Walchand Linux Users' Group

Presents Club Service On

Topics Covered :

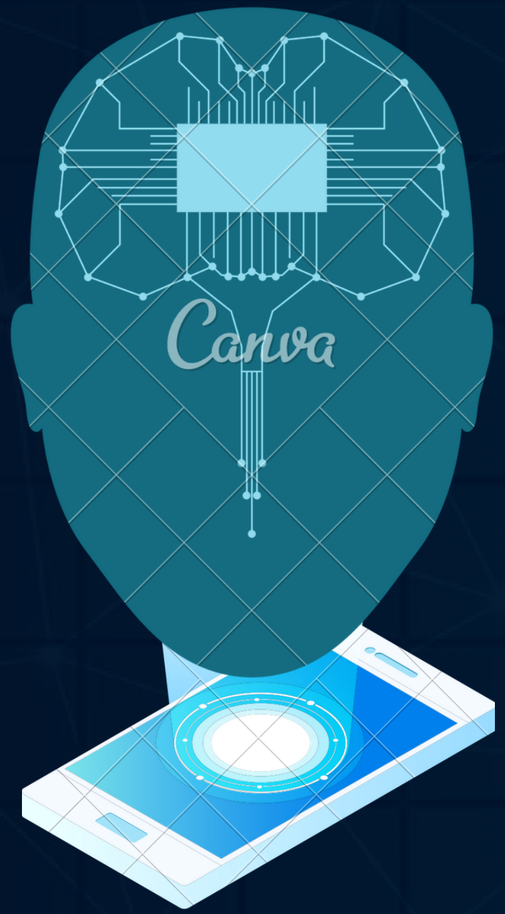
- Big data
- Hadoop intro
- Architecture
- Working
- Advantages

CONNECT WITH US

www.wcewlug.org

15th Sept 6:00 PM
N2 Classroom

WLUG
COMMUNITY | KNOWLEDGE | SHARE



HADOOP TUTORIAL

PRESENTED BY :

BRIJESH
UMESH

TABLE OF CONTENT

WLUG
COMMUNITY | KNOWLEDGE | SHARE

Update package:
sudo apt get update

Accept required certificates:
sudo apt-get install apt-transport-https ca-certificates

Install docker
sudo apt install docker.io

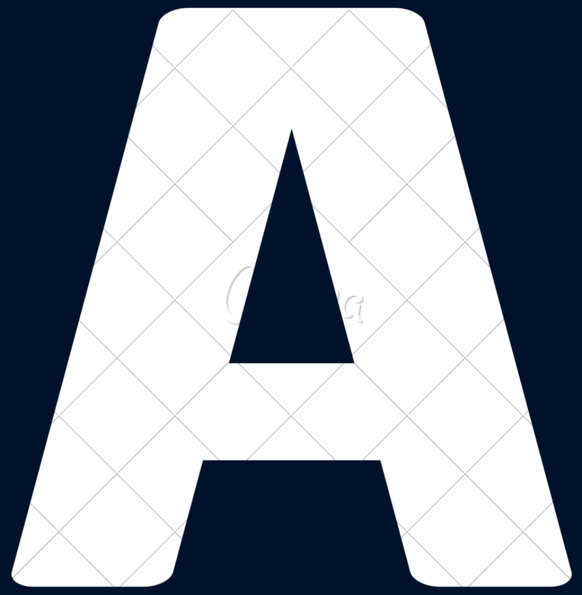
Pull the hadoop image:
docker pull sequenceiq/hadoop-docker:2.7.0

WLUG
COMMUNITY | KNOWLEDGE | SHARE

BIG DATA

- Since the Dawn of time to 2005 - Humans have created 130 exabytes of data.

WLUG
COMMUNITY | KNOWLEDGE | SHARE

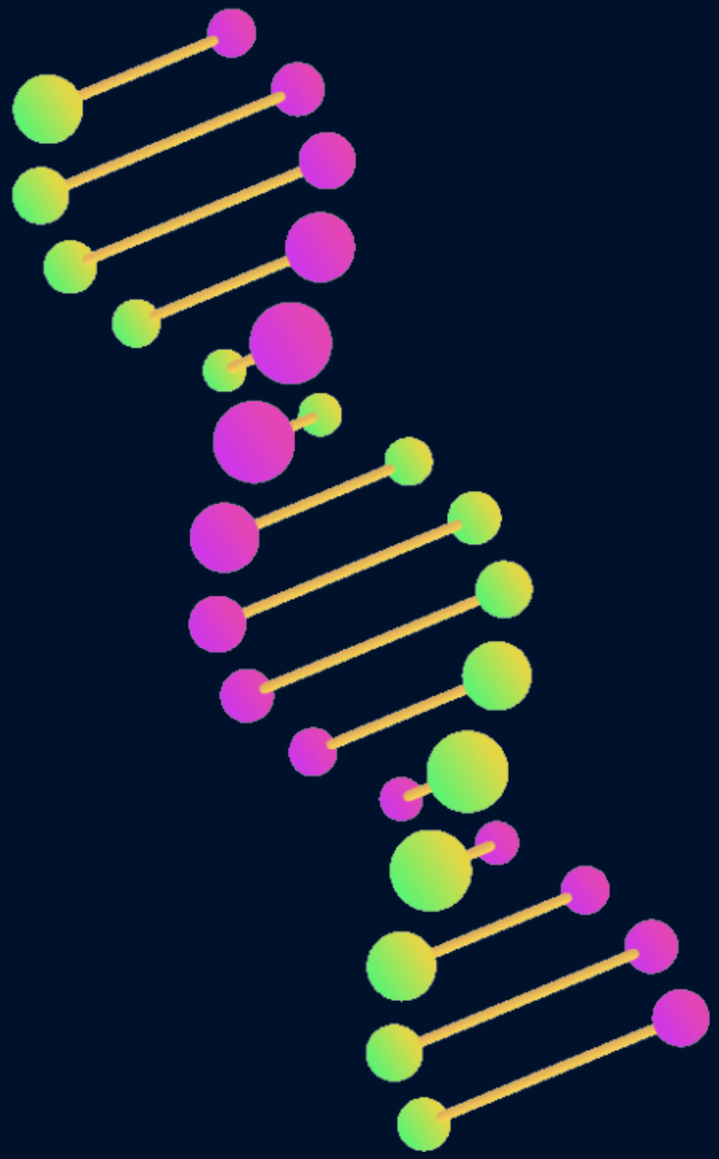


1 Byte

WLUG
COMMUNITY | KNOWLEDGE | SHARE

1 kiloBytes = 1000 Bytes

1 MegaBytes ~ 1000 KiloBytes



WLUG
COMMUNITY | KNOWLEDGE | SHARE

Human GNOME can be Encoded in Digital memory of around 1 GB size.

1 GigaBytes ~ 1000 MegaBytes

WLUG
COMMUNITY | KNOWLEDGE | SHARE

1 TeraBytes ~ 1000 GigaBytes

1 PetaBytes ~ 1000 TeraBytes

390 billion trees

7.9 billion people on earth

WLUG
COMMUNITY | KNOWLEDGE | SHARE

1 ExaBytes ~ 1000 PetaBytes

WLUG
COMMUNITY | KNOWLEDGE | SHARE

BIG DATA

- Since the Dawn of time to 2005, Humans have created 130 exabytes of data.

WLUG
COMMUNITY | KNOWLEDGE | SHARE

BIG DATA

- Since the Dawn of time, Humans have created 130 exabytes of data.
- Till 2010 - 1,200 ExaBytes

WLUG
COMMUNITY | KNOWLEDGE | SHARE

BIG DATA

- Since the Dawn of time, Humans have created 130 exabytes of data.
- Till 2010 - 1,200 ExaBytes
- Till 2015 - 7,900 ExaBytes

WLUG
COMMUNITY | KNOWLEDGE | SHARE

BIG DATA

- Since the Dawn of time, Humans have created 130 exabytes of data.
- Till 2010 - 1,200 ExaBytes
- Till 2015 - 7,900 ExaBytes
- Till 2020 -49,900 ExaBytes

SOLUTIONS ?

WLUG
COMMUNITY | KNOWLEDGE | SHARE

TRADITIONAL SOLUTIONS

WLUG
COMMUNITY | KNOWLEDGE | SHARE

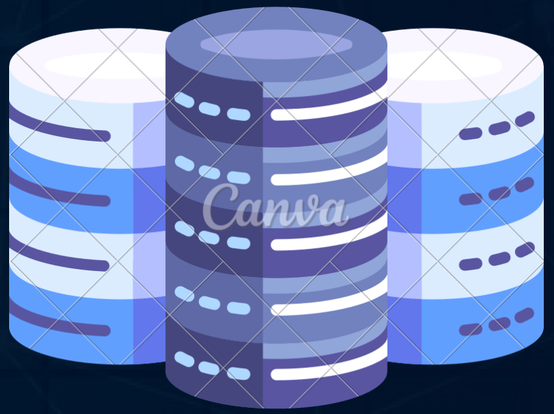
IBM AND ORACLE

SQL

MongoDB

User

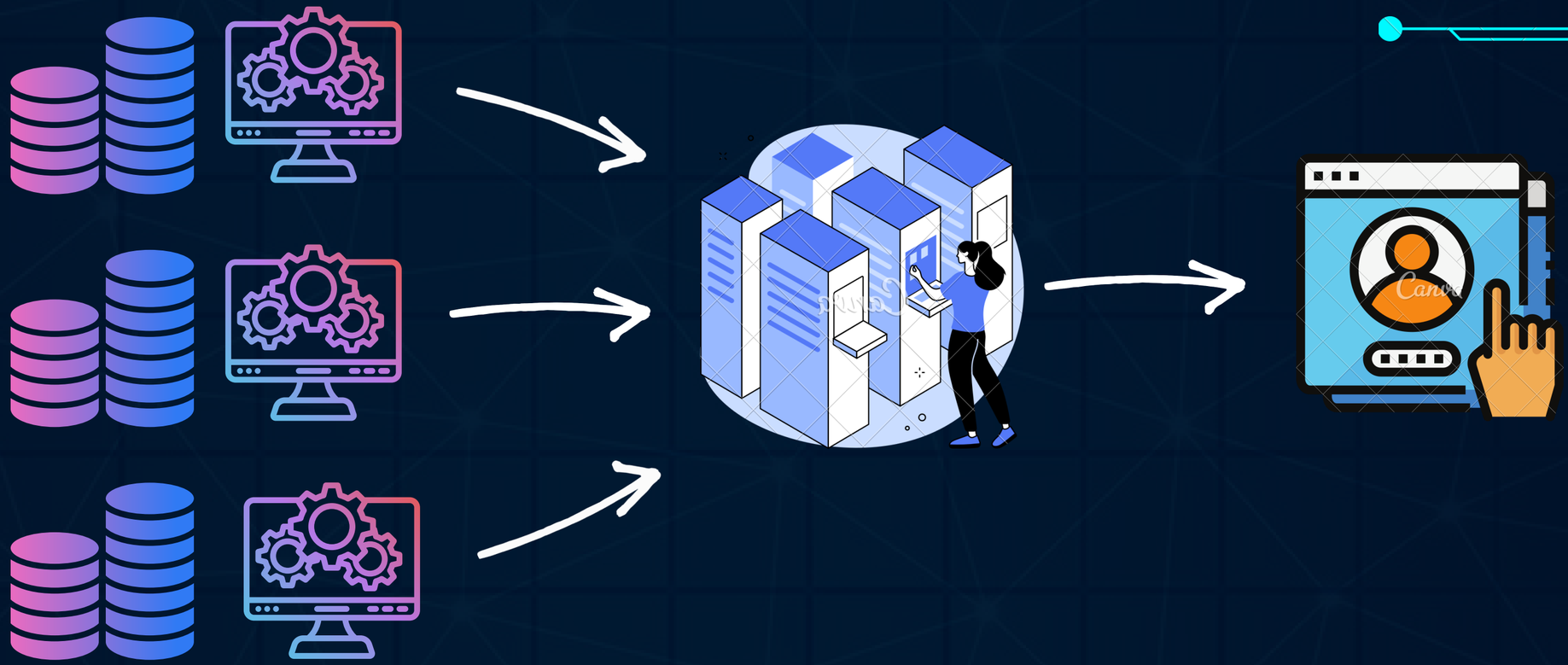
Centralized system

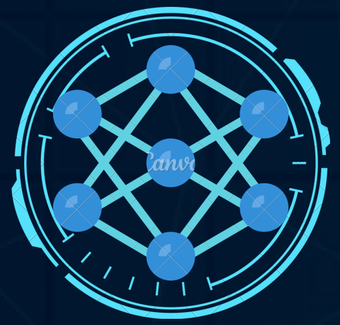


DataBases

GOOGLE SOLUTIONS

WLUG
COMMUNITY | KNOWLEDGE | SHARE





HADOOP

WLUG
COMMUNITY | KNOWLEDGE | SHARE

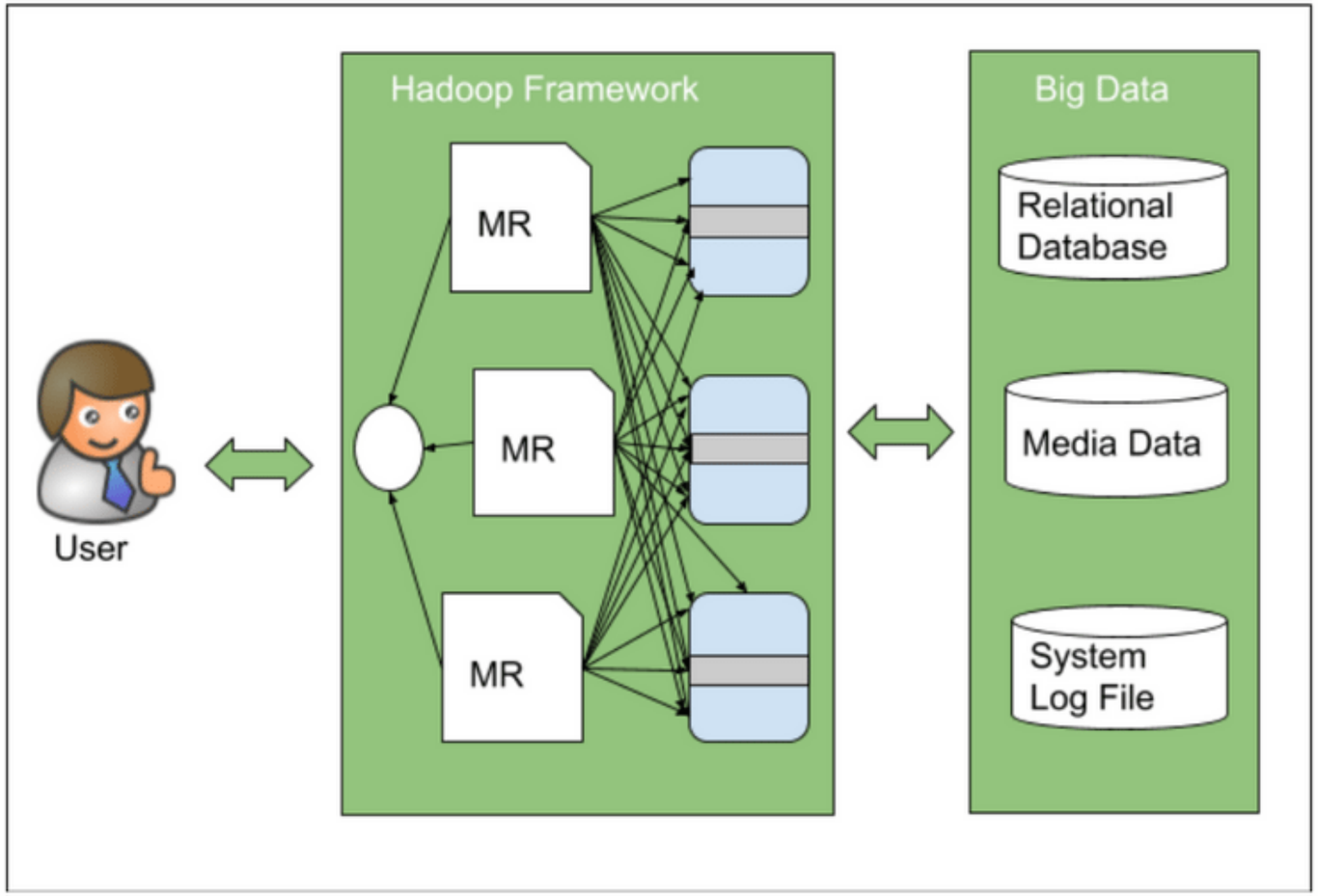
Using the solution provided by Google, Doug Cutting and his team developed an Open Source Project called HADOOP.

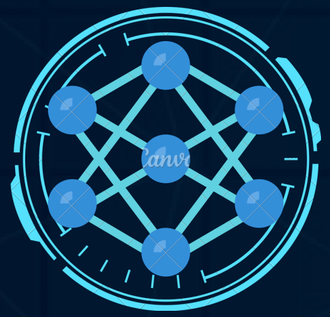
WLUG
COMMUNITY | KNOWLEDGE | SHARE

INTRODUCTION

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

HADOOP

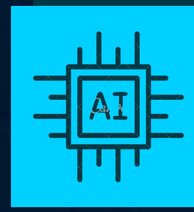

Hadoop Framework
MR
MR
MR
Big Data
Relational Database
Media Data
System Log File
User



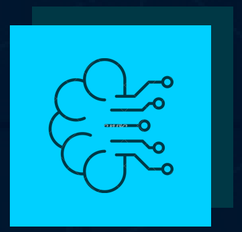
HADOOP

WLUG
COMMUNITY | KNOWLEDGE | SHARE

APACHE OPEN SOURCE FRAMEWORK



WRITTEN IN JAVA



MAP REDUCE ALGORITHM

WLUG
COMMUNITY | KNOWLEDGE | SHARE

HADOOP ARCHITECTURE

HADOOP ARCHITECTURE

WLUG
COMMUNITY | KNOWLEDGE | SHARE

HDFS

Hadoop Distributed File System

MAP REDUCE

Processing/ Computation layer

WLUG
COMMUNITY | KNOWLEDGE | SHARE

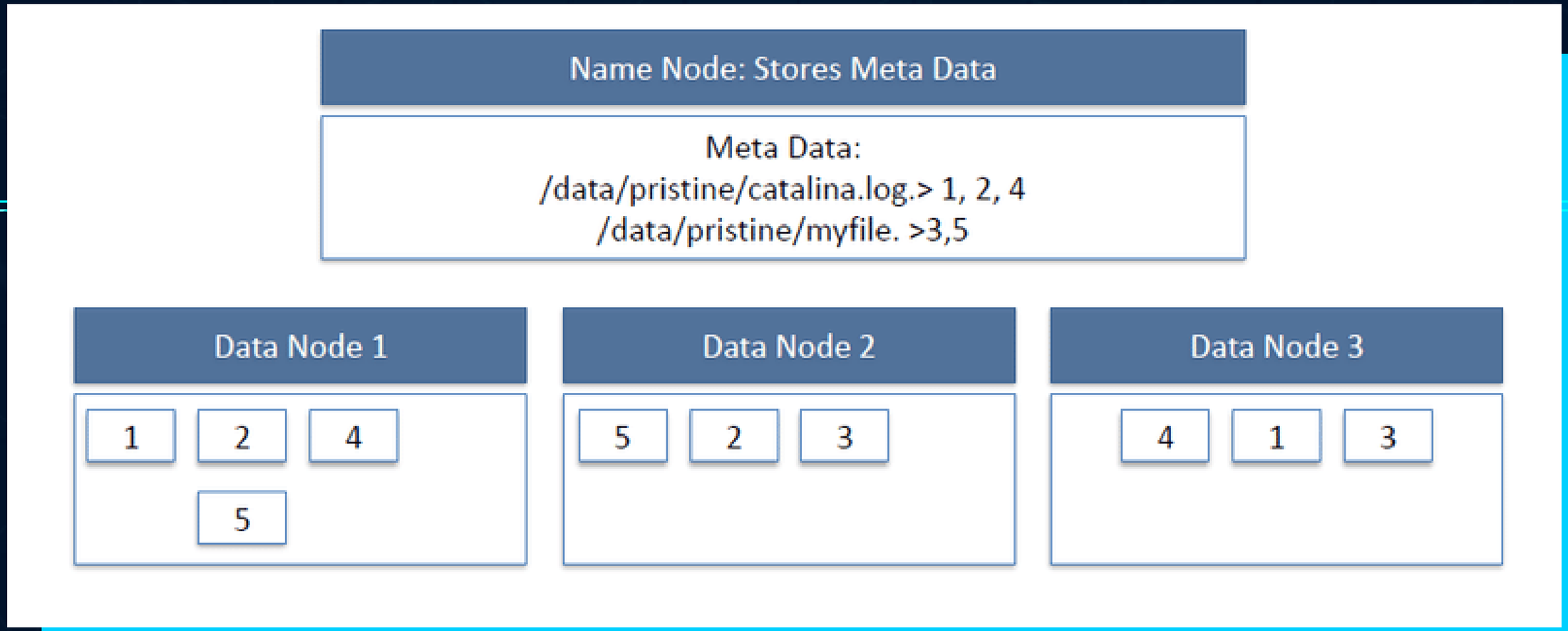
HADOOP DISTRIBUTED FILE SYSTEM

WLUG
COMMUNITY | KNOWLEDGE | SHARE

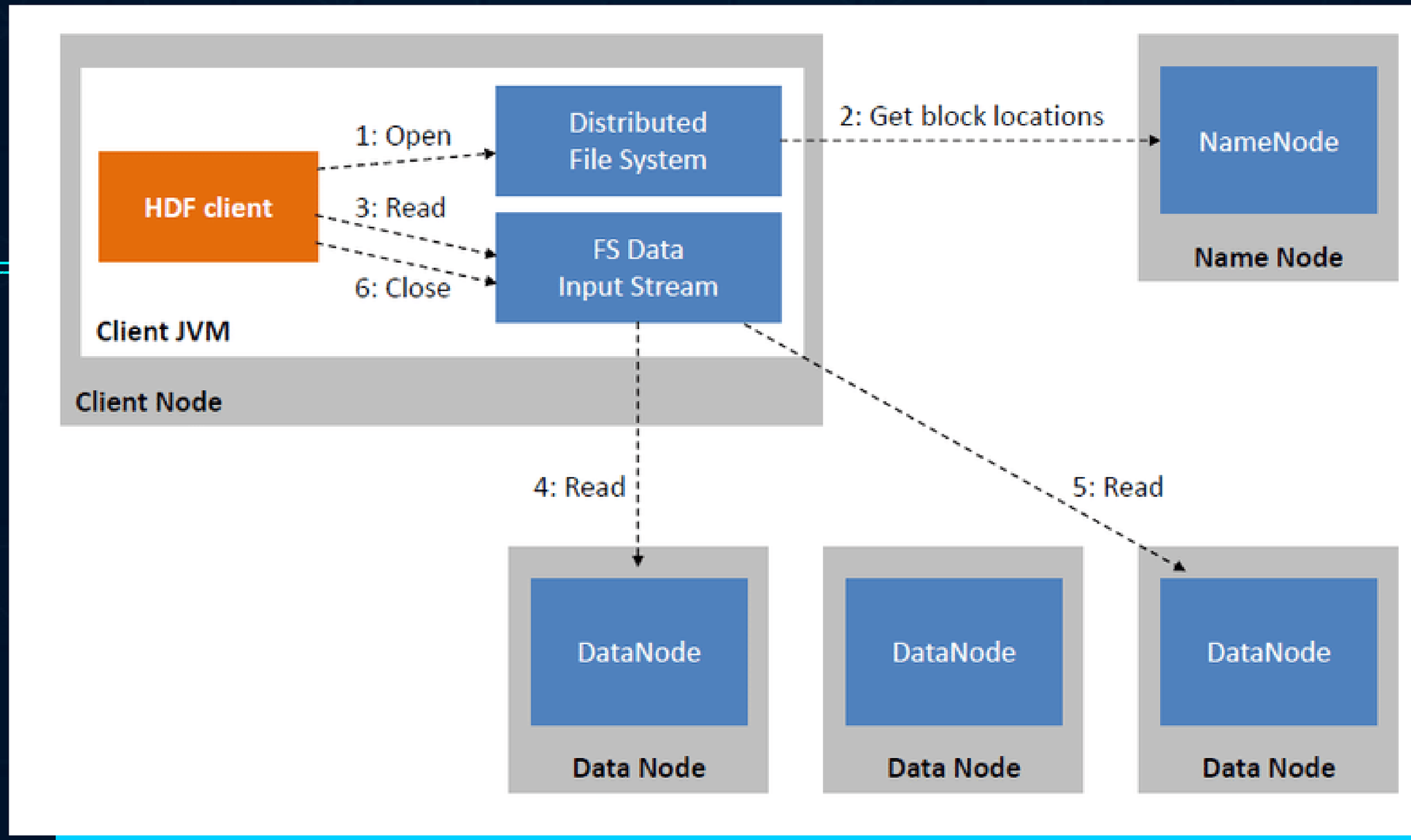
HDFS

- Based on Google File System.
- Some similarities with Existing Distributed File System
- Highly fault tolerant
- Can be deployed on low cost hardware
- High throughput access to application data
- Suitable for application having large datasets
- Provides a command line interface to interact with HDFS

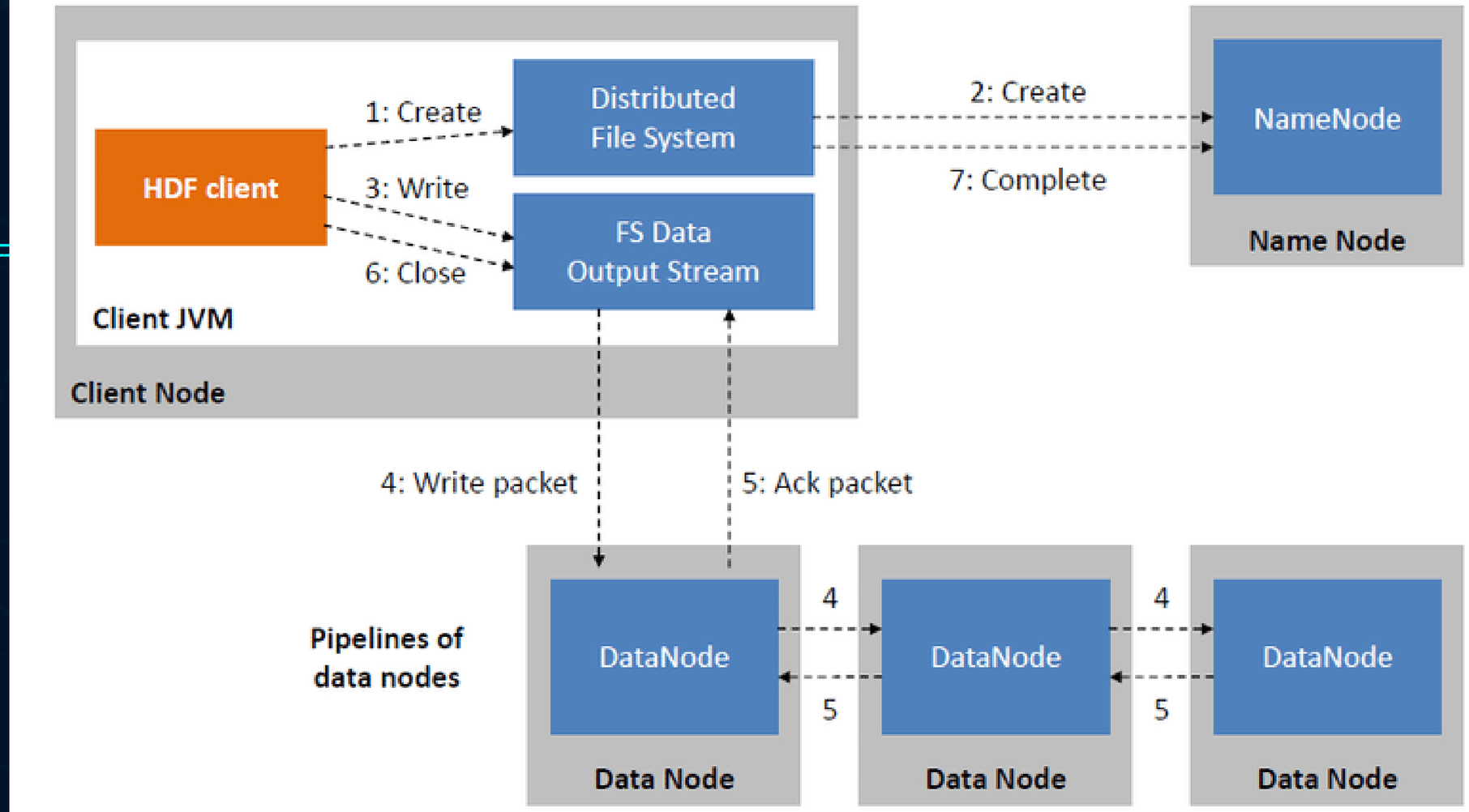
HDFS CONCEPTS



HDFS READ


HDF client
1: Open
3: Read
6: Close
Distributed File System
FS Data Input Stream
Client JVM
Client Node
2: Get block locations
NameNode
Name Node
4: Read
5: Read
DataNode
Data Node
DataNode
Data Node
DataNode
Data Node

HDFS WRITE

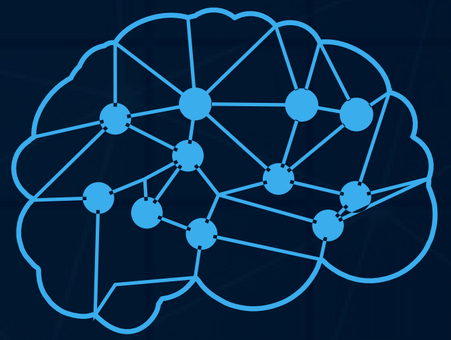

HDF client
1: Create
3: Write
6: Close
Distributed File System
FS Data Output Stream
Client JVM
Client Node
2: Create
7: Complete
NameNode
Name Node
4: Write packet
5: Ack packet
Pipelines of data nodes
DataNode
4
5
Data Node
DataNode
4
5
Data Node
DataNode
Data Node

WLUG
COMMUNITY | KNOWLEDGE | SHARE

HADOOP MAPREDUCE

WLUG
COMMUNITY | KNOWLEDGE | SHARE

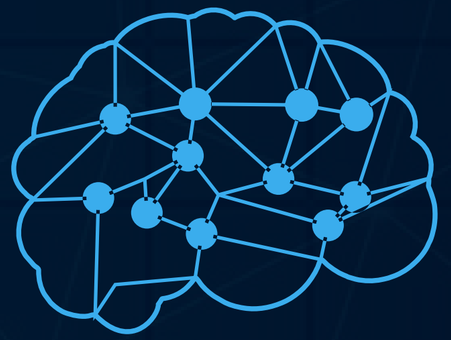
HADOOP MAPREDUCE



MapReduce is used for parallel processing of the Big Data, which is stored in HDFS

WLUG
COMMUNITY | KNOWLEDGE | SHARE

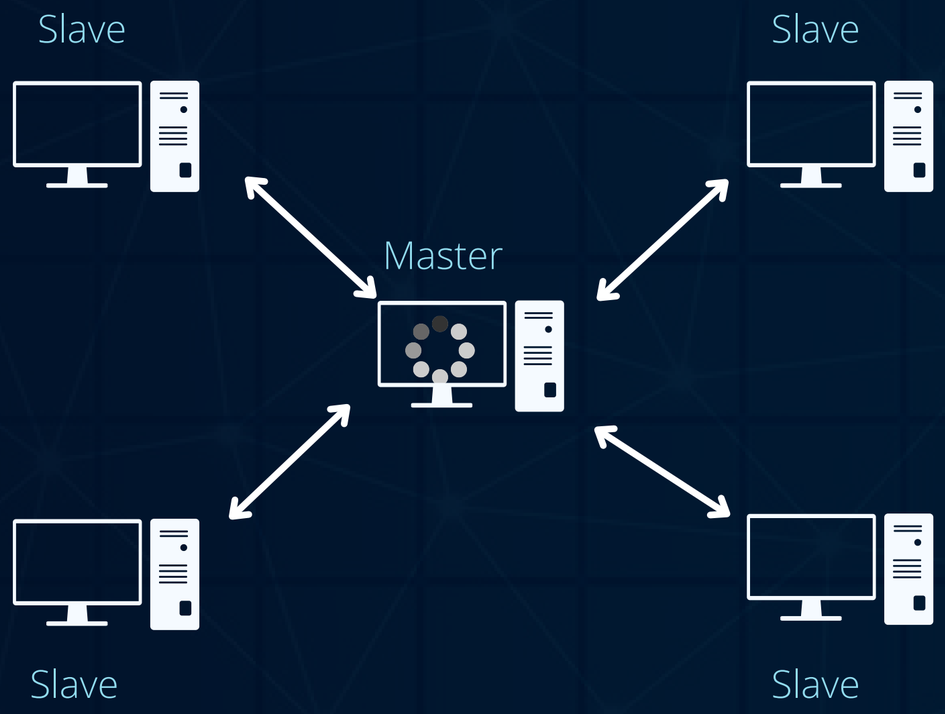
HADOOP MAPREDUCE



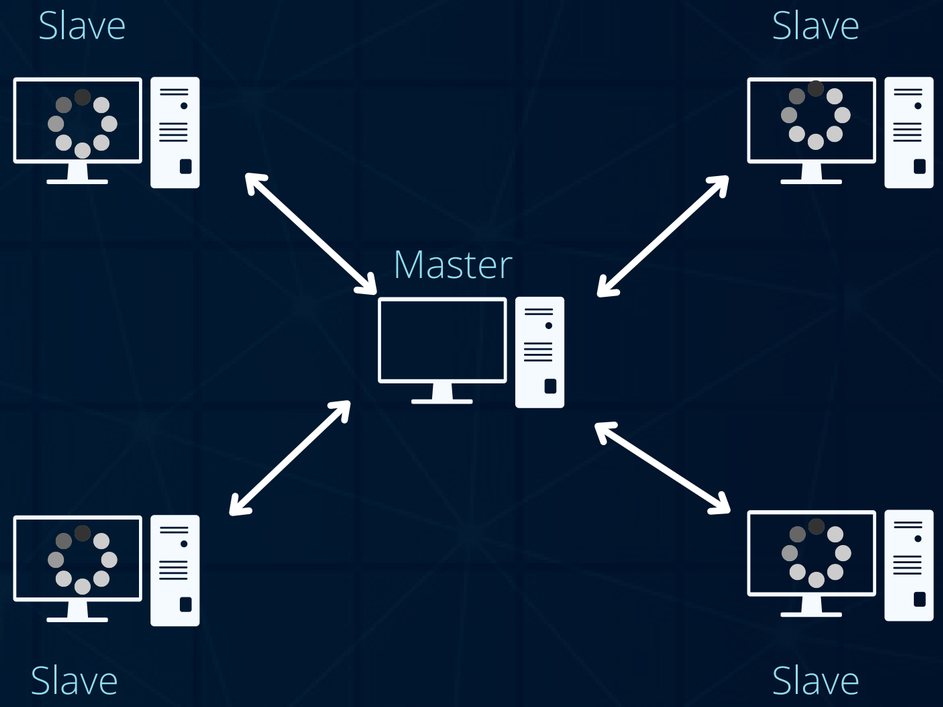
- Hadoop Mapreduce is a processing component for Hadoop
- It Processes Data in Distributed Environment

WLUG
COMMUNITY | KNOWLEDGE | SHARE

Traditional Approach
Data is processed at the Master

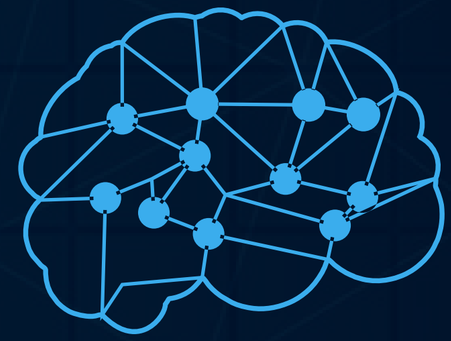


MapReduce Approach
Data is processed at the Slave Node



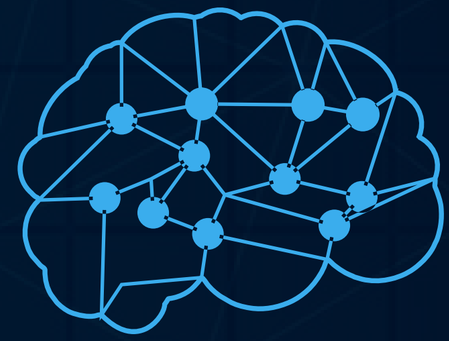
WLUG
COMMUNITY | KNOWLEDGE | SHARE

ADVANTAGE 1: PARALLEL PROCESSING



- Data is Processed in Parallel
- Processing becomes Faster

ADVANTAGE 2: DATA LOCALITY- PROCESSING TO STORAGE



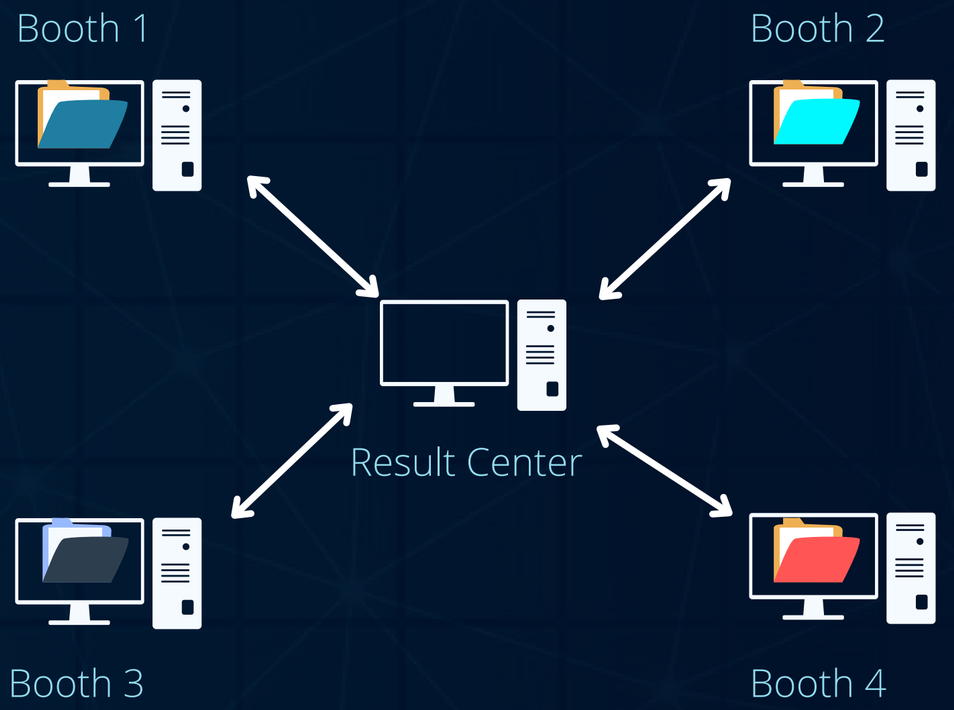
- Moving Data to Processing is Very Costly
- In MapReduce, we move Processing to Data

ELECTION VOTES COUNTING

Election Votes Casting-

- Votes are stored in Different Booths
- Result Center has result of all the booths.

Data

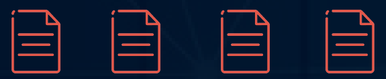


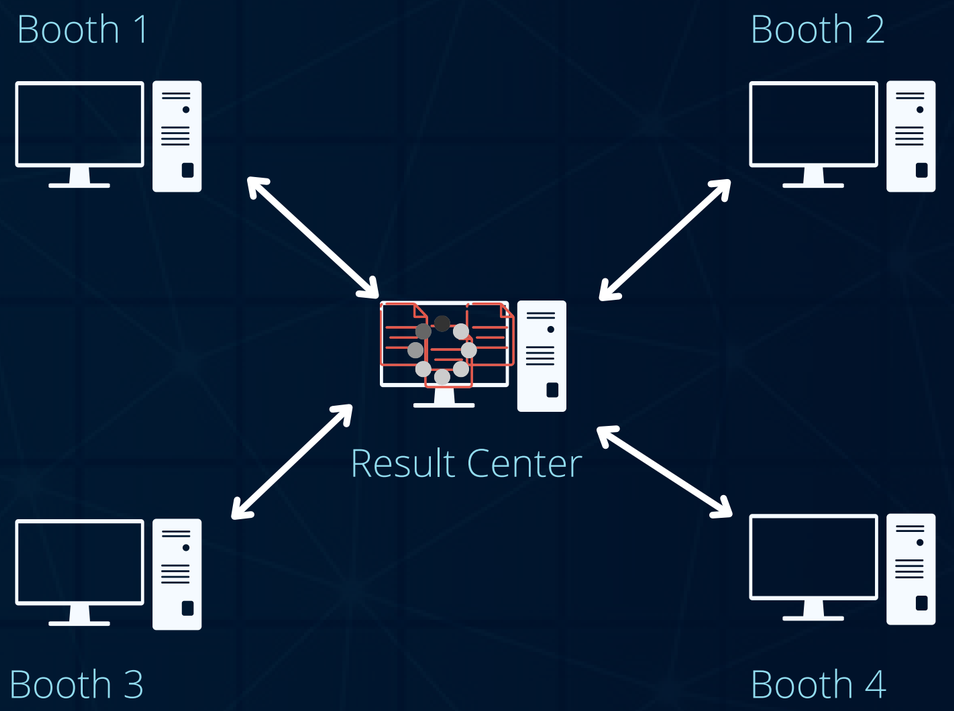
ELECTION VOTES COUNTING TRADITIONAL WAY

Counting - Traditional Approach

- Votes are moved to Result Center for counting
- Moving all the votes to Center is costly
- Result Center is over-burdened
- Counting takes time

Data



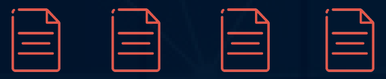


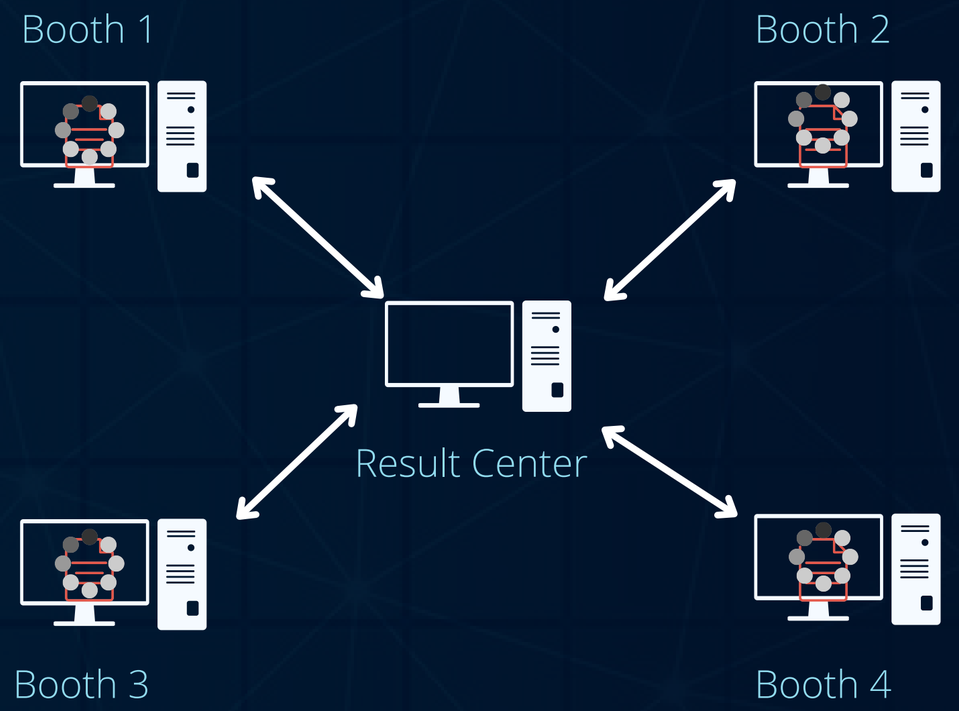
ELECTION VOTES COUNTING MAPREDUCE WAY

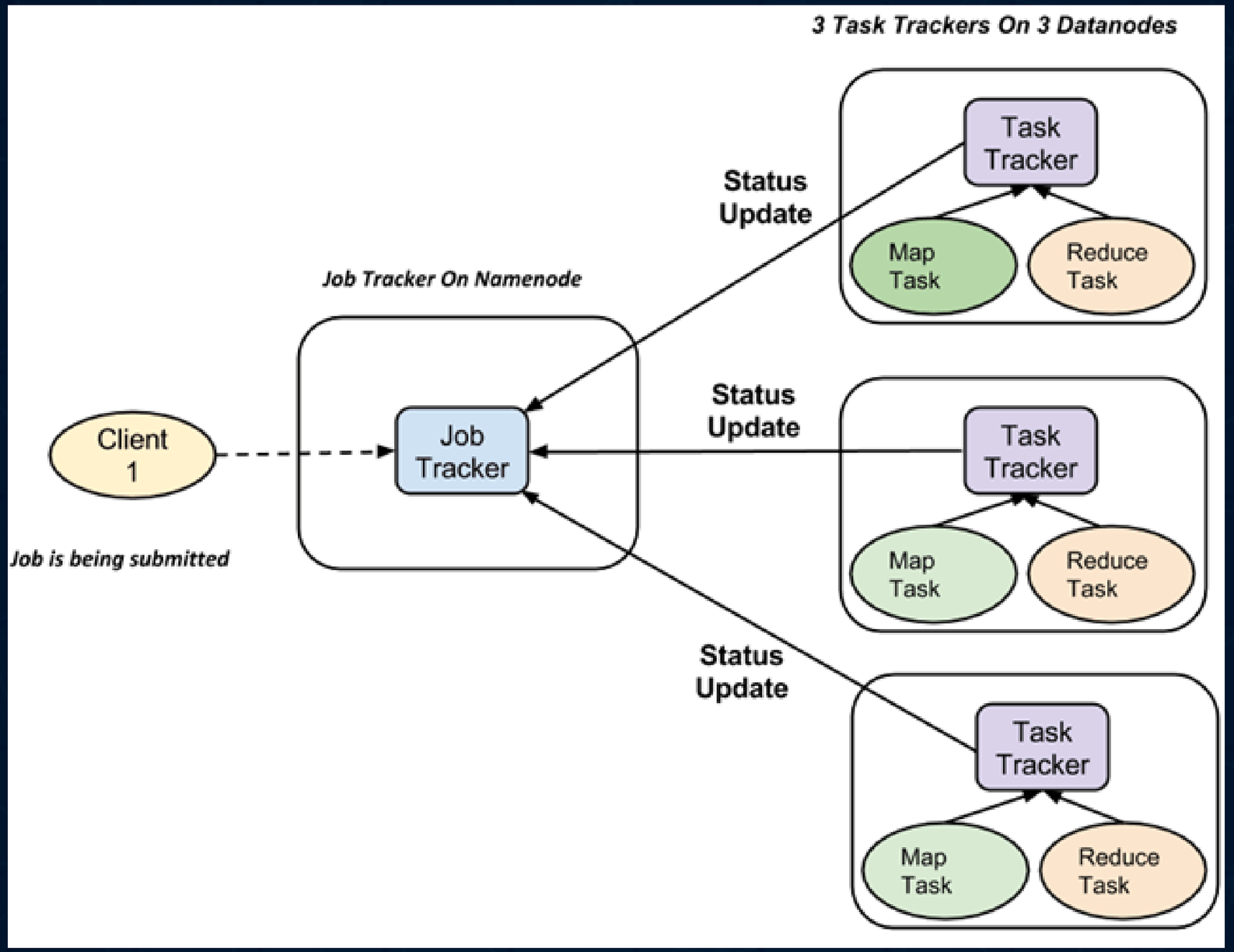
Counting - MapReduce Approach

- Votes are counted at individual booths
- Booth-wise results are sent back to the result center
- Final Result is declared easily and quickly using this way

Data

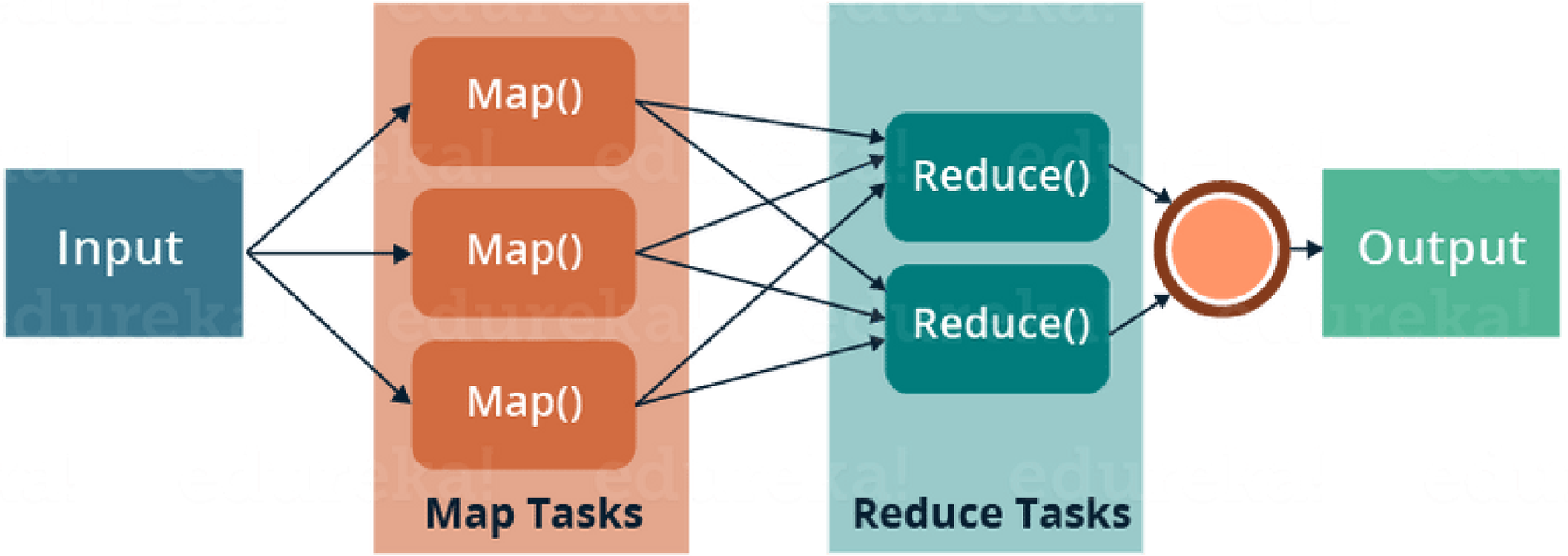




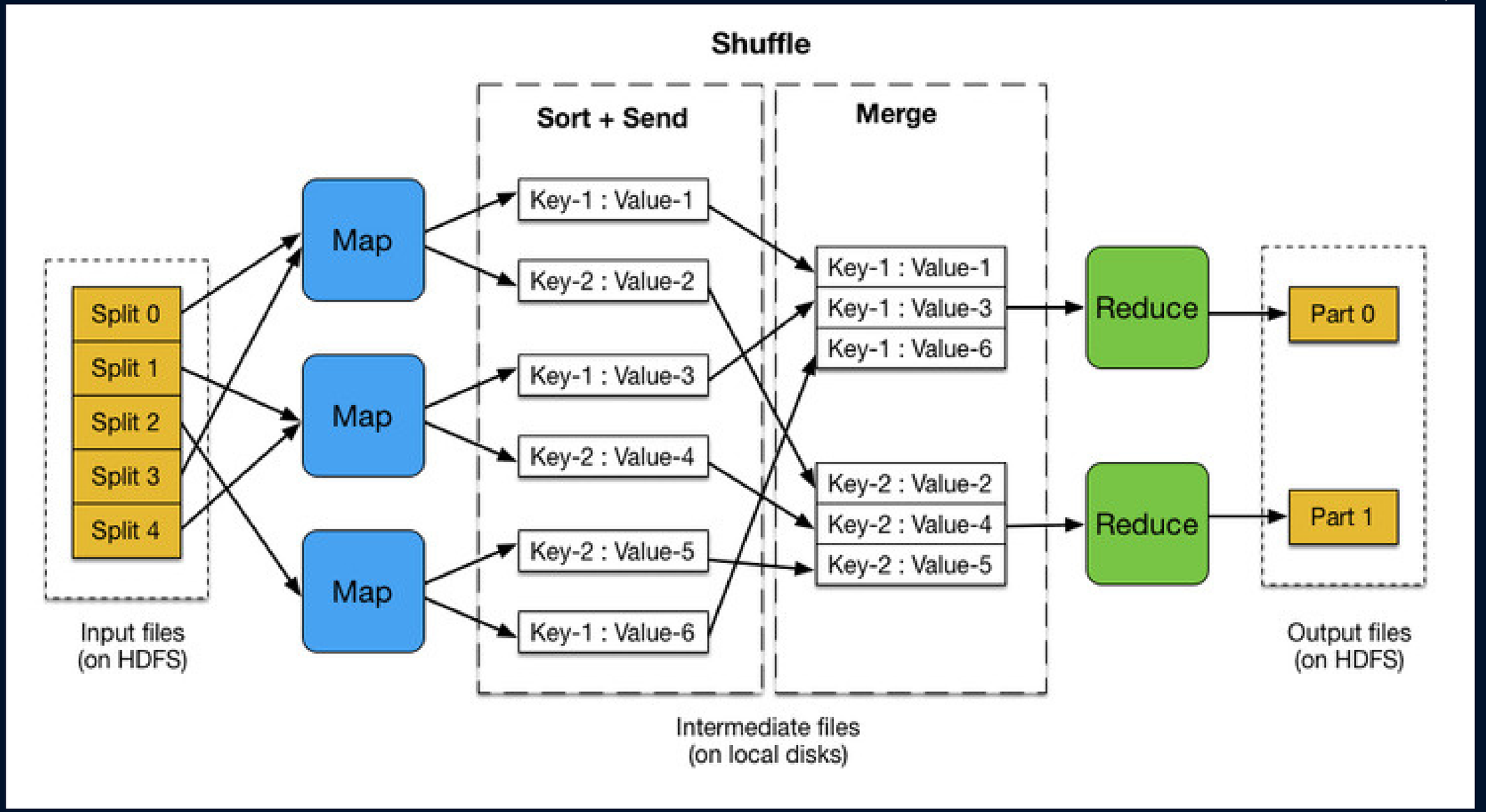


WLUG
COMMUNITY | KNOWLEDGE | SHARE

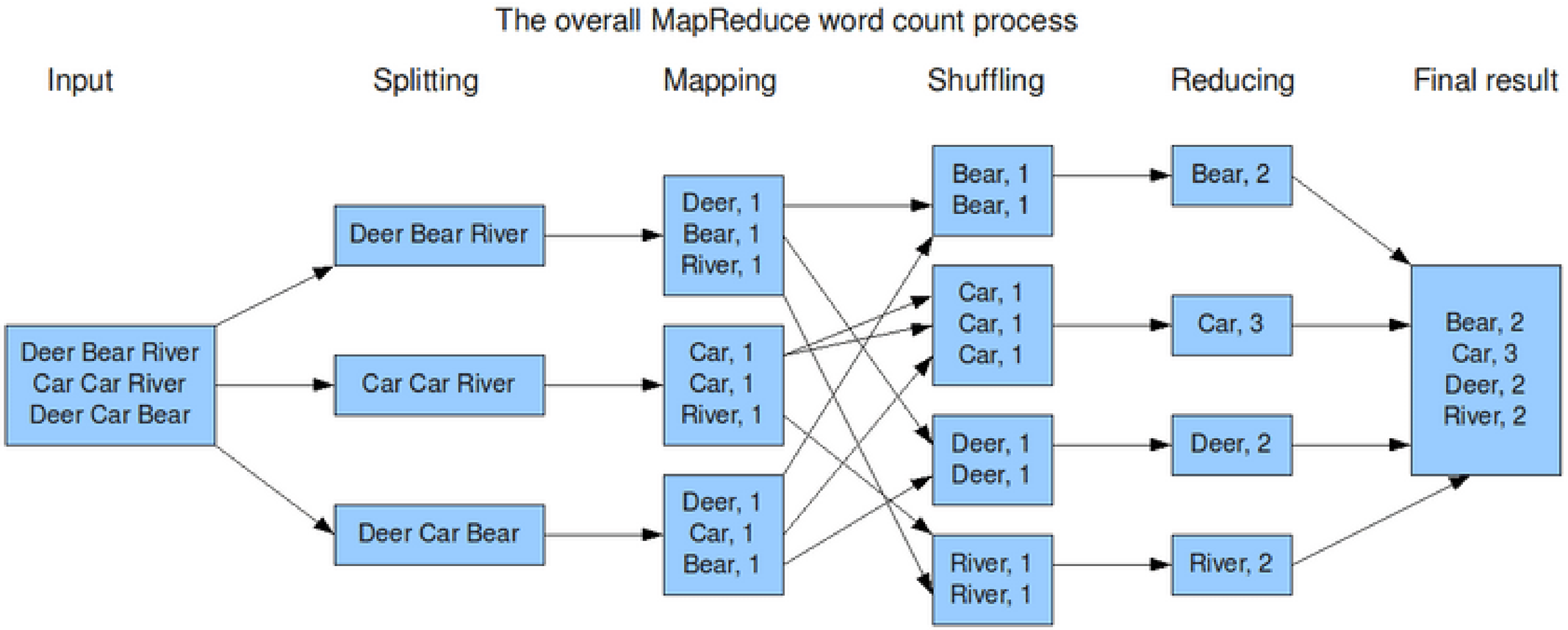
WLUG
COMMUNITY | KNOWLEDGE | SHARE



WLUG
COMMUNITY | KNOWLEDGE | SHARE



WLUG
COMMUNITY | KNOWLEDGE | SHARE

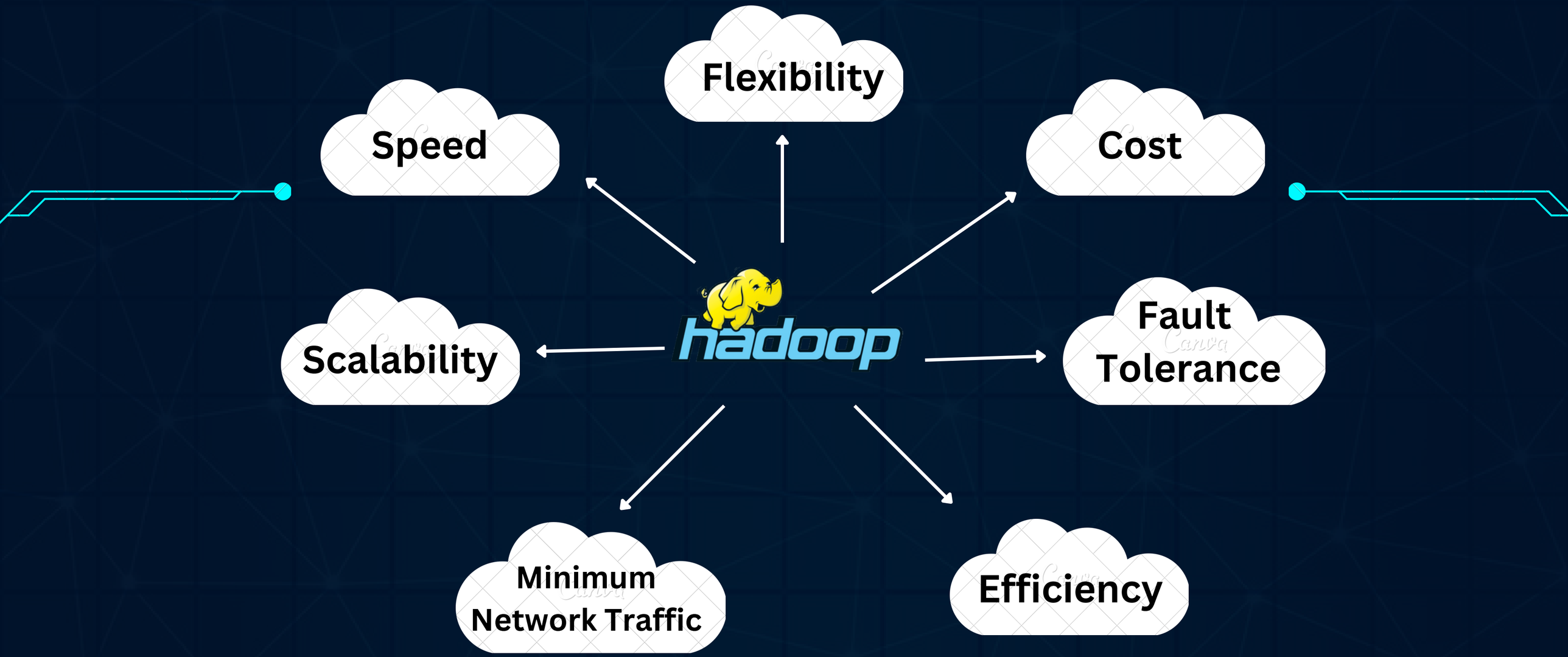


WLUG
COMMUNITY | KNOWLEDGE | SHARE

HADOOP ADVANTAGE

ADVANTAGES

WLUG
COMMUNITY | KNOWLEDGE | SHARE



WLUG
COMMUNITY | KNOWLEDGE | SHARE

THANK YOU

QUIZZ TIME!