**HDFS Data from Apache Flume**

Copy CSV file to Docker container

Copy the Apache flume binary

Copy flume config - flume.conf

Run flume to load the data

**bin/flume-ng agent --conf conf --conf-file /opt/flume.conf -name agent1 -Dflume.root.logger=DEBUG,console**

**Hive and HDFS**

**Create the Hive Table**

CREATE TABLE retail (

RecordNo INT,

Invoice STRING,

StockCode STRING,

Description STRING,

Quantity INT,

InvoiceDate STRING,

Price DOUBLE,

CustomerID INT,

Country STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/flume';

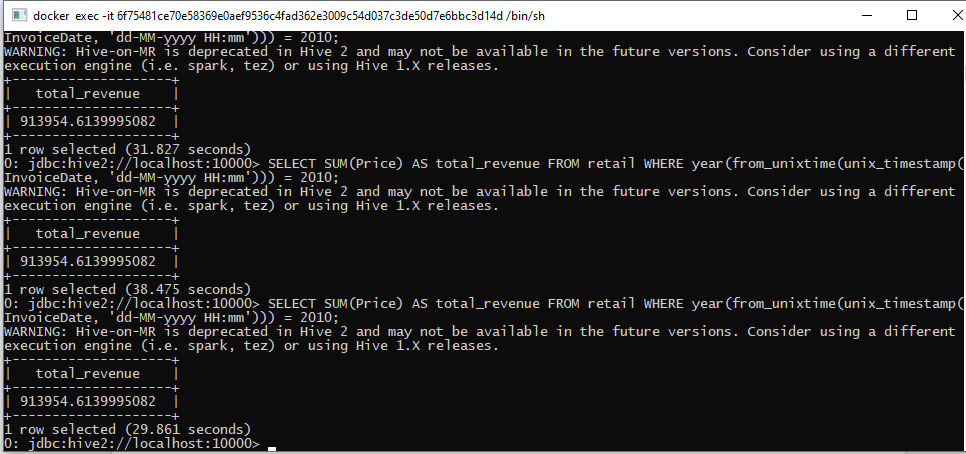
**Query 1**

Total revenue (Aggregation of Price) received in the year 2010.

Hive Query:

**SELECT SUM(Price) AS total\_revenue FROM retail WHERE year(from\_unixtime(unix\_timestamp(InvoiceDate, 'dd-MM-yyyy HH:mm'))) = 2010;**

|  |  |  |
| --- | --- | --- |
| **Query** | **Framework** | **Time** |
| Query 1 | Hive Query with Hadoop | 37.708s |
| Query 1 | Hive Query with Hadoop | 34.037s |
| Query 1 | Hive Query with Hadoop | 31.918s |
| Query 1 | Hive Query with Hadoop | 30.013s |
| Query 1 | Hive Query with Hadoop | 38.442s |
| Query 1 | Hive Query with Hadoop | 31.995s |
| Query 1 | Hive Query with Hadoop | 31.827s |
| Query 1 | Hive Query with Hadoop | 38.475s |
| Query 1 | Hive Query with Hadoop | 29.861s |



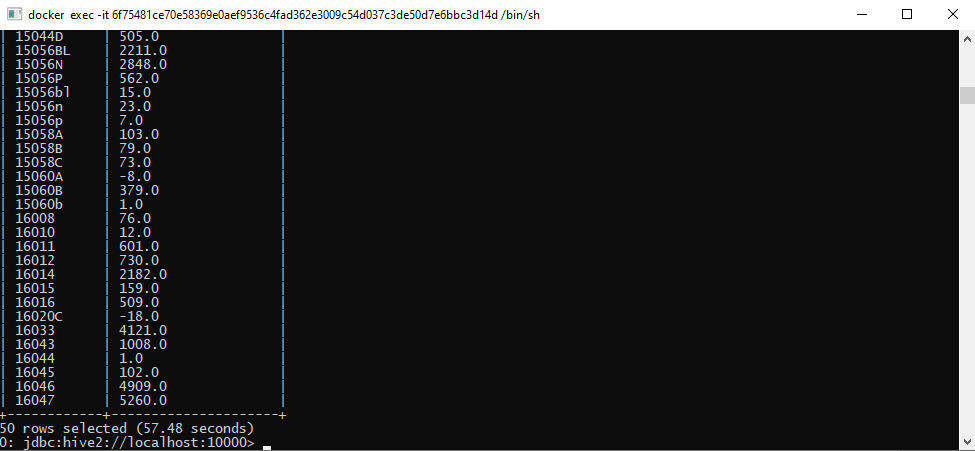
Query 2

**List of unique items sold (With same StockCode) and their total sales volume (Aggregation of Quantity) in the year 2010 sorted in ascending order of StockCode.**

Hive Query 2

**SELECT StockCode, SUM(Quantity) AS total\_sales\_volume FROM retail WHERE year(from\_unixtime(unix\_timestamp(InvoiceDate, 'dd-MM-yyyy HH:mm'))) = 2010 GROUP BY StockCode ORDER BY StockCode ASC LIMIT 50;**

|  |  |  |
| --- | --- | --- |
| **Query** | **Framework** | **Time** |
| Query 2 | Hive Query | 58.912s |
| Query 2 | Hive Query | 69.011s |
| Query 2 | Hive Query | 55.796s |
| Query 2 | Hive Query | 65.935s |
| Query 2 | Hive Query | 64.882s |
| Query 2 | Hive Query | 105.332s |
| Query 2 | Hive Query | 63.8s |
| Query 2 | Hive Query | 61.285s |
| Query 2 | Hive Query | 57.48.s |

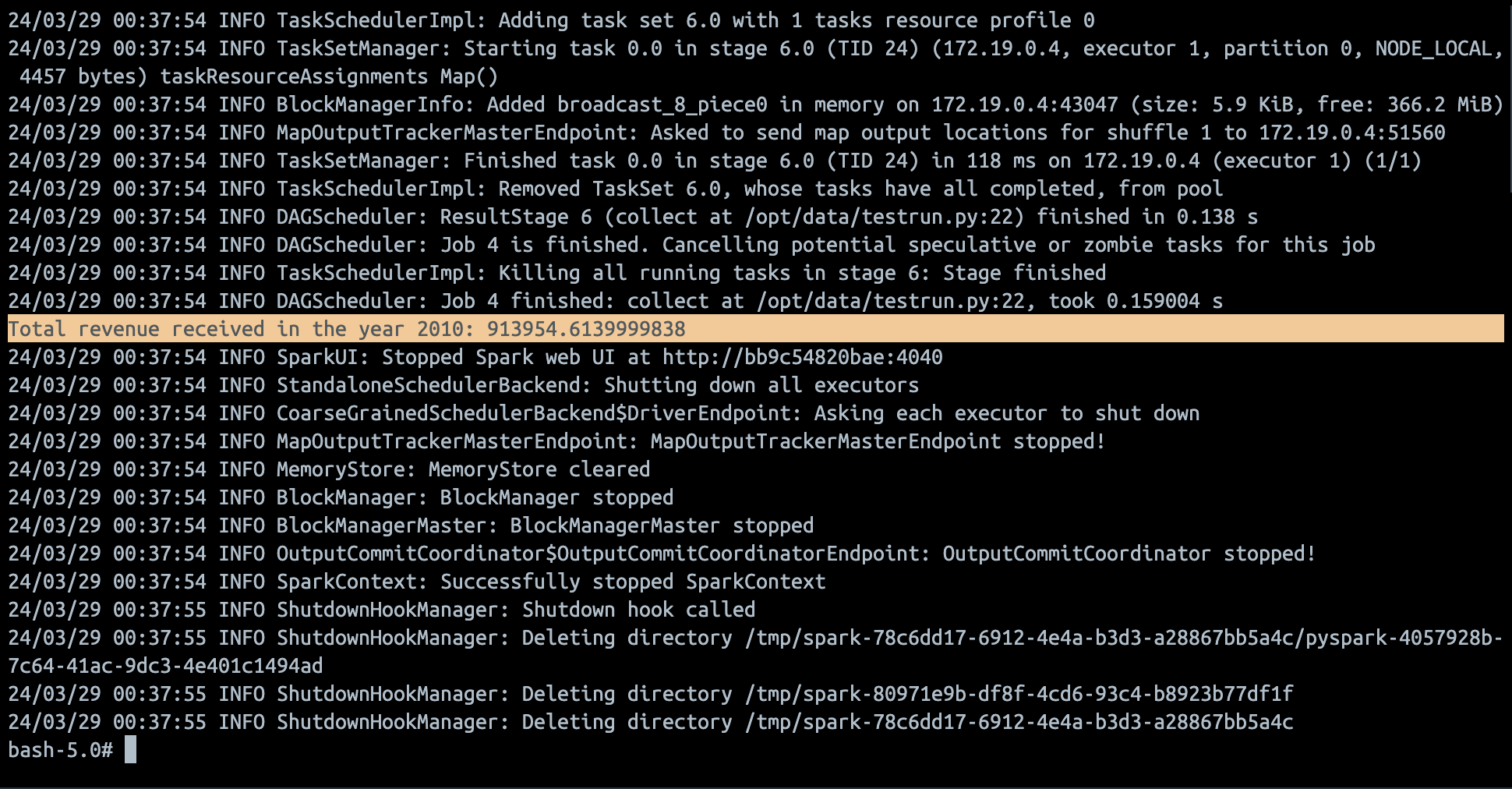


Spark Data frame

Query 1

A screenshot of a computer

Description automatically generated



Query 2

A screenshot of a computer

Description automatically generated

