

Practical 8 - Write a Pig Script for solving counting problems.

Step 1: Start Necessary Services

Before starting the practical, ensure that the necessary services, including HDFS,YARN are running on the Cloudera cluster.

The screenshot shows the Cloudera Management Console (CMF) home page. The browser address bar displays 'quickstart.cloudera:7180/cmf/home'. The page header includes navigation links for Cloudera, Hue, Hadoop, HBase, and Impala. Below the header, there are tabs for 'Status', 'All Health Issues' (with 21 issues), and 'Configuration' (with 12 configurations). The main content area lists various services and their status:

Service	Status	Issues	Actions
Cloudera QuickStart (CDH 5.13.0, Packages)	Running	0	1
Hosts	Running	2	1
HBase	Running	4	0
HDFS	Running	4	2
Hive	Running	0	2
Hue	Running	0	1
Impala	Running	0	0
Key-Value Store	Running	0	0
Oozie	Running	0	0
Solr	Running	0	0
Spark	Running	1	0
Sqoop 1 Client	Running	0	0
Sqoop 2	Running	1	1
YARN (MR2 ...)	Running	4	0

Step 2: Create Input File input.csv

Open a new terminal and create a CSV file named input.csv:

```
[cloudera@quickstart ~]$ cat > /home/cloudera/input.csv
People die when they are killed.
Don't talk, it makes you sound stupid.
The ocean is so salty because everyone pees in it.
If you see a stranger, follow him.^Z
[1]+  Stopped                  cat > /home/cloudera/input.csv
```

View the data –

```
[cloudera@quickstart ~]$ cat /home/cloudera/input.csv
People die when they are killed.
Don't talk, it makes you sound stupid.
The ocean is so salty because everyone pees in it.
[cloudera@quickstart ~]$ █
```

Step 3: View Data Through Pig

To view the data through Pig, enter the Pig shell:

```
[cloudera@quickstart ~]$ pig -x local
```

Step 4: Pig Shell Prompt

After the above command, your prompt will change to (grunt>).

```
grunt> █
```

Step 5: Pig Code for Word Count

Enter the following Pig code to get the word count of your file:

```
grunt> lines = LOAD '/home/cloudera/input.csv' AS (line:chararray);
grunt> words = FOREACH lines GENERATE FLATTEN(TOKENIZE(line)) as word;
grunt> grouped = GROUP words BY word;
grunt> wordcount = FOREACH grouped GENERATE group, COUNT(words);
grunt> DUMP wordcount;
```

Step 6: View the Output

Once you enter after writing the above code, you will see the output as:

HadoopVersion	PigVersion	UserId	StartedAt	FinishedAt	Features	
2.6.0-cdh5.13.0	0.12.0-cdh5.13.0	cloudera	2024-02-22 00:32:13	2024-02-22 00:32:27	GROUP_BY	

Success!

Job Stats (time in seconds):

JobId	Alias	Feature	Outputs	
job_local1793176200_0001		grouped,lines,wordcount,words	GROUP_BY,COMBINER	file:/tmp/temp443132024/tmp2122179645,

Input(s):

Successfully read records from: "/home/cloudera/input.csv"

Output(s):

Successfully stored records in: "file:/tmp/temp443132024/tmp2122179645"

Job DAG:

job_local1793176200_0001

```
(in,1)
(is,1)
(it,1)
(so,1)
(The,1)
(are,1)
(die,1)
(it.,1)
(you,1)
(pees,1)
(talk,1)
(they,1)
(when,1)
(makes,1)
(ocean,1)
(salty,1)
(sound,1)
(People,1)
(Don't,1)
(because,1)
(killed.,1)
(stupid.,1)
(everyone,1)
grunt> █
```