COMMISSIONED BY

ACTIAN
a division of HCLSoftware

MCKNIGHT CONSULTING GROUP

DATA WAREHOUSE PERFORMANCE

PRODUCT EVALUATION:

Actian Data Platform, Google BigQuery, Databricks and Snowflake

RELATED RESEARCH

Enterprise data integration landscape: Top capabilities and vendors exposed

MCG Stream Processing Ease of Use and TCO

William McKnight

Art Vancil

Prepared by

MCKNIGHT CONSULTING GROUP

McKnight Consulting Group
www.mcknightcg.com
September 2024

Contents

Executive Summary

Data-driven organizations rely on analytic databases to load, store, and analyze volumes of data at high speed to derive timely insights. Data volumes within modern organization's information ecosystems are rapidly expanding—placing significant performance demands on legacy architectures. Today, to fully harness their data to gain competitive advantage, businesses need modern scalable architectures and high levels of performance and reliability to provide timely analytical insights.

This report outlines the results from a McKnight Consulting Group Analytic Field Test derived from the industry standard TPC Benchmark™ H (TPC-H)[1] to compare Actian Data Platform, Google Big Query, Databricks and Snowflake. This test produced interesting results that reveal some of the performance characteristics of the four platforms.

With comparable specifications, Actian performs almost 8 times faster than Databricks, over 6 times faster than Snowflake, and over 12 times faster than BigQuery with TPC-H-like tests. When 5-user concurrency is considered, Actian performs 3 times better than Databricks, over 7 times better than Snowflake, and 9.6 times better than BigQuery.

Actian provided a price-performance that was over 8 times better than that of Snowflake and BigQuery.

These results suggest that Actian is a great choice for anyone looking to access large analytic data sets quickly and economically. Additionally, given the significant speed and cost savings that Actian provides, it is an excellent solution for organizations with large complex data sets that need to be accessed quickly and affordably.

[1] More can be learned about the TPC-H benchmark at http://www.tpc.org/tpch/.

Platform Summary

Big data analytics platforms load, store, and analyze volumes of data at high speed, providing timely insights to businesses. Data-driven organizations leverage this data, for example, for advanced analysis to market new promotions, operational analytics to drive efficiency, or for predictive analytics to evaluate credit risk and detect fraud. Customers are leveraging a mix of relational analytical databases and data warehouses to gain analytic insights.

This report focuses on relational analytical databases in the cloud, because deployments are at an all-time high and poised to expand dramatically. The cloud enables enterprises to differentiate and innovate with these database systems at a much more rapid pace than was ever possible before. The cloud is a disruptive technology, offering elastic scalability vis-à-vis on-premises deployments, enabling faster server deployment and application development, and allowing less costly storage. For these reasons and others, many companies have leveraged the cloud to maintain or gain momentum as a company.

This report compares Actian Data Platform, Google Big Query, Databricks and Snowflake—relational analytical databases based on scale-out cloud data warehouses and columnar-based database architectures. Despite these similarities, there are some distinct differences in the platforms.

Actian Data Platform

Actian Data Platform is based on its underlying technology, known as Vector. The basic architecture of Actian Data Platform is the Actian patented X100 engine, which utilizes a concept known as "vectorized query execution" where processing of data is done in chunks of cache-fitting vectors. Vector performs “single instruction, multiple data” processes by leveraging the same operation on multiple data simultaneously and exploiting the parallelism capabilities of modern hardware. It reduces overhead found in conventional "one-row-at-a-time processing" found in other platforms. Additionally, the compressed column-oriented format uses a scan-optimized buffer manager.

The measure of Actian Data Platform compute power is known as Actian Units (AU). At the time of this writing, the Actian Data Platform is priced at $2.50 per AU per hour. This price includes both compute and cluster storage.

Google BigQuery

Google BigQuery is a managed service with some interesting distinctions. Google abstracts the details of the underlying hardware, database, and all configurations. BigQuery is a hands-off database without indexes or column constraints. Defragmentation and system tuning are not required. It is serverless. Google Cloud manages the servers in a fully hands-off manner to the customer, dynamically allocating storage and compute resources. The customer does not define nodes and capacity of the BigQuery instance. The provisioning of compute is particularly fast and seamless.

You pay for the amount of data you query and store. Customers can pre-purchase computation "slots" for as short as one minute and billed by the hour. There is a separate charge for active storage of data.

Databricks

Databricks is a unified analytics platform that combines data warehousing, data lakes, and machine learning capabilities. It offers a scalable and collaborative environment for data teams to analyze, process, and extract insights from large datasets.

Databricks uses a pay-as-you-go pricing model, charging based on the resources consumed, such as compute power, storage, and network bandwidth. The exact pricing depends on various factors like the type of workload, region, and usage patterns.

Snowflake

As a cloud-only, fully managed solution, Snowflake has a clear separation between compute and storage. For Snowflake on GCP, which is what we used for the queries, data is stored in GCP GCS and is cached when queries are executed to bring the data in closer proximity to compute resources. Snowflake essentially offers two configuration "levers" — the size of the warehouse cluster and how many clusters are permitted to spin up to handle concurrency. Snowflake scales by cluster server count in powers of 2 (i.e., 1, 2, 4, 8, 16, and so on.) If enabled, Snowflake will spin up additional clusters to handle multi-user concurrent query workloads. Snowflake would automatically spin the additional clusters down once demand has passed. If not enabled, it will place paused queries in a queue until resources free up.

For Snowflake, you pay a flat hourly fee for when hourly compute resources are being used. We paid $3.00 per hour for the Enterprise tier. Once the compute warehouse goes inactive, you no longer pay. However, there is a separate charge for data storage.

Test Setup

The setup for this Field Test was informed by the TPC Benchmark™ H (TPC-H)[2] spec validation queries. This is not an official TPC benchmark. The queries were executed using the following setup, environment, standards, and configurations.

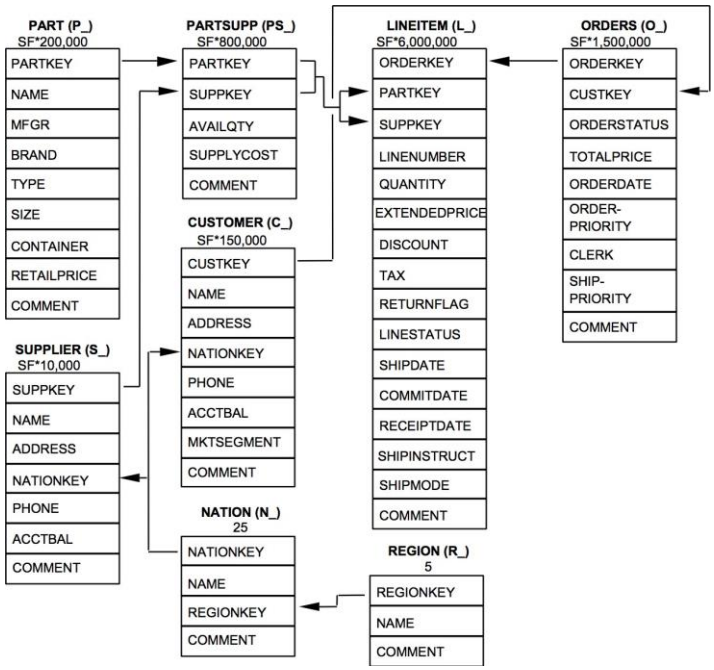
Benchmark Data

The data sets used in the benchmark were a workload derived from the well-recognized industry standard TPC Benchmark™ H (TPC-H).

From tpc.org: "The TPC-H is a decision support benchmark. It consists of a suite of business-oriented ad-hoc queries and concurrent data modifications. The queries and the data populating the database have been chosen to have broad industry-wide relevance. This benchmark illustrates decision support systems that examine large volumes of data, execute queries with a high degree of complexity, and give answers to critical business questions."

To show the data model, the following diagram was taken from page 13 of the TPC-H Revision 2.17.3 specification document.

Figure 1. TPC-H Data Model



[2] More can be learned about the TPC-H benchmark at http://www.tpc.org/tpch/.