# INVESTMENT RECOMMENDATION SYSTEM PROJECT

APAN5400 Group 8

# Background & Definition

## Background

- The stock market is a complex and dynamic environment where investors constantly seek opportunities to maximize returns and minimize risks.
- Effective and informed investment decision-making in the stock market is important

## Definition

- Extract, transform, and load financial data from the Yahoo Financial
- Analyze stock market trends and patterns using MongoDB for JSON data storage, PostGRE for structured data storage, and Flask for API service
- Generate investment recommendations for users

# Data Source Specification & Procurement Details

## 1 Data Source Specification

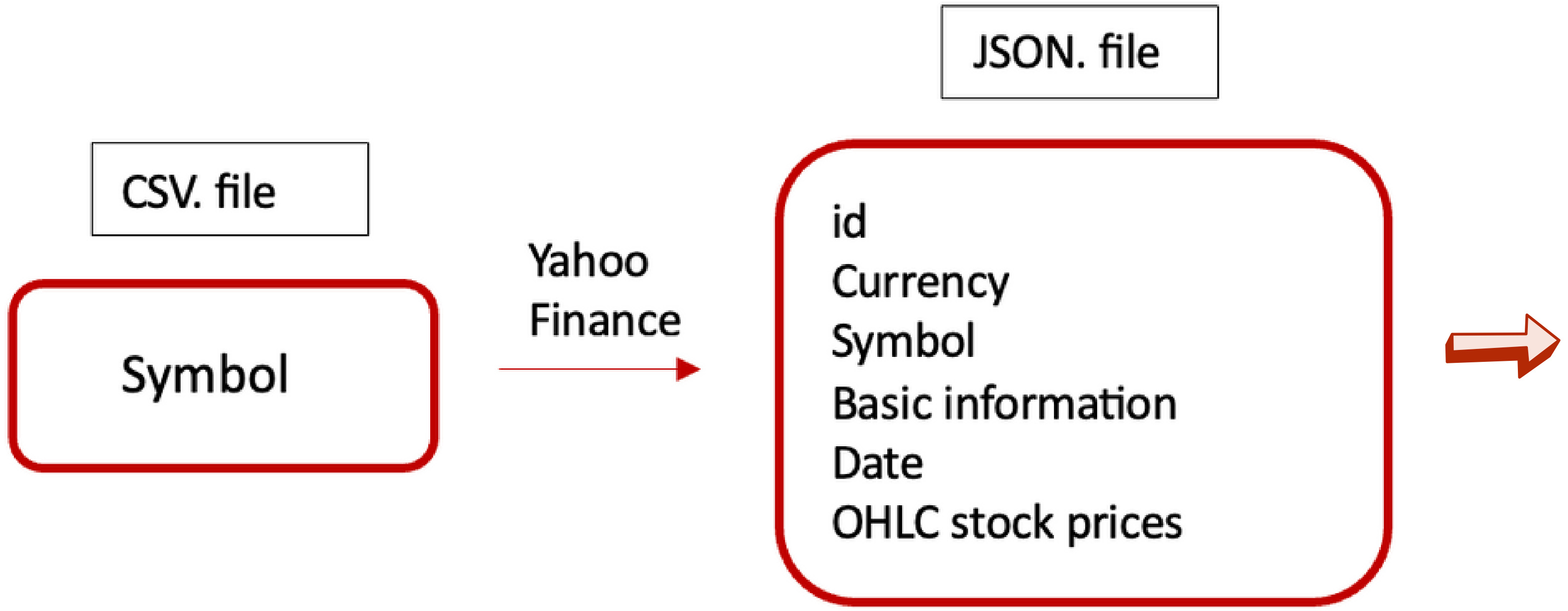
**"File" 1:**

Collect symbol from ***https://stockanalysis.com/stocks/*** .
Symbol, as the representative of the company, is an important connection during the integration of its basic information and its OHLC stock prices.

**"File" 2:**

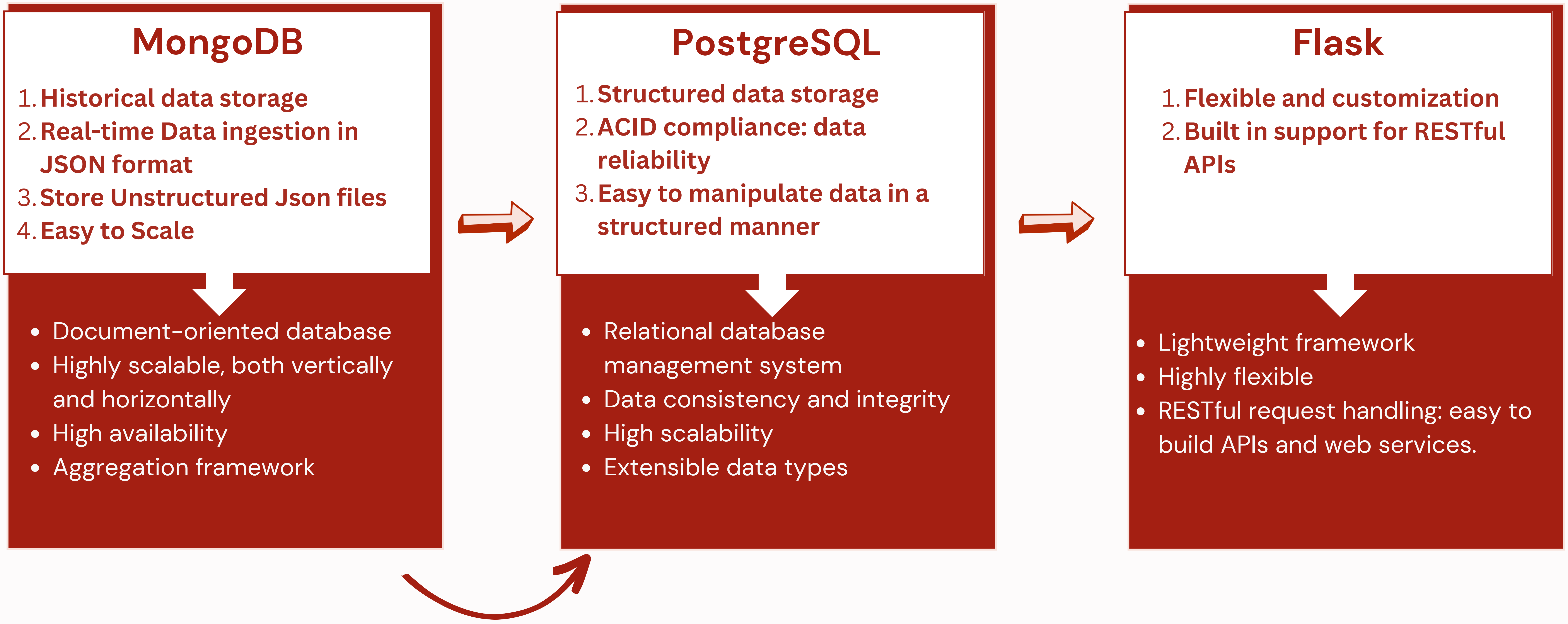
The basic information about each company based on symbol and the daily OHLC of each company from Yahoo Finance in JSON format, **ranging from 2013 to 2023**.

## 2 Procurement Details



Contains **5602** companies and the latest history caught before the deadline for the submission is: **2023-4-19**.

# Proposed Design Choices &
# Rationale for Selected Technologies



## MongoDB

1. **Historical data storage**
2. **Real-time Data ingestion in JSON format**
3. **Store Unstructured Json files**
4. **Easy to Scale**

- Document-oriented database
- Highly scalable, both vertically and horizontally
- High availability
- Aggregation framework

## PostgreSQL

1. **Structured data storage**
2. **ACID compliance: data reliability**
3. **Easy to manipulate data in a structured manner**

- Relational database management system
- Data consistency and integrity
- High scalability
- Extensible data types

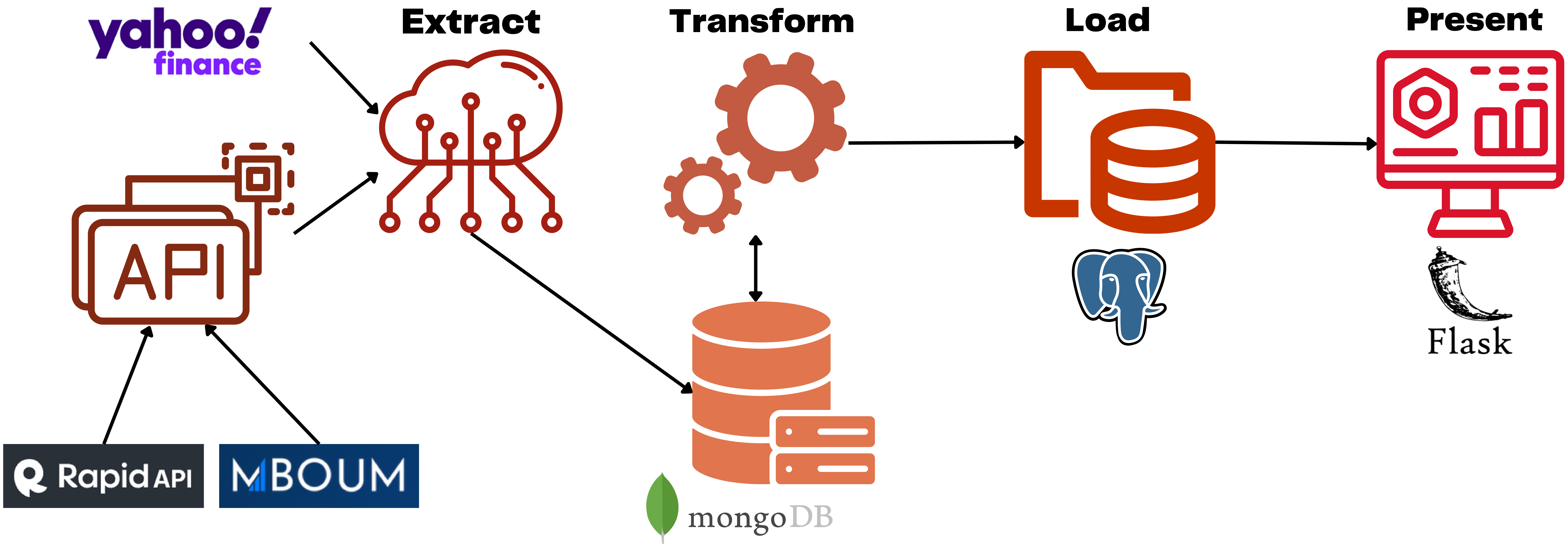
## Flask

1. **Flexible and customization**
2. **Built in support for RESTful APIs**

- Lightweight framework
- Highly flexible
- RESTful request handling: easy to build APIs and web services.

**Transform to Structured Data**

# ETL PipDiagram



# Scalability and Cost Implications

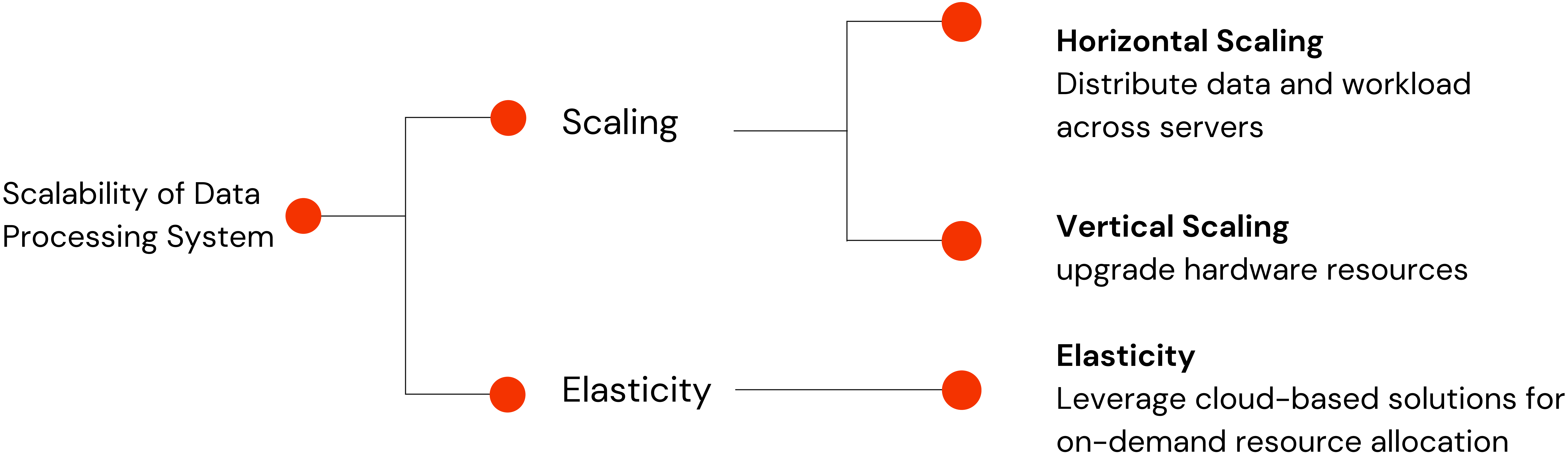
## *Cost Implications* (Monthly)*

| | | | |
|---|---|---|---|
| **1. Infrastructure Costs** | (1) Cloud Service Provider | Compute instances, Storage, Database services | $100~$200 |
| | (2) On premises Hardware | Server hardware, Network equipment (one time cost) | $300~$1500 |
| **2. Data Storage and Processing Costs** | (1) Data Storage | SQL databases (PostgreSQL) | $100~$200 |
| | (2) Data Processing | Real-time analytics services | $100~$200 |
| **3. API Usage Costs** | NASDAQ Financial Data API | Basic / Premier tier | $50~$100 |

→ ***$650 ~ $2200***

*Since this is currently an academic project, its monetary costs will be relatively small. A large portion of the percentage will be the cost of the R&D member's time. If the decision is made to expand it into a corporate project, the monetary cost will increase significantly.

# Scalability and Cost Implications



Scalability of Data Processing System

Scaling

**Horizontal Scaling**
Distribute data and workload across servers

**Vertical Scaling**
upgrade hardware resources

Elasticity

**Elasticity**
Leverage cloud-based solutions for on-demand resource allocation

**<u>Data Quality dimension</u> ensures**

- **ACCURACY:** data cleaning & validation checks
- **COMPLETENESS:** fill in missing values
- **CONSISTENCY:** data schema & data normalization
- **TIMELINESS:** real-time data sources
- **RELEVANCY:** exclude irrelevant data

**<u>Licensing</u>:** use permissive open-source data that can be accessed, used, shared, and modified

**<u>Scalability</u> ensures:**

(1) adapt to changing demands& requirements
(2) handle increasing data volumes & processing demands

# Conclusion

## Success Metric & Evaluation Criteria

### Data Extraction

**Success Metric: Extraction Efficiency**

**Rationale**

It is crucial because it directly impacts the speed and accuracy of downstream data processing and analysis

**Evaluation Criteria**

Data extraction rate of **5000** records per minute or higher.

### Data Storage

**Success Metric: Data Reliability and Integrity**

**Rationale**

If the data stored is corrupted, incomplete, or inaccurate, it will affect the accuracy of analysis, and any insights based on this analysis may be flawed.

**Evaluation Criteria**

Data loss rate of **0.5%** or less

### Data Processing

**Success Metric: Data Processing Time**

**Rationale**

- Affects the speed at which insights can be gained from the data.
- Affects the cost of data processing. Longer processing time = More resources(CPU, memory) = Higher costs.

**Evaluation Criteria**

Data processing time of **10 seconds** or less

**Success Metric: Scalability**

**Rationale**

Our data updates each day. Without scalability, we may not be able to handle the increasing data volume= slower processing times + increased resource usage + potential data loss.

**Evaluation Criteria**

Process **30 years** of stock data without a significant decrease in performance

**>5000**
Extraction Rate

**<0.5%**
Data loss rate

**<10s**
Processing Time

**30Y**
data process

# References

https://stockanalysis.com/stocks/
https://finance.yahoo.com/
https://rapidapi.com/sparior/api/mboum-finance

## Group Members

- Julia Yang (yy3276)
- Jianing Yang (jy3229)
- Jiehui Ding (jd3894)
- Xinyang Tang (xt2275)
- Xinyi Liu (xl3206)