RAISONI GROUP
— a vision beyond —

Assignment No. 1

Title: To install VMware for Ubunty operating system and perform basic ope commands.

Objective:

1] To install VMware for Ubunty Operating System.

2] Perform basic commands.

Theory:

Here are steps to install VMWare on Ubuntu along with commands:

1] Download VMware Workstation:
Go to VMware website and download VMWare workstation for Linux.

2] Install Dependencies:-
Open terminal and install necessary dependencies by running:

```
sudo apt update
sudo apt install build-essential gcc
linux-headers-$(uname -r)
```

RAISONI GROUP
— a vision beyond —

3] Make Installer Executable: Navigate to directory.
chmod +x VMWare-Workstation-*.bundle

4] Run installer:
Run installer using studio-
sudo ./VMWare-Workstation-*.bundle

5] Follow installation wizard:
Accepting license agreement and choosing installation directory.

6] Start VMWare Workstation

7] Basic Commands-
* vmrun- to control virtual machines from command line.

* vmware- main interface for starting VMware workstation.

* vmware-config-tools.pl- Configure VMware Tools after installing them in a guest OS.

Conclusion:
We have installed vmware for ubuntu and performed basic operations.

RAISONI GROUP
a vision beyond

Assignment No. 1

Aim: To install Hadoop framework configure it and set up a single node cluster, use web based tools to monitor your hadoop setup.

Objective:

1) To learn Hadoop distribution file system and its application.

2) Introduction of basic concept of Big data.

3) Understand different Big data tools and framework.

Theory:

Hadoop is an open source framework designed for distributed storage and processing of large datasets.

It provides a scalable, fault tolerance and cost effective solution for handling big data.

Hadoop is composed of several core components.

- Hadoop Distributed File System [HDFS]
- Yet Another Resource Negotiator [YARN]
- MapReduce
- Hadoop Eco-system

RAISONI GROUP
— a vision beyond —

* Install Hadoop

- Download latest release of Apache Hadoop.
- Extract downloaded archieve to your installation directory
- Setup Java Development as Hadoop require Java.
- Configure environment variable to path.

* Configure Hadoop

Configure Hadoop navigating all necessary steps including specifying JAVA and any other environment variable that you may need.

* Monitor Hadoop

- Hadoop provides web based tool for monitoring cluster
- Hadoop Name Node Web UI: Provides information about HDFS cluster including health of NameNode, file system metrics and more

* Access Hadoop Logs

- Monitor Hadoop logs located in log directory inside Hadoop installation directory.
- These log provides detailed information about cluster operations, errors and warning.

RAISONI GROUP
— a vision beyond —

* To verify Hadoop installation:

- Setup namenode using command hdfs namenode format
- Verify Hadoop dfs
- Verify Yarn script
- Access Hadoop on Browser
- Verify all Application for cluster.

* To use web based tool to monitor Hadoop steset up:

- Click managed entities in navigation panel.
- Add Hadoop cluster and Hadoop Node types to managed entity section.
- Click validate current document to check configuration.
- Click save current document to apply changes.

* Download JDK

- Go to oracle JDK download page.
- Select appropriate JDK version for operating system.

* Download installer package:

Install JDK
Run installer package you downloaded.
Follow installation wizard instruction.
Accept license agreement and choose installation directory.

RAISONI GROUP
— a vision beyond —

* Set JAVA-HOME Environment Variable

After installation you need to set JAVA-HOME environment variable to point your JDK.

* On Windows :

- Right click on my computer and select properties.
- Click on Advanced system setting on left side.
- Click on Environment variable button.
- Under system variable click new and add variable named JAVA-HOME with value set to path of your JDK.
- Click on save changes.

* Verify JDK installation :

Open new terminal or command prompt window. Use given command to verify java installed correctly

```
java -version
```

Conclusion :

Hence I have successfully installed hadoop framework and set up for single cluster node. Also used web based tool for monitoring hadoop setup.

RAISONI GROUP
— a vision beyond —

Assignment No 3

Aim: To implement file management tasks in Hadoop HDFS like adding, retrieving and deleting files.

Theory:

i] Create a directory in HDFS at given path:

Usage :- hadoop fs -mkdir <path>
example :- hadoop fs - mkdir /user/source code/dir/

ii] List contents of directory.
Usage : hadoop fs -ls <arge>
example : hadoop fs -ls /user/source code

iii] Upload and download a file in HDFS

Upload : hadoop fs-put
copy single src file, or multiple src file from local file system in Hadoop file system.
Usage : hadoop fs-put /home/source code/file/
user/source code /dir/

4) See content of file:
Same as unix cat command
Usage :- hadoop fs - cat (path[filename])
Example : hadoop fs - cat /user/source code/dir.
file.txt.

RAISONI GROUP
a vision beyond

5] Copy file from source to distination :-
This command is allowing multiple source as well in which case destination must be a directory.
Usage :- hadoop fs- cp <source> <dest>

6] Copy file from /to local file system to HDFS:
copyfromlocal
Usage : hadoop fs -copyfromlocal <local.src) url
Example : hadoop fs - copyfromlocal /home /source/user /source/ file.txt

Similarly to get command, except that the destination is restricted to local file reference.

7] Move file from source to destination:
Moving file across file system is not permitted.
Usage : hadoop fs-mv <src> <dist>

8] Remove a file or directory in HDFS:
Remove file specified or argument.
Delet directory only when it is empty.
Recurssion version of delet.
Usage : hadoop fs - rmv <arg>

RAISONI GROUP
— a vision beyond —

q] Display the aggregate length of file.
Usage: hadoop fs -du /user/sourcecode
/dir/file.txt

Conclusion: Hence we have successfully understood and implement file management task in Hadoop.

RAISONI GROUP
— a vision beyond —

Assignment No 3

Aim: To implement a word count application using Map Reduce Api.

Theory:

The entire mapreduce program can be fundamentally divided into 3 parts.
1) Mapper phase code
2) Reducer phase code
3) Runner code

1) Mapper phase code:
We create a class Map that extended class mapper, which is already defined in map reduce framework
We define the datatype of input and output key value pair after class declaration using angle brakets.

Both the input and output of mapper is key value pair.

Input: The key is nothing but offset of each line in text file.

Output: The key is tokenized word.

We have hadward value in our code: IntWritable.
Example: Pearl, Perl, etc.

2) Reducer code:
We created a class Reduce which extends class Reducer like that of Mapper.
We defined data type of input and output key/value pair after class declaration using angle brackets as done after Mapper.

3) Runner code:
In the runner class, we get configuration of our map reduce job to run in hadoop.
We also specify name of mapper and reducer.
The path of input and output folder is also specified.
The main method is entry point for driver. In this method we instantiable a new configuration object for job.

RAISONI GROUP
— a vision beyond —

Run MapReduce code:
The command for running MapReduce code is :-

hadoop-mapreduce-example-jar

Example:

hadoop jar hadoop-mapreduce-ex.jar word count /sample input /sample/output

Conclusion:

In this practical, we successfully implemented program to count the word in MapReduce program to understand in paradigm.

RAISONI GROUP
a vision beyond

Assignment No 5

Aim: Creating HDFS tables and loading them in Hive.

Theory:

Hive tables provide us schema to store data in various format (like csv). Hive provides multiple ways to add data to tables. We can use DML queries in Hive to import or add data to table. Once can also directly put table into hive with HDFS command.

In case we have data in Relational Database like Mysql, Oracle, IBM DB2, etc, then we can use Sqoop to efficiently transfer Petabyes of data between Hadoop and Hive. To perform above operation make sure your hive is running. Below are steps to launch hive on local system.

RAISONI GROUP
a vision beyond

Step 1: Start all your Hadoop Daemon.

start-dfs.sh - this will start namenode, datanode and secondary namenode

start-yarn.sh - this will start node manager and resource manager

jps - to check running elements

Step 2: Launch hive from terminal.

type hive - to launch hive in your local system

In hive with DML statement, we can add data.

- Using INSERT command
- Load Data Statement

Using INSERT Command.

Syntax: Insert into table <table name>
values(<add value as per column entry>);

Example: To insert data into table lets create table with name student.

Command:

```
Create table if not exist student (
  student_name string,
  student_rollno int,
  student marks float)
```

RAISONI GROUP
a vision beyond

ROW FORMAT DELIMITED
FIELDS TERMINATED BY ",";

We have successfully created student table in hive default database.

INSERT Query:

insert into table student values ('Pranav', '63', '90'), ('Sanskar', '67', '92'),

We can check data of student table with help of below command.

Select * from students;

2) Load data statement.
Hive provides us functionality to load pre created table either from our local file system or from HDFS. The load data statement is used to load data into hive table.

Syntax:
Load data [Local] inpath <The table data locations> [over write] into table <table-name>

RAISONI GROUP
— a vision beyond —

Commands:

cd/home/yuji/documents - To change directory
touch data.csv - To create data.csv file
nano data.csv - nano is linux command line editor to edit file.
cat data.csv - to see content in file.

Load data to student hive table with help of below command.

Load data Local path '/home/yuji/Documents/data.csv/ into table student;

Now lets see student table content to observe effect with help of below command.
select * from student.

We can observe that we have successfully added data to student table.

Conclusion:

Hence we have successfully understood and implemented creating HDFS table and load them in Hive.

RAISONI GROUP
— a vision beyond —

Assignment No 5

Title: To perform graph analysis and visualization using tableau.

Aim: To apply graph analytics and visualization using Tableau for comprehensive data exploration and insights.

Theory:

Graph analytics involves examination of relationship and pattern within interconnected data. Tableau a powerful data visualization tool, facilitates representation of graph data through interactive dashboard aiding in identification of trends and anomalies.

Steps:

1) Data preparation: Cleanse and structure graph data for optimal analysis within Tableau.

2) Connectivity: Explor Tableau connectivity option to steamlessly integrate diverse graph data source.

RAISONI GROUP
— a vision beyond —

3] Graph Analytics implementation: Apply advanced Graph analysis algorithm within Tableau for in depth pattern identification.

4] Dashboard creation: Develop visually compiling dashboard and reports to effectively communicate key graph analytics finding.

5] Interactivity: Leverage Tableau interactive feature to enable dynamic exploration of graph data and user driven insight.

6] Optimization: Ensure scalability by optimizing tableau performance to handle large scale graph dataset.

7] Collaboration: Integrate Tableau dashboard with extends platform to fascilitate collaborative decision making.

Objective:- Enhance team proficiency through user training session on graph analytic in Tableau.

Establish a feedback loop for continuous improvement in Tableau driven graph analytics process.

RAISONI GROUP
a vision beyond

Evaluating impact of Tableau driven graph analytics on decision making.

* Data Support:
Utilize diverse data sources, including social network, supply chain or any interconnected dataset to showcase versatility of graph analytics.

* Data Representation Format:
Visualize relationship through node link diagram, force directed layout, and other graph specific visualization within Tableau providing a comprehsive view of complex connection.

Conclusion:
Through integration of graph analytic and Tableau this approach enable a deeper understanding of interconnectio data empowering decision making with visually rich insight and foestering a data driven culture within the organization.

RAISONI GROUP
a vision beyond

Assignment No 7

Aim: To implement basic functions and commands in R programming better visualization than a data table.

Software Requirements: 1] R
2] R Studio
3] Windows/MAC/Linux

Theory:

R is an open source programming language that is widely used as a statistical software and data analysis tool.

R generally comes with command line interface. It is available across widely used platforms like windows, linux, macOS.

R programming is latest cutting edge tool.

Step 1: Install R

1] Download R installer from https://cran.rproject.org/

RAISONI GROUP
a vision beyond

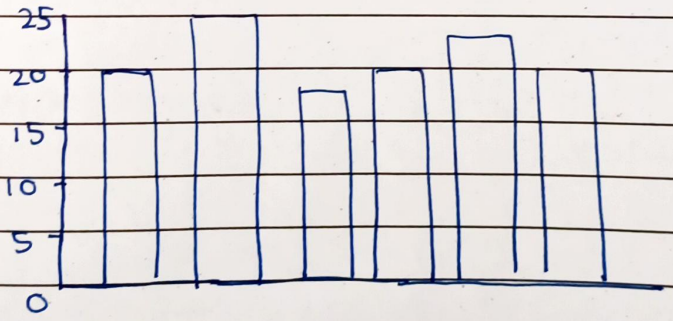
Step 2: Write a R / Python program to create a simple plot of five subjects marks.

For creating different type of barplot in R programming using both vector and matrix.

Barplots can be created in R using barplot() function.

We have vector of max temperature for seven days as
max.temp <— c(22, 27, 26, 25, 24, 23, 21)

Now we can make a bar plot out of this
barplot (max.temp)



we use main - to give title,
x lab & y lab - labels for axes,
names.arg — naming each bar
col — define color

We can plot horizontally by providing the argument horiz = TRUE

Plotting Categorical data:

Sometimes we have to plot count of each item as bar plots from categorical data.

For ex - vector of age of 10 clg freshmen

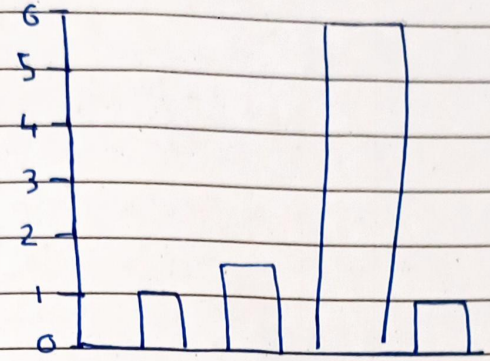
age <— c(17,18,18,17,18,19,18,16,18,18)

This count can be done using table() function

```
> table(age)
age
16 17 18 19
 1  2  6  1
```

RAISONI GROUP
a vision beyond

Age count of 10 students



Plotting Higher dimensional tables -

Sometimes the data is in form of contigency table.
For example, let us take built in Titanic dataset.
This dataset provides info on fate of passangers on fatal maiden voyage of ocean linear 'Titanic', summarized according to economic status (class), sex, age and survival R documentation.

margin.table(Titanic) - gives total count if index is not.

barplot(margin.table(Titanic, 4) - survival

barplot(margin.table(Titanic, 2) - male vs female count.

RAISONI GROUP
— a vision beyond —

Plot barplot with matrix:

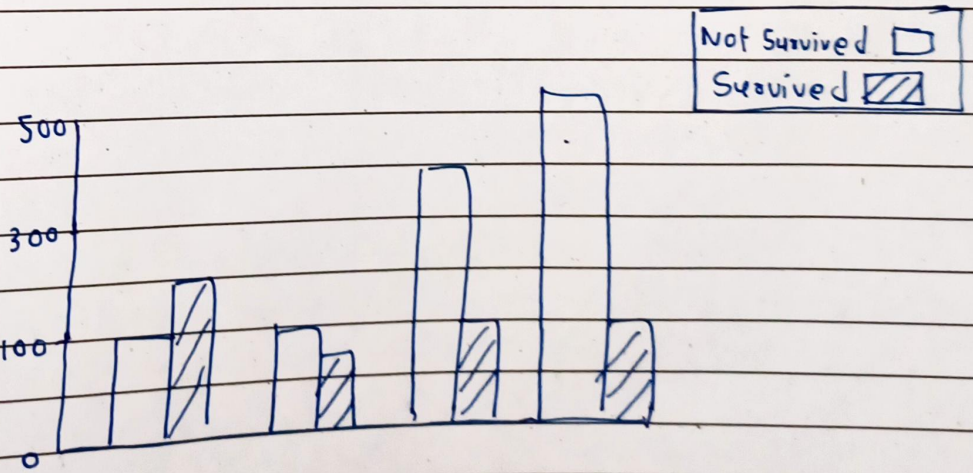
Each column of matrix will be represented by a stocked bar.

```
main = "Survival of each class")
xlab = "class"
col = c("red", "green")

legend ("topleft", c("Not Survived," "survived"),
fill = c ("red", "green")
```

legend() function is used to appropriately display legend.

Column justaposed by specifying parameter beside = TRUE

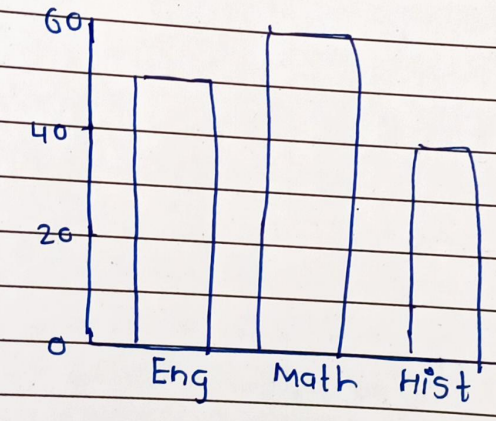


RAISONI GROUP
a vision beyond

Program -

```
marks = c(70,95,80,74)
barplot(marks,
main = "Comparing marks of 5 subject"
xlab = "Marks"
ylab = "Subject"
names.arg = c("English","Math","Hist")
col = "darkred"
horiz = FALSE)
```

O/p :



Conclusion:

We have implement basic functions and commands in R programming better visualization than data table.

RAISONI GROUP
— a vision beyond —

Assignment No. 8

Aim: To use following platform for solving only big data analytics problem of your choice, Amazon Web Service, Microsoft Azure, Google.

Objective: To solve big data analytics problem using cloud platform like Amazon web service, Microsoft Azure, Google.

Theory:

Big data analytics deals with extracting insights from massive datasets that aree too voluminous and complex for traditional dates processing methods.

Cloud platform like AWS offer s calable and cost efficient solution for big data needs.

They provide scite of services for data ingestion (Eg: S3 Storage), S3 glacier, processing.

Conclusion:

In this way, we have successfully implemented AWS in data analytics problem