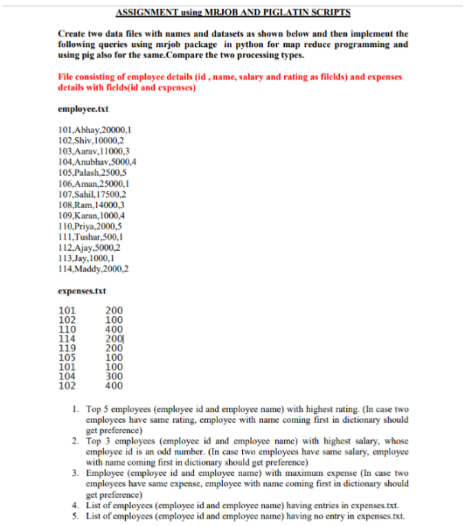
# Assignment of MRjob and Pig script

Assignment:



ASSIGNMENT using MRJOB AND PIGLATIN SCRIPTS

**Create two data files with names and datasets as shown below and then implement the following queries using mrjob package in python for map reduce programming and using pig also for the same.Compare the two processing types.**

**File consisting of employee details (id , name, salary and rating as filelds) and expenses details with fields(id and expenses)**

**employee.txt**

101,Abhay,20000,1
102,Shiv,10000,2
103,Aarav,11000,3
104,Anubhav,5000,4
105,Palash,2500,5
106,Aman,25000,1
107,Sahil,17500,2
108,Ram,14000,3
109,Karan,1000,4
110,Priya,2000,5
111,Tushar,500,1
112,Ajay,5000,2
113,Jay,1000,1
114,Maddy,2000,2

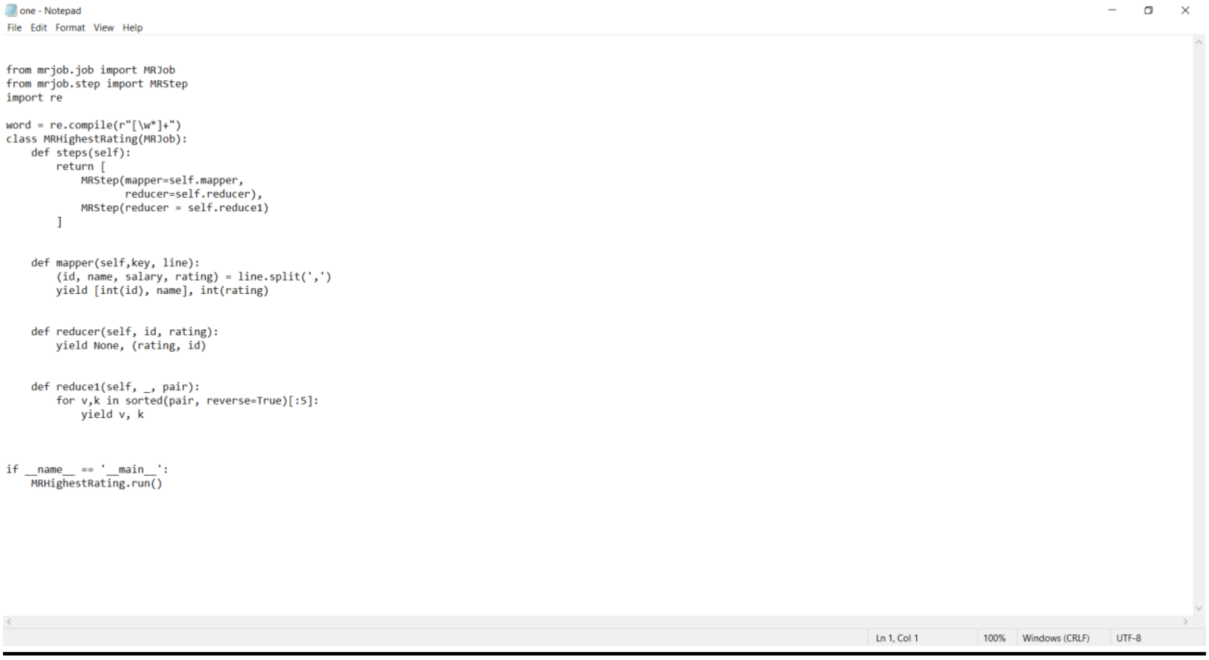
**expenses.txt**

| | |
|---|---|
| 101 | 200 |
| 102 | 100 |
| 110 | 400 |
| 114 | 200 |
| 119 | 200 |
| 105 | 100 |
| 101 | 100 |
| 104 | 300 |
| 102 | 400 |

1. 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)
2. 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)
3. 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)
4. List of employees (employee id and employee name) having entries in expenses.txt.
5. List of employees (employee id and employee name) having no entry in expenses.txt.

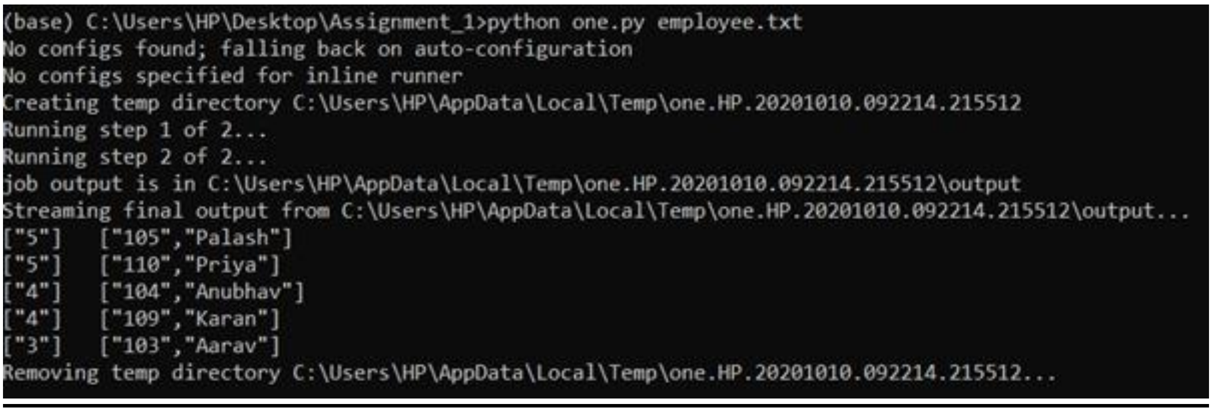
By: Chayan Gulati 18csu054

Q1.



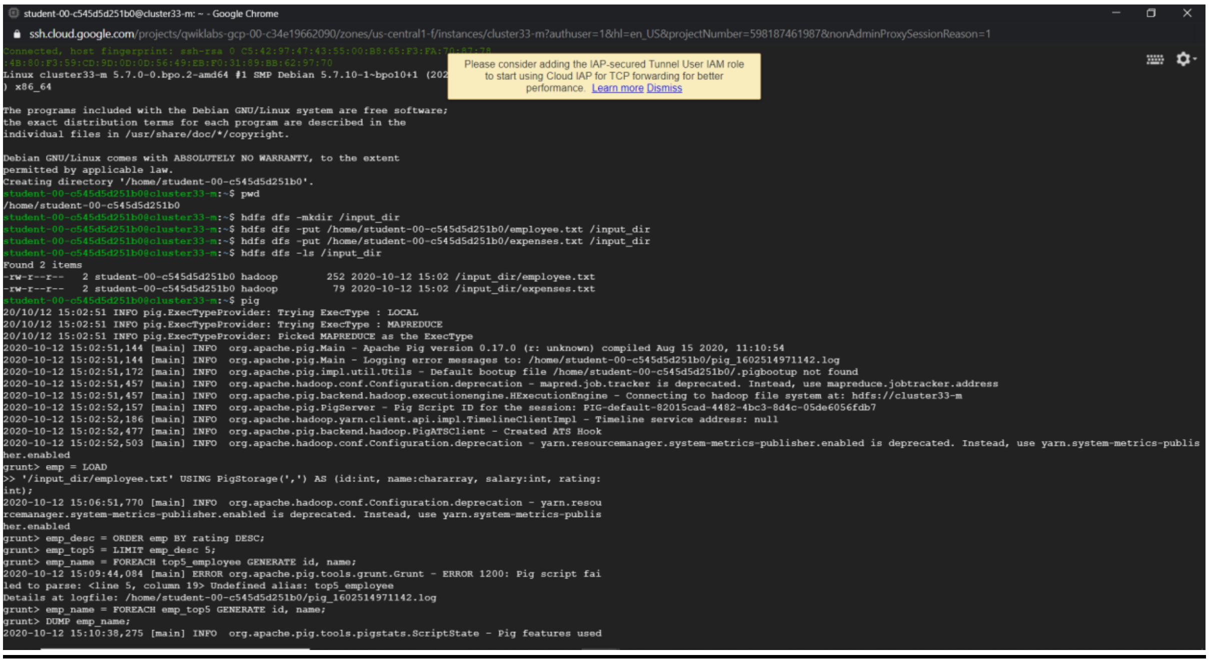
```
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

word = re.compile(r"[\w*]+")
class MRHighestRating(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper,
                   reducer=self.reducer),
            MRStep(reducer = self.reduce1)
        ]


    def mapper(self,key, line):
        (id, name, salary, rating) = line.split(',')
        yield [int(id), name], int(rating)


    def reducer(self, id, rating):
        yield None, (rating, id)


    def reduce1(self, _, pair):
        for v,k in sorted(pair, reverse=True)[:5]:
            yield v, k


if __name__ == '__main__':
    MRHighestRating.run()
```

o/p:



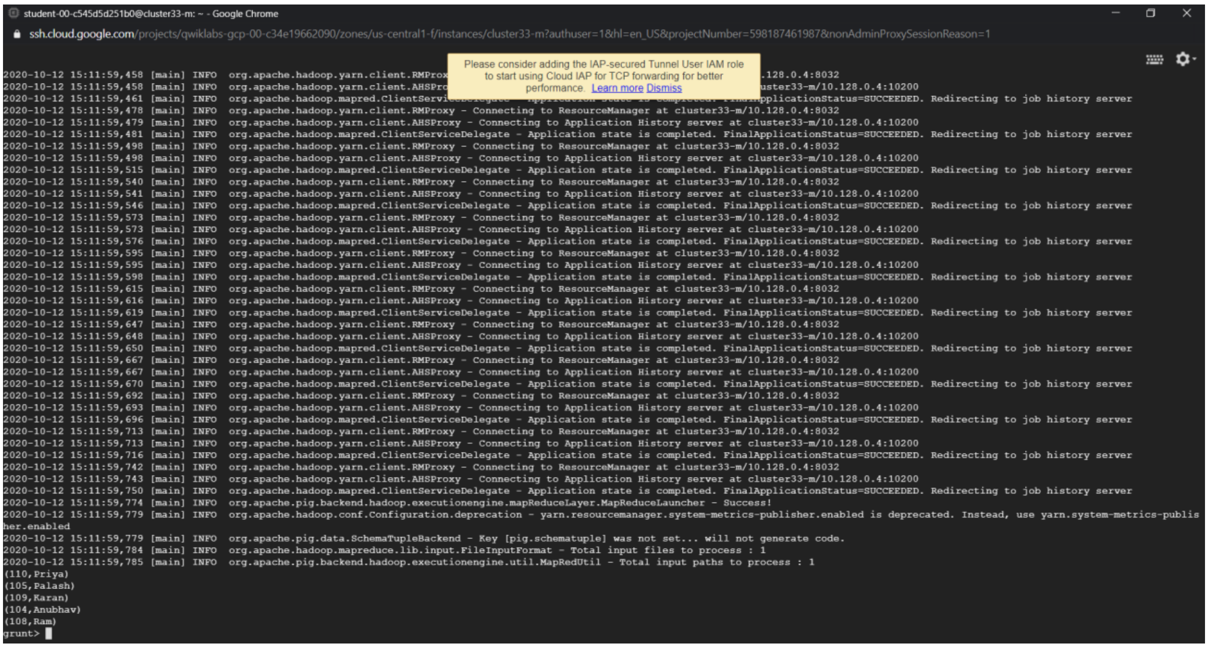
```
(base) C:\Users\HP\Desktop\Assignment_1>python one.py employee.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\HP\AppData\Local\Temp\one.HP.20201010.092214.215512
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\HP\AppData\Local\Temp\one.HP.20201010.092214.215512\output
Streaming final output from C:\Users\HP\AppData\Local\Temp\one.HP.20201010.092214.215512\output...
["5"]   ["105","Palash"]
["5"]   ["110","Priya"]
["4"]   ["104","Anubhav"]
["4"]   ["109","Karan"]
["3"]   ["103","Aarav"]
Removing temp directory C:\Users\HP\AppData\Local\Temp\one.HP.20201010.092214.215512...
```

Using Pig:


student-00-c545d5d251b0@cluster33-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-00-c34e19662090/zones/us-central1-f/instances/cluster33-m?authuser=1&hl=en_US&projectNumber=598187461987&nonAdminProxySessionReason=1

Please consider adding the IAP-secured Tunnel User IAM role to start using Cloud IAP for TCP forwarding for better performance. Learn more Dismiss

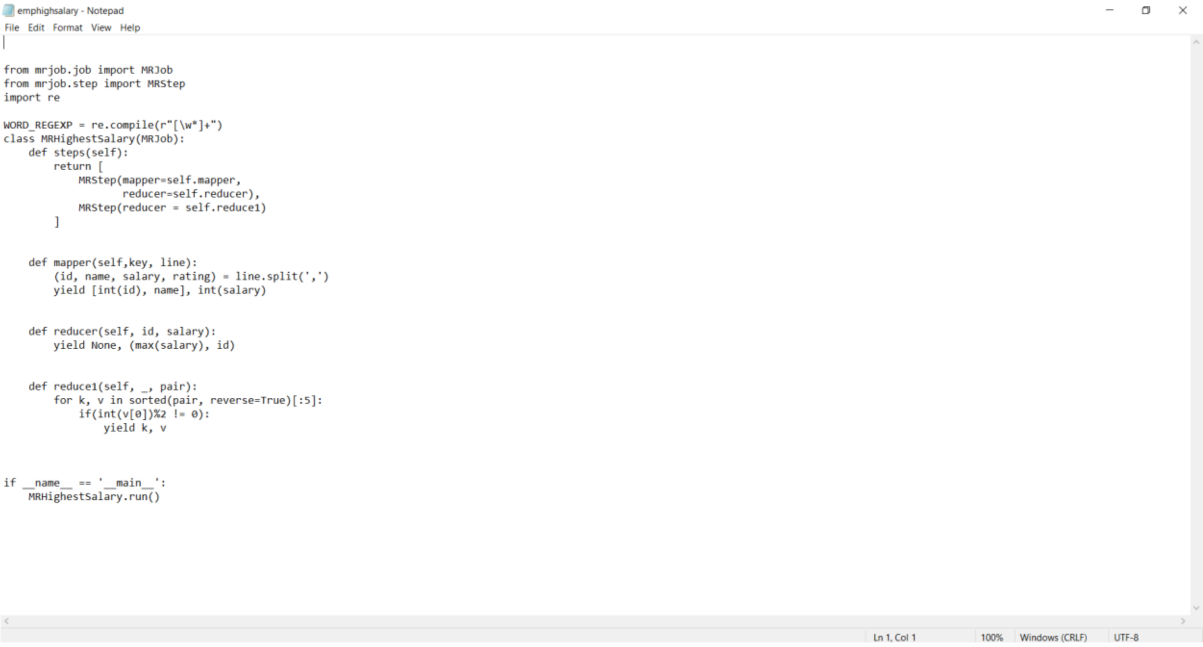
```
Connected, host fingerprint: ssh-rsa 0 C5:42:97:47:43:55:00:B8:65:F3:FA:70:87:78
:4B:80:F3:59:CD:9D:0D:0D:56:49:EB:F0:31:89:BB:62:97:70
Linux cluster33-m 5.7.0-0.bpo.2-amd64 #1 SMP Debian 5.7.10-1~bpo10+1 (202
) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Creating directory '/home/student-00-c545d5d251b0'.
student-00-c545d5d251b0@cluster33-m:~$ pwd
/home/student-00-c545d5d251b0
student-00-c545d5d251b0@cluster33-m:~$ hdfs dfs -mkdir /input_dir
student-00-c545d5d251b0@cluster33-m:~$ hdfs dfs -put /home/student-00-c545d5d251b0/employee.txt /input_dir
student-00-c545d5d251b0@cluster33-m:~$ hdfs dfs -put /home/student-00-c545d5d251b0/expenses.txt /input_dir
student-00-c545d5d251b0@cluster33-m:~$ hdfs dfs -ls /input_dir
Found 2 items
-rw-r--r--   2 student-00-c545d5d251b0 hadoop        252 2020-10-12 15:02 /input_dir/employee.txt
-rw-r--r--   2 student-00-c545d5d251b0 hadoop         79 2020-10-12 15:02 /input_dir/expenses.txt
student-00-c545d5d251b0@cluster33-m:~$ pig
20/10/12 15:02:51 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
20/10/12 15:02:51 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
20/10/12 15:02:51 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2020-10-12 15:02:51,144 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r: unknown) compiled Aug 15 2020, 11:10:54
2020-10-12 15:02:51,144 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/student-00-c545d5d251b0/pig_1602514971142.log
2020-10-12 15:02:51,172 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/student-00-c545d5d251b0/.pigbootup not found
2020-10-12 15:02:51,457 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-10-12 15:02:51,457 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://cluster33-m
2020-10-12 15:02:52,157 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-82015cad-4482-4bc3-8d4c-05de6056fdb7
2020-10-12 15:02:52,186 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: null
2020-10-12 15:02:52,477 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2020-10-12 15:02:52,503 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
grunt> emp = LOAD
>> '/input_dir/employee.txt' USING PigStorage(',') AS (id:int, name:chararray, salary:int, rating:
int);
2020-10-12 15:06:51,770 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
grunt> emp_desc = ORDER emp BY rating DESC;
grunt> emp_top5 = LIMIT emp_desc 5;
grunt> emp_name = FOREACH top5_employee GENERATE id, name;
2020-10-12 15:09:44,084 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script fai
led to parse: <line 5, column 19> Undefined alias: top5_employee
Details at logfile: /home/student-00-c545d5d251b0/pig_1602514971142.log
grunt> emp_name = FOREACH emp_top5 GENERATE id, name;
grunt> DUMP emp_name;
2020-10-12 15:10:38,275 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used
```


student-00-c545d5d251b0@cluster33-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-00-c34e19662090/zones/us-central1-f/instances/cluster33-m?authuser=1&hl=en_US&projectNumber=598187461987&nonAdminProxySessionReason=1

Please consider adding the IAP-secured Tunnel User IAM role to start using Cloud IAP for TCP forwarding for better performance. Learn more Dismiss

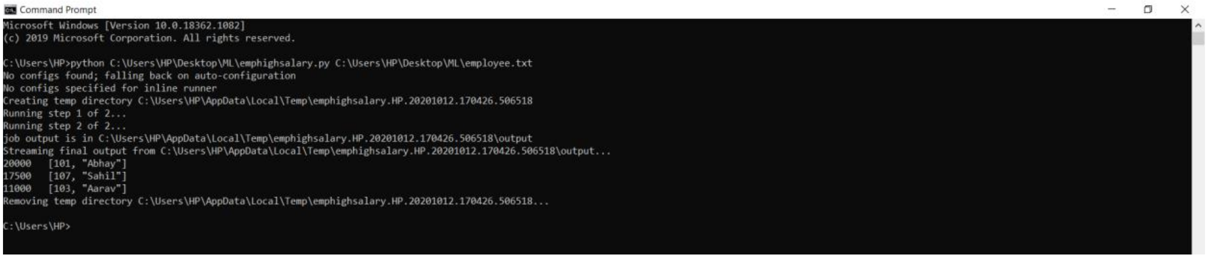
```
2020-10-12 15:11:59,458 [main] INFO  org.apache.hadoop.yarn.client.RMProx                                                .128.0.4:8032
2020-10-12 15:11:59,458 [main] INFO  org.apache.hadoop.yarn.client.AHSPro                                                uster33-m/10.128.0.4:10200
2020-10-12 15:11:59,461 [main] INFO  org.apache.hadoop.mapred.ClientServi                                                pplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,478 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,479 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,481 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,498 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,498 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,515 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,540 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,541 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,546 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,573 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,573 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,576 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,595 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,595 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,598 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,615 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,616 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,619 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,647 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,648 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,650 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,667 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,667 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,670 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,692 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,693 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,696 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,713 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,713 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,716 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,742 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:11:59,743 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:11:59,750 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,774 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-10-12 15:11:59,779 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:11:59,779 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 15:11:59,784 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 15:11:59,785 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priya)
(105,Palash)
(109,Karan)
(104,Anubhav)
(108,Ram)
grunt>
```

Q2.



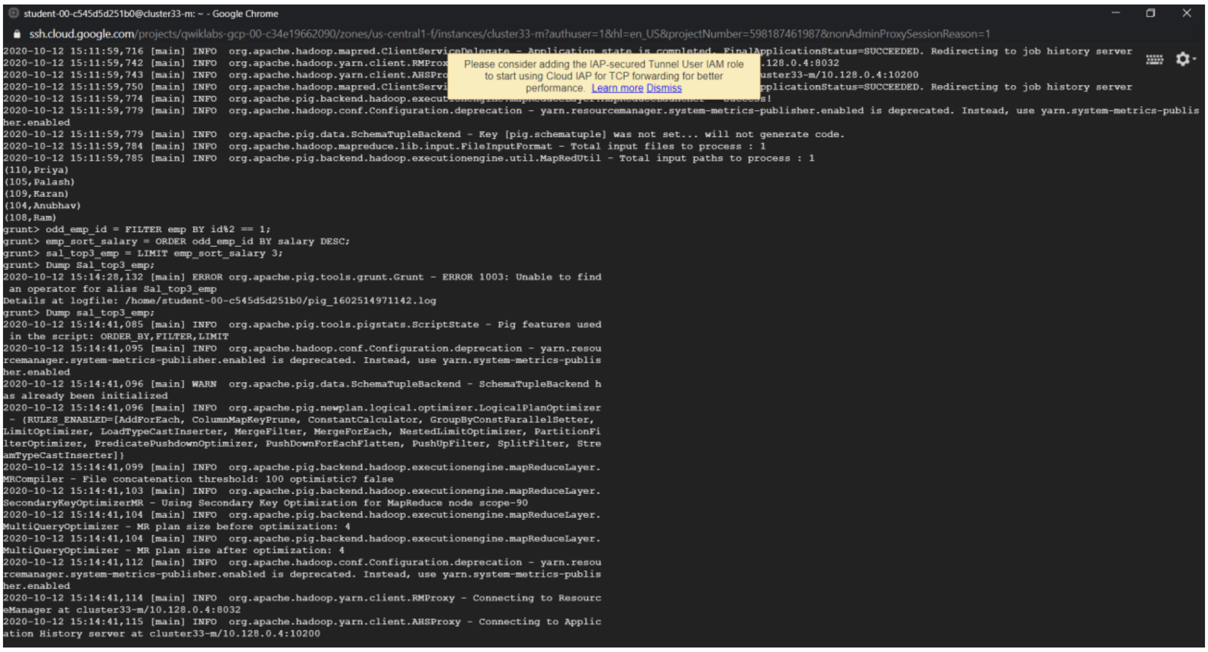
```
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

WORD_REGEXP = re.compile(r"[\w*]+")
class MRHighestSalary(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper,
                   reducer=self.reducer),
            MRStep(reducer = self.reduce1)
        ]


    def mapper(self,key, line):
        (id, name, salary, rating) = line.split(',')
        yield [int(id), name], int(salary)


    def reducer(self, id, salary):
        yield None, (max(salary), id)


    def reduce1(self, _, pair):
        for k, v in sorted(pair, reverse=True)[:5]:
            if(int(v[0])%2 != 0):
                yield k, v



if __name__ == '__main__':
    MRHighestSalary.run()
```

o/p:



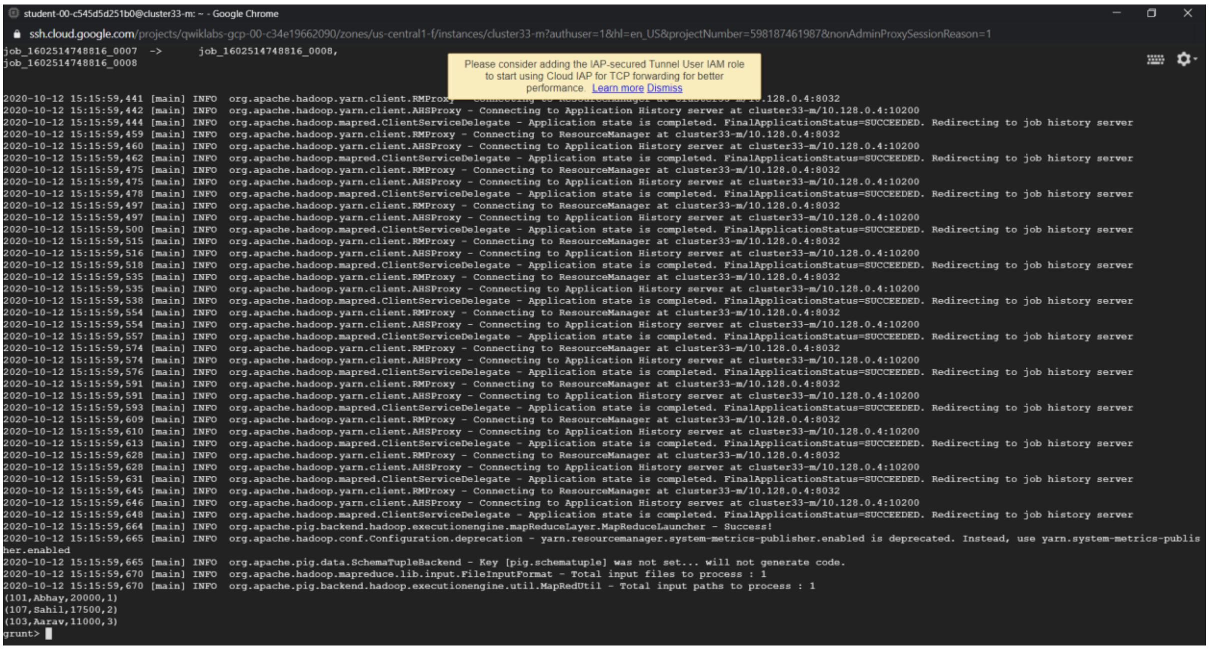
```
Microsoft Windows [Version 10.0.18362.1082]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\HP>python C:\Users\HP\Desktop\ML\emphighsalary.py C:\Users\HP\Desktop\ML\employee.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\HP\AppData\Local\Temp\emphighsalary.HP.20201012.170426.506518
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\HP\AppData\Local\Temp\emphighsalary.HP.20201012.170426.506518\output
Streaming final output from C:\Users\HP\AppData\Local\Temp\emphighsalary.HP.20201012.170426.506518\output...
20000   [101, "Abhay"]
17500   [107, "Sahil"]
11000   [103, "Aarav"]
Removing temp directory C:\Users\HP\AppData\Local\Temp\emphighsalary.HP.20201012.170426.506518...

C:\Users\HP>
```

## Using Pig:


student-00-c545d5d251b0@cluster33-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-00-c34e19662090/zones/us-central1-f/instances/cluster33-m?authuser=1&hl=en_US&projectNumber=598187461987&nonAdminProxySessionReason=1

Please consider adding the IAP-secured Tunnel User IAM role to start using Cloud IAP for TCP forwarding for better performance. Learn more Dismiss

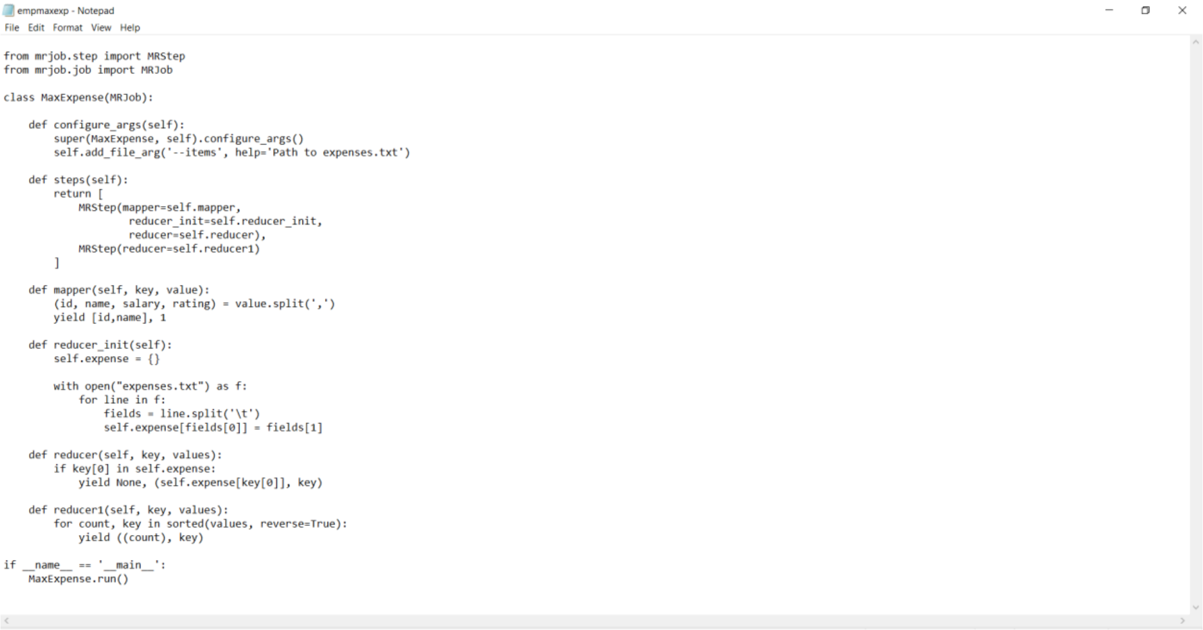
```
2020-10-12 15:11:59,716 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,742 [main] INFO  org.apache.hadoop.yarn.client.RMProx                                      .128.0.4:8032
2020-10-12 15:11:59,743 [main] INFO  org.apache.hadoop.yarn.client.AHSPro                                      uster33-m/10.128.0.4:10200
2020-10-12 15:11:59,750 [main] INFO  org.apache.hadoop.mapred.ClientServi                                      pplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:11:59,774 [main] INFO  org.apache.pig.backend.hadoop.executi                                     ss!
2020-10-12 15:11:59,779 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:11:59,779 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 15:11:59,784 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 15:11:59,785 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(110,Priya)
(105,Palash)
(109,Karan)
(104,Anubhav)
(108,Ram)
grunt> odd_emp_id = FILTER emp BY id%2 == 1;
grunt> emp_sort_salary = ORDER odd_emp_id BY salary DESC;
grunt> sal_top3_emp = LIMIT emp_sort_salary 3;
grunt> Dump Sal_top3_emp;
2020-10-12 15:14:28,132 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1003: Unable to find
 an operator for alias Sal_top3_emp
Details at logfile: /home/student-00-c545d5d251b0/pig_1602514971142.log
grunt> Dump sal_top3_emp;
2020-10-12 15:14:41,085 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used
 in the script: ORDER_BY,FILTER,LIMIT
2020-10-12 15:14:41,095 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:14:41,096 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend h
as already been initialized
2020-10-12 15:14:41,096 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer
 - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter,
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFi
lterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, Stre
amTypeCastInserter]}
2020-10-12 15:14:41,099 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MRCompiler - File concatenation threshold: 100 optimistic? false
2020-10-12 15:14:41,103 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-90
2020-10-12 15:14:41,104 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MultiQueryOptimizer - MR plan size before optimization: 4
2020-10-12 15:14:41,104 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MultiQueryOptimizer - MR plan size after optimization: 4
2020-10-12 15:14:41,112 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:14:41,114 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to Resourc
eManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:14:41,115 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Applic
ation History server at cluster33-m/10.128.0.4:10200
```


student-00-c545d5d251b0@cluster33-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-00-c34e19662090/zones/us-central1-f/instances/cluster33-m?authuser=1&hl=en_US&projectNumber=598187461987&nonAdminProxySessionReason=1

Please consider adding the IAP-secured Tunnel User IAM role to start using Cloud IAP for TCP forwarding for better performance. Learn more Dismiss

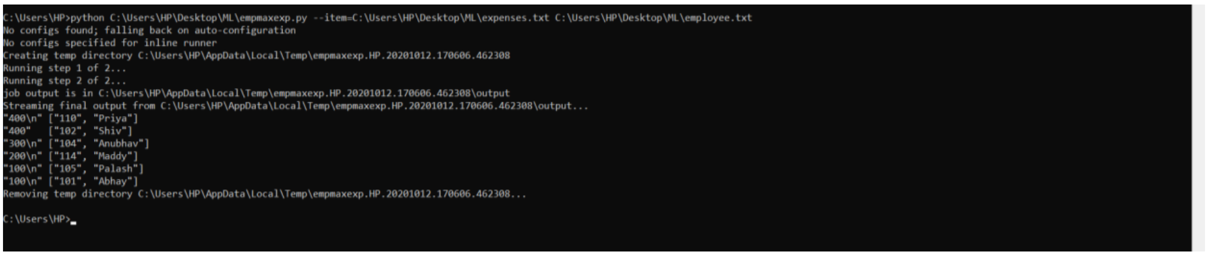
```
job_1602514748816_0007  ->      job_1602514748816_0008,
job_1602514748816_0008


2020-10-12 15:15:59,441 [main] INFO  org.apache.hadoop.yarn.client.RMProx                                      .128.0.4:8032
2020-10-12 15:15:59,442 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,444 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,459 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,460 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,462 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,475 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,475 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,478 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,497 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,497 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,500 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,515 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,516 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,518 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,535 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,535 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,538 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,554 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,554 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,557 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,574 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,574 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,576 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,591 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,591 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,593 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,609 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,610 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,613 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,628 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,628 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,631 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,645 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster33-m/10.128.0.4:8032
2020-10-12 15:15:59,646 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster33-m/10.128.0.4:10200
2020-10-12 15:15:59,648 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:15:59,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-10-12 15:15:59,665 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:15:59,665 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 15:15:59,670 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 15:15:59,670 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Abhay,20000,1)
(107,Sahil,17500,2)
(103,Aarav,11000,3)
grunt> █
```

Q3:



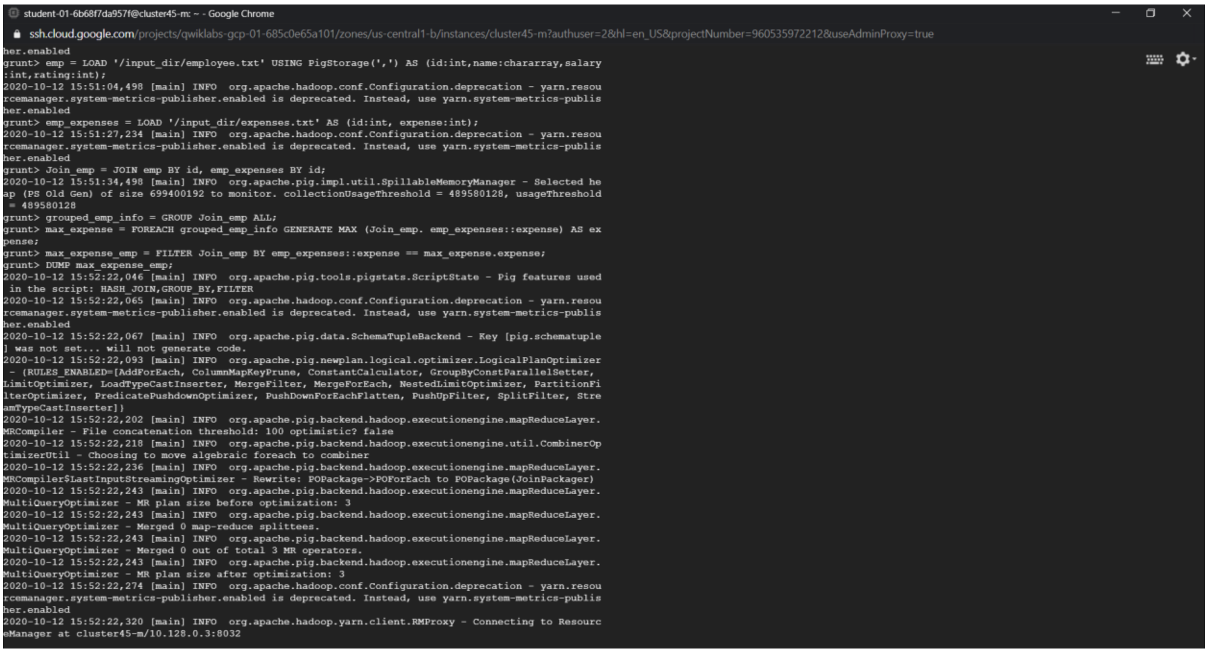
```
from mrjob.step import MRStep
from mrjob.job import MRJob

class MaxExpense(MRJob):

    def configure_args(self):
        super(MaxExpense, self).configure_args()
        self.add_file_arg('--items', help='Path to expenses.txt')

    def steps(self):
        return [
            MRStep(mapper=self.mapper,
                    reducer_init=self.reducer_init,
                    reducer=self.reducer),
            MRStep(reducer=self.reducer1)
        ]

    def mapper(self, key, value):
        (id, name, salary, rating) = value.split(',')
        yield [id,name], 1

    def reducer_init(self):
        self.expense = {}

        with open("expenses.txt") as f:
            for line in f:
                fields = line.split('\t')
                self.expense[fields[0]] = fields[1]

    def reducer(self, key, values):
        if key[0] in self.expense:
            yield None, (self.expense[key[0]], key)

    def reducer1(self, key, values):
        for count, key in sorted(values, reverse=True):
            yield ((count), key)

if __name__ == '__main__':
    MaxExpense.run()
```

o/p:

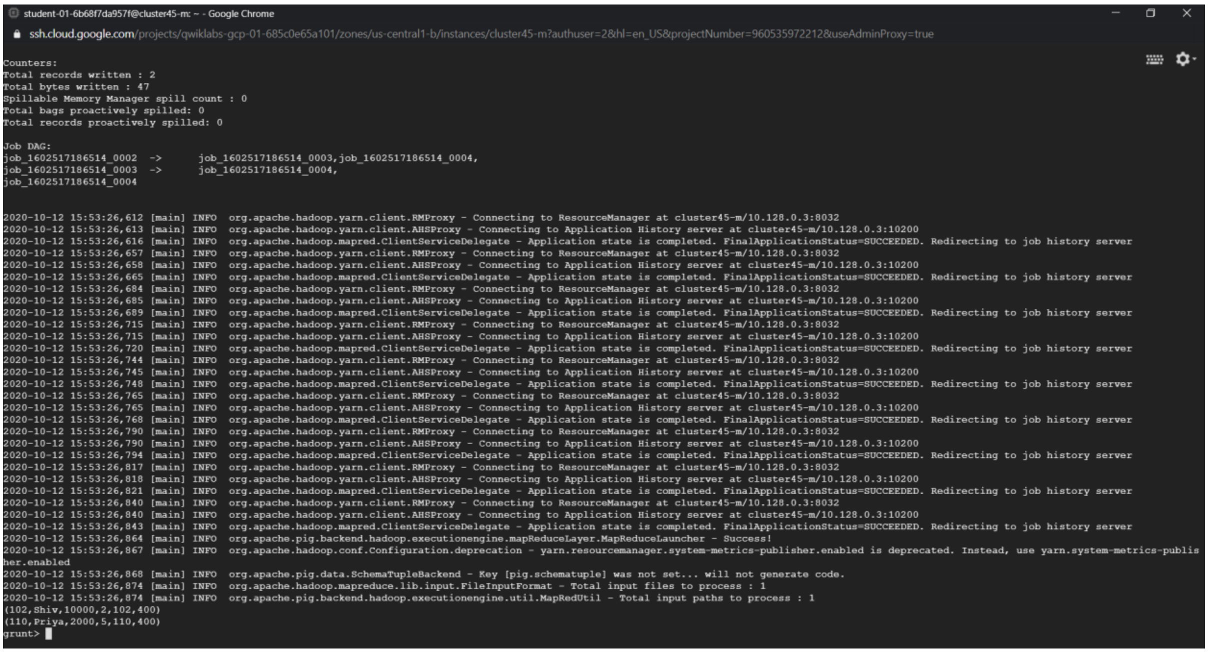


```
C:\Users\HP>python C:\Users\HP\Desktop\ML\empmaxexp.py --item=C:\Users\HP\Desktop\ML\expenses.txt C:\Users\HP\Desktop\ML\employee.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\HP\AppData\Local\Temp\empmaxexp.HP.20201012.170606.462308
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\HP\AppData\Local\Temp\empmaxexp.HP.20201012.170606.462308\output
Streaming final output from C:\Users\HP\AppData\Local\Temp\empmaxexp.HP.20201012.170606.462308\output...
"400\n" ["110", "Priya"]
"400"   ["102", "Shiv"]
"300\n" ["104", "Anubhav"]
"200\n" ["114", "Maddy"]
"100\n" ["105", "Palash"]
"100\n" ["101", "Abhay"]
Removing temp directory C:\Users\HP\AppData\Local\Temp\empmaxexp.HP.20201012.170606.462308...

C:\Users\HP>
```

## Using Pig:

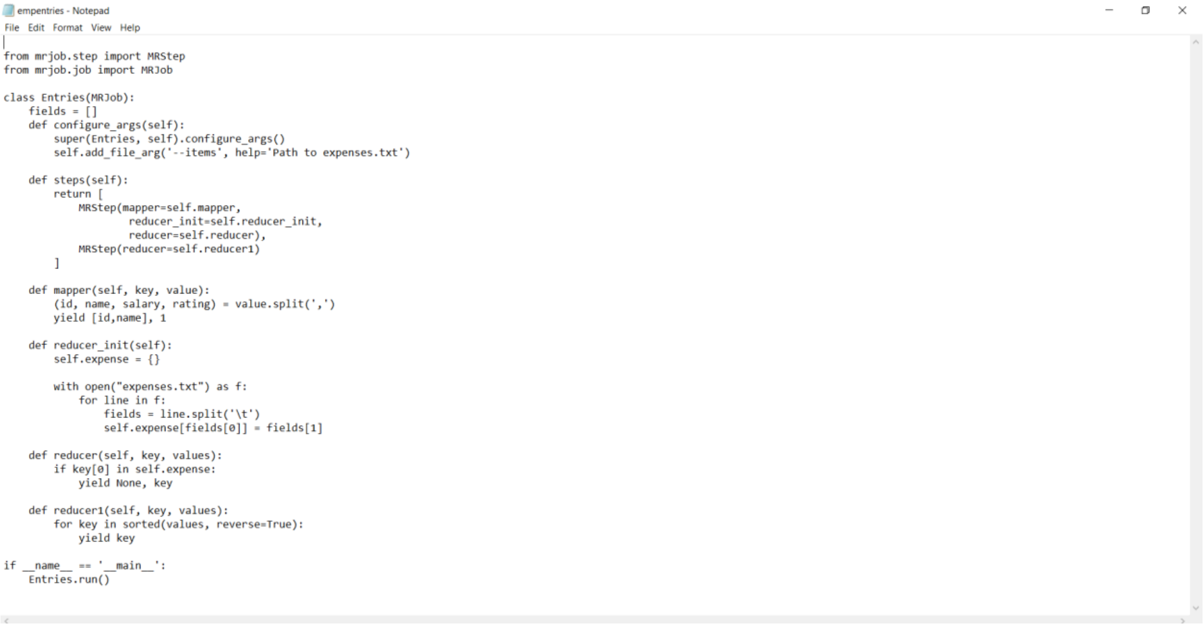

student-01-6b68f7da957f@cluster45-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-01-685c0e65a101/zones/us-central1-b/instances/cluster45-m?authuser=2&hl=en_US&projectNumber=960535972212&useAdminProxy=true

```
her.enabled
grunt> emp = LOAD '/input_dir/employee.txt' USING PigStorage(',') AS (id:int,name:chararray,salary
:int,rating:int);
2020-10-12 15:51:04,498 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
grunt> emp_expenses = LOAD '/input_dir/expenses.txt' AS (id:int, expense:int);
2020-10-12 15:51:27,234 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
grunt> Join_emp = JOIN emp BY id, emp_expenses BY id;
2020-10-12 15:51:34,498 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected he
ap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold
 = 489580128
grunt> grouped_emp_info = GROUP Join_emp ALL;
grunt> max_expense = FOREACH grouped_emp_info GENERATE MAX (Join_emp. emp_expenses::expense) AS ex
pense;
grunt> max_expense_emp = FILTER Join_emp BY emp_expenses::expense == max_expense.expense;
grunt> DUMP max_expense_emp;
2020-10-12 15:52:22,046 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used
 in the script: HASH_JOIN,GROUP_BY,FILTER
2020-10-12 15:52:22,065 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:52:22,067 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple
] was not set... will not generate code.
2020-10-12 15:52:22,093 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer
 - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter,
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFi
lterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, Stre
amTypeCastInserter]}
2020-10-12 15:52:22,202 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MRCompiler - File concatenation threshold: 100 optimistic? false
2020-10-12 15:52:22,218 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.CombinerOp
timizerUtil - Choosing to move algebraic foreach to combiner
2020-10-12 15:52:22,236 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2020-10-12 15:52:22,243 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MultiQueryOptimizer - MR plan size before optimization: 3
2020-10-12 15:52:22,243 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MultiQueryOptimizer - Merged 0 map-reduce splittees.
2020-10-12 15:52:22,243 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MultiQueryOptimizer - Merged 0 out of total 3 MR operators.
2020-10-12 15:52:22,243 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MultiQueryOptimizer - MR plan size after optimization: 3
2020-10-12 15:52:22,274 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:52:22,320 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to Resourc
eManager at cluster45-m/10.128.0.3:8032
```


student-01-6b68f7da957f@cluster45-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-01-685c0e65a101/zones/us-central1-b/instances/cluster45-m?authuser=2&hl=en_US&projectNumber=960535972212&useAdminProxy=true

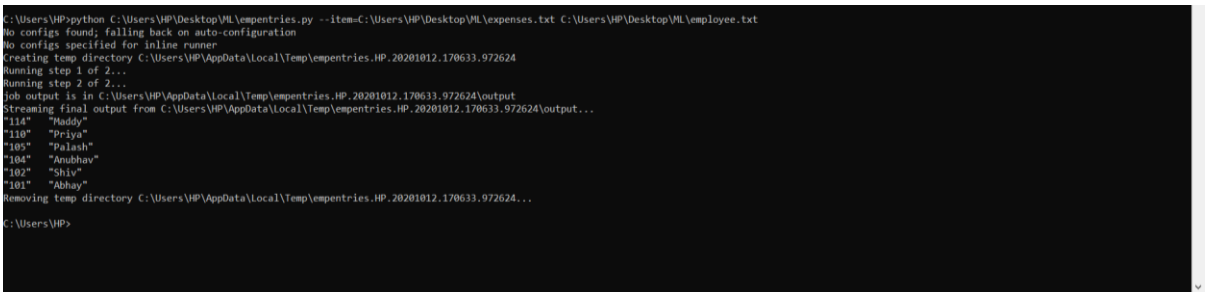
```
Counters:
Total records written : 2
Total bytes written : 47
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1602517186514_0002  ->      job_1602517186514_0003,job_1602517186514_0004,
job_1602517186514_0003  ->      job_1602517186514_0004,
job_1602517186514_0004


2020-10-12 15:53:26,612 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,613 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,616 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,657 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,658 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,665 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,684 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,685 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,689 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,715 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,715 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,720 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,744 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,745 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,748 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,765 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,765 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,768 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,790 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,790 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,794 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,817 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,818 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,821 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,840 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,840 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,843 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,864 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-10-12 15:53:26,867 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:53:26,868 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 15:53:26,874 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 15:53:26,874 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(102,Shiv,10000,2,102,400)
(110,Priya,2000,5,110,400)
grunt> 
```

Q4:



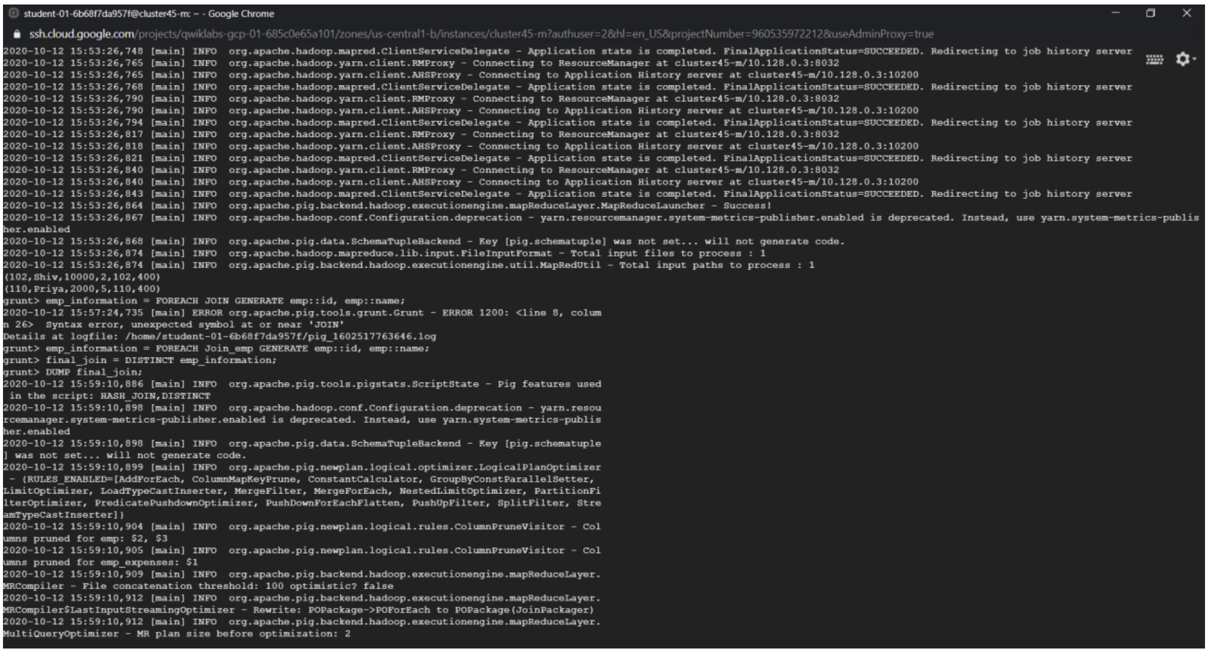
```
from mrjob.step import MRStep
from mrjob.job import MRJob

class Entries(MRJob):
    fields = []
    def configure_args(self):
        super(Entries, self).configure_args()
        self.add_file_arg('--items', help='Path to expenses.txt')

    def steps(self):
        return [
            MRStep(mapper=self.mapper,
                    reducer_init=self.reducer_init,
                    reducer=self.reducer),
            MRStep(reducer=self.reducer1)
        ]

    def mapper(self, key, value):
        (id, name, salary, rating) = value.split(',')
        yield [id,name], 1

    def reducer_init(self):
        self.expense = {}

        with open("expenses.txt") as f:
            for line in f:
                fields = line.split('\t')
                self.expense[fields[0]] = fields[1]

    def reducer(self, key, values):
        if key[0] in self.expense:
            yield None, key

    def reducer1(self, key, values):
        for key in sorted(values, reverse=True):
            yield key

if __name__ == '__main__':
    Entries.run()
```

o/p:

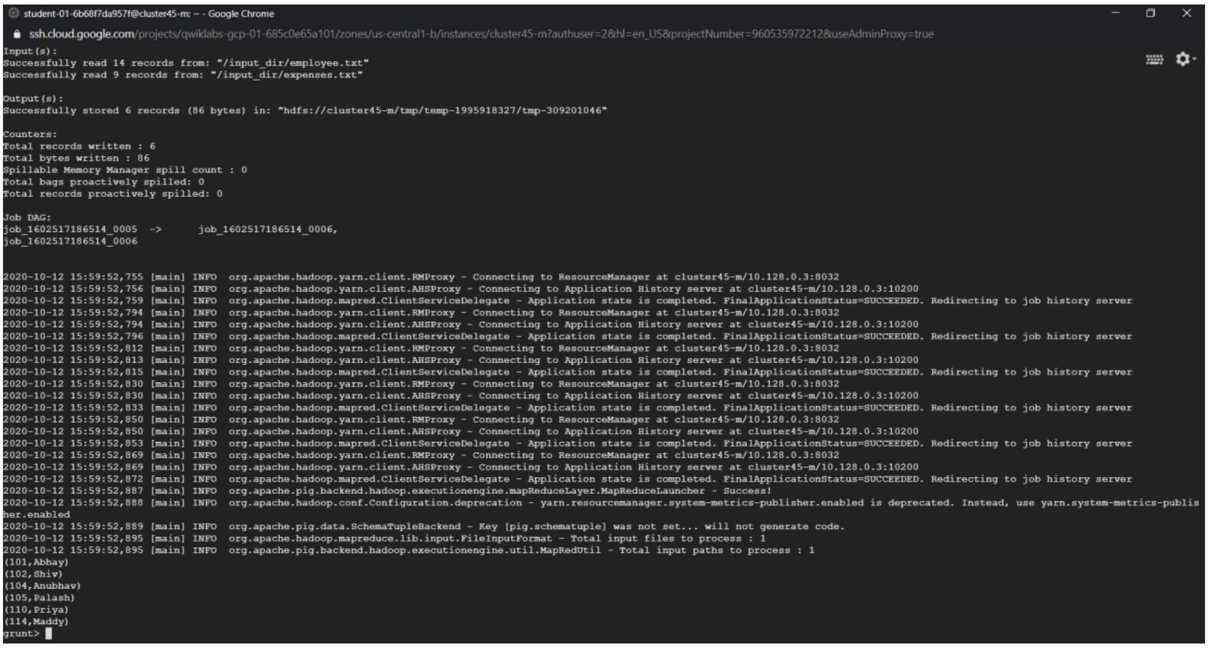


```
C:\Users\HP>python C:\Users\HP\Desktop\ML\empentries.py --item=C:\Users\HP\Desktop\ML\expenses.txt C:\Users\HP\Desktop\ML\employee.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\HP\AppData\Local\Temp\empentries.HP.20201012.170633.972624
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\HP\AppData\Local\Temp\empentries.HP.20201012.170633.972624\output
Streaming final output from C:\Users\HP\AppData\Local\Temp\empentries.HP.20201012.170633.972624\output...
"114"   "Maddy"
"110"   "Priya"
"105"   "Palash"
"104"   "Anubhav"
"102"   "Shiv"
"101"   "Abhay"
Removing temp directory C:\Users\HP\AppData\Local\Temp\empentries.HP.20201012.170633.972624...

C:\Users\HP>
```

Using Pig:

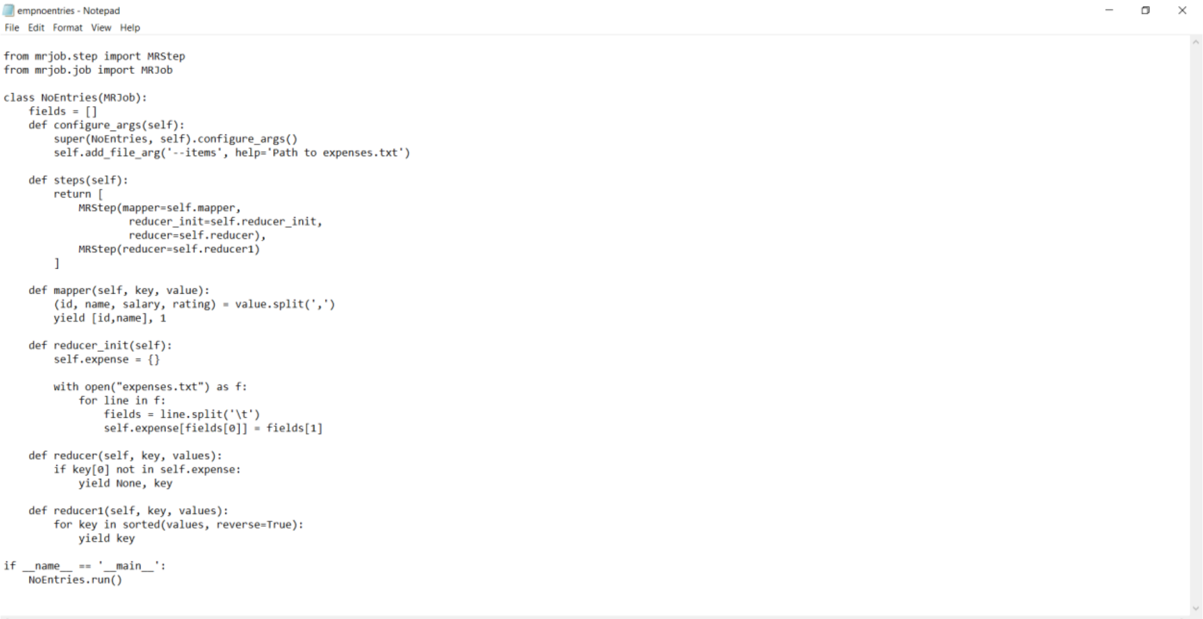

student-01-6b68f7da957f@cluster45-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-01-685c0e65a101/zones/us-central1-b/instances/cluster45-m?authuser=2&hl=en_US&projectNumber=960535972212&useAdminProxy=true

```
2020-10-12 15:53:26,748 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,765 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,765 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,768 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,790 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,790 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,794 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,817 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,818 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,821 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,840 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:53:26,840 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:53:26,843 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:53:26,864 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-10-12 15:53:26,867 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:53:26,868 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 15:53:26,874 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 15:53:26,874 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(102,Shiv,10000,2,102,400)
(110,Priya,2000,5,110,400)
grunt> emp_information = FOREACH JOIN GENERATE emp::id, emp::name;
2020-10-12 15:57:24,735 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 8, colum
n 26>  Syntax error, unexpected symbol at or near 'JOIN'
Details at logfile: /home/student-01-6b68f7da957f/pig_1602517763646.log
grunt> emp_information = FOREACH Join_emp GENERATE emp::id, emp::name;
grunt> final_join = DISTINCT emp_information;
grunt> DUMP final_join;
2020-10-12 15:59:10,886 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used
 in the script: HASH_JOIN,DISTINCT
2020-10-12 15:59:10,898 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:59:10,898 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple
] was not set... will not generate code.
2020-10-12 15:59:10,899 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer
 - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter,
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFi
lterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, Stre
amTypeCastInserter]}
2020-10-12 15:59:10,904 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Col
umns pruned for emp: $2, $3
2020-10-12 15:59:10,905 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Col
umns pruned for emp_expenses: $1
2020-10-12 15:59:10,909 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MRCompiler - File concatenation threshold: 100 optimistic? false
2020-10-12 15:59:10,912 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MRCompiler$LastInputStreamingOptimizer - Rewrite: POPackage->POForEach to POPackage(JoinPackager)
2020-10-12 15:59:10,912 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.
MultiQueryOptimizer - MR plan size before optimization: 2
```


student-01-6b68f7da957f@cluster45-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-01-685c0e65a101/zones/us-central1-b/instances/cluster45-m?authuser=2&hl=en_US&projectNumber=960535972212&useAdminProxy=true

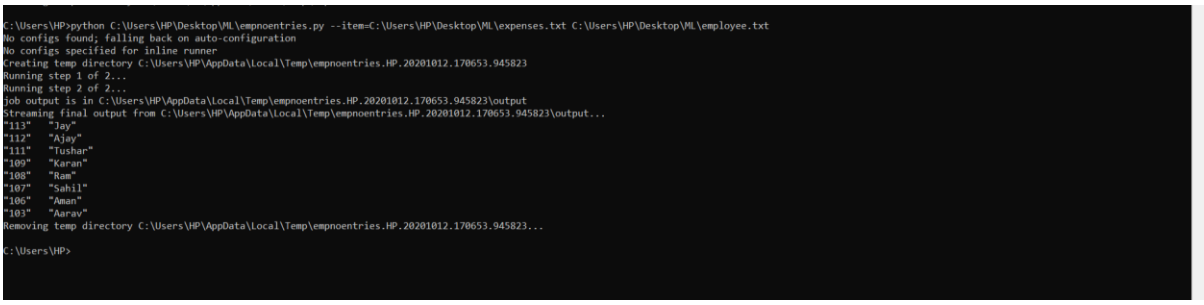
```
Input(s):
Successfully read 14 records from: "/input_dir/employee.txt"
Successfully read 9 records from: "/input_dir/expenses.txt"

Output(s):
Successfully stored 6 records (86 bytes) in: "hdfs://cluster45-m/tmp/temp-1995918327/tmp-309201046"

Counters:
Total records written : 6
Total bytes written : 86
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1602517186514_0005  ->      job_1602517186514_0006,
job_1602517186514_0006


2020-10-12 15:59:52,755 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,756 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,759 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,794 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,794 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,796 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,812 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,813 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,815 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,830 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,830 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,833 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,850 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,850 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,853 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,869 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,869 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,872 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,887 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-10-12 15:59:52,888 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:59:52,889 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 15:59:52,895 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 15:59:52,895 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Abhay)
(102,Shiv)
(104,Anubhav)
(105,Palash)
(110,Priya)
(114,Maddy)
grunt> 
```

## Q5:



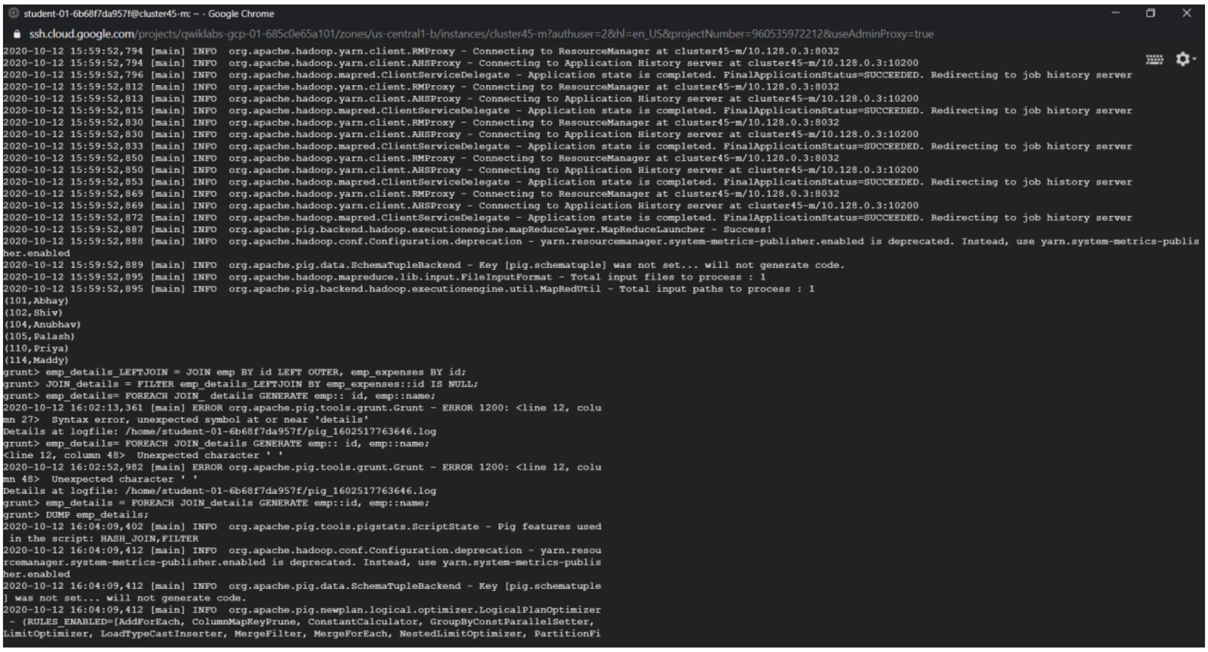
```
from mrjob.step import MRStep
from mrjob.job import MRJob

class NoEntries(MRJob):
    fields = []
    def configure_args(self):
        super(NoEntries, self).configure_args()
        self.add_file_arg('--items', help='Path to expenses.txt')

    def steps(self):
        return [
            MRStep(mapper=self.mapper,
                    reducer_init=self.reducer_init,
                    reducer=self.reducer),
            MRStep(reducer=self.reducer1)
        ]

    def mapper(self, key, value):
        (id, name, salary, rating) = value.split(',')
        yield [id,name], 1

    def reducer_init(self):
        self.expense = {}

        with open("expenses.txt") as f:
            for line in f:
                fields = line.split('\t')
                self.expense[fields[0]] = fields[1]

    def reducer(self, key, values):
        if key[0] not in self.expense:
            yield None, key

    def reducer1(self, key, values):
        for key in sorted(values, reverse=True):
            yield key

if __name__ == '__main__':
    NoEntries.run()
```

## o/p:

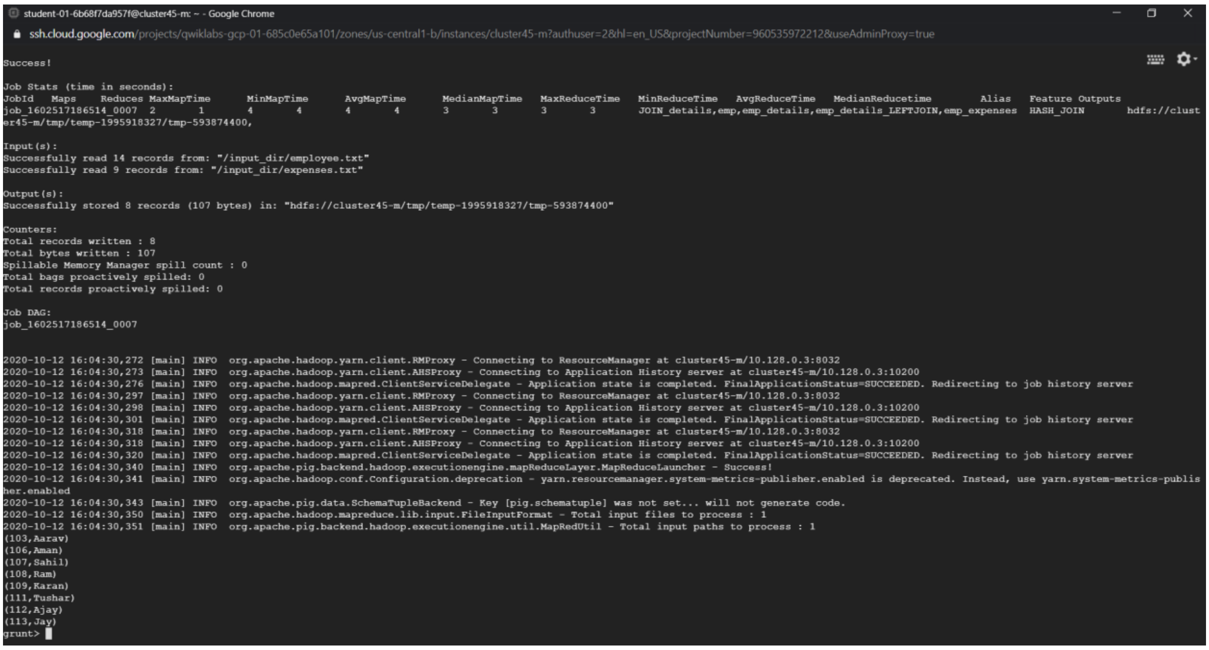


```
C:\Users\HP>python C:\Users\HP\Desktop\ML\empnoentries.py --item=C:\Users\HP\Desktop\ML\expenses.txt C:\Users\HP\Desktop\ML\employee.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\HP\AppData\Local\Temp\empnoentries.HP.20201012.170653.945823
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\HP\AppData\Local\Temp\empnoentries.HP.20201012.170653.945823\output
Streaming final output from C:\Users\HP\AppData\Local\Temp\empnoentries.HP.20201012.170653.945823\output...
"113"   "Jay"
"112"   "Ajay"
"111"   "Tushar"
"109"   "Karan"
"108"   "Ram"
"107"   "Sahil"
"106"   "Aman"
"103"   "Aarav"
Removing temp directory C:\Users\HP\AppData\Local\Temp\empnoentries.HP.20201012.170653.945823...

C:\Users\HP>
```

Using Pig:


student-01-6b68f7da957f@cluster45-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-01-685c0e65a101/zones/us-central1-b/instances/cluster45-m?authuser=2&hl=en_US&projectNumber=960535972212&useAdminProxy=true

```
2020-10-12 15:59:52,794 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,794 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,796 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,812 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,813 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,815 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,830 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,830 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,833 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,850 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,850 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,853 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,869 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 15:59:52,869 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 15:59:52,872 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 15:59:52,887 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-10-12 15:59:52,888 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 15:59:52,889 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 15:59:52,895 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 15:59:52,895 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Abhay)
(102,Shiv)
(104,Anubhav)
(105,Palash)
(110,Priya)
(114,Maddy)
grunt> emp_details_LEFTJOIN = JOIN emp BY id LEFT OUTER, emp_expenses BY id;
grunt> JOIN_details = FILTER emp_details_LEFTJOIN BY emp_expenses::id IS NULL;
grunt> emp_details= FOREACH JOIN_ details GENERATE emp:: id, emp::name;
2020-10-12 16:02:13,361 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 12, colu
mn 27>  Syntax error, unexpected symbol at or near 'details'
Details at logfile: /home/student-01-6b68f7da957f/pig_1602517763646.log
grunt> emp_details= FOREACH JOIN_details GENERATE emp:: id, emp::name;
<line 12, column 48>  Unexpected character ' '
2020-10-12 16:02:52,982 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 12, colu
mn 48>  Unexpected character ' '
Details at logfile: /home/student-01-6b68f7da957f/pig_1602517763646.log
grunt> emp_details = FOREACH JOIN_details GENERATE emp::id, emp::name;
grunt> DUMP emp_details;
2020-10-12 16:04:09,402 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used
 in the script: HASH_JOIN,FILTER
2020-10-12 16:04:09,412 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resou
rcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 16:04:09,412 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple
] was not set... will not generate code.
2020-10-12 16:04:09,412 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer
 - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter,
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFi
```


student-01-6b68f7da957f@cluster45-m: ~ - Google Chrome
ssh.cloud.google.com/projects/qwiklabs-gcp-01-685c0e65a101/zones/us-central1-b/instances/cluster45-m?authuser=2&hl=en_US&projectNumber=960535972212&useAdminProxy=true

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1602517186514_0007  2       1       4       4       4       4       3       3       3       3       JOIN_details,emp,emp_details,emp_details_LEFTJOIN,emp_expenses  HASH_JOIN       hdfs://clust
er45-m/tmp/temp-1995918327/tmp-593874400,

Input(s):
Successfully read 14 records from: "/input_dir/employee.txt"
Successfully read 9 records from: "/input_dir/expenses.txt"

Output(s):
Successfully stored 8 records (107 bytes) in: "hdfs://cluster45-m/tmp/temp-1995918327/tmp-593874400"

Counters:
Total records written : 8
Total bytes written : 107
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1602517186514_0007


2020-10-12 16:04:30,272 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 16:04:30,273 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 16:04:30,276 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 16:04:30,297 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 16:04:30,298 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 16:04:30,301 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 16:04:30,318 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster45-m/10.128.0.3:8032
2020-10-12 16:04:30,318 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster45-m/10.128.0.3:10200
2020-10-12 16:04:30,320 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2020-10-12 16:04:30,340 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-10-12 16:04:30,341 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publis
her.enabled
2020-10-12 16:04:30,343 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-10-12 16:04:30,350 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-10-12 16:04:30,351 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(103,Aarav)
(106,Aman)
(107,Sahil)
(108,Ram)
(109,Karan)
(111,Tushar)
(112,Ajay)
(113,Jay)
grunt> 
```