Program 5: Run the Pig Latin Scripts to find Word Count.

Login into Hadoop user you used while installing Hadoop , here we use hdoop user

```
hdoop@NuvobookV1:~$
```

$ su – hdoop

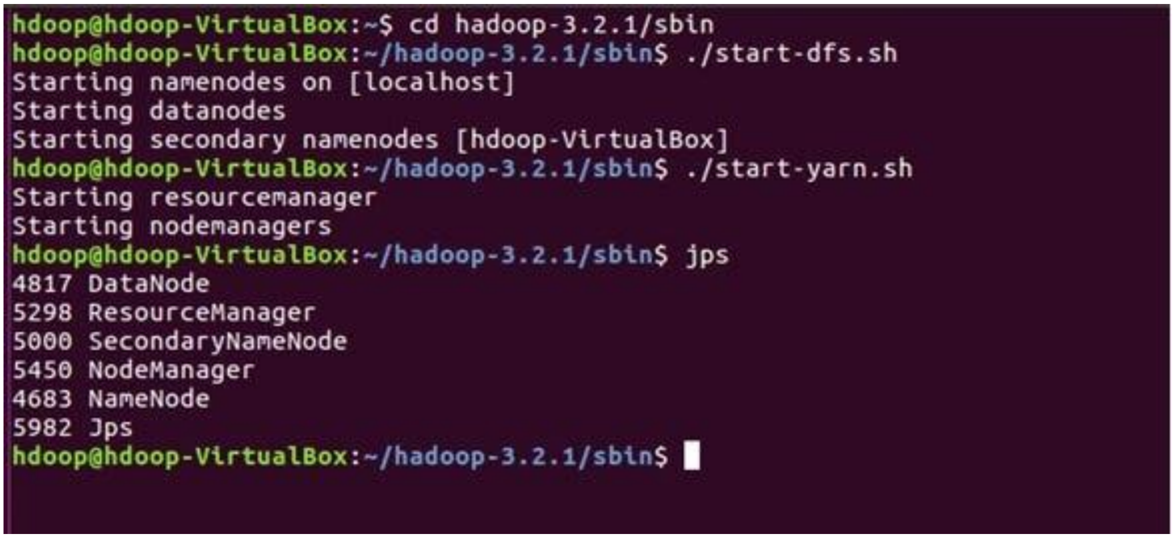
-start the Hadoop server

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

$ jps



```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

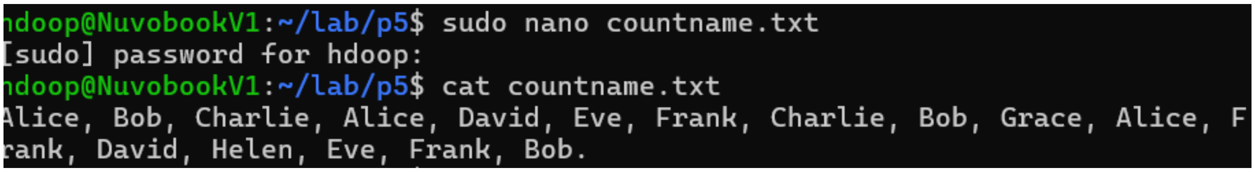
$ sudo mkdir p5/

$ cd p5/

$sudo nano countname.txt

-Insert this:

Alice, Bob, Charlie, Alice, David, Eve, Frank, Charlie, Bob, Grace, Alice, Frank, David, Helen, Eve, Frank, Bob



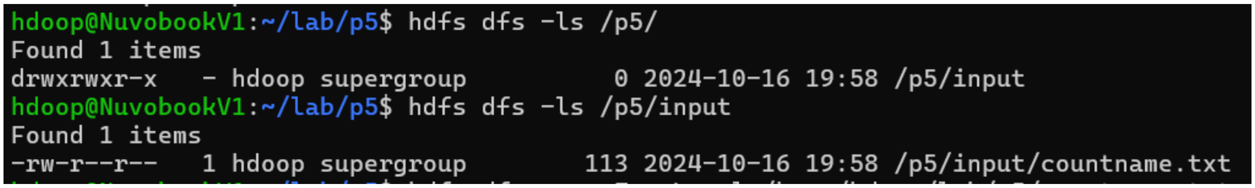
```
hdoop@NuvobookV1:~/lab/p5$ sudo nano countname.txt
[sudo] password for hdoop:
hdoop@NuvobookV1:~/lab/p5$ cat countname.txt
Alice, Bob, Charlie, Alice, David, Eve, Frank, Charlie, Bob, Grace, Alice, F
rank, David, Helen, Eve, Frank, Bob.
```

#upload the file to Hadoop:

$ hdfs dfs -mkdir /p5/

$ hdfs dfs -mkdir /p5/input

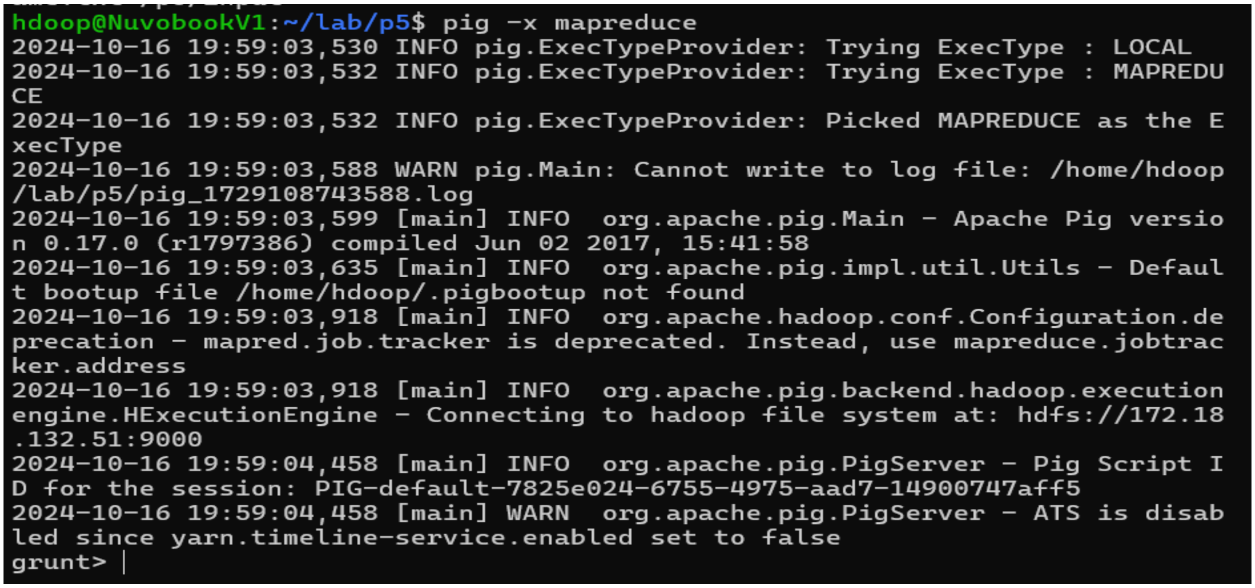
$ hdfs dfs -copyFromLocal /home/hdoop/p5/countname.txt /p5/input



```
hdoop@NuvobookV1:~/lab/p5$ hdfs dfs -ls /p5/
Found 1 items
drwxrwxr-x   - hdoop supergroup          0 2024-10-16 19:58 /p5/input
hdoop@NuvobookV1:~/lab/p5$ hdfs dfs -ls /p5/input
Found 1 items
-rw-r--r--   1 hdoop supergroup        113 2024-10-16 19:58 /p5/input/countname.txt
```

#run pig :

$ pig -x mapreduce

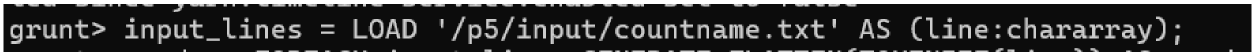


```
hdoop@NuvobookV1:~/lab/p5$ pig -x mapreduce
2024-10-16 19:59:03,530 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-16 19:59:03,532 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDU
CE
2024-10-16 19:59:03,532 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the E
xecType
2024-10-16 19:59:03,588 WARN pig.Main: Cannot write to log file: /home/hdoop
/lab/p5/pig_1729108743588.log
2024-10-16 19:59:03,599 [main] INFO  org.apache.pig.Main - Apache Pig versio
n 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-10-16 19:59:03,635 [main] INFO  org.apache.pig.impl.util.Utils - Defaul
t bootup file /home/hdoop/.pigbootup not found
2024-10-16 19:59:03,918 [main] INFO  org.apache.hadoop.conf.Configuration.de
precation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrac
ker.address
2024-10-16 19:59:03,918 [main] INFO  org.apache.pig.backend.hadoop.execution
engine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.18
.132.51:9000
2024-10-16 19:59:04,458 [main] INFO  org.apache.pig.PigServer - Pig Script I
D for the session: PIG-default-7825e024-6755-4975-aad7-14900747aff5
2024-10-16 19:59:04,458 [main] WARN  org.apache.pig.PigServer - ATS is disab
led since yarn.timeline-service.enabled set to false
grunt> |
```

#write map reduce program in pig:

-load the file countname that we upload to Hadoop:

input_lines = LOAD '/p5/input/countname.txt' AS (line:chararray);



```
grunt> input_lines = LOAD '/p5/input/countname.txt' AS (line:chararray);
```

-the rest of program which it will count the names:

Note: insert each line separately ,it should end up with “;”

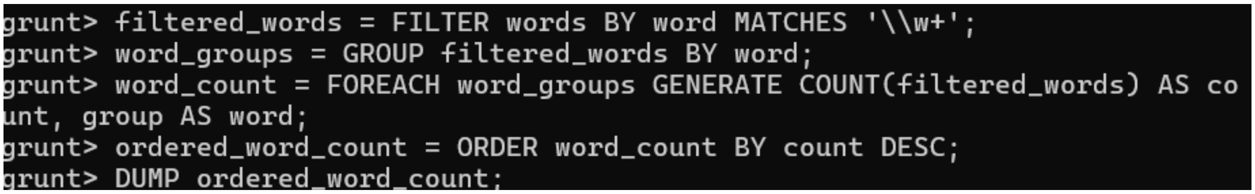
words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;

filtered_words = FILTER words BY word MATCHES ‘\\w+’;

word_groups = GROUP filtered_words BY word;

word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS word;

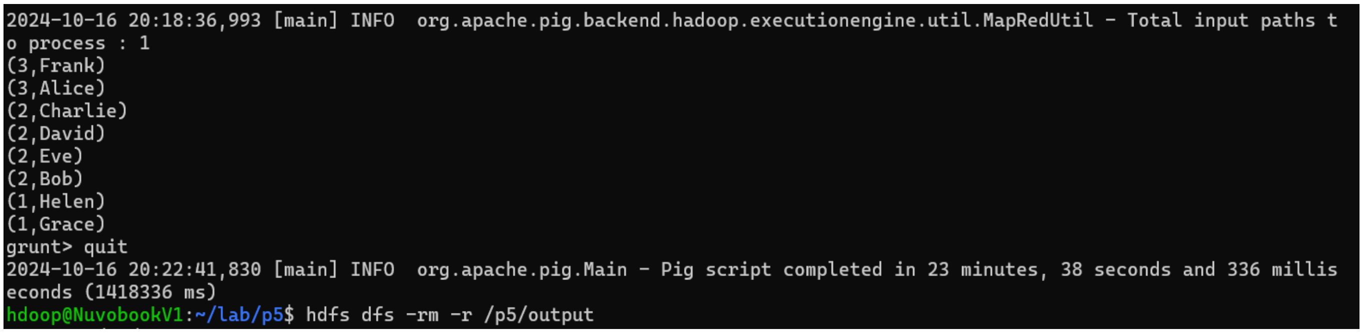
ordered_word_count = ORDER word_count BY count DESC;
DUMP ordered_word_count;



```
grunt> filtered_words = FILTER words BY word MATCHES '\\w+';
grunt> word_groups = GROUP filtered_words BY word;
grunt> word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS co
unt, group AS word;
grunt> ordered_word_count = ORDER word_count BY count DESC;
grunt> DUMP ordered_word_count;
```

Note: it will take approx. 10 minutes to compile.

Output:



```
2024-10-16 20:18:36,993 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths t
o process : 1
(3,Frank)
(3,Alice)
(2,Charlie)
(2,David)
(2,Eve)
(2,Bob)
(1,Helen)
(1,Grace)
grunt> quit
2024-10-16 20:22:41,830 [main] INFO  org.apache.pig.Main - Pig script completed in 23 minutes, 38 seconds and 336 millis
econds (1418336 ms)
hdoop@NuvobookV1:~/lab/p5$ hdfs dfs -rm -r /p5/output
```

-write quit to stop pig

#stop Hadoop:

$stop-all.sh