AI VIETNAM
All-in-One Course

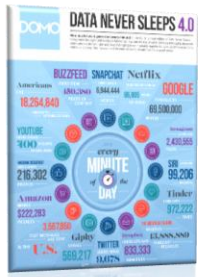
Big Data

From Zero To Mastery

Vinh Dinh Nguyen
PhD in Computer Science

Big Data
From Zero To Mastery

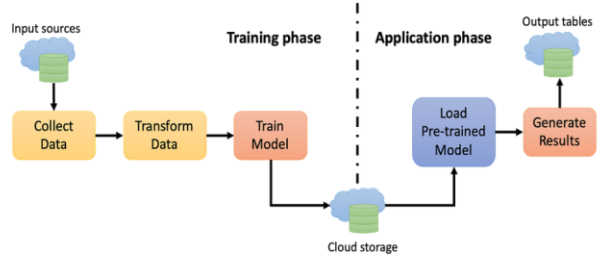
Big Data
(Zero To Mastery: Big Data Analytics Tools)



Vinh Dinh Nguyen
PhD in Computer Science

Year 2023

Big Data
(Zero To Mastery: ML Models Using PySpark in Cloud Platform)



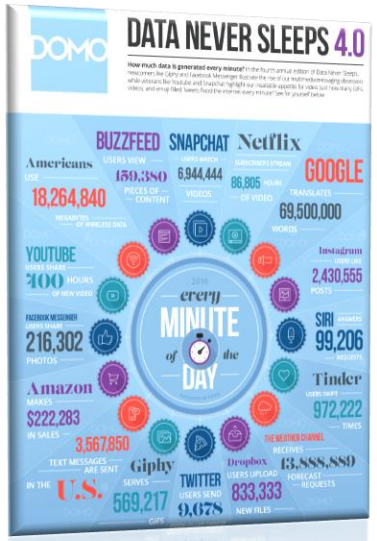
Vinh Dinh Nguyen
PhD in Computer Science

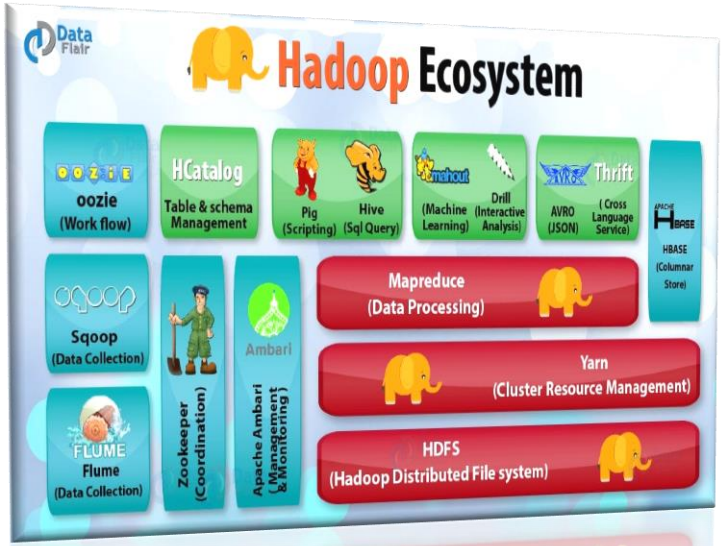
Year 2023

AI VIETNAM
All-in-One Course

Big Data 1

(Zero To Mastery: Big Data Analytics Tools)





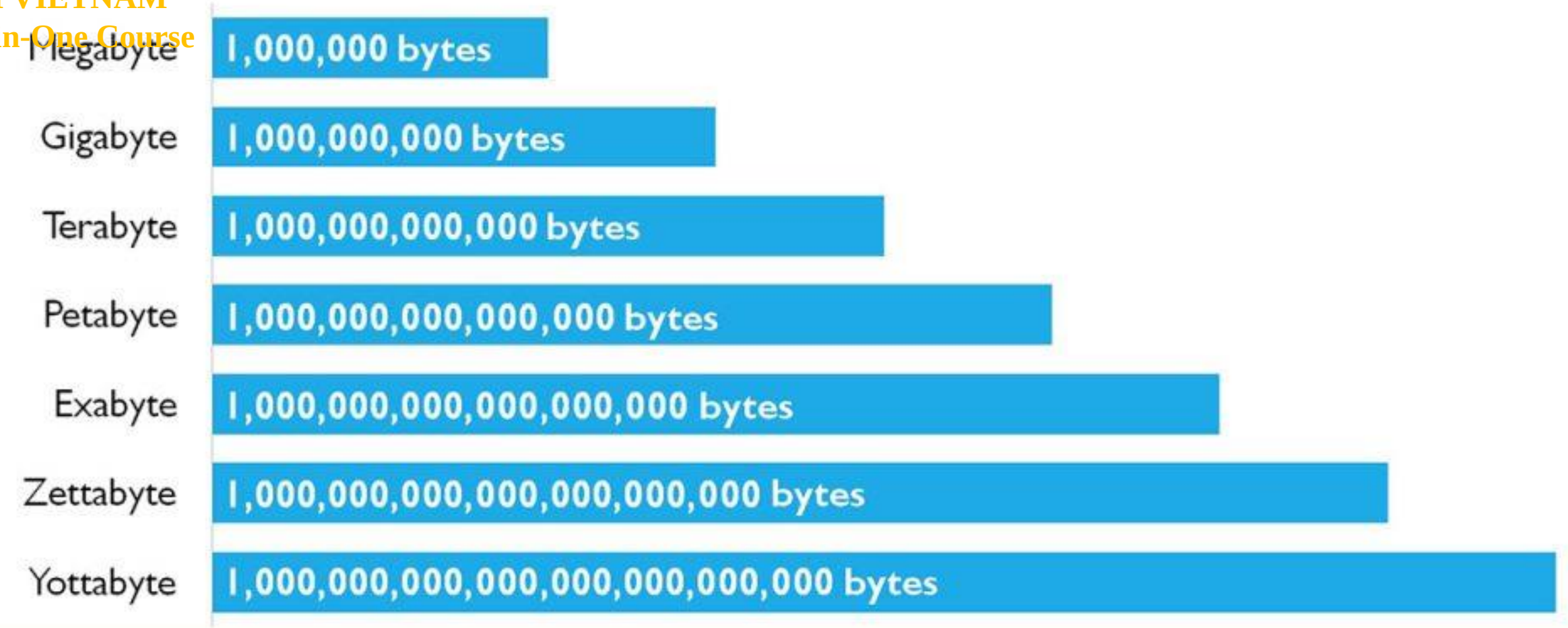
Vinh Dinh Nguyen
PhD in Computer Science

Year 2023

40 Exabytes a month for single user

6.92 billion, 2023

How big are they?



HOW BIG ARE THEY?

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

IDC

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

95m

photos and videos are shared on Instagram

Instagram Business

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

294bn

billion emails are sent

Radicati Group

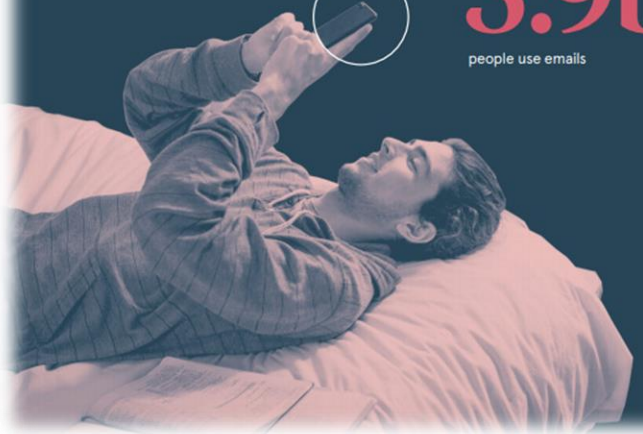
3.9bn

people use emails

4TB

of data produced by a connected car

Intel



Searches made a day — 5bn

Searches made a day from Google — 3.5bn

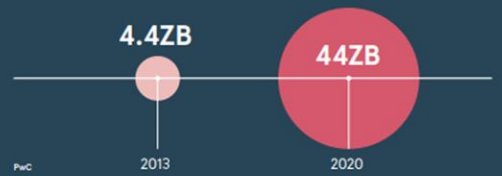
Smart insights

28PB

to be generated from wearable devices by 2020

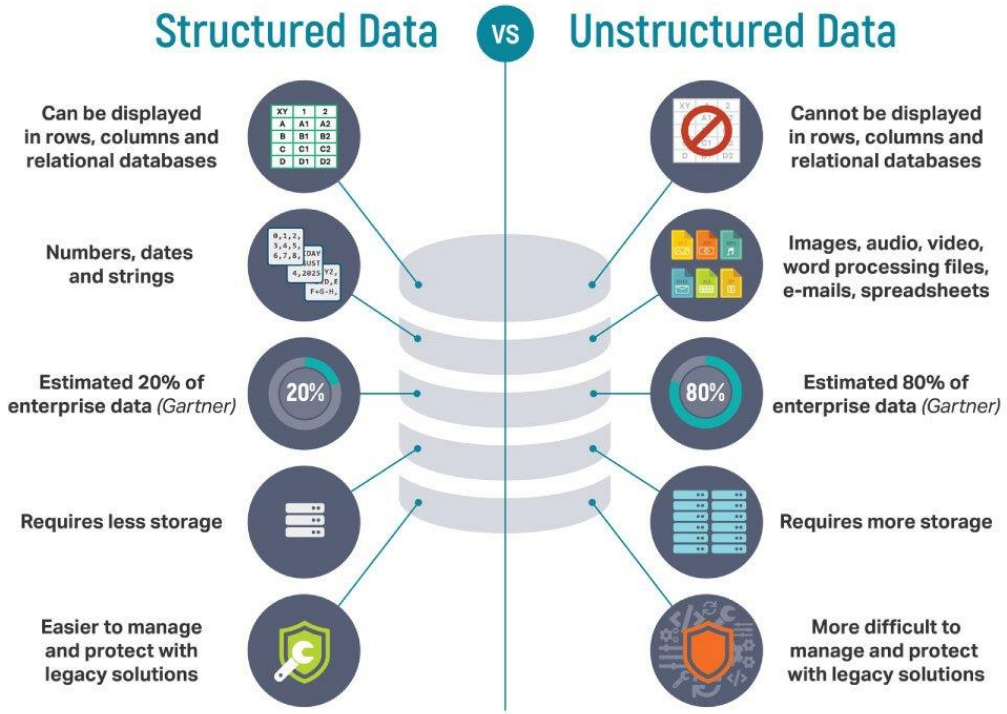
Statista

ACCUMULATED DIGITAL UNIVERSE OF DATA



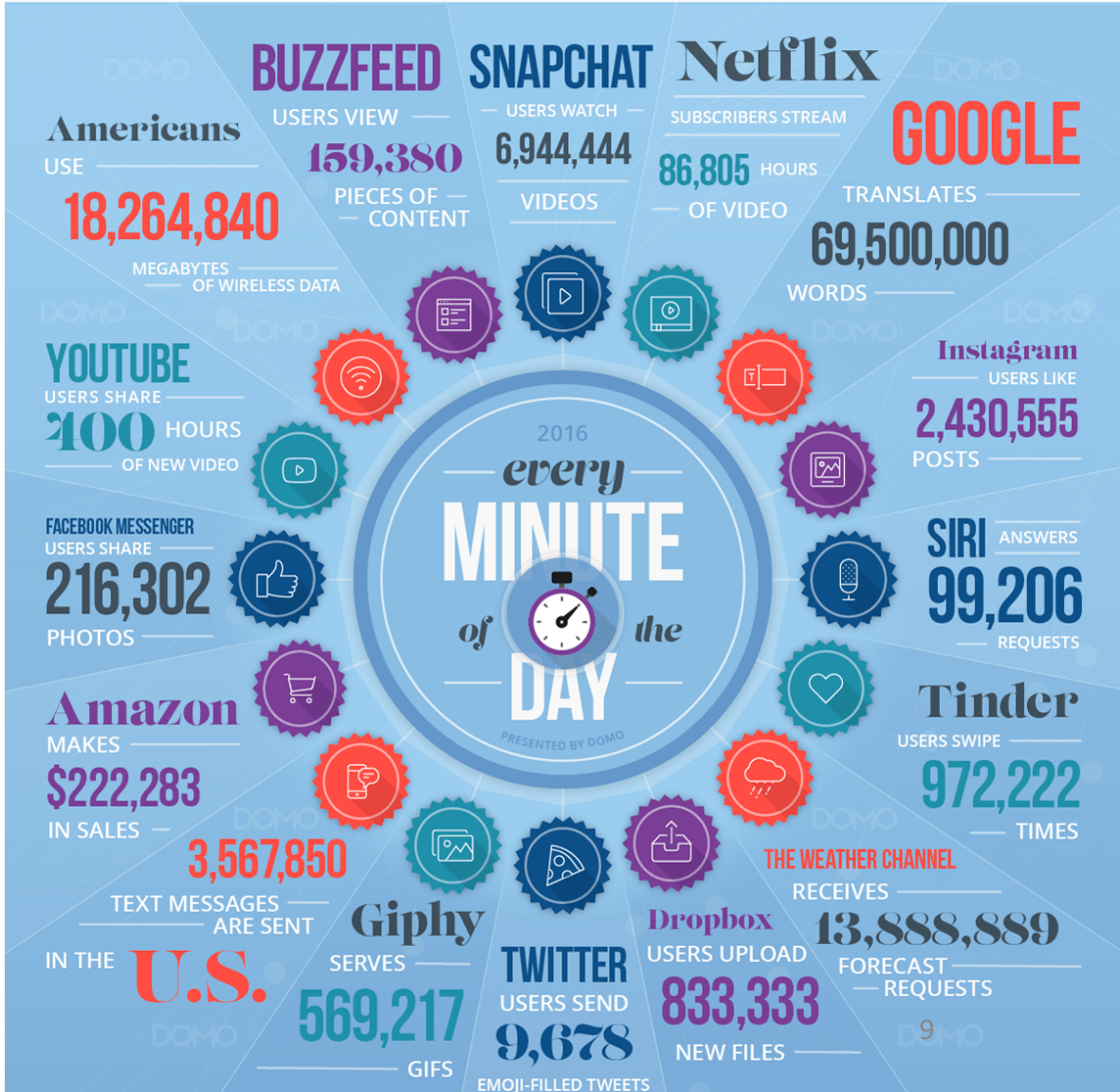
PwC


COLLECTION
STORAGE
NETWORK
RESEARCH
BIG DATA
CLOUD TECHNOLOGY
ANALYSIS
VISUALIZATION
EXABITE
VOLUME



DATA NEVER SLEEPS 4.0

How much data is generated every minute? In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like Youtube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the internet every minute? See for yourself below.



Types of Data

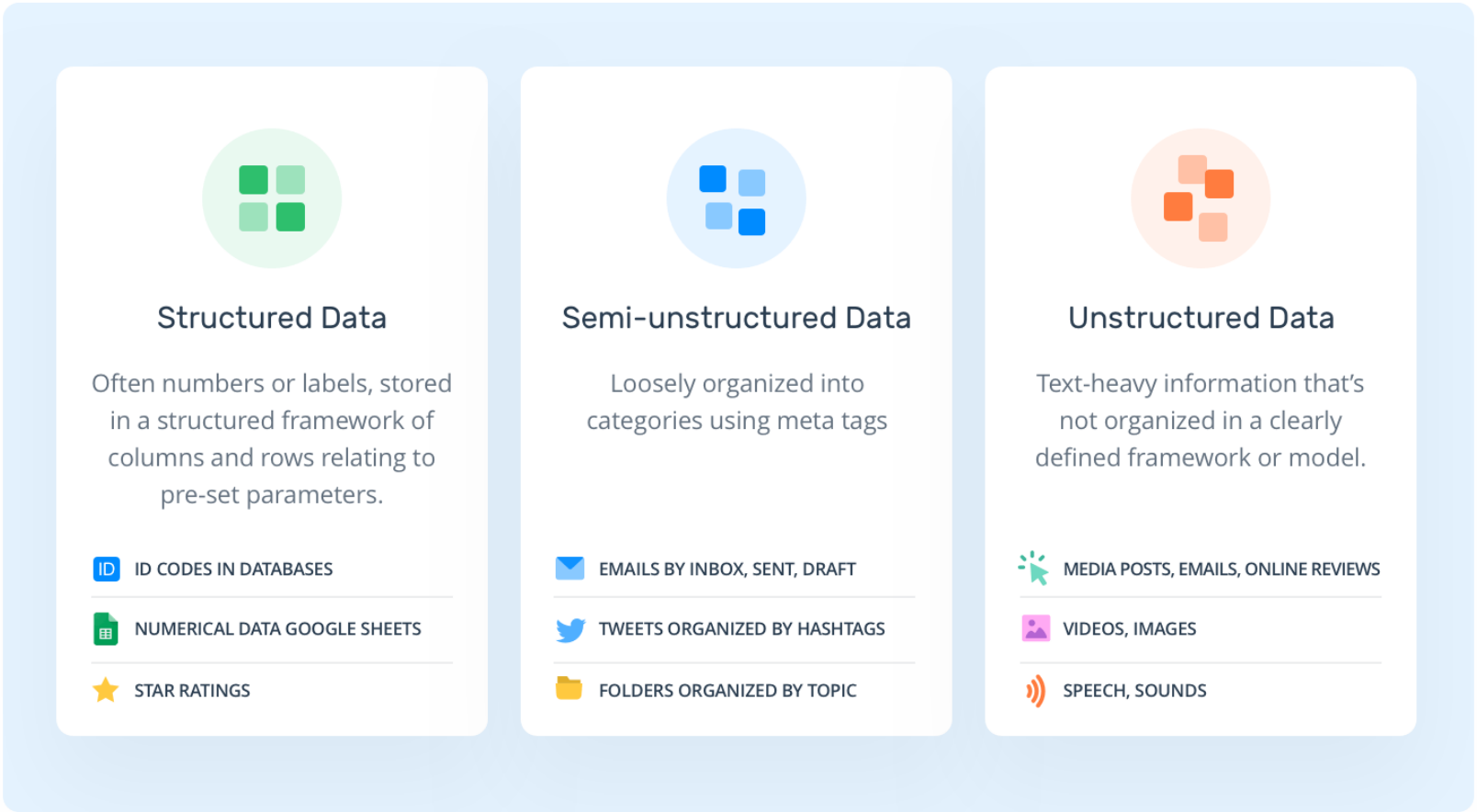
Structured

1001 1010 1110	1001 1110 0110	0101 1110 1000
1001 1010 1000	1001 0010 1110	1001 1010 1110
1001 1010 1110	1001 1010 1000	0101 1110 1000

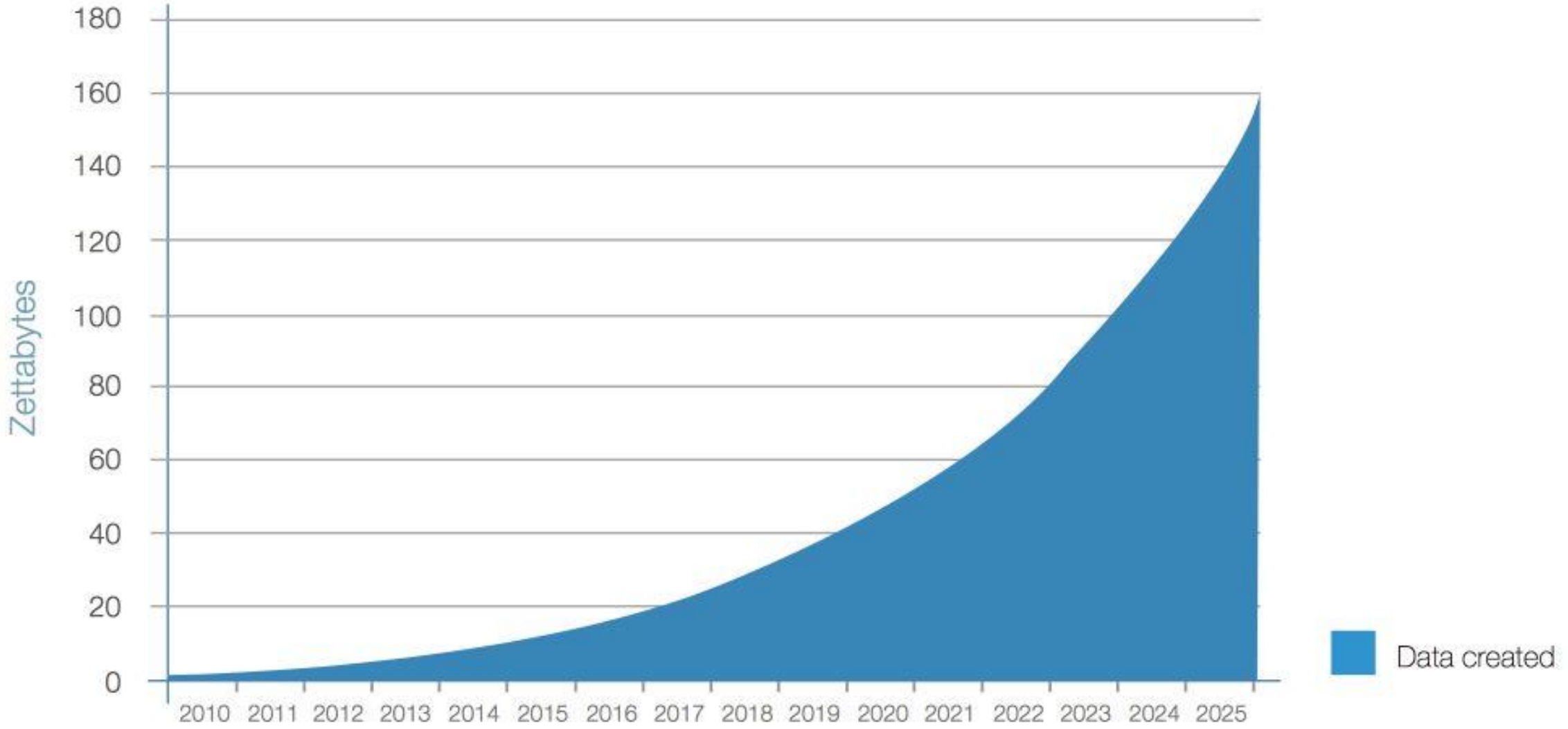
Unstructured

Semi-Structured

Unstructured vs Structured Data



The amount of data produced by us from the beginning of time until 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in 2011, and every ten minutes in 2013, and most likely every minute in 2020.



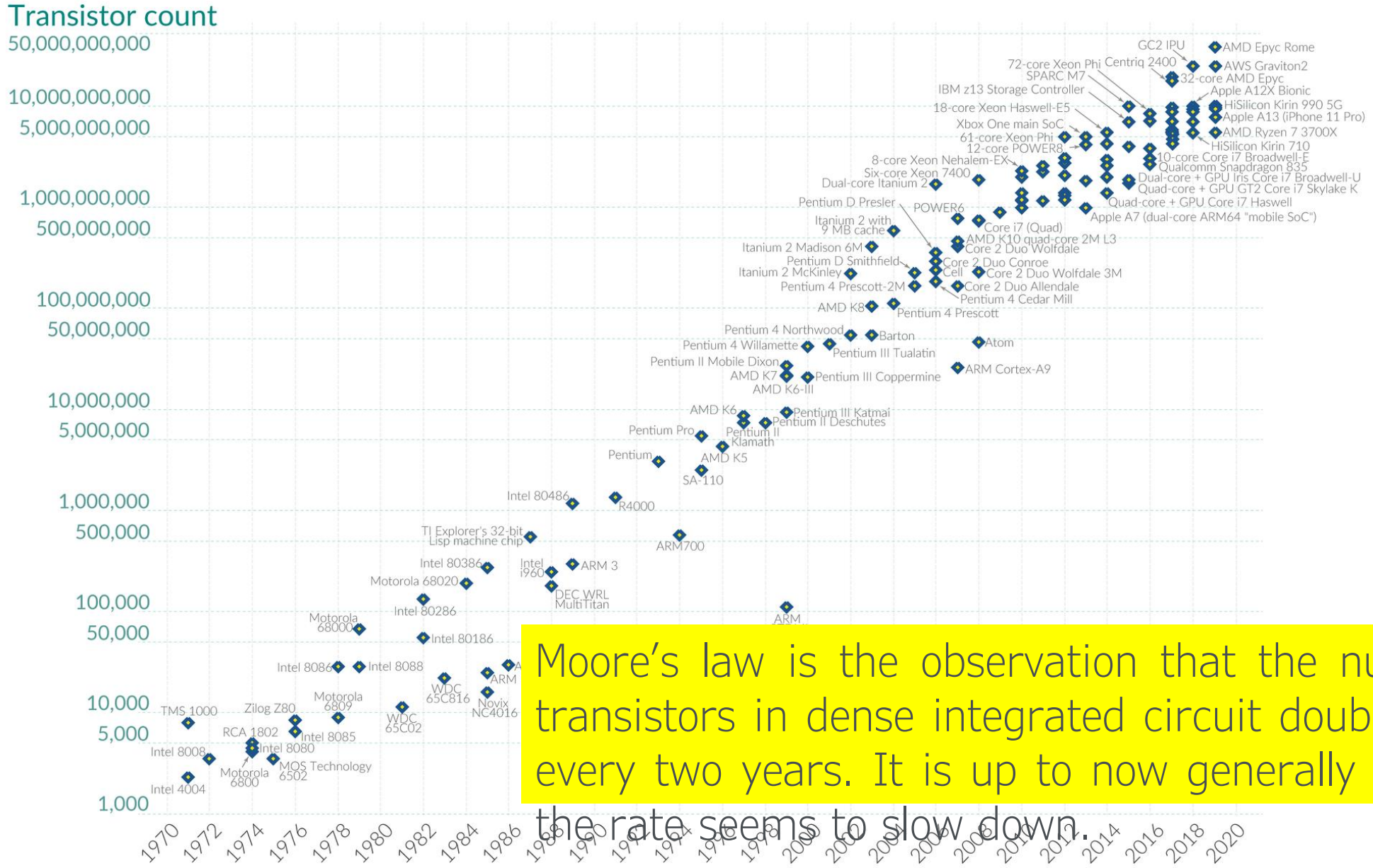
Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

Challenges arise when we need to capture, store, transform, transfer, analyze and present large volumes of data.

Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World in Data



Moore's law is the observation that the number of transistors in dense integrated circuit doubles about every two years. It is up to now generally true, but the rate seems to slow down.

What is Big Data?

Early in the 2010's, the word "Big Data" began to be used in Japanese media.

2011 was called the "first year of big data," and the utilization of big data progressed significantly from that year.

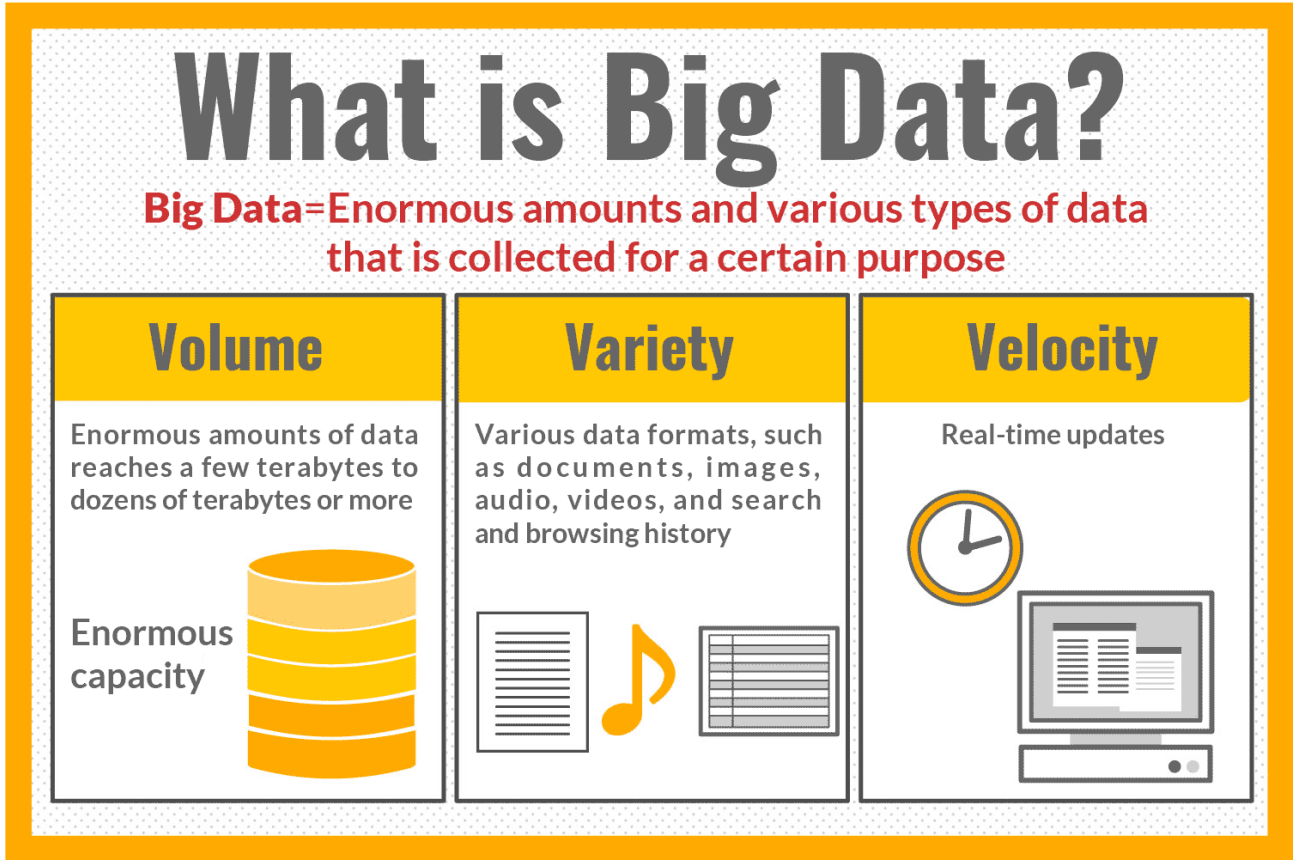


Image credit: https://en.mdv.co.jp/column/article/01.html

What is Big Data?

	Traditional Data	Big Data
Volume	The amount of data which can be processed by existing database system without any problems.	The amount of data reaches dozens of terabytes to a few petabytes or more and it cannot be processed by existing database system.
Variety	Only structured data that can be represented by tables ("columns" and "row") in a database system, such as CSV and Excel files.	Includes unstructured data in various data formats, such as documents, images, audio, videos, and even search and browsing history.
Velocity	Obtaining and analyzing data is not done in real time.	Data (e.g. traffic and financial information) is updated in real time, so it needs to be obtained and analyzed quickly.

Image credit: https://en.mdv.co.jp/column/article/01.html

VOLUME

Huge amount of data

VERACITY

Inconsistencies and uncertainty in data

VARIETY

Different formats of data from various sources

VELOCITY

High speed of accumulation of data

VALUE

Extract useful data

The 5 V's of Big Data

XENONSTACK

01. Volume

The amount of data

Big data involves a huge volume of data. The volume refers to the amount of data generated every second, minute and days.

02. Velocity

An unprecedented speed

Speed is the Big V that represents how fast data is being received and processed.

03. Variety

The types of data that are available

When working with so much data, a lot of it is unstructured and need to be further processed to structure it properly.

04. Veracity

The degree to which Big Data can be trusted

When you have a lot of data, you can actually use it for very different purposes and format it in different ways.

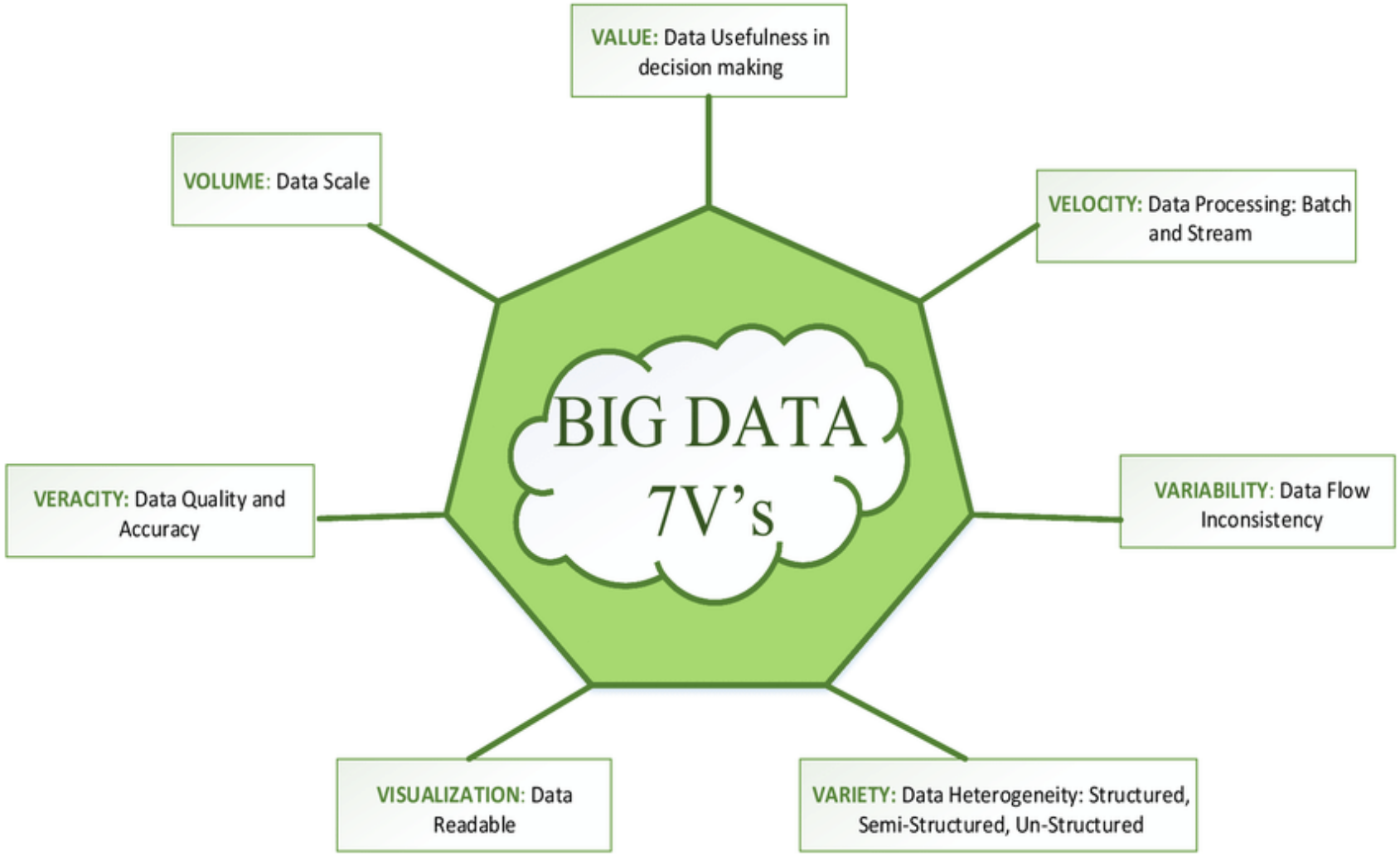
05. Value

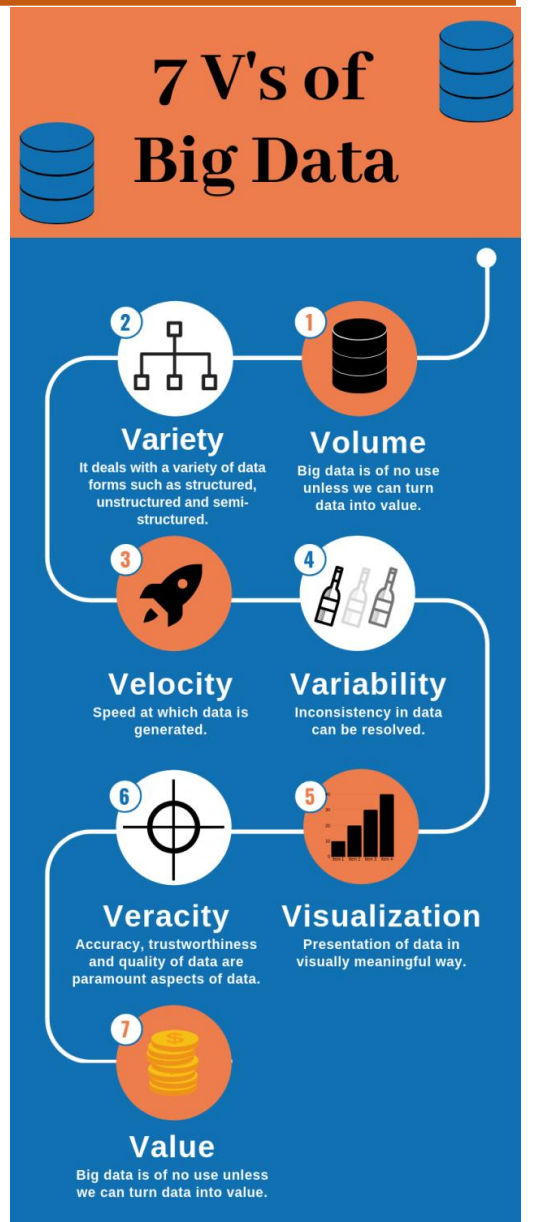
The final output from data management

The value of the data is that they are actionable, responsible for the companies to make a decision.

AI VIETNAM
All-in-One Course

What is Big Data?





AI VIETNAM
All-in-One Course

What is Big Data?

What is Big Data Processing?

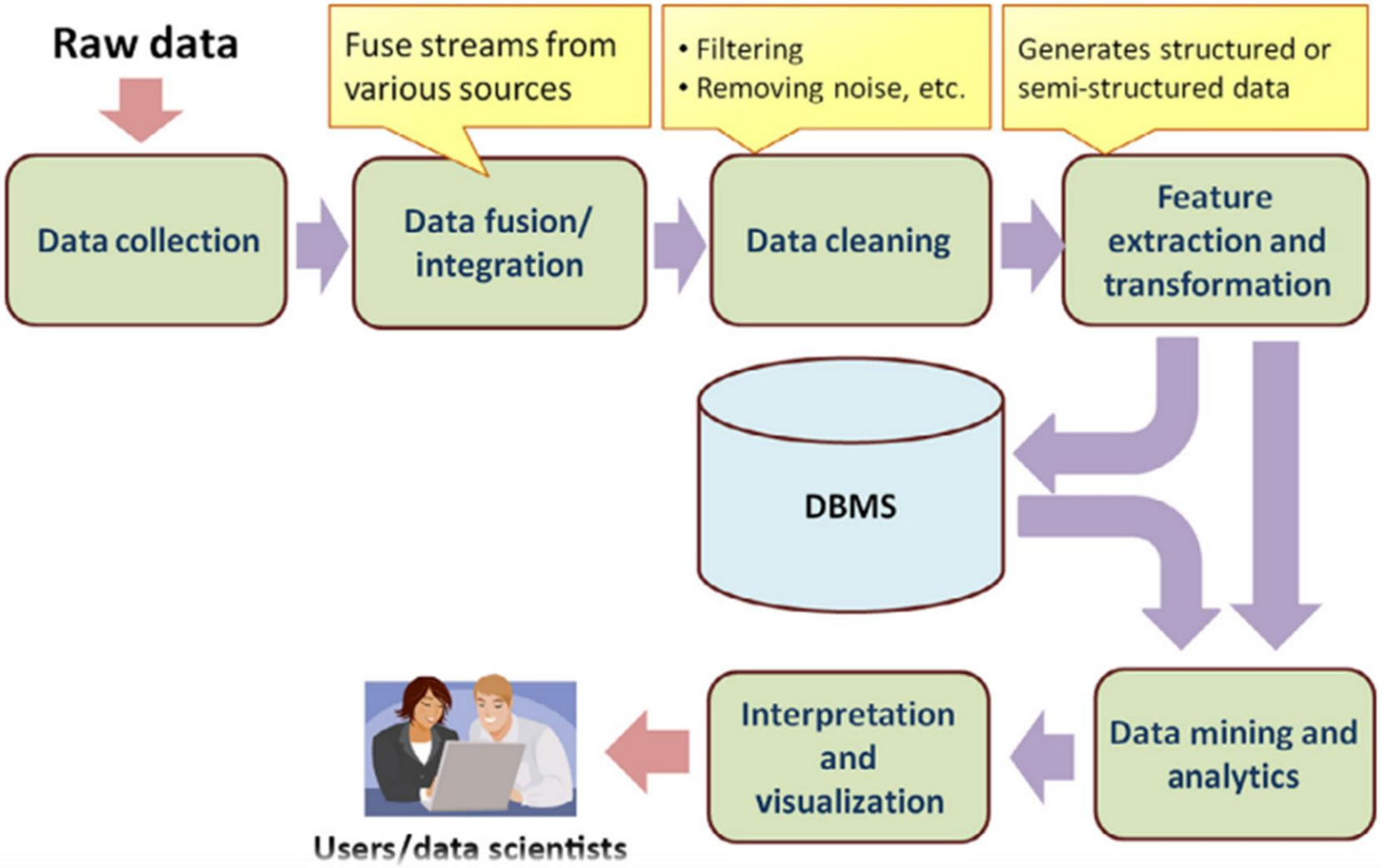
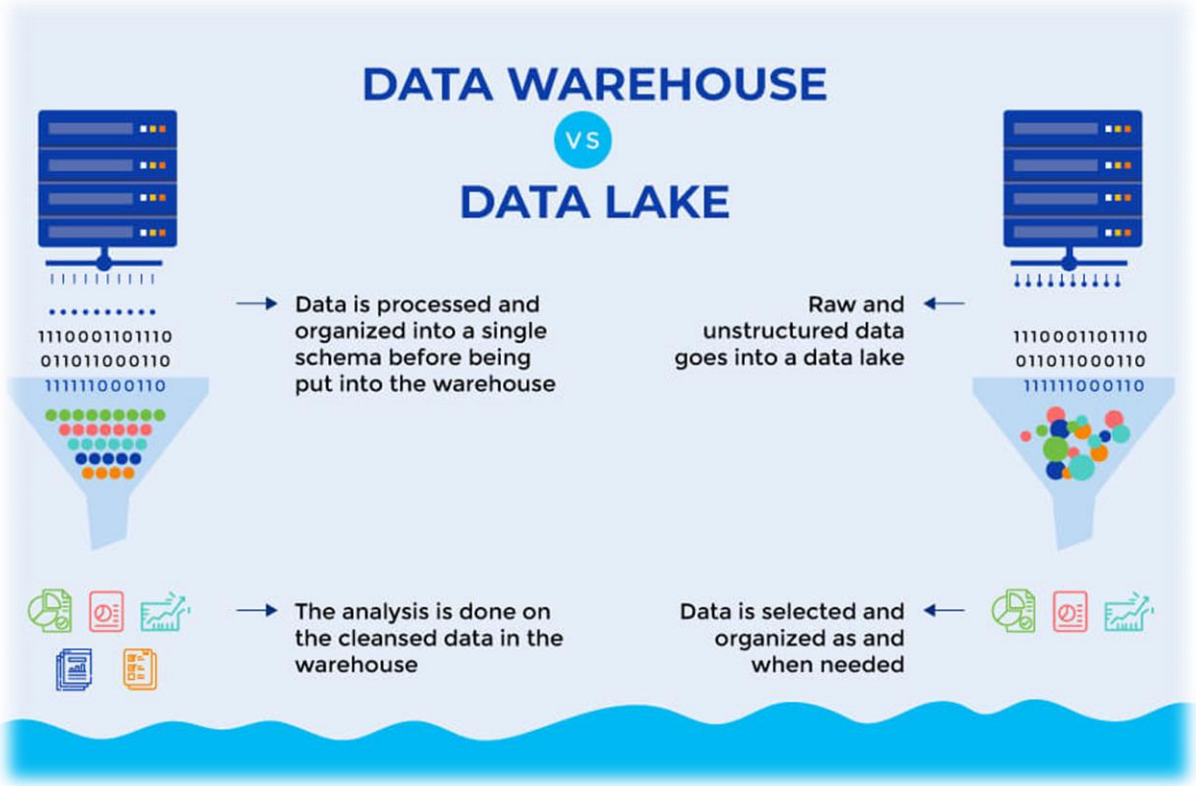
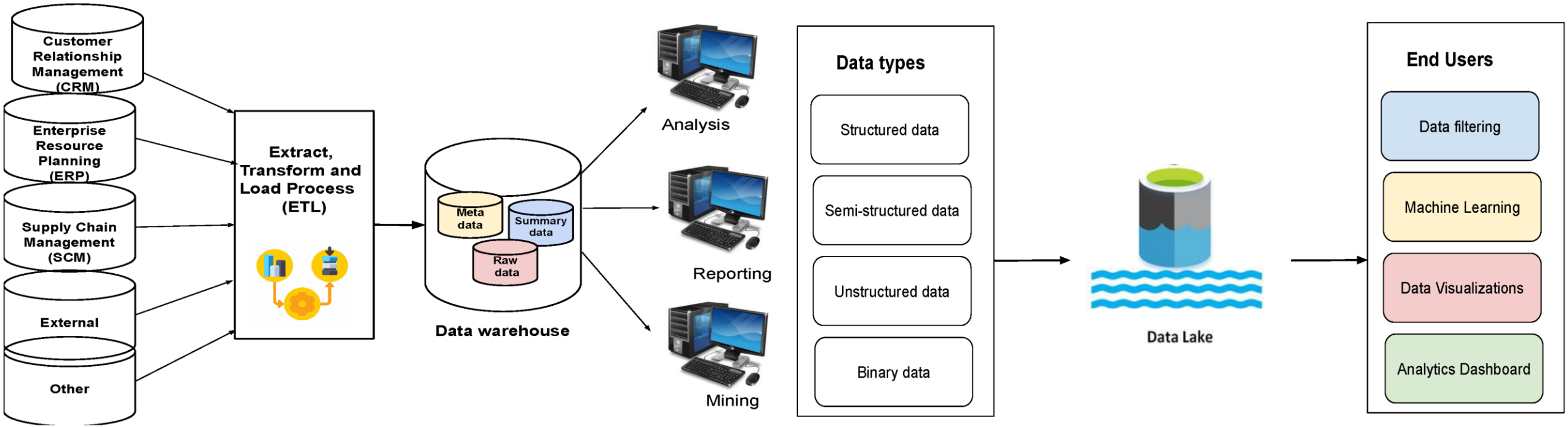


Image Credits: https://hevodata.com/learn/big-data-processing/#lg=1&slide=0

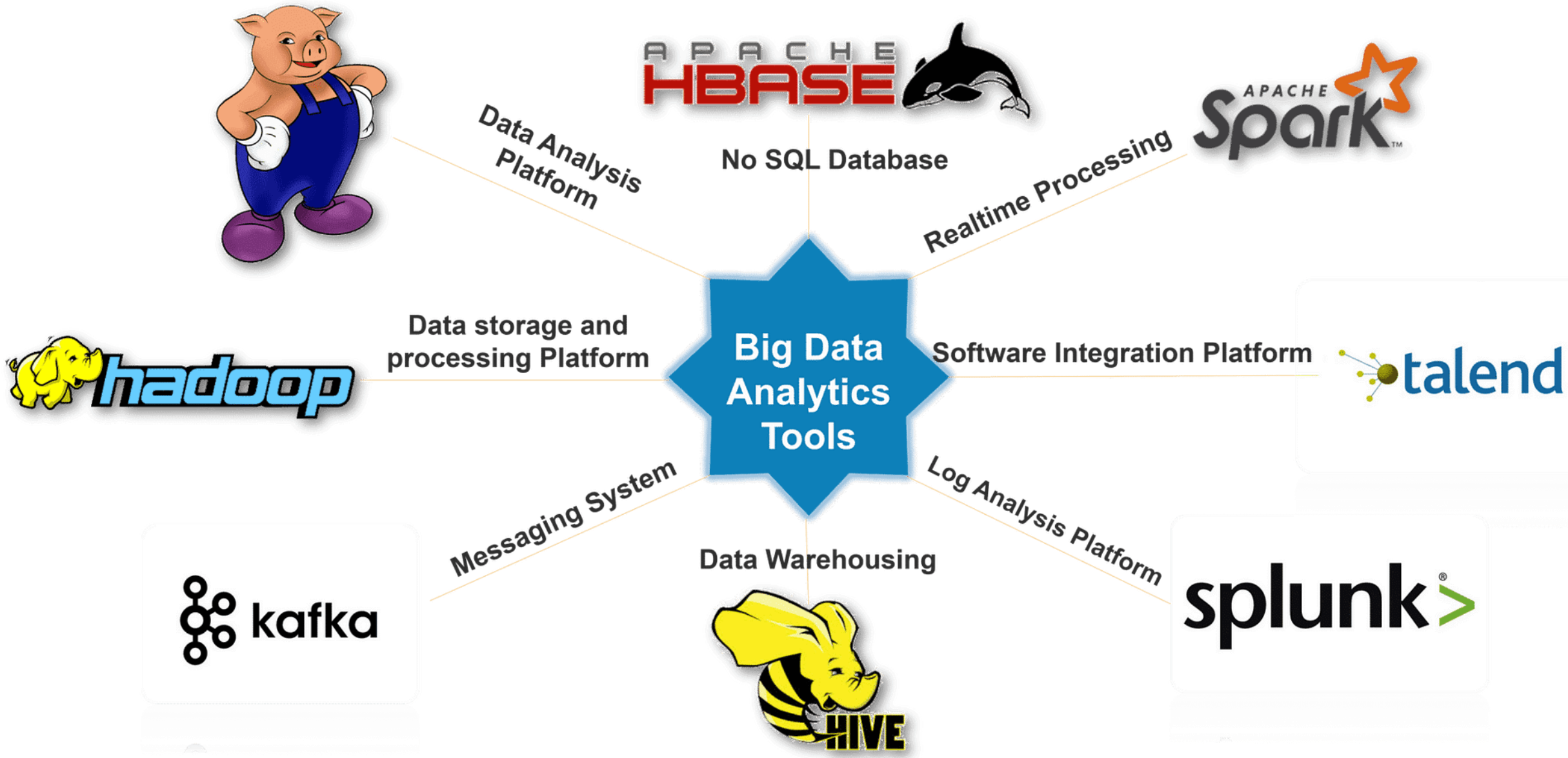
How do we store and process Big Data?





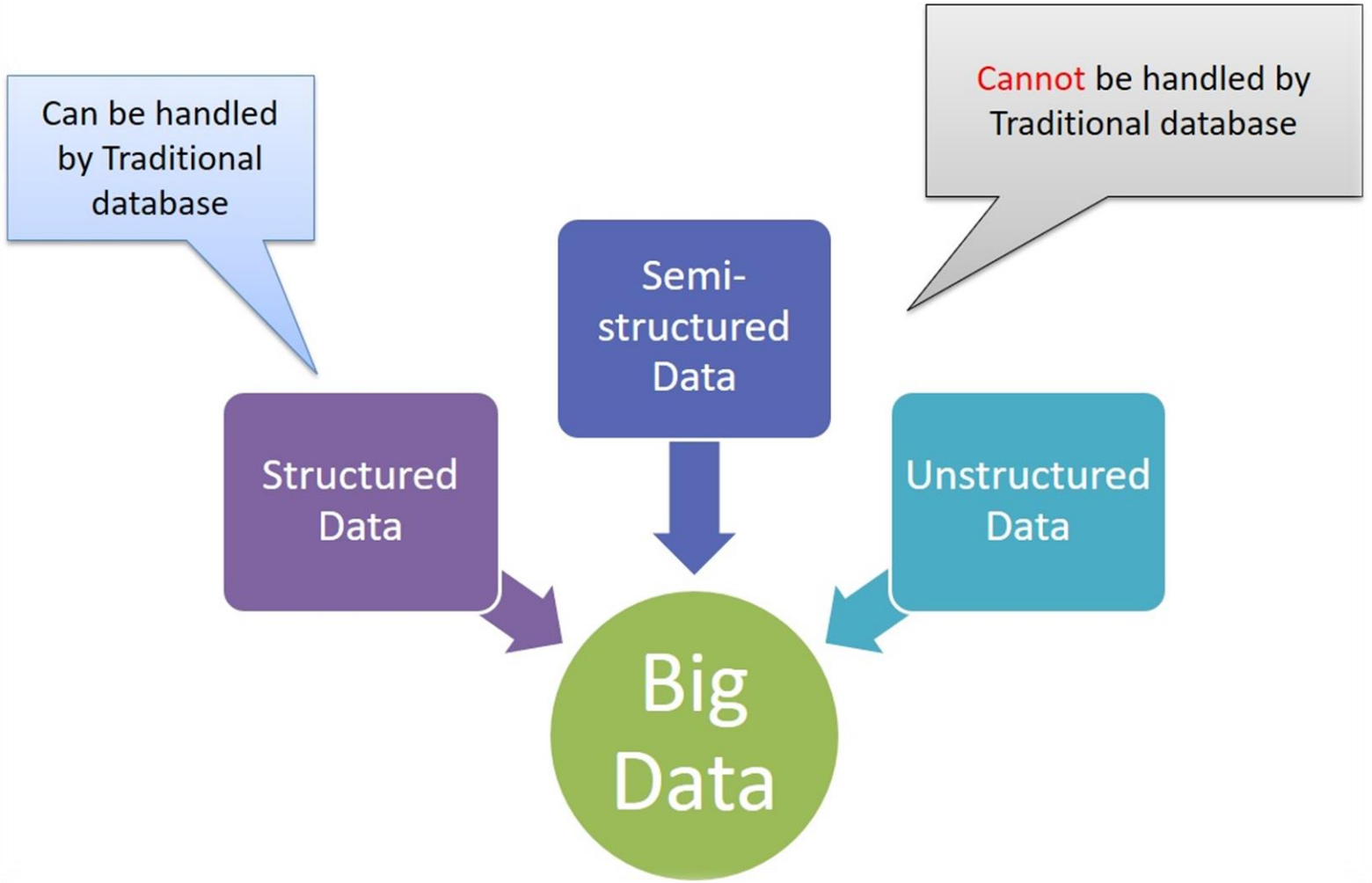
(a) Data warehouse (b) Data lake

https://www.mdpi.com/2504-2289/6/4/132



https://towardsdatascience.com/big-data-analysis-spark-and-hadoop-a11ba591c057

Challenges for traditional Database management system to handle Big Data



https://blogs.sap.com/2019/06/24/what-is-big-data-and-why-do-we-need-hadoop-for-big-data/

Challenges for traditional Database management system to handle Big Data

Traditional databases are not designed to handle
the speed at which Big Data arrives or needs to be analyzed.

https://blogs.sap.com/2019/06/24/what-is-big-data-and-why-do-we-need-hadoop-for-big-data/

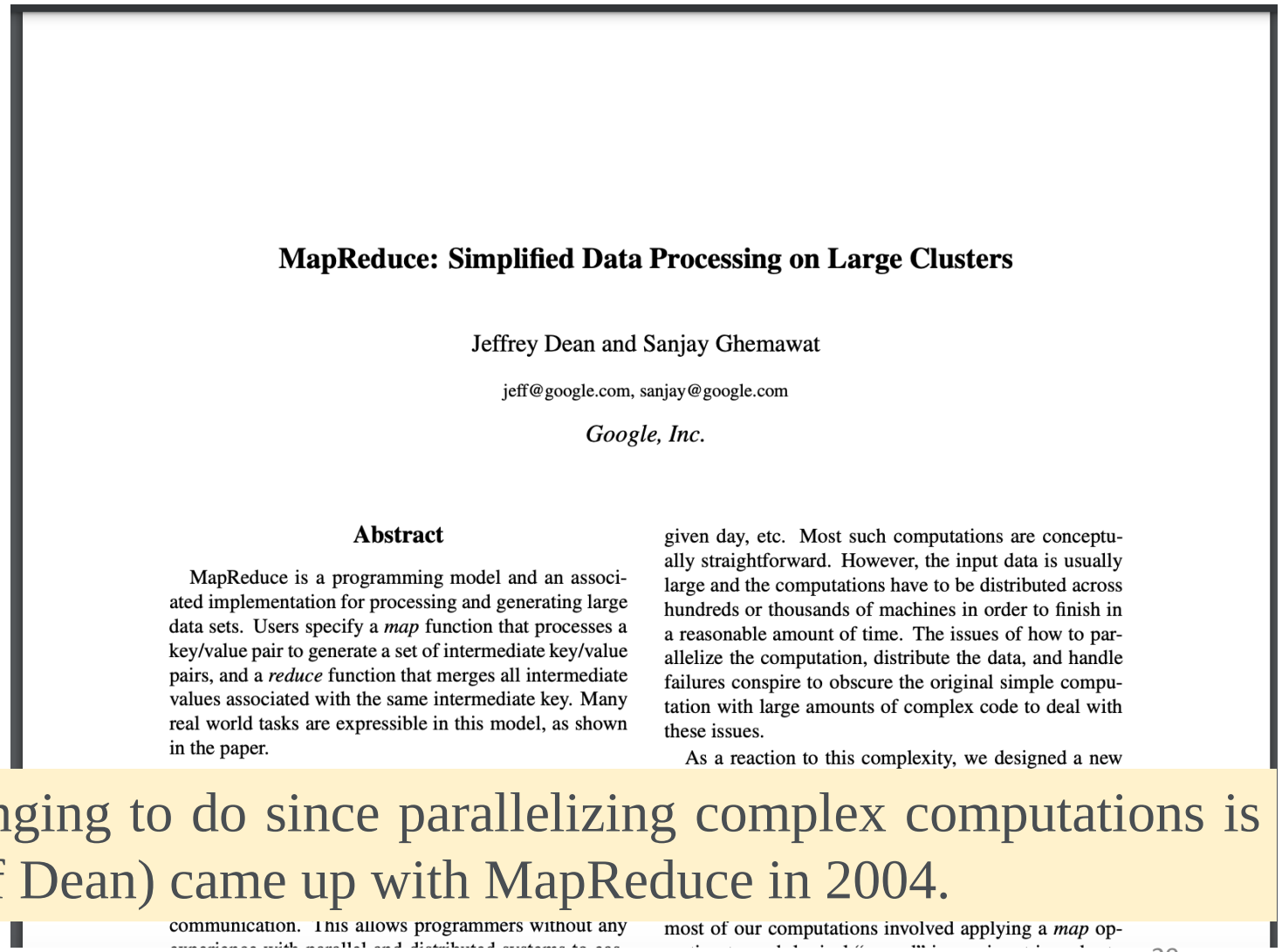
Challenges for traditional Database management system to handle Big Data

Big data in Zettabyte and Exabyte, should be processed in parallel independent tasks. Traditional database cannot handle this.

https://blogs.sap.com/2019/06/24/what-is-big-data-and-why-do-we-need-hadoop-for-big-data/

Storage and Process Big Data Technologies

MapReduce divides the task we want to do on the data into small parts and assigns them to several computers (a cluster), and aggregates the results from them.



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

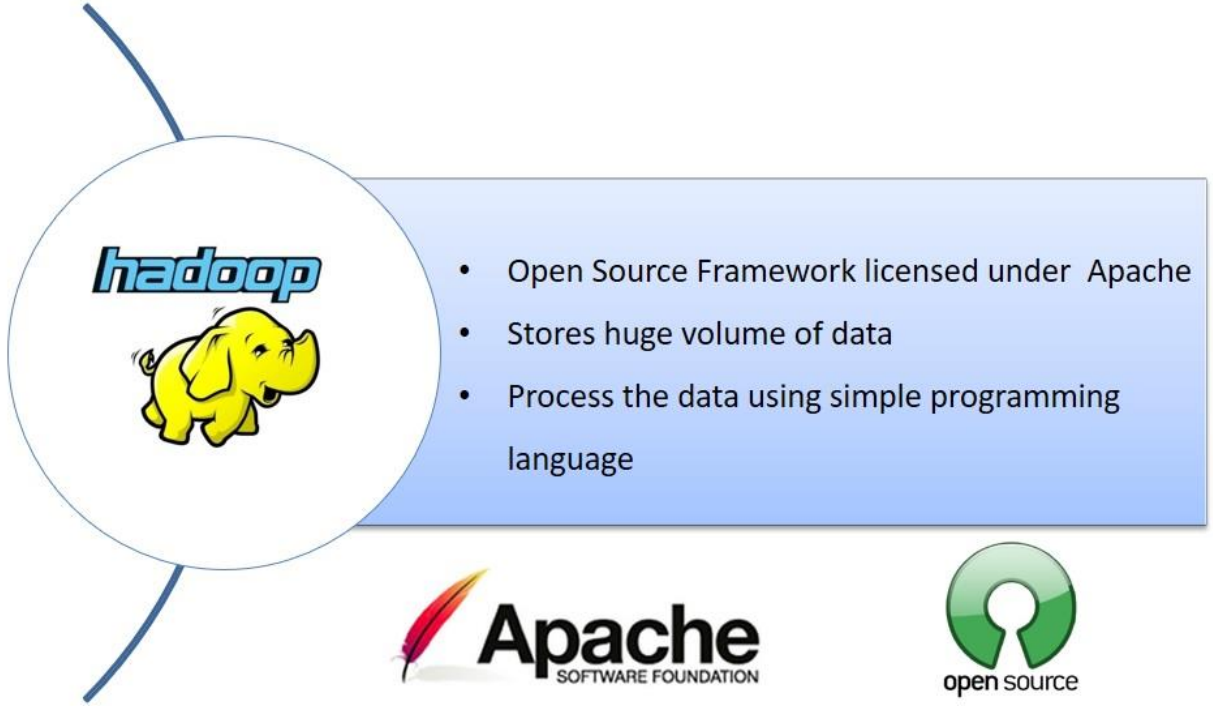
As a reaction to this complexity, we designed a new

communication. This allows programmers without any

most of our computations involved applying a *map* op-

This is however technically challenging to do since parallelizing complex computations is not trivial. This is why Google (Jeff Dean) came up with MapReduce in 2004.

Apache Hadoop is a framework that can store and process huge amounts of unstructured data ranging in size from terabytes to petabytes. It manages big data storage in a distributed way and process it parallel.



History

•2004: Google publishes the MapReduce paper

•2006: Hadoop is integrated into Apache

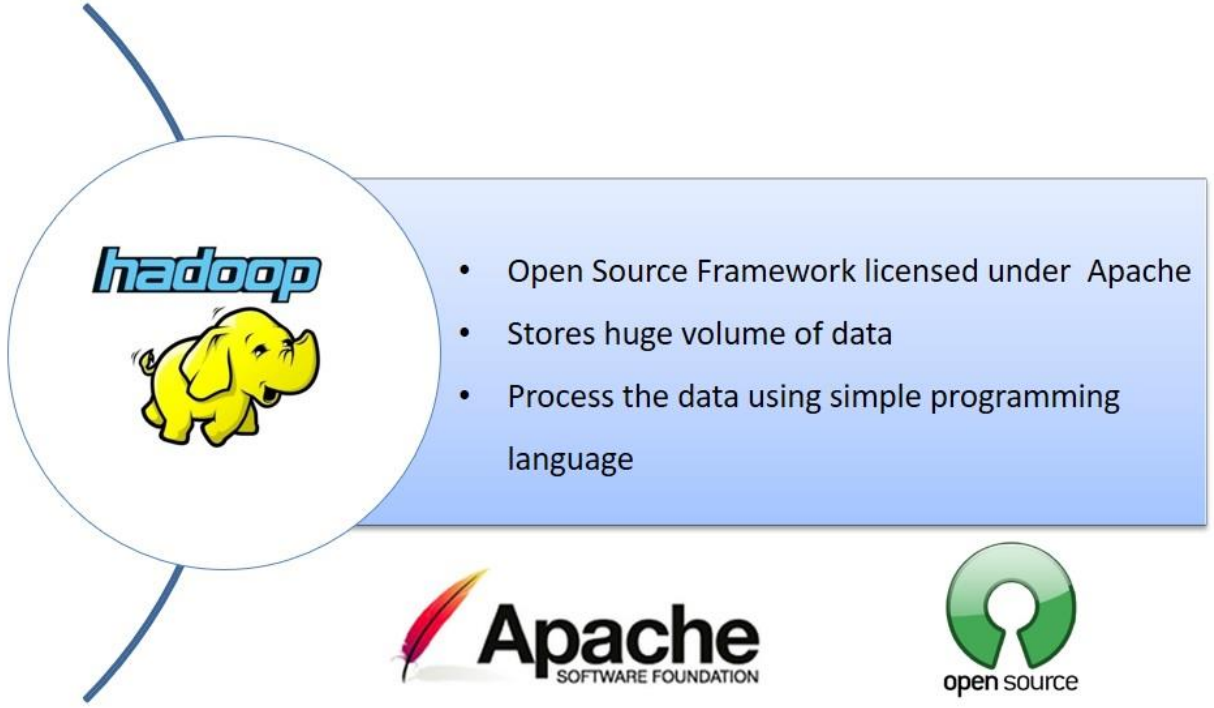
•2014: Apache launches Spark

Hadoop was created by Doug Cutting and Mike Cafarella in 2005.

Hadoop is not a database

Hadoop got its start as a Yahoo project in 2006, becoming a top-level Apache open-source project later on.

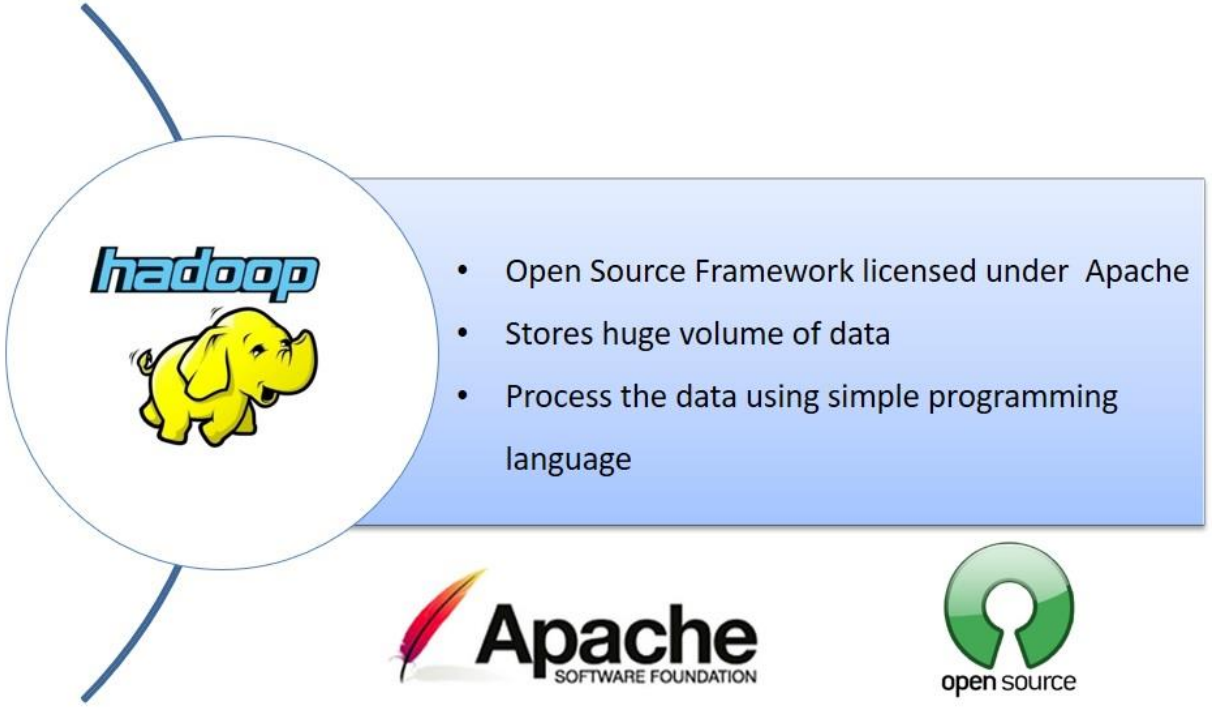
It's a general-purpose form of distributed processing that has several components: the Hadoop Distributed File System (HDFS), which stores files in a Hadoop-native format and parallelizes them across a cluster



Hadoop was created by Doug Cutting and Mike Cafarella in 2005.

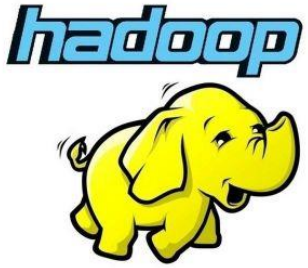
Hadoop is not a database

It's available either open-source through the Apache distribution, or through vendors such as Cloudera (the largest Hadoop vendor by size and scope), MapR, or HortonWorks.

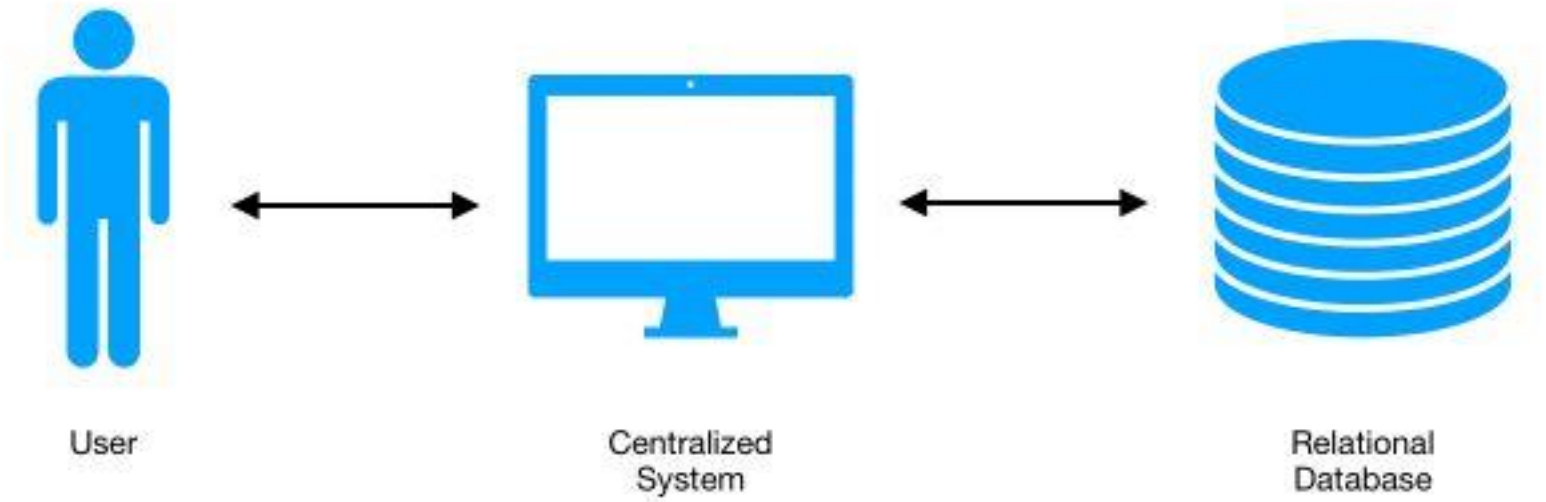


Hadoop was created by Doug Cutting and Mike Cafarella in 2005.

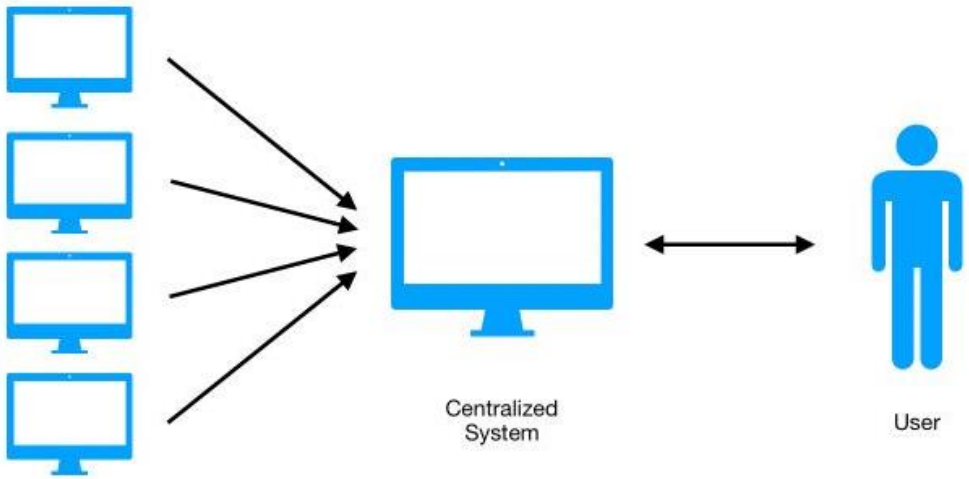
Hadoop is not a database

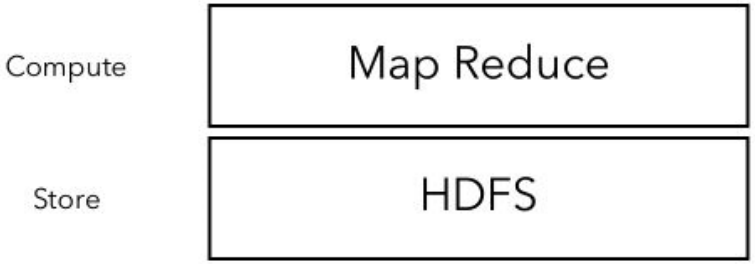


What changes with Hadoop?



In traditional approaches, the user interacts with a centralized system which queries the databases.

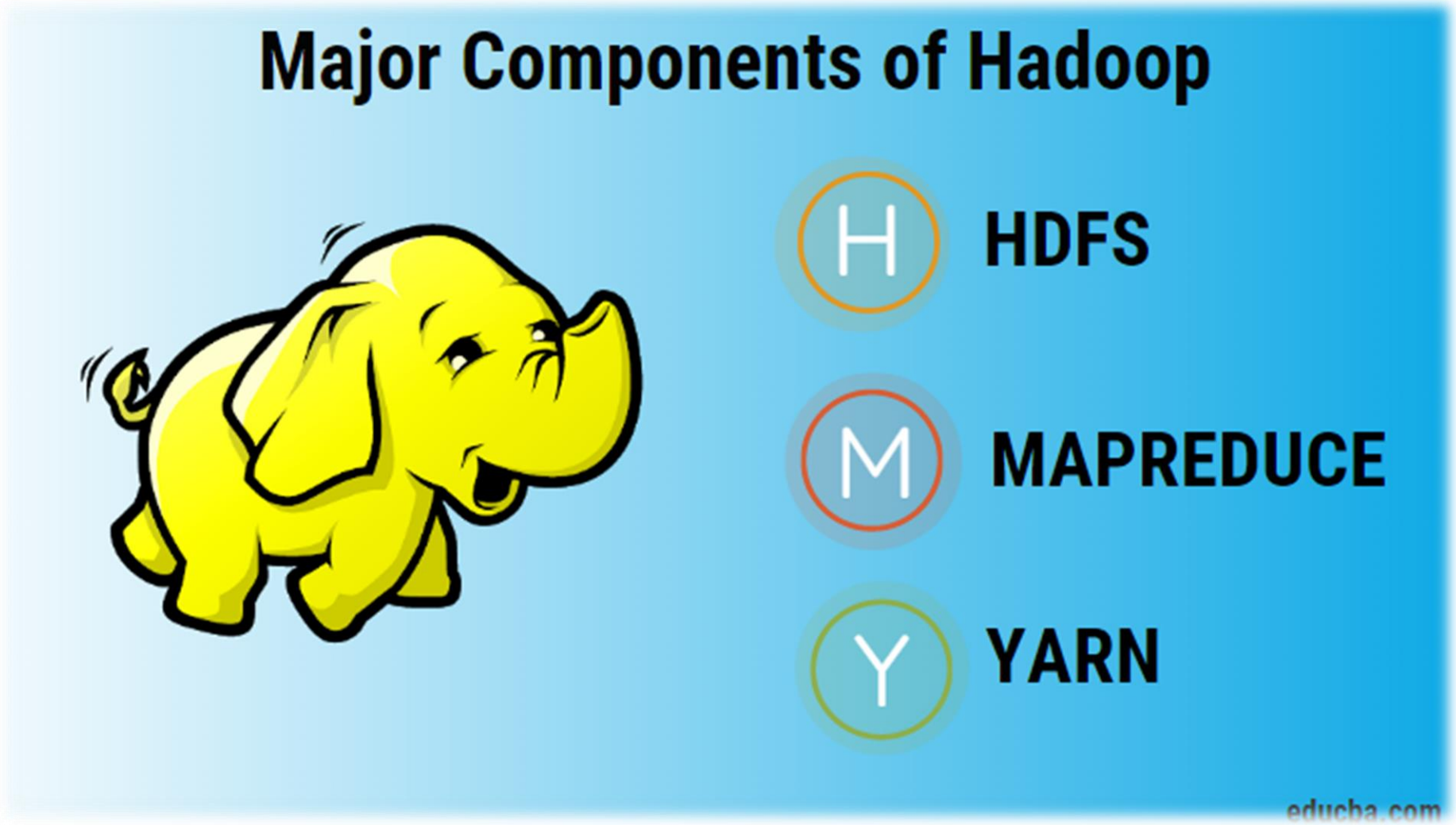




With Hadoop, this paradigm changes. The task is no longer centralized but split into several workers.

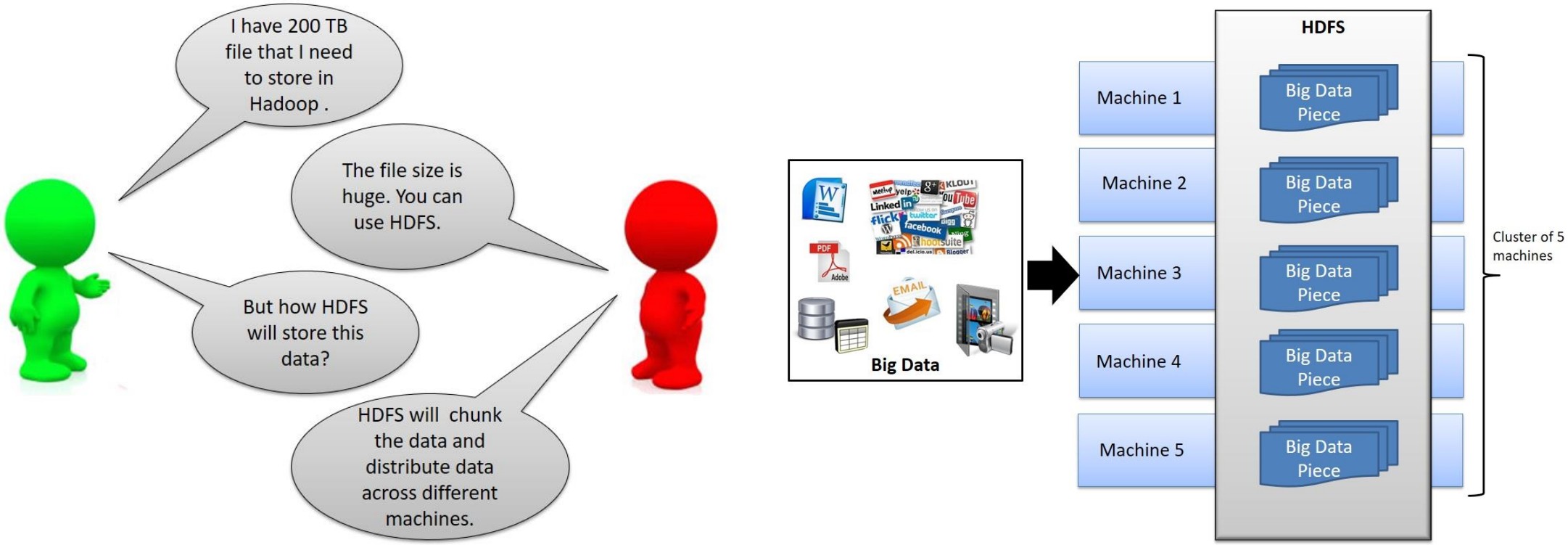
HDFS is used to store the data, and MapReduce to perform actions on the data.

Hadoop Core Components

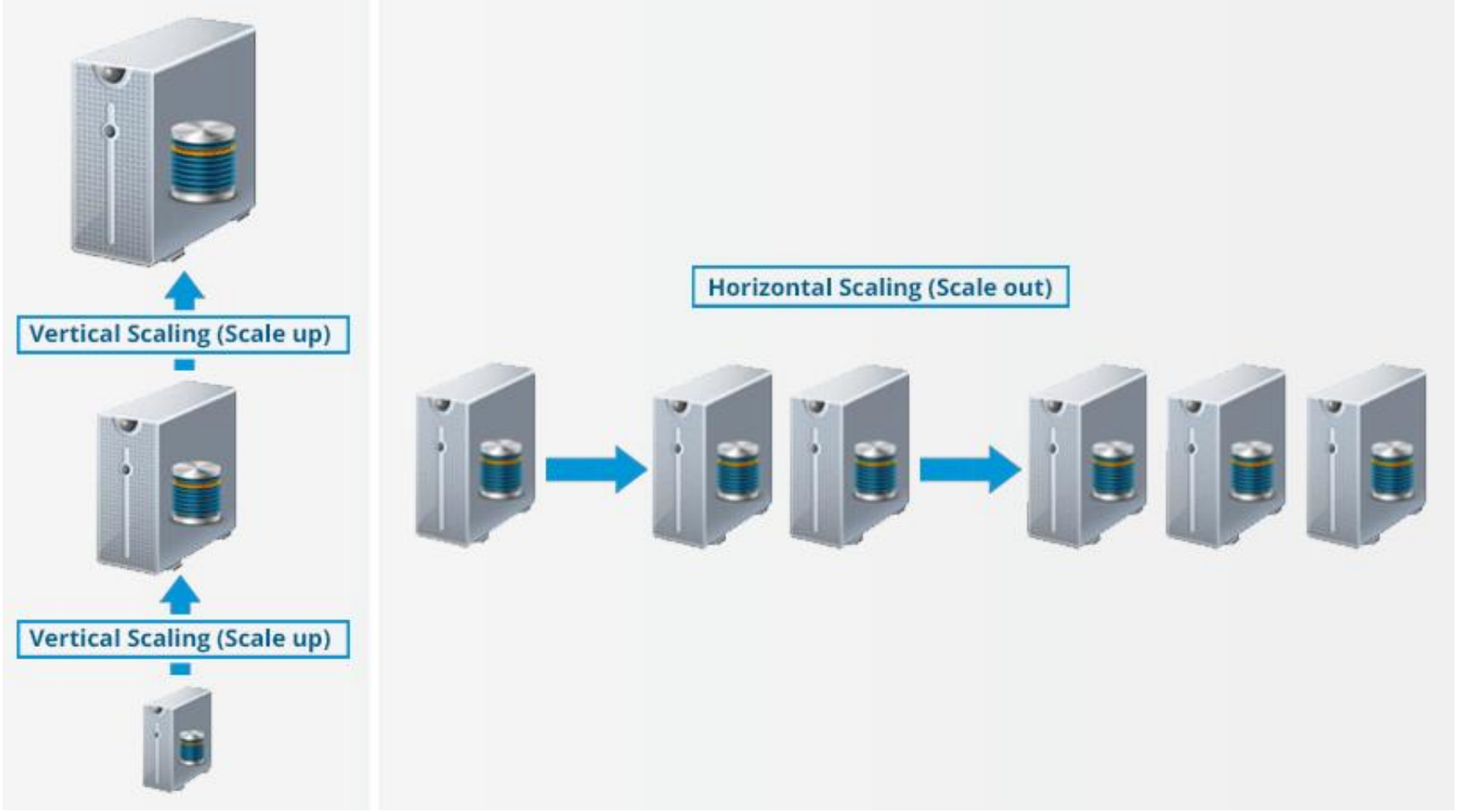


HDFS – Hadoop Distributed File System

Đây là thành phần xử lý việc lưu trữ file phân tán của Hadoop.



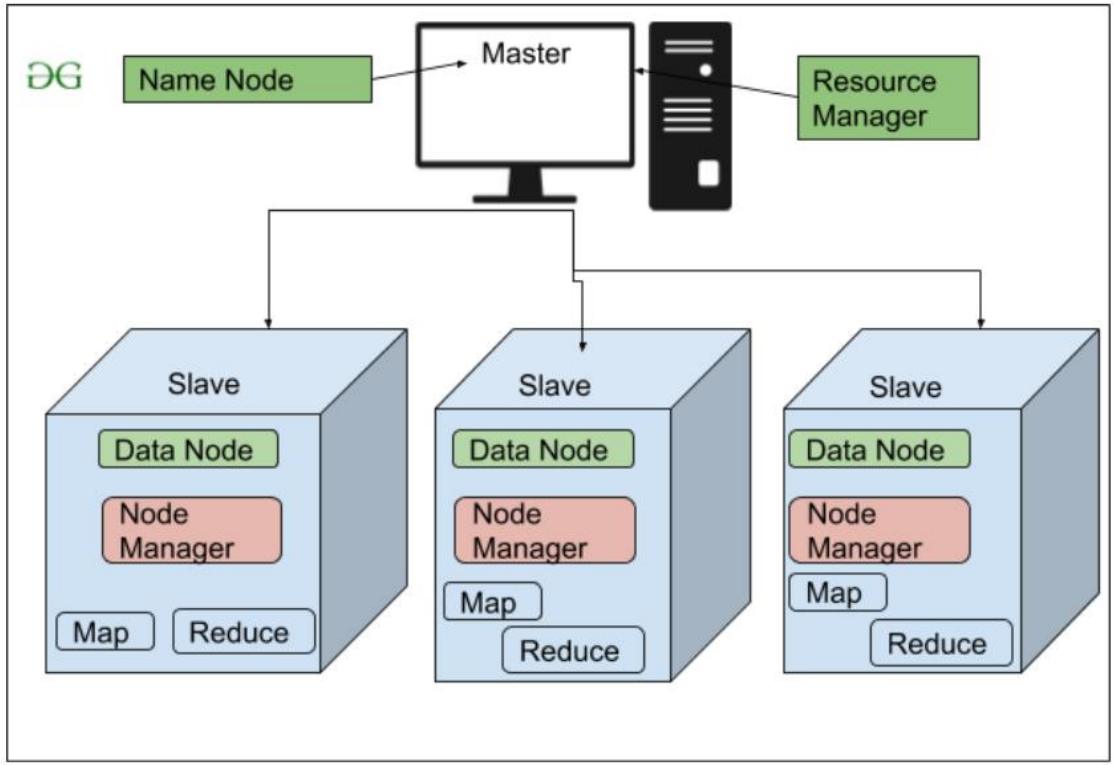
HDFS – Hadoop Distributed File System

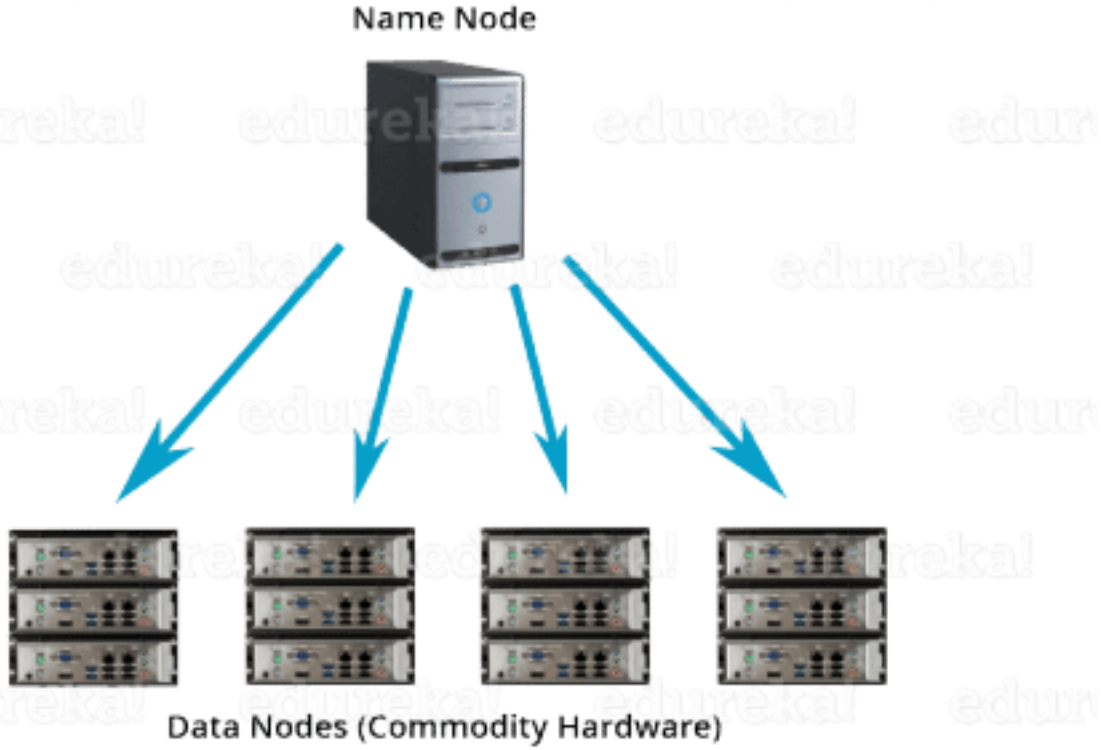


HDFS – Hadoop Distributed File System

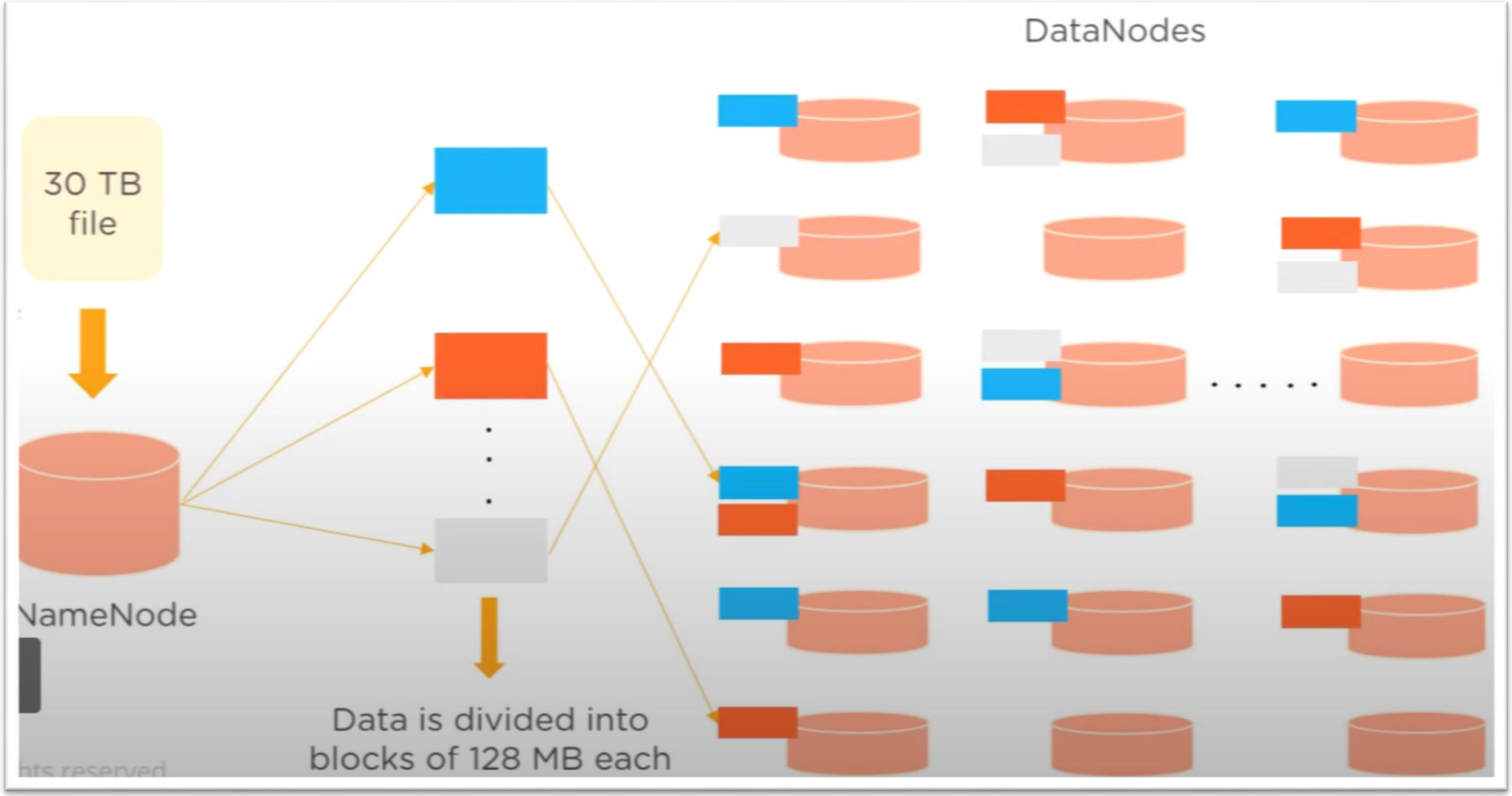
- HDFS has two core components

•Datanode sẽ là nơi lưu trữ các file dữ liệu mà bạn đưa vào. Namenode là nơi lưu địa chỉ của file đó được chia và lưu trên các datanode nào.



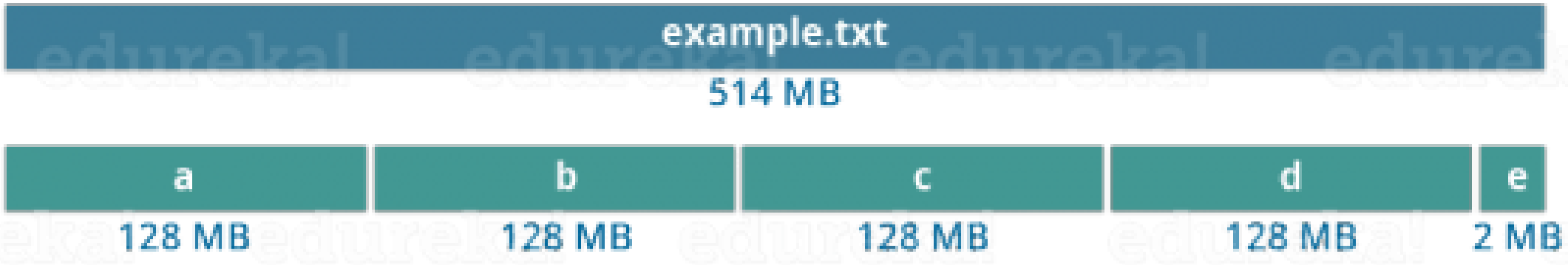


HDFS – Hadoop Distributed File System

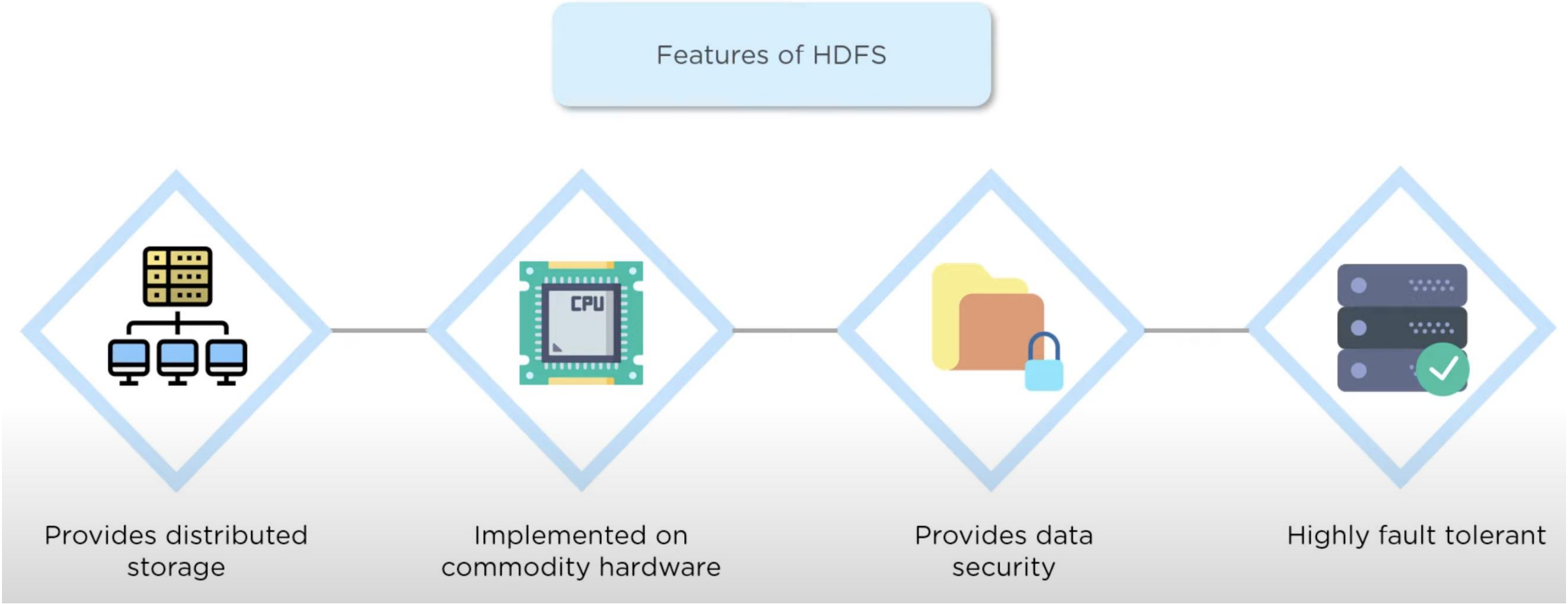


HDFS – Hadoop Distributed File System

Block trong HDFS



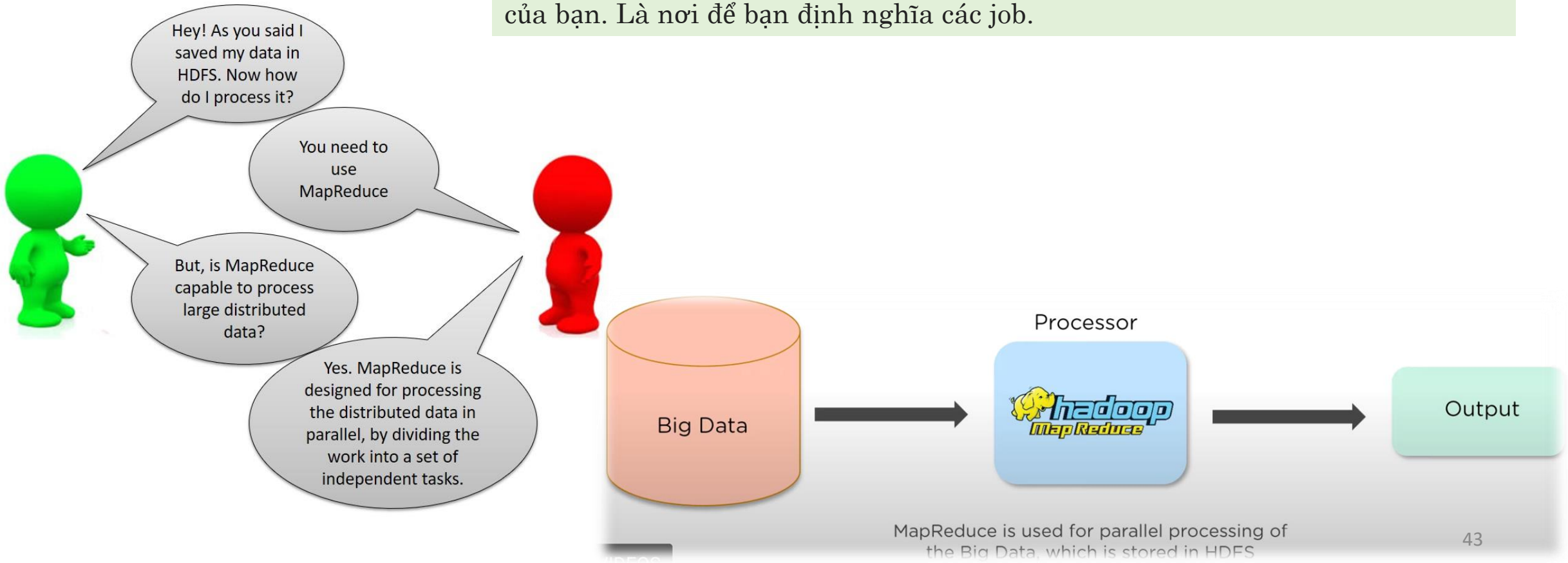
HDFS – Hadoop Distributed File System



MapReduce

It is a programming model designed for processing the distributed data stored in parallel, by dividing the work into a set of independent tasks.

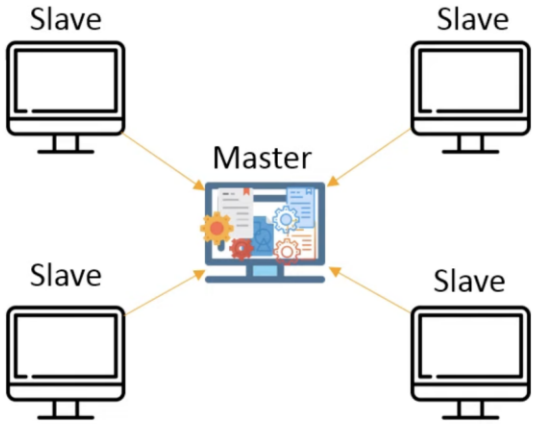
MapReduce: Đây chính là thành phần để bạn "viết code" xử lý cho bài toán của bạn. Là nơi để bạn định nghĩa các job.



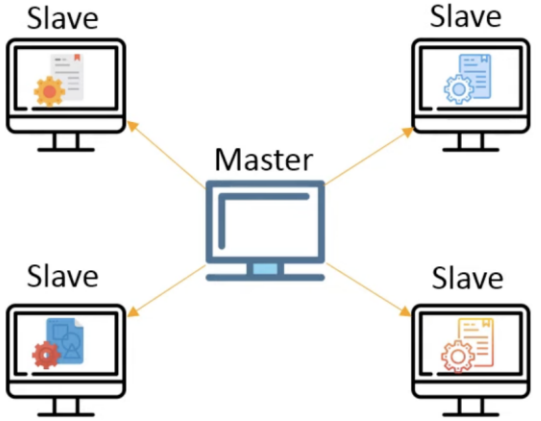
MapReduce

MapReduce is a programming paradigm that allows processing a large amount of data by initially splitting the data into blocks, sending the blocks to different clusters to perform operations, and aggregating the results.

In MapReduce approach, processing is done at the slave nodes and the final result is sent to the master node



Traditional approach – Data is processed at the Master node



MapReduce approach – Data is processed at the Slave nodes

MapReduce / WordCount Problem

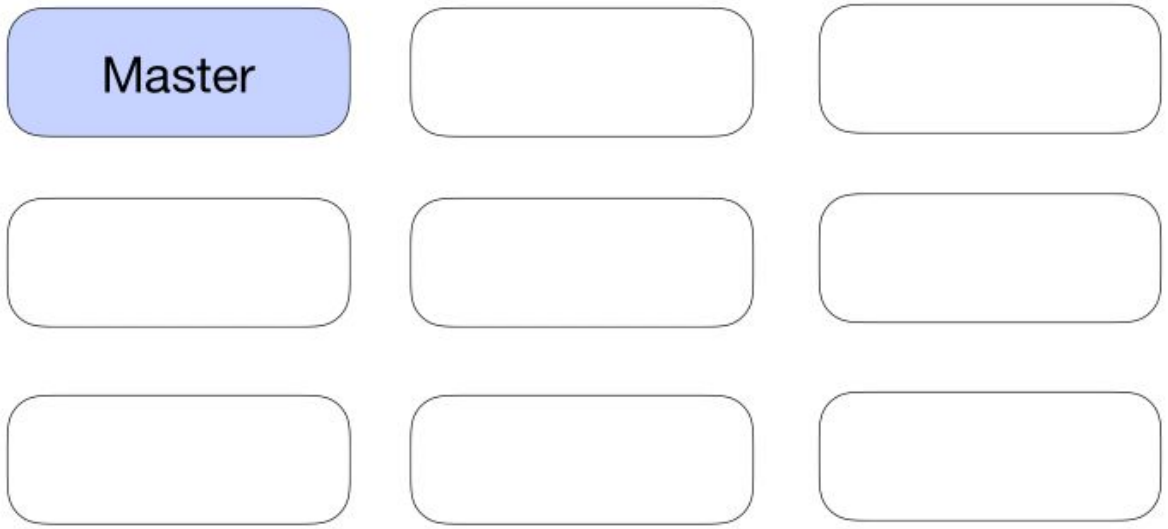
From an input text, we will count how many times each word appears, and rank the final list by occurrence.

MapReduce is of course not needed for such task, and a simple Python script on your computer would be fine. But what if your input text is a whole book? Thousands of books? Millions of books? ...

Algorithm

WordCount Problem – Elect the Master

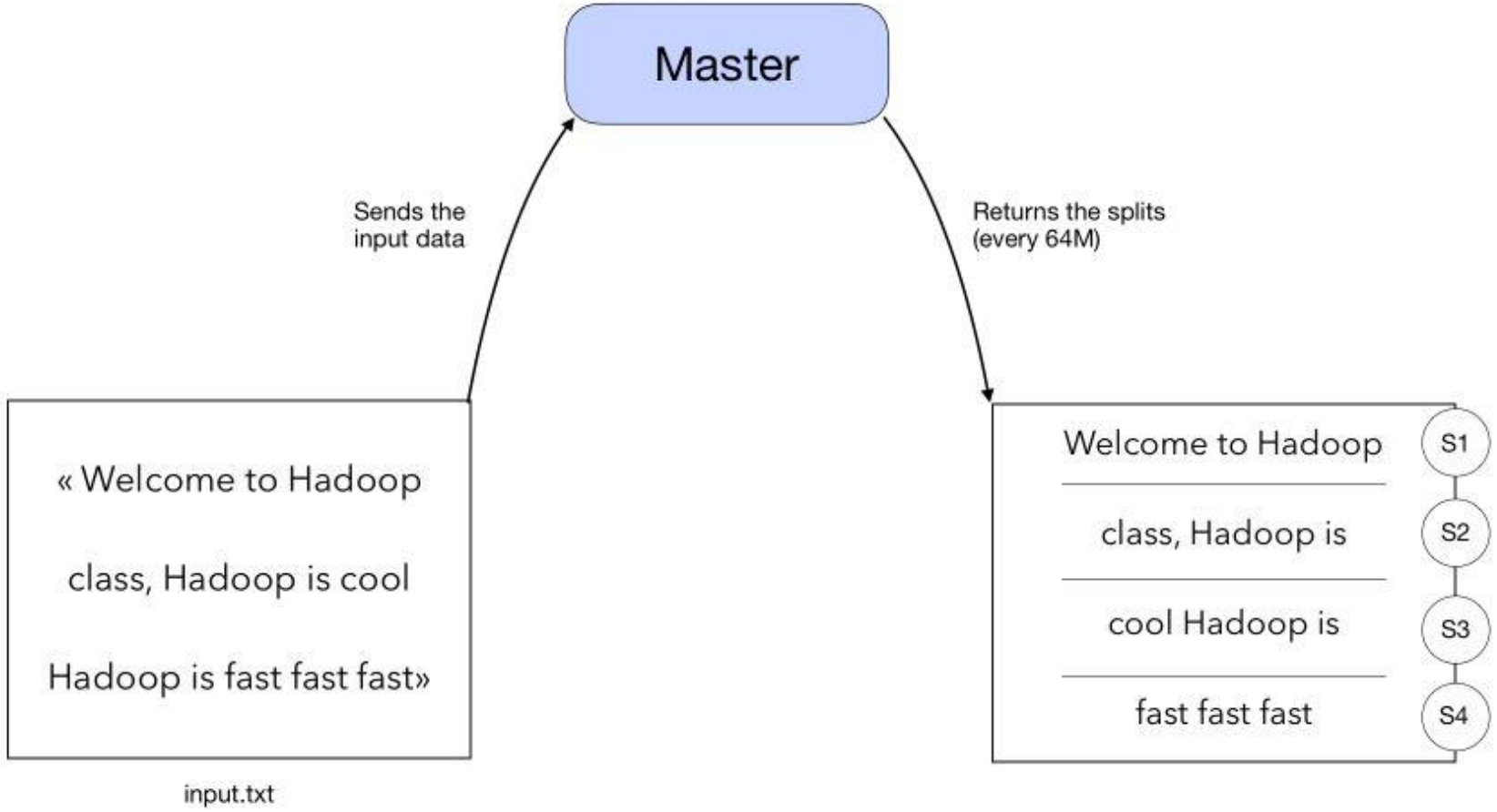
In the first step, we consider a set of machines which together form a cluster. We need to define which machine/node will become a Master, and which ones will become Workers.



The role of the Master is to :

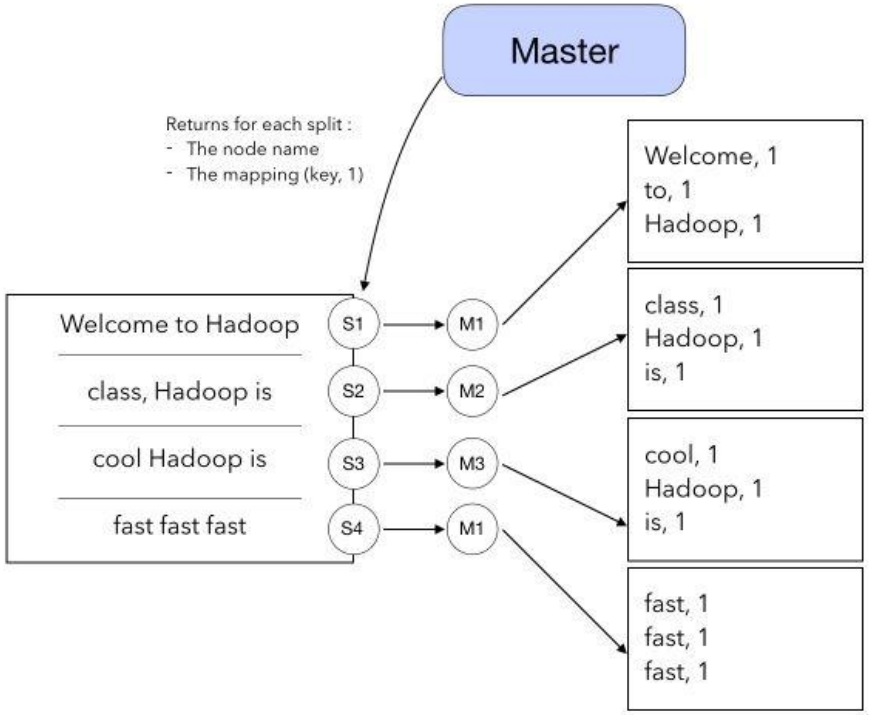
- ❑ split the input data on different machines
- ❑ remember which machine handles which part of the data
- ❑ handle duplicates to avoid loss of information if a worker fails
- ❑ aggregate the outputs and sort the final list

WordCount Problem – Split the input data



The input data is split into blocks of 64Mb. A text file might use a whole 64Mb data node inside a cluster, so it is preferable to have rather big files. Optimization tools are provided by Hadoop.

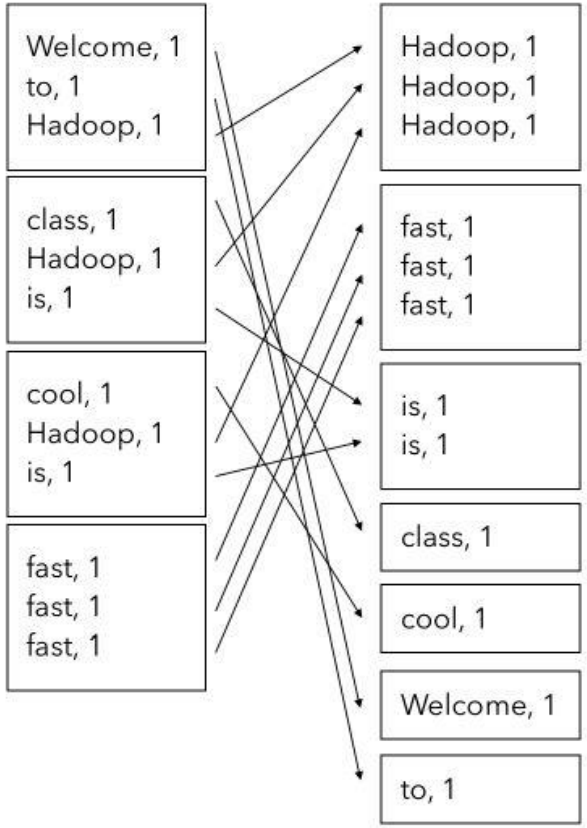
WordCount Problem – Map



In the Map, the Master returns for each split :

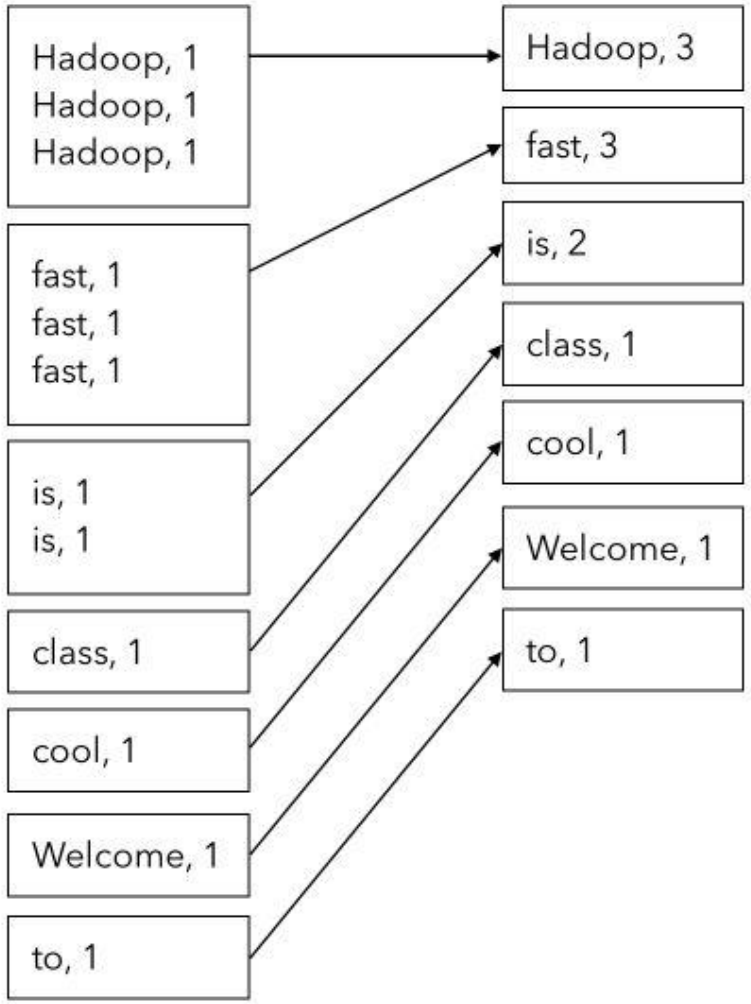
- ❑ the node name to which the split should be sent
- ❑ the mapping (Key, value) where value is equal to 1 for every word of the split

WordCount Problem – Shuffle

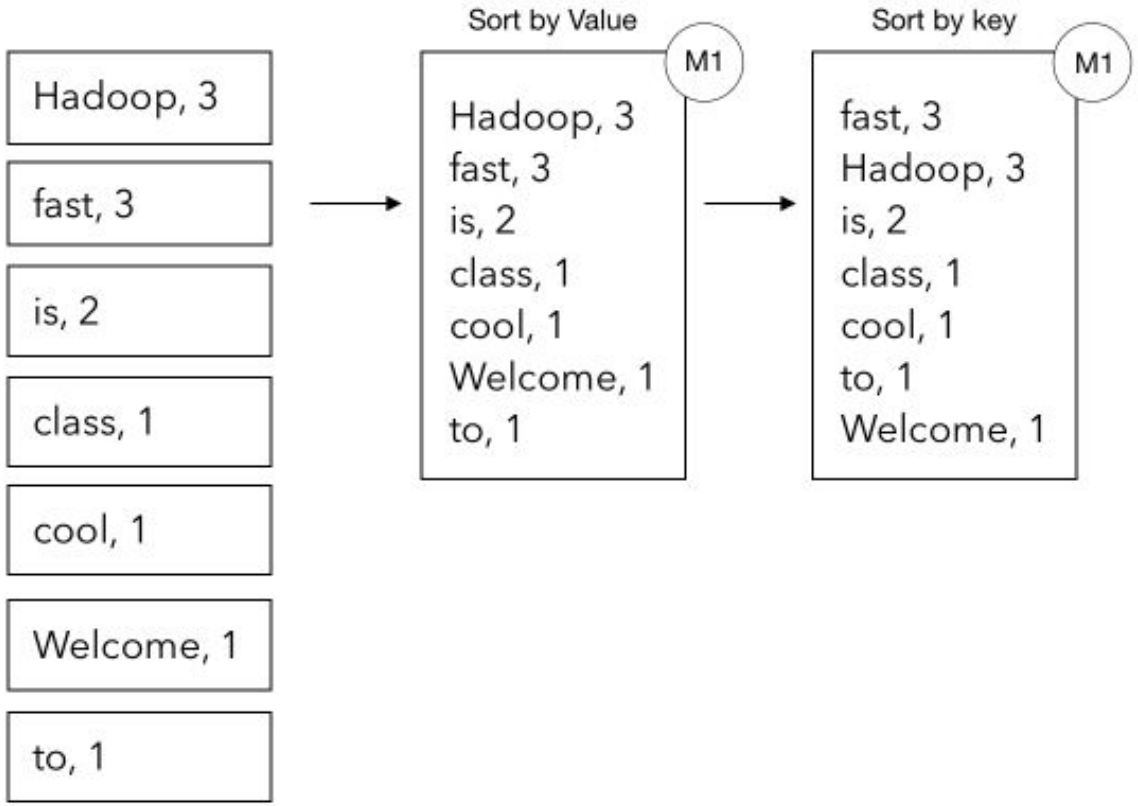


In the Shuffle step, the Map-Reduce algorithm groups the words by similarity (group a dictionary by key). It is called Shuffle because the initial splits are no longer used.

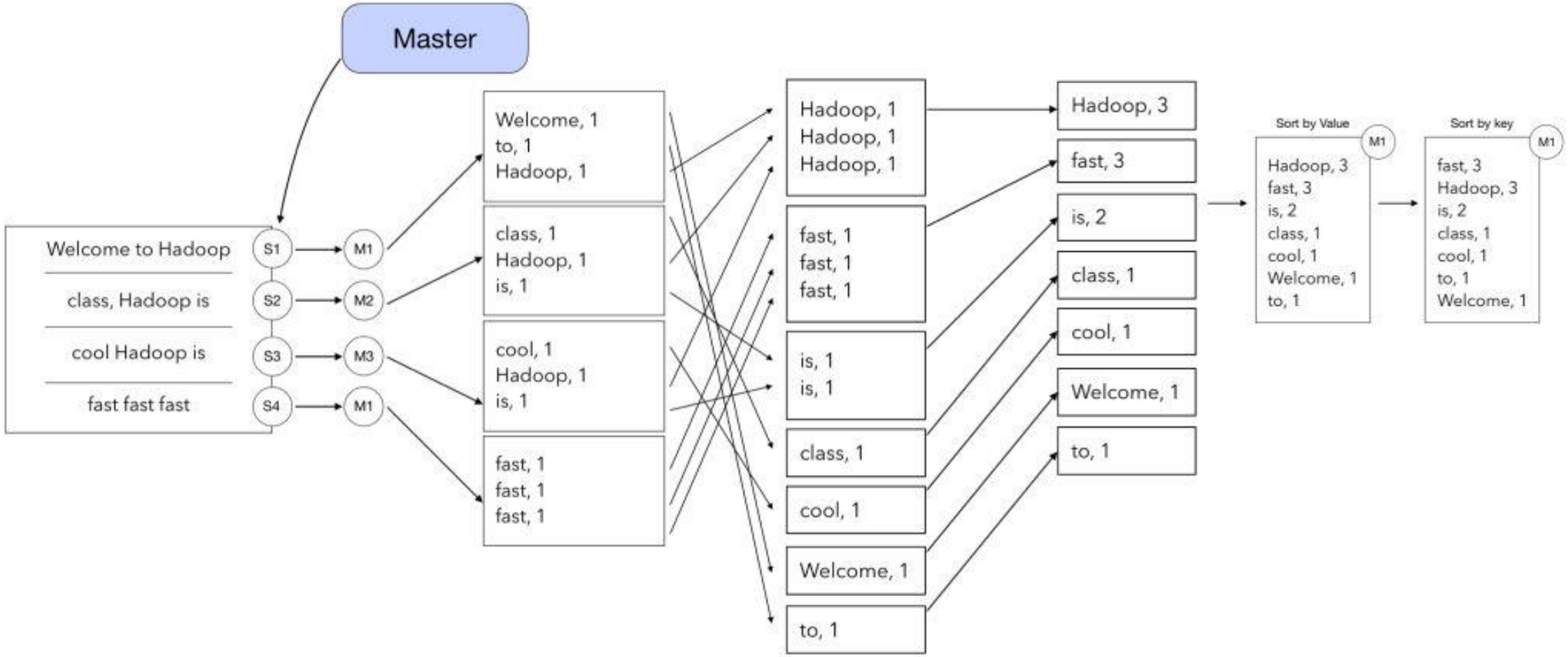
WordCount Problem – Reduce



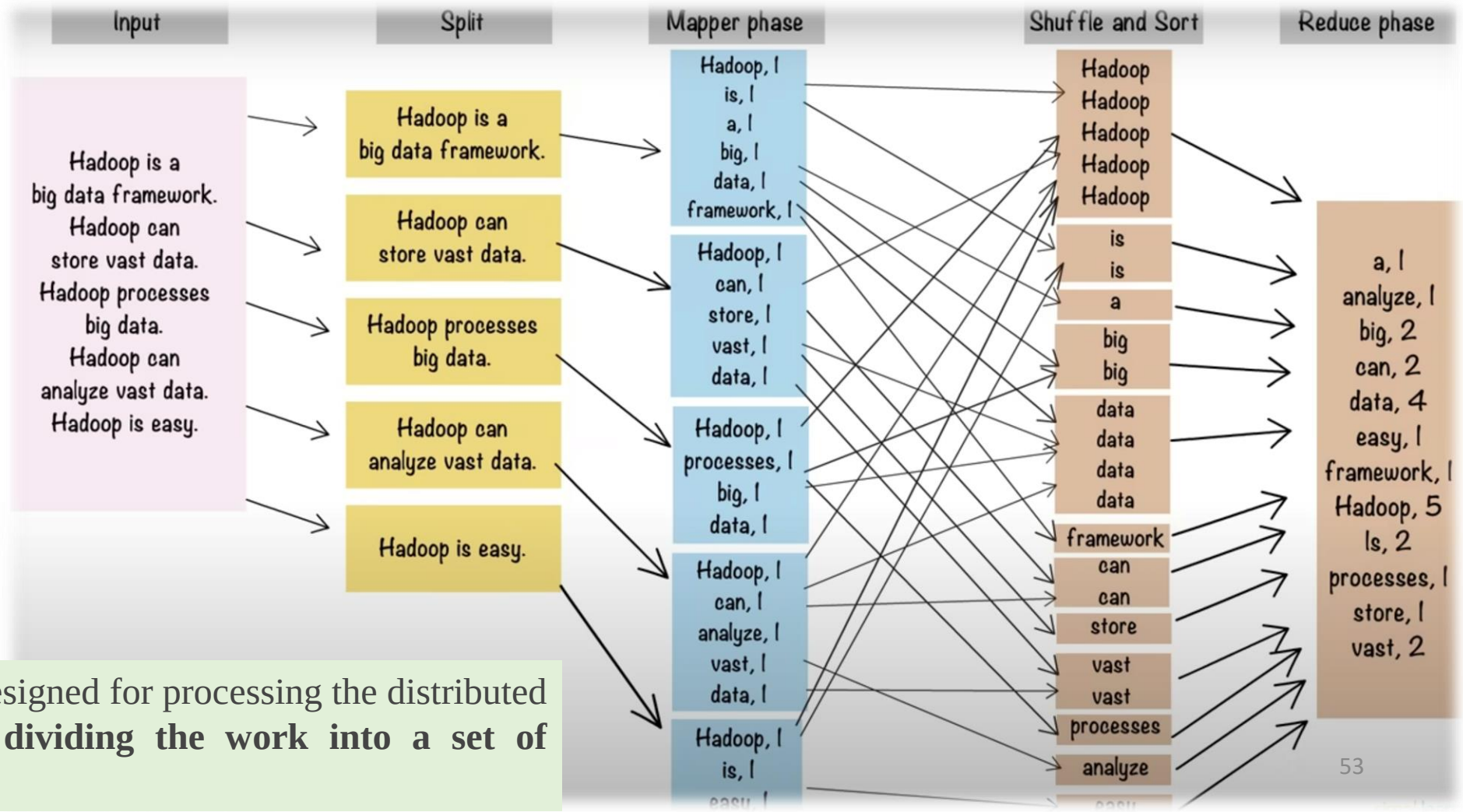
WordCount Problem – Sorting



The Big Picture

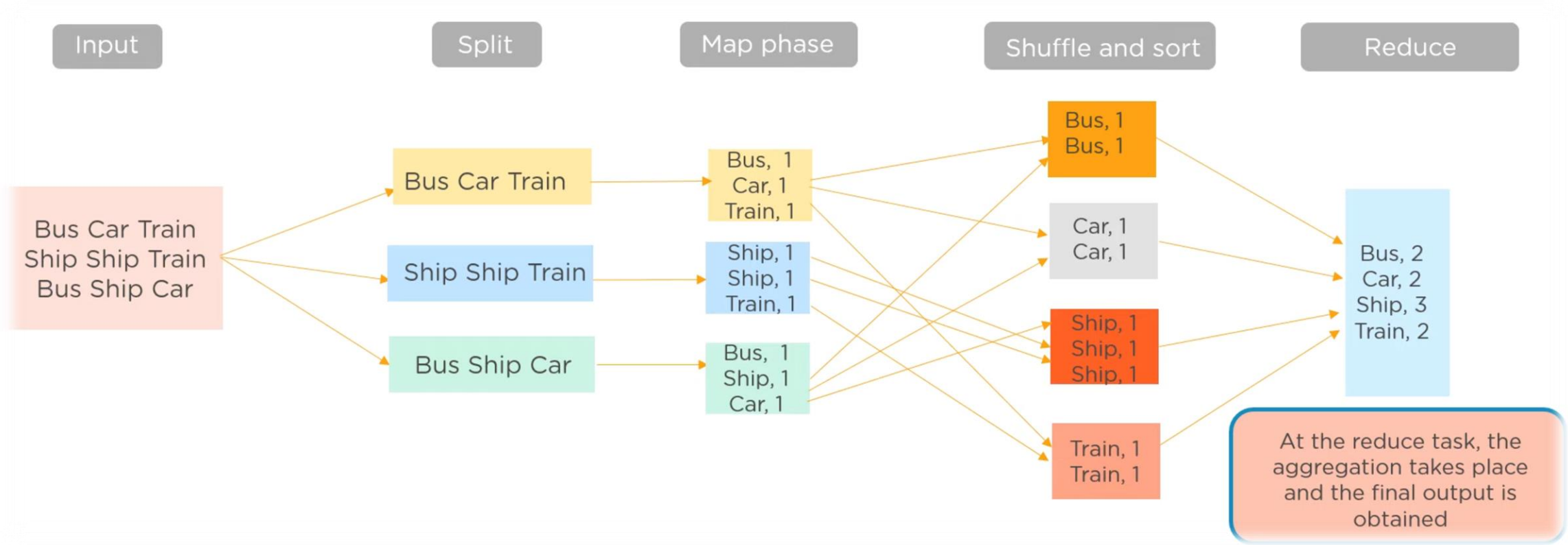


The Big Picture

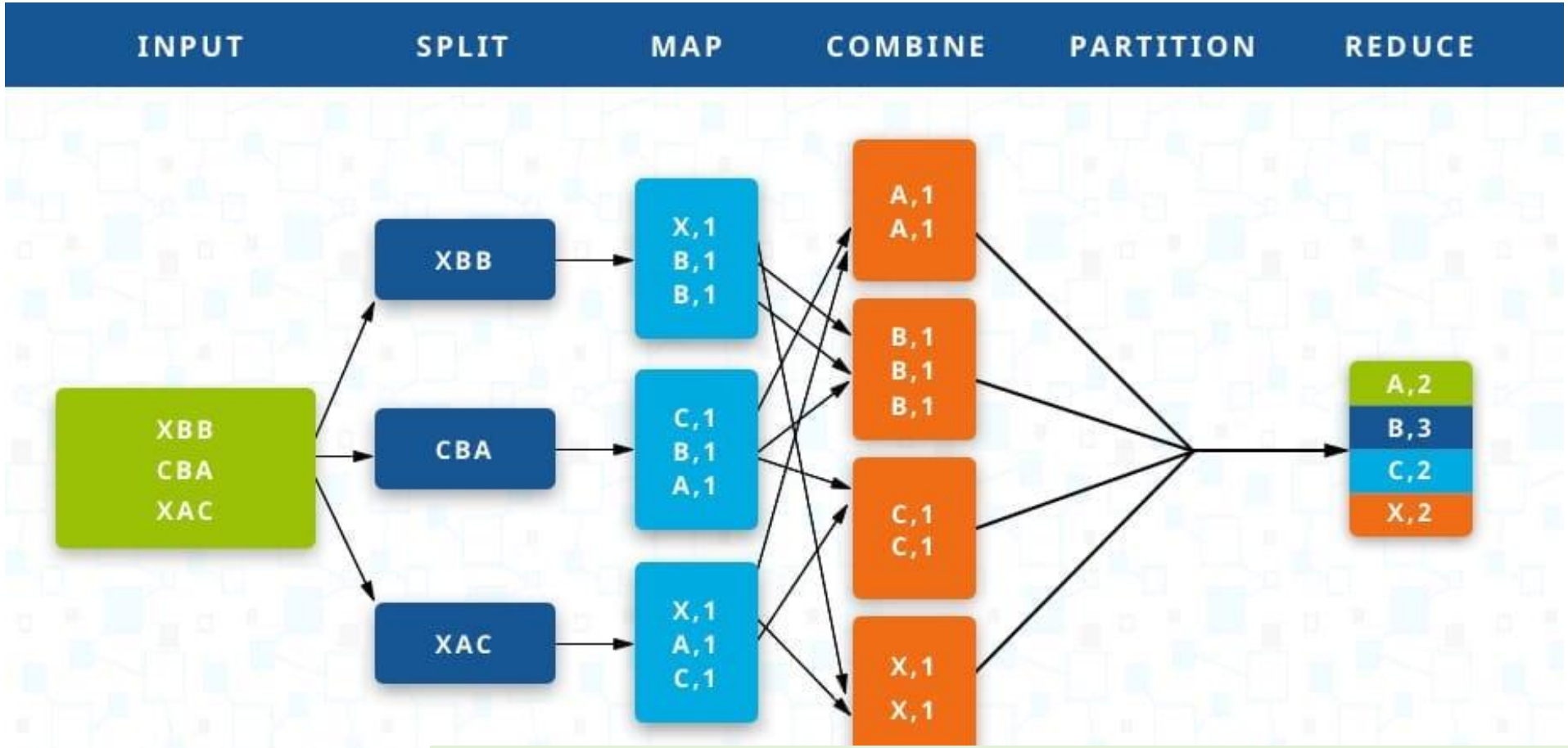


It is a programming model designed for processing the distributed data stored in parallel, by **dividing the work into a set of independent tasks**.

MapReduce

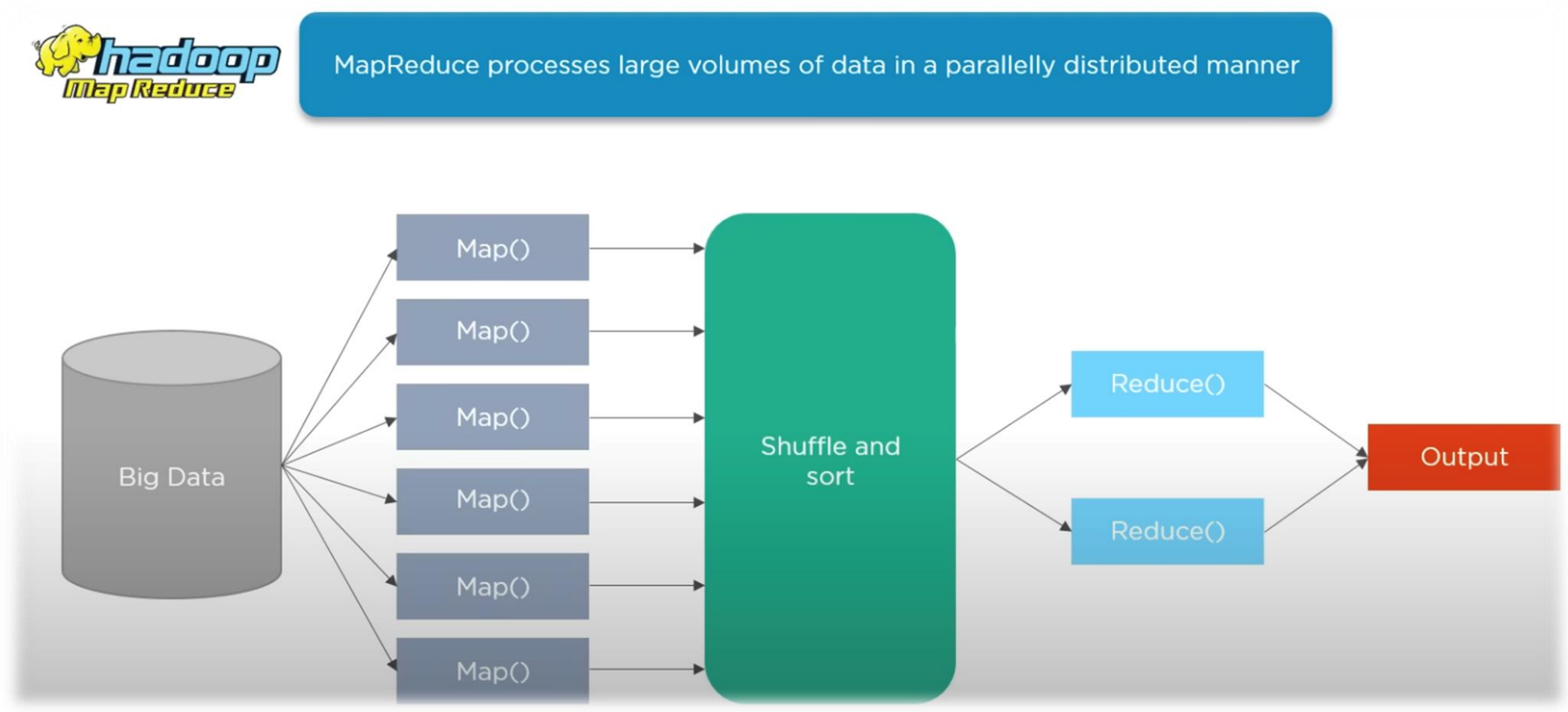

Input
Split
Map phase
Shuffle and sort
Reduce
Bus Car Train
Ship Ship Train
Bus Ship Car
Bus Car Train
Ship Ship Train
Bus Ship Car
Bus, 1
Car, 1
Train, 1
Ship, 1
Ship, 1
Train, 1
Bus, 1
Ship, 1
Car, 1
Bus, 1
Bus, 1
Car, 1
Car, 1
Ship, 1
Ship, 1
Ship, 1
Train, 1
Train, 1
Bus, 2
Car, 2
Ship, 3
Train, 2
At the reduce task, the aggregation takes place and the final output is obtained

MapReduce



It is a programming model designed for processing the distributed data stored in parallel, by **dividing the work into a set of independent tasks**.

MapReduce



It is a programming model designed for processing the distributed data stored in parallel, by **dividing the work into a set of independent tasks**.

How does Hadoop work?

The input data is divided into uniformly-sized blocks of 128Mb or 64Mb

Each file is distributed to a given cluster node, and even to several cluster nodes to handle failure of a node.

A Master node keeps track of where each file is sent.

HDFS is plugged on top of the local file system to supervise the processing.

Hadoop performs a *sort* between the map and reduces stages.

It then sends the sorted data to a given machine and displays the result.

MapReduce: Real-world example

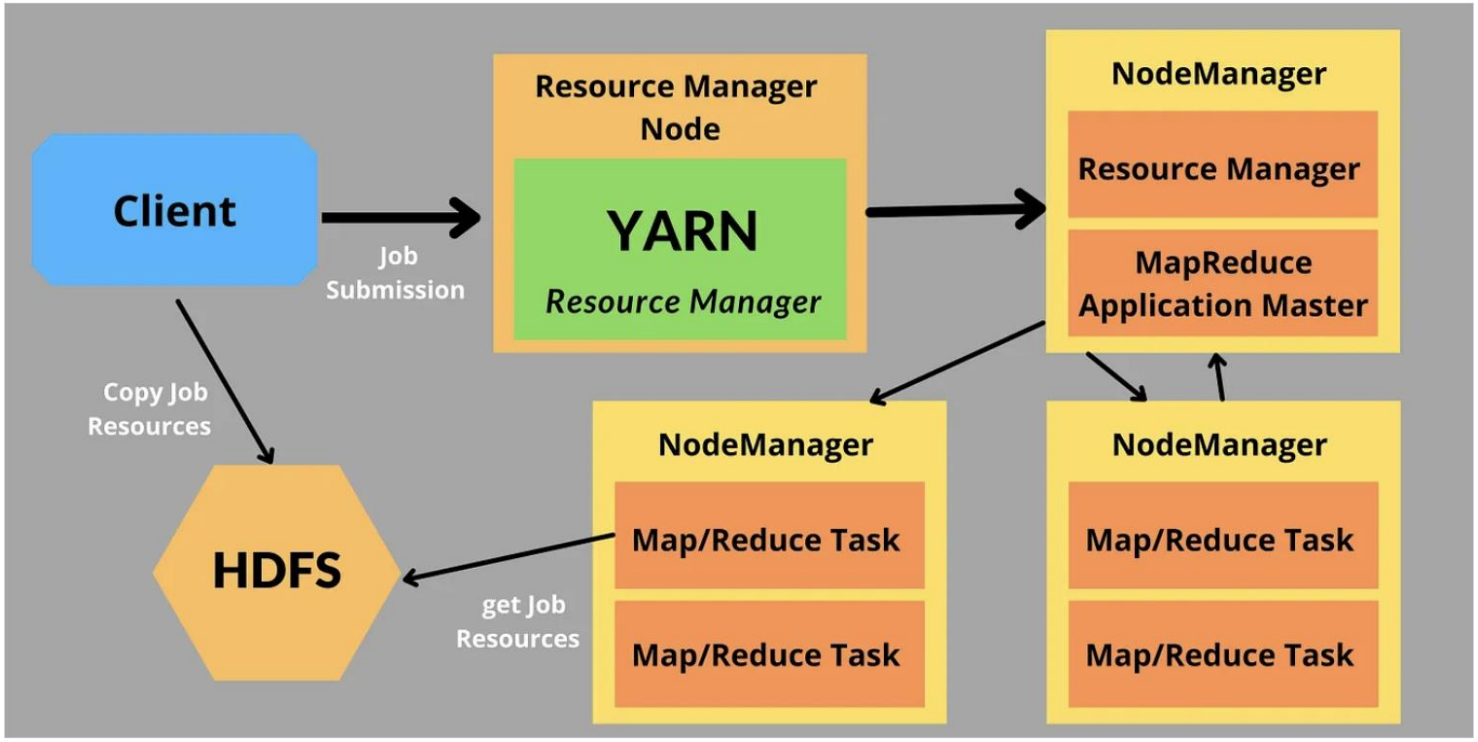
Hotel_Reviews

Review	Rating
nice hotel expensive parking got good deal stay hotel anniversary, arrived late evening took advice previous reviews did valet parking, check quick easy, little disappointed non-existent view room room clean ni	4
ok nothing special charge diamond member hilton decided chain shot 20th anniversary seattle, start booked suite paid extra website description not, suite bedroom bathroom standard hotel room, took printed	2
nice rooms not 4* experience hotel monaco seattle good hotel n't 4* level.positives large bathroom mediterranean suite comfortable bed pillowsattentive housekeeping staffnegatives ac unit malfunctioned stay	3
unique, great stay, wonderful time hotel monaco, location excellent short stroll main downtown shopping area, pet friendly room showed no signs animal hair smells, monaco suite sleeping area big striped curt	5
great stay great stay, went seahawk game awesome, downfall view building did n't complain, room huge staff helpful, booked hotels website seahawk package, no charge parking got voucher taxi, problem taxi	5
love monaco staff husband stayed hotel crazy weekend attending memorial service best friend husband celebrating 12th wedding anniversary, talk mixed emotions, booked suite hotel monte carlos, loaned bea	5
cozy stay rainy city, husband spent 7 nights monaco early january 2008. business trip chance come ride.we booked monte carlo suite proved comfortable longish stay, room 905 located street building, street no	5
excellent staff, housekeeping quality hotel chocked staff make feel home, experienced exceptional service desk staff concierge door men maid service needs work, maid failed tuck sheets foot bed instance soi	4
hotel stayed hotel monaco cruise, rooms generous decorated uniquely, hotel remodeled pacific bell building charm sturdiness, everytime walked bell men felt like coming home, secure, great single travelers, lo	5
excellent stayed hotel monaco past w/e delight, reception staff friendly professional room smart comfortable bed, particularly liked reception small dog received staff guests spoke loved, mild negative distance	5
poor value stayed monaco seattle july, nice hotel priced 100- 150 night not, hotel takes beating quotient, experience simply average, nothing exceptional paying 300+ n't ca n't terribly disappointed, wife stayed r	2
nice value seattle stayed 4 nights late 2007. looked comparable hilton marriott westin area points/miles n't gave monaco shot, pleasantly surprised nice room service quick tasty bed especially comfortable unlik	4
nice hotel good location hotel kimpton design whimsical vibe fun, staff young casual problem hotel busy stay friendly helpful, group reserved rooms gave connecting rooms fuss, not busy week.the rooms decer	4
nice hotel not nice staff hotel lovely staff quite rude, bellhop desk clerk going way make things difficult, waited forever check heavy bags no help getting throught double doors room, worst desk clerk checking	3
great hotel night quick business trip, loved little touches like goldfish leopard print robe, complaint wifi complimentary not internet access business center, great location library service fabulous,	4
horrible customer service hotel stay february 3rd 4th 2007my friend picked hotel monaco appealing website online package included champagne late checkout 3 free valet gift spa weekend, friend checked roor	1
disappointed say anticipating stay hotel monaco based reviews seen tripadvisor, definitely disppointment, decor room hotel envisioned nice, housekeeping staff impressive extremely polite cheery helpful, desk	2
fantastic stay monaco seattle hotel monaco holds high standards kimpton hotel line, having stayed kimpton hotels cities easily say seattle hotel monaco best seen, service attentive prompt, based member kimp	5
good choice hotel recommended sister, great location room nice, comfortable bed- quiet- staff helpful recommendations restaurants, pike market 4 block walk, stay,	5
hmmmmm say really high hopes hotel monaco chose base girlfriend shopping trip seattle, stay say given competition seattle just okay, hotel lot nice features little things detract bedding super soft luxurious co	3
service service service spent week g-friend labor day bumbershoot, gray line airporter drops corner hotel 10 person cab 28 total make sure flat rate town car 38. location central downtown street w. it__Ç_é_ qui	5
excellent stay, delightful surprise stay monaco, thoroughly enjoyed stay, room comfortable lovely amenities friendly staff, especially enjoyed hour indulgence, definitely come,	5

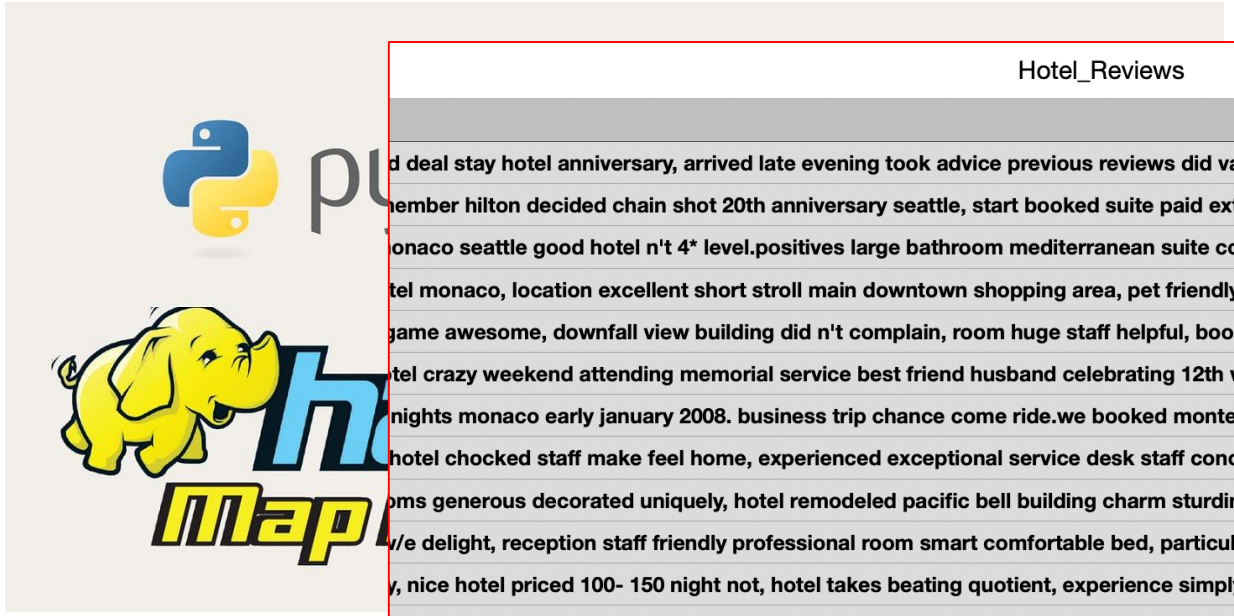
You have a dataset that consists of hotel reviews and ratings. Now you need to find out how many reviews each rating.

MapReduce: Real-world example

So what you will do is you will go through the Rating columns and count how many reviews are there for 1,2,3,4,5. It is pretty easy to do if we do have a small amount of data but when it comes to big data it can be billions or trillions of data. Therefore we can use MapReduce.



MapReduce with Python



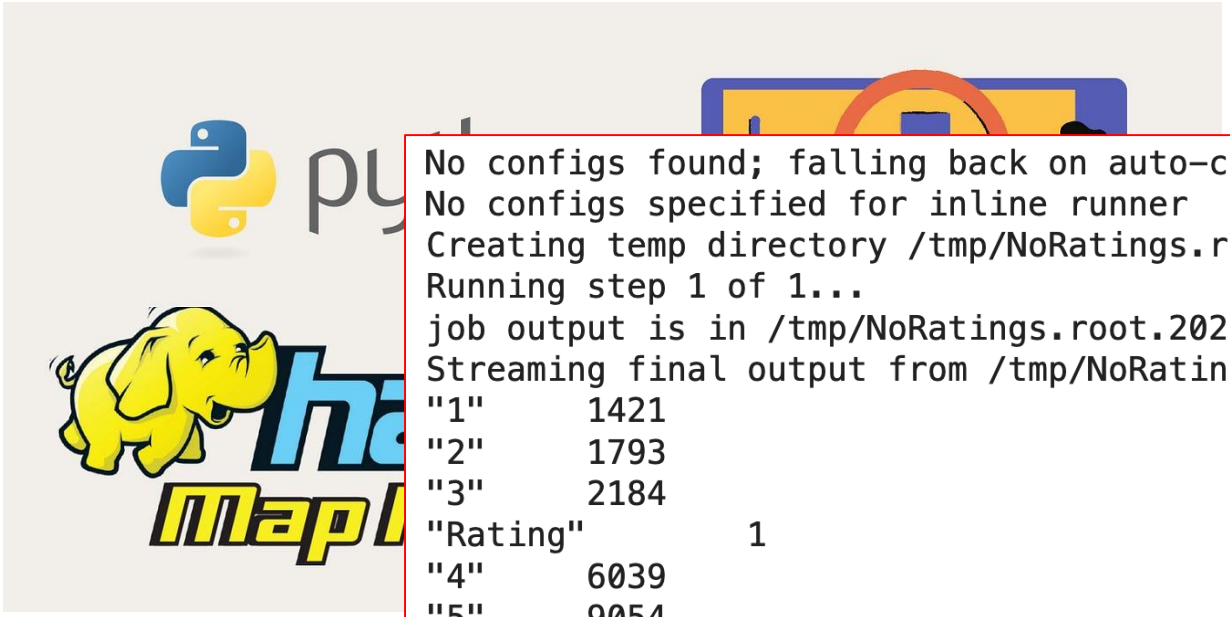
Hotel_Reviews

	Rating
d deal stay hotel anniversary, arrived late evening took advice previous reviews did valet parking, check quick easy, little disappointed non-existent view room room clean ni	4
ember hilton decided chain shot 20th anniversary seattle, start booked suite paid extra website description not, suite bedroom bathroom standard hotel room, took printed	2
onaco seattle good hotel n't 4* level.positives large bathroom mediterranean suite comfortable bed pillowsattentive housekeeping staffnegatives ac unit malfunctioned stay	3
tel monaco, location excellent short stroll main downtown shopping area, pet friendly room showed no signs animal hair smells, monaco suite sleeping area big striped curt	5
game awesome, downfall view building did n't complain, room huge staff helpful, booked hotels website seahawk package, no charge parking got voucher taxi, problem taxi	5
tel crazy weekend attending memorial service best friend husband celebrating 12th wedding anniversary, talk mixed emotions, booked suite hotel monte carlos, loaned bea	5
nights monaco early january 2008. business trip chance come ride.we booked monte carlo suite proved comfortable longish stay, room 905 located street building, street no	5
hotel chocked staff make feel home, experienced exceptional service desk staff concierge door men maid service needs work, maid failed tuck sheets foot bed instance soi	4
oms generous decorated uniquely, hotel remodeled pacific bell building charm sturdiness, everytime walked bell men felt like coming home, secure, great single travelers, lo	5
/e delight, reception staff friendly professional room smart comfortable bed, particularly liked reception small dog received staff guests spoke loved, mild negative distance	5
, nice hotel priced 100- 150 night not, hotel takes beating quotient, experience simply average, nothing exceptional paying 300+ n't ca n't terribly disappointed, wife stayed r	2
2007. looked comparable hilton marriott westin area points/miles n't gave monaco shot, pleasantly surprised nice room service quick tasty bed especially comfortable unlik	4
n design whimsical vibe fun, staff young casual problem hotel busy stay friendly helpful, group reserved rooms gave connecting rooms fuss, not busy week.the rooms decer	4
taff quite rude, bellhop desk clerk going way make things difficult, waited forever check heavy bags no help getting throught double doors room, worst desk clerk checking	3
loved little touches like goldfish leopard print robe, complaint wifi complimentary not internet access business center, great location library service fabulous,	4
ebruary 3rd 4th 2007my friend picked hotel monaco appealing website online package included champagne late checkout 3 free valet gift spa weekend, friend checked roor	1
tel monaco based reviews seen tripadvisor, definitely disppointment, decor room hotel envisioned nice, housekeeping staff impressive extremely polite cheery helpful, desk	2

[MapReduce](#) will t
data into key/value
input, and **aggregating** data together by **Reduce**.
MapReduce will deal with all your cluster failures.

```
columns = 'Review,Rating'.split(',')
                                                    atings,
                                                    t_ratings)
                                                    ues):
        yield key, sum(values)
if __name__ == "__main__":
    NoRatings.run()
```

MapReduce with Python

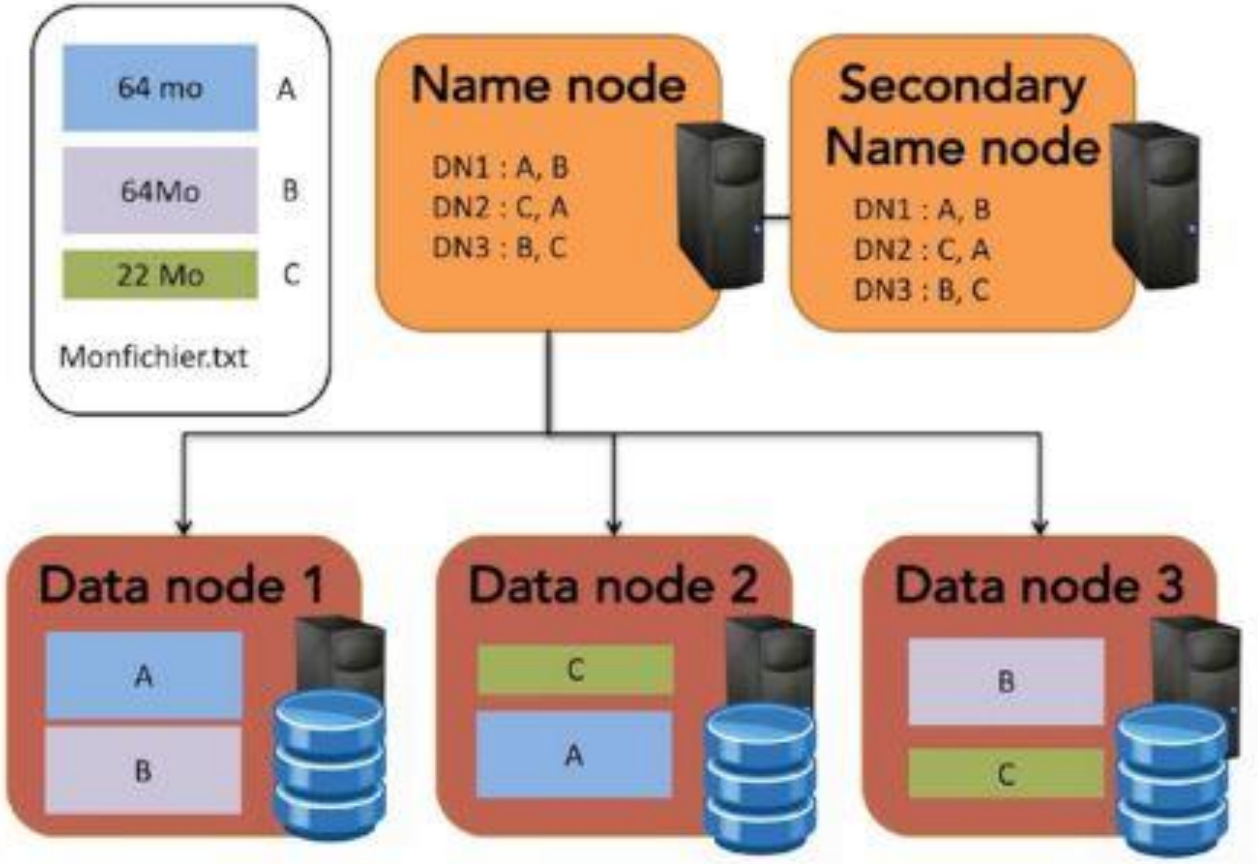


```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/NoRatings.root.20230802.072136.358628
Running step 1 of 1...
job output is in /tmp/NoRatings.root.20230802.072136.358628/output
Streaming final output from /tmp/NoRatings.root.20230802.072136.358628/output...
"1"     1421
"2"     1793
"3"     2184
"Rating"        1
"4"     6039
"5"     9054
Removing temp directory /tmp/NoRatings.root.20230802.072136.358628...
```

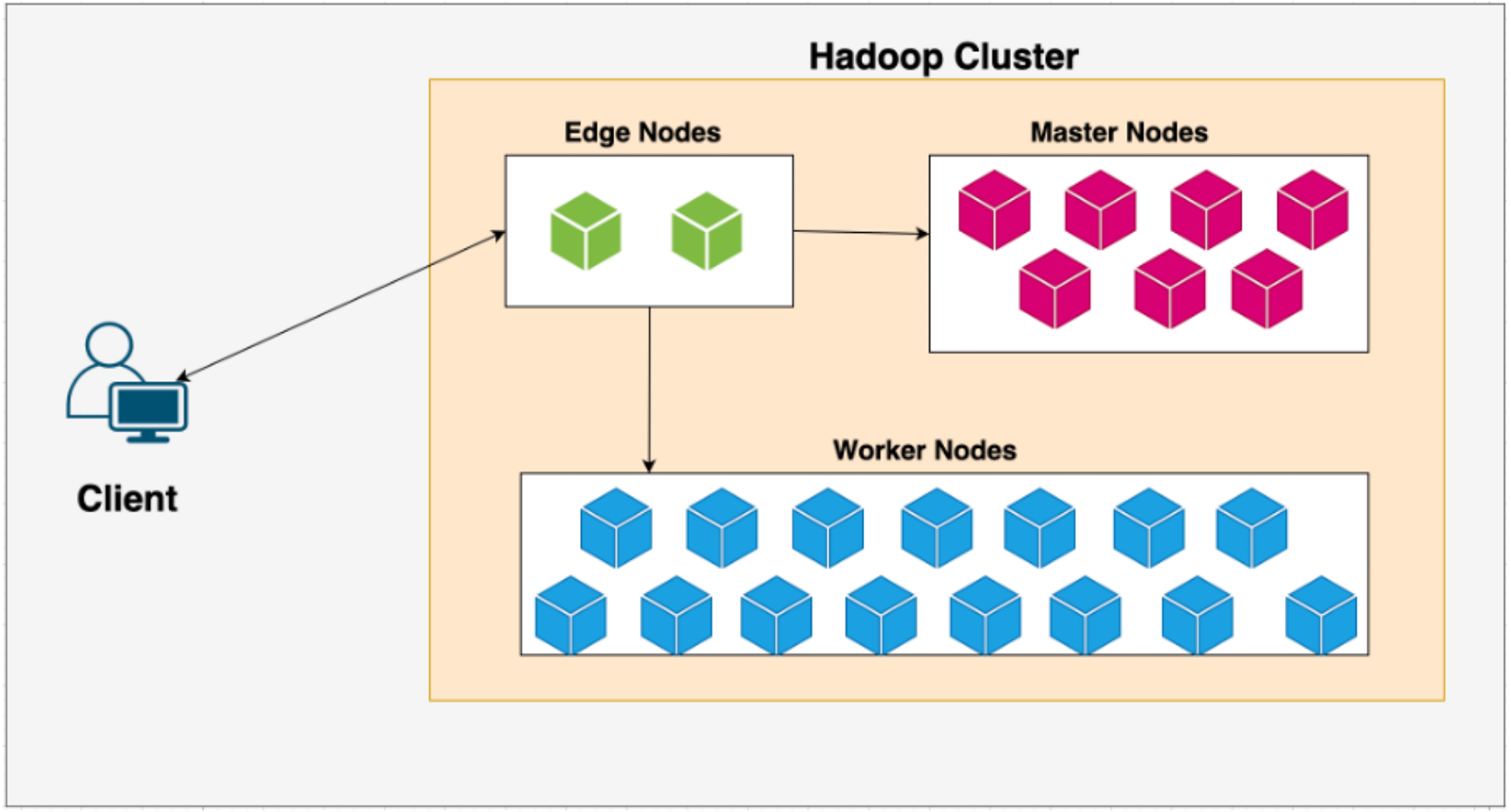
```
columns = 'Review,Rating'.split(',')
class NoRatings(MRJob):
    def steps(self):
                                                    er_get_ratings,
                                                    cer_count_ratings)


                                                    line):


                                                    pairs
            yield ratings, 1
    #Reducer function
    def reducer_count_ratings(self, key, values):
        yield key, sum(values)
if __name__ == "__main__":
    NoRatings.run()
```

MapReduce will **transform** the data using **Map** by dividing data into key/value pairs, getting the output from a map as an input, and **aggregating** data together by **Reduce**. MapReduce will deal with all your cluster failures.

Replication Factor

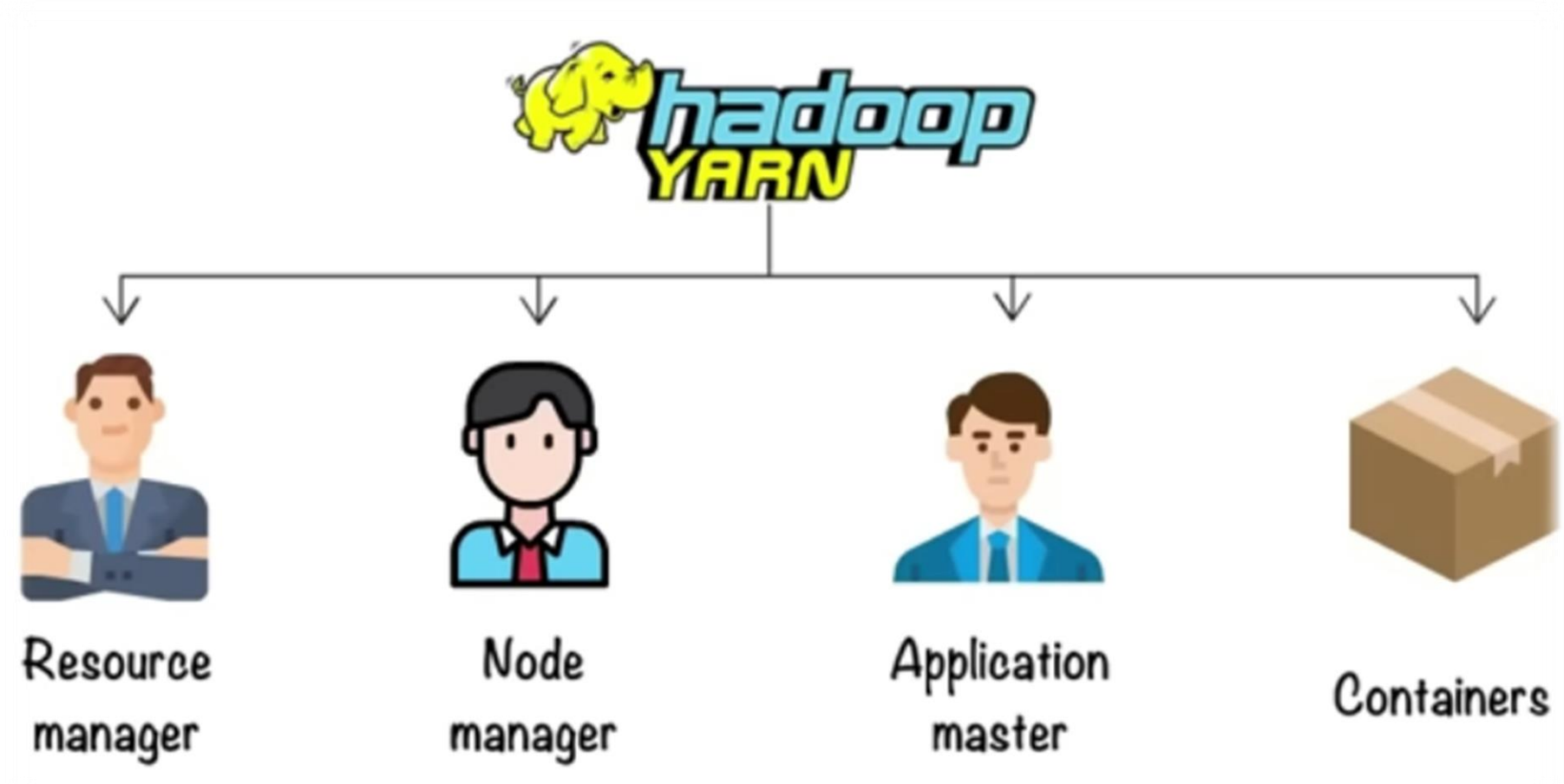


Cluster Types



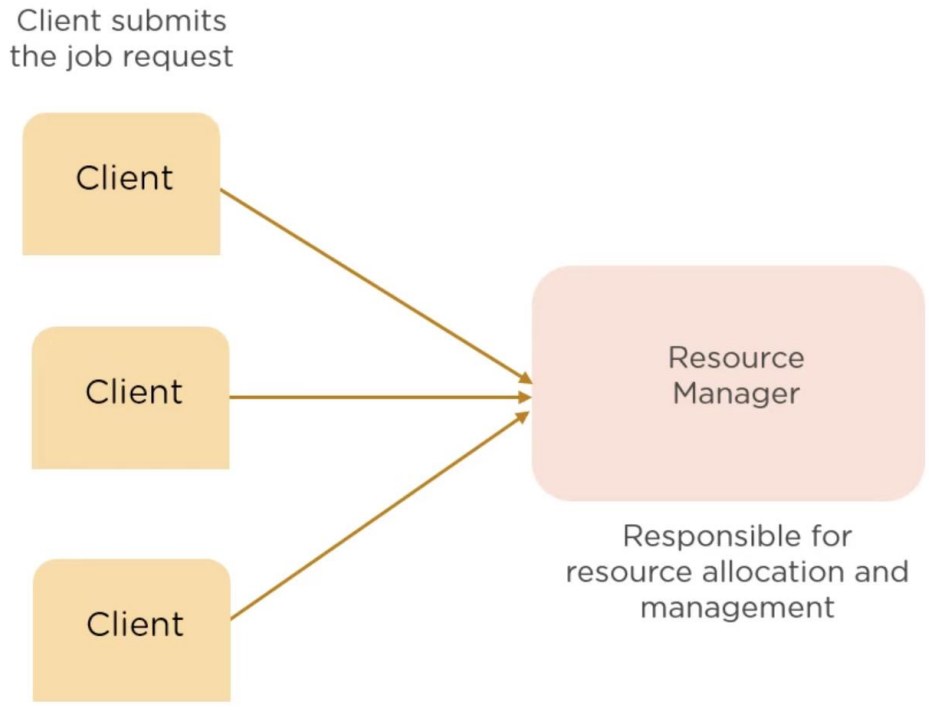
Yarn

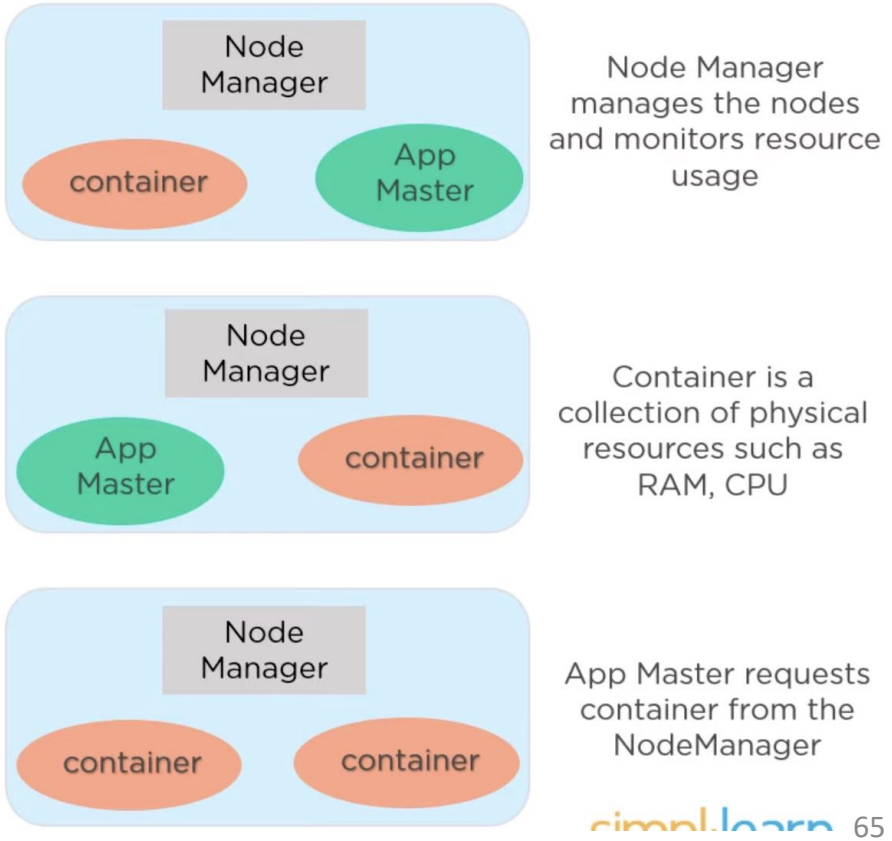
Yet Another Resource Negotiator is responsible for **Resource management** and **Job Schezzduling**.



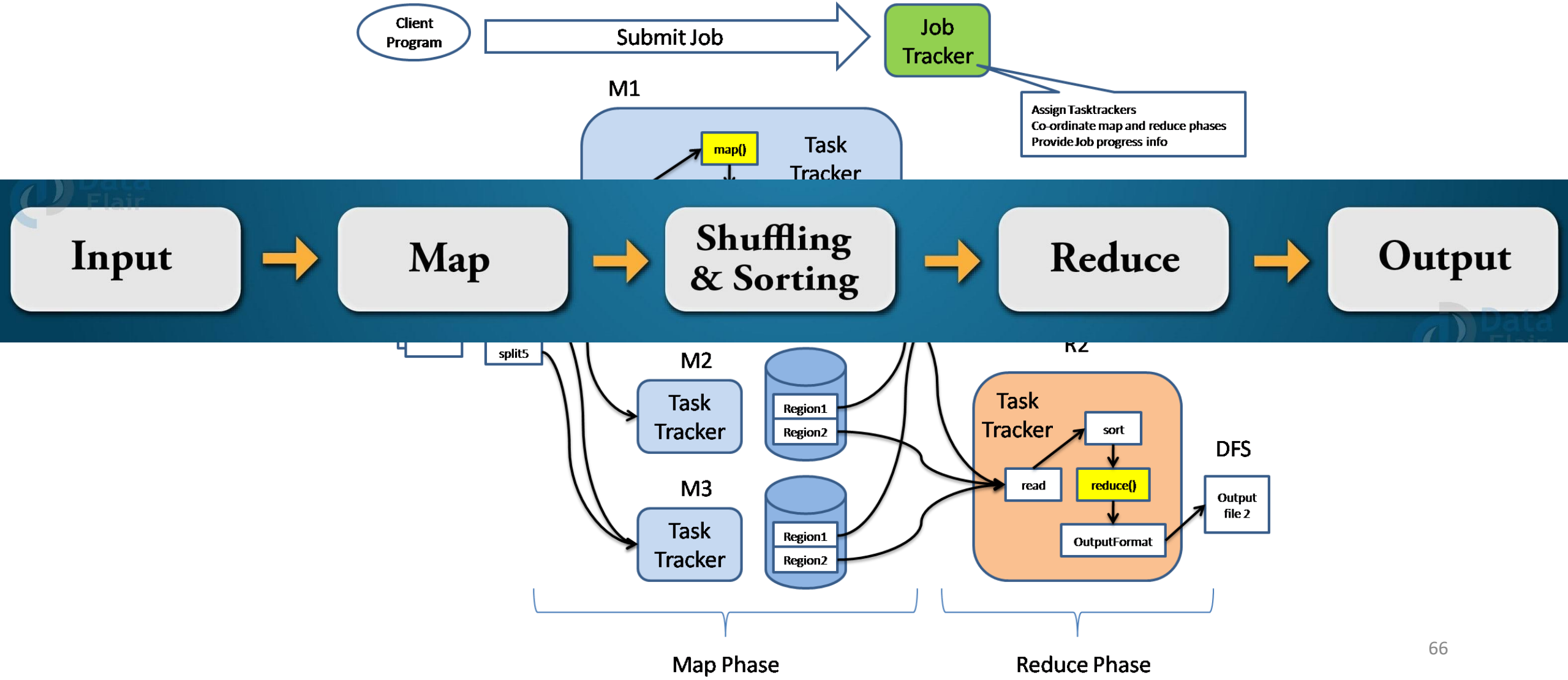
Yarn

Yet Another Resource Negotiator is responsible for **Resource management** and **Job Schezzduling**.

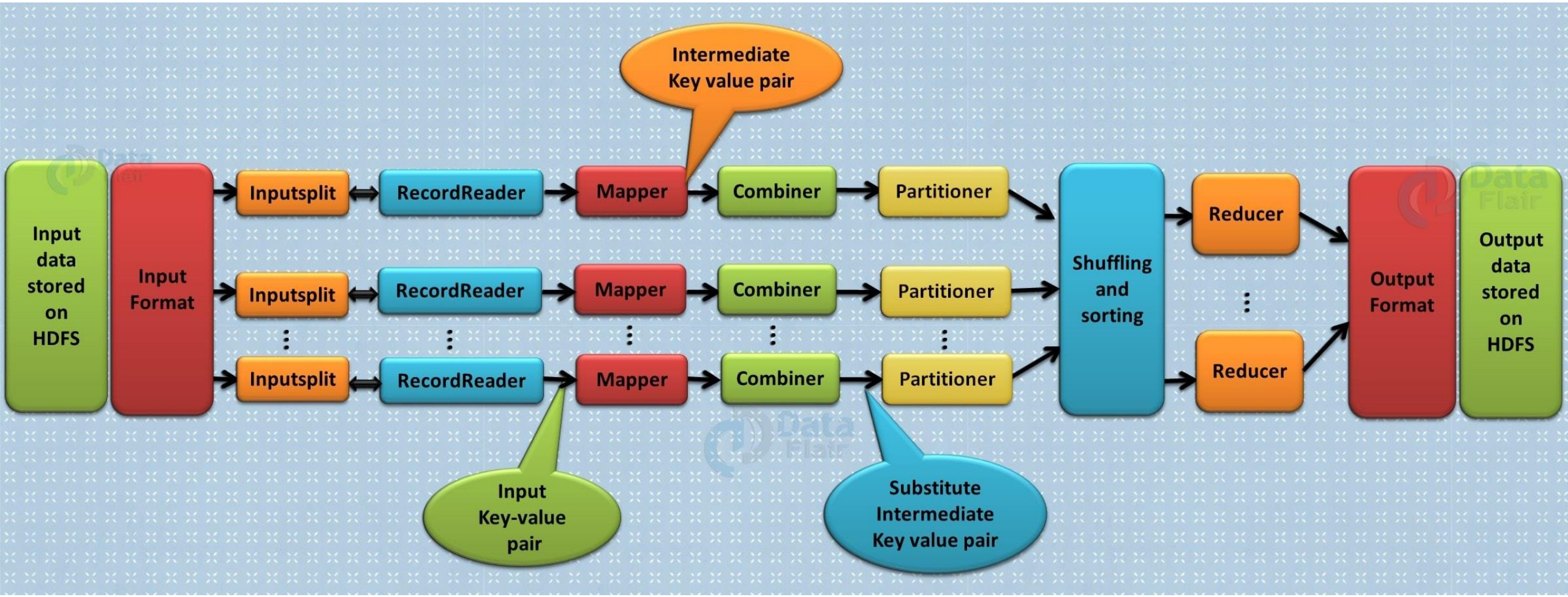




MapReduce architecture



Combiner phase



How to install and use Hadoop?

Using packaged solutions developed by Cloudera, Hortonworks or MapR. Hadoop Distributions pull together all the enhancement projects present in the Apache repository and present them as a unified product so that organizations don't have to spend time on assembling these elements into a single functional component.

Hadoop Distribution	Advantages	Disadvantages
Cloudera Distribution for Hadoop (CDH)	CDH has a user friendly interface with many features and useful tools like Cloudera Impala	CDH is comparatively slower than MapR Hadoop Distribution
MapR Hadoop Distribution	It is one of the fastest hadoop distribution with multi node direct access.	MapR does not have a good interface console as Cloudera
Hortonworks Data Platform (HDP)	It is the only Hadoop Distribution that supports Windows platform.	The Ambari Management interface on HDP is just a basic one and does not have many rich features.

How to install and use Hadoop?

Hortonworks and Cloudera, the two tech giants of Big Data, have merged in October 2018, and are now worth over 3 billions $ combined.

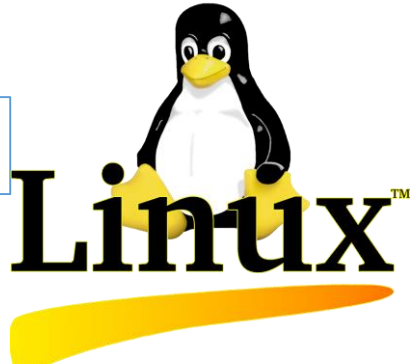
Hadoop Distribution	Advantages	Disadvantages
Cloudera Distribution for Hadoop (CDH)	CDH has a user friendly interface with many features and useful tools like Cloudera Impala	CDH is comparatively slower than MapR Hadoop Distribution
MapR Hadoop Distribution	It is one of the fastest hadoop distribution with multi node direct access.	MapR does not have a good interface console as Cloudera
Hortonworks Data Platform (HDP)	It is the only Hadoop Distribution that supports Windows platform.	The Ambari Management interface on HDP is just a basic one and does not have many rich features.

How to install and use Hadoop?

https://towardsdatascience.com/installing-hadoop-on-a-mac-ec01c67b003c

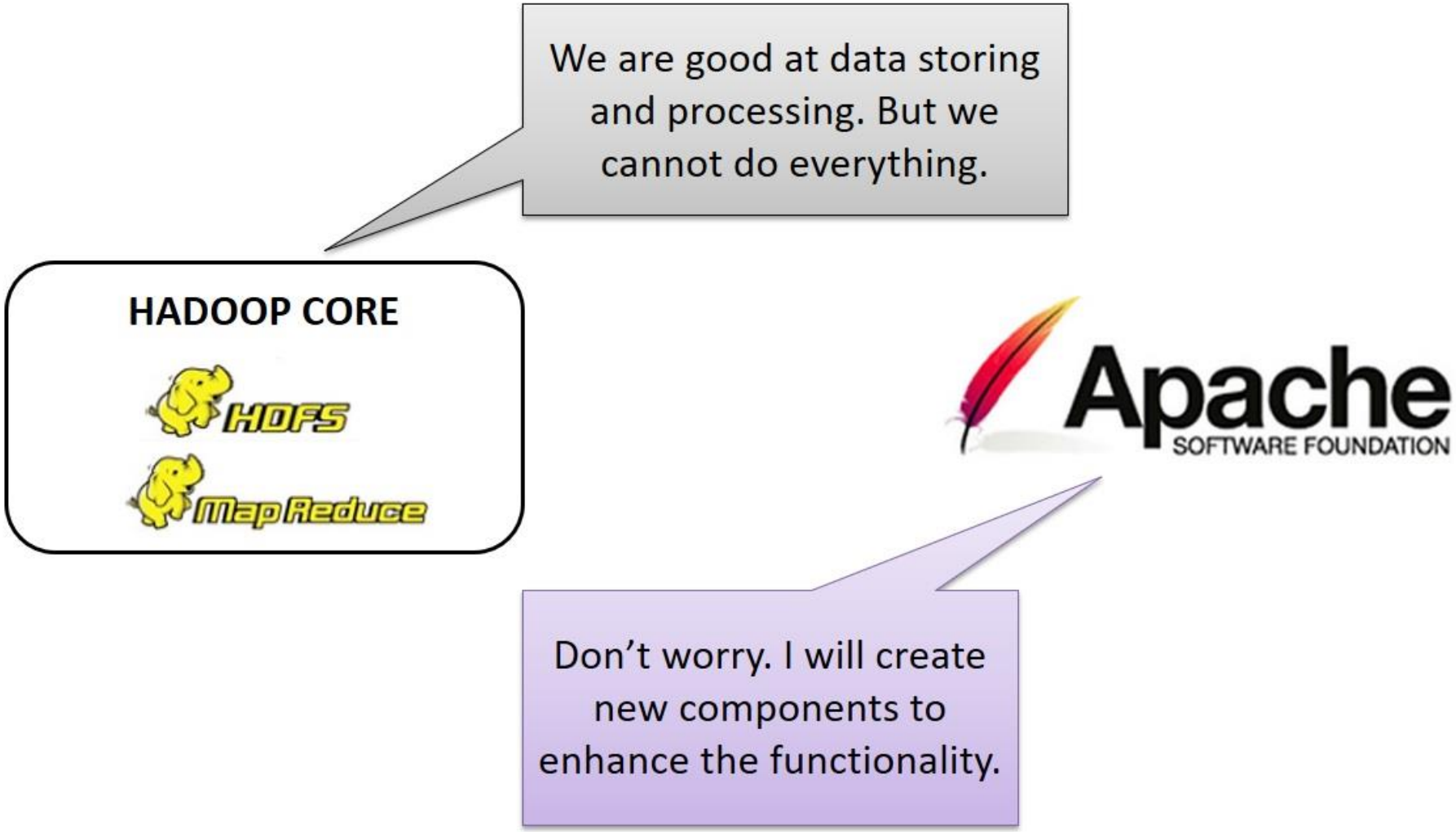
https://kontext.tech/article/377/latest-hadoop-321-installation-on-windows-10-step-by-step-guide

https://intellipaat.com/blog/tutorial/hadoop-tutorial/hadoop-installation/

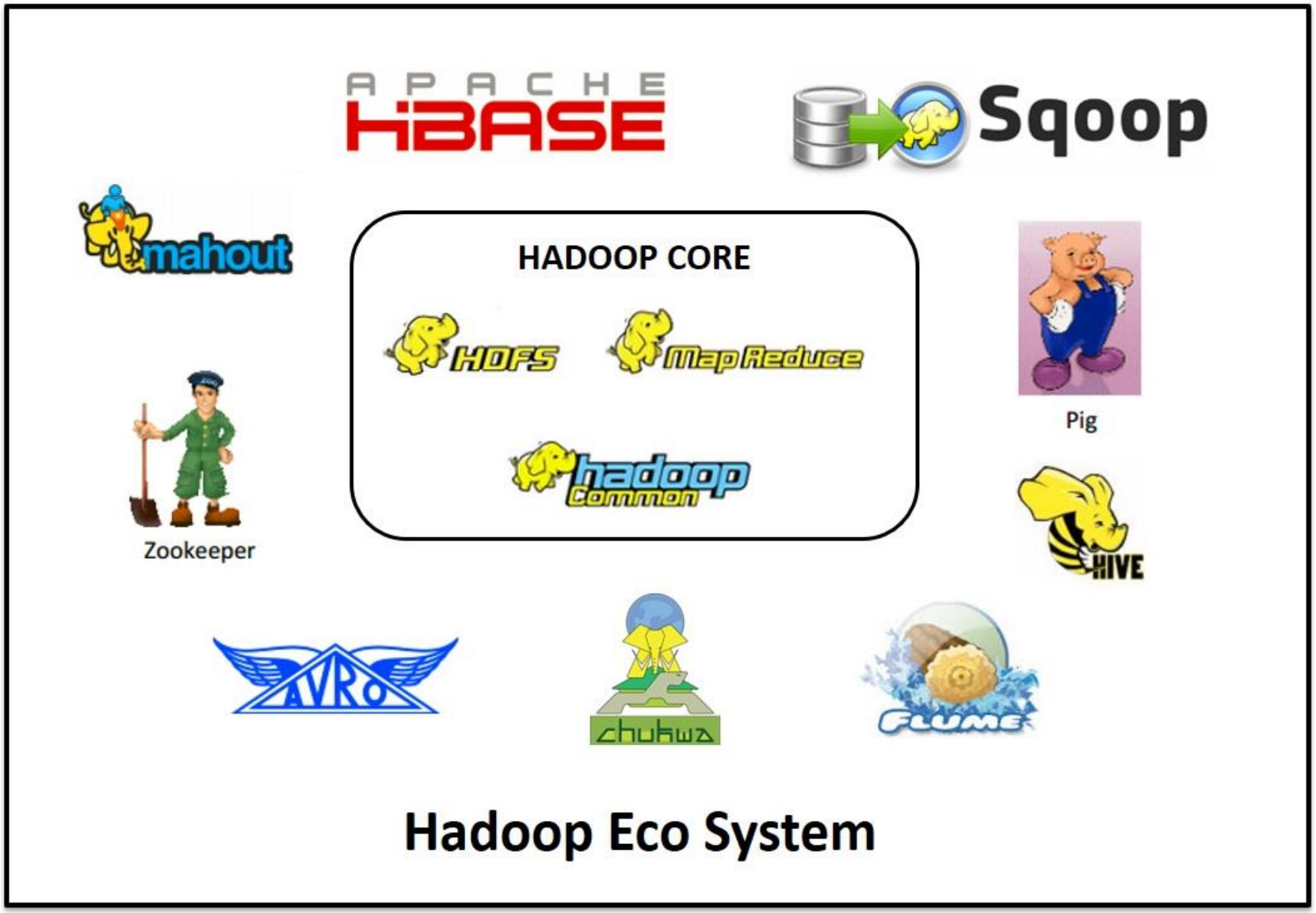



hadoop
GAME CHANGER
GAME CHANGER
GAME CHANGER
IBM
ebay
amazon
There are several applications of Hadoop like

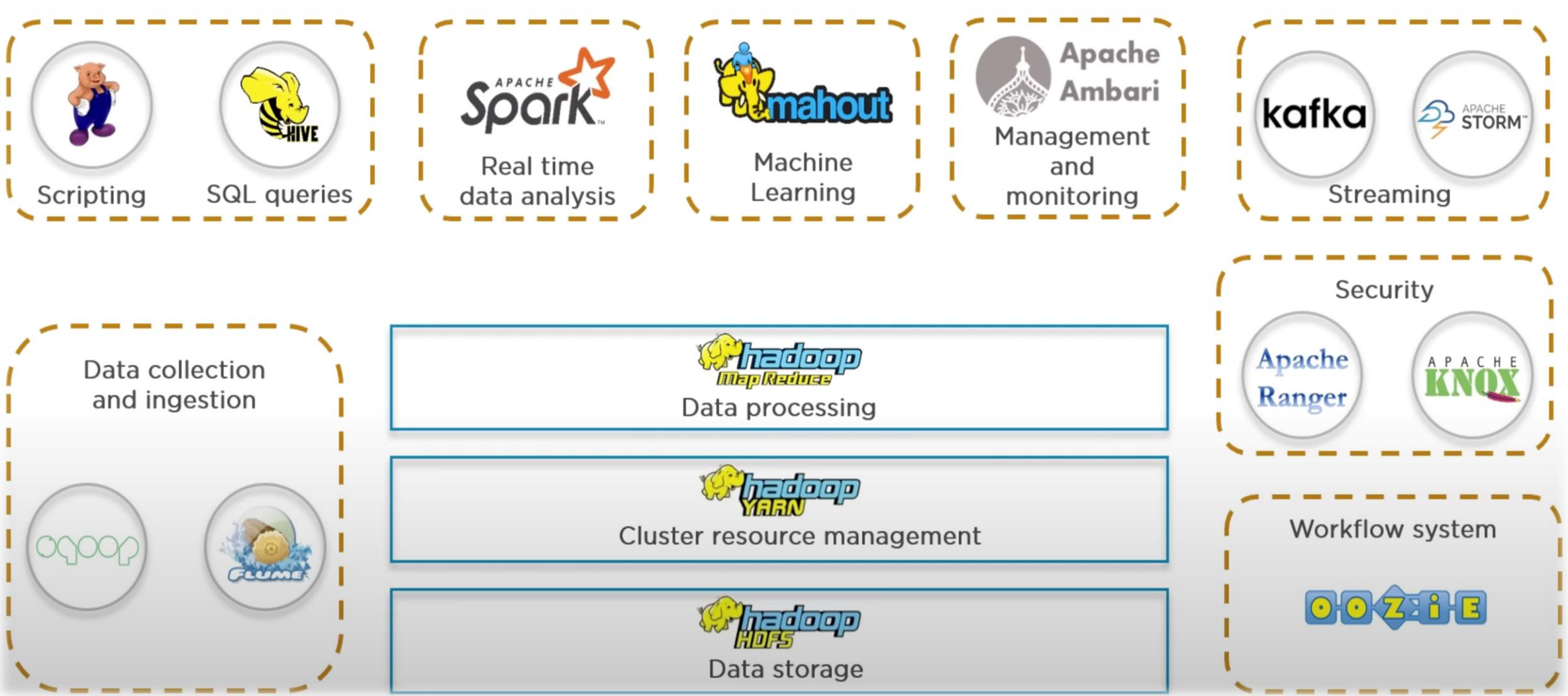
Hadoop Ecosystem



Hadoop Ecosystem



Hadoop Ecosystem



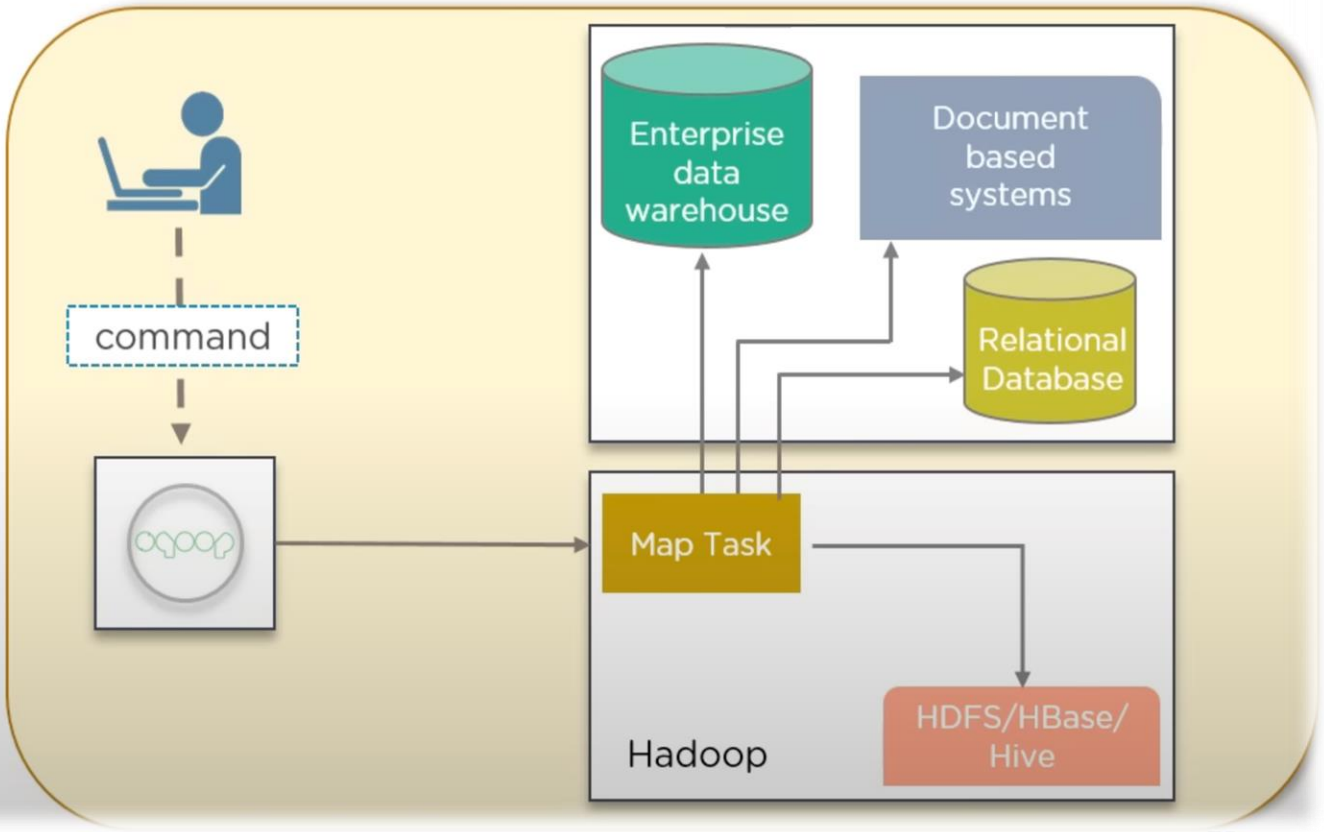
Hadoop Ecosystem

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

Hadoop data

Relational database and enterprise data warehouse

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



Hadoop Ecosystem

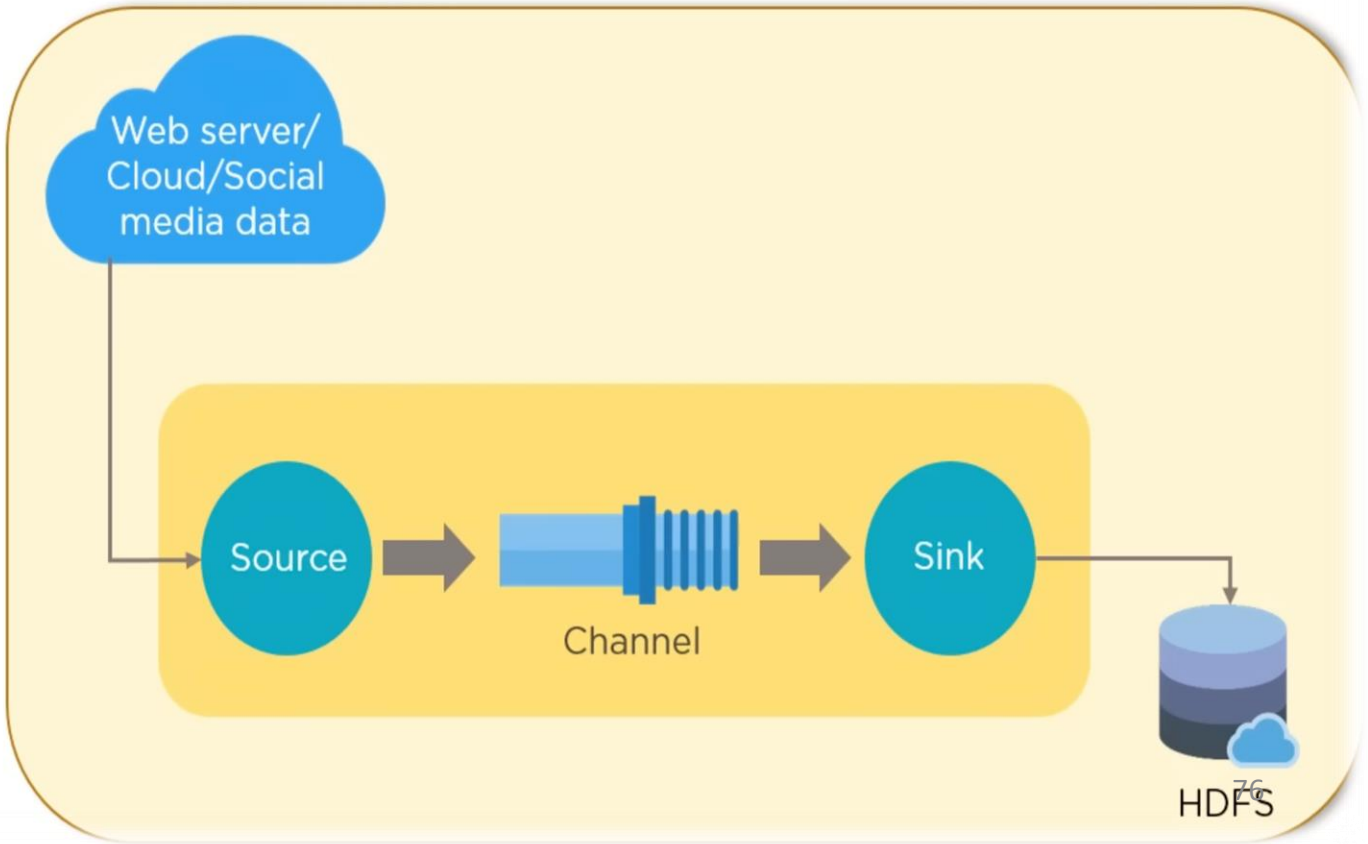
Flume is distributed service for collecting, aggregating and moving large amounts of log data

Flume

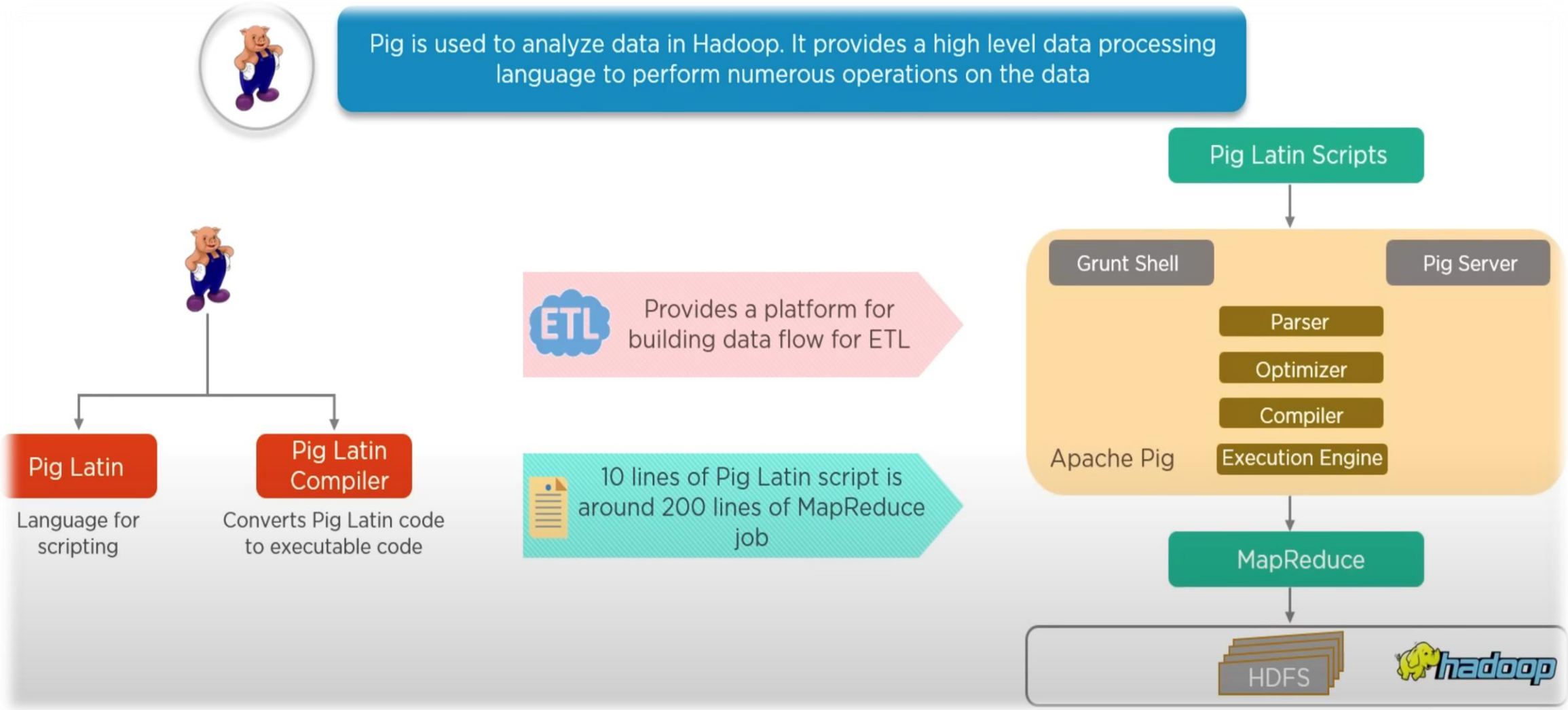
ingests

Unstructured and semi-structured data into HDFS

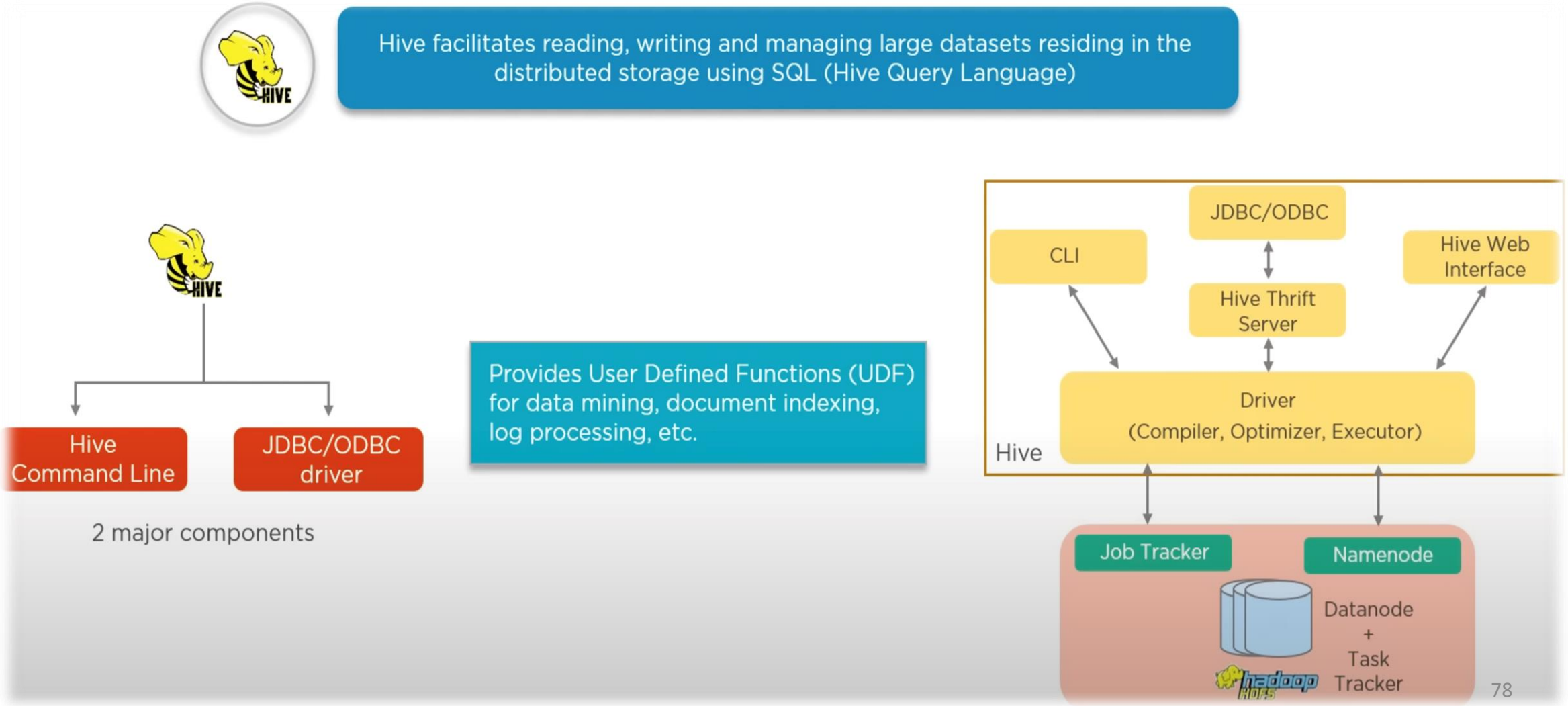
Ingests online streaming data from social media, log files, web server into HDFS



Hadoop Ecosystem



Hadoop Ecosystem



Hadoop Ecosystem

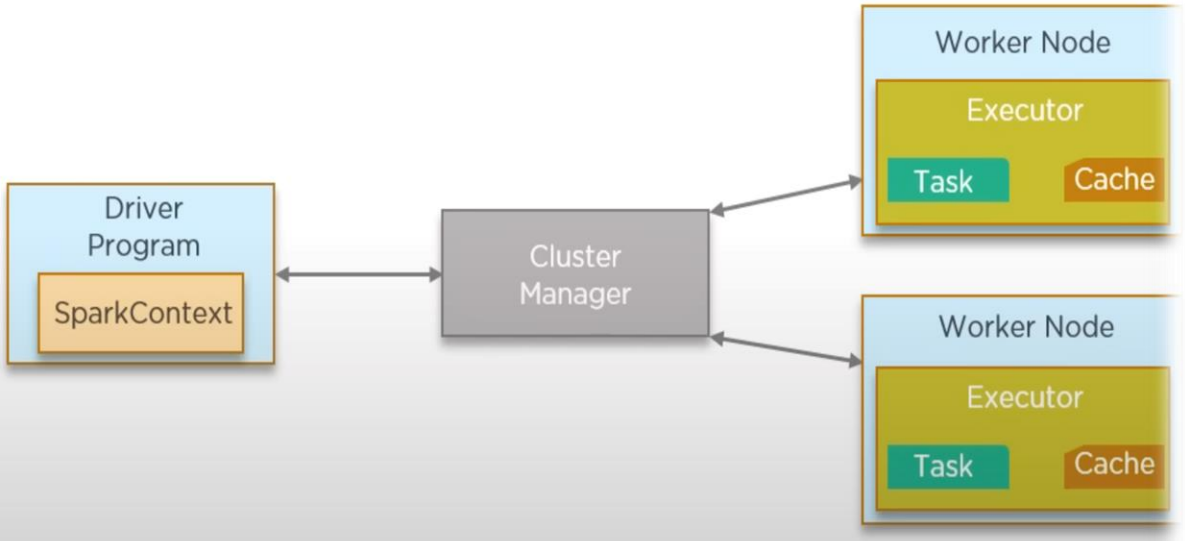
Spark is an open-source distributed computing engine for processing and analyzing huge volumes of real time data

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



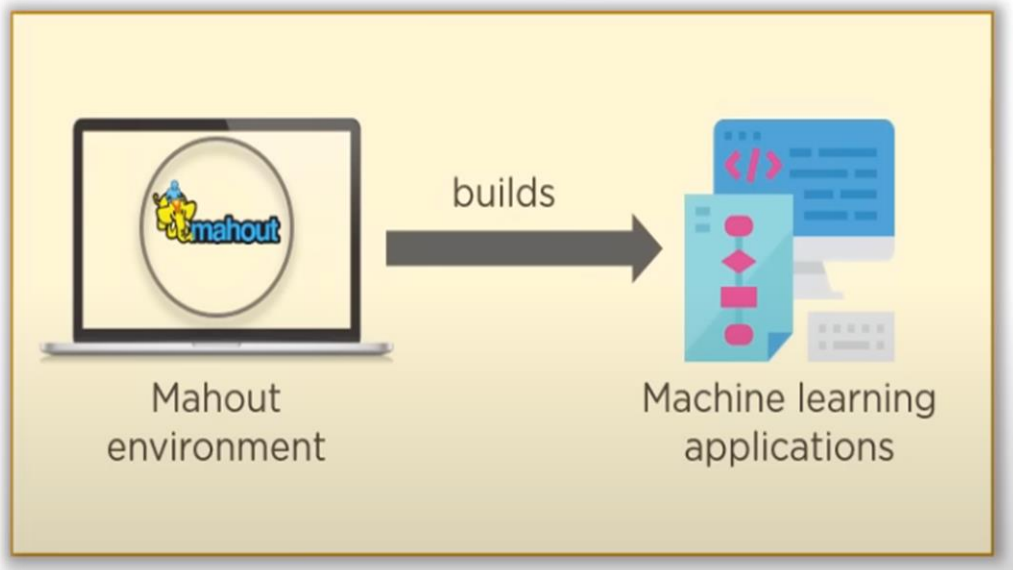
Hadoop Ecosystem

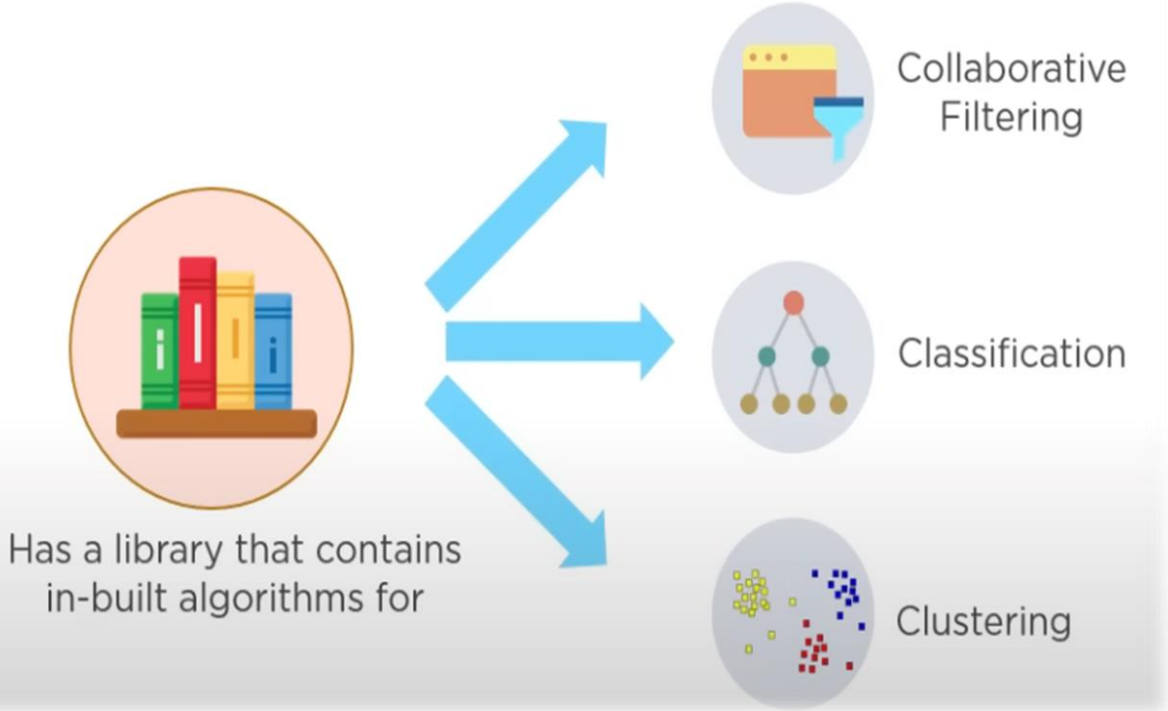
PySpark

PySpark is very well used in Data Science and Machine Learning community as there are many widely used data science libraries written in Python including NumPy, TensorFlow. Also used due to its efficient processing of large datasets

Hadoop Ecosystem

Mahout is used to create scalable and distributed machine learning algorithms

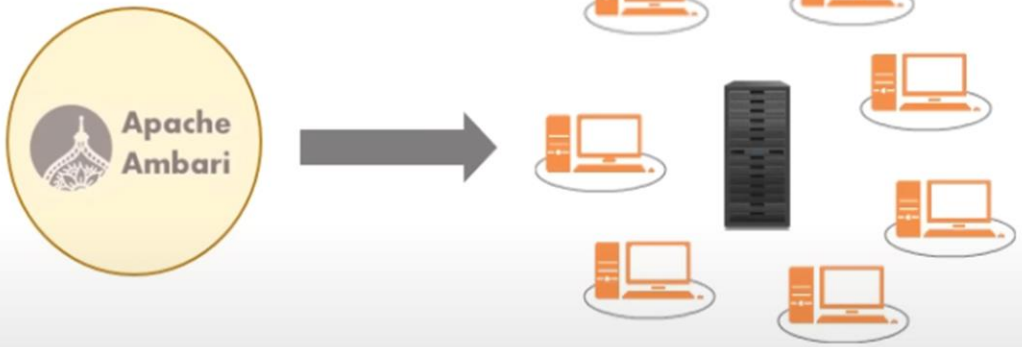




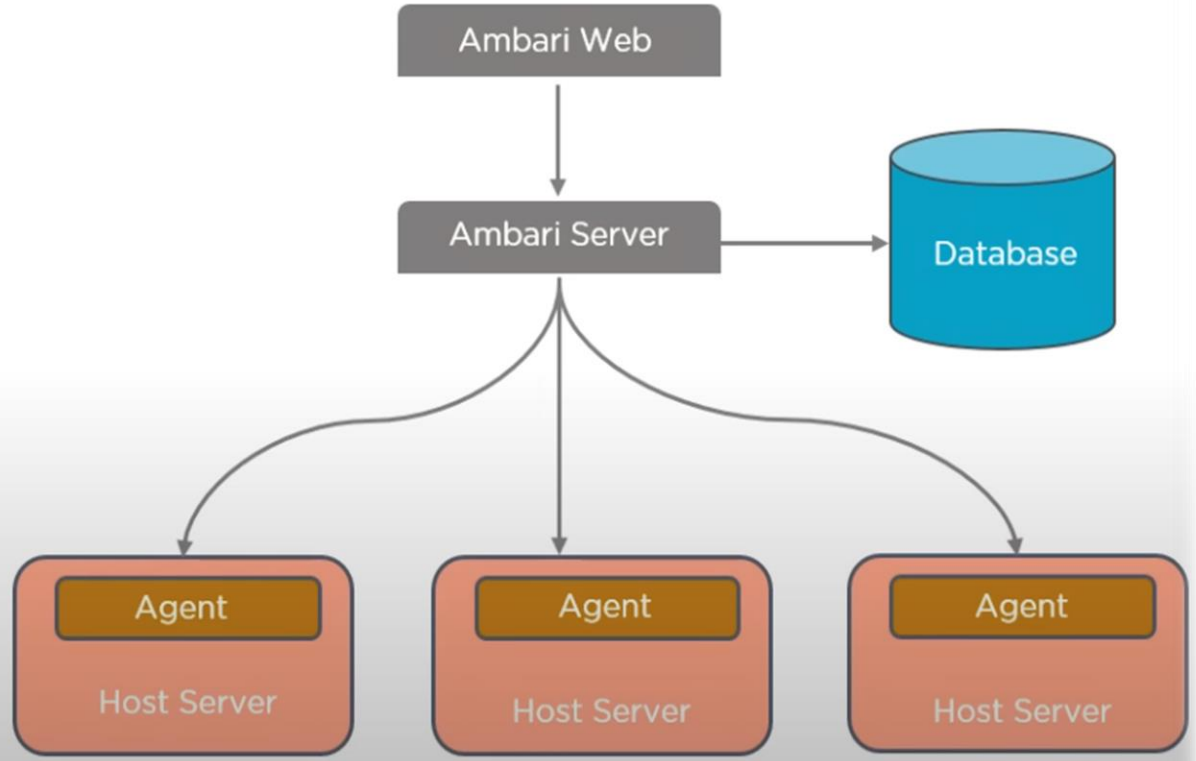
Hadoop Ecosystem

Apache Ambari

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



- Manages, monitors and provisions Hadoop clusters
- Provides a central management service to start, stop and configure Hadoop services



Hadoop Ecosystem

kafka

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

Written in Scala Java

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

Hadoop Ecosystem

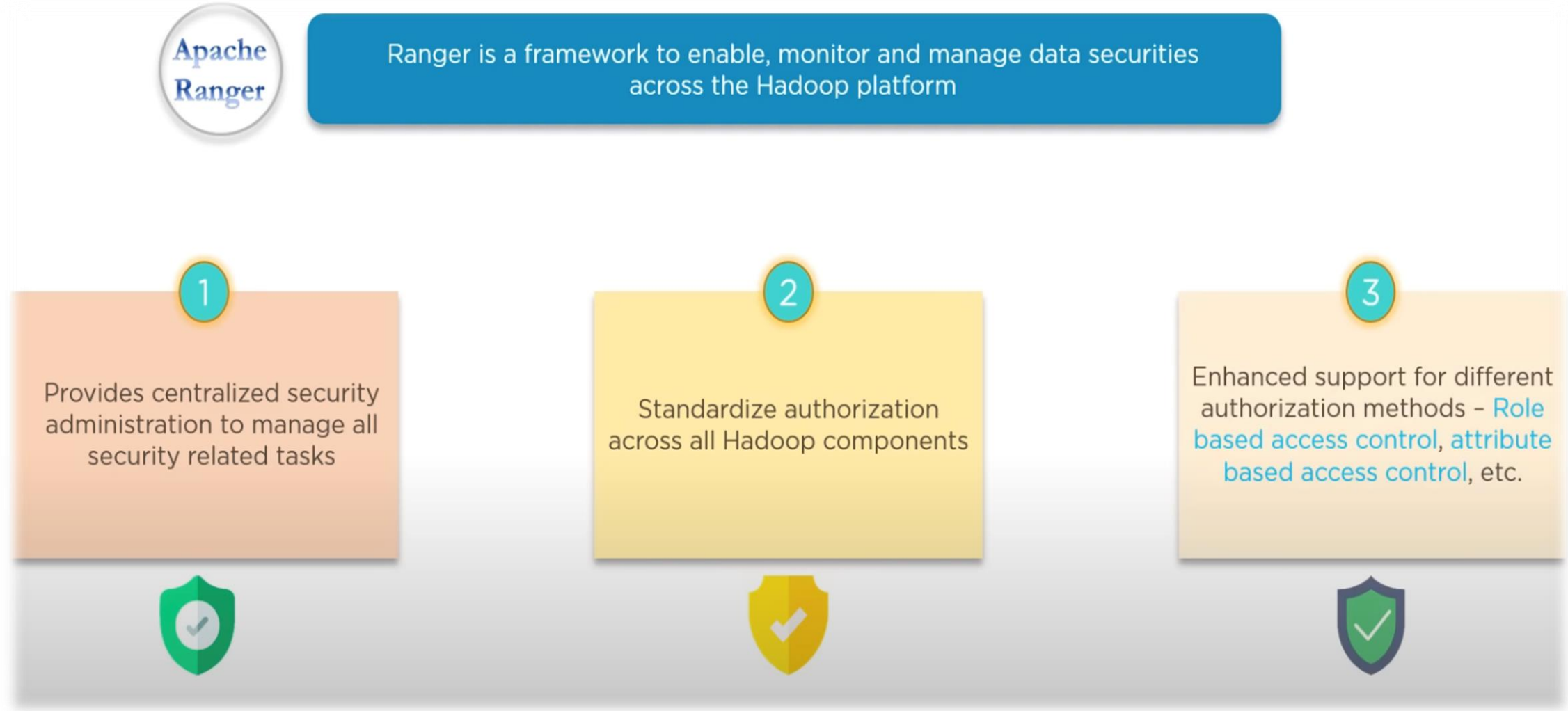
Storm is a processing engine that processes real-time streaming data at a very high speed

Written in Clojure

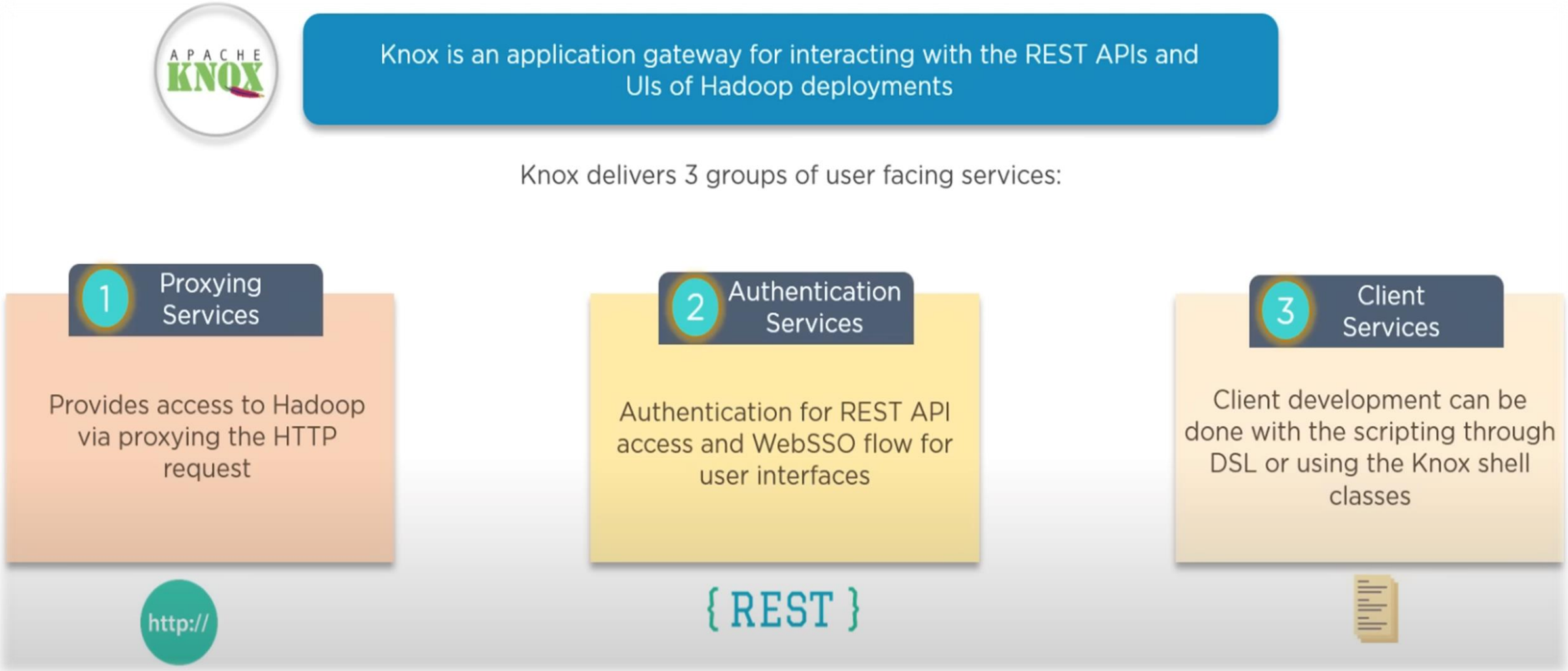
Ability to process over a million jobs in a fraction of seconds on a node

It is integrated with Hadoop to harness higher throughputs

Hadoop Ecosystem

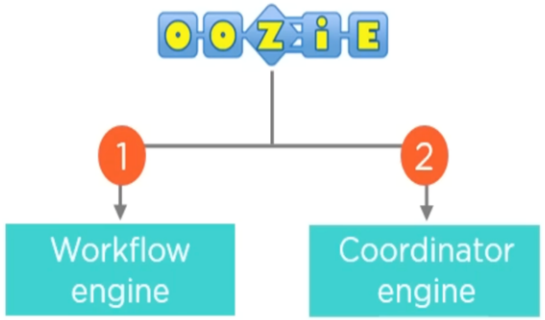


Hadoop Ecosystem



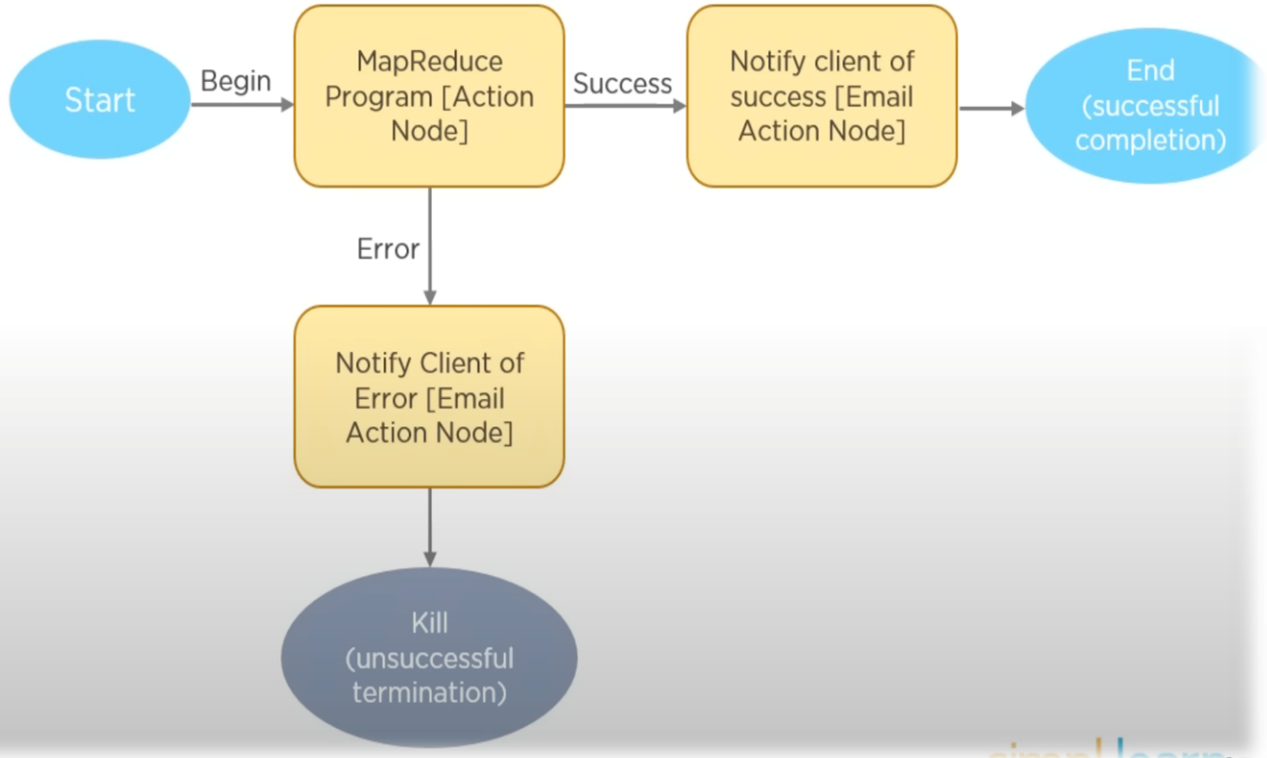
Hadoop Ecosystem

Oozie is a workflow scheduler system to manage Hadoop jobs



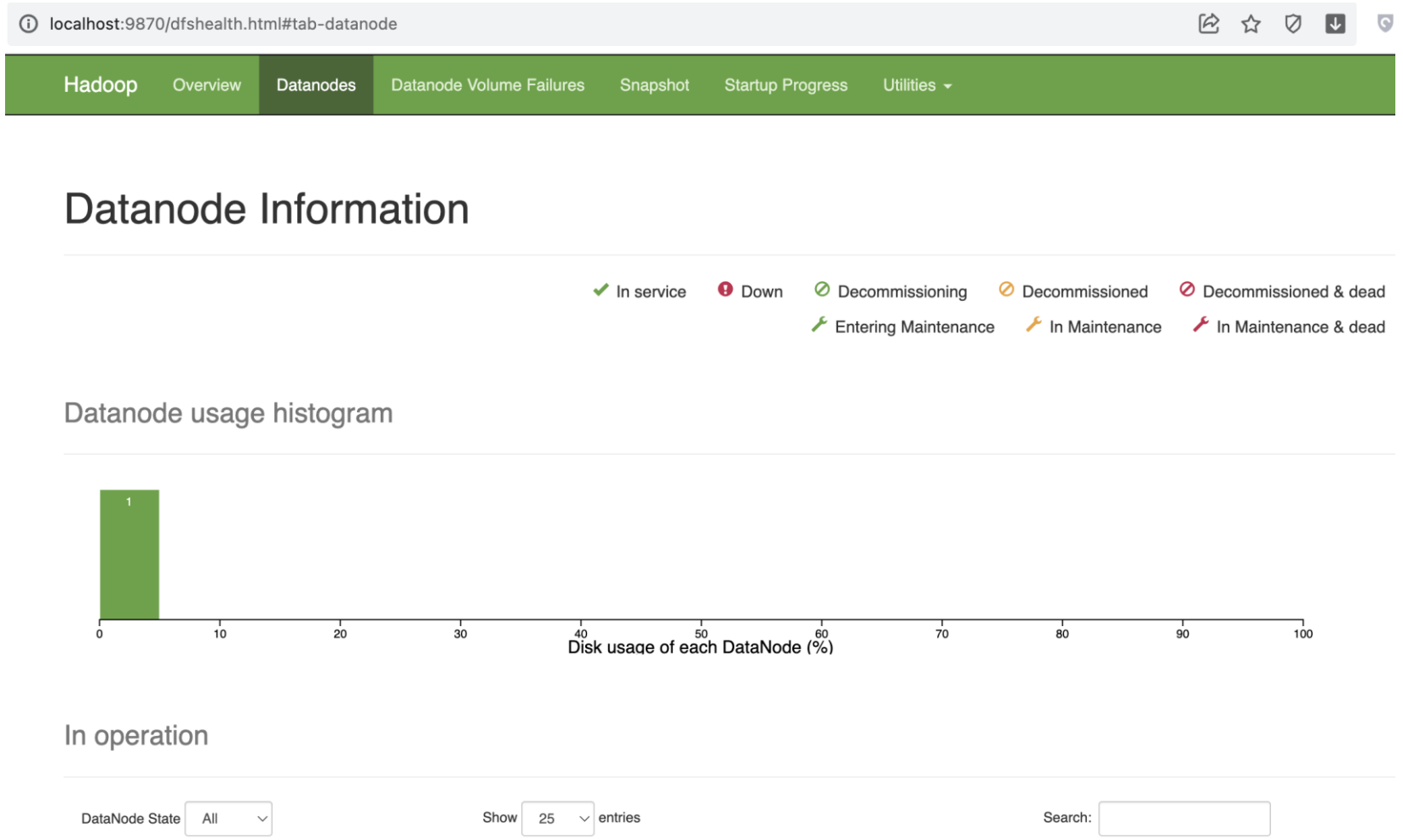
Consists of 2 parts

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



AI VIETNAM
All-in-One Course

Hadoop Example



AI VIETNAM
All-in-One Course

Hadoop Example