SOLUTION SHEET
IBM BIG SQL

Make Better Decisions Fast With Powerful SQL-On-Hadoop For Big Data Analytics

THE POWER OF APACHE™ HADOOP®, AND APACHE SPARK™

Apache Hadoop is an open source software project that enables distributed processing of large structured, semi-structured and unstructured data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a high degree of fault tolerance. With Hadoop, organizations can run large-scale, distributed analytics jobs on clusters of cost-effective server hardware. This infrastructure uses the Hadoop MapReduce framework to tackle large data sets by breaking up the data across many nodes and coordinating data processing across a massively parallel environment.

Apache Spark is an open source, in-memory compute engine combined with a stack of advanced analytics capabilities. Its in-memory processing capabilities enable extremely fast processing speeds and virtually any type of iterative processing, because it persists immediate data in memory for reuse. The Spark core—the underlying processing engine—supports a range of applications plus Java, Scala, R and Python application programming interfaces (APIs). Additionally, the Spark analytics stack supports near-real-time streaming, graph computation, built-in machine learning libraries and SQL querying of unstructured data. Spark was developed for use with Hadoop, and does not have a data store. Investing in both technologies enables a broad set of big data analytics and application development use cases, which translates into solving bigger problems and faster time to value.

HDP: A NO-CHARGE, OPEN SOURCE DISTRIBUTION OF HADOOP AND SPARK

Hortonworks Data Platform (HDP®) is built on 100 percent open source Apache ecosystem components as if you had downloaded components from Apache.org and built it yourself. HDP is a Hadoop and Spark distribution provided at no charge, and it is certified by the Open Data Platform initiative (ODPi). HDP is designed with analytics, operational excellence and security empowerment in mind. As a result, it offers an optimal combination of Apache components to maximize the development of big data applications.

IBM BIG SQL: MAKING YOUR BIG DATA SQL ACCESSIBLE

With HDP as its foundation, IBM Big SQL enables users to query Hive and HBase data using ANSI-compliant SQL. While Hadoop is highly scalable, the Big SQL advanced cost-based optimizer and massively parallel processing (MPP) architecture can run queries smarter, not harder, supporting more concurrent users and more complex SQL with less hardware compared to other SQL solutions for Hadoop. Big SQL is also an effective platform for data warehouse offloading and consolidation—a key use case for many Hadoop users. Big SQL is the first and only SQL-on-Hadoop solution to understand commonly used SQL syntax from other vendors and products such as Oracle, IBM Db2 and IBM Netezza®. And where data can’t be moved to Hadoop, Big SQL provides federated access to relational database management system (RDBMS) sources outside of Hadoop with IBM Fluid Query.
A FULL RANGE OF ANALYTICS FOR HADOOP, SPARK AND SQL

- Big SQL supports virtually all forms of data, batch and real-time processing, with unmatched scalability and performance.
- Big SQL helps improve Hadoop users’ productivity with:
  - Superior SQL-on-Hadoop performance using elastic boost technology
  - The ability to understand SQL dialects from other vendors and products for Oracle, Db2 and Netezza, which helps make it faster and easier to offload old data from existing enterprise data warehouses or data marts to free up capacity, while preserving most of the SQL from those platforms
  - A hybrid engine for Hadoop that exploits Apache Hive™, Apache HBase™ and Spark concurrently for best-in-class analytic capabilities
  - Deeper integration with Spark than other SQL-on-Hadoop technologies, enabling new use cases
  - Advanced row and column security: The integration of Big SQL and Spark enhances “shopping for data” security because sensitive attributes can be masked by default with no backdoors to the data, which empowers self-service access to data in a safe and governed manner
- Access the power of the Spark 2.1 analytics stack, including integrated analytics capabilities such as invoking Spark jobs from Big SQL and obtaining interactive and rapid in-Hadoop analytics from integrated DSX and Jupyter Notebooks.
- Get more value from freeform text faster and easier with intelligent, web-based tooling powered by Spark and SQL.

STREAMLINED AND SCALABLE BIG DATA MANAGEMENT

Big SQL and HDP help improve operational efficiency by modernizing—not replacing—the data warehouse environment. Big SQL and HDP were designed for ease of use and automation, helping IT administrators work efficiently to manage workloads, schedule jobs and reduce risk.

The solution can be used as a queryable archive, enabling organizations to store and analyze large volumes of poly-structured data without straining the data warehouse. It can also be used as a preprocessing hub, sometimes referred to as a “landing zone” for data.

Other benefits include:

- Object store support (Swift, S3)
- Automated tuning to help ease the pressure on organizations to add specialized resources as big data use cases grow
- Workload management tools that enable automated admission control for classes of concurrent queries to avoid over stressing the CPU, memory resources and disk I/O
- A self-tuning memory management infrastructure that frees administrators from managing and tuning each worker node for performance and concurrency
- A low-latency parallel execution infrastructure that accesses Hadoop data natively for fast reading and writing
- In-memory caching that places data closer to compute resources
- An ecosystem of powerful, enterprise-class components to complement Big SQL, including modules for data integration, data cleansing, metadata management, advanced statistical analytics, data replication and more.

OPEN AND HYBRID: THE MODERNIZED HADOOP PLATFORM

A mature Hadoop platform is a critical analytics foundation, and a big part of modernizing a Hadoop environment involves evolving to a hybrid architecture for agility and efficiency. However, many of today’s Hadoop distribution platforms include elements that limit migration. The very nature of exceedingly large, multi-petabyte Hadoop clusters makes them difficult to move between vendor installations simply because of the storage volumes. Shifting data across on-premises and cloud implementations presents another challenge.

Big SQL and HDP provide flexibility, elasticity and production-grade performance with container technologies on fine-tuned systems both on premises and in the cloud. This flexibility enables organizations to have a seamless hybrid implementation for data ingest, high availability and disaster recovery. The solution also lets organizations use a wide range of services and data on the cloud for self-service analytics and application development.

INTEGRATED AND SEAMLESS: BIG SQL AND HDP FIT WITHIN YOUR LARGER DATA FABRIC

Big SQL and HDP architecture is based on many years of experience managing large-scale Hadoop clusters within complex data environments. Features include:

- Configurations to suit and perform on virtually every type of Hadoop workload
- Highly optimized service distribution and management node configurations with high-availability options
- APIs and integration designed for ease of use
- Tooling to federate queries across multiple types of data management systems, such as traditional data warehouses and cloud-based data marts
- Support on IBM Power®, which enables faster SQL, Spark, machine learning and artificial intelligence results from a rich family of Linux servers to match workloads, plus three times the compute and storage infrastructure reduction with Power and IBM Elastic Storage™.
ENTERPRISE-READY

Available on premises, in the cloud and integrated with other systems in use today, Big SQL is leading the way in open source analytics.

To learn more about IBM Big SQL, contact your Hortonworks representative or visit here.

BIG SQL — ENABLES COMPLEX ANALYTIC SQL AND DATA VIRTUALIZATION

About Hortonworks

Hortonworks is an industry leading innovator that creates, distributes and supports enterprise-ready open and Connected Data Platforms and Modern Data Applications that deliver actionable intelligence from all data: data-in-motion and data-at-rest. Hortonworks is focused on driving innovation in open source communities such as Apache Hadoop, Apache NiFi and Apache Spark. Along with its 1,800+ partners, Hortonworks provides the expertise, training and services that allow customers to unlock transformational value for their organizations across any line of business.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Contact
For further information visit www.hortonworks.com
+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

© 2011-2016 Hortonworks Inc. All Rights Reserved.
Privacy Policy | Terms of Service