DWBI Project: Architect, Populate and Explore a Data Warehouse for Stock Market Analysis

ITI - Data Visualization Track - Graduation Project

1 Overview

This project will assess your ability to apply the theories, methodologies and strategies tackled in the diploma to successfully implement a data warehouse to support business intelligence queries.
You are asked to build a data warehouse model, your project should meet the following minimum
requirements:

- At least 4 sources of relatable data must be included.(3 sources for teams of 4)
- A data preparation process is required to extract data from the original sources, clean the data, transform it into a suitable structure for your data warehouse data model, and load your data warehouse.
- The data warehouse must support 5 non-trivial BI queries.
- A Dashboard that answers those queries.

2 Deliverables

For the project, there are 2 deliverables: 1) a written report, and 2) a 15 minutes presentation

The report should be structured as follows:

- **Introduction**: description of the setting of your data warehouse. This section should document: the objectives of the project, and lay out three business requirements for the data warehouse to meet. Suggested length: 1 page.

- **Data Sources**: A formal description of the sources of data, and a motivation for these sources of data.
You should describe how the data sets have been collected, note their age (i.e., when they were made available, not when they were last modified). If you have fabricated your own data for this project, describe how this has been done, and what assumptions and/or data artefacts are interwoven in this process. Suggested length: 1 page per dataset, (max 4 pages).

- **Data Warehouse Data Model**: description of the schema(s) built. What dimensions do you have, and why? Also include a discussion on why they are composed in this manner. Noteworthy, also is how the data model has been prepared for the BI queries. You MUST include how each of the data sources are included in the data model and describe how they populate the data model. Include both of the following: your star/snowflake/galaxy schema(s); and which sources of data contribute to your schema including where they are present in the schema. For example, if a dimension is populated from multiple data sets/sources explain where they contribute and why they were included in this manner.

Suggested length: 2 pages.

• **Application of Data Warehouse**: Execute five non-trivial queries that high- light the business value of your data warehouse, and which meet your requirements. For each query, you must note the sources of data contributing to the answer of the query and how. Each BI query should be articulated as a question that is answered using the data warehouse. You can use Tableau or Power BI for building the Dashboard.

• **Conclusion**: how well did your data warehouse do at addressing your requirements Suggested length: 1 page.

• **Appendix**: listings of all code (as text **NOT** as images) used in the project. If you have used code from online tutorials, include links to these sources instead noting briefly how they have been used. The final report should not be more than 20 pages.

The presentation should be no more than 5 minutes in length. It should showcase core functionality of the data ware- house and preparation process. It should also show the execution of 4 or more of the BI queries discussed in the report. The voice commentary should explain key aspects of what is being shown, and noteworthy implications as well as interesting / surprising results or outcomes.

3 DataSources

You are expected to independently source relevant data for use in the case studies. Note: data sets will likely need cleaning and you need to ensure that appropriate relationships exist between your data sets, such that they can be transformed and loaded into the warehouse for meaningful case studies.
The data warehouse
must incorporate the following sources:
• https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
• https://www.kaggle.com/camnugent/sandp500
• https://datahub.io/core/s-and-p-500-companies-financials#resource-constituents-financials

Extra credit will be given for incorporating specifically challenging or innovative sources of data.

The report should detail the sources of data, how they were generated or extracted, and the steps taken to load and transform the data for storage in the warehouse.
Possible sources of datasets include, but are not limited to:
• European Data Portal, EU Open Data Portal, and other http://data.europa.eu/
• UK's open government data repository: http://data.gov.uk
• Kaggle: http://www.kaggle.com
• Run My Code: http://www.runmycode.org/
• Amazon's public dataset repository: https://aws.amazon.com/datasets
• Google's Public Data Directory: http://www.google.com/publicdata/directory
• The UCI machine learning repository: http://archive.ics.uci.edu/ml/
• Google Data Search: https://toolbox.google.com/datasetsearch
• Data.gov https://www.data.gov/
• Quandl https://www.quandl.com
• Datahub: https://datahub.io/

4 BI Queries

The BI queries represent five (4 queries for the team of 4) exemplary non-trivial knowledge discovery exercises facilitated through your prototypical data warehouse implementation. Appropriate presentation of the results should be provided in the report. Their implications should be

appropriately discussed referencing relevant literature where applicable.
BI Queries should be:
1. Reporting stock market daily statistics
2. Reporting top and lowest performers
3. Tracking stock market changes

5 Dashboard

The Dashboard should be interactive and ready to answer the business questions that would provide Stock Traders/ Managers the ability to take powerful decisions.

6 Recommended Milestones

- Data Model – 2 Days
- Data Preparation – 2 Days
- Queries – 1 Day
- Dashboard – 2 Days

7 Stuck on the project?

Watch these two videos, they tend to help understand what Stock Market Exchange is:

- https://www.youtube.com/watch?v=p7HKvqRI_Bo
- https://www.youtube.com/watch?v=F3QpgXBtDeo