PRACTICAL 10

Creating And Using Pig Latin Script

What is Pig in Hadoop?

Pig is a scripting platform that runs on Hadoop clusters designed to process and analyze large datasets. Pig is extensible, self-optimizing, and easily programmed.

Programmers can use Pig to write data transformations without knowing Java. Pig uses both structured and unstructured data as input to perform analytics and uses HDFS to store the results.

Components of Pig

There are two major components of the Pig:

- Pig Latin script language
- A runtime engine

Pig Latin script language:

The Pig Latin script is a procedural data flow language. It contains syntax and commands that can be applied to implement business logic. Examples of Pig Latin are LOAD and STORE.

A runtime engine:

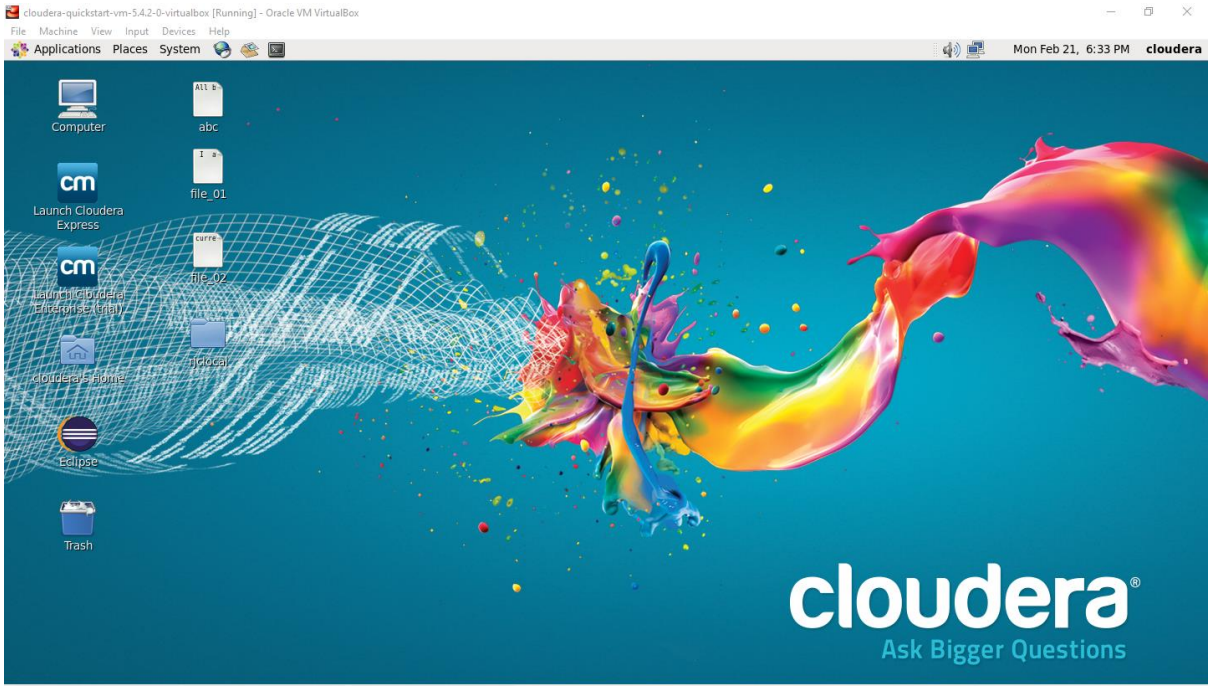
The runtime engine is a compiler that produces sequences of MapReduce programs. It uses HDFS to store and retrieve data. It is also used to interact with the Hadoop system (HDFS and MapReduce).

The runtime engine parses, validates, and compiles the script operations into a sequence of MapReduce jobs.

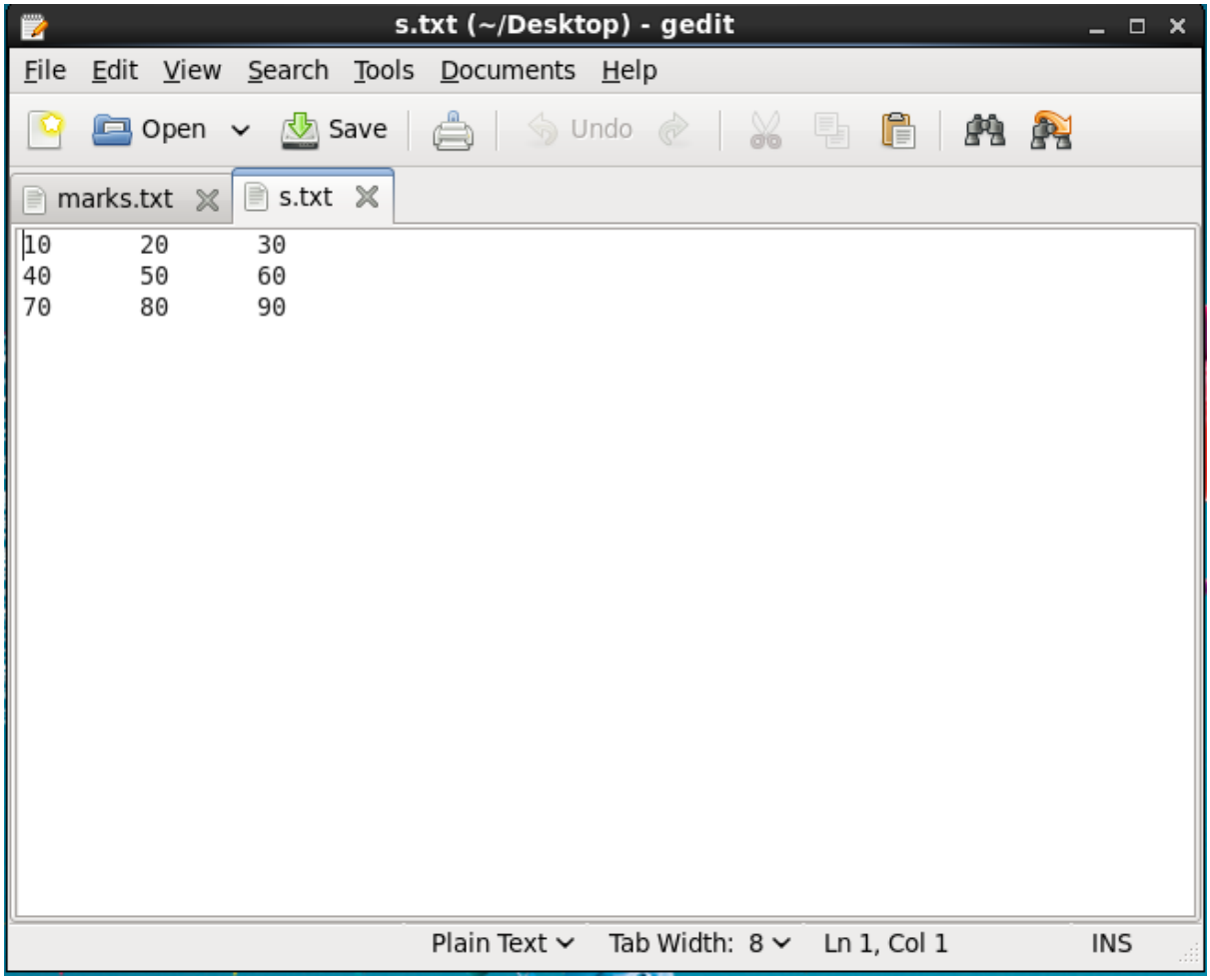
Steps:

Example -1

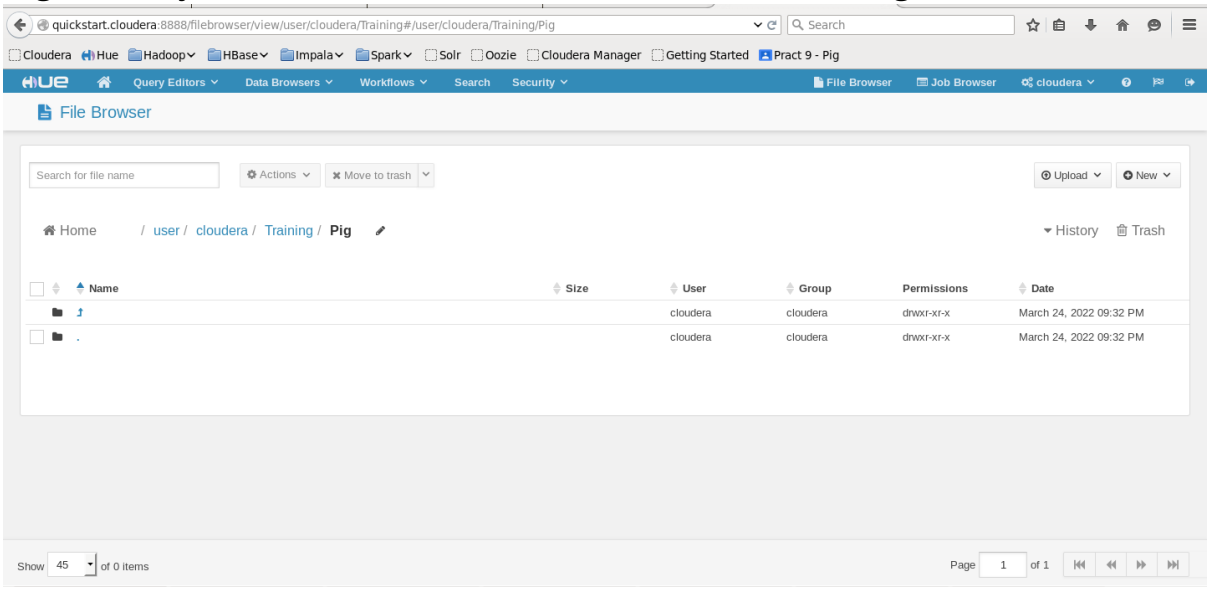
1) Start the cloudera.



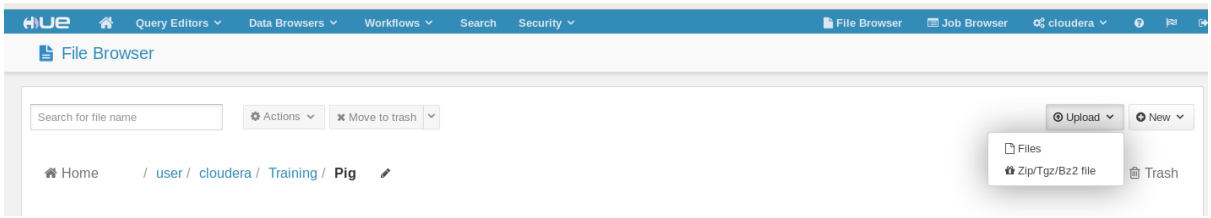
2) Here we are created text file which having data as integer values

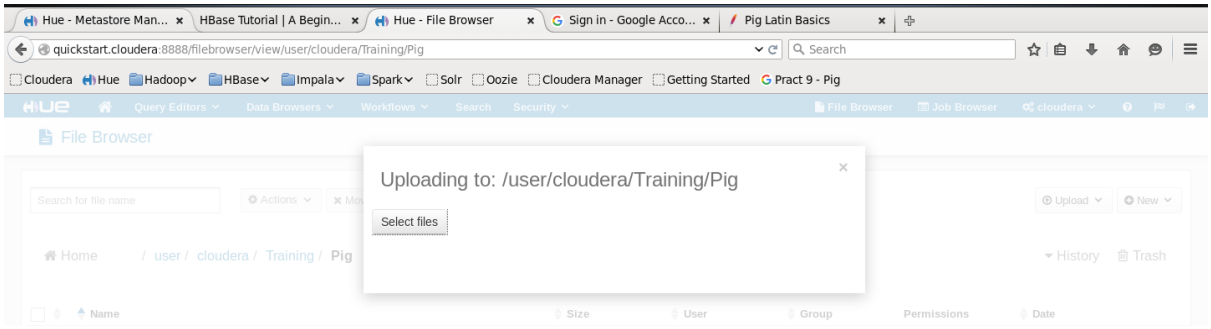


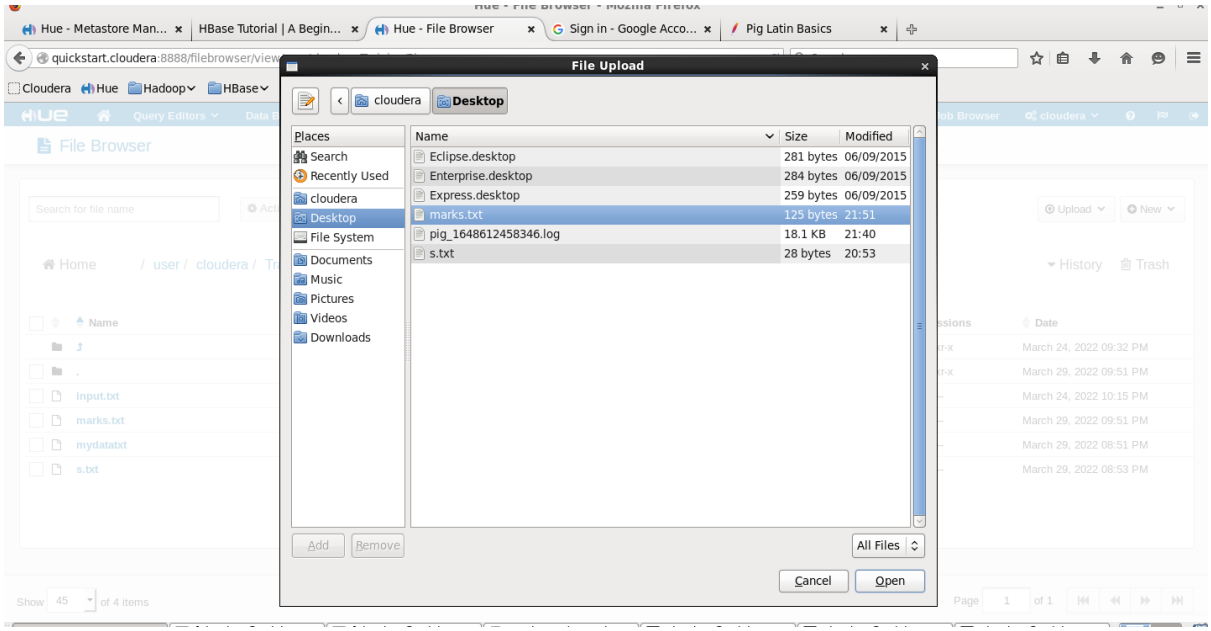
3) Now open the directory /user/cloudera
4) Now we are creating the directory as Training inside /user/cloudera
5) After creating Training directory now creating the Pig directory inside Training.
6) Pig directory has been created inside /user/cloudera/Training

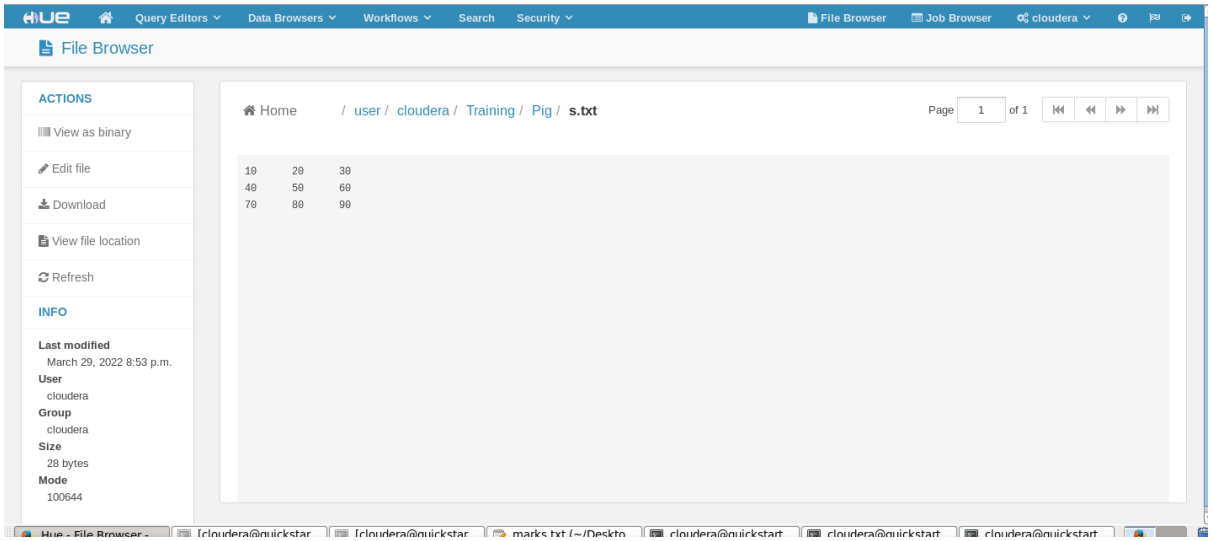


7) Uploading s.txt file inside /user/cloudera/Training/Pig directory









8) Now Open the terminal. And start Pig by typing pig on terminal.
9) Now we have to load that input file where ever it is stored.
10) Now we are dumping the data. It will done by the mapreduce task.

```
grunt> m = LOAD '/user/cloudera/Training/Pig/s.txt';
grunt> dump m;
2022-03-29 21:41:01,763 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2022-03-29 21:41:01,802 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-03-29 21:41:01,883 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2022-03-29 21:41:01,927 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:41:01,927 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:41:02,074 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2022-03-29 21:41:02,523 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig script settings are added to the job
2022-03-29 21:41:02,594 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use m
apreduce.reduce.markreset.buffer.percent
2022-03-29 21:41:02,594 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
t is not set, set to default 0.3
2022-03-29 21:41:02,594 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.output.compress is deprecated. Instead, use mapreduce.output.file
outputformat.compress
2022-03-29 21:41:03,186 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - creating jar file Job8124450571141603823.
jar
2022-03-29 21:41:06,570 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job8124450571141603823.jar created
2022-03-29 21:41:06,570 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.jar is deprecated. Instead, use mapreduce.job.jar
2022-03-29 21:41:06,597 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-03-29 21:41:06,642 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-03-29 21:41:06,649 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracker.http.add
ress
2022-03-29 21:41:06,649 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
```

```
2022-03-29 21:41:28,236 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-03-29 21:41:28,239 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.4.2  0.12.0-cdh5.4.2 cloudera        2022-03-29 21:41:02     2022-03-29 21:41:28     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature O
utputs
job_1644894610889_0043  1       0       5       5       5       5       n/a     n/a     n/a     n/a     m       MAP_ONLY        hdfs://quickstart.cloudera:8020/tmp/temp-16407194
62/tmp1940945813,

Input(s):
Successfully read 3 records (411 bytes) from: "/user/cloudera/Training/Pig/s.txt"

Output(s):
Successfully stored 3 records (49 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1640719462/tmp1940945813"

Counters:
Total records written : 3
Total bytes written : 49
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1644894610889_0043


2022-03-29 21:41:28,325 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-29 21:41:28,327 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-29 21:41:28,327 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-29 21:41:28,346 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-29 21:41:28,346 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10,20,30)
(40,50,60,)
(70,80,90)
grunt>
```

11) A schema is specified using the AS keyword.

```
grunt>  m = LOAD '/user/cloudera/Training/Pig/s.txt' AS (c1:int, c2:int, c3:int);
grunt> describe m;
m: {c1: int,c2: int,c3: int}
grunt>
```

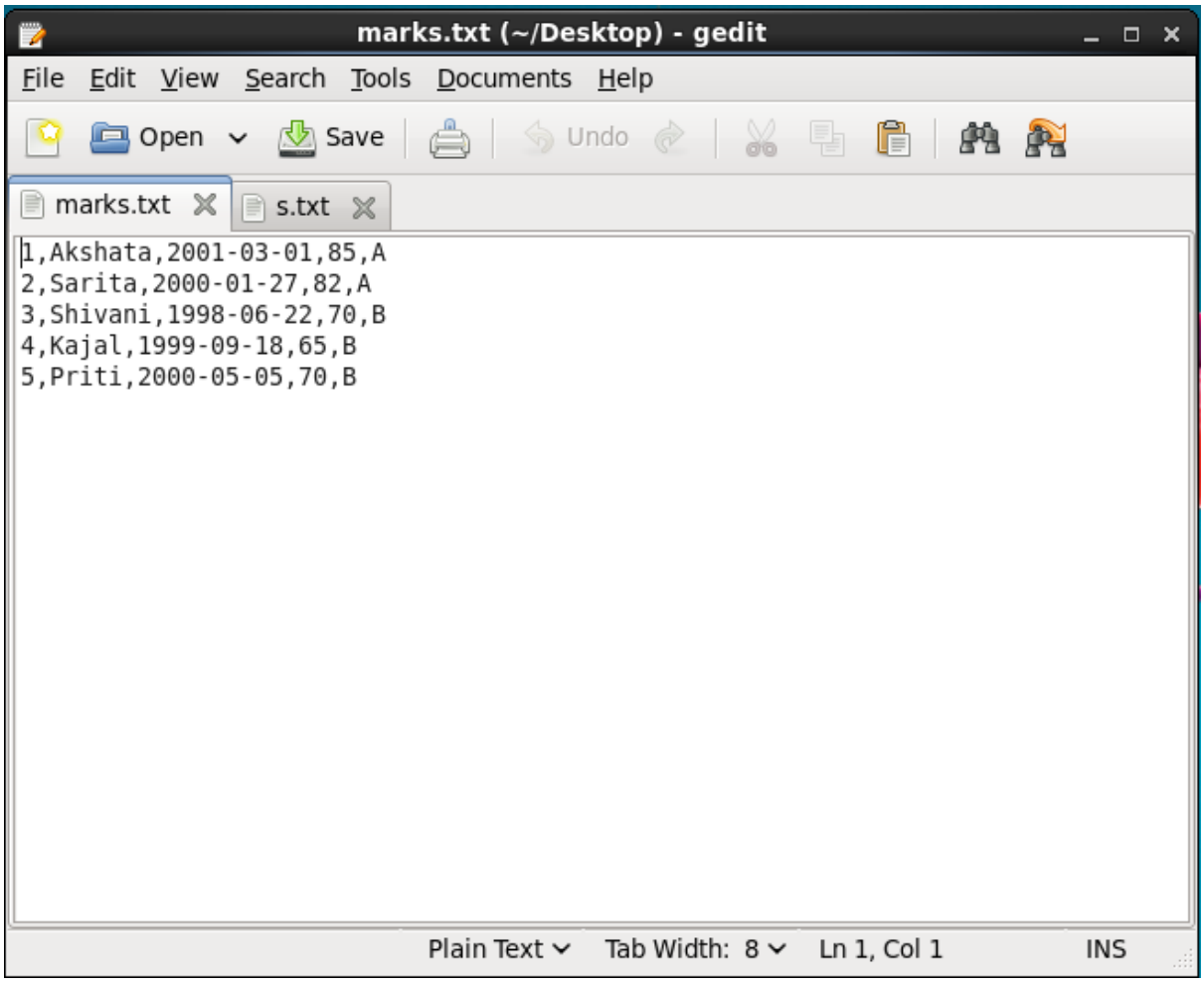
12) We can use the DESCRIBE and ILLUSTRATE operators to view the schema.

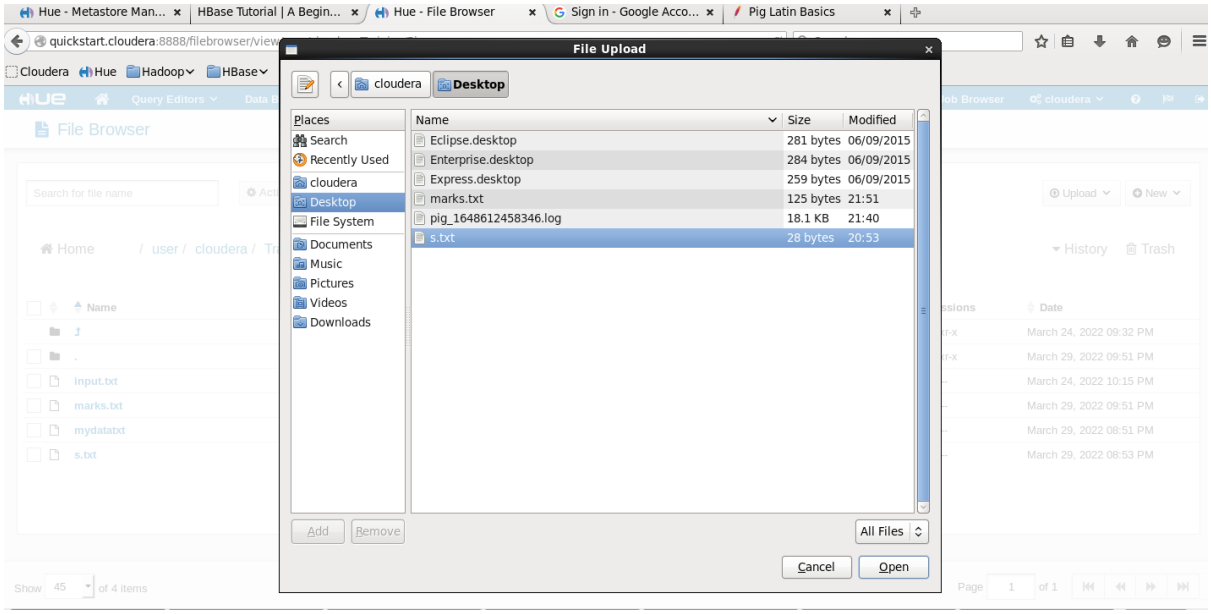
```
grunt> illustrate m;
2022-03-29 21:43:55,500 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-29 21:43:55,500 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-29 21:43:55,501 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2022-03-29 21:43:55,501 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2022-03-29 21:43:55,509 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[DuplicateForEachColumnRewrite, ImplicitSplitInserter, LoadTy
peCastInserter, NewPartitionFilterOptimizer, StreamTypeCastInserter], RULES_DISABLED=[AddForEach, ColumnMapKeyPrune, FilterLogicExpressionSimplifier, GroupByConstParallelSetter,
 LimitOptimizer, MergeFilter, MergeForEach, PartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter]}
2022-03-29 21:43:55,518 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:43:55,519 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:43:55,519 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:43:55,520 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:43:55,520 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:43:55,585 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:43:55,598 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset])
: M: m[3,4] C:  R:
2022-03-29 21:43:55,602 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-29 21:43:55,602 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-03-29 21:43:55,630 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:43:55,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:43:55,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:43:55,631 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:43:55,632 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:43:55,682 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:43:55,699 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset])
: M: m[3,4] C:  R:
2022-03-29 21:43:55,703 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:43:55,705 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:43:55,705 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:43:55,705 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:43:55,705 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:43:55,747 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:43:55,755 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset])
: M: m[3,4] C:  R:
2022-03-29 21:43:55,764 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:43:55,764 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
```

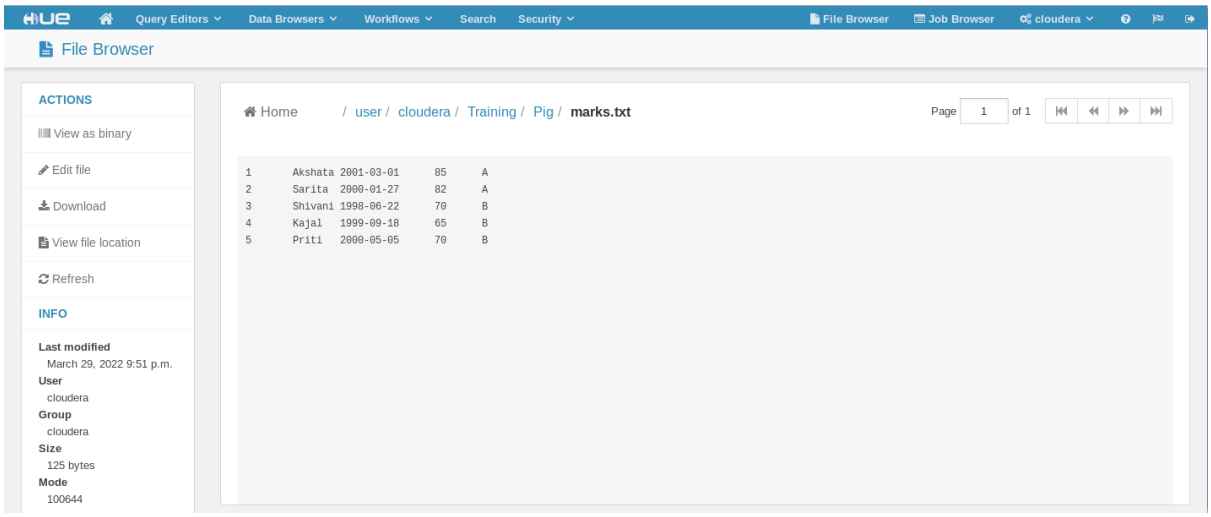
```
 to default 0.3
2022-03-29 21:43:55,946 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:43:55,949 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset]
: M: m[3,4] C:  R:
2022-03-29 21:43:55,949 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:43:55,950 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:43:55,950 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:43:55,951 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:43:55,951 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:43:55,965 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:43:55,968 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset]
: M: m[3,4] C:  R:
2022-03-29 21:43:55,968 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:43:55,969 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:43:55,969 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:43:55,969 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:43:55,970 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:43:55,993 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:43:55,996 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset]
: M: m[3,4] C:  R:
-----------------------------------------
| m     | c1:int    | c2:int    | c3:int    |
-----------------------------------------
|       | 70        | 80        | 90        |
-----------------------------------------

grunt>
```

Example-2

1) Here we are again creating input text file on cloudera







2) And we are again Loading that input file where ever it is stored.
3) also dumping the data

```
grunt> m = LOAD '/user/cloudera/Training/Pig/marks.txt';
grunt> dump m;
2022-03-29 21:46:41,315 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-03-29 21:46:41,315 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite
, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUp
Filter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-03-29 21:46:41,317 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:46:41,317 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:46:41,317 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:46:41,340 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-29 21:46:41,343 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:46:41,344 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:46:41,609 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job4514347734481447344.jar
2022-03-29 21:46:44,671 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job4514347734481447344.jar created
2022-03-29 21:46:44,677 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-03-29 21:46:44,685 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-03-29 21:46:44,685 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-29 21:46:44,688 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2022-03-29 21:46:44,694 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-29 21:46:44,820 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-29 21:46:44,820 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-03-29 21:46:44,822 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2022-03-29 21:46:44,846 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2022-03-29 21:46:44,926 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1644894610889_0044
2022-03-29 21:46:44,978 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1644894610889_0044
2022-03-29 21:46:44,987 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://quickstart.cloudera:8088/proxy/application_1644894610889_0044/
2022-03-29 21:46:45,187 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1644894610889_0044
2022-03-29 21:46:45,187 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases m
2022-03-29 21:46:45,187 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: m[4,4] C:  R:
2022-03-29 21:46:45,187 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?
jobid=job_1644894610889_0044
2022-03-29 21:46:45,261 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2022-03-29 21:46:58,532 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2022-03-29 21:47:00,227 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2022-03-29 21:47:00,227 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.4.2  0.12.0-cdh5.4.2 cloudera        2022-03-29 21:46:41     2022-03-29 21:47:00     UNKNOWN
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.4.2  0.12.0-cdh5.4.2 cloudera        2022-03-29 21:46:41     2022-03-29 21:47:00     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature O
utputs
job_1644894610889_0044  1       0       4       4       4       4       n/a     n/a     n/a     n/a     m       MAP_ONLY        hdfs://quickstart.cloudera:8020/tmp/temp-16495493
14/tmp-210229479,

Input(s):
Successfully read 5 records (512 bytes) from: "/user/cloudera/Training/Pig/marks.txt"

Output(s):
Successfully stored 5 records (150 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-1649549314/tmp-210229479"

Counters:
Total records written : 5
Total bytes written : 150
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1644894610889_0044


2022-03-29 21:47:00,301 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-29 21:47:00,302 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-29 21:47:00,302 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-29 21:47:00,302 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:47:00,309 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-29 21:47:00,309 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Akshata,2001-03-01,85,A)
(2,Sarita,2000-01-27,82,A)
(3,Shivani,1998-06-22,70,B)
(4,Kajal,1999-09-18,65,B)
(5,Priti,2000-05-05,70,B)
grunt>
```

4) A schema is specified using the AS keyword.

```
grunt> m = LOAD '/user/cloudera/Training/Pig/marks.txt' AS (ID:int, Name:chararray, DateofBirth:chararray, MarksPerPercentage:int, Grade:chararray);
grunt> describe m;
m: {ID: int,Name: chararray,DateofBirth: chararray,MarksPerPercentage: int,Grade: chararray}
grunt>
```

```
grunt> illustrate m;
2022-03-29 21:54:24,497 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-29 21:54:24,497 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-03-29 21:54:24,497 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2022-03-29 21:54:24,499 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2022-03-29 21:54:24,506 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[DuplicateForEachColumnRewrite, ImplicitSplitInserter, LoadTy
peCastInserter, NewPartitionFilterOptimizer, StreamTypeCastInserter], RULES_DISABLED=[AddForEach, ColumnMapKeyPrune, FilterLogicExpressionSimplifier, GroupByConstParallelSetter,
 LimitOptimizer, MergeFilter, MergeForEach, PartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter]}
2022-03-29 21:54:24,516 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:54:24,518 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:54:24,518 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:54:24,518 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:54:24,518 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:54:24,592 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:54:24,604 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset])
: M: m[3,4] C:  R:
2022-03-29 21:54:24,608 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-29 21:54:24,608 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-03-29 21:54:24,630 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:54:24,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:54:24,631 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:54:24,632 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:54:24,632 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:54:24,683 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:54:24,696 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset])
: M: m[3,4] C:  R:
2022-03-29 21:54:24,704 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:54:24,706 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:54:24,706 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 21:54:24,707 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 21:54:24,707 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 21:54:24,756 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 21:54:24,769 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset])
: M: m[3,4] C:  R:
2022-03-29 21:54:24,775 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 21:54:24,776 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 21:54:24,776 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
```

```
2022-03-29 22:01:15,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-03-29 22:01:15,957 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-03-29 22:01:15,957 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-03-29 22:01:15,957 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-29 22:01:15,957 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set
 to default 0.3
2022-03-29 22:01:15,970 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-03-29 22:01:15,972 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset])
: M: m[3,4] C:  R:
---------------------------------------------------------------------------------------------------------------
| m     | ID:int    | Name:chararray    | DateofBirth:chararray    | MarksPerPercentage:int    | Grade:chararray    |
---------------------------------------------------------------------------------------------------------------
|       | 1         | Akshata           | 2001-03-01               | 85                        | A                  |
---------------------------------------------------------------------------------------------------------------

grunt>
```