Instructions For Running Map Reduce Program

**How to log in by SSH terminal?**

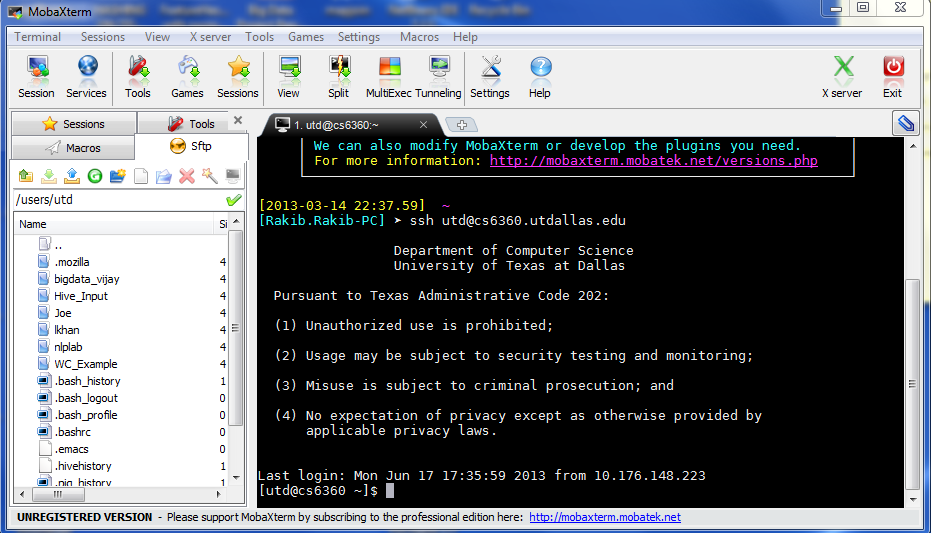
Please download an SSH client on your Windows laptops before you arrive to the workshop.  Here are the download URL's:  
  
      <http://mobaxterm.mobatek.net/MobaXterm_Setup_6.3.msi>  
      <http://www.hlt.utdallas.edu/MobaXterm_Setup_6.3.msi>

Log in **cs6360.utdallas.edu** with

Username: **utd**

Password: **hadoop.**

Run **MobaXterm** and type **utd@cs6360.utdallas.edu**

****

After log in verify your current directory by **pwd** command on ssh terminal. **pwd** shows your current directory **/users/utd**

Make sure that you have the **hadoop-core-1.0.4.jar** file in **/users/utd**. you can list all files and folder in **/users/utd** by **ls**command on terminal.

**Create a Folder by mkdir command with your name(e.g. Joe) inside /users/utd . Invoke your directory by issuing cd command on terminal .**

**In this document examples are shown using folder name 'Joe'. Please use your own name of folder in place of 'Joe' inside the commands.**

Pig Latin Hands on Exercise

**Dataset**:

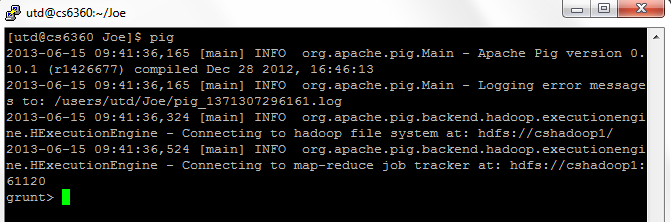
We will use the White House datasets located under **/Pig/top10/input** in the HDFS in the Programming/Master Node CS6360.utdallas.edu. Please use this folder and don’t copy to any other folder on the server. All datasets are comma separated and each line has the following 11 columns NAMELAST,NAMEFIRST,NAMEMID,UIN,BDGNBR,ACCESS\_TYPE,TOA,POA,TOD,POD,APPT\_MADE\_DATE.

**Requirement:**

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

1. Log in cs6360.utdallas.edu and Run pig

* **pig**



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.

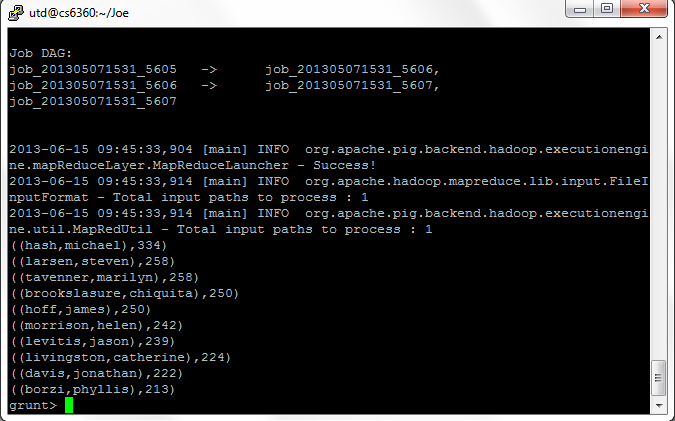
**D = order C by num\_of\_visits desc;**

* Take the first 10 visitors.

**E = limit D 10;**

* See the output

**dump E;**

****

* Store the result to a file.

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

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

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

**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

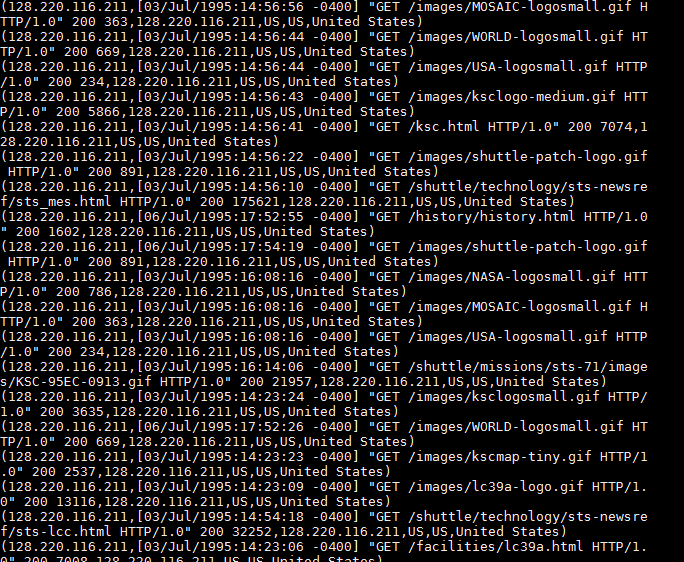
* **D = join A by IP, B by IP;**

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;**

****

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

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

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

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

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

****