AEROSPIKE

# Aerospike and Apache Spark Power a Real-time Online Profile Store at a Global Ad Tech Company

USE CASES

## Challenges with the original implementation

A Global Ad Tech company was building an Online Profile Store to hold customer data which consisted of current, historical and derived consumer data combined with behavioral attributes. This data can be used to enhance effective consumer engagement utilizing ML, Business Intelligence and other technologies.

The original implementation of the online profile store had unique challenges the Ad Tech company had to overcome. It was developed internally utilizing custom code which resulted in limitations in the data integration component. This ultimately led to data quality issues. The data infrastructure was so complex that it was incredibly inefficient in updating the profile stores. When implementing Spark jobs, they were taking around 12 hours to complete which was limiting the number of updates that could be made to the profile stores. The AdTech company had to consider their data growth roadmap and it was obvious that the original data infrastructure would not be able to handle the predicted workloads.

### Benefits with Aerospike

- 10x Reduction in insight times due to the massive parallelization of data processing
- Operational reliability at large scale with Aerospike and Spark clusters in the cloud to process 13B objects and 150 TB of unique data
- Enabling technology for defining new product offerings and business opportunities
- Efficient utilization of hardware resources by Aerospike to provide cost savings

Digital marketing companies face similar challenges to Ad Tech companies as the struggle to achieve high scale with low latency is just as mission-critical to their business. A digital marketing company began to develop an innovative system for creating customer 360 profiles of its customers for targeting purposes. The company receives millions of customer hits on their digital assets every second. Each hit triggers millions of queries, across disparate datasets such as digital identities, segmentation, device types and brings to bear complex AI models. The challenge was to achieve it at scale without making any tradeoffs around latency for reads and processing.

## Goals of the new implementation

There were a number of architectural alternatives that were evaluated to better support the deployment and maintenance of the online profile store. A part of the new implementation was recognizing the needs and limitations that needed to be innovated to address the new requirements were:

- Faster update times for the profile store by moving large datasets between Aerospike and Spark
- Scalable platform to handle the predicted data growth
- Data platform consolidation and reduction of infrastructure spend

The new implementation also had new operational requirements:

- Ease of deployment and management
- Operational reliability in large clustered environments
- Significant reduction in the operational footprint

## Why Aerospike connect for Spark

Aerospike Connect for Spark outpaced the other solutions that were being evaluated because it provided:

- Lower TCO by combining the cost efficiencies of Aerospike and Spark
- Reduced time to insight by leveraging massive parallelism of both Aerospike and Spark
- Developing applications rapidly while using Spark supported language of your choice with a rich ecosystem of open source libraries
- Access Aerospike through SQL with Zeppelin/Juypter notebook experience
- Support low latency streaming use cases

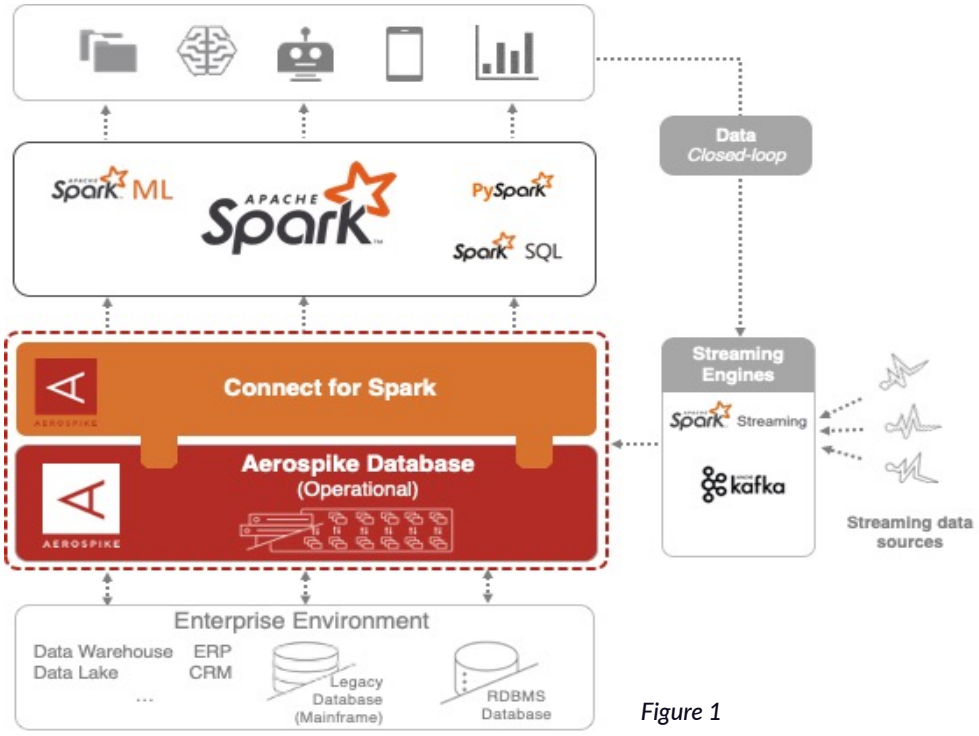


*Figure 1*

Utilizing the data platform built on Aerospike and Spark allows faster updates of the profile store. Aerospike Connect for Spark (Figure1) provides parallelization of data transfer between the Aerospike Database and Spark which ultimately enables faster analysis of the data in Spark.

Aerospike Connect for Spark provides the following capabilities:

- Loads Aerospike data into the Spark DataFrame to enable processing such as ETL and ML utilizing SparkML, and other open source libraries and frameworks and also supports low latency DataFrame writes back to the Aerospike DB.
- Leverages Spark Structured Streaming to support streaming writes to Aerospike DB
- Supports multiple languages (Python, Java and Scala)
- Supports massive parallelism by allowing you to use up to 32,768 Spark partitions to read data from an Aerospike namespace. Each namespace could store up to 549,755,813,888 records per node
- Leverages Spark SQL (ANSI SQL 2003 standard) to allow SQL access to Aerospike DB

# Data platform utilizing Aerospike and Spark

With the aforementioned goals for the project in mind, the Aerospike team worked closely with the customer to design a solution that combined Aerospike's speed at scale and the massively parallel processing capability of Spark.

It was deployed in the cloud to process 13 billion objects, that included 150 TB of unique data. There were 33 nodes in the Aerospike Cluster and approximately 300 nodes in the Spark Cluster. This new solution reduced the time to run the job to update the profile store from 12 hours to 2.4 hours, which freed up resources to run other jobs in the same cluster, thereby achieving operational efficiency. Further, this solution was rolled out into production in months, not to mention the huge cost savings.

As shown in Figure 2, the data in the online profile store is updated and utilized the following ways:

1. Online operations and customer interactions update the profile store continuously on the edge system
2. Data from the Blob store (customer lists, lead lists, etc.) is enriched with online data coming from the profile store using the spark cluster
3. The data transfer is extremely fast which is achieved by mapping 4,096 Aerospike partitions with up to 32,768 Spark partitions per namespace. This massive parallelization was translated to 8192 Spark tasks on the Spark side, handled by 7500 cores allowing the rest of the process to be parallelized even further.
4. The output data from Spark is pushed back into a Data Lake (Blob store) for Correlations, Business Intelligence, Machine Learning, Dashboarding and Statistics. Alternatively, it could be written to Aerospike DB via fast Aerospike Connect for Spark streaming API.
5. Data quality has improved significantly due to the new infrastructure in place
6. The analyzed data can provide insight on what new products/services can be offered to customers
7. Data Lakes generated by the process above are also available for customers

Besides meeting and exceeding all functional and operational requirements, this architecture also allows significant data growth in the future while maintaining a low TCO.

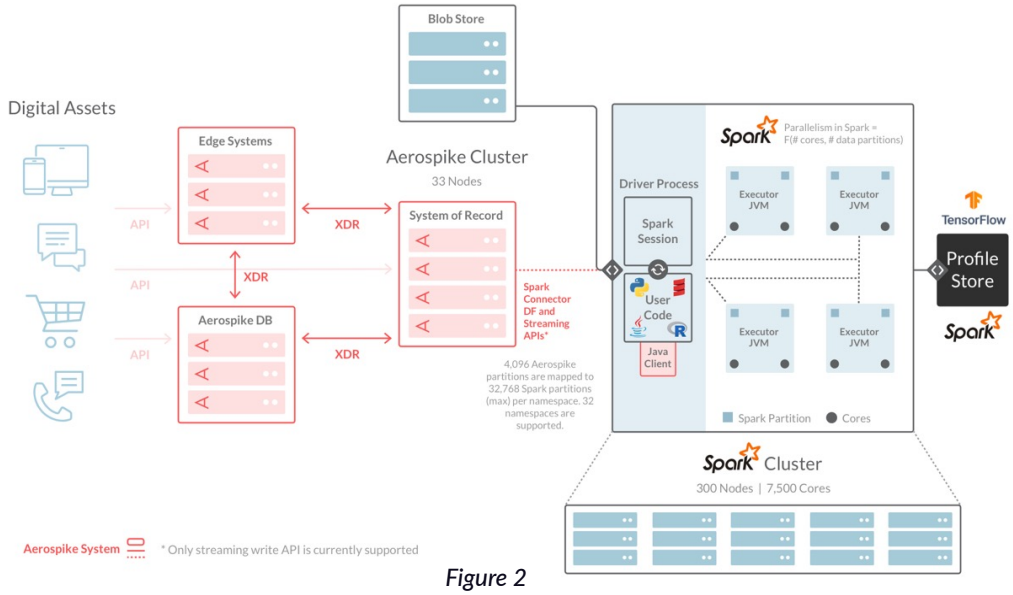


*Figure 2*

The Aerospike Real-time Data Platform enables organizations to act instantly across billions of transactions with predictable performance from terabytes to petabytes of data, while reducing server footprint by up to 80 percent.