# Part A Assignment 3

Title :-

write an appliction using HBase and Hive QL for flight information system which will include

a1> Creating, dropping and alterning database tables

2> Creating an external Hive table to connect to the HBase for customer information table

3> Load table with data insert new values and field in the table Join tables with hive

4> Create index on flight information table

5> Find the average departure delay per day in 2008.

objective :- to apply Hbase and HiveQL on various application

prequisite :- Java programming

theory :-

what is the defination of Hive? what is the present version of hive and explain about ACID transactions in Hive?

Hive : it is a datawarehouse & database where data is typically loaded from batch perfossing for analytical purposes

* the last version of Hive 3.1.2 released on August 26, 2019

* use cases that require transation with ACID properties in have

(i) Streaming data :- with ACID we can insert update our same have pattern without affecting performance of table

Date

(ii) data counceting :- ~~Found~~ Sometimes data Stored may be micouert and some changes are requred which ~~ar~~ can easely achered by insent /update delete operations

(iii) Bullc updates using SQL merge statement - with bluk marge, small rules can be marged into file without affecting read performance

2) what laind of data warehouse is suitable for hive what are types of tables in hive?

* hive is most suitable for data warehouse application because

- Analya relatively static data
- has less responsive table
- does not malce rapid changes in data

* there are two types of tables,

- managed tables
- external tables

3) is it possible to add 100 nodes when we have 100 nodes allready in hive?

yes we can add by following steps

* taice new System, erectte usernane and pass

* install SSH and makes node setup SSH connetions

* Add ssh public id key to authougre key files

* Add new datanode hostname, ip adres, and other log in to new node

* Start HDFS of newly added slave and output the sps comd

4> what is metastore in hive? what is difference between local and remote metastore?

metastore is central repository of hive meta data

local Node :- the hive metastore reservice are in the same process but make hive metastore database runs in seperate process as it can be on seprate host in local mode

Remote mode :-
hive metadatastore resources are in its own Jvm process

Conclusion :-
Studied and implemented Hbase and hiveQL on various applications.

## Part B : Assignment 1

Title :- Perform the following operations using python on Facebook metrics data set & a) Create data subset (b) merge data (c) sort data (d) transposing data (e) shape and reshape data

Objective :- To use python concept

Prequisite :- python programming.

theory :-

what are key features of python?

- python is integrated languge it does not need to compiled before run
- python is dynamically typed we don't need to state type of variable
- it object orieanted programming language
- writing python code is quick

2) what type of language is python?
python is capable of scripting but it is considered as general purpose programming language

3) what is ~~ma~~ namespace in python?
A namespace refers to name which is assigned to each object in python the namespaces are mantained like a dictionary where key is namespace and value is adress

4> What are local and global variable in python?

local variable :- any variable inside a function is called local variable

Global variable :- variable declared outside a function in global space this can be accessed by any function in programe

5> whate are benifits of using python

- easy to use
- Free and open source
- it can run on any platform
- It provides ~~no~~ .
- Python has large library support
- the data structure used in python are user fraindly

Conclusion :-

Studied and implemented concepts

Part B :- Assignment No: 2

Title :- perform the following operations using python on the Air quality and heart Dieases data sets

(a) data cleaning (b) data integration
(c) data transformation (d) Error correcting
(e) data model building.

Prerequisite :- python programming

theory :-

what type of language python ? programming or scripting.

python is general perpose programming language but scripting is also possible

2> python an interpreated language Explain

An interpreted language is any programming language which is not in machine level code before structure

python is intrepreted language

3> what are common build in data types in python

- Numbers :- they include int, flot and complex
- list : an ordered sequence of iteam
- tuple :- an ordered immutable sequence
- string : sequence of characters
- set :- collection of unique items
- dictornory :- stores value in key and value pair
- Boolean :- true / false

Conclusion :-

Studied and implemented python concepts

PART B :- Assignment No :- 3

Title :- 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

Objective :- To use python concept

Prerequisite :- Python programming

theory :-

① Explain what is Name Node in Hadoop?

NameNode works as master in Hadoop cluster the main function are

* stores metadata of actual data
* manages filesystem namespaces
* regulate client access request for actual file data

2> what is Jobtracker in hadoop & what are active followed by hadoop?

In hadoop for submitting and tracking map reduce poles, Jobtracker is used Jobtracker for on its own Jvm process

Following actions perfound by Hadoop:-

- Client application submit Job to Jobtracker
- Jobtracker communecate to NameNode to determine data location
- Near location Jobtracker locates track tracker nodes
- an chosen node it submit work

3) mention what are most Common input format defined by Hadoop ?

the most Common input formats is Hadoop are

- SequenceFile input format
- text input format
- * key value input format

conclusion :-

Studed and applied concepts of python

Part B : Assignment 4

Title :- visualize the data using python libraries matplotlib, seaborn by plotting the graphs for assignment no 2 and 3 (Group B)

Objective :- to use python concepts

theory :-

① How to create multiple subplots in maltplotlib in python ?

to create multiple plots use matplotlib pyplot subplot method which structures the figure along with axes object or axery of axes object n row ; in code attribute of subplot () method determine the m· of source end columns subplot grid

.by default it returns figure with single subplot

How it works :-

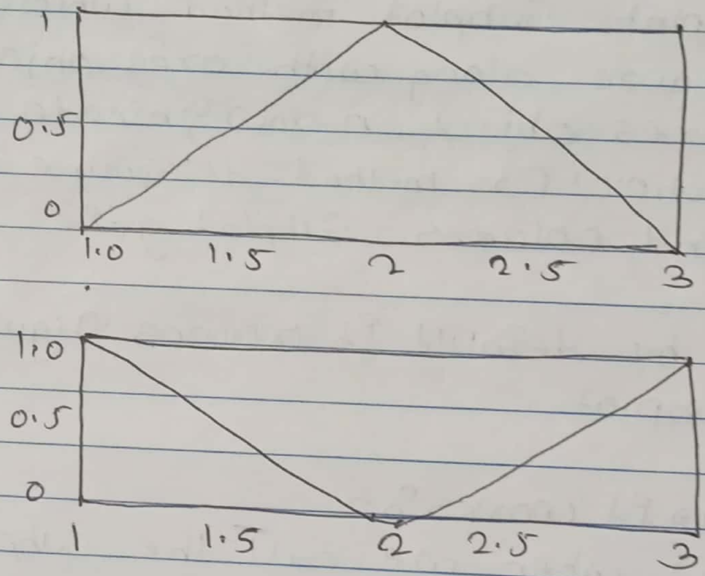
① when we call the subplots() method by stacking only in on chirection it returns axary of axes objects i.e subplots

② we can access these axes object using ndicess to create specific subplot call matplotlib-pyplot.plot() on corresponding index of axes example :-

1-D array of subplot

```
import matplotlib.pyplot as pt
x = [1,2,3]
y = [0,1,0]
z =[1,0,1]
Fig, ax = plt.subplot(2)
ax[0]. plot (x,y)
ax[1]. plol.(x,y]
```

output 5



Conclusion :-

Studied and implemented concepts of python

Date

Part B :- Assignment 5

Title :- perform the Following data visualisation operations using tableau on Adult and Iris datasets ⓐ 1D (linear) data visualization ⓑ 2D (planner) Data visualization ⓒ 3D visualization ⓕ tree/hirreachical data visualization ⓖ Network data visualization

objective :- To use python concept

theory :-

what is TABLEAU.

TABLEAU is a visual analyties platform transformating the way we use data to solve problems empouring people and organization to make most of them data

what is data visualization ?

it is graphical representation of information and data by using visual elements like chart graphs and maps provides acceptble way to see understand trends outless pattexers in data

3> list out TABLEAU file extensions :-

workbooks (.twb) - tabes workbook hold one move worksheet

Bookwaks (.tbm) - it is easy way to quickly share your work

- package workbook (twbx) :-
  it is single zip file that contains a workbook along with any supporting file data.
- extract (.hyper or .tde) -
  extract file are local copy of subset or entire data set

4> what are file size limitations with tableau?
A site convers with 100 GB storage capacity workbooks published data source individual workbook pubished to your site can maximum size of 15GB

Conclusion:-
Studied and implemented concepts of python programing