EXECUTIVE BRIEF

Sponsored by

IBM

Successful Data Warehouse Approaches to Meet Today's Analytics Demands

In this Paper

- Organizations are adopting increasingly sophisticated analytics methods
- Analytics usage trends are placing new demands on rigid data warehouses
- What's needed is hybrid data warehouse architecture that supports all deployment models

CIO INSIGHT

A QuinStreet EXECUTIVE BRIEF © 2017

Introduction

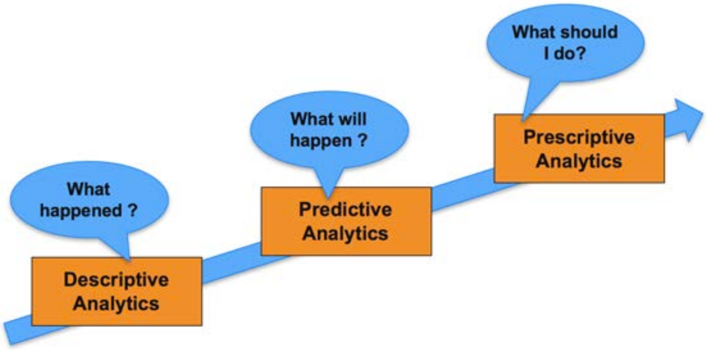
Analytics usage trends are placing new demands on rigid data warehouses. Organizations need to incorporate more relevant data and have self-service access to that data. Additionally, they must be able to do more than just spot historical trends. It is increasingly essential to add predictive and prescriptive functionality in applications to improve and automate decision making.

Specifically, organizations want to make use of more than just traditional relational data. They now must include new data sources such as unstructured data generated from mobile, web, and internet applications. They also want to apply a range of analytics (descriptive, predictive, and prescriptive) across larger, relevant data sets to uncover newer and better insights. Finally, they want speed to continuously refine the insight and applications supporting it.

The ultimate goal is to extract greater value from the data to facilitate, accelerate, or automate business decisions.

Business drivers elevate importance of analytics

Organizations are adopting increasingly sophisticated analytics methods. The evolution that is occurring builds on the success of enterprise reporting and dynamic dashboards that extract and summarize customer, inventory, financial, and other data to provide visibility into what has happened or is happening based on historical data. To complement these descriptive analysis tools and methods, companies are now using predictive and prescriptive analytics, which provide a view into what might happen next and even what business decisions should be derived based on that information.



Organizations use these sophisticated analytics techniques to gain insights into their financial and operational performance and their customers' behaviors. With these insights, organizations can make accurate predictions and better-informed decisions on emerging opportunities, competitive threats, and shifts in their markets to increase competitive advantage.

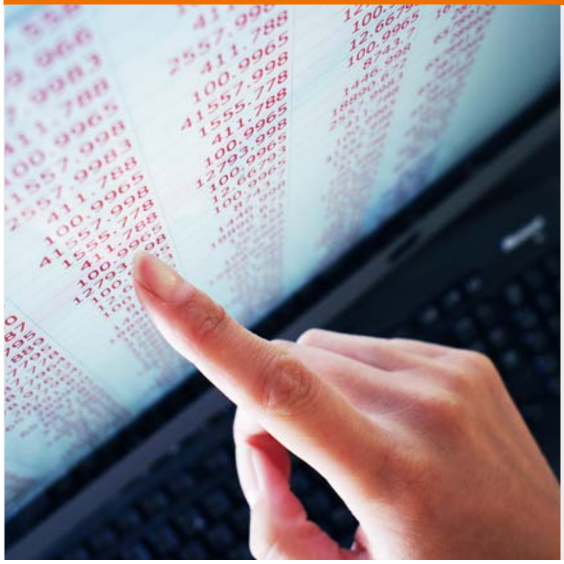
While these sophisticated techniques have long been available, there are many compelling business reasons driving today's adoption of advanced analytics to augment BI.

CIOs rank analytics as a major contributing factor[1] to an organization's competitiveness. Organizations that embrace analytics perform better than their peers. Furthermore, top performing organizations use advanced business analytics more often than low performers.[2]

Additionally, there is a direct correlation between the use of sophisticated analytics and the organization's bottom line.

What's needed?

When it comes to analytics efforts, organizations are facing a number of issues including limitations on underlying storage, lack of processing power to complete analyses in a timely manner, and the need to support older analytics software. These issues are limiting the value of on-premises data warehouses.



Traditional relational warehouses are still the much better choice for descriptive analytics because this is what they are designed and built for – the analysis of structured data to answer questions about what happened so far. Their SQL engines have great scalability and deliver results in a timely manner to be useful for business decisions. Today, predictive and prescriptive analytics are often done with popular open source frameworks such as Apache Spark and languages such as R and Python. These frameworks and languages are very feature rich and scalable to perform advanced analytics on both structured and unstructured data and to uncover answers to unknown questions that impact the business and answers to questions about the future.

To support traditional descriptive analytics on relational data as well as predictive and prescriptive analytics on structured and unstructured data requires a hybrid analytics architecture. That architecture must blend a relational analytics engine with a big data engine, such as Spark Apache, in order to cover the entire range of analytics. The hybrid analytics architecture must also support hybrid deployments of the analytic workloads on premises or in the cloud to accommodate both the massive amount of data untapped on-premises within an enterprise as well as the important new data sources on the Internet.

Organizations need this type of hybrid approach because the alternatives typically add complexity and cost and can limit what exactly can be done with analytics efforts. For example, one choice would be to replace a legacy data warehouse with a Hadoop-based big data system. This approach basically accepts the capability limitations of using SQL for advanced analytics and does not offer the required performance.

> "Analytics usage trends are placing new demands on rigid data warehouses."

A second alternative would be to operate separate systems for data warehouse and Hadoop-based big data analytics and replicate data between them. This introduces data currency and data governance problems. Some of these issues can be eased by remotely reading data into a data warehouse with each advanced analytic computation of the Hadoop system. But this approach can be time consuming, adds complexity, and lacks the flexibility needed to fine tune and modify analytics workflows.

What really helps an organization support today's demanding analytics requirements is a hybrid data analytics approach where advanced analytics compute, relational compute, and all data are collocated. This would allow an organization to run the entire range of analytics on a single copy of the large datasets in a timely manner without compromises for data currency or integrity. And this in turn would help the organization get value from the data in time where actions can be taken to meet the business requirements.

Other characteristics are needed to ensure analytics workflows can optimally run in an organization. What's needed includes:

- A rapid, data-driven application development environment that supports development and testing in the cloud
- Flexible scalability via hybrid on-premises and cloud infrastructure architecture

- Specialized and optimized data management software based on the various data types and query workloads

When selecting and deploying hybrid analytics architectures there are several challenges that also must be addressed. Organizations must have ways to deal with governance issues around self-service data access. They also need to address data integration, federation, virtualization, and quality issues. Finally, giving data scientists more analytics choices means they need new IT and business skills to use various approaches and get the most they have to offer.

IBM as your technology partner

The need for a hybrid data warehouse architecture that supports all deployment models requires a solution with special characteristics. This is an area where IBM can help.

IBM offers private and public cloud database solutions for transactional and analytic workloads, with IBM fully managed or client managed options. For analytic workloads, Db2 Warehouse on Cloud is a fully managed solution available on IBM Bluemix and AWS on the public cloud, and Db2 Warehouse is a client managed solution for Software Defined Environments in a private cloud.

IBM offers a common analytics engine with data virtualization across Db2 Warehouse products and on-premises data warehouses such as IBM Db2®, IBM PureData® System for Analytics (Netezza®), and Oracle SQL to enable the integration of Db2 Warehouse products into the data warehouse environment to support new data sources and analytic techniques users demand.

Db2 Warehouse products provide time-tested Netezza in-database analytic libraries and deep integrations with analytic languages such as R and Python. To meet today's more demanding analytics requirements and the desire to use more advanced (i.e., predictive and prescriptive) analytics techniques on data, IBM has enhanced the offering by integrating Apache Spark in the engine. (This offering is available today with Db2 Warehouse and will be released with Db2 Warehouse on Cloud in the future.)

There are several significant benefits with this integrated solution. To start, the solution delivers high performance when running analytics against a dataset. it offers a workload-optimized database that can handle either transaction or analytic workloads. It has built-in analytics, uses in-memory and massively parallel processing (MPP), and supports seamless connectivity.

With Db2 Warehouse, each node contains its own data and is overlaid with a local Apache Spark executor process. The co-location of the executors with the database engine processes minimizes the latency of accessing the data.

Combining Db2 Warehouse and Apache Spark speeds up analytics calculations and delivers fast answers to critical questions. Comparing a remote Spark cluster setup with a co-located setup can get a speed up factor of 3 to 5 times.

In addition to the performance benefits, the integrated Db2 Warehouse/Spark solution offers other advantages.

Users can leverage Spark's machine learning library to train and evaluate predictive models interactively and leverage visualization mechanisms to get graphical presentation of models.

A Jupyter notebook container is included that works with the Db2 Warehouse container as the Spark kernel to execute interactive code entered in the notebook. Using this container, a data scientist can explore the data in Db2 Warehouse. In this way, users, working in the familiar environment of their notebooks, data scientists, and business managers get out-of-the-box data exploration and visualization.

Once an analytics procedure or workflow has been created on a notebook, the Jupyter notebook container for Db2 Warehouse allows one-click-deployment in which the code on the notebook is transformed into a compiled and deployed Spark application inside Db2 Warehouse. The entire deployment process is automated, turning the code into deployed Spark applications.

When deploying the Spark application to Db2 Warehouse, it can be invoked in three different ways:

- From a command line or a script anywhere remotely using the spark-submit.sh command-line tool
- Running the application via a REST API
- Running the application via an SQL connection, which allows a data scientist to extend any existing SQL application with Spark logic, such as Microstrategy, Tableau, or Cognos reports

Another area where Db2 Warehouse helps is data management. Accessing data and transforming it into the right format for analytics routines are some of the greatest challenges when increasing the use of analytics and expanding analytics to more parts of an organization.

The range of new functional capabilities that are enabled through the built-in Spark in Db2 Warehouse go way beyond what traditional enterprise data warehouses can do, including integrated machine learning, sophisticated in-place data transformations, integration of other data types and sources (such as Parquet data and object storage), and integration of data in motion (through Spark Streaming).

Spark has an excellent framework to perform sophisticated data transformations. This comes in handy for situations where an organization has column values in its tables that need some form or feature extraction. With Db2 Warehouse, an organization can use the integrated Spark environment to run those types of transformations and extractions that are hard or even impossible to express in SQL and write the results back to the transformed table. (This type of in-warehouse transformation is often referred to as ELT.) Since the operations are run in an integrated Spark application the data does not have to leave the data warehouse at all during the entire transformation. This helps address the governance issues mentioned above.

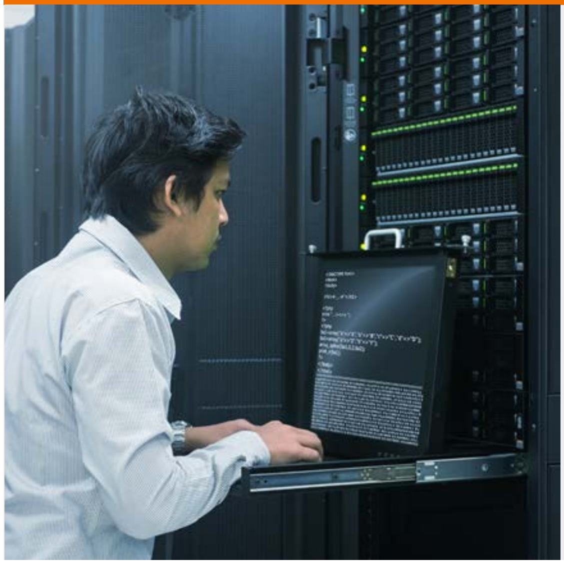
> "Organizations are adopting increasingly sophisticated analytics methods."

Spark can also be used to read source data not just from a table inside Db2 Warehouse but from any other remote source as well. Once the appropriate data is accessed and put into the right format, it can then be written into a relational table inside Db2 Warehouse. So an organization can use Spark effectively as a parallelized ETL mechanism in Db2 Warehouse.

Another capability of Db2 Warehouse is that Spark can be used for processing streaming data. Using Spark's streaming API, a user can deploy and run applications in Db2 Warehouse that directly subscribe to some message hub and permanently process and insert the relevant messages into Db2 Warehouse tables.

Taken together, the features and capabilities of a hybrid analytic architecture with Db2 Warehouse offers significant benefits:

- Db2 Warehouse lets an organization dramatically **modernize its data warehouse environment through a hybrid analytics architecture**, as Db2 Warehouse can apply advanced



analytics based on Spark against data managed in Db2 Warehouse and on-premises data warehouse through federated queries.

- Spark applications processing relational data gain significant **performance and operational QoS benefits** from being deployed and running inside Db2 Warehouse
- Db2 Warehouse enables **end-to-end creation of analytics solution**, from interactive exploration and machine learning experiments, verification of analytic flows, easy operationalization by creating deployed Spark applications, up to hosting Spark applications in a multi-tenant enterprise warehouse system and integrating them with other applications via various invocation APIs
- Db2 Warehouse allows you to invoke Spark logic via **SQL connections**
- Db2 Warehouse can **land streaming data** directly into tables via deployed Spark applications
- Db2 Warehouse can run complex **data transformations and feature extractions** that cannot be expressed with SQL using integrated Spark

Summary

In today's instant-everything world, the speed at which business decisions are made can have a huge impact on the success or failure of a project, product, or service. Many organizations are turning to advanced analytics and big data tools to accelerate the transition from data to decision.

Since the inception of data warehouses, their use has been the backbone of supported descriptive analytics efforts on relational data. Unfortunately, today's analytics requirements mean more is needed. Expansive use of predictive and prescriptive analytics on unstructured data and streaming data are best tackled with tools and capabilities offered by Apache Spark and Python and R.

Rather than running two distinct analytics operations, IBM Db2 Warehouse offers a hybrid analytics approach that integrates the Db2 Warehouse relational engine with a Spark analytics engine, and such a Db2 Warehouse node contains its own data.

This makes Db2 Warehouse ideally suited to serve as next-generation data warehousing and analytics technology for use in private clouds, virtual private clouds, and other container-supported infrastructures.

For more information, visit IBM Db2 Warehouse.

[1] http://www-935.ibm.com/services/c-suite/study/studies/cio-study/

[2] http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=XB&infotype=PM&htmlfid=GBE03729USEN&attachment=GBE03729USEN.PDF