Instructions For Running Map Reduce Program

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

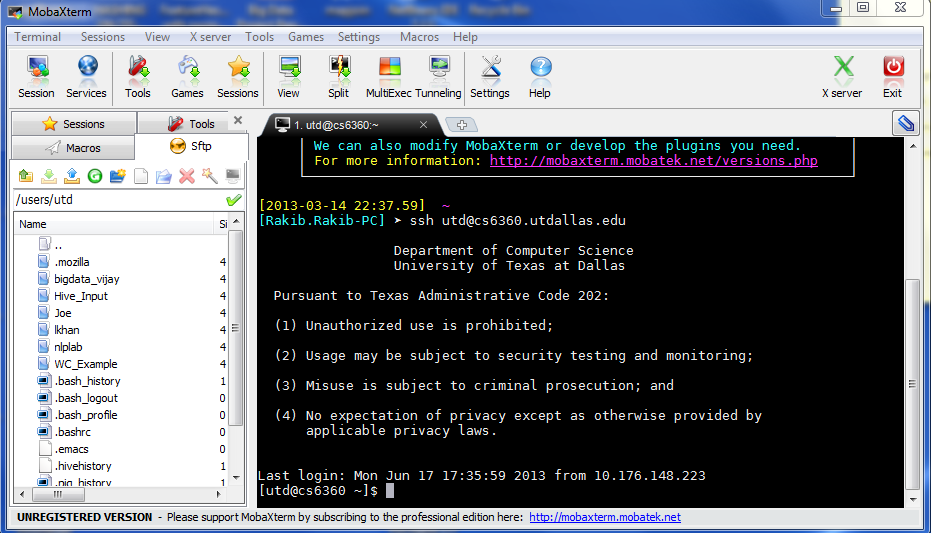
Please download an SSH client on your Windows laptops before you arrive to the workshop.  Here are the download URL's:  
  
      <http://mobaxterm.mobatek.net/MobaXterm_Setup_6.3.msi>  
      <http://www.hlt.utdallas.edu/MobaXterm_Setup_6.3.msi>

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.**

Hive Hands on Exercise

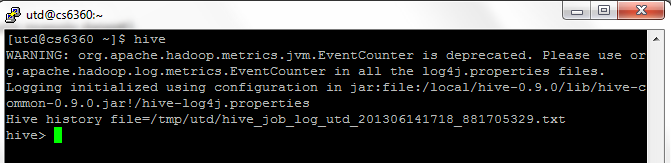
**Dataset**:

We will use the White House datasets located under **/users/utd/Hive\_Input/top10** 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 Hive commands, find the 10 most frequent visitors (NAMELAST, NAMEFIRST) to the White House (use Pig Latin example dataset).

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



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

2. Run the following commands sequentially.

* Create table visitor

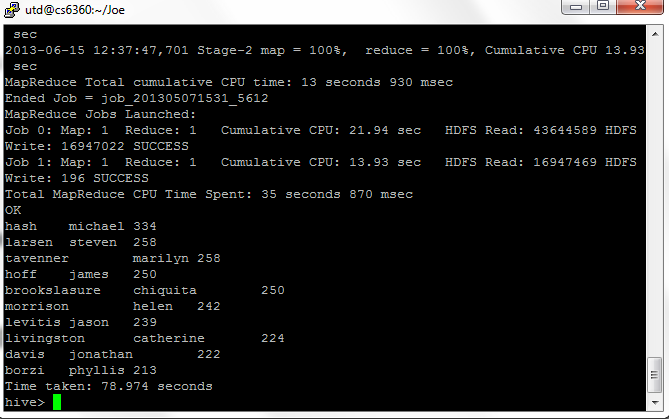
**create table visitor (NAMELAST STRING,NAMEFIRST STRING,NAMEMID STRING,UIN STRING,BDGNBR STRING,ACCESS\_TYPE STRING,TOA STRING,POA STRING,TOD STRING,POD STRING,APPT\_MADE\_DATE STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';**

* Load data from input CSV file into a table visitor

**LOAD DATA local INPATH '/users/utd/Hive\_Input/top10/White\_House.txt' into table visitor;**

* Sort visitors by descending order with their number of visits and fetch top 10 visitors

**select NAMELAST, NAMEFIRST, count(\*) as visitor\_count from visitor group by NAMELAST, NAMEFIRST order by visitor\_count desc limit 10;**



**Datasets**:

The three datasets (**/users/utd/Hive\_Input/nasa**) that will be used are as follows:

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

The three datasets are located under **/users/utd/Hive\_Input/nasa** in the **Local** Unix System. Please use this folder and don’t copy to any other folder on the server.

**Requirement**:

Write hive 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***.)

* Create table nasa and load data from NASA\_HTTP.txt

**create table nasa (IP STRING,VALUE STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';**

**LOAD DATA local INPATH '/users/utd/Hive\_Input/nasa/NASA\_HTTP.txt' into table nasa;**

* Create table **host** and load data from **HOST\_COUNTRY**.txt

**create table host (IP STRING,COUNTRY\_ABBREVIATION STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';**

**LOAD DATA local INPATH '/users/utd/Hive\_Input/nasa/HOST\_COUNTRY.txt' into table host;**

* Create table **country** and load data from **COUNTRY\_NAME**.txt

**create table country(COUNTRY\_ABBREVIATION STRING, COUNTRY\_NAME STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';**

**LOAD DATA local INPATH '/users/utd/Hive\_Input/nasa/COUNTRY\_NAME.txt' into table country;**

* Join **nasa** and **host** by **IP** and **country** and **host** by **country\_abbreviation**

**select \* from nasa join host on (nasa.ip = host.ip) join country on (country.country\_abbreviation = host.country\_abbreviation);**

