Eliminate data silos: Query many systems as one

Data virtualization in IBM Cloud Pak for Data

Highlights

- Query across multiple databases and big data repositories, individually or collectively
- Centralize access control and governance
- Make many databases—even globally distributed—appear as one to an application
- Simplify data analytics with a scalable and powerful platform

Background

Data is everywhere and the best businesses in the world today are data-driven. Businesses are collecting data from more and increasingly diverse sources to analyze and run their operations, with those sources perhaps numbering in the thousands or millions. The complexity, cost, time and risk of error in collecting, governing, storing, processing and analyzing that data centrally is increasing exponentially. In parallel, the databases and repositories that are the sources of all of this data are more powerful, with abundant processing and data storage capability of their own available and at hand.

Data virtualization overview

Data virtualization in IBM® Cloud™ Pak for Data (formerly IBM Cloud Private for Data), is a unique new technology that connects all these data sources into a single self-balancing collection of data sources or databases, referred to as a *constellation*. See Figure 1. No longer are analytics queries performed on data copied and stored in a centralized location. The analytics application submits a query that's processed on the server where the data source exists. Results of the query are consolidated within the constellation and returned to the original application. No data is copied and it exists only at the source.

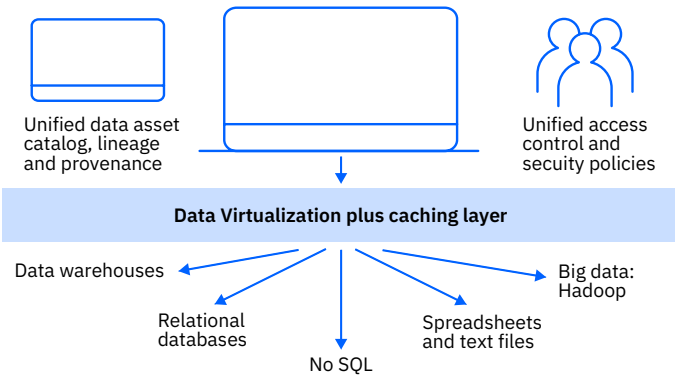

Figure 1: Data Virtualization in Cloud Pak for Data

How data virtualization works

Applications connect to IBM Data Virtualization as if they are connecting to a single IBM Db2® database. When connected, applications can submit queries against the system as if they were querying a single data source database. The workload will be collaboratively distributed and computed by all participating data sources that have data relevant to the query.

Features that matter

There are a number of important features in IBM Data Virtualization that enable businesses to more effectively work with their data.

Collaborative computing

By using the processing power of every data source and accessing the data that each data source has physically stored, latency from moving and copying data is avoided. In addition, all repository data is accessible in real time, and governance and erroneous data issues are virtually eliminated. There's no need for extract, transform and load (ETL) and duplicate data storage, accelerating processing times. This process brings real-time insights to decision-making applications or analysts more quickly and dependably than existing methods. It also remains highly complementary with existing methods and can easily coexist when it remains necessary to copy and move some data for historical, archival or regulatory purposes.

Schema folding

A common scenario in distributed data systems is that many databases store data in a common schema. For example, you may have multiple databases storing sales data or transactional data, each for a set of tenants or a region. IBM Data Virtualization can automatically detect common schemas across systems and allow them to appear as a single schema in data virtualization—a process known as schema folding. For example, a SALES table that exists in each of 20 databases can now appear as a single SALES table and be queried through Structured Query Language (SQL) as one virtual table.

Simple join view tools

Elegant inline tools makes it possible to define table views across databases of different types and perhaps geographically distributed as shown in Figure 2.

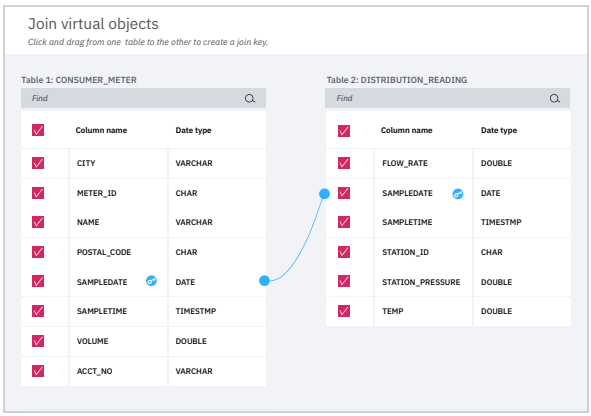

Figure 2: Intuitive interface makes it simple to join table views

Security

All communication within the constellation and back to the application is encrypted with security-rich, robust and powerful IBM technology, and Secure Sockets Layer (SSL) and Transport Layer Security (TLS) encryption using standard protocols.

Performance

IBM Data Virtualization's design and architecture of peer-to-peer computational mesh, lends a significant advantage over traditional federation architecture. Using advancements from IBM Research, the data virtualization engine is able to rapidly deliver query results from multiple data sources by leveraging advanced parallel processing and optimizations. Collaborative highly paralleled compute models provide superior query performance compared to federation, up to 430% faster against 100TB datasets[1]. IBM Data Virtualization has unmatched scaling of complex queries with joins and aggregates across dozens of live systems.

Not only is IBM Data Virtualization fast, it automatically finds databases and tables, making querying information from multiple data sources simpler. Queries can easily combine data from multiple sources including relational databases, NoSQL sources, spreadsheets and flat files.

Platform support

IBM Data Virtualization appears to an application as a single instance of a Db2 database. As a result, popular Db2 connection clients and applications can attach to IBM Data Virtualization and work without modification. This is the case even if the collection of data sources under query includes a mix of many types of data sources, such as:

- PostgreSQL
- Oracle
- Netezza®
- Microsoft SQL Server

The IBM Data Virtualization technology converts to and from all the SQL dialects. Therefore, your applications can freely code SQL, procedural language/SQL (PL/SQL) and SQL PL as if they are working directly on the Db2 database without trying to determine if the syntax is supported by the target data system. For example, popular tools are able to connect to IBM Data Virtualization without any modification or upg ade, including:

- IBM Cognos® Business Intelligence (BI) software
- Tableau
- MicroStrategy
- Looker
- Plotly
- R
- Jupyter

The data virtualization service node that applications connect to is a microservice within Cloud Pak for Data.

Apache Hive	IBM Informix® database server
Cloudera Impala	MariaDB
Db2 software	MySQL
IBM Db2 Big SQL	Netezza
IBM Db2 Event Store	Oracle
DerbyDB	PostgreSQL
Excel and Comma Separated Values (CSV) file	SQL Server
Hortonworks Data Platform (HDP) with Apache Hive	

Table 1: Supported data sources

Minimum hardware requirement

Data Virtualization in Cloud Pak for Data requires the following configuration:

- Processor with 16 (v) cores
- At least 64 gigabytes of physical random access memory (RAM)
- Recommended 200 gigabytes of disk space

Common scenarios for IBM Data Virtualization

IBM Data Virtualization is well suited to perform analytics on highly distributed data sets where the data and the analytics results are time-sensitive. It's also effective where the analytics may be a one-time operation on that specific set of data. Plus, it's applicable to scenarios where the latency for batch copying from some data sources exceeds the business need for analytics results.

Many organizations duplicate data and create new data repositories to satisfy the needs of the lines of business (LOB) for analytics. This process requires configuring physical assets and creating and maintaining new ETLs to load and transform the data to those repositories. Often the data is out of date by the time it becomes available to the LOB.

Existing approaches are reaching the saturation point for many IT organizations. With the number and diversity of data sources and need for analytics increasing, this approach is no longer scalable. IBM Data Virtualization can increase the productivity of IT organizations and provide a scalable approach for LOBs to access enterprise-wide data.

In many instances, there are policy or legal issues with copying or moving data, for example, personal information. These restrictions can get in the way of a business need for demographic analytics results. IBM Data Virtualization helps resolve these issues by leaving the protected data at the source and only returning the demographic query result.

Today, a data scientist must create a data lake, copy data from the sources of interest and integrate that data before being able to test out hypotheses with analytics. IBM Data Virtualization eliminates the need for the data lake, allowing data scientists to federate the data they require to test hypotheses by connecting tools like IBM Watson® Studio directly to the data sources.

Deliver agility to key analytical projects

The simplicity offered by IBM Data Virtualization allows users to acquire actionable, unified data when they want, in the way they want, at the speed matching their analytical needs. This technology leads to faster integration speed and performance, and improved decision-making that helps you adapt to changing business demands.

IBM Data Virtualization in Cloud Pak for Data supports a range of key initiatives, including:

- Modernization for faster, easier delivery of modern systems of engagement
- Real-time analytics that meet the immediate needs of the business
- Optimization to reduce the cost and complexity of accessing organizational data

IBM Data Virtualization enables self-service BI. The virtual, reusable data assets provide a business-friendly representation of data, allowing the user to interact with data without having to know the complexities of the physical data layer or where the data is stored. It also allows multiple BI and reporting tools to acquire data from a data virtualization layer.

IBM Data Virtualization provides a unified 360-degree view. The virtualized data asset delivers a complete view of data in real time. The virtual data layer serves as a unified, integrated view of business information that improves a user's ability to understand and use organizational data.

IBM Data Virtualization provides agile service-oriented architecture (SOA) data services. A data virtualization layer delivers the data services layer to SOA applications. It speeds the creation of virtual assets without the need to touch underlying sources and by autodiscovery and mapping of metadata that encapsulate the data access logic. Data virtualization also allows multiple business services to acquire data from a centralized location and provides loose coupling between business services and physical data sources.

IBM Data Virtualization provides improved control of information. It improves data quality through centralized access control, a robust security infrastructure and reduction in physical copies of data, thus decreasing risk. The metadata repository catalogs an organization's data stores and the relationships between the data in various data stores, enabling transparency and visibility.

Summary: Transform and expedite decision-making

Data Virtualization in Cloud Pak for Data is ideal for organizations seeking:

- Profitability, growth and risk reduction
- Agility and productivity boosts
- Optimization of existing IT investments

It improves the use of existing server and storage investments while reducing unnecessary data replication and the associated costs of duplication and infrastructure management. With simplified admini tration and a set of SQL application programming interfaces (APIs), it enables your business to derive benefits from real-time analytics.

For more information, try out Cloud Pak for Data at no cost, or schedule a consultation. We also invite you to dive deeper into our product details by visiting us on the web.

1. Performance measurements were gathered within a controlled test environment at IBM Silicon Valley Labs using IBM data virtualization against various 100TB data sources. The measurements taken in May 2019 and performance gains are compared to IBM federation.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
January 2019