

314457: Data Science and Big Data Analytics Laboratory

Third Year – Information Technology
(2019 Course)
Academic Year 2023-24 Semester – II

Teaching Scheme:	Credit Scheme:	Examination Scheme:	
Practical (PR): 02 Hrs./week	01 Credit	PR: 25 Marks	TW: 25 Marks



LABORATORY MANUAL V 3.0

DEPARTMENT OF INFORMATION TECHNOLOGY

Sinhgad College of Engineering, Pune

2023-2024

VISION

To provide excellent Information Technology education by building strong teaching and research environment.

MISSION

- 1) To transform the students into innovative, competent and high quality IT professionals to meet the growing global challenges.
- 2) To achieve and impart quality education with an emphasis on practical skills and social relevance.
- 3) To endeavour for continuous up-gradation of technical expertise of students to cater to the needs of the society.
- 4) To achieve an effective interaction with industry for mutual benefits.

PROGRAM EDUCATIONAL OBJECTIVES

The students of the Information Technology course after passing out will:

Sr. No.	Description
PEO1	Possess strong fundamental concepts in mathematics, science, engineering, and Technology to address technological challenges.
PEO2	Possess knowledge and skills in the field of Computer Science and Information Technology for analysing, designing, and implementing complex engineering problems of any domain with innovative approaches.
PEO3	Possess an attitude and aptitude for research, entrepreneurship, and higher studies in the field of Computer Science and Information Technology.
PEO4	Have a commitment to ethical practices, societal contributions through communities, and life-long learning.
PEO5	Possess better communication, presentation, time management, and team work skills leading to responsible & competent professionals and will be able to address challenges in the field of IT at the global level.

PROGRAM OUTCOMES

The students in the Information Technology course are expected to know and be able to:

Sr. No.	PO's	Description
PO1	Engineering knowledge	An ability to apply knowledge of mathematics, computing, science, engineering and technology.
PO2	Problem analysis	An ability to define a problem and provide a systematic solution with the help of conducting experiments, analysing the problem and interpreting the data.
PO3	Design/Development of Solutions	An ability to design, implement, and evaluate software or a software/hardware system, component, or process to meet desired need switch in realistic constraints.
PO4	Conduct Investigation of Complex Problems	An ability to identify, formulate, and provide essay schematic solutions to complex engineering /Technology problems.
PO5	Modern Tool Usage	An ability to use the techniques, skills, and modern engineering technology tools, and standard processes necessary for practice as an IT professional.
PO6	The Engineer and Society	An ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modelling and design of computer-based systems with necessary constraints and assumptions.
PO7	Environment and Sustainability	An ability to analyse and provide solutions for the local and global impact of information technology on individuals, organizations, and society.
PO8	Ethics	An ability to understand professional, ethical, legal, security and social issues and responsibilities.
PO9	Individual and Team Work	An ability to function effectively as an individual or a sate am member to accomplish a desired goal(s).
PO10	Communication Skills	An ability to engage in life-long learning and continuing professional development to cope up with fast changes in the technologies /tools with the help of electives, profession along animations and extra- curricular activities.
PO11	Project Management and Finance	An ability to communicate effectively in engineering community at large by means of effective presentations, report writing, paper publications, demonstrations.
PO12	Life-long Learning	An ability to understand engineering, management, financial aspects, performance, optimizations and time complexity necessary for professional practice.

PROGRAM SPECIFIC OUTCOMES

A graduate of the Information Technology Program will demonstrate-

Sr. No.	Description
PSO1	An ability to apply the theoretical concepts and practical knowledge of Information Technology in analysis, design, development and management of information processing systems and applications in the interdisciplinary domain.
PSO2	Decision making skills through the use of modern IT tools to make ready for professional responsibilities.

DOCUMENT CONTROL

Reference Code	SCOE-IT / Lab Manual Procedures
Version No	3.0
Compliance Status	Complete
Date of Compliance	01-12-2023
Security Classification	Department Specific
Document Status	Definitive
Review Period	Yearly

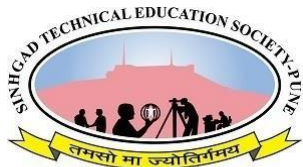
	Author
Signature	
Name	Mrs. S. S. Gadekar, Mrs. T. H. Patil
Designation	Assistant Professor

Document History

Revision No.	Revision Date	Reason For Change
1	01-07-2021	Update
2	01-07-2022	Update
3	01-02-2023	Updating the theory content of a few assignments

Summary of Changes to Data Science & Big Data Analytics Laboratory

Sr. No	Changes	Change type
1	Assignment Group A - 2	Theory content added & Changes in Code
2	Assignment Group B - 5	Theory content added



Sinhgad Institutes

Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE
S. No. 44/1, Off Sinhgad Road, Vadgaon(BK), Pune- 411041
Accredited by NAAC with Grade 'A+'

DEPARTMENT OF INFORMATION TECHNOLOGY

LABORATORY CODE

Sr. No.	Laboratory Code
1	Students should report to the concerned laboratory as per the time table.
2	Keep your bags in rack.
3	While entering in lab remove your shoes and keep it in shoe stand.
4	Turn computer monitors off when asked by your teacher
5	Do not go on banned websites.
6	No food or drinks near the keyboard.
7	Only use your assigned computer and workstation.
8	Do not change the settings on the computer
9	Ask permission to download.
10	Ask permission to print documents
11	Save your work often.
12	If you are the last class of the day, please POWER DOWN all computers and monitors.



Sinhgad Institutes

**Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE**

S. No. 44/1, Off Sinhgad Road, Vadgaon(BK), Pune- 411041

Accredited by NAAC with Grade 'A+'

ACADEMIC YEAR 2023-24, SEMESTER-II

DEPARTMENT OF INFORMATION TECHNOLOGY

SYLLABUS

Savitribai Phule Pune University, Pune

Third Year Information Technology (2019 Course)

314457: DS & BDA Lab

Teaching Scheme:	Credit Scheme:	Examination Scheme:
Practical (PR) : 02 hrs./week	01 Credit	PR :25 Marks TW : 25 Marks

Prerequisite Courses:

- Discrete mathematics
- Database Management Systems, Data warehousing, Data mining
- Programming in Python

Course Objectives:

1. To understand Big data primitives and fundamentals.
2. To understand the different Big data processing techniques.
3. To understand and apply the Analytical concept of Big data using Python.
4. To understand different data visualization techniques for Big Data.
5. To understand the application and impact of Big Data.
6. To understand emerging trends in Big data analytics.

Course Outcomes:

On completion of the course, students will be able to–

CO1: Apply Big data primitives and fundamentals for application development.

CO2: Explore different Big data processing techniques with use cases.

CO3: Apply the Analytical concept of Big data using Python.

CO4: Visualize the Big Data using Tableau.

CO5: Design algorithms and techniques for Big data analytics.

CO6: Design and develop Big data analytic application for emerging trends.

List of Laboratory Assignments

Group A: Assignments based on the Hadoop

1. Single node/Multiple node Hadoop Installation.
2. Design a distributed application using MapReduce (Using Java) which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet and process it using a pseudo distribution mode on Hadoop platform.
3. Write an application using HiveQL for flight information system which will include
 - A. Creating, Dropping, and altering Database tables.
 - B. Creating an external Hive table.
 - C. Load table with data, insert new values and field in the table, Join tables with Hive
 - D. Create index on Flight Information Table
 - E. Find the average departure delay per day in 2008.

Group B: Assignments based on Data Analytics using Python

1. Perform the following operations using Python on the Facebook metrics data sets
 - a. Create data subsets
 - b. Merge Data
 - c. Sort Data
 - d. Transposing Data
 - e. Shape and reshape Data
2. Perform the following operations using Python on the Air quality and Heart Diseases data sets
 - a. Data cleaning
 - b. Data integration
 - c. Data transformation
 - d. Error correcting
 - e. Data model building
3. Integrate Python and Hadoop and perform the following operations on forest fire dataset
 - a. Data analysis using the Map Reduce in PyHadoop
 - b. Data mining in Hive
4. Visualize the data using Python libraries matplotlib, seaborn by plotting the graphs for assignment no. 2 and 3 (Group B)
5. Perform the following data visualization operations using Tableau on Adult and Iris datasets.
 - a. 1D (Linear) Data visualization
 - b. 2D (Planar) Data Visualization

- c. 3D (Volumetric) Data Visualization
- d. Temporal Data Visualization
- e. Multidimensional Data Visualization
- f. Tree/ Hierarchical Data visualization
- g. Network Data visualization

Group C: Model Implementation

1. Create a review scrapper for any ecommerce website to fetch real time comments, reviews, ratings, comment tags, customer name using Python.
2. Develop a mini project in a group using different predictive models techniques to solve any real life problem. (Refer link dataset- <https://www.kaggle.com/tanmoyie/us-graduate-schools-admission-parameters>)

Reference Books:

1. Big Data, Black Book, DT Editorial services, 2015 edition.
2. Data Analytics with Hadoop, Jenny Kim, Benjamin Bengfort, OReilly Media, Inc.
3. Python for Data Analysis by Wes McKinney published by O' Reilly media, ISBN: 978-1-449-31979-3.
4. Python Data Science Handbook by Jake Vander Plas
<https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf>
5. Alex Holmes, Hadoop in practice, Dream tech press.
6. Online References for data set
 - a. <http://archive.ics.uci.edu/ml/>
 - b. <https://www.kaggle.com/tanmoyie/us-graduate-schools-admission-parameters>
 - c. <https://www.kaggle.com>



Sinhgad Technical Education Society's SINHGAD COLLEGE OF ENGINEERING, PUNE

S. No. 44/1, Off Sinhgad Road, Vadgaon(BK), Pune- 411041

Accredited by NAAC with Grade 'A+'

Sinhgad Institutes ACADEMIC YEAR 2023-24, SEMESTER-II

DEPARTMENT OF INFORMATION TECHNOLOGY

Name of Student :		PRN No. :	
Student Roll No. :		Class:	Third Year
Subject :	Data Science & Big Data Analytics Laboratory	Batch:	

INDEX

Group	Title of Assignment	Pg. No.	Given Date	Submission Date	Re-Mark	Sign
A	1. Single node/Multiple node Hadoop Installation.					
	2. Design a distributed application using MapReduce (Using Java) which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet and process it using a pseudo distribution mode on Hadoop platform.					
	3. Write an application using HiveQL for flight information system which will include a. Creating, Dropping, and altering Database tables. b. Creating an external Hive table. c. Load table with data, insert new values and field in the table, and Join tables with Hive. d. Create index on Flight					

C	1. Create a review scrapper for any ecommerce website to fetch real time comments, reviews, Ratings, comment tags, customer name using Python.								
	2. Develop a mini project in a group using different predictive models techniques to solve any real life problem. (Refer link dataset- https://www.kaggle.com/tanmoyie/us-graduate-schools-admission-parameters)								



Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE
S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041
Accredited by NAAC with Grade 'A+'

GROUP A ASSIGNMENT NO. 01	Single node/Multiple node Hadoop Installation.
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 01(A)

AIM:

To Perform Hadoop Installation (Configuration) on

- Single Node
- Multiple Node

OBJECTIVES:

- To understand Big Data Fundamentals.
- To understand Different Big Data Processing Techniques.

OUTCOMES:

- To apply Big Data Primitives & Fundamentals for application development.

THEORY:

Cluster Computing

A computer cluster is a set of connected computers (nodes) that work together as if they are a single (much more powerful) machine. Unlike grid computers, where each node performs a different task, computer clusters assign the same task to each node.

Homogeneous Cluster

In homogeneous clusters, all machines are assumed to be the same; however, in the heterogeneous type, machines have different computing and consumption power. All-in strategy (AIS) [70] is a framework for energy management in MapReduce clusters by powering down all nodes in the cluster during a low utilization period.

Heterogeneous Cluster

A heterogeneous cluster environment can contain processors and devices with different bandwidth and computational capabilities. Symmetric MPI applications will assign identical workloads to all participants in the application, which can cause load imbalance, as the execution time might be shorter on some devices due to their higher computational performance.

Hadoop

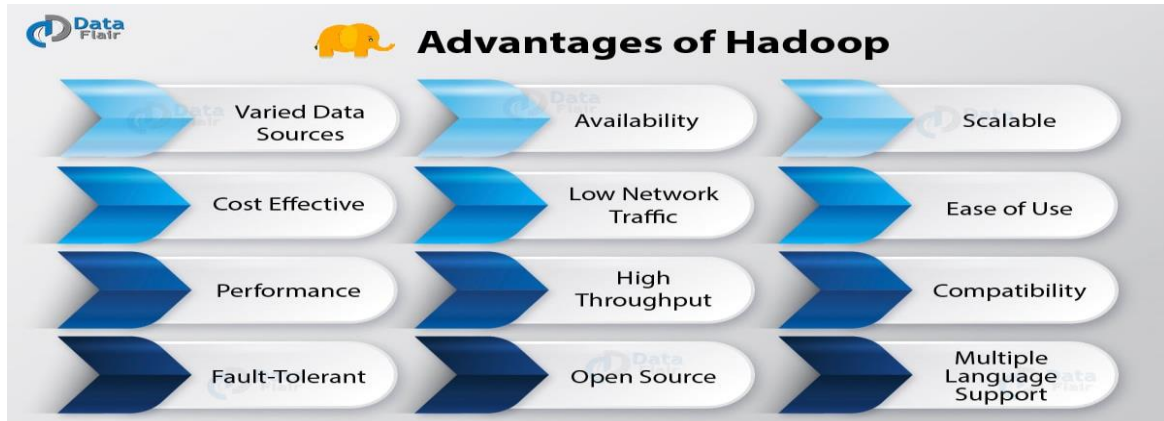
Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. This approach takes advantage of data locality— nodes manipulating the data they have access to— to allow the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

The base Apache Hadoop framework is composed of the following modules:

- **Hadoop Common** – contains libraries and utilities needed by other Hadoop modules;
- **Hadoop Distributed File System (HDFS)** – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;
- **Hadoop YARN** – a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications; and

- **Hadoop MapReduce** – an implementation of the MapReduce programming model for large scale data processing.

Benefits of using Hadoop



Installing Hadoop

Prerequisites

- **VIRTUAL BOX:** it is used for installing the operating system on it.
- **OPERATING SYSTEM:** You can install Hadoop on Linux-based operating systems. Ubuntu and Cent OS are very commonly used. In this tutorial, we are using Cent OS.
- **JAVA:** You need to install the Java 8 package on your system.
- **HADOOP:** You require Hadoop 2.7.3 package.

Step 1: Download the Java 8 Package & Save the file in your home directory.

Step 2: Extract the Java Tar File.

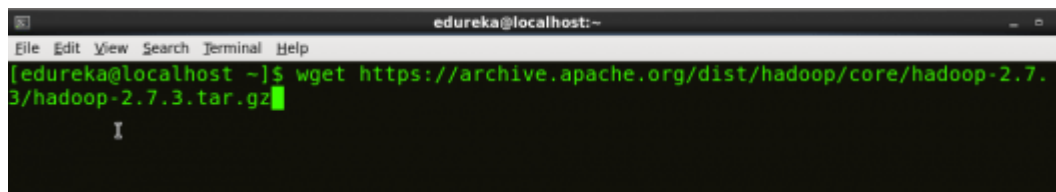
Command: `tar -xvf jdk-8u101-linux-i586.tar.gz`

```
edureka@localhost:~$ tar -xvf jdk-8u101-linux-i586.tar.gz
```

Fig: Hadoop Installation – Downloading Hadoop

Step 3: Download the Hadoop 2.7.3 Package.

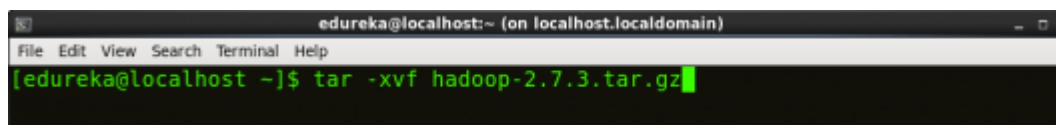
Command: `wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz`



```
edureka@localhost:~  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz
```

Step 4: Extract the Hadoop tar File.

Command: `tar -xvf hadoop-2.7.3.tar.gz`



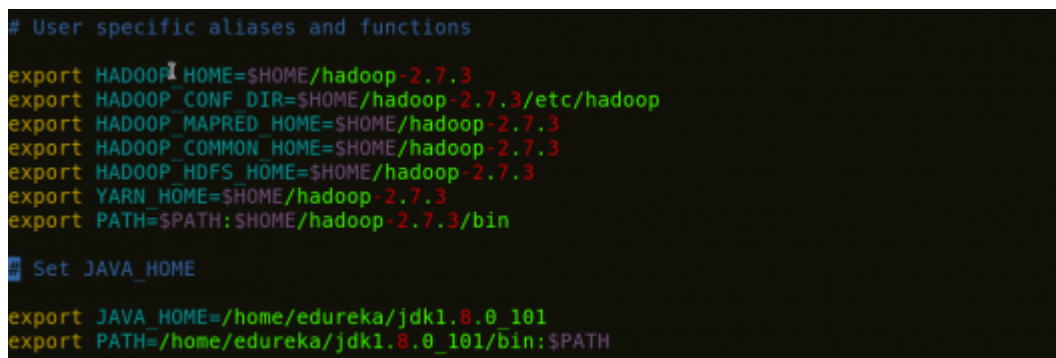
```
edureka@localhost:~ (on localhost.localdomain)  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ tar -xvf hadoop-2.7.3.tar.gz
```

Fig: Hadoop Installation – Extracting Hadoop Files

Step 5: Add the Hadoop and Java paths in the bash file (.bashrc).

Open. bashrc file. Now, add Hadoop and Java Path as shown below.

Command: `vi .bashrc`



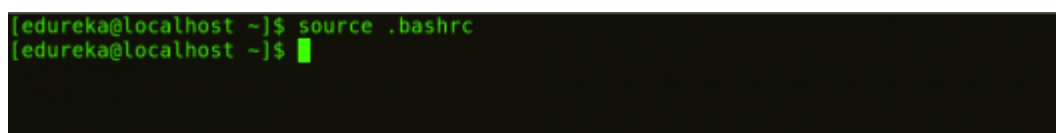
```
# User specific aliases and functions  
  
export HADOOP_HOME=$HOME/hadoop-2.7.3  
export HADOOP_CONF_DIR=$HOME/hadoop-2.7.3/etc/hadoop  
export HADOOP_MAPRED_HOME=$HOME/hadoop-2.7.3  
export HADOOP_COMMON_HOME=$HOME/hadoop-2.7.3  
export HADOOP_HDFS_HOME=$HOME/hadoop-2.7.3  
export YARN_HOME=$HOME/hadoop-2.7.3  
export PATH=$PATH:$HOME/hadoop-2.7.3/bin  
  
# Set JAVA_HOME  
export JAVA_HOME=/home/edureka/jdk1.8.0_101  
export PATH=/home/edureka/jdk1.8.0_101/bin:$PATH
```

Fig: Hadoop Installation – Setting Environment Variable

Then, save the bash file and close it.

For applying all these changes to the current Terminal, execute the source command.

Command: `source .bashrc`

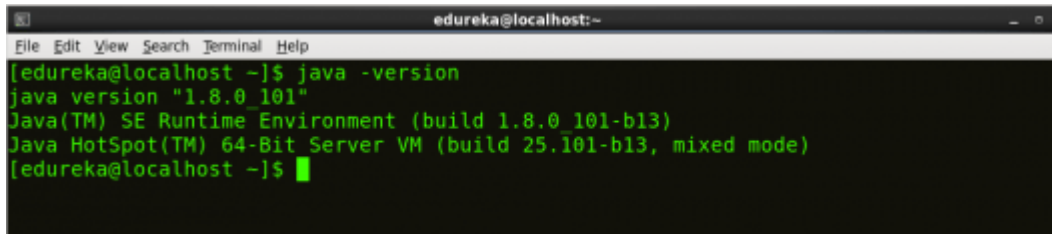


```
[edureka@localhost ~]$ source .bashrc  
[edureka@localhost ~]$
```

Fig: Hadoop Installation – Refreshing environment variables

To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and Hadoop version commands.

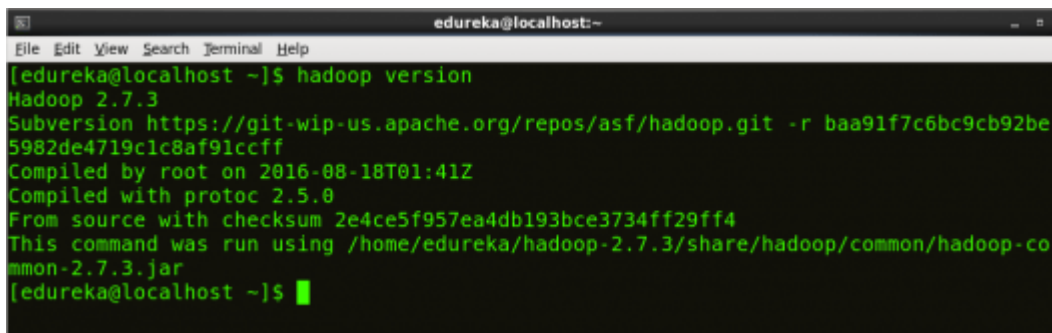
Command: java -version



```
edureka@localhost:~  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ java -version  
java version "1.8.0_101"  
Java(TM) SE Runtime Environment (build 1.8.0_101-b13)  
Java HotSpot(TM) 64-Bit Server VM (build 25.101-b13, mixed mode)  
[edureka@localhost ~]$
```

Fig: Hadoop Installation – Checking Java Version

Command: Hadoop version



```
edureka@localhost:~  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ hadoop version  
Hadoop 2.7.3  
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r baa91f7c6bc9cb92be5982de4719c1c8af91ccff  
Compiled by root on 2016-08-18T01:41Z  
Compiled with protoc 2.5.0  
From source with checksum 2e4ce5f957ea4db193bce3734ff29ff4  
This command was run using /home/edureka/hadoop-2.7.3/share/hadoop/common/hadoop-common-2.7.3.jar  
[edureka@localhost ~]$
```

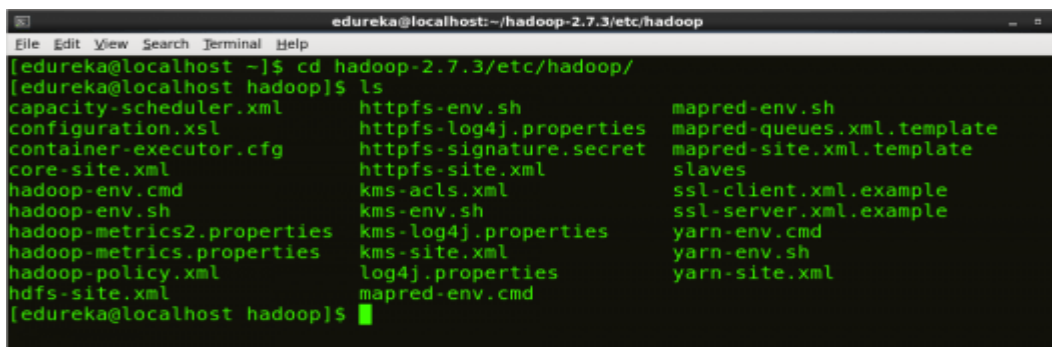
Fig: Hadoop Installation – Checking Hadoop Version

Step 6: Edit the Hadoop Configuration files.

Command: cd hadoop-2.7.3/etc/hadoop/

Command: ls

All the Hadoop configuration files are located in hadoop-2.7.3/etc/hadoop directory as you can see in the snapshot below:



```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop  
File Edit View Search Terminal Help  
[edureka@localhost ~]$ cd hadoop-2.7.3/etc/hadoop/  
[edureka@localhost hadoop]$ ls  
capacity-scheduler.xml      httpfs-env.sh               mapred-env.sh  
configuration.xsl           httpfs-log4j.properties    mapred-queues.xml.template  
container-executor.cfg      httpfs-signature.secret    mapred-site.xml.template  
core-site.xml               httpfs-site.xml            slaves  
hadoop-env.cmd              kms-acls.xml               ssl-client.xml.example  
hadoop-env.sh               kms-log4j.properties       ssl-server.xml.example  
hadoop-metrics2.properties kms-site.xml               yarn-env.cmd  
hadoop-metrics.properties kms-site.xml               yarn-env.sh  
hadoop-policy.xml           log4j.properties          yarn-site.xml  
hdfs-site.xml               mapred-env.cmd  
[edureka@localhost hadoop]$
```

Fig: Hadoop Installation – Hadoop Configuration Files

Step 7: Open core-site.xml and edit the property mentioned below inside configuration tag: Core-site.xml informs Hadoop daemon where Name Node runs in the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & MapReduce.

Command: vi core-site.xml

```
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

```
1      <?xml version="1.0" encoding="UTF-8"?>
2      <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3          <configuration>
4              <property>
5                  <name>fs.default.name</name>
6                  <value>hdfs://localhost:9000</value>
7              </property>
8          </configuration>
```

Step 8: Edit hdfs-site.xml and edit the property mentioned below inside configuration tag: Hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, and Secondary NameNode). It also includes the replication factor and block size of HDFS.

Command: vi hdfs-site.xml

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.permission</name>
<value>>false</value>
</property>
```

Fig: Hadoop Installation – Configuring hdfs-site.xml

```
1      <?xml version="1.0" encoding="UTF-8"?>
2      <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3          <configuration>
4              <property>
5                  <name>dfs.replication</name>
6                  <value>1</value>
7              </property>
8              <property>
9                  <name>dfs.permission</name>
10                 <value>>false</value>
11             </property>
12         </configuration>
```

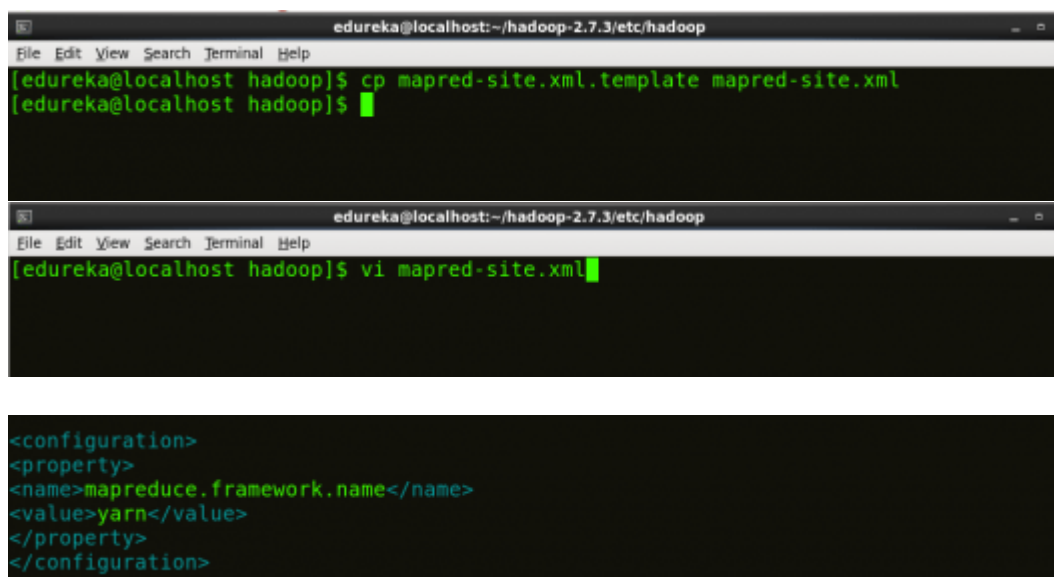
Step 9: Edit the mapred-site.xml file and edit the property mentioned below inside configuration tag:

Mapred-site.xml contains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.

In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template.

Command: cp mapred-site.xml.template mapred-site.xml

Command: vi mapred-site.xml.



```
edureka@localhost:~/hadoop-2.7.3/etc/hadoop
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ cp mapred-site.xml.template mapred-site.xml
[edureka@localhost hadoop]$

edureka@localhost:~/hadoop-2.7.3/etc/hadoop
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ vi mapred-site.xml

<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

Fig: Hadoop Installation – Configuring mapred-site.xml

```
1      <?xml version="1.0" encoding="UTF-8"?>
2      <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3          <configuration>
4              <property>
5                  <name>mapreduce.framework.name</name>
6                  <value>yarn</value>
7              </property>
8          </configuration>
```

Step 10: Edit yarn-site.xml and edit the property mentioned below inside configuration tag:

Yarn-site.xml contains configuration settings of Resource Manager and NodeManager like application memory management size, the operation needed on program & algorithm, etc.

You can even check out the details of Big Data with the Azure Data Engineering Certification in Hyderabad.

Command: vi yarn-site.xml

```

<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>

```

Fig: Hadoop Installation – Configuring yarn-site.xml

```

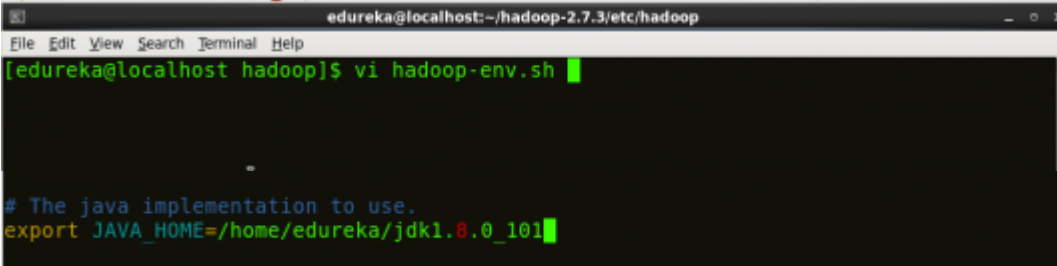
1      <?xml version="1.0">
2      <configuration>
3      <property>
4      <name>yarn.nodemanager.aux-services</name>
5      <value>mapreduce_shuffle</value>
6      </property>
7      <property>
8      <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
9      <value>org.apache.hadoop.mapred.ShuffleHandler</value>
10     </property>
11     </configuration>

```

Step 11: Edit `hadoop-env.sh` and add the Java Path as mentioned below:

`Hadoop-env.sh` contains the environment variables that are used in the script to run Hadoop like Java home path, etc.

Command: `vi hadoop-env.sh`



```

edureka@localhost:~/hadoop-2.7.3/etc/hadoop
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ vi hadoop-env.sh
# The java implementation to use.
export JAVA_HOME=/home/edureka/jdk1.8.0_101

```

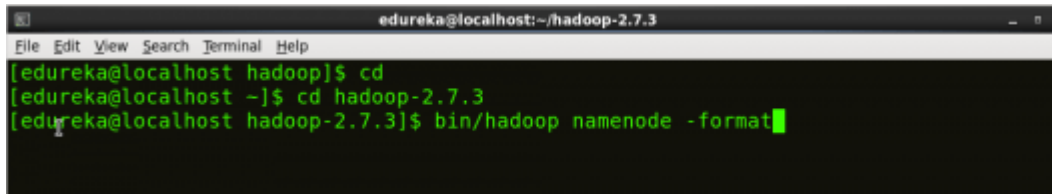
Fig: Hadoop Installation – Configuring hadoop-env.sh

Step 12: Go to Hadoop home directory and format the NameNode.

Command: `cd`

Command: `cd hadoop-2.7.3`

Command: `bin/hadoop namenode -format`

A terminal window titled 'edureka@localhost:~/hadoop-2.7.3' with a menu bar (File, Edit, View, Search, Terminal, Help). The command history shows: [edureka@localhost hadoop]\$ cd, [edureka@localhost ~]\$ cd hadoop-2.7.3, and [edureka@localhost hadoop-2.7.3]\$ bin/hadoop namenode -format. The cursor is at the end of the last command.

```
edureka@localhost:~/hadoop-2.7.3
File Edit View Search Terminal Help
[edureka@localhost hadoop]$ cd
[edureka@localhost ~]$ cd hadoop-2.7.3
[edureka@localhost hadoop-2.7.3]$ bin/hadoop namenode -format
```

Fig: Hadoop Installation – Formatting NameNode

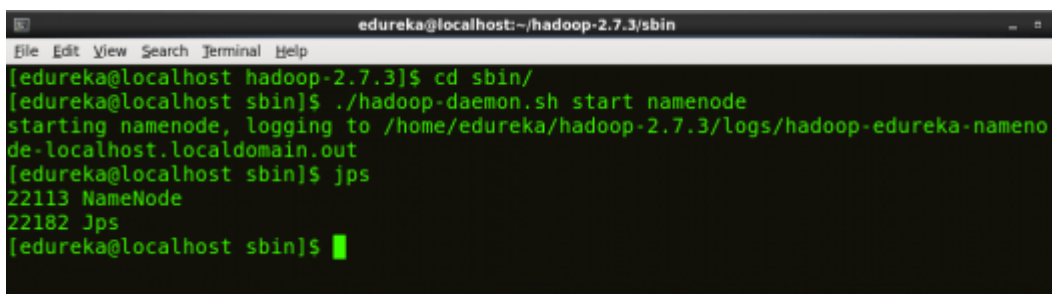
Step 13: Once the NameNode is formatted, go to hadoop-2.7.3/sbin directory and start all the daemons.

Command: cd hadoop-2.7.3/sbin

Either you can start all daemons with a single command or do it individually.

Command: ./start-all.sh

The above command is a combination of start-dfs.sh, start-yarn.sh & mr-jobhistory-daemon.sh Or you can run all the services individually as below:

A terminal window titled 'edureka@localhost:~/hadoop-2.7.3/sbin' with a menu bar (File, Edit, View, Search, Terminal, Help). The command history shows: [edureka@localhost hadoop-2.7.3]\$ cd sbin/, [edureka@localhost sbin]\$./hadoop-daemon.sh start namenode, starting namenode, logging to /home/edureka/hadoop-2.7.3/logs/hadoop-edureka-namenode-localhost.localdomain.out, [edureka@localhost sbin]\$ jps, 22113 NameNode, 22182 Jps, and [edureka@localhost sbin]\$ with a cursor.

```
edureka@localhost:~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
[edureka@localhost hadoop-2.7.3]$ cd sbin/
[edureka@localhost sbin]$ ./hadoop-daemon.sh start namenode
starting namenode, logging to /home/edureka/hadoop-2.7.3/logs/hadoop-edureka-namenode-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22182 Jps
[edureka@localhost sbin]$
```

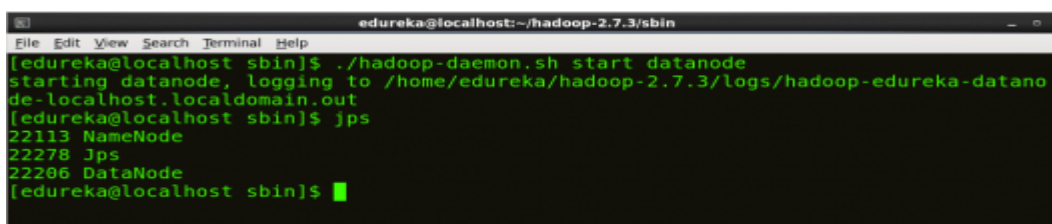
Start NameNode:

The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and tracks all the file stored across the cluster.

Command: ./hadoop-daemon.sh start namenode

Start DataNode: On startup, a DataNode connects to the Namenode and it responds to the requests from the Namenode for different operations.

Command: ./hadoop-daemon.sh start datanode.

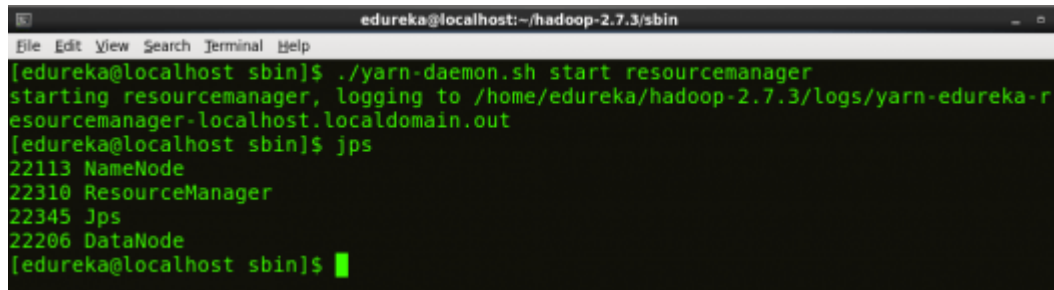
A terminal window titled 'edureka@localhost:~/hadoop-2.7.3/sbin' with a menu bar (File, Edit, View, Search, Terminal, Help). The command history shows: [edureka@localhost sbin]\$./hadoop-daemon.sh start datanode, starting datanode, logging to /home/edureka/hadoop-2.7.3/logs/hadoop-edureka-datanode-localhost.localdomain.out, [edureka@localhost sbin]\$ jps, 22113 NameNode, 22278 Jps, 22206 DataNode, and [edureka@localhost sbin]\$ with a cursor.

```
edureka@localhost:~/hadoop-2.7.3/sbin
File Edit View Search Terminal Help
[edureka@localhost sbin]$ ./hadoop-daemon.sh start datanode
starting datanode, logging to /home/edureka/hadoop-2.7.3/logs/hadoop-edureka-datanode-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22278 Jps
22206 DataNode
[edureka@localhost sbin]$
```

Start Resource Manager:

Resource Manager is the master that arbitrates all the available cluster resources and thus helps in managing the distributed applications running on the YARN system. Its work is to manage each Node Managers and the each application's Application Master.

Command: `./yarn-daemon.sh start resource manager`



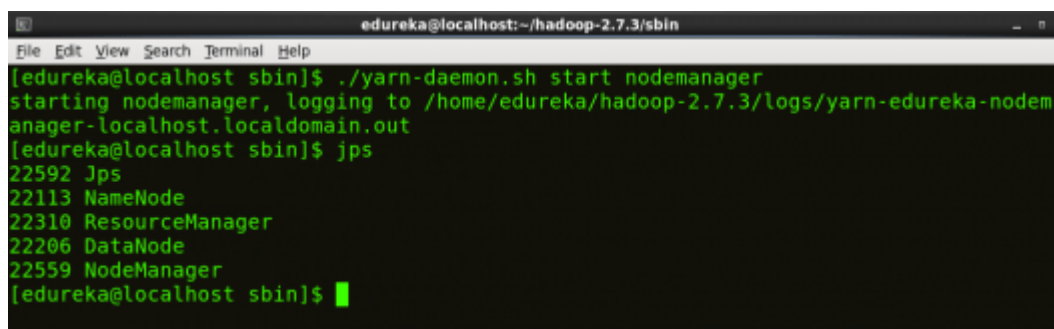
```
edureka@localhost:~/hadoop-2.7.3/sbin
[edureka@localhost sbin]$ ./yarn-daemon.sh start resourcemanager
starting resourcemanager, logging to /home/edureka/hadoop-2.7.3/logs/yarn-edureka-r
esourcemanager-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22310 ResourceManager
22345 Jps
22206 DataNode
[edureka@localhost sbin]$
```

Fig: Hadoop Installation – Starting Resource Manager

Start Node Manager:

The Node Manager in each machine framework is the agent which is responsible for managing containers, monitoring their resource usage and reporting the same to the Resource Manager.

Command: `./yarn-daemon.sh start node manager.`



```
edureka@localhost:~/hadoop-2.7.3/sbin
[edureka@localhost sbin]$ ./yarn-daemon.sh start nodemanager
starting nodemanager, logging to /home/edureka/hadoop-2.7.3/logs/yarn-edureka-nodem
anager-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22592 Jps
22113 NameNode
22310 ResourceManager
22206 DataNode
22559 NodeManager
[edureka@localhost sbin]$
```

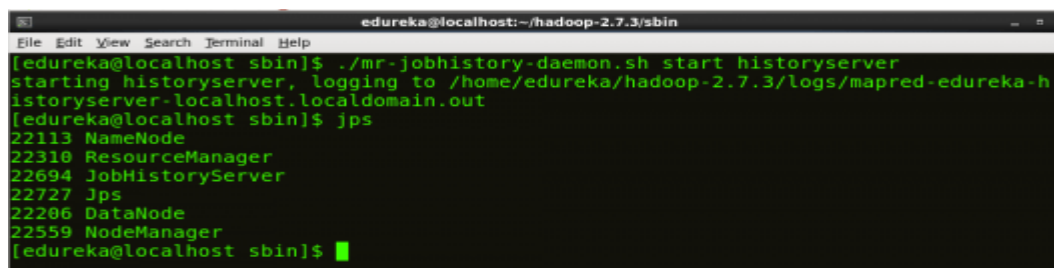
Start JobHistoryServer:

JobHistoryServer is responsible for servicing all job history related requests from client.

Command: `./mr-jobhistory-daemon.sh start historyserver`

Step 14: To check that all the Hadoop services are up and running, run the below command.

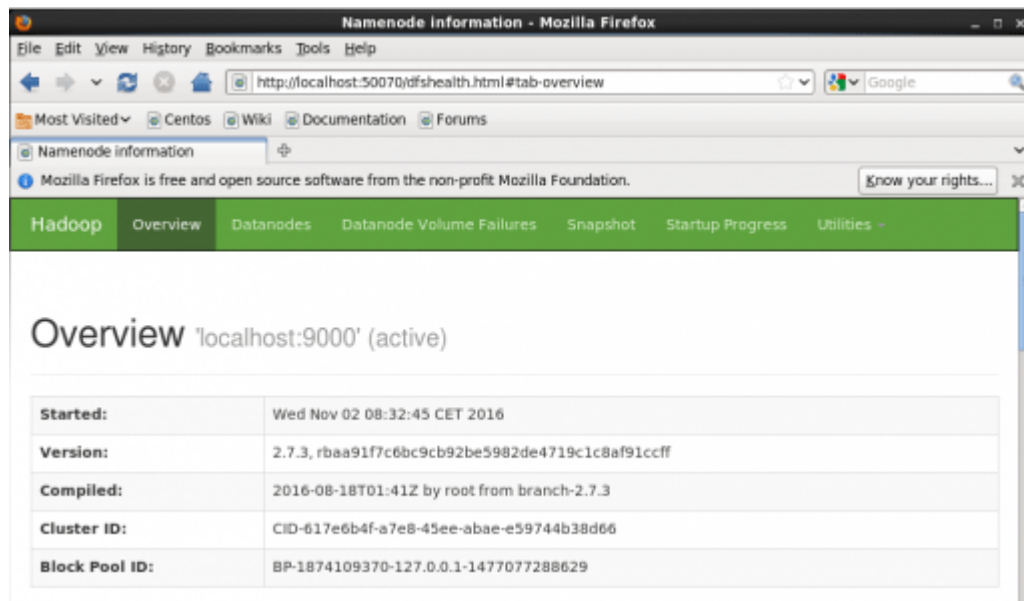
Command: `jps`



```
edureka@localhost:~/hadoop-2.7.3/sbin
[edureka@localhost sbin]$ ./mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /home/edureka/hadoop-2.7.3/logs/mapred-edureka-h
istoryserver-localhost.localdomain.out
[edureka@localhost sbin]$ jps
22113 NameNode
22310 ResourceManager
22694 JobHistoryServer
22727 Jps
22206 DataNode
22559 NodeManager
[edureka@localhost sbin]$
```

Fig: Hadoop Installation – Checking Daemons

Step 15: Now open the Mozilla browser and go to local host: 50070/dfshealth.html to check the NameNode interface.





Sinhgad Institutes

Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE

S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041

Accredited by NAAC with Grade 'A+'

GROUP A ASSIGNMENT NO. 02	Design a distributed application using MapReduce (Using Java) which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet and process it using a pseudo distribution mode on Hadoop Platform.
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 02(A)

AIM:

Design a distributed application using MapReduce (Using Java) which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet and process it using a pseudo distribution mode on Hadoop platform.

OBJECTIVES:

1. To understand Big Data Primitives & Fundamentals.
2. To understand different Big Data Processing Techniques.

OUTCOMES:

1. To apply Big Data primitives & fundamentals for application development.
2. To design algorithm & Techniques for Big Data Analysis.

THEORY:

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from

a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job. The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model. MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

Inserting Data into HDFS:

- The MapReduce framework operates on pairs, that is, the framework views the input to the job as a set of pairs and produces a set of pairs as the output of the job, conceivably of different types.
- The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface. Additionally, the key classes have to implement the Writable-Comparable interface to facilitate sorting by the framework.
- Input and Output types of a MapReduce job: (Input) -> map ->> reduce -> (Output).

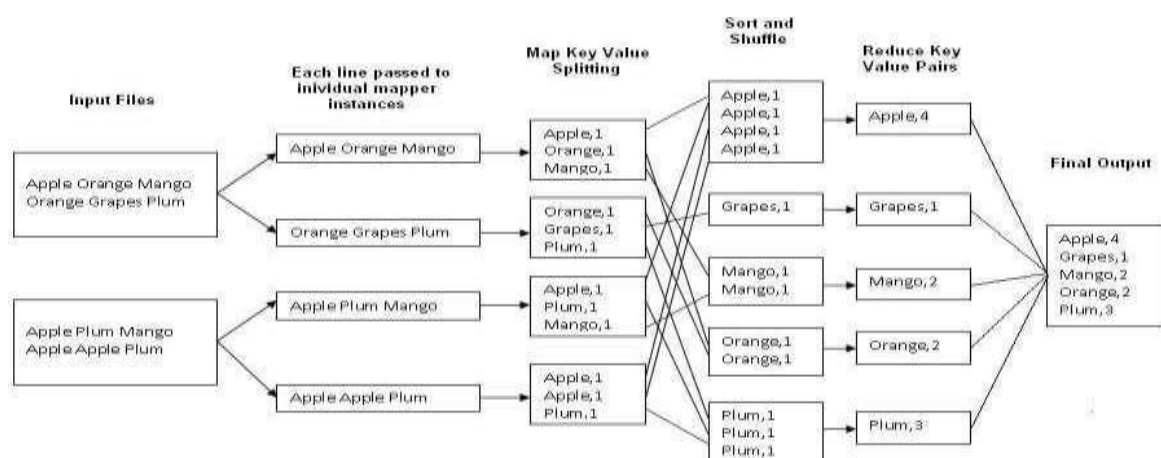


Fig.1: An Example Program to understand working of MapReduce Program.

Steps for Compilation & Execution of Program:

```
Su hadoopuser #sudomkdiranalyzelogs ls
#sudochmod -R 777 analyzelogs/ cd
ls cd .. (to move to home directory)
pwd ls cd pwd
#sudochown -R hadoop1 analyzelogs/ cd
ls
#cd analyzelogs/ ls cd ..
Copy the Files (Mapper.java,Reduce.java,Driver.java to Analyzelogs Folder)
```

```
#sudocp /home/priyanka/Desktop/assignment3/* -/analyzelogs/ (Convert
access_log_short.txt into access_log_short.csv)
```

```
Start HADOOP #start-dfs.sh #start-yarn.sh #jps cd
Cd
analyzelogs
ls pwd
ls
#ls -ltr
#ls -al
#sudochmod +r *.* pwd
#export
```

```
CLASSPATH="$HADOOP_HOME/share/hadoop/mapreduce/hadoo p-mapreduceclient-
core-
2.9.0.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop- mapreduce-clientcommon-
2.9.0.jar:$HADOOP_HOME/share/hadoop/common/hadoop-common-
2.9.0.jar:~/analyzelogs/SalesCountry/*:$HADOOP_HOME/lib/*"
```

(This should be PWD)

Compile Java Files

```
#javac-d. SalesMapper.java SalesCountryReducer.java
SalesCountryDriver.java ls #cd Sales Country/

ls cd ..
#sudo get it Manifest.txt
Main-class:
SalesCountry.SalesCountryDriver
(Press enter)
```

```
#jar -cfm analyzelogs.jar Manifest.txt SalesCountry/*.class
ls cd
#cdanalyzelogs/
```

Create Directory on Hadoop

```
#sudomkdir ~/input2000 ls pwd
#sudocp access_log_short.csv ~/input2000/
# $HADOOP_HOME/bin/hdfsdfs -put ~/input2000 /
#$HADOOP_HOME/bin/hadoop jar analyzelogs.jar/input2000/output2000
# $HADOOP_HOME/bin/hdfsdfs -cat /output2000/part- 00000
# stop-all.sh
# jps For GUI
Go to browser (local host: 50070) Go to utilities (browse directory)
```

Conclusion: Thus we have learnt how to design a distributed application using MapReduce and process a log file of a system

Output: Please attached output after conclusion page



Sinhgad Institutes

Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE

S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041

Accredited by NAAC with Grade 'A+'

GROUP A ASSIGNMENT NO. :03	Write an application using HiveQL for flight informationsystem which will include a. Creating, Dropping, and altering Database tables. b. Creating an external Hive table. c. Load table with data, insert new values and fieldin the table, and Join tables with Hive. d. Create index on Flight Information Table. e. Find the average departure delay per day in 2008.
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 03(A)

AIM:

Write an application using HiveQL for flight information system which will include

- Creating, Dropping, and altering Database tables.
- Creating an external Hive table.
- Load table with data, insert new values and field in the table, and Join tables with Hive.
- Create index on Flight Information Table.
- Find the average departure delay per day in 2008.

OBJECTIVES:

- To understand Big Data Primitives & Fundamentals.
- To understand different Big Data Processing Techniques.

OUTCOMES:

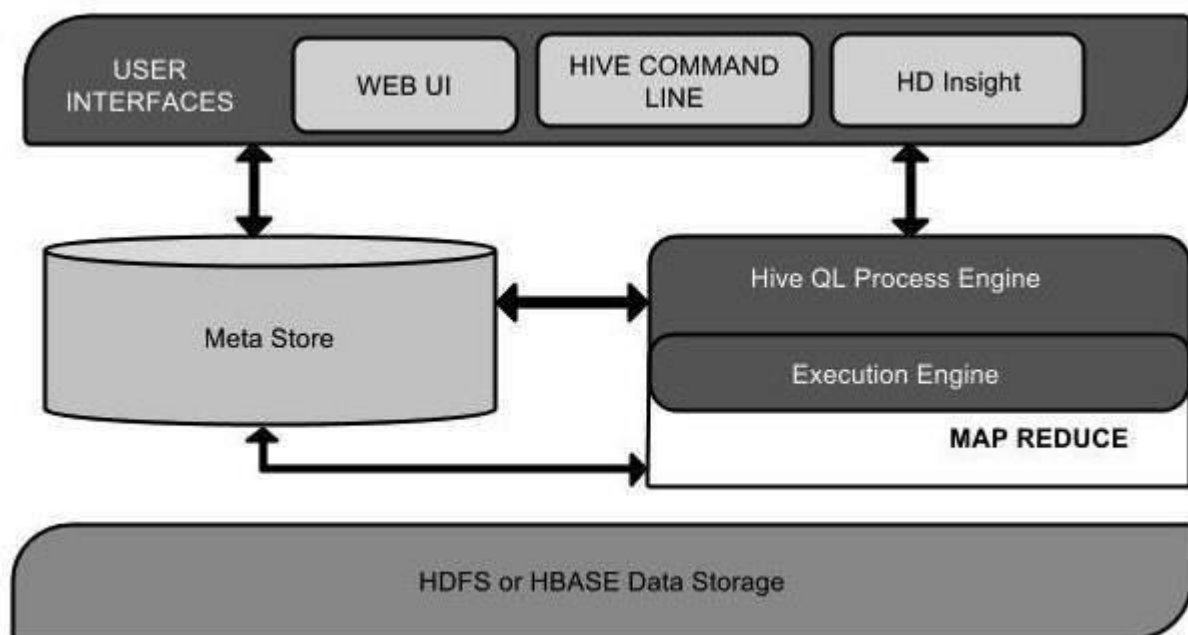
1. To apply Big Data primitives & fundamentals for application development.
2. To design algorithm & Techniques for Big Data Analysis.

THEORY:

Hive:

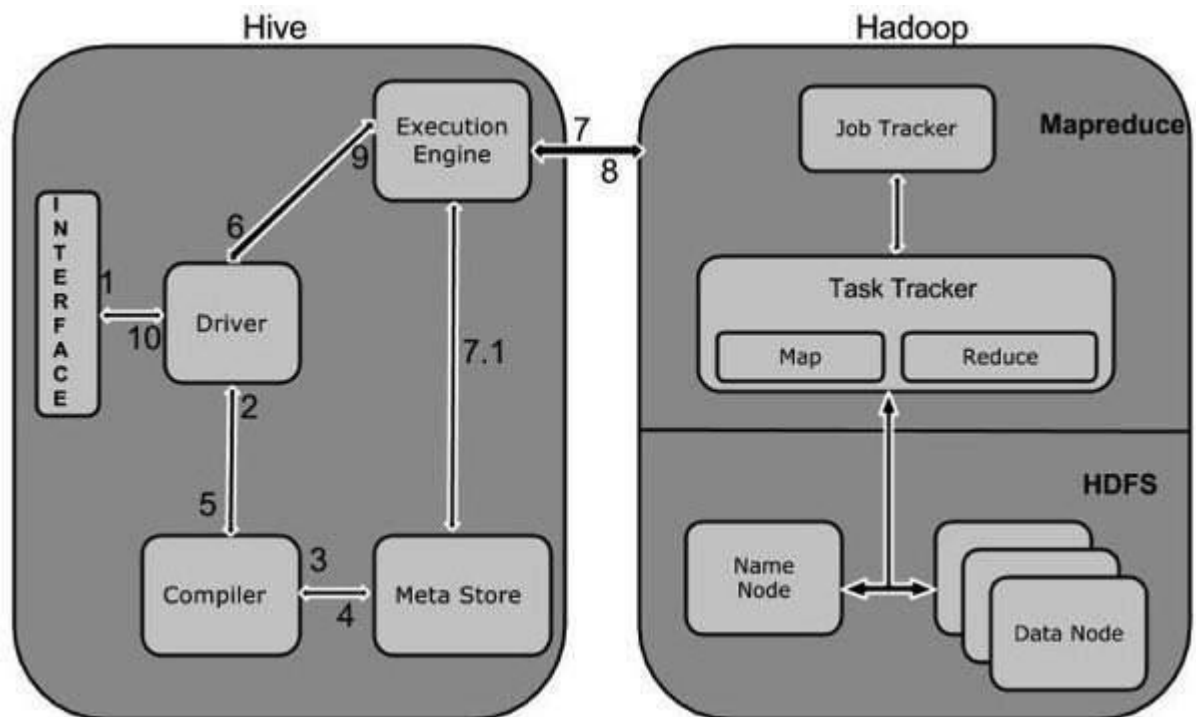
Apache Hive is a data warehouse system developed by Facebook to process a huge amount of structure data in Hadoop. We know that to process the data using Hadoop, we need to right complex map-reduce functions which is not an easy task for most of the developers. Hive makes this work very easy for us. It uses a scripting language called HiveQL which is almost similar to the SQL. So now, we just have to write SQL-like commands and at the backend of Hive will automatically convert them into the map-reduce jobs.

Hive architecture:



Hive is a data warehouse infrastructure software that can create interaction between user and HDFS. Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping. HiveQL is similar to SQL for querying on schema info on the Meta store. Execution engine processes the query and generates results as same as MapReduce results. It uses the flavour of MapReduce. Hadoop distributed file system or HBASE are the data storage techniques to store data into file system.

Working of Hive:



Data Types in Apache Hive

Hive data types are divided into the following 5 different categories:

1. Numeric Type: TINYINT, SMALLINT, INT, BIGINT
2. Date/Time Types: TIMESTAMP, DATE, INTERVAL
3. String Types: STRING, VARCHAR, CHAR
4. Complex Types: STRUCT, MAP, UNION, ARRAY
5. Misc Types: BOOLEAN, BINARY

Conclusion:

Hence we have created an application using HiveQL for flight information system which will include

1. Creating, Dropping, and altering Database tables.
2. Creating an external Hive table.

Output: Please attached output after conclusion page



Sinhgad Institutes

Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE
S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041
Accredited by NAAC with Grade 'A+'

GROUP B ASSIGNMENT NO. 01	Perform the following operations using Python on the Facebook metrics data sets a. Create data subsets b. Merge Data c. Sort Data d. Transposing Data e. Shape and reshape Data
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 01(B)

AIM:

Perform the following operations using Python on the Facebook metrics data sets

- Create data subsets
- Merge Data
- Sort Data
- Transposing Data
- Shape and reshape Data

OBJECTIVES:

- To understand & apply the analytical concept of Big Data using R/Python.

OUTCOMES:

- To apply the analytical concept of Big Data using R/Python.
- To design Big Data analytic application for Emerging Trends.

THEORY:

Python

It is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source code is also available under the GNU General Public License (GPL). This tutorial gives enough understanding on Python programming language.

Python's features

- Easy-to-learn – Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- Easy-to-read – Python code is more clearly defined and visible to the eyes.
- Easy-to-maintain – Python's source code is fairly easy-to-maintain.
- A broad standard library – Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- Interactive Mode – Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- Portable – Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- Extendable – you can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- Databases – Python provides interfaces to all major commercial databases.
- GUI Programming – Python supports GUI applications that can be.
- Created and ported to many system calls, libraries and windows systems.

Installation steps -

Python is available on a wide variety of platforms including Linux and Mac OS.

Python distribution is available for a wide variety of platforms. You need to download only the binary code applicable for your platform and install Python.

UNIX and LINUX Installation

Here are the simple steps to install Python on Unix/Linux machine.

- Open a Web browser and go to <https://www.python.org/downloads/>
- Follow the link to download zipped source code available for Unix/Linux.
- Download and extract files.
- Editing the Modules/Setup file if you want to customize some options.
- Run ./configure script
- Make
- Make install

This installs Python at standard location /usr/local/bin and its libraries at /usr/local/lib/pythonXX where XX is the version of Python.

Note- You can install Python on Windows or any other operating system.

Jupyter notebook:

With your virtual environment active, install Jupyter with the local instance of pip. Pip install jupyter

Run your notebook- jupyter notebook

Dataset: Download Facebook metrics data set.

Sample operation statements-

1. Create a data subset having 5 columns and 50 rows.(Type, Post Weekday, Post Hour, like, share)
2. Create a data subset having 4 columns and 25 rows. {Type, Total Interactions, like, share)
3. Create a subset having Post Hour>3hr.
4. Sort the subset 1 on like and share column in descending order.
5. Merge first two subset on Type and sort them on no of shares(share column)
6. Merge first two subsets on different left (like) and right columns (share).
7. Transpose first two subsets and sort them in descending order.
8. Show all data frames in wide and long formats. Convert wide to long and vice versa.

Conclusion:

Thus we have learnt different operations using Python on the Facebook metrics data sets.



Sinhgad Institutes

Sinhgad Technical Education Society's

SINHGAD COLLEGE OF ENGINEERING, PUNE

S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041

Accredited by NAAC with Grade 'A+'

GROUP B ASSIGNMENT NO.02	Perform the following operations using Python on the Air quality and Heart Diseases data sets a. Data cleaning b. Data integration c. Data transformation d. Error correcting e. Data model building
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 02(B)

AIM:

Perform the following operations using Python on the Air quality and Heart Diseases data sets

- Data cleaning
- Data integration
- Data transformation
- Error correcting
- Data model building

OBJECTIVES:

- To understand & apply the analytical concept of Big Data using R/Python.
- To understand different Data Visualization techniques for Big Data.

OUTCOMES:

- To apply the analytical concept of Big Data using R/Python.
- To design Big Data analytic application for Emerging Trends.

THEORY:

Download the datasets Air Quality and heart diseases available at [kaggle.com](https://www.kaggle.com).

Data cleaning

Data cleaning means fixing bad data in your data set. Bad data could be:

- Empty cells
- Data in wrong format
- Wrong data
- Duplicates

When working with multiple data sources, there are many chances for data to be incorrect, duplicated, or mislabeled. If data is wrong, outcomes and algorithms are unreliable, even though they may look correct. Data cleaning is the process of changing or eliminating garbage, incorrect, duplicate, corrupted, or incomplete data in a dataset. There's no such absolute way to describe the precise steps in the data cleaning process because the processes may vary from dataset to dataset. Data cleansing, data cleansing, or data scrub is the initiative among the general data preparation process. Data cleaning plays an important part in developing reliable answers and within the analytical process and is observed to be a basic feature of the info science basics. The motive of data cleaning services is to construct uniform and standardized data sets that enable data analytical tools and business intelligence easy access and perceive accurate data for each problem.

Data cleaning is the most important task that should be done as a data science professional. Having wrong or bad quality data can be detrimental to processes and analysis. Having clean data will ultimately increase overall productivity and permit the very best quality information in your decision-making.

- Error-Free Data
- Data Quality
- Accurate and Efficient
- Complete Data
- Maintains Data Consistency

Data Integration:

So far, we've made sure to remove the impurities in data and make it clean. Now, the next step is to combine data from different sources to get a unified structure with more meaningful and valuable information. This is mostly used if the data is segregated into different sources.

Data Transformation:

Now, we have a lot of columns that have different types of data. Our goal is to transform the data into a machine-learning-digestible format. All machine learning algorithms are based on mathematics. So, we need to convert all the columns into numerical format.

Handling Categorical Data:

There are some algorithms that can work well with categorical data, such as decision trees. But most machine learning algorithms cannot operate directly with categorical data. These algorithms require the input and output both to be in numerical form. If the output to be predicted is categorical, then after prediction we convert them back to categorical data from numerical data. Let's discuss some key challenges that we face while dealing with categorical data.

Encoding:

To address the problems associated with categorical data, we can use encoding. This is the process by which we convert a categorical variable into a numerical form. Here, we will look at three simple methods of encoding categorical data.

Replacing:

This is a technique in which we replace the categorical data with a number. This is a simple replacement and does not involve much logical processing. Let's look at an exercise to get a better idea of this

Error Correction:

There are many reasons such as noise, cross-talk etc., which may help data to get corrupted during transmission. Most of the applications would not function expectedly if they receive erroneous data. Thus error correction is important to do before any analysis.

- **Gauge min and max values:** For continuous variables, checking the minimum and maximum values for each column can give you a quick idea of whether your values are falling within the correct range.
- **Look for missing values:** The easiest way to find missing is to perform a count or sorting your columns. It helps in finding missing values which can be replaced/removed to get expected analysis.

Model Building:

In this phase, the data science team needs to develop data sets for training, testing, and production purposes. These data sets enable data scientists to develop analytical methods and train it, while holding aside some data for testing the model.

- Normalization
- Simple and Multiple Linear Regression
- Model Evaluation Using Visualization
- Polynomial Regression and Pipelines
- R-squared and MSE for In-Sample Evaluation
- Prediction and Decision Making

Conclusion:

Thus we have learnt different operations using Python on the Air quality data sets.



Sinhgad Institutes

Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE
S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041
Accredited by NAAC with Grade 'A+'

GROUP B ASSIGNMENT NO. 03	Integrate Python and Hadoop and perform the following operations on forest fire dataset a. Data analysis using the Map Reduce in PyHadoop. b. Data mining in Hive
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 03(B)

AIM:

Integrate Python and Hadoop and perform the following operations on forest fire dataset

- Data analysis using the Map Reduce in PyHadoop.
- Data mining in Hive

OBJECTIVES:

- To understand & apply the Analytical concept of Big Data using R/Python.
- To understand different Data Visualization techniques for Big Data.

OUTCOMES:

- To apply the analytical concept of Big Data using R/Python.
- To design Big Data analytic application for Emerging Trends.

THEORY:

Python:

Python is a popular high-level programming language known for its simplicity and readability. It has a large standard library and a vast ecosystem of third-party packages that make it suitable for a wide range of applications. Python supports multiple programming paradigms, including procedural, object-oriented, and functional programming. It is widely used in various domains, such as web development, data analysis, machine learning, artificial intelligence, scientific computing, and automation.

Hadoop:

Hadoop is an open-source framework designed for distributed storage and processing of large datasets across clusters of computers. It provides a scalable and fault-tolerant solution for processing and analyzing big data. The core components of Hadoop are Hadoop Distributed File System (HDFS) for distributed storage and MapReduce for distributed processing. Hadoop allows for parallel processing of data by breaking it into smaller chunks and distributing them across multiple nodes in a cluster. It is widely used in big data analytics and has become the de facto standard for processing large-scale datasets.

Integration of Python and Hadoop:

Python can be integrated with Hadoop through various libraries and frameworks, such as PySpark, Hadoop Streaming, and PyHive. PySpark provides a Python API for Apache Spark, a fast and distributed data processing engine that runs on top of Hadoop. Hadoop Streaming allows you to write MapReduce programs in any language, including Python, by reading input from standard input and writing output to standard output. PyHive provides a Python interface to interact with Apache Hive, a data warehouse infrastructure built on top of Hadoop, allowing you to perform data mining and analysis using SQL-like queries. By integrating Python with Hadoop, you can leverage the power of distributed computing to process and analyze large datasets efficiently. Python's simplicity and rich ecosystem make it a popular choice for data analysis and manipulation, while Hadoop provides the infrastructure for distributed processing and storage. To integrate Python and Hadoop, you can use the PySpark library, which provides a Python API for Apache Spark, a fast and distributed data processing engine that runs on Hadoop. PySpark allows you to write MapReduce programs in Python and execute them on a Hadoop cluster.

Here's a step-by-step guide on how to perform data analysis using MapReduce in PyHadoop:

1. **Install PySpark:** Install PySpark on your machine by following the instructions provided in the official PySpark documentation:
https://spark.apache.org/docs/latest/api/python/getting_started/installation.html
2. **Set up Hadoop cluster:** Set up a Hadoop cluster or use an existing cluster that you have access to.
3. **Import necessary modules:** In your Python script, import the necessary modules for PySpark: `from pyspark import SparkContext, SparkConf`

4. Create a SparkContext: Create a SparkContext object to connect to the Spark cluster:

```
conf = SparkConf().setAppName("ForestFireAnalysis") sc =  
SparkContext(conf=conf)
```

5. Load the forest fire dataset: Load the forest fire dataset into an RDD (Resilient Distributed Dataset) using the `textFile()` method. Assuming the dataset is stored in HDFS, you can specify the HDFS path to the dataset file:

```
dataset_path = "hdfs://<hdfs_path_to_dataset>" dataset_rdd =  
sc.textFile(dataset_path)
```

6. MapReduce operations: Perform your desired data analysis using MapReduce operations, such as `map()`, `reduceByKey()`, `filter()`, etc. Here's an example of counting the number of forest fires by year:

```
def extract_year(record): fields = record.split(",") return fields[2].split("-")[0]  
  
# Map operation: extract year from each record year_rdd =  
dataset_rdd.map(extract_year)  
  
# Reduce operation: count the number of forest fires by year fire_count_by_year =  
year_rdd.map(lambda year: (year, 1)).reduceByKey(lambda a, b: a + b)  
  
# Print the results  
  
for year, count in fire_count_by_year.collect(): print(f"Year: {year}, Count:  
{count}")
```

7. Stop the SparkContext: After performing the analysis, stop the SparkContext to release the cluster resources:

```
sc.stop()
```

For data mining in Hive, you can use the PyHive library, which provides a Python interface to interact with Hive. Here's an example of how to perform data mining operations in Hive using PyHive:

1. Install PyHive: Install PyHive on your machine by running the following command:

```
pip install pyhive
```

2. Import necessary modules: In your Python script, import the necessary modules for PyHive:

```
from pyhive import hive
```

3. Connect to Hive: Connect to the Hive server using the `connect` function:


```
connection = hive.connect(host='<hive_host>', port=<hive_port>,  
                           username='<hive_username>')
```

Replace <hive_host>, <hive_port>, and <hive_username> with the appropriate values for your Hive setup.

4. Create a cursor: Create a cursor object to execute Hive queries:
 `cursor = connection.cursor()`
5. Execute data mining queries: Use the cursor to execute data mining queries in Hive. For example, you can run a query to find the average forest fire area by month:

```
query = "SELECT MONTH(date), AVG(area) FROM forest_fire_data GROUP BY  
        MONTH(date)"  
cursor.execute(query)
```

```
# Fetch the results  
results = cursor.fetchall()
```

Conclusion:

Hence we have integrated Python and Hadoop and performed the following operations on forest fire dataset:

- a. Data analysis using the Map Reduce in PyHadoop.
- b. Data mining in Hive



Sinhgad Institutes

Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE
S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041
Accredited by NAAC with Grade 'A+'

GROUP B ASSIGNMENT NO. 04	Visualize the data using Python libraries matplotlib, seaborn by plotting the graphs for assignment no. 2 and 3 (Group B)
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 04(B)

AIM:

Visualize the data using Python libraries matplotlib, seaborn by plotting the graphs for assignment no. 2 and 3 (Group B)

OBJECTIVES:

1. To understand & apply the Analytical concept of Big Data using R/Python.
2. To understand different data visualization techniques for Big Data.

OUTCOMES:

1. To apply the analytical concept of Big Data using R/Python.
2. To design Big Data Analytic application for Emerging Trends.

THEORY:

It may sometimes seem easier to go through a set of data points and build insights from it but usually this process may not yield good results. There could be a lot of things left undiscovered as a result of this process. Additionally, most of the data sets used in real life are too big to do any analysis manually.

Data visualization is an easier way of presenting the data, however complex it is, to analyze trends and relationships amongst variables with the help of pictorial representation.

The following are the advantages of Data Visualization

1. Easier representation of compels data
2. Highlights good and bad performing areas
3. Explores relationship between data points

Identifies data patterns even for larger data points Visualization should have:

1. Appropriate usage of shapes, colors, and size while building visualization
2. Plots/graphs using a co-ordinate system are more pronounced
3. Knowledge of suitable plot with respect to the data types brings more clarity to the information
4. Usage of labels, titles, legends and pointers passes seamless information the wider audience.

Visualization libraries in python:

There are a lot of python libraries which could be used to build visualization like matplotlib, vispy, bokeh, seaborn, pygal, folium, plotly, cufflinks, and networkx. Of the many, matplotlib and seaborn seems to be very widely used for basic to intermediate level of visualizations.

1. Matplotlib

It is an library in Python for 2D plots of arrays, It is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

It is well maintained visualization output with high quality graphics draws a lot of users to it. Basic as well as advanced charts could be very easily built from the users/developers point of view, since it has a large community support, resolving issues and debugging becomes much easier.

2. Seaborn

This library sits on top of matplotlib. Means, it has some flavors of matplotlib while from the visualization point, it is much better than matplotlib and has added features as well.

Benefits:

- Built-in themes aid better visualization
- Statistical functions aiding better data insights
- Better aesthetics and built-in plots
- Helpful documentation with effective examples

Conclusion:

Hence we have successfully visualized the data using Python libraries matplotlib, seaborn by plotting the graphs for assignment no. 1 and 2 (Group B)



Sinhgad Institutes

Sinhgad Technical Education Society's
SINHGAD COLLEGE OF ENGINEERING, PUNE
S. No. 44/1, Off Sinhgad Road, Vadgaon (BK), Pune- 411041
Accredited by NAAC with Grade 'A+'

GROUP B ASSIGNMENT NO. 05	Perform the following data visualization operations using Tableau on Adult and Iris datasets. a. 1D (Linear) Data visualization b. 2D (Planar) Data Visualization c. 3D (Volumetric) Data Visualization d. Temporal Data Visualization e. Multidimensional Data Visualization f. Tree/ Hierarchical Data visualization g. Network Data visualization
GIVEN DATE:	
SUBMISSION DATE:	
SIGN. OF FACULTY:	

ASSIGNMENT NO. : 05(B)

AIM:

Perform the following data visualization operations using Tableau on Adult and Iris datasets.

- 1D (Linear) Data visualization
- 2D (Planar) Data Visualization
- 3D (Volumetric) Data Visualization
- Temporal Data Visualization
- Multidimensional Data Visualization
- Tree/ Hierarchical Data visualization
- Network Data visualization

OBJECTIVES:

1. To understand Application & Impact of Big Data.
2. To understand Emerging Trends in Big Data Analytics.
3. To understand different data visualization techniques for Emerging Trends.

OUTCOMES:

1. To design Big Data Analytic Application of Emerging Trends.
2. To visualize the Big Data using Tableau.
3. To design algorithms & techniques for Big Data Analytics.

THEORY:

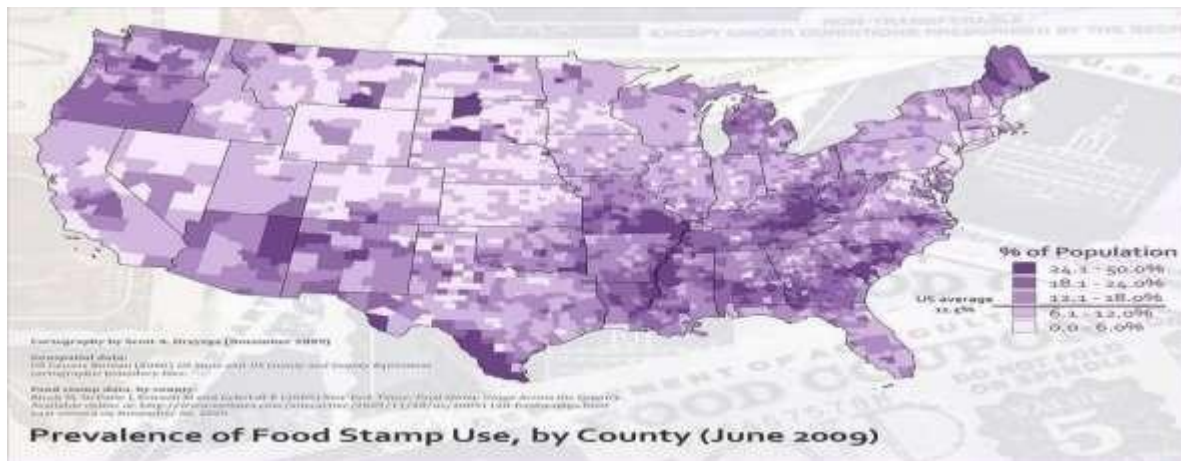
Data visualization or data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information".

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science.

Examples:

- Lists of data items, organized by a single feature (e.g., alphabetical order) (not commonly visualized)

Examples (geospatial):



- Choropleth:

Broadly, examples of scientific visualization:

- 3D computer models

In 3D computer graphics, 3D modeling (or three-dimensional modeling) is the process of developing a mathematical representation of any surface of an object (either inanimate or living) in three dimensions via specialized software. The product is called a 3D model. Someone who works with 3D models may be referred to as a 3D artist. It can be displayed as

a two- dimensional image through a process called 3D rendering or used in a computer simulation of physical phenomena. The model can also be physically created using 3D printing devices.

- Surface and volume rendering

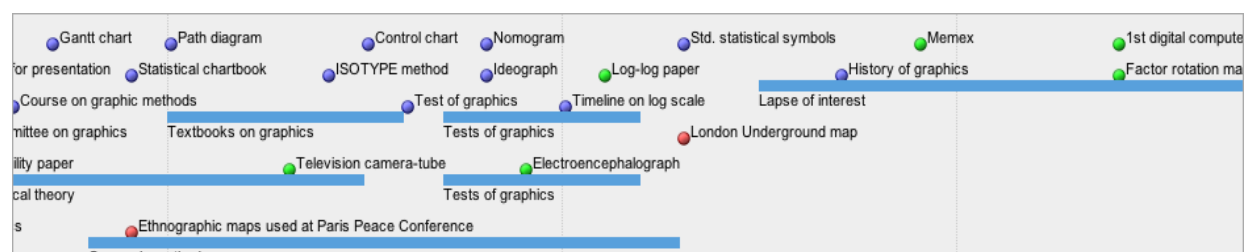
Rendering is the process of generating an image from a model, by means of computer programs. The model is a description of three-dimensional objects in a strictly defined language or data structure. It would contain geometry, viewpoint, texture, lighting, and shading information. The image is a digital image or raster graphics image. The term may be by analogy with an "artist's rendering" of a scene. 'Rendering' is also used to describe the process of calculating effects in a video editing file to produce final video output.

Volume rendering is a technique used to display a 2D projection of a 3D discretely sampled data set. A typical 3D data set is a group of 2D slice images acquired by a CT or MRI scanner. Usually these are acquired in a regular pattern (e.g., one slice every millimeter) and usually have a regular number of image pixels in a regular pattern. This is an example of a regular volumetric grid, with each volume element, or voxel represented by a single value that is obtained by sampling the immediate area surrounding the voxel.

- Computer simulations

Computer simulation is a computer program, or network of computers, that attempts to simulate an abstract model of a particular system. Computer simulations have become a useful part of mathematical modeling of many natural systems in physics, and computational physics, chemistry and biology; human systems in economics, psychology, and social science; and in the process of engineering and new technology, to gain insight into the operation of those systems, or to observe their behavior. [6] The simultaneous visualization and simulation of a system is called visualization.

Examples:

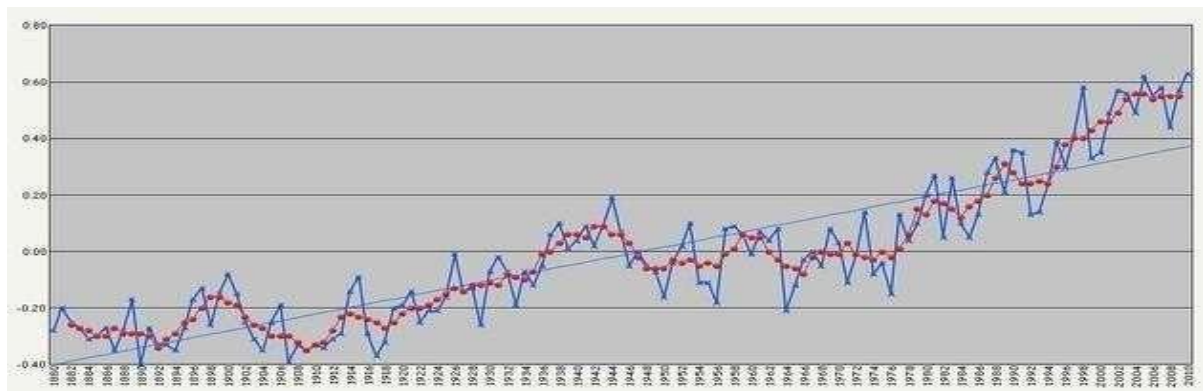


Tools: SIMILE Timeline, TimeFlow, Timeline JS, Excel Image:

Friendly, M. & Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. Web

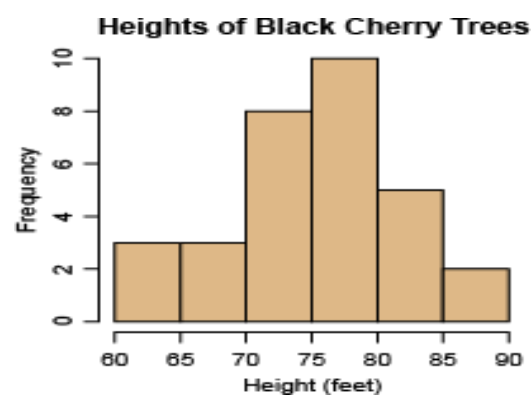
document, <http://www.datavis.ca/milestones/>. Accessed: August 30, 2012.

- Time series:

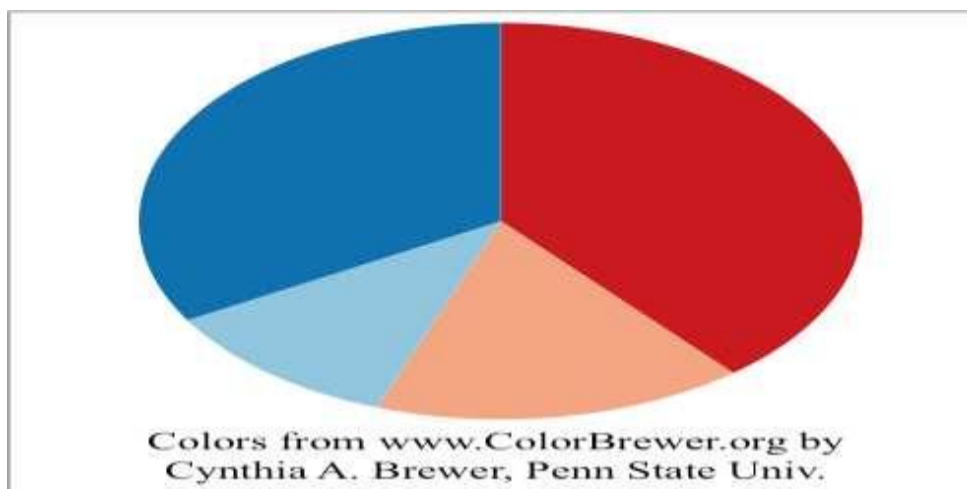


Examples (category proportions, counts)

- Histogram:

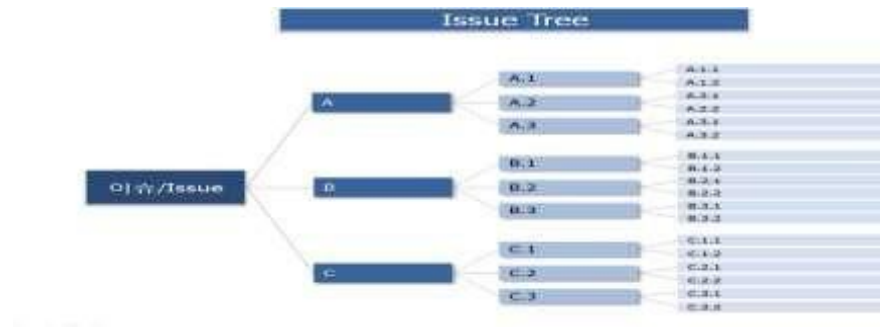


- Pie chart:

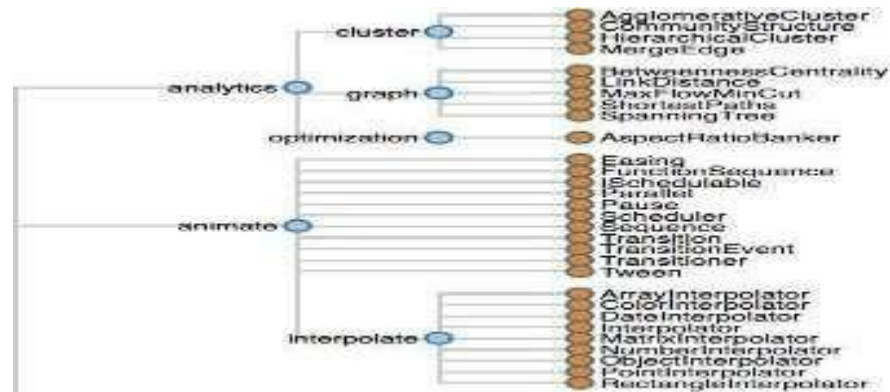


Examples:

- General tree visualization:



- Dendrogram



Examples:

- Matrix



- Node-link diagram (link-based layout algorithm)

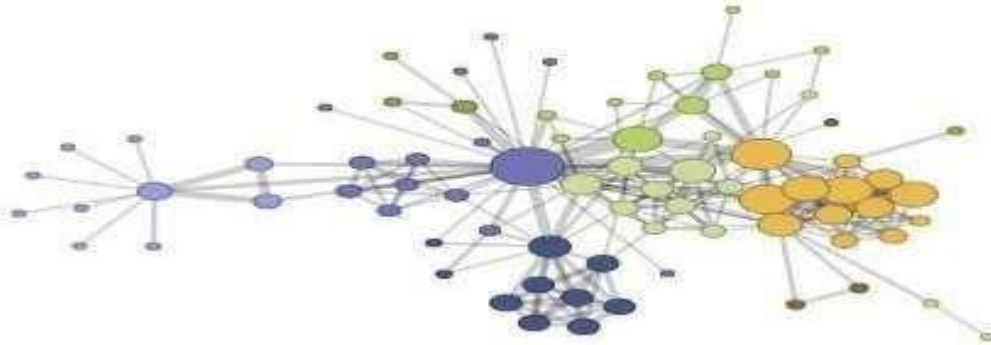


Tableau:

Tableau is a Business Intelligence tool for visually analyzing the data. Users can create and distribute an interactive and shareable dashboard, which depict the trends, variations, and density of the data in the form of graphs and charts. Tableau can connect to files, relational and Big Data sources to acquire and process data. The software allows data blending and real-time collaboration, which makes it very unique. It is used by businesses, academic researchers, and many government organizations for visual data analysis. It is also positioned as a leader Business Intelligence and Analytics Platform in Gartner Magic Quadrant.

Tableau Features:

Tableau provides solutions for all kinds of industries, departments, and data environments. Following are some unique features which enable Tableau to handle diverse scenarios.

- Speed of Analysis – as it does not require high level of programming expertise, any user with access to data can start using it to derive value from the data.
- Self-Reliant – Tableau does not need a complex software setup. The desktop version which is used by most users is easily installed and contains all the features needed to start and complete data analysis.
- Visual Discovery – The user explores and analyzes the data by using visual tools like colors, trend lines, charts, and graphs. There is very little script to be written as nearly everything is done by drag and drop.
- Blend Diverse Data Sets – Tableau allows you to blend different relational, semi structured and raw data sources in real time, without expensive up-front integration costs. The users don't need to know the details of how data is stored.
- Architecture Agnostic – Tableau works in all kinds of devices where data flows. Hence, the user need not worry about specific hardware or software requirements to use Tableau.
- Real-Time Collaboration – Tableau can filter, sort, and discuss data on the fly and embed a live dashboard in portals like SharePoint site or Salesforce. You can save your view of data and allow colleagues to subscribe to your interactive dashboards so they see the very latest data just by refreshing their web browser.
- Centralized Data – Tableau server provides a centralized location to manage all of the organization's published data sources. You can delete, change permissions, add tags, and manage schedules in one convenient location. It's easy to schedule extract refreshes

and manage them in the data server. Administrators can centrally define a schedule for extracts on the server for both incremental and full refreshes.

There are three basic steps involved in creating any Tableau data analysis report.

These three steps are –

- Connect to a data source – It involves locating the data and using an appropriate type of connection to read the data.
- Choose dimensions and measures – this involves selecting the required columns from the source data for analysis.
- Apply visualization technique – This involves applying required visualization methods, such as a specific chart or graph type to the data being analyzed.

For convenience, let's use the sample data set that comes with Tableau installation named sample – superstore.xls. Locate the installation folder of Tableau and go to My Tableau Repository. Under it, you will find the above file at Data sources\9.2\en_US-US.

Connect to a Data Source

On opening Tableau, you will get the start page showing various data sources. Under the header “Connect”, you have options to choose a file or server or saved data source. Under Files, choose excel. Then navigate to the file “Sample – Superstore.xls” as mentioned above. The excel file has three sheets named Orders, People and Returns. Choose Orders.

Choose the Dimensions and Measures

Next, choose the data to be analyzed by deciding on the dimensions and measures. Dimensions are the descriptive data while measures are numeric data. When put together, they help visualize the performance of the dimensional data with respect to the data which are measures. Choose Category and Region as the dimensions and Sales as the measure. Drag and drop them as shown in the following screenshot. The result shows the total sales in each category for each region.

Apply Visualization Technique

In the previous step, you can see that the data is available only as numbers. You have to read and calculate each of the values to judge the performance. However, you can see them as graphs or charts with different colors to make a quicker judgment.

We drag and drop the sum (sales) column from the Marks tab to the Columns shelf. The table showing the numeric values of sales now turns into a bar chart automatically.

Conclusion:

Thus we have learnt how to visualize the data in different types (1 1D (Linear) Data visualization, 2D (Planar) Data Visualization, 3D (Volumetric) Data Visualization, Temporal Data Visualization, Multidimensional Data Visualization, Tree/ Hierarchical Data visualization, Network Data visualization) by using Tableau Software.