AIM: Design a distributed application using MapReduce.

PROBLEM STATEMENT /DEFINITION: Design a distributed application using MapReduce 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.

OBJECTIVE: ● To understand the concept of Map Reduce. ● To understand the details of Hadoop File system ● To understand the technique for log file processing ● Analyze the performance of hadoop file system ● To understand use of distributed processing

THEORY: What is MapReduce?

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

The Algorithm

● Generally MapReduce paradigm is based on sending the computer to where the data resides!

● MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage. o 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.

o 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.

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.

Terminology

PayLoad - Applications implement the Map and the Reduce functions, and form the core of the job.

Mapper - Mapper maps the input key/value pairs to a set of intermediate key/value pair.

NamedNode - Node that manages the Hadoop Distributed File System (HDFS).

DataNode - Node where data is presented in advance before any processing takes place.

MasterNode - Node where JobTracker runs and which accepts job requests from clients.

SlaveNode - Node where Map and Reduce program runs.

To see the status of job

```
$ $HADOOP_HOME/bin/hadoop job -status e.g. $
$HADOOP_HOME/bin/hadoop job -status job_201310191043_0004
```

To kill the job

```
$ $HADOOP_HOME/bin/hadoop job -kill e.g. $
$HADOOP_HOME/bin/hadoop job -kill job_201310191043_0004
```

CONCLUSION:

Understand the uses of distributed data processing using Map reduce.

AIM: Write an application using HiveQL for flight information system

PROBLEM STATEMENT /DEFINITION 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.

OBJECTIVE ● To understand various NOSQL database ● To understand the integration of NOSQL database with Hadoop. ● To analyze the performance of distributed processing with NOSQL.

THEORY: · Hive and HBase Architecture · Explanation of Hive Architecture · HBase Architecture · Explanation of HBase Architecture · List and details of DDL and DML Commands in HBase and Hive

SQL queries are submitted to Hive and they are executed as follows: 1. Hive compiles the query.

2. An execution engine, such as Tez or MapReduce, executes the compiled query.

3. The resource manager, YARN, allocates resources for applications across the cluster.

4. The data that the query acts upon resides in HDFS (Hadoop Distributed File System). Supported data formats are ORC, AVRO, Parquet, and text.

5. Query results are then returned over a JDBC/ODBC connection.

Hive Clients

You can connect to Hive using a JDBC/ODBC driver with a BI tool, such as Microstrategy, Tableau, BusinessObjects, and others, or from another type of application that can access Hive over a JDBC/ODBC connection. In addition, you can also use a command-line tool, such as Beeline, that uses JDBC to connect to Hive.

SQL in Hive

Hive supports a large number of standard SQL dialects. In a future release, when SQL:2011 is adopted, Hive will support ANSI-standard SQL.

HiveServer2

Clients communicate with HiveServer2 over a JDBC/ODBC connection, which can handle multiple user sessions, each with a different thread. HiveServer2 can also handle long-running sessions with asynchronous threads

Security

HiveServer2 performs standard SQL security checks when a query is submitted, including connection authentication. After the connection authentication check, the server runs authorization checks to make sure that the user who submits the query has permission to access the databases, tables, columns, views, and other resources required by the query.

File Formats

Hive supports many file formats. You can write your own SerDes (Serializers, Deserializers) interface to support new file formats.

HBase architecture has 3 important components- HMaster, Region Server and ZooKeeper. i. HMaster HBase HMaster is a lightweight process that assigns regions to region servers in the Hadoop cluster for load balancing. Responsibilities of HMaster –

- Manages and Monitors the Hadoop Cluster
- Performs Administration (Interface for creating, updating and deleting tables.)
- Controlling the failover
- Block Cache – This is the read cache. Most frequently read data is stored in the read cache and whenever the block cache is full, recently used data is evicted.
- MemStore- This is the write cache and stores new data that is not yet written to the disk. Every column family in a region has a MemStore.

HBase uses ZooKeeper as a distributed coordination service for region assignments and to recover any region server crashes by loading them onto other region servers that are functioning. ZooKeeper is a centralized monitoring server that maintains configuration information and provides distributed synchronization

ASSIGNMENT NO. : 01(B)

AIM:

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

a. Create data subsets

b. Merge D

d. Transposing Data

e. Shape and reshape Data

OBJECTIVES:

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

OUTCOMES:

1. To apply the analytical concept of Big Data using R/Python.

2. 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.
- 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

Conclusion

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

ASSIGNMENT NO. : 02(B)

AIM:

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

a.Data cleaning

a. Data integration
b. Data transformation
c. Error correcting
d. 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:

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

THEORY:

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

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

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

Conclusion:

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

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

• Easier representation of compels data

• Highlights good and bad performing areas

• Explores relationship between data points

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

• Appropriate usage of shapes, colors, and size while building visualization

• Plots/graphs using a co-ordinate system are more pronounced

• Knowledge of suitable plot with respect to the data types brings more

clarity to the information

• 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, its 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 example

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)

Assignment B(4)

Aim:

Perform the following data visualization operations using Tableau onAdult 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

- OBJECTIVES:

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,

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.

• 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.

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.

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

ASSIGNMENT NO: 01 grp:C

TITLE: Create a review scrapper for any ecommerce website to fetch real time comments, reviews, ratings comment tags, customer name using Python.

AIM: To create a review scrapper for ecommerce website.

OBJECTIVE:

1. To scrap an ecommerce website.
2. https://www.thewhiskyexchange.com/c/35/japanese-whisky

We are going to scrape this website to retrieve the information from there.

Software used: Python Shell 3.7.3

THEORY:

Scraping product reviews from ecommerce websites has become one of the most vital competitive intelligence activities in and around the ecommerce space. Product reviews on ecommerce websites are a great source of unbiased reviews organically left by actual consumers. This means, if you are a manufacturer trying to gather deep insights about your product, look no further than ecommerce product review pages. While the availability of product reviews on various eCommerce sites is vast and deep, not every company has the expertise, infrastructure, and resources to crawl and extract reviews from eCommerce sites in an automated manner.

Luckily, PromptCloud specializes in large-scale web scraping solutions and can help you

with scraping product reviews from eCommerce portals of your choice. Our fully managed service comes with end-to-end crawler management which would insulate you from the nuances of web data extraction activity and help you focus on the application of the delivered data.

A very simple pie-chart is created using just the input vector and labels. The below script will create and save the pie chart in the current R working directory.

Applications of eCommerce product reviews scraping: eCommerce is taking over the world by storm and there’s no scarcity of product reviews on these eCommerce portals. The main advantage to manufacturing companies is the unbiased nature of these reviews which will help them understand their consumers so as to serve them better. Here are the most popular applications of product reviews scraping.

Understand your consumer preferences: Staying up-to-date with customer preferences is crucial to successful products. If your product isn't addressing what the consumer wants, you are leaving money on the table and contributing to customer dissatisfaction. To avoid this, understanding the demands and needs of your target customer is vital. This is where product reviews scraping can fill in by helping you listen to the customer's voice. By extracting product reviews and analyzing it with the right goals, your business can understand the crucial factors that are driving sales in your niche and tweak the products accordingly.

Brand monitoring: Reputation or brand image is a huge factor when it comes to customer loyalty and business growth for any organization. Maintaining a positive brand image is crucial and this means your business should be all ears to the customer grievances. Brand monitoring can help you detect unaddressed issues of your customers that can escalate to bigger issues if not detected early on. Scraping reviews from eCommerce sites can help you monitor this and preserve your positive brand image.

Competitor analysis: Listening to your customers alone is just not enough to wade through this ever so competitive business world of today. Sometimes, data associated to your competitors can help you detect low hanging fruits which you can leverage before they do. For example, if reviews on your competitors products indicate the demand for a particular feature, be the first one to incorporate that to your own product. This would help you stay ahead of the curve and garner more users.

Natural language processing: Natural language processing (NLP) is all about enabling machines to understand the context behind human languages. Systems developed using NLP helps run voice assistant platforms like Siri, Google Now and Cortana, translation services and artificial intelligence systems. Huge amounts of data is a must to cater to the requirements of a NLP system. Since product reviews scraped from eCommerce sites is user-generated content, it makes for the perfect data for training Natural language processing systems.

Fraud Detection: Counterfeit products have always been a threat to brands. Not only do they affect sales figures, but also leave a bad impression among customers who may never realize the faulty product they received was in fact a fake. Scraping product reviews helps you access this data which might have hints about some ongoing fraud. An alarming number of negative reviews are definitely worth investigating further to rule out counterfeit products being sold by the supplier. Near real-time crawls or live crawls can be performed to identify the ecommerce partners who don't stick to the agreement.

Setup the Scraping Project

Our setup is pretty simple. Just create a folder and install Beautiful Soup, pandas, and requests. To create a folder and install the libraries, enter the commands given below.

```
mkdir scraper

pip install beautifulsoup4

pip install requests

pip install pandas
```

Now, create a file inside that folder and name it anything you like. I am using the name scraper.py. We are going to import requests, pandas, and bs4.

```
import requests

from bs4 import BeautifulSoup

import pandas as pd
```

Now, we are going to set the base URL of the main page because we'll need that when we construct our URLs for each of the individual products.

Also, we will send a user-agent on every HTTP request, because if you make GET request using requests then by default the user-agent is Python which might get blocked.

So, to override that, we will declare a variable which will store our user-agent.

```
baseurl = "https://www.thewhiskyexchange.com"

headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.82 Safari/537.36'}
```

Now we need to investigate the page so that we can figure out where the links are and

how we're going to get them.

- **CONCLUSION:** After the study of this assignment we have learnt Scrape the website by

using python libraries.