IBM Enterprise Data Analytics with Cloudera

Modern data platform with advanced SQL on Hadoop, optimized for the cloud, machine learning and analytics

Highlights

– One platform with many applications for data science & engineering, data warehousing and operational databases.

– Enterprise-grade SQL-on-Hadoop for ad hoc and complex queries, high performance, security and federation.

– Ecosystem of IBM and Cloudera products and services available.

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, organizations are deploying different storage solutions for Big Data. To offer a standardized and robust platform, IBM and Cloudera have partnered to offer a single solution that includes the Cloudera Enterprise Data Hub and IBM ® Db2 ® Big SQL platforms.

Cloudera Enterprise Data Hub (EDH)

EDH is a modern data platform for machine learning and analytics optimized for the cloud. This enterprise-grade platform provides the scale and performance required for today's modern data workloads, security and governance demands of today's IT. Cloudera's modern platform makes it easy to bring more users—thousands of them—to petabytes of diverse data and provides industry-leading engines to process and query data, as well as develop and serve models quickly. The platform also provides several layers of fine-grained security and complete audibility for companies to prevent unauthorized data access and demonstrate accountability for actions taken.

Cloudera Enterprise Data Hub key features:

Shared data experience. Eliminate costly and unnecessary application silos by bringing your data warehouse, data science, data engineering, and operational database workloads together on a single, integrated data platform. Cloudera SDX enables these diverse analytic processes to operate against a shared data catalog that preserves business context like security and governance policies and schema. This common services framework persists even in transient cloud environments and makes it easier for IT to set and enforce policies while enabling business access to self-service analytics.
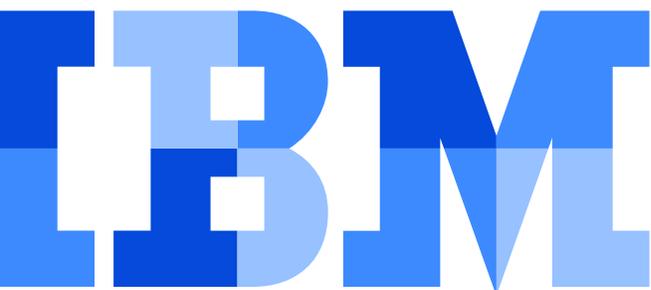
Hybrid deployment. Work where and how it's most convenient, affordable, and elective. Cloudera Enterprise can read direct from and write direct to cloud object stores like Amazon S3 and Azure Data Lake (ADLS) as well as on-premises storage environments, or HDFS and Kudu on IaaS. This provides flexibility to work on the data that you want wherever it lives, with zero copies or moves. Cloudera also provides the most popular data warehouse and machine learning engines that can run on any compute resource for ultimate deployment flexibility. Its hybrid control means users can self-service via PaaS offering, or opt for more configurability and management via IaaS, private cloud, or on-premises deployments.

Powerful open source. Cloudera develops and validates the best of open source innovations into one seamless, rock-solid platform. Key features include:

– In-Memory Data Processing: The longest and deepest experience with Apache Spark
– Open Source Leadership: Constant open source development and curation, with the most rigorous testing, for trusted innovation

Db2 Big SQL key advantages

Db2 Big SQL is an advanced SQL engine optimized for advanced analytics in a Big Data environment. IBM infuses the power of industry leading Db2 in open source technologies like machine learning, interactive and batch analytics. With Big Data, there is overwhelming amount of data but unless you plan and execute the search for meaning in that data, it is impossible to reap the benefits of modernization. Some key aspects Db2 Big SQL accelerates in digital transformation journey are discussed here.

Drive advanced data fabric for analytics with data virtualization

Enterprises are recognizing the importance of digital transformation. Digital transformation is about the need to derive new business value from digitizing and optimizing operations rather than finding the needle in the haystack. This involves unlocking hidden silos of data that was never available for analysis before. Thanks to better insights, businesses can enjoy better decision-making and product innovation.

Together, Cloudera EDH and Db2 Big SQL form an advanced data fabric designed to help manage large-scale clusters within complex data environments. With this better management, users can improve predictions, find new business opportunities, and implement improved workflows, including:

– Query data where it resides, including Hive, HBase, Spark, traditional RDBMS and NoSQL databases.
– Employ smart federation to enable enhanced data virtualization with various data warehouses and also provide access to S3 object storage.
– Perform analytics across the entire enterprise via a single database connection that accesses data from Hadoop and other sources (whether on cloud, on premises, or both).
– Seamlessly exploit capabilities such as querying a table in a database using polymorphic table function.
– Operationalize machine learning models with fresh data, not only from Hadoop but also from disparate sources using SQL.

Seize cost savings opportunities with ANSI SQL dialects and compatibility

While Hadoop provides high scalability, Db2 Big SQL's advanced cost-based optimizer and massively parallel processing (MPP) architecture can execute queries faster and smarter. Its ability to run more concurrent users and complex SQL queries with less hardware makes it attractive solution to cut costs while reaping the benefits of data warehousing on commodity hardware. Db2 Big SQL is the only SQL-on-Hadoop solution that understands different dialects of SQL from vendors and products, such as Oracle, Db2 and Netezza. It has high compatibility with ANSI SQL standards with support for PL/SQL as well. With these rich capabilities, users are more able to:

– Migrate existing applications without major rewrites
– Port BI applications for business intelligence tools including Cognos® BI, Tableau and others
– Offload costly ETL processing to free EDWs to perform analytics and operations
– Archive data from an EDW that is running out of capacity

Db2 Big SQL provides elastic scalability with its ability to successfully run all 99 TPCDS queries up to 100TB with numerous concurrent users. It also can run multiple workers per node for efficient CPU and memory utilization.

Empower your data scientists and engineers with simplified data access for your end users

Business analysts and data scientists run interactive queries to explore and understand gigabytes to petabytes of data before building models or charts. With its unmatched scalability and performance, Db2 Big SQL empowers users and applications to execute complex queries smarter, unlocking insights from data with their analytic tools of choice while achieving high concurrency for BI workloads by executing complex queries smarter. Administrators can provide widespread data access to users, while still ensuring only necessary data can be accessed, by anonymizing or masking data for users who do not have correct privileges. Such advanced capabilities enables self-service analytics in a governed and safe manner. While keeping data secure, Db2 Big SQL opens up the following opportunities:

– Advanced row and column security with no backdoors to data is possible thanks to low-level integration with Apache Spark
– Data access is permitted with specific products or tooling that only allow Open Database Connectivity (ODBC) or Java Database Connectivity (JDBC)—a usage pattern of particular interest to developers
– Robust role-based access control (RBAC), row-based dynamic filtering, column-based dynamic masking, and Apache Ranger integration are included. These provide centralized security administration and auditing for data lakes
– Short rapid queries can be used to search by key words or key word ranges. Db2 Big SQL uses HBase when random, real-time read/write access to data is needed
– Data scientists can access data directly using their tool of choice and build, test and deploy models seamlessly

One platform. Many uses. Designed for your needs.

Cloudera EDH and Db2 Big SQL provide one platform for many uses and is specifically built for your enterprise's needs.

For more information

To learn more about IBM Enterprise Data Analytics with Cloudera, visit ibm.com/analytics/partners/cloudera or contact an IBM data management expert.