Citus Data prepares CitusDB 4.0, now a massively parallel PostgreSQL analytic database

Analyst: Matt Aslett

9 Mar, 2015

Citus Data has changed its positioning since our last update, evolving CitusDB from being a scalable analytics database predominantly designed to bring SQL analytics to Hadoop to offering a stand-alone massively parallel columnar analytics database that is PostgreSQL-compatible.

The 451 Take

We noted that Citus Data was entering a crowded market in 2013, and therefore see the change of direction as a good thing. While the MPP analytic-database market is no less crowded, Citus Data is differentiated by its focus on extending, rather than forking, PostgreSQL. Making the cstore_fdw and pg-shard projects open source should grow the company's profile in the PostgreSQL user community and lay the foundation for potential CitusDB adoption. The competitive situation is likely to heat up, given that Pivotal's open source Greenplum strategy appears to be dependent on making it the default MPP choice for PostgreSQL, but we agree with Citus Data that Greenplum, having forked from PostgreSQL several years ago, will be a challenge. Either way, the PostgreSQL community will decide.

Context

The first time we encountered Citus Data, almost two years ago, the company had just launched CitusDB 2.0, bringing real-time SQL analytics to the Apache Hadoop data-processing framework. The SQL-on-Hadoop party was already in full swing, and got very crowded very quickly. While Citus Data had planned to differentiate itself by bringing SQL-based analytics to other nonrelational data

platforms, including NoSQL databases, the company instead took a more radical change of direction, turning CitusDB into a stand-alone PostgreSQL-compatible massively parallel columnar analytics database.

CitusDB 2.0 was already based on PostgreSQL – colocating the open source database on each node in a distributed Hadoop cluster and taking advantage of PostgreSQL's foreign data wrapper technology to query data in HDFS via the local query execution – and Citus Data added its own intellectual property in terms of distributed and local query-planner and distributed and local query-execution capabilities. The company has now brought that distributed approach to PostgreSQL with a fully parallelized architecture, transparent sharding and the ability to create columnar tables alongside PostgreSQL's existing row-based architecture.

Several companies have used PostgreSQL as the basis for commercial analytic databases in the past (for example, Netezza, ParAccel, Greenplum and Aster Data were all based on PostgreSQL to some degree), but while those were all forks of the PostgreSQL codebase, Citus Data insists that its approach is better described as extending PostgreSQL, taking advantage of the extensible architecture that was introduced with PostgreSQL 9.1. The result is that CitusDB is fully compatible with other PostgreSQL extensions, such as PostGIS.

Additionally, Citus Data made the code behind its column store and transparent sharding capabilities available as open source projects in April and December 2014, respectively, so PostgreSQL users are able to use cstore_fdw and pg-shard projects to add columnar tables and horizontal scalability to the open source 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. What Citus Data will add with the forthcoming CitusDB 4.0 is the combination of cstore_fdw and pg-shard with its own massively parallel multi-node, multi-core architecture; built-in replication; and high availability. CitusDB 4.0 is based on PostgreSQL 9.4 and is due to be generally available at the end of March.

CitusDB itself is not open source, but is free up to six nodes, after which it is licensed per node with additional paid-for support services. The company has fewer than 10 paying customers at this stage, but reports good interest for use cases in areas such as network analytics, retail, logistics and mobile analytics, as well as ad-tech and security. The sweet spot is essentially a use case that involves a large volume of machine-generated data, but a human user looking for sub-second query response.

The company now has 15 employees, compared with fewer than 10 in April 2013, and 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 formerly at Amazon). Citus Data had planned to take a series A funding round in late 2013, but ended up extending its seed funding round from $1.6m to closer to $5m and adding Bullpen Capital to a list of seed investors that already included Data Collective, Trinity Ventures, SV Angel and Digital Garage. Another interesting addition to the company: in July 2014, PostgreSQL core contributor Josh Berkus joined Citus Data's executive board.

Competition

In its initial guise, the company was competing against Cloudera's Impala and Pivotal's HAWQ, but that space got very crowded very quickly, and every Hadoop and database vendor now seems to offer its own take on SQL-on-Hadoop. As such, Citus Data's change of direction seems like a wise move, although the analytic-database market is also crowded, and dominated by the likes of Oracle, IBM, Microsoft, Teradata and SAP.

Given its focus on PostgreSQL, Citus Data's closest competition can be expected to come from other approaches to adding massively parallel processing to PostgreSQL. The company itself mentions the Postgres-XL project, which was launched by TransLattice in May 2014 and is based on the technology it acquired along with StormDB in late 2013, although Citus Data maintains that it rarely actually encounters it in competitive bake-offs.

Additionally, Pivotal recently announced plans to open-source its Greenplum MPP database, which was originally based on PostgreSQL. As we have noted, Pivotal is hopeful that the PostgreSQL community will embrace Greenplum's massively parallel architecture as a next major milestone in the development of PostgreSQL itself. Citus Data is somewhat skeptical about its chances, noting that Greenplum forked off from PostgreSQL several years ago, and may prove difficult to reintegrate. Nevertheless, the company does expect to see Pivotal Greenplum in competitive situations along with other MPP databases, such as HP's Vertica, Teradata's Aster Database and Amazon Web Services' Redshift service.

SWOT Analysis

Strengths

Citus Data's founders have proven experience in distributed data management and cutting-edge distributed systems design.

Weaknesses

The company has moved from one crowded space to another, although its focus on extending PostgreSQL, rather than forking it, should give it an installed base to target.

Opportunities

Threats

Making the cstore_fdw and pg-shard projects open source should grow the company's profile in the PostgreSQL user community and lay the foundation for potential CitusDB adoption.

Pivotal's open source Greenplum strategy appears to be dependent on making it the default MPP choice for PostgreSQL, which could overshadow CitusDB.