Enhanced Data Extraction using Gen AI

EY and Elastic Collaboration

The better the question. The better the answer. The better the world works.

Shape the future with confidence

Abstract

The growing accessibility of diverse types of data including structured databases, unstructured text, and multimedia, pose significant challenges for organizations that want to derive meaningful insights from complex data. Conventional search and retrieval methods are increasingly inadequate for managing the complexity and immense volume of data today. Let's take a look at how generative AI (gen AI) can enhance retrieval strategies through language embeddings and source grounding, focusing on optimizing performance, speed, and scalability to effectively address these challenges.

To assess the effectiveness of these gen AI-driven strategies, we'll explore a critical intersection between financial services and environmental, social, and governance (ESG).

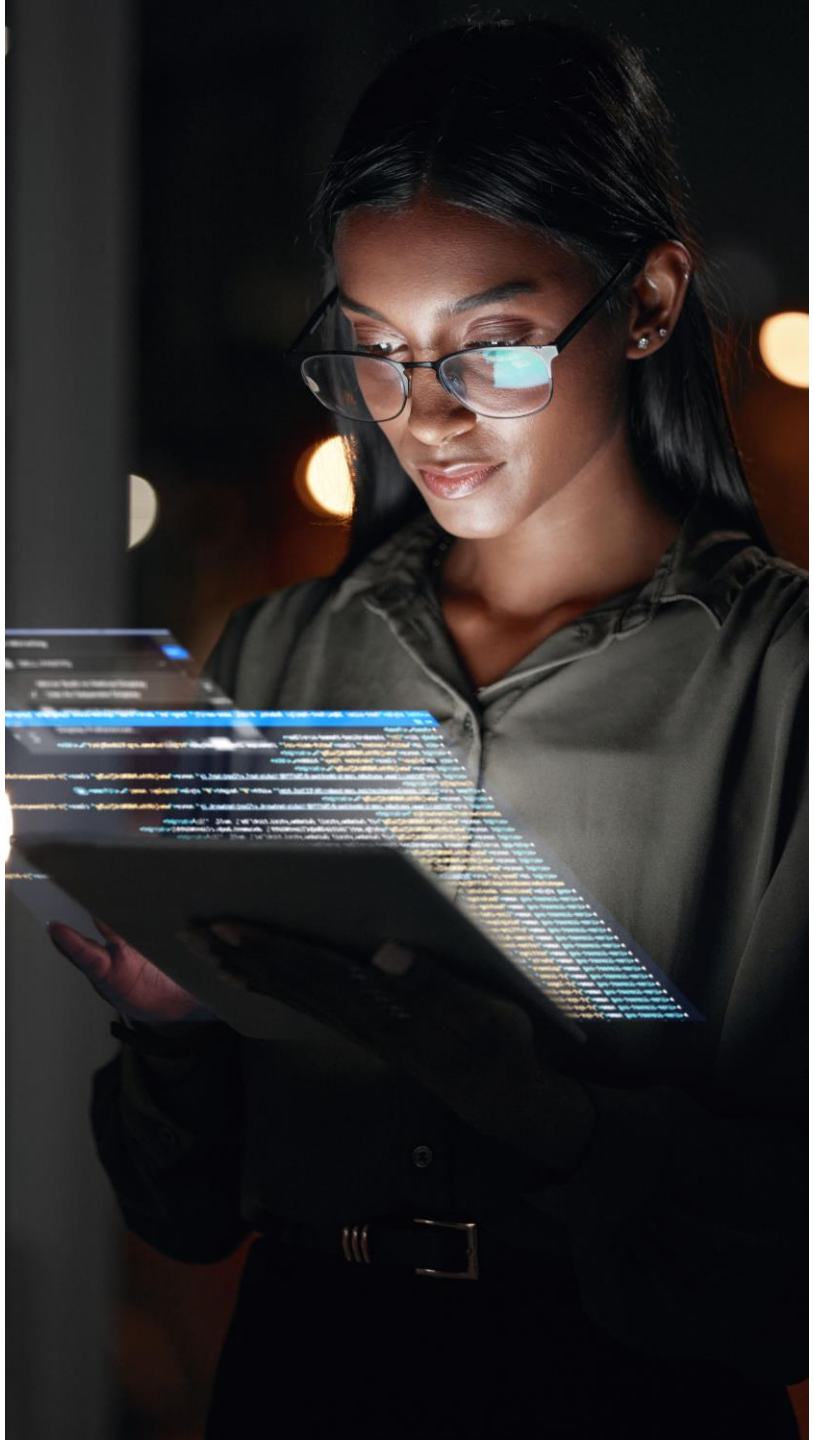
We'll specifically focus on extracting data from unstructured documents, such as banks' emissions reports and quarterly reports, and constructing a database from these data points that were previously difficult to access, demonstrating the practical applications and benefits of advanced data retrieval in the financial services sector.

Introduction

Data extraction has always been challenging, particularly when dealing with unstructured, inconsistent, and notably large amounts of data. Organizations have often relied on external data providers, which was not only costly but also not always up-to-date or live.

Alternatively, organizations had to build their own extraction pipelines, an endeavour that came with its own challenges. But with the advent of gen AI, the entire financial services industry has been disrupted, resulting in a lasting change in the field of data extraction.

Gen AI can autonomously analyze and interpret vast amounts of unstructured data with unprecedented accuracy and speed, using natural language processing and machine learning algorithms. These innovative capabilities include contextual understanding, pattern recognition and the generation of coherent data summaries, which significantly reduce the time and resources required to extract data.



Organizations that have attempted to implement gen AI solutions have quickly encountered new challenges, including:

Large language models (LLMs) may generate hallucinations – responses that are out of context – that result in unreliable outcomes.

Cost and speed constraints can result in limited scalability across extensive source databases.

Out-of-the-box LLMs and search engines are difficult to set up for the most suitable parameters

Let's take a look at varying retrieval and language model strategies that can offer innovative information retrieval methods for the financial services sector.

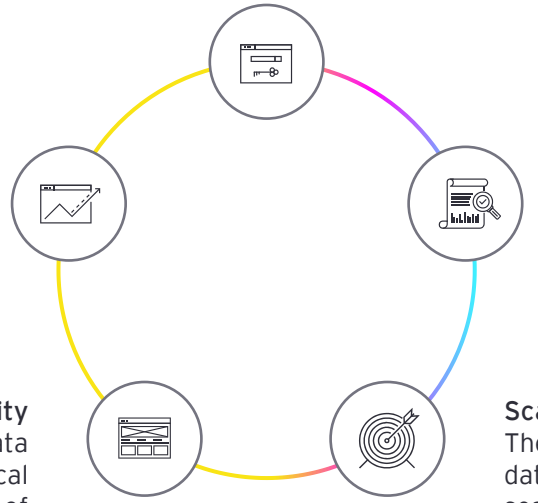
Current state and main challenges

The recent surge in data availability has rendered traditional methods of data extraction and analysis obsolete. These legacy systems, once reliant on manual keyword searches and static queries, struggle when confronted with today's vast, dynamic, and diverse data streams.

Key challenges in information retrieval include:

Keyword dependency
Limited to exact keyword matches, traditional systems suffer with the nuances of language, failing to capture context and semantic variations.

Velocity and volume barriers
Traditional architectures struggle to maintain performance and scalability amid the rapid growth of data speed and volume.



Static queries
Predefined queries lack the flexibility to adapt to new data types or unexpected querying requirements, hindering the discovery of insights in evolving datasets.

Unstructured data complexity
The predominance of unstructured data necessitates advanced analytical techniques beyond the scope of traditional methods.

Scalability challenges
The increasing volume and complexity of data outpace the capabilities of conventional search tools, leading to slower search responses and a strain on resources.

These challenges highlight the need for a sophisticated solution to data extraction and analysis. Such solutions should be designed to handle the intricacies of language, adapt to evolving data types, and scale in response to the increasing volume and complexity of data.

Gen AI and retrieval strategies

The process of data search, storage, and analysis is being revolutionized using advanced retrieval systems enabled by gen AI. These systems, characterized by their scalability and high- levels of performance, excel in real-time processing of various data types, including structured, unstructured text, numerical, and geospatial information. The use of sophisticated domain specific queries in these systems enable intricate and detailed searches, unlocking profound insights from extensive datasets. These strategies are integral for a wide array of applications including log and event data analysis, full-text searches, security intelligence, business analytics, and operational intelligence.

The pipeline of these retrieval systems is a comprehensive collection of tools that enhance the core functionalities. It combines language embedding models and source groundings, data transformation and storage (including vectors), and data search and retrieval, all within a single ecosystem. It also encompasses tools for data security and provides integration capabilities with other software, including various data sources and LLMs. This integration is particularly valuable in addressing the financial services industry's very nuanced challenges.

Throughout the pipeline, the technology was developed by EY's gen AI professionals and the technology was enabled by Elasticsearch[1]. For comparison, we'll compare the efficiency, cost, and speed between EY's approach and a naïve retrieval pipeline. Due to the distributed systems approach and overall design, EY's solution showed superior performance alongside Elastic's technology stack.

[1] Elasticsearch is an open- source distributed, RESTful search and analytics engine, scalable data store, and vector database capable of addressing a growing number of use cases. As the heart of the Elastic Stack, it centrally stores your data for lightning-fast search, fine-tuned relevancy, and powerful analytics that scale with ease.

Gen AI and retrieval strategies (Cont'd)

The end-to-end pipeline of the retrieval system is outlined below:

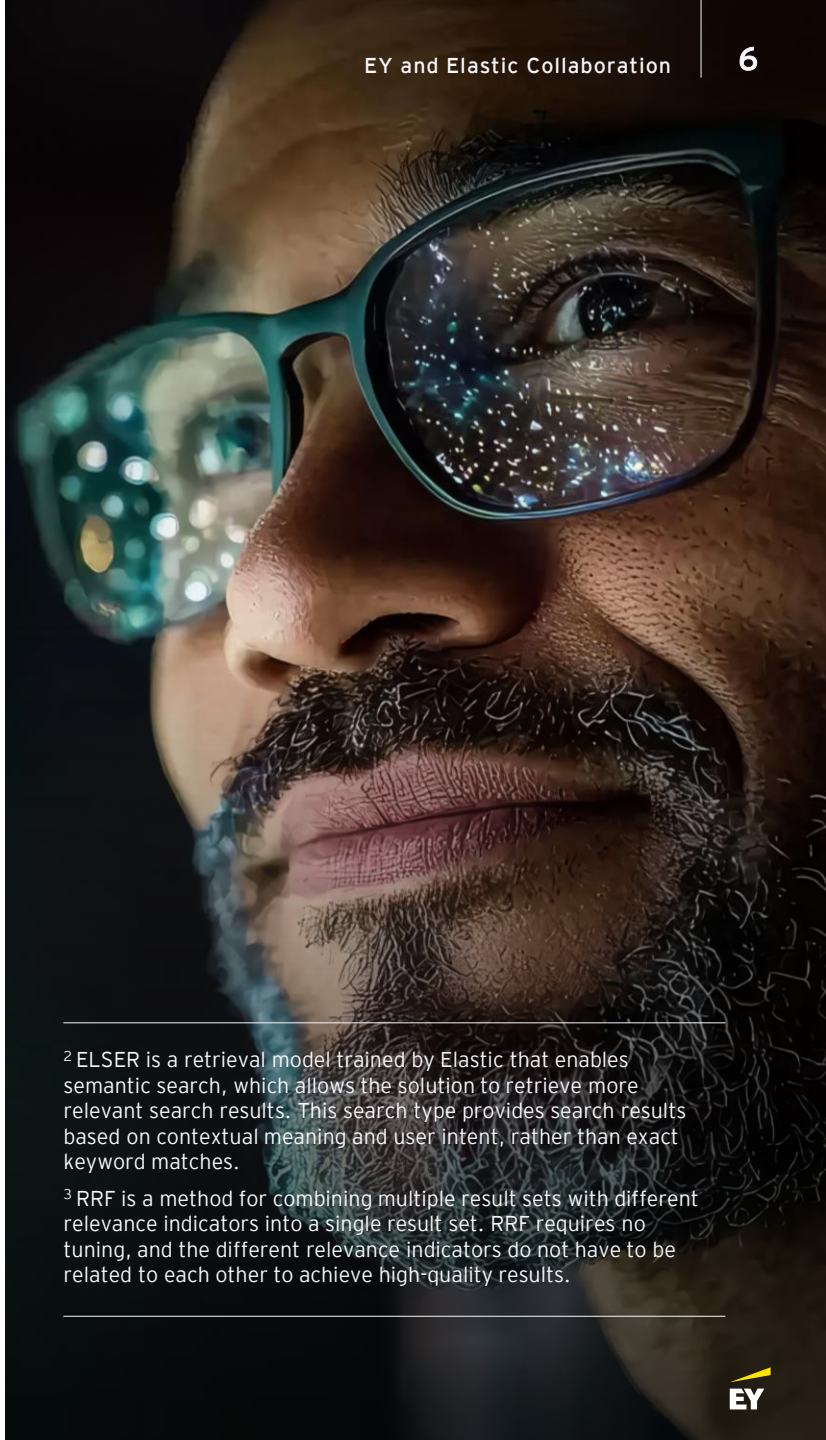
Vector store

A crucial component of the retrieval pipeline is the vector store, which is essentially a robust data storage system that can handle a diverse array of data types. This includes unstructured text, structured data, and dense vectors (embeddings). The vector store is designed to accommodate data both before and after it has been transformed by embedding models, making it a versatile tool in the pipeline.

Embedding models

These retrieval systems often employ embedding models, such as the Elastic Learned Sparse EncodeR (ELSER)[2] , to facilitate a retrieval model that offers enterprises the capability to execute precise semantic searches. These models help convert normal language into a space of vectors that are understood by retriever systems, emphasizing the understanding of context and the user's intent, and transcending the limitations of traditional keyword matching.

Harnessing a rich training dataset composed of high-quality question-answer pairs, these models enhance the efficiency of calculating similarity between queries and documents. This not only refines the accuracy of information retrieval, but also expedites the indexing process, improving the search experience for users



[2] ELSER is a retrieval model trained by Elastic that enables semantic search, which allows the solution to retrieve more relevant search results. This search type provides search results based on contextual meaning and user intent, rather than exact keyword matches.

[3] RRF is a method for combining multiple result sets with different relevance indicators into a single result set. RRF requires no tuning, and the different relevance indicators do not have to be related to each other to achieve high-quality results.

Gen AI and retrieval strategies (Cont'd)

Ranking models

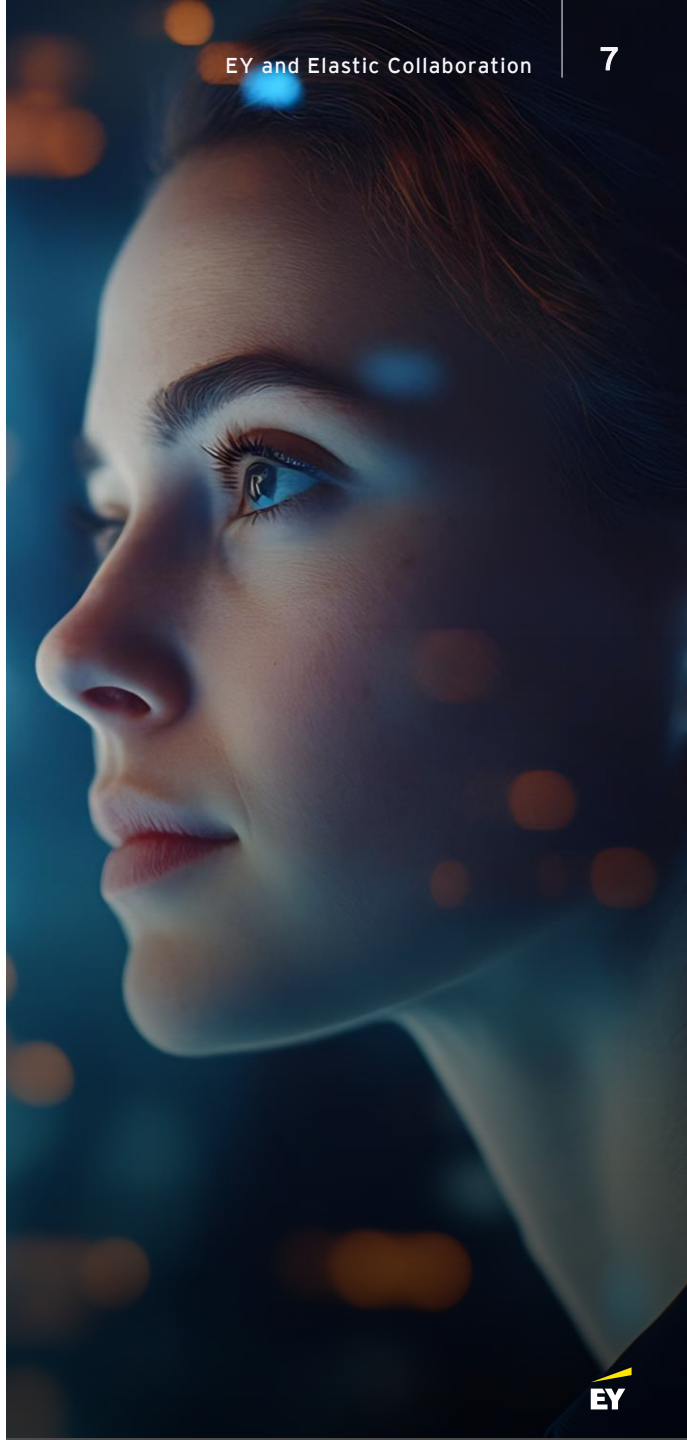
Ranking models, such as Reciprocal Rank Fusion (RRF)[3], are a key part of the retrieval pipeline. These models combine the strengths of multiple search strategies to improve result relevance. By merging outcomes from various queries and algorithms, ranking models improve the retrieval process for complex documents, making it easier to find pertinent information amid vast datasets.

Similarity searches

The use of methods like k-Nearest Neighbors (kNN) in the retrieval pipeline, allows for fast and efficient similarity searches. These methods find the "nearest" documents to a given query in the embedding space. This is especially useful for identifying documents with similar patterns or themes, even if the exact keywords are not matched, enabling a more intuitive search experience.

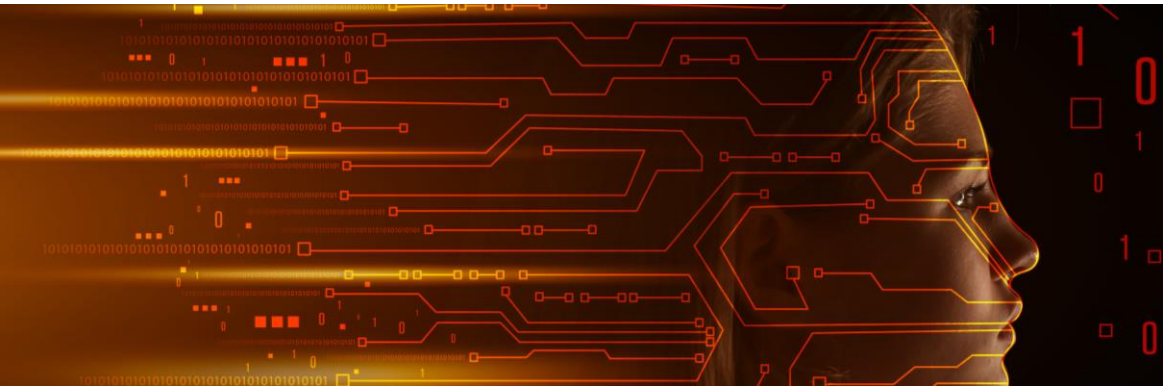
The components within the retrieval system pipeline allow for enhanced search capability that grasps user intent, stores and retrieves the most relevant data for complex similarity searches, and the enables the fusion of multiple search strategies to further refine the search relevance. The search engine's ability to effectively handle metadata and employ different retrieval methods for intuitive similarity searches further underscores its capacity to efficiently navigate and interpret complex data structures.

This process provides a powerful foundation for handling complex data, enabling businesses to more fully capitalize on their data's potential. Combined with solution engineering and industry expertise, it represents a transformative approach to overcoming the multifaceted challenges of data analysis in the digital age.



Use case implementation evaluation

1. Extracting ESG variables from annual ESG reports

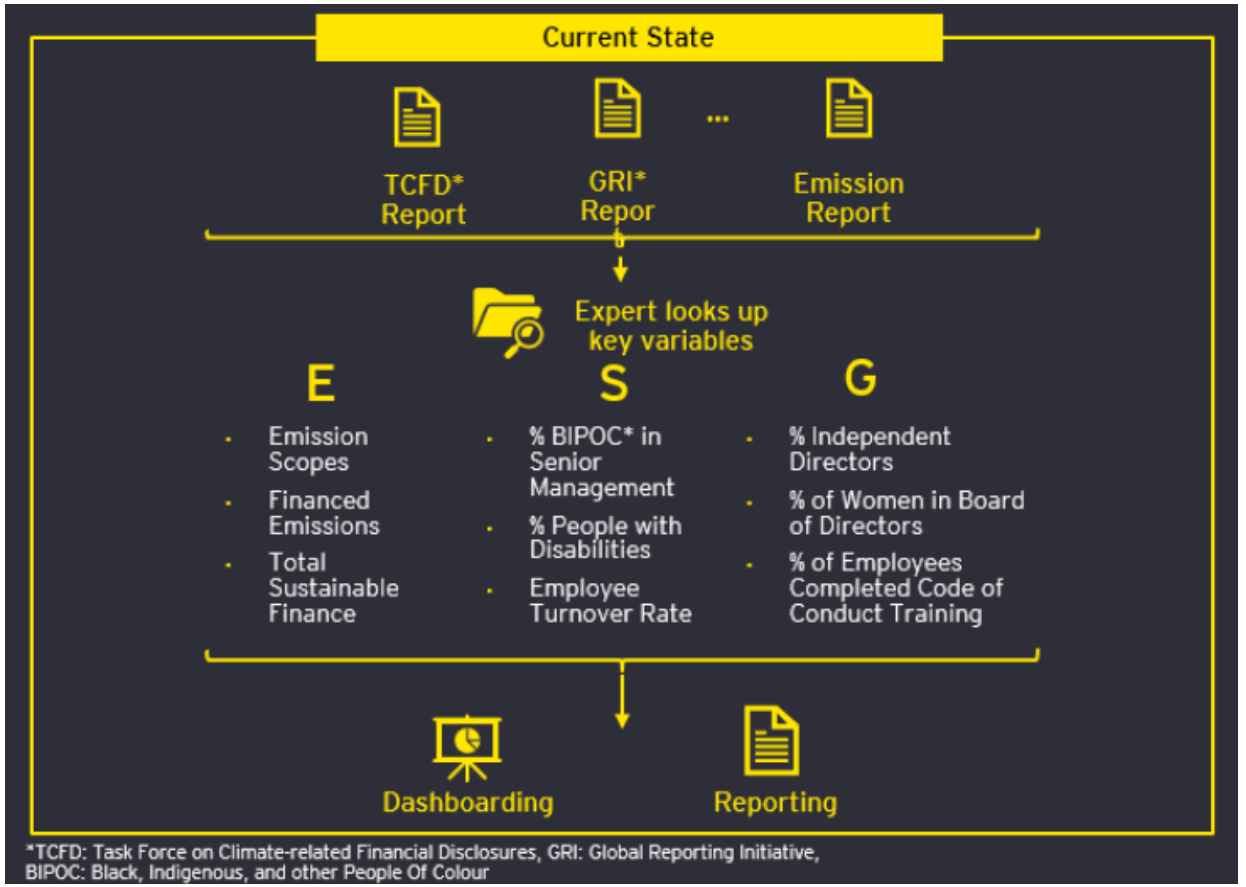


ESG reporting in the financial services industry is becoming an essential component of corporate transparency and accountability, reflecting a company's commitment to sustainable and ethical operations. These reports provide invaluable insights into a company's approach to managing risks and opportunities related to environmental, social, and governance issues.

As such, ESG data has become a critical metric for investors, regulators, and the public, who are increasingly factoring non-financial elements into their evaluations to discern material risks and identify potential for growth due to its varied nature and nuanced information.

The EY Climate Risk Stress Testing Survey underscores the challenges organizations face in integrating diverse external and internal data sources. Inconsistencies in report formats and varying levels of detail create substantial barriers to the effective indexing and retrieval of essential ESG data points, such as Scope 1, 2, and 3 emissions. These challenges are exacerbated by data gaps, the variability in data quality over time, and the lack of adaptability in ESG platforms, which struggle to keep up with the ever-evolving regulatory landscape.

In this use case study, we aim to extract ESG variables from publicly available annual ESG reports from the top Canadian banks. The key objectives in refining ESG data extraction include achieving high accuracy and speed to support critical financial and compliance decisions, help enabling scalability to handle the growing volume and complexity of ESG data, and maintaining flexibility to adapt to evolving reporting standards and frameworks .



Use case implementation evaluation (Cont'd)

1. Extracting ESG variables from annual ESG reports

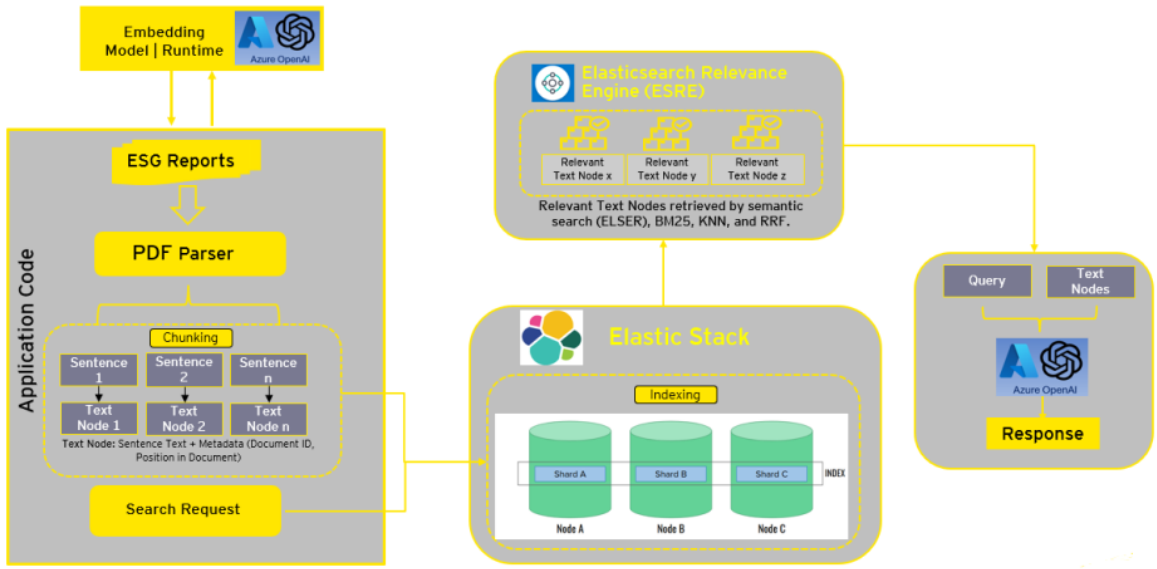


Figure 1: Workflow of EY's solution

EY's solution, enhanced by Elastic's advanced technology, offers a sophisticated solution to these challenges. By deploying the retrieval strategies, a robust and context-aware solution is crafted that not only streamlines data extraction from a variety of semi-structured reports but also provides the necessary flexibility and scalability to meet growing data demands and the rigor of comprehensive analysis. The advantages of this approach are significant. It improves the data retrieval process by improving accuracy and speed, boosts the flexibility and scalability of solutions derived from ESG metrics, and strengthens the foundation for informed sustainable business practices.

We establish a benchmark for ESG reporting that is in line with the ultimate objectives of efficiency, flexibility, scalability, and providing actionable insights for sustainable business strategies.

Figure 1 showcases the workflow of the EY solution by integrating the Elastic's stack with LLMs, highlighting the various components and their interconnections.

Retrieval-augmented generation (RAG) is an innovative AI technology that boosts the precision of information retrieval for complex documents like ESG reports using language embeddings and grounded sources. It achieves this by analyzing PDF documents, indexing, and chunking, – meaning dividing text into smaller "chunks" – ensuring comprehensive analysis and accessibility of detailed information within such documents, This enables precise identification and extraction of relevant information.

One limitation is that Naïve RAG's data processing (indexing and chunking) can slow down retrieval speeds. To address this, employing optimized search technologies can significantly reduce response times by refining the indexing strategy and enhancing the efficiency of data retrieval through sophisticated algorithms and distributed computing. As demonstrated in Figure 2, Elastic RAG can deliver responses up to three times faster than Naive RAG without compromising performance.

Use case implementation evaluation (Cont'd)

1. Extracting ESG variables from annual ESG reports

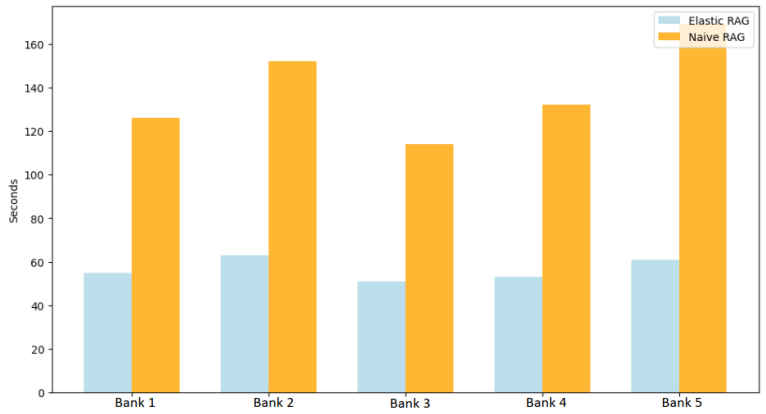


Figure 2: Processing time comparison: Elastic RAG vs Naive RAG

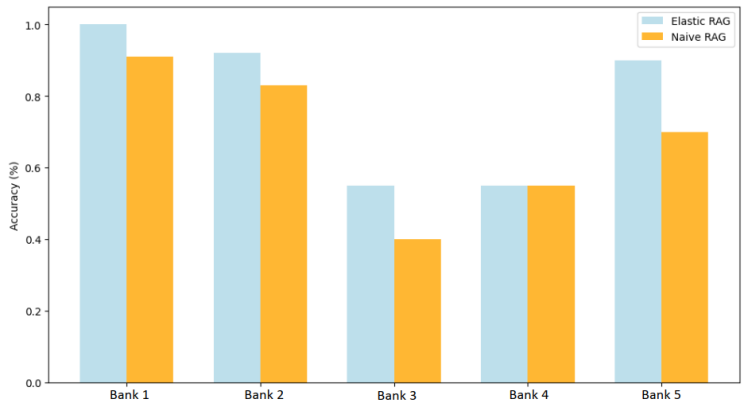


Figure 3: Accuracy Test comparison: Elastic RAG vs Naïve RAG

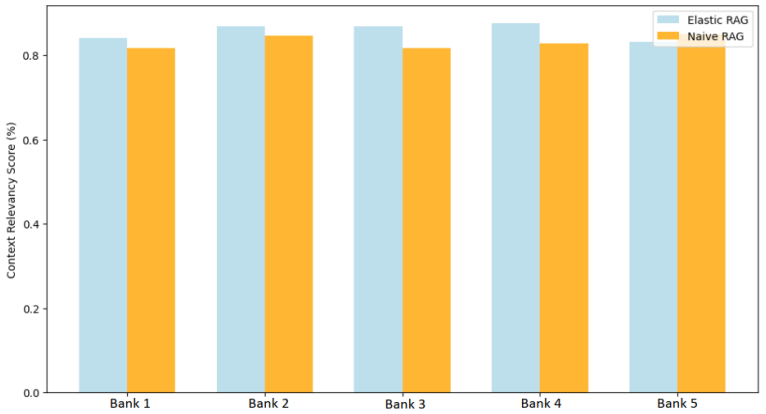


Figure 4: Context relevancy score comparison: Elastic RAG vs Naive RAG

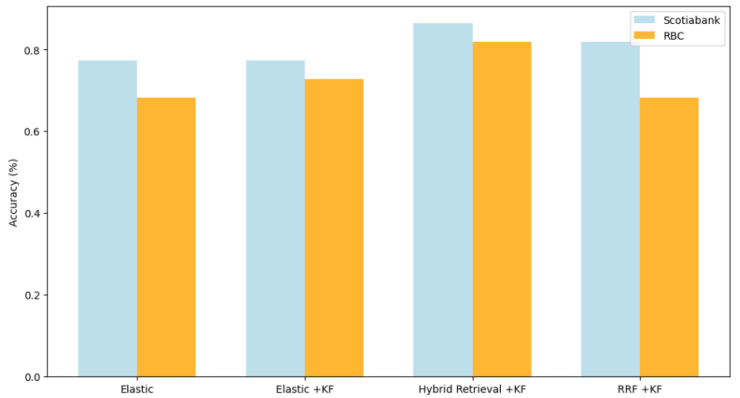


Figure 5: Scalability of Elastic solution

RAGFigures 3 and 4 illustrate a comparative analysis of Elastic RAG against a baseline Naive RAG across five Canadian banks in terms of context relevancy and accuracy. In both metrics, Elastic RAG consistently outperforms Naive RAG. The context relevancy score, which assesses how well the retrieved information matches the query's context, is notably higher with Elastic RAG for all banks. Similarly, Elastic RAG achieves superior accuracy, suggesting that it retrieves more relevant data with greater precision.

Figure 5 displays that various data retrieval methods, such as Elastic RAG with keywords filter (KF) and Hybrid Retrieval (vector search and BM25) with keyword filter, consistently maintain high accuracy levels, which can effectively address the challenges posed by growing data volumes. This shows the system's robustness and adaptability, with Elastic RAG with keyword filter being particularly effective for refining searches. The integration of different search techniques within the search framework ensures that the system remains efficient and accurate while handling more ESG reports, proving that scalability doesn't compromise quality.

The fusion of search technologies with generative AI, represents a significant advancement in the field, providing a scalable and precise method for data extraction from these reports. This approach doesn't just keep pace with current demands but is forward-thinking, positioning it as a leader in innovation and setting high standards for future developments in ESG data processing

Use case implementation evaluation (Cont'd)

2. Extracting financial variables from financial reports

In the realm of financial analysis, extracting more than 40 financial variables from quarterly reports is an intricate endeavour, which is further magnified when deploying LLMs– which are typically attuned to unstructured text–to analyze structured table data found in financial documents. These reports feature tables filled with data organized in rows and columns, representing a unique challenge for LLMs that are adept at navigating textual data but less so with the intricate relationships and numerical nuances of tables. The extraction of financial variables from these reports is a more intricate task than that of ESG reporting, as it typically involves a greater number of variables. This represents a significant increase in data complexity when compared to ESG data extraction, making it a particularly challenging endeavor due to the sheer volume and intricacy of the financial information presented.

The EY approach to these challenges employs powerful search capabilities along with advanced table summarization techniques. This includes employing chain-of-thought and chain-of-verification processes to enhance the accuracy of the extracted data. Drawing on the experiences from ESG reporting, we developed a hybrid retrieval system that combines vector search with the BM25 algorithm, significantly increasing the reliability and precision of the data extraction process.

Use case implementation evaluation (Cont'd)

2. Extracting financial variables from financial reports

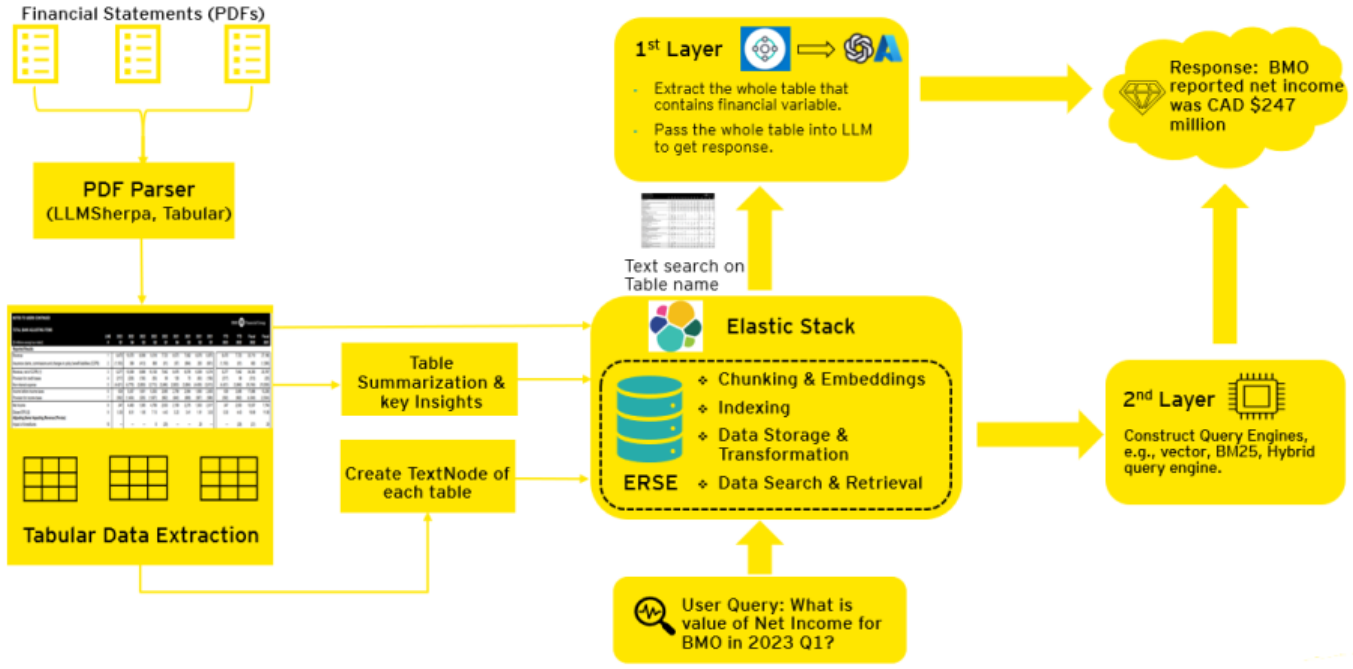


Figure 6: EY's solution of financial Variables extraction

Figure 6 illustrates the workflow of the financial data extraction solution. The process begins with the analysis of PDF Financial reports to accurately capture table-based data. It then transforms that data into a structure that is more suitable to LLM processing, using sophisticated prompting techniques to enable complete interpretation of table information. This solution incorporates a two-tiered retrieval mechanism. The initial phase involves the extraction of complete tables, guided by contextual cues, which are then analyzed by LLMs. Advanced query mechanisms are deployed, including a hybrid query engine that is part of an advanced search ecosystem.

Use case implementation evaluation (Cont'd)

2. Extracting financial variables from financial reports

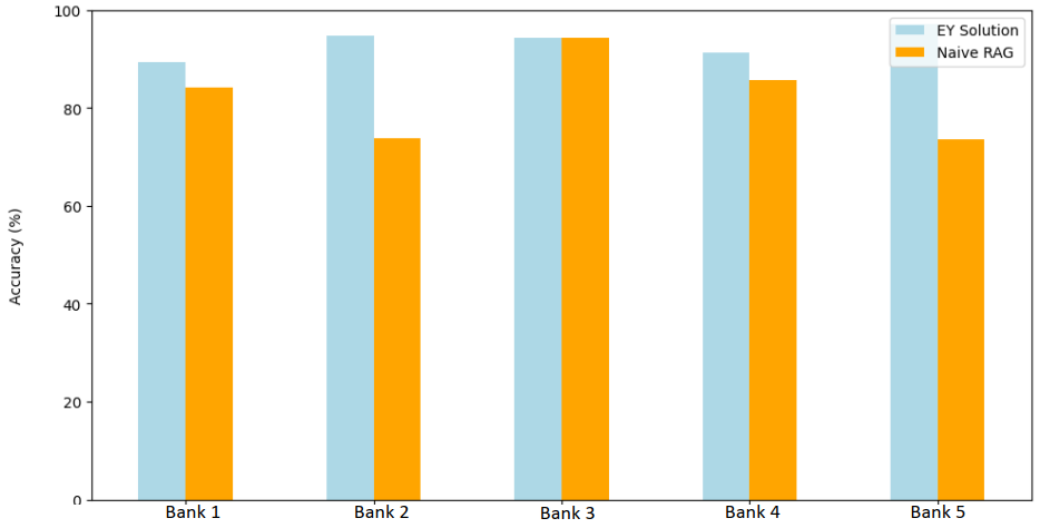


Figure 7: The performance of the EY solution in 2023 Q1 Supplementary Financial Reports

Figure 7 showcases the performance of the EY solution in 2023 Q1 Supplementary Financial Reports. In summary, the EY solution has made remarkable strides in accuracy enhancement, illustrating significant performance boosts where some instances reveal accuracy enhancements of nearly 24% compared to traditional RAG methods. This advancement not only refines the data extraction process but also elevates the quality of insights extracted from financial reports, setting a new standard for both efficiency and robustness in the realm of financial data analytics

Conclusion

Generative AI plays a pivotal role in revolutionizing data retrieval in the financial services sector. It showcases how integrating advanced search technologies with AI can greatly improve the extraction of complex data from diverse documents, setting new standards for accuracy, speed and scalability. The demonstrated solutions not only address current data analysis challenges but also pave the way for future innovation, emphasizing the importance of adopting AI-driven strategies to fully capitalize on the growing volumes of data in informed decision-making processes.

Authors

EY

Yara Elias

EY Canada Senior Manager,
Risk Consulting

Kiranjot Dhillon

EY Canada Senior Manager,
Risk Consulting

Vishaal Venkatesh

EY Canada Manager,
Risk Consulting

Morgan Wang

EY Canada Senior Consultant,
Risk Consulting

Ahmad Ghawanmeh

EY Canada Senior Consultant,
Risk Consulting

Elastic

Ian Santee

Elastic, Senior Enterprise
Account Executive

Thaddeus Walsh

Elastic, Principal
Solutions Architect

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multi-disciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

4713104

ey.com/ca