Assignment 10: Hadoop

Bhargavanarasimhan Lakshminarasimhan: G01093525

Please answer the following questions

1) ***Create a Hive table based on columns present in the csv file. Include the create table command and show the table schema using the describe <tablename> command [10]***

```
Time taken: 0.861 seconds
hive> describe salesrecordstable;
OK
region                  string
country                 string
itemtype                string
saleschannel            string
orderpriority           string
orderdate               date
orderid                 int
shipdate                date
unitssold               int
unitprice               float
unitcost                float
totalrevenue            float
totalcost               float
totalprofit             float
Time taken: 0.527 seconds, Fetched: 14 row(s)
```

2) ***Visit the NameNode web UI interface and report the number of blocks for the dataset file. Also include all the block information [10 points]***

```
hduser@virtualmachine:~$ hdfs fsck /user/hive/warehouse/salesrecords -files -blocks -locations
Connecting to namenode via http://localhost:9870/fsck?ugi=hduser&files=1&blocks=1&locations=1&path=%2Fuser%2Fhive%2Fwarehouse%2Fsalesrecords
FSCK started by hduser (auth:SIMPLE) from /127.0.0.1 for path /user/hive/warehouse/salesrecords at Fri Apr 19 19:16:39 EDT 2019
/user/hive/warehouse/salesrecords <dir>
/user/hive/warehouse/salesrecords/salesrecords.csv 257687552 bytes, replicated: replication=1, 2 block(s):  OK
0. BP-1648874202-127.0.1.1-1538344637229:blk_1073741826_1002 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-3fc667d0-a13a-4e03-b039-d8e0a203f11f,DISK]]
1. BP-1648874202-127.0.1.1-1538344637229:blk_1073741827_1003 len=123469824 Live_repl=1  [DatanodeInfoWithStorage[127.0.0.1:9866,DS-3fc667d0-a13a-4e03-b039-d8e0a203f11f,DISK]]


Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    1
 Total symlinks:                0

Replicated Blocks:
 Total size:    257687552 B
 Total files:   1
 Total blocks (validated):      2 (avg. block size 128843776 B)
 Minimally replicated blocks:   2 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
FSCK ended at Fri Apr 19 19:16:39 EDT 2019 in 14 milliseconds


The filesystem under path '/user/hive/warehouse/salesrecords' is HEALTHY
```

3) Write a Hive SQL query and show output of the following statements

a) Sort the Region based on their count in Descending order [10 points]

select region, count(*) from salesrecordstable group by region order by count(*) desc;

```
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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>
Job running in-process (local Hadoop)
2019-04-19 22:34:32,059 Stage-1 map = 0%,  reduce = 0%
2019-04-19 22:34:40,215 Stage-1 map = 100%,  reduce = 0%
2019-04-19 22:34:41,239 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local588691577_0003
Launching Job 2 out of 2
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>
Job running in-process (local Hadoop)
2019-04-19 22:34:43,744 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1539616268_0004
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 1546149888 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 1030766592 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Sub-Saharan Africa      536654
Europe  535440
Asia    301318
Middle East and North Africa    256772
Central America and the Caribbean       223070
Australia and Oceania   166824
North America   44904
Time taken: 27.253 seconds, Fetched: 7 row(s)
```

b) Count the number of Item Type and Sales Channel. Sort the output in Descending order with respect to count [10 points]

select itemtype,saleschannel,count(itemtype),count(saleschannel)
 from salesrecordstable
group by itemtype, saleschannel
order by count(itemtype),count(saleschannel) desc;

```
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>
Job running in-process (local Hadoop)
2019-04-19 22:37:36,638 Stage-2 map = 100%,  reduce = 0%
2019-04-19 22:37:37,644 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local429796468_0006
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 2319224832 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 1546149888 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Cereal  Online  85674   85674
Vegetables      Online  85716   85716
Household       Offline 85812   85812
Office Supplies Online  85822   85822
Clothes Offline 85850   85850
Meat    Offline 85878   85878
Beverages       Online  85946   85946
Clothes Online  86000   86000
Office Supplies Offline 86000   86000
Vegetables      Offline 86036   86036
Fruits  Offline 86040   86040
Baby Food       Offline 86040   86040
Household       Online  86058   86058
Snacks  Online  86068   86068
Meat    Online  86072   86072
Cosmetics       Offline 86078   86078
Personal Care   Online  86094   86094
Beverages       Offline 86150   86150
Cosmetics       Online  86162   86162
Cereal  Offline 86168   86168
Baby Food       Online  86276   86276
Personal Care   Offline 86340   86340
Fruits  Online  86344   86344
Snacks  Offline 86358   86358
Time taken: 9.707 seconds, Fetched: 24 row(s)
```

c) Find the Item Type with count where Total Cost is less than Total Profit [10 points]

```
hive> select itemtype,count(itemtype) from salesrecordstable where totalcost<totalprofit group by itemtype;
Query ID = hduser_20190419224050_7aeae25b-f5c3-4946-846a-0bcef89ed85a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 2
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>
Job running in-process (local Hadoop)
2019-04-19 22:40:52,542 Stage-1 map = 0%,  reduce = 0%
2019-04-19 22:40:56,565 Stage-1 map = 100%,  reduce = 50%
2019-04-19 22:40:57,599 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local764338268_0007
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 3092299776 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Clothes 171850
Time taken: 7.441 seconds, Fetched: 1 row(s)
```