**Introduction and Objective**

This project demonstrates the application of cloud computing using Spark and Python to perform big data analytics and analysis on financial data. It processes large stock and (Exchange Traded Fund) ETF datasets, calculates financial metrics like Sharpe ratios, volatility, and moving averages, and visualizes trends like stock correlations and technical indicators.

**Why is this project important?**

The cloud has allowed trading desks all over the world to access the same data at the concurrently and perform their analysis. As financial data is continuously accumulating and growing in volume, the scalability of the cloud will allow smaller trading firms to compete in the industry without having to invest in physical infrastructure.

**What objective and features does this project have?**

The objective of this project is to demonstrate that large financial datasets can be hosted in a cloud environment and analysis and processing can be performed by a user from a remote machine to gather insights on financial instruments like stocks and ETFs and improve the operations of a trading desk.  
Key features of this project include:

1. **Cloud Integration:** This project is hosted on the Google Cloud Platform and uses HDFS to store the data in a distributed file system.
2. **Data Preprocessing & Cleaning:** Removing noise, handling missing data, and aggregating large datasets.
3. **Querying & Analysis:** Performing advanced queries such as Sharpe ratio computation, volatility during market crashes, price breakout analysis, and moving average crossovers.
4. **Visualization:** Generating insightful graphs such as correlation heatmaps, Sharpe ratio trends, and price/volatility trends.

**Why doesn’t traditional computing solve the problem well?**  
Traditional Computing struggles with the parallel processing and distributed storage needed for efficient querying and visualization of large datasets like those in the stock market. Traditional computing also makes collaboration more difficult as the dataset wouldn’t be shared between users making maintaining and updating the data more complex.

**How do cloud computing technologies fit into this project?**  
By leveraging the cloud, trading firms can operate in a leaner fashion by not maintaining expensive physical infrastructure which also enables smaller trading firms to better compete in the financial industry.

Since employees don’t need to be on site, trading firms can also expand their horizons by finding traders who reside in other cities or countries and would work remotely from home and open offices in other jurisdictions.

**Technical Solutions**

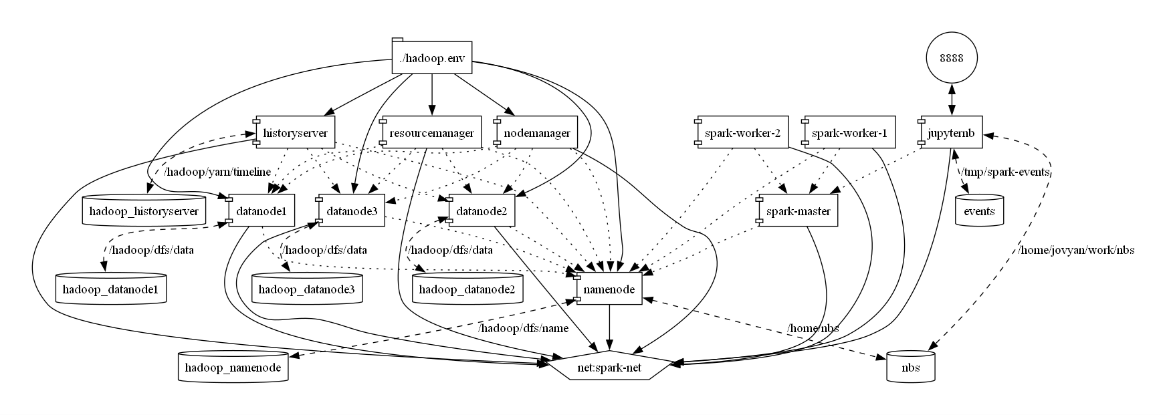
This project leverages cloud and container technologies to process and analyze large datasets. It uses Google Cloud Platform with an Ubuntu instance (8 CPU cores, 16GB RAM) to host and process data. Apache Spark is used to distribute the data processing and complex queries. Docker and Docker Compose containerize services like Spark, HDFS, and the frontend Jupyter Notebook. The financial dataset, sourced from [Kaggle](https://www.kaggle.com/datasets/jacksoncrow/stock-market-dataset)., contains stock market data for the NASDAQ. Stock Market Data for the NASDAQ.

**Cost Estimation of Cloud Resources**  
Using the Google Cloud Platform Cost Estimater with one N1 instance with 6 vCPUs and 12 GiB of RAM. The server would also be running continuously as traders from different time zones would want to access the data set at different times of the day. (*Google Cloud Pricing Calculator*, n.d.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| service\_display\_name | name | quantity | region | Total Price |
| Compute Engine | Custom Instance Core running in Americas | 182.5 | us-central1 | $145.38 |
| Compute Engine | Custom Instance Ram running in Americas | 365 | us-central1 | $38.95 |
| Compute Engine | Storage PD Capacity | 20 | us-central1 | 0 |
| Cloud Storage | Standard Storage US Regional | 20 | us-central1 | $0.30 |
|  |  |  | Total | $184.63 |

**Architecture Design (1 mark)**

The project architecture uses a containerized setup with Docker, Docker-Compose and Spark for distributed data processing running on a Google Cloud Instance. Stock and ETF data is stored in HDFS, a distributed file system. Spark handles the queries like calculating moving averages, volatility, and Sharpe ratios. Processed data is then visualized using Pandas and Matplotlib.

Top of Form

Hutchinson, J. (2020, November 16). *Trading system functionality made possible by the cloud – Insights from Trading Technologies | AWS for Industries*. Aws.amazon.com. https://aws.amazon.com/blogs/industries/trading-system-functionality-made-possible-by-the-cloud-insights-from-trading-technologies/

*Google Cloud Pricing Calculator*. (n.d.). Cloud.google.com. https://cloud.google.com/products/calculator?hl=enBottom of Form