****

**DataStax Enterprise Final Project**

**INTRODUCTION**

The goal of the project is to create an analysis system for the financial services industry. The system will provide analysts with historical time-series tick data for a specific financial instrument. Tick data represents the time, price and quantity traded (shares for stocks or contracts for commodities) for a single trade. From this data, an analyst can create price and trade data in various durations of time (typically referred to as “price bars”, e.g. 1 minute bars, 5 minute bars, etc.). Parameters of interest typically include the open, high, low, and close prices for each price bar. The total volume traded during each bar is also of interest. A time-series of price bars can then be created to form a graphical “bar chart”. Tabular representation is common for analysis and will be used throughout this exercise.

News events and their impact on stock prices is another area of interest. Therefore, the system must provide news. RSS feeds and FTP drop-boxes are common implementations

From the perspective of the analyst, the system must collect time-series tick data for each financial instrument of interest (the Crude-oil continuous commodities contract, in this case). Each entry must contain the date, time, price, and quantity traded so that the open, high, low, and close prices can be determined in order to create price bars. The total volume traded for each price bar must also be determined. Analysis of the price bar time-series and/or the price/quantity data with any given price bar must be facilitated. News must be available in a textual format that can be parsed, stored, and queried in a flexible manner.

From the perspective of the data provider, the system must contain all data that allows an analyst complete flexibility to create a time-series of price bars in any duration. The provider must ensure data sequence, stability, bandwidth and availability at all times.

Feel free to use any and all available references throughout the project.

Notes:

1. Open is the first tick in the price bar. Close is the last tick in the price. High is the highest price in the price bar. Low is the lowest price in the price bar.
2. Approach to a problem is as meaningful as the correct answer. A “correct answer” may not exist.
3. Transform and parse source data into any required intermediate format.
4. Teamwork is encouraged.

**DATA MODELING**

The source data for the system consists of tick data in a .csv file and news data in a single .xml file. (formats for each are shown below).

Task:

Develop an efficient means to ingest the source data .csv/.xml files into the cluster. All APIs, Cassandra tools, and/or third-party tools are acceptable. Please provide a description of all procedures used to accomplish this task along with any issues encountered.

From this data, please provide a data model including associated keyspaces and column-families to yield the information listed in Table 1 for the financial analysis system described above. All performance related best practices should be considered, throughout. Also provide trace information for each.

Deliverables:

Please provide a detailed description of the solution including assumptions and/or difficulties encountered. Also provide the result-sets, keyspace description output, and any scripts and/or application code.

**Table 1 – Cassandra Data Modeling Project Objectives**

1. All tick data within a given 1-minute interval for a specified date.
2. All tick data within a given 10-minute interval for a specified date.
3. Total quantity traded at a given price on for a specified date (e.g. quantity traded at 98.01 on 1/10/2014).
4. Total quantity traded at a given price on for a specified 1-minute interval on specified date (e.g. quantity traded at 98.01 for the 1-minute interval beginning at 10:01:00 on 1/10/2014).
5. Total quantity traded at a given price on for a specified 10-minute interval on specified date (e.g. quantity traded at 98.01 for the 10-minute interval beginning at 13:10:00 on 1/10/2014).

**DATA MODELING EXTRA CREDIT**

Create open, high, low, close, total quantity trade price bars for 1-minute and 10-minute intervals.

**SOLR**

An application of Solr in the DataStax Enterprise stack is to perform free-text and faceted searches within data fields. A noted area of interest in financial news events are changes in a company’s executive office -- for example, a CEO is replaced or the Board of Directors adds new members.

Task:

Determine changes in the executive office and Board of Directors for the news articles provided.

Deliverables:

Please provide a detailed description of the solution including assumptions and/or difficulties encountered. Also provide solrconfig.xml, solrschema.xml, description of required column-families, and all results files.

**Table 3 – Solr Project Objectives**

1. Develop a schema for news data.
2. Validate schema.
3. Load all news data.
4. Determine the names of all replaced and/or new C-Suite executives and/or Board members along with the Exchange:Symbol designation for the affected company.

**SOLR EXTRA CREDIT**

The letter ‘s/S’ is said to be the most commonly occurring letter in the English language. Validate or disprove this in the news events. Whitespace characters are not to be considered.

**DEMO**

Please create a recorded demo or presentation of the project solution (time permitting).

****

**APPENDIX – Sample Data Format and Specification**

=================

Data is provided for greater than one million tick data points.

File:

cl\_ticks..csv

Format:

date, date, hhmm.sequence, price, quantity

Sample Record:

|  |
| --- |
| 1/10/2014,1/10/2014,1040.90640,92.50,1 |

=================

Stock News is provided in a single XML file.

File:

Stocknews.xml

Format:

<stocknews>

<page>

<title></title>

<id></id>

<symbol></symbol>

<date></date>

<text xml:space="preserve"></text>

</page>

</stocknews>

Sample Record:

<stocknews>

<page>

<title>Fitch: Activist Investor Activity Distracts U.S. Restaurants</title>

<id>1</id>

<symbol>EXCH:SYM</symbol>

<date>Jan 13, 2014</date>

<text xml:space="preserve">

NEW YORK, Jan 13, 2014 (BUSINESS WIRE) --

Increased activist investor activity in the restaurant space portends change for

the U.S. restaurant industry and is a testament to the fact that shareholder

demands continue to rise despite challenging operating conditions, according to

Fitch Ratings. Darden Restaurants Inc. (Darden), Bob Evans Farms, Inc., and …

</text>

</page>

</stocknews>

=================