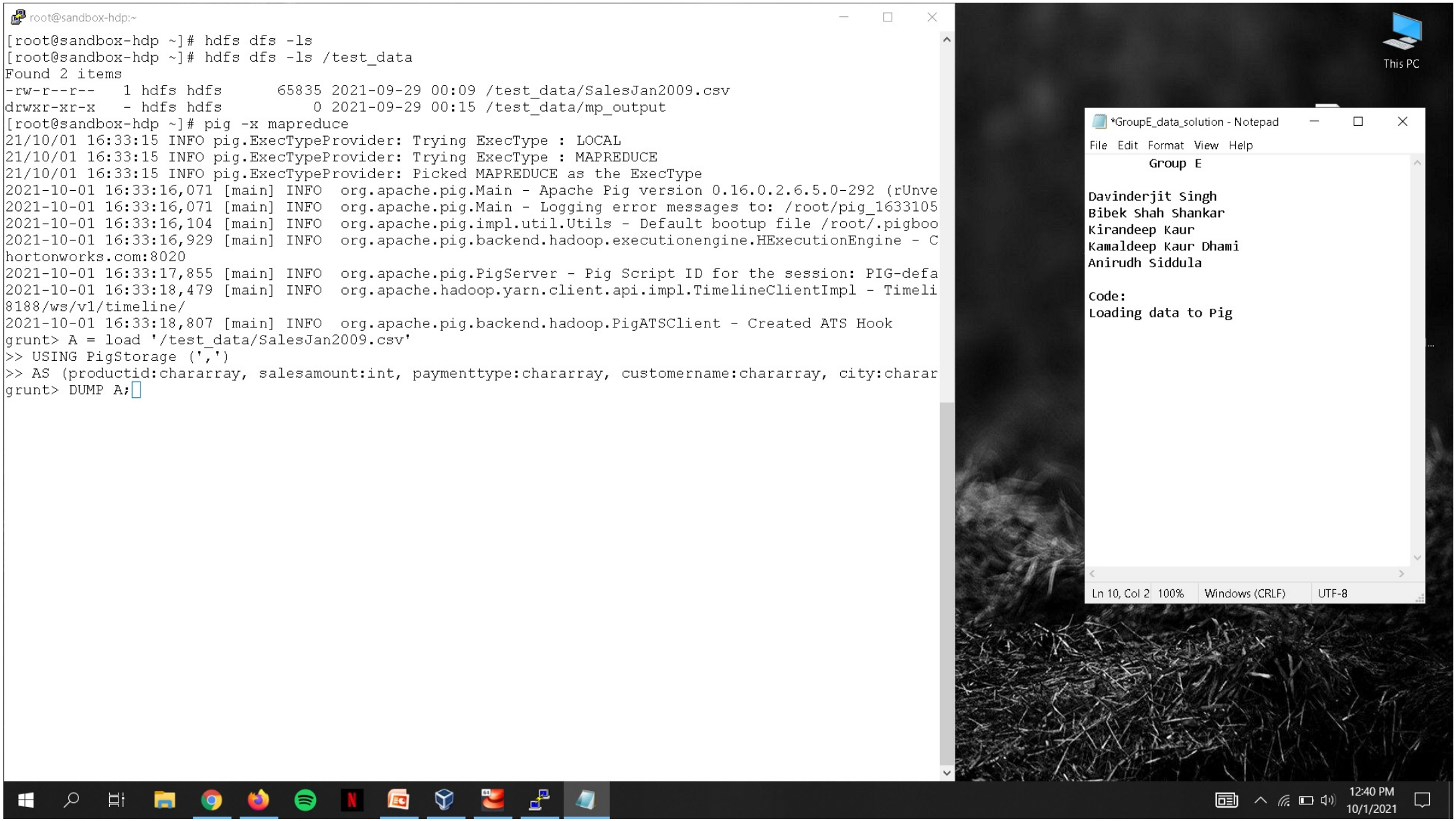
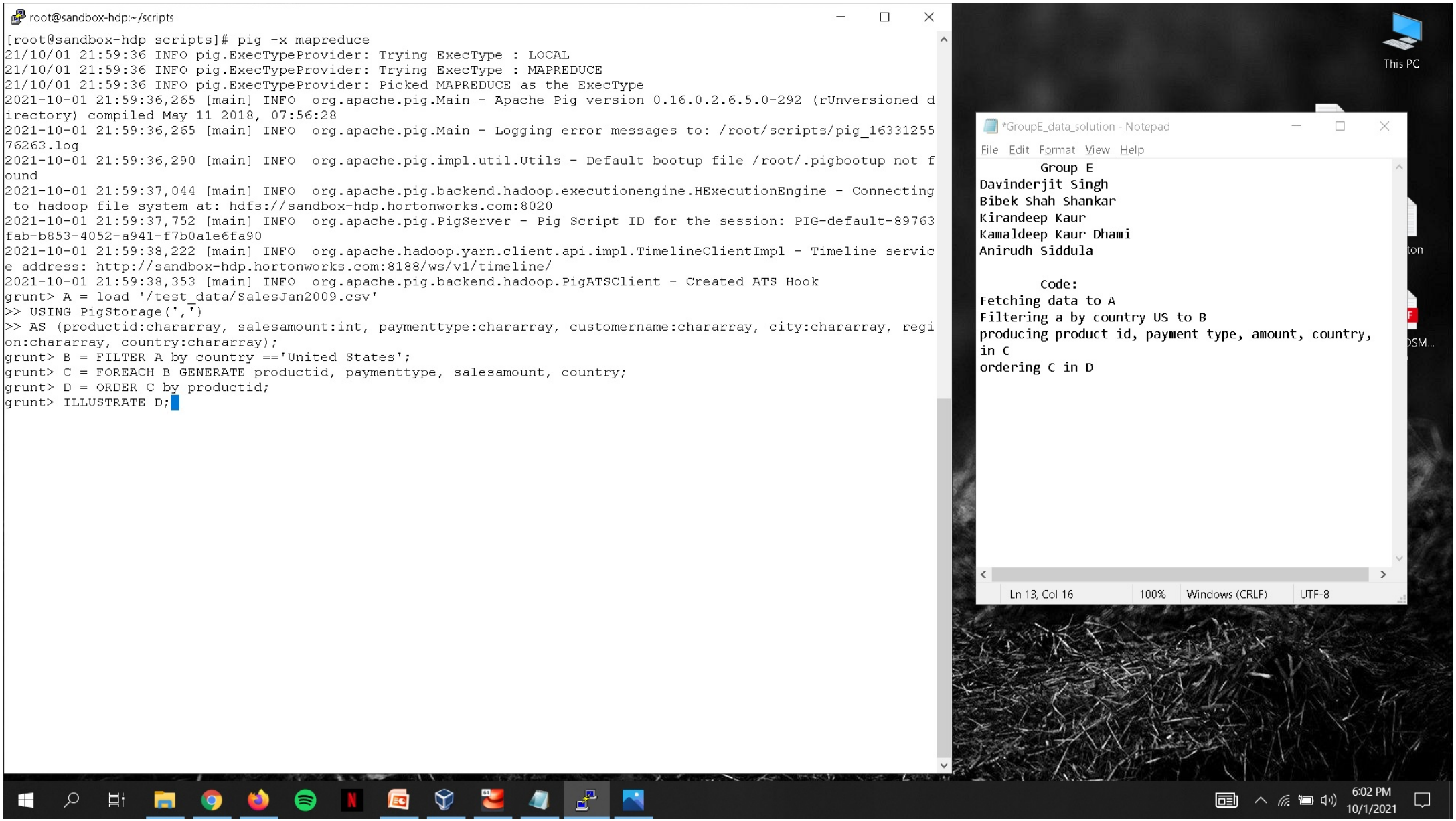
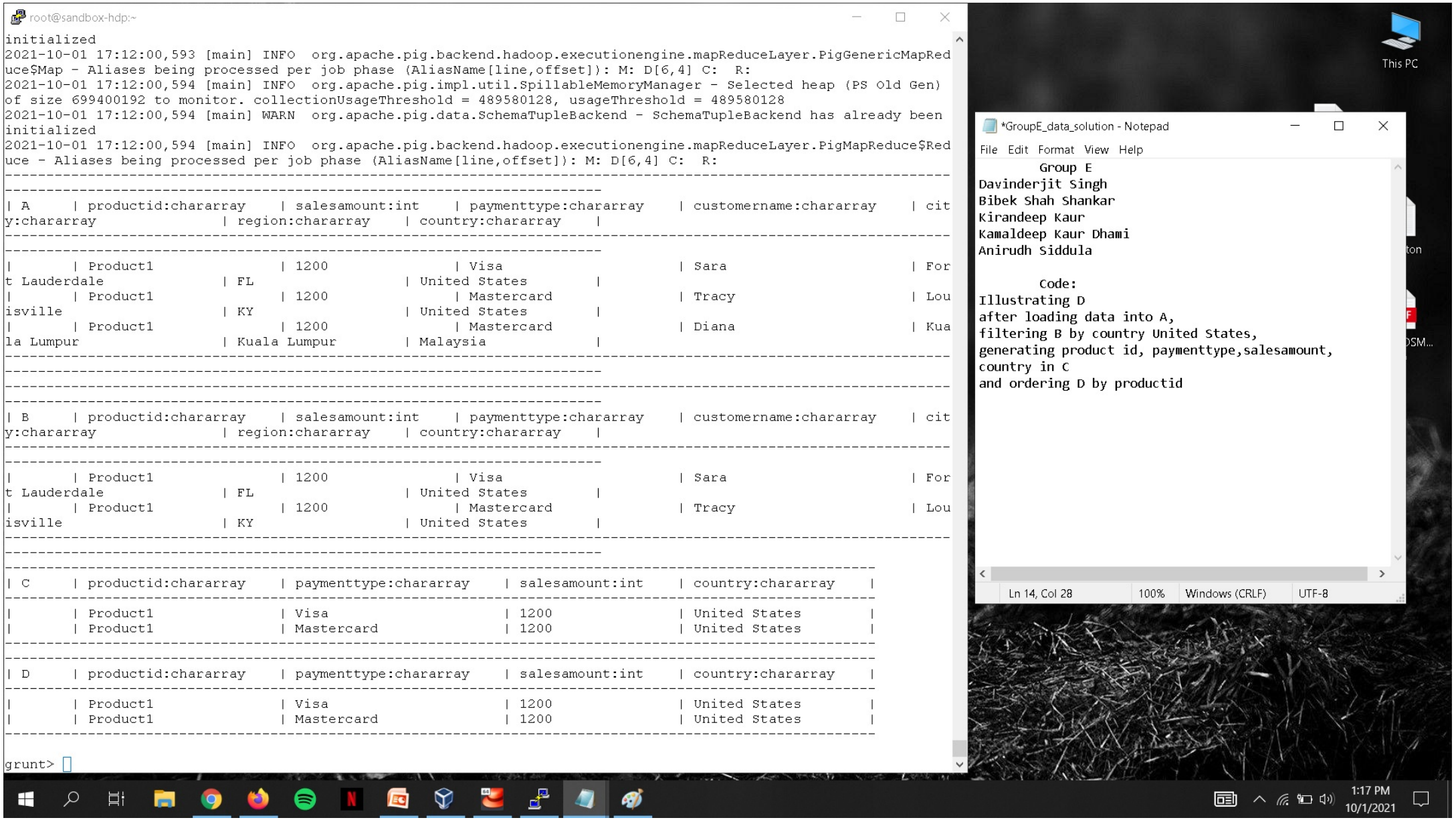
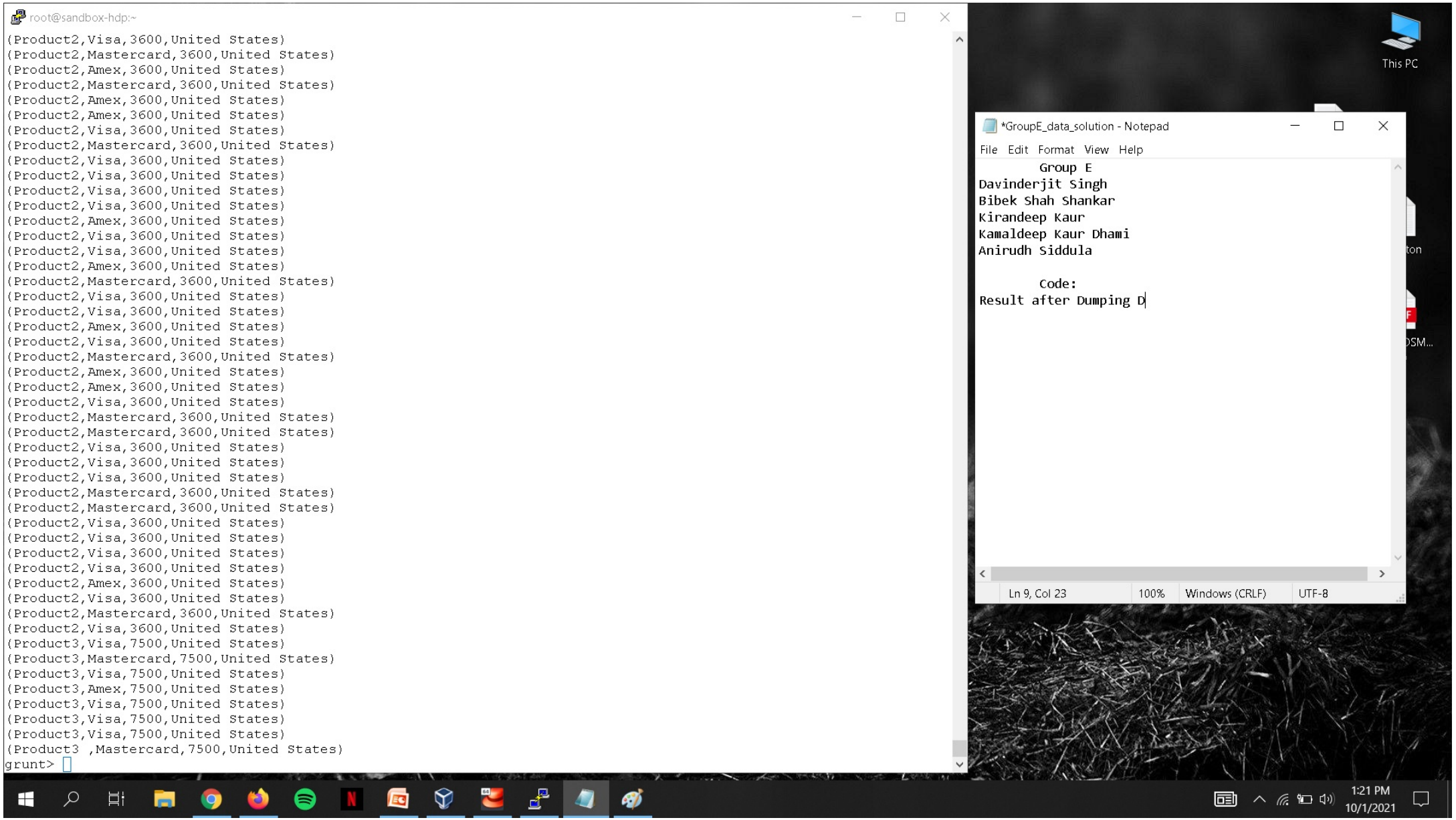
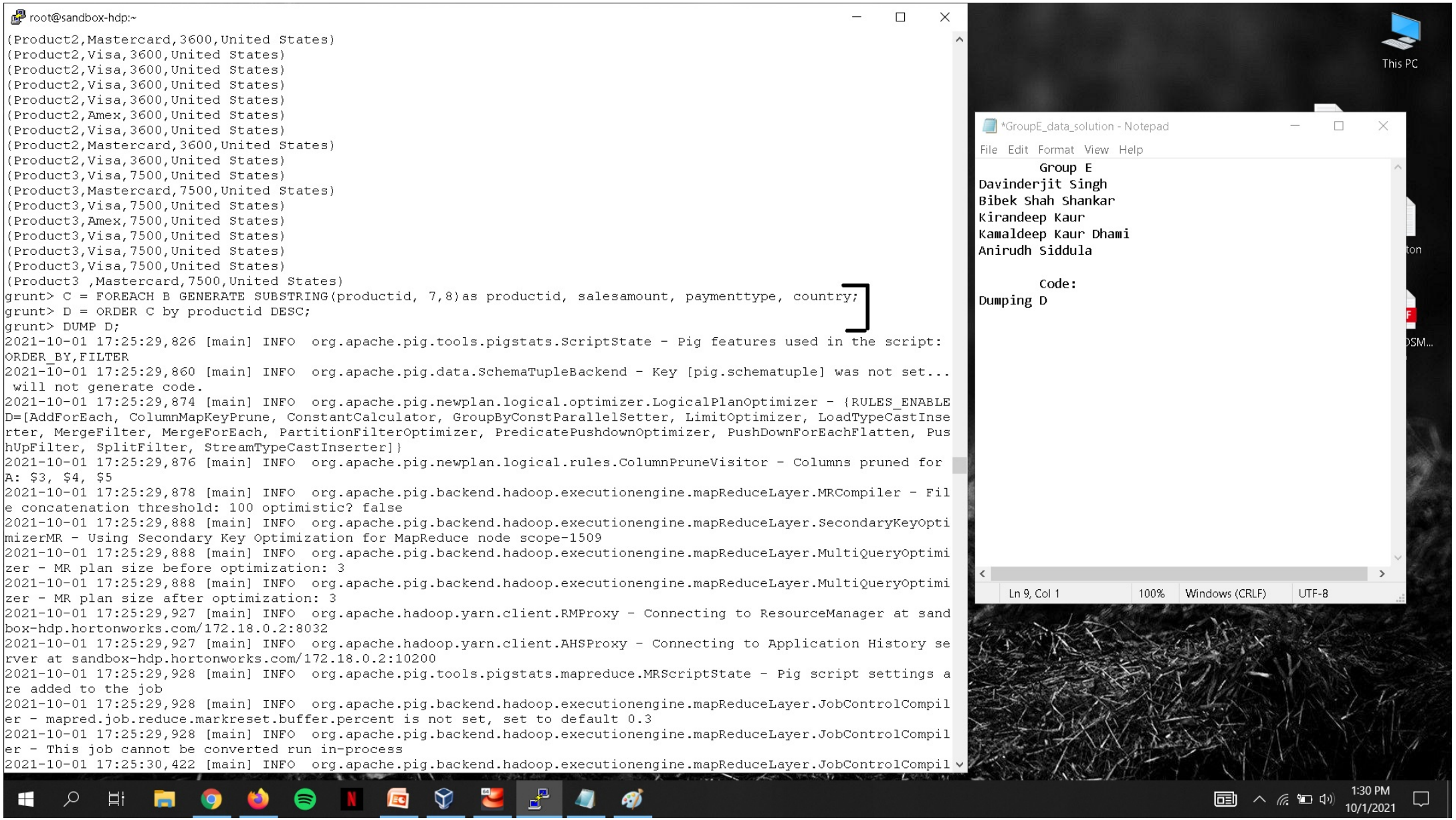
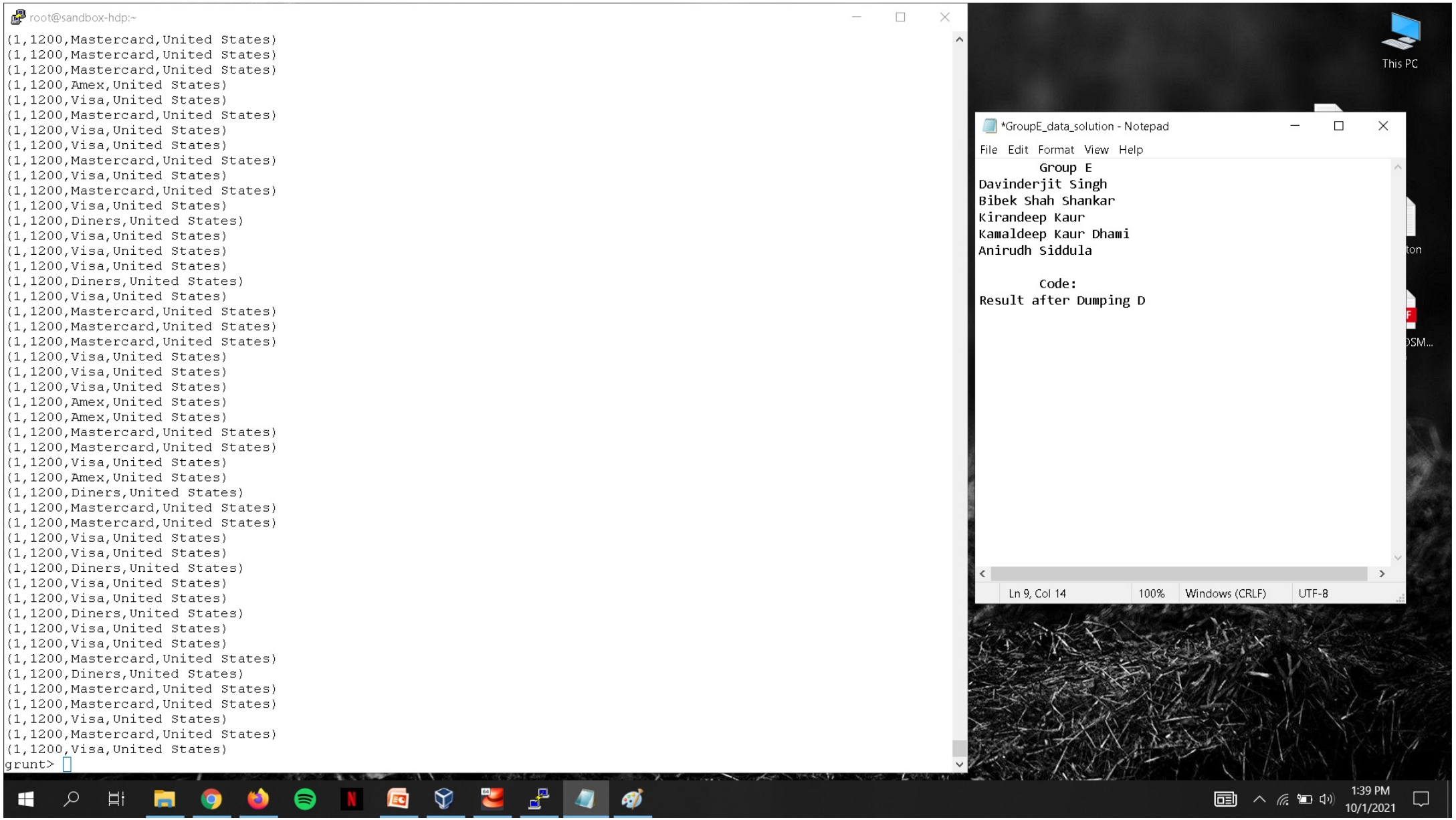

root@sandbox-hdp:~
[root@sandbox-hdp ~]# hdfs dfs -ls
[root@sandbox-hdp ~]# hdfs dfs -ls /test_data
Found 2 items
-rw-r--r--   1 hdfs hdfs   65835 2021-09-29 00:09 /test_data/SalesJan2009.csv
drwxr-xr-x   - hdfs hdfs   0 2021-09-29 00:15 /test_data/mp_output
[root@sandbox-hdp ~]# pig -x mapreduce
21/10/01 16:33:15 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
21/10/01 16:33:15 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
21/10/01 16:33:15 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-01 16:33:16,071 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnve
2021-10-01 16:33:16,071 [main] INFO org.apache.pig.Main - Logging error messages to: /root/pig_1633105
2021-10-01 16:33:16,104 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /root/.pigboo
2021-10-01 16:33:16,929 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - C
hortonworks.com:8020
2021-10-01 16:33:17,855 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-defa
2021-10-01 16:33:18,479 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeli
8188/ws/v1/timeline/
2021-10-01 16:33:18,807 [main] INFO org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
grunt> A = load '/test_data/SalesJan2009.csv'
>> USING PigStorage (',')
>> AS (productid:chararray, salesamount:int, paymenttype:chararray, customername:chararray, city:charar
grunt> DUMP A;
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Loading data to Pig
Ln 10, Col 2
100%
Windows (CRLF)
UTF-8
12:40 PM
10/1/2021

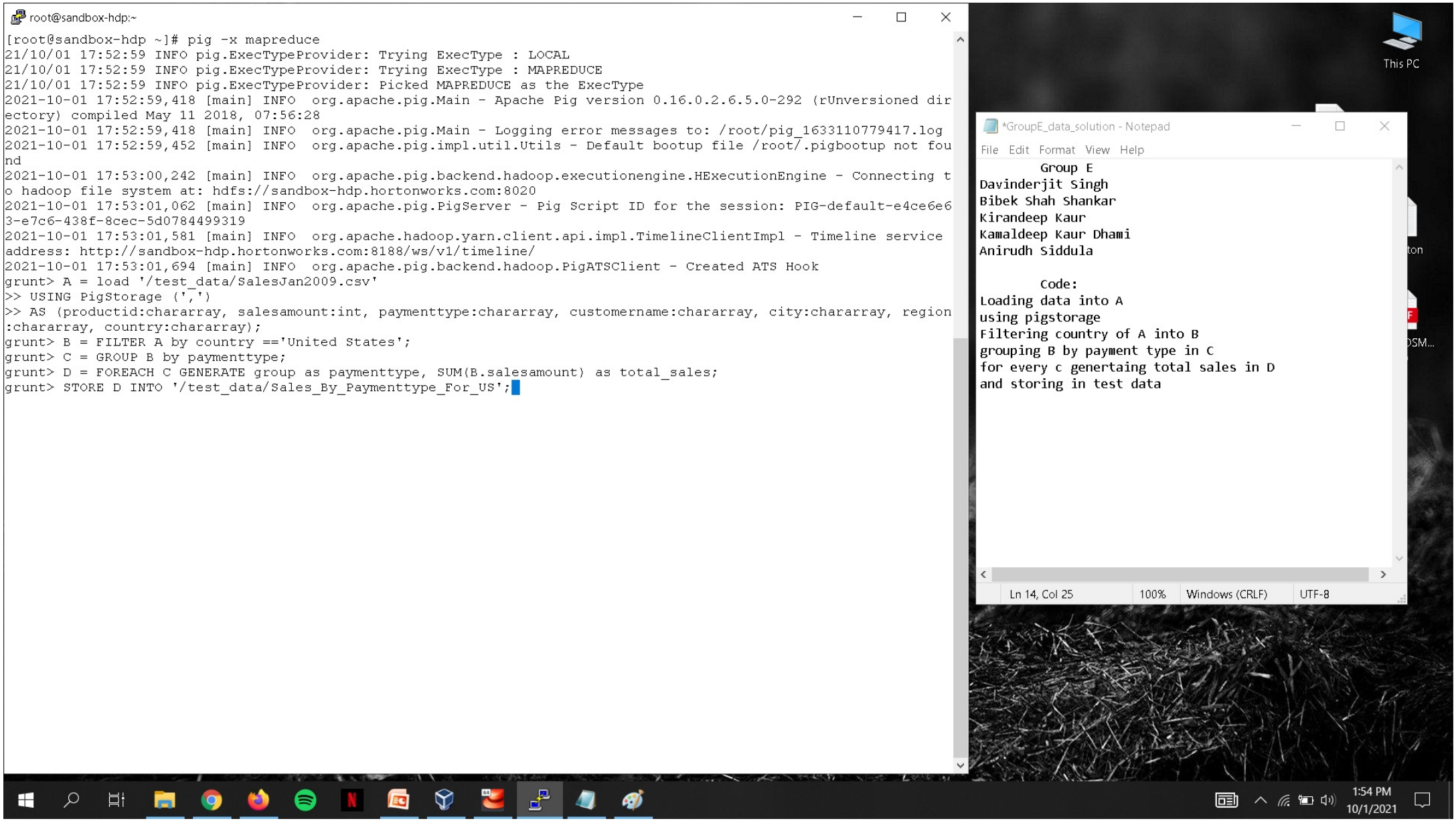

root@sandbox-hdp:~/scripts
[root@sandbox-hdp scripts]# pig -x mapreduce
21/10/01 21:59:36 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
21/10/01 21:59:36 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
21/10/01 21:59:36 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-01 21:59:36,265 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory) compiled May 11 2018, 07:56:28
2021-10-01 21:59:36,265 [main] INFO org.apache.pig.Main - Logging error messages to: /root/scripts/pig_1633125576263.log
2021-10-01 21:59:36,290 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2021-10-01 21:59:37,044 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://sandbox-hdp.hortonworks.com:8020
2021-10-01 21:59:37,752 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-89763fab-b853-4052-a941-f7b0a1e6fa90
2021-10-01 21:59:38,222 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://sandbox-hdp.hortonworks.com:8188/ws/v1/timeline/
2021-10-01 21:59:38,353 [main] INFO org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
grunt> A = load '/test_data/SalesJan2009.csv'
>> USING PigStorage(',')
>> AS (productid:chararray, salesamount:int, paymenttype:chararray, customername:chararray, city:chararray, region:chararray, country:chararray);
grunt> B = FILTER A by country =='United States';
grunt> C = FOREACH B GENERATE productid, paymenttype, salesamount, country;
grunt> D = ORDER C by productid;
grunt> ILLUSTRATE D;
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Fetching data to A
Filtering a by country US to B
producing product id, payment type, amount, country, in C
ordering C in D
Ln 13, Col 16
100%
Windows (CRLF)
UTF-8
6:02 PM
10/1/2021

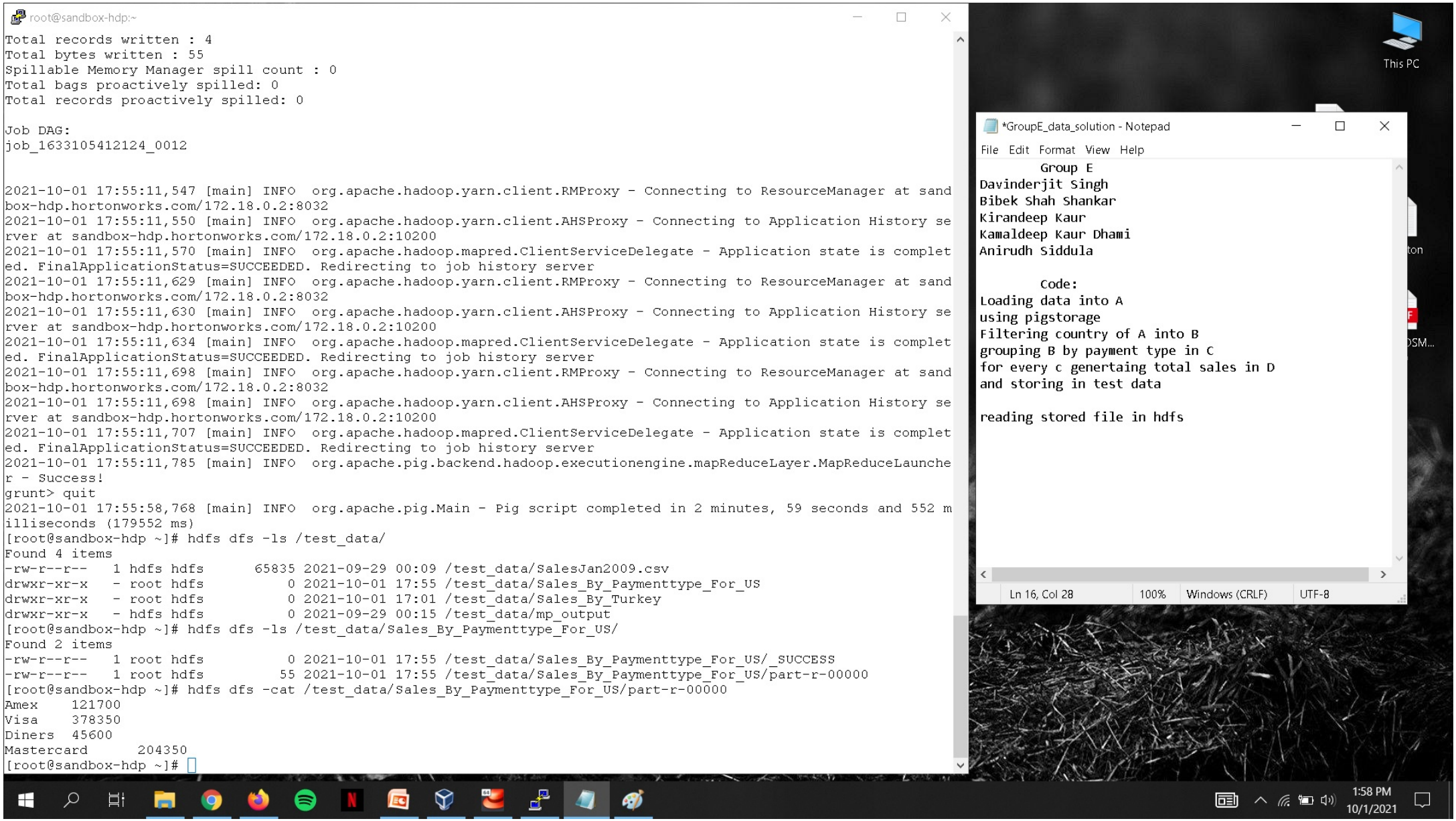

root@sandbox-hdp:~
initialized
2021-10-01 17:12:00,593 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigGenericMapReduce$Map - Aliases being processed per job phase (AliasName[line,offset]): M: D[6,4] C: R:
2021-10-01 17:12:00,594 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-10-01 17:12:00,594 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-10-01 17:12:00,594 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapReduce$Reduce - Aliases being processed per job phase (AliasName[line,offset]): M: D[6,4] C: R:
A	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Visa	Sara	Fort Lauderdale	FL	United States
	Product1	1200	Mastercard	Tracy	Louisville	KY	United States
	Product1	1200	Mastercard	Diana	Kuala Lumpur	Kuala Lumpur	Malaysia
B	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Visa	Sara	Fort Lauderdale	FL	United States
	Product1	1200	Mastercard	Tracy	Louisville	KY	United States
C	productid:chararray	paymenttype:chararray	salesamount:int	country:chararray
	Product1	Visa	1200	United States
	Product1	Mastercard	1200	United States
D	productid:chararray	paymenttype:chararray	salesamount:int	country:chararray
	Product1	Visa	1200	United States
	Product1	Mastercard	1200	United States
grunt>
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Illustrating D
after loading data into A,
filtering B by country United States,
generating product id, paymenttype,salesamount,
country in C
and ordering D by productid
Ln 14, Col 28
100%
Windows (CRLF)
UTF-8
1:17 PM
10/1/2021

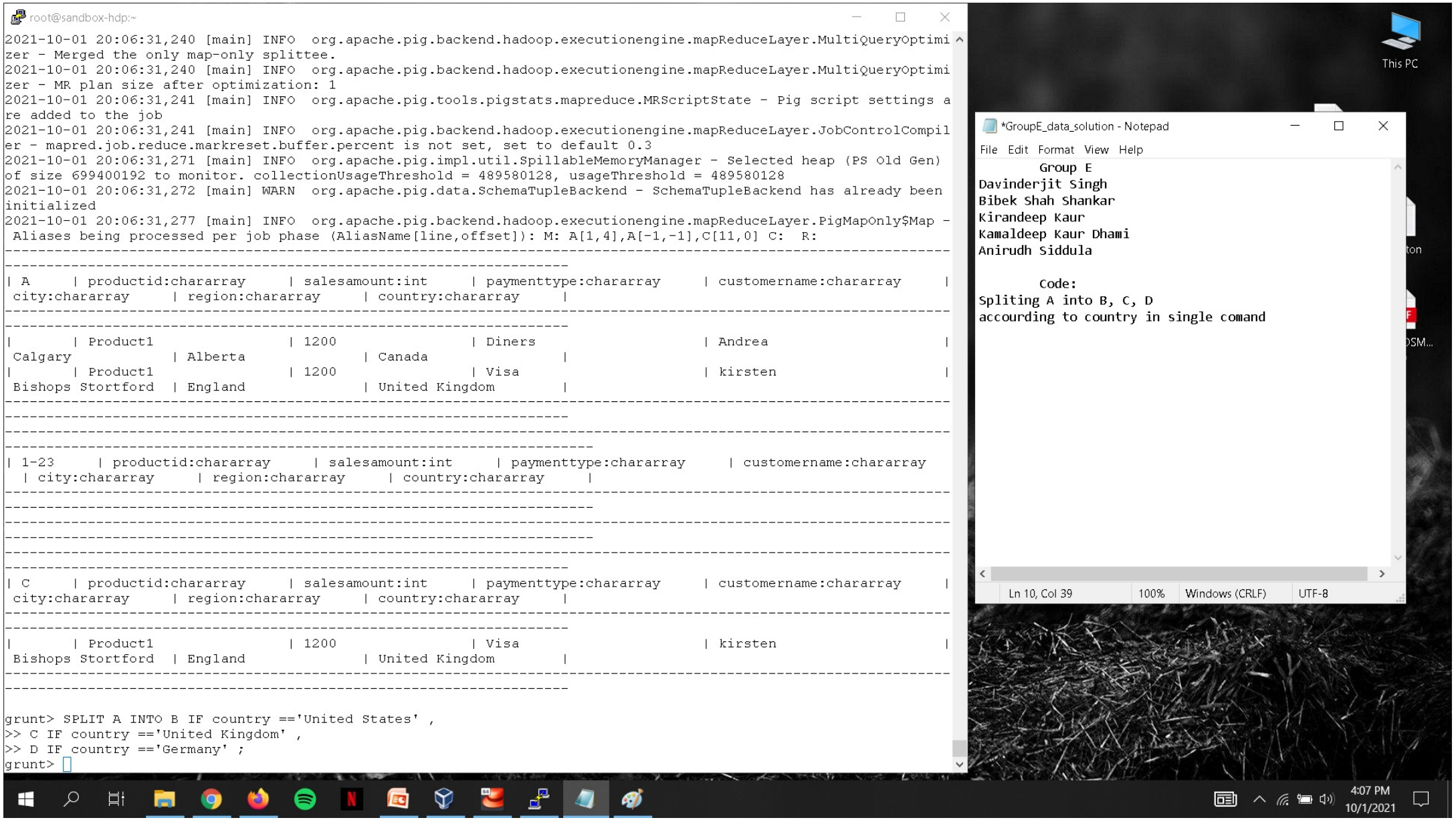

root@sandbox-hdp:~
(Product2,Visa,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Visa,3600,United States)
(Product3,Visa,7500,United States)
(Product3,Mastercard,7500,United States)
(Product3,Visa,7500,United States)
(Product3,Amex,7500,United States)
(Product3,Visa,7500,United States)
(Product3,Visa,7500,United States)
(Product3,Visa,7500,United States)
(Product3 ,Mastercard,7500,United States)
grunt>
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Result after Dumping D
Ln 9, Col 23
100%
Windows (CRLF)
UTF-8
1:21 PM
10/1/2021

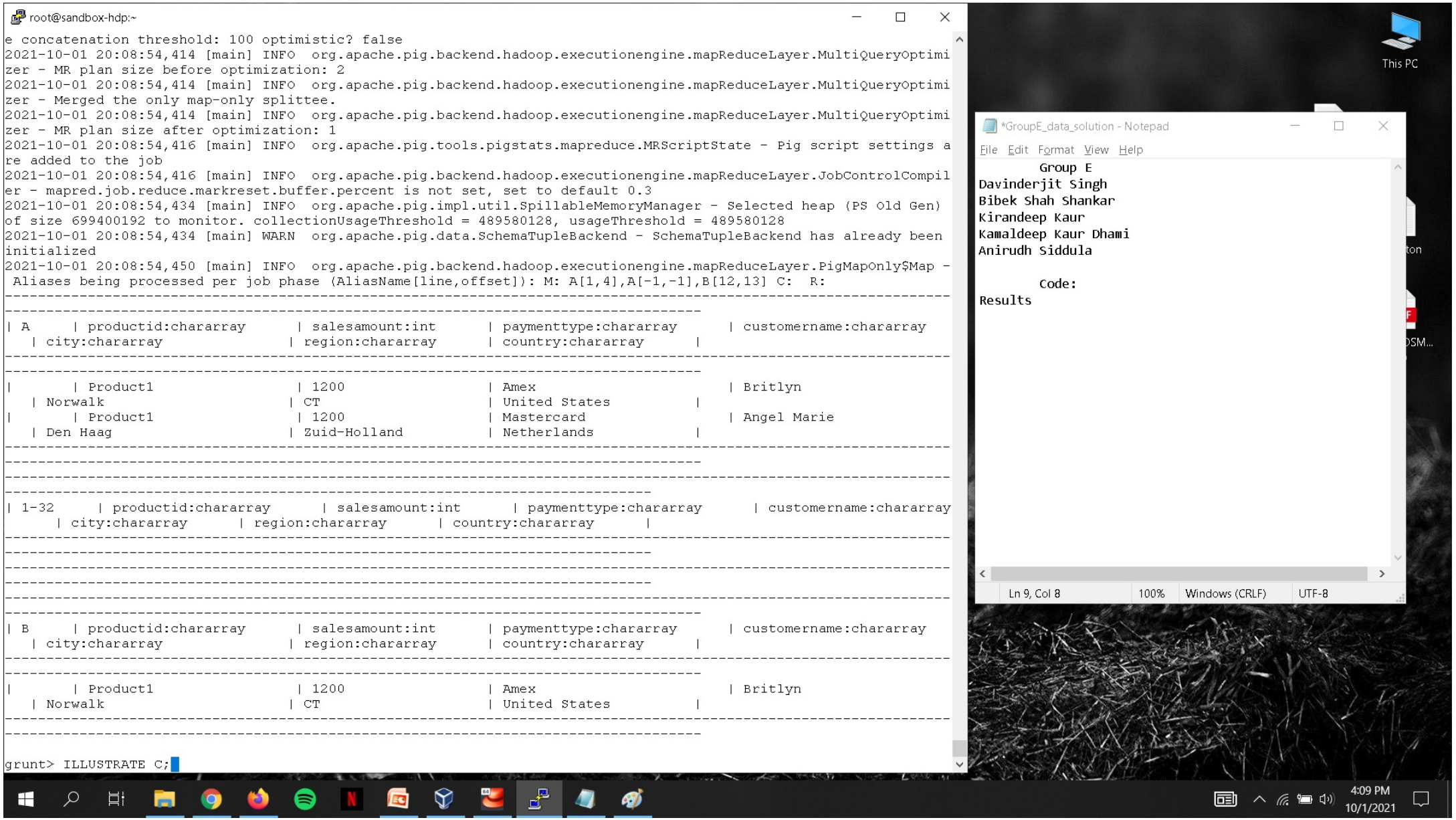

root@sandbox-hdp:~
(Product2,Mastercard,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Amex,3600,United States)
(Product2,Visa,3600,United States)
(Product2,Mastercard,3600,United States)
(Product2,Visa,3600,United States)
(Product3,Visa,7500,United States)
(Product3,Mastercard,7500,United States)
(Product3,Visa,7500,United States)
(Product3,Amex,7500,United States)
(Product3,Visa,7500,United States)
(Product3,Visa,7500,United States)
(Product3,Visa,7500,United States)
(Product3 ,Mastercard,7500,United States)
grunt> C = FOREACH B GENERATE SUBSTRING(productid, 7,8)as productid, salesamount, paymenttype, country;
grunt> D = ORDER C by productid DESC;
grunt> DUMP D;
2021-10-01 17:25:29,826 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: ORDER_BY,FILTER
2021-10-01 17:25:29,860 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-10-01 17:25:29,874 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2021-10-01 17:25:29,876 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for A: $3, $4, $5
2021-10-01 17:25:29,878 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2021-10-01 17:25:29,888 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.SecondaryKeyOptimizerMR - Using Secondary Key Optimization for MapReduce node scope-1509
2021-10-01 17:25:29,888 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2021-10-01 17:25:29,888 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2021-10-01 17:25:29,927 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
2021-10-01 17:25:29,927 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
2021-10-01 17:25:29,928 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-10-01 17:25:29,928 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-01 17:25:29,928 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2021-10-01 17:25:30,422 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompil
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Dumping D
Ln 9, Col 1
100%
Windows (CRLF)
UTF-8
1:30 PM
10/1/2021

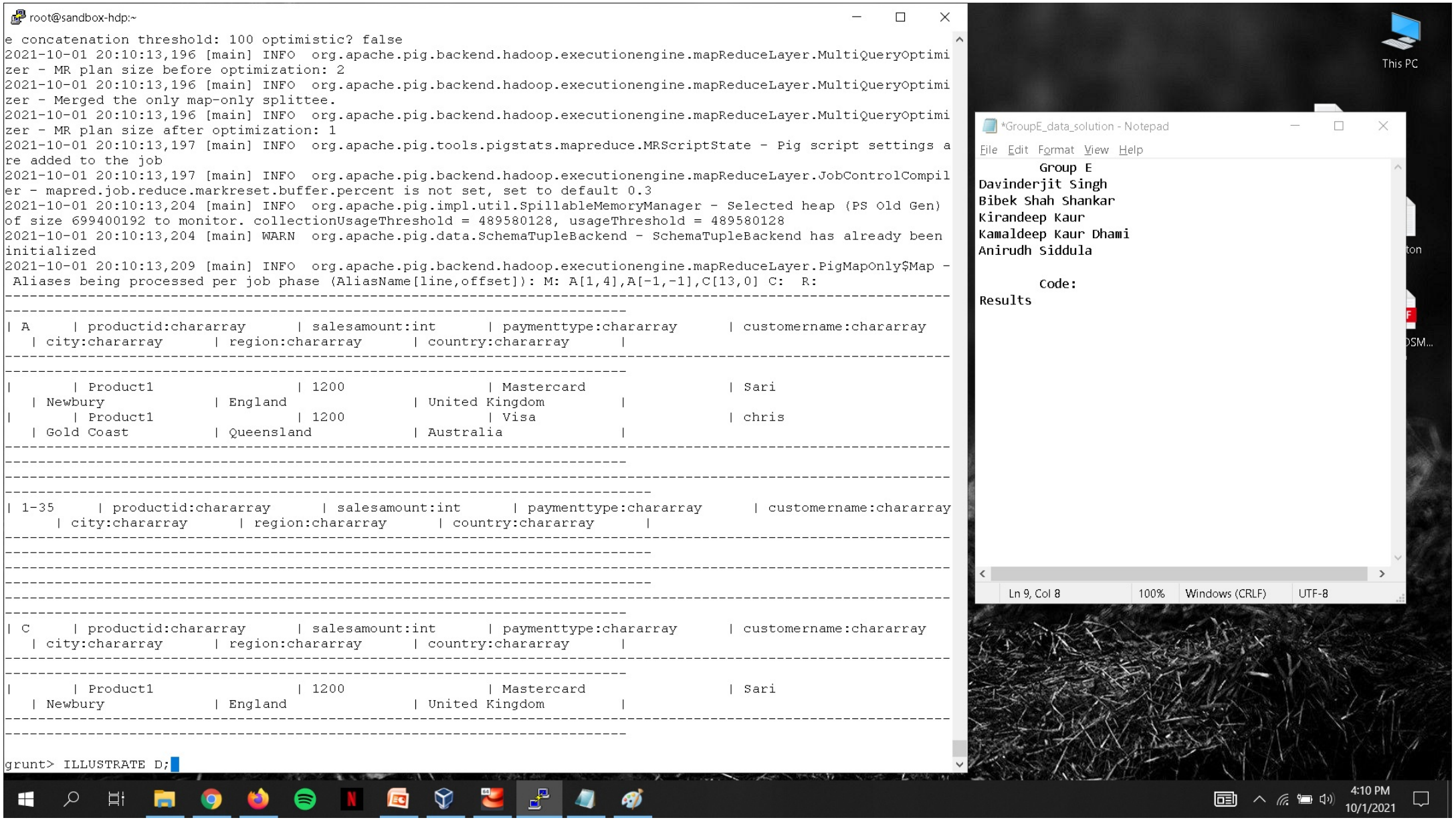

root@sandbox-hdp:~
(1,1200,Mastercard,United States)
(1,1200,Mastercard,United States)
(1,1200,Mastercard,United States)
(1,1200,Amex,United States)
(1,1200,Visa,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
(1,1200,Diners,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Diners,United States)
(1,1200,Visa,United States)
(1,1200,Mastercard,United States)
(1,1200,Mastercard,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Amex,United States)
(1,1200,Amex,United States)
(1,1200,Mastercard,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
(1,1200,Amex,United States)
(1,1200,Diners,United States)
(1,1200,Mastercard,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Diners,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Diners,United States)
(1,1200,Visa,United States)
(1,1200,Visa,United States)
(1,1200,Mastercard,United States)
(1,1200,Diners,United States)
(1,1200,Mastercard,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
(1,1200,Mastercard,United States)
(1,1200,Visa,United States)
grunt>
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Result after Dumping D
Ln 9, Col 14
100%
Windows (CRLF)
UTF-8
1:39 PM
10/1/2021

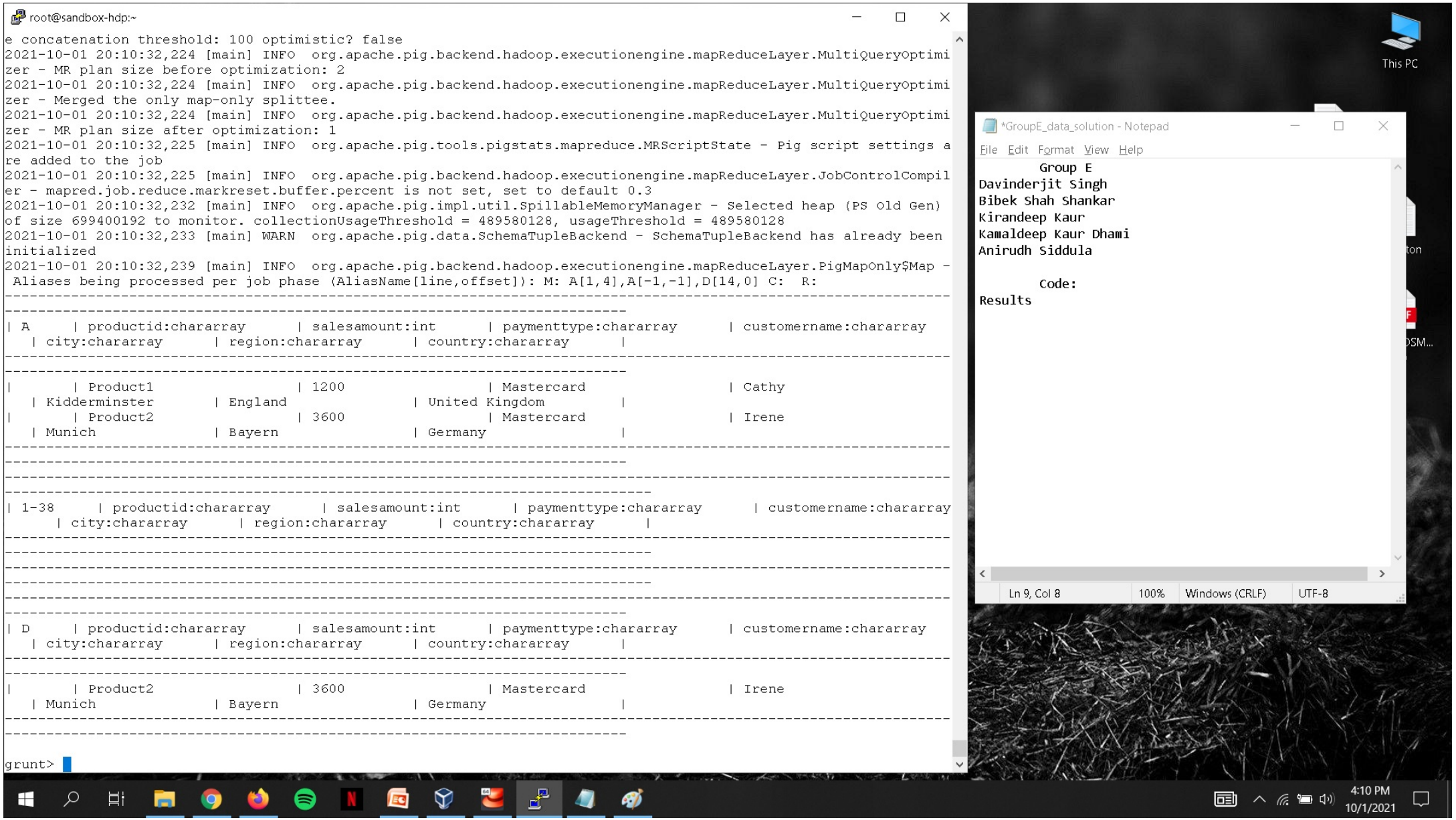

root@sandbox-hdp:~
[root@sandbox-hdp ~]# pig -x mapreduce
21/10/01 17:52:59 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
21/10/01 17:52:59 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
21/10/01 17:52:59 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-01 17:52:59,418 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0.2.6.5.0-292 (rUnversioned directory) compiled May 11 2018, 07:56:28
2021-10-01 17:52:59,418 [main] INFO org.apache.pig.Main - Logging error messages to: /root/pig_1633110779417.log
2021-10-01 17:52:59,452 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2021-10-01 17:53:00,242 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://sandbox-hdp.hortonworks.com:8020
2021-10-01 17:53:01,062 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-e4ce6e63-e7c6-438f-8cec-5d0784499319
2021-10-01 17:53:01,581 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://sandbox-hdp.hortonworks.com:8188/ws/v1/timeline/
2021-10-01 17:53:01,694 [main] INFO org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
grunt> A = load '/test_data/SalesJan2009.csv'
>> USING PigStorage (',')
>> AS (productid:chararray, salesamount:int, paymenttype:chararray, customername:chararray, city:chararray, region:chararray, country:chararray);
grunt> B = FILTER A by country =='United States';
grunt> C = GROUP B by paymenttype;
grunt> D = FOREACH C GENERATE group as paymenttype, SUM(B.salesamount) as total_sales;
grunt> STORE D INTO '/test_data/Sales_By_Paymenttype_For_US';
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Loading data into A
using pigstorage
Filtering country of A into B
grouping B by payment type in C
for every c genertaing total sales in D
and storing in test data
Ln 14, Col 25
100%
Windows (CRLF)
UTF-8
1:54 PM
10/1/2021

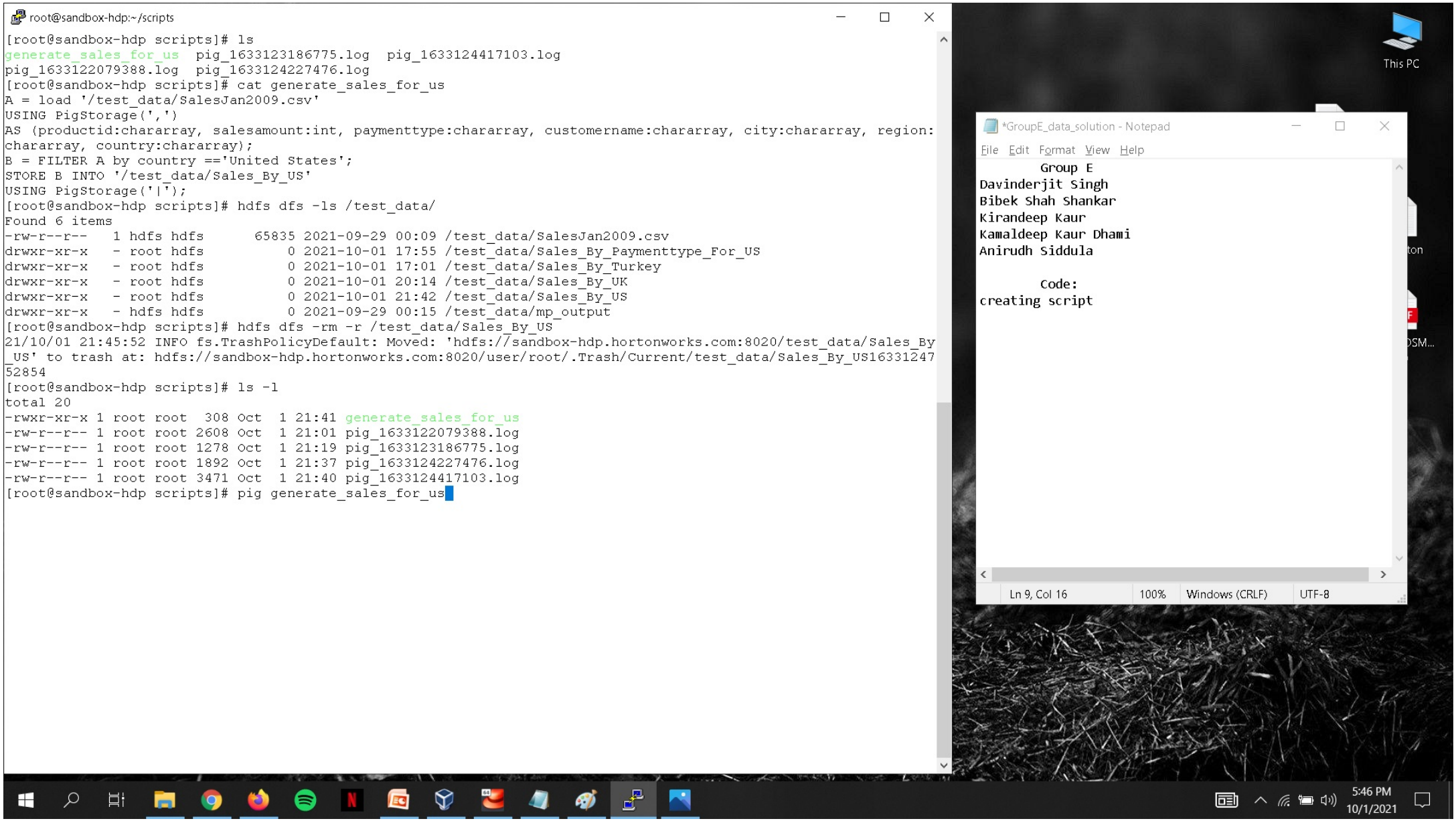

root@sandbox-hdp:~
Total records written : 4
Total bytes written : 55
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_1633105412124_0012
2021-10-01 17:55:11,547 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
2021-10-01 17:55:11,550 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
2021-10-01 17:55:11,570 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-01 17:55:11,629 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
2021-10-01 17:55:11,630 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
2021-10-01 17:55:11,634 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-01 17:55:11,698 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
2021-10-01 17:55:11,698 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
2021-10-01 17:55:11,707 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-01 17:55:11,785 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> quit
2021-10-01 17:55:58,768 [main] INFO org.apache.pig.Main - Pig script completed in 2 minutes, 59 seconds and 552 milliseconds (179552 ms)
[root@sandbox-hdp ~]# hdfs dfs -ls /test_data/
Found 4 items
-rw-r--r-- 1 hdfs hdfs 65835 2021-09-29 00:09 /test_data/SalesJan2009.csv
drwxr-xr-x - root hdfs 0 2021-10-01 17:55 /test_data/Sales_By_Paymenttype_For_US
drwxr-xr-x - root hdfs 0 2021-10-01 17:01 /test_data/Sales_By_Turkey
drwxr-xr-x - hdfs hdfs 0 2021-09-29 00:15 /test_data/mp_output
[root@sandbox-hdp ~]# hdfs dfs -ls /test_data/Sales_By_Paymenttype_For_US/
Found 2 items
-rw-r--r-- 1 root hdfs 0 2021-10-01 17:55 /test_data/Sales_By_Paymenttype_For_US/_SUCCESS
-rw-r--r-- 1 root hdfs 55 2021-10-01 17:55 /test_data/Sales_By_Paymenttype_For_US/part-r-00000
[root@sandbox-hdp ~]# hdfs dfs -cat /test_data/Sales_By_Paymenttype_For_US/part-r-00000
Amex 121700
Visa 378350
Diners 45600
Mastercard 204350
[root@sandbox-hdp ~]#
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Loading data into A
using pigstorage
Filtering country of A into B
grouping B by payment type in C
for every c genertaing total sales in D
and storing in test data
reading stored file in hdfs
Ln 16, Col 28
100%
Windows (CRLF)
UTF-8
1:58 PM
10/1/2021

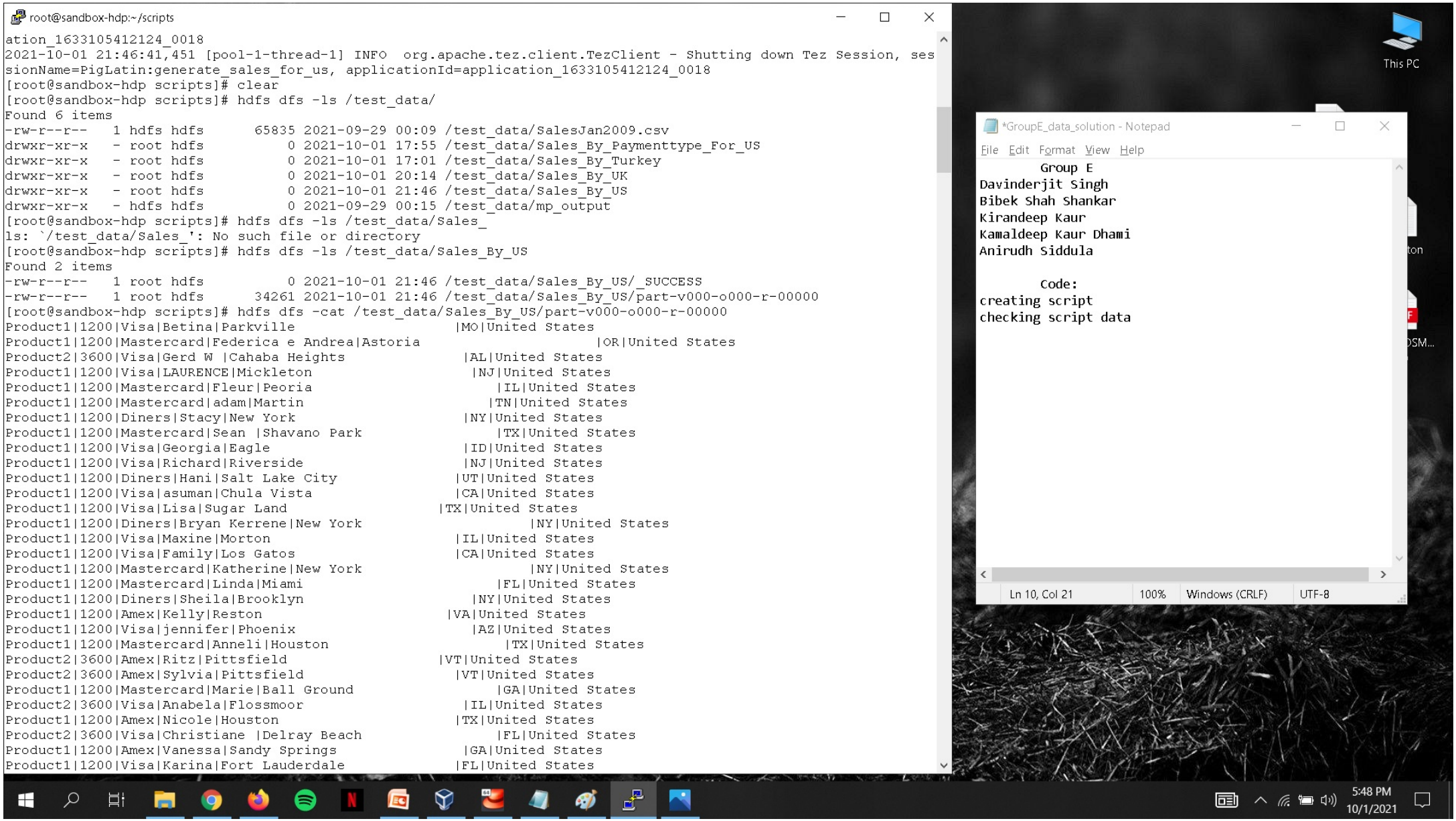

root@sandbox-hdp:~
2021-10-01 20:06:31,240 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged the only map-only splittee.
2021-10-01 20:06:31,240 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-10-01 20:06:31,241 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-10-01 20:06:31,241 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-01 20:06:31,271 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-10-01 20:06:31,272 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-10-01 20:06:31,277 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset]): M: A[1,4],A[-1,-1],C[11,0] C: R:
A	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Diners	Andrea	Calgary	Alberta	Canada
	Product1	1200	Visa	kirsten	Bishops Stortford	England	United Kingdom
1-23	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
C	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Visa	kirsten	Bishops Stortford	England	United Kingdom
grunt> SPLIT A INTO B IF country =='United States' ,
>> C IF country =='United Kingdom' ,
>> D IF country =='Germany' ;
grunt>
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Spliting A into B, C, D
accourding to country in single comand
Ln 10, Col 39
100%
Windows (CRLF)
UTF-8
4:07 PM
10/1/2021


root@sandbox-hdp:~
e concatenation threshold: 100 optimistic? false
2021-10-01 20:08:54,414 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2021-10-01 20:08:54,414 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged the only map-only splittee.
2021-10-01 20:08:54,414 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-10-01 20:08:54,416 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-10-01 20:08:54,416 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-01 20:08:54,434 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-10-01 20:08:54,434 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-10-01 20:08:54,450 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset]): M: A[1,4],A[-1,-1],B[12,13] C: R:
A	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Amex	Britlyn	Norwalk	CT	United States
	Product1	1200	Mastercard	Angel Marie	Den Haag	Zuid-Holland	Netherlands
1-32	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
B	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Amex	Britlyn	Norwalk	CT	United States
grunt> ILLUSTRATE C;
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Results
Ln 9, Col 8
100%
Windows (CRLF)
UTF-8
4:09 PM
10/1/2021


root@sandbox-hdp:~
e concatenation threshold: 100 optimistic? false
2021-10-01 20:10:13,196 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2021-10-01 20:10:13,196 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged the only map-only splittee.
2021-10-01 20:10:13,196 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-10-01 20:10:13,197 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-10-01 20:10:13,197 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-01 20:10:13,204 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-10-01 20:10:13,204 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-10-01 20:10:13,209 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset]): M: A[1,4],A[-1,-1],C[13,0] C: R:
A	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Mastercard	Sari	Newbury	England	United Kingdom
	Product1	1200	Visa	chris	Gold Coast	Queensland	Australia
1-35	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
C	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Mastercard	Sari	Newbury	England	United Kingdom
grunt> ILLUSTRATE D;
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Results
Ln 9, Col 8
100%
Windows (CRLF)
UTF-8
4:10 PM
10/1/2021


root@sandbox-hdp:~
e concatenation threshold: 100 optimistic? false
2021-10-01 20:10:32,224 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2021-10-01 20:10:32,224 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - Merged the only map-only splittee.
2021-10-01 20:10:32,224 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-10-01 20:10:32,225 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2021-10-01 20:10:32,225 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2021-10-01 20:10:32,232 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128
2021-10-01 20:10:32,233 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-10-01 20:10:32,239 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being processed per job phase (AliasName[line,offset]): M: A[1,4],A[-1,-1],D[14,0] C: R:
A	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product1	1200	Mastercard	Cathy	Kidderminster	England	United Kingdom
	Product2	3600	Mastercard	Irene	Munich	Bayern	Germany
1-38	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
D	productid:chararray	salesamount:int	paymenttype:chararray	customername:chararray	city:chararray	region:chararray	country:chararray
	Product2	3600	Mastercard	Irene	Munich	Bayern	Germany
grunt>
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
Results
Ln 9, Col 8
100%
Windows (CRLF)
UTF-8
4:10 PM
10/1/2021


root@sandbox-hdp:~/scripts
[root@sandbox-hdp scripts]# ls
generate_sales_for_us pig_1633123186775.log pig_1633124417103.log
pig_1633122079388.log pig_1633124227476.log
[root@sandbox-hdp scripts]# cat generate_sales_for_us
A = load '/test_data/SalesJan2009.csv'
USING PigStorage(',')
AS (productid:chararray, salesamount:int, paymenttype:chararray, customername:chararray, city:chararray, region:
chararray, country:chararray);
B = FILTER A by country =='United States';
STORE B INTO '/test_data/Sales_By_US'
USING PigStorage('|');
[root@sandbox-hdp scripts]# hdfs dfs -ls /test_data/
Found 6 items
-rw-r--r--   1 hdfs hdfs   65835 2021-09-29 00:09 /test_data/SalesJan2009.csv
drwxr-xr-x   - root hdfs   0 2021-10-01 17:55 /test_data/Sales_By_Paymenttype_For_US
drwxr-xr-x   - root hdfs   0 2021-10-01 17:01 /test_data/Sales_By_Turkey
drwxr-xr-x   - root hdfs   0 2021-10-01 20:14 /test_data/Sales_By_UK
drwxr-xr-x   - root hdfs   0 2021-10-01 21:42 /test_data/Sales_By_US
drwxr-xr-x   - hdfs hdfs   0 2021-09-29 00:15 /test_data/mp_output
[root@sandbox-hdp scripts]# hdfs dfs -rm -r /test_data/Sales_By_US
21/10/01 21:45:52 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/test_data/Sales_By
_US' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/root/.Trash/Current/test_data/Sales_By_US16331247
52854
[root@sandbox-hdp scripts]# ls -l
total 20
-rwxr-xr-x 1 root root 308 Oct 1 21:41 generate_sales_for_us
-rw-r--r-- 1 root root 2608 Oct 1 21:01 pig_1633122079388.log
-rw-r--r-- 1 root root 1278 Oct 1 21:19 pig_1633123186775.log
-rw-r--r-- 1 root root 1892 Oct 1 21:37 pig_1633124227476.log
-rw-r--r-- 1 root root 3471 Oct 1 21:40 pig_1633124417103.log
[root@sandbox-hdp scripts]# pig generate_sales_for_us
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
creating script
Ln 9, Col 16
100%
Windows (CRLF)
UTF-8
5:46 PM
10/1/2021


root@sandbox-hdp:~/scripts
ation_1633105412124_0018
2021-10-01 21:46:41,451 [pool-1-thread-1] INFO org.apache.tez.client.TezClient - Shutting down Tez Session, ses
sionName=PigLatin:generate_sales_for_us, applicationId=application_1633105412124_0018
[root@sandbox-hdp scripts]# clear
[root@sandbox-hdp scripts]# hdfs dfs -ls /test_data/
Found 6 items
-rw-r--r--   1 hdfs hdfs   65835 2021-09-29 00:09 /test_data/SalesJan2009.csv
drwxr-xr-x   - root hdfs   0 2021-10-01 17:55 /test_data/Sales_By_Paymenttype_For_US
drwxr-xr-x   - root hdfs   0 2021-10-01 17:01 /test_data/Sales_By_Turkey
drwxr-xr-x   - root hdfs   0 2021-10-01 20:14 /test_data/Sales_By_UK
drwxr-xr-x   - root hdfs   0 2021-10-01 21:46 /test_data/Sales_By_US
drwxr-xr-x   - hdfs hdfs   0 2021-09-29 00:15 /test_data/mp_output
[root@sandbox-hdp scripts]# hdfs dfs -ls /test_data/Sales_
ls: `/test_data/Sales_': No such file or directory
[root@sandbox-hdp scripts]# hdfs dfs -ls /test_data/Sales_By_US
Found 2 items
-rw-r--r--   1 root hdfs   0 2021-10-01 21:46 /test_data/Sales_By_US/_SUCCESS
-rw-r--r--   1 root hdfs   34261 2021-10-01 21:46 /test_data/Sales_By_US/part-v000-o000-r-00000
[root@sandbox-hdp scripts]# hdfs dfs -cat /test_data/Sales_By_US/part-v000-o000-r-00000
Product1|1200|Visa|Betina|Parkville|MO|United States
Product1|1200|Mastercard|Federica e Andrea|Astoria|OR|United States
Product2|3600|Visa|Gerd W |Cahaba Heights|AL|United States
Product1|1200|Visa|LAURENCE|Mickleton|NJ|United States
Product1|1200|Mastercard|Fleur|Peoria|IL|United States
Product1|1200|Mastercard|adam|Martin|TN|United States
Product1|1200|Diners|Stacy|New York|NY|United States
Product1|1200|Mastercard|Sean |Shavano Park|TX|United States
Product1|1200|Visa|Georgia|Eagle|ID|United States
Product1|1200|Visa|Richard|Riverside|NJ|United States
Product1|1200|Diners|Hani|Salt Lake City|UT|United States
Product1|1200|Visa|asuman|Chula Vista|CA|United States
Product1|1200|Visa|Lisa|Sugar Land|TX|United States
Product1|1200|Diners|Bryan Kerrene|New York|NY|United States
Product1|1200|Visa|Maxine|Morton|IL|United States
Product1|1200|Visa|Family|Los Gatos|CA|United States
Product1|1200|Mastercard|Katherine|New York|NY|United States
Product1|1200|Mastercard|Linda|Miami|FL|United States
Product1|1200|Diners|Sheila|Brooklyn|NY|United States
Product1|1200|Amex|Kelly|Reston|VA|United States
Product1|1200|Visa|jennifer|Phoenix|AZ|United States
Product1|1200|Mastercard|Anneli|Houston|TX|United States
Product2|3600|Amex|Ritz|Pittsfield|VT|United States
Product2|3600|Amex|Sylvia|Pittsfield|VT|United States
Product1|1200|Mastercard|Marie|Ball Ground|GA|United States
Product2|3600|Visa|Anabela|Flossmoor|IL|United States
Product1|1200|Amex|Nicole|Houston|TX|United States
Product2|3600|Visa|Christiane |Delray Beach|FL|United States
Product1|1200|Amex|Vanessa|Sandy Springs|GA|United States
Product1|1200|Visa|Karina|Fort Lauderdale|FL|United States
This PC
*GroupE_data_solution - Notepad
File Edit Format View Help
Group E
Davinderjit Singh
Bibek Shah Shankar
Kirandeep Kaur
Kamaldeep Kaur Dhami
Anirudh Siddula
Code:
creating script
checking script data
Ln 10, Col 21
100%
Windows (CRLF)
UTF-8
5:48 PM
10/1/2021