IDC ANALYZE THE FUTURE

IDC VENDOR SPOTLIGHT

Sponsored by: DataStax

Data management technology must be flexible to support just about any data format and enable enterprises to define and analyze data, with features that ensure scalability, reliability, and data security.

Technology to Manage Always-On, Always-Connected Data in a Hybrid Cloud World

September 2018

Written by: Carl W. Olofson, Research Vice President, Data Management Software

Executive Summary

Enterprises are faced with a number of challenges as they struggle to compete in an always-on, digitally connected world. The range of types and shapes of data to be collected and analyzed is constantly growing. The process of digital transformation is applying pressure on older systems to evolve and newer systems to ensure that the data is consistent, current, comprehensible, trustworthy, and secure. On top of this, enterprises must formulate policies with respect to the cloud: what data will move, what data will stay, what applications will be enhanced, what applications will be replaced, and what applications must be added?

AT A GLANCE

WHAT'S IMPORTANT

Data management systems can be uncoordinated with each other, can lack consistent policy enforcement for data access or updates, and don't provide a coherent way for users to find the data they need. To leverage data effectively, organizations need better data management technology.

Behind all of this must be data management technology that is flexible enough to support almost any data format and manageable enough to enable enterprises to define and analyze that data, with features that ensure scalability, reliability, and data security. At the same time, such a system must support agile development by allowing for independent data development by project groups with the autonomy to build data systems as they see fit and then to reconcile and integrate the data that is needed for the broader enterprise. What are the required elements of such a system? This IDC Vendor Spotlight examines these matters and looks at DataStax Enterprise (DSE) 6 from DataStax as a technology to consider in confronting this world of challenges and choices.

Where We Are Today

Today, enterprises use a range of data management technologies, each designed to serve the specific data management and access needs of various kinds of applications. The technologies include large-scale data collection; customer experience management; mobile device management; coordination, collection, and analysis of data from the Internet of Things (IoT); data warehousing; and back-office transaction processing.

These systems are often uncoordinated with each other, lack consistent policy enforcement for data access or update, and fail to provide a coherent means for users to find the data they need and view it in context. This means that much, or even most, of the data is not being leveraged to its fullest effect, and it isn't even clear how much of it ought to be moved to the cloud or retained at all.

The Demands and Challenges of Digital Transformation

The digital transformation that is expanding enterprise data management from systems of record to systems of engagement, from internal systems with external interfaces to blends of internal and external systems, is changing how we see and use data on a fundamental level. This transformation is beginning to impact IT organizations at all levels, moving enterprises from using data only in an internally controlled and managed form for the automation of fixed business processes to using data that is highly variable; that may come from external, uncontrolled sources; and that enables new modes of operation and analysis. This transformation is being driven, at the data level, by a host of very different data management technologies, including such schema-optional (also called NoSQL) data management systems as key-value (KV) stores, document databases, graph databases, and wide column stores.

Data from these operational data stores is sometimes collected, cleansed, organized, and transformed in scalable data lake platforms such as Hadoop. It arrives not only in files but also in streams, and it sometimes demands immediate processing. Data comes from sensors, logs, mobile devices, and the IoT. The data lake is a useful ingestion point and can serve as a base for deep but narrow queries and analytics. For more complex analytics, it may be moved to an environment that supports scalable column-oriented queries, such as a wide column store. For a complex and well-managed combination of this data with other enterprise historical data, it is often moved into a relational data warehouse. A data lake is suitable for collecting, cleansing, sorting, and deduplicating data, and for focused analysis but is generally not appropriate for either operational data management or complex query and analysis.

The operational data is used for a variety of purposes as well. Some operational data enhances data already in the data warehouse that is used for business analysis. Some operational data is used for more specialized analytics, such as network failure detection and fraud detection. In other situations, operational data manages state for highly interactive applications including gaming applications, ecommerce, set-top boxes, and trading applications. Data is also used to drive real-time processes such as fleet management or shop floor control. Further, operational data is used for predictive analysis and machine learning.

In the past, all the data in the enterprise was kept in well-defined files and databases. It served the needs of the applications for which it was designed. It changed format and structure only slowly. The digital transformation has brought about a new class of data that may come from the outside or may be born in an application but repurposed for analytics. Its volume is enormous. Its structure and format change frequently. Often, a NoSQL database is the right technology for this kind of data. This enormous shift offers significant potential but comes with demands and challenges that have never been seen before.

Cloud, Multicloud, and Hybrid Cloud

Cloud is not just another deployment platform. Efficient use of cloud resources requires a radically different way of architecting data management for many database management systems (DBMSs), especially those that come from a centralized, legacy background. Cloud DBMSs must have built into them the ability to reference all resources symbolically (rather than by physical name or location), to allocate and release resources dynamically (including processors, memory, and storage), to handle multitenant issues transparently and, in conjunction with that, to self-adjust performance to avoid the "noisy neighbor" problem. These capabilities enable DBMSs to satisfy the core requirements of the cloud: virtualization and elastic scalability. DBMSs built for this environment are designed to deal with these issues.

A multicloud environment brings with it an added level of complexity. Most enterprises are likely to have application data in more than one cloud. Coordinating that data across clouds is a significant challenge. In the absence of a system designed to address this challenge, enterprises are looking at considerable manual effort in coordinating data across clouds. These challenges also exist with respect to hybrid cloud, where a portion of enterprise data is managed in a private cloud on the premises, and the rest is managed in one or more managed cloud services. The problem is made worse in this case by the issues of process scheduling and data latency between the on-premises data and the cloud data.

The Data Security Imperative

The rise in security-based regulations, especially GDPR, in recent years has highlighted the need for better, more comprehensive security policy definition and enforcement across the enterprise. It is critical for the enterprise to have a consistent security system that not only identifies data that requires immediate protection but also includes role-based access at the data item level and the ability to mask or redact sensitive data.

The Need for a Self-Managing Data Service

Databases used for mission-critical activities present special challenges. Mission-critical databases not only must always be available but also must deliver excellent transaction throughput. These operational databases can become massive in size, supporting large-scale queries, yet are expected to deliver subsecond responses to those queries.

Database administrators (DBAs) are tuning the database and monitoring the database for any problems and to perform preventive maintenance. In addition, the processes involved in applying software upgrades and patches can be highly disruptive and time consuming. Despite many self-managing features in enterprise DBMSs, it remains the case that, at the highest level of size, complexity, and performance criticality, meeting the service-level agreement (SLA) for the database requires a set of skills generally regarded as something of a black art. And of course, even the best of DBAs can make mistakes — and those mistakes can cause poor performance and unscheduled downtime.

With a database cloud service, the nuts and bolts of applying patches and basic tuning are taken care of by the service itself. In addition, self-managing features are called for so that the database system can optimize its own configuration and performance with minimal human intervention.

A Summary of Requirements

A system that provides management of data across the enterprise in an always-on, always-available manner handles widely distributed data across on-premise, private, hybrid, public, and multicloud deployment scenarios; ensures consistent data security; and provides automated self-management features is required to enable coherent data management across the process of digital transformation. Such a system should also feature the following:

- Contextual data support, enabling users to see data in a meaningful context
- Operation in an always-on manner, with failover and automatic recovery features
- Capability of handling real-time streaming data and the low-latency tolerances involved
- Dynamic scalability without requiring downtime
- Full, robust data security

Considering DSE 6 from DataStax

DataStax Enterprise is a distributed DBMS solution designed to offer data management functionality that fits into the agile development model while retaining support for the structure and rigor demanded by enterprise data systems. Built on top of the widely embraced Apache Cassandra open source data management system, DSE includes Apache Spark and Apache Solr integration for analytics and search, as well as needed facilities for managing the same data in multiple contexts, with a multimodel development approach that supports key-value, tabular, JSON, and graph forms of data organization and access. DSE is commonly used for digital operational and customer experience applications in which immediate, continuous, and robust transactional data management is required.

DSE's access languages include Cassandra's native Cassandra Query Language (CQL) as well as SQL, and DSE Studio enables developers to incorporate Spark SQL and graph data as well, all supported by an easy-to-understand visual interface. Supported programming languages and access APIs include C/C++, C#, Java, Node.js, ODBC/JDBC, Python, PHP, and Ruby as well as a range of synchronous and asynchronous APIs. DSE can be deployed on a single node or cluster in an on-premises datacenter or in a cloud, across multiple clouds, and in a hybrid cloud.

DSE 6 satisfies the previously mentioned requirements as follows:

- **Contextual data support.** DSE supports multiple workloads (operational, analytical, search) in the same cluster as well as multiple data models, thus enabling full support for all data workflows and models. DSE also connects to Hadoop data lakes for historical analytical functions. On the developer side, DSE Studio operates in a context-aware manner, offering suggestions and validations as users write Cassandra, Spark, and graph queries, obviating the need for users to refer to definitional metadata.
- **Operation in an always-on manner.** DSE is built on Cassandra and designed to deliver virtually limitless uptime for applications. In addition to continuous operation, DSE offers continuously available components such as its AlwaysOn SQL Engine, which ensures constant and uninterrupted analytic query support.

- **Ability to handle real-time streaming data.** In addition to Cassandra's operational data support, DSE offers integrated Spark Streaming and enables DSE users to query data at rest and data in motion in a straightforward fashion. Moreover, DSE 6 supplies a 2–3x read/write performance gain over open source Cassandra.
- **Dynamic data distribution, scalability, and manageability.** DSE's masterless architecture ensures maximum deployment flexibility and predictable linear scale performance, along with smart automatic and visual management capabilities such as DSE 6's new NodeSync functionality that eliminates the need for manual Cassandra repair operations.
- **Robust data security.** In addition to Cassandra's basic security support, DSE's security features include auditing, external authentication support, data encryption, row-level access control, unified authentication, and support for standard security protocols.

Challenges

In a field crowded with NoSQL contenders, competing claims, and confusion regarding data management requirements in an always-connected world, DataStax will be challenged to rise above the din with a message that is clear and compelling. Also, changes in vendor alignments, technologies, and user requirements demand a flexible product development plan and nimble execution.

Conclusion

Digital transformation, driven by the need for the full utilization of all relevant data by the enterprise, the desire to better serve the customer and keep the customer engaged, and the opportunity to realize greater levels of efficiency, productivity, and innovation, has created a mandate for new ways to manage data and leverage its value. In place of the incumbent data management technologies that tend to require formalism that stifles agility while lacking scalability and flexibility, enterprises are looking to new technologies in a class that is commonly described as NoSQL.

With its support for a wide variety of data types, its multimodel approach, its easy-to-manage scalability and data distribution, and its foundation in the well-regarded and battle-tested open source platforms of Apache Cassandra, Apache Spark, and Apache Solr, DSE 6 from DataStax has carved out a critical part of this emerging data environment. The system is designed to deliver scalable, high-speed data management and analytics both on-premises and in the cloud.

About the analyst:

Carl W. Olofson, *Research Vice President, Data Management Software*

Carl Olofson manages IDC's Database Management Software service. Mr. Olofson's research involves following sales and technical developments in the structured data management (SDM) markets, including database management systems (DBMSs), dynamic data management systems, database development and management software, and dynamic data grid managers.

IDC Custom Solutions

IDC Corporate USA
5 Speen Street
Framingham, MA
01701, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com