IBM Db2 Big SQL

Make better decisions fast with a powerful
Db2 engine for big data analytics

Highlights

- Infuse the power of Db2 on big data for SQL analytics
- Enterprise-grade ANSI SQL engine for data on Cloudera Data Platform or object store with high concurrency and performance
- Low latency support for ad-hoc and complex queries, security and federation capabilities
- Robust granular row and column-level security for accessing data
- Sub-second latency for star-schema queries by enabling MQTs on aggregated data
- Easy scaling out or in based on workloads for parallel execution in the containerized Big SQL

The Internet of Things (IoT), artificial intelligence (AI), social media and mobile applications are driving an increase in data volume, velocity and variety. To capitalize on this trend and obtain faster actionable insights, many organizations are either deploying a data platform from Cloudera or a containerized private cloud platform for complete data lifecycle use cases with data in object storage like IBM COS or Ceph.

IBM Db2 Big SQL is a hybrid ANSI-compliant data virtualization tool for accessing, querying and summarizing data across the enterprise that:

- Provides a massively parallel processing (MPP) architecture
- Time-tested ANSI SQL compliant SQL engine optimized for open source file formats
- Provides easy application portability (e.g. Cognos, Tableau, MicroStrategy, etc.)
- Instantly combines data from disparate sources across the enterprise in a single query using Federation
- Provides high performance and supports high concurrency for business intelligence workloads
- Enables granular access to data and can be controlled using advanced row and column-level security
- Understands dialects from other vendors and various products from Oracle, IBM® Db2® and IBM Netezza®

IBM Db2 Big SQL is now available in two formats: as a cloud-native service on IBM Cloud Pak for Data, IBM's unified data and AI platform, or as an integrated service on Cloudera Data Platform.

Use cases

- Get deep and timely insights into credit exposures across customer, industries and geographies
- Simplify data supply chains and provide accessibility to data for actionable insights
- Bring the SQL processing power of Db2 to big data by reusing SQL skills with no retooling costs
- Empower power SQL users to query data in data lakes
- Provide an analytics workspace to support hundreds of users to concurrently query data with high performance
- Employ a big data query engine that can seamlessly process SQL workloads generated by tools as-is

Features of IBM Db2 Big SQL include:

High performance

IBM Db2 Big SQL is built with an advanced SQL compiler and cost-based optimizer, so that complex and analytical SQL workloads maintain high performance, concurrency and scalability. It provides an ANSI-compliant SQL parser that can run 99 TPC-DS queries and structured streaming with new APIs.

Integration with Apache Spark

IBM Db2 Big SQL integrates with Spark for easier insight delivery and faster processing. This integration will also enable operationalizing of machine learning models with fresh data, not only from Hadoop but also from disparate sources. This feature is limited to IBM Db2 Big SQL on Cloudera Data Platform. IBM Cloud Pak for Data has its native Spark service, which shares metadata information with Big SQL.

Federation capabilities

IBM Db2 Big SQL uses a single database connection, enabling you to access data across Hadoop and relational databases, whether they are on the cloud, on premises or both. It also includes Fluid Query capabilities to enhance virtualization with various data warehouses. You can also federate with S3 Object Storage and WebHDFS (technical preview only).

Compatibility with multiple SQL dialects

Being compatible with a number of SQL dialects makes the platform well-suited for RDBMS (including Db2, Netezza and Oracle) offload as well as fast and easy consolidation. Data can be offloaded from existing enterprise data warehouses or data marts to free up capacity while preserving most of the familiar SQL.

Enhanced performance using MQTs

IBM Db2 Big SQL offers an array of performance features to help improve the performance of queries in the Hadoop ecosystem. One of these features is the ability to create and automatically match materialized query tables (MQTs) to incoming queries. MQTs can simplify query processing and improve performance because expensive joins or aggregation operations can be calculated and cached in MQTs. Once these MQTs are created, Big SQL will automatically rewrite queries when possible to improve performance by taking advantage of the newly created MQTs.

Advanced workload management

Workload management (WLM) offers the ability to monitor and control statements executing on the cluster to make efficient use of cluster resources, ensuring that the cluster is not over- or underutilized. At a high level, a certain number of concurrent queries can be executed so as to not oversaturate the cluster resources. When this threshold is exceeded, incoming work is queued until some of the earlier queries have completed.

Elastic scalability

IBM Db2 Big SQL offers the only SQL engine able to successfully run 99 TPCDS queries up to 100TB with numerous concurrent users. It also has the ability to run multiple workers per node for efficient CPU and memory utilization. With its cloud-native capabilities, IBM Db2 Big SQL on IBM Cloud Pak for Data can instantly scale out/in or up/down based on workload needs.

Security-rich SQL

Robust role-based access control (RBAC), row-based dynamic filtering, column-based dynamic masking, and Apache Ranger integration are included with IBM Db2 Big SQL to provide centralized security administration and auditing for data lakes.

Standards-compliant Open Database Connectivity and Java Database Connectivity

For developers, the usage pattern allows you to access the database with specific products or tooling that allow only Open Database Connectivity (ODBC) or Java Database Connectivity (JDBC).

Access to JSON and XML data using built-in SQL JSON functions

One of the challenges when accessing JSON data is sometimes referencing the fields in the JSON itself. The acknowledged strengths of JSON can also contribute to its complexity, particularly in the area of parsing. In addition to built-in support using SerDes, IBM Db2 Big SQL now has built-in JSON SQL functions that will allow the user to store and query JSON data in the same manner that XML data can be handled using the set of XML functions provided in IBM Db2 Big SQL.

The value of IBM Db2 Big SQL on IBM Cloud Pak for Data

As Db2's high-performant engine on IBM Cloud Pak for Data, IBM Db2 Big SQL integrates with a unified data and AI platform that modernizes how businesses collect, organize and analyze their data to infuse AI throughout their organization. Utilizing an open source architecture and containerized services for data management, analytics and AI/ML, IBM Cloud Pak for Data provides the information architecture needed to flexibly build, deploy and manage AI solutions across any cloud.

Big SQL helps data engineers and data scientists analyze structured and unstructured data already in open source file formats in big data stores. An elastically scalable, cloud-native engine, it can help organizations drive down IT costs and speed innovation throughout their digital transformation.

By introducing Big SQL on the platform, users can:

- Drive additional workloads based on Hadoop use cases
- Exploit different formats of common SQL engines on a single platform
- Query data where it resides and scale up or down based on workload or concurrency needs

IBM Db2 Big SQL is ideal for:

- Enterprises that are modernizing their information architecture and looking for a solution to instantly access data that is already in object stores or other big data stores
- Organizations looking to leverage the benefits of separating compute and storage so they can scale out or in independently
- Companies whose data is already moved to cloud platforms like AWS or Azure but find that the SQL solutions on these platforms are not meeting analytics SLAs
- IT operators who want to reuse tools or applications seamlessly without the need to reinvest when modernizing data

The value of IBM Db2 Big SQL on Cloudera Data Platform

As an advanced SQL engine on Hadoop, IBM Db2 Big SQL on Cloudera Data Platform supercharges your analytical workloads on data lakes (physical and virtual) with no vendor lock-in.

The core capabilities of IBM Db2 Big SQL focus on data virtualization, SQL compatibility, scalability, performance and enterprise security and governance, making it a desirable query engine to seek insights from disparate data sources including Hadoop.

IBM Db2 Big SQL on Cloudera Data Platform brings value when:

- Analytics use cases are not fulfilled by Hive or Impala on the platform
- Use cases were exploiting traditional SQL capabilities and organizations want to extend it to big data
- Performance suffers when hundreds of concurrent users access the data in HDFS
- Companies want to reuse tools or applications that were written for traditional data warehouses without investing more time in rewriting them
- IT staff or developers want to reuse ANSI SQL knowledge to write new applications and also use ML in SQL

For more information

Learn more about Db2 Big SQL on IBM Cloud Pak for Data and Cloudera Data Platform, or contact your IBM representative or IBM Business Partner.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
November 2020