PIG PRAC FILE

**GRUNT SHELL- IN LOCAL MODE**

**grunt> A = LOAD '/etc/passwd' using PigStorage(':');**

**grunt> B = FOREACH A GENERATE $0;**

**grunt> DUMP B;**

**grunt> A = LOAD '/etc/passwd' using PigStorage(':');**

**grunt> B = FOREACH A GENERATE $0 , $2;**

**grunt> STORE B INTO 'OUTPUT';**

**grunt> A = LOAD '/etc/passwd' using PigStorage(':');**

**grunt> B = FOREACH A GENERATE $0 , $2;**

**grunt> STORE B INTO 'OUTPUTNEW' using PigStorage('|');**

**grunt> iniData = LOAD '/home/gopalkrishna/PRAC/PIG/empdata.log' using PigStorage('|') as (cid:int,cname:chararray,cheadcnt:int,cloc:chararray);**

**grunt> filData = FILTER iniData BY cloc == 'Pune';**

**grunt> ordData = ORDER filData BY cname;**

**grunt> STORE filData INTO 'FILTEROUT';STORE ordData INTO 'ORDERBYOUT';**

**grunt> iniData = LOAD '/home/gopalkrishna/PRAC/PIG/empdata.log' using PigStorage('|') as (cid:int,cname:chararray,cheadcnt:int,cloc:chararray);**

**grunt> SPLIT iniData INTO A IF cheadcnt > 300000,B IF cloc == 'Pune';**

**grunt> STORE A INTO 'FIRSTSPLIT';STORE B INTO 'SECONDSPLIT';**

**NOTE: The Process Name of PIG is “RunJar”**

**gopalkrishna@ubuntu:~/PRAC/PIG$ jps**

**8907 DataNode**

**9379 NodeManager**

**9426 Jps**

**8790 NameNode**

**9248 ResourceManager**

**9064 SecondaryNameNode**

**8527 RunJar**

**gopalkrishna@ubuntu:~/PRAC/PIG$**

**NOTE: To Know the PIG Version:**

**gopalkrishna@ubuntu:~$ pig --version**

**Apache Pig version 0.15.0 (r1682971)**

**compiled Jun 01 2015, 11:44:35**

**gopalkrishna@ubuntu:~$**

**GRUNT SHELL- IN HDFS MODE**

**gopalkrishna@ubuntu:~$ start-all.sh**

**gopalkrishna@ubuntu:~$ mr-jobhistory-daemon.sh start historyserver**

**gopalkrishna@ubuntu:~$ jps**

**8907 DataNode**

**9379 NodeManager**

**10358 Jps**

**10322 JobHistoryServer**

**8790 NameNode**

**9248 ResourceManager**

**9064 SecondaryNameNode**

**gopalkrishna@ubuntu:~$**

**gopalkrishna@ubuntu:~$ hadoop fs -mkdir /piginput**

**gopalkrishna@ubuntu:~$**

**gopalkrishna@ubuntu:~$ hadoop fs -put /home/gopalkrishna/PRAC/PIG/empdata.log /piginput**

**gopalkrishna@ubuntu:~$**

**grunt> A = LOAD '/piginput/empdata.log' using PigStorage('|') as (cid:int,cname:chararray,cheadcnt:int,cloc:chararray);**

**grunt> B = GROUP A BY cloc;**

**grunt> STORE B INTO '/GROUPBYOUT';**

**grunt> A = LOAD '/piginput/empdata.log' using PigStorage('|') as (cid:int,cname:chararray,cheadcnt:int,cloc:chararray);**

**grunt> B = LIMIT A 3;**

**grunt> C = DISTINCT A;**

**grunt> STORE B INTO '/LIMITOUT';STORE C INTO '/DISTOUT';**

**SCRIPT MODE- IN LOCAL MODE**

**gopalkrishna@ubuntu:~/PRAC/PIG$ cat crossScript.pig**

**A = LOAD 'data1.log' using PigStorage('\t') as (id:int,name:chararray,sal:int);**

**B = LOAD 'data2.log' using PigStorage('\t') as (Id:int,Name:chararray,amt:int);**

**crsData = CROSS A , B;**

**STORE crsData INTO 'CROSS-OUT';**

**gopalkrishna@ubuntu:~/PRAC/PIG$**

**gopalkrishna@ubuntu:~/PRAC/PIG$ cat unionScript.pig**

**A = LOAD 'data1.log' using PigStorage('\t') as (id:int,name:chararray,sal:int);**

**B = LOAD 'data2.log' using PigStorage('\t') as (Id:int,Name:chararray,amt:int);**

**uniData = UNION A , B;**

**DUMP uniData;**

**gopalkrishna@ubuntu:~/PRAC/PIG$**

**gopalkrishna@ubuntu:~/PRAC/PIG$ cat joinScript.pig**

**A = LOAD 'data1.log' using PigStorage('\t') as (id:int,name:chararray,sal:int);**

**B = LOAD 'data2.log' using PigStorage('\t') as (Id:int,Name:chararray,amt:int);**

**JoinData = JOIN A BY id , B BY Id;**

**LJoin = JOIN A BY id LEFT , B BY Id;**

**RJoin = JOIN A BY id RIGHT , B BY Id;**

**FJoin = JOIN A BY id FULL , B BY Id;**

**--FilFJoinOut = FOREACH FJoin GENERATE id,amt;**

**--STORE FilFJoinOut INTO 'FULLOUT69NEW';**

**STORE JoinData INTO 'INNERJOIN-OUT';**

**STORE LJoin INTO 'LEFTJOIN-OUT';**

**STORE RJoin INTO 'RIGHTJOIN-OUT';**

**STORE FJoin INTO 'FULLJOIN-OUT';**

**gopalkrishna@ubuntu:~/PRAC/PIG$**

**PARAMETER SUBSTITION IN PIG**

**INPUT:**

**gopalkrishna@ubuntu:~/PRAC/PIG$ cat NYSE\_Data**

**NYSE CPO 1981-05-12 12.5 25.5 7.2 13.45 25 13.25**

**NYSE CPO 1982-06-12 11.5 28.5 6.2 11.56 43 12.75**

**NYSE DPO 1981-05-12 16.5 35.5 9.9 19.25 42 18.25**

**NYSE FPO 1984-02-27 12.5 22.5 6.2 13.11 31 14.25**

**NYSE GPO 1986-05-09 19.5 25.5 9.2 21.50 34 20.25**

**NYSE VPO 1981-05-12 22.5 55.5 5.2 27.27 18 28.25**

**gopalkrishna@ubuntu:~/PRAC/PIG$**

**SCRIPT:**

**gopalkrishna@ubuntu:~/PRAC/PIG$ cat nyse\_dailyTrade.pig**

**daily = LOAD 'NYSE\_Data' using PigStorage('\t') as (exchange:chararray,symbol:chararray,date:chararray,open:float,high:float,low:float,close:float,volume:int,adj\_close:float);**

**yesterday = FILTER daily BY date == '$DATE';**

**grpdata = GROUP yesterday ALL;**

**minmax = FOREACH grpdata GENERATE MAX(yesterday.high),MIN(yesterday.low);**

**STORE yesterday INTO 'NYSE-YESTERDAY-DATA';**

**STORE grpdata INTO 'NYSE-GRPDATA';**

**STORE minmax INTO 'NYSE-MINMAXDATA';**

**gopalkrishna@ubuntu:~/PRAC/PIG$**

**COMMAND TO RUN: pig -param "DATE=1981-05-12" -x local nyse\_dailyTrade.pig**

**To PIG Script with Parameter substitution to read a param file - below is the command**

**gopalkrishna@ubuntu:~/PRAC/PIG/RT-SCRIPT-RELATED$ pig -param\_file param.properties script.pig**