Big Data Lab (Assignment)

Name: Yash Jaiswal
USN: 1NT18IS185
Semester: 6 (C2 Batch)

Hadoop Brief:

Hadoop is an open-source framework from Apache and is used to store, process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing. It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover, it can be scaled up just by adding nodes in the cluster.

Modules of Hadoop:

- HDFS: Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
- Yarn: Yet another Resource Negotiator is used for job scheduling and managing the cluster.
- Map Reduce: This is a framework which helps Java programs to do the parallel computation on data using key value pairs. The Map task takes input data and converts it into a data set which can be computed in Key value pairs. The output of Map task is consumed by reduce task and then the output of reducer gives the desired result.
- Hadoop Common: These Java libraries are used to start Hadoop and are used by other Hadoop modules.

MAPREDUCE BRIEF:

A MapReduce is a data processing tool which is used to process the data parallely in a distributed form. It was developed in 2004, on the basis of a paper titled as "MapReduce: Simplified Data Processing on Large Clusters," published by Google.

The MapReduce is a paradigm which has two phases, the mapper phase, and the reducer phase. In the Mapper, the input is given in the form of a key-value pair. The output of the Mapper is fed to the reducer as input. The reducer runs only after the Mapper is over. The reducer too takes input in key-value format, and the output of the reducer is the final output.

Usage of MapReduce:

- It can be used in various applications like document clustering, distributed sorting, and web link-graph reversal.
- It can be used for distributed pattern-based searching.
- We can also use MapReduce in machine learning.

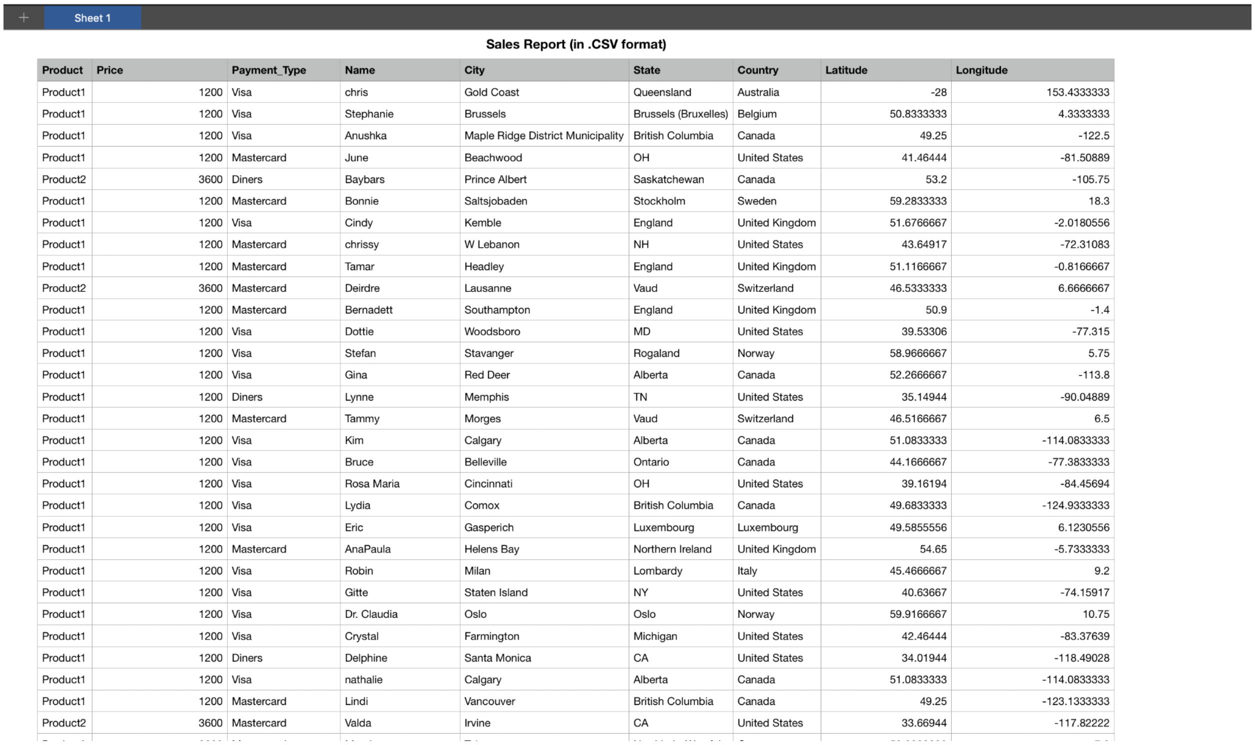
Exercise 1: Implement a map-reduce program in JAVA or Python using any Data of your choice

Dataset Preparation

1. Download CSV from here.

2. Format the CSV to remove **Transaction_date**, **Last_Login**, and **Account_Created**.

3. Final CSV should look something like this.



Sales Report (in .CSV format)

Product	Price	Payment_Type	Name	City	State	Country	Latitude	Longitude
Product1	1200	Visa	chris	Gold Coast	Queensland	Australia	-28	153.4333333
Product1	1200	Visa	Stephanie	Brussels	Brussels (Bruxelles)	Belgium	50.8333333	4.3333333
Product1	1200	Visa	Anushka	Maple Ridge District Municipality	British Columbia	Canada	49.25	-122.5
Product1	1200	Mastercard	June	Beachwood	OH	United States	41.46444	-81.50889
Product2	3600	Diners	Baybars	Prince Albert	Saskatchewan	Canada	53.2	-105.75
Product1	1200	Mastercard	Bonnie	Saltsjobaden	Stockholm	Sweden	59.2833333	18.3
Product1	1200	Visa	Cindy	Kemble	England	United Kingdom	51.6766667	-2.0180556
Product1	1200	Mastercard	chrissy	W Lebanon	NH	United States	43.64917	-72.31083
Product1	1200	Mastercard	Tamar	Headley	England	United Kingdom	51.1166667	-0.8166667
Product2	3600	Mastercard	Deirdre	Lausanne	Vaud	Switzerland	46.5333333	6.6666667
Product1	1200	Mastercard	Bernadett	Southampton	England	United Kingdom	50.9	-1.4
Product1	1200	Visa	Dottie	Woodsboro	MD	United States	39.53306	-77.315
Product1	1200	Visa	Stefan	Stavanger	Rogaland	Norway	58.9666667	5.75
Product1	1200	Visa	Gina	Red Deer	Alberta	Canada	52.2666667	-113.8
Product1	1200	Diners	Lynne	Memphis	TN	United States	35.14944	-90.04889
Product1	1200	Mastercard	Tammy	Morges	Vaud	Switzerland	46.5166667	6.5
Product1	1200	Visa	Kim	Calgary	Alberta	Canada	51.0833333	-114.0833333
Product1	1200	Visa	Bruce	Belleville	Ontario	Canada	44.1666667	-77.3833333
Product1	1200	Visa	Rosa Maria	Cincinnati	OH	United States	39.16194	-84.45694
Product1	1200	Visa	Lydia	Comox	British Columbia	Canada	49.6833333	-124.9333333
Product1	1200	Visa	Eric	Gasperich	Luxembourg	Luxembourg	49.5855556	6.1230556
Product1	1200	Mastercard	AnaPaula	Helens Bay	Northern Ireland	United Kingdom	54.65	-5.7333333
Product1	1200	Visa	Robin	Milan	Lombardy	Italy	45.4666667	9.2
Product1	1200	Visa	Gitte	Staten Island	NY	United States	40.63667	-74.15917
Product1	1200	Visa	Dr. Claudia	Oslo	Oslo	Norway	59.9166667	10.75
Product1	1200	Visa	Crystal	Farmington	Michigan	United States	42.46444	-83.37639
Product1	1200	Diners	Delphine	Santa Monica	CA	United States	34.01944	-118.49028
Product1	1200	Visa	nathalie	Calgary	Alberta	Canada	51.0833333	-114.0833333
Product1	1200	Mastercard	Lindi	Vancouver	British Columbia	Canada	49.25	-123.1333333
Product2	3600	Mastercard	Valda	Irvine	CA	United States	33.66944	-117.82222

Source Code

SalesCountryDriver.java

```
package sales;

import org.apache.hadoop.fs.Path;

public class SalesCountryDriver{

    public static void main(String[] args) {

        JobClient my_client = new JobClient();

        // Create a configuration object for the job
        JobConf job conf = new JobConf (SalesCountryDriver.class);
        // Set a name of the Job
        job_conf.setJobName("SalePerCountry");

        // Specify data type of output key and value
        job_conf.setOutputKeyClass (Text.class);
        job_conf.setOutputValueClass (IntWritable.class);

        // Specify names of Mapper and Reducer Class
        job_conf.setMapperclass (sales.SalesMapper.class);
        job_conf.setReducerClass(sales.SalesCountryReducer.class):

        // Specify formats of the data type of Input and output
        job_conf.setInputFormat (TextInputFormat.class);
        job_conf.setOutputFormat (TextOutputFormat.class);

        // Set input and output directories using command line arguments.
        // arg[0] = name of input directory on HOFS, and arg[1]= name of output directory to be created to store the output
        FileInputFormat.setInputPaths (job conf, new Path(args[e]));
        FileOutputFormat.setOutputPath(job conf, new Path(args[1]));

        my_client.setConf(job_conf);

        try {
            // Run the job
            JobClient.runJob(jab_conf);
        } catch (Exception e) {
            e.printStackTrace();
        }

    }
}
```

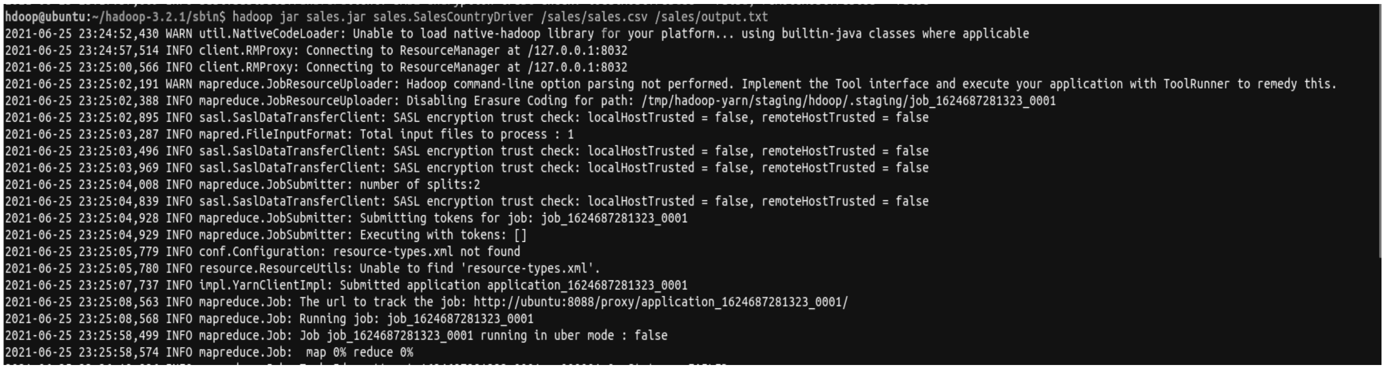
SalesMapper.java

```
package sales;

import java.io.IOException;

public class SalesMapper extends MapReduceBase implements Mapper <LongWritable, Text, Text, IntWritable> {

        private final static IntWritable one = new IntWritable(1);

        public void map (LongWritable key, Text value, OutputCollector <Text, IntWritable> output, Reporter reporter)
throws IOException {

                String valueString= value.toString();
                String[] SingleCountryData = valueString.split(".");
                output.collect (new Text(SingleCountryData[7]), one);


        }

}
```

SalesCountryReducer.java

```
package sales;

import java.io.IOException;

public class SalesCountryReducer extends MapReduceBase implements Reducer<Text, IntWritable, Text,
IntWritable> {

        public void reduce (Text t_key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output,
Reporter report  throws IOException {

                Text key = t_key;

                int frequencyForCountry = 0;
                while (values.hasNext()) {

                        // replace type of value with the actual type of our value
                        IntWritable value = (IntWritable) values.next();
                        frequencyForCountry += value.get();

                }
                output.collect(key, new Intwritable (frequencyForCountry));
        }
}
```

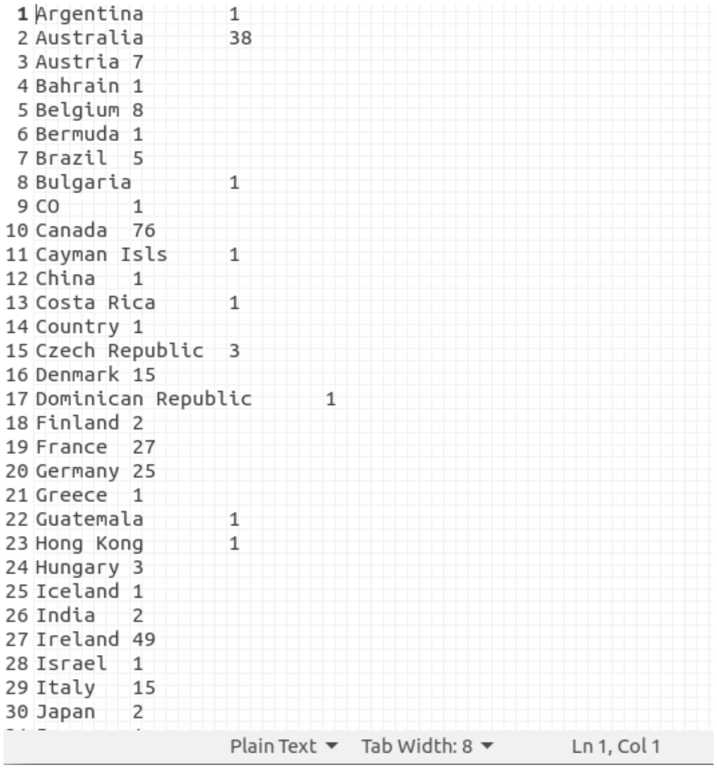
Execution



```
hdoop@ubuntu:~/hadoop-3.2.1/sbin$ hadoop jar sales.jar sales.SalesCountryDriver /sales/sales.csv /sales/output.txt
2021-06-25 23:24:52,430 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-06-25 23:24:57,514 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-06-25 23:25:00,566 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-06-25 23:25:02,191 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-06-25 23:25:02,388 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1624687281323_0001
2021-06-25 23:25:02,895 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-06-25 23:25:03,287 INFO mapred.FileInputFormat: Total input files to process : 1
2021-06-25 23:25:03,496 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-06-25 23:25:03,969 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-06-25 23:25:04,008 INFO mapreduce.JobSubmitter: number of splits:2
2021-06-25 23:25:04,839 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-06-25 23:25:04,928 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1624687281323_0001
2021-06-25 23:25:04,929 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-06-25 23:25:05,779 INFO conf.Configuration: resource-types.xml not found
2021-06-25 23:25:05,780 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-06-25 23:25:07,737 INFO impl.YarnClientImpl: Submitted application application_1624687281323_0001
2021-06-25 23:25:08,563 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1624687281323_0001/
2021-06-25 23:25:08,568 INFO mapreduce.Job: Running job: job_1624687281323_0001
2021-06-25 23:25:58,499 INFO mapreduce.Job: Job job_1624687281323_0001 running in uber mode : false
2021-06-25 23:25:58,574 INFO mapreduce.Job:  map 0% reduce 0%
```

MapReduce Output

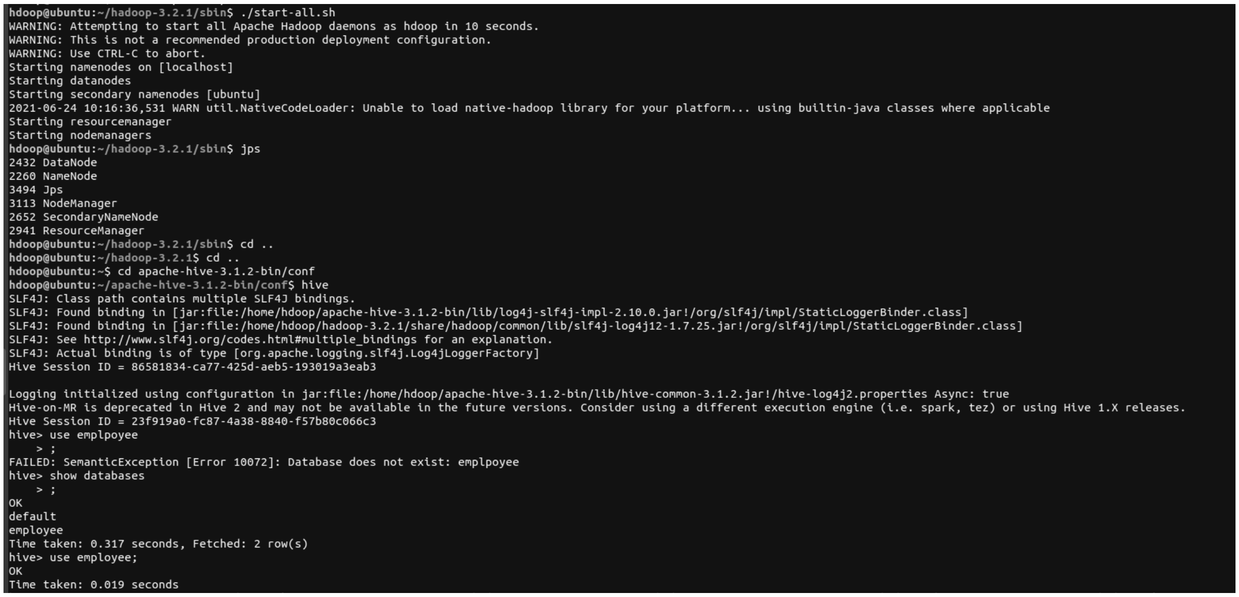
The output displays the total number of products sold for each country.



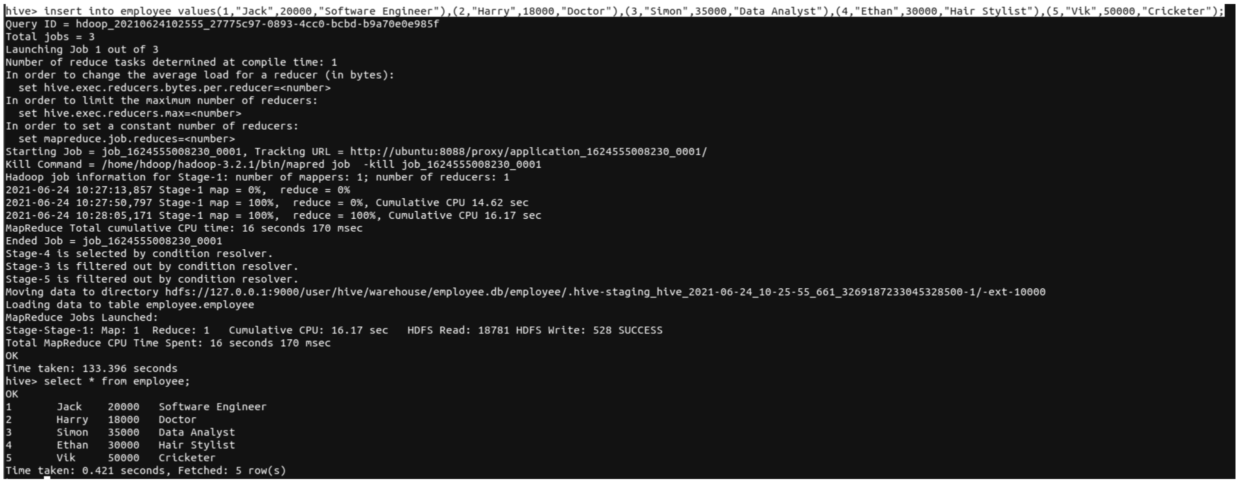
```
Argentina       1
Australia       38
Austria 7
Bahrain 1
Belgium 8
Bermuda 1
Brazil  5
Bulgaria        1
CO      1
Canada  76
Cayman Isls     1
China   1
Costa Rica      1
Country 1
Czech Republic  3
Denmark 15
Dominican Republic      1
Finland 2
France  27
Germany 25
Greece  1
Guatemala       1
Hong Kong       1
Hungary 3
Iceland 1
India   2
Ireland 49
Israel  1
Italy   15
Japan   2
```

<u>Exercise 2</u>: Implement a map-reduce program in JAVA or Python using any Data of your choice.

1. Insert 5 records using the INSERT command.

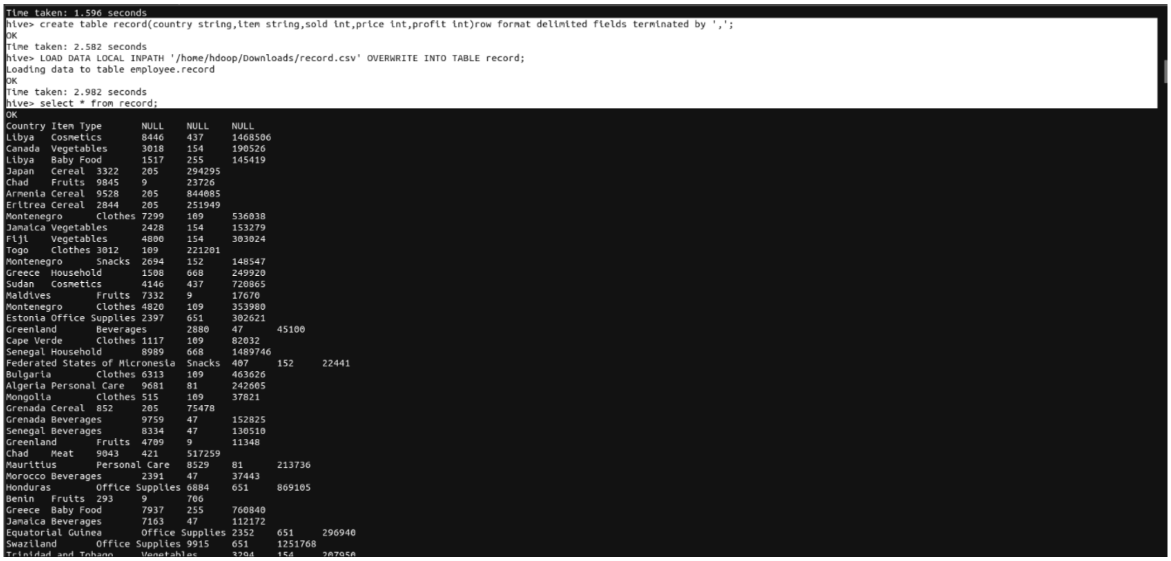


```
hdoop@ubuntu:~/hadoop-3.2.1/sbin$ ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hdoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu]
2021-06-24 10:16:36,531 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting resourcemanager
Starting nodemanagers
hdoop@ubuntu:~/hadoop-3.2.1/sbin$ jps
2432 DataNode
2260 NameNode
3494 Jps
3113 NodeManager
2652 SecondaryNameNode
2941 ResourceManager
hdoop@ubuntu:~/hadoop-3.2.1/sbin$ cd ..
hdoop@ubuntu:~/hadoop-3.2.1$ cd ..
hdoop@ubuntu:~$ cd apache-hive-3.1.2-bin/conf
hdoop@ubuntu:~/apache-hive-3.1.2-bin/conf$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 86581834-ca77-425d-aeb5-193019a3eab3

Logging initialized using configuration in jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Hive Session ID = 23f919a0-fc87-4a38-8840-f57b80c066c3
hive> use emplpoyee
    > ;
FAILED: SemanticException [Error 10072]: Database does not exist: emplpoyee
hive> show databases
    > ;
OK
default
employee
Time taken: 0.317 seconds, Fetched: 2 row(s)
hive> use employee;
OK
Time taken: 0.019 seconds
```



```
hive> insert into employee values(1,"Jack",20000,"Software Engineer"),(2,"Harry",18000,"Doctor"),(3,"Simon",35000,"Data Analyst"),(4,"Ethan",30000,"Hair Stylist"),(5,"Vik",50000,"Cricketer");
Query ID = hdoop_20210624102555_27775c97-0893-4cc0-bcbd-b9a70e0e985f
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1624555008230_0001, Tracking URL = http://ubuntu:8088/proxy/application_1624555008230_0001/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624555008230_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-24 10:27:13,857 Stage-1 map = 0%,  reduce = 0%
2021-06-24 10:27:50,797 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 14.62 sec
2021-06-24 10:28:05,171 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 16.17 sec
MapReduce Total cumulative CPU time: 16 seconds 170 msec
Ended Job = job_1624555008230_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://127.0.0.1:9000/user/hive/warehouse/employee.db/employee/.hive-staging_hive_2021-06-24_10-25-55_661_3269187233045328500-1/-ext-10000
Loading data to table employee.employee
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 16.17 sec   HDFS Read: 18781 HDFS Write: 528 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 170 msec
OK
Time taken: 133.396 seconds
hive> select * from employee;
OK
1       Jack    20000   Software Engineer
2       Harry   18000   Doctor
3       Simon   35000   Data Analyst
4       Ethan   30000   Hair Stylist
5       Vik     50000   Cricketer
Time taken: 0.421 seconds, Fetched: 5 row(s)
```

2. Load the data(text or CSV) into the table.

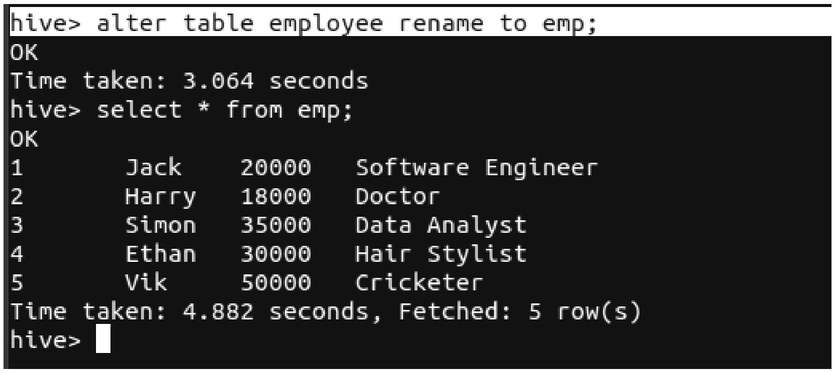


```
Time taken: 1.596 seconds
hive> create table record(country string,item string,sold int,price int,profit int)row format delimited fields terminated by ',';
OK
Time taken: 2.582 seconds
hive> LOAD DATA LOCAL INPATH '/home/hdoop/Downloads/record.csv' OVERWRITE INTO TABLE record;
Loading data to table employee.record
OK
Time taken: 2.982 seconds
hive> select * from record;
OK
Country Item Type       NULL    NULL    NULL
Libya   Cosmetics       8446    437     1468506
Canada  Vegetables      3018    154     190526
Libya   Baby Food       1517    255     145419
Japan   Cereal  3322    205     294295
Chad    Fruits  9845    9       23726
Armenia Cereal  9528    205     844085
Eritrea Cereal  2844    205     251949
Montenegro      Clothes 7299    109     536038
Jamaica Vegetables      2428    154     153279
Fiji    Vegetables      4800    154     303024
Togo    Clothes 3012    109     221201
Montenegro      Snacks  2694    152     148547
Greece  Household       1508    668     249920
Sudan   Cosmetics       4146    437     720865
Maldives        Fruits  7332    9       17670
Montenegro      Clothes 4820    109     353980
Estonia Office Supplies 2397    651     302621
Greenland       Beverages       2880    47      45100
Cape Verde      Clothes 1117    109     82032
Senegal Household       8989    668     1489746
Federated States of Micronesia  Snacks  407     152     22441
Bulgaria        Clothes 6313    109     463626
Algeria Personal Care   9681    81      242605
Mongolia        Clothes 515     109     37821
Grenada Cereal  852     205     75478
Grenada Beverages       9759    47      152825
Senegal Beverages       8334    47      130510
Greenland       Fruits  4709    9       11348
Chad    Meat    9043    421     517259
Mauritius       Personal Care   8529    81      213736
Morocco Beverages       2391    47      37443
Honduras        Office Supplies 6884    651     869105
Benin   Fruits  293     9       706
Greece  Baby Food       7937    255     760840
Jamaica Beverages       7163    47      112172
Equatorial Guinea       Office Supplies 2352    651     296940
Swaziland       Office Supplies 9915    651     1251768
Trinidad and Tobago     Vegetables      3294    154     207950
```

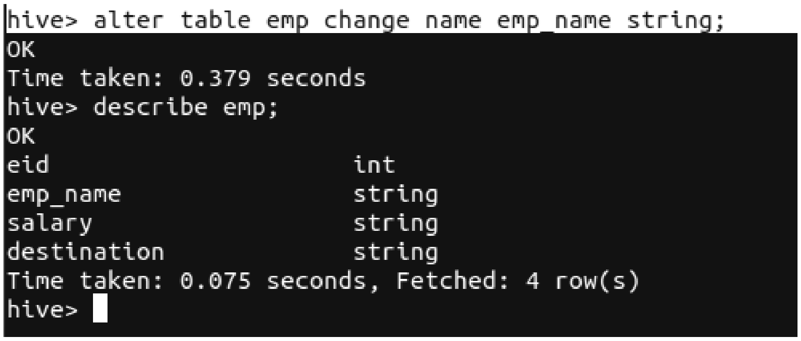
3. Demonstrate the Alter command for the following cases:

I. Rename the table name

II. Rename the column name “C1” to “C2”

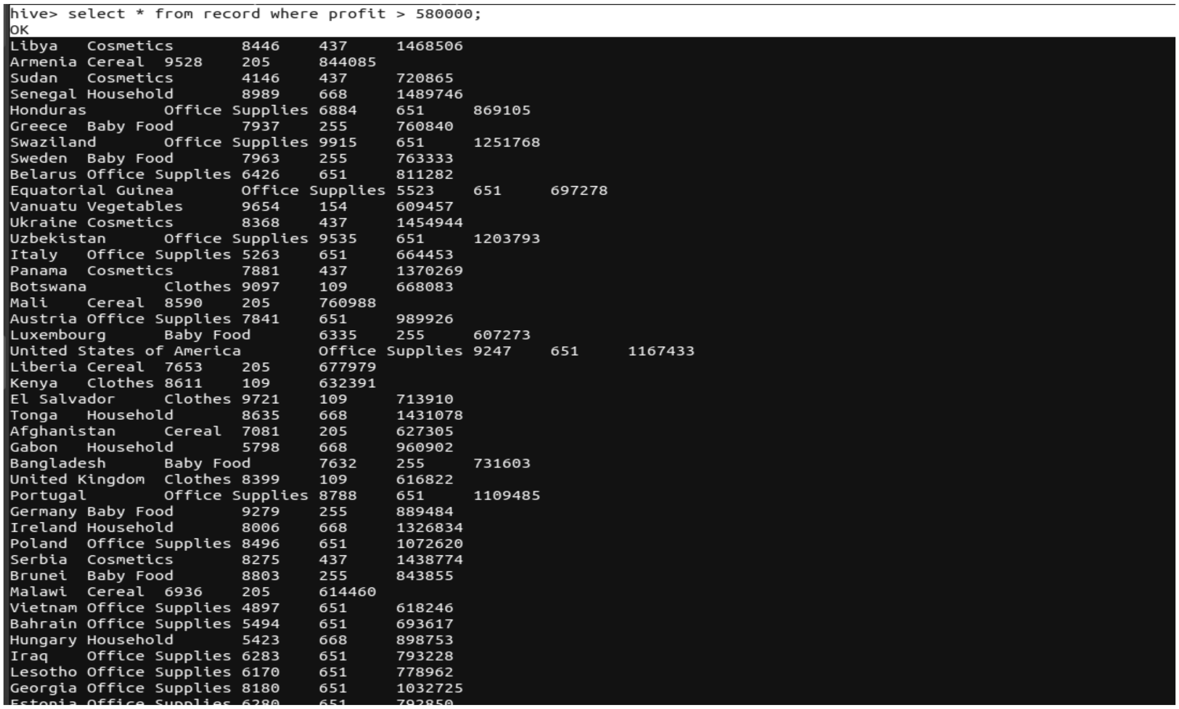


```
hive> alter table employee rename to emp;
OK
Time taken: 3.064 seconds
hive> select * from emp;
OK
1       Jack    20000   Software Engineer
2       Harry   18000   Doctor
3       Simon   35000   Data Analyst
4       Ethan   30000   Hair Stylist
5       Vik     50000   Cricketer
Time taken: 4.882 seconds, Fetched: 5 row(s)
hive> 
```

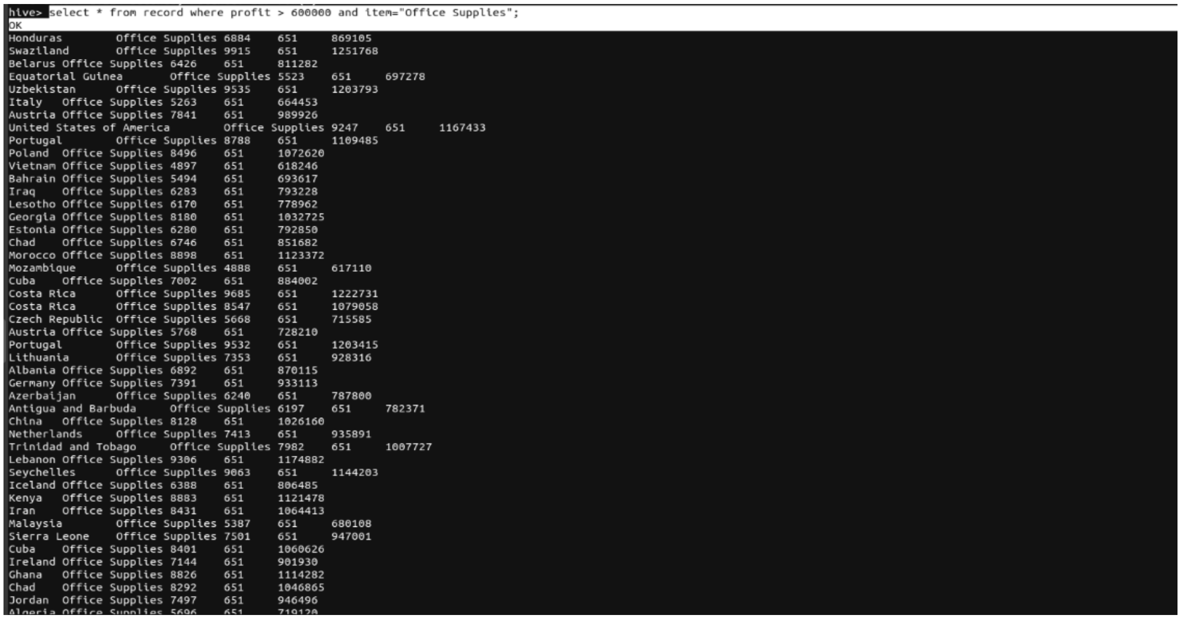


```
hive> alter table emp change name emp_name string;
OK
Time taken: 0.379 seconds
hive> describe emp;
OK
eid                     int
emp_name                string
salary                  string
destination             string
Time taken: 0.075 seconds, Fetched: 4 row(s)
hive> 
```

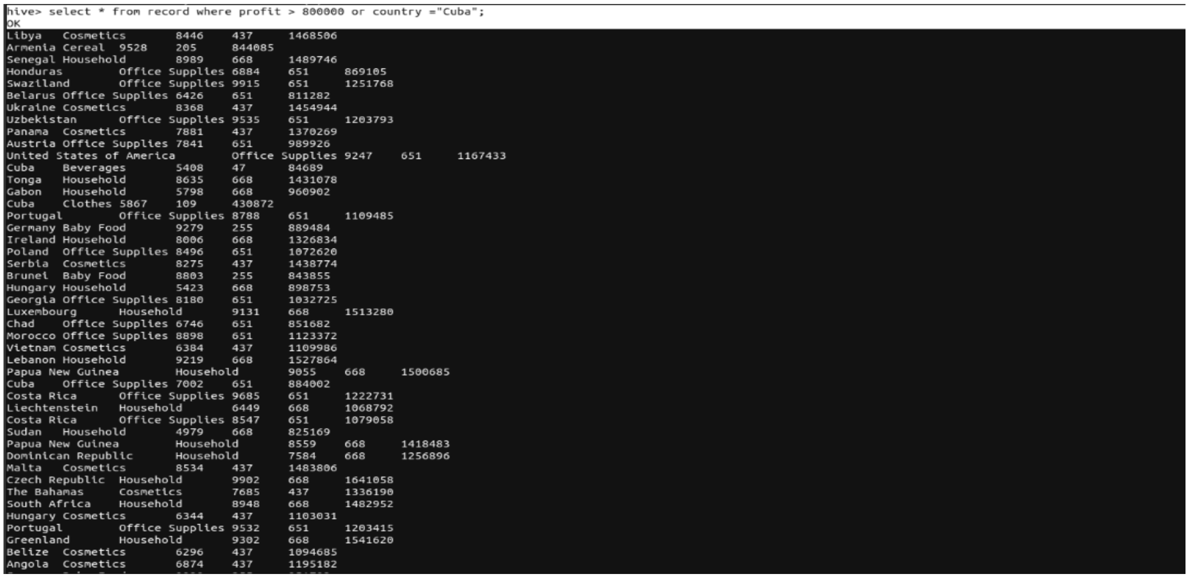
4. AND, OR, IN, NOTIN, SUBSTR, CONCAT, Case operators



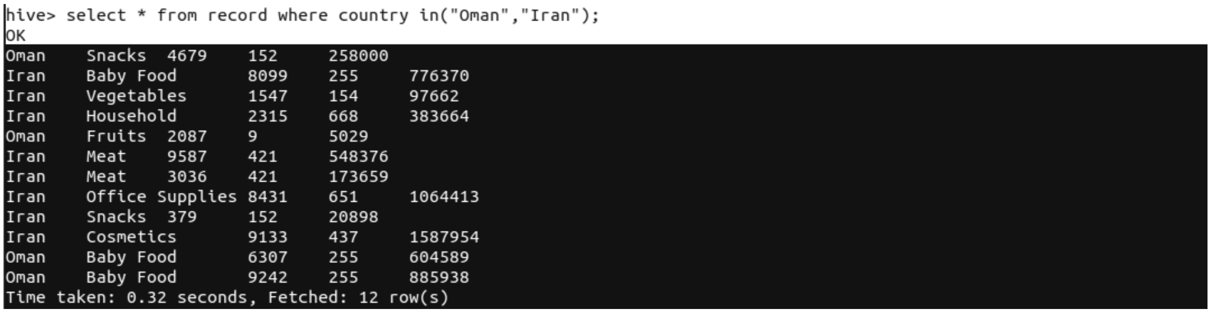
```
hive> select * from record where profit > 580000;
OK
Libya   Cosmetics       8446    437     1468506
Armenia Cereal  9528    205     844085
Sudan   Cosmetics       4146    437     720865
Senegal Household       8989    668     1489746
Honduras        Office Supplies 6884    651     869105
Greece  Baby Food       7937    255     760840
Swaziland       Office Supplies 9915    651     1251768
Sweden  Baby Food       7963    255     763333
Belarus Office Supplies 6426    651     811282
Equatorial Guinea       Office Supplies 5523    651     697278
Vanuatu Vegetables      9654    154     609457
Ukraine Cosmetics       8368    437     1454944
Uzbekistan      Office Supplies 9535    651     1203793
Italy   Office Supplies 5263    651     664453
Panama  Cosmetics       7881    437     1370269
Botswana        Clothes 9097    109     668083
Mali    Cereal  8590    205     760988
Austria Office Supplies 7841    651     989926
Luxembourg      Baby Food       6335    255     607273
United States of America        Office Supplies 9247    651     1167433
Liberia Cereal  7653    205     677979
Kenya   Clothes 8611    109     632391
El Salvador     Clothes 9721    109     713910
Tonga   Household       8635    668     1431078
Afghanistan     Cereal  7081    205     627305
Gabon   Household       5798    668     960902
Bangladesh      Baby Food       7632    255     731603
United Kingdom  Clothes 8399    109     616822
Portugal        Office Supplies 8788    651     1109485
Germany Baby Food       9279    255     889484
Ireland Household       8006    668     1326834
Poland  Office Supplies 8496    651     1072620
Serbia  Cosmetics       8275    437     1438774
Brunei  Baby Food       8803    255     843855
Malawi  Cereal  6936    205     614460
Vietnam Office Supplies 4897    651     618246
Bahrain Office Supplies 5494    651     693617
Hungary Household       5423    668     898753
Iraq    Office Supplies 6283    651     793228
Lesotho Office Supplies 6170    651     778962
Georgia Office Supplies 8180    651     1032725
Estonia Office Supplies 6280    651     792850
```



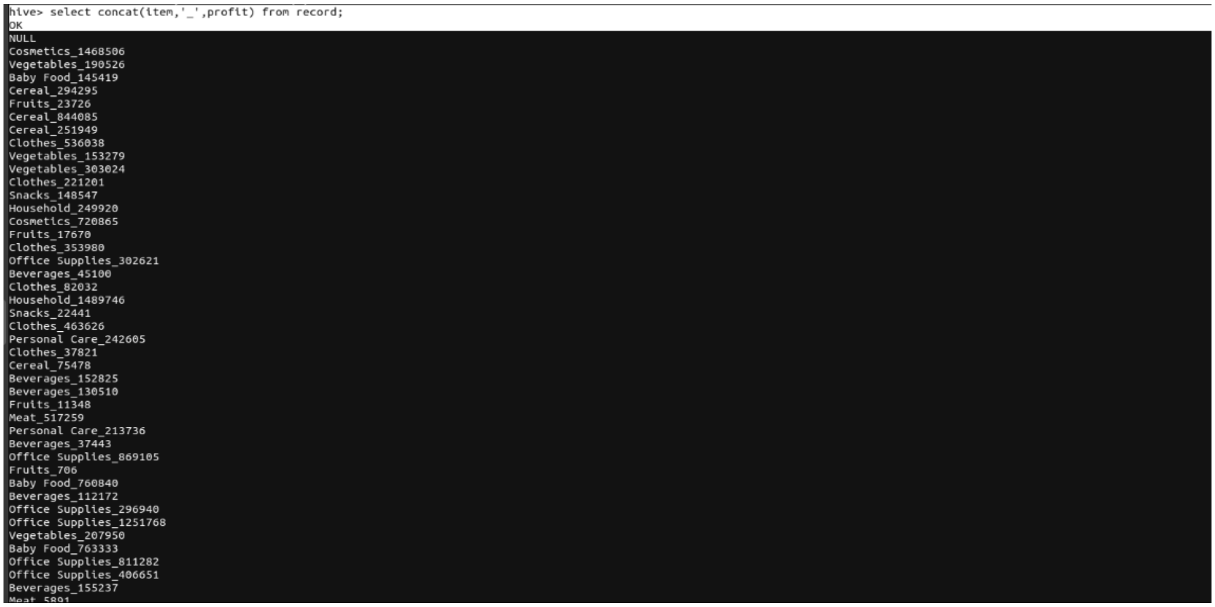
```
hive> select * from record where profit > 600000 and item="Office Supplies";
OK
Honduras        Office Supplies 6884    651     869105
Swaziland       Office Supplies 9915    651     1251768
Belarus Office Supplies 6426    651     811282
Equatorial Guinea       Office Supplies 5523    651     697278
Uzbekistan      Office Supplies 9535    651     1203793
Italy   Office Supplies 5263    651     664453
Austria Office Supplies 7841    651     989926
United States of America        Office Supplies 9247    651     1167433
Portugal        Office Supplies 8788    651     1109485
Poland  Office Supplies 8496    651     1072620
Vietnam Office Supplies 4897    651     618246
Bahrain Office Supplies 5494    651     693617
Iraq    Office Supplies 6283    651     793228
Lesotho Office Supplies 6170    651     778962
Georgia Office Supplies 8180    651     1032725
Estonia Office Supplies 6280    651     792850
Chad    Office Supplies 6746    651     851682
Morocco Office Supplies 8898    651     1123372
Mozambique      Office Supplies 4888    651     617110
Cuba    Office Supplies 7002    651     884002
Costa Rica      Office Supplies 9685    651     1222731
Costa Rica      Office Supplies 8547    651     1079058
Czech Republic  Office Supplies 5668    651     715585
Austria Office Supplies 5768    651     728210
Portugal        Office Supplies 9532    651     1203415
Lithuania       Office Supplies 7353    651     928316
Albania Office Supplies 6892    651     870115
Germany Office Supplies 7391    651     933113
Azerbaijan      Office Supplies 6240    651     787800
Antigua and Barbuda     Office Supplies 6197    651     782371
China   Office Supplies 8128    651     1026160
Netherlands     Office Supplies 7413    651     935891
Trinidad and Tobago     Office Supplies 7982    651     1007727
Lebanon Office Supplies 9306    651     1174882
Seychelles      Office Supplies 9063    651     1144203
Iceland Office Supplies 6388    651     806485
Kenya   Office Supplies 8883    651     1121478
Iran    Office Supplies 8431    651     1064413
Malaysia        Office Supplies 5387    651     680108
Sierra Leone    Office Supplies 7501    651     947001
Cuba    Office Supplies 8401    651     1060626
Ireland Office Supplies 7144    651     901930
Ghana   Office Supplies 8826    651     1114282
Chad    Office Supplies 8292    651     1046865
Jordan  Office Supplies 7497    651     946496
Algeria Office Supplies 5696    651     719120
```



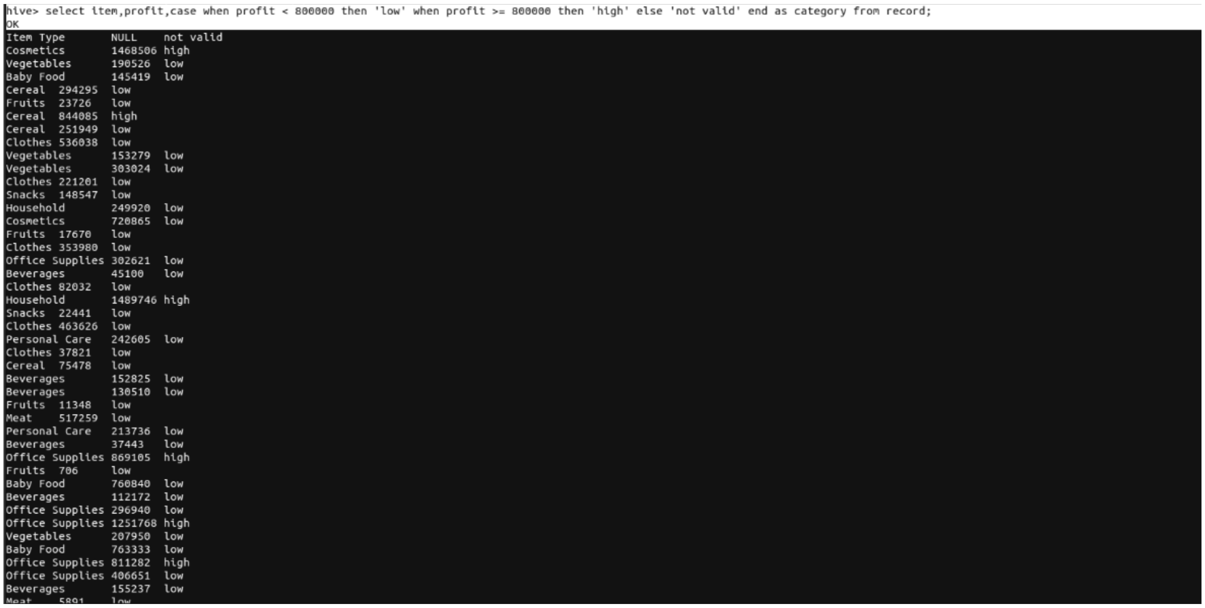
```
hive> select * from record where profit > 800000 or country ="Cuba";
OK
Libya   Cosmetics       8446    437     1468506
Armenia Cereal  9528    205     844085
Senegal Household       8989    668     1489746
Honduras        Office Supplies 6884    651     869105
Swaziland       Office Supplies 9915    651     1251768
Belarus Office Supplies 6426    651     811282
Ukraine Cosmetics       8368    437     1454944
Uzbekistan      Office Supplies 9535    651     1203793
Panama  Cosmetics       7881    437     1370269
Austria Office Supplies 7841    651     989926
United States of America        Office Supplies 9247    651     1167433
Cuba    Beverages       5408    47      84689
Tonga   Household       8635    668     1431078
Gabon   Household       5798    668     960902
Cuba    Clothes 5867    109     430872
Portugal        Office Supplies 8788    651     1109485
Germany Baby Food       9279    255     889484
Ireland Household       8006    668     1326834
Poland  Office Supplies 8496    651     1072620
Serbia  Cosmetics       8275    437     1438774
Brunei  Baby Food       8803    255     843855
Hungary Household       5423    668     898753
Georgia Office Supplies 8180    651     1032725
Luxembourg      Household       9131    668     1513280
Chad    Office Supplies 6746    651     851682
Morocco Office Supplies 8898    651     1123372
Vietnam Cosmetics       6384    437     1109986
Lebanon Household       9219    668     1527864
Papua New Guinea        Household       9055    668     1500685
Cuba    Office Supplies 7002    651     884002
Costa Rica      Office Supplies 9685    651     1222731
Liechtenstein   Household       6449    668     1068792
Costa Rica      Office Supplies 8547    651     1079058
Sudan   Household       4979    668     825169
Papua New Guinea        Household       8559    668     1418483
Dominican Republic      Household       7584    668     1256896
Malta   Cosmetics       8534    437     1483806
Czech Republic  Household       9902    668     1641058
The Bahamas     Cosmetics       7685    437     1336190
South Africa    Household       8948    668     1482952
Hungary Cosmetics       6344    437     1103031
Portugal        Office Supplies 9532    651     1203415
Greenland       Household       9302    668     1541620
Belize  Cosmetics       6296    437     1094685
Angola  Cosmetics       6874    437     1195182
```



```
hive> select * from record where country in("Oman","Iran");
OK
Oman    Snacks  4679    152     258000
Iran    Baby Food       8099    255     776370
Iran    Vegetables      1547    154     97662
Iran    Household       2315    668     383664
Oman    Fruits  2087    9       5029
Iran    Meat    9587    421     548376
Iran    Meat    3036    421     173659
Iran    Office Supplies 8431    651     1064413
Iran    Snacks  379     152     20898
Iran    Cosmetics       9133    437     1587954
Oman    Baby Food       6307    255     604589
Oman    Baby Food       9242    255     885938
Time taken: 0.32 seconds, Fetched: 12 row(s)
```

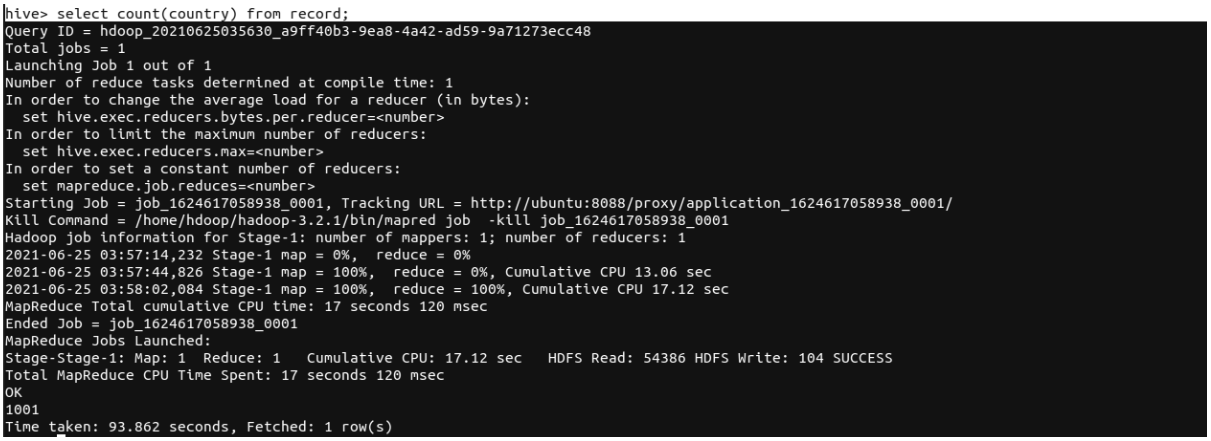


```
hive> select concat(item,'_',profit) from record;
OK
NULL
Cosmetics_1468506
Vegetables_190526
Baby Food_145419
Cereal_294295
Fruits_23726
Cereal_844085
Cereal_251949
Clothes_536038
Vegetables_153279
Vegetables_303024
Clothes_221201
Snacks_148547
Household_249920
Cosmetics_720865
Fruits_17670
Clothes_353980
Office Supplies_302621
Beverages_45100
Clothes_82032
Household_1489746
Snacks_22441
Clothes_463626
Personal Care_242605
Clothes_37821
Cereal_75478
Beverages_152825
Beverages_130510
Fruits_11348
Meat_517259
Personal Care_213736
Beverages_37443
Office Supplies_869105
Fruits_706
Baby Food_760840
Beverages_112172
Office Supplies_296940
Office Supplies_1251768
Vegetables_207950
Baby Food_763333
Office Supplies_811282
Office Supplies_406651
Beverages_155237
Meat_5891
```

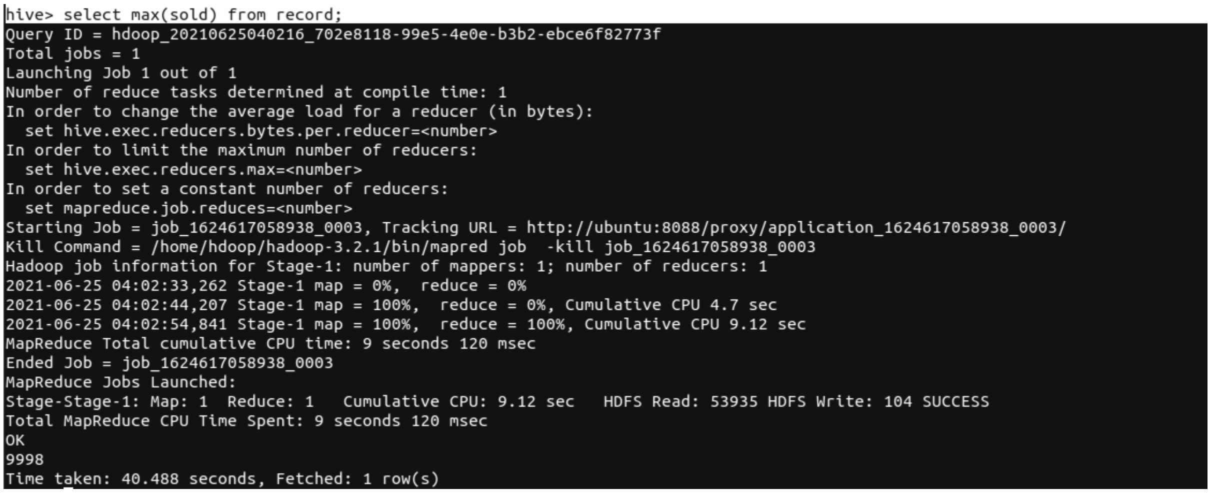


```
hive> select item,profit,case when profit < 800000 then 'low' when profit >= 800000 then 'high' else 'not valid' end as category from record;
OK
Item Type       NULL    not valid
Cosmetics       1468506 high
Vegetables      190526  low
Baby Food       145419  low
Cereal  294295  low
Fruits  23726   low
Cereal  844085  high
Cereal  251949  low
Clothes 536038  low
Vegetables      153279  low
Vegetables      303024  low
Clothes 221201  low
Snacks  148547  low
Household       249920  low
Cosmetics       720865  low
Fruits  17670   low
Clothes 353980  low
Office Supplies 302621  low
Beverages       45100   low
Clothes 82032   low
Household       1489746 high
Snacks  22441   low
Clothes 463626  low
Personal Care   242605  low
Clothes 37821   low
Cereal  75478   low
Beverages       152825  low
Beverages       130510  low
Fruits  11348   low
Meat    517259  low
Personal Care   213736  low
Beverages       37443   low
Office Supplies 869105  high
Fruits  706     low
Baby Food       760840  low
Beverages       112172  low
Office Supplies 296940  low
Office Supplies 1251768 high
Vegetables      207950  low
Baby Food       763333  low
Office Supplies 811282  high
Office Supplies 406651  low
Beverages       155237  low
Meat    5891    low
```

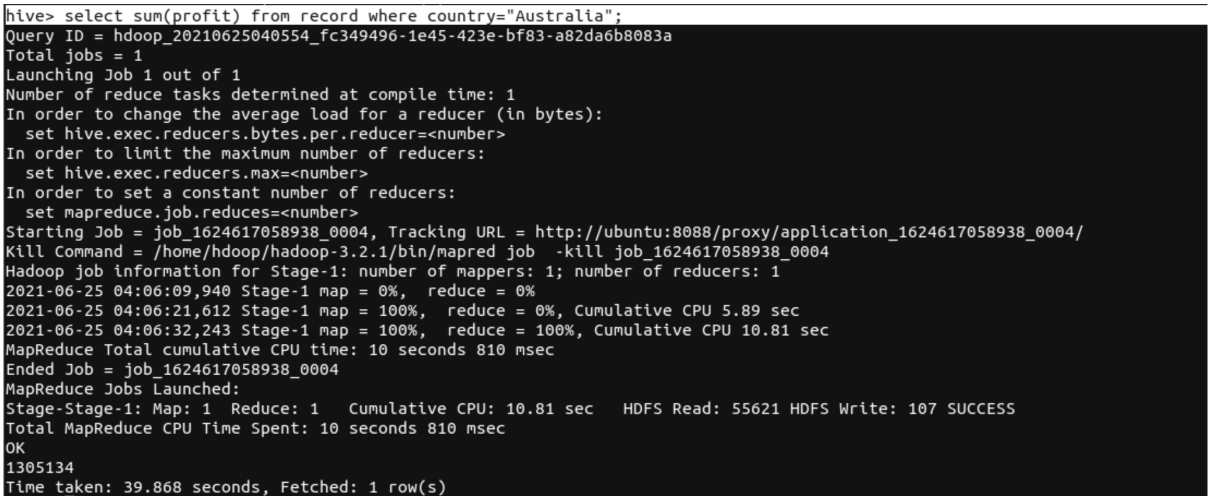
5. Aggregate functions COUNT, MIN, MAX, SUM, AVG



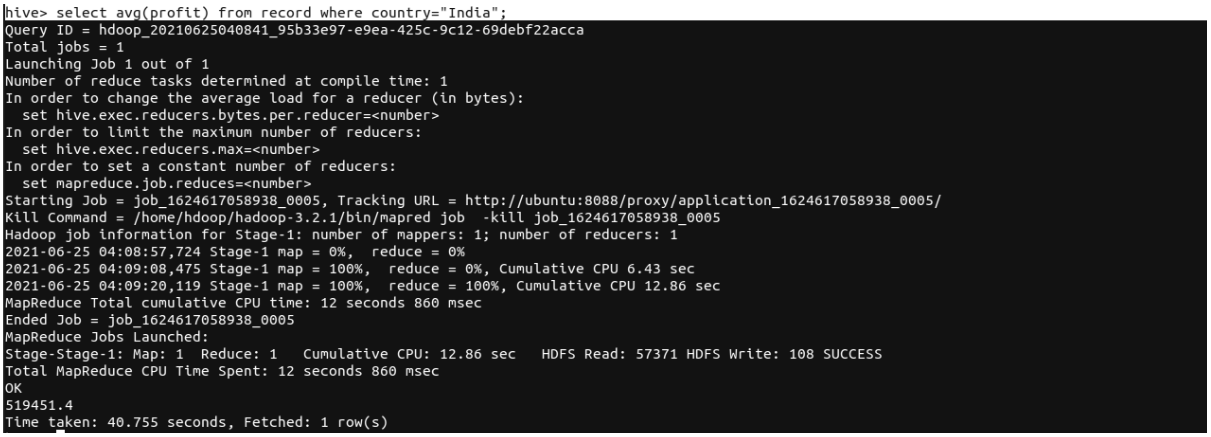
```
hive> select count(country) from record;
Query ID = hdoop_20210625035630_a9ff40b3-9ea8-4a42-ad59-9a71273ecc48
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1624617058938_0001, Tracking URL = http://ubuntu:8088/proxy/application_1624617058938_0001/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624617058938_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-25 03:57:14,232 Stage-1 map = 0%,  reduce = 0%
2021-06-25 03:57:44,826 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.06 sec
2021-06-25 03:58:02,084 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.12 sec
MapReduce Total cumulative CPU time: 17 seconds 120 msec
Ended Job = job_1624617058938_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 17.12 sec   HDFS Read: 54386 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 120 msec
OK
1001
Time taken: 93.862 seconds, Fetched: 1 row(s)
```



```
hive> select max(sold) from record;
Query ID = hdoop_20210625040216_702e8118-99e5-4e0e-b3b2-ebce6f82773f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1624617058938_0003, Tracking URL = http://ubuntu:8088/proxy/application_1624617058938_0003/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624617058938_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-25 04:02:33,262 Stage-1 map = 0%,  reduce = 0%
2021-06-25 04:02:44,207 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.7 sec
2021-06-25 04:02:54,841 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.12 sec
MapReduce Total cumulative CPU time: 9 seconds 120 msec
Ended Job = job_1624617058938_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.12 sec   HDFS Read: 53935 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 120 msec
OK
9998
Time taken: 40.488 seconds, Fetched: 1 row(s)
```

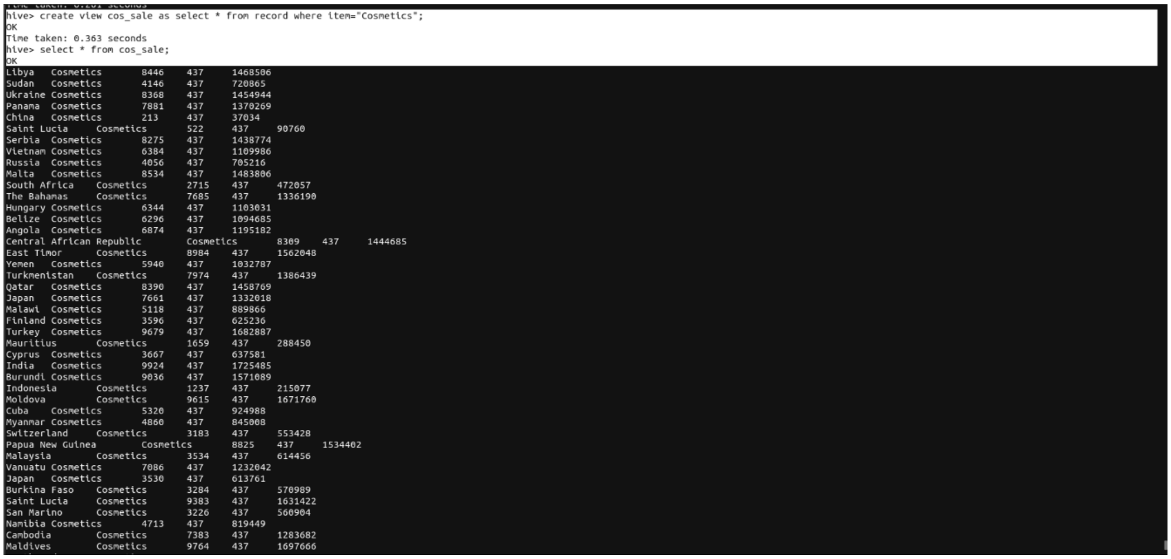


```
hive> select sum(profit) from record where country="Australia";
Query ID = hdoop_20210625040554_fc349496-1e45-423e-bf83-a82da6b8083a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1624617058938_0004, Tracking URL = http://ubuntu:8088/proxy/application_1624617058938_0004/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624617058938_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-25 04:06:09,940 Stage-1 map = 0%,  reduce = 0%
2021-06-25 04:06:21,612 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.89 sec
2021-06-25 04:06:32,243 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.81 sec
MapReduce Total cumulative CPU time: 10 seconds 810 msec
Ended Job = job_1624617058938_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.81 sec   HDFS Read: 55621 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 810 msec
OK
1305134
Time taken: 39.868 seconds, Fetched: 1 row(s)
```



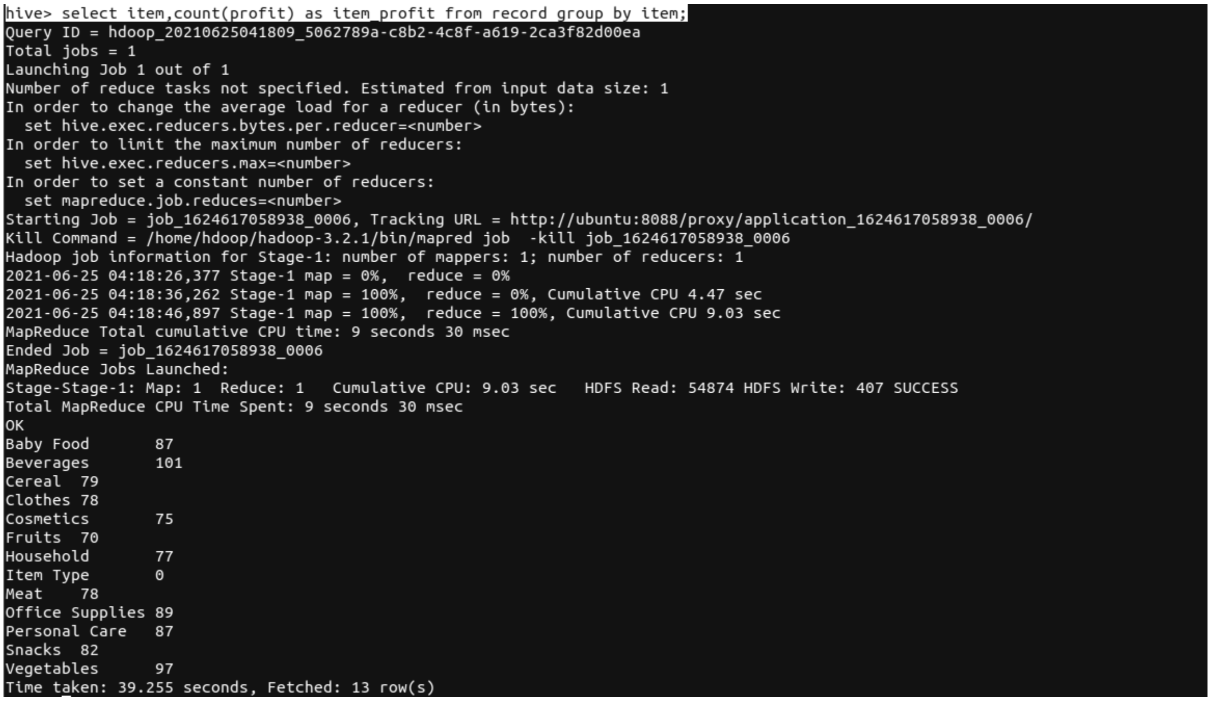
```
hive> select avg(profit) from record where country="India";
Query ID = hdoop_20210625040841_95b33e97-e9ea-425c-9c12-69debf22acca
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1624617058938_0005, Tracking URL = http://ubuntu:8088/proxy/application_1624617058938_0005/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624617058938_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-25 04:08:57,724 Stage-1 map = 0%,  reduce = 0%
2021-06-25 04:09:08,475 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.43 sec
2021-06-25 04:09:20,119 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.86 sec
MapReduce Total cumulative CPU time: 12 seconds 860 msec
Ended Job = job_1624617058938_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.86 sec   HDFS Read: 57371 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 860 msec
OK
519451.4
Time taken: 40.755 seconds, Fetched: 1 row(s)
```

6. Create a separate view



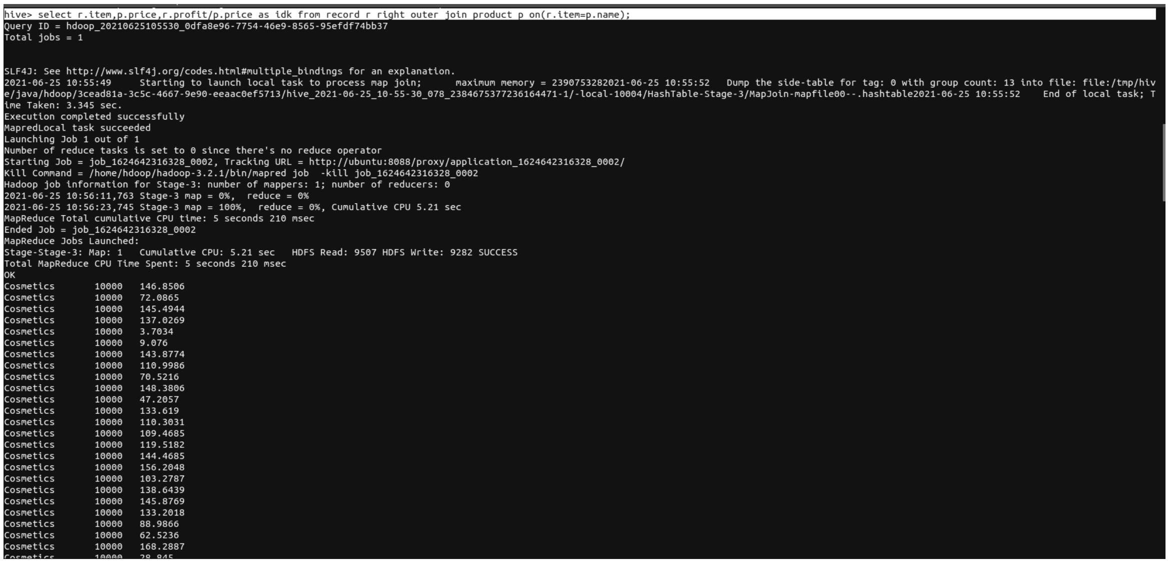
```
Time taken: 0.201 seconds
hive> create view cos_sale as select * from record where item="Cosmetics";
OK
Time taken: 0.363 seconds
hive> select * from cos_sale;
OK
Libya   Cosmetics       8446    437     1468506
Sudan   Cosmetics       4146    437     720865
Ukraine Cosmetics       8368    437     1454944
Panama  Cosmetics       7881    437     1370269
China   Cosmetics       213     437     37034
Saint Lucia     Cosmetics       522     437     90760
Serbia  Cosmetics       8275    437     1438774
Vietnam Cosmetics       6384    437     1109986
Russia  Cosmetics       4056    437     705216
Malta   Cosmetics       8534    437     1483806
South Africa    Cosmetics       2715    437     472057
The Bahamas     Cosmetics       7685    437     1336190
Hungary Cosmetics       6344    437     1103031
Belize  Cosmetics       6296    437     1094685
Angola  Cosmetics       6874    437     1195182
Central African Republic        Cosmetics       8309    437     1444685
East Timor      Cosmetics       8984    437     1562048
Yemen   Cosmetics       5940    437     1032787
Turkmenistan    Cosmetics       7974    437     1386439
Qatar   Cosmetics       8390    437     1458769
Japan   Cosmetics       7661    437     1332018
Malawi  Cosmetics       5118    437     889866
Finland Cosmetics       3596    437     625236
Turkey  Cosmetics       9679    437     1682887
Mauritius       Cosmetics       1659    437     288450
Cyprus  Cosmetics       3667    437     637581
India   Cosmetics       9924    437     1725485
Burundi Cosmetics       9036    437     1571089
Indonesia       Cosmetics       1237    437     215077
Moldova         Cosmetics       9615    437     1671760
Cuba    Cosmetics       5320    437     924988
Myanmar Cosmetics       4860    437     845008
Switzerland     Cosmetics       3183    437     553428
Papua New Guinea        Cosmetics       8825    437     1534402
Malaysia        Cosmetics       3534    437     614456
Vanuatu Cosmetics       7086    437     1232042
Japan   Cosmetics       3530    437     613761
Burkina Faso    Cosmetics       3284    437     570989
Saint Lucia     Cosmetics       9383    437     1631422
San Marino      Cosmetics       3226    437     560904
Namibia Cosmetics       4713    437     819449
Cambodia        Cosmetics       7383    437     1283682
Maldives        Cosmetics       9764    437     1697666
```

7. GROUP BY

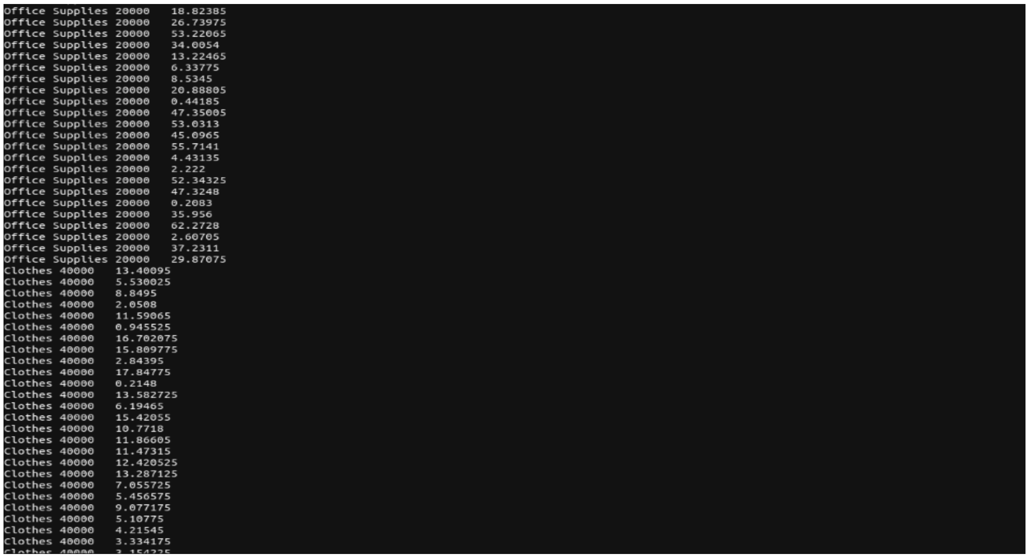


```
hive> select item,count(profit) as item_profit from record group by item;
Query ID = hdoop_20210625041809_5062789a-c8b2-4c8f-a619-2ca3f82d00ea
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1624617058938_0006, Tracking URL = http://ubuntu:8088/proxy/application_1624617058938_0006/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624617058938_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-25 04:18:26,377 Stage-1 map = 0%,  reduce = 0%
2021-06-25 04:18:36,262 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.47 sec
2021-06-25 04:18:46,897 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.03 sec
MapReduce Total cumulative CPU time: 9 seconds 30 msec
Ended Job = job_1624617058938_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.03 sec   HDFS Read: 54874 HDFS Write: 407 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 30 msec
OK
Baby Food       87
Beverages       101
Cereal  79
Clothes 78
Cosmetics       75
Fruits  70
Household       77
Item Type       0
Meat    78
Office Supplies 89
Personal Care   87
Snacks  82
Vegetables      97
Time taken: 39.255 seconds, Fetched: 13 row(s)
```

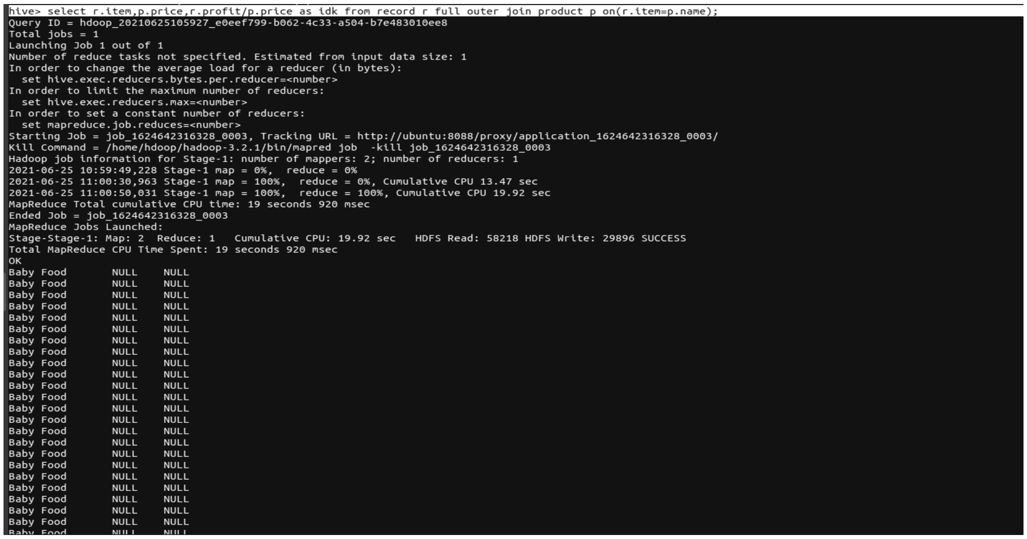
8. Perform the following joins (outer, left outer, right outer)



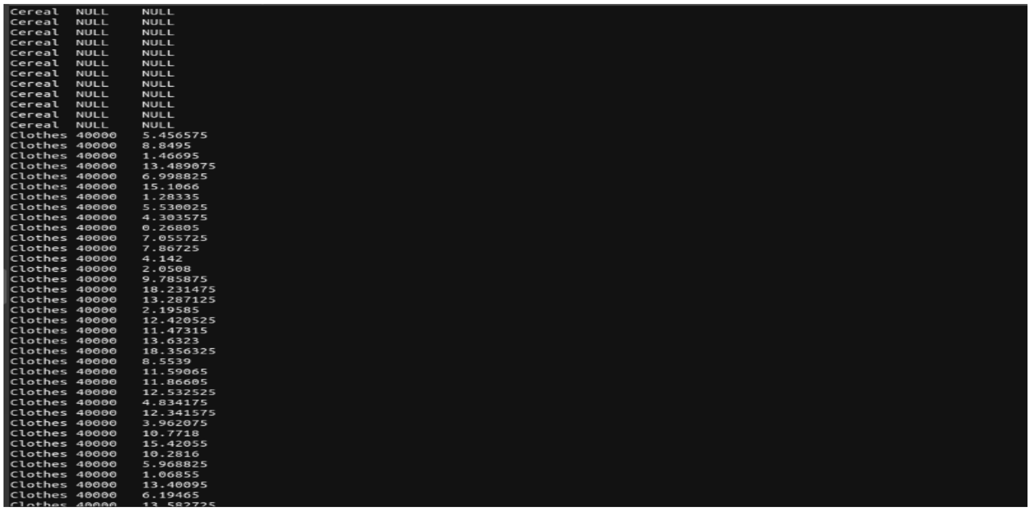
```
hive> select r.item,p.price,r.profit/p.price as idk from record r right outer join product p on(r.item=p.name);
Query ID = hdoop_20210625105530_0dfa8e96-7754-46e9-8565-95efdf74bb37
Total jobs = 1


SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2021-06-25 10:55:49     Starting to launch local task to process map join;      maximum memory = 2390753282021-06-25 10:55:52   Dump the side-table for tag: 0 with group count: 13 into file: file:/tmp/hiv
e/java/hdoop/3cead81a-3c5c-4667-9e90-eeaac0ef5713/hive_2021-06-25_10-55-30_078_2384675377236164471-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile00--.hashtable2021-06-25 10:55:52    End of local task; T
ime Taken: 3.345 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1624642316328_0002, Tracking URL = http://ubuntu:8088/proxy/application_1624642316328_0002/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624642316328_0002
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-06-25 10:56:11,763 Stage-3 map = 0%,  reduce = 0%
2021-06-25 10:56:23,745 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 5.21 sec
MapReduce Total cumulative CPU time: 5 seconds 210 msec
Ended Job = job_1624642316328_0002
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 5.21 sec   HDFS Read: 9507 HDFS Write: 9282 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 210 msec
OK
Cosmetics       10000   146.8506
Cosmetics       10000   72.0865
Cosmetics       10000   145.4944
Cosmetics       10000   137.0269
Cosmetics       10000   3.7034
Cosmetics       10000   9.076
Cosmetics       10000   143.8774
Cosmetics       10000   110.9986
Cosmetics       10000   70.5216
Cosmetics       10000   148.3806
Cosmetics       10000   47.2057
Cosmetics       10000   133.619
Cosmetics       10000   110.3031
Cosmetics       10000   109.4685
Cosmetics       10000   119.5182
Cosmetics       10000   144.4685
Cosmetics       10000   156.2048
Cosmetics       10000   103.2787
Cosmetics       10000   138.6439
Cosmetics       10000   145.8769
Cosmetics       10000   133.2018
Cosmetics       10000   88.9866
Cosmetics       10000   62.5236
Cosmetics       10000   168.2887
Cosmetics       10000   28.845
```



```
Office Supplies 20000   18.82385
Office Supplies 20000   26.73975
Office Supplies 20000   53.22065
Office Supplies 20000   34.0054
Office Supplies 20000   13.22465
Office Supplies 20000   6.33775
Office Supplies 20000   8.5345
Office Supplies 20000   20.88805
Office Supplies 20000   0.44185
Office Supplies 20000   47.35005
Office Supplies 20000   53.0313
Office Supplies 20000   45.0965
Office Supplies 20000   55.7141
Office Supplies 20000   4.43135
Office Supplies 20000   2.222
Office Supplies 20000   52.34325
Office Supplies 20000   47.3248
Office Supplies 20000   0.2083
Office Supplies 20000   35.956
Office Supplies 20000   62.2728
Office Supplies 20000   2.60705
Office Supplies 20000   37.2311
Office Supplies 20000   29.87075
Clothes 40000   13.40095
Clothes 40000   5.530025
Clothes 40000   8.8495
Clothes 40000   2.0508
Clothes 40000   11.59065
Clothes 40000   0.945525
Clothes 40000   16.702075
Clothes 40000   15.809775
Clothes 40000   2.84395
Clothes 40000   17.84775
Clothes 40000   0.2148
Clothes 40000   13.582725
Clothes 40000   6.19465
Clothes 40000   15.42055
Clothes 40000   10.7718
Clothes 40000   11.86605
Clothes 40000   11.47315
Clothes 40000   12.420525
Clothes 40000   13.287125
Clothes 40000   7.055725
Clothes 40000   5.456575
Clothes 40000   9.077175
Clothes 40000   5.10775
Clothes 40000   4.21545
Clothes 40000   3.334175
Clothes 40000   3.154225
```



```
hive> select r.item,p.price,r.profit/p.price as idk from record r full outer join product p on(r.item=p.name);
Query ID = hdoop_20210625105927_e0eef799-b062-4c33-a504-b7e483010ee8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1624642316328_0003, Tracking URL = http://ubuntu:8088/proxy/application_1624642316328_0003/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1624642316328_0003
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2021-06-25 10:59:49,228 Stage-1 map = 0%,  reduce = 0%
2021-06-25 11:00:30,963 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.47 sec
2021-06-25 11:00:50,031 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.92 sec
MapReduce Total cumulative CPU time: 19 seconds 920 msec
Ended Job = job_1624642316328_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 19.92 sec   HDFS Read: 58218 HDFS Write: 29896 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 920 msec
OK
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
Baby Food       NULL    NULL
```



```
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Cereal  NULL    NULL
Clothes 40000   5.456575
Clothes 40000   8.8495
Clothes 40000   1.46695
Clothes 40000   13.489075
Clothes 40000   6.998825
Clothes 40000   15.1066
Clothes 40000   1.28335
Clothes 40000   5.530025
Clothes 40000   4.303575
Clothes 40000   0.26805
Clothes 40000   7.055725
Clothes 40000   7.86725
Clothes 40000   4.142
Clothes 40000   2.0508
Clothes 40000   9.785875
Clothes 40000   18.231475
Clothes 40000   13.287125
Clothes 40000   2.19585
Clothes 40000   12.420525
Clothes 40000   11.47315
Clothes 40000   13.6323
Clothes 40000   18.356325
Clothes 40000   8.5539
Clothes 40000   11.59065
Clothes 40000   11.86605
Clothes 40000   12.532525
Clothes 40000   4.834175
Clothes 40000   12.341575
Clothes 40000   3.962075
Clothes 40000   10.7718
Clothes 40000   15.42055
Clothes 40000   10.2816
Clothes 40000   5.968825
Clothes 40000   1.06855
Clothes 40000   13.40095
Clothes 40000   6.19465
Clothes 40000   13.582725
```

GitHub Access Link: https://github.com/yashjaiswal1/bd_lab_assignment

References:

- https://youtu.be/K0aDh_sfVrc
- https://youtu.be/U3fkWvaqgl8
- https://youtu.be/SAX8b3AN3Uc