**PIG CONCEPT LABS**

**LAB 1. GROUPING INPUT DATA**

1. Copy Input Data to Hadoop fs -put /user/root

2.iniData = load 'empdata.log' using PigStorage ('|') as (<cid:int,cname:chararray,cheadcnt:int,cloc:chararray>);

3.grunt> grpData = Group iniData BY cloc;

4.grunt> STORE grpData INTO 'GROUPBY-OUTN';

5.root@ubuntu:/home/batch40/PIG# hadoop fs -ls /user/root/GROUPBY-OUTN

Found 2 items

drwxr-xr-x - root supergroup 0 2014-10-31 00:31 /user/root/GROUPBY-OUTN/\_logs

-rw-r--r-- 1 root supergroup 740 2014-10-31 00:31 /user/root/GROUPBY-OUTN/part-r-00000

***LAB 2.Split Commands used to split the data as per location***

grunt> iniData = LOAD '/PIG12/empdata.log' using PigStorage ('|') as (<cid:int,cname:chararray,cheadcnt:int,cloc:chararray>);

grunt> SPLIT iniData INTO A IF cloc == 'mumbai' , B IF cloc == 'hyderabad';

grunt> STORE A INTO '/FIRST-SPLITVIKAS'; STORE B INTO 'SECONDSPLITVIKAS' using PigStorage ('#');

root@ubuntu:/home/batch40/PIG# hadoop fs -ls /FIRST-SPLITVIKAS

Found 2 items

drwxr-xr-x - root supergroup 0 2014-10-31 01:05 /FIRST-SPLITVIKAS/\_logs

-rw-r--r-- 1 root supergroup 73 2014-10-31 01:05 /FIRST-SPLITVIKAS/part-m-00000

root@ubuntu:/home/batch40/PIG# hadoop fs -cat /FIRST-SPLITVIKAS/part-m-00000

10501 WiproTechnologies 178000 mumbai

10201 infosystechltd 740000 mumbai

root@ubuntu:/home/batch40/PIG# hadoop fs -ls /user/root/SECONDSPLITVIKAS

Found 2 items

drwxr-xr-x - root supergroup 0 2014-10-31 01:05 /user/root/SECONDSPLITVIKAS/\_logs

-rw-r--r-- 1 root supergroup 82 2014-10-31 01:05 /user/root/SECONDSPLITVIKAS/part-m-00000

root@ubuntu:/home/batch40/PIG# hadoop fs -cat /user/root/SECONDSPLITVIKAS/part-m-00000

10101#UnitedHealthGroup#300000#hyderabad

10501#WiproTechnologies#200000#hyderabad

root@ubuntu:/home/batch40/PIG#

***Input PIG Script for Grouping on Input Data***

1. Copy Input Data to Hadoop fs -put /user/root

2.iniData = load 'empdata.log' using PigStorage ('|') as (cid:int,cname:chararray,cheadcnt:int,cloc:chararray);

3.grunt> grpData = Group iniData BY cloc;

4.grunt> STORE grpData INTO 'GROUPBY-OUTN';

**Output :**

5.root@ubuntu:/home/batch40/PIG# hadoop fs -ls /user/root/GROUPBY-OUTN

Found 2 items

drwxr-xr-x - root supergroup 0 2014-10-31 00:31 /user/root/GROUPBY-OUTN/\_logs

-rw-r--r-- 1 root supergroup 740 2014-10-31 00:31 /user/root/GROUPBY-OUTN/part-r-00000

**Split Commands used to split the data as per location**

grunt> iniData = LOAD '/PIG12/empdata.log' using PigStorage ('|') as (cid:int,cname:chararray,cheadcnt:int,cloc:chararray);

grunt> SPLIT iniData INTO A IF cloc == 'mumbai' , B IF cloc == 'hyderabad';

grunt> STORE A INTO '/FIRST-SPLITVIKAS'; STORE B INTO 'SECONDSPLITVIKAS' using PigStorage ('#');

root@ubuntu:/home/batch40/PIG# hadoop fs -ls /FIRST-SPLITVIKAS

Found 2 items

drwxr-xr-x - root supergroup 0 2014-10-31 01:05 /FIRST-SPLITVIKAS/\_logs

-rw-r--r-- 1 root supergroup 73 2014-10-31 01:05 /FIRST-SPLITVIKAS/part-m-00000

root@ubuntu:/home/batch40/PIG# hadoop fs -cat /FIRST-SPLITVIKAS/part-m-00000

10501 WiproTechnologies 178000 mumbai

10201 infosystechltd 740000 mumbai

root@ubuntu:/home/batch40/PIG# hadoop fs -ls /user/root/SECONDSPLITVIKAS

Found 2 items

drwxr-xr-x - root supergroup 0 2014-10-31 01:05 /user/root/SECONDSPLITVIKAS/\_logs

-rw-r--r-- 1 root supergroup 82 2014-10-31 01:05 /user/root/SECONDSPLITVIKAS/part-m-00000

root@ubuntu:/home/batch40/PIG# hadoop fs -cat /user/root/SECONDSPLITVIKAS/part-m-00000

10101#UnitedHealthGroup#300000#hyderabad

10501#WiproTechnologies#200000#hyderabad

root@ubuntu:/home/batch40/PIG#

-------------------------------------------------------------------------

**/home/batch40/PIG/nano SampleScript.pig**

A = LOAD '/home/batch40/PIG/data1.log' using PigStorage('\t') as (id:int,name:chararray,sal:int);

B = FOREACH A GENERATE id,sal + 2000;

C = ORDER B BY id ;

DUMP C;

Output :

(1000,14000)

(1001,16000)

(1002,18000)

(1003,20000)

(1004,22000)

---------------------------------------------------

Ex : /home/batch40/PIG/SampleSaveScript.pig

A = LOAD '/home/batch40/PIG/data1.log' using PigStorage('\t') as (id:int,name:chararray,sal:int);

B = FOREACH A GENERATE id,sal + 2000;

C = ORDER B BY id ;

STORE C INTO 'Test';

root@ubuntu:/home/batch40/PIG# cd Test

root@ubuntu:/home/batch40/PIG/Test# ll

total 20

drwxr-xr-x 2 root root 4096 2014-10-31 08:55 ./

drwxr-xr-x 10 root root 4096 2014-10-31 08:55 ../

-rwxrwxrwx 1 root root 55 2014-10-31 08:55 part-r-00000\*

-rw-r--r-- 1 root root 12 2014-10-31 08:55 .part-r-00000.crc

-rwxrwxrwx 1 root root 0 2014-10-31 08:55 \_SUCCESS\*

-rw-r--r-- 1 root root 8 2014-10-31 08:55 .\_SUCCESS.crc

root@ubuntu:/home/batch40/PIG/Test# cat part-r-00000

1000 14000

1001 16000

1002 18000

1003 20000

1004 22000

root@ubuntu:/home/batch40/PIG/Test#

-----------------------------------------

**extrascript.pig**

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

--DESCRIBE A;

ILLUSTRATE A;

------------------------------------------------------------

| A | id: bytearray | name: bytearray | sal: bytearray |

------------------------------------------------------------

| | 1001 | DEF | 14000 |

------------------------------------------------------------

------------------------------------------------

| A | id: int | name: chararray | sal: int |

------------------------------------------------

| | 1001 | DEF | 14000 |

------------------------------------------------

----------------------------------------------------------------------------

**Flatten Script:**

root@ubuntu:/home/batch40/PIG# cat flattenscript.pig

A = LOAD '/PIGFlatten/empdata.log' using PigStorage('|') as (cid:int,name:chararray,cheadcnt:int,cloc:chararray);

B = GROUP A BY cloc;

C = FOREACH B GENERATE group , FLATTEN(A);

STORE C INTO '/FLATTENB';

root@ubuntu:/home/batch40/PIG# hadoop fs -ls /FLATTENB

Found 2 items

drwxr-xr-x - root supergroup 0 2014-10-31 23:24 /FLATTENB/\_logs

-rw-r--r-- 1 root supergroup 763 2014-10-31 23:24 /FLATTENB/part-r-00000

root@ubuntu:/home/batch40/PIG# hadoop fs -cat /FLATTENB/part-r-00000

pune 10101 UnitedHealthGroup 320000 pune

pune 10201 infosystechltd 800000 pune

pune 10410 Cognizant Technology Solutions 123000 pune

mumbai 10201 infosystechltd 740000 mumbai

mumbai 10501 WiproTechnologies 178000 mumbai

chennai 10410 Cognizant Technology Solutions 200000 chennai

chennai 10301 TataConsultancyLimited 320000 chennai

chennai 10301 TataConsultancyLimited 330000 chennai

chennai 10101 UnitedHealthGroup 400000 chennai

kolkatta 10301 TataConsultancyLimited 100000 kolkatta

bangalore 10201 infosystechltd 700000 bangalore

bangalore 10501 WiproTechnologies 1000 bangalore

bangalore 10410 Cognizant Technology Solutions 150000 bangalore

hyderabad 10101 UnitedHealthGroup 300000 hyderabad

hyderabad 10501 WiproTechnologies 200000 hyderabad

-------------------------------------------------------------------------

**INNER JOIN**

root@ubuntu:/home/batch40/PIG# cat JoinS.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);

IJoin = JOIN A BY id , B BY id;

STORE IJoin INTO 'INNERJOIN-OUT14';

root@ubuntu:/home/batch40/PIG/INNERJOIN-OUT14# cat part-r-00000

1000 ABC 12000 1000 ABC 12000

1001 DEF 14000 1001 DEF 14000

root@ubuntu:/home/batch40/PIG/INNERJOIN-OUT14# cd ..

root@ubuntu:/home/batch40/PIG# cat data1.log

1000 ABC 12000

1001 DEF 14000

1002 XYZ 16000

1003 PQR 18000

1004 MNO 20000

root@ubuntu:/home/batch40/PIG# cat data2.log

1000 ABC 12000

1001 DEF 14000

2002 TTT 16000

3003 PPP 18000

4004 GGG 20000

root@ubuntu:/home/batch40/PIG#

-------------------------------------------------------------------------

root@ubuntu:/home/batch40/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,sal:int);

uniData = UNION A, B;

--STORE uniData INTO 'UNION-OUT12';

DUMP uniData;

(1000,ABC,12000)

(1001,DEF,14000)

(1002,XYZ,16000)

(1003,PQR,18000)

(1004,MNO,20000)

(1000,ABC,12000)

(1001,DEF,14000)

(2002,TTT,16000)

(3003,PPP,18000)

(4004,GGG,20000)

root@ubuntu:/home/batch40/PIG# cat data1.log

1000 ABC 12000

1001 DEF 14000

1002 XYZ 16000

1003 PQR 18000

1004 MNO 20000

root@ubuntu:/home/batch40/PIG# cat data2.log

1000 ABC 12000

1001 DEF 14000

2002 TTT 16000

3003 PPP 18000

4004 GGG 20000

root@ubuntu:/home/batch40/PIG#

-------------------------------------------------------------------

**LEFT-RIGHT-FULL JOIN**

root@ubuntu:/home/batch40/PIG# cat JoinLRF.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,sal:int);

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;

STORE LJoin INTO 'LEFT-JOIN';

STORE RJoin INTO 'RIGHT-JOIN';

STORE FJoin INTO 'FJOIN';

Input(s):

Successfully read records from: "file:///home/batch40/PIG/data2.log"

Successfully read records from: "file:///home/batch40/PIG/data1.log"

Output(s):

Successfully stored records in: "file:///home/batch40/PIG/FJOIN"

Successfully stored records in: "file:///home/batch40/PIG/RIGHT-JOIN"

Successfully stored records in: "file:///home/batch40/PIG/LEFT-JOIN"

root@ubuntu:/home/batch40/PIG/FJOIN# cat part-r-00000

1000 ABC 12000 1000 ABC 12000

1001 DEF 14000 1001 DEF 14000

1002 XYZ 16000

1003 PQR 18000

1004 MNO 20000

2002 TTT 16000

3003 PPP 18000

4004 GGG 20000

root@ubuntu:/home/batch40/PIG/FJOIN# cd ..

root@ubuntu:/home/batch40/PIG# cd LEFT-JOIN

root@ubuntu:/home/batch40/PIG/LEFT-JOIN# cat part-r-0000

cat: part-r-0000: No such file or directory

root@ubuntu:/home/batch40/PIG/LEFT-JOIN# cat part-r-00000

1000 ABC 12000 1000 ABC 12000

1001 DEF 14000 1001 DEF 14000

1002 XYZ 16000

1003 PQR 18000

1004 MNO 20000

root@ubuntu:/home/batch40/PIG/LEFT-JOIN# cd ..

root@ubuntu:/home/batch40/PIG# cd RIGHT-JOIN

root@ubuntu:/home/batch40/PIG/RIGHT-JOIN# cat part-r-00000

1000 ABC 12000 1000 ABC 12000

1001 DEF 14000 1001 DEF 14000

2002 TTT 16000

3003 PPP 18000

4004 GGG 20000

root@ubuntu:/home/batch40/PIG/RIGHT-JOIN#

-----------------------------------------------------

**tokenscript.pig**

root@ubuntu:/home/batch40/PIG# hadoop fs -mkdir /PIGT

root@ubuntu:/home/batch40/PIG# nano tokendata.log

root@ubuntu:/home/batch40/PIG# nano tokendata.log

root@ubuntu:/home/batch40/PIG# cat tokenscript.pig

A = LOAD '/PIGT/tokendata.log' as (line:chararray);

B = FOREACH A GENERATE TOKENIZE(line);

DUMP B;

root@ubuntu:/home/batch40/PIG#

root@ubuntu:/home/batch40/PIG# pig tokenscript.pig

Job DAG:

job\_201410160209\_0001

2014-11-01 06:00:44,261 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2014-11-01 06:00:44,334 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2014-11-01 06:00:44,334 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

({(Hadoop),(is),(one),(of),(the),(bigdata),(tool)})

-------------------------------------------------------------------------

**PIG WORD COUNT EXAMPLE USING PIG SCRIPT**

root@ubuntu:/home/batch40/PIG# cat wordcountpigscript.pig

inputdata = load '/user/root/input.txt' as (line:chararray);

words = FOREACH inputdata GENERATE FLATTEN(TOKENIZE(line)) AS word;

filtered\_words = FILTER words BY word MATCHES '\\w+';

word\_groups = GROUP filtered\_words BY word;

word\_count = FOREACH word\_groups GENERATE group AS word , COUNT(filtered\_words) AS count;

ordered\_word\_count = ORDER word\_count BY count DESC;

DUMP words;

DUMP filtered\_words;

DUMP word\_groups;

--STORE ordered\_word\_count INTO 'PIG30/wcountputInPig';

DUMP ordered\_word\_count;

root@ubuntu:/home/batch40/PIG#

root@ubuntu:/home/batch40/PIG# hadoop fs -put input.txt /user/root