Distributed Financial Risk Assessment

Sujana Maithili Chindam (Group 7)

Introduction

This report provides overview of data profiling, cleaning, and ingestion for Distributed Financial Risk Assessment.

Data Source

This analysis focuses on stocks listed on the Hong Kong Stock Exchange (HKEX). The list of companies and their tickers was obtained from the HKEX website.

Once the tickers were collected, historical stock data from 01 January 2000 to 31 October 2024 was downloaded using Yahoo Finance. The workflow for downloading the data is outlined below:

```
# Create a directory to store the stock data
mkdir stocks

# Iterate through each ticker and download its historical data
while read -r line; do
    TICKER=$line python3 download-ticker.py
    sleep 1
done < ticker.txt
```

The data was retrieved using the `yfinance` Python library.

```
import os
import yfinance as yf

ticker = os.getenv("TICKER")

start_date = "2000-01-01"
end_date = "2024-10-31"

data = yf.download(ticker, start=start_date, end=end_date)

if not data.empty:
    csv_file = f"stocks/{ticker}.csv"
    data.to_csv(csv_file)
    print(f"Historical data saved to {csv_file}")
```

The downloaded data is stored in the `stocks/` directory, with each CSV file corresponding to a specific stock. Each file contains detailed daily historical data, including key metrics such as opening prices, closing prices, high and low prices, and trading volume.

Table 1: Input data Visualisation

Date	Price	Adj Close	Close	High	Low	Volume
2000-01-04	24.67	70.04	70.04	72.19	70.04	3194413
2000-01-05	22.97	65.22	65.22	67.90	64.86	6058531

Data Profiling

Data profiling is a crucial step in the data analysis. It involves examining and summarizing datasets to understand their structure, content, and quality.

Following are the Statistics measured for each stock exchange date:

- **Start Date**: The earliest available date in the dataset for each stock, representing when trading data begins.
- **Minimum Close Price**: The lowest recorded closing price for the stock.
- **Maximum Close Price**: The highest recorded closing price for the stock.
- **Average Close Price**: The average of all closing prices for the stock.
- **Standard Deviation of Close Prices**: The variability of closing prices over the recorded period.
- **Null Values**: The count of missing (null) values in the `Close` column.
- **Zero Values**: The count of rows where the `Close` price is recorded as zero.

Table 2: Data Profiling Metrics for Stocks

Stock	Start Date	Min Close	Max Close	Avg Close	Std Dev Close	Null Count	Zero Count
0001.HK	2000-01-04	28.9461	123.0745	68.4965	20.3846	0	0
0002.HK	2000-01-04	28.9500	96.9500	59.3026	16.4811	0	0
0003.HK	2000-01-04	2.4421	16.3100	7.4452	3.3786	0	0
0004.HK	2000-01-04	3.4526	33.3500	14.3006	7.0777	0	0
0005.HK	2000-01-03	28.2000	152.8000	83.1612	28.5090	0	0
0006.HK	2000-01-04	22.8000	82.2500	48.1553	14.8687	0	0
0007.HK	2000-09-11	0.0160	11.2000	1.7957	1.7044	0	0
0008.HK	2003-01-08	1.7096	6.4780	4.0214	0.8109	0	0
0009.HK	2001-09-12	0.0100	23.6685	4.8352	6.3664	0	0
0010.HK	2000-01-04	4.8750	54.1000	24.0057	13.6892	0	0

Data Cleaning

These are the three cleanings implemented:

- **Removal of Unnecessary Columns:** Removing colums such as opening prices, high and low prices, and trading volume and keep only closing prices.
- **Duplicate Date:** Duplicate rows based on the `Date` column were removed to ensure that each date in the dataset corresponds to a unique stock entry.
- **Forward and Backward Filling:** For rows where the closing price was missing, forward and backward filling techniques were used. Null values in the `Close` column were replaced by the most recent non-null value encountered in previous rows (forward filling). If any null values remained at the beginning of the dataset, they were filled using the first available non-null value from subsequent rows (backward filling).

The below table shows the data after cleaning.

Table 3: Date and Closing Price of a Stock

Date	Close
2000-01-04	70.0424
2000-01-05	65.2180
2000-01-06	62.0018
2000-01-07	63.2526
2000-01-10	63.7886
2000-01-11	65.3967
2000-01-12	64.3246
2000-01-13	63.4312
2000-01-14	62.5378
2000-01-17	63.2526
2000-01-18	66.4688
2000-01-19	65.7541
2000-01-20	65.5754
2000-01-21	65.9327
2000-01-24	66.8261

Data Ingestion

After cleaning data for each stock, the data is ingested by writing the processed DataFrames back to HDFS and is stored as a CSV file in the `stocks_processed/` directory. Each file retains only the essential columns: `Date` and `Close`.