**Term Project Checkpoint A**

**Greg Roberts and Flavio Mota**

**MSDS 459 - Knowledge Engineering**

**Apr 27, 2025**

**Abstract**

This research report presents the creation of a consumer discretionary sector knowledge graph that integrates relationships among publicly traded firms, their product categories, consumer segments, transaction records, economic indicators, and diverse data sources. To date, we have defined core node and edge types, assembled automated pipelines for data ingestion from indices, government surveys, web sources, financial APIs, and regulatory filings, and constructed an initial EdgeDB schema that reflects both graph semantics and strong typing. Preliminary data loading and entity extraction processes have been validated, and a prototype graph visualization and query interface has been developed to demonstrate multi-hop exploration of sector dynamics.

**Introduction**

The impetus for this research arises from the critical role that consumer discretionary spending plays in signaling economic cycles and informing strategic decisions in investment and marketing. By creating a knowledge base that captures granular transaction patterns and firm attributes, we aim to enable equity analysts, portfolio managers, marketing strategists, and data science teams to navigate complex interdependencies without manual data wrangling. Potential users include senior executives seeking an executive view of sub-sector performance, data scientists performing cluster analysis on spending velocity and ticket size, and application developers building RAG-enhanced chat interfaces for natural language enquiry. Our planned applications encompass targeted stock recommendation systems, precision marketing model development, and interactive dashboards that translate user queries into graph queries.

**Literature review**

Knowledge graphs have been applied in domains as varied as biomedical research and web search, with foundational work such as the Never Ending Language Learning project demonstrating large scale fact harvesting and self-supervised relation extraction. The textbook Knowledge Graphs Fundamentals Techniques and Applications outlines best practices in ontology design and graph database implementation. In finance, academic studies have explored graph-based modeling of corporate networks and sentiment analysis driven by social media feeds. Industry white papers by leading AI vendors describe frameworks for RAG augmented knowledge retrieval that combine embedding models with graph query engines. Collectively these sources inform our approach to schema definition, data pipeline architecture, entity disambiguation, and graph powered insight generation.

**Methods**

Our methodology addresses the five core research questions. First, we chose the consumer discretionary sector due to its sensitivity to economic cycles and Greg's professional vantage point at a payments company, enabling access to rich transaction data. Second, we designed an EdgeDB graph relational schema guided by RDF Schema principles in which node types such as Company, Product-Category, Consumer-Segment, Economic-Indicator, Transaction-Record, and Data-Source are defined with strong typing and explicit relationship semantics. Third, to populate the graph we automated ingestion from the Fiserv Small Business Index, BEA expenditure surveys, Census Bureau retail reports, FRED time series via API, Wikipedia structured lists, Yahoo Finance real time data, SEC EDGAR filings, and industry news via a focused crawler.

Scrapy spiders and Python ETL scripts normalize JSON CSV and HTML into EdgeDB upserts, while LangChain loaders and text splitters map unstructured documents into nodes and edges using Named Entity Recognition and Relationship Extraction. Fourth, we identified likely users and framed sample queries such as Which firms lead growth among mid income cohorts or Which product categories align with shifts in consumer confidence. Finally, we outlined our application stack comprising an interactive graph exploration dashboard, a natural language interface powered by RAG augmented retrieval, and a recommendation engine that applies node centrality and community detection algorithms.

**Results**

Thus far our initial ingestion pipelines have successfully loaded thousands of transaction records alongside macroeconomic indicators and equity fundamentals. The Fiserv index data proved invaluable in capturing real time small business spending trends, while BEA and Census datasets provide robust anchoring to macro series. Wikipedia scrapes offered reliable lists of sector constituents, and our SEC filing parser extracted firm level narrative insights that enrich edge semantics. During schema trials we discovered additional node types such as SubSectorCluster and PeerGroup emerged organically from cluster analysis. Entity resolution challenges around ticker symbol changes and corporate breakups highlighted the need for sustained record linkage processes. Early visualizations demonstrate that multi hop paths, linking a drop in consumer confidence through CPI changes to firm market reactions that can be surfaced in seconds.

**Conclusions**

Our findings demonstrate that the assembled data assets address the management problem of building a knowledge graph for competitive intelligence in the consumer discretionary sector. The integration of transaction level records with macroeconomic indicators and firm level fundamentals creates a rich multidimensional view of market dynamics. Real time small business spending trends from the FSBI combined with BEA and Census data anchor our internal records to established benchmarks. Wikipedia scrapes and SEC filings enrich the graph with context around sector constituents and corporate strategy. Early visual explorations confirm that queries such as identifying firms that gain when mid income cohort spending rises or uncovering peer clusters by transaction velocity can be executed in seconds rather than days. This capability directly supports equity analysts and marketing strategists by surfacing competitive signals and demand drivers without manual data assembly.

At the same time, we recognize areas that require ongoing attention to ensure the knowledge graph remains a reliable tool for decision makers. Entity resolution challenges around ticker symbol changes and corporate reorganizations call for robust record linkage processes. The focused crawler must be scaled and governed to capture evolving qualitative insights from industry news and trade journals. Schema governance will be essential to manage new node types that emerge from data mining and to preserve strong typing as the graph grows. Overall, the breadth and depth of the collected data lay a solid foundation for a competitive intelligence platform. With careful data quality management and an extensible schema design, the knowledge graph will empower users to generate actionable insights and maintain a strategic edge in the consumer discretionary market.

**References:**

T. Mitchell and E. Fredkin, "Never-ending language learning," 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2014, pp. 1-1, doi: 10.1109/BigData.2014.7004203.

**Appendix: Questions to address**

**1. What is your topic and why did you choose it?**

Our project focuses on building a **consumer discretionary sector knowledge graph** that integrates rich relationships between publicly traded firms, their products, and observed consumer-spending patterns. We chose this for a variety of reasons including a strong interest in how consumer discretionary spending is closely linked to economic cycles. Greg works as a data scientist at a payments company where he has direct exposure to granular transaction data across small and medium-sized merchants, insights that are invaluable for understanding demand drivers in discretionary spending.

Embedding these data into a graph allows us to surface non-obvious links (for example, between macroeconomic indicators and sub-sector performance) and powers applications such as targeted stock recommendations or marketing-strategy optimization. This is done through "Named Entity Recognition," (NER). Nodes, representing patterns gleaned from the web scraping process, are connected with lines (these lines are the edges).

These relationships are mined from the data using Relationship Extraction ("RE").

Additionally, the knowledge graphs provide a visual representation of the available information and the relationships among the nodes. This visualization provides a useful framework that makes it incredibly easy to explain to an audience (e.g., the C Suite, stakeholders, investors, and so on).

**2. Initial thoughts on the database schema (EdgeDB graph-relational model)**

Building on RDF Schema principles for ontology design, our EdgeDB schema will define object types (nodes) and links (edges) with both strong typing and graph semantics. Likely **node types** include:

- **Company** (ticker, headquarters, sector classification)

- **ProductCategory** (e.g., apparel, leisure equipment)

- **ConsumerSegment** (household income bands, demographics)

- **EconomicIndicator** (CPI, consumer confidence)

- **TransactionRecord** (timestamp, amount, merchant_id)

- **DataSource** (BEA, BLS, FRED, FSBI, Wikipedia, Yahoo Finance, internal acquirer data)

The web scraping and data mining processes will likely reveal other nodes. And the new information will be used to adjust the algorithm to make the same robust and useful in gathering new data.

**3. Identified information sources and initial data collection**

To populate the graph, we'll combine:

- **Fiserv Small Business Index (FSBI)**: a rolling index that tracks small business revenue and consumer spending trends across key discretionary categories.

- **Government datasets**: consumer expenditure surveys (BEA), retail sales reports (U.S. Census Bureau), and labor statistics (BLS) to anchor our internal data to macro trends.

- **FRED series**: key consumer spending time-series (e.g., Personal Consumption Expenditures, Retail Sales) fetched directly from the Federal Reserve Economic Data API.

- **Web sources**:

- **Wikipedia** for structured lists (e.g., sector constituents, historical index values).
    - https://en.wikipedia.org/wiki/Amazon
    - https://en.wikipedia.org/wiki/Home_Depot
    - https://en.wikipedia.org/wiki/AutoZone
    - https://en.wikipedia.org/wiki/Booking_Holdings
    - https://en.wikipedia.org/wiki/McDonald%27s
    - https://en.wikipedia.org/wiki/Nike,_Inc.
    - https://en.wikipedia.org/wiki/TJX_Companies
    - https://en.wikipedia.org/wiki/Lowe%27s
    - https://en.wikipedia.org/wiki/Mercado_Libre
- **Yahoo Finance** APIs for real-time price and fundamental data on consumer discretionary stocks.

- **Company disclosures**: SEC EDGAR filings (10-Ks, earnings calls) for firm-level fundamentals and strategic initiatives.

- **Industry publications and news feeds**: trade journals and targeted web crawls (using a focused-crawler pipeline with distiller modules) to capture qualitative context.

**Automated ingestion:**
We'll implement **Scrapy** spiders to scrape HTML tables and JSON endpoints from FSBI, FRED, Wikipedia, and Yahoo Finance; ingest SEC filings with BeautifulSoup/XML parsers; and pull CSV/Excel data via Python ETL scripts. Extracted data—whether JSON, CSV, or HTML—is then normalized and upserted into EdgeDB. For unstructured documents (PDFs and web pages), we'll integrate LangChain loaders and text-splitters to extract and map facts into nodes and edges.

**4. Likely users and user questions**

Primary users include:

- **Equity analysts and portfolio managers**, querying "Which consumer discretionary firms show leading growth among mid-income segments?"

- **Marketing strategists**, asking "Which product categories correlate with upticks in disposable-income spending during holidays?"

- **Data science teams**, performing advanced analytics such as "What sub-sector clusters emerge when segmenting firms by both transaction velocity and average ticket size?"

By traversing the knowledge graph, users can quickly answer multi-hop questions (e.g., identify companies whose stock prices align with shifts in consumer confidence), rather than assembling data manually.

**5. Proposed application and KG utility**

We envision a **web-based dashboard and API** that layers:

1. **Graph-powered search & exploration**: interactive graph visualizations for exploratory analysis.

2. **Natural-language question answering**: a RAG-augmented chatbot that translates user queries into graph queries for on-the-fly insights.

3. **Recommendation engine**: using graph analytics (e.g., node centrality, community detection) to surface high-potential stock picks or marketing targets.

In this stack, the knowledge base underpins information retrieval (via schema-driven queries), information extraction (through automated ETL and Scrapy-based ingestion),

question-answering (via embedding-augmented retrieval over node properties), and recommendations (through graph algorithms that detect patterns in spending behavior and firm fundamentals).