

## Pig Latin Hands on Exercise

### Exercise 1:

#### **Dataset:**

We will use the White House datasets located under **/Pig/top10/input** in the HDFS in the Programming/Master Node CS6360.utdallas.edu. Please use this folder and don't copy to any other folder on the server. All datasets are comma separated.

This dataset contain sample White House visitor information. Each line has the following 11 columns:

NAMELAST,NAMEFIRST,NAMEMID,UIN,BDGNBR,ACCESS\_TYPE,TOA,POA,TOD,POD,APPT\_MADE\_DATE.

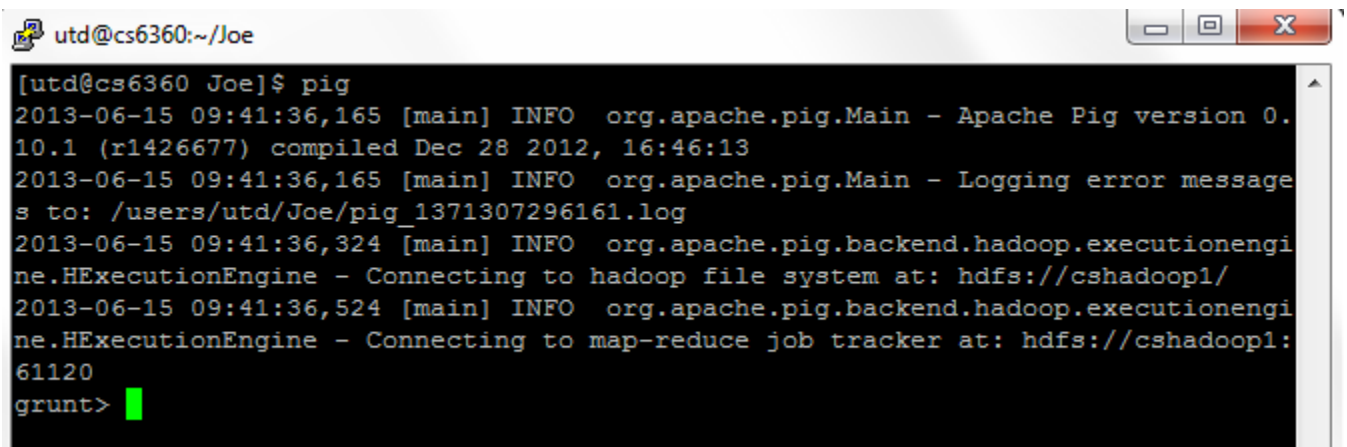
#### **Requirement:**

Using Pig Latin commands, find the top 10 most frequent visitors (NAMELAST, NAMEFIRST) to the White House.

1. Log in cs6360.utdallas.edu and change directory to **Joe** (you should put your name instead of Joe) by following commands:  
**cd /users/utd/Joe**

Then run the following command:

**pig**

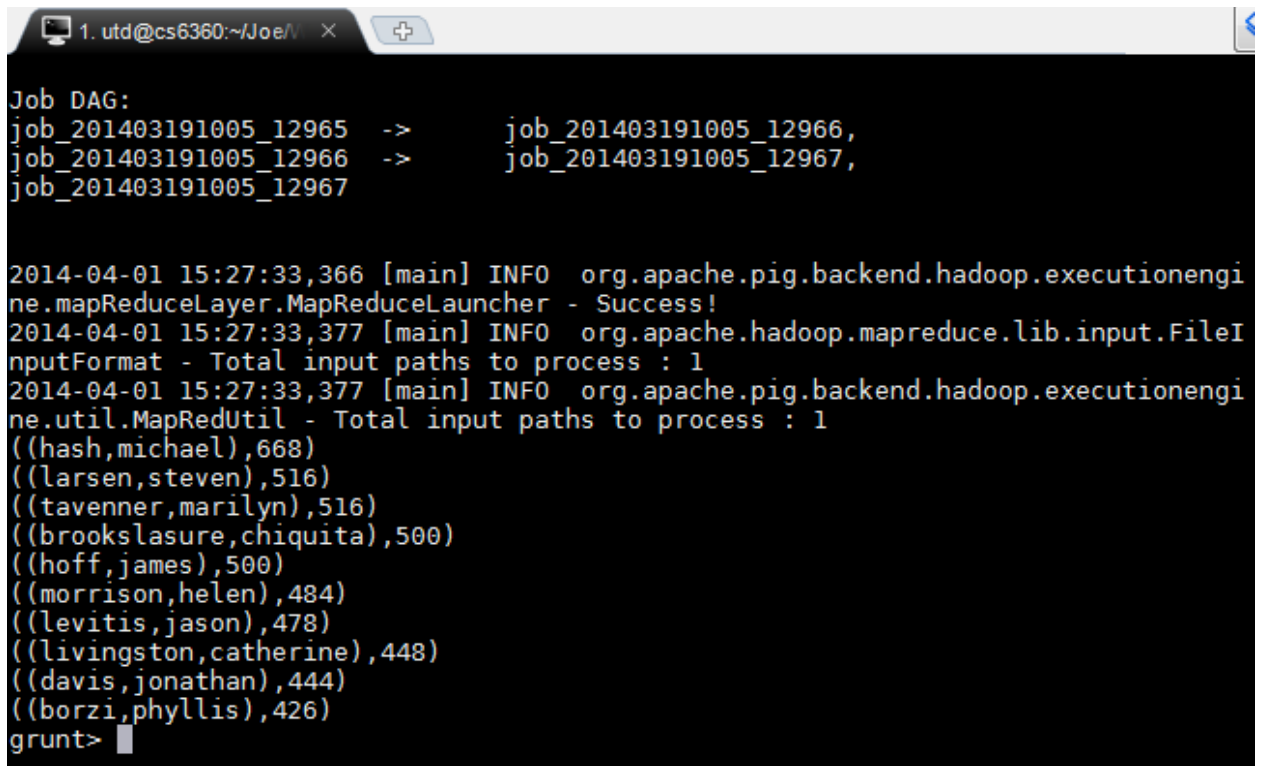


```
utd@cs6360:~/Joe
[utd@cs6360 Joe]$ pig
2013-06-15 09:41:36,165 [main] INFO org.apache.pig.Main - Apache Pig version 0.10.1 (r1426677) compiled Dec 28 2012, 16:46:13
2013-06-15 09:41:36,165 [main] INFO org.apache.pig.Main - Logging error messages to: /users/utd/Joe/pig_1371307296161.log
2013-06-15 09:41:36,324 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://cshadoop1/
2013-06-15 09:41:36,524 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: hdfs://cshadoop1:61120
grunt> █
```

The above dialog shows the interactive mode. In this mode you can execute pig commands one by one.

2. Run the following commands sequentially.

- Load the input CSV file into a variable A  
**A = load '/Pig/top10/input' using PigStorage(',') as (NAMELAST,NAMEFIRST,NAMEMID,UIN,BDGNBR,ACCESS\_TYPE,TOA,POA,TOD,POD,APPT\_MADE\_DATE);**
- Group A by NAMELAST,NAMEFIRST and put it in to variable B  
**B = group A by (NAMELAST,NAMEFIRST);**
- For each group find the number of visits and put it to variable C  
**C = foreach B generate group, COUNT(A.(NAMELAST,NAMEFIRST)) as num\_of\_visits;**
- Sort visitors by descending order with their number of visits and put these to variable D.  
**D = order C by num\_of\_visits desc;**
- Take the first 10 visitors.  
**E = limit D 10;**
- See the output  
**dump E;**



The screenshot shows a terminal window with a dark background. At the top, a browser-like tab is visible with the text '1. utd@cs6360:~/Joe/V'. The terminal output displays a 'Job DAG' with three jobs: 'job\_201403191005\_12965', 'job\_201403191005\_12966', and 'job\_201403191005\_12967'. Below this, there are several log messages from the Apache Pig backend, including 'INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!' and 'INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1'. The final output of the 'dump E;' command is a list of 10 tuples, each containing a name and a count, such as '((hash,michael),668)' and '((larsen,steven),516)'. The prompt 'grunt>' is visible at the bottom.

```
Job DAG:
job_201403191005_12965 -> job_201403191005_12966,
job_201403191005_12966 -> job_201403191005_12967,
job_201403191005_12967

2014-04-01 15:27:33,366 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2014-04-01 15:27:33,377 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2014-04-01 15:27:33,377 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
((hash,michael),668)
((larsen,steven),516)
((tavenner,marilyn),516)
((brookslasure,chiquita),500)
((hoff,james),500)
((morrison,helen),484)
((levitis,jason),478)
((livingston,catherine),448)
((davis,jonathan),444)
((borzi,phyllis),426)
grunt>
```

- Store the result to a file to your directory  
**store E into '/home/<your name>/Pig/output/top10/top\_visitors';**

For example:

**store E into '/home/Joe/Pig/output/top10/top\_visitors';**

N.B. If you have the error “File already exists” then remove the file by following commands from grunt shell in Pig and execute the above command.

**fs -rmr /home/<your name>/Pig/output/top10/top\_visitors**

For example:

**fs -rmr /home/Joe/Pig/output/top10/top\_visitors**

The file is stored at HDFS. The file can be viewed by

**fs -cat /home/<your name>/Pig/output/top10/top\_visitors/part-r-00000**

For example:

**fs -cat /home/Joe/Pig/output/top10/top\_visitors/part-r-00000**

## **Exercise 2:**

### **Datasets:**

The three datasets (**/Pig/join/input**) that will be used are as follows:

- NASA\_HTTP.txt: The delimiter is tab and each line has the following 2 columns IP, VALUE.
- HOST\_COUNTRY.txt: The delimiter is tab and each line has the following 2 columns IP, COUNTRY ABBREVIATION.
- COUNTRY\_NAME.txt: The delimiter is tab and each line has the following 2 columns COUNTRY ABBREVIATION, COUNTRY NAME

### **Requirement:**

1. Write Pig Latin commands to do multiple tables inner join for the above mentioned datasets (*the join attribute is (IP) for the first two datasets and country abbreviation for the second and third datasets.*)

- Load the input text file (tab delimited) file into a variable A, B and C

**A = load '/Pig/join/input/NASA\_HTTP.txt' using PigStorage('\t') as (IP,VALUE);**

**B = load '/Pig/join/input/HOST\_COUNTRY.txt' using PigStorage('\t') as (IP,  
COUNTRY\_ABBREVIATION);**

**C= load '/Pig/join/input/COUNTRY\_NAME.txt' using PigStorage('\t') as  
(COUNTRY\_ABBREVIATION, COUNTRY\_NAME);**

- Join A and B by IP and put the result into D  
**D = join A by IP, B by IP;**
- Join D and C by COUNTRY\_ABBREVIATION and put the result into E  
**E = join D by COUNTRY\_ABBREVIATION, C by COUNTRY\_ABBREVIATION;**
- Show the output  
**dump E;**

```
(128.220.116.211,[03/Jul/1995:14:56:56 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:56:44 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:56:44 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 200 234,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:56:43 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:56:41 -0400] "GET /ksc.html HTTP/1.0" 200 7074,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:56:22 -0400] "GET /images/shuttle-patch-logo.gif HTTP/1.0" 200 891,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:56:10 -0400] "GET /shuttle/technology/sts-newsref/sts_mes.html HTTP/1.0" 200 175621,128.220.116.211,US,US,United States)
(128.220.116.211,[06/Jul/1995:17:52:55 -0400] "GET /history/history.html HTTP/1.0" 200 1602,128.220.116.211,US,US,United States)
(128.220.116.211,[06/Jul/1995:17:54:19 -0400] "GET /images/shuttle-patch-logo.gif HTTP/1.0" 200 891,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:16:08:16 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:16:08:16 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:16:08:16 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 200 234,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:16:14:06 -0400] "GET /shuttle/missions/sts-71/images/KSC-95EC-0913.gif HTTP/1.0" 200 21957,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:23:24 -0400] "GET /images/ksclogosmall.gif HTTP/1.0" 200 3635,128.220.116.211,US,US,United States)
(128.220.116.211,[06/Jul/1995:17:52:26 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:23:23 -0400] "GET /images/kscmap-tiny.gif HTTP/1.0" 200 2537,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:23:09 -0400] "GET /images/lc39a-logo.gif HTTP/1.0" 200 13116,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:54:18 -0400] "GET /shuttle/technology/sts-newsref/sts-lcc.html HTTP/1.0" 200 32252,128.220.116.211,US,US,United States)
(128.220.116.211,[03/Jul/1995:14:23:06 -0400] "GET /facilities/lc39a.html HTTP/1.0" 200 7008,128.220.116.211,US,US,United States)
```

- Store the result to a file to your directory  
**store E into '/home/<your name>/Pig/output/join/nasa';**

For example:

**store E into '/home/Joe/Pig/output/join/nasa';**

- N.B. If you have the error “File already exists” then remove the file by following commands from grunt shell in Pig and execute the following command.  
**fs -rmr /home/<your name>/Pig/output/join/nasa**

Example:

**fs -rmr /home/Joe/Pig/output/join/nasa**

- The file is stored at HDFS. The file can be viewed by  
**fs -cat /home/<your name>/Pig/output/join/nasa/part-r-00000**

For example:

**fs -cat /home/Joe/Pig/output/join/nasa/part-r-00000**

2. Implement Co-group command on IP for the datasets NASA\_HTTP and HOST\_COUNTRY

- Load the input text file (tab delimited) file into a variable A, B and C

**A = load '/Pig/join/input/NASA\_HTTP.txt' using PigStorage('\t') as (IP,VALUE);**

**B = load '/Pig/join/input/HOST\_COUNTRY.txt' using PigStorage('\t') as (IP, COUNTRY\_ABBREVIATION);**

- Co-group A and B by IP and put the result into D  
**C = cogroup A by IP, B by IP;**

- Show the output  
**dump C;**

- Store the result to a file to your directory  
**store C into '/home/<your name>/Pig/output/cogroup/nasa';**

For example:

**store C into '/home/Joe/Pig/output/cogroup/nasa';**

- N.B. If you have the error “File already exists” then remove the file by following commands from grunt shell in Pig and execute the above command.  
**fs -rmr /home/<your name>/Pig/output/cogroup/nasa**

Example:

**fs -rmr /home/Joe/Pig/output/cogroup/nasa**

- The file is stored at HDFS. The file can be viewed by  
**fs -cat /home/<Your name>/Pig/output/cogroup/nasa/part-r-00000**

For example:

**fs -cat /home/Joe/Pig/output/cogroup/nasa/part-r-00000**

```
/images/et-intertank_1-small.gif HTTP/1.0" 200 79791),(205.254.162.108,[14/Aug/1995:20:24:26 -0400] "GET /icons/text.xbm HTTP/1.0" 200 527),(205.254.162.108,[14/Aug/1995:20:24:26 -0400] "GET /icons/image.xbm HTTP/1.0" 200 509),(205.254.162.108,[14/Aug/1995:20:24:03 -0400] "GET /shuttle/missions/51-l/51-l-crew.gif HTTP/1.0" 200 172498)} {}
205.254.180.217 {(205.254.180.217,[30/Aug/1995:13:16:27 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 200 234),(205.254.180.217,[30/Aug/1995:13:16:30 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669),(205.254.180.217,[30/Aug/1995:13:16:55 -0400] "GET /shuttle/missions/sts-70/movies/movies.html HTTP/1.0" 200 2979),(205.254.180.217,[30/Aug/1995:13:16:57 -0400] "GET /shuttle/missions/sts-70/sts-70-patch-small.gif HTTP/1.0" 200 5978),(205.254.180.217,[30/Aug/1995:13:17:47 -0400] "GET /shuttle/movies/TDRSopen.mpg HTTP/1.0" 200 401595),(205.254.180.217,[30/Aug/1995:13:23:55 -0400] "GET / HTTP/1.0" 200 7089),(205.254.180.217,[30/Aug/1995:13:23:56 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866),(205.254.180.217,[30/Aug/1995:13:23:57 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786),(205.254.180.217,[30/Aug/1995:13:23:58 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363),(205.254.180.217,[30/Aug/1995:13:23:59 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 200 234),(205.254.180.217,[30/Aug/1995:13:24:01 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 200 669),(205.254.180.217,[30/Aug/1995:13:24:33 -0400] "GET /shuttle/missions/sts-70/movies/movies.html HTTP/1.0" 200 2979),(205.254.180.217,[30/Aug/1995:13:24:35 -0400] "GET /shuttle/missions/sts-70/sts-70-patch-small.gif HTTP/1.0" 200 5978),(205.254.180.217,[30/Aug/1995:13:26:37 -0400] "GET /shuttle/missions/sts-70/movies/sts-70-tdrs-deploy.mpg HTTP/1.0" 200 692912),(205.254.180.217,[30/Aug/1995:13:16:23 -0400] "GET / HTTP/1.0" 200 7089),(205.254.180.217,[30/Aug/1995:13:16:25 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866),(205.254.180.217,[30/Aug/1995:13:16:25 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786),(205.254.180.217,[30/Aug/1995:13:16:26 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363)} {}
205.254.180.221 {(205.254.180.221,[29/Aug/1995:14:45:59 -0400] "GET /shuttle/missions/sts-71/sts-71-patch-small.gif HTTP/1.0" 200 12054),(205.254.180.221,[29/Aug/1995:14:46:02 -0400] "GET /shuttle/missions/sts-71/sts-71-patch-small.gif HTTP/1.0" 200 12054),(205.254.180.221,[29/Aug/1995:14:46:01 -0400] "GET /shuttle/missions/sts-71/movies/movies.html HTTP/1.0" 200 3381),(205.254.180.221,[29/Aug/1995:14:51:38 -0400] "GET /shuttle/missions/sts-71/movies/sts-71-launch.mpg HTTP/1.0" 200 1121554)} {}
205.254.180.237 {(205.254.180.237,[26/Jul/1995:16:52:21 -0400] "GET /images/launchmedium.gif HTTP/1.0" 200 11853),(205.254.180.237,[26/Jul/1995:16:52:11 -0400] "GET /shuttle/missions/missions.html HTTP/1.0" 200 8677),(205.254.180.237,[26/Jul/1995:16:54:24 -0400] "GET /shuttle/technology/sts-newsref/stsref-toc.html HTTP/1.0" 200 84905),(205.254.180.237,[26/Jul/1995:16:53:54 -0400] "GET /images/shuttle-patch-small.gif HTTP/1.0" 200 4179),(205.254.180.237,[26/Jul/1995:16:52:40 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204),(205.254.180.237,[26/Jul/1995:16:52:22 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786)} {(205.254.180.237,US)}
```