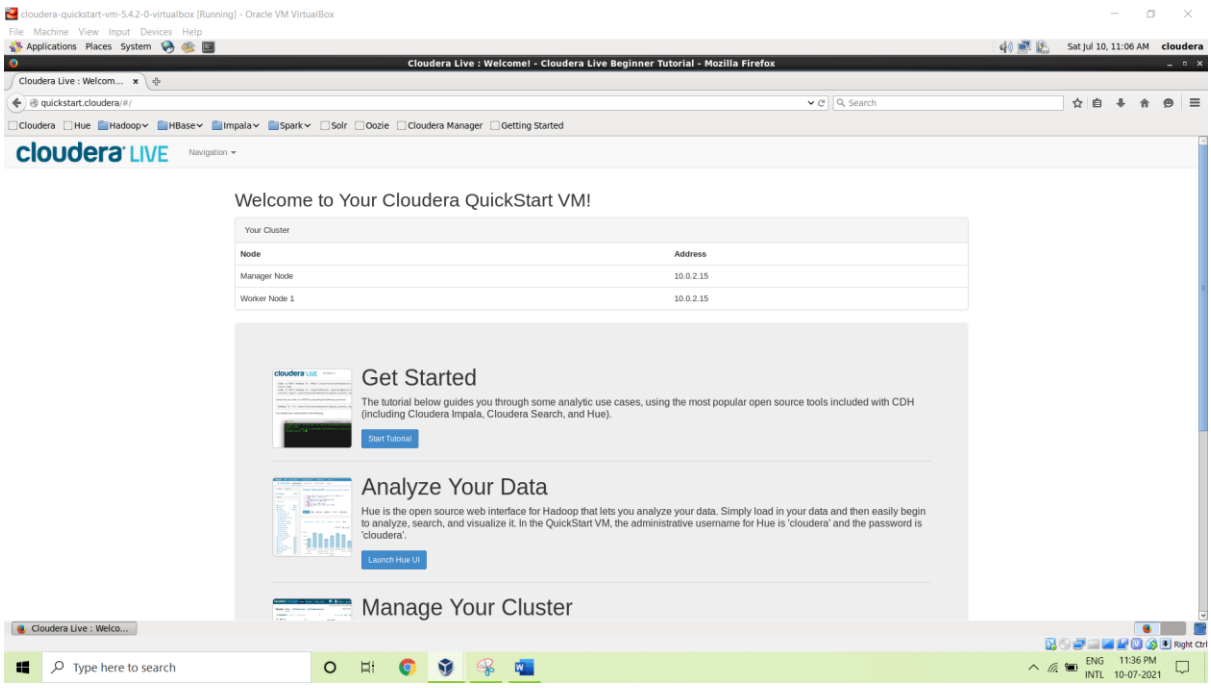
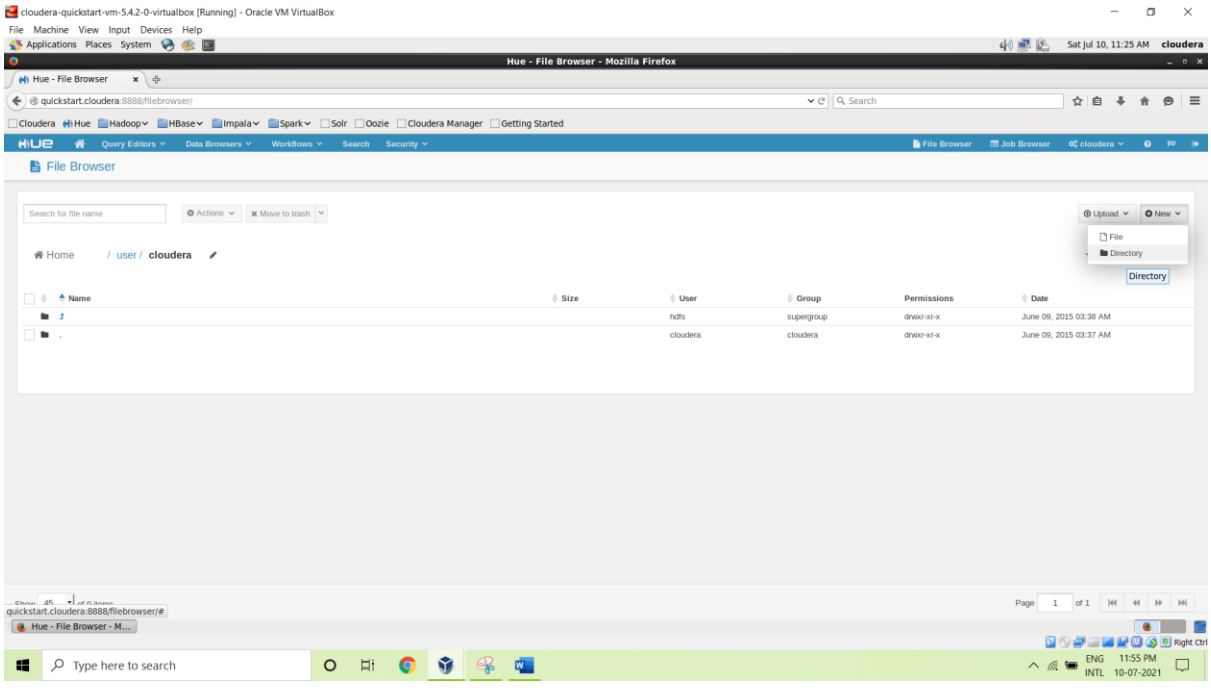
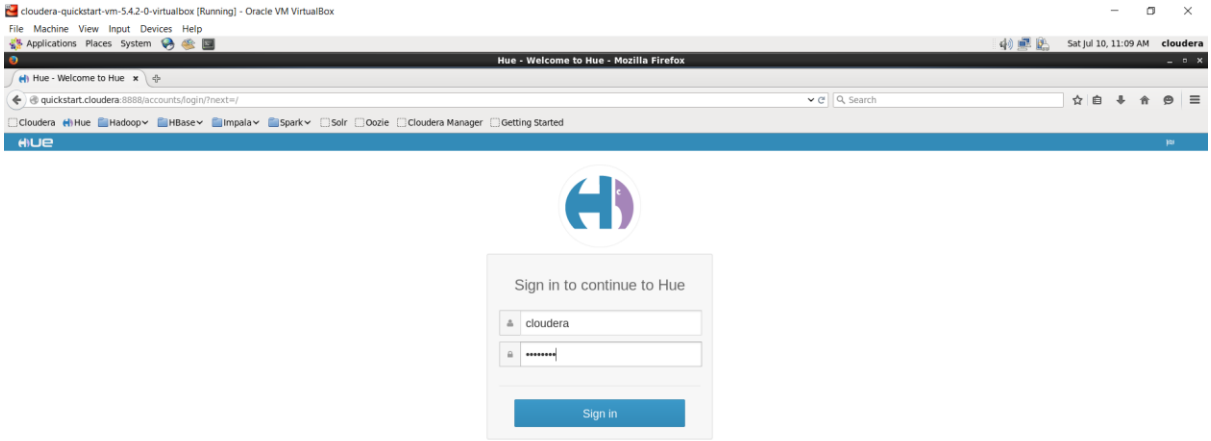
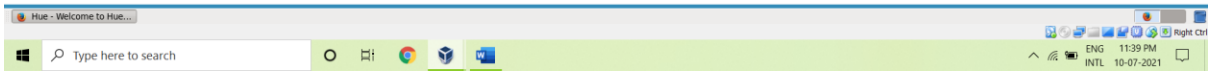
Experiment : 06

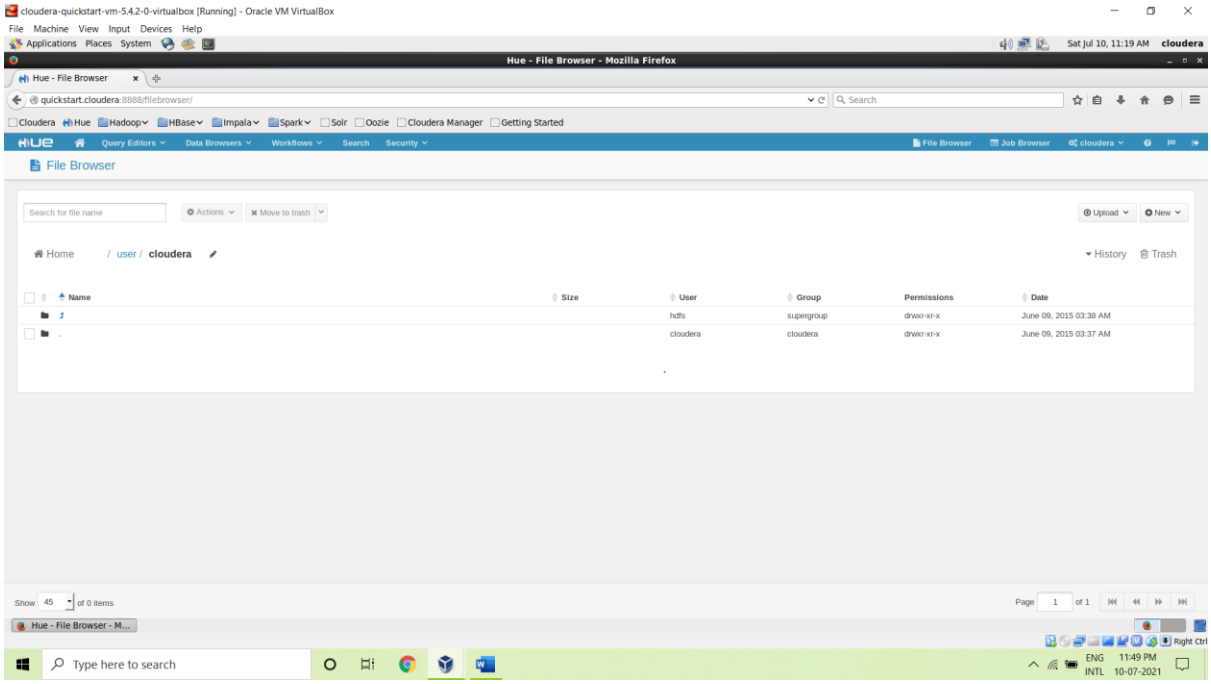
Implement word count problem using pig

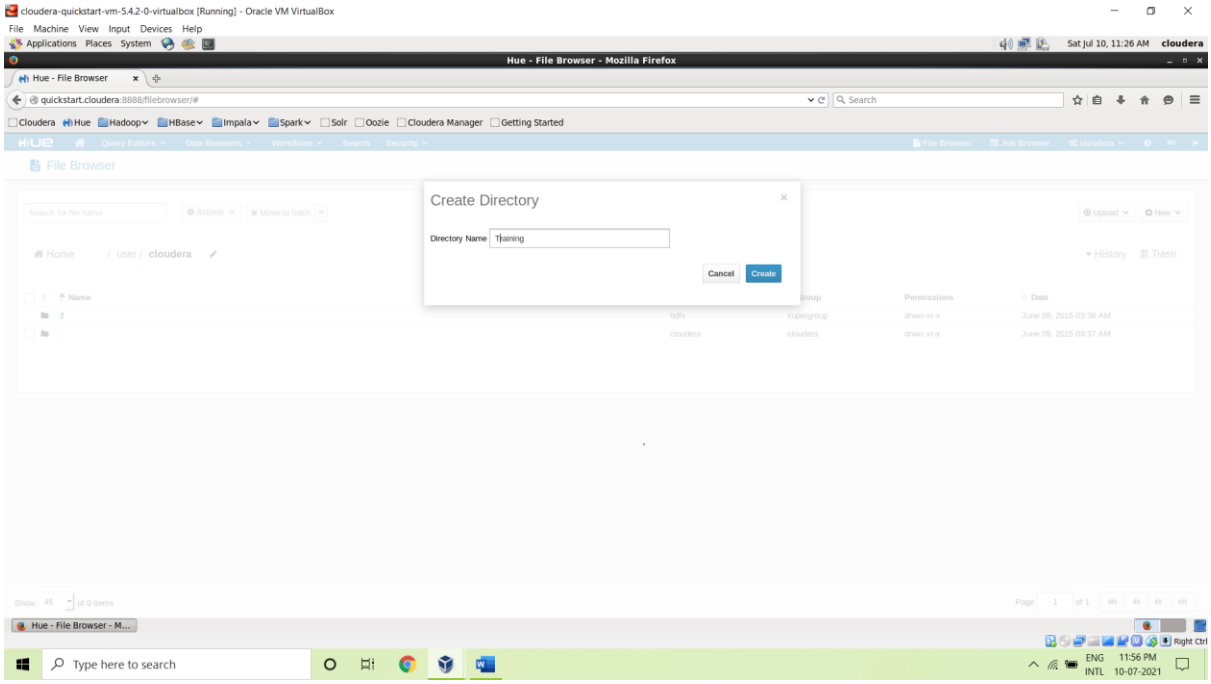


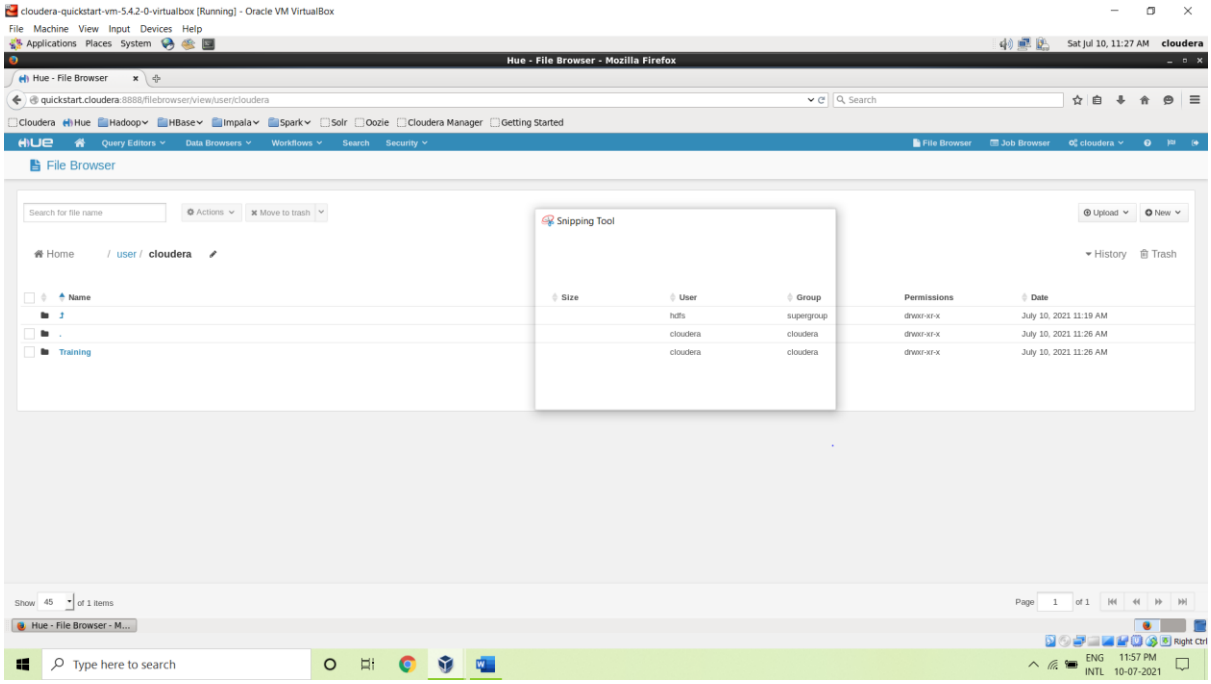


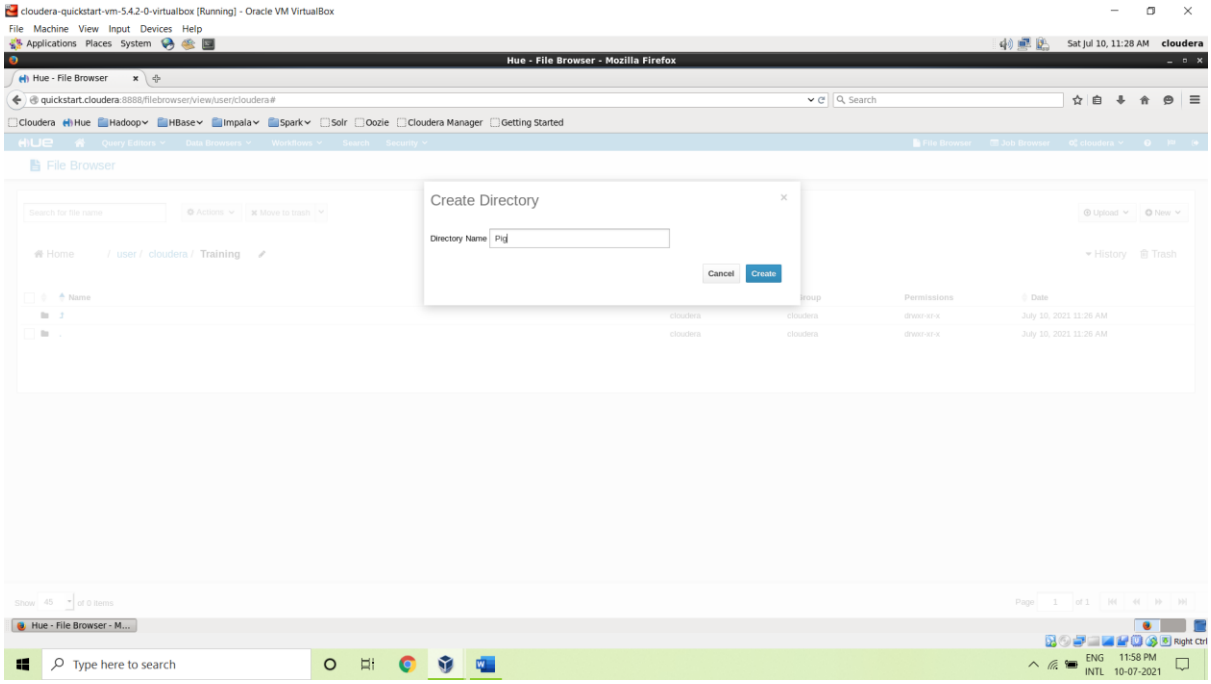

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:09 AM cloudera
Hue - Welcome to Hue - Mozilla Firefox
Hue - Welcome to Hue
quickstart.cloudera:8888/accounts/login/?next=/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Sign in to continue to Hue
cloudera
Sign in

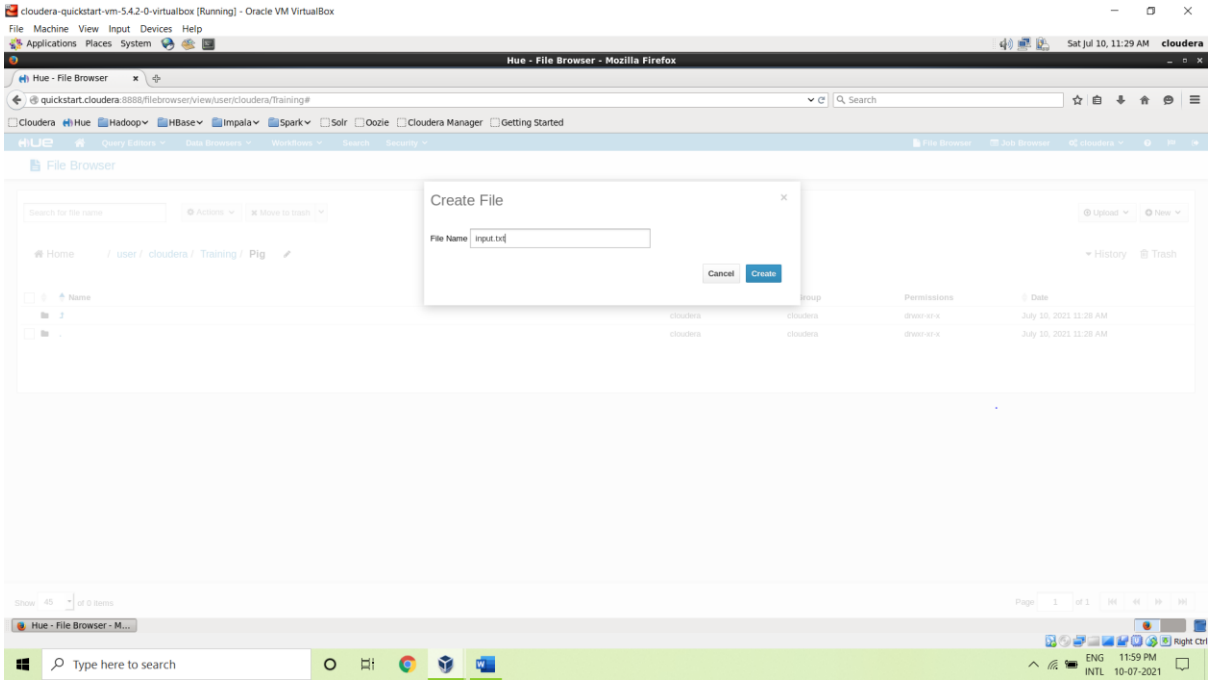

Hue - Welcome to Hue...
Type here to search
ENG
INTL
11:39 PM
10-07-2021

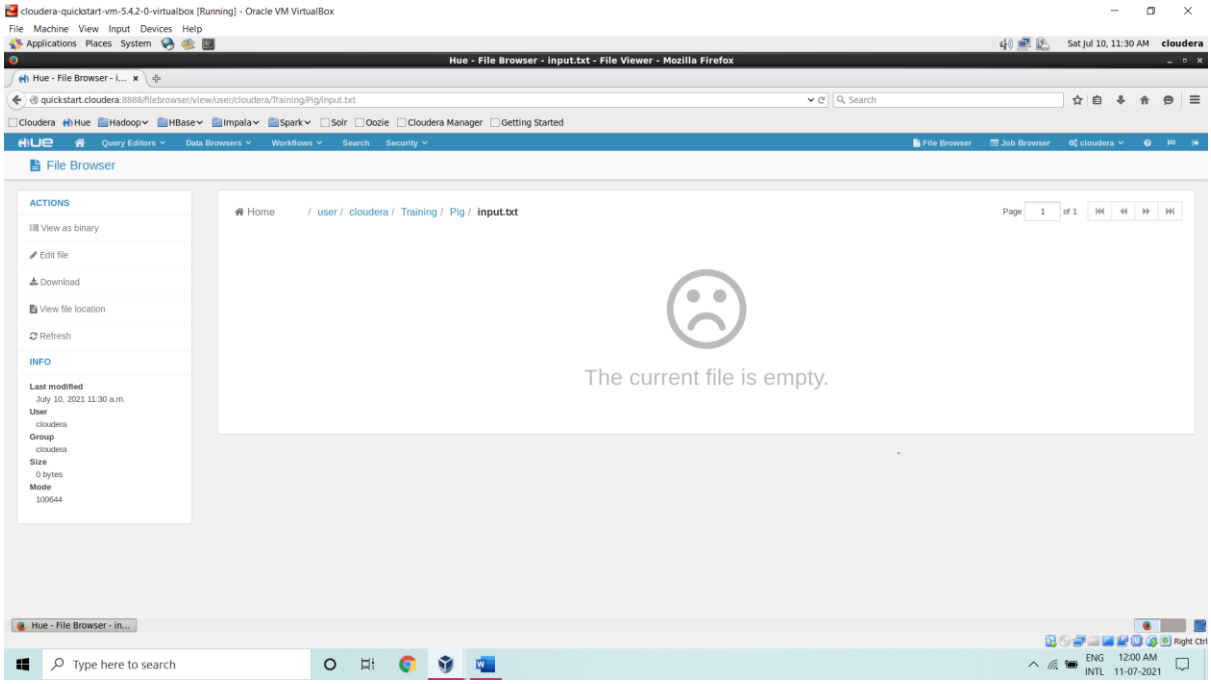

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:19 AM cloudera
Hue - File Browser - Mozilla Firefox
Hue - File Browser
quickstart.cloudera:8888/filebrowser/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security
File Browser Job Browser cloudera
File Browser
Search for file name
Actions
Move to trash
Upload
New
Home / user / cloudera
History
Trash
Name
Size
User
Group
Permissions
Date
hdfs
supergroup
drwxr-xr-x
June 09, 2015 03:38 AM
cloudera
cloudera
drwxr-xr-x
June 09, 2015 03:37 AM
Show 45 of 0 items
Page 1 of 1
Hue - File Browser - M...
Type here to search
ENG
INTL
11:49 PM
10-07-2021

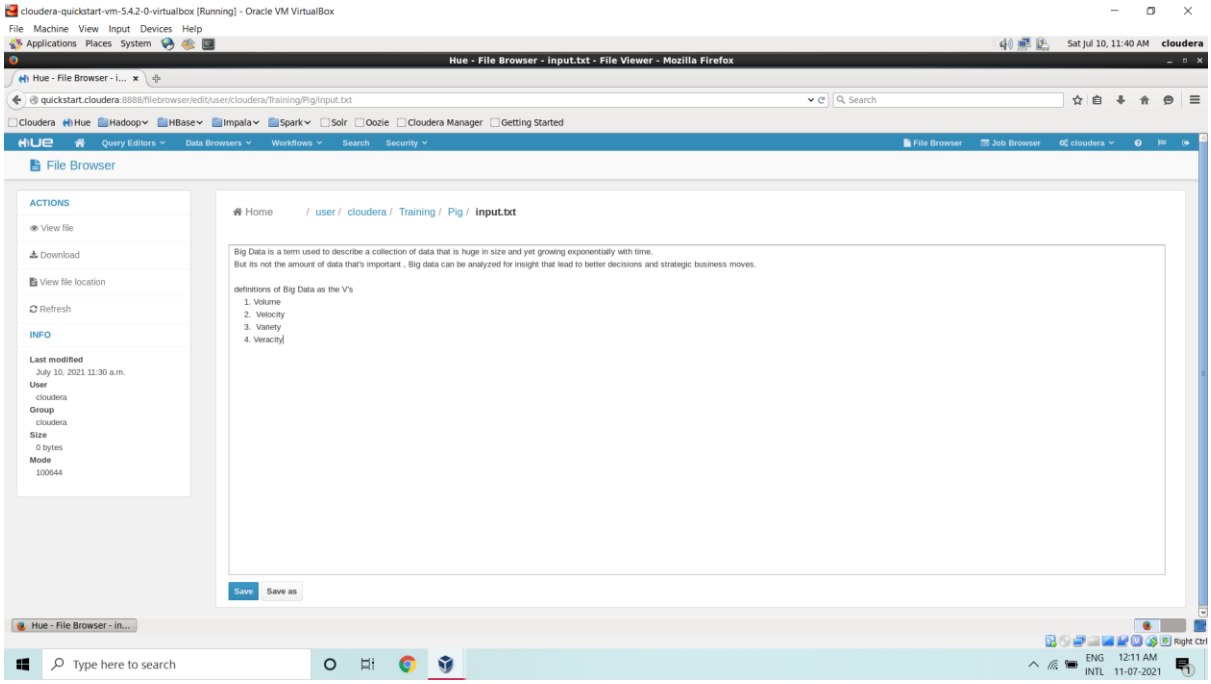

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:26 AM cloudera
Hue - File Browser - Mozilla Firefox
Hue - File Browser
quickstart.cloudera:8888/filebrowser/#
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security
File Browser Job Browser cloudera
File Browser
Create Directory
Directory Name Training
Cancel Create
Home / user / cloudera
Upload New
History Trash
Name Size User Group Permissions Date
hdfs supergroup drwxr-xr-x June 09, 2015 03:38 AM
cloudera cloudera drwxr-xr-x June 09, 2015 03:37 AM
Show 45 of 0 items
Page 1 of 1
Hue - File Browser - M...
Type here to search
ENG INTL 11:56 PM 10-07-2021

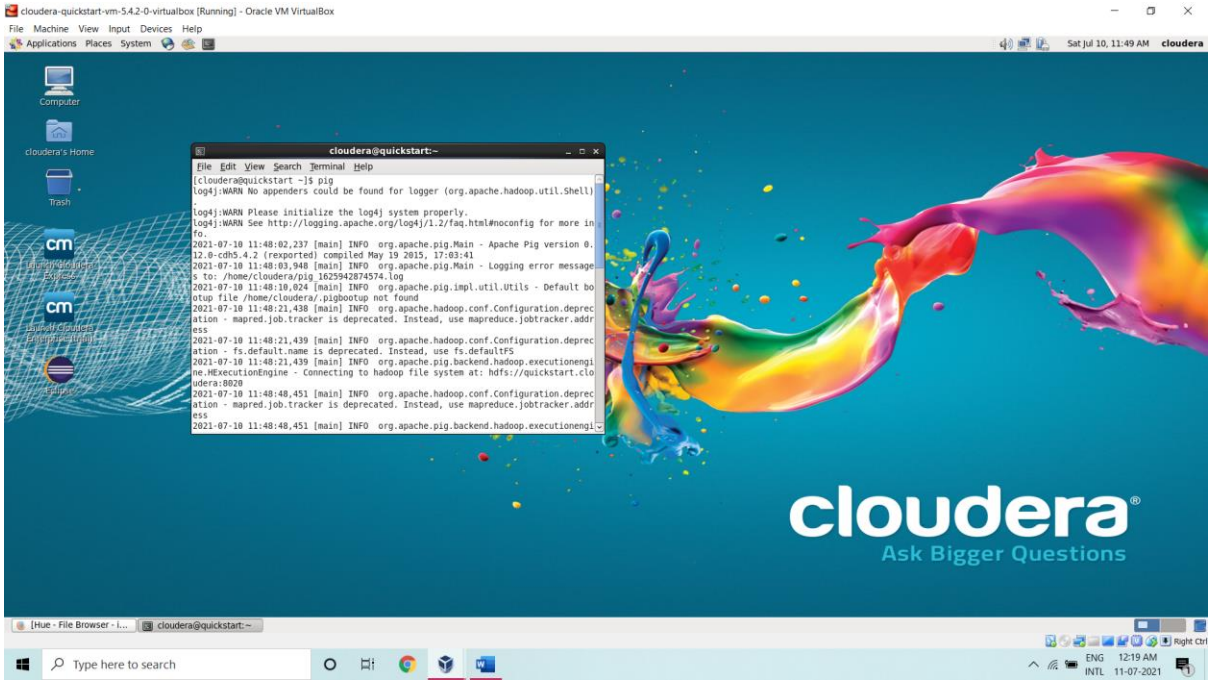

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:27 AM cloudera
Hue - File Browser - Mozilla Firefox
Hue - File Browser
quickstart.cloudera:8888/filebrowser/view/user/cloudera
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security
File Browser Job Browser cloudera
File Browser
Snipping Tool
Search for file name
Actions Move to trash
Upload New
Home / user / cloudera
History Trash
Name Size User Group Permissions Date
hdfs supergroup drwxr-xr-x July 10, 2021 11:19 AM
cloudera cloudera drwxr-xr-x July 10, 2021 11:26 AM
Training cloudera cloudera drwxr-xr-x July 10, 2021 11:26 AM
Show 45 of 1 items
Page 1 of 1
Hue - File Browser - M...
Type here to search
ENG INTL 11:57 PM 10-07-2021

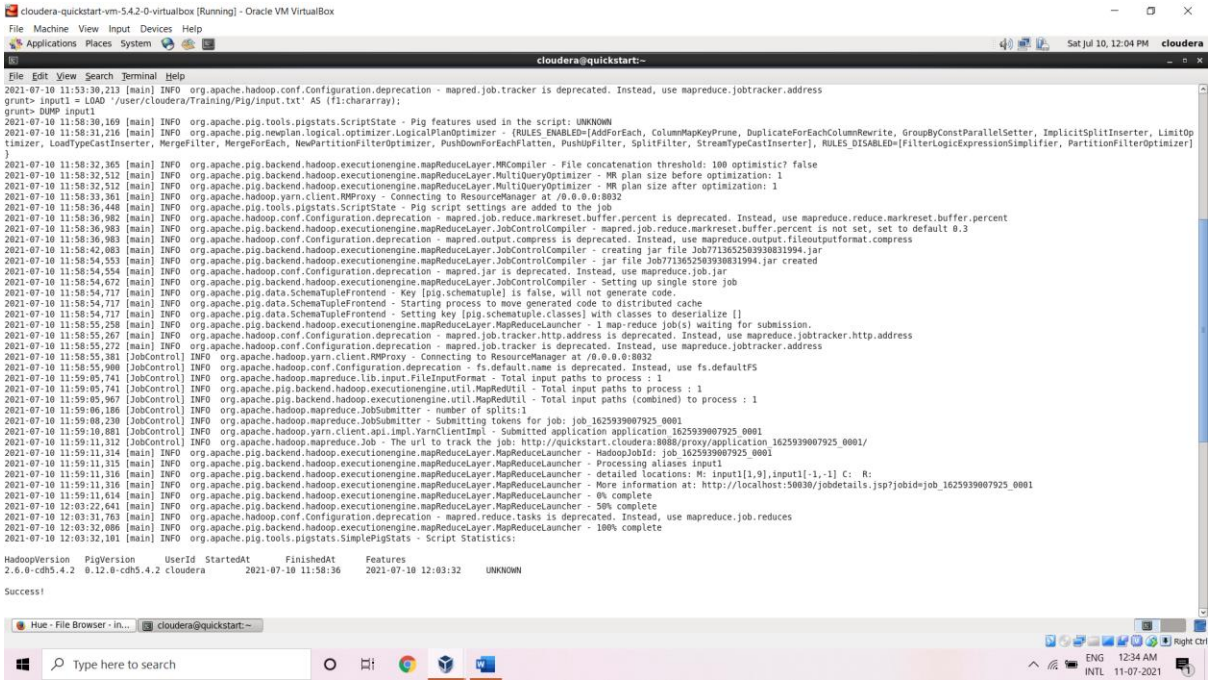

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:28 AM cloudera
Hue - File Browser - Mozilla Firefox
Hue - File Browser
quickstart.cloudera:8888/filebrowser/view/user/cloudera#
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
File Browser
Create Directory
Directory Name Pig
Cancel
Create
Home / user / cloudera / Training
Upload New
History Trash
Name
Permissions
Date
Hue - File Browser - M...
Type here to search
ENG INTL
11:58 PM
10-07-2021


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:29 AM cloudera
Hue - File Browser - Mozilla Firefox
Hue - File Browser
quickstart.cloudera:8888/filebrowser/view/user/cloudera/Training#
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
File Browser
Create File
File Name input.txt
Cancel
Create
Home / user / cloudera / Training / Pig
Upload New
History Trash
Name
Permissions
Date
Hue - File Browser - M...
Type here to search
ENG INTL
11:59 PM
10-07-2021

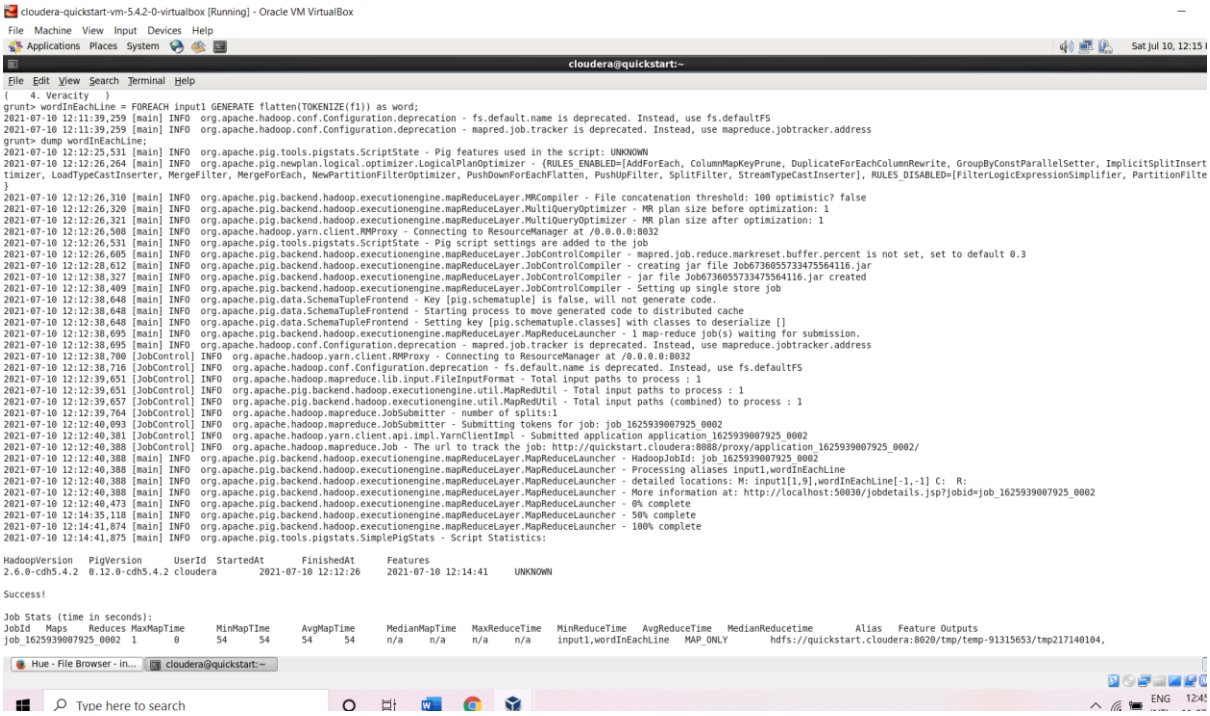

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:30 AM cloudera
Hue - File Browser - input.txt - File Viewer - Mozilla Firefox
Hue - File Browser - i...
quickstart.cloudera:8888/filebrowser/view/user/cloudera/Training/Pig/input.txt
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security
File Browser Job Browser cloudera
File Browser
ACTIONS
View as binary
Edit file
Download
View file location
Refresh
INFO
Last modified
July 10, 2021 11:30 a.m.
User
cloudera
Group
cloudera
Size
0 bytes
Mode
100644
Home / user / cloudera / Training / Pig / input.txt
Page 1 of 1
The current file is empty.
Hue - File Browser - in...
Type here to search
ENG INTL 12:00 AM 11-07-2021

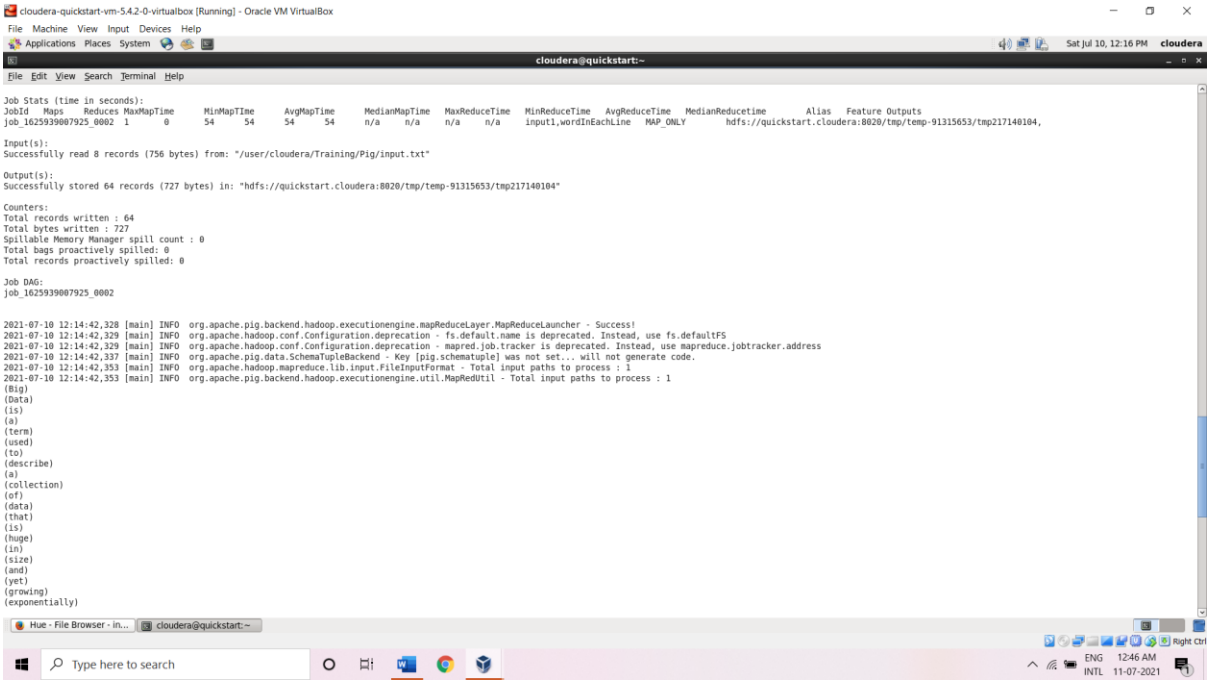

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:40 AM cloudera
Hue - File Browser - input.txt - File Viewer - Mozilla Firefox
Hue - File Browser - i...
quickstart.cloudera:8888/filebrowser/edit/user/cloudera/Training/Pig/input.txt
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security
File Browser Job Browser cloudera
File Browser
ACTIONS
View file
Download
View file location
Refresh
INFO
Last modified
July 10, 2021 11:30 a.m.
User
cloudera
Group
cloudera
Size
0 bytes
Mode
100644
Home / user / cloudera / Training / Pig / input.txt
Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.
But its not the amount of data that's important , Big data can be analyzed for insight that lead to better decisions and strategic business moves.
definitions of Big Data as the V's
1. Volume
2. Velocity
3. Variety
4. Veracity
Save
Save as
Hue - File Browser - in...
Type here to search
ENG INTL 12:11 AM 11-07-2021

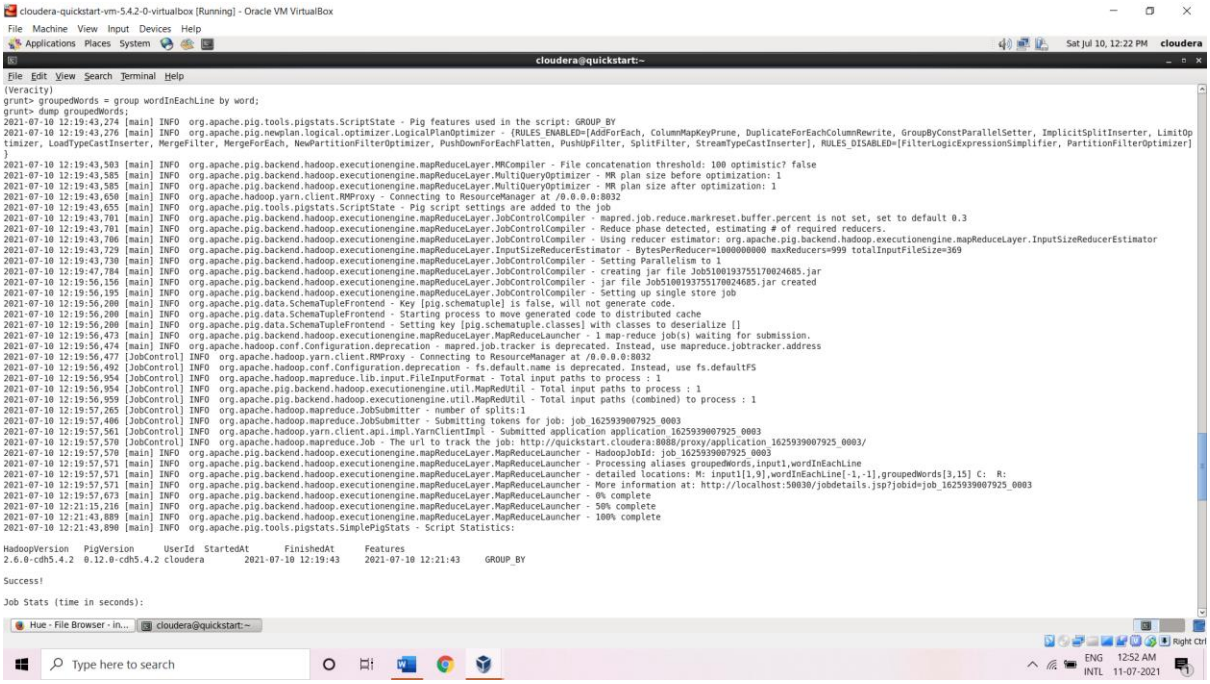

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 11:49 AM cloudera
Computer
cloudera's Home
Trash
Launch Cloudera Express
Launch Cloudera Enterprise
Eclipse
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2021-07-10 11:48:02,237 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2021-07-10 11:48:03,948 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1625942874574.log
2021-07-10 11:48:10,024 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/cloudera/.pigbootup not found
2021-07-10 11:48:21,438 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-10 11:48:21,439 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-10 11:48:21,439 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2021-07-10 11:48:48,451 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-10 11:48:48,451 [main] INFO org.apache.pig.backend.hadoop.executionengi
cloudera®
Ask Bigger Questions
[Hue - File Browser - i...
cloudera@quickstart:~
Type here to search
Right Ctrl
ENG INTL 12:19 AM 11-07-2021


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 12:04 PM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
2021-07-10 11:53:30,213 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
grunt> input1 = LOAD '/user/cloudera/Training/Pig/input.txt' AS (f1:chararray);
grunt> DUMP input1
2021-07-10 11:58:30,169 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2021-07-10 11:58:31,216 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-07-10 11:58:32,365 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-07-10 11:58:32,512 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-07-10 11:58:32,512 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-07-10 11:58:33,361 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-10 11:58:36,448 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-07-10 11:58:36,982 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2021-07-10 11:58:36,983 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-07-10 11:58:36,983 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputformat.compress
2021-07-10 11:58:42,083 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job7713652503930831994.jar
2021-07-10 11:58:54,553 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job7713652503930831994.jar created
2021-07-10 11:58:54,554 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.jar is deprecated. Instead, use mapreduce.job.jar
2021-07-10 11:58:54,672 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-07-10 11:58:54,717 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-07-10 11:58:54,717 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2021-07-10 11:58:54,717 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2021-07-10 11:58:55,258 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-07-10 11:58:55,267 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracker.http.address
2021-07-10 11:58:55,272 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-10 11:58:55,381 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-10 11:58:55,900 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-10 11:59:05,741 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-10 11:59:05,741 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-07-10 11:59:05,967 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2021-07-10 11:59:06,186 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2021-07-10 11:59:08,230 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1625939007925_0001
2021-07-10 11:59:10,881 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1625939007925_0001
2021-07-10 11:59:11,312 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://quickstart.cloudera:8088/proxy/application_1625939007925_0001/
2021-07-10 11:59:11,314 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1625939007925_0001
2021-07-10 11:59:11,315 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases input1
2021-07-10 11:59:11,316 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: input1[1,9],input1[-1,-1] C: R:
2021-07-10 11:59:11,316 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?jobid=job_1625939007925_0001
2021-07-10 11:59:11,614 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2021-07-10 12:03:22,641 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2021-07-10 12:03:31,763 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2021-07-10 12:03:32,086 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-07-10 12:03:32,101 [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 2021-07-10 11:58:36 2021-07-10 12:03:32 UNKNOWN
Success!
Hue - File Browser - in...
cloudera@quickstart:~
Type here to search
Right Ctrl
ENG INTL 12:34 AM 11-07-2021

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1625939007925_0001  1       0       70      70      70      70      n/a     n/a     n/a     n/a     input1  MAP_ONLY        hdfs://quickstart.cloudera:8020/tmp/temp-91315653/tmp-1510987010,

Input(s):
Successfully read 8 records (756 bytes) from: "/user/cloudera/Training/Pig/input.txt"

Output(s):
Successfully stored 8 records (416 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-91315653/tmp-1510987010"

Counters:
Total records written : 8
Total bytes written : 416
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1625939007925_0001


2021-07-10 12:03:32,310 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-07-10 12:03:32,449 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-10 12:03:32,449 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-10 12:03:32,473 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-07-10 12:03:32,842 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-10 12:03:32,842 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. )
(But its not the amount of data that's important , Big data can be analyzed for insight that lead to better decisions and strategic business moves.)
()
(definitions of Big Data as the V's )
(    1. Volume)
(    2.  Velocity)
(    3.  Variety)
(    4. Veracity   )
grunt>
```




cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 12:16 PM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1625939007925_0002 1 0 54 54 54 n/a n/a n/a n/a input1,wordInEachLine MAP_ONLY hdfs://quickstart.cloudera:8020/tmp/temp-91315653/tmp217140104,
Input(s):
Successfully read 8 records (756 bytes) from: "/user/cloudera/Training/Pig/input.txt"
Output(s):
Successfully stored 64 records (727 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-91315653/tmp217140104"
Counters:
Total records written : 64
Total bytes written : 727
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_1625939007925_0002
2021-07-10 12:14:42,328 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-07-10 12:14:42,329 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-10 12:14:42,329 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-10 12:14:42,337 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-07-10 12:14:42,353 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-10 12:14:42,353 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Big)
(Data)
(is)
(a)
(term)
(used)
(to)
(describe)
(a)
(collection)
(of)
(data)
(that)
(is)
(huge)
(in)
(size)
(and)
(yet)
(growing)
(exponentially)
Hue - File Browser - in...
cloudera@quickstart:~
Type here to search
ENG INTL 12:46 AM 11-07-2021


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sat Jul 10, 12:22 PM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
(Veracity)
grunt> groupedWords = group wordInEachLine by word;
grunt> dump groupedWords;
2021-07-10 12:19:43,274 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2021-07-10 12:19:43,276 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-07-10 12:19:43,503 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-07-10 12:19:43,585 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-07-10 12:19:43,585 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-07-10 12:19:43,650 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-10 12:19:43,655 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-07-10 12:19:43,701 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-07-10 12:19:43,701 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2021-07-10 12:19:43,706 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2021-07-10 12:19:43,729 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=369
2021-07-10 12:19:43,730 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2021-07-10 12:19:47,784 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job5100193755170024685.jar
2021-07-10 12:19:56,156 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job5100193755170024685.jar created
2021-07-10 12:19:56,195 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-07-10 12:19:56,200 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-07-10 12:19:56,200 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2021-07-10 12:19:56,200 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2021-07-10 12:19:56,473 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-07-10 12:19:56,474 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-10 12:19:56,477 [JobControl] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-10 12:19:56,492 [JobControl] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-10 12:19:56,954 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-10 12:19:56,954 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-07-10 12:19:56,959 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2021-07-10 12:19:57,265 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2021-07-10 12:19:57,406 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1625939007925_0003
2021-07-10 12:19:57,561 [JobControl] INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1625939007925_0003
2021-07-10 12:19:57,570 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://quickstart.cloudera:8088/proxy/application_1625939007925_0003/
2021-07-10 12:19:57,570 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1625939007925_0003
2021-07-10 12:19:57,571 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases groupedWords,input1,wordInEachLine
2021-07-10 12:19:57,571 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: input1[1,9],wordInEachLine[-1,-1],groupedWords[3,15] C: R:
2021-07-10 12:19:57,571 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?jobid=job_1625939007925_0003
2021-07-10 12:19:57,673 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2021-07-10 12:21:15,216 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2021-07-10 12:21:43,889 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-07-10 12:21:43,890 [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 2021-07-10 12:19:43 2021-07-10 12:21:43 GROUP_BY
Success!
Job Stats (time in seconds):
Hue - File Browser - in...
cloudera@quickstart:~
Type here to search
ENG INTL 12:52 AM 11-07-2021

cloudera@quickstart:~

File Edit View Search Terminal Help

```
(3.,{(3.)})
(4.,{(4.)})
(as,{(as)})
(be,{(be)})
(in,{(in)})
(is,{(is),(is)})
(of,{(of),(of),(of)})
(to,{(to),(to)})
(Big,{(Big),(Big),(Big)})
(But,{(But)})
(V's,{(V's)})
(and,{(and),(and)})
(can,{(can)})
(for,{(for)})
(its,{(its)})
(not,{(not)})
(the,{(the),(the)})
(yet,{(yet)})
(Data,{(Data),(Data)})
(data,{(data),(data),(data)})
(huge,{(huge)})
(lead,{(lead)})
(size,{(size)})
(term,{(term)})
(that,{(that),(that)})
(used,{(used)})
(with,{(with)})
(time.,{(time.)})
(Volume,{(Volume)})
(amount,{(amount)})
(better,{(better)})
(moves.,{(moves.)})
(that's,{(that's)})
(Variety,{(Variety)})
(growing,{(growing)})
(insight,{(insight)})
(Velocity,{(Velocity)})
(Veracity,{(Veracity)})
(analyzed,{(analyzed)})
(business,{(business)})
(describe,{(describe)})
(decisions,{(decisions)})
(important,{(important)})
(strategic,{(strategic)})
(collection,{(collection)})
(definitions,{(definitions)})
(exponentially,{(exponentially)})
(,{()})
grunt>
```

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox

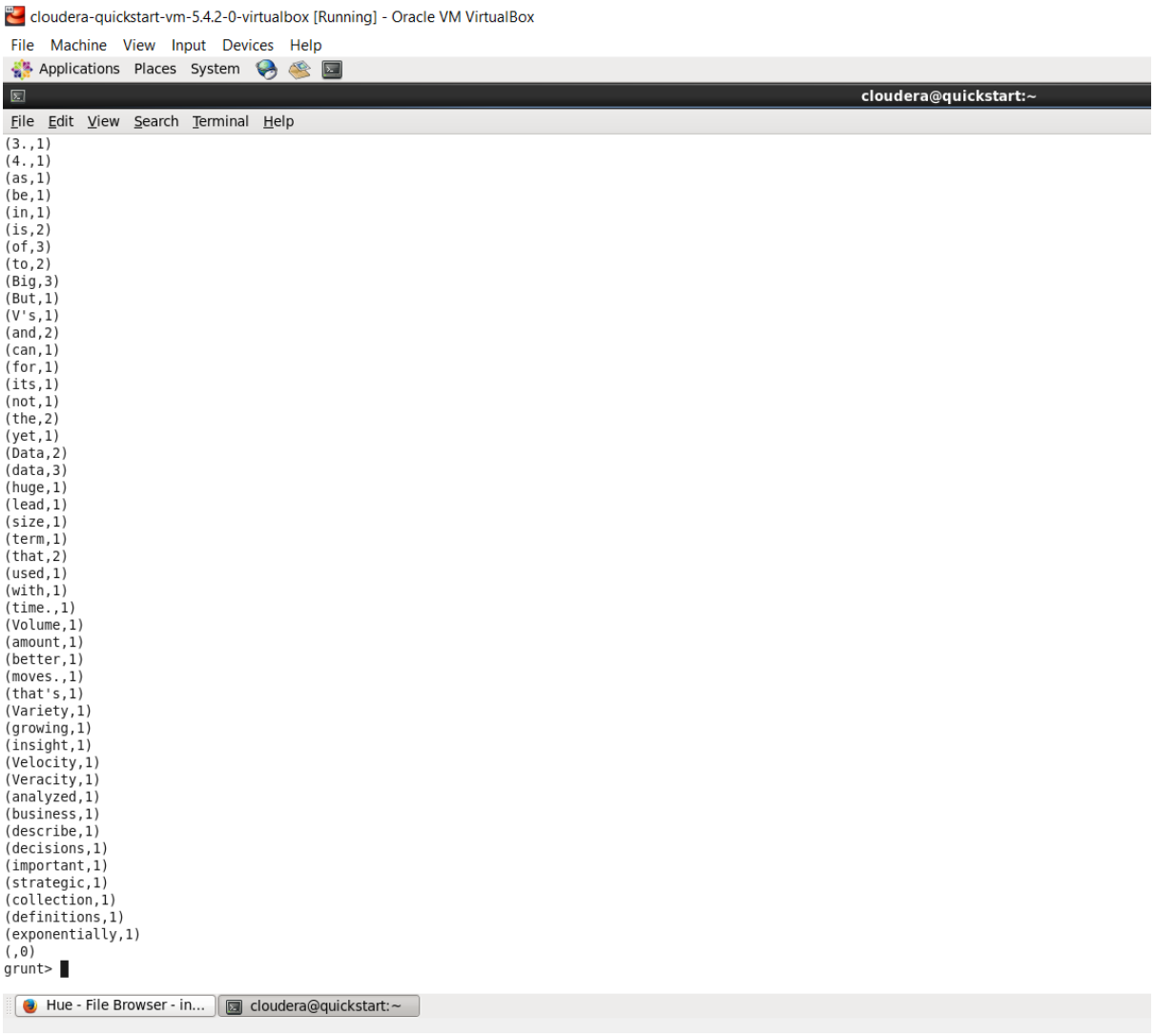
File Machine View Input Devices Help

Applications Places System

cloudera@quickstart:~

File Edit View Search Terminal Help

```
(exponentially,{(exponentially)})
(,{()})
grunt> countedWords = foreach groupedWords generate group, COUNT(wordInEachLine);
grunt> dump countedWords;
2021-07-10 12:25:44,631 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2021-07-10 12:25:44,635 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, Impl
timizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier,
}
2021-07-10 12:25:44,653 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-07-10 12:25:44,700 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to combiner
2021-07-10 12:25:45,058 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-07-10 12:25:45,058 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-07-10 12:25:45,124 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-10 12:25:45,130 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-07-10 12:25:45,179 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-07-10 12:25:45,180 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of required reducers.
2021-07-10 12:25:45,180 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputS
2021-07-10 12:25:45,191 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=369
2021-07-10 12:25:45,192 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2021-07-10 12:25:56,704 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job4270952580853334023.jar
2021-07-10 12:25:56,315 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job4270952580853334023.jar created
2021-07-10 12:25:56,360 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2021-07-10 12:25:56,370 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2021-07-10 12:25:56,370 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2021-07-10 12:25:56,370 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2021-07-10 12:25:56,483 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2021-07-10 12:25:56,483 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-07-10 12:25:56,487 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2021-07-10 12:25:56,499 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-07-10 12:25:56,916 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2021-07-10 12:25:56,916 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-07-10 12:25:56,922 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2021-07-10 12:25:57,022 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2021-07-10 12:25:57,310 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1625939007925_0004
2021-07-10 12:25:57,467 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1625939007925_0004
2021-07-10 12:25:57,483 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://quickstart.cloudera:8088/proxy/application_1625939007925_0004/
2021-07-10 12:25:57,483 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1625939007925_0004
2021-07-10 12:25:57,483 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases countedWords,groupedWords,input1,wordInEachLine
2021-07-10 12:25:57,483 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: input1[1,9],wordInEachLine[-1,-1],countedWords[4,15],groupedWords[3,15
oupedWords[3,15] R: countedWords[4,15]
2021-07-10 12:25:57,483 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information at: http://localhost:50030/jobdetails.jsp?jobid=job_1625939007925_0004
2021-07-10 12:25:57,636 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2021-07-10 12:27:09,118 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2021-07-10 12:27:44,238 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-07-10 12:27:44,239 [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        2021-07-10 12:25:45     2021-07-10 12:27:44     GROUP_BY
```


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
(3.,1)
(4.,1)
(as,1)
(be,1)
(in,1)
(is,2)
(of,3)
(to,2)
(Big,3)
(But,1)
(V's,1)
(and,2)
(can,1)
(for,1)
(its,1)
(not,1)
(the,2)
(yet,1)
(Data,2)
(data,3)
(huge,1)
(lead,1)
(size,1)
(term,1)
(that,2)
(used,1)
(with,1)
(time.,1)
(Volume,1)
(amount,1)
(better,1)
(moves.,1)
(that's,1)
(Variety,1)
(growing,1)
(insight,1)
(Velocity,1)
(Veracity,1)
(analyzed,1)
(business,1)
(describe,1)
(decisions,1)
(important,1)
(strategic,1)
(collection,1)
(definitions,1)
(exponentially,1)
(,0)
grunt>
Hue - File Browser - in...
cloudera@quickstart:~