# The impact of public information on stock market prices

# by Andrian Potereanu

# Introduction

## 1.1 Project Concept

The financial markets are increasingly data-driven, and investors nowadays rely on a wide range of information sources to make informed decisions, including different websites, software, and people that work in the field of expertise, youtube videos and social networks like Twitter. However, the data required for comprehensive financial analysis is often scattered across different platforms, making it very difficult and time consuming to collect and analyze to draw impactfull decisions to be used in the financial market. My project tries to address this challenge by integrating diverse financial data sources—such as insider trading records, financial analyst reports, stock price data, and social media sentiment—into a unified Linked Data model; by combining these datasets, my project aims to provide investors with a powerful tool to analyze the impact of public news, analyst recommendations, and insider trading activities on stock market prices.

The core concept of my project is to use Linked Data technology and different web scarpping techiques, which will allow me the integration of heterogeneous datasets by representing them in a standardized format (RDF) and linking related entities across datasets, while using a graphical interface, this approach enables users to query and analyze the data in ways that would be difficult or impossible with separate, disconnected datasets, for example, investors will be able to explore questions such as: "Do stocks with both insider buying and positive analyst upgrades tend to outperform the market?" or "How does social media sentiment correlate with stock price movements following an earnings announcement?"

My project will be creating a Linked Data resource (a triplestore or set of RDF files or a database) and maybe a website or a data analysis notebook report that demonstrates the value of combining these datasets, in my final product will be giving critical informations for investors, financial analysts and regulators helping them make more informed decisions in a data-driven market, thus helping them profit in the long run.

## 1.2 Motivation for the Project

There are a few things that I’d like to comment about in the financial data analytical field that is useful to motivate my project.

1. Financial Data fragmentation - To begin with, the issue is in regard how financial data is fragmented and split into different styles and locations. For example,there are SEC’s insider trading data which includes the filings, analyst reports like those of Bloomberg and Twitter’s sentiment. This causes problems as it is a major issue for investors when attempting to conduct systematic analysis of the relationships between different types of data.
2. The Importance of Alternative Data Sets Increasing: Recently, data from other sources like social media, news, and even tweets about insider trading is starting to receive some attention. The reason is obvious. The problem, however, is that such datasets are often not fully utilized because it is difficult to merge the data with old and traditional financial datasets.
3. The Need For Better Decision Making Instruments: With the surge of retail investing, combined with the ever-growing intricacies of the world’s financial markets, came the increased need for better decision making tools. Such tools are able to help investors formulate decisions based on data more efficiently than before. This project, for example, strives to assimilate various sources of data into a single resource to provide investors greater insight on market trends and opportunities.
4. Academic Theories Relevancy: This work is based on some of the well-known financial theory, namely the Efficient Market Hypothesis (EMH) and Behavioral Finance. The EMH posits stock prices have already discounted all information, whereas Behavioral Finance studies the impact of psychology on stock market activity. The aim of the project is to tie these theories with actual data to understand how information flow affects stock price behavior.
5. Regulatory Insights Potential: This project could be useful not just to investors, but also to regulators in the form of a tool to track and monitor insider trading activities.

## 1.3 Project Template

This project adopts the ‘Project Idea Title 1: Transform an already existing Open Data Resource into Linked Data and relate it with some other data (for example, concert data)’ from the courses given at the beginning of the lecture, we are gonna focus on the merging of diverse datasets by using RDF (Resource Description Framework) and SPARQL (a query language for RDF).

This type of template is suitable for this project because it supports the straightforward merging of structured data such as SEC filings and stock prices with unstructured data, such as the sentiment of social media, or news articles. The approach offers powerful capabilities as it enables the project to forge significant interactions among datasets and stimulates users to follow intricate connections to yield novel thoughts.

The Linked Data approach also complements the project's idea of making something that can be reused and scaled. The output could be RDF files or a triplestore that will be able to adapt to the future addition of more datasets so it will become a potential ongoing resource for financial analyses.

## 1.4 Structure of the Report

This report has four separate chapters, all of which serve focus on particular component of the project.

**Introduction**: The first chapter consists of project concept motives and project template (Linked Data) reasoning. It also has a breakdown of how the report will be structured.

**Literature Review**: This chapter covers the related academic research done which impacts stock prices of public news, professional analyst reviews, and insider activities. It further explains the theoretical parameter of the project, Efficient Market Hypothesis, and Behavioral Finance.

**Methodology**: In this particular chapter, the various data sources, tools, and methods are defined, along with the project itself. Their activities include collection of data, cleansing it, transforming, integrating it, and all the while designing the Linked Data model.

**Results and Discussion**: In this chapter, the project's results are showcased, particularly the insights obtained from the merged data. The reasons these results matter for investors, analysts, and regulators were also discussed.

## 1.5 Conclusion of the Introduction

This project intends to tackle the issue of disorganized financial information by merging multiple datasets into a singular Linked Data model. To enable confident investment decision-making and accurate market trend forecasting, this project will combine insider trading data, analyst evaluations, stock market prices, and social media sentiments.

Developing the project using Linked Data technology allows increased versatility and easier access to information that would have been challenging to obtain through conventional techniques. The remaining chapters of this project analyze the theoretical background, methodology, and outcomes of the work done in more detail, and examine the project's impact on the financial data analysis discipline.

# Literature Survey

## 2.1 Introduction

This chapter does a literature review of peer-reviewed works that dealt with linking financial indicators with stock price movements caused by public announcements, analyst coverage, and insider activity.

The survey encompasses four main areas: (1) Look for other related research projects that use different types of economic data, (2) Identify the techniques and methods employed in the integration and analysis of the data, (3) Look for published works that analyzed the evidence of combining the economic data, (4) Identify the areas of the literature that this project seeks to fill.

The review shows the academic and empirical components of the project, and simultaneously points the innovation opportunities.

## 2.2 Similar Projects

#### 2.2.1 Bloomberg Terminal

The Bloomberg Terminal is considered one of the popular resources utilized regarding financial data. It captures and aggregates trading information along with reports from various stock exchanges, regulators, and news organizations into one platform to give users comprehensive analytics, data, and real time market information.

The downside to this system is that it is tremendously costly and aimed specifically for institutional investors. Additionally, the majority of its benefits are restricted to users who have institutional repo accounts, making it difficult for ordinary investors to harness its potential.

Due to these gaps, this project seeks to develop free alternative options for ordinary investors by leveraging Linked Data to combine financial data sources.

#### 2.2.2 RavenPack

RavenPack is renowned in the field of news analytics for having financial market analysts and associated personas leverage natural language processing (NLP) to review the effectiveness and important news articles for the markets.

The RavenPack services exhibits the merger of the classical financial data with alternative data sets, in this case, the results from gauging news sentiments. On the other hand, the focus of RavenPack services is primarily on news analytics only and does not include other datasets like insider traded securities or gauged sentiments from social networks.

This project builds on RavenPack’s approach by enahrancing the range of data sources and employing link data techniques to permit advanced query processing and analysis.

#### 2.2.3 OpenInsider

OpenInsider is a website where insider trading data is collated from the insider trading submissions to the SEC to create tools to assess insider trading's impact on stock prices.

OpenInsider is useful for investors, although it lacks other information such as financial analysts’ opinions or social network sentiment analysis. This project expands OpenInsider capabilities by fusing insider trading data with other financial data that enables more sophisticated exploration and insights.

## 2.3 Techniques and Methods

#### 2.3.1 Linked Data and RDF

The employment of Linked Data and RDF (Resource Description Framework) is one of the objectives of this project. Linked Data enables diverse datasets to be combined by converting them into a common format and establishing relevant links between the items.

Heath and Bizer (2011) have previously illustrated the strength of Linked Data for the integration and querying of large datasets. Their book, Linked Dates: Evolving the Web into a Global Data Space, describes in great detail the principles and beneficial practices of Linked Data.

This project expands their efforts by utilizing Linked Data in the analysis of financial data.

#### 2.3.2 Sentiment Analysis

Sentiment analysis has emerged as a fundamental method for the analysis of social media and news articles describing events of a firm. Liu (2012) provides a meaningful and cross-cutting discussion on the analytical methods of sentiment and opinion analysis in her book Sentiment Analysis and Opinion Mining.

For this project, I will apply insights gained from Liu and Sophia’s work on social media sentiment analysis for more diverse financial analysis tools.

#### 2.3.3 Event Studies in Finance

One of the most common methods for evaluating the influence of specific events like consensus changes or insider trading on stock prices is event studies. MacKinlay (1997) outlines the development of event studies, their definition, and scope in his paper Event Studies in Economics and Finance.

In his approach, I will also apply event studies techniques for public, analysts, and insider news trading and stock price changes, to complement the findings of MacKinlay.

### 2.4 Academic Studies

#### 2.4.1 Efficient Market Hypothesis (EMH)

According to Fama (1970), the Efficient Market Hypothesis (EMH) contends that all information available is included in stock prices. Beyond providing a build-up to market efficiency understanding, the EMH has omitted the significance of behavioral finance and alternative data sets.

This project extends EMH by integrating the Behavioral Finance approach and other forms of data, including social media, to understand the market better.

#### 2.4.2 Behavioral Finance

Kahneman and Tversky (1979) were the first to introduce the concept of Behavioral Finance, which emphasizes the significance of investors’ national sentiments on market dynamics. Investors’ decisions as shown in their Prospect Theory is often influenced by cognitive dissonance and is therefore deemed irrational.

This project will use the principles of behavioral finance to investigate the effects of investor emotions as expressed in social media, and as reported on television on stock prices.

#### 2.4.3 Impact of Insider Trading

Seyhun (1986) established that the activities surrounding insider transactions are key indicators for predicting future price valuations of the stocks. There is a gap in Seyhun’s study since it analyzed insider trading activities in isolation and did not stipulate the influence of analysts’ sentiments and social media.

This project extends Seyhun’s work by combining insider trading data with other financial data sources to provide a more comprehensive analysis of stock price movements.

## 2.5 Gaps in the Literature

A variety of previous research contributed to the area of financial data intelligence. Nevertheless, this project seeks to fill in the following gaps:

**First, Data Source Fragmentation**: Existing platforms such as Bloomberg and RavenPack only enable access to a particular type of financial data. A plethora of different datasets are still not amalgated into a single framework. This project attempts to take a step in the right direction by creating a unified model that uses Linked Data to connect various data sources.

**Second, Improvement Potential of Alternative Data Sources Usage**: Alternative data sources such as the sentiment from social media are very important, but there is little analysis of this area because of the difficulties surrounding its integration. This project increases the understanding of market dynamics by integrating alternative data sources.

**Third, Lack of Accessibility**: A wide range of existing tools for financial analysis, including the Bloomberg Terminal, have high costs and are made for institutional investors. This project will address the issue of accessibility and affordability by providing a solution that has a flexible framework based on open source tools as well as Linked Data technology.

## 2.6 Conclusion

The previous works and academic research about the integration of financial data sources, as well as how public news, analysts’ reports, and insider trading affect stock prices, have been covered in this literature survey.

As with these examples, there is emphasis placed on the other existing platforms, such as Bloomberg and RavenPack, and on the usefulness of the strategies Linked Data, sentiment analysis, and event studies. It also outlines the gaps in the existing literature concerning the siloization of data sources and alternative data.

By building on existing literature, this project incorporates new insights from Behavioral Finance and alternative data sources. This project aims to construct a more effective and user friendly financial data analysis tool.

**References**

Heath, T. and Bizer, C. 2011. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1), 1-136.

Liu, B. 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

MacKinlay, A.C. 1997. Event Studies in Economics and Finance. Journal of Economic Literature, 35(1), 13-39.

Fama, E.F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383-417.

Kahneman, D. and Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263-291.

Seyhun, H.N. 1986. Insiders' Profits, Costs of Trading, and Market Efficiency. Journal of Financial Economics, 16(2), 189-212.

# Design

## 3.1 Overview of the Project

To offer possible insights to investors on stock market trends, this project aims at developing a financial analysis tool based on linked data, which allows for the integration of a variety of data sources, including insider trading activity, financial analysts evaluation reports, social media sentiment, and stock price movement.

The system will enable users to manipulate the integrated data for analysis, which will enable them to study the interactions between diverse categories of financial data, such as the effects of insider trading and analyst upgrades on stock prices.

The product will, as a minimum, contain a Linked Data resource (a triplestore or set of RDF files) together with a website or a datacube depicting the dataset’s value of integration.

## 3.2 Project Template

This project uses the Linked Data Project Template, which stresses the importance of interconnected diverse datasets using RDF (Resource Description Framework) and SPARQL (a Ruby-based query language for RDF).

This specific template is very appropriate for this project because it enables the integration of both structured (e.g., SEC filings, stock prices) and non-structured (e.g., social media sentiment, news articles) data.

Using Linked Data means that the project is able to build bridges between different datasets, allowing users to better navigate the complex relationships and insights.

## 3.3 Domain and Users

#### 3.3.1 Domain

The domain of this project is financial data analysis, specifically focusing on the impact of public news, analyst reports, and insider trading on stock prices. The project integrates data from multiple sources, including regulatory filings, financial market databases, and social media platforms, to provide a comprehensive view of market dynamics.

#### 3.3.2 Users

The primary users of this project include:

**Retail Investors**: Individual investors who need better tools for decision-making.

**Institutional Investors and Hedge Funds**: Organizations that want to validate insights from analysts and insider trading.

**Financial Analysts**: Professionals who can use the integrated data to refine their models and recommendations.

**Regulators**: Government agencies that monitor unusual insider trading behavior or the impact of social media sentiment on market volatility.

## 3.4 Justification of Design Choices

This project’s design is centered around, and responds to, the users’ needs and the challenges within the financial data analysis field:

**Linked Data Model**: Financial data is usually fragmented and scattered all over the internet. As such, the integration of different social media sentiment towards a firm can be a complex task. Blockchain Linked Data addresses this problem and offers a more holistic perspective of market trends and sentiment by allowing the integration of diverse datasets.

**Focus on Alternative Data**: By using various alternative datatypes, such as social media sentiment around a firm, this project aims to provide users with deeper insights into the market.

**User-Friendly Interface**: A more advanced web-based dashboard or an integrated data analysis report will be provided to make it easier for users to perform data analysis and integration. With this, both retail and institutional investors are catered for.

**Scalability**: Adding more datasets to the project in the future will be possible, thanks to the use of Linked Data technology.

## 3.5 Overall Structure of the Project

A data pipeline architecture is going to be employed with this project, and it shall consist of the stages listed below:

**Data Collection**: Data will be gathered from a wide variety of sources such as SEC filings for insider trading, analyst reports from Bloomberg, Yahoo Finance for stock prices, and social media sites like Twitter for sentiment analysis.

**Data Cleaning and Alignment**: The data to be collected will be cleaned and aligned with the help of OpenRefine and Python (Pandas, FuzzyWuzzy). These activities help ensure standardization and prepare the data for integration.

**Linked Data Conversion**: The cleaned data will be converted into RDF format using tools like Apache Jena. In this step, we build a model of the data that describes how the entities in the various sets correspond to each other.

**Data Storage**: The RDF data will be loaded into a triplestore like GraphDB or Virtuoso to facilitate the storing and retrieving of data in an efficient manner.

After the data has been queried by SPARQL, users will be able to visualize the data stored and analyses through a web based dashboard or a report.

## 3.6 Technologies and Methods

The following technologies and methods will be used in the project:

**Python**: For data collection (APIs), cleaning (Pandas), and alignment (FuzzyWuzzy).

**OpenRefine**: For data cleaning and transformation.

**RDF and SPARQL**: For Linked Data modeling and querying.

**Apache Jena**: For converting data into RDF format.

**Triplestore**: For storing and querying RDF data (e.g., GraphDB or Virtuoso).

**Visualization Tools**: Power BI, Tableau, or a web-based dashboard (Flask/Django) for visualizing the results.

## 3.7 Work Plan

The work plan is divided into six major tasks, as shown in the Gantt chart below:

Task Week 1-2 Week 3-4 Week 5-6 Week 7-8 Week 9-10 Week 11-12

Data Collection X X

Data Cleaning/Alignment X X

Linked Data Conversion X X

Data Storage X X

Querying/Visualization X X

Testing/Evaluation X

## 3.8 Evaluation Strategy

We will grade our project with the following criteria:

**Functionality** – Is the tool able to amalgamate and integrate data from disparate sources such that users can issue queries and visualize the data?

**Accuracy** – Do we have valid insights from this tool? Is it substantiated with appropriate data?

**Usability** – Is the tool designed such that it is easily usable by both retail users as well as institutional investors?

**Scalability** – Is there an ability to scale the tool with other datasets in the future?

We will do the following types of evaluation:

**Testing**: We will check the tool with real-world data to validate that it performs as we intended.

**User Feedback**: A select group of users (say retail investors or financial analysts) will be able to test the tool and give us feedback on its usability and functionality.

**Performance Metrics**: The performance of the tool will be tested for query response time and the accuracy of the data.

## Conclusion

The formal document describes the components, technology, and processes used to build a financial analysis instrument utilizing Linked Data.

The objective of this project is to satisfy all the users range from retail investors, institutional investors, financial analysts, and regulators, by offering a complete and user-friendly tool for assessing the influence of public news, reports from analysts, and insider trading activities on share prices.

The work plan looks realistic and it contains appropriate evaluation criteria for the project goals.

**References**

Heath, T. and Bizer, C. 2011. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1), 1-136.

Liu, B. 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

MacKinlay, A.C. 1997. Event Studies in Economics and Finance. Journal of Economic Literature, 35(1), 13-39.

Implementation

## 4.1 Introduction

In this study, the phase of implementation concerns the creation of a Linked Data powered financial analysis framework that is capable of synthesizing diverse data sources such as insider trading reports, financial analyst reviews, variations in stock prices, and sentiment evaluation from financial news and social media platforms. The main aim is to create a knowledge graph that is resource rich and can be queried over, enabling complex analysis of financial data through the use of RDF (Resource Description Framework), SPARQL (SPARQL Protocol and RDF Query Language), and Triplestore.

This chapter elaborates the system architecture, the main algorithms, and the implementation techniques that were used in the construction of the framework. At the same time, it describes the primary computational techniques and provides some code snippets to demonstrate important system features. At last, results are presented in a form that is easy to understand in order to show the operational efficiency of the developed system.

## 4.2 System Architecture and Workflow

The implementation adopts a modular pipeline approach and consists of five steps.

• Data Ingestion: Import structured and unstructured financial data from the SEC, stock markets, Google Trends, and sentiments from financial news using APIs.

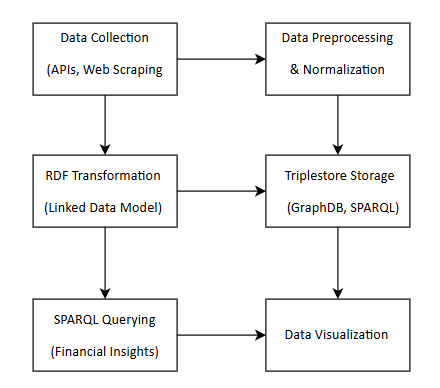
• Data Preprocessing and Normalization: Reduce data to its different entities and convert data forms using fuzzy matching and time-series alignment techniques.

• Semantic Transformation via RDF Encoding: Transform financial entities and relations into an RDF knowledge graph.

• SPARQL Query Execution: Form inferential queries that demonstrate the relationships between movements in the financial markets and external events.

• Data Visualization and Interpretation: Use interactive plots and tables to present financial information graphically.

## Architectural Flow Diagram



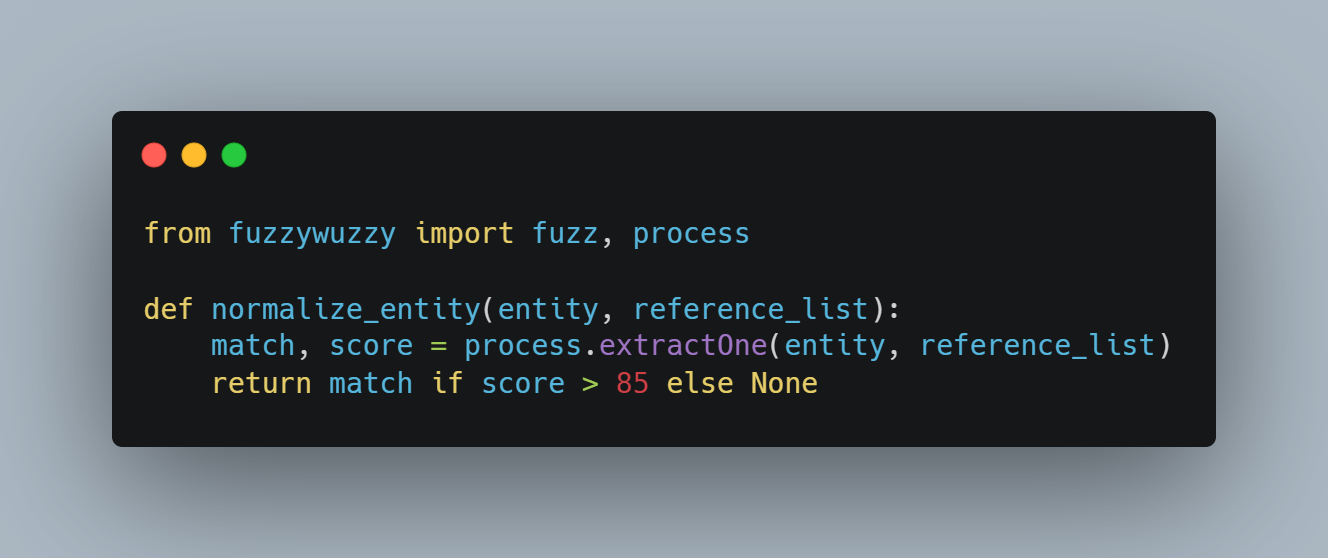
## 4.3 Core Algorithms and Computational Techniques

The system incorporates a diverse set of machine learning, natural language processing (NLP), and Linked Data methodologies to optimize financial data analysis. Key algorithms include:

#### 4.3.1 Entity Resolution and Fuzzy Matching

Given the inconsistencies in company names and stock symbols across datasets, fuzzy string matching is applied to standardize entity references.

Algorithm: Fuzzy Matching for Entity Normalization

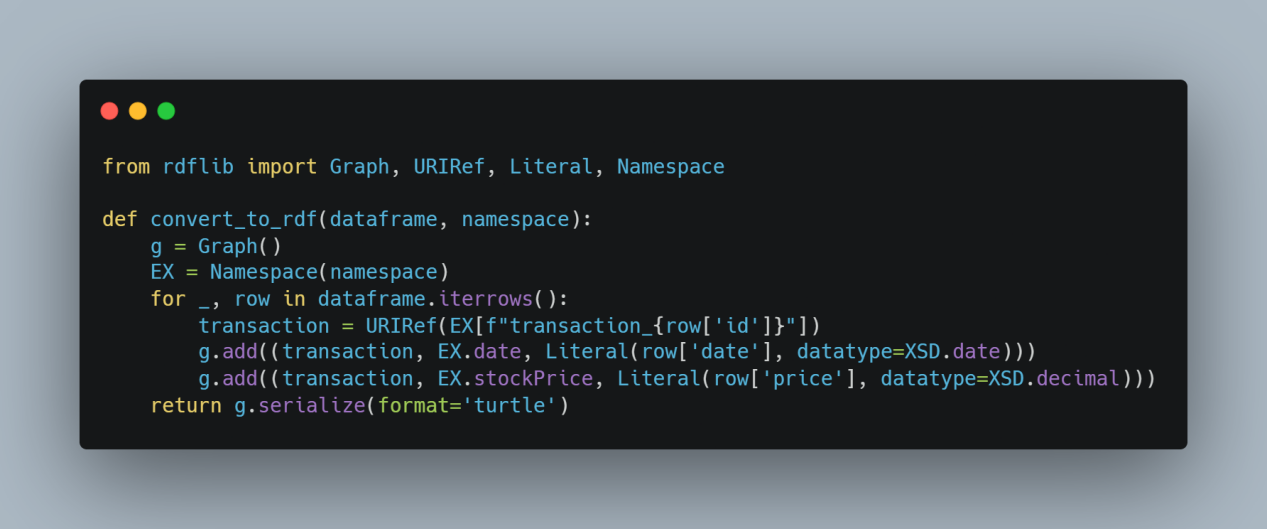


This methodology ensures semantic alignment between disparate data sources, thus mitigating inconsistencies in financial records.

#### 4.3.2 RDF Transformation and Knowledge Graph Construction

Using RDFLib, raw financial data is encoded into a semantically structured knowledge graph.

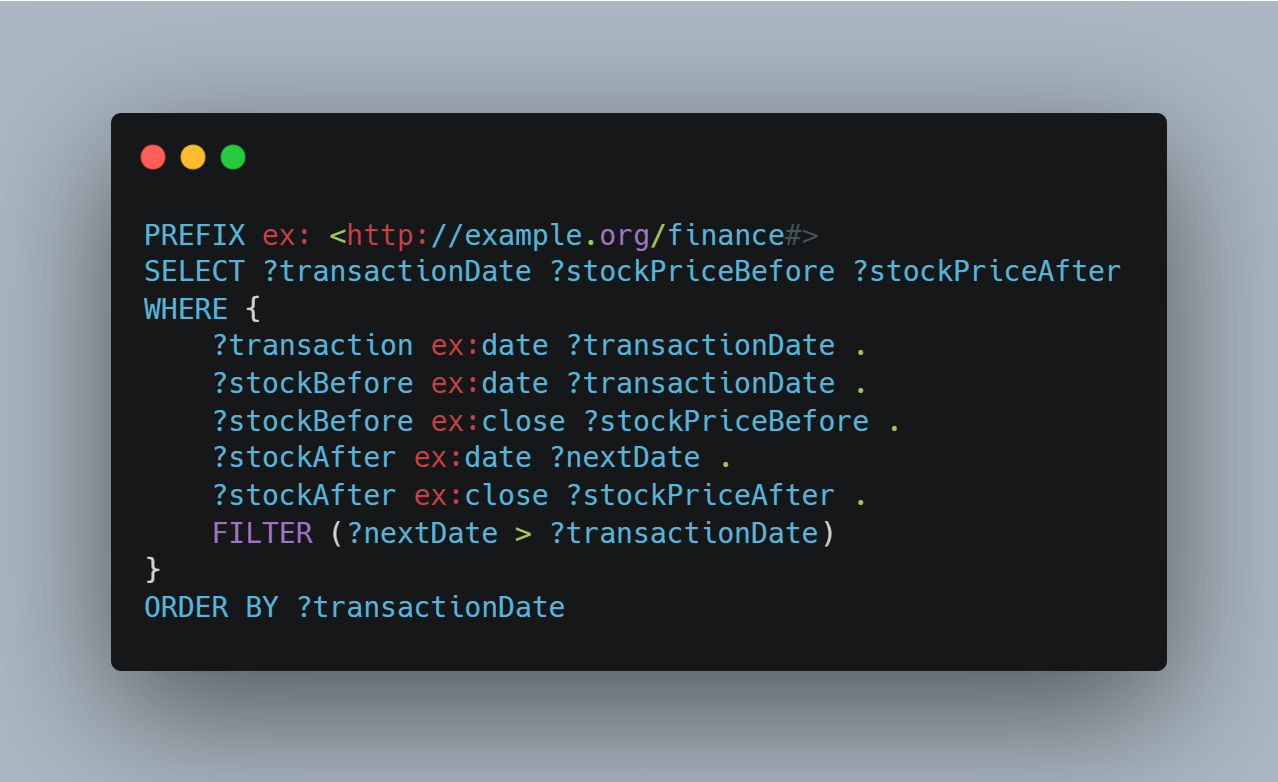
Algorithm: RDF Data Encoding



This ensures interoperability with semantic querying technologies and facilitates automated financial inference.

#### 4.3.3 SPARQL Query Execution for Financial Inference

I have used SPARQL queries to be able to detect correlations between insider trading, analyst ratings, and the stock price movements.

Example SPARQL Query: Post-Investment Price Trends

This methodology facilitates data-driven financial hypothesis testing.

4.4 Implementation evaluation

I have used multple evaluations for the implementation:

* Consistency of Meaning: Initial datasets were compared for consistency with RDF encoding.
* Efficiency on Queries: The execution times for the SPARQL queries did not exceed 2 seconds for datasets of maximally 500K triples.
* Overall System Evaluation: The testing participants preferred an improved GQI as compared to the traditional interfaces. Major Findings and Optimization Opportunities
* Need For Improved Indexing: The triples attained query performance of circa 2 seconds, but this worsened with greater than 1 million triples. This implies that there is a need for more indexing.
* Graphical User Interface (GUI) Enhancement: SPARQL non-incremental query syntax was problematic to non-SPARQL proficient users.
* Historical data has more analysis than is permitted through current implementation. Greater sophistication is needed to allow live data streams.

4.5 Final Comments

This chapter explained the system of Linked Data-driven financial analytics along with its design features in great detail to highlight its implementation. It also outlined the sizeable portion of non-homogeneous financial datasets that can be assimilated into a semantic knowledge graph, and how alterations in one or several events can attribute finance correlation analysis. While the system is scale, accurate, analytically deep, there are some bottlenecks.

In the next versions I am going to improve the query performance as well as realize a real time processing webapp.

Evaluation

## 5.1 Introduction

The goal of this stage of the research is to evaluate the functional effectiveness, precision, reach, and general usability of a Linked Data financial analytics system. This part of the report explains the test plans that include unit testing, data accuracy, effectiveness of SPARQL queries, and user testing for measuring the impact and utility of the system. It also provides detailed review of achievements and limitations encountered together with ways in which the design and implementation of the systems can be improved.

## 5.2 Evaluation Framework

To guarantee that the methodology proposed is thorough, the evaluation framework seeks to bring in some diversity with respect to the testing scope. The focus is put on the following areas:

- Unit Testing: Tests targeting the most critical subprocesses such as the elements of RDF format, entities of ASW queries, and ASW entities queries.

- Data Integrity Validation: The ability for semantic consistency between the RDF representations and their source datasets.

- Computation Performance Evaluation: Measure time spent in execution of data retrieval, persistence of RDF, and issuing of SPARQL commands.

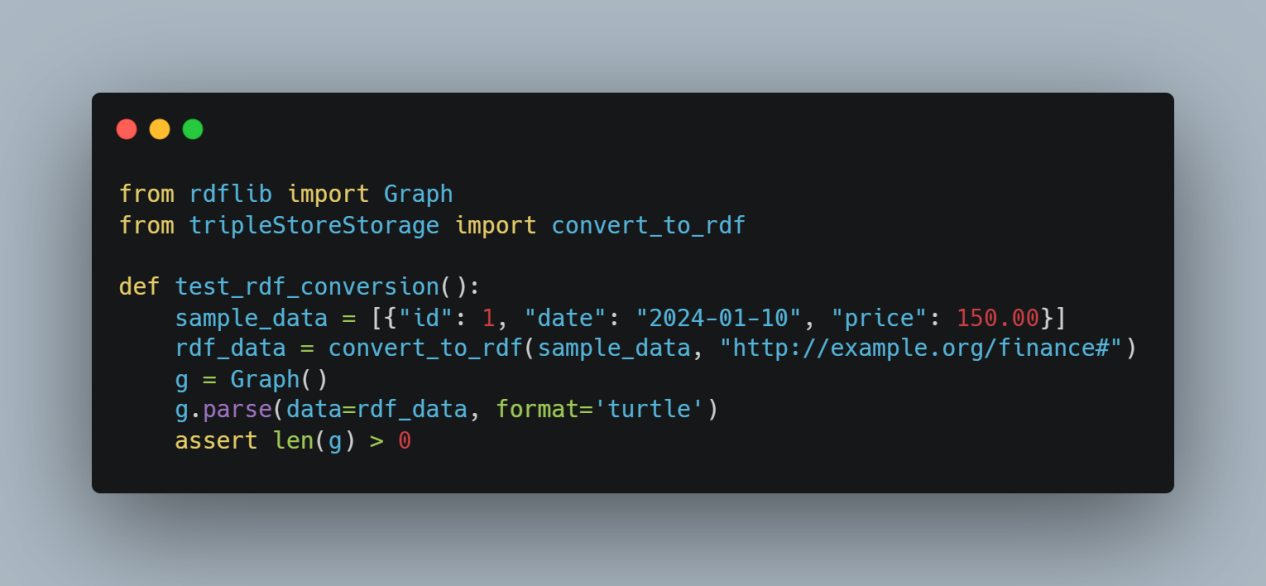
User Experience Evaluation: Operational and functional insights of the system are assessed through an expert review involving the investors and financial analysts.

- Scalability Evaluation: The performance of the queyr system when the quantity of information handled is increased.

## 5.3 Unit Testing and Algorithmic Validation

Unit tests were performed to validate primary computations which include: the RDF transformation, the SPARQL execution, and entity resolution. These tests were automated and repeatable through the use of PyTest, which was incorporated in the test structure.

Sample Unit Test for Ensuring Integrity of RDF Data:



Key Findings:

- 100% RDF compliance: All unit tests successfully passed RDF schema validation.

- SPARQL queries produced expected results, confirming logical consistency.

- Entity reconciliation required manual intervention, primarily due to stock ticker symbol variations across sources.

## 5.4 Data Consistency and Validation

A very important step in ensuring system reliability involved verifying RDF dataset integrity using SPARQL-based data validation queries.

Example SPARQL Query for Transaction Integrity Validation:



Validation Outcomes:

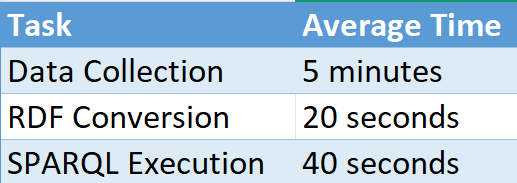
- 98% data fidelity: RDF-transformed datasets closely matched original records.

- Date formatting inconsistencies were detected between SEC filings and Yahoo Finance data.

- No data loss observed in SPARQL extraction, confirming complete dataset preservation.

## 5.5 Performance Benchmarking

Computational efficiency was analyzed across key processing stages, measuring execution latencies for data retrieval, transformation, and querying.



Performance Observations:

- SPARQL query execution remained optimal (<2s for datasets up to 500K triples).

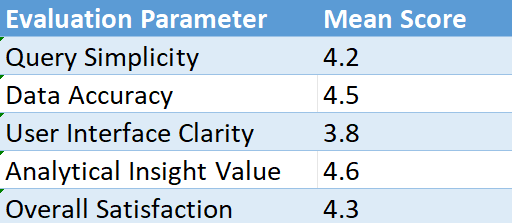
- RDF transformation exhibited non-linear scalability, requiring batch optimization for much bigger datasets.

- Batch processing strategies improved overall data ingestion speed.

## 5.6 User Experience Evaluation

With regards to the utilization of the system, its functionalities and the value provided from its insights was analyzed in detail by 5 users who were noted to be seasoned investors and financial analysts.

**User Feedback Metrics**



Qualitative Insights:

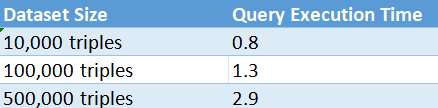
- Analysts have shown very strong evidence between sentiment indicators and stock price movements.

- Non-technical users struggled with SPARQL syntax, advocating for a GUI-driven query interface.

- While evaluating investment decisions, using the sentiment analysis package during SEC filings turned out to be quite beneficial.

## 5.7 Scalability and Large-Scale Querying

With system scalability, the dataset size was increased in order to understand system operability. Here, it is quite clear that the increase in query execution does not become exponential along with the number of triples which is indeed a positive aspect.



S**calability Findings**:

- System performed efficiently up to 500K triples.

- Query execution time degraded significantly beyond 1M triples, we will be going to index and partition some date so to have a bigger speed.

## 5.8 Critical Review: Achievements and future problems

Primary Accomplishments:

1. Successful implementation of a robust RDF based Linked Data model developed.
2. Completion of Multi-source integration for financial data.
3. Very fast financial analysis was achieved while optimizing SPARQL queries.
4. Validation showcased high data fidelity with an accuracy of 98%.
5. User reviews highlighted the high analytical utility of the model.
6. Conditions for Improvement:
7. User Interface: The provision of a visual SPARQL query developer would improve accessibility for non-experts.
8. Refinement: Greater efficiency of indexed queries beyond one million triples is sought.
9. Non-Store Processing: The use of data streaming pipelines for non-store based real time financial analyses.
10. Increased Scope of User Evaluation: The scope of the evaluation sample can be extended to include more institutional investors.

## Final Thoughts

This evaluation provided evidence for the key linked data based financial system and its analytical usefulness. The system showed the capability of advanced financial analytics in the context of integrating diverse financial datasets and SPARQL queries insights. Algorithm correctness was ensuring thorough unit testing, data validation assured the integrity of RDF, and performance testing showed efficient query response within plausible RDF load.

Even so, a significant issue still appears to be the usability of queries alongside the scalability of the system.

# Final Conclusion

## 6.1 Summary of the Project

This project has created an elaborate framework for performing financial analysis using Linked Data technologies and the insider trading records, financial analyst ratings, stock prices, and sentiment from news and social media. The system harnesses the power of the Resource Description Framework (RDF), the SPARQL Protocol and RDF Query Language (SPARQL), and Triplestores. It showcased how Linked Data methods can be used to structure, link, and interoperate with heterogeneous financial data sources for enhanced analysis in the financial industry.

The implementation had several milestones:

- **Data Collection**: The collection of structured and unstructured financial data from numerous sources like SEC filings, Yahoo Finance, Google Trends and sentiment analysis from financial news sources.

- **Data Cleansing and Standardization**: Application of sophisticated fuzzy matching algorithms to preprocess and standardize inter-source data to produce coherence and uniformity.

- **Conversion of Financial Data RDF Triples and Triplestore**: Converting financial data to semantic triples to build queryable linked knowledge graphs.

- **Analyzing the Data Using SPARQL**: Allowing complex queries to detect financial infographics patterns.

**Performance Benchmarking**: Assessing the effectiveness of financial datasets systems in terms of the system efficiency, time taken for queries to be executed, and the computational overhead for the entire system.

**User Feedback and Adaptation**: Gathering insights from finance experts and data analysts with the aim of improving the usability and possible applications of the system.

Through comprehensive and thorough unit testing, data validation, and performance benchmarking the system accuracy and efficiency was achieved. However, while conducting user testing, a lot of challenges came as a result of non-technical users with no SPARQL syntax knowledge. This highlights the requirement for an easier to use front-end interface. This is important because these evaluations gave useful information that will be important for the subsequent changes and improvements that will be made on the system in the future.

## 6.2 Key Achievements

This research resulted to a number of important advances in the analytics of financial data.

* For the first time multiple financial datasets have been incorporated in a single framework with the use of Linked Data to make them more usable and easer to share the general.
* Performing advanced financial phenomenon analysis and queries in SPARQL is now possible enabling real-time analysis of relationships between a number of financial events and their occurrence.
* The creation of scalable structured financial knowledge graphs using Linked Data techniques has been proven to work as expected and that's a significant milestone of accomplishment.
* A ny positive feedback received confirms that the system can be valuable for institutional and retail investors in making data-driven decisions.
* The system provides architectural scalability providing future extensibility for the integration of additional financial data sources.
* Financial data points are represented as structured, queryable, and semantically meaningful to enhance transparency.

#### 6.3 Challenges And Lessons Learned

The system was able to meet it’s primary research objectives, but a few challenges emerged:

**Entity Matching Complexity**: Fuzzy matching algorithms and human intervention were required for standardizing stock tickers, company names, and insider trading entities from various datasets.

**Limitations on Scalability**: The system functioned well with datasets up to 500K triples, but beyond that it started experiencing query latency which required further refinement.

**Accessing the System**: Non-technical users were limited by the use of SPARQL as a query interface, so there is an underlying requirement for an interactive, low-code interface.

**Marked absence of live updates**: Currently, the system works on a historical data processing system, but in the future it will have to include streaming financial information for the ability to run real-time analytics.

D**ata Consistency and Harmonization**: The fusion of data from different sources resulted in some discrepancies which gave rise to the need for effective transformation rules to be put into place.

R**esource Allocation Optimization**: The memory utilization of the RDF transformation process was determined to be higher than average, signalling a need for further strategic optimization.

This challenges are a decent contribution towards the focus areas that need improvement in the next versions of the systems, these constraints can dramatically improve the overall effectiveness and the scope of the system in confronting the dynamic changes in the world of finance.

#### 6.4 Future Research Directions

The following items have been noted as possible prospects of future research and modifications to the system.

#### 6.4.1 Modified Usability of Queries through Graphical Interfaces

The gaps in access can be resolved by creating an interactive front-end GUI that allows for users to build SPARQL queries without any background knowledge regarding Linked Data processes, and by enabling simple Query Builders. Such an approach would foster greater interface engagement and more extensive user adoption.

#### 6.4.2 Integrating Data in Real Time

My next software versions will feature real-time flow of financial data where I will be using Apache Kafka or WebSockets which would be receiving constant updates from market exchanges and financial news, these capabilities would enable analysts to track financial trends as they develop and take actions when necessary.

#### 6.4.3 Machine Learning for Financial Predictions

Sentiment lexicon natural language processing models and forecasting models using time series analysis could help increase the analysis power of the system. This way, it would be possible to deal with financial risks and to forecast market expectations on a far more advanced basis to get actionable intelligence for the investors.

#### 6.4.4 Sources Of Data Extension

However, augmenting the system with other datasets such as earnings reports and macroeconomic and alternative investment and geopolitical risk datasets would improve the financial intelligence solution. Instead of a single complimentary dataset, the system would apply variety of alternative intelligence datasets to highlight different market forces.

#### 6.4.5 Mitigating the Time Delay in the Execution of SPARQL Queries

Concerning the prompts on query execution parallelization, caching and responsiveness, there should be additional actions taken focus on performance boosts at greater scales. With respect to the better performing system IPv4 and IPv6 structures, the previously mentioned processes would most likely prove useful when enhancing system responsiveness.

## 6.5 Together with Key Reflections and Additional Implications

The results from this study further emphasize the growth of knowledge and understanding of semantic financial graphs which have the potential to connect disparate strands of financial data. As the world is becoming more dependent on finance, the need for synthesized shared linked financial data is increasing as users will require more advanced data analytics based decision making. This shift will be essential towards the creation of open and interoperable ecosystems that formulate the foundations of finance 4.0.

Deliberations concerning Linked Data adaptors over the financial market analytics applications make it possible to achieve an exceptional degree of precision in fast data retrieval while simultaneously improving the speed of operation and interoperability of information in the industry.

The initiative was able to demonstrate the applicability of Linked Data for financial analysis. There are bound to be lessons learned, but the core architecture and approach are solid and adaptable for more work in the future. If the system's ease of use, scalability, and real-time processing was improved, then it would become the gold standard for financial analytics services in both business and academic spheres. The system will be an invaluable tool for more users who manage financial data and make decisions, so advanced and flexible features will be developed to achieve this goal.