Solution Brief:
Aerospike Connect for Spark

Powering Advanced SQL Analytics Using Aerospike and Spark

Highlights

- New: Now supports Aerospike Database 6 massively parallel secondary indexes for real-time SQL analytics
- Reduce time to insight by leveraging massive parallelism of Spark and Aerospike
- Rapidly develop applications using Spark supported language of your choice with a rich ecosystem of open-source libraries
- Access Aerospike through SQL with Zeppelin/Jupyter notebooks
- Support low latency streaming use cases
- Lower TCO by combining the cost efficiencies of Spark and Aerospike

Overview

Advanced analytics includes predictive data analytics across massive amounts of multi-modal data, usually ingested from disparate sources. This encompasses analytics from the edge to the core, streaming or batch, SQL analytics, and AI/ML training and inference, for example. Use cases span industries, from payments systems detecting fraud, advertising technology companies conducting real-time bidding, ecommerce sites creating customer 360 insights, or manufacturers conducting predictive maintenance with IoT.

The commonality for these use cases all involves processing data in-situ at the edge with the ability to both read and right to the database in real-time at extremely high rates. They all take data from the edge, replicate it to the core, and serve to their AI/ML models for training and then in production inference pipelines. These use cases require a solution that would help them achieve the desired scale and latency without blowing up their infrastructure budget. Traditionally, they dealt with this challenge by using an assortment of disparate open source and/or proprietary products with high complexity and costs. These solutions included Hadoop, Kafka, Spark, and Cassandra, to name a few.

Aerospike Connect for Spark can help enterprises answer these challenges by harnessing best of breed technologies: i.e. Aerospike for storage and Spark for computation. Aerospike with its strong consistency guarantees, higher IOPs and performance with hybrid memory architecture, and lower TCO, could be used as a system of record database, edge database, and query and reporting database. Spark then could be used for massively parallel and in-memory computations that support AI/ML and other complex analytics workloads.

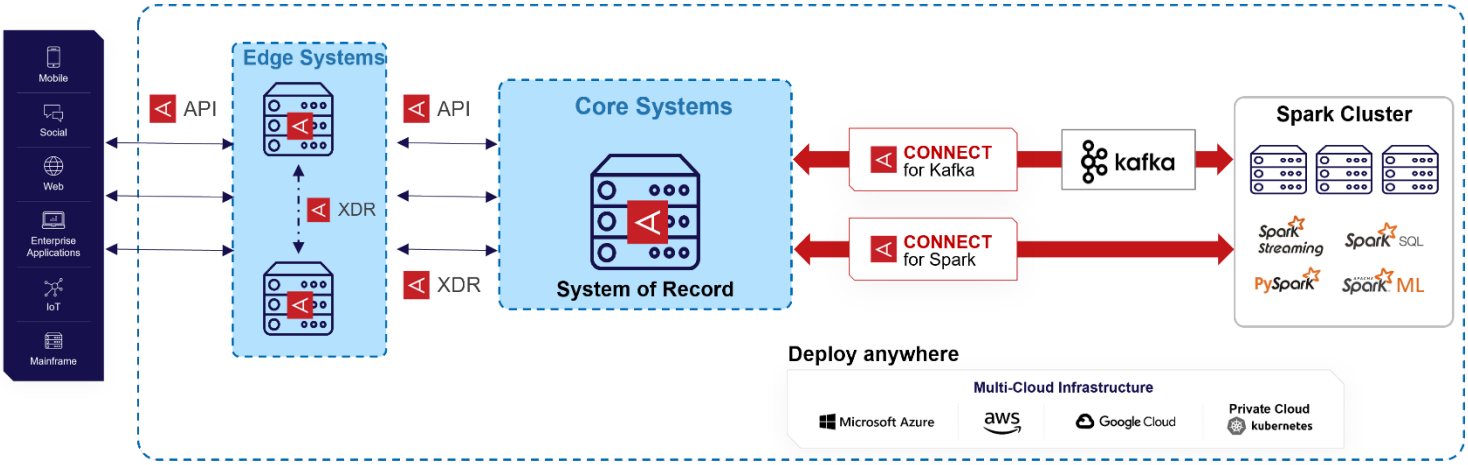

Figure 1. Aerospike Connect for Spark

Solution Brief:

Aerospike Connect for Spark

Aerospike Connect for Spark offers the following core capabilities to support your use cases:

- Loads Aerospike data into both Spark DataFrame and DataSets to enable further complex processing in Spark such as ETL and AI/ML using SparkML and other open source libraries and frameworks that support PySpark
- Leverages Aerospike 6 massively parallel secondary indexes for greater performance and scale.
- Leverages Structured Spark Streaming to support streaming reads (change notifications) from and writes to Aerospike
- Supports multiple languages (Python, Java, Scala, etc.)
- Supports massive parallelism by allowing you to use up to 32,768 Spark partitions to read data from an Aerospike namespace. Each namespace can store up to 32 billion records across 4,096 partitions
- Leverages Spark SQL (ANSI SQL 2003 standard) to allow SQL access to Aerospike

Benefits

- Lower TCO by enabling analysis of massive, larger datasets with a significantly smaller storage cluster footprint. Additionally, running Spark and Aerospike in separate clusters provides the ability to right-size them independently based on capacity needs. Further, Aerospike Connect for Spark can leverage existing investments in Spark.
- Drastically reduce time to insight by combining massively parallel computations in Spark with the massively parallel reads from a distributed database such as Aerospike. This capability can be extended to reduce training and inference times for AI/ML applications as well.
- Save time developing analytics and AI/ML applications that use data in Aerospike by using a Spark supported language of your choice and the rich ecosystem of libraries that are already available with Spark.
- Access Aerospike using the all-too-familiar SQL for batch queries and exploratory data analysis (EDA) using notebooks such as Zeppelin or Jupyter.
- Analyze data written to Aerospike at the edge in near-real time using the Spark streaming read API and rapidly write the results back to Aerospike using the Spark streaming write API for future processing.

Typical Use Cases for Aerospike Connect for Spark

Financial Services and FinTech: Bridge legacy systems such as Mainframes to modern services such as Fraud detection, risk management, etc. using Kafka and Aerospike, to enable new banking financial services applications

e-Commerce and Retail / CPG: Behavior data integration, clickstream integration with product data

Telco: Stream data form the Aerospike system that is used for high velocity and high volume ingest at the edge system in the IoT architecture to the back-end core systems for persistent storage, training AI/ML models, and analysis

Industrial Internet: Edge & device data synchronization with back end systems etc.

Data Lake / EDW Integration for real-time Analytics: Integrating operational data with datalakes, data warehouses for analytics

AdTech: Real-time clickstream data synchronization and integration

About Aerospike

Aerospike is the global leader in next-generation, real-time NoSQL data solutions for any scale. Aerospike enterprises overcome seemingly impossible data bottlenecks to compete and win with a fraction of the infrastructure complexity and cost of legacy NoSQL databases. Aerospike's patented Hybrid Memory Architecture™ delivers an unbreakable competitive advantage by unlocking the full potential of modern hardware, delivering previously unimaginable value from vast amounts of data at the edge, to the core and in the cloud. Aerospike empowers customers to instantly fight fraud; dramatically increase shopping cart size; deploy global digital payment networks; and deliver instant, one-to-one personalization for millions of customers. Aerospike customers include Airtel, Banca d'Italia, Nielsen, PayPal, Snap, Verizon Media and Wayfair. The company is headquartered in Mountain View, Calif., with additional locations in London; Bengaluru, India; and Tel Aviv, Israel.

2020 Aerospike, Inc. All rights reserved. Aerospike and the Aerospike logo are trademarks or registered trademarks of Aerospike. All other names and trademarks are for identification purposes and are the property of their respective owners.