SIM GLOBAL EDUCATION

UNIVERSITY OF WOLLONGONG AUSTRALIA

ISIT312 – Big Data Management

Introduction to Hadoop

Sionggo Japit

sjapit@uow.edu.au

7 October 2022

ISIT312-Big Data Management

Introduction to Hadoop

What is Big data?

- With the advancement of technology, the data available for processing are becoming more complex, huge in volume, and of diverse types; they form multitude of unstructured, semi-structured, as well as structured data.

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Veracity

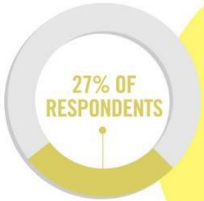
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR



in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

What is big data?

- The terminology used to refer to these multitude of data is big data.

Definition:

- Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just cannot manage them.

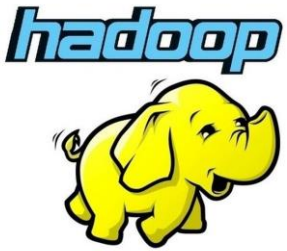
https://www.oracle.com/in/big-data/what-is-big-data/

What is Hadoop?

- Because of the voluminous size and complex form, big data is hard to capture, store, search, share, analyse, and visualize.
- What is the solution?

What is Hadoop?

- Because of the voluminous size and complex form, big data is hard to capture, store, search, share, analyse, and visualize.
- What is the solution?

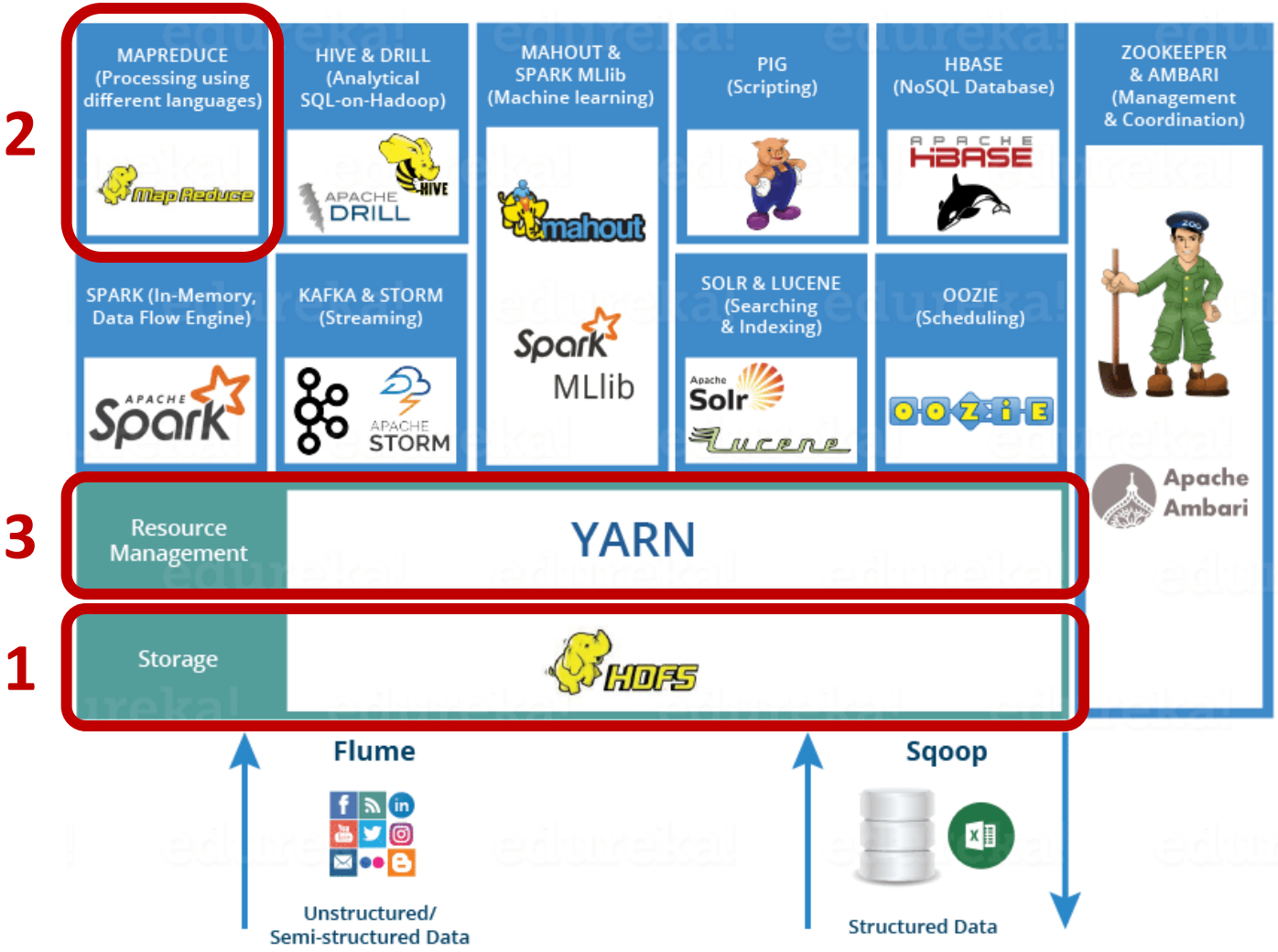


Hadoop to the rescue!!!

How Hadoop Work?

- Hadoop is a framework that can store and process huge amounts of unstructured data ranging in size from terabytes to petabytes. It is a highly fault-tolerant and highly available system.
- Hadoop consisted of three components that were specifically designed to work on big data.
    1. Storage unit – Hadoop **HDFS** (Hadoop Distributed File System)
    2. Hadoop **MapReduce**, and
    3. Hadoop **YARN** (Yet Another Resource Negotiator)

Hadoop Ecosystem



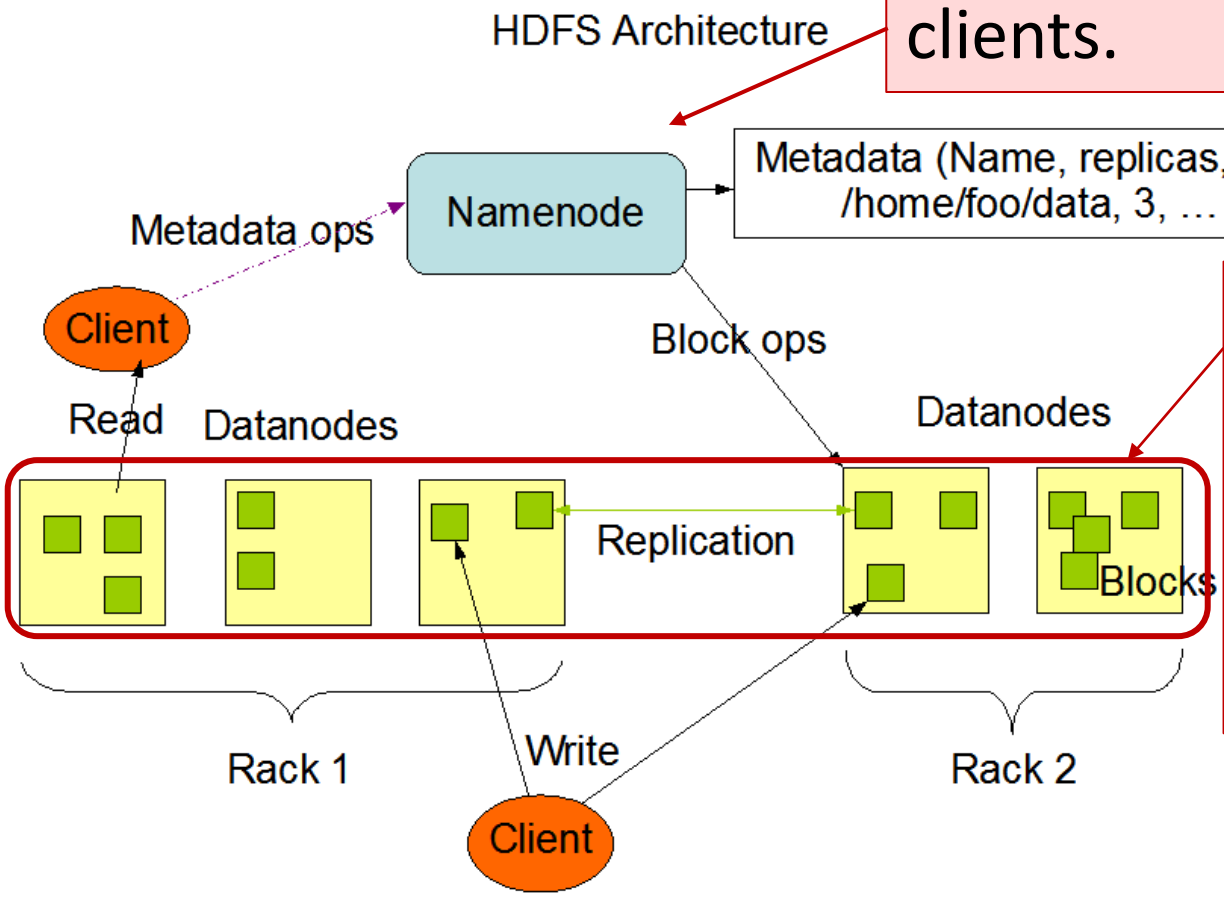
How Hadoop Work?

Hadoop HDFS

- The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.
- It divides the data into blocks and stores them across multiple systems.
- In this way, data is not lost at any cost, even if one data node crashes. Hence, this makes HDFS a highly fault-tolerant.
- The block size is 128 MB by default.

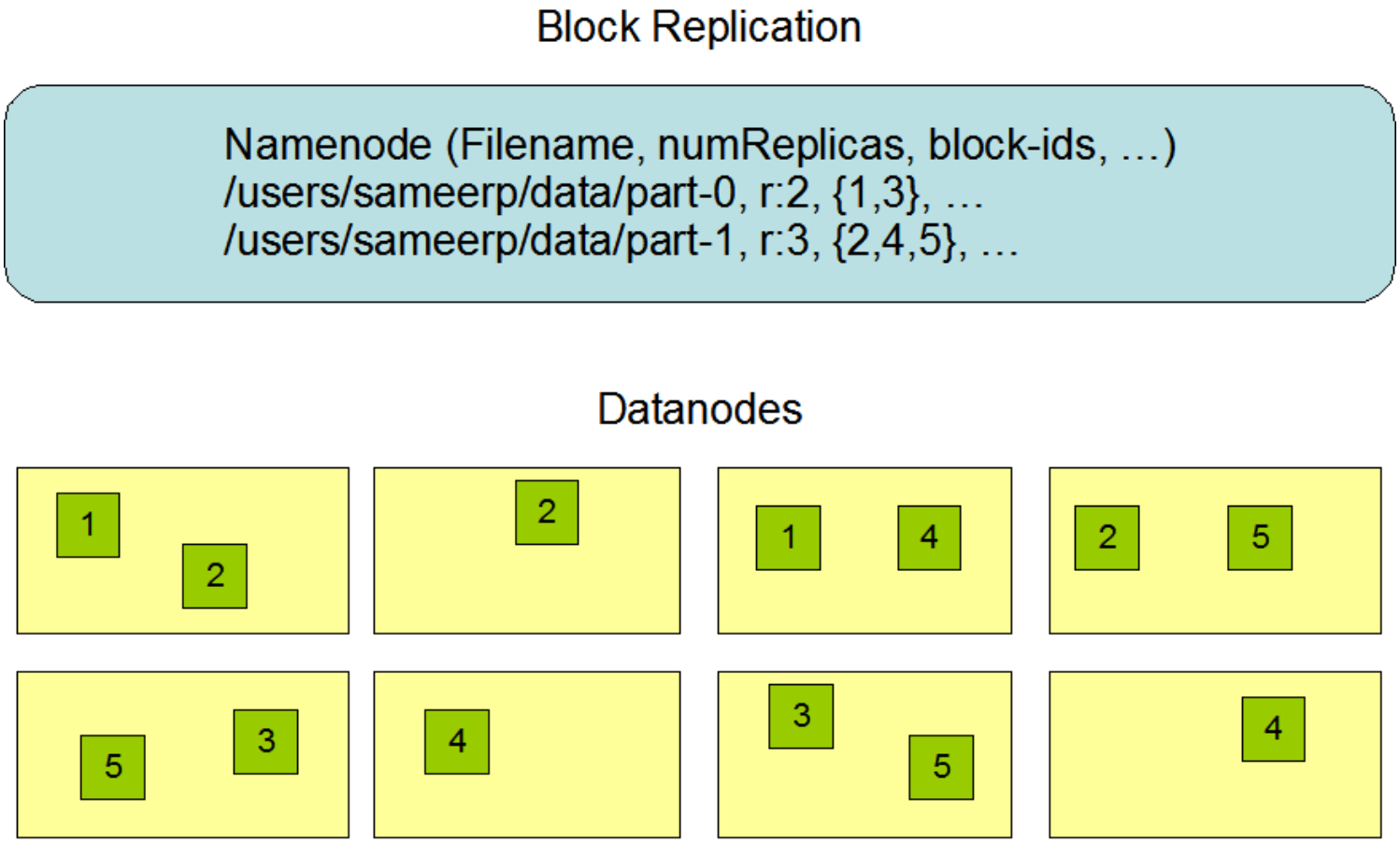
Hadoop HDFS

NameNode – A master server that manages the file system namespaces and regulates access to files by clients.



DataNodes – manage storage attached to the nodes that they run on.

Hadoop HDFS

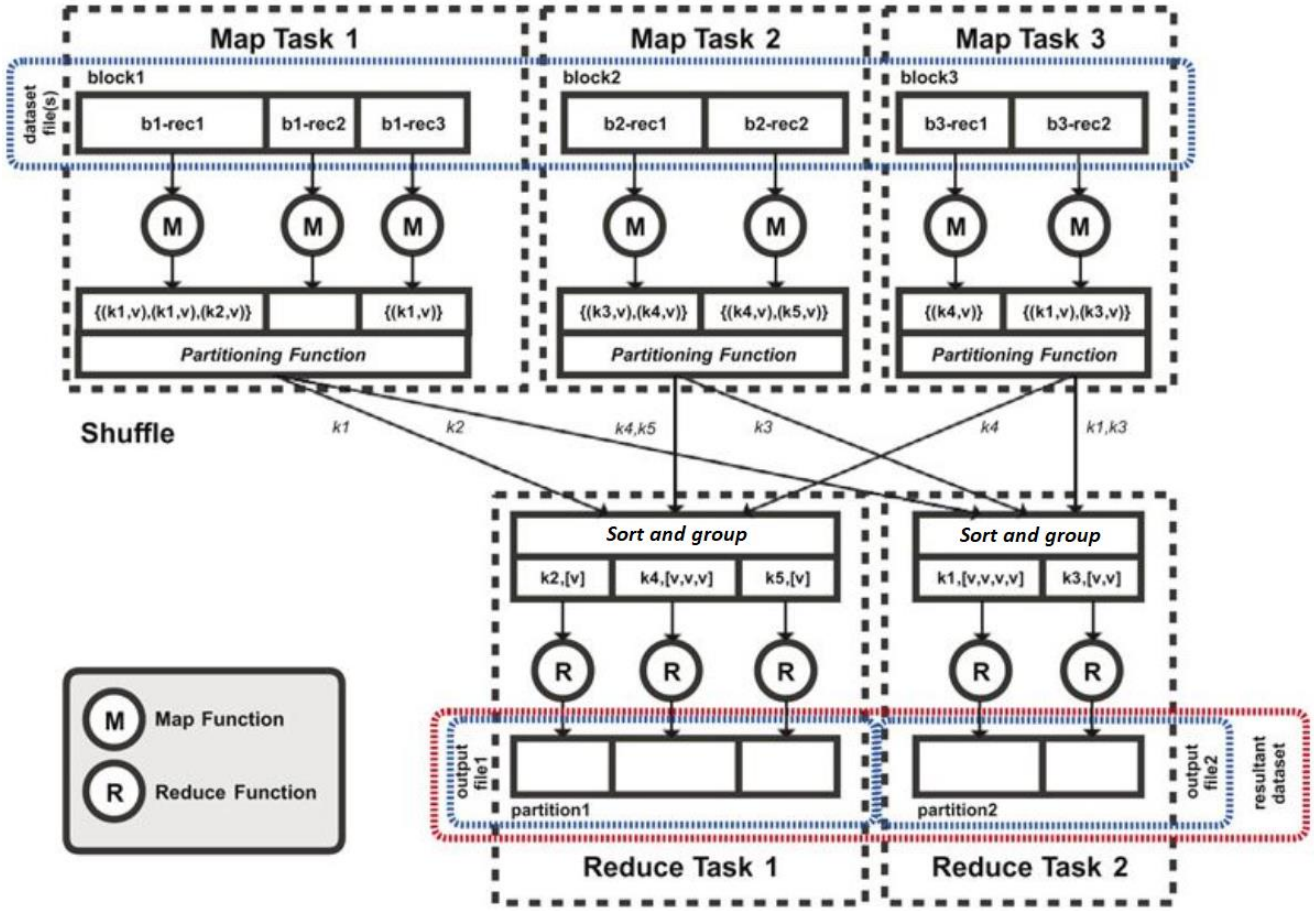

Block Replication
Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …
Datanodes
1
2
2
1
4
2
5
5
3
4
3
5
4

How Hadoop Work?

Hadoop MapReduce

- Hadoop MapReduce processes the data stored in Hadoop HDFS in parallel across various nodes in the cluster.
- It works in two phases:
  i. It splits data into parts and processes each of them separately on different data nodes.
  ii. The individual results are then aggregated to give the final output.

MapReduce


Map Task 1
Map Task 2
Map Task 3
dataset file(s)
block1
b1-rec1
b1-rec2
b1-rec3
block2
b2-rec1
b2-rec2
block3
b3-rec1
b3-rec2
M
{(k1,v),(k1,v),(k2,v)}
{(k1,v)}
{(k3,v),(k4,v)}
{(k4,v),(k5,v)}
{(k4,v)}
{(k1,v),(k3,v)}
Partitioning Function
Shuffle
k1
k2
k4,k5
k3
k4
k1,k3
Sort and group
k2,[v]
k4,[v,v,v]
k5,[v]
k1,[v,v,v,v]
k3,[v,v]
R
output file1
partition1
partition2
output file2
resultant dataset
Reduce Task 1
Reduce Task 2
Map Function
Reduce Function

MapReduce

Input Splits:

- Data set is divided into fixed-size chunk (block) that is consumed by a single map.

Mapping:

- Data in each chunk is passed to a mapping function to produce counts of occurrences of each word, and prepare a list of key-value pair where key is the word, and value is the frequency of occurrences.

Shuffling

- Shuffling process will consolidate the relevant records from Mapping phases by clubbing together the same words and accumulate their frequency.

MapReduce

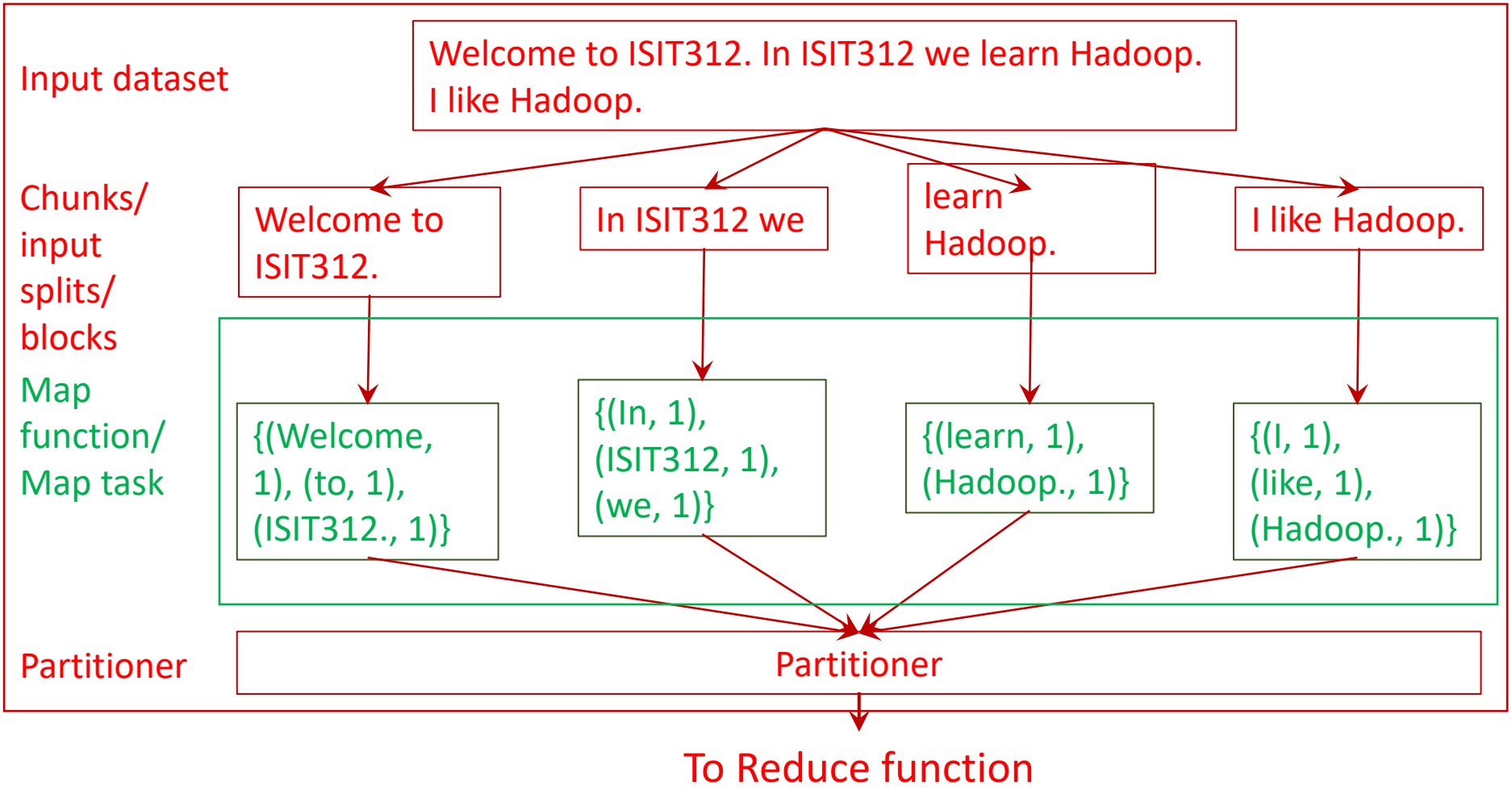
Reducing

- In this phase, the output values from the Shuffling phase are aggregated, by combining all the words into a single output, that is, producing a complete dataset.

MapReduce (Without Combiner) – Map Phase

- The following example depicts a MapReduce function without a combiner in the Map function.
- Dataset is split into blocks/chunks of fixed size.
- Map function is then assigned to process each block, that is, each Map function operates only one block.
- Each Map function outputs (produces) sets of key-value pair records.
- If without combiner, the sets of key-value pair records are passed to *Patitioner*, which ensures each key-value pair record is passed to one and only one Reducer.

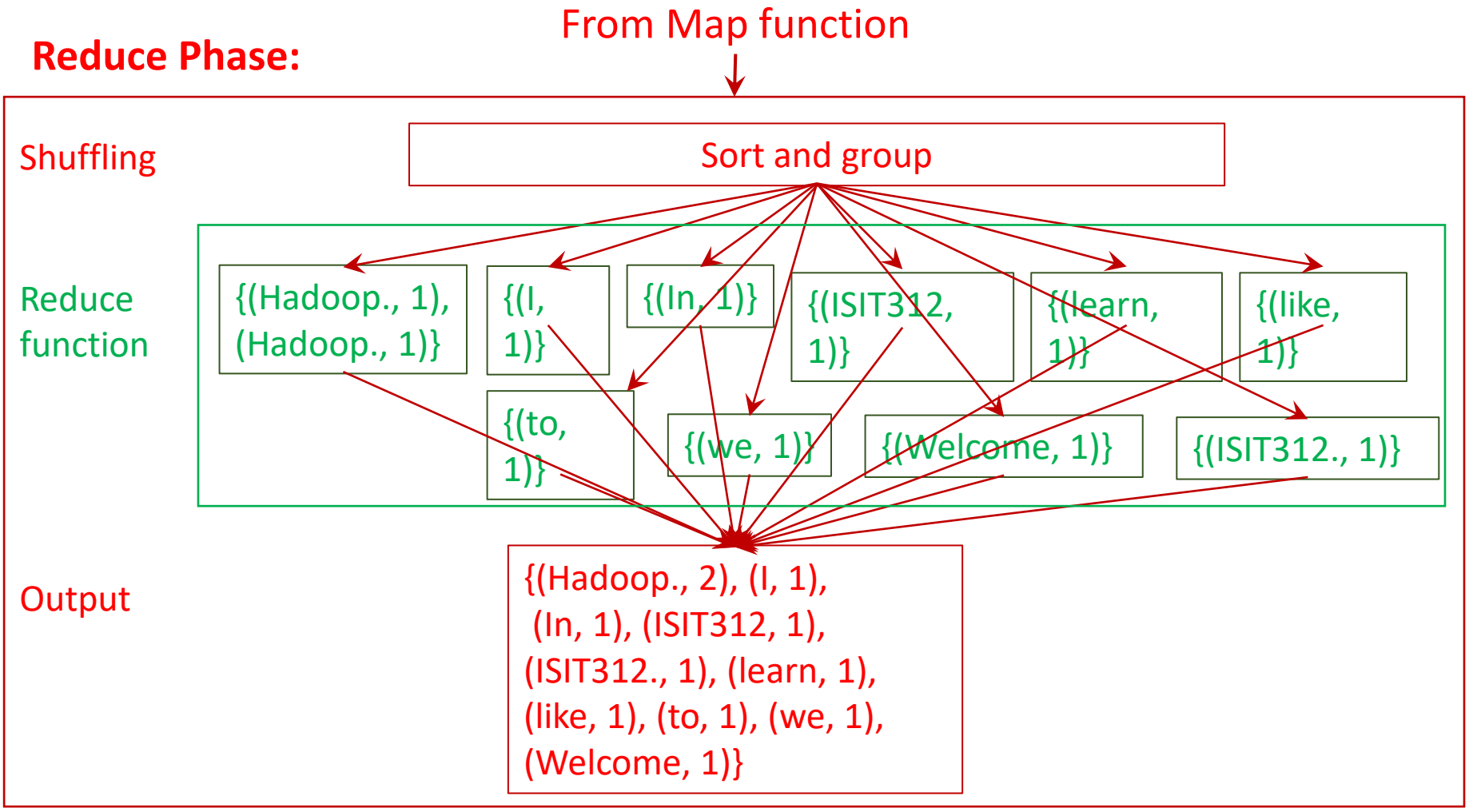
Map Phase:


Input dataset
Welcome to ISIT312. In ISIT312 we learn Hadoop.
I like Hadoop.
Chunks/
input
splits/
blocks
Welcome to
ISIT312.
In ISIT312 we
learn
Hadoop.
I like Hadoop.
Map
function/
Map task
{(Welcome,
1), (to, 1),
(ISIT312., 1)}
{(In, 1),
(ISIT312, 1),
(we, 1)}
{(learn, 1),
(Hadoop., 1)}
{(I, 1),
(like, 1),
(Hadoop., 1)}
Partitioner
Partitioner
To Reduce function

MapReduce (Without Combiner) – Reduce Phase

- Input to Reduce Phase is the output from Map Phase, that is, the *Partitioner* ensures each key-value pair is passed to one and only one Reducer.
- The Reduce will perform a sort and group function to sort and group the key-value pair by the key and accumulate the values for each key.
- The Reduce function will then output a set consisting of all the key-value pairs.
- Note: A reducer may be receiving input from multiple Map functions.

Reduce Phase:



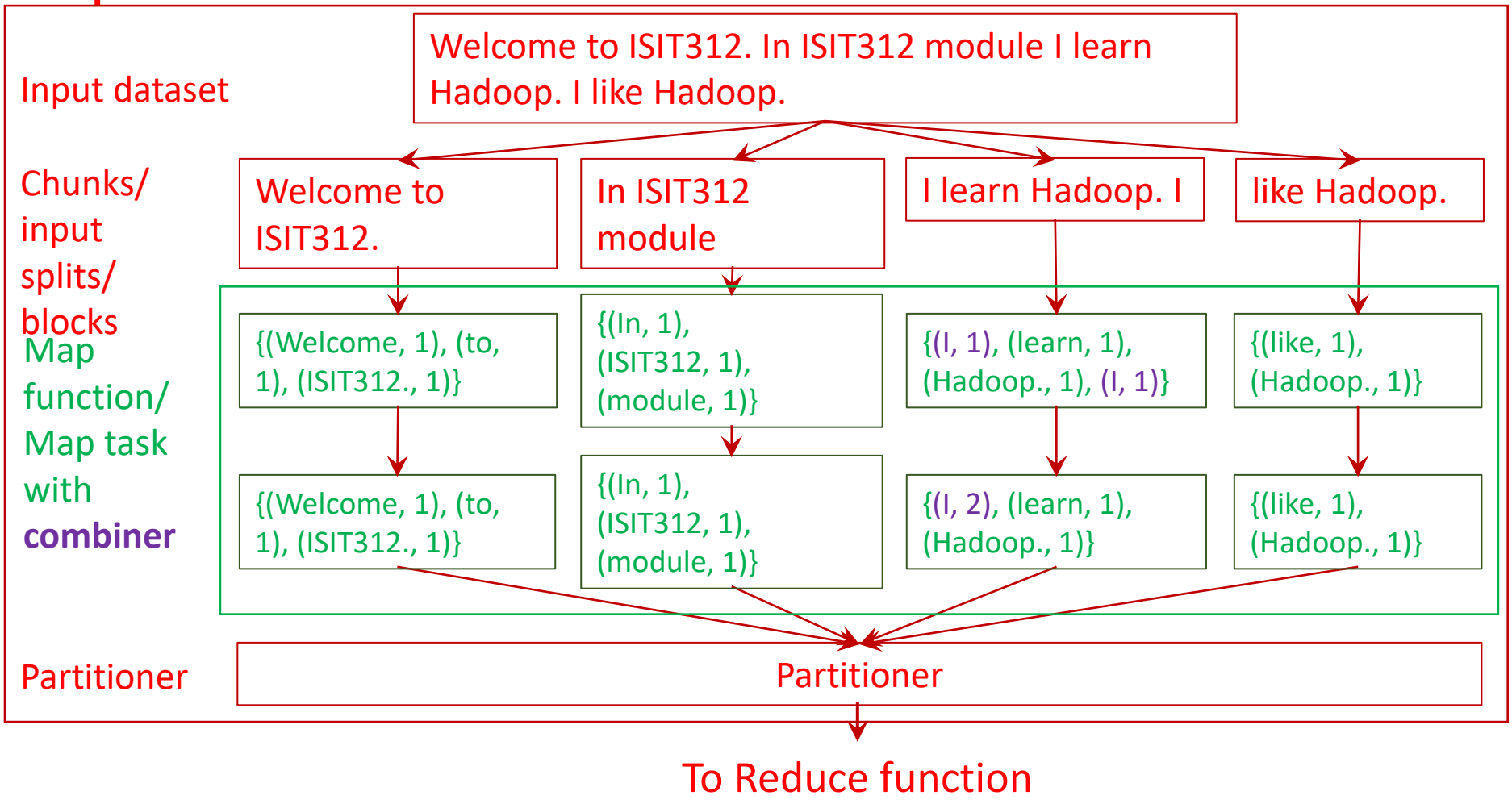
MapReduce (With Combiner) – Map Phase

- The following example depicts a MapReduce function with a combiner in the Map function.
- Dataset is split into blocks/chunks of fixed size.
- Map function is then assigned to process each block, that is, each Map function operates only one block.
- Each Map function outputs (produces) sets of key-value pair records.

MapReduce (With Combiner) – Map Phase

- Combiner performs sum or count function to combine each key and its value before passing the key-aggregateValue pairs to *Patinioner*, which ensures each key-aggregateValue pair record is passed to one and only one Reducer.

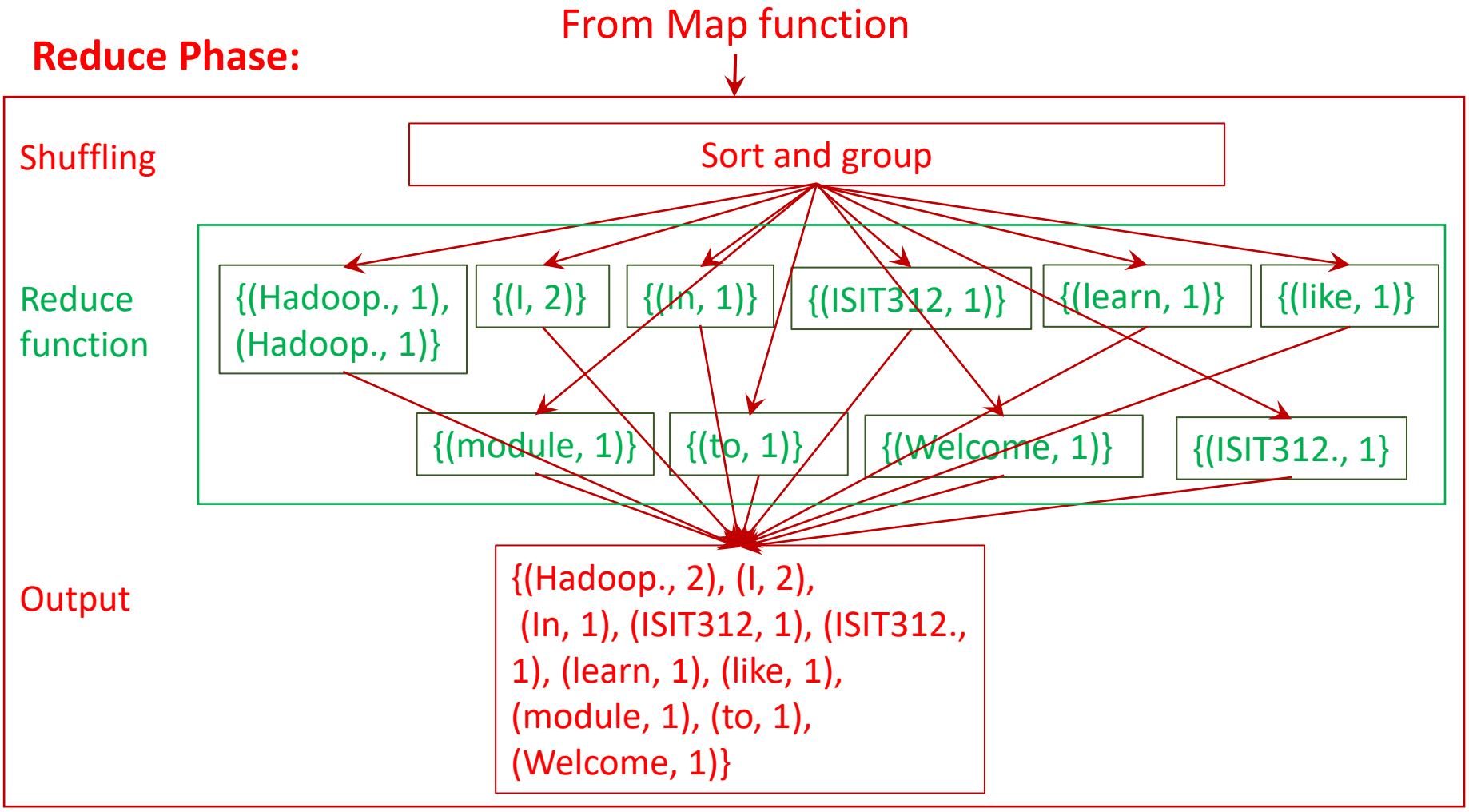
Map Phase:


Input dataset
Welcome to ISIT312. In ISIT312 module I learn Hadoop. I like Hadoop.
Chunks/ input splits/ blocks
Welcome to ISIT312.
In ISIT312 module
I learn Hadoop. I
like Hadoop.
Map function/ Map task with combiner
{(Welcome, 1), (to, 1), (ISIT312., 1)}
{(In, 1), (ISIT312, 1), (module, 1)}
{(I, 1), (learn, 1), (Hadoop., 1), (I, 1)}
{(like, 1), (Hadoop., 1)}
{(Welcome, 1), (to, 1), (ISIT312., 1)}
{(In, 1), (ISIT312, 1), (module, 1)}
{(I, 2), (learn, 1), (Hadoop., 1)}
{(like, 1), (Hadoop., 1)}
Partitioner
Partitioner
To Reduce function

MapReduce (With Combiner) – Reduce Phase

- Input to Reduce Phase is the output from Map Phase, that is, the *Partitioner* ensures each key-value pair is passed to one and only one Reducer.
- The Reduce will perform a sort and group function to sort and group the key-value pair by the key and accumulate the values for each key.
- The Reduce function will then output a set consisting of all the key-value pairs.
- Note: A reducer may be receiving input from multiple Map functions.

Reduce Phase:



How Hadoop Work?

- The MapReduce processes (shown earlier), improve load balancing and saves a considerable amount of time.
- Now that we have the output of MapReduce jobs ready, it is time for us to run them on the Hadoop cluster. This is done with the help of a set of resources such as RAM, network bandwidth and CPU.
- Multiple jobs are run on Hadoop simultaneously, and each of them needs some resources to complete the task successfully.

How Hadoop Work?

- To efficiently manage these resources, we have the third component of Hadoop, that is YARN.

YARN (Yet Another Resource Negotiator)

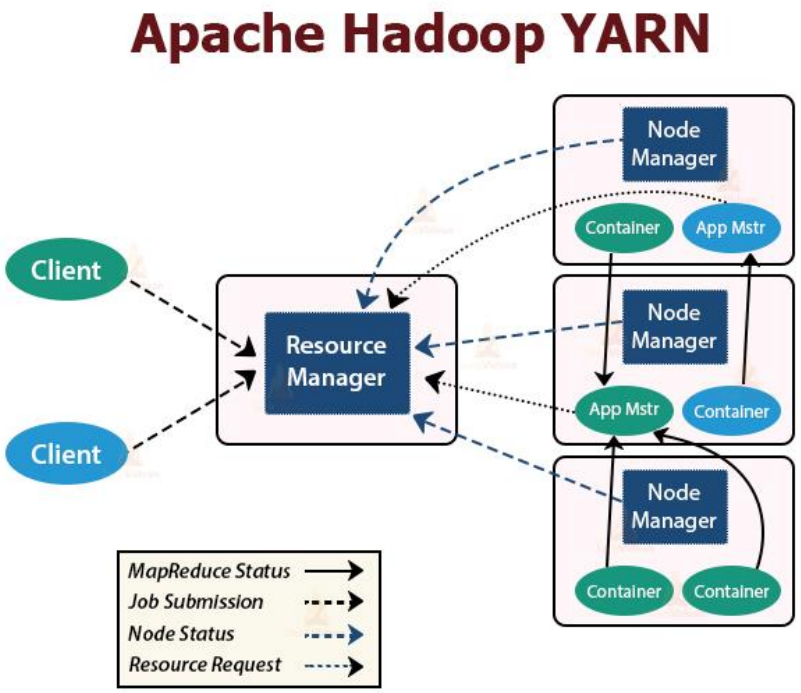
- YARN is responsible for sharing resources amongst the applications running in the cluster and scheduling the task in the cluster.
- It consists of a resource manager, node manager, application master, and containers.

How Hadoop Work?

- The resource manager assigns resources.
- The node managers handle the nodes and monitor the resources usage in the node.
- The containers hold a collection of physical resources.
- Now, suppose we want to process the MapReduce jobs, we had created,
    i. The application master requests the container from the node manager
    ii. Once the node manager gets the resources, it sends the jobs to the resource manager.

How Hadoop Work?

- This YARN processes job requests and manages cluster resources in Hadoop.



How Hadoop Work?

- In addition to these components, Hadoop also has various data tools and frameworks dedicated to managing, processing, and analysing data.
- These tools and frameworks are also commonly referred to as the Hadoop echosystem.

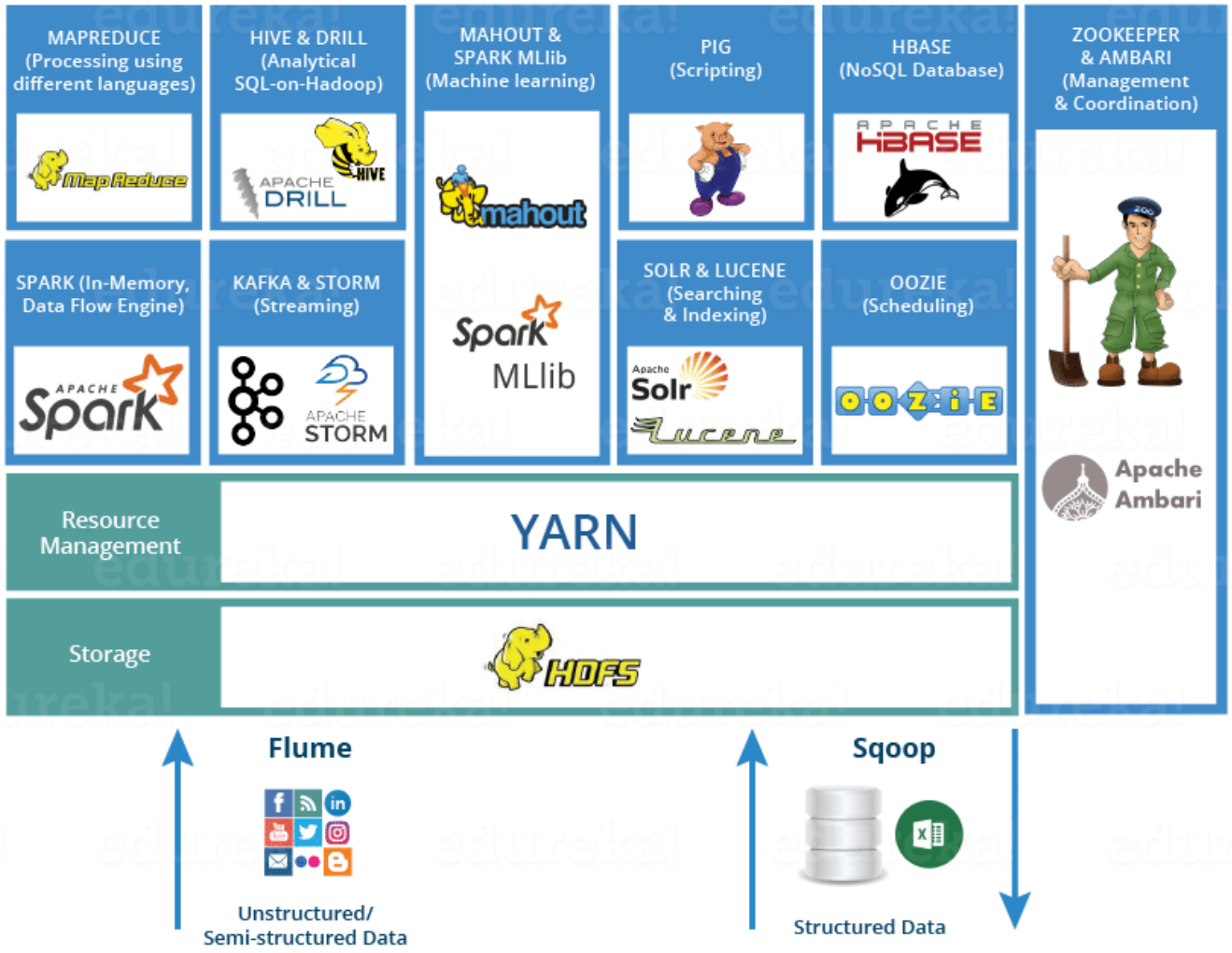
Hadoop Ecosystem

- The Hadoop echosystem comprises of several other components such as:
  - Hive (Analytical SQL on Hadoop),
  - Pig (Scripting),
  - Apache Spark (In-Memory, Data flow engine)
  - Spark Mlib (Machine Learning),
  - Apache Hbase (NoSQL database),
  - Apache Zookeeper and Ambari (Management and coordination),
  - KAFKA and STORM (Streaming),

Hadoop Ecosystem

- The Hadoop echosystem comprises of several other components such as: (continue…)
  - SOLR and LUCENE (Searching and Indexing),
  - OOZIE (Scheduling)
  - Flume (Unstructured/semi-structured data), and
  - Sqoop (Structured data)

Hadoop Ecosystem



References

- https://www.oracle.com/in/big-data/what-is-big-data/
- https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Other useful links:

- https://hadoop.apache.org/docs/r3.1.0/api/org/apache/hadoop/fs/FileSystem.html
- https://www.techtarget.com/searchdatamanagement/definition/Hadoop-Distributed-File-System-HDFS