AEROSPIKE

Aerospike Connect for Presto/Trino

SOLUTION BRIEF

Unlimited scale sql operations with aerospike and presto

Overview

Sound business decisions lean heavily on insights generated from data such as device metrics, user behavior tracking, business transactions, location data, and much more. A Business Intelligence (BI) tool such as Tableau, MicroStrategy, Looker, or Qlik, etc. is a staple in a business or data analyst's toolbox and SQL is their language of choice for data mining. However, they care not only about premier dashboarding and analytics capabilities, but also the speed of analysis, which is usually measured as time-to-insight.

Dashboarding and interactive queries demand near real-time response times to be impactful. Such workloads are typically read-heavy, low latency, and high throughput. Large datasets such as a table with a million records or more are typically stored in a database and queried on a needs basis. Under the covers, BI tools generate SQL queries and send it via a JDBC interface to the SQL query engine, which is backed by a database. No matter how optimized your SQL query engine is, most of the performance gains are wiped out by slow reads from the underlying database. Slower queries can lead to significant losses in productivity, which can become very pronounced for large enterprises where millions of queries are run every day. This can severely impact your decision-making ability, not to mention revenues.

Highlights

- **New: Now supports Aerospike Database 6 massively parallel secondary indexes for real-time SQL**
- **Run ANSI SQL queries securely on massive amounts of data stored in Aerospike.**
- **Federate queries across multiple Aerospike clusters or between Aerospike and other databases**
- **Analyze Aerospike data via Business Intelligence (BI) tools e.g., Tableau, Qlik, Looker, etc.**
- **Leverage massive parallelism of Aerospike for speeding up Presto queries.**

Aerospike Connect for Presto

Aerospike Connect for Presto is a connector that enables business and data analysts to use ANSI SQL to query data stored in Aerospike via Presto - a highly parallel and distributed SQL query engine. Furthermore, it is multi-tenant and capable of concurrently running hundreds of memory, I/O, and CPU-intensive queries and can scale to thousands of workers. Figure 1 below depicts a high-level architecture of a typical deployment.

With its multi-site clustering capability, the Aerospike data platform enables you to extend the below architecture to provide a globally consistent view of data to a geographically distributed team of business and data analysts.

Benefits

- **Easy Integration with BI Tools**
  Analyze data stored in Aerospike via your favorite Business Intelligence tools such as Tableau, Qlik, Looker, and more.
- **Build AI/ML models**
  Move quickly from proof-of-concept to production of AI/ML models by pre-processing data stored in Aerospike.
- **Accelerate time to Insight**
  Combine the massive parallelism of Aerospike with ad-hoc Presto queries to dramatically reduce time-to-insight.
- **Lower TCO**
  Analyze massive datasets with a small storage cluster footprint.

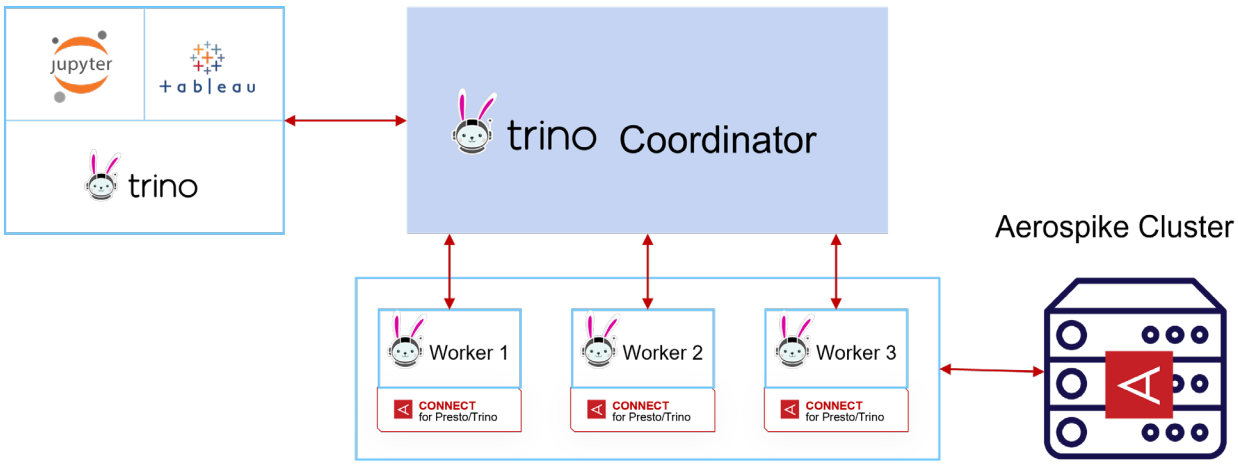

Figure 1: High-level architecture

Aerospike Connect for Presto offers the following core capabilities:

Run ANSI SQL queries on massive amounts of data in-place

- Query data stored in Aerospike without the need for complex and error-prone processes of copying data over to other analytics systems, which significantly helps with governance and compliance.
- Presto data types, including complex types such as maps and arrays are supported, along with both read and write statements.
- Presto Aggregate functions such as min/max, sum, avg, etc. are supported.

Federate queries across multiple Aerospike clusters or between Aerospike and other databases

- Enterprise database deployments are typically polyglot in nature, but you can now deploy Aerospike into an ecosystem of DBs that consist of Cassandra, PostgreSQL, Oracle, etc.

Analyze Aerospike data via Business Intelligence (BI) tools

- Create insightful dashboards using Tableau, Qlik, Looker, etc., by accessing Presto over JDBC to analyze data stored in Aerospike.
- Secure your deployment with TLS and LDAP support
- Use TLS to secure connection between Presto and the Aerospike clusters and LDAP for authenticating Presto users with the Aerospike database.

Query records with different schemas within the same set in Aerospike

- Aerospike is a NoSQL schemaless database, but Connect for Presto reconciles those differences and offers a SQL experience that you are familiar with, while leveraging the benefits of a NoSQL database.
- It supports schema inference so that you have the option to not define the table definition in the Presto Catalog.

Leverage massively parallel secondary indexes of Aerospike for speeding up queries

- It can scan 4,096 partitions in parallel to load data into and up to 32K Presto splits across your Presto cluster and uses the recently released Aerospike expressions for pushdowns to the database.

Leverage Presto Cost Based Optimization (CBO) via row count for query optimization

- Aerospike Connect for Presto is one of the few Presto connectors that supporting CBO for speeding up Joins in Presto.

Deploy anywhere

- Deploy in cloud or Kubernetes environments to help you leverage Managed Presto Services offered by multiple Cloud providers.

Aerospike Connect for Presto powers the following use cases:

Interactive Queries

Business analysts run 100's of interactive queries on a daily basis with millisecond-to-seconds query latency to generate insights for business-critical decisions - such as "count the number of users that have clicked the new banner ad" or "what are some categories of ads our users have seen?". These queries are characterized by low response times and usually involve smaller data sets. You can either use Presto CLI or any SQL editor that support Presto JDBC drivers such as Hue, Zeppelin, Quix, etc. Connect for Presto is designed to not cause any performance degradation for the database at query time.

BI Dashboarding

Data analysts can use a BI tool such as Tableau, Looker, Qlik, etc., to connect to Presto over JDBC (using the Presto JDBC driver for the BI tool) and analyze data stored in Aerospike at scale for visualization and reporting. They can create real-time dashboards and deliver high-quality data insights quicker than before.

Data Preparation for AI/ML

Python is extensively used by data scientists looking to prep, process, and analyze data for analytics and machine learning use cases. Jupyter Notebook is an open-source, interactive, and web-based notebook used for data analysis and visualization. You can now analyze data stored in the Aerospike Database via Presto using the Jupyter Notebook with the PyHive Presto Python library. The advantage of using a Jupyter Presto notebook is that you can not only explore data stored in Aerospike, but also pre-process it to create AI/ML models using popular Python libraries such as Pandas, NumPy, Scikit-learn, etc., and quickly progress from proof-of-concept to production.

AEROSPIKE

The Aerospike Real-time Data Platform enables organizations to act instantly across billions of transactions while reducing server footprint by up to 80 percent. The Aerospike multi-cloud platform powers real-time applications with predictable sub-millisecond performance from terabytes to petabytes of data with five nines uptime with globally distributed, strongly consistent data. Applications built on the Aerospike Real-time Data Platform fight fraud, provide recommendations that dramatically increase shopping cart size, enable global digital payments, and deliver hyper-personalized user experiences to tens of millions of customers. Customers such as Airtel, Experian, Nielsen, PayPal, Snap, Wayfair and Yahoo rely on Aerospike as their data foundation for the future. Headquartered in Mountain View, California, the company also has offices in London, Bangalore and Tel Aviv.

2525 E Charleston Road, Mountain View, CA, 94043 | (408) 462-2376 | aerospike.com