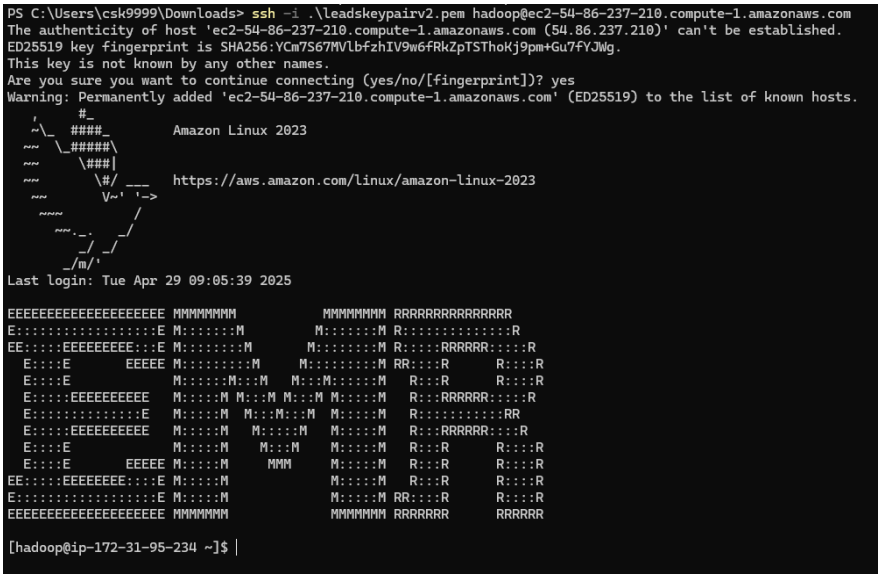
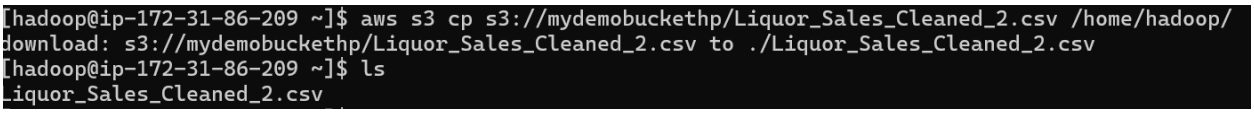
 # Connect to your EMR cluster using SSH and a key pair

ssh -i .\leadskeypairv2.pem hadoop@ec2-54-86-237-210.compute-1.amazonaws.com



 # Copy the CSV file from S3 to your EMR local filesystem

aws s3 cp s3://mydemobuckethp/Liquor_Sales_Cleaned_2.csv /home/hadoop/



 # Download and extract MySQL JDBC connector to enable Sqoop to connect to RDS

wget https://cdn.mysql.com/Downloads/Connector-J/mysql-connector-j-9.2.0.tar.gz

gunzip mysql-connector-j-9.2.0.tar.gz

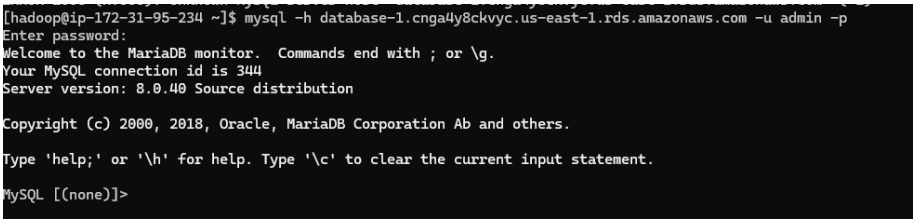
tar -xvf mysql-connector-j-9.2.0.tar

 # Copy JDBC JAR into Sqoop's lib directory so Sqoop can use it

sudo cp mysql-connector-j-9.2.0/mysql-connector-j-9.2.0.jar /usr/lib/sqoop/lib

 # Connect to the AWS RDS MySQL instance

mysql -h database-1.cnga4y8ckvyc.us-east-1.rds.amazonaws.com -u admin -p



 # Create the Database and liquor_sales table schema in RDS

Create database liquoreSales;

Use liquoreSales;

```
MySQL [(none)]> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
4 rows in set (0.018 sec)

MySQL [(none)]> create database liquoreSales;
Query OK, 1 row affected (0.020 sec)

MySQL [(none)]> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| liquoreSales       |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
5 rows in set (0.001 sec)
```

```
CREATE TABLE liquor_sales (
 invoice_item_number VARCHAR(50),
 sale_date DATE,
 store_number INT,
 store_name VARCHAR(100),
 address VARCHAR(200),
 city VARCHAR(100),
 zip_code VARCHAR(20),
 store_location VARCHAR(50),
 county_number INT,
 county VARCHAR(100),
 category INT,
 category_name VARCHAR(100),
 vendor_number INT,
 vendor_name VARCHAR(100),
 item_number INT,
 item_description VARCHAR(200),
 pack INT,
 bottle_volume_ml INT,
 state_bottle_cost FLOAT,
 state_bottle_retail FLOAT,
```

```
 bottles_sold INT,
 sale_dollars FLOAT,
 volume_sold_liters FLOAT,
 volume_sold_gallons FLOAT
);
```

```
MySQL [(none)]> use liquorSales;
ERROR 1049 (42000): Unknown database 'liquorSales'
MySQL [(none)]> use liquoreSales;
Database changed
MySQL [liquoreSales]> CREATE TABLE liquor_sales (
    ->     invoice_item_number     VARCHAR(50),       -- Can be string or integer
    ->     sale_date               DATE,
    ->     store_number            INT,
    ->     store_name              VARCHAR(100),
    ->     address                 VARCHAR(200),
    ->     city                    VARCHAR(100),
    ->     zip_code                VARCHAR(20),       -- Supports both integer and string ZIPs (like ZIP+4)
    ->     store_location          VARCHAR(50),       -- Stored as a string, e.g., "41.5868,-93.6250"
    ->     county_number           INT,
    ->     county                  VARCHAR(100),
    ->     category                INT,
    ->     category_name           VARCHAR(100),
    ->     vendor_number           INT,
    ->     vendor_name             VARCHAR(100),
    ->     item_number             INT,
    ->     item_description        VARCHAR(200),
    ->     pack                    INT,
    ->     bottle_volume_ml        INT,               -- Stored as INT; can change to FLOAT if needed
    ->     state_bottle_cost       FLOAT,
    ->     state_bottle_retail     FLOAT,
    ->     bottles_sold            INT,
    ->     sale_dollars            FLOAT,
    ->     volume_sold_liters      FLOAT,
    ->     volume_sold_gallons     FLOAT
    -> );
Query OK, 0 rows affected (0.063 sec)

MySQL [liquoreSales]> exit
Bye
[hadoop@ip-172-31-95-234 ~]$ aws s3 cp s3://mydemobuckethp/Liquor_Sales_Cleaned_2.csv /home/hadoop/
download: s3://mydemobuckethp/Liquor_Sales_Cleaned_2.csv to ./Liquor_Sales_Cleaned_2.csv
[hadoop@ip-172-31-95-234 ~]$ ls
Liquor_Sales_Cleaned_2.csv  mysql-connector-j-9.2.0  mysql-connector-j-9.2.0.tar  wget-log  wget-log.1
[hadoop@ip-172-31-95-234 ~]$
```

 # Load data from local file into the MySQL liquor_sales table

```
LOAD DATA LOCAL INFILE '/home/hadoop/Liquor_Sales_Cleaned_2.csv'

INTO TABLE liquor_sales

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES

(invoice_item_number, sale_date, store_number, store_name, address, city, zip_code,
store_location, county_number, county, category, category_name, vendor_number, vendor_name,
item_number, item_description, pack, bottle_volume_ml, state_bottle_cost, state_bottle_retail,
bottles_sold, sale_dollars, volume_sold_liters, volume_sold_gallons);
```

```
MySQL [(none)]> use liquoreSales
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [liquoreSales]> LOAD DATA LOCAL INFILE '/home/hadoop/Liquor_Sales.csv' INTO TABLE liquor_sales FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES (invoice_item_number,sale_date,store_number
,store_name,address,city,zip_code,store_location,county_number,county,category,category_name,vendor_number,vendor_name,item_number,item_description,pack,bottle_volume_ml,state_bottle_cost,state_bottle_retail,b
ottles_sold,sale_dollars,volume_sold_liters,volume_sold_gallons);
ERROR 2 (HY000): File '/home/hadoop/Liquor_Sales.csv' not found (Errcode: 2)
MySQL [liquoreSales]> LOAD DATA LOCAL INFILE '/home/hadoop/Liquor_Sales_Vleaned_2.csv' INTO TABLE liquor_sales FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES (invoice_item_number,sale_date,st
ore_number,store_name,address,city,zip_code,store_location,county_number,county,category,category_name,vendor_number,vendor_name,item_number,item_description,pack,bottle_volume_ml,state_bottle_cost,state_bottl
e_retail,bottles_sold,sale_dollars,volume_sold_liters,volume_sold_gallons);
ERROR 2 (HY000): File '/home/hadoop/Liquor_Sales_Vleaned_2.csv' not found (Errcode: 2)
MySQL [liquoreSales]> LOAD DATA LOCAL INFILE '/home/hadoop/Liquor_Sales_Cleaned_2.csv' INTO TABLE liquor_sales FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES (invoice_item_number,sale_date,st
ore_number,store_name,address,city,zip_code,store_location,county_number,county,category,category_name,vendor_number,vendor_name,item_number,item_description,pack,bottle_volume_ml,state_bottle_cost,state_bottl
e_retail,bottles_sold,sale_dollars,volume_sold_liters,volume_sold_gallons);
Query OK, 17687883 rows affected, 65535 warnings (6 min 52.079 sec)
Records: 17687883  Deleted: 0  Skipped: 0  Warnings: 14473060

MySQL [liquoreSales]> select * from liquor_sales limit 5;
+---------------------+------------+--------------+-----------------------------------+------------------------+-------------+-----------+--------------------------------+---------------+--------+-----
-----+--------------------------------+---------------+---------------+-------------+-------------------------------+------+------------------+-------------------+---------------------+--------------+-
-------------+-------------------+---------------------+
| invoice_item_number | sale_date  | store_number | store_name                        | address                | city        | zip_code  | store_location                 | county_number | county | cate
gory | category_name                  | vendor_number | vendor_name   | item_number | item_description              | pack | bottle_volume_ml | state_bottle_cost | state_bottle_retail | bottles_sold |
sale_dollars | volume_sold_liters | volume_sold_gallons |
+---------------------+------------+--------------+-----------------------------------+------------------------+-------------+-----------+--------------------------------+---------------+--------+-----
-----+--------------------------------+---------------+---------------+-------------+-------------------------------+------+------------------+-------------------+---------------------+--------------+-
-------------+-------------------+---------------------+
| S24127700024        | 2015-02-19 |         3678 | "Smoke Shop                       |  The"                  | 1918 SE 14TH ST | DES MOINES | 50320                       |             0 | 77     |
   0 | 1031200                        |             0 | 380           |           0 | 41783                         |    0 |                6 |               500 |                4.89 |            7 |
           2 |              14.68 |                   1 |
| S15066200002        | 2013-10-10 |         2633 | Hy-Vee #3 / BDI / Des Moines      | 3221 SE 14TH ST        | DES MOINES  | 50320     | POINT (-93.596754 41.554101) |            77 | Polk   |  108
2900 | MISC. IMPORTED CORDIALS & LIQUEURS |         305 | MHW Ltd       |      904969 | Sabe Premiom Sake Double Barrel |    6 |              750 |             14.99 |               22.49 |            6 |
      134.94 |                4.5 |                1.19 |
| S19323500030        | 2014-06-03 |         2607 | Hy-Vee Wine and Spirits / Shenandoah | 520 SO FREMONT      | SHENANDOAH  | 51601     | POINT (-95.385111 40.761736) |            73 | Page   |  106
2200 | PUERTO RICO & VIRGIN ISLANDS RUM   |           434 | Luxco-St Louis |      45277 | Paramount White Rum           |   12 |             1000 |              4.34 |                6.51 |           12 |
       78.12 |                 12 |                3.17 |
| S23334500013        | 2015-01-06 |         4810 | Kum & Go #518 / Ankeny            | 3603 NE OTTERVIEW CIRCLE | ANKENY    | 50021     | POINT (-93.572458 41.760989) |            77 | Polk   |  106
2200 | PUERTO RICO & VIRGIN ISLANDS RUM   |            35 | "Bacardi U.S.A. |           0 | 43121                       |    0 |               12 |               500 |                5.54 |            8 |
           1 |               8.31 |                 0.5 |
| S09742200010        | 2012-12-27 |         4025 | Karam Kaur Khasriya Llc           | 702 13TH ST            | BELLE PLAINE | 52208    | POINT (-92.277759 41.897052) |             6 | Benton |  101
2100 | CANADIAN WHISKIES                  |           260 | Diageo Americas |       11298 | Crown Royal Canadian Whisky |    6 |             1750 |               8.2 |               12.36 |            2 |
       92.98 |                3.5 |                0.92 |
+---------------------+------------+--------------+-----------------------------------+------------------------+-------------+-----------+--------------------------------+---------------+--------+-----
-----+--------------------------------+---------------+---------------+-------------+-------------------------------+------+------------------+-------------------+---------------------+--------------+-
-------------+-------------------+---------------------+
5 rows in set (0.015 sec)

MySQL [liquoreSales]>
```

Sqoop import from MySQL to HDFS (one mapper)

sqoop import \

--connect jdbc:mysql://database-1.cnga4y8ckvyc.us-east-1.rds.amazonaws.com/liquoreSales \

--username admin \

--password admin123 \

--table liquor_sales \

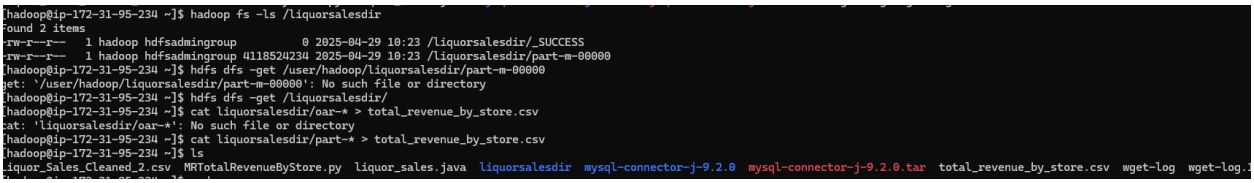
--target-dir /liquorsalesdir \

--m 1


```
MySQL [(none)]> use liquoreSales
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [liquoreSales]> LOAD DATA LOCAL INFILE '/home/hadoop/Liquor_Sales.csv' INTO TABLE liquor_sales FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES (invoice_item_number,sale_date,store_number
,store_name,address,city,zip_code,store_location,county_number,county,category,category_name,vendor_number,vendor_name,item_number,item_description,pack,bottle_volume_ml,state_bottle_cost,state_bottle_retail,b
ottles_sold,sale_dollars,volume_sold_liters,volume_sold_gallons);
ERROR 2 (HY000): File '/home/hadoop/Liquor_Sales.csv' not found (Errcode: 2)
MySQL [liquoreSales]> LOAD DATA LOCAL INFILE '/home/hadoop/Liquor_Sales_Vleaned_2.csv' INTO TABLE liquor_sales FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES (invoice_item_number,sale_date,st
ore_number,store_name,address,city,zip_code,store_location,county_number,county,category,category_name,vendor_number,vendor_name,item_number,item_description,pack,bottle_volume_ml,state_bottle_cost,state_bottl
e_retail,bottles_sold,sale_dollars,volume_sold_liters,volume_sold_gallons);
ERROR 2 (HY000): File '/home/hadoop/Liquor_Sales_Vleaned_2.csv' not found (Errcode: 2)
MySQL [liquoreSales]> LOAD DATA LOCAL INFILE '/home/hadoop/Liquor_Sales_Cleaned_2.csv' INTO TABLE liquor_sales FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES (invoice_item_number,sale_date,st
ore_number,store_name,address,city,zip_code,store_location,county_number,county,category,category_name,vendor_number,vendor_name,item_number,item_description,pack,bottle_volume_ml,state_bottle_cost,state_bottl
e_retail,bottles_sold,sale_dollars,volume_sold_liters,volume_sold_gallons);
Query OK, 17687883 rows affected, 65535 warnings (6 min 52.079 sec)
Records: 17687883  Deleted: 0  Skipped: 0  Warnings: 14473060

MySQL [liquoreSales]> select * from liquor_sales limit 5;
+---------------------+------------+--------------+-----------------------------------+------------------------+-------------+-----------+--------------------------------+---------------+--------+-----
-----+--------------------------------+---------------+---------------+-------------+-------------------------------+------+------------------+-------------------+---------------------+--------------+-
-------------+-------------------+---------------------+
| invoice_item_number | sale_date  | store_number | store_name                        | address                | city        | zip_code  | store_location                 | county_number | county | cate
gory | category_name                  | vendor_number | vendor_name   | item_number | item_description              | pack | bottle_volume_ml | state_bottle_cost | state_bottle_retail | bottles_sold |
sale_dollars | volume_sold_liters | volume_sold_gallons |
+---------------------+------------+--------------+-----------------------------------+------------------------+-------------+-----------+--------------------------------+---------------+--------+-----
-----+--------------------------------+---------------+---------------+-------------+-------------------------------+------+------------------+-------------------+---------------------+--------------+-
-------------+-------------------+---------------------+
| S24127700024        | 2015-02-19 |         3678 | "Smoke Shop                       |  The"                  | 1918 SE 14TH ST | DES MOINES | 50320                       |             0 | 77     |
   0 | 1031200                        |             0 | 380           |           0 | 41783                         |    0 |                6 |               500 |                4.89 |            7 |
           2 |              14.68 |                   1 |
| S15066200002        | 2013-10-10 |         2633 | Hy-Vee #3 / BDI / Des Moines      | 3221 SE 14TH ST        | DES MOINES  | 50320     | POINT (-93.596754 41.554101) |            77 | Polk   |  108
2900 | MISC. IMPORTED CORDIALS & LIQUEURS |         305 | MHW Ltd       |      904969 | Sabe Premiom Sake Double Barrel |    6 |              750 |             14.99 |               22.49 |            6 |
      134.94 |                4.5 |                1.19 |
| S19323500030        | 2014-06-03 |         2607 | Hy-Vee Wine and Spirits / Shenandoah | 520 SO FREMONT      | SHENANDOAH  | 51601     | POINT (-95.385111 40.761736) |            73 | Page   |  106
2200 | PUERTO RICO & VIRGIN ISLANDS RUM   |           434 | Luxco-St Louis |      45277 | Paramount White Rum           |   12 |             1000 |              4.34 |                6.51 |           12 |
       78.12 |                 12 |                3.17 |
| S23334500013        | 2015-01-06 |         4810 | Kum & Go #518 / Ankeny            | 3603 NE OTTERVIEW CIRCLE | ANKENY    | 50021     | POINT (-93.572458 41.760989) |            77 | Polk   |  106
2200 | PUERTO RICO & VIRGIN ISLANDS RUM   |            35 | "Bacardi U.S.A. |           0 | 43121                       |    0 |               12 |               500 |                5.54 |            8 |
           1 |               8.31 |                 0.5 |
| S09742200010        | 2012-12-27 |         4025 | Karam Kaur Khasriya Llc           | 702 13TH ST            | BELLE PLAINE | 52208    | POINT (-92.277759 41.897052) |             6 | Benton |  101
2100 | CANADIAN WHISKIES                  |           260 | Diageo Americas |       11298 | Crown Royal Canadian Whisky |    6 |             1750 |               8.2 |               12.36 |            2 |
       92.98 |                3.5 |                0.92 |
+---------------------+------------+--------------+-----------------------------------+------------------------+-------------+-----------+--------------------------------+---------------+--------+-----
-----+--------------------------------+---------------+---------------+-------------+-------------------------------+------+------------------+-------------------+---------------------+--------------+-
-------------+-------------------+---------------------+
5 rows in set (0.015 sec)

MySQL [liquoreSales]>
```

```
[hadoop@ip-172-31-86-209 ~]$ sudo cp mysql-connector-j-9.2.0/mysql-connector-j-9.2.0.jar /usr/lib/sqoop/lib
[hadoop@ip-172-31-86-209 ~]$ sqoop import \
  --connect jdbc:mysql://database-1.cnga4y8ckvyc.us-east-1.rds.amazonaws.com/liquoreSales \
  --username admin \
  --password admin123 \
  --table liquor_sales \
  --target-dir /liquorsalesdir \
  --m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/log4j-slf4j-impl-2.17.2.jar!/org/slf4j/impl/StaticLogge
rBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
2025-04-29 15:58:21,795 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2025-04-29 15:58:21,837 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2025-04-29 15:58:21,984 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2025-04-29 15:58:21,985 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automati
cally registered via the SPI and manual loading of the driver class is generally unnecessary.
2025-04-29 15:58:22,776 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `liquor_sales` AS t LIMIT 1
2025-04-29 15:58:22,868 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `liquor_sales` AS t LIMIT 1
2025-04-29 15:58:22,891 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
2025-04-29 15:58:27,978 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/c82ec54a13ea551e2a366295dec5997f/liq
```

```
[hadoop@ip-172-31-95-234 ~]$ hadoop fs -ls /liquorsalesdir
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-29 10:23 /liquorsalesdir/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup 4118524234 2025-04-29 10:23 /liquorsalesdir/part-m-00000
[hadoop@ip-172-31-95-234 ~]$
```



 # create a table in HBase and verify it, you can use the HBase shell. Here's how:

 # Start HBase shell to create and inspect tables

hbase shell

 # Create HBase table with column family 'info'

create 'liquor_sales_hbase', 'info'

 # List all HBase tables

list

 # Show structure of the HBase table

describe 'liquor_sales_hbase'

```
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.5.5-amzn-0, r32843145930543cc64da58161007ab4b5dd34e9d, Wed Oct 23 05:06:50 AM UTC 2024
Took 0.0016 seconds
hbase:001:0> create 'liquor_sales_hbase', 'info'
Created table liquor_sales_hbase
Took 1.5024 seconds
=> Hbase::Table - liquor_sales_hbase
hbase:002:0> list
TABLE
liquor_sales_hbase
1 row(s)
Took 0.0386 seconds
=> ["liquor_sales_hbase"]
hbase:003:0> describe 'liquor_sales_hbase'
Table liquor_sales_hbase is ENABLED
liquor_sales_hbase, {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}
COLUMN FAMILIES DESCRIPTION
{NAME => 'info', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL =
> 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKC
ACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

1 row(s)
Quota is disabled
Took 0.1355 seconds
hbase:004:0>
```

 # Import data from MySQL to HBase using Sqoop with invoice_item_number as row key

```
sqoop import \
 --connect jdbc:mysql://database-1.cnga4y8ckvyc.us-east-1.rds.amazonaws.com/liquoreSales \
 --username admin \
 --password admin123 \
 --table liquor_sales \
 --hbase-table liquor_sales_hbase \
 --column-family info \
 --hbase-row-key invoice_item_number \
 --m 1
```

```
[hadoop@ip-172-31-91-178 ~]$ sqoop import \
  --connect jdbc:mysql://database-1.cnga4y8ckvyc.us-east-1.rds.amazonaws.com/liquoreSales \
  --username admin \
  --password admin123 \
  --table liquor_sales \
  --hbase-table liquor_sales_hbase \
  --column-family info \
  --hbase-row-key invoice_item_number \
  --m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hbase/lib/client-facing-thirdparty/log4j-slf4j-impl-2.17.2.jar!/org/slf4j/impl/StaticLogg
rBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
2025-04-30 04:39:25,318 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2025-04-30 04:39:25,359 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2025-04-30 04:39:25,488 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
```

 # Scan first 5 rows of the HBase table to verify records

```
scan 'liquor_sales_hbase', LIMIT => 5
```

 # Count total records in HBase table

count 'liquor_sales_hbase'

```
Took 0.5937 seconds
hbase:002:0> count 'liquor_sales_hbase'
Current count: 1000, row: S03447500019
Current count: 2000, row: S03538800004
Current count: 3000, row: S03625500072
Current count: 4000, row: S03724500044
Current count: 5000, row: S03823200005
Current count: 6000, row: S03919700013
Current count: 7000, row: S04004400036
Current count: 8000, row: S04094100024
Current count: 9000, row: S04186100095
Current count: 10000, row: S04274400082
Current count: 11000, row: S04364400094
Current count: 12000, row: S04455300008
Current count: 13000, row: S04538600105
Current count: 14000, row: S04628500013
Current count: 15000, row: S04726000011
Current count: 16000, row: S04812000022
Current count: 17000, row: S04896100042
```

 # View location in HDFS where HBase table data is stored (HBase-managed)

(e.g., /hbase/data/default/liquor_sales_hbase)

```
[hadoop@ip-172-31-91-178 ~]$ hdfs dfs -ls /user/hbase/data/default/
Found 1 items
drwxr-xr-x   - hbase hbase          0 2025-04-30 04:51 /user/hbase/data/default/liquor_sales_hbase
[hadoop@ip-172-31-91-178 ~]$ hdfs dfs -ls /user/hbase/data/default/liquor_sales_hbase/
Found 3 items
drwxr-xr-x   - hbase hbase          0 2025-04-30 04:32 /user/hbase/data/default/liquor_sales_hbase/.tabledesc
drwxr-xr-x   - hbase hbase          0 2025-04-30 04:46 /user/hbase/data/default/liquor_sales_hbase/6a2b4073906409fcb59f0c8f6bd7dac5
drwxr-xr-x   - hbase hbase          0 2025-04-30 04:46 /user/hbase/data/default/liquor_sales_hbase/6e779b0aa2fc97567c299cc50771a6fe
[hadoop@ip-172-31-91-178 ~]$
```

 # Install development tools and libraries required for happybase/thrift

sudo yum install python3-devel

```
[hadoop@ip-172-31-91-178 ~]$ sudo -i

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE:::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR:::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM            MMMMMMM RRRRRRR      RRRRRR

[root@ip-172-31-91-178 ~]# jps
4834 Main
9570 KMSWebServer
15650 RunJar
8579 NameNode
13126 Bootstrap
16552 Main
15112 RunJar
6569 AgentHttpServer
```

```
[root@ip-172-31-91-178 ~]# sudo yum install python3-devel
Last metadata expiration check: 1:07:55 ago on Wed Apr 30 04:24:26 2025.
Dependencies resolved.
================================================================================
 Package                Architecture          Version                    Repository                Size
================================================================================
Installing:
 python3-devel          x86_64                3.9.21-1.amzn2023.0.3      amazonlinux               206 k

Transaction Summary
================================================================================
Install  1 Package

Total download size: 206 k
Installed size: 764 k
Is this ok [y/N]: y
Downloading Packages:
python3-devel-3.9.21-1.amzn2023.0.3.x86_64.rpm                  4.7 MB/s | 206 kB     00:00
--------------------------------------------------------------------------------
Total                                                           2.8 MB/s | 206 kB     00:00
Running transaction check
Transaction check succeeded.
Running transaction test
Transaction test succeeded.
```

 # Install thriftpy2 (required for happybase)

pip install thriftpy2

```
[root@ip-172-31-91-178 ~]# pip install thriftpy2
Collecting thriftpy2
  Using cached thriftpy2-0.5.2.tar.gz (782 kB)
  Installing build dependencies ... done
  WARNING: Missing build requirements in pyproject.toml for thriftpy2 from https://files.pythonhosted.org/packages/f8/3a/d983b26df175
83a3cc865a9e1737bb8faacfa1e16e3ed17353ef48847e6b/thriftpy2-0.5.2.tar.gz#sha256=cefcb2f6f8b12c00054c6f942dd2323a53b48b8b6862312d03b677
dcf0d4a6da.
  WARNING: The project does not specify a build b                    o setuptools without 'wheel'.
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Collecting Cython>=3.0.10
  Using cached Cython-3.0.12-cp39-cp39-manylinux_                    hl (3.6 MB)
Requirement already satisfied: ply<4.0,>=3.4 in /usr/lib/python3.9/site-packages (from thriftpy2) (3.11)
Collecting six~=1.15
  Using cached six-1.17.0-py2.py3-none-any.whl (11 kB)
Building wheels for collected packages: thriftpy2
  Building wheel for thriftpy2 (pyproject.toml) ... done
  Created wheel for thriftpy2: filename=thriftpy2-0.5.2-cp39-cp39-linux_x86_64.whl size=1766116 sha256=6855bb47bd4ac65ecd020f9d8a835e
5edb3d7b01e661d8b3e5cec8c276915e6e
  Stored in directory: /root/.cache/pip/wheels/95/51/1d/d7303cb7c6b02c5793595b3138311bc81c4b5470ed7d306aa2
```

https://files.pythonhosted.org/packages/f8/3a/d983b26df17583a3cc865a9e1737bb8faacfa1e16e3ed17353ef48847e6b/thriftpy2-0.5.2.tar.gz#sha256=cefcb2f6f8b12c00054c6f942dd2323a53b48b8b6862312d03b677dcf0d4a6da

Ctrl+Click to follow link

 # Install happybase for Python HBase access

pip install happybase

```
[root@ip-172-31-91-178 ~]# pip install happybase
Collecting happybase
  Using cached happybase-1.2.0.tar.gz (40 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: six in /usr/local/lib/python3.9/site-packages (from happybase) (1.17.0)
Requirement already satisfied: thriftpy2>=0.4 in /usr/local/lib64/python3.9/site-packages (from happybase) (0.5.2)
Requirement already satisfied: Cython>=3.0.10 in /usr/local/lib64/python3.9/site-packages (from thriftpy2>=0.4->happybase) (3.0.12)
Requirement already satisfied: ply<4.0,>=3.4 in /usr/lib/python3.9/site-packages (from thriftpy2>=0.4->happybase) (3.11)
Using legacy 'setup.py install' for happybase, since package 'wheel' is not installed.
Installing collected packages: happybase
    Running setup.py install for happybase ... done
Successfully installed happybase-1.2.0
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. I
t is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
[root@ip-172-31-91-178 ~]# jps
```

 # Create a Python script that connects and lists data from HBase using happybase

(e.g., python my_hbase_reader.py)

1. List all the tables present in HBase.

```
#Listing table
import happybase

print("connecting to HBase")
con = happybase.Connection('localhost')


con.open()
print("Connected")

print("Listing tables...")
print(con.tables())

print("Closing the connection")
con.close()
```

 # Upload CSV to HDFS to run MRJob

hdfs dfs -put liquorsales.csv /user/hadoop/

 # List files in HDFS to confirm upload

hdfs dfs -ls /user/hadoop/

Liquor Sale Analysis Question and Answers:

 # Run MapReduce job using MRJob over HDFS CSV file

```
[hadoop@ip-172-31-95-234 ~]$ nano MRTopSellingLiquorCategories.py
```

python MRTotalRevenueByStore.py \

-r hadoop \

hdfs:///user/hadoop/liquorsales.csv \

--output-dir hdfs:///user/hadoop/output/total_revenue_by_store/

```
[hadoop@ip-172-31-86-209 ~]$ ls
Liquor_Sales_Cleaned_2.csv  hdfs:              liquorsales.csv  mysql-connector-j-9.2.0
MRTotalRevenueByStore.py    liquor_sales.java  liquorsalesdir   mysql-connector-j-9.2.0.tar
[hadoop@ip-172-31-86-209 ~]$ hdfs dfs -put liquorsales.csv /user/hadoop/
[hadoop@ip-172-31-86-209 ~]$ hdfs dfs -ls /user/hadoop/
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup 4118524234 2025-04-29 16:48 /user/hadoop/liquorsales.csv
[hadoop@ip-172-31-86-209 ~]$ python MRTotalRevenueByStore.py \
    -r hadoop \
    hdfs:///user/hadoop/liquorsales.csv \
    --output-dir hdfs:///user/hadoop/output/total_revenue_by_store/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.4.0
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/MRTotalRevenueByStore.hadoop.20250429.164900.795548
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/MRTotalRevenueByStore.hadoop.20250429.164900.795548/files/wd..
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/MRTotalRevenueByStore.hadoop.20250429.164900.795548/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.0-amzn-0.jar] /tmp/streamjob10937752844739905224.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-86-209.ec2.internal/172.31.86.209:8032
  Connecting to Application History server at ip-172-31-86-209.ec2.internal/172.31.86.209:10200
  Connecting to ResourceManager at ip-172-31-86-209.ec2.internal/172.31.86.209:8032
  Connecting to Application History server at ip-172-31-86-209.ec2.internal/172.31.86.209:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1745941213278_0002
```

```
  map 71% reduce 0%
  map 73% reduce 0%
  map 75% reduce 0%
  map 77% reduce 0%
  map 84% reduce 0%
  map 87% reduce 0%
  map 90% reduce 0%
  map 96% reduce 0%
  map 97% reduce 0%
  map 100% reduce 0%
  map 100% reduce 29%
  map 100% reduce 33%
  map 100% reduce 62%
  map 100% reduce 66%
  map 100% reduce 67%
  map 100% reduce 95%
  map 100% reduce 100%
 Job job_1745941213278_0002 completed successfully
 Output directory: hdfs:///user/hadoop/output/total_revenue_by_store/
ounters: 55
        File Input Format Counters
                Bytes Read=4120490314
        File Output Format Counters
                Bytes Written=36929
        File System Counters
                FILE: Number of bytes read=39518264
                FILE: Number of bytes written=91084702
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
```

 # Copy job output from HDFS to local filesystem on EMR node

hdfs dfs -get /user/hadoop/output/total_revenue_by_store .

```
Liquor_Sales_Cleaned_2.csv       MRTotalRevenueByStore.py  liquorsalesdir          mysql-connector-j-9.2.0.tar  wget-log
MRTopSellingLiquorCategories.py  liquor_sales.java         mysql-connector-j-9.2.0  total_revenue_by_store.csv   wget-log.1
```

 # Merge all HDFS part files into a single local CSV file

hdfs dfs -getmerge /user/hadoop/output/total_revenue_by_store/ total_revenue_by_store.csv

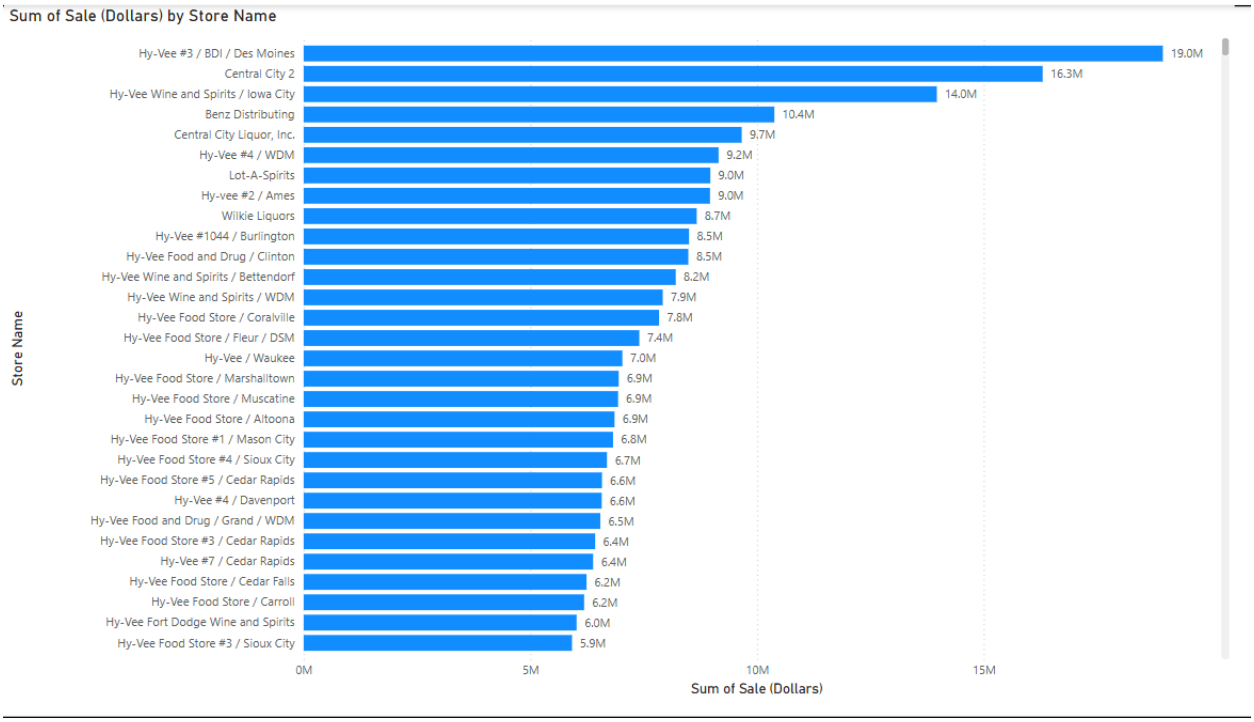
```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/output/total_revenue_by_store/
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MRTotalRevenueByStore.hadoop.20250429.164900.795548...
Removing temp directory /tmp/MRTotalRevenueByStore.hadoop.20250429.164900.795548...
[hadoop@ip-172-31-86-209 ~]$ hadoop fs -ls /user/hadoop/output/total_revenue_by_store/
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-29 16:55 /user/hadoop/output/total_revenue_by_store/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup      12158 2025-04-29 16:54 /user/hadoop/output/total_revenue_by_store/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup      12471 2025-04-29 16:54 /user/hadoop/output/total_revenue_by_store/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup      12300 2025-04-29 16:55 /user/hadoop/output/total_revenue_by_store/part-00002
[hadoop@ip-172-31-86-209 ~]$ hdfs dfs -get /user/hadoop/output/total_revenue_by_store .
[hadoop@ip-172-31-86-209 ~]$ hdfs dfs -getmerge /user/hadoop/output/total_revenue_by_store/ total_revenue_by_store.csv
[hadoop@ip-172-31-86-209 ~]$ ls
Liquor_Sales_Cleaned_2.csv  hdfs:              liquorsales.csv  mysql-connector-j-9.2.0      total_revenue_by_store
MRTotalRevenueByStore.py    liquor_sales.java  liquorsalesdir   mysql-connector-j-9.2.0.tar  total_revenue_by_store.csv
[hadoop@ip-172-31-86-209 ~]$ chmod 700 total_revenue_by_store.csv
[hadoop@ip-172-31-86-209 ~]$ chmod 700 total_revenue_by_store.csv
```

 # Download final result to your personal machine using scp or SFTP

scp hadoop@<your-node-public-ip>:~/total_revenue_by_store.csv .

After Downloading the total_revenue_by_store.csv file , I have created visualization in Power BI for better understanding of our Analysis.

1. Total Revenue by Store



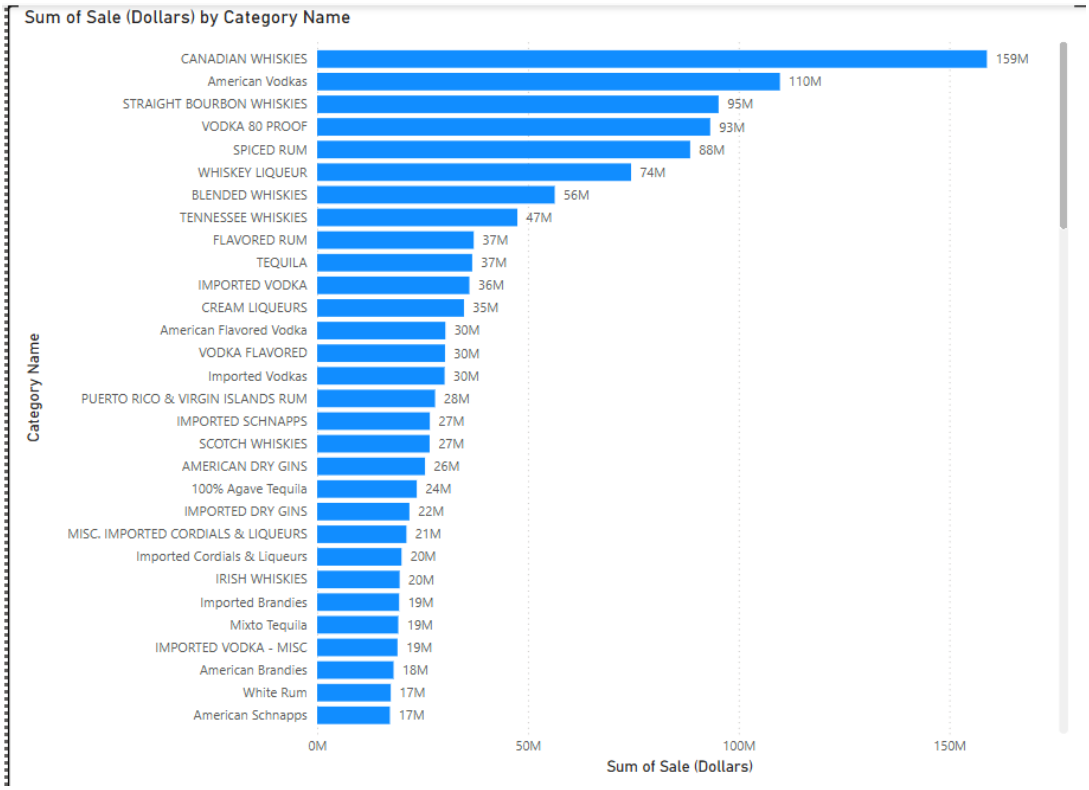
1. Top-Selling Categories:

```
[hadoop@ip-172-31-86-9 ~]$ python MRTopSellingLiquorCategories.py \
    -r hadoop \
    hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned_2.csv \
    --output-dir hdfs:///user/hadoop/output/top_selling_categories_total_bottles_sales/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.4.0
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/MRTopSellingLiquorCategories.hadoop.20250430.153156.607434
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/MRTopSellingLiquorCategories.hadoop.20250430.153156.607434/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/MRTopSellingLiquorCategories.hadoop.20250430.153156.607434/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.0-amzn-0.jar] /tmp/streamjob7974346819842575520.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-86-9.ec2.internal/172.31.86.9:8032
  Connecting to Application History server at ip-172-31-86-9.ec2.internal/172.31.86.9:10200
  Connecting to ResourceManager at ip-172-31-86-9.ec2.internal/172.31.86.9:8032
  Connecting to Application History server at ip-172-31-86-9.ec2.internal/172.31.86.9:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1746024827769_0001
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:33
  Submitting tokens for job: job_1746024827769_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1746024827769_0001
  The url to track the job: http://ip-172-31-86-9.ec2.internal:20888/proxy/application_1746024827769_0001/
  Running job: job_1746024827769_0001
```

```
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/output/top_selling_categories_total_bottles_sales/
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MRTopSellingLiquorCategories.hadoop.20250430.153156.607434...
Removing temp directory /tmp/MRTopSellingLiquorCategories.hadoop.20250430.153156.607434...
[hadoop@ip-172-31-86-9 ~]$ ls
Liquor_Sales_Cleaned_2.csv  MRTopSellingLiquorCategories.py  mysql-connector-j-9.2.0  mysql-connector-j-9.2.0.tar
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -ls /user/hadoop/output/top_selling_categories_bottles_sales/
ls: `/user/hadoop/output/top_selling_categories_bottles_sales/': No such file or directory
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -ls /user/hadoop/output/top_selling_categories_total_bottles_sales/
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-30 15:39 /user/hadoop/output/top_selling_categories_total_bottles_sales/_SUCC
ESS
-rw-r--r--   1 hadoop hdfsadmingroup       3067 2025-04-30 15:38 /user/hadoop/output/top_selling_categories_total_bottles_sales/part-
00000
-rw-r--r--   1 hadoop hdfsadmingroup       2992 2025-04-30 15:38 /user/hadoop/output/top_selling_categories_total_bottles_sales/part-
00001
-rw-r--r--   1 hadoop hdfsadmingroup       3312 2025-04-30 15:39 /user/hadoop/output/top_selling_categories_total_bottles_sales/part-
00002
[hadoop@ip-172-31-86-9 ~]$
```

```
Liquor_Sales_Cleaned_2.csv  MRTopSellingLiquorCategories.py  mysql-connector-j-9.2.0  mysql-connector-j-9.2.0.tar
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -ls /user/hadoop/output/top_selling_categories_total_bottles_sales/
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-30 15:39 /user/hadoop/output/top_selling_categories_total_bottles_sales/_SUCC
ESS
-rw-r--r--   1 hadoop hdfsadmingroup       3067 2025-04-30 15:38 /user/hadoop/output/top_selling_categories_total_bottles_sales/part-
00000
-rw-r--r--   1 hadoop hdfsadmingroup       2992 2025-04-30 15:38 /user/hadoop/output/top_selling_categories_total_bottles_sales/part-
00001
-rw-r--r--   1 hadoop hdfsadmingroup       3312 2025-04-30 15:39 /user/hadoop/output/top_selling_categories_total_bottles_sales/part-
00002
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -get /user/hadoop/output/top_selling_categories_total_bottles_sales/ .
[hadoop@ip-172-31-86-9 ~]$ cat top_selling_categories_bottles_sales/part-* > top_selling_categories_bottles_sales.csv
cat: 'top_selling_categories_bottles_sales/part-*': No such file or directory
[hadoop@ip-172-31-86-9 ~]$ cat top_selling_categories_total_bottles_sales//part-* > top_selling_categories_bottles_sales.csv
[hadoop@ip-172-31-86-9 ~]$ cat top_selling_categories_total_bottles_sales/part-* > top_selling_categories_bottles_sales.csv
[hadoop@ip-172-31-86-9 ~]$
```

This csv file is exported to local and imported into Power bi for Analysis



Top Recommendations Based on Sales Performance

- Canadian Whiskies lead significantly with $159M, followed by American Vodkas ($110M) and Straight Bourbon Whiskies ($95M)

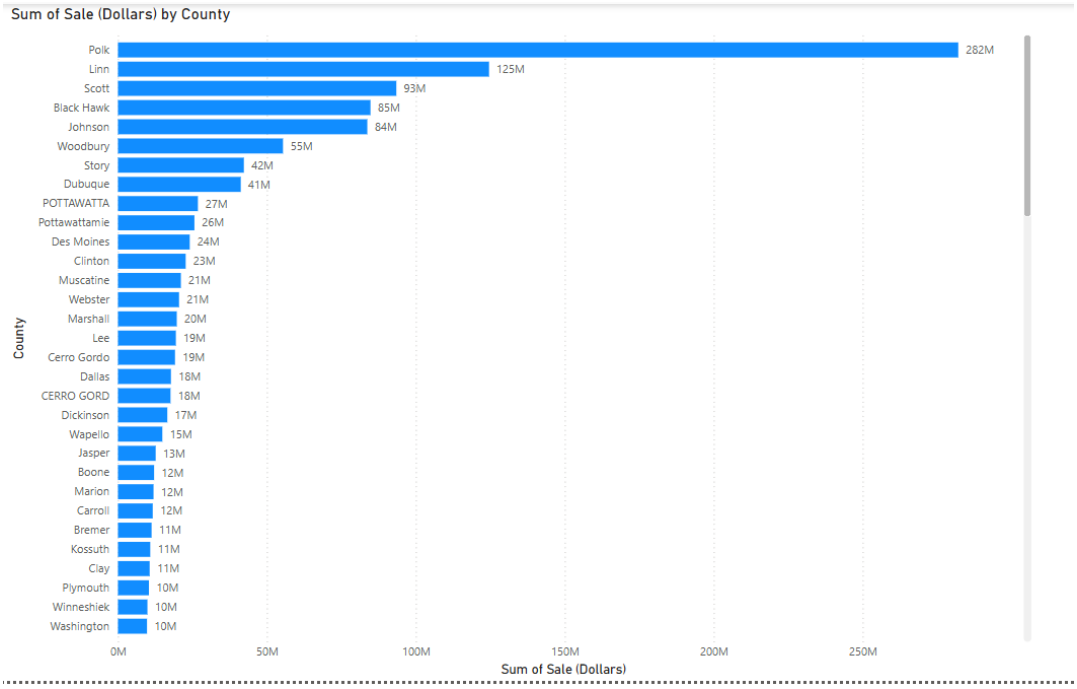
- Allocate more marketing budget and shelf space to these top 3 categories, as they are driving the majority of revenue.
- Products like American Schnapps, White Rum, Mixto Tequila, etc., are under $20M in sales.
- Consider either repositioning or bundling these items to increase volume. Analyze regional trends to identify where demand could be boosted.

2. County-Level Sales Analysis

```
[hadoop@ip-172-31-86-9 ~]$ nano MRCountyLevelSalesAnalysis.py
[hadoop@ip-172-31-86-9 ~]$ python MRCountyLevelSalesAnalysis.py \
    -r hadoop \
    hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned_2.csv \
    --output-dir hdfs:///user/hadoop/output/county_level_sales_litres_gallons/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.4.0
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/MRCountyLevelSalesAnalysis.hadoop.20250430.155118.707713
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/MRCountyLevelSalesAnalysis.hadoop.20250430.155118.707713/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/MRCountyLevelSalesAnalysis.hadoop.20250430.155118.707713/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.0-amzn-0.jar] /tmp/streamjob4603727125573640 40.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-86-9.ec2.internal/172.31.86.9:8032
  Connecting to Application History server at ip-172-31-86-9.ec2.internal/172.31.86.9:10200
  Connecting to ResourceManager at ip-172-31-86-9.ec2.internal/172.31.86.9:8032
  Connecting to Application History server at ip-172-31-86-9.ec2.internal/172.31.86.9:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1746024827769_0002
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:33
  Submitting tokens for job: job_1746024827769_0002
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1746024827769_0002
  The url to track the job: http://ip-172-31-86-9.ec2.internal:20888/proxy/application_1746024827769_0002/
```

```
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -ls /user/hadoop/output/county_level_sales_litres_gallons/
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-30 15:58 /user/hadoop/output/county_level_sales_litres_gallons/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       1581 2025-04-30 15:57 /user/hadoop/output/county_level_sales_litres_gallons/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup       2188 2025-04-30 15:58 /user/hadoop/output/county_level_sales_litres_gallons/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup       2087 2025-04-30 15:58 /user/hadoop/output/county_level_sales_litres_gallons/part-00002
[hadoop@ip-172-31-86-9 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-86-9 ~]$ hadoop fs -ls /home/hadoop/county_level_sales_litres_gallons/
ls: '/home/hadoop/county_level_sales_litres_gallons/': No such file or directory
[hadoop@ip-172-31-86-9 ~]$ cat county_level_sales_litres_gallons/part-* > county_level_sales_litres_gallons.csv
[hadoop@ip-172-31-86-9 ~]$ ls
Liquor_Sales_Cleaned_2.csv       county_level_sales_litres_gallons      mysql-connector-j-9.2.0.tar
MRCountyLevelSalesAnalysis.py    county_level_sales_litres_gallons.csv  top_selling_categories_bottles_sales.csv
MRTopSellingLiquorCategories.py  mysql-connector-j-9.2.0                top_selling_categories_total_bottles_sales
[hadoop@ip-172-31-86-9 ~]$
```

top_selling_categories_bottles_sales.csv is used in Power BI for Analysis , below is the screenshot



Top Recommendations Based on County-Level Sales

- Focus marketing and inventory efforts on top-performing counties like Polk, Linn, and Scott, which together account for a significant share of total sales. These regions are key revenue drivers and should be prioritized for promotions and new product launches.
- Target growth opportunities in mid-to-lower tier counties such as Dallas, Dickinson, and Boone by launching localized campaigns or distribution enhancements to increase visibility and capture untapped potential.

3. Store Performance Analysis

```
[hadoop@ip-172-31-86-9 ~]$ nano MRStorePerformanceAnalysis.py
[hadoop@ip-172-31-86-9 ~]$ python MRStorePerformanceAnalysis.py \
    -r hadoop \
    hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned_2.csv \
    --output-dir hdfs:///user/hadoop/output/store_performance_volume_avg_sale/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 3.4.0
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/MRStorePerformanceAnalysis.hadoop.20250430.161333.726320
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/MRStorePerformanceAnalysis.hadoop.20250430.161333.726320/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/MRStorePerformanceAnalysis.hadoop.20250430.161333.726320/files/
Running step 1 of 1...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.4.0-amzn-0.jar] /tmp/streamjob17011514719943790473.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-86-9.ec2.internal/172.31.86.9:8032
  Connecting to Application History server at ip-172-31-86-9.ec2.internal/172.31.86.9:10200
  Connecting to ResourceManager at ip-172-31-86-9.ec2.internal/172.31.86.9:8032
  Connecting to Application History server at ip-172-31-86-9.ec2.internal/172.31.86.9:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1746024827769_0003
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 1
  number of splits:33
  Submitting tokens for job: job_1746024827769_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1746024827769_0003
```

```
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/output/store_performance_volume_avg_sale/
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MRStorePerformanceAnalysis.hadoop.20250430.161333.726320...
Removing temp directory /tmp/MRStorePerformanceAnalysis.hadoop.20250430.161333.726320...
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -ls /user/hadoop/output/store_performance_volume_avg_sale/
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-30 16:21 /user/hadoop/output/store_performance_volume_avg_sale/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup      41082 2025-04-30 16:20 /user/hadoop/output/store_performance_volume_avg_sale/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup      41674 2025-04-30 16:20 /user/hadoop/output/store_performance_volume_avg_sale/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup      41334 2025-04-30 16:21 /user/hadoop/output/store_performance_volume_avg_sale/part-00002
[hadoop@ip-172-31-86-9 ~]$ cat store_performance_volume_avg_sale/part-* > store_performance_volume_avg_sale.csv
cat: 'store_performance_volume_avg_sale/part-*': No such file or directory
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -get /user/hadoop/output/store_performance_volume_avg_sale/ .
[hadoop@ip-172-31-86-9 ~]$ cat store_performance_volume_avg_sale/part-* > store_performance_volume_avg_sale.csv
[hadoop@ip-172-31-86-9 ~]$ ls
Liquor_Sales_Cleaned_2.csv       county_level_sales_litres_gallons      store_performance_volume_avg_sale
MRCountyLevelSalesAnalysis.py    county_level_sales_litres_gallons.csv  store_performance_volume_avg_sale.csv
MRStorePerformanceAnalysis.py    mysql-connector-j-9.2.0                top_selling_categories_bottles_sales.csv
MRTopSellingLiquorCategories.py  mysql-connector-j-9.2.0.tar            top_selling_categories_total_bottles_sales
[hadoop@ip-172-31-86-9 ~]$
```

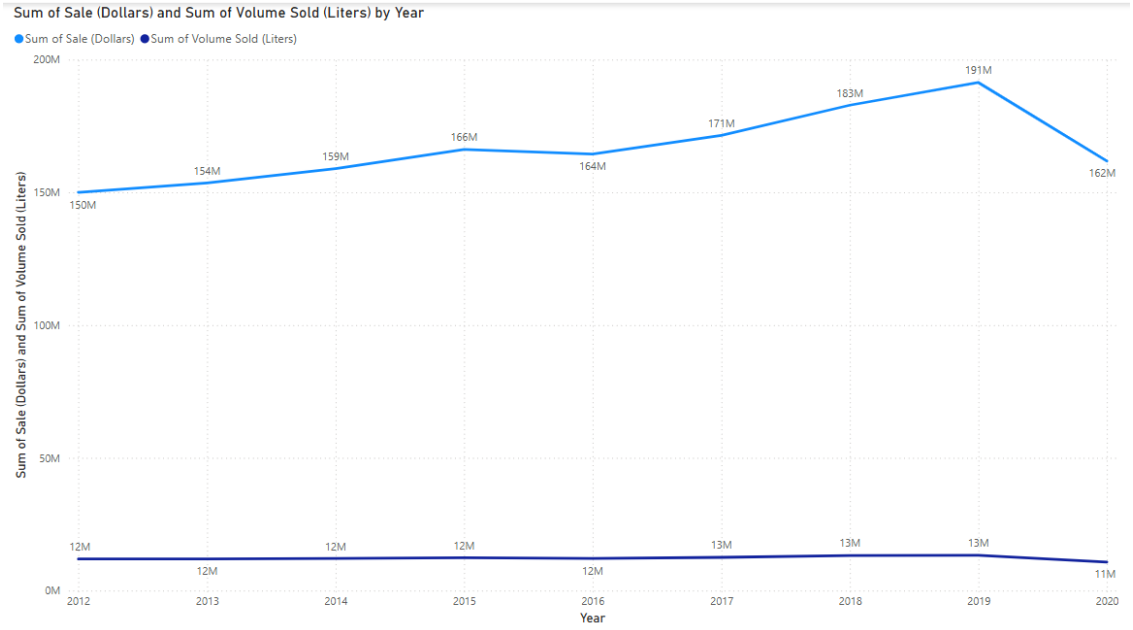
Store Name	Sum of Sale (Dollars)	Sum of Bottles Sold	Sum of Volume Sold (Liters)
Hy-Vee #3 / BDI / Des Moines	18,951,530.16	1572637	1,420,817.98
Central City 2	16,299,714.51	1407106	1,254,165.01
Hy-Vee Wine and Spirits / Iowa City	13,966,009.03	1172046	1,036,316.67
Benz Distributing	10,381,865.58	793480	735,097.93
Central City Liquor, Inc.	9,661,522.97	751590	627,331.11
Hy-Vee #4 / WDM	9,151,753.25	657151	571,016.85
Lot-A-Spirits	8,969,790.53	758281	695,274.71
Hy-vee #2 / Ames	8,962,681.98	710593	623,054.88
Wilkie Liquors	8,669,420.22	709439	666,330.44
Hy-Vee #1044 / Burlington	8,500,264.58	693655	629,086.50
Hy-Vee Food and Drug / Clinton	8,484,379.03	730852	672,800.24
Hy-Vee Wine and Spirits / Bettendorf	8,206,991.78	626129	562,513.05
Hy-Vee Wine and Spirits / WDM	7,920,432.14	608732	525,598.55
Hy-Vee Food Store / Coralville	7,839,186.21	640092	572,711.89
Hy-Vee Food Store / Fleur / DSM	7,403,129.57	602138	506,578.63
Hy-Vee / Waukee	7,029,511.78	521706	453,933.97
Hy-Vee Food Store / Marshalltown	6,948,821.99	554815	521,958.14
Hy-Vee Food Store / Muscatine	6,935,302.64	543133	521,745.19
Hy-Vee Food Store / Altoona	6,857,009.08	568206	488,817.50
Hy-Vee Food Store #1 / Mason City	6,824,491.59	530454	509,471.46
Hy-Vee Food Store #4 / Sioux City	6,689,814.39	556242	496,089.63
Hy-Vee Food Store #5 / Cedar Rapids	6,580,194.12	545109	458,981.41
Hy-Vee #4 / Davenport	6,573,916.73	489681	433,959.51
Hy-Vee Food and Drug / Grand / WDM	6,542,861.31	535670	459,459.48
Hy-Vee Food Store #3 / Cedar Rapids	6,429,650.47	544653	466,108.49
Hy-Vee #7 / Cedar Rapids	6,384,243.94	467734	436,160.06
Hy-Vee Food Store / Cedar Falls	6,239,674.48	469571	424,372.95
Hy-Vee Food Store / Carroll	6,185,919.60	515527	487,833.37
Hy-Vee Fort Dodge Wine and Spirits	6,019,823.12	501817	462,632.14
Hy-Vee Food Store #3 / Sioux City	5,918,861.19	515880	460,981.18
Hy-Vee Drugstore / University / DSM	5,733,753.93	504464	397,068.68
Charlie's Wine and Spirits,	5,690,305.62	484219	427,043.29
I-80 Liquor / Council Bluffs	5,638,356.71	484686	435,864.31
Total	**1,500,497,152.36**	**127198322**	**110,159,349.01**

Top Recommendations:

- Focus on top-performing stores like Hy-Vee #3 (Des Moines), Central City 2, and Hy-Vee Wine & Spirits (Iowa City) — these locations collectively account for a major share of total revenue and should be prioritized for new launches, seasonal promotions, and premium inventory.
- Analyze conversion efficiency by comparing volume vs. revenue — stores like Benz Distributing and Hy-Vee #2 (Ames) show high sales per liter, indicating strong pricing/mix performance worth replicating.

3. Trends in Liquor Sales Over Time

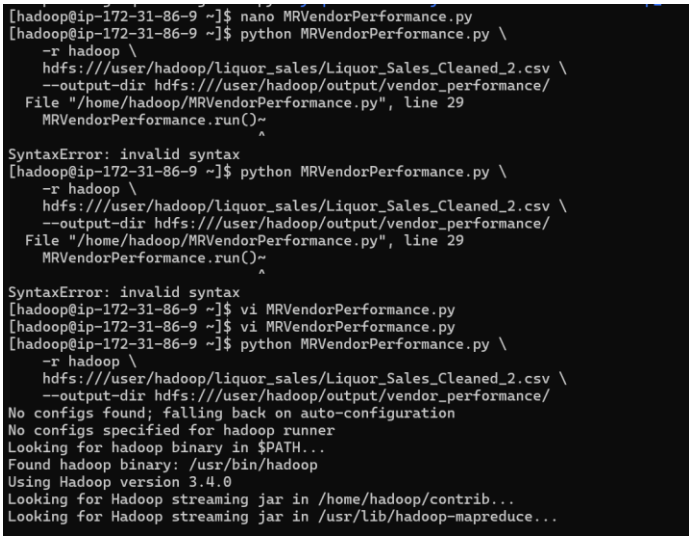
```
job output is in hdfs:///user/hadoop/output/liquor_sales_trends/
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MRLiquorSalesTrends.hadoop.20250430.163804.923581...
Removing temp directory /tmp/MRLiquorSalesTrends.hadoop.20250430.163804.923581...
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -get /user/hadoop/output/liquor_sales_trends/ .
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -ls /user/hadoop/output/liquor_sales_trends/
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-30 16:45 /user/hadoop/output/liquor_sales_trends/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       5815 2025-04-30 16:44 /user/hadoop/output/liquor_sales_trends/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup       5773 2025-04-30 16:45 /user/hadoop/output/liquor_sales_trends/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup       5780 2025-04-30 16:45 /user/hadoop/output/liquor_sales_trends/part-00002
[hadoop@ip-172-31-86-9 ~]$ cat liquor_sales_trends/part-* > liquor_sales_trends.csv
[hadoop@ip-172-31-86-9 ~]$ ls
Liquor_Sales_Cleaned_2.csv       county_level_sales_litres_gallons      mysql-connector-j-9.2.0.tar
MRCountyLevelSalesAnalysis.py    county_level_sales_litres_gallons.csv  store_performance_volume_avg_sale
MRLiquorSalesTrends.py           liquor_sales_trends                    store_performance_volume_avg_sale.csv
MRStorePerformanceAnalysis.py    liquor_sales_trends.csv                top_selling_categories_bottles_sales.csv
MRTopSellingLiquorCategories.py  mysql-connector-j-9.2.0                top_selling_categories_total_bottles_sales
[hadoop@ip-172-31-86-9 ~]$
```



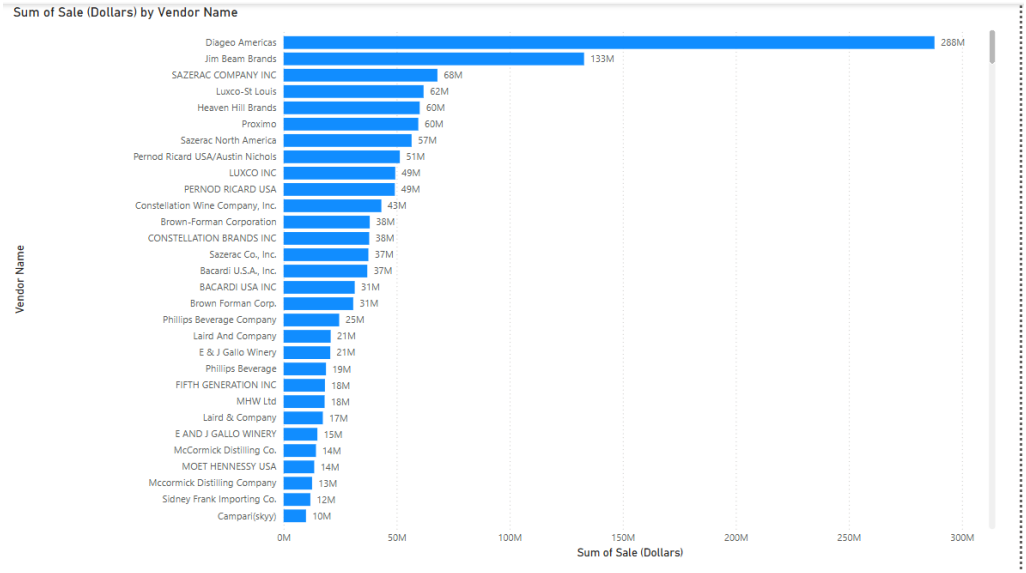
Recommendations:

- Leverage momentum from the growth period (2014–2019) where both sales and volume steadily increased, peaking in 2019. Analyze key drivers (e.g., promotions, new products, vendor performance) from these years to replicate success strategies.
- Investigate the significant drop in 2020, where sales fell from $191M to $162M and volume also declined. Evaluate potential external factors (e.g., COVID-19 impact, supply chain issues) and adjust forecasting, inventory, and marketing accordingly.
- Despite stable volume from 2012–2019 (~12–13M liters), revenue increased, indicating improved pricing or premium product mix. Continue to explore and promote high-margin products to maintain revenue growth even if volume stays flat.
- Build resilience for future downturns by diversifying channels (e.g., online sales), expanding product categories, and strengthening vendor contracts to avoid revenue dips like in 2020.

4. Vendor Performance



```
                Reduce input groups=472
                Reduce input records=17687883
                Reduce output records=472
                Reduce shuffle bytes=109128178
                Shuffled Maps =99
                Spilled Records=35375766
                Total committed heap usage (bytes)=164322344496
                Virtual memory (bytes) snapshot=119310790656
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/hadoop/output/vendor_performance/
Removing HDFS temp directory hdfs:///user/hadoop/tmp/mrjob/MRVendorPerformance.hadoop.20250430.170414.727264...
Removing temp directory /tmp/MRVendorPerformance.hadoop.20250430.170414.727264...
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -get /user/hadoop/output/vendor_performance/ .
[hadoop@ip-172-31-86-9 ~]$ hdfs dfs -ls /user/hadoop/output/vendor_performance/
Found 4 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-04-30 17:11 /user/hadoop/output/vendor_performance/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       7763 2025-04-30 17:10 /user/hadoop/output/vendor_performance/part-00000
-rw-r--r--   1 hadoop hdfsadmingroup       7462 2025-04-30 17:11 /user/hadoop/output/vendor_performance/part-00001
-rw-r--r--   1 hadoop hdfsadmingroup       8804 2025-04-30 17:11 /user/hadoop/output/vendor_performance/part-00002
[hadoop@ip-172-31-86-9 ~]$ ls
Liquor_Sales_Cleaned_2.csv       county_level_sales_litres_gallons      store_performance_volume_avg_sale
MRCountyLevelSalesAnalysis.py    county_level_sales_litres_gallons.csv  store_performance_volume_avg_sale.csv
MRLiquorSalesTrends.py           liquor_sales_trends                    top_selling_categories_bottles_sales.csv
MRStorePerformanceAnalysis.py    liquor_sales_trends.csv                top_selling_categories_total_bottles_sales
MRTopSellingLiquorCategories.py  mysql-connector-j-9.2.0                vendor_performance
MRVendorPerformance.py           mysql-connector-j-9.2.0.tar
[hadoop@ip-172-31-86-9 ~]$
```



Top Recommendation Based on Vendor Sales Chart:

- Leverage and deepen strategic partnerships with top-performing vendors like Diageo Americas and Jim Beam Brands, who collectively contribute over 40% of total vendor sales. These vendors are critical revenue drivers and should be prioritized for joint marketing initiatives, preferential shelf placement, and exclusive promotional campaigns to further boost overall sales performance.