MIT CSAIL ALLIANCES

Technology Use Case Study

Aurum

Aurum AllowsCompanies to Scan Multiple Data Sets for Answers

CSAIL PhD researcher Raul Castro Fernandez is helping companies including CSAIL Alliance members BASF, Dell EMC, BT, and Merck answer complex questions by accessing thousands of data spreads within data sources.

As companies are growing, collecting data becomes an essential task in day to day business operations. However, the challenge awaits when analysts are left with extracting tremendous amounts of data. This step is crucial to the survival of the business because it holds a value that can provide insights into the company's growth. Raul Castro Fernandez, PhD researcher at MIT CSAIL and founder of data discovery tool Aurum, explains how a problem needs to be addressed within companies about the hardship of locating, extracting and translating important data that provides a solution to the problem. He states, "One of the biggest problems that many companies face is that even though they now appreciate that data is valuable,they have a lot of internal questions that they could solve with data. But it will be hard to understand which data will support your question. Hence, the point of Aurum is that it will serve as a tool, so it makes it easy for you to detect which data might contain the answer." Aurum seeks to revolutionize the industry and heighten business performance by extracting data more efficiently and enabling employees to utilize their time on other tasks.

What is Aurum?

Aurum is a system to tackle data discoveryproblems. It introduces a new discovery algebra, called the Source Retrieval Query Language (SRQL), that lets users directly search for relevant data sources through a set of primitives that expose the relative insights of the underlying data. Aurum solves two problems in a business setting i) merging data that is stored across multiple storage systems, from databases to data lakes; ii) answeringcomplex questions by accessing thousands of data spreads within data sources.At the start of solving basic or complex questions, most of one's energy will be spent on understanding how to find the data based on specific needs of the question. Fernandez states, "By talking to the people that are using Aurum this is largely the case, they struggle to even find the data and this is why they are so interested because it's a way for them to shorten their time." Aurum's specific focus is to surface the datasets that are most relevant. However, the outcome may be that one will find the data that contains the answer to thequestion but it will come with some metadata that helps to clean the data or filter the data to solve the question. In other words, Aurum depicts the datasets that are the most relevant and there is some metadata that also helps with the tasks that are associated with this data preparation by an application. This will create a positive impact in the industry by reducing the time that an analyst spends on collecting, extracting, and analyzing data.

(continued)

One of Aurum's characteristics is its ability to merge data that is stored across multiple storage systems, from databases to data lakes. Aurum connects to various data sources (such as databases, spreadsheets,etc.) within an organization and represents a model that holds true of the data. For instance, a user on the system seeks to find data that is related to "X"; the user has the correct API's or functions needed to search for "X". Therefore, Aurum can build a set of APIs or a set of functions that the user can see. To implement this in a real business setting, Fernandez illustrates how this situation can be applied. He explains, "Let's say that you have data that could be more useful if you could join data with a different department. Would merging the data provide value or become fatal? Aurum will help you determine whether merging data is possible and what steps are required to execute the process." In other cases, researchers that are using data in a very ad hoc way: 1. copy data, 2. make some changes, 3. copy that data over again, 4. repeat. The result of this creates a number of different copies that are useful for the researcher, however, they reduce the quality of data over time. Hence, Aurum becomes a useful resource in efforts to determine which of those data sets are copies of each other so that one can remove them or reconcile if needed.

Another important characteristic of Aurum is its ability to answer complex questions by accessing thousands of data spreads within data sources. These days a company's biggest problem is finding data that is valuable but also answering internal questions. Most companies do not know how to troubleshoot therefore they juggle ideas and seek out different methods to conduct. The trouble awaits when sorting out which data answer specific questions. For instance, a researcher wants to investigate the gender gap distribution in each department at MIT. Although, this seems like a fairly easy task to solve this gets tricky when understanding which data will solve this question. An employee in the database could read as a full-time employee, part-time employee, or even a contractor. So, it is important to establish all of these rules into the databases in order to run an efficient scan. Aurum is the tool to make it easier for the user to detect which data will answer the question correctly.

The future of Aurum is bright and ready to revolutionize the industry one Aurum at a time. The benefits that industries and companies will receive from using Aurum will increase overall business performance. Since most companies are drawn to the traditional route of extracting data; Aurum will serve to summarize the different data sources and provide answers within a few hours depending on the complexity of the question. The amount of data and complexity will determine how long Aurum isable to conduct the scans. Therefore, this leaves companies to allocate their time, energy, and resources to other projects.

Putting Aurum to Work

In recent news, Aurum has partnered with industries like Koser Bank, BASF, Dell EMC, BT, and Merck to track key performances of the product. With the help of CSAIL Alliances, Fernandez was able to successfully connect with these companies to experiment trial runs of Aurum. He emphasizes how the process was smooth and stress-free because of the positive supportprovided by the Alliances team. Fernandez says, "With CSAIL Alliances, I was provided with many opportunities to showcase my work to Alliance members. I would deliver a presentation during an Alliances event and connect with interested company representatives. In speaking with the members after my presentation, we would determine if Aurum could be useful for their specific data challenge. If Aurum could help with their problem, I would send them a demo after the event. We specifically built this Aurum prototype that can be deployed so that even if I cannot access the data from the company, I can just send them the prototype and they can spend about 20 minutes setting it up and see if it brings value to the company... the prototype was useful for companies trying to determine if Aurum would be helpful in their organization."

(continued)

With the Alliance member connections, Fernandez was able to track the impact and value of Aurum while addressing any implications. In the case of Merck, the pharmaceutical company is responsibl for an immense amount of internal data that is produced on a daily basis. However, they require data from external sources which is in the public domain that is not accessible. Merck wants to understand the connection of external data in relation totheir internal data sources. Aurum helps to merge the external and internal sources and present meaningful data to the user. This is beneficial because it helps the user to see the connections and strategically help them narrow down research to the inquiries and research that need to be answered. Although there have been challenges along the way, the team was able to benefit from the research in efforts to improve Aurum. Once the team receives the feedback from the users, they can reevaluate which specificbeats need more fundamental alteration.

“It is a value proposition working with big companies. Aurum is able to help them determine the best way to find the answer they are looking for and I can learn the many approaches they use to try to discover the data. It is mutually beneficial”, says Fernandez. Aurum has a unique approach to users in comparison to its’ competition: value and versatility. For most companies, there is a barrier for employees to tap into a software and solve the problem. Aurum’s feature is that it can be run on API so it makes it easier to maneuver the system for those who are not familiar with the interface. Aurum provides versatility in the sense that it is easy for users to understand the functions ofAurum and solve the problem.

With that said, Aurum is taking data discovery to the next level by helping industries to locate, extract, and analyze tremendous amounts of data. Companies will benefit from this tool by solving questions effortlessly and allocating their time more efficiently. Although there will be some engineering challenges along the way; the research team aims to utilize the research angle to solve complex questions. Theresearch team is workingtowards an open source release. The vision is to build a community beyondthe companies that are already using it, so that current engineering issues can be resolved.