Julie Bach & Samuel Doherty
Professor Disha Shende
DSDA 310
Dec 9th, 2025

FINAL REPORT - CASE STUDY 2

1. Introduction

This project implements a Company Intelligence Agent that processes corporate documents, converts them into searchable text, retrieves relevant information, and answers user questions using a retrieval-augmented generation (RAG) framework. Delta Air Lines serves as the focal company for this implementation. The system is composed of two core components:

- A data processing pipeline that performs document extraction, cleaning, chunking, and storage
- An application layer integrating a TF-IDF model, a Groq LLM, and a Streamlit interface that enables non-technical users to interact with the processed knowledge base.

2. Document Collection

The system integrates information from Delta Airline and Yahoo Finance:

- **Delta Airline's 2024 Form 10-K (SEC Filing)** with financial performance, operations, and risk disclosures
- **Delta Airline's 2024 ESG Report** with environmental metrics, climate goals, and sustainability initiatives
- **Major Shareholders Summary (Yahoo Finance)** with ownership concentration and investor profile.

3. Data Processing Pipeline

The data processing pipeline performs the transformation of raw PDFs to a structured and searchable document store. The workflow proceeds as follows:

3.1 PDF Text Extraction

The script first attempts extraction using PDFplumber, reading each page and preserving structural content. While this method works well for Form 10-K filings, Delta's ESG report contains images, decorative layouts, and scan-like formatting, which disrupted PDFplumber's text extraction. This might have resulted in wrong-order extraction or blank pages. Hence, the final text extraction function will always apply Optimal Character Recognition—a tool that is used to convert images of text (like scanned documents or photos) into machine-readable, editable text data—on the ESG file. If a page yields empty or near-empty text, the script falls back to multi-column extraction, as these PDFs might use a two-column layout, which might extract texts in the wrong order. This approach could enable recovery of most content but might introduce mirror errors. On the other hand, the other two files are purely used with regular PDFplumber's text extraction. Each extracted page is appended to a running text body.

3.2 Text Cleaning and Standardization

After extraction, the text is standardized and cleaned. The cleaning logic includes:

- Removing page headers, footers, excess whitespace, extra hyphens, and duplicated line breaks
- Normalizing whitespace, punctuation, hyphens, and line breaks
- Repairing common PDF artifacts (e.g., split words, broken numbering).

This ensures that the downstream retrieval system receives high-quality and consistent text.

3.3 Text Chunking and Data Store

The cleaned text is divided into overlapping segments using a sliding-window chunking strategy. Each chunk is roughly 600 words long with a 100-word overlap. This overlap improves retrieval quality by ensuring that important sentences are not split in disruptive ways. Once chunked, the text segments from all documents are combined into a unified DataFrame. Each row contains the chunk ID, the company name, the source document, and the chunk text. The completed store is saved as both a CSV and a pickle file for loading during the application phase.

4. Retrieval System

The retrieval layer is built around a TF-IDF weighting system trained on all text chunks. When a user enters a question, it is transformed into a TF-IDF vector and compared against all chunk vectors using cosine similarity. The top-k most relevant chunks are returned. This method ensures accurate retrieval and serves as the basis for the RAG system.

5. LLM Integration

The question-answering model integrates the Groq LLaMA-3.3-70B Versatile model. Retrieved chunks and user queries are combined into a structured prompt with the following rules:

- The model must answer only using the retrieved evidence
- It must cite chunk IDs
- It must avoid speculation or adding external knowledge
- If the documents do not contain the answer, the model must explicitly state so.

For the testing phase, we use this sample set of questions and later verify again in the input files.

```
company_snap_questions = [
    "What is Delta Airlines?",
    "What are the main business segments?",
    "How does Delta generate revenue?"]

risk_factors_questions = ["What are the main risks?"]

esg_questions = [
    "What environmental goals does the Delta describe?",
    "What sustainability targets does it mention?"]

custom_questions = [
    "How is Delta's DEI strategy?",
    "What is the major holder breakdown?",
    "What are some awards that Delta achieves?"
    "What is Delta known for"]
```

Below is an example of the Q&A:

==========

QUESTION: What is Delta Airlines?

ANSWER:

Delta Airlines is described in the provided chunks as a major airline company that operates in a highly competitive industry, with significant competition from other carriers, including American Airlines and United Airlines (chunk_032). The company has a strong focus on customer service, with a goal of increasing customer loyalty through its award-winning SkyMiles program (chunk_005). Delta also has a diversified business with high-margin revenue streams, including premium products, partnerships, and complementary businesses such as cargo operations and maintenance services (chunk_008). The company has made significant investments in its people, product, and reliability, and has achieved differentiated performance through margin expansion, durable earnings, and free cash flow (chunk_005). (chunk_032, chunk_014, chunk_013, chunk_008, chunk_005)

Top Chunks:

	chunk_id	**source_file**	**score**
31	**chunk_032**	annual_report_2024	0.224142
13	**chunk_014**	annual_report_2024	0.223418
12	**chunk_013**	annual_report_2024	0.201537
7	**chunk_008**	annual_report_2024	0.188648
4	**chunk_005**	annual_report_2024	0.155097

By using these constraints, we could ensure transparency and reliability in generated responses.

6. KPI Extraction and Summary

To support the Company Snapshot tab, a subset of 6 key performance indicators (KPIs) was extracted from Delta Air Lines' Form 10-K and ESG Report, including:

- Total operating revenue
- Total cargo revenue
- YoY revenue growth rate, compared to 2023
- Number of reportable segments
- Number of served customers
- Total fuel savings
- Key ESG targets and sustainability commitments

Because the document does not contain perfectly-formatted texts, the KPIs extraction process combined automated regex-assisted searching and human verification, ensuring accuracy while still leveraging the document processing pipeline.

6.1 Loading Cleaned Documents and Previewing Snippet.

We first loaded the cleaned text files produced in the data pipeline and the unified text chunk store. Because financial statements' frequency varies in structure, a preview function is used to print the surrounding context for any keyword or regex pattern. The function allowed us to rapidly search the cleaned text and manually inspect relevant passages.

6.2 Extraction and Verification

For example, we want to extract the total operating revenue, and a preview_matches is used to print snippets around each match, as shown below. The same process is repeated for the other five KPIs. All these values were manually inspected again before being added to the KPI JSON file used by the Streamlit app.

```
preview_matches(
    text_10k,
    r"cargo revenue",
    window=250
)
```

```
Found 4 matches for pattern: 'cargo revenue'
```

[1] ...a fixed dollar or percentage division of revenues for tickets sold to passengers traveling on connecting flight itineraries. Cargo Through our global network, our cargo operations are able to connect the world's major freight gateways. We generate cargo revenues in domestic and international markets through the use of cargo space on regularly scheduled passenger aircraft. We are a member of SkyTeam Cargo, an international airline cargo alliance with eight other airlines that offer a network spanning six co...

[2] ...mber of SkyTeam Cargo, an international airline cargo alliance with eight other airlines that offer a network spanning six continents, through which we provide global solutions to our customers by connecting our network with those partners. In 2024, cargo revenues were approximately $822 million. Other Complementary Businesses We have various other businesses arising from our airline operations, including the following: + In addition to providing maintenance and engineering support for our fleet of 1,292 m...

[3] ...ts our estimate of revenue expected to be recognized in the next twelve months based on projected redemptions, while the balance classified as a noncurrent liability represents our estimate of revenue expected to be recognized beyond twelve months. Cargo Revenue Cargo revenue is recognized when we provide the transportation. Delta Air Lines, Inc. | 2024 Form 10-K 65 --- PAGE 68 END --- Notes to the Consolidated Financial Statements Other Revenue Year Ended December 31, (in millions) 2024 2023 2022 Ref...

[4] ... of revenue expected to be recognized in the next twelve months based on projected redemptions, while the balance classified as a noncurrent liability represents our estimate of revenue expected to be recognized beyond twelve months. Cargo Revenue Cargo revenue is recognized when we provide the transportation. Delta Air Lines, Inc. | 2024 Form 10-K 65 --- PAGE 68 END --- Notes to the Consolidated Financial Statements Other Revenue Year Ended December 31, (in millions) 2024 2023 2022 Refinery $ 4,642 $...

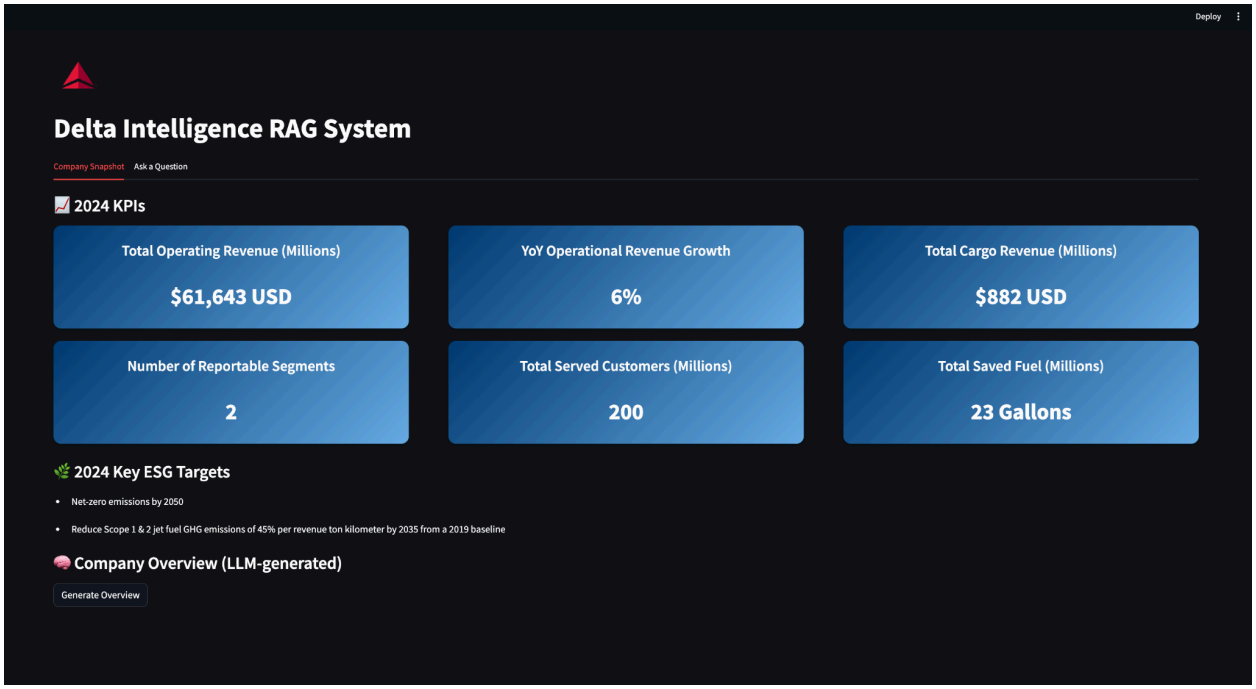
7. Streamlit Application

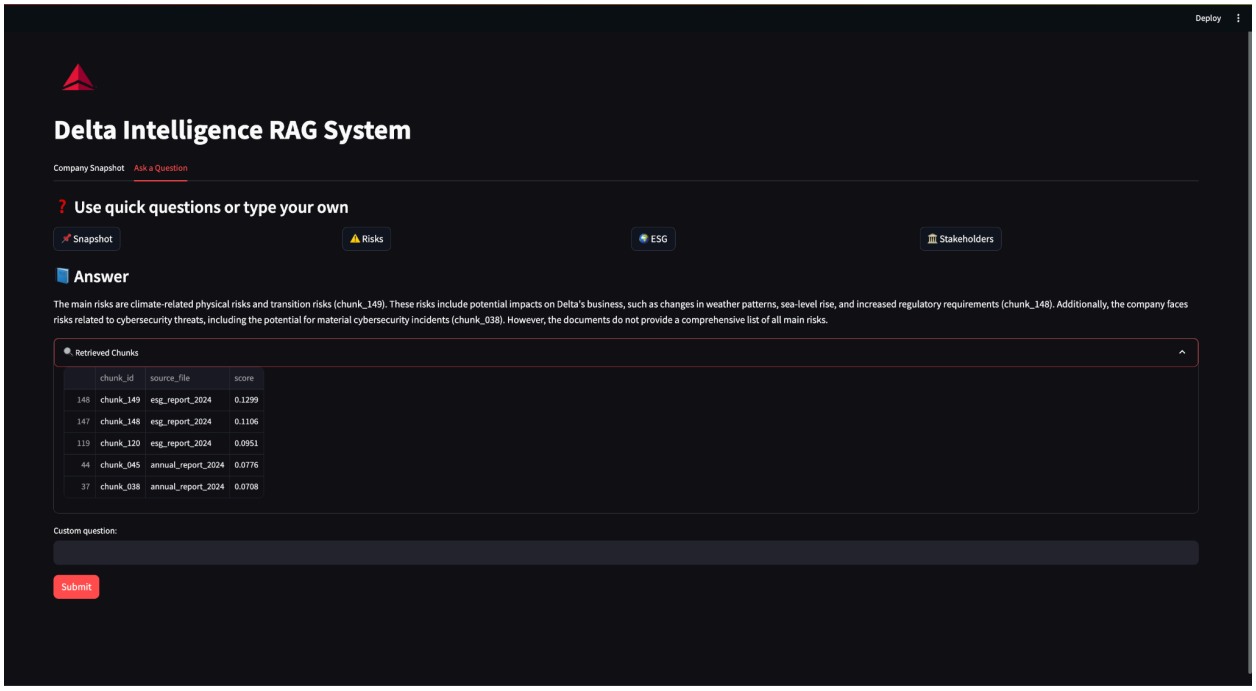
The user-facing component of the system is an interactive Streamlit dashboard. Upon launch, the application:

- Loads the chunk store and KPI summary
- Initializes the TF-IDF retriever
- Connect to the Groq LLM
- Renders an interactive, Delta-themed dashboard

The dashboard contains two main tabs. The Company Snapshot tab displays the KPI cards, verified ESG targets, and a button that generates a 3-4 sentence company overview using the LLM. The Ask-a-Question tab, on the other hand, allows users to select predefined queries (e.g., Snapshot, Risks, ESG, Stakeholders), enter a customer question, view LLM-generated answers with chunk citations, and expand a panel to view retrieved chunks and their similarity scores.

The dashboard can be accessed either by running the app.py file or using this link. Below is the screenshots:





8. Limitations and Difficulties

There were three main limitations we encountered:

- First, we were unable to extract the PDF with 100% accuracy because of the format of the ESG report. We were forced to rely on OCR. This likely affects the accuracy of the retrieval system.
- Second, the system is not as automated as it could be, as the KPI extraction is still done manually.
- Third, the API for the Groq model we are using has limited query uses, so we encountered difficulty testing the agent fully.
- Fourth, there are other chunking strategies we could have used, such as recursive splitting. There may be other methods that are more applicable to this model and may have yielded better results.

Hence, for future improvements, we would like to discover more robust text extraction and chunking techniques to enhance accuracy while using other paid API that allows unlimited calls at low cost to challenge the agent with different questions and identify hallucinations.

9. Conclusion

This project demonstrates a functional Company Intelligence Agent capable of PDF extraction, text cleaning, chunking, retrieval, and LLM-based question answering. The resulting Streamlit application offers a clean, intuitive interface for exploring Delta's financial, operational, and environmental information. Despite remaining limitations around OCR reliability and automated KPI extraction, we were able to design and deploy a practical RAG system for internal corporate use.

10. Acknowledgment of AI Use

Generative AI tools, including ChatGPT and GroqLLMs, were used to support code generation, debugging, and interface design in accordance with the course guidelines.