

AS13 Report Naman Kedia
Financial Statement Question-Answering with a RAG System

1. Processing Details

This section describes how raw financial filings were processed, cleaned, and organized.

1.1 Data Collection:

- Financial reports for 10 sampled companies (tickers) over 10 years 2010–2019 (*both inclusive*) were downloaded programmatically using **sec_edgar_downloader** pypi package.
- The reports were stored in a structured directory format:
`data/original/sec-edgar-filings/{ticker}/10-K/{timestamp-year}/full-submisison.txt`
- `Tickers = ["BA", "BIIB", "D", "FFIV", "GOOG", "HLT", "HRL", "IDXX", "STZ", "TSLA"]`

1.2 Data Cleaning:

- Raw HTML or text files were parsed using **BeautifulSoup** (*'lxml' parser*)
- **Basic Processing** was done to maintain a folder structure
 - `PROCESSED_DATA_DIR/{ticker}_{earliest year}/content.txt`
- **Advanced Processing:**
 - Boilerplate content such as headers, footers, style and irrelevant sections (e.g., post "XBRL INSTANCE DOCUMENT") were removed.
 - Non-ASCII characters and gibberish encoding were filtered using regular expressions.
 - While we needed to create vectors for only 5 sampled tickers, a vector data store(*table name='final'*) was made for all the 10 tickers. Except Google(GOOG)(*omitted*) all others had completed 10 years of filings.

2. Vector Store Construction

This section explains how document embeddings were created and stored in the vector database.

2.1 Chunking:

- Each financial report was divided into nodes based on logical boundaries:
 - Initially basic chunking used paragraph splitting, trailing space splitting with a maximum chunk size of 500 characters.
 - Finally advanced chunking method was used
 - **RecursiveCharacterTextSplitter()** from langchain
 - It has the effect of trying to keep all paragraphs (and then sentences, and then words) together as long as possible, as those would generically seem to be the strongest semantically related pieces of text.
 - Metadata **{ticker and year}** was added to each chunk to enable efficient filtering during retrieval.
 - **LlamaIndex** was used to convert all chunks to TextNodes for db insert

2.2 Vector Database:

- **chromaDB** was the vector store index used as it is open source
 - Batch processing was done to insert chunks as throttling limit of 40K on chromaDb and our total chunks were ~300K

- Embeddings: Dense vectors representing each chunk. Encoded using **'all-MiniLM-L6-v2'** sentence transformer
- Metadata: Key-value pairs (**ticker**, **year**) for filtering.
- Documents: Original text content for context during retrieval.

3. RAG Baseline Implementation

This section details the components of the Retrieval-Augmented Generation (RAG) system.

3.1 Pipeline Overview:

The RAG pipeline combines retrieval and generation to answer user queries based on company financial statements.

1. Retriever:

- Retrieves relevant chunks from the vector store using cosine similarity between query embeddings and stored embeddings.
- Metadata filters ensure that only chunks matching the specified company (ticker) and (year) are considered.

2. Re-ranker:

- A cross-encoder model ("**cross-encoder/ms-marco-MiniLM-L-6-v2**") ranks retrieved chunks based on relevance to the query.
- Typically between 0 and 1 (when using sigmoid activation)
- Higher scores indicate better semantic matches
- Best Score returned as our selected node for LLM input

3. Generator:

- Uses a pre-trained language model from Together API (**meta-llama/Llama-3.3-70B-Instruct-Turbo**) to generate responses based on the retrieved chunks. (*Using free credits*)
- Input format
 - i. "Answer the question given the context when possible. Keep in mind to return correct units when returning numerical data. Question: {query}\n Context: {retrieved_chunk}\n Answer:"

4. Evaluation

This section evaluates the system's performance in terms of retrieval relevance and response quality.

- We should try to **ask questions of the previous financial year**. For eg 2014 answers would be in 2015 metadata
- We run **5 queries per ticker** based on various relevant questions
- Average response time was **1-3 seconds** per query

4.1 Retriever Performance:

- For each query, we assessed whether the retrieved chunks contained relevant information.
- Example Query: "What were Boeing's revenues in 2014?"
 - Retrieved Chunk: "Notes to the Consolidated Financial Statements Summary of Business Segment Data (Dollars in millions) Years ended December 31, 2014 2013 2012 Revenues: Commercial Airplanes \$59,990 \$52,981 \$49,127 Defense, Space & Security: Boeing Military

Aircraft 13,511 15,285 15,388 Network & Space Systems 8,003 8,512 7,911 Global Services & Support 9,367 9,400 9,308 Total Defense, Space & Security 30,881 33,197 32,607 Boeing Capital”

- Relevance Score: High

4.2 Response Performance:

- The generated responses were evaluated for accuracy, completeness, and coherence.
- Example Response: “Boeing's revenue in 2014 was \$90,287 million (or \$90.3 billion). This is calculated by adding the revenues of Commercial Airplanes (\$59,990 million), Defense, Space & Security (\$30,881 million), and Boeing Capital (\$416 million).”
 - Strengths: Accurate numerical data, concise phrasing with units due to proper prompting.
- **Limitations:** Due to our chunk size being less than 500, the queries for numerical data work really well. However, on asking questions which are more subjective the contexts are less likely to be valid.
 - We have also prompted the LLM to give answers only relevant to the context to make sure the responses we receive are verified and genuine. This is also why the responses fail on subjective questions.

5. Streamlit App

A simple web application was developed using Streamlit to allow users to interact with the RAG system.

Features:

1. Dropdown menus for selecting company ticker and year.
2. Text box for entering queries.
3. Button to submit queries and display results (response + relevant context + response time).
4. Any errors will be handled and shown with an alert dialog saying no context was found.
5. All Responses were generated from the actual RagSystem() class in the backend

Sample Screen Shot

Financial Statement Q&A

Select a Company Ticker:

BA

Select a Year:

2016

Enter your question:

What was the net profit for 2015

Submit

Response:

The net profit for 2015 was \$33 billion.

Relevant Context:

Airline financial performance also plays a role in the demand for new capacity. Airlines continue to focus on increasing revenue through alliances, partnerships, new marketing initiatives, and effective leveraging of ancillary services and related revenues. Airlines are also relentlessly focusing on reducing costs by renewing fleets to leverage more efficient airplanes and in 2015 benefited significantly from lower fuel costs. Net profits for the global airline industry are estimated to total *33 billion in 2015 compared to 17 billion in 2014*. We expect airline profits to continue to grow in 2016, driven by strong passenger demand and relatively low oil prices.

Example Queries & Results

Query	Ticker	Year	Retrieved Context	Generated Response
What were Boeing's revenues in 2014?	BA	2015	"..... Revenues: Commercial Airplanes \$59,990 \$52,981 \$49,127 Defense, Space & Security: Boeing Military Aircraft 13,511 15,285 15,388 Network & Space Systems 8,003 8,512."	"Boeing's revenues in 2015 were \$96 billion."
What were Tesla's R&D expenses in 2015?	AAPL	2016	"R&D expenses for the year ended December 31, 2015 were \$717.9 million, an increase from \$464.7 million for the year ended December 31, 2014."	"Tesla's R&D expenses in 2015 were \$717.9 million"

How much came from companion animal diagnostics in 2013?	IDEXX	2014	"Developing and introducing new proprietary diagnostic tests and services that provide valuable medi....."	"The context provided does not specify the amount that came from companion animal diagnostics in 2013. It only mentions the company's goals and strategies related to companion animal products and services, but does not provide any numerical data or specific information about the year 2013. Therefore, the answer is: No specific amount is provided in the context."
--	-------	------	--	--

Sample Screenshots of Manual Validation

For BA (Boeing Airlines)

```
Query: What was Boeing's revenue in 2014?
Best retrieved node:Notes to the Consolidated Financial Statements Summary of Business Segment Data (Dollars in millions) Years ended December 31, 2014 2013 2012 Revenues: Commercial A
irplanes $59,990 $52,981 $49,127 Defense, Space & Security: Boeing Military Aircraft 13,511 15,285 15,388 Network & Space Systems 8,003 8,512 7,911 Global Services & Support 9,367 9,40
0 9,308 Total Defense, Space & Security 30,881 33,197 32,607 Boeing Capital 416 408 468
Response: Boeing's revenue in 2014 was $90,287 million (or $90.3 billion). This is calculated by adding the revenues of Commercial Airplanes ($59,990 million), Defense, Space & Securit
y ($30,881 million), and Boeing Capital ($416 million).

Query: What were Boeing's net earnings in 2012?
Best retrieved node:See Notes to the Consolidated Financial Statements on pages 54 – 108 . 51 Table of contents The Boeing Company and Subsidiaries Consolidated Statements of Cash Flow
s (Dollars in millions) Years ended December 31, 2012 2011 2010 Cash flows – operating activities: Net earnings $3,900 $4,018 $3,307 Adjustments to reconcile net earnings to net cash p
rovided by operating activities: Non-cash items – Share-based plans expense 193 186 215 Depreciation and amortization 1,811
Response: Boeing's net earnings in 2012 were $3.9 billion.

Query: What was Boeing's commercial airplanes segment revenue in 2013?
Best retrieved node:Notes to the Consolidated Financial Statements Summary of Business Segment Data (Dollars in millions) Years ended December 31, 2013 2012 2011 Revenues: Commercial A
irplanes $52,981 $49,127 $36,171 Defense, Space & Security: Boeing Military Aircraft 15,936 16,019 14,585 Network & Space Systems 8,512 7,911 8,964 Global Services & Support 8,749 8,67
7 8,427 Total Defense, Space & Security 33,197 32,607 31,976 Boeing Capital 408 468 547 Other segment 102 106 123
Response: Boeing's commercial airplanes segment revenue in 2013 was $52,981 million, or $52.981 billion.
```

For Biogen and Dominion

```
Query: How much cash did Biogen have at the end of 2012?
Best retrieved node:Cash and cash equivalents, end of the year $ 570,721 $ 514,542 $ 759,598 See accompanying notes to these consolidated financial statements. F-5 Table of Contents BI
OGEN IDEC INC. AND SUBSIDIARIES CONSOLIDATED STATEMENTS OF EQUITY (In thousands) Preferred stock Common stock Additional paid-in capital Accumulated other comprehensive income (loss) R
etained earnings Treasury stock Total Biogen Idex Inc. shareholders' equity Noncontrolling interests Total equity Shares
Response: Biogen had $759,598 thousand (or $759.6 million) in cash and cash equivalents at the end of 2012.

Query: What was Dominion's operating revenue in 2014?
Best retrieved node:/s/ Deloitte & Touche LLP Richmond, Virginia February 27, 2015 57 Table of Contents Dominion Resources, Inc. Consolidated Statements of Income Year Ended December 3
1, 2014 2013 2012 (millions, except per share amounts) Operating Revenue $ 12,436 $ 13,120 $ 12,835 Operating Expenses Electric fuel and other energy-related purchases 3,400 3,885 3,64
5 Purchased electric capacity 361 358 387 Purchased gas 1,355 1,331 1,177 Other operations and maintenance 2,765 2,459
Response: Dominion's operating revenue in 2014 was $12,436 million, or $12.436 billion.
```

Observations

Strengths:

- Accurate Retrieval:
 - The retriever consistently identifies relevant chunks based on metadata filtering and similarity search
 - Company and year-specific information is accurately filtered using metadata
 - Vector similarity search effectively finds semantically relevant content
 - Handles financial terminology and numerical data well
- Quality Response Generation:
 - The generator produces coherent responses that align with retrieved context

- Maintains numerical accuracy when reporting financial figures
- Preserves important contextual information from financial statements
- Successfully combines information from structured financial data
- Effective Re-ranking:
 - Re-ranking helps get the best search results using cross-encoder scores
 - Improves relevance by considering both query and document content simultaneously
 - Reduces noise from initial embedding-based retrieval
- Robust Data Processing:
 - Effective chunking strategy that preserves document structure
 - Good handling of numerical and tabular financial data

Limitations:

- Context Integration Challenges:
 - Responses lack depth when multiple nodes are required for answers
 - Difficulty in using information across different sections of financial reports
 - Limited ability to combine historical trends across multiple years as single best node retrieved
 - Struggles with complex analytical queries requiring multiple data points
- Scope Constraints:
 - Subjective answers are not possible as context is usually not enough to capture them at our max chunk size
 - Cannot make forward-looking statements or predictions
 - Limited ability to compare across companies or sectors
- Technical Limitations:
 - Fixed chunk size may not be optimal for all types of financial data
 - Embedding model may not capture all nuances of financial terminology
 - Re-ranking process adds latency to query responses

Conclusion

The implemented RAG system demonstrates strong performance in retrieving and generating answers based on financial statements. It provides a solid foundation for question-answering tasks over structured financial data especially numerical data due to accurate chunk filtering with cosine similarity and metadata filtering.

***** End Of Main Report *****

Appendix

1. Subjective Output

Financial Statement Q&A

Select a Company Ticker:

HLT

Select a Year:

2016

Enter your question:

Is hilton planning to expand soon?

Submit

Response:

Yes, Hilton is planning to expand soon. They may continue to build their portfolio by launching new hotel and non-hotel brands in the future and expand existing brands into new international markets.

Relevant Context:

We launched a new midscale brand, Tru by Hilton, in January 2016. We introduced a new brand, Canopy by Hilton, in October 2014, opened our first Curio - A Collection by Hilton hotel in August 2014, opened the first Herb N' Kitchen Restaurant in 2013 and opened our first Home2 Suites by Hilton hotel in 2011. We may continue to build our portfolio by launching new hotel and non-hotel brands in the future. In addition, the Hilton Garden Inn, DoubleTree by Hilton and Hampton by Hilton brands have been expanding into new jurisdictions outside the United States in recent years. We may continue to expand existing brands into new international markets. New hotel products or concepts or brand expansions may not be accepted by hotel owners, franchisees or customers and we cannot guarantee the level of acceptance any new brand will have in the development and consumer marketplaces. If new branded hotel products, non-hotel branded concepts or brand expansions are not as successful as we anticipate, we may not recover the costs we incurred in their development or expansion, which could have a material adverse effect on our business, financial condition or results of operations.

2. No output due to missing context

Financial Statement Q&A

Select a Company Ticker:

HLT

Select a Year:

2019

Enter your question:

What will happen in 2024 with the hotel

Submit

Response:

There is no specific information provided about what will happen in 2024 with the hotel. The context discusses the development pipeline and the risks associated with it, but it does not provide a specific timeline or prediction for 2024.

Relevant Context:

hotels in our development pipeline, which we define as hotels under construction or approved for development under one of our brands. The commitments of owners and developers with whom we have contracts are subject to numerous conditions, and the eventual development and construction of our development pipeline not currently under construction is subject to numerous risks, including, in certain cases, the owner's or developer's ability to obtain adequate financing and obtaining governmental or regulatory approvals. As a result, not every hotel in our development pipeline may develop into a new hotel that enters our system.

3. Handling missing contexts

Financial Statement Q&A

Select a Company Ticker:

TSLA

Select a Year:

2010

Enter your question:

What will happen in 2003

Submit

Response:

There is no information provided in the context to determine what will happen in 2003. If you could provide more context or details, I would be happy to try and assist you further.

No relevant context found.

4. Output with response time

Financial Statement Q&A

Select a Company Ticker:

BA

Select a Year:

2012

Enter your question:

What was the net profit in 2011

Submit

Response:

The net profit in 2011 was \$7 billion.

Relevant Context:

airline industry were impacted in 2011 by high oil prices, which were approximately 40% higher, on average, than 2010 prices. Net profits for the global airline industry are expected to be modest this year, totaling *7billion* in 2011, *reduced from* 16 billion in 2010. Airlines continue to focus on boosting revenue through alliances and partnerships and ancillary fees and services, while also

Approx Response Time: 2.73 seconds

5. Other Sample Retrieval/Response

Query: How much beer revenue did Constellation generate in 2013?

Best retrieved node:Fiscal 2012 . Fiscal 2013 Fiscal 2012 % Increase (Decrease) (in millions) Beer \$ 2,588.1 \$ 2,469.5 5 % Wine and Spirits: Wine 2,495.8 2,386.8 5 % Spirits 300.3 267.5 12 % Total Wine and Spirits 2,796.1 2,654.3 5 % Total Reportable Segments 5,384.2 5,123.8 5 % Consolidation and Eliminations (2,588.1) (2,469.5) (5 %) Consolidated Net Sales \$ 2,796.1 \$ 2,654.3 5 % Net sales increased to \$2,796.1 million for Fiscal 2013 from \$2,654.3 million

Response: Constellation generated \$2,588.1 million in beer revenue in 2013.

Query: What were Constellation's spirits sales in 2015?

Best retrieved node:annual report on Form 10-K of Constellation Brands, Inc. /s/ KPMG LLP Rochester, New York April 25, 2016 EX-31.1 10 stzex311_229201610k.htm EXHIBIT 31.1 Exhibit Exh ibit 31.1 RULE 13a-14(a)/15d-14(a) CERTIFICATION OF CHIEF EXECUTIVE OFFICER Constellation Brands, Inc. Form 10-K for Fiscal Year Ended February 29, 2016 I, Robert Sands, certify that: 1. I have reviewed this report on Form 10-K of Constellation Brands, Inc.;

Response: The provided context does not contain specific information about Constellation's spirits sales in 2015. It appears to be a certification section of the annual report, which d oes not include detailed sales data. To find the spirits sales for 2015, you would need to look at the financial sections or tables within the full annual report that specifically outl ine sales performance by category, including spirits, for that year.

Reference

- **ChatGPT was used for generating a few portions of the code such as: eg. Streamlit UI, Folder Traversal For processing and download, few prompt generation for financial queries etc.**