Citus Data looks to a PostgreSQL future, seeks second round of funding

Analyst: Jason Stamper

1 Jul, 2015

Citus Data has raised $5m in venture capital to date, but says it expects to raise another round in the next couple of quarters. Meanwhile, it has just announced new versions of the proprietary and open source technology it has developed, and won a big new customer in MixRank, which has standardized on CitusDB.

The 451 Take

The Citus Data technical team has proven experience in distributed data management and distributed systems design, which should help as Citus looks to raise another funding round. The company's focus on extending PostgreSQL, rather than forking it, gives it a solid installed base to target with its CitusDB product. By making the pg_shard and cstore_fdw projects open source, Citus expanded its interaction with the PostgreSQL user community, and we think it's laying the foundation for increased CitusDB adoption. The latest version of CitusDB, 4.0, adds out-of-the-box integration with pg_shard, too. For companies looking to scale out PostgreSQL, CitusDB should make most short lists.

Context

Citus Data was founded in 2011 and has so far raised $5m in venture capital from major investors Data Collective, SV Angel and Bullpen Capital, as well as some Silicon Valley angels. Management recently told us that the company is hoping to close another investment round in the next couple of quarters. Citus continues to be led by its founders, Umur Cubukcu (former director of business development at supply-chain vendor TrueDemand Software), Ozgun Erdogan and Sumedh Pathak (both previously worked at Amazon). In July 2014, PostgreSQL core contributor Josh Berkus joined

Citus Data's executive board, which is certainly noteworthy. Citus Data has about 20 staff – up five from March – with employees located in the US, Turkey, the Netherlands and Germany. It also says it has about 20 paying customers.

Citus Data's business model includes both proprietary and open source projects. Its CitusDB database is proprietary – it is built on PostgreSQL, which enables the open source database underpinnings to be clustered across commodity hardware; this isn't something that's possible with a standard PostgreSQL instance. CitusDB also adds high-availability and resilience capabilities while continuing to support standard SQL queries. However, Citus Data also supports two Apache open source projects that anyone can download and use with PostgreSQL for free: cstore_fdw and pg_shard.

The cstore_fdw and pg_shard projects add columnar tables and horizontal scalability to the open source PostgreSQL database. In addition to horizontal scaling, pg_shard supports real-time inserts and updates across multiple nodes, enabling PostgreSQL to be used for real-time analytics as well as operational workloads; this is what Citus Data means by a 'hybrid' database. Of course, it hopes that companies putting the technologies into production will want the CitusDB database – which is where it makes its money – as well as support for the open source projects.

Technology

CitusDB has recently come out with a new version of its CitusDB database – aimed at mixed transactional and analytic workloads – as well as new versions of the cstore_fdw and pg_shard Apache technologies it backs. Most important for the company is the CitusDB update, which is now on version 4.0 and based on PostgreSQL 9.4. It is worth noting at this point that Citus insists that it is not forking PostgreSQL, merely maintaining compatibility with it. As each new version of PostgreSQL comes out, Citus promises to maintain compatibility. Another change with CitusDB 4.0 is out-of-the-box compatibility with pg_shard, reducing implementation and configuration times.

The recently announced cstore_fdw 1.2 includes new 'INSERT' and 'COPY' features and enhanced memory usage; pg_shard, which was launched at the same time, has improved performance and new features for shard repair and import from CSV files from the command line. Cstore_fdw, which creates a columnar store for data in PostgreSQL, helps compress data by four to six times, thereby speeding up analytics dramatically.

So what is the difference between licensing CitusDB and simply using those two open source technologies with postgreSQL? The company says CitusDB 4.0 combines cstore_fdw and pg_shard

with its own massively parallel multi-node, multi-core architecture; moreover, it adds built-in replication and high availability. Those aren't features that users get automatically when piecing together PostgreSQL, cstore_fdw and pg_shard on their own.

Citus offers phone or email support to companies that license CitusDB; it can also offer phone or email support for production deployments of cstore_fdw and pg_shard, although it concedes that these types of support agreements are relatively few and far between. While CitusDB is not open source, it is free up to six nodes, after which it is licensed per node with additional paid-for support services.

As for where the company started – offering SQL-based analytic access to data stored in Hadoop – Citus says that ship has largely sailed. It argues that companies want real-time analytic access to data combined with it being a single transactional database, and it says HDFS just isn't up to the job. Citus says the reason for pushing this 'hybrid' approach by combining analytics and transactional data is all about simplicity and cost reduction. If companies really want to, CitusDB can still integrate with Hadoop, but Citus sees little demand. The company told us it has no plans to offer support for any databases other than PostgreSQL (along with Hadoop integration); it says PostgreSQL is the only one that can handle the sharding and scale-out needs of CitusDB.

Customers

CitusDB has roughly 20 paying customers, which is up from just 10 three months ago. One reference customer is Agari Data, which uses CitusDB for analyzing more than 2.5 billion emails a day for threat identification. Agari develops data-driven security offerings that power real-time cyber-threat detection and prevention for global companies and their customers. Another is CloudFlare, which offers a real-time analytics portal for its two million+ websites running on a 40-50TB database, using CitusDB at the back end.

MixRank is a newly signed CitusDB customer. It has standardized on CitusDB to power its Customer Discovery Platform, which helps B2B inside sales teams find new customers on a real-time basis. It applies the technology across a 160TB database, and Citus says the firm achieved 50-100 times cost savings compared with rival approaches.

Competition

Another company that is attempting to scale open source databases is Percona, which may be competitive in some situations given its recent acquisition of Tokutek. Tokutek's TokuDB is a high-performance storage engine for MySQL, MariaDB and Percona Server. Another competitor,

albeit not in the PostgreSQL space, is Deep Information Sciences, which says its storage engine drastically speeds up MySQL and Percona database performance, enabling them to become hybrid transactional and analytical engines (as Citus claims to do for PostgreSQL).

One of the larger vendors that says it can offer combined OLAP and OLTP in the same database is SAP with HANA. It's so confident of its capability to do analytics and transaction in the same database – and hence reduce complexity – that the latest release of its Business Suite enterprise applications platform only runs on HANA.

A smaller rival targeting both OLTP and OLAP workloads is JustOne Database, which is also based on PostgreSQL. Another possibility is the Postgres-XL project, which was launched by TransLattice in May 2014. If companies are primarily looking for an MPP analytic database, competitors include Pivotal Greenplum, HP Vertica, Teradata Aster and Amazon Web Services' Redshift.

SWOT Analysis

Strengths

The Citus Data technical team has proven experience in distributed data management and distributed systems design. Improved out-of-the-box integration with its open source technologies should be a boon for implementers.

Weaknesses

The company's technology only supports PostgreSQL, which it says is because rival open source databases such as MySQL wouldn't allow automatic sharding at the database layer.

Opportunities

We still think making the cstore_fdw and pg-shard projects open source was a good move that should help it to build up a list of potential customers for its commercial software and support services (assuming companies take the open source tools into production, which we think they increasingly will).

Threats

Pivotal's open source Greenplum strategy looks competitive, and Pivotal certainly has deeper pockets. We also note that there are even larger companies – notably SAP with HANA – talking up a 'hybrid' OLAP and OLTP approach, albeit without open source roots.