

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

Assignment #: 2

Done by: Esraa Khaled Ahmed Fouad Omar # ID: 20399123

January 23, 2023

Part 2

Copy sales_data_samples.csv from local virtual machine to HDFS. As you can see, we copied the file successfully.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
[osboxes@quickstart-bigdata ~]$ hdfs dfs -put /home/osboxes/Documents/sales_data_sample.csv /user/osboxes/inputdata/sales_data_sample.csv
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user/osboxes/inputdata/
Found 2 items
-rw-r--r-- 3 osboxes osboxes 526283 2023-01-26 15:36 /user/osboxes/inputdata/sales_data_sample.csv
-rw-r--r-- 3 osboxes osboxes 723331 2023-01-24 17:26 /user/osboxes/inputdata/whquestion.txt
[osboxes@quickstart-bigdata ~]$ #Name: Esraa Ahmed Fouad Omar # ID :20399123
```

Open the terminal and write hive in it to be in hive mode. As you can see, we got in the hive mode successfully.

```
[osboxes@quickstart-bigdata ~]$ # Name: Esraa Ahmed Fouad Omar # ID:20399123
[osboxes@quickstart-bigdata ~]$ hdfs dfs -put /home/osboxes/Downloads/sales_data_sample.csv /user/osboxes/inputdata/sales_data_sample.csv
[osboxes@quickstart-bigdata ~]$ hive
WARNING: Use "yarn jar" to launch YARN applications.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/jars/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/jars/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]

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/jars/hive-common-2.1.1-cdh6.3.2.jar!/hive-log4j2.properties Async: false

WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive>
```

Create Database and call it data. As you can see, we create database successfully.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
hive> CREATE DATABASE IF NOT EXISTS data ;
OK
Time taken: 1.047 seconds
hive> USE data ;
OK
Time taken: 0.13 seconds
hive> Name: Esraa Ahmed Fouad Omar # ID :20399123
```

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

Use database which we were created from previous step. As you can see, everything is ok.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
hive> USE data ;
OK
Time taken: 0.13 seconds
hive> Name: Esraa Ahmed Fouad Omar # ID :20399123
```

Create table and insert the name and data type of columns into it as shown below.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
hive> CREATE EXTERNAL TABLE IF NOT EXISTS sale_updated_lab2 (
> ORDERNUMBER INT,
> QUANTITYORDERED INT,
> PRICEEACH FLOAT,
> ORDERLINENUMBER INT,
> SALES FLOAT,
> ORDERDATE STRING,
> STATUS STRING,
> QTR_ID INT,
> MONTH_ID INT,
> YEAR_ID INT,
> PRODUCTLINE STRING,
> MSRP INT,
> PRODUCTCODE STRING,
> CUSTOMERNAME STRING,
> PHONE STRING,
> ADDRESSLINE1 STRING,
> ADDRESSLINE2 STRING,
> CITY STRING,
> STATE STRING,
> POSTALCODE INT,
> COUNTRY STRING,
> TERRITORY STRING,
> CONTACTLASTNAME STRING,
> CONTACTFIRSTNAME STRING,
> DEALSIZE STRING
> )
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;
OK
Time taken: 0.16 seconds
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123
```

Load data(sales_data_sample.csv) into the sale_updated_lab2 table as shown below.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
hive> LOAD DATA INPATH '/user/osboxes/inputdata/sales_data_sample.csv' INTO TABLE sale_updated_lab2;
Loading data to table data.sale_updated_lab2
OK
Time taken: 0.524 seconds
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123
```

Skip/remove the header from the table as shown below

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
Time taken: 0.079 seconds
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive> ALTER TABLE sale_updated_lab2 SET TBLPROPERTIES ("skip.header.line.count"="1");
OK
Time taken: 0.069 seconds
hive>
```

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

Let's see sample of our data.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
Time taken: 17.118 seconds, Fetched: 1 row(s)
hive> SELECT * FROM sale_updated LIMIT 10;
OK
10107 30 95.7 2 2871.0 2/24/2003 0:00 Shipped 1 2 2003 Motorcycles 95 S10_1678 Lan
d of Toys Inc. 2125557818 897 Long Airport Avenue NYC NY 10022 USA NA Yu Kwai Small
10121 34 81.35 5 2765.9 5/7/2003 0:00 Shipped 2 5 2003 Motorcycles 95 S10_1678 Rei
ms Collectables 26.47.1555 59 rue de l'Abbaye Reims 51100 France EMEA Henriot Paul Small
10134 41 94.74 2 3884.34 7/1/2003 0:00 Shipped 3 7 2003 Motorcycles 95 S10_1678 Lyo
n Souveniers +33 1 46 62 7555 27 rue du Colonel Pierre Avia Paris 75508 France EMEA Da CunhaDanie
l Medium
10145 45 83.26 6 3746.7 8/25/2003 0:00 Shipped 3 8 2003 Motorcycles 95 S10_1678 Toy
s4GrownUps.com 6265557265 78934 Hillside Dr. Pasadena CA 90003 USA NA Young Julie
Medium
10159 49 100.0 14 5205.27 10/10/2003 0:00 Shipped 4 10 2003 Motorcycles 95 S10_1678 Cor
porate Gift Ideas Co. 6505551386 7734 Strong St. San Francisco CA NULL USA Brown Julie Mediu
m
10168 36 96.66 1 3479.76 10/28/2003 0:00 Shipped 4 10 2003 Motorcycles 95 S10_1678 Tec
hnics Stores Inc. 6505556809 9408 Furth Circle Burlingame CA 94217 USA NA Hirano Juri
Medium
10180 29 86.13 9 2497.77 11/11/2003 0:00 Shipped 4 11 2003 Motorcycles 95 S10_1678 Dae
dalus Designs Imports 20.16.1555 184 chausse de Tournai Lille 59000 France EMEA Rance Martine Small
10188 48 100.0 1 5512.32 11/18/2003 0:00 Shipped 4 11 2003 Motorcycles 95 S10_1678 Her
kku Gifts +47 2267 3215 Drammen 121 PR 744 Sentrum Bergen NULL Norway EMEA Oeztan Veysel Mediu
m
10201 22 98.57 2 2168.54 12/1/2003 0:00 Shipped 4 12 2003 Motorcycles 95 S10_1678 Min
i Wheels Co. 6505555787 5557 North Pendale Street San Francisco CA NULL USA NA Murphy Julie
Small
10211 41 100.0 14 4708.44 1/15/2004 0:00 Shipped 1 1 2004 Motorcycles 95 S10_1678 Aut
o Canal Petit (1) 47.55.6555 25 rue Lauriston Paris 75016 France EMEA Perrier Dominique Mediu
m
Time taken: 0.071 seconds, Fetched: 10 row(s)
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive>
```

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

1. Number of orders per country

```
Name: Esraa Ahmed Fouad Omar # ID :20399123

hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive> SELECT sale_updated_lab2.COUNTRY , COUNT(sale_updated_lab2.ORDERLINENUMBER) count FROM sale_updated_lab2
> GROUP BY sale_updated_lab2.COUNTRY;
Query ID = osboxes_20230126160226_c6931d9c-d852-4900-b862-4aebfd9dc43e
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>
23/01/26 16:02:26 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
23/01/26 16:02:26 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
Starting Job = job_1674722277352_0022, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1674722277352_0022/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job -kill job_1674722277352_0022
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-26 16:02:35,094 Stage-1 map = 0%, reduce = 0%
2023-01-26 16:02:39,193 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.42 sec
2023-01-26 16:02:44,313 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.19 sec
MapReduce Total cumulative CPU time: 3 seconds 190 msec
Ended Job = job_1674722277352_0022
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.19 sec HDFS Read: 538949 HDFS Write: 524 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 190 msec
OK
Australia      185
Austria        55
Belgium        33
Canada         70
Denmark        63
Finland        92
France         314
Germany        62
Ireland        16
Italy          113
Japan          52
Norway         85
Philippines    26
Singapore      79
Spain          342
Sweden         57
Switzerland    31
UK             144
USA            1004
Time taken: 19.374 seconds, Fetched: 19 row(s)
hive> |
```

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

2. Total amount for the orders that are shipped (status=shipped) per country

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
hive> Select sale_updated lab2.COUNTRY, SUM(sale_updated lab2.QUANTITYORDERED) from sale_updated_lab2 Where sale_updated_lab2.STATUS = 'Shipped' group by sale_u
updated lab2.COUNTRY;
Query ID = osboxes_20230126155919_96ca6451-8fd4-4ca4-83f8-4d087808b841
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>
23/01/26 15:59:19 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
23/01/26 15:59:19 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
Starting Job = job_1674722277352_0021, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1674722277352_0021/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job -kill job_1674722277352_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-26 15:59:26,039 Stage-1 map = 0%, reduce = 0%
2023-01-26 15:59:31,139 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.84 sec
2023-01-26 15:59:35,213 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.73 sec
MapReduce Total cumulative CPU time: 4 seconds 730 msec
Ended Job = job_1674722277352_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.73 sec HDFS Read: 539883 HDFS Write: 555 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 730 msec
OK
Australia 5550
Austria 1686
Belgium 963
Canada 2293
Denmark 1770
Finland 3192
France 10663
Germany 2148
Ireland 490
Italy 3773
Japan 1842
Norway 2842
Philippines 961
Singapore 2760
Spain 10646
Sweden 1239
Switzerland 1078
UK 4584
USA 32923
Time taken: 17.886 seconds, Fetched: 19 row(s)
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
```

3. Total how many Small, Medium and Large deal have been made.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive> Select sale_updated lab2.DEALSIZE, COUNT(sale_updated lab2.DEALSIZE) from sale_updated_lab2 group by sale_updated_lab2.DEALSIZE ;
Query ID = osboxes_20230126160428_8a95a33d-9deb-48e8-96d2-7388de00f578
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>
23/01/26 16:04:28 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
23/01/26 16:04:28 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
Starting Job = job_1674722277352_0023, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1674722277352_0023/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job -kill job_1674722277352_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-01-26 16:04:36,200 Stage-1 map = 0%, reduce = 0%
2023-01-26 16:04:40,276 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.44 sec
2023-01-26 16:04:45,362 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.19 sec
MapReduce Total cumulative CPU time: 3 seconds 190 msec
Ended Job = job_1674722277352_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.19 sec HDFS Read: 538946 HDFS Write: 156 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 190 msec
OK
Large 157
Medium 1384
Small 1282
Time taken: 19.039 seconds, Fetched: 3 row(s)
hive>
```


4. create separate partition for order STATUS

- create table with the same columns name and data type and partition it using status column as shown below.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive> CREATE TABLE IF NOT EXISTS partitioned_sale_updated_lab2
  > (
  > ORDERNUMBER INT,
  > QUANTITYORDERED INT,
  > PRICEEACH FLOAT ,
  > ORDERLINENUMBER INT,
  > SALES FLOAT ,
  > ORDERDATE STRING ,
  > QTR ID INT ,
  > MONTH ID INT ,
  > YEAR ID INT ,
  > PRODUCTLINE STRING ,
  > MSRP INT ,
  > PRODUCTCODE STRING ,
  > CUSTOMERNAME STRING ,
  > PHONE STRING ,
  > ADDRESSLINE1 STRING ,
  > ADDRESSLINE2 STRING ,
  > CITY STRING ,
  > STATE STRING ,
  > POSTALCODE INT ,
  > COUNTRY STRING ,
  > TERRITORY STRING ,
  > CONTACTLASTNAME STRING ,
  > CONTACTFIRSTNAME STRING ,
  > DEALSIZE STRING )
  > partitioned by (STATUS STRING);
OK
Time taken: 0.096 seconds
hive>
```

- Change your configuration from strict mode to non-strict mode. As you can see, we changed it successfully

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive>
```

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

- Overwrite into the previous created table as shown below.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive> insert overwrite table partitioned_sale_updated_lab2 partition(status) select * from sale_updated_lab2;
Query ID = osboxes_20230126160941_9105c403-2786-4e4a-b0d3-f037d6d556f8
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
23/01/26 16:09:42 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
23/01/26 16:09:42 INFO client.RMProxy: Connecting to ResourceManager at quickstart-bigdata/192.168.114.134:8032
Starting Job = job_1674722277352_0024, Tracking URL = http://quickstart-bigdata:8088/proxy/application_1674722277352_0024/
Kill Command = /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/lib/hadoop/bin/hadoop job -kill job_1674722277352_0024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-01-26 16:09:49,577 Stage-1 map = 0%, reduce = 0%
2023-01-26 16:09:54,673 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.04 sec
MapReduce Total cumulative CPU time: 3 seconds 40 msec
Ended Job = job_1674722277352_0024
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://quickstart-bigdata:8020/user/hive/warehouse/data.db/partitioned_sale_updated_lab2/.hive-staging_hive_2023-01-26_16-09-41_498_6044276227058385254-1/-ext-10000
Loading data to table data.partitioned_sale_updated_lab2 partition (status=null)

Time taken to load dynamic partitions: 0.171 seconds
Time taken for adding to write entity : 0.001 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 3.04 sec HDFS Read: 535794 HDFS Write: 468114 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 40 msec
OK
Time taken: 14.805 seconds
hive>
```

- Describe your table.

```
Name: Esraa Ahmed Fouad Omar # ID :20399123
Time taken: 14.805 seconds
hive> DESCRIBE partitioned_sale_updated_lab2;
OK
ordernumber          int
quantityordered      int
pricceach             float
orderlinenumber      int
sales                float
orderdate            string
qtr_id               int
month_id             int
year_id              int
productline          string
msrp                 int
productcode          string
customername         string
phone                string
addressline1         string
addressline2         string
city                 string
state                string
postalcode            int
country              string
territory             string
contactlastname       string
contactfirstname     string
dealsize              string
status               string

# Partition Information
# col_name            data_type            comment
status                string
Time taken: 0.099 seconds, Fetched: 30 row(s)
hive> # Name: Esraa Ahmed Fouad Omar # ID :20399123;
hive>
```