Individual Contribution Report

IRISH Stock Market Analysis

Student:

Neeraj Chavan

Teacher:

Dr. Shazia A. Afzal

Course:

Data Storage Solutions for Data Analytics

# Brainstorming

I have the habit checking the news while I’m having my morning coffee. And one fine day, when I was sitting with my friend Bharat, discussing the recent news and developments, I came across the news about how Brexit will impact the Irish businesses. It was at this moment Bharat proposed an interesting idea of analyzing the Irish Stock Market as a part of our Data Storage Solutions project. We wanted to see how real-world events impact the stock market. This is when we started to think about how to approach this and how to go about developing a data storage solution for this analysis.

# Understanding Concepts

|  |  |  |
| --- | --- | --- |
| We began by scraping the Irish Stock market data from multiple sources. Bharat created the code for scraping the Irish Times and PESTLE data in Python. While Prasad scraped the data for Irish Stock Market, over the past 15 years.  However, the thing about scraping data from multiple sources is that the data is very unclean. So, I had to clean all the scraped data using R. I and Aayush then split the cleaning work and finally we had the cleaned dataset.  During this phase I understood how to scrape data using Python and Rvest(R) and how to clean the data to remove null and N/A values in R.  Then we created the ERD model and decided to go with Star Schema for our project.  After that we created the source database in SQL Server. This was my first experience of working in SQL and it was quite an exciting one for me. |  | From understanding how to create databases & learning about basic data manipulation statements to using Joins and Stored Procedures, I familiarized myself with the SQL Server Management environment. The examples taught in the class helped me grasp the fundamental concepts of Relational Databases and Data Warehousing.  After that, I learnt about ETL using SSIS in Visual Studio. The basic process of how to load source database and the dimension tables and fact tables into the data warehouse. From using various SSIS tools like Merge, Data Conversion, Lookup etc., I was able to understand SSIS in a more elaborate manner.  Reporting is one of the most important aspects of any data storage solution. And I was able to understand how to develop reports using SSRS and visualizations in R.  Also, I had the opportunity to learn about XML Schema and explored the concept of graph databases using Neo4j. |

# Implementation of Aforementioned Concepts in the Project

## Relational Databases

I contributed in creating the ERD model and the dimensional model based on the things I had learnt in class.

## Data Warehousing

I contributed in loading source tables and dimension tables in SSIS. And, helped with developing reports in SSRS.

## Extensible Markup Language(XML)

Learnt about how to create XML schema from SQL and used .NET console in Visual Studio to develop and generate the XSD schema for the dimensional tables and fact tables.

## Graph Databases (Neo4j)

Understood how to load CSV into Graph Databases.

# Challenges Faced

|  |  |  |
| --- | --- | --- |
| Getting Used To SQL |  | Cleaning of Data |
| Being a novice to SQL, I had no previous experience with it. So initially, I had to dedicate quite a chunk of my time in understanding how SQL queries are developed and how to write optimized queries. |  | One of the primary challenges I had to face initially was during cleaning of data. The scraped data was unclean with null/NA/garbage values. I had to ensure that the dataset was totally clean before using it for analysis. I did so by using R |
|  |  |  |

# Conclusion

Overall, it was an insightful experience working on this project. I am now much more confident about data storage concepts and their applications in real world. From understanding SQL, SSIS, SSRS to generating XML Schema and exploring graph databases in NEO4J, it was a good learning curve for me.

The further scope of this project is develop it in such a manner so as to try use it to analyze real time data and then later generate predictions based on that.

|  |  |
| --- | --- |
|  |  |