DataWalk Performance Test: Exceptional Results With Vast Amounts Of Data

Executive Summary

In a performance test using a very large data set, the DataWalk software platform exhibited exceptional results in the time it required for loading, indexing, querying, and processing of the data into a usable analytical product. Notably, it accomplished these steps with a very low utilization of resources. Especially given the unusual size and characteristics of the data set used, the test clearly demonstrated DataWalk's ability to handle large amounts of data—quickly and cost-efficiently.

Performance Test and Configuration

DataWalk recently conducted a performance assessment of its software platform using a large dataset of uniform distribution, which represents the most challenging scenario in terms of large data volume with high cardinality (with very little repetition in values). For this test, 495 billion objects (13TB of data) were loaded into DataWalk's database for testing of the platform's data loading, indexing, querying, and computation performance.

The testing was conducted on a hardware configuration of commodity servers, consisting of:

- One application server (AWS m5a.2xlarge with 8 CPUs and 32 GB RAM)
- One integration server (AWS m5a.2xlarge with 8 CPUs and 32 GB RAM)
- Sixteen compute servers (AWS r6i.4xlarge each with 16 CPUs and 128 GB RAM)

With the above configuration, the compute component had an aggregate total of 256 CPUs and 2.048 TB of RAM. The total AWS infrastructure cost for the entire configuration would be $15,825 per month, or $189.9K per year.

As detailed below, the performance test results reflected DataWalk's exceptional capabilities for handling vast amounts of data.

Rapid Loading & Indexing

Loading and indexing this large dataset for the first time into DataWalk demonstrated remarkable speed. In the benchmark test, 20 datasets were loaded in parallel, with a total of 495 billion objects, in just under 6 hours. This included a 185 billion-record dataset that drove the 6-hour completion time. The primary constraint was the speed of the network, rather than the software itself.

Loading Summary	Measurement
Data Volume	495B objects (13 TB)
Total Load & Index Time	21,431.10 seconds (5.95 hours)
Effective Load Rate	8,598,991 objects/s
	600 MB/s

Fast Querying

After loading and indexing the dataset, DataWalk's query performance was measured. The dataset used in the test could not be partitioned due to its characteristics, so for each query the entire dataset had to be scanned in its entirety.

Despite this worst case scenario, querying the dataset was very fast: 20 entity types with 495 billion objects were scanned to generate a 360º view of a 2,500-object population in less than 2 minutes. This is analogous to a scenario of scanning 20 fact tables with 495 billion objects against 2,500 dimensions in under 2 minutes.

Query Summary	Measurement
Data Volume	495B objects (13 TB)
Queries	20
Query Time	1 m 56 s

Fast Time to Usable Results

As is typically the case in DataWalk, the analytic results it generated during the performance test required no additional processing. This resulted in data that was immediately usable in its final form as link-chart data. Had the querying been done on an alternative system with a traditional stand-alone database, the output would have been a series of raw data (e.g. CSV files) requiring additional processing time to be combined into analytically usable data—which could require several hours.

Because DataWalk outputs the data into an immediately analytical form, it eliminates the computational burden posed by stand-alone databases, resulting in dramatically faster time-to-results. Since most cloud database services are charged according to access time, DataWalk's low computation time (on top of its rapid loading, indexing, and querying) translates into significant cost savings in the long run.

Low Resource Utilization

As illustrated in figure 1 below, from DataWalk's monitoring service, during the execution of the above queries, CPU and RAM utilization on each server was remarkably low. CPU utilization was typically only a few percent (with a single peak of 25%), and RAM utilization was consistently less than 5%.

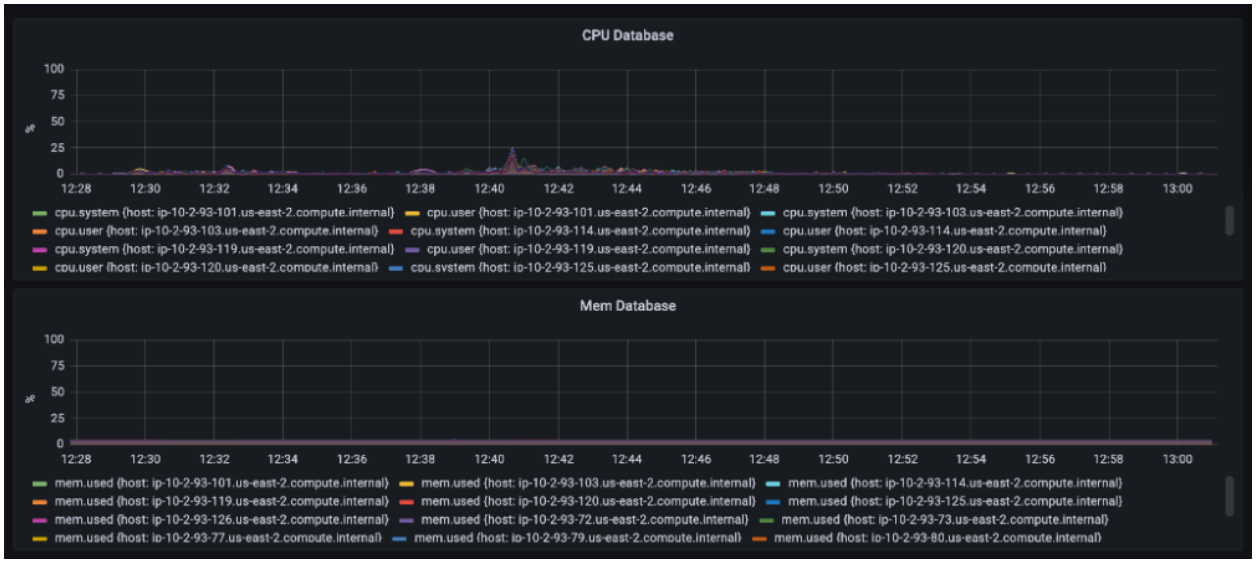

Figure 1. Screenshot of CPU and memory utilization measurements.

Also worth noting is that, unlike some alternative data analysis systems, DataWalk does not require the entire data set to reside in memory. As a simple example, DataWalk executed these tests with an aggregate RAM configuration of just over 2TB, while an alternative in-memory system would require at the very least 13TB of RAM just to hold the data. In practice, in-memory systems often require far more RAM than just the volume of the data being analyzed, so 13TB is actually an understatement.

DataWalk's low utilization of processing and memory resources provides significant advantages:

- More headroom for additional workloads
- Reduced requirements for hardware infrastructure, translating into lower costs

Conclusion

As these recent performance tests demonstrated, the DataWalk software platform has a number of exceptional capabilities for handling large amounts of data. With its combination of rapid loading, indexing, and querying, fast time to results, and low resource utilization, DataWalk offers a high degree of operational efficiency. Since most cloud services are charged by the hour, this efficiency means that fewer infrastructure resources are consumed, which translates into lower operational costs and ultimately lower total cost of ownership.

In addition, as a commercial-off-the-shelf (COTS) solution, DataWalk requires no custom coding whatsoever, eliminating the overhead otherwise required to maintain additional code and correct it for errors. DataWalk is immediately ready to use, fully operational, and highly reliable. This plug-and-play functionality further lowers the operational costs of its use and ownership.

For further specifics on the performance test that was conducted, contact your DataWalk representative.