Problem description:

In agriculture sector where farmers and agribusinesses have to make innumerable decisions every day and intricate complexities involves the various factors influencing them. An essential issue for agricultural planning intention is the accurate yield estimation for the numerous crops involved in the planning. Data mining techniques are necessary approach for accomplishing practical and effective solutions for this problem. Agriculture has been an obvious target for big data. Environmental conditions, variability in soil, input levels, combinations and commodity prices have made it all the more relevant for farmers to use information and get help to make critical farming decisions. Mining the large amount of existing crop, soil and climatic data, and analysing new, non-experimental data optimizes the production and makes agriculture more resilient to climatic change.

Dataset Used: crop_production.csv

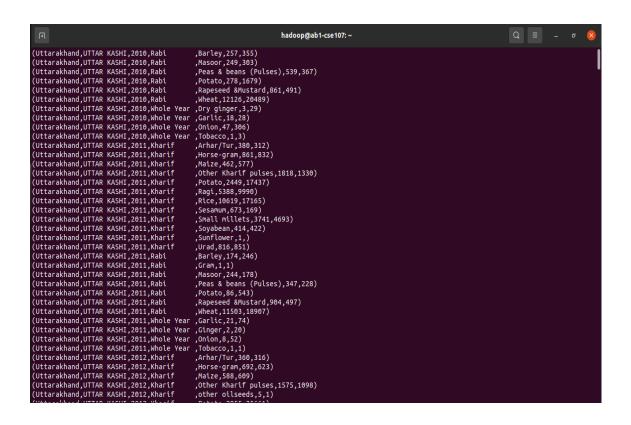
State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254	2000
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2	1
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102	321
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176	641
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720	165
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Coconut	18168	65100000
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Dry ginger	36	100
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sugarcane	1	2
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sweet potato	5	15
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Tapioca	40	169
Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Arecanut	1254	2061
Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Other Kharif pulses	2	1

```
pig -x local
agriculture= load '/home/hadoop/Documents/miniproject/crop_production.csv' using
PigStorage(',') as
(State_Name:chararray,District_Name:chararray,Crop_Year:int,Season:chararray,Crop:chara
rray,Area:int,Production:int);
describe agriculture;
dump agriculture;
statewise_group = group agriculture by State_Name;
dump statewise_group;
store statewise_group into '/home/hadoop/Documents/miniproject/statewise_output';
filter_district = filter agriculture by District_Name == 'MADURAI';
dump filter_district;
store filter_district into '/home/hadoop/Documents/miniproject/filter_output;
order_year = order agriculture by Crop_Year desc;
dump order_year;
foreach_crop = foreach agriculture generate Crop;
dump foreach_crop;
```

Pig commands:

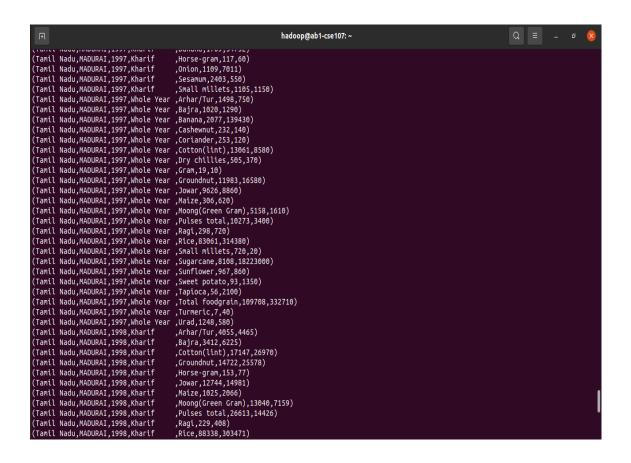
Output:

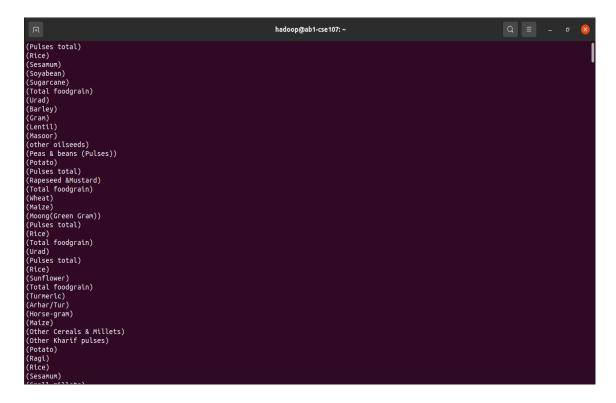
```
Q = _ 0
                                                                  hadoop@ab1-cse107: ~
 hadoop@ab1-cse107:~$ pig -x local
2022-11-05 09:35:56,630 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2022-11-05 09:35:56,631 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2022-11-05 09:35:56,665 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2022-11-05 09:35:56,665 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1667621156663.log
2022-11-05 09:35:56,679 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found 2022-11-05 09:35:56,740 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapredu
ce.jobtracker.address
2022-11-05 09:35:56,742 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: fil
e:///
2022-11-05 09:35:56,804 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.
bytes-per-checksum
2022-11-05 09:35:56,817 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-92cc4cb0-d19c-45f0-9874-cad10aa0fd5
2022-11-05 09:35:56,817 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> agriculture= load '/home/hadoop/Documents/miniproject/crop_production.csv' using PigStorage(',') as (State_Name:chararray,District_Name
:chararray,Crop_Year:int,Season:chararray,Crop:chararray,Area:int,Production:int);
2022-11-05 09:36:05,908 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.
bytes-per-checksum
grunt> describe agriculture;
agriculture: {State_Name: chararray,District_Name: chararray,Crop_Year: int,Season: chararray,Crop: chararray,Area: int,Production: int}
grunt> dump agriculture;
```

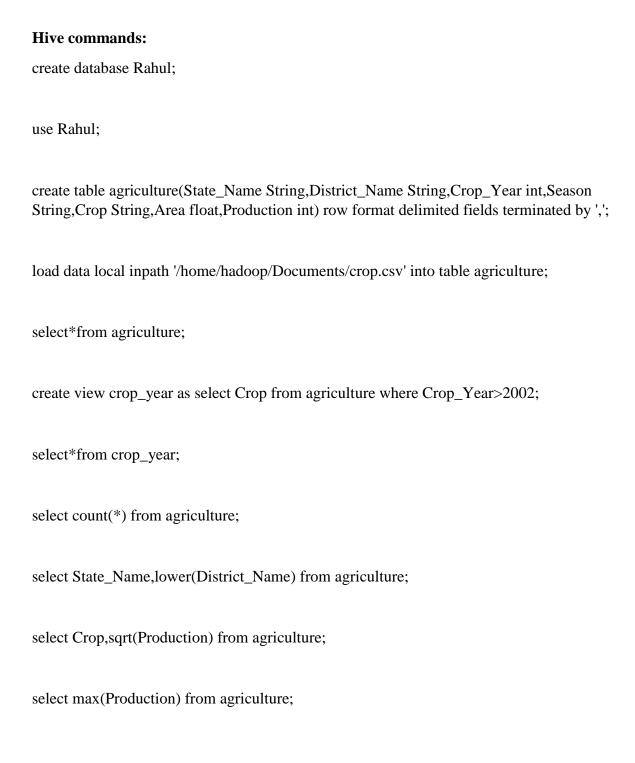


ABS, 2010, Rabl , Cashewnut, 1000, 200), (Andaman and Nicobar Islands, NICOBARS, 2010, Rabl , Arecanut, 944, 1610), (Andaman and Nicobar Islands, NICOBARS, 2010, Authurn , Sugarcame, 13, 41), (Andaman and Nicobar Islands, NICOBARS, 2010, Authurn , Rice, 31, 01), (Andaman and Nicobar Islands, NICOBARS, 2006, Mhole Vear , Taploca, 52, 450), (Andaman and Nicobar Islands, NICOBARS, 2006, Mhole Vear , Taploca, 52, 450), (Andaman and Nicobar Islands, NICOBARS, 2006, Mhole Vear , Taploca, 52, 450), (Andaman and Nicobar Islands, NICOBARS, 2006, Mhole Vear , Suer , 220)}) grunt> [

hadoop@ab1-cse107: ~ Q = bytes-per-checksum grunt> filter_district = filter agriculture by District_Name == 'MADURAI'; grunt> dump filter_district; 2022-11-05 10:48:53,584 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs. bytes-per-checksum 2022-11-05 10:48:53,597 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER 2022-11-05 10:48:53,611 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs. bytes-per-checksum 2022-11-05 10:48:53,635 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES ENABLED=[AddForEach, ColumnMapKeyP 2022-11-05 10:48:55,655 [main] INFO org.apache.pig.newplan.logical.optimizer.logicalrlanoptimizer - {MULES_EMABLED=[AddForEach, ColumnMapkeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-11-05 10:48:53,682 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor . collectionUsageThreshold = 489580128, usageThreshold = 489580128
2022-11-05 10:48:53,720 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 1 00 optimistic? false 2022-11-05 10:48:53,734 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before op timization: 1 2022-11-05 10:48:53.734 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLaver.MultiQuervOptimizer - MR plan size after opt imization: 2022-11-05 10:48:53,749 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs. bytes-per-checksum 2022-11-05 10:48:53,800 [main] INFO org.apache.hadoop.metrics2.impl.MetricsConfig - Loaded properties from hadoop-metrics2.properties 2022-11-05 10:48:53,892 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2022-11-05 10:48:53,892 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2022-11-05 10:48:53,909 [main] INFO org.apache.hadoop.netrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2022-11-05 10:48:53,909 [main] INFO org.apache.hadoop.conf.configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecat ed. Instead. use mapreduce.reduce.markreset.buffer.percent ed. Instead, use mapreduce.reduce.markreset.burrer.percent 2022-11-05 10:48:53,912 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markr eset.buffer.percent is not set, set to default 0.3 2022-11-05 10:48:53,914 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use map reduce.output.fileoutputformat.compress 2022-11-05 10:48:53,922 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store iob 2022-11-05 10:48:53,931 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-11-05 10:48:53,931 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2022-11-05 10:48:53,931 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Settin
g key [pig.schematuple.local.dir] with code temp directory: /tmp/1667625533931-0
2022-11-05 10:48:53,961 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) wait







Output:

