ASSIGNMENT PIG

TASK 1: WORD COUNT

EXPLANATION: BEFORE STARTING TH PIG SHELL WE NEED TO MAKE A FILE INSIDE HDFS AND ENTER SOME DATA. SO MADE A FILE CALLED "wordcount.txt" WHICH IS SAVED IN PATH '/user/acadgild/wordcount.txt'. AFTER THAT WE LOGIN INTO PIG SHELL BY GIVING COMMAND IN HDFS SHELL AS "pig"

COMMAND: pig

SOLUTION REPORT:

```
[acadgild@localhost ~]$ pig
19/01/14 14:12:58 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/01/14 14:12:58 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/01/14 14:12:58 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2019-01-14 14:12:58,283 [main] INFO  org.apache.pig.Main - Apache Pig version
0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2019-01-14 14:12:58,283 [main] INFO  org.apache.pig.Main - Logging error
messages to: /home/acadgild/pig_1547455378277.log
2019-01-14 14:12:58,405 [main] INFO  org.apache.pig.impl.util.Utils - Default
bootup file /home/acadgild/.pigbootup not found
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-
2.6.5/share/hadoop/common/lib/slf4j-log4j12-
1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-
1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2019-01-14 14:13:00,136 [main] WARN  org.apache.hadoop.util.NativeCodeLoader -
Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
2019-01-14 14:13:00,213 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is
deprecated. Instead, use mapreduce.jobtracker.address
2019-01-14 14:13:00,213 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 14:13:00,213 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to
hadoop file system at: hdfs://localhost:8020
2019-01-14 14:13:01,988 [main] INFO  org.apache.pig.PigServer - Pig Script ID
for the session: PIG-default-a107aea0-9d4c-475b-bd5a-e0c4ae64d5d3
2019-01-14 14:13:01,988 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt>
```

COMMAND: A = LOAD '/user/acadgild/wordcount.txt' as (line:chararray);

EXPLANATION: HERE THE ABOVE COMMAND WILL LOAD THE REQUIRED INPUT DATA FROM THE HDFS PATH. AS INSIDE THE FILE THE DATA IS HAVING ONLY TEXT FILE SO WE GIVE SCHEMA FOR DATA AS CHARARRAY. ALL THESE WILL BE SAVED IN HEADER A.

SOLUTION REPORT:

```
grunt> A = LOAD '/user/acadgild/wordcount.txt' as  (line:chararray);
2019-01-13 22:05:17,511 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
```

COMMAND: Words = foreach A generate FLATTEN (TOKENIZE(line,' ')) as word;

EXPLANATION: AS THE INPUT DATA PRESENT IS IN SENTENCES SO FOR THE WORD COUNT OPERATION. FIRST WE NEED TO CONVERT THE SENTENCES INTO WORDS. "FOREACH" KEYWORD

IS USED TO SELECT THE WHATEVER DATA IS LOADED INTO A. "FLATTEN(TOKENIZE)" KEYWORD IS USED TO CONVERT THE DATA PRESENT AS SENTENCES INTO MULTIPLE ROWS OF WORDS. THUS USING FLATTEN FUNCTION THE BAG IS CONVERTED INTO TUPLE, MEANS THE ARRAY OF STRINGS CONVERTED INTO MULTIPLE ROWS

AS WE HAD PROVIDED INPUT DATA AS "HELLO WORLD HOW ARE YOU". THIS WILL BE CONVERTED AS -
(HELLO)
(WORLD)
(HOW)
(ARE)
(YOU)

EXPLANATION: NOW WE NEED TO COUNT THE OCCURANCES OF THE WORDS. FOR THAT WE FIRST WE HAVE TO GROUP ALL THE WORDS USING "GROUP BY" KEYWORD. AFTER THIS WE SELECT ALL THESE WORDS AND LOAD IT HEADER "word_count" AND WITH THE KEYWORD "COUNT" WE COUNT THE NUMBER OF WORDS. THEN USED THE "ORDER BY" KEYWORD TO ORDER THE WORDS IN DESCENDING ORDER. THEN LATER WE SAVED THIS IN A OUTPUT DIRECTORY "/wordcount_output" WHICH WILL BE IN HDFS. THEN THE JAVA PROGRAM WILL RUN AND AFTER THE SUCCESS MESSAGE WE CAN CHECK THE OUTPUT IN HDFS PATH "/user/acadgild/wordcount_output".

COMMAND: word_groups = group Words by word;
word_count = foreach word_groups generate group, COUNT(Words);
ordered_word_count = order word_count by group desc;
store ordered_word_count into '/wordcount_output';

SOLUTION REPORT:
grunt> Words = foreach A generate FLATTEN (TOKENIZE(line,' ')) as word;
2019-01-13 22:25:43,763 [main] INFO
org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752

grunt> word_groups = group Words by word;
grunt> word_count = foreach word_groups generate group, COUNT(Words);
grunt> ordered_word_count = order word_count by group desc;
grunt> store ordered_word_count into '/wordcount_output';
erver: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2019-01-13 22:39:44,879 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2019-01-13 22:41:59,521 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2019-01-13 22:42:00,526 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2019-01-13 22:42:01,527 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2019-01-13 22:42:02,527 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2019-01-13 22:42:03,529 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)

```
2019-01-13 22:42:04,529 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:05,530 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:06,530 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:06,637 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:42:07,638 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:08,639 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:09,640 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:10,641 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:11,641 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:12,642 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:13,643 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:14,643 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:15,644 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:16,644 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:16,746 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Unable to get job related diagnostics
2019-01-13 22:42:16,747 [main] INFO  org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at localhost/127.0.0.1:8032
2019-01-13 22:42:16,764 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:42:17,788 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
```

```
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:18,789 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:19,790 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:20,792 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:21,792 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:22,793 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:23,794 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:24,795 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:25,796 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:26,797 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:26,907 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:42:27,908 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:28,908 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:29,910 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:30,910 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:31,911 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:32,911 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
```

```
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:33,913 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:34,914 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:35,915 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:36,917 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:37,027 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:42:38,027 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:39,029 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:40,030 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:41,032 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:42,034 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:43,035 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:44,037 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:45,039 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:46,042 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:47,044 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:47,147 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Unable to retrieve job to compute warning aggregation.
```

```
2019-01-13 22:42:47,165 [main] INFO  org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at localhost/127.0.0.1:8032
2019-01-13 22:42:47,176 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:42:48,201 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:49,203 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:50,204 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:51,206 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:52,208 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:53,217 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:54,220 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:55,221 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:56,224 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:57,226 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:57,332 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:42:58,333 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:42:59,339 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:00,340 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:01,342 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
```

```
2019-01-13 22:43:02,343 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:03,345 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:04,346 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:05,348 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:06,349 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:07,351 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:07,461 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:43:08,463 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:09,464 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:10,464 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:11,465 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:12,466 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:13,468 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:14,469 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:15,469 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:16,470 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:17,472 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
```

```
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:17,574 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Unable to get job related diagnostics
2019-01-13 22:43:17,576 [main] INFO  org.apache.hadoop.yarn.client.RMProxy -
Connecting to ResourceManager at localhost/127.0.0.1:8032
2019-01-13 22:43:17,605 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:43:18,688 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:19,689 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:20,690 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:21,691 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:22,691 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:23,692 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:24,692 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:25,693 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:26,693 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:27,694 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:27,801 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:43:28,801 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:29,802 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:30,803 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
```

```
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:31,805 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:32,805 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:33,806 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:34,807 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:35,808 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:36,808 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:37,810 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:37,927 [main] INFO
org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2019-01-13 22:43:38,929 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 0 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:39,931 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 1 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:40,933 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 2 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:41,935 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 3 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:42,937 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 4 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:43,938 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 5 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:44,939 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 6 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:45,941 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 7 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
```

```
MILLISECONDS)
2019-01-13 22:43:46,942 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:47,943 [main] INFO  org.apache.hadoop.ipc.Client - Retrying
connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time(s); retry policy
is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000
MILLISECONDS)
2019-01-13 22:43:48,046 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Unable to retrieve job to compute warning aggregation.
2019-01-13 22:43:48,046 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Success!
```

COMMAND: hadoop fs -cat /wordcount_output/part-r-00000

SOLUTUON REPORT:

```
[acadgild@localhost ~]$ hadoop fs -ls /wordcount_output
19/01/13 22:38:22 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2019-01-13 22:36
/wordcount_output/_SUCCESS
-rw-r--r--   1 acadgild supergroup         34 2019-01-13 22:36
/wordcount_output/part-r-00000
[acadgild@localhost ~]$ hadoop fs -ls /wordcount_output/part-r-00000
19/01/13 22:40:41 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
-rw-r--r--   1 acadgild supergroup         34 2019-01-13 22:36
/wordcount_output/part-r-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /wordcount_output/part-r-00000
19/01/13 22:40:55 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
you   1
world 1
how   1
hello 1
are   1
```

OUTPUT:

REFER SCREENSHOT PIG0.png

TASK 2:

EXPLANATION: HERE WE NEED TO START THE PIG IN LOCAL MODE SO WE USE THE ABOVE BELOW COMMAND.

COMMAND: pig -x local

SOLUTION REPORT:

```
[acadgild@localhost ~]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-
2.6.5/share/hadoop/common/lib/slf4j-log4j12-
1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-
1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
```

```
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
19/01/14 00:14:47 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/01/14 00:14:47 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2019-01-14 00:14:47,468 [main] INFO  org.apache.pig.Main - Apache Pig version
0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2019-01-14 00:14:47,469 [main] INFO  org.apache.pig.Main - Logging error
messages to: /home/acadgild/pig_1547405087466.log
2019-01-14 00:14:47,648 [main] INFO  org.apache.pig.impl.util.Utils - Default
bootup file /home/acadgild/.pigbootup not found
2019-01-14 00:14:48,234 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is
deprecated. Instead, use mapreduce.jobtracker.address
2019-01-14 00:14:48,235 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:14:48,241 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to
hadoop file system at: file:///
2019-01-14 00:14:48,408 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:14:48,503 [main] INFO  org.apache.pig.PigServer - Pig Script ID
for the session: PIG-default-3c09ca30-1854-4965-ac09-24c01499981d
2019-01-14 00:14:48,503 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false

EXPLANATION: AS THE DATASET FOR EMPLOYEE DETAILS AND EMPLOYEE EXPENSES ARE SAVED
IN HDFS IN A FILE AT PATH "/home/acadgild/employee_details.txt" AND
"/home/acadgild/employee_expenses.txt". SO WE LOAD THESE TWO FILES IN PIG. SO
HERE WE USE HEADER "emp_details" FOR LOADING EMPLOYEE DETAILS DATASET WITH THE
BAG CONTAINING FIELDS AS ID, NAME, SALARY AND RATING. THEN FOR EMPLOYEE EXPENSES
DATASET WE LOAD IT IT "emp_expenses" WITH SCHEMA FOR FIELDS AS ID AND EXPENSES.

COMMAND: emp_details = LOAD '/home/acadgild/employee_details.txt' USING
PigStorage(',') AS (id:int, name:chararray, salary:int, rating:int);

COMMAND: emp_expenses = LOAD '/home/acadgild/employee_expenses.txt' USING
PigStorage() AS (id:int, expenses:int);

SOLUTION REPORT:
grunt> emp_details = LOAD '/home/acadgild/employee_details.txt' USING
PigStorage(',') AS (id:int, name:chararray, salary:int, rating:int);
2019-01-14 00:25:52,923 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:25:52,923 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
grunt>  emp_expenses = LOAD '/home/acadgild/employee_expenses.txt' USING
PigStorage() AS (id:int, expenses:int);
2019-01-14 00:23:57,028 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:23:57,031 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS

EXPLANATION: USING DUMP COMMAND WE CAN SEE THE DATASET LOADED.

COMMAND: dump emp_details;
COMMAND: dump emp_expenses;

SOLUTION:
```

```
grunt> dump emp_details;
2019-01-14 00:26:27,585 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
Pig features used in the script: UNKNOWN
2019-01-14 00:26:27,691 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:26:27,691 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:26:27,691 [main] WARN  org.apache.pig.data.SchemaTupleBackend -
SchemaTupleBackend has already been initialized
2019-01-14 00:26:27,694 [main] INFO
org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer -
{RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator,
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter,
MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer,
PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2019-01-14 00:26:27,709 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File
concatenation threshold: 100 optimistic? false
2019-01-14 00:26:27,726 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size before optimization: 1
2019-01-14 00:26:27,727 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size after optimization: 1
2019-01-14 00:26:27,768 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:26:27,769 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:26:27,782 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:26:27,783 [main] INFO
org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are
added to the job
2019-01-14 00:26:27,789 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2019-01-14 00:26:27,798 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Setting up single store job
2019-01-14 00:26:27,798 [main] INFO  org.apache.pig.data.SchemaTupleFrontend -
Key [pig.schematuple] is false, will not generate code.
2019-01-14 00:26:27,798 [main] INFO  org.apache.pig.data.SchemaTupleFrontend -
Starting process to move generated code to distributed cacche
2019-01-14 00:26:27,798 [main] INFO  org.apache.pig.data.SchemaTupleFrontend -
Distributed cache not supported or needed in local mode. Setting key
[pig.schematuple.local.dir] with code temp directory: /tmp/1547405787798-0
2019-01-14 00:26:27,823 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
1 map-reduce job(s) waiting for submission.
2019-01-14 00:26:27,830 [JobControl] INFO
org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
2019-01-14 00:26:27,875 [JobControl] WARN
org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User
classes may not be found. See Job or Job#setJar(String).
2019-01-14 00:26:27,930 [JobControl] INFO  org.apache.pig.builtin.PigStorage -
Using PigTextInputFormat
2019-01-14 00:26:27,935 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
```

```
process : 1
2019-01-14 00:26:27,935 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths to process : 1
2019-01-14 00:26:27,936 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths (combined) to process : 1
2019-01-14 00:26:27,998 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2019-01-14 00:26:28,081 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job:
job_local544345434_0002
2019-01-14 00:26:28,486 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The
url to track the job: http://localhost:8080/
2019-01-14 00:26:28,486 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
HadoopJobId: job_local544345434_0002
2019-01-14 00:26:28,486 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Processing aliases emp_details
2019-01-14 00:26:28,486 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
detailed locations: M: emp_details[5,14],emp_details[-1,-1] C:  R:
2019-01-14 00:26:28,491 [Thread-47] INFO
org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2019-01-14 00:26:28,524 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
0% complete
2019-01-14 00:26:28,524 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Running jobs are [job_local544345434_0002]
2019-01-14 00:26:28,576 [Thread-47] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is
deprecated. Instead, use mapreduce.jobtracker.address
2019-01-14 00:26:28,583 [Thread-47] INFO
org.apache.hadoop.conf.Configuration.deprecation -
mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use
mapreduce.reduce.markreset.buffer.percent
2019-01-14 00:26:28,584 [Thread-47] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:26:28,584 [Thread-47] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:26:28,585 [Thread-47] INFO
org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter is
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigOutputCommitter
2019-01-14 00:26:28,640 [Thread-47] INFO
org.apache.hadoop.mapred.LocalJobRunner - Waiting for map tasks
2019-01-14 00:26:28,641 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - Starting task:
attempt_local544345434_0002_m_000000_0
2019-01-14 00:26:28,740 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task -  Using ResourceCalculatorProcessTree : [ ]
2019-01-14 00:26:28,743 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.MapTask - Processing split: Number of splits :1
Total Length = 263
Input split[0]:
   Length = 263
   ClassName: org.apache.hadoop.mapreduce.lib.input.FileSplit
   Locations:

-----------------------
```

```
2019-01-14 00:26:28,785 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-01-14 00:26:28,786 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigRecordReader -
Current split being processed file:/home/acadgild/employee_details.txt:0+263
2019-01-14 00:26:28,832 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of
size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold
= 489350752
2019-01-14 00:26:28,833 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set...
will not generate code.
2019-01-14 00:26:28,854 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map -
Aliases being processed per job phase (AliasName[line,offset]): M:
emp_details[5,14],emp_details[-1,-1] C:  R:
2019-01-14 00:26:28,869 [LocalJobRunner Map Task Executor #0] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigHadoopLogger -
org.apache.pig.backend.hadoop.executionengine.physicalLayer.expressionOperators.
POProject(ACCESSING_NON_EXISTENT_FIELD): Attempt to access field which was not
found in the input
2019-01-14 00:26:28,869 [LocalJobRunner Map Task Executor #0] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigHadoopLogger -
org.apache.pig.backend.hadoop.executionengine.physicalLayer.expressionOperators.
POProject(ACCESSING_NON_EXISTENT_FIELD): Attempt to access field which was not
found in the input
2019-01-14 00:26:28,869 [LocalJobRunner Map Task Executor #0] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigHadoopLogger -
org.apache.pig.backend.hadoop.executionengine.physicalLayer.expressionOperators.
POProject(ACCESSING_NON_EXISTENT_FIELD): Attempt to access field which was not
found in the input
2019-01-14 00:26:28,872 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner -
2019-01-14 00:26:28,872 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task:attempt_local544345434_0002_m_000000_0 is
done. And is in the process of committing
2019-01-14 00:26:28,883 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner -
2019-01-14 00:26:28,884 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task attempt_local544345434_0002_m_000000_0 is
allowed to commit now
2019-01-14 00:26:28,904 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of
task 'attempt_local544345434_0002_m_000000_0' to file:/tmp/temp2096933922/tmp-
325758746/_temporary/0/task_local544345434_0002_m_000000
2019-01-14 00:26:28,906 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - map
2019-01-14 00:26:28,906 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task 'attempt_local544345434_0002_m_000000_0'
done.
2019-01-14 00:26:28,906 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - Finishing task:
attempt_local544345434_0002_m_000000_0
2019-01-14 00:26:28,906 [Thread-47] INFO
org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2019-01-14 00:26:29,033 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:26:29,053 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:26:29,054 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
```

```
2019-01-14 00:26:29,073 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
100% complete
2019-01-14 00:26:29,074 [main] INFO
org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion     PigVersion  UserId      StartedAt   FinishedAt  Features
2.6.5 0.16.0      acadgild    2019-01-14 00:26:27     2019-01-14 00:26:29
      UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps  Reduces     MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime
      MaxReduceTime     MinReduceTime     AvgReduceTime     MedianReducetime
      Alias Feature     Outputs
job_local544345434_0002 1     0     n/a   n/a   n/a   n/a   0     00    0
      emp_details MAP_ONLY    file:/tmp/temp2096933922/tmp-325758746,

Input(s):
Successfully read 16 records from: "/home/acadgild/employee_details.txt"

Output(s):
Successfully stored 16 records in: "file:/tmp/temp2096933922/tmp-325758746"

Counters:
Total records written : 16
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local544345434_0002


2019-01-14 00:26:29,081 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:26:29,082 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:26:29,086 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:26:29,111 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Encountered Warning ACCESSING_NON_EXISTENT_FIELD 6 time(s).
2019-01-14 00:26:29,111 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Success!
2019-01-14 00:26:29,112 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:26:29,112 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:26:29,112 [main] WARN  org.apache.pig.data.SchemaTupleBackend -
SchemaTupleBackend has already been initialized
2019-01-14 00:26:29,141 [main] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
process : 1
2019-01-14 00:26:29,141 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
```

```
paths to process : 1
(101,Amitabh,20000,1)
(102,Shahrukh,10000,2)
(103,Akshay,11000,3)
(104,Anubhav,5000,4)
(105,Pawan,2500,5)
(106,Aamir,25000,1)
(107,Salman,17500,2)
(108,Ranbir,14000,3)
(109,Katrina,1000,4)
(110,Priyanka,2000,5)
(111,Tushar,500,1)
(112,Ajay,5000,2)
(113,Jubeen,1000,1)
(114,Madhuri,2000,2)
(,,,)
(,,,)

grunt> dump emp_expenses;
2019-01-14 00:27:11,486 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
Pig features used in the script: UNKNOWN
2019-01-14 00:27:11,573 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:27:11,573 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:27:11,573 [main] WARN  org.apache.pig.data.SchemaTupleBackend -
SchemaTupleBackend has already been initialized
2019-01-14 00:27:11,574 [main] INFO
org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer -
{RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator,
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter,
MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer,
PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2019-01-14 00:27:11,582 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File
concatenation threshold: 100 optimistic? false
2019-01-14 00:27:11,594 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size before optimization: 1
2019-01-14 00:27:11,594 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer
- MR plan size after optimization: 1
2019-01-14 00:27:11,629 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:27:11,630 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:27:11,641 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:27:11,642 [main] INFO
org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are
added to the job
2019-01-14 00:27:11,645 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2019-01-14 00:27:11,651 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler
- Setting up single store job
2019-01-14 00:27:11,657 [main] INFO  org.apache.pig.data.SchemaTupleFrontend -
Key [pig.schematuple] is false, will not generate code.
```

```
2019-01-14 00:27:11,658 [main] INFO  org.apache.pig.data.SchemaTupleFrontend -
Starting process to move generated code to distributed cacche
2019-01-14 00:27:11,659 [main] INFO  org.apache.pig.data.SchemaTupleFrontend -
Distributed cache not supported or needed in local mode. Setting key
[pig.schematuple.local.dir] with code temp directory: /tmp/1547405831657-0
2019-01-14 00:27:11,685 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
1 map-reduce job(s) waiting for submission.
2019-01-14 00:27:11,694 [JobControl] INFO
org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
2019-01-14 00:27:11,744 [JobControl] WARN
org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set.  User
classes may not be found. See Job or Job#setJar(String).
2019-01-14 00:27:11,762 [JobControl] INFO  org.apache.pig.builtin.PigStorage -
Using PigTextInputFormat
2019-01-14 00:27:11,763 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
process : 1
2019-01-14 00:27:11,766 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths to process : 1
2019-01-14 00:27:11,768 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
paths (combined) to process : 1
2019-01-14 00:27:11,823 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2019-01-14 00:27:11,871 [JobControl] INFO
org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job:
job_local871729234_0003
2019-01-14 00:27:12,136 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The
url to track the job: http://localhost:8080/
2019-01-14 00:27:12,138 [Thread-70] INFO
org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter set in config null
2019-01-14 00:27:12,163 [Thread-70] INFO
org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is
deprecated. Instead, use mapreduce.jobtracker.address
2019-01-14 00:27:12,163 [Thread-70] INFO
org.apache.hadoop.conf.Configuration.deprecation -
mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use
mapreduce.reduce.markreset.buffer.percent
2019-01-14 00:27:12,163 [Thread-70] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:27:12,163 [Thread-70] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:27:12,173 [Thread-70] INFO
org.apache.hadoop.mapred.LocalJobRunner - OutputCommitter is
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigOutputCommitter
2019-01-14 00:27:12,190 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
HadoopJobId: job_local871729234_0003
2019-01-14 00:27:12,190 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Processing aliases emp_expenses
2019-01-14 00:27:12,190 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
detailed locations: M: emp_expenses[4,15],emp_expenses[-1,-1] C:  R:
2019-01-14 00:27:12,199 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
0% complete
2019-01-14 00:27:12,200 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
```

```
Running jobs are [job_local871729234_0003]
2019-01-14 00:27:12,212 [Thread-70] INFO
org.apache.hadoop.mapred.LocalJobRunner - Waiting for map tasks
2019-01-14 00:27:12,213 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - Starting task:
attempt_local871729234_0003_m_000000_0
2019-01-14 00:27:12,297 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task -  Using ResourceCalculatorProcessTree : [ ]
2019-01-14 00:27:12,302 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.MapTask - Processing split: Number of splits :1
Total Length = 73
Input split[0]:
   Length = 73
   ClassName: org.apache.hadoop.mapreduce.lib.input.FileSplit
   Locations:

-----------------------

2019-01-14 00:27:12,338 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2019-01-14 00:27:12,339 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigRecordReader -
Current split being processed file:/home/acadgild/employee_expenses.txt:0+73
2019-01-14 00:27:12,391 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenured Gen) of
size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold
= 489350752
2019-01-14 00:27:12,395 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set...
will not generate code.
2019-01-14 00:27:12,416 [LocalJobRunner Map Task Executor #0] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map -
Aliases being processed per job phase (AliasName[line,offset]): M:
emp_expenses[4,15],emp_expenses[-1,-1] C:  R:
2019-01-14 00:27:12,429 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner -
2019-01-14 00:27:12,429 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task:attempt_local871729234_0003_m_000000_0 is
done. And is in the process of committing
2019-01-14 00:27:12,436 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner -
2019-01-14 00:27:12,439 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task attempt_local871729234_0003_m_000000_0 is
allowed to commit now
2019-01-14 00:27:12,445 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of
task 'attempt_local871729234_0003_m_000000_0' to file:/tmp/temp2096933922/tmp-
991445478/_temporary/0/task_local871729234_0003_m_000000
2019-01-14 00:27:12,446 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - map
2019-01-14 00:27:12,446 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.Task - Task 'attempt_local871729234_0003_m_000000_0'
done.
2019-01-14 00:27:12,446 [LocalJobRunner Map Task Executor #0] INFO
org.apache.hadoop.mapred.LocalJobRunner - Finishing task:
attempt_local871729234_0003_m_000000_0
2019-01-14 00:27:12,447 [Thread-70] INFO
org.apache.hadoop.mapred.LocalJobRunner - map task executor complete.
2019-01-14 00:27:12,654 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:27:12,659 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
```

```
2019-01-14 00:27:12,665 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:27:12,689 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
100% complete
2019-01-14 00:27:12,693 [main] INFO
org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion     PigVersion  UserId      StartedAt   FinishedAt  Features
2.6.5 0.16.0      acadgild    2019-01-14 00:27:11     2019-01-14 00:27:12
      UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps  Reduces     MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime
      MaxReduceTime     MinReduceTime     AvgReduceTime     MedianReducetime
      Alias Feature     Outputs
job_local871729234_0003 1     0     n/a   n/a   n/a   n/a   0     00    0
      emp_expenses      MAP_ONLY    file:/tmp/temp2096933922/tmp-991445478,

Input(s):
Successfully read 9 records from: "/home/acadgild/employee_expenses.txt"

Output(s):
Successfully stored 9 records in: "file:/tmp/temp2096933922/tmp-991445478"

Counters:
Total records written : 9
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local871729234_0003


2019-01-14 00:27:12,694 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:27:12,704 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:27:12,708 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-01-14 00:27:12,737 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
Success!
2019-01-14 00:27:12,742 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is
deprecated. Instead, use dfs.bytes-per-checksum
2019-01-14 00:27:12,743 [main] INFO
org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is
deprecated. Instead, use fs.defaultFS
2019-01-14 00:27:12,745 [main] WARN  org.apache.pig.data.SchemaTupleBackend -
SchemaTupleBackend has already been initialized
2019-01-14 00:27:12,797 [main] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to
process : 1
2019-01-14 00:27:12,797 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input
```

```
paths to process : 1
(101,200)
(102,100)
(110,400)
(114,200)
(119,200)
(105,100)
(101,100)
(104,300)
(102,400)
```

TASK A: Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference)

EXPLANATION: HERE WE USING ORDER FUNCTION TO ORDER THE RATING AND THE ID BY ASCENDING ORDER AND LIMIT COMMAND TO LIMIT THE EMPLOYEE BY TOP 5 WITH ID AND THE NAME. THEN USING DUMP COMMAND WE CAN SEE THE OUTPUT.

COMMAND:
```
employee = ORDER emp_details BY rating asc, id asc;
limit_employee = LIMIT employee 5;
top_employee = FOREACH limit_employee GENERATE id,name;
```

COMMAND:
```
dump top_employee;
```

SOLUTION REPORT:
```
grunt> employee = ORDER emp_details BY rating asc, id asc;
grunt> limit_employee = LIMIT employee 5;
grunt> top_employee = FOREACH limit_employee GENERATE id,name;

grunt> dump top_employee;

(101,Amitabh,20000,1)
(106,Aamir,25000,1)
(111,Tushar,500,1)
(113,Jubeen,1000,1)
(102,Shahrukh,10000,2)
```

OUTPUT: REFER SCREENSHOT PIG1.png

TASK B: Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number. (In case two employees have same salary, employee with name coming first in dictionary should get preference).

EXPLANATION: HERE WE USING ORDER FUNCTION TO ORDER THE employee_details BY ARRANGING THE SALARY IN DESCENDING ORDER AND FILTERING THE EVEN NYMBER USING THE LOGIC (ID%2)!=0 AND THEN LIMITING EMPLOYEE BY 3, THEN BRINGING THE RESULT ACCORDING TO THE EMPLOYEE ID AND NAME. THEN BY USING DUMP COMMAND WE GET THE OUTPUT.

COMMAND:
```
order_employee = ORDER emp_details BY salary DESC;
filter_employee = FILTER order_employee BY (id%2)!=0;
limit_employee = LIMIT filter_employee 3;
top_3employee = FOREACH limit_employee GENERATE id,name;
```

COMAND:
```
DUMP top_3employee;
```

SOLUTION REPORT:
```
grunt> order_employee = ORDER emp_details BY salary DESC;
grunt> filter_employee = FILTER order_employee BY (id%2)!=0;
grunt> limit_employee = LIMIT filter_employee 3;
grunt> top_3employee = FOREACH limit_employee GENERATE id,name;
```

```
grunt> DUMP top_3employee;

(101,Amitabh)
(107,Salman)
(103,Akshay)
```

OUTPUT: REFER SCREENSHOT PIG2.png

TASK C: Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference).

EXPLANATION: HERE WE JOIN BOTH THE TABLES USING THE JOIN FUNCTION AND SORT THE EXPENSES BY DESCENDING ORDER THEN LIMIT THE EMPLOYEE INTO FIRST TWO ROWS AND BRING THE ID AND NAME OF THE EMPLOYEE AS THE OUTPUT. THEN USING DUMP COMMAND WE GET THE OUTPUT.

```
COMMAND:  emp_det_exp = JOIN emp_details BY id, emp_expenses by id;
        order_exp = ORDER emp_det_exp BY expenses DESC;
        limit_exp = LIMIT order_exp 2;
        max_exp = FOREACH limit_exp GENERATE $0,$1;

COMMAND: dump max_exp;
SOLUTION REPORT:
grunt> emp_det_exp = JOIN emp_details BY id, emp_expenses by id;
grunt> order_exp = ORDER emp_det_exp BY expenses DESC;
grunt>limit_exp = LIMIT order_exp 2;
grunt> max_exp = FOREACH limit_exp GENERATE $0,$1;

grunt> dump max_exp;

(102,Shahrukh)
(110,Priyanka)
```

OUTPUT: REFER SCREENSHOT PIG3.png

TASK D: List of employees (employee id and employee name) having entries in employee_expenses file.

EXPLANATION: HERE WE JOIN THE BOTH TABLES USING JOIN FUNCTION AND REMOVE THE REDUNDANT TUPLES FROM RELATION AND DISPLAY THE ID AND NAME OF THE EMPLOYEE. USING DUMP COMMAND WE GET OUTPUT.

```
COMMAND: emp_det_exp = JOIN emp_details BY id, emp_expenses by id
       emp_names = FOREACH emp_det_exp GENERATE($0,$1);
       names = DISTINCT emp_names;

COMMAND: DUMP names;

grunt> emp_det_exp = JOIN emp_details BY id, emp_expenses by id;
grunt> emp_names = FOREACH emp_det_exp GENERATE($0,$1);
grunt> names = DISTINCT emp_names;

grunt> DUMP names;

((101,Amitabh))
((102,Shahrukh))
((104,Anubhav))
((105,Pawan))
((110,Priyanka))
((114,Madhuri))
```

OUTPUT: REFER SCREENSHOT PIG4.png

TASK E: List of employees (employee id and employee name) having no entry in employee_expenses file.

EXPLANATION: USING LEFT OUTER JOIN, WE JOIN BOTH THE TABLES AND FILTER THE EMPTY VALUES IN THE COLUMNS ($4: emp_expenses.id AND $5: emp_expenses.expenses) AND GENERATE THE OUTPUT BY ID AND NAME OF THE EMPLOYEE.

COMMAND:
```
emp_det_exp = JOIN emp_details BY id LEFT OUTER, emp_expenses BY id;
filter_emp_det_exp = FILTER emp_det_exp BY $4 Is NULL and $5 Is NULL;
gen_emp_det_exp = FOREACH filter_emp_det_exp GENERATE $0,$1;
```

COMMAND:
```
DUMP gen_emp_det_exp;
```

SOLUTION REPORT:
```
grunt> emp_det_exp = JOIN emp_details BY id LEFT OUTER, emp_expenses BY id;
grunt> filter_emp_det_exp = FILTER emp_det_exp BY $4 Is NULL and $5 Is NULL;
grunt> gen_emp_det_exp = FOREACH filter_emp_det_exp GENERATE $0,$1;

grunt> DUMP gen_emp_det_exp;

(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
(,)
(,)
```

OUTPUT: REFER SCREENSHOT PIG5.png

TASK 3

AS HERE I AM NOT ABLE TO GET THE DATASET FOR Delayed FLIGHTS AND AIRPORT FROM THE GIVEN LINK IN THE BLOG AS IT SAYS THAT THE FILE DOES NOT EXIST IN THE GOOGLE DRIVE. SO BELOW I HAVE JUST GIVEN THE STEPS TO BE DONE.

1) Find out the top 5 most visited destinations.

STEPS:

COMMANDS:

```
REGISTER '/home/acadgild/airline_usecase/piggybank.jar';
A = load '/home/acadgild/airline_usecase/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP
_INPUT_HEADER');
B = foreach A generate (int)$1 as year, (int)$10 as flight_num, (chararray)$17
as origin,(chararray) $18 as dest;
C = filter B by dest is not null;
D = group C by dest;
E = foreach D generate group, COUNT(C.dest);
F = order E by $1 DESC;
Result = LIMIT F 5;
A1 = load '/home/acadgild/airline_usecase/airports.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP
_INPUT_HEADER');
A2 = foreach A1 generate (chararray)$0 as dest, (chararray)$2 as city,
```

```
(chararray)$4 as country;
joined_table = join Result by $0, A2 by dest;
dump joined_table;
```

2) Which month has seen the most number of cancellations due to bad weather?

OUTPUT: REFER SCREENSHOT PIG1.png
COMMANDS:

```
REGISTER '/home/acadgild/airline_usecase/piggybank.jar';
A = load '/home/acadgild/airline_usecase/DelayedFlights.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP
_INPUT_HEADER');
B = foreach A generate (int)$2 as month,(int)$10 as flight_num,(int)$22 as
cancelled,(chararray)$23 as cancel_code;
C = filter B by cancelled == 1 AND cancel_code =='B';
D = group C by month;
E = foreach D generate group, COUNT(C.cancelled);
F= order E by $1 DESC;
Result = limit F 1;
dump Result;
```