

Aim :- Install and run Pig then write Pig latin scripts to sort, and filter your data.

Theory :-

Apache Pig :- It is a high level platform for creating programs that run on hadoop. It is used to analyze large datasets using a scripted language called Pig latin.

Procedure :-

1) Creating two CSV files :

1st file :

-> nano data.csv

add sample data

1, John, 50

2, Ram, 45

3, Kavin, 80

2nd file :

-> nano others data.csv

1, HR

2, Finance

3, I.T

Teacher's Signature : _____

2) Load data into Pig

→ pig -x local

a) Load the first dataset:

→ A = LOAD 'data.csv' USING PigStorage(',') AS
(id:int, name:chararray, age:int);

b) Load the second dataset:

→ D = LOAD 'other-data.csv' USING PigStorage(',') AS
(id:int, dep:chararray);

3) Sort the data:

→ B = ORDER A BY age DESC;
DUMP B;

4) Group the data:

→ C = GROUP A BY age;
DUMP C;

5) Join the data;

→ E = JOIN A BY id, D BY id;
DUMP E;

Teacher's Signature : _____

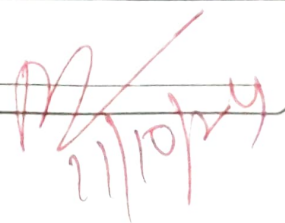
6) Filter the data:

⇒ F = FILTER A BY age > 25 ;
DUMP F;

Learning Outcomes :-

- Understanding of Pig latin
- learn how to load data from CSV file into Apache Pig using the LOAD statement
- Gains the ability to sort datasets based on specific fields using ORDER BY clause.

Teacher's Signature :


21/10/24

Experiment - 7

Aim :- Implement Sqoop commands for importing and exporting files.

Theory :- Apache Sqoop is a tool designed for efficient data transfer between Hadoop and relational databases, allowing users to import data from a relational database into Hadoop's HDFS or export data from HDFS back to the database.

Procedure :-

i) Open MySQL for content creation and display :-

ii) Creation of Table using commands :-
create table reg (id varchar(5), name varchar(20));

iii) Inserting values using commands :-
insert into reg (id, name) values ('3a', 'Gaurima');

iv) Display the contents using :-
mysql> select * from reg;

id	name
3a	Gaurima

2) Open sqoop and connect MySQL with it to import the table created or viewed into HDFS

i) For importing files :-
sqoop import \

ii) For exporting files :-
sqoop export \

Learning Outcomes :-

- I learnt about Apache Sqoop in Hadoop
- I learnt about the import and export commands
- I learnt about the MySQL and its uses.

Experiment - 8

Aim :- Analyzing data with watson studio, run through a sample notebook in Watson Studio

Theory :- IBM Watson Studio is a cloud-based data science and machine learning platform that enables users to prepare, analyze, and visualize data using various tools and libraries like Jupyter, Python, R and Scala.

Software Required :- IBM id, Internet, Web Browser

Procedure :-

1) Accessing Watson Studio:

- Go to IBM cloud and log in your account
- Open Watson Studio from dashboard in the catalog.
- Create a new project in Watson Studio.

2) Import the Sample Notebook:

- Go to Asset tab and click to add to Project.
- Select Notebook from option.
- Create new notebook or import existing sample notebook

3) Running the Notebook :

- Open the notebook by clicking on its name.
- To run a cell, click on cell and press shift + Enter or click run.

4) Visualizing Results :

After executing the notebook, review the visualization and output to analyze insights from the data.

Learning Outcomes :-

- I learnt about IBM Watson studio.
- I learnt about the different services in Watson.
- I learnt about the collaboration for the project.

Experiment - 9

Aim :- Using Fluid Query with Big SQL.

Theory :- Fluid Query is an IBM tool designed to facilitate data movement and analysis across different data storage systems within the IBM ecosystem, particularly in hybrid cloud environments.

Procedure :-

- First, enable the `!sqsh` command line using '`!table`' command which is used for displaying all the already present schemas and table names.
- Then, create a table with column names & types so we can describe the table using some commands.
- Furthermore, we create a folder in `bigsql` folder called `sample data` and transfer dataset in the file called `sample` on the Desktop.

- Then, this sample data in bigsql folder is overwritten in the table created in Big sql earlier.

Learning Outcomes :-

- I learnt about the Fluid Query.
- I learnt about the BigSql.
- I learnt how to query and analyze data from different sources.

Experiment - 10

Aim :- Implement HBase commands with dataset.

Theory :- HBase is a distributed, non-relational, column oriented database built on top of Hadoop HDFS. It is designed to handle large amount of sparse data in a scalable and fault-tolerant way.

Procedure :-

1. Type 'hbase shell' to open hbase.
2. Create table t1 by typing the command line create 't1', 'name', 'marks'.
3. Type command, put 't1', '1', 'name:fname', 'Anandita'.
4. To view the table timestamp, use command scan 't1'.
5. In order to disable the table or check the table is disable, we use disable 't1' or is_disabled 't1'.

Teacher's Signature : _____

6. To count the no. of rows and check the status of hbase node we use commands like ; count 't1', status 'summary'

7. In order to disable the tables starting with specific letter and drop a table we use commands ~~like~~ ; disable-all 't.*'

Learning Outcomes :-

- I learnt about HBase commands.
- I learnt how to query large-scale data.
- I learnt about Hadoop ecosystem for large scale batch processing.