Ovum
TMT intelligence | informa

IBM Watson Studio and Watson Knowledge Catalog make data and AI more accessible, with governance

A single environment to build, train, and deploy machine-learning and deep learning models

Publication Date: 03 Apr 2018 | Product code: INT002-000086

Paige Bartley

Ovum view

Summary

At the IBM Think 2018 conference in March, the company announced the availability of IBM Watson Studio, a single, cloud-based environment aimed primarily at data scientists, data engineers, and developers, providing a suite of tools to leverage data and build, train, and deploy machine-learning and deep-learning models at scale. You would be right to assume that these capabilities, focused on operationalizing data science within the enterprise, sound familiar; the product is an evolution of IBM's former Data Science Experience (DSX) workbench product that adds in functionality around governance and data prep as well as increased capabilities around deep-learning-as-a-service, automated model testing, drag-and-drop neural network design using popular open frameworks, and embedded, pretrained, customizable Watson APIs. Most notable are Watson Studio's close integration with Watson Knowledge Catalog, the company's cloud-based catalog for data and artificial intelligence (AI) assets, and Data Refinery, the company's cloud-native self-service data prep capabilities, which are available with both Watson Studio and Watson Knowledge Catalog. Packaging all these capabilities together in a single environment speaks to IBM's platform strategy for AI and increases the utility and governance of the self-service ecosystem.

A collaborative ecosystem enabling self-service data science

When IBM originally built DSX, an integrated development environment (IDE) for data science, the goal was to provide the enterprise with an environment in which the process of building, training, and deploying machine-learning models could be scaled and operationalized within the enterprise. This meant not only providing a workbench of tools for conducting data science – think access to familiar open source notebooks and libraries – but also building a collaborative ecosystem where users could share work products and piggyback on existing efforts. It also meant providing intuitive, graphical functionality for the growing number of business power users beginning to dabble in data science. Watson Studio, in inheriting and absorbing DSX, has maintained these collaborative capabilities as a core element of its DNA. With expanded functionality and integration with the new IBM Cloud Private for Data, Watson Studio ties into an ecosystem of diverse data types and end users. Watson Studio is built to be an enabler of self-service, helping data scientists (as well as some subject matter experts) participate in the workflows required to bring AI, machine learning, and deep learning to life.

But as data science becomes scaled and operationalized within the enterprise, there is an increased need for governance. Data science is no longer a process confined to individual laptops and lone data scientists toiling away in locked rooms; it spans the enterprise and involves multiple end-user personas. Users need access to data while ensuring that data policies are enforced, such as the masking of personally identifiable information (PII), so that sensitive data is not inadvertently exposed. Users need to blend and prep data for input into analytics tools and machine-learning models while having underlying data governance policies consistently enforced. Watson Studio was built to accelerate the creation and application of AI within the enterprise, and for this to happen, users need to access and manipulate data freely, without worrying about compliance or the exposure of sensitive data. Therefore, it was imperative that the Watson Studio environment include governance capabilities, and IBM delivered it with close integration with Watson Knowledge Catalog. Embedded

capabilities for self-service data prep with an integrated data catalog ensure that users can readily navigate and prepare data for use in models while existing governance policies are respected and enforced.

Data Refinery enables ready access to quality, prepped data

Machine-learning models need data, and preferably lots of it. But the quality of the data is also of critical importance to the success and accuracy of the models. In an end-to-end environment built for bringing machine-learning and AI models to production, it makes sense to provide users with a way to refine and blend their data sets without ever leaving the platform. Data Refinery, IBM's self-service data prep functionality, is embedded in Watson Studio and Watson Knowledge Catalog. This approach ensures that data prep is integrated with Watson Studio's analytics and collaborative capabilities, and that users have a single platform from which to navigate, access, prep, and leverage their data.

Embedding self-service data prep in the Watson Studio ecosystem provides opportunities for better data governance. In particular, the tight integration of Watson Studio with Watson Knowledge Catalog (discussed below) ensures that data retains the policies that have been assigned. Pairing data catalogs with data prep environments is a current market trend, and for good reason. Catalogs help users find and navigate the data they need to prep and provide tools for governance. The Policy Activation Engine in Watson Knowledge Catalog pushes into Data Refinery, ensuring that policies that have been assigned in the catalog, such as the masking of sensitive data, are maintained as users blend and transform their data.

IBM, to be sure, is late to the self-service data prep market: the general availability of the Data Refinery feature in Watson Studio and Watson Knowledge catalog was announced at the IBM Think 2018 conference. It lacks some of the deep technical data prep functionality offered by more mature, purpose-built, and standalone data prep solutions on the market. But make no mistake: Data Refinery is a powerful tool. Its value comes from its embedded nature in the Watson Studio environment, which equals more than the sum of its parts. Its status as an embedded tool, alongside Watson Knowledge Catalog, allows for integration with governance capabilities and provides users with a single interface for finding and navigating the data they need, blending and prepping it, and then inputting it into models. IBM is also placing a lot of emphasis on the development of AI-based functionality for Data Refinery, to help guide users through the data prep process. IBM's approach reflects a market trend to make data prep a feature, rather than a standalone tool, within broader information management or analytics platforms.

What about visualization and analysis for data that has been prepped? Watson Studio has native visualization capabilities that allow users to do preliminary analysis on their prepped data directly within the platform. But if an enterprise wishes to use a dedicated business intelligence (BI) platform, that can also be accommodated. Data Refinery provides a direct connector to Watson Analytics, IBM's visualization and BI platform. Direct connectors to Tableau and other visualization tools are on the immediate product roadmap, meaning the enterprise will not be restricted to the IBM ecosystem of products.

Watson Knowledge Catalog enables key governance features

As Watson Studio is a cloud-native platform, it needed a cloud-native data catalog. Watson Knowledge Catalog is the cloud-based counterpart to IBM's on-premises catalog offering, the Information Governance Catalog. Embedded in Watson Studio, alongside Data Refinery, Watson Knowledge Catalog not only enables policy controls for data and AI assets, but also helps users find and navigate the information they need by providing a search engine-like experience, with machine-learning-enabled features such as automatic ratings for data sets (available in the Enterprise edition of Watson Knowledge Catalog). Having these two capabilities side by side in the Watson Studio ecosystem facilitates self-service, making sure users can find high-quality data quickly and reliably while ensuring that governance policies are consistently maintained across the data prep and data science lifecycle.

Watson Knowledge Catalog is a metadata catalog that points to different data sources in the enterprise, allowing data to remain in place in its native repository. It currently supports connectivity to over 30 data sources (both IBM and non-IBM), allowing users to access data wherever it may reside, such as in Hadoop clusters. The governance process begins as soon as a data asset is added to the catalog: all incoming data, both structured and unstructured, is profiled upon being added. The profiling process uses AI capabilities to automatically assess the data based on 180+ predefined classifiers, detecting common formats such as phone numbers, Social Security numbers (SSN), credit card numbers, and addresses. Uniquely, unstructured data is also profiled: the catalog is integrated with Watson Natural Language Understanding, which can analyze the content of documents and classify them based on business terms that have been defined in the catalog's business glossary. Sentiment and emotion can also be assessed in unstructured documents such as PDF reports.

Once data has been classified, granular rules can be set in the Policy Activation Engine to govern the data. A common use case in Watson Studio might be to mask all SSNs or other sensitive types of information so that the data sets they reside in may still be used for analysis and modeling. This centralized policy engine, which governs all the data being leveraged in Watson Studio, enables end users to ultimately access and leverage more data; as rules are applied consistently across the entire Watson Studio ecosystem, data does not have to be siloed simply because it contains sensitive fields. These policies, when crafted thoroughly by the enterprise, reduce much of the regulatory risk exposure that can come with building machine-learning and AI models.

The "Watson" moniker of Watson Knowledge Catalog does more than pay lip service to AI. The catalog is fundamentally AI-powered, leveraging Watson-based capabilities throughout to help classify data, rate data assets, and help users find data via natural language query. These capabilities are on par with or slightly ahead of the market trends that Ovum has identified in the data catalog space. Additionally, IBM's experience and engineering momentum in AI give it an edge in building out further functionality. Given modern volumes of data, AI-powered functionality in data catalogs will be critical in ensuring governance.

Appendix

Further reading

SWOT Assessment: IBM Analytics Suite, IT0014-003313 (July 2017)

"IBM's data science IDE – not your father's analytic tool," IT0014-003128 (June 2016)

"Hortonworks and IBM to OEM each other's big-data products," IT0014-003288 (June 2017)

Author

Paige Bartley, Senior Analyst, Data and Enterprise Intelligence

paige.bartley@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

Ovum
TMT intelligence | informa

CONTACT US

ovum.informa.com
askananalyst@ovum.com

INTERNATIONAL OFFICES

Beijing
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
San Francisco
Sao Paulo
Tokyo