Ovum
TMT intelligence | informa

DataStax Enterprise 6 extends Apache Cassandra

Ovum view

Summary

As one of the most popular NoSQL databases, Apache Cassandra has been known for its ability to handle extremely massive scales of structured and variably-structured data. DataStax, the company that has been known for driving much of the Casandra project contributions as well as providing commercial support to Apache Cassandra has in recent years adopted a new strategy to expand beyond the limitations of the Apache project with a more multi-faceted platform featuring unique content. DataStax Enterprise 6 (DSE) takes that differentiation to new levels with features that accelerate performance, simplify database management, improve Spark connectivity and resiliency, and enhance the embedded search and graph features. In so doing, DataStax is embracing the open core strategy that has become de facto standard for most open source providers, enabling it to meet the challenge of emerging multi-schema offerings that will be redefining the NoSQL database market.

DataStax is creating unique value-add beyond Apache Cassandra

Apache Cassandra is a “wide column” NoSQL database that combines the versatility of a table-oriented database with the speed and efficiency of a key/value data store. Cassandra has proven a good choice where the primary need is the ability to handle huge volumes of uploads, such as at Instagram, where it ingests over 80 million photos per day.

DataStax is the commercial company that was originally formed to deliver professional support to Apache Cassandra. While DataStax began adding its own content to extend the core open source Apache Cassandra database since the company’s initial 2011 commercial release, differentiation picked up steam in the wake of its 2015 acquisition of Aurelius, a specialized provider of a graph database (and the creators of Apache TinkerPop, the de facto standard for modern property graph databases), DataStax has been on a path to differentiate its platform and expand beyond pure Cassandra open source.

In so doing, DataStax is following what has become the common path among open source companies, a strategy that used to be termed “open core.” That is where a company supporting a commercial distribution of an open source platform is adding proprietary features atop it. But, as shown in its new 6 DSE release, DataStax is evolving its offering toward a more multifaceted platform that won’t strictly be defined by Cassandra. In so doing, it is embracing a multi-model path that others, such as Microsoft’s cloud-only Cosmos DB platform, are introducing to redefine the NoSQL database market.

Extensibility is a key theme of DSE 6

There are two broad themes to the v6 release. The first is the steady march of adding “enterprise-grade” capabilities that are the theme of any data platform provider that is targeting business-critical deployments. The other major focus is extensibility, binding search, analytics, and graph more tightly into the platform.

Enterprise-grade improvements encompass performance and manageability. Performance has been optimized for compute processor-intensive hardware that, at minimum, double throughput. Another

key improvement is a new bulk loading capability that provides up to a 4x increase in bulk loading and unloading performance.

For manageability, DSE 6 eliminates a major administration complaint that users have had about open source Cassandra for years – repair operations. DSE's new NodeSync feature automatically handles repair operations, eliminating what had been a very labor-intensive process. DSE 6 provides the option for administrators to monitor NodeSync operations visually through the refresh of OpsCenter, which accompanies the DSE 6 release. OpsCenter 6.5 (DataStax has not yet sync'ed the version numbering) adds a new automated upgrade service that handles routine patches. Additionally, a Docker image of DSE is now available as a development platform; a Docker image for production will arrive shortly after DSE 6 starts shipping.

For extensibility, DSE 6 goes beyond the original Cassandra database's operational use case roots to make analytics and search first class citizens. Specifically, it unifies analytics and search and makes them available to all schema/models that are stored in DSE.

It includes a new enterprise connectivity layer for Apache Spark that combines a new SQL engine (it is *not* Spark SQL) with high availability functionality and more transparent analytics support for graph. Graph is an area of strength for DSE, as the team it added through the 2015 Aurelius acquisition leads, and is responsible for 90% of the commits, to the Apache TinkerPop project. TinkerPop is the framework that has become the de facto standard for modern graph property databases (most of them support it). With the 6.0 release, you can now load graph data along with Cassandra data concurrently; you no longer must load graph tables separately. But full graph integration for now is a work in progress, as the graph schema is still maintained separately within DSE; that will change in an upcoming dot release when DataStax unifies graph schema into the mother ship, meaning that you'll only need to maintain a single schema.

Search is also more tightly bound into the core engine with DSE 6 as it gets unified with the Cassandra CQL query language; that will facilitate performing search operations with familiar SQL keywords such as LIKE.

The unified analytics support in DSE and growing convergence of graph tables reflects a trend toward multi-model databases that support multiple paths for query. Examples include addition of interactive SQL query to Hadoop, JSON support in household name relational databases, and the inclusion of SQL-like query languages with Cassandra and Couchbase. The driver is that, while specific databases will still be targeted at specific use cases or data models, there is the need to extend them for edge cases (e.g., adding SQL query to JSON databases or vice versa) where organizations would rather not implement a second data platform. The new features in DSE 6 provide such flexibility.

Courting the Cloud

With Ovum forecasting that cloud will become the default target for *new* big data workloads starting in 2019, a managed DSE cloud service is essential for DataStax to plug into this growth. This is an area where incumbent (read: on premise) data platforms have been playing catch-up with Amazon, Azure, and Google Cloud.

DataStax is ramping up cloud support. DataStax has always positioned DSE as a distributed database suited for hybrid on-premise and cloud deployment. It has been certified for each of the major public clouds (AWS, Azure, and Google Cloud), and notably, has become one of the first third-party databases to support the Oracle Public Cloud. Now it is taking the next step by inaugurating its

own *managed* cloud service. The DataStax Managed Cloud is available on AWS, and has just added presence on Microsoft Azure.

The brains behind the managed service largely come from DataScale, a consulting firm that DataStax picked up in 2016 that specialized in large-scale Cassandra deployment. While the DataStax Managed Cloud automates core functions of managed cloud such as deployment, self-healing, and upgrade/patching, the expertise from DataScale allows DataStax to offer professional services for onboarding or other scenarios requiring high-touch. Our take, however, is that professional services will be the side show. The expertise that came with this team will be reflected more in the level of automation in DataStax Managed Cloud.

Bolstered data privacy

Security and data privacy have become strategic issues for the enterprise, with data protection directives such as the EU's General Data Protection Regulation mandating granular control of data. Traditionally, NoSQL databases have been thought of as the less-secure younger sibling to relational databases; because of the flexibility of the schema, it's easier for sensitive data to inadvertently be ingested, and harder to detect and manage it once it's in there. Nevertheless, it is unrealistic to simply avoid placing sensitive data types in NoSQL environments – much of the world's personal and sensitive information exists in structured, unstructured and semi-structured formats not suited to relational systems. Instead, vendors need to step up their security and privacy capabilities to meet the growing enterprise demand for these features.

DataStax has responded to this increased enterprise need by gradually adding in more security, governance, and data protection features. Already available prior to DSE 6 was encryption between the nodes and clients, role-based access controls, transparent data encryption, data auditing, row-based security, and integration with Kerberos, LDAP and active directory. For the enterprise, though, these are becoming check-box requirements: not true product differentiators. So to address this, DSE 6 is adding in more granular privacy features that allow for the separation of duties, ensuring that more than one individual can be required to perform critical changes. There's still more work to be done on the security and governance front; data masking capabilities, a key facilitator for compliance with regulations such as GDPR, is still conspicuously absent. This, however, is on the product roadmap, as are more granular security management features.

Appendix

Further reading

On the Radar: DataStax, IT014-00286 (December 2013)

"DataStax adds JSON and graph computing to its Cassandra distribution," IT0014-003141 (July 2016)

Open Source and Big Data, IT0014-003193 (February 2017)

Author

Tony Baer, Principal Analyst, Information Management

tony.baer@ovum.com

Paige Bartley, Senior Analyst, Information Management

paige.bartley@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

CONTACT US

www.ovum.com
askananalyst@ovum.com

INTERNATIONAL OFFICES

Beijing
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
San Francisco
Sao Paulo
Tokyo