Liquor Sales Analysis Solution Instruction (Step by Step):

✅ Step 1: Load Data from S3 to HDFS

1. **Download the cleaned CSV from S3**:
2. aws s3 cp s3://liquor-sales-assignment-150489/Liquor_Sales_Cleaned.csv .
3. **Create HDFS directory**:
4. hdfs dfs -mkdir -p /user/hadoop/liquor_sales/
5. **Upload to HDFS**:
6. hdfs dfs -put Liquor_Sales_Cleaned.csv /user/hadoop/liquor_sales/

✅ Step 2: Install Required Python Libraries

pip install mrjob --user

pip install boto3 --user

📊 Analysis 1: Total Revenue per Store

Script: MRTotalRevenueByStore.py

Command:

python MRTotalRevenueByStore.py -r hadoop hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned.csv --output-dir hdfs:///user/hadoop/output/total_revenue_by_store/

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

cat total_revenue_by_store/part-* > total_revenue_by_store.csv

---

📊 Analysis 2: Top-Selling Liquor Categories (By Bottles & Revenue)

Script: MRTopSellingLiquorCategories.py

Command:

python MRTopSellingLiquorCategories.py -r hadoop hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned.csv --output-dir hdfs:///user/hadoop/output/top_selling_categories/

hdfs dfs -get /user/hadoop/output/top_selling_categories/

cat top_selling_categories/part-* > top_selling_categories.csv

---

📊 Analysis 3: County-Level Sales (Sale $, Litres, Gallons)

Script: MRCountyLevelSalesAnalysis.py

Command:

python MRCountyLevelSalesAnalysis.py -r hadoop hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned.csv --output-dir hdfs:///user/hadoop/output/county_level_sales/

hdfs dfs -get /user/hadoop/output/county_level_sales/

cat county_level_sales/part-* > county_level_sales.csv

📊 Analysis 4: Store Performance (Revenue, Volume, Avg Sale)

Script: MRStorePerformanceAnalysis.py

Command:

python MRStorePerformanceAnalysis.py -r hadoop hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned.csv --output-dir hdfs:///user/hadoop/output/store_performance/

hdfs dfs -get /user/hadoop/output/store_performance/

cat store_performance/part-* > store_performance.csv

📊 Analysis 5: Vendor Performance (By Revenue & Volume)

Script: MRVendorPerformance.py

Command:

python MRVendorPerformance.py -r hadoop hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned.csv --output-dir hdfs:///user/hadoop/output/vendor_performance/

hdfs dfs -get /user/hadoop/output/vendor_performance/

cat vendor_performance/part-* > vendor_performance.csv

📊 Analysis 6: Monthly & Yearly Sales Trends

Script: MRLiquorSalesTrends.py

Command:

python MRLiquorSalesTrends.py -r hadoop hdfs:///user/hadoop/liquor_sales/Liquor_Sales_Cleaned.csv --output-dir hdfs:///user/hadoop/output/liquor_sales_trends/

hdfs dfs -get /user/hadoop/output/liquor_sales_trends/

cat liquor_sales_trends/part-* > liquor_sales_trends.csv

📦 Final Step: Copy Output to Local

(Repeat for each output folder)

scp -i Drunken_Master.pem hadoop@ec2-52-0-19-246.compute-1.amazonaws.com:/home/hadoop/<output_file>.csv .
