Important Commands for BDM Lab Test

Part I: Utility Commands (filename.ext for instance thisissick.csv)

.....

1) pwd (Check current working directory)

2) cd .. (Go to parent directory)

3) ls (List files)

4) sudo apt-get update

5) sudo apt-get upgrade

6) sudo apt-get install openssh-server

7) echo $PATH

8) echo "text" >> filename.ext (append text to file)

9) nano .bashrc

10) touch filename.ext (create file)

11) source ~/.bashrc

12) cp /mnt/c/Users/"Seth Woo"/Downloads/filename.ext . (copy file from local Downloads folder to your Linux file system)

13) cat filename.ext (read contents of file)

14) rm filename.ext (delete file)

15) clear

.....

Part II: HDFS

.....

1) hdfs namenode -format (if Datanode disappears, stop-all.sh then delete the Datanode and Namenode files in the Data folder, then run this, followed by start-all.sh, and jps to verify)

2) start-all.sh

3) jps

4) hdfs dfs -mkdir /user/hdfs (create new directory on hdfs)

5) hdfs dfs -ls /user/hdfs/ (list files in directory on hdfs)

6) hdfs dfs -put filename.ext /user/hdfs/filename.ext (import file from local Linux file system to hdfs)

7) hdfs dfs -get /user/hdfs/filename.ext

8) hdfs dfs -cat /user/hdfs/filename.ext (read content of file in hdfs)

9) hdfs dfs -appendToFile filename.ext /user/hdfs/filename2.ext

10) hdfs dfs -rm /user/hdfs/filename.ext (delete file in hdfs)

11) hdfs dfs -rmdir /user/hdfs/ (remove directory on hdfs)

12) stop-all.sh

.....

Part III: MySQL (remove header from csv before importing file from local Linux file system to table in mysql, data loaded from local file system to mysql)

.....

1) mysql -uroot -proot (start mysql shell)

2) CREATE DATABASE WQD7007;

3) show databases;

3) Use WQD7007;

4) CREATE TABLE churn (customerID varchar(20), PaperlessBilling varchar(3), PaymentMethod varchar(30), MonthlyCharges numeric(8,2), Churn varchar(3));

5) show tables;

6) SET GLOBAL local_infile=ON;

7) mysql -uroot -proot --local_infile=1 <yourdatabase> -e "LOAD DATA LOCAL INFILE '~/filename.ext' INTO TABLE churn FIELDS TERMINATED BY ',"" (import file from local Linux file system to table in mysql)

8) SELECT * from churn;

9) DROP TABLE churn;

10) sqoop import -connect jdbc:mysql://localhost/WQD7007 -username root -password root -table churn -m 1 (import table in mysql to hdfs)

11) hdfs dfs -cat /user/sethwoo/churn/* (read contents of table in hdfs)

12) exit;

.....

Part IV: Hive (remove header from csv before copying file from local Downloads folder to your Linux file system, data is loaded from hdfs to hive)

.....

1) sudo service ssh start (run this command in the beginning just in case)

2) create database wqd7007;

3) show databases;

4) hdfs dfs -mkdir /user/hdfs/batting

5) hdfs dfs -put Batting.csv /user/hdfs/batting/

6) hdfs dfs -ls /user/hdfs/batting/

7) use wqd7007;

8) CREATE EXTERNAL TABLE IF NOT EXISTS batting(

playerID STRING, yearID INT, stint INT, teamID STRING, lgID STRING, G INT, G_batting INT,

AB INT, R INT, H INT, B2 INT, B3 INT, HR INT, RBI INT, SB INT, CS INT, BB INT, SO INT, IBB INT,

HBP INT, SH INT, SF INT, GIDP INT, G_old INT)

COMMEND 'Batting stats'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/hdfs/batting';

9) select * from batting;

10) select * from batting limit 5; (select first 5 rows from batting table)

11) SELECT yearID, max(R) FROM batting GROUP BY yearID;

12) SELECT a.yearID, a.playerID, a.R FROM batting a

JOIN (SELECT yearID year_ID, max(R) max_r FROM batting GROUP BY yearID) b

ON (a.yearID = b.year_ID AND a.R = b.max_r);

13) SELECT stddev(R) FROM batting; (standard deviation of R column)

14) SELECT yearID, stddev(R) FROM batting GROUP BY yearID;

15) SELECT min(R) FROM batting; (minimum of R column)

16) SELECT max(R) FROM batting; (maximum of R column)

17) exit;

.....

Part V: Pig (not necessary to remove header from csv before copying file from local Downloads folder to Linux file system, data is loaded from hdfs to pig)

.....

1) batting = load '/user/hdfs/batting/Batting.csv' using PigStorage(','); (import file from hdfs to structured text file or relation called batting, filename is case-sensitive ya)

2) raw_runs= FILTER batting BY $1>0; (exclude the header of the csv i.e. the first row)

3) DUMP raw_runs (run Pig Latin statements and display results on screen, note this command is not terminated by ;)

4) Runs = FOREACH raw_runs GENERATE $0 as playerID, $1 as year, $8 as runs; ($0 corresponds to the first column, column 0 in the original csv)

5) grp_data = GROUP Runs by (year);

6) max_runs = FOREACH grp_data GENERATE group as grp, MAX(Runs.runs) as max_runs;

7) DUMP max_runs

8) join_max_run = JOIN max_runs by ($0, max_runs), Runs by (year,runs);

9) join_data = FOREACH join_max_run GENERATE $0 as year, $2 as playerID, $1 as run;

10) DUMP join_data

11) quit

.....