

Task No: 10

Date:

i. Construct the Pig Latin Scripts to find character Count

ii. Construct the Pig Latin Scripts to find a max temp for each and every year.

Tools: Pig, LINUX / WINDOWS

CO4

K3

AIM:

To Implement the Pig Latin scripts to find character Count, max temp for each and every year.

PROCEDURE:

Construct the Pig Latin Scripts to find character Count

- 1) Load the data from LOCAL file system
- 2) Convert the Sentence into words using TOKENIZE Function with delimiter
- 3) Convert Column into Rows
- 4) Apply GROUP BY and Generate word count
- 5) Dump the word count

Construct the Pig Latin Scripts to find a max temp for each and every year

- 1) Load your data (assuming it's in a tab-delimited file)
- 2) Extract the year from the date column (assuming date is in 'YYYY-MM-DD' format)
- 3) Group the data by year
- 4) Calculate the maximum temperature for each year
- 5) Store the result in an output file

Pig Latin Scripts:

(i) Construct the Pig Latin Scripts to find character Count

File Name: wordcount.pig

```
data = LOAD './countwords.txt' USING PigStorage() AS (line:chararray);
```

```
Words = FOREACH data GENERATE FLATTEN(TOKENIZE(line,' ')) AS word;
```

```
Grouped = GROUP Words BY word;
```

```
wordcount = FOREACH Grouped GENERATE group, COUNT(Words);
```

```
DUMP wordcount;
```

(ii) Construct the Pig Latin Scripts to find a max temp for each and every year

File Name: tempanalysis.pig

```
raw_data = LOAD './temperature.txt' USING PigStorage('\t') AS (date:chararray, temperature:double);
```

```
data_with_year = FOREACH raw_data GENERATE SUBSTRING(date, 0, 4) AS year, temperature;
```

```
grouped_data = GROUP data_with_year BY year;
```

```
max_temp_per_year = FOREACH grouped_data GENERATE group AS year,  
MAX(data_with_year.temperature) AS max_temperature;  
  
STORE max_temp_per_year INTO 'your_output_file.txt' USING PigStorage('\t');
```

Input Files:

countwords.txt

This is a hadoop post
hadoop is a bigdata technology

temperature.txt

9/12/2020	33.4
7/13/2021	30.45
9/14/2020	28.6
8/15/2023	30.4
6/16/2022	33.5
9/17/2023	38.6
9/18/2020	38.5
5/19/2023	36.4
9/20/2023	35.7
9/21/2021	36.7
2/22/2023	39.4
9/23/2023	40.2
3/24/2022	40.6
1/25/2022	40.8
9/26/2023	42.8

OUTPUT:

grunt>exec wordcount.pig

(a,2)
(is,2)
(This,1)
(post,1)
(hadoop,2)
(bigdata,1)
(technology,1)

grunt>exec tempanalysis.pig

(2020,38.5)
(2021,36.7)
(2022,40.8)
(2023,42.8)

Result:

Thus the Pig Latin script, to find character Count, max temp for each and every year was executed successfully.