GENERATIVE BI

A TRANSFORMATIONAL APPROACH TO CONVERSATIONAL ANALYTICS

Ian Macdonald, SNR Director, Product

Introduction

Imagine the following conversation between a business user and an AI. It might go something like this:

User: How can I double our company's revenue?

AI: To double revenue, consider focusing on the following areas: Conversion Rate, Activation Rate, Retention Rate, expanding product offerings, entering new markets, and or changing pricing strategy.

User: Extract the Conversion Rate.

AI: I'm sorry, could you please provide more context? Which conversion rate data are you referring to? Currently, we have already predefined these metrics: Web to Signup Conversion, Signup to Paid Conversion.

User: Show me Web to Signup Conversion Rate.

AI: Our average Web to Signup Conversion Rate for the last 3 months is currently 5%. By increasing the conversion rate to 12%, you have the potential to double your revenue.

User: Add this chart to my dashboard "My KPI" and set up a daily alert to be sent to me via Teams.

AI: Sure, I have added the chart to your dashboard "My KPI" and set up a daily alert to be sent to you via Slack. You will receive a notification every day with the latest conversion rate data for San Francisco, New York, and Seattle.

From the above conversation, it is apparent that the AI Assistant knows about the "available metrics" and the causal relationship between company revenue and these metrics. It knows how to generate a database query given the user input, and how to run the metrics query to get the output data. It can unravel the visual output in natural language, and it can save the chart output individually or incorporate it into a dashboard.

This sounds like something that might seem fanciful and found only in the pages of a science fiction novel, but Pyramid delivers this capability right now.

We call this "Generative BI" or "Gen BI", the application of Generative AI technology, Large Language Models (LLMs) and other Machine Learning techniques to the Business Intelligence and analytics process.

These characteristics and other requirements for conversational analytics to be effective in an enterprise can be summarized as five key capabilities.

- **An AI strategy that works with all your Data:** Gen BI that works directly off any data source.
- **Multi-LLM Strategy:** bring the right LLM's to the right problem.
- **Question Without Concern:** no private data is sent to the LLM, just the semantic model meta data.
- **Sophisticated Answers to Complex Questions**: complex analytics handled through the simplicity of conversational natural speech.
- **Embedded Gen-BI:** Gen BI that flows through to embedded content in other applications without coding.

This paper outlines how the Pyramid Platform achieves these goals and the underlying technologies that make Gen BI conversational analytics available today.

Background: LLMs and Private Data

Limitations and restrictions of LLMs

The past eighteen months have seen an extremely rapid evolution of the capabilities of Generative AI and Large Language Models that are able to interpret natural language questions and respond in a humanlike fashion with great accuracy. While these LLMs work well with qualitative questions and answers, their ability to handle more quantitative and analytic enquiries effectively is questionable.

Using natural language to formulate analytic questions of data requires a subtly different approach. While the language capability of LLMs needs to be exploited, additional processing of the analytic component is required to produce accurate results that are consistent and repeatable.

As their name implies, Large Language Models are predominantly concerned with parsing and generating text, extracting meaning, and generating responses that mimic human interactions. But ask an LLM to aggregate data or perform more sophisticated math, calculations or analysis and its limitations become all too apparent. The results are often inaccurate, inconsistent, or just downright wrong.

Then consider the data. Most organizations today are dealing with huge volumes of data and often the interesting facts are only revealed by examining the detail. Moving large volumes of data to and fro, between the data repositories and the LLM services would be slow and inefficient, not to mention highly expensive – assuming it was even possible.

In addition, sensitive, private data has no place being sent to a public service for analysis. Data that reveals a company's standing sent to a public service (hosted by another company) could potentially compromise that organization's reputation or credit rating. Personal details or health data often have legal constraints on how that data is handled and where it may be stored. Sending that data to an LLM would almost certainly result in lawsuits from both government regulatory authorities and individual or class actions.

At the same time, accurately interpreting questions about data held in private databases (on premise or private cloud) requires information about that data to be made available to the LLM. This is more feasible as no true, actual data is sent, just the data that describes the data to be analyzed, in database germs the schema or semantic model of the data.

We need to find an approach that handles the limitations of LLMs when it comes to analytic processing: that does not require actual data to be sent to the LLM; can provide semantic information about the data to the LLM; and can access private data for analysis with a full range of analytic processing capability.

The Analytic Recipe

The solution to the above limitations is take the analytic processing of the actual data out of the LLM and instead use the LLM for what it is good at: interpreting and generating natural language in a humanlike fashion.

For the LLM to do that effectively, it needs information about the data that is being questioned. In the Pyramid Platform, this is available as a Semantic Model, either generated by the Pyramid Platform itself or directly consumed from other analytic databases with their own semantics, like SAP BW.

This meta data can be safely sent to the LLM as it does not contain any actual data, just a description of the database structures that contain the actual data. This helps the LLM to better interpret the users' questions about numeric measures (such as sales, margin, profit etc.) as well as what to analyze those measures by attribute (such as country, year, location etc.)

What the LLM is asked to produce is a series of steps or a 'recipe' for how to produce the analytic results requested by accurately interpreting the users' questions. To achieve this the LLM must be asked to respond in a particular way using a technique known as "prompt engineering". Pyramid's sophisticated prompt engineering ensures that the analytic steps returned by the LLM are consistent and intelligible.

What is then needed is a capable 'robot' able to execute the 'recipe' – following the steps defined, generate queries and visualizations that are called for, and assemble those into interactive dashboard and/or formatted reports. **The more powerful the robot, the more complex of a recipe the LLM can produce**.

Of course, Pyramid is exactly such a platform able to perform these steps and deliver to the user a finished analysis or assembled dashboard or report.

Gen BI: Pyramid and AI Integration

The “analytic recipe” approach described above bypasses the analytic limitations of LLMs that have been identified. But what about the other key features of an enterprise conversational analytics solution?

Here is where both existing and new capabilities in the Pyramid Platform combine to satisfy these additional needs.

An AI strategy that works Directly on Data.

It has already been established that moving large amounts of data for analysis to an LLM would be slow, expensive, and fraught with governance risk. The same can be said if the data must be moved to a specialist data framework for the analysis to be performed. It would be far better to query the data where it resides, whether that be a standard relational database like Oracle, a cloud-based data service like SnowFlake or even an existing analytic database like SAP BW or SAP HANA.

Of course, this is precisely what the Pyramid Platform provides with its intelligent and highly performant PYRANA query engine. PYRANA takes the analytic requirements defined in the analytic receipt by the LLM and generates the queries and operations required to return to correct answer.

PYRANA achieves this by leveraging a shared semantic model of the data that is being addressed. The semantic model describes the database, tables, and relationships that the user is querying, as well as additional functional aspects such as hierarchies, security, and flexible meta-data management.

PYRANA uses the semantic model to generate the SQL required to satisfy the analytic needs of the user. This may be a single SQL query, or in more sophisticated scenarios, multiple SQL queries that are executed in parallel against the underlying relational database. PYRANA then orchestrates the returned answer sets to complete the analytics needed before presenting the results through the Pyramid UI as dynamic visualizations of the data concerned.

This approach allows the Pyramid Platform to scale to whatever the underlying database technology can handle, **working directly on the data source**, with no limitations or restrictions on the analytic sophistication on offer.

Where the underlying database has its own semantic model, such as SAP BW, SAP HANA, or Microsoft SQL Server Analysis Services, the PYRANA engine consumes the underlying native semantic mode directly and uses that to generate MDX, the native querying language of multidimensional analytic databases, or SQL in the case of HANA.

By using the above techniques, PYRANA can directly query ALL your data in place, without loss of analytic functionality or sophistication, capitalizing on and making full use of the time, effort and money spent creating these data resources.

No more moving of data, copy managing, ingestion into proprietary data engines or recreation of analytic models is required.

All this is driven by the users original, natural language question.

Multi-LLM Strategy

When consulting an expert, it makes sense to talk to the right one for the right problem. For health matters - a doctor, for legal advice - a lawyer, or for oil and gas processing - a chemical engineer. It makes sense then, that an LLM designed for that domain will perform better than a general purpose LLM. So that the data to be analyzed is specific to verticals or functions, the questions asked, and the responses given will contain domain specific concepts, terms and even acronyms.

Pyramid provides the ability to connect to multiple LLMs and even to bind a particular LLM to a particular data model. This ensures that the right LLM is used for the right kind of data. It also allows the Pyramid ChatBot to use a specified LLM for language interpretation and the generation of the response.

The Pyramid platform also supports full speech to text and text to speech support and again, the multi provider option is available, allowing a choice of engines to perform this function. It is even possible to define the “mood” of the generated output from formal to a more casual conversational aspect.

While domain specific LLMs are perhaps still in their infancy, Pyramid already provides the ability to connect to multiple LLMs, thus future proofing conversational analytics implementation.

In addition, it means a Pyramid customer is not locked into a particular LLM provider and offers a flexible and independent LLM strategy.

Question Without Concern

It is important to emphasize that Pyramid at no time sends the actual data to be analyzed to any public LLM service. Pyramid breaks down the question asked into specific prompts for the LLM to process. It will send this prompt with the semantic model that describes the data to generate a response that is accurate and predictable in terms of the analytic steps that are then to be executed by the Pyramid Platform.

Pyramid users can be assured that the governance and security of their valuable and confidential data is always preserved.

Of course, this is assuming that the LLM concerned is a public service. It is highly likely in the not-so-distant future that private or locally installable LLMs will become available, in which case Pyramid's multi LLM strategy already has this covered.

Sophisticated Answers to Complex Questions

Pyramid's natural language interface functionality now includes speech input and output. This doesn't just simply apply to asking a question about a specific data set, but works across the Pyramid Platform, extending to designer and analytical capabilities.

This means that not only can a user ask analytic questions of their data in any database, but also ask Pyramid to construct new content such as dashboards and formatted, scheduled reports through the same natural language interface.

This allows instructions like:

> "Create a sales analysis slide showing me quantities sold by occupation and education, a different comparison of returns and customer happiness by state and last an analysis of country sales performance by promotion using revenues, expenses and margin, and I want it sliced by years with slide insights."

This generates the queries, visualizations, KPIs, filters and insights and assembles them into a working, interactive dashboard. This is a game changer for self-service, conversational analytics.

Once the dashboard is running, the ChatBot switches to run time mode and allows the user to interrogate the analyses presented for further analysis. This includes refinement and follow up questions in the context of the answer, changing of the output visualizations and many other capabilities. Something not offered by any other solution.

Of course, the user might not have English as their first language. The Pyramid natural language interface delivered through the ChatBot respects the language the user is using and will understand and respond in that language. In fact, almost any language is supported.

Embedded Gen-BI

With the rapidly growing market for embedding analytics into online applications and microsites, there is an over-arching need to access powerful functionality, but without the weight of a full application. Pyramid's embedded content delivers a collection of deep AI capabilities DIRECTLY in the hosting app, anchored by the unique ability to deliver conversational analytics through the same embedded content.

Pyramid's embedding approach injects the analytic content directly into the webpage, rather than making use of cumbersome and resource intensive Iframes. This allows many analytic objects to be included into one page whilst maintaining exceptional runtime performance. Pyramid will auto generate the required code, which simply needs copying and pasting into your Web application.

The best bit, though, is that any associated Gen-BI capability follows along automatically. This includes the use of Pyramid's LLM enabled ChatBot, immediately delivering cutting edge AI driven conversational analytics into your application with minimum effort.

Gen-AI vs Gen-BI

What is the difference between Generative BI and Generative AI? While they both operate off the power of AI technologies like large language models, there is a big difference in how they operate. The best way to understand this is by illustrating the following scenarios on how one would go about implementing generative AI in a data and analytics platform.

SCENARIO 1: Gen-AI driven analysis

The user would log into a chat or natural language interface and enter in his/her question that they want answered. The interface in this case utilizes the Python programming language and a large language model (LLM) to generate the desired result.

An example question could be, "Please analyze sales by country and list the top 3 countries."

The question would then be sent to a large language model (LLM) to understand and interpret the question. This interpretation would then be generated into python code/script to be able to execute on the request.

While this may sound fairly simple, it's not. This script would need to include all the Python code to do the following:

- Turn the question into Python Script to be able to query the data.
- Perform the analysis with regards to the question that had originally been asked.
- Visualize the results back to the user.

Once the script is generated, it would need to be compiled and run, connecting with the data source, returning the results back and then visualizing the content in the form of a static set of analysis and visuals.

This is not where it ends as there are a few more hurdles one has to overcome to get this to work. The database to be connected to needs to be specified, along with the tables or schema required. This somehow needs to be entered into the code and managed.

The Python code would then need to be executed and where is that going to happen? What if you wanted to change the database being connected to?

Assuming this is all in place, and the user was comfortable with the process, there is an even bigger question and that is the "Next Question".

What if the user wanted to ask another question based on the results from the initial question? And what about the question after that?

This would have to be followed by creating some manual code and passing that back into the process that was just described all over again.

Scenario 1 is complicated, manual and can't be utilized by your everyday user.

SCENARIO 2: Gen-BI driven analysis

The second scenario showcases how the Pyramid platform goes from turning the generative AI process described in Scenario 1 and delivering Generative BI where all the steps are handled automatically.

To begin with, the user would log into the Pyramid Platform and utilize the natural language interface through typing or through speech to text.

Once the request is received, Pyramid automatically retrieves all the details for how to connect to the database required and its structures.

Pyramid will then utilize the Large Language Model, which can be interchangeable depending on the task, that will then generate what Pyramid terms a 'recipe' with all the information that Pyramid requires on how to execute the request on the platform.

The recipe is then passed to Pyramid's query engine (PYRANA) which includes all the capabilities to handle and execute the request.

Critically though, this is dynamic and can be done seamlessly time and again allowing users to engage with follow up questions and produce and reproduce interactive content as visualizations, text or speech.

Not only does the flow of events happen seamlessly to the user, it happens over and over again without any intervention – in other words, it iterates flawlessly until the user is satisfied with what they require.

Summary

This paper has attempted to convey the high degree of integration and leverage of Generative AI Large Language Models in the context of sophisticated analytics and analytic content building that the Pyramid Platform now supports.

The imagined business user and AI conversation outlined in the introduction is now a reality with the Pyramid Platform.

The ideas and implementation described above are simply the start of a revolution in conversational analytics, a revolution that Pyramid will remain at the forefront of for the foreseeable future.