Eckerson Group | White Paper

When a Data Catalog is Not Enough

The Case for the Self-Service Analytics Workbench

BY JOE HILLEARY

JANUARY 2021

RESEARCH SPONSORED BY PROMETHIUM

Promethium

About the Author

Joe Hilleary is a writer and a data enthusiast. He believes that we are living through a pivotal moment in the evolution of data technology and is dedicated to helping organizations find the best ways to leverage their information. With a background in both analytics and the liberal arts, he crafts clear, articulate narratives on technical topics that empower stakeholders to make informed decisions. Hilleary is a Research Analyst at Eckerson Group.

About Eckerson Group

Eckerson Group is a global thought leader that helps organizations get more value from data. Our research and consulting experts think critically, write clearly, and present persuasively about data analytics. They have substantial experience in the field and specialize in data strategy, data architecture, data management, data governance, data science, and data analytics. Organizations rely on them to demystify data and analytics and develop business-driven strategies that harness the power of data. **Learn what Eckerson Group can do for you!**

About This Report

To conduct research for this report, Eckerson Group interviewed industry experts and vendors. The report is sponsored by Promethium who has exclusive permission to syndicate its content.

Table of Contents

Executive Summary

Data catalogs provide valuable information about data assets, but businesspeople want answers to questions, not just metadata. An emerging technology —the self-service analytics workbench—extends the data catalog with other self-service functionality and built-in workflows to meet this need. It supports the entire lifecycle of a business question from the moment a business user poses it until a data analyst or engineer presents a visualization of the answer.

A self-service analytics workbench integrates a variety of analytics capabilities into a single platform. These include data catalogs, data virtualization, data preparation, and data visualization. These functionalities support every step in the process a data analyst or engineer goes through to answer business questions: search and examine data, extract and join tables, format and transform data sets, and visualize and publish results. In addition, self-service workbenches provide built-in workflows that connect business users who have questions with data analysts and engineers who can answer them.

By consolidating multiple tools into a single workflow-enhanced platform, self-service analytics workbenches speed time to insight and improve communication between business users and the data analysts and engineers who support them. They eliminate the need to buy, learn, and integrate multiple disparate tools to enable the question-to-answer workflow, making the process of answering ad hoc questions more natural and efficient.

Key Takeaways

> Data catalogs are not a complete solution for self-service analytics.

> Merging the tools required for self-service into a single platform increases efficiency.

> A question-based workflow enables technical teams and business users to collaborate better and answer questions faster.

Recommendations

> Investigate the steps in your team's process for answering ad hoc questions.

> Identify bottlenecks in your question-to-answer workflow that slow the process of finding answers.

> Consider replacing purpose-built analytics tools (i.e., data catalog, data preparation, data analysis) with a self-service analytics workbench.

Introduction

Every company wants a data catalog these days, but what they really need is a self-service analytics workbench. Data catalogs provide information about data assets, which is great, but business users want answers to questions, not metadata. Answering business questions currently requires numerous purpose-built tools: a query engine, a data warehouse, tools for data preparation and visualization, and, yes, a data catalog. Although important, a data catalog is just one piece of the puzzle. On the other hand, a self-service analytics workbench consolidates five or six tools into a single platform that supports the entire self-service analytics workflow.

Self-service analytics workbenches put business questions front and center. They recognize that data's value comes from providing insights to the business. Their built-in support for the question-to-answer workflow reflects this emphasis by supporting all the steps needed to answer a question. The alignment between structure and function increases analyst efficiency, so they can answer more questions in less time.

This report demonstrates the shortcomings of a multi-tool approach to self-service. It identifies the limitations of traditional data catalogs and compares question-to-answer workflows with and without a self-service analytics workbench. Finally, it describes the characteristics of a self-service analytics workbench and introduces one of the leading vendors in the space—Promethium.

The Role of a Data Catalog

Use Cases

Data catalogs gather and store metadata about data assets. Data analysts use catalogs to find and evaluate sources for their analyses, while data stewards and curators employ them to govern and manage data sets and other artifacts. Although these functions are important, they make up only two steps in the question-to-answer workflow.

Much of the information in a data catalog comes from auto-population. Data catalogs crawl an organization's data sources and automatically capture metadata, such as file and field names, locations, usage, and lineage. Many catalogs also pull relevant information from business glossaries and use machine learning (ML) to auto-tag data and suggest relationships between assets. Analysts add detail through tags, comments, and reviews, while data stewards contribute information about data ownership, certification status, and permissions.

Data catalogs crawl an organization's data sources and automatically capture metadata.

Data exploration. Data catalogs enable analysts to explore multiple data assets quickly to find those that best meet their needs. They can search populated catalogs by keywords, business terms, facets, tags, or file names. They can evaluate data sets by looking at attributes such as field names, field values, cardinality, and other statistical information. Sample data, curator notes, user ratings, and comments provide additional detail, while links and tags enable analysts to move rapidly between related assets.

Data governance. Data catalogs serve an increasingly important role in data governance. Data curators can flag data sets of dubious quality within a catalog and add comments about the proper use of particular assets. They can also use certifications to guide analysts from across the organization to the same sources for key performance indicators (KPIs). This reduces data silos and helps create a standard set of enterprise metrics. Finally, catalogs allow stewards to restrict access to sensitive data by roles or groups.

Shortcomings

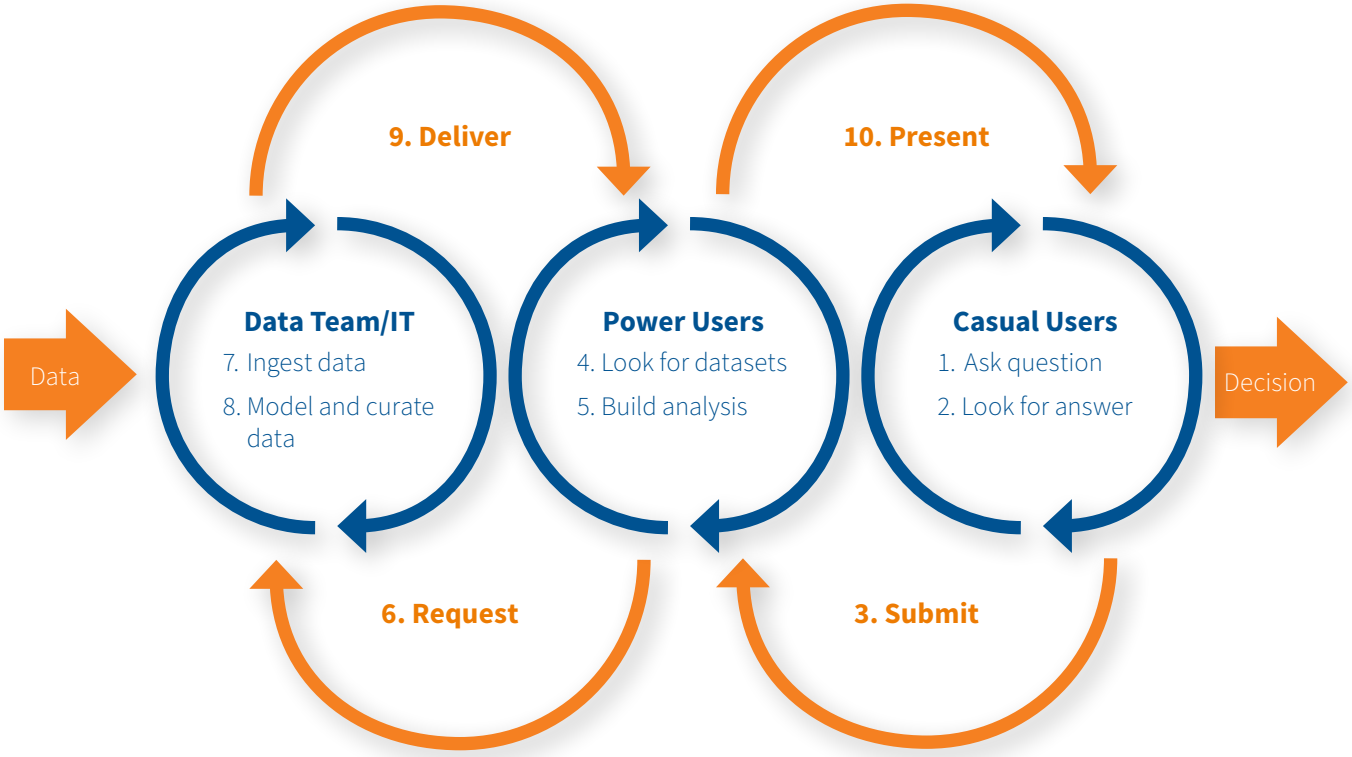
Data catalogs do not provide a complete solution for self-service analytics. They fall short because they only support part of the question-to-answer workflow—searching for data sets. They tell analysts the location of data, but generally don't connect them to it. At best, most data catalogs provide a preview or sample of data. If data analysts want to manipulate the data set, they typically need to access and query it separately. The analyst may need to obtain permission from the data owner, write SQL, or use a special tool to query the data. This prevents iterative exploration and analysis. Some catalogs now provide direct access to data sets, but this functionality is tacked on, not built in from the beginning.

Casual line-of-business users don't care about finding data sets. They care about finding answers to questions. It's not the job of an executive or manager to understand the cardinality of data or the naming schema of tables. The information provided by a data catalog simply isn't relevant to their approach. They need a tool that allows them to ask a question and get an answer.

Line-of-business users don't care about finding data sets.
They care about finding answers to questions.

Incomplete Solution. Answering a question involves a process like the one shown in figure 1.

Figure 1. The Question-to-Answer Meta-Workflow



The process of answering a question consists of multiple linked workflows. These make up the larger question-to-answer meta-workflow. In this graphic, each of the blue cycles symbolizes the workflow of a key participant in the process, while the orange represents the steps between them. The meta-workflow is both iterative and flexible. Not every question will require every step. If a given participant has what they need, they won't pass the question back to the next layer.

The rightmost workflow represents the casual line-of-business user. This individual conceives of a question and attempts to find the answer in the reports or dashboards they can access on their own. If they are unable to find what they need, they submit their question to a power user, such as an analyst or data scientist. This user builds novel analyses. They employ the data catalog as a search tool to find data sets, and either write SQL queries themselves or use a data virtualization tool to access data. If they can't find the data they need or don't have the requisite skills to access it, they submit a request to the enterprise data team to generate data on their behalf. The final workflow represents the steps the data team uses to ingest and govern data. This workflow also relies on a catalog, but here it is used to curate data and manage access.

The question-to-answer workflow requires numerous tools: the casual user needs a BI platform to access reports or dashboards, the power user needs a query engine and tools for data preparation and visualization in addition to the catalog, and the enterprise data team needs ETL and warehousing

tools to make the data available to the power user. Finally, everyone needs a shared platform on which to communicate.

The question-to-answer workflow requires numerous tools.

Each of these tools adds to the complexity of the overall solution. Teams must purchase, learn, and integrate every new tool in the workflow. This creates more work for IT and keeps them from other tasks. The lack of seamless integration also adds friction to the system, so each step takes longer. Together, these tools may facilitate self-service, but not efficiently.

Introduction to the Self-Service Analytics Workbench

Recently, vendors have started building platforms that combine the capabilities of the entire suite of self-service tools. Eckerson Group calls these platforms self-service analytics workbenches. Self-service workbenches place their emphasis on answering questions. They recognize how the question-to-answer process operates and support it end-to-end.

Self-service workbenches provide a place for the casual user to ask questions and view answers. For the power user, workbenches consolidate the functionality needed for the analyst's entire workflow into a single tool. They also put power users in direct contact with the data, reducing the need for an IT or data team to provide backend support. Workbenches further improve the efficiency of the system by connecting casual and power users via a communication layer.

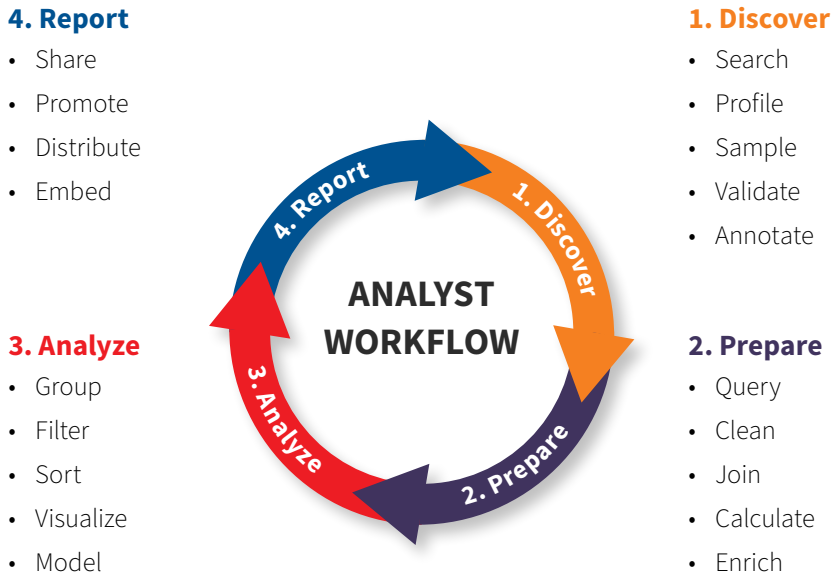
We can see the impact of a self-service workbench by focusing on the analyst or power user's workflow within the question-to-answer meta-workflow. The analyst workflow consists of four phases. (See figure 2.)

Previously, analysts used a different tool for every phase. A self-service workbench means they only need one. The workbench replaces all of their tools and also seamlessly flows from one phase to the next.

Previously, analysts used a different tool for every phase of their workflow.
A self-service workbench means they only need one.

Discovery. A self-service workbench provides the essential features of a data catalog. Users can search using natural language. They can tag, rate, and comment on assets. Analysts can view usage statistics

Figure 2. The Analyst Workflow



and preview data before querying it. If an organization has already invested in other tools, many workbenches will integrate with data catalogs and business glossaries to provide additional metadata.

Preparation. A self-service workbench will help the analyst get the data ready. Machine learning and automation features should assist in the process of cleaning and enriching the data. The workbench should identify relationships between data and suggest joins to the analyst. It should convert these suggestions into queries that retrieve the data and create tables. A self-service workbench should not require a high level of SQL knowledge. It may expose SQL syntax and allow advanced users to write or manipulate it, but it should be a low-code environment.

Some workbenches copy data into a new location where the analyst can query it. Others virtualize the data, leaving it where it lies. Virtualization has the advantage of reducing costs and time to insight because the data stays put. In either case, the workbench actually connects the analyst to the data rather than just revealing its location.

Analysis. A self-service workbench provides basic visualization capabilities. The goal of a workbench is not to create state-of-the-art, pixel-perfect, animated dashboards, but rather to give the analyst the building blocks to answer specific questions. A simple chart or graph often suffices. For more advanced needs, workbenches can forward queries to common, full-service BI tools.

Reporting/Publishing. Workbenches really shine in the final phase. Because a self-service workbench is organized around asking and answering questions, communication is one of its core functionalities.

Workbenches provide a single embedded channel of correspondence for casual and power users. They create conversations for each question, so users can return answers within a single thread. Increasingly, workbenches also integrate with tools like Slack and Microsoft Teams, allowing data-driven questions to be asked and answered directly in the main collaboration channels of the business.

Self-service analytics workbenches tend to evolve from one part of the analytics stack and expand into the others, so each has slightly different functionalities and areas of excellence. Regardless, the following are the key qualities to look for when evaluating a workbench:

> Built-in analytics workflow

> Question-based approach

> Natural language search

> Data discovery tool

> Collaborative tagging and search history

> Data virtualization engine

> Data preparation tool

> Auto-ML

> Visualization tool

> Communication platform

> Archive of old questions

Self-service analytics workbenches represent a new approach to asking and answering questions. They replace the old multi-tool method of supporting self-service with a single, unified platform. This reduces the pain of maintaining a self-service environment. Workbenches' emphasis on the question-to-answer process matches the natural workflow of business users, increasing their efficiency. As a result, organizations can answer more questions in less time.

Case Study: Before and After the Self-Service Analytics Workbench

In this section, we will examine how a company manages self-service before and after deploying a self-service analytics workbench. The organization in question is a mid-sized diagnostic instruments

company. It has a team of sales representatives dispersed across the country that travel regularly to meet with potential clients. The sales department uses a customer relationship management (CRM) tool to track trips and keep up with their sales. A dedicated team of three data analysts use this data to build dashboards and reports for management. The company also has a centralized IT department responsible for maintaining data infrastructure across the entire enterprise.

The vice president of sales, while examining travel budgets, wants to know if sales representatives who traveled farther afield from their headquarters generated enough additional sales to justify their travel expenses. In this case study, we will follow this question through the entire question-to-answer workflow before and after the company acquired a self-service analytics workbench.

Before

The VP begins by looking at their dashboard. The dashboard displays revenue by representative but doesn't include metrics about travel costs. The VP is not technically proficient enough to write their own SQL queries to gather additional data, so they send an email to their local data analyst asking for data. The two meet and converse to ensure the analyst fully understands the VP's requirements.

To answer the VP's question, the data analyst knows that they need to find a table containing travel records for the company's employees and a table containing sales by representative. The analyst searches the company's data catalog to find and evaluate the data sets needed, but discovers they don't have access to the right data. They email a request to IT to move the data into a warehouse where they can query it. This request sits in the queue for IT until they have time to deal with it. IT uses an ETL tool to move the data from the CRM tool to the warehouse and notifies the analyst. The analyst writes a SQL query to retrieve and join the data before importing it into a BI tool. Finally, the analyst builds a visualization to show the sales in dollars per mile traveled for each representative and presents the findings to the VP.

This workflow requires numerous tools and layers of correspondence even for a simple question. The data catalog saves some time, but it's only one step in the analyst's workflow. The process requires a BI tool, a SQL query engine, a data warehouse, and an ETL tool in addition to the catalog. It will take weeks to generate the answer because each step is disjointed and every time a participant hands off the question, it must sit in a queue. Moreover, when the VP sees the results, they may have follow-up questions that cause the whole process to repeat itself.

After

In order to accelerate time to insight, the company purchased a self-service analytics workbench and rolled it out to the sales department. Now, for the same question, the workflow looks as follows:

When the VP can't find an answer to their question in their dashboard, they turn to the workbench. They search for their question in the question archive using natural language. The tool browses its index to determine if the question was previously asked. If so, it returns the answer immediately. If not, the VP submits the question, which places it in a queue for a data analyst. The analyst then uses the same platform to generate an answer.

The data analyst uses the workbench's catalog-like search function to find relevant data sets. They retrieve and join the tables using data virtualization and preparation features, and then build a visualization to send to the VP, all within the platform and without recourse to IT. If the VP has additional questions, they send comments to the analyst via the workbench. The two repeat this process until the VP is satisfied with the result.

The sales team can now handle all of their questions and answers in a single platform in a highly efficient manner that supports both the natural meta-workflow between businessperson and data analyst and the power user workflow performed by the data analyst. The self-service workbench also reduces the number of tools needed to support self-service from five in the first scenario to one in the second, truly delivering on the mantra of faster, better, cheaper.

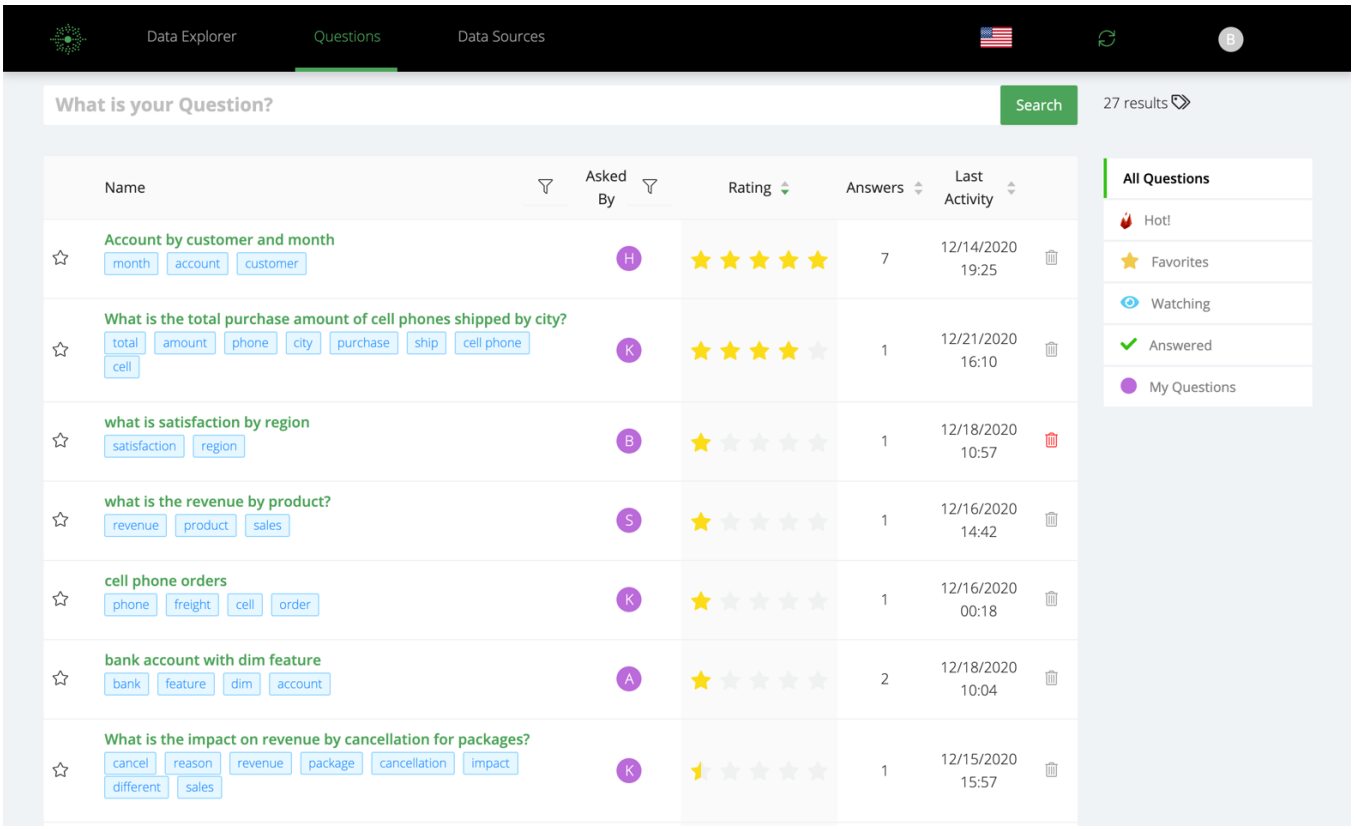
Promethium

One of the leading vendors in the emerging space of self-service workbenches is Promethium, which was founded in 2018 by CEO Kaycee Lai, formerly the president of the data catalog company Waterline Data. Lai conceived of Promethium after hearing customer complaints that traditional data catalogs could tell analysts where the data was, but not what to do with it. Promethium provides much of the functionality of a data catalog, including search and both automatic and manual tagging. It also integrates with third-party data catalogs. What Promethium does differently is to go beyond simply locating data to provide support within a single platform for the entire question-to-answer lifecycle.

Promethium goes beyond simply locating data to provide support within a single platform for the entire question-to-answer lifecycle.

The platform provides solutions to help two types of users: casual business users who need answers to questions and power users or engineers. For the casual user, the cycle begins when a business user types a question into a search bar using everyday language. Promethium relies on natural language processing (NLP) to interpret the question and returns a list of similar questions that others in the company have already asked. Users can click on those questions to see answers, and their data journey might end there. (See figure 3.)

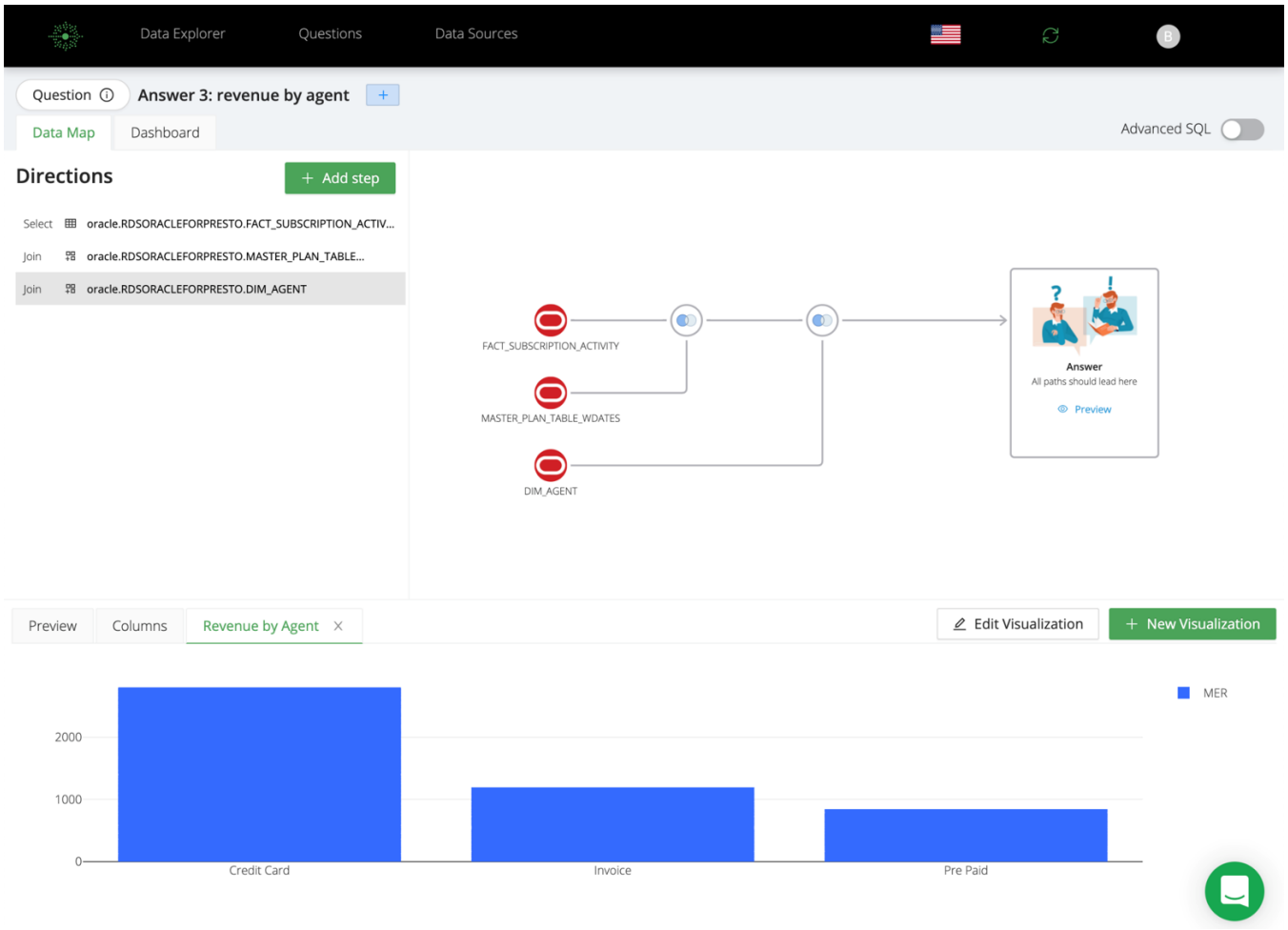
Figure 3. The Promethium Business User Interface for Finding Answers to Questions



If the business user doesn't see a relevant question or suitable answer, they can submit a question using natural language. Promethium then places the question in a queue for a power user or data engineer to answer. The data engineer can use a chat window to communicate with the business user to clarify the information they seek and perhaps what they are looking to accomplish. After the question and requirements have been made clear, the data engineer or analyst uses Promethium to conduct an analysis on behalf of the user. The engineer can search for data sets, examine metadata, scan data samples, and see what sets were used to answer similar questions.

Promethium returns search results quickly because it never actually touches or moves the data. Instead, it uses a data virtualization tool powered by Trino (formerly Presto). Promethium is hosted in the cloud by Amazon Web Services (AWS) and sifts through the metadata of sources via API connectors to find the appropriate tables and suggest relationships between them. Once a user has chosen the sets they want, Promethium recommends joins and auto-generates a SQL query. This query can also be edited directly, should the user want. Promethium saves the query as a "view," a visual model of the proposed combination of data, which can be extracted via Trino by any major BI tool like Tableau or Power BI. In addition to making it easy to export the view into third party tools, Promethium also provides charts and dashboards that allow the analyst to create simple visualizations quickly within the platform. (See figure 4.)

Figure 4. Answering a Question with Promethium



The requester can immediately see the answer, and both question and answer are cataloged for future searches. In addition, Promethium saves the queries, joins, and data model that went with the question, allowing analysts to reuse the work when answering similar questions.

Target customer. The ideal customer for Promethium is a company whose data is dispersed over multiple repositories. Promethium removes the need to manually consolidate and integrate data before any questions can be asked. This reduces the amount of time needed to gather data. The real value of Promethium, however, lies in its support of the question-to-answer lifecycle. As a self-service workbench, Promethium contains end-to-end functionality for the data analyst workflow coupled with a collaboration interface that bridges the gap between business users and data analysts.

Conclusion

Data catalogs help users locate data assets and provide valuable features for data governance. They are, however, just one step in the larger process of analyzing data—what we've been calling the question-to-answer meta-workflow. To fully support that workflow, companies should consider implementing self-service analytic workbenches.

These products not only integrate all the functionality required to support self-service analytics, increasing power user productivity, they also bridge the communication gap between business users and technical teams. Self-service analytics workbenches don't require casual users to do gymnastics to get answers—they can just type a question and look for existing answers. If an answer doesn't exist, users can activate the built-in workflow.

Self-service analytics workbenches don't require casual users to do gymnastics to get answers—they can just type a question.

Self-service analytics workbenches provide a refreshing change in approach by putting business questions front and center. They recognize the failure of previous self-service solutions to meet the needs of business users and, in response, locate data analytics within the end-to-end cycle of asking and answering questions.

About Eckerson Group

Wayne Eckerson, a globally-known author, speaker, and consultant, formed **Eckerson Group** to help organizations get more value from their data. His goal was to provide organizations with expert guidance during every stage of their data and analytics journey.

Today, Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps data analytics leaders abreast of the latest trends, techniques, and tools in the field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate business requirements into compelling strategies and solutions.
- **Our advisors** provide one-on-one coaching and mentoring to data leaders and help software vendors develop go-to-market strategies.

Eckerson Group is a global research and consulting firm that focuses on data and analytics. Our experts have substantial experience in the field and specialize in data governance, self-service analytics, data architecture, data science, data management, and business intelligence.

Our clients say we are hard-working, insightful, and humble. It all stems from our love of data and desire to help organizations harness the power of data. We are a family of continuous learners, interpreting the world of data and analytics for you.

Get more value from your data. Put an expert on your side. **Learn what Eckerson Group can do for you!**

Eckerson Group

RESEARCH CONSULTING ADVISORY

GET • MORE • VALUE • FROM • YOUR • DATA

About Promethium

Always be ready to answer tomorrow's questions today. Promethium is on a mission to help every business to be data driven by enabling every employee to make data driven decisions in real time without the technical complexity of data management. Promethium is the only SaaS-based data management solution that uses AI to answer questions for analytics. So you can spend your time taking action on insights, Promethium automates data discovery, preparation, query, and visualization without the need to move, migrate, or ETL data. Founded in 2018 and headquartered in Menlo Park, Calif. For more information visit **www.pm61data.com.**