Hadoop Use-Case

Datasets: stocks.txt, dividends.txt

1. Copy the stocks and dividends files into HDFS location of your choice.

2. Create 'Stocks' and ‘Dividends’ tables as Hive managed tables using tab as field separator.

3. Load the above files into the managed tables created in Step 2 using appropriate format.

4. Create 'StocksBucket' table as Hive bucketed table clustered by stock-symbol into 5 buckets. Load the stocks data into the bucketed table from the managed table created in step 3.

5. Create a SparkSQL program to load 'stocks' and 'dividends' data into two different dataframes and create a new dataframe that contains the following data fields (shown below) and save it to a CSV file.

    - New CSV File Data: Exchange, Symbol, AverageDividend, AverageClosingPrice, AverageVolume

6. Create a Kafka topic with 3 partitions and a replication-factor of 3. Create this topic by starting three brokers in single-node multiple-broker configuration. Name the topic as "stock_data"

7. Write a Kafka Producer using Kafka Producer API that reads data from the file created in step 5, and produce the data to a Kafka topic "stock_data" created in step 6.