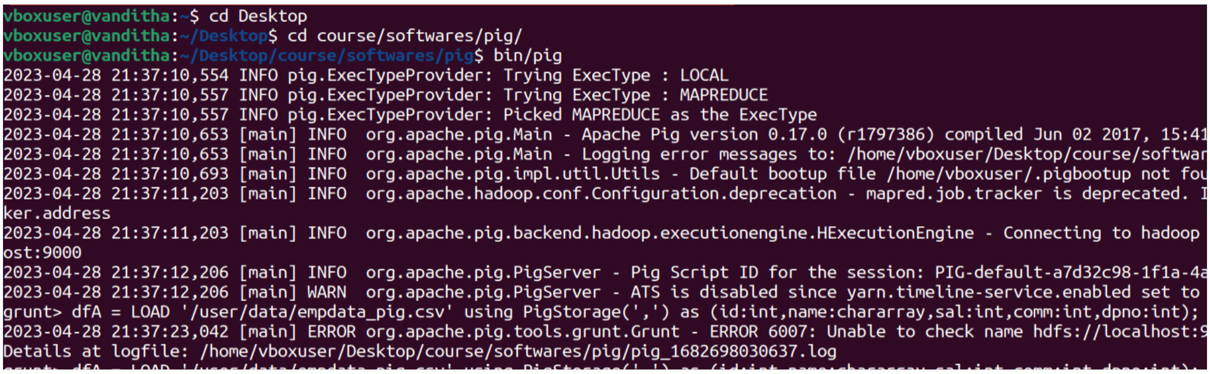
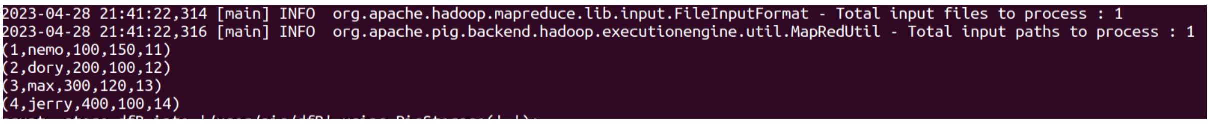
PIG SCRIPTING CODES:

Load Data: loading the csv file from HDFS to pig with a given schema
Dump the loaded data dfA.



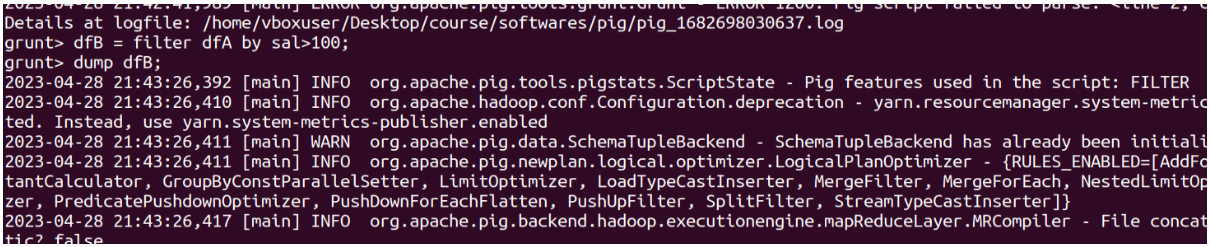
```
vboxuser@vanditha:~$ cd Desktop
vboxuser@vanditha:~/Desktop$ cd course/softwares/pig/
vboxuser@vanditha:~/Desktop/course/softwares/pig$ bin/pig
2023-04-28 21:37:10,554 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2023-04-28 21:37:10,557 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2023-04-28 21:37:10,557 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-04-28 21:37:10,653 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41
2023-04-28 21:37:10,653 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/vboxuser/Desktop/course/softwar
2023-04-28 21:37:10,693 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/vboxuser/.pigbootup not fou
2023-04-28 21:37:11,203 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. I
ker.address
2023-04-28 21:37:11,203 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop
ost:9000
2023-04-28 21:37:12,206 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-a7d32c98-1f1a-4a
2023-04-28 21:37:12,206 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to
grunt> dfA = LOAD '/user/data/empdata_pig.csv' using PigStorage(',') as (id:int,name:chararray,sal:int,comm:int,dpno:int);
2023-04-28 21:37:23,042 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 6007: Unable to check name hdfs://localhost:9
Details at logfile: /home/vboxuser/Desktop/course/softwares/pig/pig_1682698030637.log
grunt> dfA = LOAD '/user/data/empdata_pig.csv' using PigStorage(',') as (id:int,name:chararray,sal:int,comm:int,dpno:int);
```



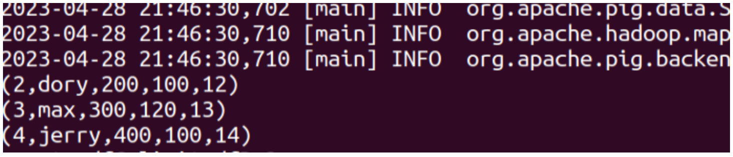
```
2023-04-28 21:41:22,314 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-04-28 21:41:22,316 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,nemo,100,150,11)
(2,dory,200,100,12)
(3,max,300,120,13)
(4,jerry,400,100,14)
```

Aggregate (by row):

Filter the contents of dfA by condition salary>100.
Dump dfB;

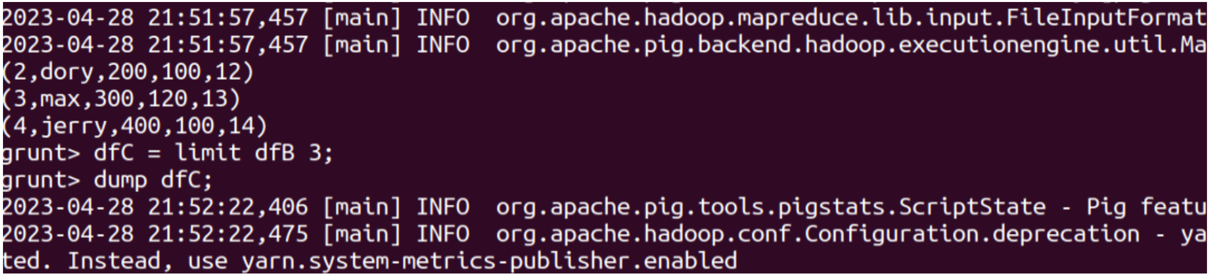


```
2023-04-28 21:42:41,989 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse: <line 2, c
Details at logfile: /home/vboxuser/Desktop/course/softwares/pig/pig_1682698030637.log
grunt> dfB = filter dfA by sal>100;
grunt> dump dfB;
2023-04-28 21:43:26,392 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2023-04-28 21:43:26,410 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metric
ted. Instead, use yarn.system-metrics-publisher.enabled
2023-04-28 21:43:26,411 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initiali
2023-04-28 21:43:26,411 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddFo
tantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOp
zer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2023-04-28 21:43:26,417 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concat
tic? false
```



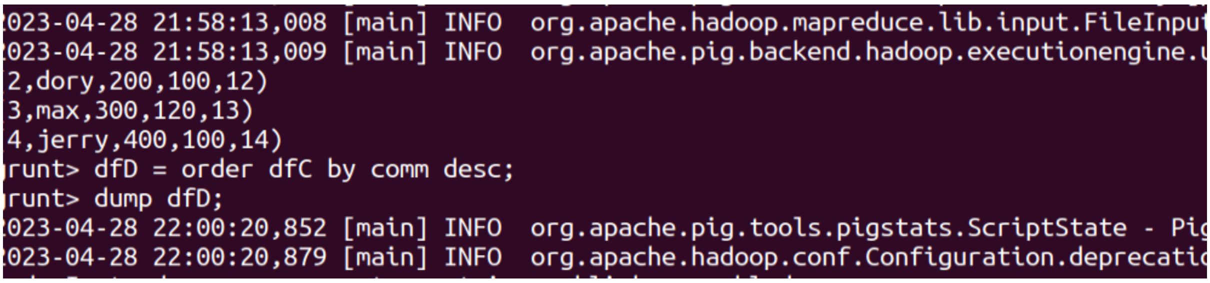
```
2023-04-28 21:46:30,702 [main] INFO  org.apache.pig.data.S
2023-04-28 21:46:30,710 [main] INFO  org.apache.hadoop.map
2023-04-28 21:46:30,710 [main] INFO  org.apache.pig.backen
(2,dory,200,100,12)
(3,max,300,120,13)
(4,jerry,400,100,14)
```

Get the first 3 rows of dfC using limit operation.
Dump dfC.



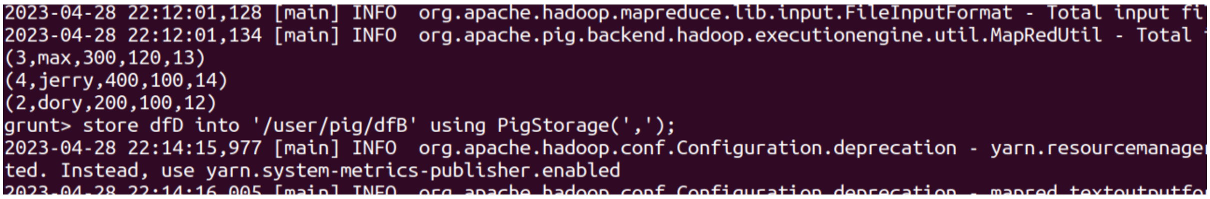
```
2023-04-28 21:51:57,457 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat
2023-04-28 21:51:57,457 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.Ma
(2,dory,200,100,12)
(3,max,300,120,13)
(4,jerry,400,100,14)
grunt> dfC = limit dfB 3;
grunt> dump dfC;
2023-04-28 21:52:22,406 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig featu
2023-04-28 21:52:22,475 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - ya
ted. Instead, use yarn.system-metrics-publisher.enabled
```

dfD contains all Sorted rows from dfC in descending order
Dump dfD

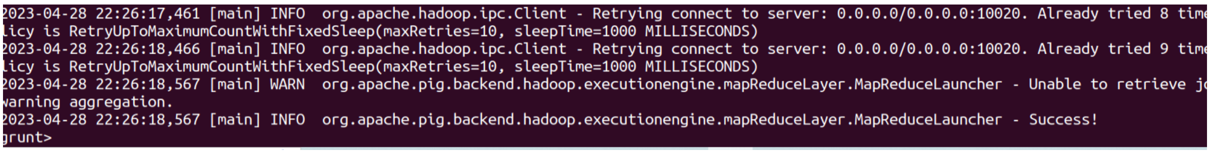


```
023-04-28 21:58:13,008 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInput
023-04-28 21:58:13,009 [main] INFO  org.apache.pig.backend.hadoop.executionengine.u
2,dory,200,100,12)
3,max,300,120,13)
4,jerry,400,100,14)
runt> dfD = order dfC by comm desc;
runt> dump dfD;
023-04-28 22:00:20,852 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig
023-04-28 22:00:20,879 [main] INFO  org.apache.hadoop.conf.Configuration.deprecatio
```

Store the dfD data



```
2023-04-28 22:12:01,128 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input fi
2023-04-28 22:12:01,134 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total
(3,max,300,120,13)
(4,jerry,400,100,14)
(2,dory,200,100,12)
grunt> store dfD into '/user/pig/dfB' using PigStorage(',');
2023-04-28 22:14:15,977 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanage
ted. Instead, use yarn.system-metrics-publisher.enabled
```

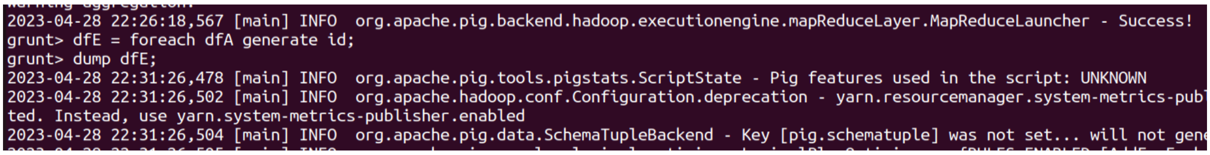


```
2023-04-28 22:26:17,461 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 8 time
licy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2023-04-28 22:26:18,466 [main] INFO  org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. Already tried 9 time
licy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2023-04-28 22:26:18,567 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Unable to retrieve jo
warning aggregation.
2023-04-28 22:26:18,567 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

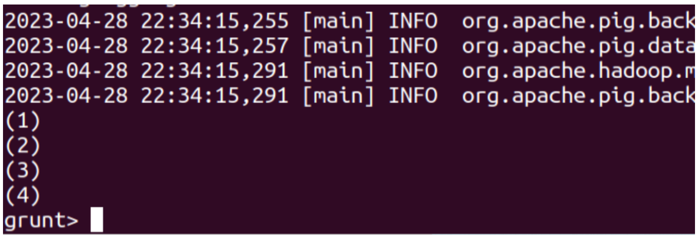
Transform (by column):

Select existing column

Select only the id column from dfA and load it into dfE
Dump dfE



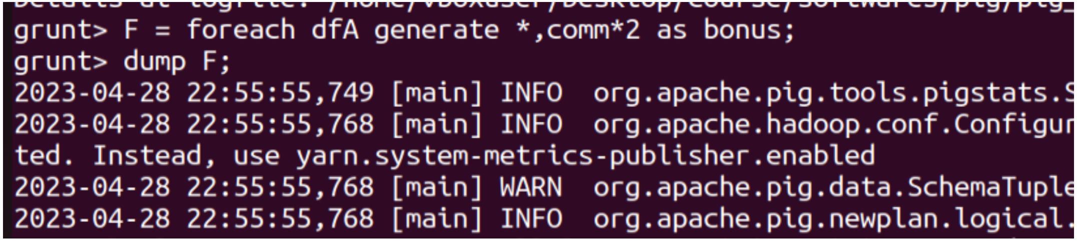
```
2023-04-28 22:26:18,567 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> dfE = foreach dfA generate id;
grunt> dump dfE;
2023-04-28 22:31:26,478 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2023-04-28 22:31:26,502 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-pub
ted. Instead, use yarn.system-metrics-publisher.enabled
2023-04-28 22:31:26,504 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not gene
```



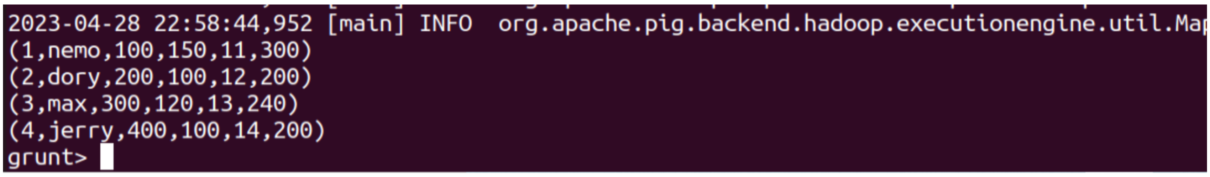
```
2023-04-28 22:34:15,255 [main] INFO  org.apache.pig.back
2023-04-28 22:34:15,257 [main] INFO  org.apache.pig.data
2023-04-28 22:34:15,291 [main] INFO  org.apache.hadoop.m
2023-04-28 22:34:15,291 [main] INFO  org.apache.pig.back
(1)
(2)
(3)
(4)
grunt> █
```

Create new column based on existing column

Create a table F where new columns are created based on some operation on existing table dfA
Dump F

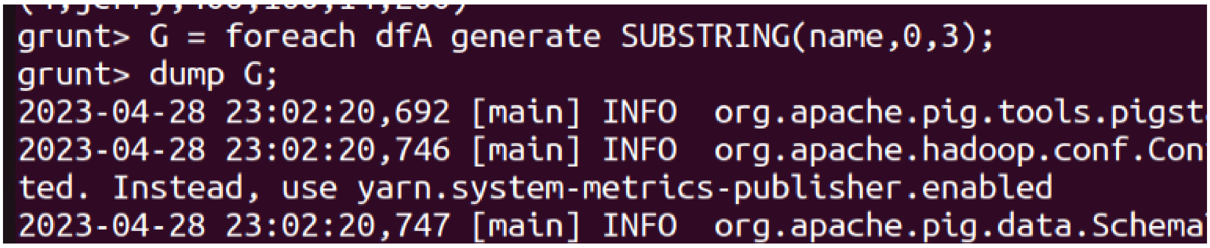


```
grunt> F = foreach dfA generate *,comm*2 as bonus;
grunt> dump F;
2023-04-28 22:55:55,749 [main] INFO  org.apache.pig.tools.pigstats.S
2023-04-28 22:55:55,768 [main] INFO  org.apache.hadoop.conf.Configur
ted. Instead, use yarn.system-metrics-publisher.enabled
2023-04-28 22:55:55,768 [main] WARN  org.apache.pig.data.SchemaTuple
2023-04-28 22:55:55,768 [main] INFO  org.apache.pig.newplan.logical.
```

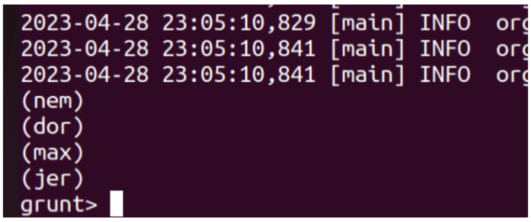


```
2023-04-28 22:58:44,952 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.Map
(1,nemo,100,150,11,300)
(2,dory,200,100,12,200)
(3,max,300,120,13,240)
(4,jerry,400,100,14,200)
grunt>
```

Transform columns:

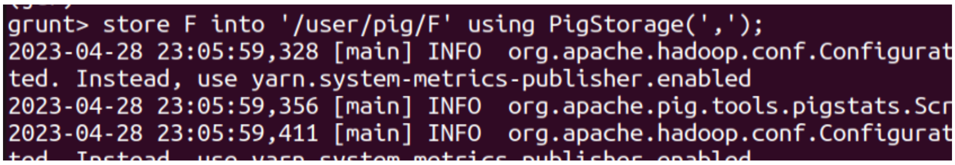


```
grunt> G = foreach dfA generate SUBSTRING(name,0,3);
grunt> dump G;
2023-04-28 23:02:20,692 [main] INFO  org.apache.pig.tools.pigst
2023-04-28 23:02:20,746 [main] INFO  org.apache.hadoop.conf.Con
ted. Instead, use yarn.system-metrics-publisher.enabled
2023-04-28 23:02:20,747 [main] INFO  org.apache.pig.data.Schema
```

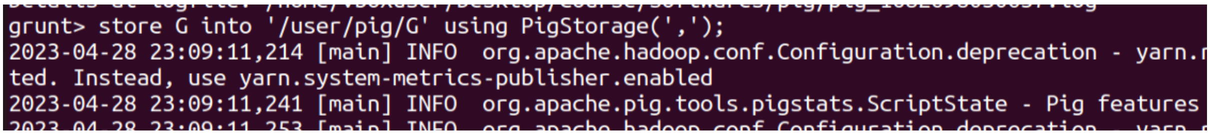


```
2023-04-28 23:05:10,829 [main] INFO  org
2023-04-28 23:05:10,841 [main] INFO  org
2023-04-28 23:05:10,841 [main] INFO  org
(nem)
(dor)
(max)
(jer)
grunt>
```

Store Data:



```
grunt> store F into '/user/pig/F' using PigStorage(',');
2023-04-28 23:05:59,328 [main] INFO  org.apache.hadoop.conf.Configurat
ted. Instead, use yarn.system-metrics-publisher.enabled
2023-04-28 23:05:59,356 [main] INFO  org.apache.pig.tools.pigstats.Scr
2023-04-28 23:05:59,411 [main] INFO  org.apache.hadoop.conf.Configurat
```



```
grunt> store G into '/user/pig/G' using PigStorage(',');
2023-04-28 23:09:11,214 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.
ted. Instead, use yarn.system-metrics-publisher.enabled
2023-04-28 23:09:11,241 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
```