NUCLEUS RESEARCH

IBM COURTS THE ELUSIVE DATA SCIENTIST

ANALYST
Anne Moxie

THE BOTTOM LINE

While the analytics market has been a focus for heavy investment, the needs of the data scientist have largely been overlooked. Nucleus expects that IBM's new Data Science Experience platform will help data scientists to experience at least 20 percent more productivity through the availability of collaboration tools and curated datasets. Over time, the open-source nature of the platform will also help to encourage greater participation and talent growth in the field of data science.

•••

THE ANNOUNCEMENT

On June 6, IBM announced not only a much needed new platform built specifically for the data scientist, but also a massive shift in the structure of how it attracts and retains customers. This platform, unlike any of IBM's other offerings, is open-source, which will allow for more users to leverage its functionality and participate in a community style environment.

The Data Science Experience platform offers a number of features including:

- Curated datasets. Through the IBM Cloud Bluemix platform, data scientists will have access to 250 curated datasets for analysis.
- Open source resources. Users will also have access to RSudio, Jupyter Notebooks, and H2O for open source options.

- Collaboration. With access to collaborative capabilities, data scientists will be able to share code and provide feedback on analyses.

All of these features are available via Apache Spark, which is an engine for data processing that has built-in functionality for streaming, SQL, graph processing, and machine learning.

WHY IT MATTERS

With this announcement, IBM is addressing some of the most pressing issues for data scientists. In particular, the Data Science Experience platform allows for team participation on projects through collaboration, accelerates the access to data by providing clean datasets, and incorporates data visualization, which allows data scientists to more effectively share their insights.

A COLLABORATIVE ENVIRONMENT

Collaboration on data science projects is critical because they often involve a team of users to develop a complete analysis. This team can consist of data scientists, application developers, data engineers, and business people.

The reason for the reliance on teams is largely because data science is a relatively new area of expertise and there are few people trained specifically to be data scientists. As a result, they primarily come from two main backgrounds; computer science or statistics. A data scientist is naturally better at one of these topics and may need some support on the other. For support on data management aspects, data engineers are often pulled into a project as well. In addition, projects almost always require input from business people to facilitate getting to the actual business need of the problem.

With this in mind, IBM has created a platform that supports ongoing communication and collaboration. This will help data scientists to function and operate as more of a collective team, which will ultimately allow users to perform more advanced and accurate analyses by combining their areas of expertise.

ACCELERATED DATA ACCESSIBILITY

Data scientists report that they often spend 70-80 percent of their time on data preparation and, being a high-priced employee, this is clearly not an effective use of their time. Some organizations tackle this by deciding to hire a data engineer to manage the data preparation aspect. However, not every organization can afford an additional hire and data scientists often have to wear the data engineer hat.

The new Data Experience platform helps to facilitate faster access to clean data by providing curated datasets. IBM has been investing in data resources, such as the Weather Company, that can provide data scientists with external information that will augment their analyses. As a result, data scientists will have faster access to more complete information. Combining the benefits of speedier data preparation with easier collaboration, Nucleus expects the Data Science Experience platform will increase a data scientist's productivity by at least 20 percent. As IBM continues to invest in it data offerings, the productivity improvements will continue to grow over time.

Nucleus expects that easier collaboration and reduced data preparation time will increase overall data scientist productivity by at least 20 percent.

Data is a moving object because at every point in time there is new data being created and collected. In order to facilitate what is a constantly shifting information-based climate, IBM uses Spark for data streaming. Spark is particularly valuable in this case because its functionality is well suited to distributed processing and faster ongoing analysis. As IoT applications are starting to be used more often, the distributed computing capabilities will be a valuable tool for users that are struggling to manage incoming data that is constantly being collected from many different sensors.

VISUALIZATION

With the inclusion of Shiny, which is a Web application framework for R, the Data Science Experience platform allows data scientists to build interactive graphics and visualizations that can be distributed to a wider audience. This is critical to driving a higher ROI because data science is complex and often requires an understanding of statistics to interpret and utilize results of analyses. Therefore, to communicate results effectively, data scientists need a tool such as Shiny to best display insights and promote a greater understanding and use of analyses for every day decision making.

LOOKING AHEAD

The majority of companies in the analytics space have been laser focused on compensating for the lack of available data scientists. As a result, the cloud analytics market is full of eager new vendors that hope to put self-service statistical capabilities in the hands of business users who want to access interactive graphics for improved daily decision making. While this has resulted in a few highly successful

applications in an overall noisy space, the data democratization movement has distracted the market from developing tools for the data scientists that do exist.

There are only a few data scientists available because data science is hard and requires extensive training to leverage properly. Vendors, such as IBM with its new Data Science Experience platform, are now working toward making their lives a bit easier by building solutions to address their unique pain points. As data science continues to be a highly sought after skill, IBM's audience for its new platform will continue to grow and will simultaneously encourage a community, cultivating talent in the field.