Term Project Checkpoint B

Greg Roberts and Flavio Mota

MSDS 459 - Knowledge Engineering

May 11, 2025

Abstract

This paper presents the design and current status of a consumer-discretionary knowledge graph built in Neo4j. We integrate a manually curated list of 10 key companies and a structured product taxonomy with time-series and financial metrics ingested via APIs (FRED, Fiserv FSBI, Yahoo Finance) and unstructured corporate metadata and filings scraped from Wikipedia and EDGAR. Our prototype pipelines clean and transform these diverse sources into a graph schema that captures entities such as Company, Industry, Product, Index, SpendingSeries, and Document, connected through meaningful relationships. The resulting graph will support exploratory and competitive-intelligence queries, demonstrating the value of a unified data model for stock-recommendation analytics.

Introduction

In the consumer-discretionary sector, decision makers require a holistic view that brings together corporate metadata, market indices, consumer-spending trends, and product relationships. Traditional siloed tables obscure cross-entity insights, while a graph model naturally expresses the interconnections among companies, industries, and time-series indicators. This research constructs an extensible knowledge graph in Neo4j to support competitive intelligence and stock-recommendation analyses. By programmatically ingesting both structured and unstructured data alongside curated reference files, we aim to deliver a data-driven platform that enables portfolio managers and analysts to traverse relationships, perform sector comparisons, and uncover latent patterns.

Literature Review

Foundational work by Kejriwal, Knoblock, and Szekely (2021) describes semi-automated pipelines for relation extraction and graph construction, illustrating how knowledge graphs

can bind heterogeneous data into a coherent structure. In financial applications, Yao et al. (2018) demonstrated the use of graph embeddings and traversal techniques for risk analysis, while Paulheim (2017) explored recommendation systems built on graph representations of corporate and market data. These studies highlight the benefits of graph queries for uncovering non-obvious correlations—such as linking consumer-spending anomalies to specific corporate events—and informed our choice of Neo4j for its flexible schema and rich query language.

Methods

Our ingestion workflow combines manual curation with automated pipelines:

- Manual curation: We maintain a CSV file of 10 consumer-discretionary companies (symbol, name, sector) and a JSON file defining product categories and subcategories.

Company	% Wtg
Amazon.com, Inc.	21.3%
Tesla, Inc.	16.0%
The Home Depot, Inc.	7.3%
Booking Holdings Inc.	4.8%
McDonald's Corporation	4.8%
The TJX Companies, Inc.	4.1%
Lowe's Companies, Inc.	3.6%
Starbucks Corporation	2.6%
O'Reilly Automotive, Inc.	2.3%
Chipotle Mexican Grill, Inc.	2.0%

Figure 1 - The Consumer Discretionary Select Sector SPDR Fund (XLY): https://finance.yahoo.com/quote/XLY/holdings/

- API integration:
    - FRED API: Retrieves U.S. consumer-spending series (e.g., Retail Trade, Food Services) as time-series data.
    - Fiserv FSBI API: Downloads monthly small-business index values for the consumer-discretionary sector.
    - Yahoo Finance API: Pulls key financial metrics (revenue, market capitalization, P/E ratio, EBITDA) for each company.
- Web scraping:

- Scrapy: Crawls and downloads Wikipedia pages for each company, extracting CEO, headquarters, founding date, and sector classification, etc.
- BeautifulSoup: Parses EDGAR HTML filings (e.g., 10-K and 8-K), capturing document metadata (type, date, URL) for downstream NLP.

Data are cleaned and normalized in Pandas, then ingested into Neo4j via the official Python driver over the Bolt protocol.

Sample Cypher output used in our Neo4j prototype:

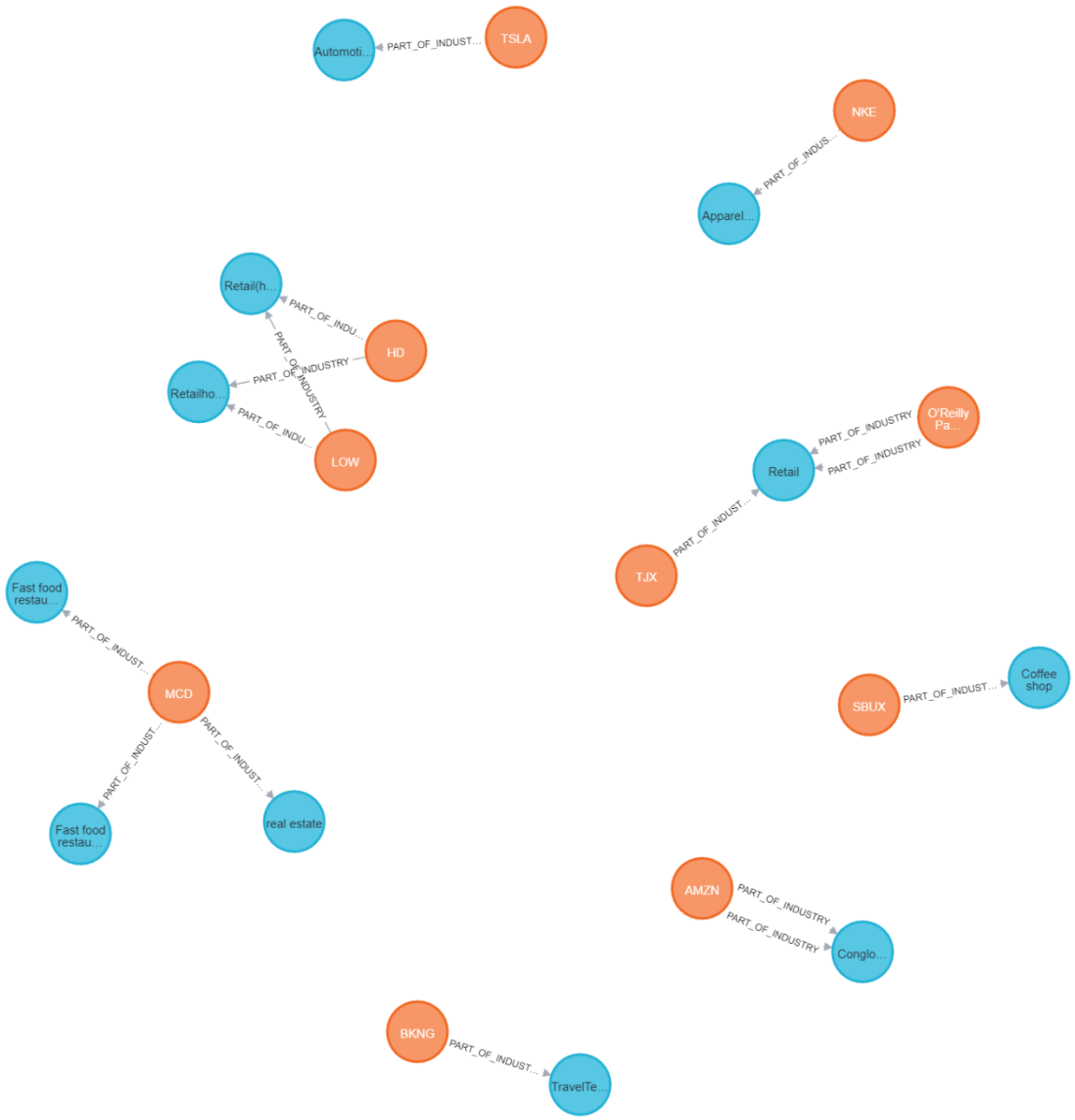

Figure 2: Company graph - Neo4j

Figure 3 - HD and LOW product graph - Neo4j

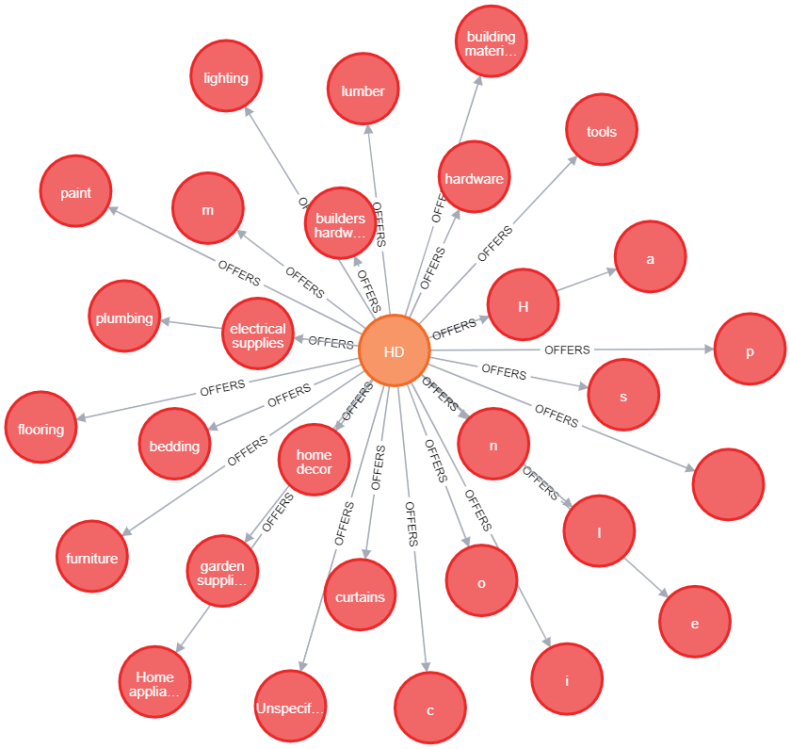

Figure 4 - HD product graph - Neo4j

Results

Data Sources Identified

- Manually curated: A list of 10 consumer-discretionary companies; a hierarchical product taxonomy.
- Structured via APIs:
    - Consumer-spending series (FRED: Retail Trade, Food Services).
    - Monthly FSBI metrics (Fiserv).
    - Corporate financial metrics (Yahoo Finance).
- Unstructured via scraping:
    - Company metadata from Wikipedia (CEO, location, sector).
    - Filings metadata from EDGAR (document type, date, URL).

Information Collected

- A CSV table of 10 companies with symbol, name, and sector classification.
- JSON files for FSBI monthly values spanning January 2015 through April 2025.
- Scraped HTML pages stored under /data/wikipages/, each parsed into JSON records containing structured metadata fields.
- EDGAR document catalogs stored under /data/edgar/, with metadata tables for future full-text analysis.

Proposed Graph Schema

- Node labels:
    - Company (attributes: symbol, name, sector, CEO, headquarters)
    - Industry (name)
    - Product (category, subcategory)
    - Index (name)
    - IndexPoint (date, value)
    - SpendingSeries (seriesName)
    - SpendingPoint (date, value)

    - Document (type, date, url)
- Relationships:
    - (c:Company)-[:PART_OF_INDUSTRY]->(i:Industry)
    - (c:Company)-[:OFFERS]->(p:Product)
    - (idx:Index)-[:HAS_POINT]->(pt:IndexPoint)
    - (ss:SpendingSeries)-[:MEASURED_AT]->(sp:SpendingPoint)
    - (c:Company)-[:HAS_METRIC]->(m:FinancialMetric)
    - (d:Document)-[:RELATED_TO]->(c:Company)

Conclusions

The current prototype demonstrates how a partially curated, programmatically enriched knowledge graph can unify disparate data sources into a single analytical framework. By linking corporate metadata, market indices, consumer-spending trends, and document archives, the graph supports multidimensional queries—enabling competitive-intelligence analyses such as peer-group comparisons, sector-wide trend spotting, and risk-event correlations. This integrated model directly addresses the management problem of constructing a data-driven platform for stock recommendations in the consumer discretionary sector. Future work will expand company coverage, automate continuous EDGAR ingestion with NLP-based relation extraction, integrate sentiment and analyst-rating feeds, and deploy scheduled pipelines to keep the graph current. Ultimately, this knowledge graph will serve as a robust foundation for strategic decision support and predictive modeling.