# Dataset and Business Requirement

## Nguồn dữ liệu (Data Source)

Dữ liệu trong dự án này được thu thập từ nhiều API chuyên cung cấp thông tin về thị trường chứng khoán, bao gồm:

1. sec-api.io (Danh sách các công ty đang giao dịch trên thị trường chứng khoán)

* Cung cấp thông tin về các công ty niêm yết trên NYSE, NASDAQ và AMEX.
* Dữ liệu được cập nhật hàng tháng.

1. Alpha Vantage API (Market Status)

* Cung cấp thông tin về trạng thái hoạt động của các thị trường chứng khoán trên toàn cầu.
* Dữ liệu được cập nhật hàng tháng.

1. Alpha Vantage API (News Sentiment)

* Cung cấp tin tức liên quan đến các cổ phiếu kèm theo phân tích cảm xúc (sentiment analysis).
* Dữ liệu được cập nhật liên tục theo khung giờ giao dịch.

1. Polygon API (OHLC Data)

* Cung cấp dữ liệu giá cổ phiếu bao gồm giá mở cửa, cao nhất, thấp nhất và đóng cửa (OHLC).
* Dữ liệu được cập nhật hàng ngày.

## Yêu cầu nghiệp vụ (Business Requirement)

Dự án này có mục tiêu thu thập, xử lý và phân tích dữ liệu chứng khoán nhằm hỗ trợ các quyết định đầu tư. Cụ thể, dự án thực hiện các nhiệm vụ sau:

1. Xác định danh sách công ty niêm yết: Truy vấn sec-api.io vào cuối mỗi tháng để cập nhật danh sách các công ty mới niêm yết hoặc bị hủy niêm yết.
2. Theo dõi trạng thái thị trường: Truy vấn Alpha Vantage API để kiểm tra trạng thái mở cửa/đóng cửa của thị trường chứng khoán toàn cầu.
3. Phân tích tin tức và sentiment: Thu thập tin tức trong các khung giờ giao dịch để phân tích xu hướng tâm lý của thị trường.
4. Thu thập dữ liệu giá cổ phiếu (OHLC): Truy vấn Polygon API hàng ngày để cập nhật giá cổ phiếu.

Dữ liệu thu thập được sẽ được sử dụng để xây dựng các mô hình phân tích và dự báo biến động thị trường.

# Understand Data

## Cấu trúc dữ liệu từ API

1. **sec-api.io (Danh sách công ty giao dịch trên thị trường chứng khoán)**

* ticker: Mã giao dịch của công ty trên sàn chứng khoán.
* name: Tên đầy đủ của công ty.
* exchange: Sàn giao dịch mà công ty đang niêm yết (NYSE, NASDAQ, AMEX).
* sector: Ngành công nghiệp chính của công ty.
* industry: Ngành nghề cụ thể của công ty.
* sicCode: Mã ngành SIC của công ty.
* sicDescription: Mô tả ngành SIC.
* entityType: Loại thực thể của công ty (ví dụ: Corporation, LLC, Partnership).
* isDelisted: Trạng thái niêm yết của công ty (True nếu đã bị hủy niêm yết, False nếu vẫn đang giao dịch).
* category: Loại hình công ty (Public, Private, Investment Trust, v.v.).

1. **Alpha Vantage API (Market Status - Trạng thái thị trường)**

* market\_type: Loại thị trường chứng khoán (ví dụ: Stock Market, Commodity Market).
* region: Khu vực của thị trường (ví dụ: US, EU, Asia).
* primary\_exchanges: Các sàn giao dịch chính hoạt động trong khu vực.
* local\_open: Thời gian mở cửa của thị trường tại múi giờ địa phương.
* local\_close: Thời gian đóng cửa của thị trường tại múi giờ địa phương.
* current\_status: Trạng thái hiện tại của thị trường (Open, Closed, Pre-market, After-hours).

1. **Alpha Vantage API (News Sentiment - Phân tích tin tức và cảm xúc thị trường**

* title: Tiêu đề bài viết.
* url: Liên kết đến bài viết gốc.
* time\_published: Thời gian bài viết được xuất bản.
* authors: Danh sách tác giả của bài viết.
* summary: Tóm tắt nội dung bài viết.
* overall\_sentiment\_score: Điểm số cảm xúc tổng thể của bài viết (từ -1 đến 1, với -1 là tiêu cực, 0 là trung lập, 1 là tích cực).
* overall\_sentiment\_label: Nhãn cảm xúc tổng thể của bài viết (Negative, Neutral, Positive).
* ticker\_sentiment: Danh sách cổ phiếu được nhắc đến trong bài viết và điểm số sentiment tương ứng, bao gồm:
  + ticker: Mã cổ phiếu.
  + relevance\_score: Độ liên quan của cổ phiếu với nội dung bài viết.
  + ticker\_sentiment\_score: Điểm số cảm xúc đối với cổ phiếu (từ -1 đến 1).
  + ticker\_sentiment\_label: Nhãn cảm xúc của cổ phiếu (Negative, Neutral, Positive).

1. **Polygon API (OHLC Data - Giá cổ phiếu)**

* T: Mã cổ phiếu.
* o: Giá mở cửa của cổ phiếu trong phiên giao dịch.
* h: Giá cao nhất trong phiên.
* l: Giá thấp nhất trong phiên.
* c: Giá đóng cửa của cổ phiếu trong phiên giao dịch.
* v: Khối lượng giao dịch (số lượng cổ phiếu được giao dịch).
* t: Timestamp của dữ liệu giao dịch.
* n: Số lượng giao dịch đã thực hiện.

## Tần suất cập nhật dữ liệu

* Danh sách công ty (sec-api.io): Cập nhật hàng tháng.
* Tình trạng thị trường (Alpha Vantage API - Market Status): Cập nhật hàng tháng.
* Tin tức & sentiment (Alpha Vantage API - News Sentiment): Cập nhật liên tục trong khung giờ giao dịch.
* Dữ liệu giá cổ phiếu (Polygon API - OHLC Data): Cập nhật hàng ngày.

## Ước tính khối lượng dữ liệu

* **Dữ liệu OHLC**: ~10,000 dòng/ngày.
* **Tin tức & sentiment**: ~100 bài viết/ngày.
* **Danh sách công ty**: ~28,000 dòng (cập nhật theo tháng).

## Thiết kế Database lưu trữ dữ liệu từ sec-api.io và Alpha Vantage API for market status

* 2 API **sec-api.io** và **Alpha Vantage API for market status** không cần phải query thường xuyên như 2 API còn lại vì dữ liệu trong 2 API này hiếm khi thay đổi và chỉ cần cập nhật vào cuối tháng.
* Do đó, ta sẽ lưu dữ liệu từ 2 API này vào Database nhằm giả lập Backend Database làm Data Source thêm đa dạng
* Database được chuẩn hóa 3NF gồm 5 table sau:

A diagram of a company

AI-generated content may be incorrect.

## Thiết kế Data Warehouse

Thiết kế bảng Dimension và bảng Fact cho các Business Requirement từ Data Source là 2 API và 1 Database

**Bảng Dimension:**

* 1. **Bảng Dim Companies:**
* companyId(PK)
* name: Tên công ty
* ticker: ký hiệu công ty trên sàn giao dịch chứng khoán
* isDelisted: kiểm tra công ty có bị loại khỏi sàn hay không
* category: mô tả hạng mục cổ phiếu của công ty
* currency: đơn vị tiền tệ mà cổ phiếu của công ty được giao dịch
* location: địa điểm của trụ sở công ty
* exchange: tên sàn giao dịch
* region: khu vực
* industry: ngành cụ thể mà công ty hoạt động
* sector: ngành của công ty
* sicIndustry: xác định ngành dựa trên mã SIC
* sicSector: mô tả ngành dựa trên mã SIC
  1. **Bảng Dim News:**
* newId (PK)
* title: tiêu đề của bài viết
* url: đường dẫn đến bài viết
* timePublished: thời gian xuất bản
* author: nhóm tác giả
* summary: tóm tắt nội dung
* source: nguồn cung cấp
* source\_domain: tên miền của nguồn
* overall\_sentiment\_score: điểm cảm xúc tổng thể của bài viết
* overall\_sentiment\_lable: nhãn mô tả cảm xúc tổng thể của bài viết
  1. **Bảng Dim Topics:**
* topicId (PK)
* topicName: tên topic
  1. **Bảng Dim Times:**
* timeId (PK)
* dayOfWeek: ngày trong tuần
* day: ngày
* month: tháng
* quarter: quý
* year: năm
  1. **Bảng Fact News\_Topics:**
* topicNewId (PK)
* newId (FK)
* topicId (FK)
* relevance\_scrore: điểm độ liênn quan của bài viết đến chủ đề đó
  1. **Bảng Fact News\_Companies**
* newCompanyId (PK)
* newId (FK)
* companyId (FK)
* relevance\_score: độ liên quan của bài báo đến cổ phiếu đó
* ticker\_sentiment\_score: điểm cảm xúc của bài báo cho cổ phiếu đó
* ticker\_sentiment\_label: nhãn mô tả cảm xúc của bài báo cho cổ phiếu đó
  1. **Bảng Fact Candles:**
* candleId (PK)
* companyId (FK)
* open: giá mở cửa
* high: giá cao nhất
* low: giá thấp nhất
* close: giá đóng cửa
* volume: khối lượng giao dịch

## Galaxy Schema

A diagram of a computer code

AI-generated content may be incorrect.