Assignment No: 1

Aim :- To install VMware for ~~Ubu~~ ubuntu operating system and perform basic commands.

Objective :-

1) To install VMware for Ubuntu operating system. perform basic commands.

Theory :-

Here are the steps to install VMware on ~~Ubu~~ Ubu along with ~~&~~ some commands :-

① Download VMWare workstation -
Go to VMware website and download the VMware workstation for Linux.

② Install Dependencies :-
open a terminal and install necessary dependencies running :

```
sudo apt update
sudo apt install
```

③ Make Installer Executable : Navigate to directory

```
chmod + x VMware - Workstation - x bundle.
```

④ Run installer :-
Run the installer using studo -

```
sudo. /VM ware - Workstation - * . bundle
```

⑤ Follow Installation Wizard :
accepting license agreement & choosing installat directory

⑥ start VMware workstation

⑦ Basic commands :-

Assignment No : 1 .

Aim :- To install VMware for ~~tbut~~ ubuntu operating system and perform basic commands.

Objective :-
① To install VMWare for Ubuntu operating system.
② perform basic commands.

Theory :-
Here are the steps to install VMware on ~~Ubt~~ Ubuntu along with ~~&~~ some commands :-
① Download VMWare workstation -
Go to VMware website and download the VMware workstation for Linux.
② Install Dependencies :-
open a terminal and install necessary dependencies by running :
sudo apt update
sudo apt install
③ Make Installer Executable : Navigate to directory
chmod + x VMware - Workstation - x bundle.
④ Run installer :-
Run the installer using studo -
sudo . VMware - Workstation - * . bundle
⑤ Follow Installation Wizard :
accepting license agreement & 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.

Practical 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 solutions for handling big data,

Hadoop is composed of several core components

- Hadoop Distributed File System [HDFS]
- Yet Another Resource Negotiation [YARN]
- Map Reduce
- Hadoop Eco-system.

*Install Hadoop.

- Download latest release of Apache Hadoop.
- Extract download 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 includling 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.

* To verify Hadoop installation :-
- Set up 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 set up

- Click managed entities in navigatior 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 installer package.

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

* 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 now 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 & set up for single cluster ~~me~~ node. Also used web based tool for monitoring Hadoop setup.

Practical No.3.

Aim :- File Management task in Hadoop.

Theory :-

i) Create a directory in HDFS at given path :
usuage :- hadoop Fs- mkelir <path>
example : hadoop Fs- mkelir/user/sourcecode/dir/

ii) List the contents of a directory
usuage : hadoop Fs- ls <are>
example :- hadoop Fs- ls /user/source code

iii) Upload and download a file in HDFS
upload :- hadoop Fs-put
copy single src file, or multipule src files from local file system in the Hadoop file system
usage :- hadoop . Fs-put /home/source code/ file.txt/
user/source code./ clirt

, Download : hadoop Fs - get
copies/ Download Files to local file system
usage : hadoop Fs - ger <hdfs-src><local det>
example : hadoop fs- ger /user/source code/client/
home/ source code/

4) See content of a file :-
same as unix cat command
usage : hadoop Fs - car< path [Filename]>
example : hadoop Fs - car [user/source code/clin/ file.txt

5] Copy the File from source to clistination :-
This command is allowing multiple sources as well in which case the distination must be a directory

usage :- hadoop Fs - cp <source><dest>

6) Copy File from / to local File system to HDFS.

- copy fromlocal

usage : hadoop Fs - copyfromlocal <local.src> url
example : hadoop Fs - copyFromlocal /home/ source/user/ /source / abc.txt

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

7) Move File from source to distination :-
Moving File across Filesystem is not permitted
usage : hadoop Fs - mv <src><dist>

8) Remove a file or ~~elim~~ directory in HDFS
Remove Files specified as argument. Delet directory only when it is empty.
usage : hadoop Fs - mv <arg>
Recursion version of delet
usage : hadoop Fs - rmv <arg>

9) Display the agg regate dength of a file :
usage :- hadoop fe - du/user/sourcecode/dir/
file.txt

Conclusion :- Hence, we have successfully understood and implemented the file management & rast in Hadoop.

Practical NO 4

Aim :- Implement word count mapreduce program to understand mapreduce paradigm.

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 extend the class mapper, which is already defined in mapreduce Frameworks
- We define the datatype of input and output key/value per after the class declaration using angle brakets.
- Both the input and output of Mapper is a key value pair.

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

• Output : The key is tokanized word.

• We have the hardwed value in our code 1:Int writable. Example : Dear/, Ber/, etc.

(II) Reducer code :-

- We created a class Reduce which extends class Reducer like that of Mapper.

-we defined the data type of input and output key/value pair after the class declaration using angle brackets as done after for Mapper.

(iii) Runner code :-

- In the runner class, we get configuration of our mapreduce Job to run in hadoop.
- We also specify the name of mapper and reducer.
- The path of input & output folders is also specified.
- The main method is the entry point for the driver, In this method we instantiable a new configuration object for Job.

Run the MapReduce Code :-

The command for running a 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 the program to count the word in MapReduce program to understand in paraligm.

Assignment No : 5

Aim :- Creating HDFS tables & 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, them we can use Sqoop to effeciently transfer Petabytes of data between Hadoop & Hive. To perform above operation make sure your hive is running. Below are steps to launch hive on local system.

Step1 : Start all your Hadoop Daemon
start-dfssh - this will start namenode, datanode and secondary namenode.
start-yarn-sh - this will start node manager and resources manager.
jps - to check running daemons.

Step2 : Launch hive from terminal
type hive - to launch hive in you 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 )

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', '7', '92'),
```

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

```
Select * from students,
```

* load data statement :-

Hive provides us functionality to load precreated 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_names>

commands :-

cd/home/yuji/documents - To change directory

touch data.csv - To create data.csv file

naro 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 Localpath '/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.

Assignment No. 6.

Title : To perform graph analysis and visualization using tableau.

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

Theory :-
Graph analytics involves examination of relationship & 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.

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 yeature to enable dynamic exploration of graph data and user driven insight

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

⑦ Collaboration :- Integrate Tableau dashboard with extends platyorm to yacilitate collaborative decision making.

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

Establish a yeedback loop yor contineous improvement in Tableau driven graph analytics process.
Evaluating impact oy Tableau graph analytics on decision.

- 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 interconnection data empowering decision making with visually rich insight & fostering a data driven culture within the organization.

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

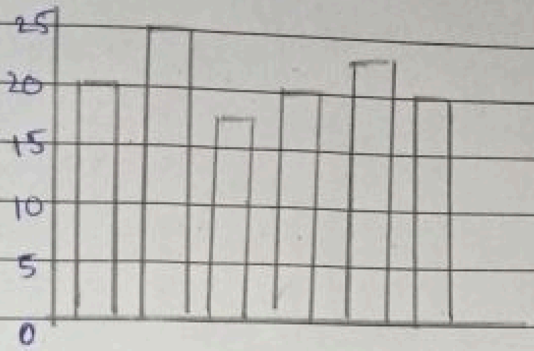
i] Download R installer from : https://cran.rproject.org/

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)
```

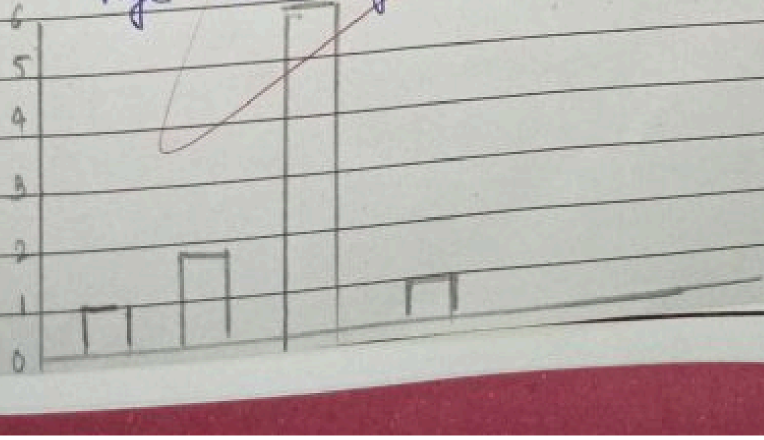
Now we can make a bar plot out of this
barplot (max. temp)



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 12 61

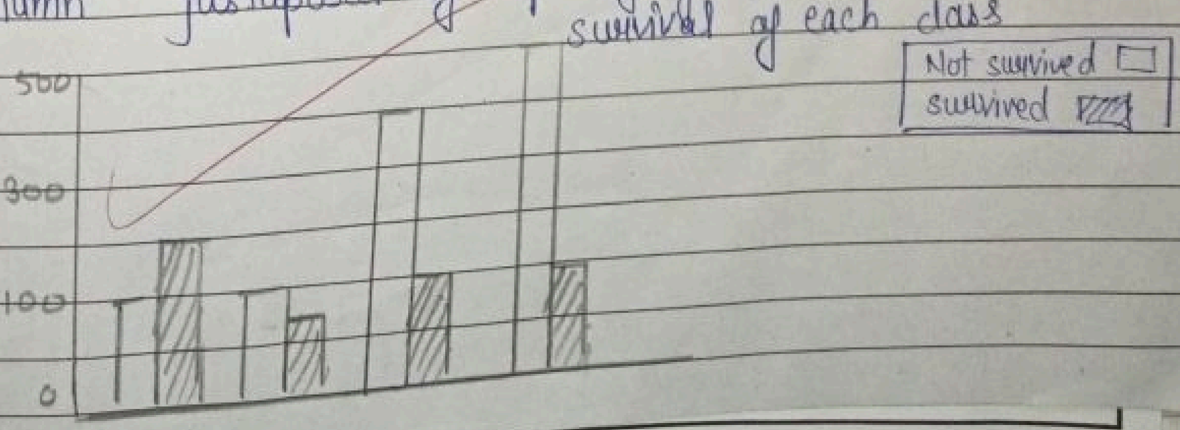
Age count of 10 students.



Sometimes the data is in form of contigency table for example, let us take built in Titanic dataset.
This dataset provides info on fate or passangers on fatal maiden vayage 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.

Plot barplot with matrix :-
Each column of matrix will be represented by a stacked 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

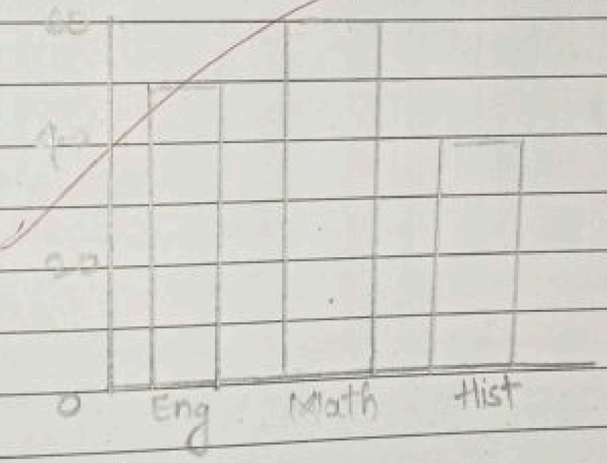


Program :-

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

O/p :-

comparing marks of 5 subjects.



Conclusion :-

We have implement basic functions

Assigoment No : 8

Aim : To use following paltform for solving only big data analytics problem of your choice, Amazon web services, Microsoft Azure, Google.

Objective :- To solve a big data analytics problem using an cloud platform like Amazon web services, Microsoft Azure, Google.

Theory :-

Big data analytics deals with extracting insights from massive dataset that are too voluminous and explain for traditional dates processing methods cloud platform like AWS offer scalable & cost efficient solution for big data needs. They provide a sate of services for data ingestion (eg. s3 storage), s3 glacier, processing.

Conclusion :-

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