Student Data Analysis under Distributed Environment(HADOOP)

Pig Installation Steps:

- To get a Pig distribution, download a recent stable release from one of the Apache Download Mirrors.

  Extract pig-0.11.1.tar.gz file

- Setting up Pig related environment variables into $HOME/.bashrc file of user root.

```
# Set Pig related environment variables
export PIG_INSTALL=/root/Downloads/pig-0.11.0
export PATH=$PATH:$PIG_INSTALL/bin
export PIG_CLASSPATH=/root/hadoop-1.1.1/conf
```

- Start the Grunt shell: $ **pig**

  It displays following messages on terminal:

```
root@ubuntu:~/hadoop-1.1.1# cd /
root@ubuntu:/# pig

2014-04-28 00:13:41,817 [main] INFO  org.apache.pig.Main - Apache Pig version 0.11.0 (r1444328) compiled Feb 09 2013, 02:59:16
2014-04-28 00:13:41,818 [main] INFO  org.apache.pig.Main - Logging error messages to: //pig_1398669221809.log
2014-04-28 00:13:41,921 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2014-04-28 00:13:42,515 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:54310
2014-04-28 00:13:43,097 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:54311
Grunt >
```

- **Input File: (studentdata.txt)**

Record structure: (Enrollment_No-Class-Name-Attendance_Percentage)

1-TE-Ron-58	44-SE-Adam-26	87-TE-Gibrel-58
2-FE-Dock-34	45-TE-Rooney-42	88-FE-Dock-65
3-TE-Sri-87	46-TE-Jack-92	89-FE-Ron-95
4-SE-Harry-76	47-TE-Sri-40	90-SE-Gibrel-3
5-FE-Watson-78	48-SE-Sri-4	91-FE-Rooney-66
6-TE-Ron-70	49-SE-Watson-98	92-FE-Harry-5
7-TE-Jack-93	50-SE-Adam-36	93-TE-Rooney-79
8-SE-Harry-50	51-BE-Rooney-77	94-TE-Sri-59
9-TE-Rooney-38	52-BE-Harrison-26	95-TE-Adam-33
10-FE-Harrison-56	53-BE-Rooney-4	96-SE-Dock-38
11-SE-Ron-53	54-SE-Gibrel-51	97-TE-Harry-56
12-TE-Gibrel-5	55-FE-Harrison-5	98-SE-Rooney-53
13-FE-Jack-66	56-SE-Harry-26	99-BE-Rooney-48
14-BE-Rooney-74	57-SE-Dock-0	100-SE-Sri-20
15-TE-Sri-69	58-TE-Jack-61	
16-TE-Jack-42	59-BE-Gibrel-86	
17-BE-Harry-75	60-BE-Rooney-23	
18-TE-Jack-49	61-BE-Gibrel-88	
19-BE-Sri-91	62-FE-Watson-36	
20-TE-Gibrel-72	63-BE-Harrison-20	
21-BE-Gibrel-73	64-SE-Sri-98	
22-TE-Ron-81	65-BE-Sri-97	
23-TE-Harrison-13	66-TE-Dock-2	
24-BE-Harry-74	67-TE-Jack-86	
25-FE-Gibrel-82	68-SE-Jack-17	
26-SE-Harry-44	69-TE-Dock-17	
27-FE-Sri-12	70-TE-Harrison-29	
28-TE-Adam-2	71-TE-Harrison-25	
29-FE-Dock-66	72-TE-Harrison-55	
30-BE-Harrison-6	73-FE-Watson-24	
31-BE-Sri-54	74-BE-Jack-78	
32-FE-Ron-66	75-SE-Dock-33	
33-BE-Dock-23	76-BE-Gibrel-63	
34-TE-Watson-68	77-SE-Rooney-90	
35-SE-Ron-35	78-BE-Harrison-81	
36-BE-Sri-21	79-TE-Rooney-11	
37-BE-Gibrel-76	80-SE-Watson-39	
38-TE-Adam-81	81-SE-Sri-48	
39-FE-Harry-37	82-BE-Jack-64	
40-BE-Watson-2	83-FE-Watson-89	
41-SE-Ron-89	84-TE-Ron-89	
42-FE-Jack-5	85-FE-Harrison-51	
43-FE-Jack-47	86-BE-Gibrel-74	

- **Pig Query Execution:**

  Step 1: Copy studentdata.txt file from Unix file system to the Hadoop file system

  - root@ubuntu**:~# hadoop fs -copyFromLocal /root/Desktop/studentdata.txt /user/root/small**

  Step 2: Start grunt shell

  - By default mode of pig query execution is "mapreduce"

- root@ubuntu:/# pig
- Warning: $HADOOP_HOME is deprecated.
  2014-04-28 00:13:41,817 [main] INFO org.apache.pig.Main - Apache Pig version 0.11.0 (r1444328) compiled Feb 09 2013, 02:59:16
  2014-04-28 00:13:41,818 [main] INFO org.apache.pig.Main - Logging error messages to: //pig_1398669221809.log
  2014-04-28 00:13:41,921 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
  2014-04-28 00:13:42,515 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:54310
  2014-04-28 00:13:43,097 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:54311
- grunt>

Step 3: Pig query execution

- ***Query 1: Group student records by class.***
  *grunt>**classd = load '/user/root/small/studentdata.txt' using PigStorage('-');***
  *grunt> **stud = group classd by $1;***
  *grunt> **store stud into 'user/root/small/Classstudents';***

  2014-04-28 00:50:16,119 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
  2014-04-28 00:50:16,529 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
  2014-04-28 00:50:16,632 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Mu

```
ltiQueryOptimizer - MR plan size before optimization: 1
2014-04-28 00:50:16,632 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Mu
ltiQueryOptimizer - MR plan size after optimization: 1
2014-04-28 00:50:16,834 [main] INFO
org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added
to the job
2014-04-28 00:50:16,885 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - mapred.job.reduce.markreset.buffer.percent is not
set, set to default 0.3
2014-04-28 00:50:16,890 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - Using reducer estimator:
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Inp
utSizeReducerEstimator
2014-04-28 00:50:16,903 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Inp
utSizeReducerEstimator - BytesPerReducer=1000000000
maxReducers=999 totalInputFileSize=1575
2014-04-28 00:50:16,903 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - Setting Parallelism to 1
2014-04-28 00:50:16,904 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - creating jar file Job1647863832204725138.jar
2014-04-28 00:50:29,745 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - jar file Job1647863832204725138.jar created
2014-04-28 00:50:29,839 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - Setting up single store job
2014-04-28 00:50:30,073 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - 1 map-reduce job(s) waiting for submission.
2014-04-28 00:50:30,575 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - 0% complete
2014-04-28 00:50:30,821 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input
paths to process : 1
2014-04-28 00:50:30,822 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
2014-04-28 00:50:30,849 [JobControl] INFO
```

```
org.apache.hadoop.util.NativeCodeLoader - Loaded the native-hadoop
library
2014-04-28 00:50:30,849 [JobControl] WARN
org.apache.hadoop.io.compress.snappy.LoadSnappy - Snappy native
library not loaded
2014-04-28 00:50:30,871 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths (combined) to process : 1
2014-04-28 00:50:31,871 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - HadoopJobId: job_201404272033_0003
2014-04-28 00:50:31,871 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - Processing aliases classd,stud
2014-04-28 00:50:31,871 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - detailed locations: M: classd[1,9],stud[2,7] C:  R:
2014-04-28 00:50:31,872 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - More information at:
http://localhost:50030/jobdetails.jsp?jobid=job_201404272033_000
3
2014-04-28 00:50:47,581 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - 50% complete
2014-04-28 00:50:57,645 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - 66% complete
2014-04-28 00:51:06,760 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Ma
pReduceLauncher - 100% complete
2014-04-28 00:51:06,764 [main] INFO
org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion   UserId StartedAt   FinishedAt
Features
1.1.1   0.11.0  root    2014-04-28 00:50:16     2014-04-28
00:51:06    GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces     MaxMapTime  MinMapTIme  AvgMapTime
MedianMapTime   MaxReduceTime   MinReduceTime
AvgReduceTime   MedianReducetime   Alias   Feature
```

```
Outputs
job_201404272033_0003    1    1    9    9    9    9
12    12    12    12    classd,stud    GROUP_BY
hdfs://localhost:54310/user/root/user/root/small/Classstudents,

Input(s):
Successfully read 100 records (1948 bytes) from:
"/user/root/small/studentdata.txt"

Output(s):
Successfully stored 4 records (1695 bytes) in:
"hdfs://localhost:54310/user/root/user/root/small/Classstudents"

Counters:
Total records written : 4
Total bytes written : 1695
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201404272033_0003


2014-04-28 00:51:06,803 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

- ***<u>Query 2</u>: Filter the students with more than 70% attendance then group those student records by class.***

  grunt> **nondstud = load '/user/root/small/studentdata.txt' using PigStorage('-') as (int, chararray, chararray,int);**

  grunt> **nonstud1 = filter nondstud by val_3 > 70;**

  grunt> **ndstudent= group nonstud1 by val_1;**

  grunt> **store ndstudent into '/user/root/small/regularstud';**

```
2014-04-28 01:09:33,141 [main] INFO
org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script:
GROUP_BY,FILTER
2014-04-28 01:09:33,573 [main] INFO
```

```
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompil
er - File concatenation threshold: 100 optimistic? false
2014-04-28 01:09:33,671 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQuer
yOptimizer - MR plan size before optimization: 1
2014-04-28 01:09:33,671 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQuer
yOptimizer - MR plan size after optimization: 1
2014-04-28 01:09:33,840 [main] INFO
org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the
job
2014-04-28 01:09:33,878 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobContro
lCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to
default 0.3
2014-04-28 01:09:33,883 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobContro
lCompiler - Using reducer estimator:
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSize
ReducerEstimator
2014-04-28 01:09:33,888 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSize
ReducerEstimator - BytesPerReducer=1000000000 maxReducers=999
totalInputFileSize=1575
2014-04-28 01:09:33,888 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobContro
lCompiler - Setting Parallelism to 1
2014-04-28 01:09:33,889 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobContro
lCompiler - creating jar file Job5798419949182401988.jar
2014-04-28 01:09:45,626 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobContro
lCompiler - jar file Job5798419949182401988.jar created
2014-04-28 01:09:45,675 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobContro
lCompiler - Setting up single store job
2014-04-28 01:09:45,715 [main] INFO
org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false,
will not generate code.
2014-04-28 01:09:45,715 [main] INFO
org.apache.pig.data.SchemaTupleFrontend - Starting process to move
generated code to distributed cacche
2014-04-28 01:09:45,722 [main] INFO
org.apache.pig.data.SchemaTupleFrontend - Setting key
[pig.schematuple.classes] with classes to deserialize []
```

```
2014-04-28 01:09:45,998 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - 1 map-reduce job(s) waiting for submission.
2014-04-28 01:09:46,500 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - 0% complete
2014-04-28 01:09:46,921 [JobControl] INFO
org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths
to process : 1
2014-04-28 01:09:46,921 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
input paths to process : 1
2014-04-28 01:09:46,977 [JobControl] INFO
org.apache.hadoop.util.NativeCodeLoader - Loaded the native-hadoop library
2014-04-28 01:09:46,977 [JobControl] WARN
org.apache.hadoop.io.compress.snappy.LoadSnappy - Snappy native library
not loaded
2014-04-28 01:09:46,992 [JobControl] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
input paths (combined) to process : 1
2014-04-28 01:09:48,010 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - HadoopJobId: job_201404272033_0004
2014-04-28 01:09:48,010 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - Processing aliases ndstudent,nondstud,nonstud1
2014-04-28 01:09:48,010 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - detailed locations: M: nondstud[1,11],nondstud[-1,-
1],nonstud1[2,11],ndstudent[3,11] C:  R:
2014-04-28 01:09:48,010 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - More information at:
http://localhost:50030/jobdetails.jsp?jobid=job_201404272033_0004
2014-04-28 01:10:00,794 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - 50% complete
2014-04-28 01:10:10,852 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - 66% complete
2014-04-28 01:10:17,949 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu
ceLauncher - 100% complete
2014-04-28 01:10:17,953 [main] INFO
org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
```

```
HadoopVersion   PigVersion   UserId StartedAt   FinishedAt   Features
1.1.1   0.11.0   root   2014-04-28 01:09:33   2014-04-28 01:10:17   GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps   Reduces   MaxMapTime   MinMapTIme   AvgMapTime   MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime   Alias   Feature   Outputs
job_201404272033_0004   1   1   7   7   7   7   12   12   12   12   ndstudent,nondstud,nonstud1   GROUP_BY   /user/root/small/regularstud,

Input(s):
Successfully read 100 records (1948 bytes) from: "/user/root/small/studentdata.txt"

Output(s):
Successfully stored 4 records (539 bytes) in: "/user/root/small/regularstud"

Counters:
Total records written : 4
Total bytes written : 539
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201404272033_0004


2014-04-28 01:10:17,977 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

- **Output Files:**

1. ***Output of Query 1: Classstudents***

BE {(63,BE,Harrison,20),(61,BE,Gibrel,88),(51,BE,Rooney,77),(52,BE,Harrison,26),(53,BE,Rooney,4),(14,BE,Rooney,74),(60,BE,Rooney,23),(59,BE,Gibrel,86),(86,BE,Gibrel,74),(17,BE,Harry,75),(19,BE,Sri,91),(82,BE,Jack,64),(21,BE,Gibrel,73),(78,BE,Harrison,81),(76,BE,Gibrel,63),(24,BE,Harry,74),(74,BE,Jack,78),(30,BE,Harrison,6),(31,BE,Sri,54),(33,BE,Dock,23),(36,BE,Sri,21),(37,BE,Gibrel,76),(65,BE,Sri,97),(40,BE,Watson,2),(99,BE,Rooney,48)}
FE {(92,FE,Harry,5),(91,FE,Rooney,66),(25,FE,Gibrel,82),(88,FE,Dock,65),(55,FE,Harrison,5),(13,FE,Jack,66),(5,FE,Watson,78),(27,FE,Sri,12),(42,FE,Jack,5),(29,FE,Dock,66),(85,FE,Harrison,51),(43,FE,Jack,47),(83,FE,Watson,89),(39,FE,Harry,37),(32,FE,Ron,66),(62,FE,Watson,36),(73,FE,Watson,24),(89,FE,Ron,95),(10,FE,Harrison,56),(2,FE,Dock,34)}
SE {(100,SE,Sri,20),(4,SE,Harry,76),(8,SE,Harry,50),(11,SE,Ron,53),(26,SE,Harry,44),(35,SE,Ron,35),(41,SE,Ron,89),(44,SE,Adam,26),(48,SE,Sri,4),(49,SE,Watson,98),(50,SE,Adam,36),(54,SE,Gibrel,51),(56,SE,Harry,26),(57,SE,Dock,0),(64,SE,Sri,98),(68,SE,Jack,17),(75,SE,Dock,33),(77,SE,Rooney,90),(80,SE,Watson,39),(81,SE,Sri,48),(90,SE,Gibrel,3),(96,SE,Dock,38),(98,SE,Rooney,53)}
TE {(69,TE,Dock,17),(70,TE,Harrison,29),(71,TE,Harrison,25),(72,TE,Harrison,55),(34,TE,Watson,68),(28,TE,Adam,2),(94,TE,Sri,59),(23,TE,Harrison,13),(95,TE,Adam,33),(22,TE,Ron,81),(79,TE,Rooney,11),(3,TE,Sri,87),(97,TE,Harry,56),(20,TE,Gibrel,72),(18,TE,Jack,49),(84,TE,Ron,89),(16,TE,Jack,42),(15,TE,Sri,69),(87,TE,Gibrel,58),(12,TE,Gibrel,5),(9,TE,Rooney,38),(1,TE,Ron,58),(7,TE,Jack,93),(58,TE,Jack,61),(47,TE,Sri,40),(46,TE,Jack,92),(45,TE,Rooney,42),(6,TE,Ron,70),(38,TE,Adam,81),(66,TE,Dock,2),(67,TE,Jack,86),(93,TE,Rooney,79)}

- ***Output of Query 2: Regularstud***

BE {(65,BE,Sri,97),(74,BE,Jack,78),(78,BE,Harrison,81),(86,BE,Gibrel,74),(21,BE,Gibrel,73),(24,BE,Harry,74),(19,BE,Sri,91),(37,BE,Gibrel,76),(14,BE,Rooney,74),(17,BE,Harry,75),(51,BE,Rooney,77),(59,BE,Gibrel,86),(61,BE,Gibrel,88)}
FE {(83,FE,Watson,89),(5,FE,Watson,78),(89,FE,Ron,95),(25,FE,Gibrel,82)}
SE {(49,SE,Watson,98),(4,SE,Harry,76),(41,SE,Ron,89),(77,SE,Rooney,90),(64,SE,Sri,98)}
TE {(3,TE,Sri,87),(7,TE,Jack,93),(20,TE,Gibrel,72),(22,TE,Ron,81),(38,TE,Adam,81),(46,TE,Jack,92),(67,TE,Jack,86),(84,TE,Ron,89),(93,TE,Rooney,79)}