1.

root@ubuntu:/home/mrinmoy# cat > f1

1233 tcs 32442 hyd

3746 inf 23244 ban

9574 cts 84743 mum

6453 acc 63535 del

root@ubuntu:/home/mrinmoy# pig -x local

2015-02-06 21:15:53,157 [main] INFO org.apache.pig.Main - Logging error messages to: /home/mrinmoy/pig\_1423286153155.log

2015-02-06 21:15:53,410 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///

grunt> A =LOAD 'f1' using PigStorage(' ');

grunt> B =FOREACH A GENERATE $1,$3;

grunt> STORE B INTO 'FIL1';

Input(s):

Successfully read records from: "file:///home/mrinmoy/f1"

Output(s):

Successfully stored records in: "file:///home/mrinmoy/FIL1"

grunt> quit

root@ubuntu:/home/mrinmoy# cat FIL1

cat: FIL1: Is a directory

root@ubuntu:/home/mrinmoy# cat FIL1/part-m-00000

tcs hyd

inf ban

cts mum

acc del

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

2.

grunt> A =LOAD 'f1' using PigStorage(' ');

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

grunt> STORE B INTO 'FIL2';

Input(s):

Successfully read records from: "file:///home/mrinmoy/f1"

Output(s):

Successfully stored records in: "file:///home/mrinmoy/FIL2"

grunt> quit

root@ubuntu:/home/mrinmoy# cat FIL2/part-m-00000

1233 32442

3746 23244

9574 84743

6453 63535

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

3. FILTER

grunt> A =LOAD 'f1' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> fildata =FILTER A BY cloc=='hyd';

grunt> STORE fildata INTO 'FIL3';

Input(s):

Successfully read records from: "file:///home/mrinmoy/f1"

Output(s):

Successfully stored records in: "file:///home/mrinmoy/FIL3"

grunt> quit

root@ubuntu:/home/mrinmoy# cat FIL3/part-m-00000

1233 tcs 32442 hyd

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

4. ORDER

grunt> A =LOAD 'f1' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> orddata =ORDER A BY cname;

grunt> STORE orddata INTO 'fil4';

grunt> quit

root@ubuntu:/home/mrinmoy# cat fil4/part-m-00000

cat: fil4/part-m-00000: No such file or directory

root@ubuntu:/home/mrinmoy# cat fil4/part-r-00000

6453 acc 63535 del

9574 cts 84743 mum

3746 inf 23244 ban

1233 tcs 32442 hyd

root@ubuntu:/home/mrinmoy# cat f1

2323|kiran|45000|rjy

6453|kumar|42000|hyd

6452|phani|35000|ong

7463|karun|30000|ngl

---------------------SPLIT--------------------

grunt> A =load 'f1' using PigStorage('|') as(id:int,name:chararray,sal:int,loc:chararray);

grunt> SPLIT A INTO c IF sal>40000,D IF id==2323; //HERE WE HAVE TO GIVE MINIMUM 2 COMPARISIONS IN SPLIT

grunt> DUMP c;DUMP D;

OUTPUT:

(2323,kiran,45000,rjy)

grunt> A =load 'f1' using PigStorage('|') as(id:int,name:chararray,sal:int,loc:chararray);

grunt> SPLIT A INTO c IF sal>40000,D IF id==2325;

grunt> DUMP c;DUMP D;

OUTPUT: (nothing)

grunt> A =load 'f1' using PigStorage('|') as(id:int,name:chararray,sal:int,loc:chararray);

grunt> SPLIT A INTO c IF sal>40000,D IF id>6000;

grunt> DUMP c;DUMP D;

OUTPUT:

(6453,kumar,42000,hyd)

(6452,phani,35000,ong)

(7463,karun,30000,ngl)

----------------------------DISTINCT-------------------

cat f3

2323|kiran|45000|rjy

2323|kiran|45000|rjy

6452|phani|35000|ong

7463|karun|30000|ngl

root@ubuntu:/home/mrinmoy# pig -x local

grunt> A =load 'f3' using PigStorage('|') as(id:int,name:chararray,sal:int,loc:chararray);

grunt> B =DISTINCT A;

grunt> DUMP B;

OUTPUT :

(2323,kiran,45000,rjy)

(6452,phani,35000,ong)

(7463,karun,30000,ngl)

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

5. DISTINCT & LIMIT

grunt> A =LOAD 'f1' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> dist =DISTINCT A;

grunt> lim =LIMIT A 3;

grunt> STORE dist INTO 'FIL5.in';STORE lim INTO 'FIL6.in';

grunt> quit

root@ubuntu:/home/mrinmoy# cat FIL5.in/part-r-00000

1233 tcs 32442 hyd

3746 inf 23244 ban

6453 acc 63535 del

9574 cts 84743 mum

root@ubuntu:/home/mrinmoy# cat FIL6.in/part-r-00000

1233 tcs 32442 hyd

3746 inf 23244 ban

9574 cts 84743 mum

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

6. GROUPING

grunt> A =LOAD 'f1' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> B =GROUP A BY cloc;

grunt> STORE B INTO 'group1';

grunt> quit;

root@ubuntu:/home/mrinmoy# cat group1/part-m-00000

cat: group1/part-m-00000: No such file or directory

root@ubuntu:/home/mrinmoy# cat group1/part-r-00000

ban {(3746,inf,23244,ban)}

del {(6453,acc,63535,del)}

hyd {(1233,tcs,32442,hyd)}

mum {(9574,cts,84743,mum)}

------------------------------------------JOININNGS-----------------

root@ubuntu:/home/mrinmoy# cat f1

2323|kiran|45000|rjy

6453|kumar|42000|hyd

6452|phani|35000|ong

7463|karun|30000|ngl

root@ubuntu:/home/mrinmoy# cat f2

2323|ramesh|25000|rjy

7464|mahesh|27000|kkd

3762|rajesh|25000|mmd

7463|suresh|18000|sec

grunt> file1 =LOAD 'f1' using PigStorage('|') as (aid:int,aname:chararray,asal:int,aloc:chararray);

grunt> file2 =LOAD 'f2' using PigStorage('|') as (bid:int,bname:chararray,bsal:int,bloc:chararray);

grunt> joindata =JOIN file1 BY aid,file2 BY bid;

grunt> dump joindata;

(2323,kiran,45000,rjy,2323,ramesh,25000,rjy)

(7463,karun,30000,ngl,7463,suresh,18000,sec)

grunt> lj =JOIN file1 BY aid LEFT,file2 BY bid; //Left outer means left table comes as it is

grunt> STORE lj INTO 'leftj';

root@ubuntu:/home/mrinmoy# ls leftj;

part-r-00000 \_SUCCESS

root@ubuntu:/home/mrinmoy# cat leftj/part-r-00000

2323 kiran 45000 rjy 2323 ramesh 25000 rjy

6452 phani 35000 ong

6453 kumar 42000 hyd

7463 karun 30000 ngl 7463 suresh 18000 sec

grunt> rj =JOIN file1 BY aid RIGHT,file2 BY bid; // Right join means right table comes as it is

grunt> STORE rj INTO 'rightj';

root@ubuntu:/home/mrinmoy# ls rightj

part-r-00000 \_SUCCESS

root@ubuntu:/home/mrinmoy# cat rightj/part-r-00000

2323 kiran 45000 rjy 2323 ramesh 25000 rjy

3762 rajesh 25000 mmd

7463 karun 30000 ngl 7463 suresh 18000 sec

7464 mahesh 27000 kkd

grunt> fj =JOIN file1 BY aid FULL,file2 BY bid;

grunt> STORE fj INTO 'fullj';

root@ubuntu:/home/mrinmoy# cat leftj/part-r-00000

2323 kiran 45000 rjy 2323 ramesh 25000 rjy

6452 phani 35000 ong

6453 kumar 42000 hyd

7463 karun 30000 ngl 7463 suresh 18000 sec

root@ubuntu:/home/mrinmoy# ls fullj //Full join means both tables come as they are

part-r-00000 \_SUCCESS

root@ubuntu:/home/mrinmoy# cat fullj/part-r-00000

2323 kiran 45000 rjy 2323 ramesh 25000 rjy

3762 rajesh 25000 mmd

6452 phani 35000 ong

6453 kumar 42000 hyd

7463 karun 30000 ngl 7463 suresh 18000 sec

7464 mahesh 27000 kkd

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

7. CROSSING

root@ubuntu:/home/mrinmoy# cat f1

1233 tcs 32442 hyd

3746 inf 23244 ban

9574 cts 84743 mum

6453 acc 63535 del

root@ubuntu:/home/mrinmoy# cat > f2

64526 abc 7463 RJY

85742 gdt 6453 KKD

78746 try 7836 AMP

48347 jyf 7391 DWM

root@ubuntu:/home/mrinmoy# pig -x local

2015-02-06 22:20:28,871 [main] INFO org.apache.pig.Main - Logging error messages to: /home/mrinmoy/pig\_1423290028869.log

2015-02-06 22:20:29,020 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///

grunt> fdata =LOAD 'f1' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> sdata =LOAD 'f2' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> out =CROSS fdata,sdata;

grunt> STORE out INTO 'fil7.in';

root@ubuntu:/home/mrinmoy# cat fil7.in/part-r-00000

3746 inf 23244 ban 64526 abc 7463 RJY

3746 inf 23244 ban 48347 jyf 7391 DWM

3746 inf 23244 ban 85742 gdt 6453 KKD

3746 inf 23244 ban 78746 try 7836 AMP

6453 acc 63535 del 64526 abc 7463 RJY

6453 acc 63535 del 48347 jyf 7391 DWM

6453 acc 63535 del 85742 gdt 6453 KKD

6453 acc 63535 del 78746 try 7836 AMP

9574 cts 84743 mum 64526 abc 7463 RJY

9574 cts 84743 mum 48347 jyf 7391 DWM

9574 cts 84743 mum 85742 gdt 6453 KKD

9574 cts 84743 mum 78746 try 7836 AMP

1233 tcs 32442 hyd 64526 abc 7463 RJY

1233 tcs 32442 hyd 48347 jyf 7391 DWM

1233 tcs 32442 hyd 85742 gdt 6453 KKD

1233 tcs 32442 hyd 78746 try 7836 AMP

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

8. UNION

grunt> fdata =LOAD 'f1' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> sdata =LOAD 'f2' using PigStorage(' ') as (cid:int,cname:chararray,cemp:int,cloc:chararray);

grunt> unidata =UNION fdata,sdata;

grunt> DUMP *unidata;*

-------------------------UNION----------------

root@ubuntu:/home/mrinmoy# cat f1

1234|tata |4567|hyderabad

2354|wipro |4536|banglore

1245|accen |2657|chennai

1234|cts |4536|mumbai

8357|ibm |3272|delhi

5436|tech |3425|banglore

root@ubuntu:/home/mrinmoy# cat f1

1234|tata |4567|hyderabad

2354|wipro |4536|banglore

1245|accen |2657|chennai

1234|cts |4536|mumbai

8357|ibm |3272|delhi

5436|tech |3425|banglore

grunt> A =load 'f1' using PigStorage('|') as(aid:int,aname:chararray,aemp:int,aloc:chararray);

grunt> B =load 'f2' using PigStorage('|') as(bid:int,bname:chararray,bemp:int,bloc:chararray);

grunt> uni =UNION A,B;

grunt> dump uni;

(1234,tata ,4567,hyderabad)

(2354,wipro ,4536,banglore)

(1245,accen ,4567,chennnai)

(1234,cts ,4536,mumbai)

(8357,ibm ,3272,delhi)

(5436,tata ,3425,banglore)

(1234,tata ,4567,hyderabad)

(2354,wipro ,4536,banglore)

(1245,accen ,2657,chennai)

(1234,cts ,4536,mumbai)

(8357,ibm ,3272,delhi)

(5436,tech ,3425,banglore)

---------------------------------DESCRIBE-----------------

grunt> A =load 'f1' using PigStorage('|') as(aid:int,aname:chararray,aemp:int,aloc:chararray);

grunt> DESCRIBE A;

A: {aid: int,aname: chararray,aemp: int,aloc: chararray}

-------------------------CO GROUP---------------------- //same as group but performance is better than group

root@ubuntu:/home/mrinmoy# pig -x local

grunt> A =LOAD 'f1' using PigStorage('|') as (id:int,name:chararray,emp:int,loc:chararray);

grunt> B =COGROUP A BY name;

grunt> STORE B INTO 'cogrp';

Input(s):

Successfully read records from: "file:///home/mrinmoy/f1"

Output(s):

Successfully stored records in: "file:///home/mrinmoy/cogrp"

oot@ubuntu:/home/mrinmoy# ls cogrp

part-r-00000 \_SUCCESS

root@ubuntu:/home/mrinmoy# cat cogrp/part-r-00000

Accenture {(3734,Accenture ,337328,Banglore)}

Tech mahindra {(3636,Tech mahindra ,346647,Hyderabad)}

cognizant technology services {(3526,cognizant technology services ,343832,Chennai)}

international business machine {(3421,international business machine ,236312,Hyderabad)}

tata cosultancy services {(1243,tata cosultancy services ,193373,Banglore)}

wipro technologies {(7463,wipro technologies ,343621,Mumbai)}

//PIG IN SCRIPT MODE INPUT FROM LOCAL

root@ubuntu:/home/mrinmoy# cat f1

1243|tata cosultancy services |193373|Banglore

3421|international business machine |236312|Hyderabad

7463|wipro technologies |343621|Mumbai

3526|cognizant technology services |343832|Chennai

3734|Accenture |337328|Banglore

3636|Tech mahindra |346647|Hyderabad

root@ubuntu:/home/mrinmoy# nano first

root@ubuntu:/home/mrinmoy# pig -x local first;

Input(s):

Successfully read records from: "file:///home/mrinmoy/f1"

Output(s):

Successfully stored records in: "file:/tmp/temp-1126901759/tmp1842937463"

Job DAG:

job\_local\_0001

2015-02-11 21:24:58,593 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2015-02-11 21:24:58,619 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2015-02-11 21:24:58,619 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(tata cosultancy services ,193373)

(international business machine ,236312)

(wipro technologies ,343621)

(cognizant technology services ,343832)

(Accenture ,337328)

(Tech mahindra ,346647)

//PIG in script mode and input from hdfs

root@ubuntu:/home/mrinmoy# hadoop fs -put f1 jbk

root@ubuntu:/home/mrinmoy# hadoop fs -cat jbk/f1

1243|tata cosultancy services |193373|Banglore

3421|international business machine |236312|Hyderabad

7463|wipro technologies |343621|Mumbai

3526|cognizant technology services |343832|Chennai

3734|Accenture |337328|Banglore

3636|Tech mahindra |346647|Hyderabad

root@ubuntu:/home/mrinmoy# nano first /// IN NANO WE HAVE WRITE PIG COMMANDS LOAD, FOREACH,DUMP>..ETC

root@ubuntu:/home/mrinmoy# pig first;

Input(s):

Successfully read 6 records (673 bytes) from: "/user/root/jbk/f1"

Output(s):

Successfully stored 6 records (276 bytes) in: "hdfs://localhost:8020/tmp/temp668233278/tmp-157222441"

Counters:

Total records written : 6

Total bytes written : 276

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job\_201502112108\_0003

2015-02-11 21:31:00,882 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2015-02-11 21:31:00,899 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2015-02-11 21:31:00,899 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(tata cosultancy services ,193373)

(international business machine ,236312)

(wipro technologies ,343621)

(cognizant technology services ,343832)

(Accenture ,337328)

(Tech mahindra ,346647)

-----------------------COUNT,MAX,MIN---------------------- // IN PIG ONLY ROW OPERATIONS WILL DONE IN HIVE WE CAN MAKE COLUMN OPERATIONS

grunt> A =LOAD 'f2' using PigStorage('|') as (id:int,name:chararray,emp:int,loc:chararray);

grunt> B =GROUP A BY id;

grunt> DUMP B ;

(1,{(1,tata cosultancy services ,10,Banglore)})

(2,{(2,international business machine ,11,Hyderabad)})

(3,{(3,wipro technologies ,12,Mumbai)})

(4,{(4,cognizant technology services ,12,Chennai)})

(5,{(5,Accenture ,13,Banglore),(5,Tech mahindra ,13,Hyderabad)})

(7,{(7,wipro technologies ,14,Hyderabad)})

grunt> count10 =FOREACH B GENERATE group,COUNT(A.emp);

grunt> DUMP count10;

(1,1)

(2,1)

(3,1)

(4,1)

(5,2)

(7,1)

grunt> A =LOAD 'f2' using PigStorage('|') as (id:int,name:chararray,emp:int,loc:chararray);

grunt> B =GROUP A BY name;

grunt> DUMP B;

(Accenture ,{(5,Accenture ,13,Banglore)})

(Tech mahindra ,{(5,Tech mahindra ,13,Hyderabad)})

(cognizant technology services ,{(4,cognizant technology services ,12,Chennai)})

(international business machine ,{(2,international business machine ,11,Hyderabad)})

(tata cosultancy services ,{(1,tata cosultancy services ,10,Banglore)})

(wipro technologies ,{(3,wipro technologies ,12,Mumbai),(7,wipro technologies ,14,Hyderabad)})

grunt> A =LOAD 'f2' using PigStorage('|') as (id:int,name:chararray,emp:int,loc:chararray);

grunt> B =GROUP A BY name;

grunt> max10 =FOREACH B GENERATE group,MAX(A.emp);

grunt> DUMP max10;

(Accenture ,13)

(Tech mahindra ,13)

(cognizant technology services ,12)

(international business machine ,11)

(tata cosultancy services ,10)

(wipro technologies ,14)

grunt> A =LOAD 'f2' using PigStorage('|') as (id:int,name:chararray,emp:int,loc:chararray);

grunt> B =GROUP A BY name;

grunt> min10 =FOREACH B GENERATE group,MIN(A.emp);

grunt> DUMP min10;

(Accenture ,13)

(Tech mahindra ,13)

(cognizant technology services ,12)

(international business machine ,11)

(tata cosultancy services ,10)

(wipro technologies ,12)

untu:/home/mrinmoy# pig -x local

2015-02-13 05:29:38,309 [main] INFO org.apache.pig.Main - Logging error messages to: /home/mrinmoy/pig\_1423834178307.log

2015-02-13 05:29:38,545 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///

grunt> A =load 'f1' using PigStorage('|') as(aid:int,aname:chararray,aemp:int,aloc:chararray);

grunt> ILLUSTRATE A;

2015-02-13 05:29:57,121 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///

2015-02-13 05:29:57,302 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2015-02-13 05:29:57,306 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

2015-02-13 05:29:57,325 [main] WARN org.apache.hadoop.io.compress.snappy.LoadSnappy - Snappy native library is available

2015-02-13 05:29:57,326 [main] INFO org.apache.hadoop.util.NativeCodeLoader - Loaded the native-hadoop library

2015-02-13 05:29:57,326 [main] INFO org.apache.hadoop.io.compress.snappy.LoadSnappy - Snappy native library loaded

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

| A | aid: bytearray | aname: bytearray | aemp: bytearray | aloc: bytearray |

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

| | 1245 | accen | 2657 | chennai |

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

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

| A | aid: int | aname: chararray | aemp: int | aloc: chararray |

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

| | 1245 | accen | 2657 | chennai |

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

grunt> A =load 'f1' using PigStorage('|') as(aid:int,aname:chararray,aemp:int,aloc:chararray);

grunt> EXPLAIN A;

2015-02-13 05:30:15,805 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - pig.usenewlogicalplan is set to true. New logical plan will be used.

#-----------------------------------------------

# Logical Plan:

#-----------------------------------------------

fake: Store 1-24 Schema: {aid: int,aname: chararray,aemp: int,aloc: chararray} Type: Unknown

|

|---A: Load 1-23 Schema: {aid: int,aname: chararray,aemp: int,aloc: chararray} Type: bag

#-----------------------------------------------

# New Logical Plan:

#-----------------------------------------------

fake: (Name: LOStore Schema: aid#9:int,aname#10:chararray,aemp#11:int,aloc#12:chararray)

|

|---A: (Name: LOForEach Schema: aid#9:int,aname#10:chararray,aemp#11:int,aloc#12:chararray)

| |

| (Name: LOGenerate[false,false,false,false] Schema: aid#9:int,aname#10:chararray,aemp#11:int,aloc#12:chararray)ColumnPrune:InputUids=[9, 10, 11, 12]ColumnPrune:OutputUids=[9, 10, 11, 12]

| | |

| | (Name: Cast Type: int Uid: 9)

| | |

| | |---aid:(Name: Project Type: bytearray Uid: 9 Input: 0 Column: 0)

| | |

| | (Name: Cast Type: chararray Uid: 10)

| | |

| | |---aname:(Name: Project Type: bytearray Uid: 10 Input: 1 Column: 0)

| | |

| | (Name: Cast Type: int Uid: 11)

| | |

| | |---aemp:(Name: Project Type: bytearray Uid: 11 Input: 2 Column: 0)

| | |

| | (Name: Cast Type: chararray Uid: 12)

| | |

| | |---aloc:(Name: Project Type: bytearray Uid: 12 Input: 3 Column: 0)

| |

| |---(Name: LOInnerLoad[0] Schema: aid#9:bytearray)

| |

| |---(Name: LOInnerLoad[1] Schema: aname#10:bytearray)

| |

| |---(Name: LOInnerLoad[2] Schema: aemp#11:bytearray)

| |

| |---(Name: LOInnerLoad[3] Schema: aloc#12:bytearray)

|

|---A: (Name: LOLoad Schema: aid#9:bytearray,aname#10:bytearray,aemp#11:bytearray,aloc#12:bytearray)RequiredFields:null

#-----------------------------------------------

# Physical Plan:

#-----------------------------------------------

A: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-14

|

|---A: New For Each(false,false,false,false)[bag] - scope-13

| |

| Cast[int] - scope-2

| |

| |---Project[bytearray][0] - scope-1

| |

| Cast[chararray] - scope-5

| |

| |---Project[bytearray][1] - scope-4

| |

| Cast[int] - scope-8

| |

| |---Project[bytearray][2] - scope-7

| |

| Cast[chararray] - scope-11

| |

| |---Project[bytearray][3] - scope-10

|

|---A: Load(file:///home/mrinmoy/f1:PigStorage('|')) - scope-0

2015-02-13 05:30:15,935 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false

2015-02-13 05:30:15,959 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1

2015-02-13 05:30:15,959 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1

#--------------------------------------------------

# Map Reduce Plan

#--------------------------------------------------

MapReduce node scope-15

Map Plan

A: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-14

|

|---A: New For Each(false,false,false,false)[bag] - scope-13

| |

| Cast[int] - scope-2

| |

| |---Project[bytearray][0] - scope-1

| |

| Cast[chararray] - scope-5

| |

| |---Project[bytearray][1] - scope-4

| |

| Cast[int] - scope-8

| |

| |---Project[bytearray][2] - scope-7

| |

| Cast[chararray] - scope-11

| |

| |---Project[bytearray][3] - scope-10

|

|---A: Load(file:///home/mrinmoy/f1:PigStorage('|')) - scope-0--------

Global sort: false

----------------------------------------TOKENIZE----------------

/home/mrinmoy# cat ff

HI how are you

good morning

good evening

good night

grunt> A =LOAD 'ff' using PigStorage('\n') as (line:chararray);

grunt> B =FOREACH A GENERATE TOKENIZE(line);

grunt> dump B;

({(HI),(how),(are),(you)})

({(good),(morning)})

({(good),(evening)})

({(good),(night)})

-------------------------------------FLATTEN--------------------

root@ubuntu:/home/mrinmoy# cat f1

1234|tata |4567|hyderabad

2354|wipro |4536|banglore

1245|accen |2657|chennai

1234|cts |4536|mumbai

8357|ibm |3272|delhi

5436|tech |3425|banglore

grunt> A =load 'f1' using PigStorage('|') as(aid:int,aname:chararray,aemp:int,aloc:chararray);

grunt> gprdata =GROUP A BY aloc;

grunt> FLAT =FOREACH gr GENERATE FLATTEN(A);

grunt> DUMP FLAT;

(8357,ibm ,3272,delhi)

(1234,cts ,4536,mumbai)

(1245,accen ,2657,chennai)

(2354,wipro ,4536,banglore)

(5436,tech ,3425,banglore)

(1234,tata ,4567,hyderabad)

grunt> A =load 'f1' using PigStorage('|') as(aid:int,aname:chararray,aemp:int,aloc:chararray);

grunt> gr =GROUP A BY aloc;

grunt> FLAT =FOREACH gr GENERATE group,FLATTEN(A); //HERE WE HAVE ADDED GROUP NAME ALSO..

grunt> dump FLAT;

(delhi,8357,ibm ,3272,delhi)

(mumbai,1234,cts ,4536,mumbai)

(chennai,1245,accen ,2657,chennai)

(banglore,2354,wipro ,4536,banglore)

(banglore,5436,tech ,3425,banglore)

(hyderabad,1234,tata ,4567,hyderabad)

-----------------SPLIT----------------

//BY USING SPLIT IT DIVIDE THE DATA BASED ON CRITERIA

//THIS IS OUR INPUT.

root@ubuntu:/home/jayesh# cat Sample.txt

101:GAURAV:20000:HYDERABAD

102:VAIBHAV:10000:PUNE

103:PRAWIN:30000:SANGVI

104:HARISH:40000:NAGAR

105:VIKAS:60000:MUMBAI

105:TUSHAR:70000:PUNE

101:TILOTTAMMA:100000:PUNE

root@ubuntu:/home/jayesh# pig

2014-12-11 04:55:49,583 [main] INFO org.apache.pig.Main - Logging error messages to: /home/jayesh/pig\_1418302549582.log

2014-12-11 04:55:49,824 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020

2014-12-11 04:55:50,153 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021

grunt> inidata = Load '/PIG43/Sample.txt' using PigStorage (':') as (cid:int,cname:chararray,sal:int,clocation:chararray);

grunt> SPLIT inidata into A IF cid>101 , B IF sal<100000; //HERE WE SET CONDITION FOR SPLIT A O/P AND SPLIT B ALSO

grunt> STORE A INTO 'FIRSTSPLIT7'; STORE B INTO 'SECONDSPLIT7';

grunt> quit

//BASED ON OUR GIVEN SCHEMA IT SHOWS O/PS

root@ubuntu:/home/jayesh# hadoop fs -cat /user/root/FIRSTSPLIT7/part-m-00000

102 VAIBHAV 10000 PUNE

103 PRAWIN 30000 SANGVI

104 HARISH 40000 NAGAR

105 VIKAS 60000 MUMBAI

105 TUSHAR 70000 PUNE

root@ubuntu:/home/jayesh# hadoop fs -cat /user/root/SECONDSPLIT7/part-m-00000

101 GAURAV 20000 HYDERABAD

102 VAIBHAV 10000 PUNE

103 PRAWIN 30000 SANGVI

104 HARISH 40000 NAGAR

105 VIKAS 60000 MUMBAI

105 TUSHAR 70000 PUNE

-------UDF----------------

INPUT

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

File: epochTimeData.log

129464723

129423464

129464125

129234622

122457341

134998034

190874678

231234556

567446677

345563465

123445556

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

File: epochscript.pig

REGISTER EPOCH-UC.jar;

A = LOAD 'epochTimeData.log' as (timeline:chararray);

B = FOREACH A GENERATE timeline , com.epoch.pigusecase.EPOCH(timeline);

STORE B INTO 'EPOCH-OUTPUT1' using PigStorage('|');

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

root@ubuntu:/home/mrinmoy# pig -x local epochscript.pig

......................................

......................................

......................................

...................

Success!

Job Stats (time in seconds):

JobId Alias Feature Outputs

job\_local\_0001 A,B MAP\_ONLY file:///home/mrinmoy/EPOCH-OUTPUT1,

Input(s):

Successfully read records from: "file:///home/mrinmoy/epochTimeData.log"

Output(s):

Successfully stored records in: "file:///home/mrinmoy/EPOCH-OUTPUT1"

Job DAG:

job\_local\_0001

2014-12-22 05:05:34,839 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

root@ubuntu:/home/mrinmoy# ls EPOCH-OUTPUT1/

part-m-00000 \_SUCCESS

root@ubuntu:/home/mrinmoy# cat EPOCH-OUTPUT1/part-m-00000

129464723|1970-01-07:19

129423464|1970-01-07:07

129464125|1970-01-07:18

129234622|1970-01-05:03

122457341|1969-12-07:09

134998034|1970-01-22:03

190874678|1970-01-22:14

231234556|1969-12-23:15

567446677|1970-01-06:13

345563465|1970-01-23:10

123445556|1969-12-18:20