MIT 805 Assignment 1 Part 1
Big Data Collection and Analysis

Muphulusi Dzivhani
Student Number: u18069682
Email: u18069682@tuks.co.za

Abstract

This report details the selection and analysis of a large-scale Stock Exchange dataset for MIT 805 Big Data, sourced from Kaggle. The dataset contains historical daily price and volume data for all U.S. stocks and ETFs. The analysis is structured around the Core 3 Vs of Big Data (Volume, Velocity, Variety), the principle of Value, and four additional Vs (Veracity, Variability, Visualization, and Validity) as required by the assignment brief. The report describes the dataset's characteristics, evaluates its challenges and opportunities through this framework, and discusses the significant potential business value that can be extracted through subsequent processing with Big Data tools like Hadoop MapReduce.

1 Introduction

The global financial sector is a perfect generator of Big Data, with stock exchanges producing terabytes of data daily from trading activities, news feeds, and transactional records. Analyzing this data is critical for investors, economists, and regulators to discern trends, manage risk, and foster market stability. This report presents the collection and preliminary analysis of a historical stock market dataset [1], selected from the Stock Exchange domain as per the assignment guidelines. The dataset is evaluated through the structured lens of the 7 Vs of Big Data, beginning with the core three, then value, and finally four other selected Vs to establish its suitability for large-scale analysis and to outline the potential insights that can be derived, fulfilling the requirements for Assignment 1 of MIT 805.

2 Dataset Description

The selected dataset offers a comprehensive record of trading activity, ideal for longitudinal market analysis.

- **Domain:** Stock Exchange
- **Source:** Kaggle (Contributed by Boris Marjanovic) [1].
- **Link:** https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs
- **Content:** Historical daily Open, High, Low, Close (OHLC) prices and trading Volume for individual stocks and ETFs.
- **Size:** The dataset is approximately 15GB, meeting the project requirement of being between 10GB and 40GB for meaningful Big Data analysis.
- **Format:** Comma-Separated Values (CSV), a standard, structured format for tabular data.
- **Scope:** Encompasses tickers from major U.S. exchanges including NYSE, NASDAQ, and AMEX.
- **Time Period:** Data spans multiple decades, providing a rich history of market performance through various economic cycles.

3 Big Data Analysis : The 7 Vs

The following analysis evaluates the dataset against the Core 3 Vs (Volume, Velocity, Variety), the principle of Value, and four additional Vs (Veracity, Variability, Visualization, and Validity) as required by the assignment brief.

3.1 Core 3 Vs

3.1.1 Volume

The **Volume** of this dataset is substantial. It contains millions of records across thousands of unique ticker symbols, with each record containing multiple price points and volume figures. The size of approximately 15GB qualifies it as a large-volume dataset that is impractical to process with traditional tools like Microsoft Excel and necessitates distributed processing frameworks like Apache Hadoop [2].

3.1.2 Velocity

While this specific dataset is a historical batch download, it represents the aggregation of extremely high-**Velocity** data. In a live environment, stock market data is generated in real-time, with ticks occurring in milliseconds. Analyzing historical data at this daily granularity is the first step towards understanding systems capable of handling near-real-time data streams for algorithmic trading.

3.1.3 Variety

The dataset exhibits **Variety** in its composition. The core data is strictly structured. However, its true analytical power is unlocked through integration with other data sources, introducing semi-structured or unstructured data, thereby increasing its variety.

3.2 Value

The **Value** derived from this data is the primary justification for its analysis. Potential values include:

- **Predictive Analytics:** Building models to forecast stock price movements.
- **Risk Management:** Identifying volatile stocks and constructing diversified portfolios to mitigate risk.
- **Algorithmic Trading Backtesting:** Developing and validating automated trading strategies against historical data.
- **Economic Research:** Studying market responses to macroeconomic events.

3.3 Additional Vs

3.3.1 Veracity

Veracity refers to the data's trustworthiness and quality. While sourced from a reputable platform (Kaggle) and ultimately from reliable financial feeds, data quality is not guaranteed. Issues such as missing days , adjustment for stock splits, and possible outliers must be addressed through rigorous data cleaning and validation processes to ensure the accuracy of subsequent analysis.

3.3.2 Variability

Market data is highly **Variable**. Its meaning and the patterns within can change rapidly based on external factors like economic indicators, geopolitical events, and corporate announcements. The rate of data flow and its significance are inconsistent, peaking during market hours and around major news events, requiring robust systems to handle these fluctuations in context and volume.

3.3.3 Visualization

Visualization is crucial for interpreting the complex relationships within this data. Effective visualizations will be key in Assignment 2 and include:

- **Time-series Line Charts:** For tracking individual stock performance over time.
- **Candlestick Charts:** For detailed analysis of daily price movements.
- **Heatmaps:** For visualizing correlation matrices between different stocks or sectors.
- **Interactive Dashboards:** For exploring sector performance and portfolio composition.

3.3.4 Validity

Validity concerns whether the data is correct, accurate, and meaningful for its intended use. For stock analysis, each record must validly represent a true trading event on the specified date. Ensuring validity involves checking for anomalies like negative prices or volumes, or prices that deviate significantly from typical ranges, which could indicate data corruption and would need to be addressed before analysis.

4 Business and Societal Insights

The analysis of this dataset can yield actionable intelligence for various stakeholders:

- **Investors & Traders:** To identify trends, forecast potential price movements, and backtest trading strategies to improve returns.
- **Financial Institutions:** To enhance risk management frameworks, detect anomalous trading patterns indicative of fraud, and optimize asset allocation for clients.
- **Policy Makers & Regulators:** To monitor the market for systemic risks, understand the impact of economic policies, and ensure overall market stability and transparency.
- **Researchers & Academics:** To study market efficiency, behavioral economics, and the impact of specific events on market dynamics.

5 Conclusion

The selected historical US stocks and ETFs dataset is a strong candidate for Big Data analysis due to its significant volume, the inherent velocity of the domain it represents, and the clear potential for extracting high value. The analysis through the 7 Vs framework : the core three, value, and four others, highlights both its opportunities and the challenges related to its veracity, variability, and validity that must be managed. This foundational analysis sets the stage for Assignment 2, which will involve processing this dataset using Hadoop MapReduce to extract meaningful insights and creating visualizations to interpret the results effectively.

Appendix

- GitHub Repository: https://github.com/18069682/MIT805-big-data-stock-analysis
- The repository contains this report, the dataset description, and initial project setup. Code for MapReduce jobs and visualization will be added for Assignment 2.

References

[1] B. Marjanovic, "Price and Volume Data for All US Stocks & ETFs," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

[2] Apache Software Foundation, "Apache Hadoop," 2024. [Online]. Available: https://hadoop.apache.org/

[3] "Yahoo Finance," Yahoo Inc., 2024. [Online]. Available: https://finance.yahoo.com/

[4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.