In-memory database firm MemSQL helps users create Kafka data pipelines

JASON STAMPER
17 OCT 2016

With the release of MemSQL 5.5, the company has introduced a new SQL syntax that enables developers to create data pipelines right from their command-line interface.
In-memory database player MemSQL claims that its database is so fast that it can be used to handle not only transactional workloads but also analytics. With the latest version, the company says it has also made it easier for developers to set up their data pipelines – helping them stream data into MemSQL in real time from Apache Kafka.

**THE 451 TAKE**

It’s a good sign that the $36m in funding that MemSQL landed in April is anticipated to take it all the way to profitability. With just over 100 paying customers, including household names such as Comcast, EMC and Pinterest, MemSQL has built a solid business on the back of its fast in-memory database, which can handle both transactions and analytics. The next stage in its evolution is undoubtedly to make it easier for developers to grapple with the growing number and kind of data types that they want to move in and out of the MemSQL database. Therefore, the company’s Pipeline technology in version 5.5 is a welcome addition, helping developers build pipelines around the popular Kafka message broker. The launch of a database-as-a-service (DBaaS) version, possibly as early as this year, should give it another shot in the arm, particularly – at least initially – for smaller deployments and those looking to do development and testing.

**CONTEXT**

An in-memory database is designed to accelerate transactional (OLTP) workloads, analytic (OLAP) workloads, or both. Since memory is an order of magnitude faster than disk, running the database in-memory offers a huge performance benefit. The very largest datasets won’t fit entirely in-memory – although its price has fallen, in-memory is still more expensive than disk. For that reason, and also to add additional high availability since memory is volatile, in-memory databases write back to disk sporadically.

MemSQL was founded in 2011 in San Francisco, and came out of stealth a year later. To date, the company has raised a total of $85m, the most recent being a $36m series C announced in April. It has noted that it believes that should give it enough money to take it to profitability. The latest round included Accel Partners, Caffeinated Capital, Data Collective, First Round Capital, IA Ventures, Khosla Ventures and REV Capital. At the time of its latest funding round, CEO and cofounder Eric Frenkel told us that MemSQL didn’t need to raise any more cash, but saw it as a good opportunity to accelerate the firm’s growth.

The company is planning to open an office in London as the first step in its international expansion, and other locations are likely to follow. It says it already has customers on every continent. MemSQL reports that it has just over 100 paying customers, which is up from the 50 it counted when we caught up with the company in January. It says its headcount has also increased, from about 70 in January to just over 100 now.

**TECHNOLOGY**

The key to at least some of MemSQL’s appeal as a database is that it is multi-modal. That is, it can handle transactional duties – it supports ACID compliance, which is needed for transactional integrity – as well as analytic workloads, including streaming analytics. It achieves this in several ways. The company keeps all data consistent in a row-based format in a relational database in-memory. By converting SQL statements into native C++ instructions, it achieves a rapid pace for ad hoc queries. But it also persists to a disk-based column store that sits on traditional or SSD disks for rapid, historical data analytics.

MemSQL can be deployed on-premises or on Amazon Web Services, Microsoft Azure or Google clouds. The company says about one-third of its customers are running MemSQL in the cloud, mostly on AWS or Azure. The clients manage the database in the cloud as if it were on their own premises – MemSQL doesn’t offer DBaaS, although we understand this is something it is working on that might even be launched before year-end.
With the latest version of its in-memory database, 5.5, MemSQL says it has made a significant number of enhancements and improvements. Perhaps the most significant is the ability for developers to quickly and simply create data pipelines – essentially mapping the way that data is ingested into MemSQL in the first place – for data that is streaming in from the popular Apache Kafka open source message broker. Kafka is a distributed streaming platform that enables users to publish and subscribe to streams of data like a messaging system.

There are several reasons for wanting to move data from Kafka (and the sources of data that feed into it) and MemSQL. Using MemSQL's transactional capabilities, organizations could build applications that make use of data streaming from Kafka. The firm's in-memory streaming analytics technology could also be deployed to run analytics on data streaming from Kafka and other sources, or to compare streaming data with more historical data stored in MemSQL.

Furthermore, thanks to integration between MemSQL and the open source Apache Spark technology, which offers big-data processing with built-in modules for streaming, machine learning and graph processing, data can be moved from various sources via Kafka to Spark and then MemSQL. Analytics can be run where they most make sense – for example, real-time analytics might make use of MemSQL's in-memory pace, while deeper machine learning could be pushed back to Spark via the MemSQL Spark Connector.

Previously, this would have been a relatively time-consuming and manual process. But now, with MemSQL 5.5's simple ‘CREATE PIPELINE’ SQL syntax, users can construct real-time data-streaming pipelines via the command line. In turn, it can help give business users that want to analyze data coming from Kafka (and the other sources that may be feeding into the Kafka pub-sub message broker) almost instant visibility into their data.

Besides being able to create the pipeline from Kafka, the SQL syntax has several other features that enable developers to monitor and manage their pipelines once they have been created. Other functions include start/stop the pipeline, test the pipeline, alter the pipeline, drop the pipeline, show the pipeline, show pipeline errors, and so forth. These are all SQL functions used in a command-line interface, and hence only suitable for use by developers with SQL skills. The movement of data from Kafka into MemSQL offers what's known as exactly-once semantics, which means that data is moved once and only once, giving enterprises guarantees that all data required is moved, with guaranteed delivery and no duplication.

Other ways of loading data into MemSQL (including the ‘LOAD DATA’ syntax, which loads data from a .CSV file, or the company's Streamliner technology that loads data in parallel from multiple nodes and which is built with Apache Spark) do not offer the same level of pipeline flexibility, or the exactly-once guarantees of MemSQL's Pipeline. It's worth mentioning that with the first release of the Pipeline technology in MemSQL 5.5, the only messaging layer supported is Kafka. Others that might prove useful for users include the likes of RabbitMQ, AciveMQ and ZeroMQ, as well as Amazon Kinesis or Azure Event Hubs in the cloud. MemSQL says it is likely to add support for additional brokers based on customer demand.

Aside from the data-pipelining tool for Kafka, there are several other enhancements in version 5.5, according to the company. These include improved query performance, new query-profiling capabilities, and higher concurrency for distributed joins, with results demonstrating five times faster query performance for star-schema data models. MemSQL says it has also improved its resource-aware scheduling, and boosted its compute and network resource utilization.

As for its future direction, the company notes that besides adding other data sources to the MemSQL Pipeline tool, it will add support for Spark 2.x to its Streamliner and MemSQL Spark Connector technologies soon. It plans to demo the technology at the Spark Summit in Boston from February 7-9 next year.

CUSTOMERS
Complementing its 100 or so paying customers, recent client wins include EMC, Verizon and Pandora. EMC employs MemSQL's real-time analytics for a customer acquisition lifecycle management application; Verizon for real-time web-based video and application analytics; and Pandora to analyze end-user music-listening habits for advertising personalization. Other reference clients include Pinterest, Comcast, Akamai, Shutterstock, Ziff Davis, Tapjoy and CPXi.
**COMPETITION**

Other companies claiming to be multi-modal (supporting both transactional and analytics workloads) include VoltDB, which like MemSQL, offers a free community edition and supports ingestion from Kafka data sources. Additional competition comes from NuoDB, Altibase and Pivotal (with GemFire and Apache Geode), as well as GridGain Systems, Hazelcast, Software AG and ScaleOut Software.

Among the larger vendors, Oracle, IBM and Microsoft have deployed in-memory capabilities to enable their operational databases to handle analytics. SAP has a pure in-memory database – HANA – that can also cope with both transactions and analytics.

Clustrix offers ClustrixDB, which is a NewSQL database that achieves ACID compliance, but can also handle SQL queries. Startup LeanXcale claims that it can handle transactional and analytical workloads as well. Some of the NoSQL database providers, including Redis Labs and MarkLogic, can handle transactional consistency along with analytics on their scale-out architectures.

**SWOT ANALYSIS**

**STRENGTHS**

MemSQL is a well-respected player in the in-memory database field, and with the release of v.5.5 it has added a Pipeline tool that should make it even easier to ingest data from Kafka and other sources.

**WEAKNESSES**

The company’s Pipeline tool is currently for Kafka only, so those looking to build pipelines for the likes of RabbitMQ and ZeroMQ have to continue to build them themselves.

**OPPORTUNITIES**

There is growing interest in being able to run transactional and analytic workloads in the same place to reduce complexity and latency – something we expect to continue to grow in the era of the Internet of Things.

**THREATS**

Competition in the database space is as stiff as ever, and MemSQL and its ilk must vie not only with the relational database incumbents but also with in-memory rivals, NoSQL, NewSQL and even in-memory grid firms.