Ovum
TMT intelligence | informa

# SWOT Assessment: IBM Data Refinery

## Analyzing the strengths, weaknesses, opportunities, and threats

Publication Date: 03 Apr 2018 | Product code: INT002-000085

Paige Bartley

# Summary

## Catalyst

The self-service movement is expanding beyond analytics and visualization. Multiple end-user personas in the enterprise need access to refined data, for different purposes. Data prep products that cater primarily to business analysts are missing the bigger picture: that data science initiatives and self-service analytics initiatives often rely on the same data, and that collaboration across roles helps strengthen a data-driven culture. IBM Watson Studio, in which IBM Data Refinery data prep capabilities are embedded, is a single environment and user interface (UI) for data cataloging, data science, and data prep, which caters to multiple end-user personas. This report provides a high-level overview of the key strengths and weaknesses of the IBM offering.

## Key messages

- Data Refinery is an embedded component of IBM Watson Studio and is also accessible through IBM Watson Knowledge Catalog, an intelligent cataloging service available independently or through seamless integration with Watson Studio.
- IBM's platform approach makes the same data prep capabilities equally accessible to multiple end-user roles, including data scientists, data engineers, developers, and knowledge workers, increasing collaboration and control for governance.
- Data Refinery's embedded nature in IBM Watson Studio and Watson Knowledge Catalog centralizes control for IT and enables many compliance-ready features out of the box.
- Having entered the self-service data prep market relatively recently, IBM has some catching up to do on granular functionality. However, its platform approach is its strength.

## Ovum view

Data Refinery is not a standalone offering, nor should it be: its strength is that it embeds data prep functionality in a broad data management platform, where its capabilities can be accessed equally by multiple end-user personas for multiple uses. The packaging of the product is as follows: IBM Watson Studio is the overarching platform and UI; IBM Watson Knowledge Catalog is available both within Watson Studio and as a standalone service; and IBM Data Refinery capabilities can be accessed equally from IBM Watson Knowledge Catalog and IBM Watson Studio. Because IBM Watson Studio offers a single environment for data cataloging, data exploration, data prep, and data science, the needs of a broad variety of self-service users are met by a single product, boosting collaboration potential and consolidating control for governance. The platform approach helps operationalize the analytics effort in the enterprise by allowing multiple end users – business analysts, data scientists, and data engineers – to work collaboratively in the same environment, sharing workflows and data sets.

Although IBM is a relative latecomer to self-service data prep functionality, its platform approach providing a common interface for data scientists and business users is ahead of the curve. Today's self-service data prep users are tomorrow's citizen data scientists, and providing a single platform with data prep and data science tools together encourages a cross-functional, collaborative approach. Data Refinery certainly has some catching up to do with regard to granular data prep functionality, but

 

the challenge is not insurmountable. Its value as part of the holistic Watson Studio will gain appeal as organizations increasingly set their sights on operationalizing and unifying self-service and data science processes within the organization.

# Recommendations for enterprises

## Why consider IBM Data Refinery?

The platform approach offered by IBM Watson Studio combines data cataloging, data prep, and data science into a single UI, making it a good option for organizations that want to consolidate self-service users under a single, governable platform. Tools for business analysts, data scientists, and data engineers all exist within the same product, helping "unsilo" these roles and increase collaboration around data. The product is a good option for organizations looking for embedded data prep capabilities that can help refine data for multiple uses, not just self-service analytics.

IBM Watson Studio and Watson Knowledge Catalog, with their embedded Data Refinery capabilities, are ideal for any organization that is looking to scale up and operationalize its data science initiative. Because Data Refinery is accessible directly within the IBM Watson Studio workbench, users can blend and refine data for input into data science models without ever leaving the product or interface. Giving data scientists the data prep tools they need right on the platform that they already use to access notebooks, manage data science projects, schedule analytic compute runs, and manage access and track lineage to different sources of data means the utility of the product is increased and governance of the data science process becomes more centralized.

# SWOT analysis

## Strengths

### The platform approach unifies data prep, data catalog, and data science capabilities

Data Refinery was originally developed as a standalone offering. However, IBM quickly realized that its value would be greater as an embedded component in existing data management offerings. Its current packaging in IBM Watson Studio and IBM Watson Knowledge Catalog ensure that users of either environment have the tools they need to access and prep data without toggling between products. The UI of IBM Watson Studio and IBM Watson Knowledge Catalog is unified in a single environment, lending a single, seamless user experience and frictionless access to all data manipulation capabilities, including data cataloging and exploration, data prep, data science, and even preliminary visualization and analysis. Because data consumers in the enterprise are united under a single platform, with access to all the tools they need, governance is increased and the ability to operationalize data exploration and the data prep process is bolstered. IT has direct visibility of the platform's users and activities, offering a single point of control for important functions such as role-based access controls and data policies.

Unifying data catalog, data prep, and data science workbench capabilities in a single environment meets an important need: serving the widest possible array of self-service users in the same cohesive

 

ecosystem, where they can be managed and monitored together by IT. Instead of primarily targeting business analysts, as many data prep tools do, Data Refinery is designed to meet the needs of multiple personas: business analysts, data scientists, data engineers, and developers. Providing a single product and single UI, with which they can all work side by side, means data prep can become more collaborative and more than just a vehicle for self-service data prep. Today's business analysts are tomorrow's citizen data scientists, and IBM Watson Studio allows users to move fluidly between roles and functionality, if they wish, allowing enterprises to realize greater potential in their use of data.

### Governance capabilities bolster compliance and help meet regulatory needs

The data prep process influences the outcome of data analysis and, subsequently, business decisions. The business therefore needs full visibility of the actions taken on data during the data prep process. IBM's Data Refinery capabilities, as packaged within the IBM Watson Studio and IBM Watson Knowledge Catalog, are compliance-ready out of the box. As well as providing a complete audit trail, the product provides a number of governance and security capabilities that bolster compliance with various data protection regulations, such as the EU's General Data Protection Regulation (GDPR). Granular role-based access controls, the inheritance of role-based access controls from data source repositories, and row-, column-, and entity-level security all ensure that data is only viewed and accessed by authorized parties. When users search for data in the product interface, only the results that they are authorized to view are returned. Additionally, having a single platform for data cataloging, data science, and data prep inherently consolidates governance control and IT visibility: one platform (as opposed to disparate tools) means more streamlined monitoring, policy enforcement, and auditability.

Perhaps the most powerful tool for compliance in Data Refinery's arsenal is its machine-learning-powered ability to automatically detect and mask data that is potentially sensitive, such as social security numbers and phone numbers. Few data prep products offer functionality to automatically detect sensitive data, and Data Refinery's machine-learning-driven approach allows it to scale to enterprise volumes of data. Because potentially sensitive data can be automatically detected, and subsequently masked on the column level, the business can unlock the value of data that might have previously been siloed due to security concerns. The automatic detection and masking of sensitive data increases the enterprise's ability to protect data "by design and by default," and, perhaps counterintuitively, makes more data available for analysis.

## Weaknesses

### Direct integration with self-service BI and visualization tools is underdeveloped

In all, the capabilities of the IBM Watson Studio are designed for governance, data science, and data prep, all from a single UI and platform. It would follow, then, that self-service analytics and visualization would be the logical final step of these platform capabilities, with users completing data prep functions then being able to transition fluidly to full-powered visualization and analysis with functionality for direct integration in the native IBM analytics environment (IBM Watson Analytics) or preferred third-party business intelligence (BI) tools.

However, this is still not the case. Once data prep has been completed in the IBM Watson Data Platform environment, the process for getting data into self-service BI and analytics tools is still

 

relatively manual. Once users have discovered data with IBM Watson Knowledge Catalog, they can use the integrated Data Refinery capabilities to get the data and move it into a target of their choice, whether that be a cloud database, an on-premises database, or a CSV extract. Users then leverage their third-party or IBM tools against those targets to use with self-service BI and analytics tools. Direct connectors for popular visualization tools, such as Tableau and Qlik, have yet to be developed in the IBM Watson Data Platform environment (although a direct connector to Tableau is planned for early 2Q 2018), and native support for partner file formats (such as .tde, .qvd, and .pbi) is not yet offered. The result is a platform where users are given tools to prep data, but offered relatively little support in transitioning to self-service visualization and analytics.

There is a silver lining, though: some self-service visualization capabilities are available today in the IBM Watson Studio, with the recent addition of Dynamic Dashboard Embedded Service. After users finish their data prep task, they can use the integrated visualization capabilities to quickly find insights and create visualizations and dashboards. Additionally, direct connectors for BI tools are on the 2018 roadmap, and the recent direct connector to IBM Watson Analytics holds special potential for establishing a single, seamless user experience for governance, prep, and analysis.

## Features for the automation and guidance of the user journey are currently lacking

Self-service data prep is gradually becoming democratized, moving beyond technical users to reach a larger audience of business users. The spread of data prep functionality to an increasingly nontechnical audience parallels the development of BI and visualization tools, to which more automation and "smart" functionality were progressively added to guide nontechnical end users in the analysis process. Data prep tools are today seeing the same, with automation added in to guide users in the data prep journey and the use of machine learning and other functionality to predict next best actions and provide suggestions for actions such as transformations.

This is an area in which Data Refinery does not perform as strongly as some of the other data prep offerings, perhaps due to its relatively recent entry in the market. In general, users need to know what they want to do with the data; predictive transformations are not offered, and automated suggestions and help for actions are not provided. Other automated and single-click functionality is lacking: automated visual flagging of outliers, anomalies, and missing or mismatched content is not conducted, data is not automatically de-duplicated, there are no automated joins, and there is no single-click function to append or join multiple data sources. The product does not offer automated recommendations for data relationships and keys for combining data across multiple data sets and sources. Although data can be enriched with third-party sources, sources to enrich data are not automatically recommended; users need to seek them out.

In the current Data Refinery offering, some of the most conspicuous absences are related to machine learning. Although machine learning is used in the product to detect sensitive data and outliers, anomalies, and missing or mismatched content, it is not yet leveraged extensively in guiding the user journey. More guided and machine-learning-powered functionality is on the product roadmap, but for now, less technical users may struggle to make sense of the next best steps.

 

# Opportunities

## Data prep capabilities are well positioned to accelerate data science initiatives

By embedding Data Refinery directly in IBM Watson Studio, IBM has positioned its data prep capabilities not only to help accelerate traditional analysis and visualization efforts by business analysts, but also to help data scientists refine the data that they need to build models and get them to production. The decision to offer data prep capabilities directly in the Watson Studio environment reflects IBM's holistic approach to operationalizing data science within the enterprise. By giving data scientists a centralized workbench of tools where they can access notebooks, manage data science projects, schedule analytic compute runs, and manage access and track lineage to different sources of data, IBM provides a platform that "unsilos" the data science process from individual laptops and machines. With Data Refinery embedded in Watson Studio, data scientists benefit directly by being able to quickly and easily blend and refine the data that they need for input into models, without ever leaving the product interface. Data prep capabilities make Watson Studio more comprehensive, further facilitating adoption and making the product "stickier" for users.

The data scientist audience will only continue to grow in importance and influence in the enterprise, and platforms such as IBM Watson Studio are accelerating the changes necessary to operationalize the data science process. By pairing data prep capabilities in a full-featured data science workbench, IBM is reaching a critical audience of users that need ready access to prepped data sets for their models. As Data Refinery is equally accessible to data scientists and business analysts from its respective locations in IBM Watson Studio and IBM Watson Knowledge Catalog, it helps unify analytics efforts across the enterprise, allowing users with different skill sets and objectives to prep and leverage the same data for different uses.

## The single-platform ecosystem can potentially operationalize the analytics process

What Data Refinery may currently lack in automation and guidance of the end-user journey, it makes up for in its single-platform approach. IBM Watson Studio, under which IBM Watson Knowledge Catalog and IBM Data Refinery exist, encompasses a broad range of data-handling functions. IBM Watson Studio is meant to be a one-stop shop to leverage data, with functionality built for data scientists, business analysts, and data engineers. Providing a single platform under which all these personas can work side by side, with access to the tools they need, helps operationalize the analytics process by eliminating organizational and product silos and centralizing data governance functions. One platform, with one UI, simplifies use and increases adoption.

The missing piece here, which IBM is actively working on, is direct connectivity to BI and analytics tools (discussed above, under "weaknesses"). Once IBM Watson Studio can connect seamlessly to tools such as Tableau, users will have an end-to-end experience that allows them to go from exploring and searching data and prepping data to visualizing and analyzing data, with minimal toggling or switching between products and interfaces. This will be especially true if the enterprise chooses to leverage IBM Watson Analytics, in conjunction with IBM Watson Studio, for its visualization needs – direct connection to this environment was recently made available.

As self-service expands beyond analytics and moves toward data prep and data science, it will be an asset to have a platform that provides centralized tools for all these functions. IBM will likely see increased interest in the IBM Watson Studio as the enterprise gradually looks to unify its approach toward data governance, data prep, and data science. Seamless connectivity to IBM Watson

 

Analytics, which was recently completed, will prove to be the ultimate opportunity to create a single end-to-end environment in which analysis can be conducted immediately after data exploration and prep.

# Threats

## The late entry in the self-service data prep market means there is catching up to do

The Data Refinery module in IBM Watson Studio and Watson Knowledge Catalog is a relatively recent addition to the product ecosystem, with private beta initially made available in October 2017. The capabilities being embedded rather than provided standalone gave the company a head start in developing the necessary connectors (particularly to IBM data sources) and tapping into the existing collaborative structure of Watson Studio to build out tools for the publishing and sharing of workflows, pipelines, and other models. IBM was not building a self-service data prep ecosystem from scratch, but instead adding self-service data prep tools to an existing data management and collaborative platform.

However, its relatively late start in adding self-service data prep functionality to the environment means that it faces an uphill battle in catching up to other data prep providers, particularly those that offer standalone functionality. Standalone data prep vendors depend on having best-of-breed capabilities combined with a robust ecosystem of direct connectors, and many have products that have been generally available for several years. Data Refinery's strength is its embedded nature in the holistic IBM Watson Studio, rather than its current depth of data prep functionality. This means that IBM needs not only to play catch-up with data prep features such as automated functionality, but also to focus disproportionately on pitching the value of the entire Watson Studio.

IBM is addressing these concerns with an aggressive development roadmap that is focused on expanding functionality, with particular focus on automation, machine learning, and connectors to various repositories and analytics tools. But however aggressive the roadmap, it will be a struggle to make up for lost time. Data Refinery's capabilities will be unlikely to replace established enterprise deployments of standalone data prep tools, so the company needs to focus on selling the holistic value of the Watson Studio as a unified data management and collaboration ecosystem.

## The ecosystem of data connectors needs to expand to keep pace with market

Users of IBM's Data Refinery capabilities currently have access to 30+ connectors to IBM, non-IBM, and third-party data sources. These connectors should take care of the bulk of the data access needs of most organizations, whether data exists on-premises, in the cloud, or on the desktop. Connectivity to Hadoop (HDFS), Amazon S3, traditional databases (relational database management systems [RDBMS]), and NoSQL sources cover the wide range of data – particularly big data – that the typical enterprise may have in various locations.

However, with the number of repositories and data sources in the enterprise only becoming more diverse, IBM will need to quickly scale up the number of connectors that it offers to keep pace with the market; some vendors of data prep functionality support connectivity to 80 or more unstructured and structured data sources. Part of the data prep process is blending and joining data from disparate sources, and if self-service data prep users cannot access or connect to data in commonly used repositories or applications, then that data cannot be effectively leveraged, and downstream analytics efforts will suffer. Organizations that are looking for a way to operationalize data prep within the

 

enterprise are also looking for solutions with the widest array of connectivity, so that no data is left locked out of the data prep process. The enterprise, then, with its diverse IT ecosystem and large number of software-as-a-service (SaaS) applications, will tend to favor data prep tools that offer broad compatibility.

SaaS compatibility, in particular, is an area in which IBM faces being left behind in the market if it does not develop additional connectors. Although Salesforce compatibility is offered, Data Refinery does not connect to other SaaS applications out of the box (APIs and/or SDKs are available to enable custom integration). Connectors for additional SaaS applications are currently being developed, but IBM has a lot of catching up to do in terms of compatibility. As the enterprise moves steadily to the cloud and seeks out maximum connectivity, this could prove a threat to IBM.

# Data sheet

## Key facts about the solution

**Table 1: Data sheet: IBM**

| | | | |
|---|---|---|---|
| **Product name** | IBM Data Refinery | **Product classification** | Data preparation |
| **Version number** | Continuous delivery | **Release date** | March 2018 |
| **Industries covered** | All | **Geographies covered** | All |
| **Relevant company sizes** | Midsize and large enterprises | **Platforms supported** | Managed on IBM Cloud, accessible via common web browsers from all platforms |
| **Languages supported** | English, with more coming in 2018 | **Licensing options** | Subscription, pay as you go |
| **Deployment options** | Cloud | **Routes to market** | Direct, channel |
| **URL** | www.ibm.com | **Company headquarters** | Armonk, NY, US |
| **European headquarters** | Portsmouth, UK | **North America headquarters** | Armonk, NY, US |
| **Asia-Pacific headquarters** | Singapore | | |

Source: Ovum

 

# Appendix

## Methodology

Ovum SWOT Assessments are independent reviews carried out using Ovum's evaluation model for the relevant technology area, supported by conversations with vendors, users, and service providers of the solution concerned, and in-depth secondary research.

## Further reading

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

*Beyond Self-Serve: Expanding the End-User Audience of Data Prep,* IT0014-003213 (January 2017)

"IBM's return to revenue growth suggests strategic imperatives are paying off," INT003-000049 (February 2018)

## 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