Airflow Pipeline And Automated Data Extraction For A Smart Assistant

CASE STUDY

About the Client

Our customer is an energy technology startup based in Texas, USA. They created a device-independent smart assistant that runs on any browser and has been custom built just for upstream Oil and Gas. Any company can connect to limitless internal and public data sources, extract high-value insights, and allow for a natural conversation with the data. The smart assistant helps companies make brisk decisions, by organizing the collective knowledge within the company and making it accessible.

Client goals

The company's co-founder - Schlumberger and Shell alum got curious about data, analytics, and machine learning during the major downfall of oil prices in November 2014. It led him to ask questions about why it was happening and can it be predicted in a better way?

He wanted to simplify complex industrial knowledge that was easy to access. He and his colleagues have a long-term goal to create an ecosystem of workflows around the smart assistant so that everyone in an organization can ask the tough questions and get back the answers without any delay.

Problem - Manual and Slow Data Extraction

The smart assistant was using Natural Language Processing to integrate with multiple systems used by a company to unstructured data from documents, emails, and much more. The then-current version of the search engine had manual data extraction - which was extremely time-consuming. They were using a third party API to extract the financial data.

JTC's Collaboration

JTC has been a technology partner of the energy tech startup under the staff augmentation model since September 2020.

Building Airflow Pipelines and Automated Data Extraction

We developed an Airflow Pipeline to extract text from unstructured documents and upload it to an elastic search database and created an external API for users outside of the team network. The Airflow pipeline automated workflows and managed boring and redundant manual tasks.

We built airflow pipelines for:

- Extracting data from uploaded and emailed documents
- Extracting companies investor presentations and transcripts
- Extracting information from financial documents
- Extracting news data from peer websites

We added a new feature where anyone from the organization can compare financial data on a quarterly basis or look into the yearly growth.

We did this by designing elastic search mapping and queries for refined results. Now anyone can ask questions like: “What was NRG’s third-quarter revenue growth percentage” and they will get the reply at the same moment in a conversational way. The more questions you ask the assistant - the smarter it gets.

Another feature added by our team is to develop a process where the assistant nudges you with questions that you didn't think of asking. This can provide relevant insights and, hopefully, uncover new knowledge from the data.

The team also reinvented the backend data structure with dependency injection and storage abstraction. - increase in speed, reduced latency, restructured the backend using service based architecture. We also redesigned Data Explorer for viewing different connected data sources in an improved manner. We improved table extraction logic for extracting tables from unstructured PDFs and documents on top of existing libraries (Airflow).

Results

Response time has been reduced drastically and the UI/UX is better in the new version. With the new infrastructure development, the user is now able to ask questions based upon suggested, and typeahead questions. The user can also see financial information about his peers and compare them. And filter and sort responses on the basis of different tags.

Build your business with us,
Feel free to contact us

☎ +91 9876543210

@ enquire@jtc.com