Big Data Analysis of S&P 500 Historical Stock Trends Using Spark and MongoDB

Team Members:

Peng Ren – 2304917

Zhenliang Hao – 2304918

1. Introduction

This project analyzes the historical performance of S&P 500 companies using Apache Spark and MongoDB. The dataset consists of daily stock prices of S&P 500 companies from their listing dates to 2023, sourced from Kaggle. This data provides a rich foundation for examining annual performance, volatility, and market behavior during financial crises.

Given the size and complexity of the dataset, Spark is employed for distributed data processing, while MongoDB is used for flexible document-based storage and future retrieval.

2. Methodology

Dataset Details:

- **Source:** Kaggle
- **Format:** CSV, converted to structured documents in MongoDB

- **Size:** ~1 million rows, 10+ columns per row

Data Preprocessing:

- Converted date column to timestamp format
- Removed rows with null "close" or "ticker" values
- Extracted year from the date for grouping and filtering

System Architecture:

- **MongoDB Atlas**: Used for storing the raw and cleaned data
- **PyMongo**: Used to insert and retrieve data due to connector issues on Colab
- **Apache Spark (PySpark)**: Used for transformations, aggregations, and statistical computations

Schema & Indexing:

- Documents structured as: {ticker, date, close, year}
- Indexing managed by MongoDB Atlas defaults

3. Implementation

MongoDB Integration:

- CSV uploaded into Colab and inserted into MongoDB Atlas via PyMongo
- Read back into Colab and converted to Spark DataFrame after removing ObjectId

Spark Processing:

- Calculated annual price change percentage using row_number() window function
- Computed standard deviation (volatility) for each stock
- Isolated performance during crisis years (2008 and 2020)

Key Spark Functions Used:

- groupBy, agg, row_number, last, stddev, to_date, year, join

4. Results & Insights

Annual Performance:

- Top and bottom performing stocks by year were identified
- Data exported as CSV for dashboarding or further use

Volatility Ranking:

- Stocks ranked by standard deviation of close price
- Highlights most and least volatile equities

Financial Crisis Analysis:

- Performance drops in 2008 and 2020 compared side by side
- Visualized using horizontal bar charts

Visualizations:

- Line chart for overall S&P 500 trend
- Bar charts for year-by-year performance
- Side-by-side comparison of crisis years

5. Conclusion & Future Work

Challenges Faced:

- MongoDB Spark Connector was incompatible with Google Colab
- Resolved by using PyMongo and converting to Spark DataFrame manually

Future Improvements:

- Automate entire ETL and visualization in a dashboard
- Integrate with real-time stock data APIs
- Expand analysis by adding sector classification and market cap data

This project demonstrates an end-to-end big data pipeline using Apache Spark and MongoDB for stock market analysis. The methodology and tooling are scalable and transferable to other financial datasets.