DESCRIPTION

Objective: To use hive features for data engineering or analysis and sharing the actionable insights

Problem Statement:

NewYork stock exchange data of seven years, between 2010 to 2016, is captured for 500+ listed companies. The data set comprises of intra-day prices and volume traded for each listed company. The data serves both for machine learning and exploratory analysis projects, to automate the trading process and to predict the next trading-day winners or losers.. The scope of this project is limited to exploratory data analysis.

Domain: BFSI

Analysis to be done: Exploratory analysis to understand how MoM or YoY companies from different sectors or industries and states have progressed in a period of 7 years

Content: This data set contains prices.csv and securities.csv files having the following features:

Prices.csv:

Date: Trading date

Symbol: Ticker code or listed company code on NY stock exchange

Open: Intra-day opening price for each listed company

Close: Intra-day closing price for each listed company

Low: Intra-day lowest price for each listed company

High: Intra-day highest price for each listed company

Volume: Number of shares traded per day per company

Securities.csv:

Ticker\_Symbol: Country to which the customer belongs

Security: Legal name of the listed company

Sector: Business vertical of the listed company

Sub\_Industry: Business domain of the listed company within a Sector.

Headquarter: Headquarters address

Steps to perform:

1) Create a data pipeline using sqoop to pull the data from the table below from MYSQL server into Hive.

a. MYSQL DATABASE NAME: BDHS\_PROJECT

i. Stock\_prices

ii. Stock\_companies

Check the TABLE description: STOCK\_PRICES

Column Name Datatype

Trading\_date Date

Symbol String

Open double

Close double

Low double

High double

Volume int

TABLE: STOCK\_COMPANIES

Column Name Datatype

Symbol String

Company\_name String

Sector String

Sub\_industry String

Headquarter String

2) Create a new hive table with the following fields by joining the above two hive tables.

Please use appropriate Hive built-in functions for columns (a,b,e and h to l).

Trading\_year: Should contain YYYY for each record

Trading\_month: Should contain MM or MMM for each record

Symbol: Ticker code

CompanyName: Legal name of the listed company

State: State to be extracted from headquarters value.

Sector: Business vertical of the listed company

Sub\_Industry: Business domain of the listed company within a sector

Open: Average of intra-day opening price by month and year for each listed company

Close: Average of intra-day closing price by month and year for each listed company

Low: Average of intra-day lowest price by month and year for each listed company

High: Average of intra-day highest price by month and year for each listed company

Volume: Average of number of shares traded by month and year for each listed company

DATA ANALYSIS USING HIVE

3) Find the top five companies that are good for investment

4) Show the best-growing industry by each state, having at least two or more industries mapped.

5) For each sector find the following.

Worst year

b. Best year

c. Stable year

You can download the datasets from here