Data store convergence

How IBM AutoSQL delivers universal data access and querying at a single point across disparate data sources

Highlights

- Have a single view of data across all types and sources
- Run warehouse-grade queries against data lakes
- Execute distributed and virtualized queries 53% faster
- Use a range of data management solutions such as Db2, Netezza, Event Store, Spark and Hadoop

Rising data volume and variety have pushed organizations to store that data within a myriad of different repositories across different vendors and locations. While the principle of storing data where it fits best is a good one, the fragmented nature of such environments is taking its toll. 74% of potentially valuable data is not used.[1] The problem is even worse for big data, where 88% is not used effectively.[2]

Physical combination of this data through extract, transform and load (ETL) processes or making data copies has been used in an attempt to bridge the gaps between these silos. However, the result has often been complex, time consuming and expensive. This is particularly true when queries need to be adjusted to fit the varying types of data or their repositories. The movement of unstructured data in a data lake to a data warehouse for analysis alone can be a significant time-sink and budget killer.

IBM offers an alternative with IBM® AutoSQL, a universal distributed query engine that can converge and query all data stores without physically moving them. This technology allows users to not only see data from all repositories with a single view, but query them with no movement, replication or manual adjustments needed. The value of this simpler approach cannot be overstated; a typical Fortune 1000 company is estimated to gain USD 65 million in additional net income for a 10% increase in data accessibility. This paper dives deeper into the technologies that permit converged data queries, along with the data and AI platform and data stores that can augment their success.

Key capabilities driving data store convergence

AutoSQL enables complete and unhindered data access

AutoSQL is a key capability that simplifies access to data by leveraging data virtualization, cloud object storage, and automated governance to facilitate the convergence of data across clouds, data lakes, data warehouses, and databases. AutoSQL turns this set of technologies from a convenience to a must-have performance booster. It goes beyond simply abstracting data through data virtualization by executing warehouse-grade queries against an enterprise's data, including data lakes and streaming data, across a hybrid landscape. Without AutoSQL technology, considerable work is needed to get the data ready to be queried, or conversely, prepare the query to access data where it resides.

With the AutoSQL universal query engine, warehouse-grade queries using SQL can be run across all data repositories, whether that data happens to be structured, unstructured, or somewhere in between. In other words, there's no need to move the data or adjust the query manually as the query will automatically be adjusted to fit the data and its location. And it does this with blazing speed—executing distributed and virtualized queries 53% faster than the industry standard.[3] As such, data in Hadoop or formats used in streaming solutions such as Apache Parquet can be queried just as easily and effectively as data from awarehouse. With 80% of worldwide data estimated to be unstructured by 2025, such capabilities are not just nice, they're necessary.

The value of AutoSQL comes into sharp focus when you think about the interaction between the data lake and the data warehouse. Traditionally, unstructured data in the data lake would need to be moved into a data warehouse to take advantage of its superior querying ability. If that's no longer necessary, data can be left on much cheaper object storage and queried as though it were in a data warehouse. Because warehouse storage costs are roughly ten times those of data lakes, the enterprise will save both money and time.

AutoSQL can also independently scale compute and storage based on a company's needs thanks to its ability to leverage cloud object storage and comes with automated, embedded governance. And for even more assistance, data virtualization now offers autonomous cache creation. This feature uses a recommendation engine to identify long-running and commonly executed queries and provide a rank order list of potential queries to add to a cache. Collectively these features help reinforce the enterprise-grade querying now possible across the entire business with AutoSQL.

Other capabilities through IBM Cloud Pak for Data

Other enterprise features are available through IBM Cloud Pak® for Data, the data and AI platform through which IBM's AutoSQL and data virtualization capabilities are used. In this way, they become part of an intelligent data fabric that includes automated capabilities for data governance, data privacy and AI. With this data fabric, not only will businesses be able to move and query data stores with ease from a single self-service point, they will also be able to add metadata as data comes in, mask private data for those without a legitimate need, and make a smooth connection with the systems that turn the data into insight.

IBM Cloud Pak for Data is also built upon a Red Hat® OpenShift® foundation. This means the platform can go wherever OpenShift does, even onto other vendors' clouds. This level of openness helps users to avoid the hassle and expense of vendor lock-in. Furthermore, in instances where the platform needs to be placed in close proximity to the data, the highly containerized, mobile nature of the platform makes achieving that reality simpler than it has been previously.

As previously mentioned, the AutoSQL capability is underpinned by critical technologies such as data virtualization, automated governance, and cloud object storage.

Each of these technologies has significance for the scope of the AutoSQL capability as a whole:

Data Virtualization
IBM's data virtualization capability is built using data federation with an abstraction layer running on top. In this way, it provides direct access to databases, data warehouses, open-source repositories such as Hadoop, and streaming data stores without requiring the data to be moved. Moreover, the data can be accessed even when dispersed geographically or across on-premises and cloud deployments. It can also be accessed across multiple vendors' products, helping reduce costs typically associated with vendor lock-in.

Despite these advantages, the biggest savings from data virtualization comes from its reduction of costly ETL processes. Using data virtualization, ETL requests are expected to be reduced between 25% and 65%. Over a three-year period, this can provide savings between USD 0.9 and 2.4 million."[4] These savings are possible for several reasons. First, any data transfer fees are no longer incurred because the data stays in one place. There are also considerable efficiencies for employees who can put their time toward more valuable activities. Instead of transferring, copying and generally waiting on the data they need, they can access it right away through a single access point. Data virtualization, therefore, lays the groundwork for robust data access without the traditional hassle.

Automated data governance
IBM's data governance capability enables organizations to discover, curate, analyze, prepare, and share data. Furthermore, data governance supports the automatic application of industry-specific governance rules across an enterprise. Data governance ensures that an organization's data complies with defined rules and processes. Organizations can implement revised or new regulations with speed and precision, potentially avoiding costly fines for noncompliance.

Visit IBM's data governance page →

Cloud object storage
Cloud object storage (COS) is a solution for storing massive amounts of data. COS enables organizations to put in place data growth and resilience strategies to accommodate the explosion of data. To accommodate varying data requirements, COS offers organizations varying classes and ties that provide the required flexibility and performance necessary for data accessibility. Furthermore, COS provides high-speed data transfer to satisfy latency, bandwidth, and cost requirements. With the AutoSQL capability, data can also be left in a much cheaper object store data lake and queried without needing to move it.

Visit IBM's cloud object storage page →

The AutoSQL capability optimizes data accessibility and availability for organizations. The universal distributed query engine:

- Minimizes the need for multiple query engines
- Minimizes the need for data migration and replication
- Reduces the data warehouse footprint and associated costs
- Provides flexibility due to its vendor agnostic design and independent storage and compute scaling
- Embeds automated governance

Data management options

The goal of being able to access data no matter where it is stored and query it easily is also aided by a group of data stores that can efficiently and effectively meet the needs of a data-driven, AI-ready organization. IBM provides many data stores capable of fulfilling those needs.

IBM Db2—the AI database

IBM Db2® databases boast a long history of enterprise-grade performance and have recently been augmented with capabilities that make them both powered by AI and built for AI.

Powered by AI
Machine learning query optimization
- SQL performance is monitored over time, allowing models to be created and optimized for specific SQL statements. More efficient access paths are used, leading to faster query execution and reduced resource consumption.

Confidence-based querying
- Machine learning is used to score the accuracy of query results based on the previous accuracy of historic query results.

Adaptive workload management
- Using machine learning, workload runtimes are monitored and used to both adjust ongoing workloads and predict utilization. 30% database performance improvements have been observed.[5]

Built for AI
Native Graph functionality
- Multi-model data management in Db2 takes advantage of Graph databases' ability to support dynamic multidimensional data management while reducing the expense of having a separate database.

Native blockchain support
- The Db2 Blockchain connector presents blockchain data as a Db2 relational table, allowing it to be analyzed alongside Db2 data.

Language support
- REST APIs are supported, along with languages such as PYTHON and GO, architectures such as JSON, and collaborative development environments such as Jupyter Notebooks.

Enterprise-grade
IBM BLU Acceleration
- Combine in-memory computing, massively parallel processing (MPP), actionable compression, data skipping, and column-based shadow tables to heighten speed and performance without interfering with transaction reliability.

Backups and recovery
- Use IBM Db2 pureScale® clustering technology and geographical dispersion to avoid outages. All sync modes are supported for HADR as well as change-queue-based replication and change data capture replication.

Security and encryption
- Integrate with centralized enterprise key managers that support Key Management Interoperability Protocol 1.1 and can be hosted around the world to comply with regulatory requirements.

Read the Db2 solution brief →

IBM Db2 Warehouse

IBM Db2 Warehouse shares many of the same qualities as the Db2 database and offers extras designed to improve analytical workloads:

IBM BLU Acceleration
The BLU Acceleration technology described previously will also help speed analytics workloads.

Resiliency
Unhealthy compute nodes are automatically detected by the cloud provider's native Kubernetes service, which removes the node from the cluster and delivers a new one from a hot standby pool or provisions one just in time.

Analytical capabilities
Algorithms including Association Rules, ANOVA, k-means, Regression, and Naïve Bayes can be run against data. Native Python drivers and integration into Jupyter Notebooks is also supported.

Read the Data Warehouse ebook →

Netezza Performance Server

Netezza® Performance Server comes from a long line of data warehouse appliances built upon the idea of extremely high performance with minimal effort.

Simplicity
Out-of-the box performance with little to no indexing or tuning lowers administrative needs and ongoing maintenance. Built for resiliency; failure of a node leaves no significant performance degradation.

Speed
Unique asymmetric massively parallel processing (AMPP) and IBM's patented hybrid columnar acceleration assist provide fast results. With faster cores and advanced NVMe flash drives it can support thousands of users with heightened speed.

Smarts
A library of more than 200 prebuilt, scalable, in-database analytic functions is included. This includes in-database geospatial analytics that are compatible with industry-standard ESRI GIS formats.

Read the Netezza Performance Server solution brief →

Data lake and open-source options

IBM offers a number of options for data lake and open source data management needs, particularly for Hadoop.

Cloudera
IBM partners with Cloudera to provide Hadoop implementations that are perfect for those seeking to create or improve their data lake.

Big Match
Data lake performance is improved by this enterprise-ready technology that matches multiple fragmented or duplicate records associated with the same customer. It uses pre-configured algorithms to score similarity and match records natively within Hadoop.

MongoDB and PostgreSQL
IBM provides the support needed to take full advantage of the JSON document storage and high-volume data storage of MongoDB as well as the object-relational database PostgreSQL.

Read the data lake ebook →

IBM Db2 Event Store

Event Store is designed specifically for the ingestion and analysis of high-speed streaming data.

High-speed ingestion
Land 3 million events per second using a 3-node system and ingest more than 250 billion events in a day with just those 3 nodes if required.

Open source ready
Use Apache Parquet's columnar data format to store data for universal access and easy integration into open source stacks.

New and existing data
Analyze the data ingested by Event Store alongside historical data to provide more accurate, informed insights.

Read the Event Store solution brief →

Next steps

Don’t stand by helplessly while your data disperses into a wide variety of incompatible repositories. Instead, converge those repositories without ever moving the data and query them as though they were all a single data warehouse. The time and effort saved will allow your business to derive more valuable insight, faster, with performance improvements to spare.

Give IBM Cloud Pak for Data a try today for free so you can see the difference for yourself. Or, if you have any questions, book a free meeting with an IBM expert who would be happy to share their advice.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
June 2021

1 IDC/Segate Rethink Data survey, 2020

2 sigma study 2020

3 Based on internal testing

4 New Technology: The Projected Total Economic Impact of IBM Cloud Pak for Data https://www.ibm.com/downloads/cas/V5GNQKGE

5 Based on IBM internal testing