Big Data Application Development (CSCI-GA.2437)

Project Proposal

Team members **: Samvid Zare (sz3369), Yulu Qin (yq810) , Harsh Dubey (hd2225)**

Project idea:

The project idea is to derive features from the raw financial data associated with financial assets such as stocks and crypto currencies for U.S. markets. The features derived from the data can then be fed into Machine Learning models in the domain such as time series forecasting to predict stock prices in future or to build a custom performance dashboard displaying insights about a particular asset.

Datasets:

The project will use the following three datasets to obtain the data in the raw format.

1. **Historical stocks indicators daily data [Harsh]:** This dataset will contain historical pricing and indicator (numerical attributes for stocks indicating its price movement) data for multiple stocks and cryptos with daily frequency. There are many sources to obtain such data, some of the sources are as follows:
    a. Yahoo finance API: Yahoo Finance can provide daily stock pricing for past few years (up to 5 years of historical data) . The data can be fetched using python library for yahoo finance. It's open source and freely available.
    b. Stooq platform can be used to fetch freely available historical stock data. The data for U.S. markets can be directly downloaded from the website.
    c. Official Nasdaq website for the stocks data: Free historical stock data for stocks listed under Nasdaq is available at Nasdaq's website. We can download and store the data using web crawling scripts.

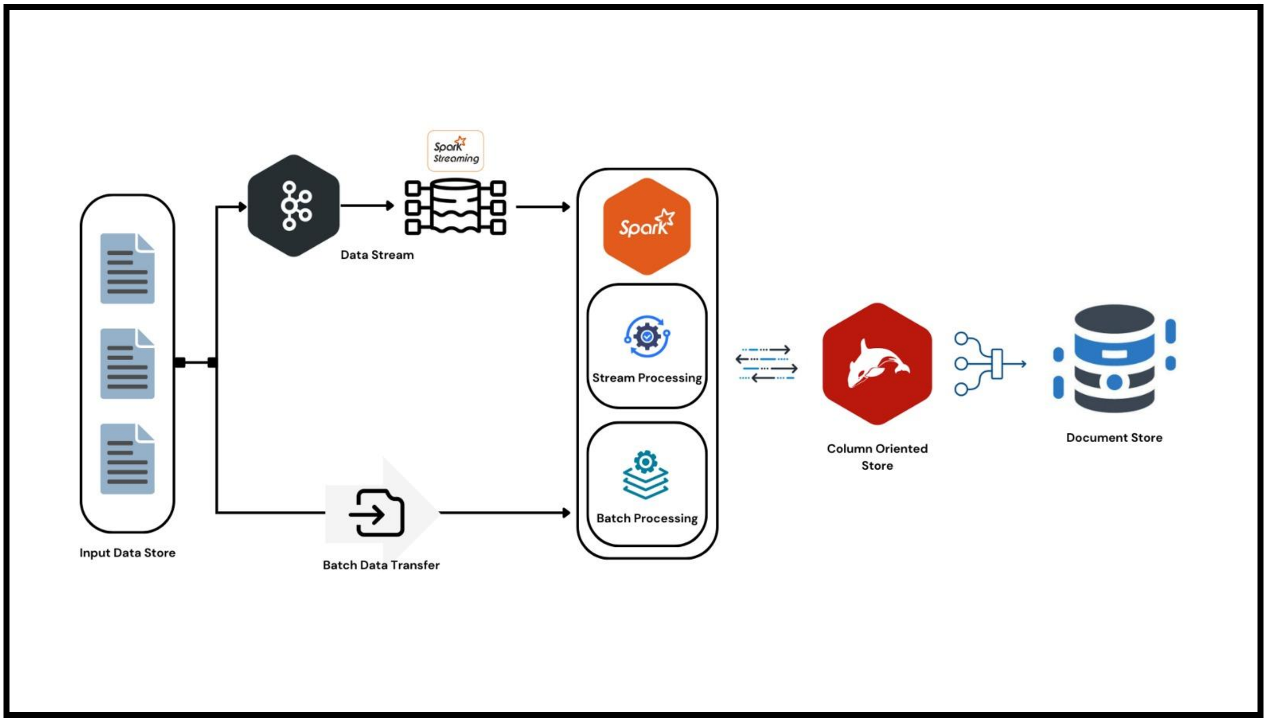
Above datasets can be used in combination with each other with appropriate data cleaning and merging methods.

2. **Public conversations about stocks data [Yulu]:** This dataset will include discussions and comments about multiple stock/cryptos in the public domain. The data resources may include:
    a. Yahoo Finance communications: This includes users and experts' comments for stocks. The data can be fetched from the Python based yahoo finance API.
    b. Twitter discussions (Tweet/Re-tweets) is an alternative resource. The Twitter API is provided to fetch comments and tweets.

3. **Company financial details data [Samvid]:** This dataset will contain publicly available financial data of companies listed in U.S. stock exchanges. Such data may include revenue, stock dividends and additional details on a per-quarter basis. Such data can be obtained from the following sources:
    a. Yahoo finance API: Yahoo Finance can provide company financial data at one place on their website under the financials tab. The data can be fetched using python library for yahoo finance. It's open source and freely available.

b. Sec website : Such open financial data can be obtained from the U.S. Securities and Exchange commission’s website as well.

Architecture:


Spark Streaming
Data Stream
Spark
Stream Processing
Batch Processing
Column Oriented Store
Document Store
Input Data Store
Batch Data Transfer