The Database Platform for Cloud Applications
DataStax Enterprise 5.0, OpsCenter 6.0, and Studio 1.0

June 2016
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TABLE OF CONTENTS</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>INTRODUCTION</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>MULTI-MODEL AND DATASTAX ENTERPRISE</strong></td>
<td>3</td>
</tr>
<tr>
<td>Introducing DataStax Enterprise Graph</td>
<td>3</td>
</tr>
<tr>
<td>Common Graph Use Cases</td>
<td>3</td>
</tr>
<tr>
<td>Visual Graph Development</td>
<td>4</td>
</tr>
<tr>
<td>Visual Graph Management and Monitoring</td>
<td>5</td>
</tr>
<tr>
<td>Support for JSON</td>
<td>5</td>
</tr>
<tr>
<td>Integrated Drivers</td>
<td>6</td>
</tr>
<tr>
<td><strong>SIMPLIFYING DATABASE MANAGEMENT AND OPERATIONS</strong></td>
<td>6</td>
</tr>
<tr>
<td>Advanced Replication</td>
<td>6</td>
</tr>
<tr>
<td>Enhanced Security</td>
<td>7</td>
</tr>
<tr>
<td>Unified Authentication</td>
<td>7</td>
</tr>
<tr>
<td>Role Based Access Control</td>
<td>7</td>
</tr>
<tr>
<td>Tiered Storage</td>
<td>7</td>
</tr>
<tr>
<td>Multi-Instance</td>
<td>7</td>
</tr>
<tr>
<td>Enhanced Analytics</td>
<td>7</td>
</tr>
<tr>
<td>Spark Streaming Fault Tolerance</td>
<td>8</td>
</tr>
<tr>
<td>Bring Your Own Spark</td>
<td>8</td>
</tr>
<tr>
<td>Faster Enterprise Search</td>
<td>8</td>
</tr>
<tr>
<td>Enterprise Provisioning</td>
<td>9</td>
</tr>
<tr>
<td>Improved Alerting</td>
<td>9</td>
</tr>
<tr>
<td>External Monitoring Integration</td>
<td>9</td>
</tr>
<tr>
<td>Certified Cassandra 3.0</td>
<td>10</td>
</tr>
<tr>
<td>Materialized Views</td>
<td>10</td>
</tr>
<tr>
<td>New Storage Engine</td>
<td>10</td>
</tr>
<tr>
<td>New Hints</td>
<td>10</td>
</tr>
<tr>
<td><strong>CONCLUSION</strong></td>
<td>10</td>
</tr>
<tr>
<td><strong>ABOUT DATASTAX</strong></td>
<td>10</td>
</tr>
</tbody>
</table>
Introduction
With DataStax Enterprise, DataStax delivers a comprehensive data management layer with a unique always-on architecture that accelerates the ability of enterprises, government agencies, and systems integrators to power the exploding number of cloud applications. DataStax Enterprise (DSE) powers cloud applications that require data distribution across datacenters and clouds, by using a secure, operationally simple platform built on Apache Cassandra™.

This paper provides a summary of the new features and functionality added as part of the DSE 5.0 release which includes updates to the associated development tools and OpsCenter 6.0. The objective of the release was to make it easier to both build and run large-scale, interactive and intelligent applications. As part of this initiative the features and capabilities added as part of the release fall into two major categories:

1. Robust multi-model database capabilities
2. Simplification and automation of management, monitoring, and development tasks

Multi-Model and DataStax Enterprise
Modern applications need to consume data using a variety of different access patterns and this has historically required using a number of different databases, one for each access method. The complexity created by stitching together multiple systems makes it difficult and costly to maintain or develop new functionality.

To solve this problem, DataStax Enterprise provides a powerful multi-model platform with support for key-value, tabular, JSON and graph. This capability allows you to write data to a single solution and access it using a variety of methods based on the needs of your application.

Introducing DataStax Enterprise Graph
A graph database is used for managing and querying data that is complex and highly connected. They make it easy to find commonalities or abnormalities and unlock the value in data relationships.

DataStax Enterprise Graph is the first graph database fast enough to power customer facing applications, capable of scaling to massive datasets and advanced integrated tools that power deep analytical queries. Because all of DataStax Enterprise is built on the core architecture of Apache Cassandra™, DataStax Enterprise Graph can scale to billions of objects, spanning hundreds of machines across multiple datacenters with no single point of failure. DataStax Enterprise delivers a full graph solution, with tools for both application developers and operations staff.

- DataStax Enterprise Server 5.0 – Supports the graph data model
- OpsCenter 6.0 – Create, manage and monitor graph databases
- DataStax Drivers – Supports the Gremlin language used for graph databases as well as all other API’s used in DataStax Enterprise (e.g. CQL, etc.)
- DataStax Studio 1.0 – Visualize and query graph databases using the new visual developer tool

Common Graph Use Cases
Graph databases are well suited for large data sets with numerous and highly complex relationships. Graph databases make it easy to discover, explore and make sense of these relationships.

Customer 360°
A company must understand the data relationships across its multiple business units to create a holistic view of its customers. A graph model is the best way to consolidate the disparate data for use by both BI tools and other business applications. An example of this application include using graph to understand the various ways a customer interacts with your company, what types of accounts do they
have, what services are they using and what are the various identities they use across the separate properties both virtual and physical.

**Recommendation and Personalization**

Almost all enterprises need to understand how they can quickly and most effectively influence customers to buy their products and recommend them to others using components in a cloud application such as recommendation, personalization, and network (people or machines) analysis engines. A graph is well suited to these and similar analytical use cases where recommending products, next actions, or advertising based on a user’s data, past behavior, and interactions are important.

**Security and Fraud Detection**

In a complex and highly interrelated network of users, entities, transactions, events, and interactions, a graph database can help determine which entity, transaction, or interaction is fraudulent, poses a security risk, or is a compliance concern. In short, a graph database assists in finding the bad needle in a haystack of relationships and events that involve countless financial interactions.

**Internet of Things**

The IoT use cases most commonly involve devices or machines that generate time-series information such as event and status data. A graph works well in this case because the streams from individual points create a high degree of complexity when blended together. Further, analytics involved in tasks such as root-cause analysis, involve numerous relationships that form among the data elements and tend to be of much greater interest when examined collectively than reviewed in isolation.

**Visual Graph Development**

DataStax Studio is a new web-based visual tool designed to help developers visually interact with and query DSE Graph databases. DataStax Studio lets developers easily write Gremlin queries (the industry standard graph query language) and visualize graph data output.
Visual Graph Management and Monitoring

DSE Graph can be visually managed and monitored through DataStax OpsCenter 6.0. Graph-enabled clusters can easily be provisioned and administered in OpsCenter, while key performance diagnostics can also be collected and monitored through the visual interface of OpsCenter.

Support for JSON

In addition to key/value, tabular, and graph, DataStax Enterprise 5.0 also supports the JSON data model. The JSON model support in DSE has the flexibility to store data with complex nested schemas and is able to easily move data to and from application tiers – both of which are the primary use cases of popular Document-oriented databases. The major difference between DSE and some other databases that support JSON is the requirement that the JSON data adhere to a database schema.

Support for JSON is present in both the DataStax Enterprise Server and DataStax DevCenter, which is a visual development tool that simplifies the effort required to develop and query JSON data models.
Integrated Drivers
DataStax provides drivers for all key development languages (e.g. Java, Python, C#, C++, Ruby, Node.js, PHP, etc.) For DSE 5.0, all drivers have been updated to include support for all data models supported by DSE. This means that a single driver can interact with all data models utilized by your application which simplifies application development.

Simplifying Database Management and Operations
Many of the enhancements in this release are aimed at simplifying the management and operations of distributed database clusters. Significant additions and improvements have been several key areas such as: replication, security, and storage management.

Advanced Replication
DSE Advanced Replication builds upon the multi-datacenter support in Apache Cassandra™ to facilitate scenarios where selective or “hub and spoke” replication is required. DSE Advanced Replication connects multiple clusters and keeps data in sync and up-to-date, whereas the native Cassandra operates only within a single cluster. By replicating at the cluster level, you can replicate data from remote deployments to a central hub to produce a global view of the entire enterprise.

DSE Advanced Replication is specifically designed to tolerate sporadic connectivity that can occur in constrained environments, such as retail, oil-and-gas remote sites and cruise ships. In these environments we need a system that allows us to prioritize our data in order to make the best use limited bandwidth. Additionally when connectivity is lost, data must be stored at the edge so that when connectivity is restored, replication resumes and no data is lost.
With DSE Advanced Replication data flows from an edge cluster can be prioritized to ensure the most important data is sent first. In the instance that connectivity is lost, DSE Advanced Replication will continue to store data so that your remote applications can continue to function, and then replicate this data once connectivity is restored. And because DSE Advanced Replication is fully integrated into the DSE platform workloads such as Search and Analytics can be used at both remote and central locations.

**Enhanced Security**
DataStax Enterprise builds upon the basic security features provided by the open source Apache Cassandra™ project by providing the tools necessary to meet stringent HIPAA, PCI and SOX compliance requirements. This functionality includes support for industry standard authentication mechanisms, role-based authentication, user activity auditing, and end-to-end encryption. With DSE 5.0, DataStax has enhanced the existing functionality to include support multiple authentication sources, encryption of the commit log, and encryption for DSE Search index data.

**Unified Authentication**
Security administrators can now configure DSE clusters for multiple sources of user credentials and groups. As an example, Unified Authentication allows utilize both ActiveDirectory and LDAP as sources of authentication. This allows administrators to provision applications via one source, and users through another. Unified Authentication supports LDAP, Kerberos, ActiveDirectory, and the users and groups native to Cassandra.

**Role Based Access Control**
Role Based Access Control has been enhanced to support role assignment in LDAP and Microsoft ActiveDirectory servers. Role definitions and database permissions are stored in the database and managed via conventional GRANT or REVOKE statements. However, the association between a role and a user can be managed externally in ActiveDirectory or LDAP, which eliminates the duplication of user data.

**Tiered Storage**
DSE Tiered Storage allows database administrators to leverage multiple storage options in order to achieve performance and cost optimizations goals. This feature allows administrators to keep recently written and hot data on more performant solid-state drives (SSD), and older less-frequently accessed data on slower but cheaper mechanical disks (HDD). This operation is done transparently and automatically, with data remaining accessible regardless of where it resides.

**Multi-Instance**
Datacenter real estate is at a premium, and it’s sometimes cheaper to run larger, more powerful machines. To help, DSE allows administrators to achieve higher overall system utilization by dividing machines among several instances of DSE. DataStax Enterprise intelligently provisions database nodes across the datacenter to ensure that cost saving goals are met without compromising availability in the instance of rack or machine failure.

**Enhanced Analytics**
DSE 5.0 continues our investment in Apache Spark by integrating Spark 1.6. With this release introduces significant performance enhancements including: Catalyst optimizer improvements for DataFrames and SQL, dynamic resource allocation, Spark Streaming backpressure, improved Spark Streaming state management, unified memory management, and significant additions to Spark’s machine learning packages.

Integration with Cassandra has also been improved by removing unnecessary data movement, pushing down processing into Cassandra, adding new Cassandra data types and supporting DSE Search queries from DataFrames and Spark SQL.
Spark Streaming Fault Tolerance
Streaming applications run for an indefinite amount of time consuming data and performing real-time analytics. However, streaming applications are not immune to system faults such as disk failure or hardware failure. Streaming applications need to tolerate these faults without having to replay the stream from the beginning. To accommodate these scenarios Spark provides a checkpoint mechanism such that data processing resumes at the last checkpoint.

With this release DSE provides a powerful new feature called the DSE File System (DSEFS) to address the checkpoint needs of Spark Streaming applications. DSEFS is built from the ground up with scalability, fault-tolerance, and performance as core principles. In this debut release, DSEFS is focused on the needs of streaming applications, integrating seamlessly into Spark as an HDFS-compatible file system and allowing a fault-tolerant way to checkpoint these Streaming applications. In addition to the checkpointing capabilities, DSEFS provides a command-line management tool to create directories, load files, and copy, move, and delete files and directories, as well as monitor file system usage.

Bring Your Own Spark
Some organizations have already made investments in their big data infrastructure, and would like to combine these investments to leverage their combined capabilities. DataStax recognizes that the data center is a lively village of technologies, and the challenge is to combine these villagers in the Big Data applications. To that end, we introduced Bring Your Own Hadoop in DSE 4.5 to allow you to connect your current Hadoop investment to DataStax Enterprise.

In DSE 5.0 we are shifting our whole focus to Spark. Not only has Spark come to be the preferred way to conduct analysis of data in Cassandra from within DSE, but it is also the mechanism we are using to allow integration with Hadoop vendors from outside of DSE. This Bring Your Own Spark capability will allow users to connect their existing Hadoop investments to DataStax Enterprise while leveraging the advanced features of DSE, such as Advanced Security. This configuration is streamlined via a simple command-line tool to export the necessary parameters and artifacts that can easily be added to Spark job submissions. Similarly, external Hadoop resources, especially HDFS, are accessible by Spark jobs run from with DSE.

Faster Enterprise Search
One of the common challenges with building search applications is minimizing the latency between the time the data is inserted into a data store and the time queries are able to search that data. DataStax introduced Live Indexing a few releases back, which provides real-time indexing capabilities that address the problem.

DSE 5.0 brings a number of enhancements to Live Indexing, which continue to set the standard for real-time indexing performance. Included are new abilities that allow the system to utilize memory outside of the native memory space, which improves the efficiency of Live Indexing and reduces latency between the insert and query time even further. In addition, an improved back pressure mechanism also gives DSE Search nodes more stability and indexing efficiency when data is being inserted into Cassandra at the high throughput rate for which Cassandra is known.
Enterprise Provisioning
The new and revamped OpsCenter Lifecycle Manager Deploy feature transforms the manual and complex provisioning process into a simplified, robust, and automated task. OpsCenter provisioning allows you to automate the deployment and management of DSE clusters on premise or in the cloud.

Improved Alerting
Managing mission-critical applications that depend on DSE require you to have granular notifications on the health of your clusters. In OpsCenter 6.0, we have improved the alerting functionality of OpsCenter by adding a number of features such as SNMP Traps, PostURL enhancements and custom email expansions.

External Monitoring Integration
If monitoring your key metrics from a centralized monitoring dashboard such as Graphite is important for you, we have some good news. In OpsCenter 6.0, OpsCenter agents can push the key metrics directly to Graphite server. With just a couple of settings, OpsCenter can successfully push its metrics and lets you monitor the health of your DSE clusters directly from Graphite.
Certified Cassandra 3.0

Materialized Views
Cassandra users are accustomed to data modeling with significant denormalization in the form of “query tables”. That is, tables that are structured with a key, enable efficient retrievals for the specific queries in their applications. These tables have always been managed manually, resulting in a fair amount of boilerplate code in applications and the extra complexity of keeping data up to date in multiple locations. Any developer that has created more than a handful of query tables off a base table will recognize the usual pattern for query table creation - include all of the columns from the first table, but change the key around. Materialized Views can be thought of as automated, database-side denormalization to enable Cassandra to handle those query table permutations by defining view definitions off of a base table. The developer is no longer responsible for managing multiple tables in their application just to be able to query efficiently.

New Storage Engine
The new storage engine is designed from the ground up to be far more efficient for CQL data and does away with column name repetition resulting in massive space savings. It also stores timestamps more efficiently and keys the data in manner more efficient for tables with clustering columns.

New Hints
Hints are no longer stored in Cassandra tables and instead are stored as files on the file system. A result is that hints can be handed off in segments very efficiently to the node expecting them. Once the hints are delivered, the file can simply be removed rather than having to issue a delete to the hints table.

Conclusion
The DSE 5.0 release delivers multi-model capabilities for cloud applications along with a simplified and streamlined set of functionality that automates difficult tasks and makes everything easy to use.

The latest software is available for download now on the DataStax Downloads Page.

About DataStax
DataStax, the leading provider of database software for cloud applications, accelerates the ability of enterprises, government agencies, and systems integrators to power the exploding number of cloud applications that require data distribution across datacenters and clouds, by using our secure, operationally simple platform built on Apache Cassandra™.

With more than 500 customers in over 50 countries, DataStax is the database technology of choice for the world’s most innovative companies, such as Netflix, Safeway, ING, Adobe, Intuit, Target and eBay. Based in Santa Clara, Calif., DataStax is backed by industry-leading investors including Comcast Ventures, Crosslink Capital, Lightspeed Venture Partners, Kleiner Perkins Caufield & Byers, Meritech Capital, Premji Invest and Scale Venture Partners. For more information, visit DataStax.com or follow us @DataStax. 06.15.16