



June 15, 2021

## DataStax Taps Pulsar for Streaming Data Platform

[Alex Woodie](#)



DataStax today unveiled Astra Streaming, a new event streaming platform based on Apache Pulsar, a publish and subscribe (pub-sub) data platform that competes with Apache Kafka. Astra Streaming is pre-integrated in the cloud with Astra DB, DataStax's serverless, multi-cloud version of the Apache Cassandra NoSQL database.

Dealing with the demands of historical data on the one hand, and real-time streams of event data on the other, is one of the key hurdles that big data architects seek to overcome. The problem is that streaming data platforms are not ideal places for holding onto stateful data that rarely changes, while the traditional databases that are designed to hold stateful data treat event data as second-rate elements. Bringing the two systems and data types together is necessary to serve emerging use data-driven experiences for users, including those involving AI and IoT technologies, but it's not simple or easy.

Now [DataStax](#) is throwing its hat into the streaming data ring with Astra Streaming, a hosted version of Pulsar that is pre-built to closely integrate with Astra DB, its flagship Cassandra database-as-a-service offering.

According to Chris Latimer, DataStax's vice president of product for streaming, Astra Streaming gives customers the same types of scale and performance that they have come to expect from Cassandra, which is widely considered to be one of the most scalable databases around. By engineering them to work together, DataStax says its customers can enjoy the data benefits of both systems while cutting costs.

"DataStax has built an open source Cassandra sink connector that is used within Astra Streaming allowing users to stream data into their Astra DB instances," Latimer tells *Datanami*. "At the same time, we're building full bidirectional capabilities to also allow developers to stream changes happening on their Astra DB databases into Astra Streaming as an event stream."

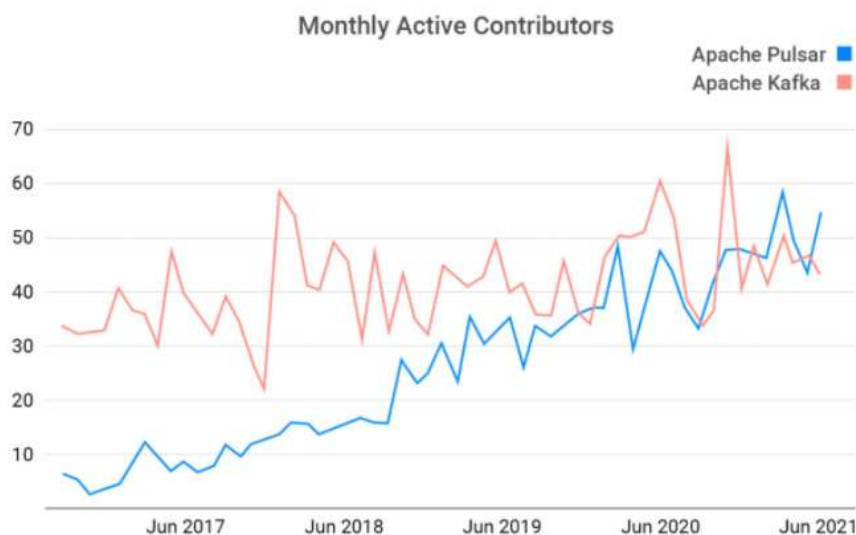


Access to a SQL-like query construct will make Astra Streaming more familiar to users accustomed to SQL, says Ed Anuff, DataStax's chief product officer.

"A key aspect of stream processing is the ability to interact with event logs in a way that feels familiar to anyone who is experienced with database technologies," Anuff says in a press release. "While existing solutions work up to a point, they generally can't compete with the scale, performance, and reliability that comes from Apache Cassandra...With Astra Streaming, you can achieve the same usability benefits associated with SQL-like stream processing interfaces with the cross-cloud, high-performance persistence capabilities of Astra DB."

Apache Pulsar was originally developed at Yahoo as a distributed messaging platform to provide data for services such as Yahoo Finance, Yahoo Mail, and Flickr. Yahoo released Pulsar as open source in 2016, Apache Software Foundation adopted it as [a top-level project](#) in 2018, and since then it has been adopted by a variety of companies for production use cases, including Tencent, Comcast, Appen, and Overstock.

Kafka is the dominant pub-sub messaging system, and is used by 70% of the Fortune 500, according to [Confluent](#), the commercial outfit behind Kafka. Confluent has soared to a \$4.5 billion valuation and [a pending IPO](#) on the back of its dominant position in streaming data. But that hasn't stopped Kafka's competitors from trying to eat into that lead.



Participation in the Pulsar project exceeded participation in the Kafka project, according to [StreamNative's 2021 Apache Pulsar User Report](#)

DataStax says that Pulsar compares favorably to Kafka in several respects, including support for messaging semantics of MQ-based solutions, Latimer says. "Pulsar also includes capabilities that Apache Kafka is missing out of the box like geo-replication and multi-tenancy and the workarounds are costly and resource intensive," he says.

[GigaOm](#) research shows that Apache Pulsar has advantages over Kafka in terms of both price and performance, according to GigaOm analyst William McKnight. "We see Pulsar becoming an increasingly popular choice for streaming applications," McKnight stated in a press release.

In addition to DataStax, there are a handful of other companies providing commercial Pulsar solutions. That includes [StreamNative](#), which launched its Pulsar-as-a-service offering in the fall of 2020 and employs committers to the Apache Pulsar project, as well as [Pandio](#), which offers a hosted version of Pulsar. [Splunk](#) acquired Pulsar service provider Streamlio in 2019, and in 2020 [adopted Pulsar](#) as the core underlying technology for its Splunk Data Stream Processor (DSP).

Meanwhile, the open source Apache Pulsar project moves on. Pulsar reached a couple of milestones this week, including accepting its 400<sup>th</sup> contributor, which exceeds the number of contributors for Kafka. It also adopted exactly once semantics with transactions in Pulsar version 2.8, according to [the StreamNative blog](#).

DataStax is now accepting requests to participate in the beta for the Astra Streaming service.