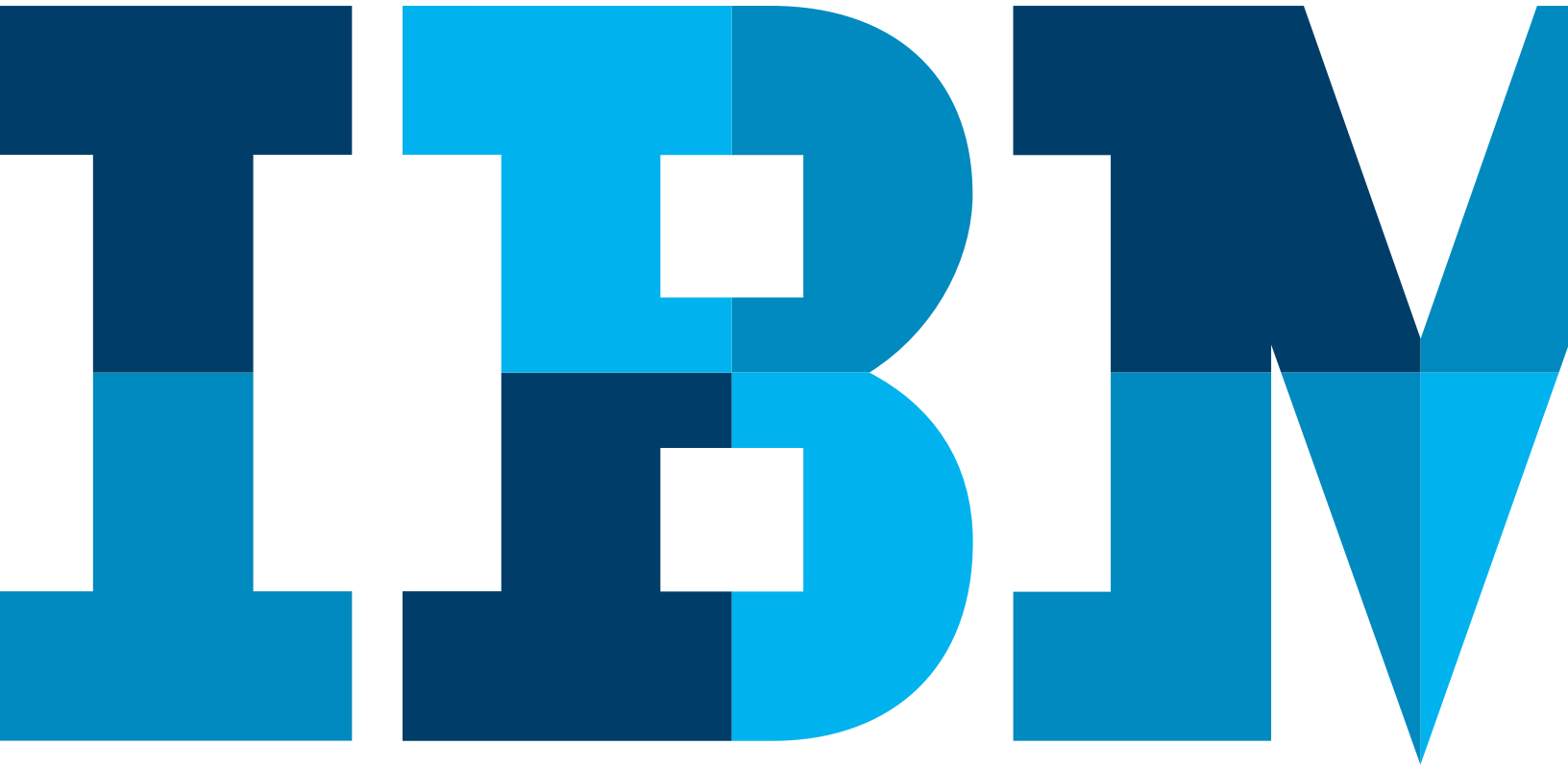
IBM Data Science Experience
White paper

IBM Data Science Experience and IBM Bluemix Data Connect

A combination with the power to reduce time-to-insight



IBM

Executive brief

To capitalize on the business value inherent in their data, organizations of all kinds are seeking faster and deeper insight as a key source of competitive advantage. This means accessing and analyzing data sets of all sizes from a growing number of increasingly diverse sources.

Technologies for analyzing data are constantly improving, enabling organizations to gain rapid insight from even the largest volumes of business data. Equally, innovative cloud-based solutions aimed at the business community have removed much of the complexity from setting up and managing an analytics platform. However, two key barriers to insight remain: the time and effort required to source the right data, and the time and effort required to prepare that data for analysis. The problem is exacerbated by the fact that much of the work must be repeated each time the data is refreshed or the business wants to ask new questions about the data.

The combination of IBM Bluemix Data Connect and IBM Data Science Experience solves these challenges, providing an all-in-one web-based platform for selecting, exploring, preparing, analyzing, managing and governing business data. With the ability to connect to both on-premises and cloud data sources, and to provide a single point of access and control across multiple open-source data-transformation and analytics tools, these IBM solutions can work together to significantly reduce time-to-insight.

The paper will explore:

- The challenges facing data scientists and business analysts as they seek to gain rapid, high-quality insight from growing volumes of data
- The shortcomings of traditional approaches to data preparation and integration
- The proposed solution: a common foundation for data access, preparation, integration, analysis, collaboration and governance.

The analytics bottleneck

In leading organizations across all industries, data is rightly seen as a key source of competitive advantage. Whether an organization is looking to optimize manufacturing and logistics processes, to improve service levels through better understanding of customer preferences, or to identify the most profitable new markets—the answers are waiting to be unearthed from vast stores of customer, sales, financial and operational data. The right insights at the right moment will drive smarter business decisions and help organizations to out-think their rivals.

Even as the number of systems in the typical organization increases, and the variety and volume of data grow, new analytics technologies are emerging that can crunch the data and produce insight at ever higher speeds. However, the best analytics tools in the world are worthless without the right data. The challenge for data scientists and business analysts is to source and prepare the relevant data for analysis. What's more, they are often under increasing time pressure, and they serve a business intelligence community that expects flexibility and the ability to ask new questions using continually updated data. In practical terms, this means that data scientists and business analysts need to source and prepare data quickly, and in a consistent, repeatable way.

In many organizations today, reserves of data sit largely untapped because of the prohibitive time and effort required to extract them from source systems and prepare them for analysis. As a result, opportunities to seize competitive advantage are going to waste, and the value locked away in enterprise data remains dormant.

Anecdotally, data professionals spend as much as 80 percent of their time just finding and refining data, leaving little time to actually perform useful analysis. A key challenge here is that much organizational data is subject to heavy governance and security controls. Since data is perhaps the most valuable of all assets in the information age, it is not surprising that IT organizations apply very tight controls over it.

Seeking efficiency and control

In practice, the need to safeguard data places barriers in the path of data scientists and business analysts. It also creates an unhelpful tension between IT staff and the community of data professionals and analysts. To get their hands on the data they need, data professionals must submit requests to IT, wait for the response, check the data to see if it is what they are expecting, and then often resubmit a refined request. Naturally, this is frustrating for the data professionals, and it's also just as frustrating for the IT staff, for whom these ad hoc requests are an unhelpful interruption to more interesting and important work.

The cycle of requesting and waiting for data is inefficient and increases administrative costs. It also increases the time-to-insight, making it harder for organizations to identify and act on competitive opportunities. Where external sources of data are relevant to analysis—for example, weather records in an analysis of summer clothing or ice cream sales across a region—this adds to the complexity.

Even when the right data has been found, it will usually need to be transformed, cleaned and joined before it can be used in analytical tools. For most data professionals today, this means importing various data sets into a spreadsheet and joining them, then matching records, eliminating duplicates, classifying data into different types, identifying and resolving errors, and transforming fields to ensure that different data sets are comparable.

Finally, once the data professional has prepared their data and run the initial analysis, they will often find that the results suggest the need to modify the source data or the preparation methods. And each time the business needs to ask the same questions on updated data—for example, for a daily, weekly or monthly report—it will usually be necessary to re-run much of the sourcing and preparation process all over again. This lack of repeatability is a major drain on productivity, and further increases the time-to-insight.

A further challenge for many organizations is that the audience for analytics is broadening. Alongside traditional power users, line-of-business users are being empowered by new self-service tools that enable analytics to be incorporated in day-to-day work. So, in addition to centralized, IT-owned business intelligence there is now also a decentralized approach owned by the lines of business, which implies an added need for governance. Policies governing access to data must be carefully managed, and there must be a deeper focus on data stewardship and curation, so that users are not working with poor-quality data or having to start from scratch each time. In many organizations, there is now a chief data officer who is responsible for ensuring that data serves the needs of all users and provides a single version of the truth. However, the split between bottom-up and top-down approaches complicates the picture, and the use of different technologies in these two areas can make it difficult and costly to govern data effectively.

Unified environment for analytics

To help organizations tackle the key bottleneck in analytics today—the time and effort involved in sourcing and preparing data for analysis—and to ensure effective governance over corporate data, IBM proposes a unified approach to data access, preparation and integration. With IBM Bluemix Data Connect acting as a single, common foundation for these functions, and integrating with IBM Data Science Experience for repeatable analytics and predictive modeling, organizations can build collaborative environments for using data to drive competitive advantage.

Bluemix Data Connect is a cloud-first self-service data preparation and integration solution that empowers data professionals to source data from multiple sources, load it, clean it, transform it and deliver it to multiple targets. The Bluemix Data Connect APIs help developers to connect easily to multiple source databases and load data into multiple targets, while IBM Bluemix Data Connect is primarily aimed at knowledge workers and helps them to select data, visualize it, enrich it and prepare it for analysis. By automating the repetitive and time-consuming tasks of cleaning, shaping, formatting and profiling data, Bluemix Data Connect accelerates the data preparation stages and makes them repeatable. Bluemix Data Connect features a growing set of connectors to other data sources, including Amazon Redshift, Apache Hive, Apache Impala, Microsoft Azure and SQL Server, MySQL, Oracle, PostgreSQL, Salesforce.com, Sybase and many more.

IBM Data Science Experience provides an interactive, collaborative, cloud-based environment in which data scientists can use multiple IBM and open source tools to gain rapid insight into data. Based on Apache Spark in-memory technology, Data Science Experience accelerates the development of iterative and machine-learning applications. By embedding Bluemix Data Connect in Data Science Experience, IBM has made it significantly faster and easier for data professionals to securely access, enhance, prepare, and visualize data from multiple sources for analysis. Using Bluemix Data Connect, data professionals can skip the process of importing data extracts in the form of spreadsheets or CSV files, and instead securely connect directly to any data source. The solution offers a broad and growing range of connectors to some of the most widely used cloud data systems, as well as secure gateway technology for accessing on-premises systems.

Iterating for speed

The tight integration between Bluemix Data Connect and Data Science Experience simplifies the creation of "pipelines"—sequences of operations that are used to build complex, repeatable data flows to support potentially long-running analyses. These pipelines, in combination with the "canvas"—a flexible, collaborative graphical interface for data modeling, enable the operationalization of analytics, allowing data scientists to work iteratively through business problems to build, test and adjust their models. Even non-technical users can create and iterate pipelines, see interim results using the power of Apache Spark, and feed them into a target environment for further analysis. The key differentiator versus traditional data science pipelines is this ability to use Spark to preview results directly in the data shaper—so that users can very rapidly check how their model works on the real sample of data, putting an end to lengthy cycles of switching back and forth between different tools for debugging.

Organizations can use IBM Bluemix Data Connect and Data Science Experience to create sets of pre-built components that non-technical users can assemble to create the pipelines they need—eliminating all of the lengthy stages of hunting for data and preparing it for use. For more advanced users, the IBM approach is highly extensible, so that data scientists can start from the same pre-built components for data connectivity and preparation, and then extend their analysis into tools such as Jupyter Notebooks. Data Science Experience provides a collaborative environment that supports all users of analytics within the organization, enabling people to leverage their own past work or that of others for greater productivity and shorter time-to-insight.

The user-friendly, spreadsheet-like interface of Bluemix Data Connect makes it easy to join data from different sources, assess its quality, filter out unwanted or sub-standard data, apply string transformations and unit conversions, and sort the data into the appropriate formats for downstream analysis. Bluemix Data Connect then transforms the data according to the defined actions and pushes it to the next stage. The interface automatically recognizes a growing range of standard data types such as dates, social security numbers, zip codes, addresses and telephone numbers, enabling these to be standardized where differences are present across different data sets.

A key benefit of using Bluemix Data Connect is that it enables a strategic approach to data preparation—versus the tactical approach of continually re-doing the same manual tasks in spreadsheets each time requirements change. Using the embedded Bluemix Data Connect functionality within Data Science Experience brings data sourcing, preparation and modeling into the same environment, empowering data professionals to iterate and refine both transformations and analyses. Once the desired results are achieved, the parameters can be preserved ready for re-use on a future occasion. Bluemix Data Connect retains all of the credentials for accessing the data sources, and all of the complexity around the transformations, so there is no need for any manual re-work when new analysis is required on updated data. In such a scenario, the data professional can simply run the desired job in Bluemix Data Connect, which then automatically loads and transforms the data ready for analysis.

Understanding and managing all data

By providing consistency and repeatability around data preparation, Bluemix Data Connect can help organizations to create a rich and valuable "data asset" as the first step towards a new kind of governance and value-add around enterprise data, in which multiple people from across the organization can collaborate on the management, enrichment and control of data.

Naturally, good governance is essential to the effective adoption of analytics within an organization, because people need to be able to trust data and trace its lineage. IBM is working intensively with the open source community to develop the open source Apache Atlas metadata solution. By providing rich metadata around connections, data sets, data profiling, classification, lineage, types, security and sensitivity, Apache Atlas will make it easier for data stewards within organizations to set and monitor policies around data governance.

For existing users of IBM Information Governance Catalog, its tight integration with Data Science Experience will enable properly governed hybrid analytics environments for the first time. Until now, the challenge for organizations with past investments in cataloging and displaying data governance policies and assets was how to extend these to cover self-service analytics tools in the cloud. By capturing metadata created in the cloud and enabling it to be managed alongside on-premises metadata, Data Science Experience enables organizations to control the risk of hybrid environments and bring analytics to a wider audience.

With end-to-end governance and pre-built connectors for multiple heterogeneous data sources, IBM Bluemix Data Connect and IBM Data Science Experience can help organizations adopt the data lake concept. This helps to drive new value from enterprise data by providing an abstraction layer that maps data to business semantics and makes it possible to manage access, lineage and versioning—regardless of where and how the data is actually stored at the back-end.

Powering up a real use-case

BlocPower, a startup based in New York City, used IBM Data Science Experience to build a predictive model for energy usage that will help building owners to cut expenses and emissions. The company imported an existing Jupyter Notebook from GitHub — one of more than 200,000 available on the service — and used the integrated tools in IBM Data Science Experience to:

- Import energy consumption data from object storage with a single click.
- Clean the data and visualize it, using the matplotlib tool in Python to explore correlations between energy usage and building characteristics such as size and age.
- Create a model to predict energy consumption in kWh for different buildings according to their unique characteristics (using a linear regression function in the scikit-learn Python library for machine learning).
- Classify buildings by efficiency using the K-means algorithm to cluster them according to their scores across four dimensions indicative of energy efficiency: gas use for heating, gas use for domestic purposes, electricity use for plugged equipment, and electricity use for air conditioning.
- Use matplotlib visualizations to help reduce from four clusters of buildings to two — interpreted as being the efficient and inefficient groups.
- Use Flexdashboard and Shiny in RStudio to create an interactive dashboard that plots buildings on a city map, giving users the ability to drill into them to check their energy efficiency compared to the average. The dashboard also provides predicted energy usage in kWh and the predicted cost.

By bringing together all of these diverse technologies in a single platform backed by significant community resources — in addition to automatically creating and managing the Apache Spark cluster used to crunch the data — IBM Data Science Experience dramatically cut the time-to-insight for BlocPower.

Fast, powerful, repeatable

By eliminating the need for a tangle of inconsistent and error-prone spreadsheets or complex procedural SQL scripts, Bluemix Data Connect can act as the central point of control for the preparation of data for analysis, managed by data professionals and supervised where appropriate by IT professionals. Bluemix Data Connect gives non-IT users the ability to securely access, clean and transform the data they need in a way that is consistent and repeatable. This enables the business to perform rapid analysis without burdening IT staff or jeopardizing governance and security around enterprise data.

Bluemix Data Connect makes it easy for data professionals to:

- Clean data – eliminating noise and inconsistencies between data sources
- Enrich data – by combining data sets for a more complete view
- Transform data – making it ready for downstream analysis.

Naturally, higher-quality data delivered faster to analytics tools will result in better insights being delivered sooner to business decision-makers. In combination with the numerous analytical tools available within IBM Data Science Experience, Bluemix Data Connect helps data professionals to:

- Explore the relationships between data sets–recognizing potentially interesting lines of inquiry
- Ask deeper questions about data–and build predictive models that enable machine-learning as a service
- Understand data in context–by enabling the recognition of causal relationships between entities.

As Bluemix Data Connect is embedded within Data Science Experience, it enables a seamless process of iteration between preparation and analysis, making it easier to refine and enhance analysis based on the results already achieved. This can create a virtuous circle or feedback loop, where the outcomes of analysis drive improved data enrichment techniques, which in turn drive better analysis. Equally, Data Science Experience is ideally suited for creating machine-learning applications for predictive analytics. And what is unique about Bluemix Data Connect and Data Science Experience is that they provide a single analytics environment that serves the needs of both highly advanced data scientists and less experienced line-of-business users.

For more information

To learn more about how Bluemix Data Connect and IBM Data Science Experience can cut the time and effort required to source business data and prepare it for analysis, visit datascience.ibm.com

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
December 2016

Please Recycle

CDW12365-USEN-00