جامعة جدة
University of Jeddah

PIG Latin!

Big Data

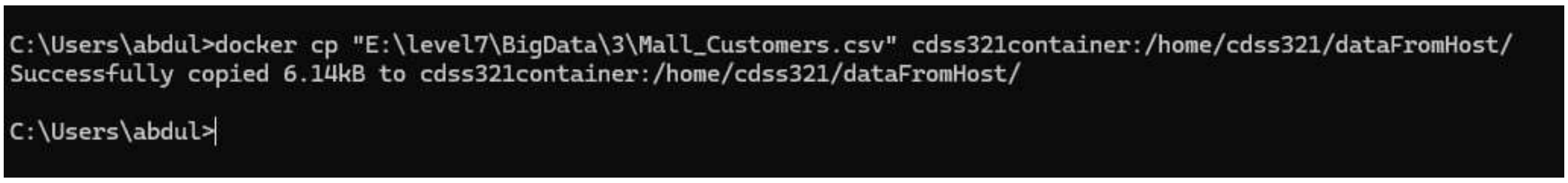
Amal Almutairi

Nouf Mansour

Noor Balfas

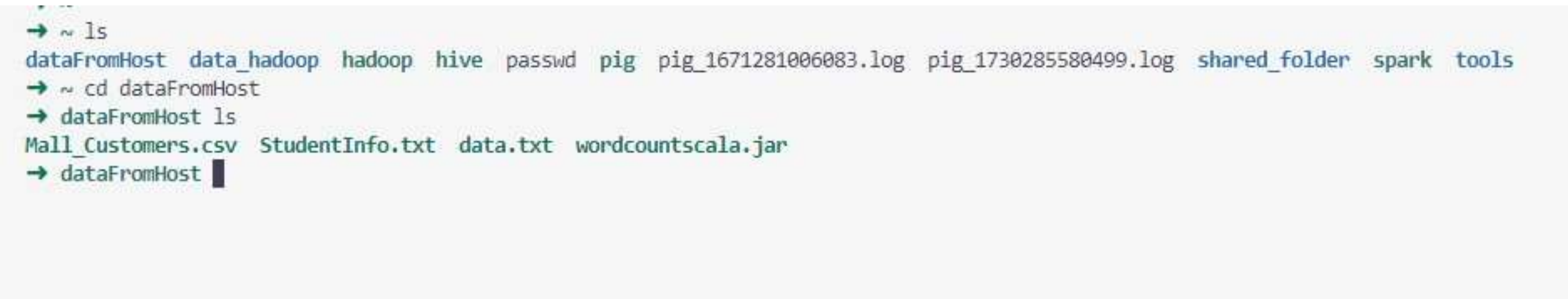
Muna Saleh Dini

- ## Uploading Files to Docker Container



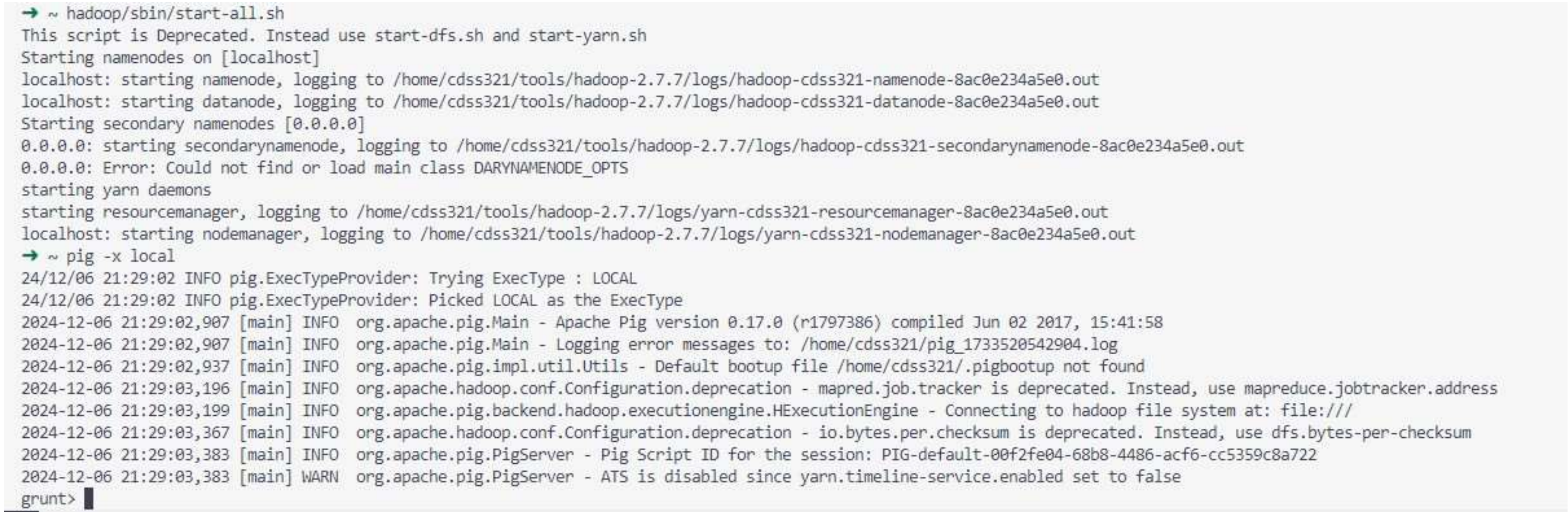
```
C:\Users\abdul>docker cp "E:\level7\BigData\3\Mall_Customers.csv" cdss321container:/home/cdss321/dataFromHost/
Successfully copied 6.14kB to cdss321container:/home/cdss321/dataFromHost/

C:\Users\abdul>
```



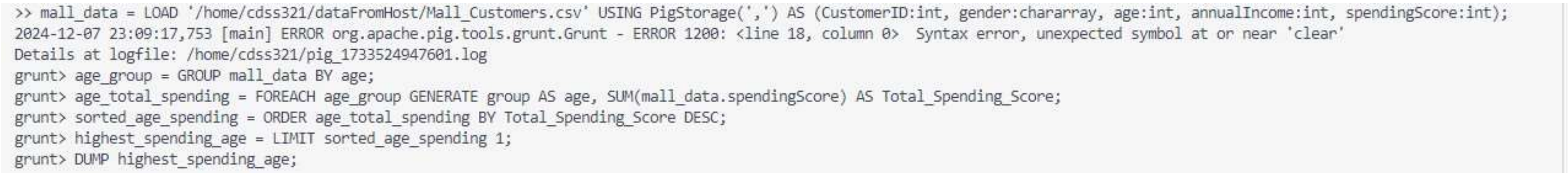
```
➜ ~ ls
dataFromHost  data_hadoop  hadoop  hive  passwd  pig  pig_1671281006083.log  pig_1730285580499.log  shared_folder  spark  tools
➜ ~ cd dataFromHost
➜ dataFromHost ls
Mall_Customers.csv  StudentInfo.txt  data.txt  wordcountscala.jar
➜ dataFromHost
```

- ## Starting Hadoop Cluster & PIG in cluster mode!



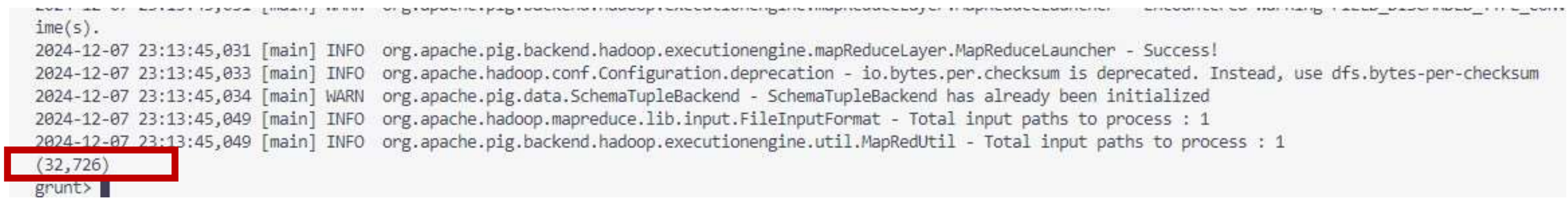
```
➜ ~ hadoop/sbin/start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/cdss321/tools/hadoop-2.7.7/logs/hadoop-cdss321-namenode-8ac0e234a5e0.out
localhost: starting datanode, logging to /home/cdss321/tools/hadoop-2.7.7/logs/hadoop-cdss321-datanode-8ac0e234a5e0.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/cdss321/tools/hadoop-2.7.7/logs/hadoop-cdss321-secondarynamenode-8ac0e234a5e0.out
0.0.0.0: Error: Could not find or load main class DARYNAMENODE_OPTS
starting yarn daemons
starting resourcemanager, logging to /home/cdss321/tools/hadoop-2.7.7/logs/yarn-cdss321-resourcemanager-8ac0e234a5e0.out
localhost: starting nodemanager, logging to /home/cdss321/tools/hadoop-2.7.7/logs/yarn-cdss321-nodemanager-8ac0e234a5e0.out
➜ ~ pig -x local
24/12/06 21:29:02 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
24/12/06 21:29:02 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2024-12-06 21:29:02,907 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-12-06 21:29:02,907 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/cdss321/pig_1733520542904.log
2024-12-06 21:29:02,937 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/cdss321/.pigbootup not found
2024-12-06 21:29:03,196 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-12-06 21:29:03,199 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2024-12-06 21:29:03,367 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-12-06 21:29:03,383 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-00f2fe04-68b8-4486-acf6-cc5359c8a722
2024-12-06 21:29:03,383 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

1. Write a PIG script that reports the age of customers who have collectively the highest total Spending_Score:



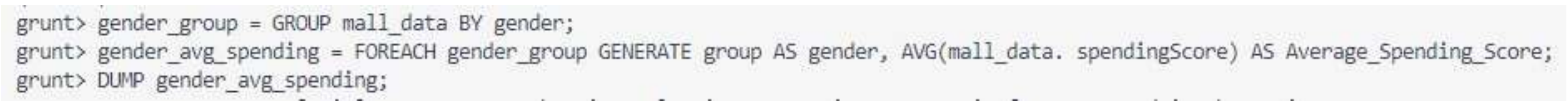
```
>> mall_data = LOAD '/home/cdss321/dataFromHost/Mall_Customers.csv' USING PigStorage(',') AS (CustomerID:int, gender:chararray, age:int, annualIncome:int, spendingScore:int);
2024-12-07 23:09:17,753 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 18, column 0>  Syntax error, unexpected symbol at or near 'clear'
Details at logfile: /home/cdss321/pig_1733524947601.log
grunt> age_group = GROUP mall_data BY age;
grunt> age_total_spending = FOREACH age_group GENERATE group AS age, SUM(mall_data.spendingScore) AS Total_Spending_Score;
grunt> sorted_age_spending = ORDER age_total_spending BY Total_Spending_Score DESC;
grunt> highest_spending_age = LIMIT sorted_age_spending 1;
grunt> DUMP highest_spending_age;
```

Output

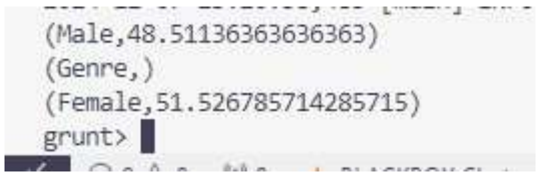


```
ime(s).
2024-12-07 23:13:45,031 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-12-07 23:13:45,033 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2024-12-07 23:13:45,034 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-12-07 23:13:45,049 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2024-12-07 23:13:45,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(32,726)
grunt>
```

2. Write a PIG script that reports the average Spending_Score for men and for women.

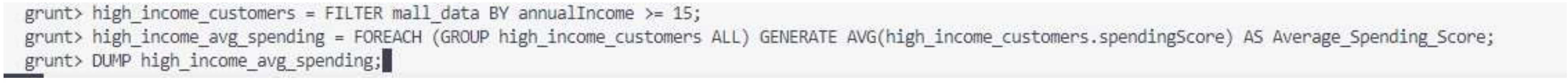


```
grunt> gender_group = GROUP mall_data BY gender;
grunt> gender_avg_spending = FOREACH gender_group GENERATE group AS gender, AVG(mall_data. spendingScore) AS Average_Spending_Score;
grunt> DUMP gender_avg_spending;
```

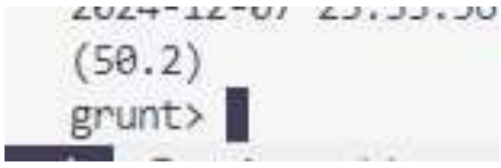


```
(Male,48.51136363636363)
(Genre,)
(Female,51.526785714285715)
grunt>
```

3. Write a PIG script that reports the average Spending_Score of customers whom annual salary is at least 15,000$.



```
grunt> high_income_customers = FILTER mall_data BY annualIncome >= 15;
grunt> high_income_avg_spending = FOREACH (GROUP high_income_customers ALL) GENERATE AVG(high_income_customers.spendingScore) AS Average_Spending_Score;
grunt> DUMP high_income_avg_spending;
```



```
2024-12-07 23:33:38
(50.2)
grunt>
```