Powering SQL on Hadoop

Evaluation of TPC-DS Query Benchmark Using Hadoop

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) White Paper

March 2017

PREPARED FOR: kognitio

EMA™

IT & DATA MANAGEMENT RESEARCH, INDUSTRY ANALYSIS & CONSULTING

Powering SQL on Hadoop: Evaluation of TPC-DS Query Benchmark Using Hadoop

Table of Contents

Powering SQL on Hadoop: Evaluation of TPC-DS Query Benchmark Using Hadoop

Organizations welcome Hadoop into their environments for a number of reasons. First, enterprises can derive additional insights from multi-structured data. Hadoop provides a reliable, scalable framework for distributed processing of large, complex, or unstructured data that arrives at significant volumes. Second, traditional database platforms bring a licensing model that can significantly increase license and maintenance costs. These costs can amount to significant barriers to entry for small organizations starting their data management efforts and ongoing expenses for larger companies.

Unlike the start of the Hadoop "revolution," business stakeholders predominantly use environments that implement Hadoop (not data scientists). Data scientists, who explore and discover new connections in the data using programmatic tools such as Java and Python, will always be involved with Hadoop. However, business stakeholders from marketing, finance, and sales use the information stored in Hadoop for analytical and business intelligence use cases. Their expectations for complex workloads, data access, and speed of response are the core drivers associated with whether Hadoop can earn its place at the table in regard to being an enterprise-grade option as an analytics environment.

IN THE KNOW

WHO: Chief Data Officers (CDO), Chief Information Officers (CIO), and data architects of data-driven organizations.

WHEN: As big data analytics supports a wider business audience.

WHAT: How to handle complex analytics workloads and business intelligence on Hadoop by implementing a mature SQL on Hadoop implementation.

EMA big data research conducted regularly from 2012-2016 shows that organizations implementing Hadoop as part of their data management environments are moving from simple exploratory use cases to complex workload-driven analytics use cases. These organizations are working for the implementation of mission-critical business goals and workloads in environments that utilize Hadoop. According to EMA panel respondents, the most commonly implemented use cases are market basket analysis, social brand management analysis, fraud analysis and risk assessment, geospatial grouping, and relationship analysis. All of these projects require access to speed of response and many different groups across the organization.

These advanced workloads are SQL-intensive, requiring a full range of current ANSI SQL query support—not just a subset. In addition, many workloads must support the type of low-latency response business stakeholders require. Finally, these platforms need to support a level of concurrent users who will use these environments. When analytical applications gain adoption, their use becomes widespread and the underlying data management platform needs to support users while providing an acceptable level of performance.

An effective SQL on Hadoop environment must support advanced ANSI SQL versions such as SQL99 and SQL2003 to effectively maintain analytical and business intelligence query workloads. This is essential for business users who are loath to make adjustments from the queries used currently. Full SQL support is also critical when an enterprise plans to utilize data analysis and data visualization tools, such as SAS and Tableau, which generate their own SQL in ways that the data consumer and IT technologies do not fully control.

TPC Benchmarks and the TPC-DS

The Transaction Processing Performance Council (TPC) was created to provide standards and benchmarks associated with the computer processing and data management industries. The intent of the TPC benchmarks is to provide a common framework for the evaluation of functional compliance, performance measurements, and price to performance comparisons for a range of domains and use cases across multiple hardware architectures, software products, and implementation strategies.

The TPC designed, refined, and provided certification criteria for a number of different benchmarks. These benchmarks include a vendor and configuration testbed for performance of processing and data management implementations. Specifically, the TPC-DS benchmark[1] was established to provide measurement for a complex workload analytics/business intelligence environment or a decision support (DS) system. The components specified in the TPC-DS benchmark include:

- Access and processing of large volumes of data
- Answer real-world business questions within an analytics/business intelligence environment
- Execute SQL queries of various operational requirements and complexities (ad-hoc, standard reporting, analytical and data mining workloads)
- Test high CPU and I/O loads associated with a complex workload environment
- Synchronize with source operational (OLTP) databases through database maintenance functions

The latest version of the TPC-DS considers how analytical and BI environments include big data implementation strategies and technologies. With this in mind, TPC-DS v2.x was designed for SQL on Hadoop big data environments with specifications for the storage of data within file types supported by Hadoop Distributed File System (HDFS), such as parquet[2] and orc.[3] With these adjustments, the TPC-DS retained all key characteristics of an analytics and business intelligence framework while benchmarking against the wider audience of big data and SQL on Hadoop.

Kognitio Testing Process

In the winter of 2017 Kognitio, a provider of in-memory database management platforms and technology, focused on the ability of SQL on Hadoop platforms to meet the specifications of TPC-DS v2.3.[4] Specifically, Kognitio evaluated how those platforms could meet the requirements of SQL query syntax, functional query execution, and concurrency load aspects required by analytics and business intelligence environments.

Kognitio compared the performance of the following platforms:

- Apache Impala 2.6.0
- Kognitio on Hadoop v8.1.50
- Apache Spark 2.0 Beta

All of these SQL on Hadoop platforms performed the TPC-DS queries in separate test runs on the same 12-node Hadoop cluster, running Cloudera CDH 5.8.2.

[1] http://www.tpc.org/tpcds/
[2] https://parquet.apache.org/
[3] https://orc.apache.org/
[4] It should be noted that Kognitio's performance of SQL syntax evaluation, functional, and concurrency load aspects of the TPC-DS benchmark was not intended to be a completely certified version of the benchmark.
That complete evaluation includes components on the pricing of the underlying platform, as well as a certified performance/price/value comparison. The goal(s) of the Kognitio TPC-DS testing was to provide an equal footing evaluation of several SQL on Hadoop platforms for SQL compliance, query-functional, and load performance.

SQL Query Syntax Support

Each platform performing a TPC-DS benchmark is allowed to make minor modifications to the auto-generated TPC-DS queries. These minor modifications account for differences in SQL query implementation on a platform-by-platform basis. In terms of SQL support for the TPC-DS, the following results were observed during the Kognitio-performed testing:

	Impala	Kognitio	Spark
Without adjustment	57	76	72
Minor changes	18	23	27
Syntax not supported	24	-	-

Figure 1: SQL Query Syntax Support

Apache Impala supported 75 of the TPC-DS queries either out-of-the-box or with the allowed minor modifications. The remaining 24 queries could not support the SQL syntax of the TPC-DS queries. Both Kognitio on Hadoop and Apache Spark supported all 99 of the TPC-DS queries.

Functional SQL Query Performance

For the initial run of the TPC-DS queries, the testing executed the generated queries (with minor allowed modifications) against a 1TB instance of the TPC-defined dataset in a single query stream. Apache Impala and Apache Spark stored their information in Hive tables using the Parquet file format. Kognitio stored its information in flat files that the Kognitio platform later ingested and stored in-memory.

TPC requires that the TPC-DS run queries randomly to simulate the ad-hoc nature of queries against the 1TB dataset. They run in a single query stream to test the functional results of the platform and its ability to return the results. This part of the testing provides validation of a platform's ability to support the functional requirements of the SQL queries or whether the platform can run the query against the dataset. Tests were performed 10 times to ensure the random nature of the TPC-DS execution did not positively or negatively impact the results for any one platform.

	Impala	Kognitio	Spark
Query ran	73	99	89
Long running	2	-	10
Fastest query speed	6	92	1

Figure 2: Functional SQL Query Performance (1 query stream)

As indicated in the chart, Apache Impala executed 73 of the queries with results. Two of the TPC-defined queries did not return results on Apache Impala within 60 minutes and were considered "long running" queries – not failed, but without results. Overall, Apache Impala returned six queries with the top performance during the functional query testing.

Kognitio executed all 99 queries. All results were returned in less than the 60-minute threshold for "long running" queries. Kognitio on Hadoop returned the fastest results on 92 of the 99 queries.

Apache Spark performed all 99 queries. However, 10 of those queries exceeded the "long running" query threshold. Apache Spark returned the fastest result in a single instance of the 99 queries.

Concurrent Query Stream Performance

To test the concurrent and more intensive nature of an analytics and business intelligence environment, 10 streams of simultaneous queries were tested. This tests is designed to evaluate the ability of the platform to serve multiple user communities in the form of data consumers running ad-hoc queries, reporting platforms, executing standard reporting workloads and data visualization, or analytical platforms executing requests to populate dashboards and/or analysis platforms. Since the results of a single query stream showed little variation in the performance results, the test performed the concurrency load exam once per platform.

	Impala	Kognitio	Spark
Queries executed	68	92	79
Long running	7	7	20
Fastest query speed	12	80	-

Figure 3: Concurrency Testing Performance (10 query streams)

Apache Impala was only able to complete 68 of the 99 TPC-DS queries during the concurrency load testing. Like Kognitio, Impala had seven queries that did not return results before the 60-minute threshold; unlike Kognitio, it provided no support for the remaining 24 queries. Under these conditions, Apache Impala returned 12 of the queries in the fastest time.

Kognitio on Hadoop executed 92 of the 99 TPC-DS queries. Kognitio also had 7 queries that failed to return results before the "long running" threshold. Kognitio had 80 of their 92 queries return the fastest results.

Apache Spark executed 79 of the TPC-DS queries, but the remaining 20 failed to return results within 60 minutes during this concurrency load testing, thus failing the "long running" test. None of the Apache Spark queries were the fastest during this component of the test.

EMA Perspective

In the world of big data analytics (specifically SQL on Hadoop), business stakeholders require their analytical applications to work the same as they do on traditional RDBMS platforms. Stakeholders will expect their SQL queries to be processed in a new environment with minimal adjustment. They will not adopt and endorse a new platform that requires significant adjustments from past platforms. Platforms that cannot support advanced SQL functionality will be isolated and used for limited use cases (such as exploration). They will not be promoted to a co-equal status for use across the organization.

The results of the Kognitio TPC-DS testing show that when Apache Impala, Kognitio on Hadoop, and Apache Spark run on the same hardware footprint and under the same TPC-designed conditions, there are significant differences in how these SQL on Hadoop platforms perform in terms of SQL syntax support, functional query performance, and more complex concurrency load performance.

For SQL syntax support, Kognitio supported all 99 of the TPC-DS queries. This shows Kognitio's maturity of SQL support required to meet the expectation of business stakeholders and the data visualization and analytical platforms that they chose to use. Apache Spark also met this functional query support test. Apache Impala still needs additional development and maturation of SQL support; it failed to support the syntax in nearly 25% of the queries designated by the TPC-DS.

For functional query execution, Apache Impala provides reasonable performance for the query syntax that it supports. Apache Spark provides support across the board for the TPC-DS queries, but often has difficulty returning result sets within the acceptable timeframes of the TPC-DS. The Kognitio platform meets the requirement of functional query execution the fastest of all three benchmarked platforms.

The concurrency load testing of the evaluation places the SQL on Hadoop platforms under a fair amount of stress to support 10 streams of queries against 1TB of data. Under this part of the testing, "long running" queries were not included in the results to avoid impacting the overall results of all 3 platforms tested. With this exclusion, the Kognitio run testing would not meet the full requirements of a validated TPC-DS benchmark.

In EMA's opinion, Kognitio and Apache Spark show the most support for SQL and the functional execution of the TPC-DS benchmark. Apache Impala, however, still must work to support the full slate of current analytical functions of the TPC-DS queries. The overall performance of Kognitio on Hadoop shows that it has an advantage over Apache Spark to not only support the SQL syntax and execution of the queries, but to perform those queries with a high level of performance that business stakeholders have come to demand and expect.

With these considerations, of the SQL on Hadoop platforms evaluated, Kognitio on Hadoop provides the most coverage and performance for analytical and business intelligence workloads for organizations implementing big data analytics environments.

With these considerations, of the SQL on Hadoop platforms evaluated,

Kognitio on Hadoop provides the most coverage and performance

for analytical and business intelligence workloads for organizations implementing big data analytics environments.

About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help EMA's clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals, and IT vendors at www.enterprisemanagement.com or blogs.enterprisemanagement.com. You can also follow EMA on Twitter, Facebook, or LinkedIn.

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com

3532.030817