Execution of Pig Latin script using Hadoop cluster

Team 4:
Myat Oo,
Mohammad Islam
Zijian Zhang

Table of Contents

1) Introduction: 2

2) Implementation: 2

a) Database Connection: 2

b) Virtualization Machines: 3

c) Apache Hadoop: 3

d) Apache Pig: 4

e) CSV file: 5

3) Results: 5

1) Introduction:

In this project, we have analyzed a database using the csv files provided by the instructor.

We have written Pig Latin scripts on Hadoop cluster of size 2 and founded specific information that is related to several csv files. Our queries were able to connect with the csv files that were need to generate the result that have been asked.

There were many csv files and the questions were the following:
A. What is the average length of films in each category? List the results in alphabetic order of categories.
B. Which categories have the longest and shortest average film lengths?
C. Which customers have rented action but not comedy or classic movies?
D. Which actor has appeared in the most English-language movies?
E. How many distinct movies were rented for exactly 10 days from the store where Mike works?
F. Alphabetically list actors who appeared in the movie with the largest cast of actors.

2) Implementation:

a) Database Connection:

We have analyzed csv files and derived the results using the csv files on Hadoop on a cluster of 2 or 3 virtual machines. Initially, we have set up the virtual machine environment an embedded the apache Hadoop framework in it. We then wrote Pig Latin scripts for each question and saved them in our folder as pig files. Then, we have implemented all the necessary methods in it to get the relevant correlations among the csv files and the specific information asked in the questions. By using the Hadoop on the virtualization machines, we were able to extract the necessary results for the questions. This has been done through the assistance of Hadoop. The Hadoop played a role as the bridge between the Pig Latin Script that we wrote and the csv files that were provided to extract specific information from.

b) Virtualization Machines:

In computing, a virtual machine (VM) is an emulation of a computer system. Virtual machines are based on computer architectures and provide functionality of a physical computer. Their implementations may involve specialized hardware, software, or a combination.

There are different kinds of virtual machines, each with different functions:

- System virtual machines (also termed full virtualization VMs) provide a substitute for a real machine. They provide functionality needed to execute entire operating systems. A hypervisor uses native execution to share and manage hardware, allowing for multiple environments which are isolated from one another, yet exist on the same physical machine. Modern hypervisors use hardware-assisted virtualization, virtualization-specific hardware, primarily from the host CPUs.

- Process virtual machines are designed to execute computer programs in a platform-independent environment.

In our project, we have used virtual machines such as Happy and Grumpy to implement the Hadoop and to analyze and extract the results from the csv files using the Pig Latin language. By embedding the csv files on Hadoop and using the Pig Latin queries on the command line, we have pulled out many information such as

- The average length of films in each category.
- Categories have the longest and shortest average film lengths.
- Customers who have rented action but not comedy or classic movies.
- Actor who has appeared in the most English-language movies.
- The number of distinct movies that were rented for exactly 10 days from the store where Mike works.
- List actors in alphabetic who have appeared in the movie with the largest cast of actors.

c) Apache Hadoop:

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver highavailability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Hadoop is supported by GNU/Linux platform and its flavors. Therefore, we have to install a Linux operating system for setting up Hadoop environment. In case you have an OS other than Linux, you can install Virtual box software in it and have Linux inside the Virtual box.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality,[3] where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

In our project, we have embedded the Apache Hadoop Software in our virtual machines to run MapReduce and to get the information using the csv files. The MapReduce we implemented in Hadoop has allowed us to get the crucial information we were looking for using the Pig Latin queries we wrote. Through the Hadoop, we executed the MapReduce function in our program to gather all the results we were looking for.

d) Apache Pig:

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark. Pig Latin abstracts the programming from the Java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQL for relational database management systems. Pig Latin can be extended using user-defined functions (UDFs) which the user can write in Java, Python, JavaScript, Ruby or Groovy and then call directly from the language.

This Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Now, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:

Ease of programming. It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.

- Optimization opportunities. The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
- Extensibility. Users can create their own functions to do special-purpose processing.

In our project, we have executed Apache Pig on Hadoop in MapReduce to pull out the information we needed by using the csv files. Using the Apache Pig was a suitable alternative of Java. It has reduced the necessity of writing codes by almost 20 times.

e) CSV file:

A CSV is a comma separated values file which allows data to be saved in a table structured format. CSVs look like a garden-variety spreadsheet but with a .csv extension (Traditionally they take the form of a text file containing information separated by commas, hence the name).

CSV files can be used with any spreadsheet program, such as Microsoft Excel, Open Office Calc, or Google Spreadsheets. They differ from other spreadsheet file types in that you can only have a single sheet in a file, they cannot save cell, column, or row styling, and cannot save formulas.

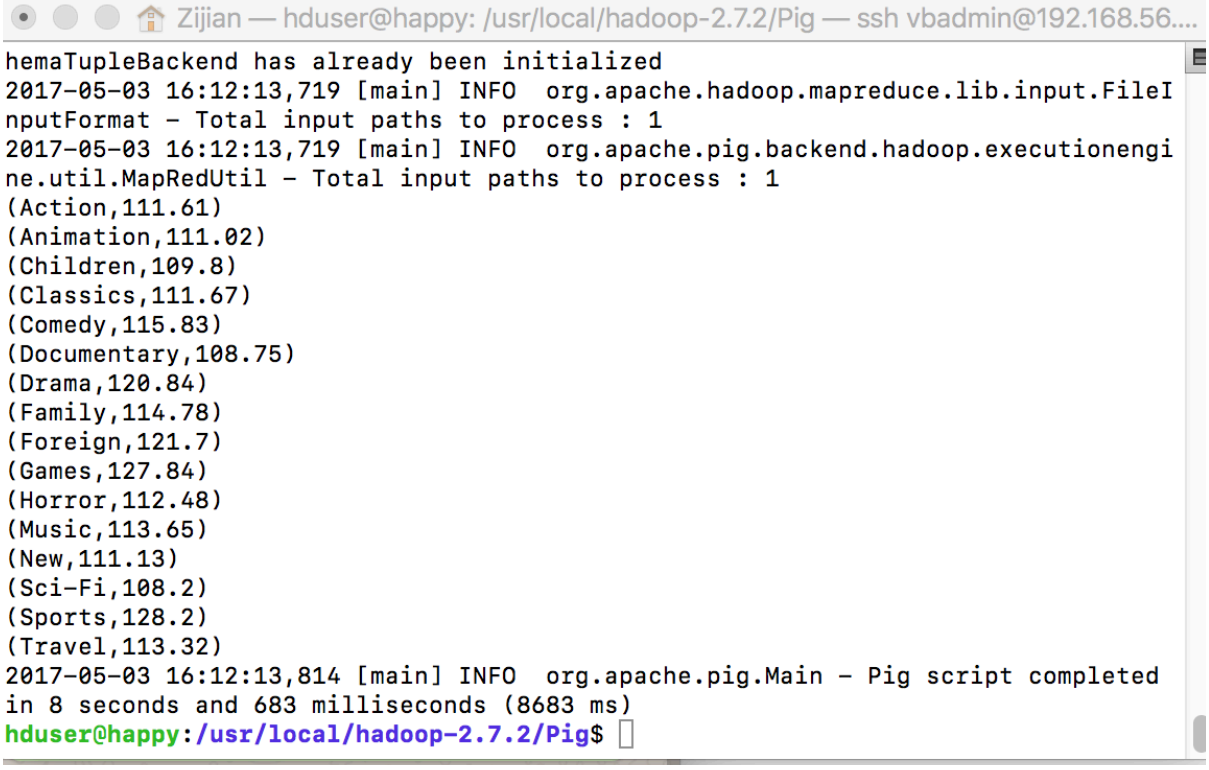
For our project, we have been given many csv files which needed to relate on another, based on the requirement of the question that is asked, and determine such information. There were many different csv files and they data like actor, address, city, film category, payment, language, rental, staff, store and etc. By relating one csv file with another, we collected information like the average length of films in each category, Categories have the longest and shortest average film lengths, Customers who have rented action but not comedy or classic movies, Actor who has appeared in the most English-language movies, The number of distinct movies that were rented for exactly 10 days from the store where Mike works, List actors in alphabetic who have appeared in the movie with the largest cast of actors.

3) Results:

Local mode

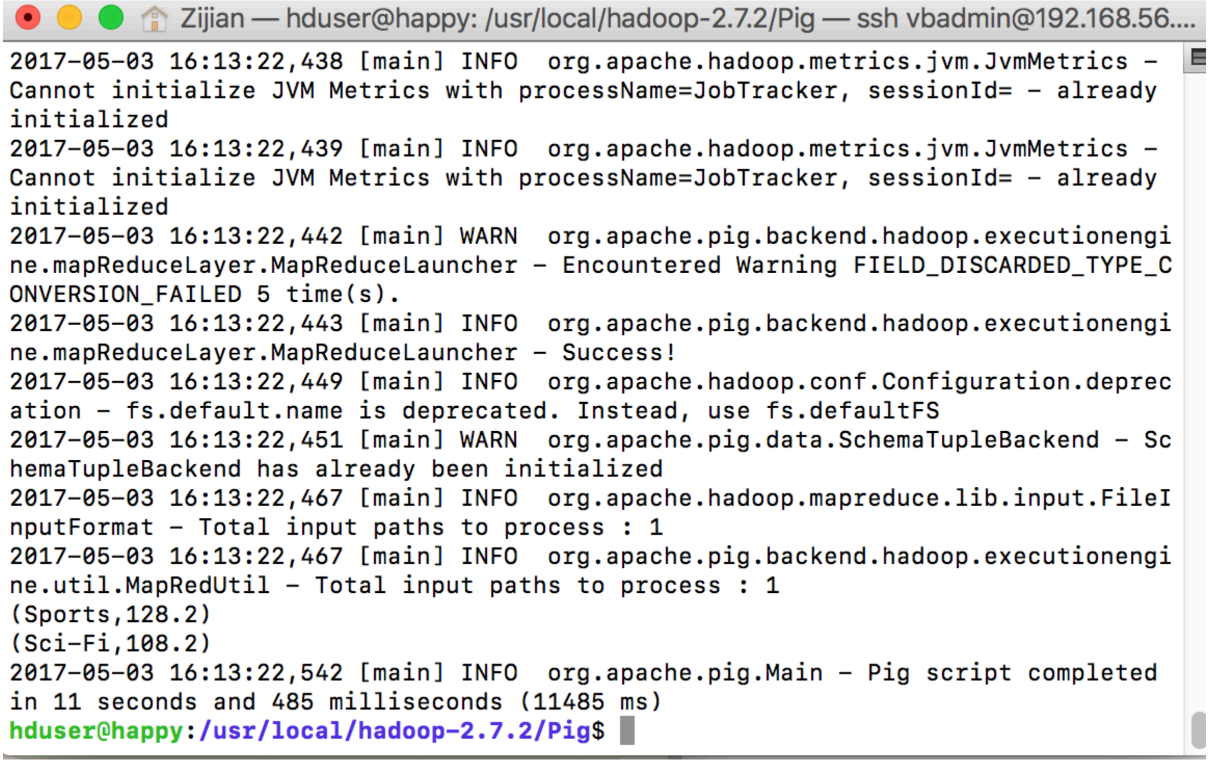
```
cd /usr/local/hadoop-2.7.2/Pig
start-dfs.sh
start-yarn.sh
mr-jobhistory-daemon.sh start historyserver
pig -x local queryA.pig
pig -x local queryB.pig
pig -x local queryC.pig
pig -x local queryD.pig
pig -x local queryE.pig
pig -x local queryD.pig
```

Query A


Zijian — hduser@happy: /usr/local/hadoop-2.7.2/Pig — ssh vbadmin@192.168.56....

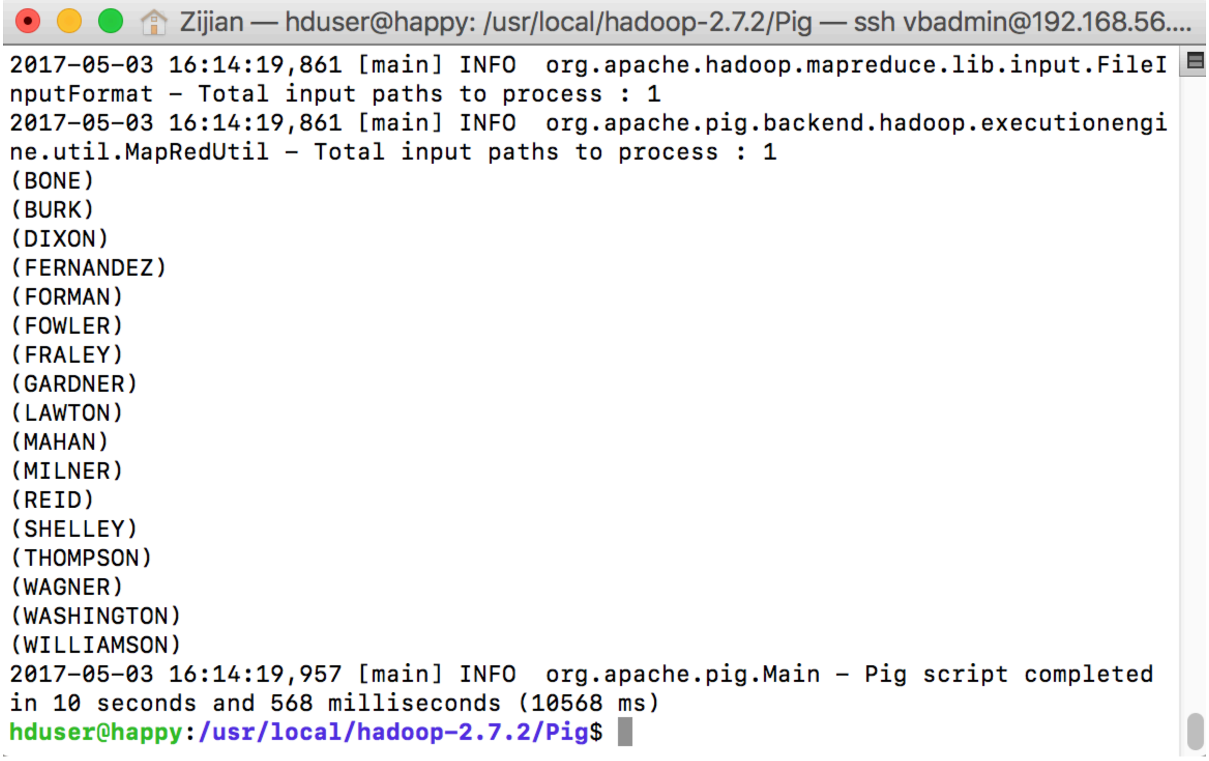
```
hemaTupleBackend has already been initialized
2017-05-03 16:12:13,719 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-05-03 16:12:13,719 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Action,111.61)
(Animation,111.02)
(Children,109.8)
(Classics,111.67)
(Comedy,115.83)
(Documentary,108.75)
(Drama,120.84)
(Family,114.78)
(Foreign,121.7)
(Games,127.84)
(Horror,112.48)
(Music,113.65)
(New,111.13)
(Sci-Fi,108.2)
(Sports,128.2)
(Travel,113.32)
2017-05-03 16:12:13,814 [main] INFO  org.apache.pig.Main - Pig script completed
in 8 seconds and 683 milliseconds (8683 ms)
hduser@happy:/usr/local/hadoop-2.7.2/Pig$ 
```

Query B


Zijian — hduser@happy: /usr/local/hadoop-2.7.2/Pig — ssh vbadmin@192.168.56....

```
2017-05-03 16:13:22,438 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-05-03 16:13:22,439 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-05-03 16:13:22,442 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 5 time(s).
2017-05-03 16:13:22,443 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2017-05-03 16:13:22,449 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-03 16:13:22,451 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2017-05-03 16:13:22,467 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-05-03 16:13:22,467 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Sports,128.2)
(Sci-Fi,108.2)
2017-05-03 16:13:22,542 [main] INFO  org.apache.pig.Main - Pig script completed
in 11 seconds and 485 milliseconds (11485 ms)
hduser@happy:/usr/local/hadoop-2.7.2/Pig$ 
```

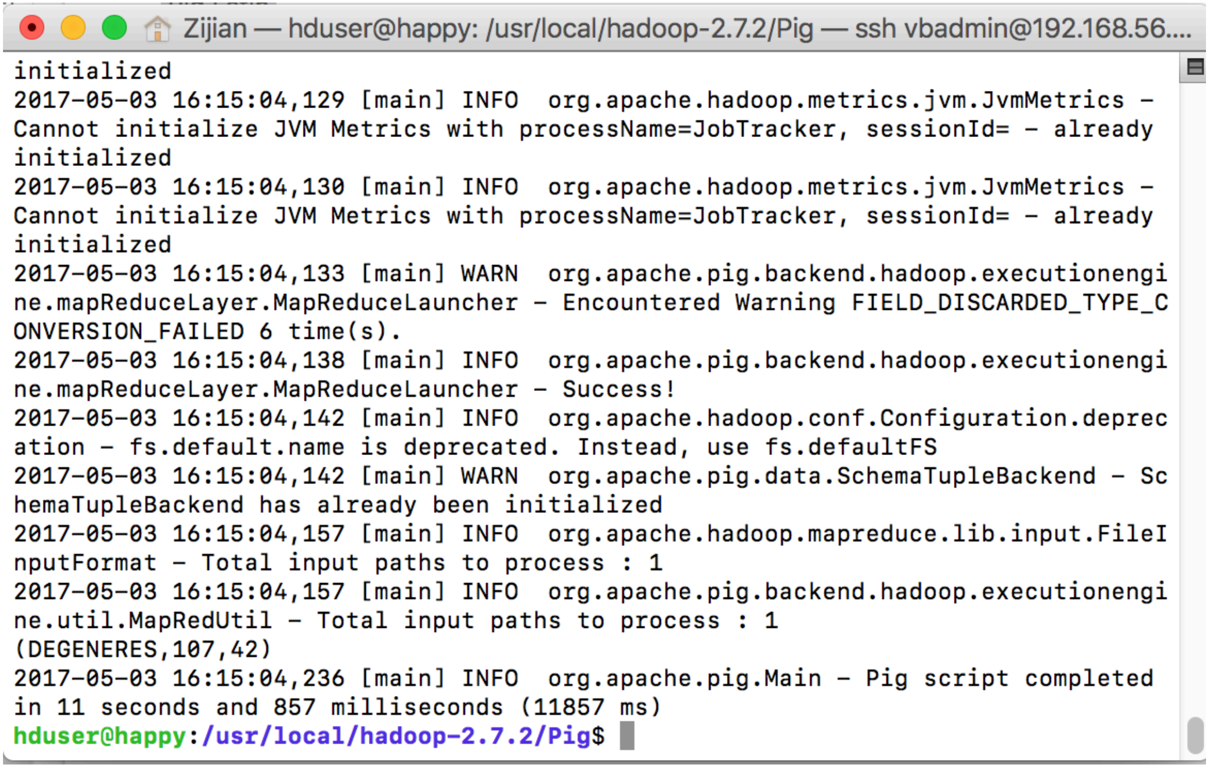
Query C



Zijian — hduser@happy: /usr/local/hadoop-2.7.2/Pig — ssh vbadmin@192.168.56....

```
2017-05-03 16:14:19,861 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-05-03 16:14:19,861 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(BONE)
(BURK)
(DIXON)
(FERNANDEZ)
(FORMAN)
(FOWLER)
(FRALEY)
(GARDNER)
(LAWTON)
(MAHAN)
(MILNER)
(REID)
(SHELLEY)
(THOMPSON)
(WAGNER)
(WASHINGTON)
(WILLIAMSON)
2017-05-03 16:14:19,957 [main] INFO  org.apache.pig.Main - Pig script completed
in 10 seconds and 568 milliseconds (10568 ms)
hduser@happy:/usr/local/hadoop-2.7.2/Pig$
```

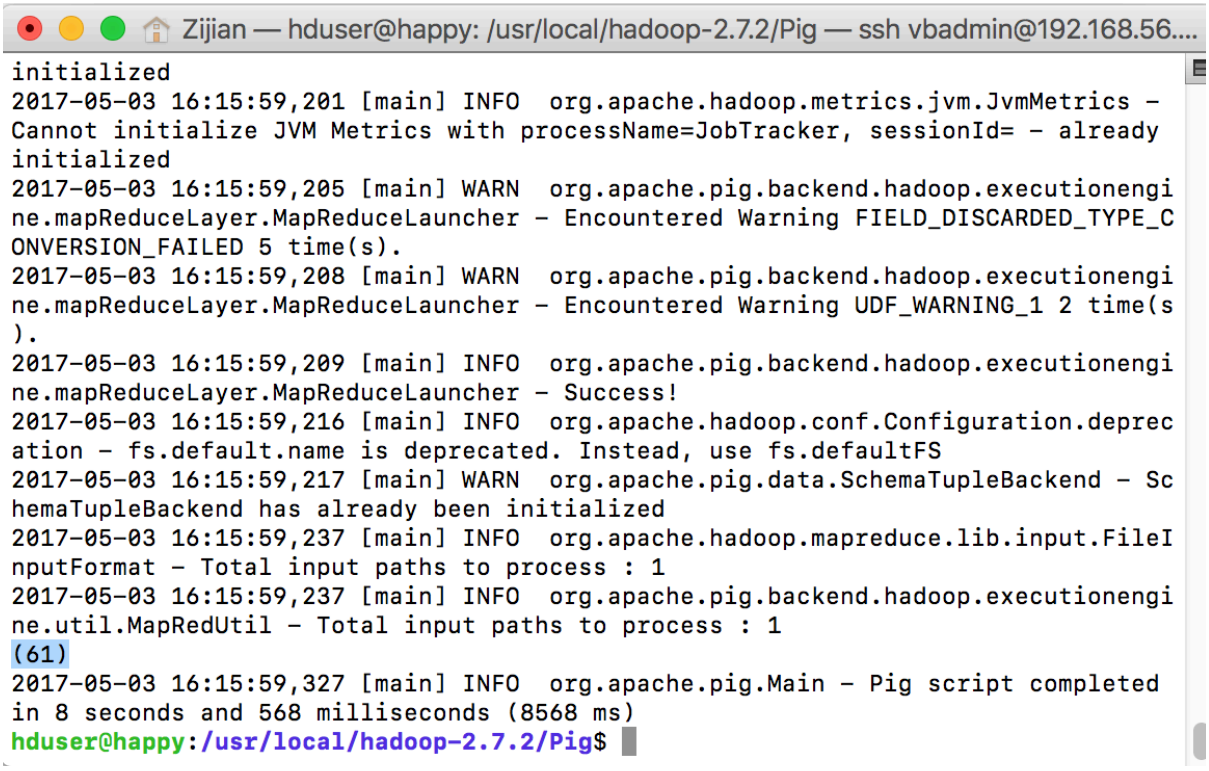
Query D



Zijian — hduser@happy: /usr/local/hadoop-2.7.2/Pig — ssh vbadmin@192.168.56....

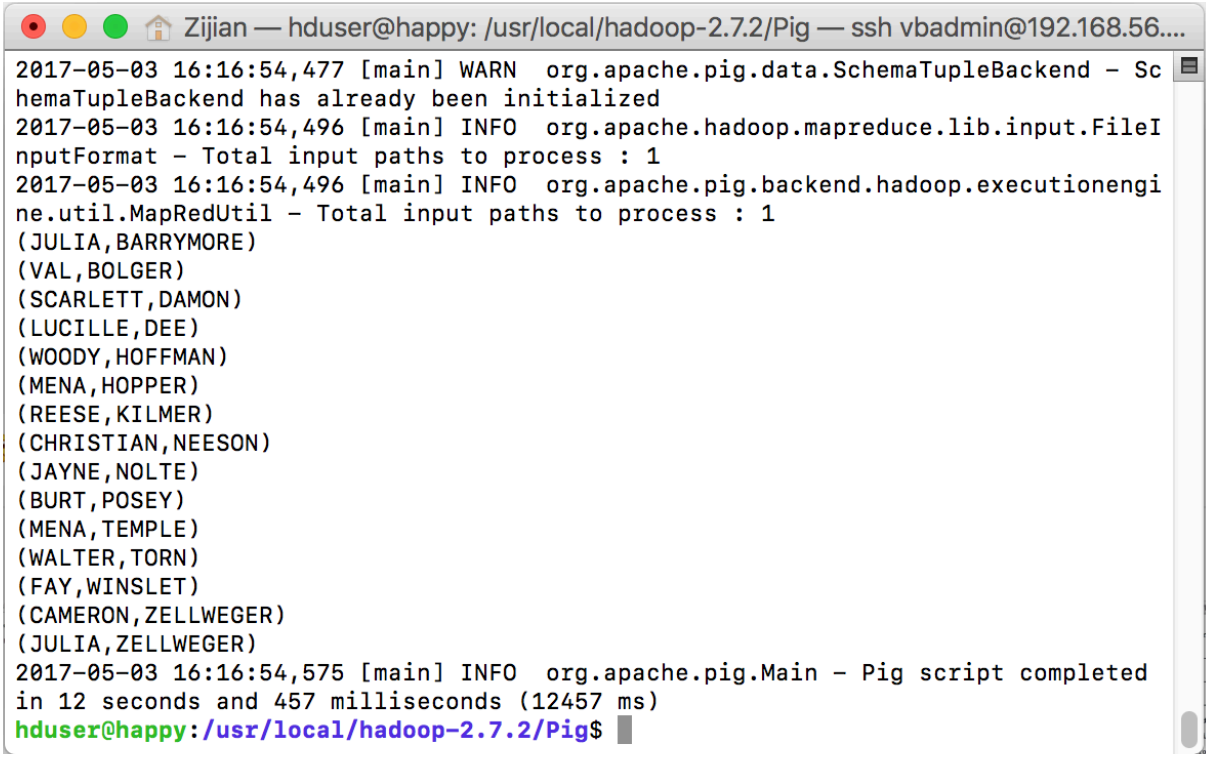
```
initialized
2017-05-03 16:15:04,129 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-05-03 16:15:04,130 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-05-03 16:15:04,133 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 6 time(s).
2017-05-03 16:15:04,138 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2017-05-03 16:15:04,142 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-03 16:15:04,142 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2017-05-03 16:15:04,157 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-05-03 16:15:04,157 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(DEGENERES,107,42)
2017-05-03 16:15:04,236 [main] INFO  org.apache.pig.Main - Pig script completed
in 11 seconds and 857 milliseconds (11857 ms)
hduser@happy:/usr/local/hadoop-2.7.2/Pig$
```

Query E


Zijian — hduser@happy: /usr/local/hadoop-2.7.2/Pig — ssh vbadmin@192.168.56....

```
initialized
2017-05-03 16:15:59,201 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2017-05-03 16:15:59,205 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 5 time(s).
2017-05-03 16:15:59,208 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning UDF_WARNING_1 2 time(s
).
2017-05-03 16:15:59,209 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2017-05-03 16:15:59,216 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-05-03 16:15:59,217 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2017-05-03 16:15:59,237 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-05-03 16:15:59,237 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(61)
2017-05-03 16:15:59,327 [main] INFO  org.apache.pig.Main - Pig script completed
in 8 seconds and 568 milliseconds (8568 ms)
hduser@happy:/usr/local/hadoop-2.7.2/Pig$
```

Query F

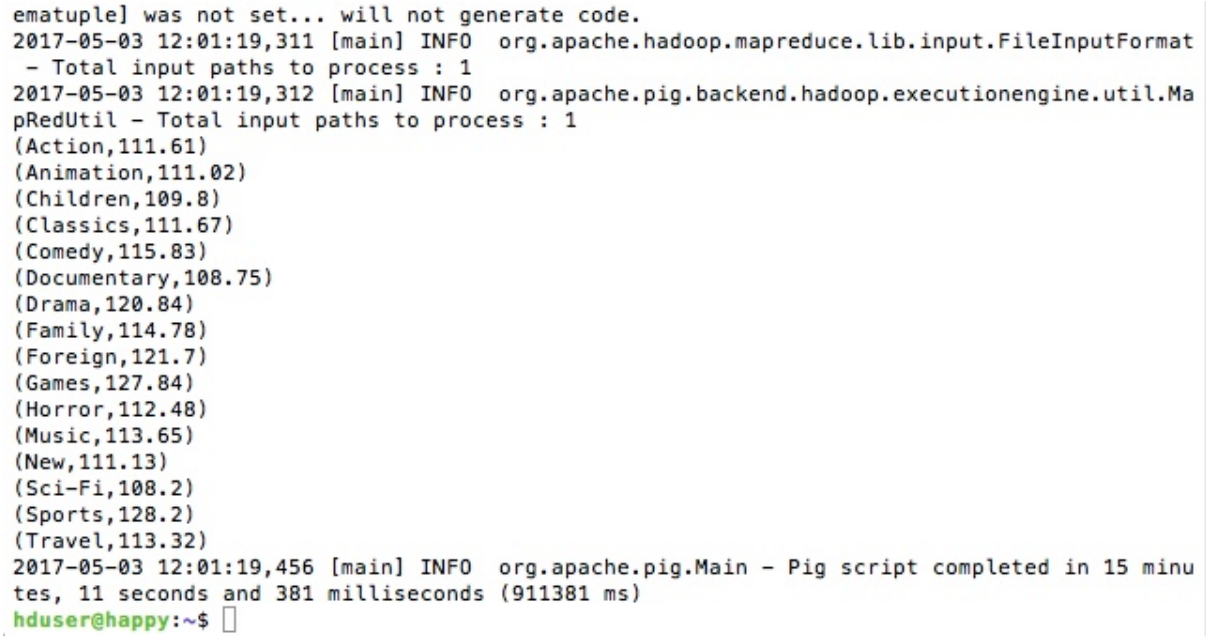

Zijian — hduser@happy: /usr/local/hadoop-2.7.2/Pig — ssh vbadmin@192.168.56....

```
2017-05-03 16:16:54,477 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2017-05-03 16:16:54,496 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2017-05-03 16:16:54,496 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(JULIA,BARRYMORE)
(VAL,BOLGER)
(SCARLETT,DAMON)
(LUCILLE,DEE)
(WOODY,HOFFMAN)
(MENA,HOPPER)
(REESE,KILMER)
(CHRISTIAN,NEESON)
(JAYNE,NOLTE)
(BURT,POSEY)
(MENA,TEMPLE)
(WALTER,TORN)
(FAY,WINSLET)
(CAMERON,ZELLWEGER)
(JULIA,ZELLWEGER)
2017-05-03 16:16:54,575 [main] INFO  org.apache.pig.Main - Pig script completed
in 12 seconds and 457 milliseconds (12457 ms)
hduser@happy:/usr/local/hadoop-2.7.2/Pig$
```

MapReduce mode

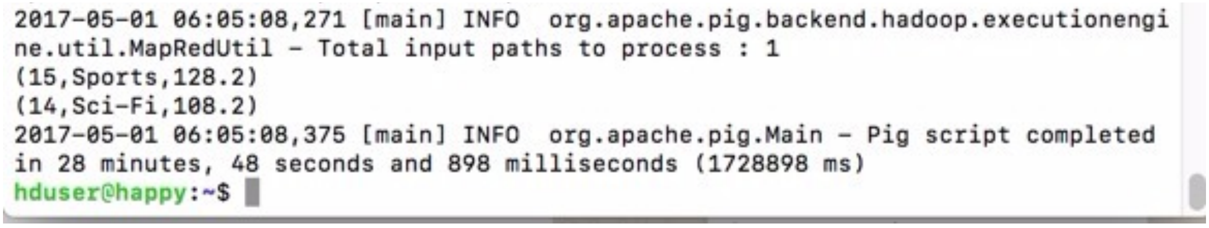
```
stop-dfs.sh
stop-yarn.sh
start-dfs.sh
start-yarn.sh
pig -x mapreduce queryA.pig
pig -x mapreduce queryB.pig
pig -x mapreduce queryC.pig
pig -x mapreduce queryD.pig
pig -x mapreduce queryE.pig
pig -x mapreduce queryF.pig
```

Query A



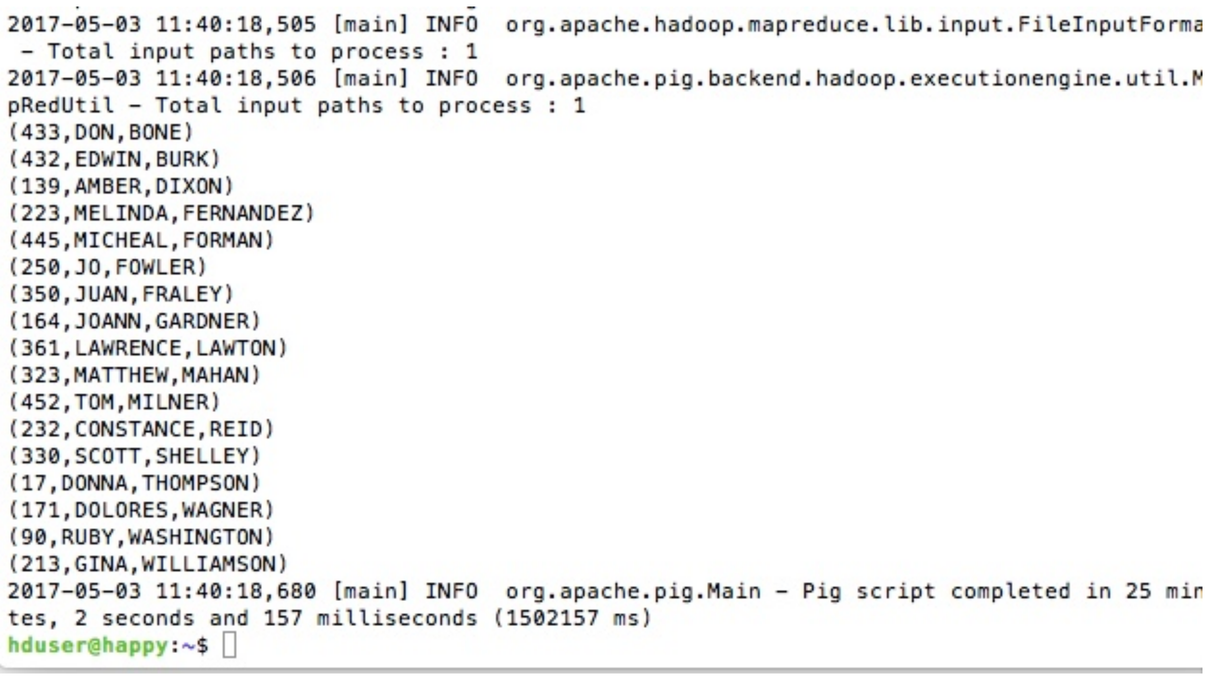
```
ematuple] was not set... will not generate code.
2017-05-03 12:01:19,311 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat
 - Total input paths to process : 1
2017-05-03 12:01:19,312 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.Ma
pRedUtil - Total input paths to process : 1
(Action,111.61)
(Animation,111.02)
(Children,109.8)
(Classics,111.67)
(Comedy,115.83)
(Documentary,108.75)
(Drama,120.84)
(Family,114.78)
(Foreign,121.7)
(Games,127.84)
(Horror,112.48)
(Music,113.65)
(New,111.13)
(Sci-Fi,108.2)
(Sports,128.2)
(Travel,113.32)
2017-05-03 12:01:19,456 [main] INFO  org.apache.pig.Main - Pig script completed in 15 minu
tes, 11 seconds and 381 milliseconds (911381 ms)
hduser@happy:~$ 
```

Query B



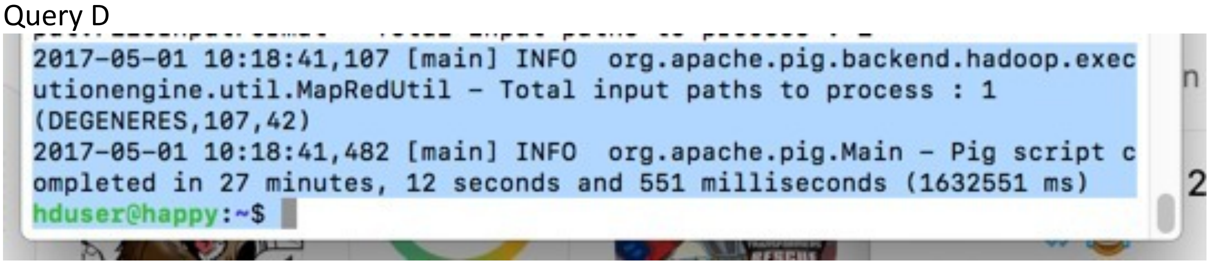
```
2017-05-01 06:05:08,271 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(15,Sports,128.2)
(14,Sci-Fi,108.2)
2017-05-01 06:05:08,375 [main] INFO  org.apache.pig.Main - Pig script completed
in 28 minutes, 48 seconds and 898 milliseconds (1728898 ms)
hduser@happy:~$ 
```

Query C



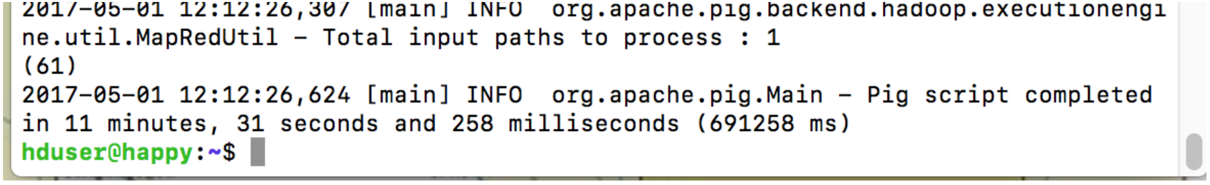
```
2017-05-03 11:40:18,505 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputForma
 - Total input paths to process : 1
2017-05-03 11:40:18,506 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.M
pRedUtil - Total input paths to process : 1
(433,DON,BONE)
(432,EDWIN,BURK)
(139,AMBER,DIXON)
(223,MELINDA,FERNANDEZ)
(445,MICHEAL,FORMAN)
(250,JO,FOWLER)
(350,JUAN,FRALEY)
(164,JOANN,GARDNER)
(361,LAWRENCE,LAWTON)
(323,MATTHEW,MAHAN)
(452,TOM,MILNER)
(232,CONSTANCE,REID)
(330,SCOTT,SHELLEY)
(17,DONNA,THOMPSON)
(171,DOLORES,WAGNER)
(90,RUBY,WASHINGTON)
(213,GINA,WILLIAMSON)
2017-05-03 11:40:18,680 [main] INFO  org.apache.pig.Main - Pig script completed in 25 mir
tes, 2 seconds and 157 milliseconds (1502157 ms)
hduser@happy:~$ 
```

Query D



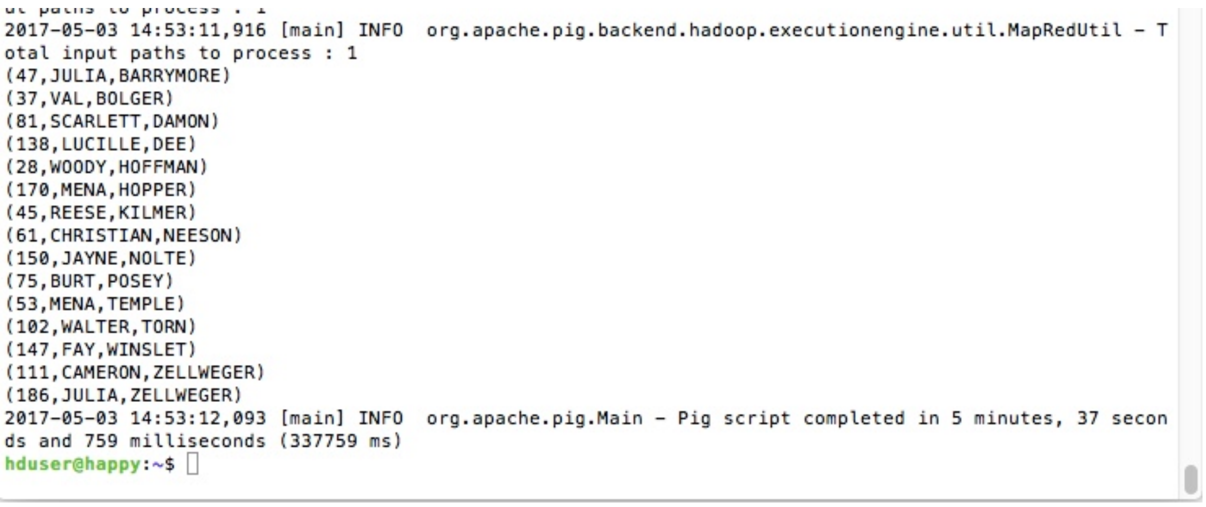
```
2017-05-01 10:18:41,107 [main] INFO  org.apache.pig.backend.hadoop.exec
utionengine.util.MapRedUtil - Total input paths to process : 1
(DEGENERES,107,42)
2017-05-01 10:18:41,482 [main] INFO  org.apache.pig.Main - Pig script c
ompleted in 27 minutes, 12 seconds and 551 milliseconds (1632551 ms)
hduser@happy:~$ 
```

Query E



```
2017-05-01 12:12:26,307 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(61)
2017-05-01 12:12:26,624 [main] INFO  org.apache.pig.Main - Pig script completed
in 11 minutes, 31 seconds and 258 milliseconds (691258 ms)
hduser@happy:~$ 
```

Query F



```
ut paths to process : 1
2017-05-03 14:53:11,916 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - T
otal input paths to process : 1
(47,JULIA,BARRYMORE)
(37,VAL,BOLGER)
(81,SCARLETT,DAMON)
(138,LUCILLE,DEE)
(28,WOODY,HOFFMAN)
(170,MENA,HOPPER)
(45,REESE,KILMER)
(61,CHRISTIAN,NEESON)
(150,JAYNE,NOLTE)
(75,BURT,POSEY)
(53,MENA,TEMPLE)
(102,WALTER,TORN)
(147,FAY,WINSLET)
(111,CAMERON,ZELLWEGER)
(186,JULIA,ZELLWEGER)
2017-05-03 14:53:12,093 [main] INFO  org.apache.pig.Main - Pig script completed in 5 minutes, 37 secon
ds and 759 milliseconds (337759 ms)
hduser@happy:~$ 
```