Data catalogues offer boost in the machine learning race

Why making corporate data more accessible and understandable to its data scientists, ultimately boosts the productivity of machine learning.

By Lindsay Clark

Originally published on www.idgconnect.com on June 06 2019

Spending on machine learning and artificial intelligence is growing by 37 percent annually and is set to reach $77.6 billion in 2022, according to IDC. But organisations keen to exploit this latest technological trend hit the same problem: data. Only 18 percent say their companies have a clear strategy in place for sourcing the data that enable AI work, McKinsey has found.

To crack the problem, a new class of enterprise data tools is emerging. Data catalogue systems are designed to help organisations manage a common weakness in the data science process, says Gartner research analyst Sanjeev Mohan.

"What happens is a data scientist finds out that their organisation has data lake. They get excited, jump in and think, 'Wow I have access to all the data I need, this is perfect'. But they get ready to train an algorithm; they don't like data lake: it's too slow. So, they make a copy of the data onto their laptop. They run until they are happy with the model. Then they go to IT and say, 'Here are the models, can you operationalise them?'"

Problems getting models into production

But typically, IT teams are not keen to operationalise models built in this way — with good reason, Mohan says. Firstly, the data has been outside the firewall. Then, the data scientists do not understand the provenance of the data: how it was collected, its quality, and what biases it may contain. As a result, there is no way to audit operational models once they are in production, he says.

"Data catalogues are trying to solve all these problems by offering curation of data from one centralised place," he says.

One data catalogue is built by Redwood, California start-up Alation. Founded six years ago, its products are used by 100 in enterprises around the world. Among them are eBay and LinkedIn. The group also includes the French bank Société Générale. Data manager, Julie Lerose, says the interest in a solution started two years ago with a discussion between marketing, finance and risk teams.

Giving users access to data at the right moment

"We talked about the problems in accessing all the data we have in Teradata, in SAS, Microsoft Excel and SharePoint and on big data platforms. We asked the finance and risk team how we can help our users access the data at the right moment. And how to improve our recommendations as it's difficult for all these people to find the right data," she says.

Discussion with Teradata, its enterprise data warehouse provider, led the bank to choose the Alation data catalogue. After an eight-week proof of concept with users in marketing, they rated the tool highly for recommendations and ease of search.

Alation uses machine learning to match the data with business language, helping users search for data using normal business terms, and recommending new data sources based on the data they have already used, like ecommerce recommendation engines. It also has a social media-style function to help users collaborate and share knowledge of data sets, as well as listing trusted sources who can verify data lineage.

Importance of defining governance and access

Lerose says: "It has saved users time accessing data because they can do more or less everything in the data catalogue. But it is important to understand that it is not just about deploying and using the tool. First, you need to define governance and rule of access; to do that we worked with the finance and risk teams."

Pascale Assémat, director of data marketing, retail, Société Générale, says the main business benefit was productivity. "We have many data instances: data warehouse, data lakes, the cloud, and so on. It is one of our biggest pain points. Someone who does not know the history of the data does not know how to start. This kind of tool gives them access to the data owners and helps them understand the data from a business point of view."

A report from 451 Research says data catalogue has come from nowhere in the past five years to become a technology allowing self-service analytics, self-service data preparation and multi-location data management. As well as the number of companies using the technology increasing, those that have already adopted data catalogues anticipate rapid expansion, both in terms of the number of users and the range of roles, the study found.

People and data are roadblocks

The most popular interpretation of the concept is the enterprise data catalogue, which offers access to any instance of data, and is used by 35 percent of adopters, the study says. Other catalogues address data lakes (16 percent), data warehouse or data mart (23 percent) and cloud data (26 percent).

Matthew Aslett, 451 Research vice president, says data catalogues are a means to make machine learning more productive. "In our research, access to and preparation of data is the number two challenge; skills are the number one. Combine the two, and it highlights that if data scientists are hard to hire, you don't want them spending time accessing and preparing data."

Providers in the market aside from Alation include Podium Data (acquired by Qlik), data lake company Zaloni, and Immuta. Enterprise data management companies such as Informatica also have products relevant to the problem, he says.

Becoming a strategic platform

Data catalogue adoption usually starts with data science and analytics teams pushing for deployments in one particular use case, Aslett says. But it requires someone such as a chief data officer to propel adoption into multiple use cases across the enterprise, he says.

"It does take a certain amount of will. That's why we initially see small, focused deployments. Driving out across the rest of organisation takes a bit more work. For a large enterprise, it can be a big-ticket purchase, but we have seen companies be successful quickly and it becomes a strategic platform," Aslett says.

Machine learning and AI is not magic. Only organisations with an understanding of their data, and how external sources augment it, can make these techniques work for them. Those looking to catalogues to boost the productivity of their hard-to-hire data scientists may be at an advantage.

Lindsay Clark

Lindsay Clark is a freelance journalist specialising in business IT, supply chain management, procurement and business transformation. He has worked as news editor at Computer Weekly and several other leading trade magazines. He has also written for The Guardian, The Financial Times and supplements to The Times.
