Checkpoint B: Data and Schema for the Knowledge Graph

Abstract

This project develops a hybrid knowledge base integrating structured financial time-series data with unstructured regulatory and media text for publicly traded healthcare companies. Data was collected from three primary sources: FDA regulatory announcements, Yahoo Finance stock price data, and global news articles via the GDELT API. A dual architecture was implemented in which PostgreSQL serves as the primary system of record for structured and semi-structured data, while Memgraph provides a graph-based analytical layer. The knowledge graph models companies, regulatory documents, news articles, trading days, and price observations as labeled nodes, with semantic relationships such as MENTIONS and HAS_PRICE represented as edges. The design follows a subject-predicate-object triple structure. This hybrid architecture enables both transactional integrity and efficient multihop traversal, supporting predictive modeling and competitive intelligence applications.

1. Introduction

The healthcare industry operates within a highly regulated environment where product approvals, safety communications, recalls, and enforcement actions can significantly influence a firm's performance. Publicly traded healthcare companies often experience market reaction to regulatory announcements and media coverage. However, these interactions are difficult to analyze using traditional isolated data sources. Financial time series data is structured and numeric, while regulatory documents and news articles are unstructured and textual. Integrating these heterogeneous data types requires a representation capable of modeling both entities and their relationships.

The objective of this project is to construct a knowledge graph that unifies regulatory documents, media coverage, and financial market behavior for major publicly traded healthcare firms. The resulting system

is designed to support relationship centric analysis, such as tracing how specific regulatory events propagate through companies and affect stock price trajectories. To accomplish this, a hybrid relational graph database architecture is implemented.

2. Literature Review and Theoretical Grounding

Knowledge graphs are commonly modeled using subject–predicate–object triples, in which entities are connected through explicitly defined relationships. Nickel et al. (2015) outline four broad approaches to constructing knowledge bases:

1. **Manually Curated Approach:** Triples are created manually by a closed group of experts, leading to high accuracy but low scalability.
2. **Collaborative Approach**: Triples are created manually by an open group of volunteers, offering better scaling than curated methods (e.g., Freebase, Wikidata).
3. **Automated Semi-structured Approach:** Triples are extracted automatically from structured or semi-structured data sources, such as Wikipedia infoboxes, using rules or machine learning.
4. **Automated Unstructured Approaches (Information Extraction):** Triples are extracted automatically from unstructured text using natural language processing (NLP) and machine learning techniques.

Following Nickel et al.'s taxonomy, this project mainly employs automated information extraction from unstructured text (approach 4) supplemented by manual curation (approach 1).

Regulatory documents are collected through a focused crawler, and company mentions are detected using structured alias matching. These detected relationships are explicitly encoded as graph edges. The initial company registry (companies.csv) represents manual curation, providing authoritative identifiers and aliases that guide the extraction process. Unlike embedding-based systems, which represent relationships implicitly in vector space, this approach maintains explicit, interpretable semantic triples such as (RegulatoryDoc, MENTIONS, Company) and (Company, HAS_PRICE, PricePoint). This explicit

symbolic structure facilitates interpretability and deterministic traversal while remaining extensible to future representation learning techniques.

The labeled property graph model implemented in Memgraph directly reflects the triple-based structure described by Nickel et al. (2015), while the relational layer in PostgreSQL supports normalized entity storage and structured indexing.

3. Hybrid Database Architecture

A dual-storage architecture was selected to balance transactional integrity with graph-based analytical flexibility. The system employs a hybrid approach that separates structured storage from relationship-centric analysis.

PostgreSQL functions as the primary system of record, storing structured financial time series and semi-structured regulatory documents using JSONB fields. This layer benefits from strong ACID guarantees, mature indexing mechanisms, and efficient aggregation capabilities over numeric data. Referential integrity is enforced through foreign keys and normalized schema design.

Memgraph serves as a complementary analytical layer optimized for graph traversal. Entities such as companies, regulatory documents, trading days, and price observations are modeled as nodes, while semantic connections such as MENTIONS and HAS_PRICE are represented as edges. Data from PostgreSQL is exported to Memgraph that transforms normalized relational records into graph nodes and edges. This design enables multihop queries and pattern discovery that would otherwise require multiple join tables in a purely relational system. By separating storage responsibilities, the graph layer is optimized specifically for traversal and inference rather than transactional durability, while PostgreSQL maintains data consistency and serves as the authoritative source.

4. Data Collection

A curated list of publicly traded healthcare companies was constructed and stored in as a CSV file (companies.csv). This registry provides consistent identifiers, including ticker symbols and aliases, that are used across both relational and graph layers. Regulatory documents were collected using a focused crawler targeting FDA domains. The crawler extracts meaningful English text from HTML pages, filters content using domain-specific keywords related to recalls, approvals, and enforcement actions, and stores cleaned documents in JSON Lines format. Each record contains metadata including URL, title, timestamp, relevance score, and detected entity mentions.

Financial time-series data were retrieved from Yahoo Finance for each company and stored as daily observations including open, high, low, close, adjusted close, and volume. These structured numeric records allow alignment of regulatory events with market responses over time.

The GDELT Document API was implemented to collect global media coverage referencing selected companies. Unfortunately I kept running into a HTTP 429 rate limiting error. Hence this is still a work in progress.

5. Relational Schema Design

The PostgreSQL schema models core entities including companies, regulatory documents, news articles, trading days, and price observations. Regulatory and news documents are stored with JSONB fields to preserve raw metadata while enabling indexed queries. Structured time-series data are stored in normalized tables with unique constraints on ticker–date combinations.

Bridge tables represent graph relationships in relational form. For example, a table linking regulatory documents to companies provides a normalized representation of the MENTIONS edge. These bridge tables ensure referential integrity and enable deterministic export to the graph layer. In this architecture, PostgreSQL serves as the authoritative storage engine, maintaining data consistency while allowing flexible semi-structured document storage.

6. Graph Schema Design

The knowledge graph is implemented as a labeled property graph in Memgraph. Core node types include Company, RegulatoryDoc, NewsArticle, TradingDay, and PricePoint. Relationships encode semantic and temporal associations, including document mentions of companies and company-level financial observations tied to specific dates.

This design follows a subject-predicate-object structure. For example, a regulatory document mentioning a company is represented as a directed edge from the document node to the company node. Financial observations are represented by linking companies to price nodes, which are in turn linked to trading-day nodes. This layered temporal structure enables traversal from regulatory events to subsequent price movements and allows exploration of multihop paths across entities.

7. Challenges

One main challenge encountered during implementation was API rate limiting from the GDELT Document API. HTTP 429 errors occurred during batch retrieval. As mitigation I tried exponential backoff and throttling strategies and refactoring the crawler to segment queries by timeline windows, Unfortunately this did not work either. As of right now this is a work in progress.

8. Conclusion

This is a project checkpoint report for our hybrid healthcare knowledge graph integrating structured financial data and unstructured regulatory documents within a coherent architecture. PostgreSQL provides reliable structured storage and JSON document support, while Memgraph enables efficient relationship-centric traversal and pattern discovery. Grounded in the symbolic triple-based knowledge representation framework described by Nickel et al. (2015), the system maintains explicit semantic relationships suitable for interpretability and predictive modeling.

References

- Nickel, Maximillian, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. "A Review of Relational Machine Learning for Knowledge Graphs." *Proceedings of the IEEE*, 104(1): 11–33. Online reference: https://arxiv.org/abs/1503.00759 Paper: https://arxiv.org/pdf/1503.00759.pdf