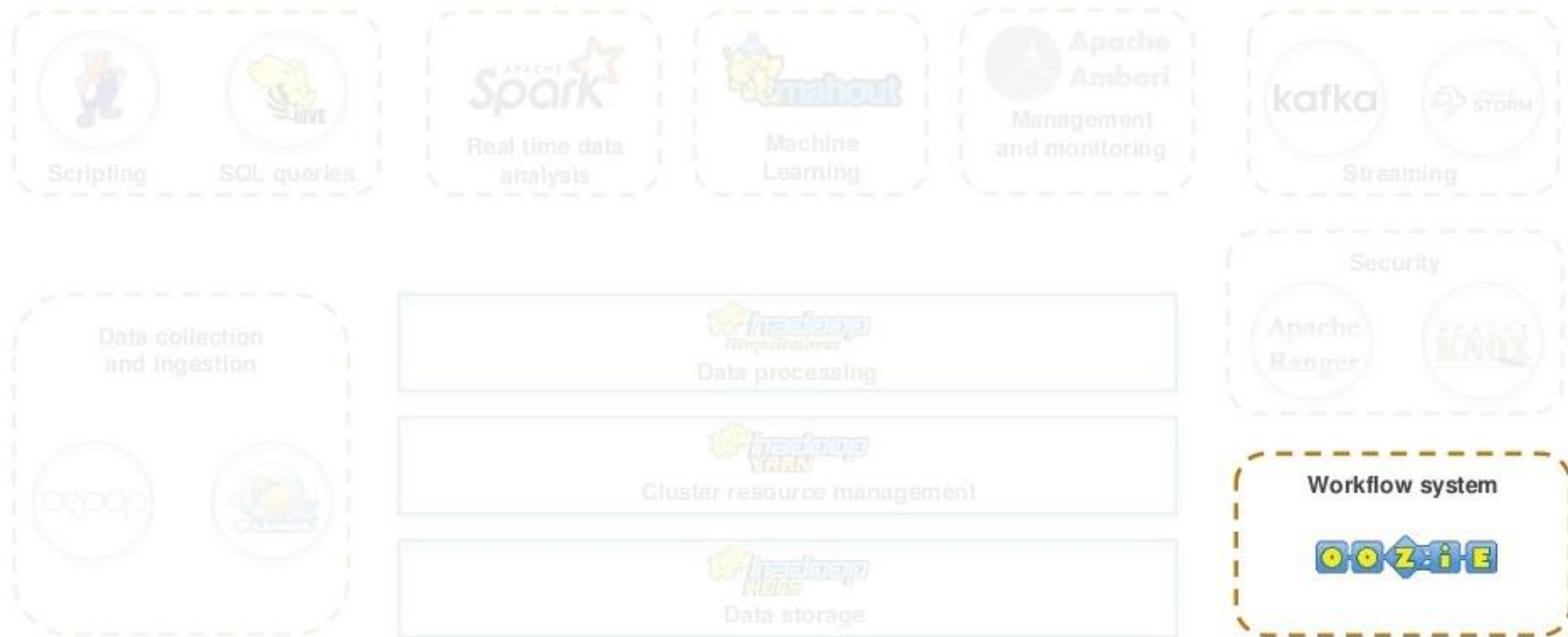
3

Client Services

Client development can be done with the scripting through DSL or using the Knox shell classes



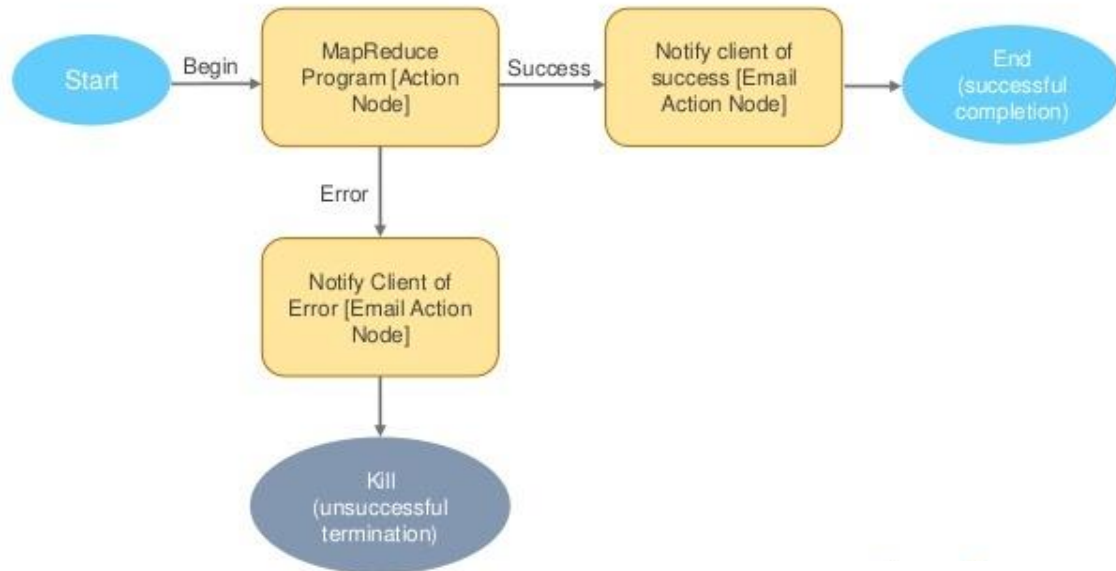
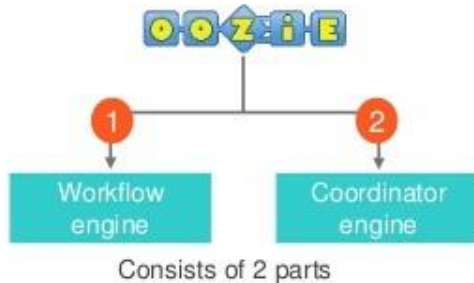
# Hadoop Ecosystem



# Oozie



Oozie is a workflow scheduler system to manage Hadoop jobs



1

Directed Acyclic Graphs (DAGs) which specifies a sequence of actions to be executed

2

These consist of workflow jobs triggered by time and data availability

Hadoop Ecosystem comprises of the following 12 components:



Hadoop HDFS



SQOOP



Apache Spark



Pig



Hadoop Hive



Oozie



HBase



Flume



Hadoop MapReduce



Impala



Cloudera Search



Hue