Why You Need the Best Distribution of Apache Cassandra™, and So Much More

CONTENTS

INTRODUCTION

Today's "Right-Now" Enterprises continue to differentiate themselves with cloud applications—any application that needs to be always-on, distributed, scalable, real-time, and contextual. With DataStax Enterprise (DSE), DataStax delivers the always-on distributed cloud database, built on Apache Cassandra™ and designed for hybrid cloud, that accelerates the ability of enterprises, government agencies, and systems integrators to power the exploding number of cloud applications.

DSE powers these cloud applications, which require data distribution across data centers and clouds, by using a secure, operationally simple platform, built on an enterprise-ready, certified version of Apache Cassandra™. DSE packs a whole suite of capabilities that enable enterprises to design, develop, deliver best-in-class cloud apps that can run seamlessly on any cloud (public, private, on-premises, or any combination thereof) run, manage, and scale these cloud apps so that they never fail; and gain contextual visibility and control across all data for transformative business decisions.

Over the years, DataStax has been instrumental in the popularity and growth of Apache Cassandra. Our six-year history can be characterized by three Eras:

- ❯ Era 1 (2010-2013): In this period, DataStax was focused on making Apache Cassandra™ a viable open-source project.

- ❯ Era 2 (2013-2016): During this period, DataStax invested time, resources and effort in driving the worldwide growth of Apache Cassandra™.

- ❯ And now in Era 3 (2016-20XX): Now, DataStax is focused on delivering the market-leading distributed cloud database to our customers: DSE.

There is often a notion in the open source world that a wide community of developers helps build the software. That is not true in the case of Apache Cassandra. More than 85% of the code came from DataStax employees. There are "company-specific" contributions from others, but the majority came from a core group of experts at DataStax. The universe of developers who understand the complexities of a distributed database deep enough to build Cassandra is small. There have been about 140 authors of Apache Cassandra code over the life of the project—about 20 are DataStax employees and 85% of the work came from 15% of the contributors.

In addition, DataStax has built many new features and capabilities into Cassandra, manages most of Cassandra issues and bug fixes, and is responsible for all the test infrastructure, execution, and documentation. All Cassandra developers rely on DataStax Enterprise drivers for development tools.

This paper provides a summary of the features and functionality of DSE that make it the best choice for companies looking to leverage the promise of Cassandra for production environments. This paper specifically highlights additional capabilities of DSE 6, including self-driving operational simplicity and robust transactional analytics.

85% of Commits from DataStax
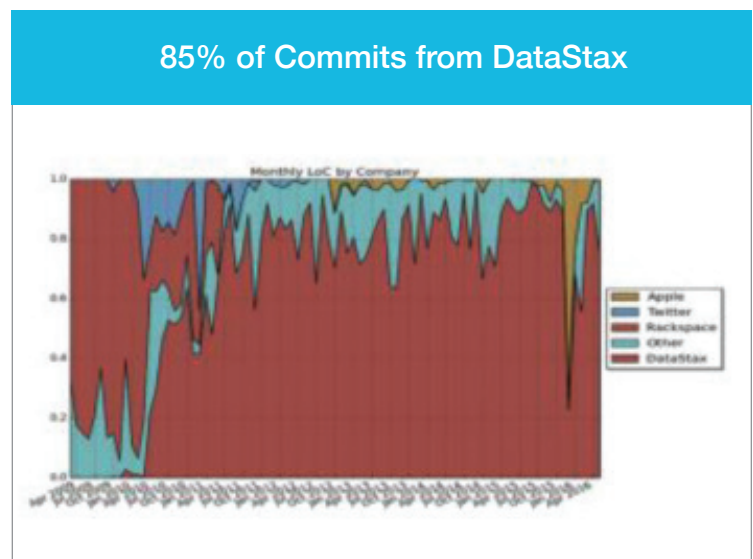
CERTIFIED APACHE CASSANDRA™ 3.0

Production Certified Apache Cassandra™

The DataStax certification process builds on the release model of Apache Cassandra and its method of using rapid development to deliver a stable data layer. The DataStax certification process includes formal quality assurance testing, scale and performance testing (up to 1,000 nodes for each major release), platform certification, third party software validation, certified software updates, back porting of bug fixes, long-term support, formal end-of-life policies, security audits, and digitally signed code.

In addition, DSE 6 offers:

> ### At Least Double the Performance of Apache Cassandra
>
> Right-Now Economy applications demand guaranteed and immediate responsiveness, no matter the surge in users or workloads. Apache Cassandra is relied upon by experts around the world as the open source database technology for real-time, geo-distributed applications at scale. DSE 6 Advanced Performance accelerates Apache Cassandra with half the latency for superior application responsiveness and twice the user capacity on the same infrastructure, and this capability has been independently verified.

> ### Self-Driving Operational Simplicity
>
> Right-Now Economy applications must be operationally simple to manage so that teams can focus on innovation and customers success. DSE 6 enables even novice DBAs and DevOps personnel to manage operations like seasoned professionals.
>
> DSE NodeSync reduces manual operations by 90% versus Apache Cassandra and eliminates cluster outages that are attributed to manual operations. That means operational cost savings, reduced support cycles, and reduced application management pain. NodeSync also makes applications run more predictably, and thus makes capacity planning easier. NodeSync's advantages for operational simplicity extend across the whole data layer including database, search, and analytics, something public cloud database services cannot achieve.
>
> DSE TrafficControl greatly reduces any bottlenecks from both client requests and internal replication.
>
> DSE Upgrade Service provides error-free patch upgrades with up to 60% less manual involvement.

> ### Highly Available Transactional Analytics
>
> Right-Now Economy enterprises demand instantly actionable insight with fresh, always available data to drive enterprise applications. DSE 6 delivers secure, highly available transactional analytics for the freshest, always available and in-the-moment insight.

> ### Hybrid Cloud Flexibility
>
> DSE 6 allows you to maintain hybrid cloud flexibility with all the benefits of a distributed cloud database on any public cloud as well as on-premises. Enjoy the best distribution of Apache Cassandra with none of the complexities for easy enterprise-wide adoption of data management at scale. Double the performance yields twice the responsiveness and throughput versus the prior version with the same hardware.

SIMPLIFYING DATABASE MANAGEMENT AND OPERATIONS

DSE is aimed at simplifying the management and operations of distributed database clusters.

DataStax OpsCenter

DataStax OpsCenter is a browser-based, visual management and monitoring solution for DSE. OpsCenter provides architects, DBAs, and operations staff with the tools necessary to intelligently and proactively ensure their databases are running well and that administration tasks are simplified so IT staff can concentrate on things other than babysitting their database systems.

DataStax OpsCenter provides advanced administration and monitoring capabilities for DSE clusters. The list of features that OpsCenter offers is elaborated below.

MONITORING	CLUSTER MANAGEMENT
• Custom Cassandra and server statistics	• Visual cluster provisioning /management for Cassandra
• Events dashboard	• Perform utility functions on each node or in mass
• View replication relationships	• Multi-cluster dashboard, support for 1000 node cluster deployment
• Proactive alerts	• Visual cluster provisioning for analytics, graph and search
• External notifications (e.g. email)	• Advanced security with role based access controls
• Analytics monitoring	• Automated node/ring balancing of data
• Search monitoring	• Visual Backup Service management
• In-memory tables monitoring	• Visual Performance Service
	• Visual restore management
	• Visual Repair Service management
	• Visual capacity planning w/ forecasting
	• Automated software update notifications
	• Automated support collection package
	• Visual Best Practice Service with built-in expert advisors
	• Built-in Auto-Failover
	• View Keyspaces
	• View Column Families

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.

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.

In-Memory Option

The in-memory option of DSE provides in-memory computing abilities to let developers, architects and administrators easily choose what parts (some or all) of a database reside fully in RAM. The in-memory option of DSE is designed to service use cases that lend themselves to in-memory computing while allowing disk-based workloads to be quickly serviced by the traditional storage model. This design allows applications with both in-memory and disk-based needs to be supported by one database platform. This capability is simple to use, an in-memory table acts and presents as any other table, making it completely transparent to an application and the overall database, there is no possibility for data loss with in-memory tables as all incoming writes go first to a disk-based commit log and are then written to the in-memory table, enables full tiered storage as data may be assigned to in-memory objects, SSD's, traditional spinning disks, cloud-based storage platforms all in the same database cluster, and in-memory table usage can easily be monitored and managed with DataStax OpsCenter.

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.

DATASTAX ENTERPRISE ANALYTICS

DataStax Enterprise Analytics integrates real-time and batch analytics capabilities directly in DSE. With DSE Analytics you can easily generate ad-hoc reports, target customers with advanced personalization and process real-time streams of data to create intelligent applications.

Because DSE Analytics is a core capability of DSE you can utilize powerful transactional, search, and graph functionality without having to implement the complicated and time consuming processes of keeping multiple data silos in sync with cumbersome ETL integrations. DSE Analytics allows you to easily stream data to or from your existing databases including: Apache Hadoop™, HDFS, HBase, Oracle, MySQL, IBM DB2, and external Apache Spark™ clusters. Because DSE Analytics also provides ODBC and JDBC compatible interfaces you can also easily integrate with your business intelligence tools.

DSE includes 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.

> ## 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.

DSE provides a powerful 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 DSE 5.0, 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. The Bring Your Own Spark capability allows 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.

DATASTAX ENTERPRISE SEARCH

DataStax Enterprise Search is the only commercial search solution that can scale for the needs to enterprises today. DSE Search is tightly integrated into DataStax Enterprise which uses the enterprise-certified version of Apache Cassandra™ to provide linear scale, continuous availability with zero downtime and native real-time multi-datacenter replication. Extending the capabilities of DSE Search intelligently partitions your data across the cluster to avoid hot spots, evenly distribute load, and route queries to provide the responsiveness that your application requires.

DSE Search includes support for all of the native data types of Apache Cassandra™ including user defined types and nested collections. In addition, DSE Search also includes indexing support for geospatial data and full text documents. You can easily analyze textual data using powerful lexical tools in multiple languages so that your users can find the data that is important to them. DSE Search also provides real-time filtering, faceting, and aggregation which makes it trivial to implement product catalogs, document repositories, and ad-hoc reporting engines. DSE sets the standard for real-time indexing performance. Included are 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, a 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.

ADVANCED SECURITY

DataStax Enterprise builds upon the basic security features provided by the open source Apache Cassandra™ 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.

> ● **End to End Encryption**
>
> DataStax Enterprise has advanced integrated encryption capabilities that protects data on disk, in transit between nodes, and between client and server. Combined with support for the industry standard KMIP encryption key management protocol, DataStax Enterprise protects your data whether it's in-flight or at rest.

> ● **User Activity Auditing**
>
> DataStax Enterprise has integrated user activity auditing functionality that allows you to record and audit all or a portion of user activity. Record all or track a subset of user activity including login attempts to discover data breaches or unauthorized behavior.

> ● **Unified Authentication**
>
> Security administrators can 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 supports 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.

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.

❯ 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.
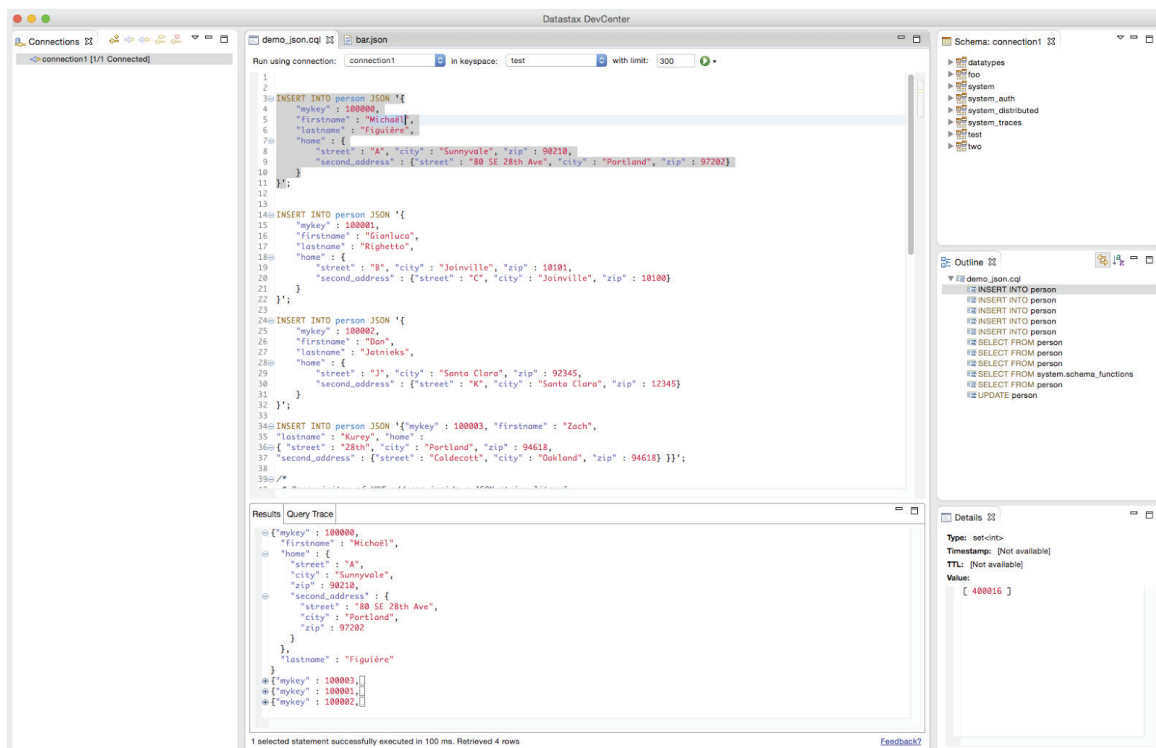
DATASTAX ENTERPRISE GRAPH

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—Supports the graph data model
- OpsCenter—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—Visualize and query graph databases using the new visual developer tool

❯ 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.
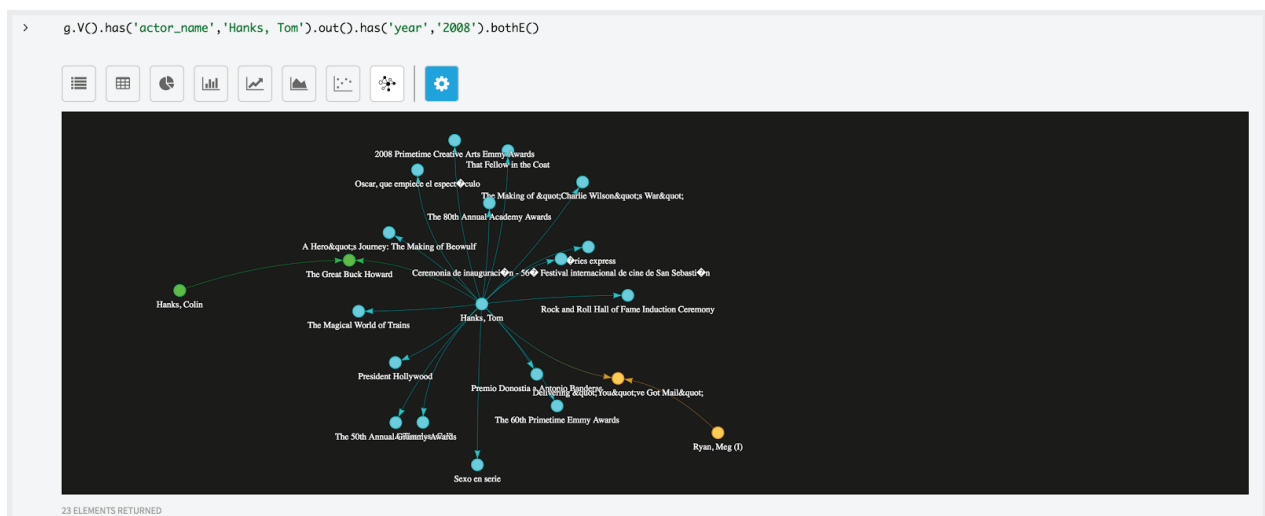
> ### Visual Graph Management and Monitoring

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.

DEVELOPMENT TOOLS

DataStax DevCenter, Studio, and database drivers are built from the ground up to enable teams to build intelligent, large-scale distributed, high-performance applications faster, while remaining familiar and easy to use.

> **Integrated Drivers**
>
> DataStax provides drivers for all key development languages (e.g. Java, Python, C#, C++, Ruby, Node.js, PHP, etc.) In DSE 5.0, a single driver can interact with all data models utilized by your application which simplifies application development.

SERVICES SUITE

DataStax also offers customers a whole host of services to help them accelerate their implementation of DataStax Enterprise, which includes our industry leading, certified version of Apache CassandraTM. The DataStax services team brings their deep product expertise and best practices knowledge and can assist across the board, from development environment set-up, architecture design, application-code reviews, production readiness, streamlining deployment and operations, and training.

CONCLUSION

DSE delivers on the promise of NoSQL technology via a distributed cloud database designed for hybrid cloud and cloud applications and with a simplified and streamlined set of functionality that automates difficult tasks and makes everything easy to use.

DataStax is driven by the demand from and expectations of our customers. As a result, we are continuing to build and extend capabilities like graph, that we believe our customers need to drive success in their enterprises. We will continue to add in newer features, integrations, extensions and tools in the DSE roadmap.

Enabling our customers' success is a key priority for us. DataStax believes in providing the training, (online self-service as well as live), documentation, services, and support that our customers need.

Stability is a major reason why customers already choose DSE. As our customers grow, their need for a true, always-on distributed cloud database is imperative. DSE delivers distributed database technology that is scalable, responsive, and intelligent, with capabilities like advanced performance, end-to-end security and the functionality necessary to build and operate enterprise implementation of DSE seamlessly and securely.

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

DATASTAX ENTERPRISE SUBSCRIPTIONS

	DSE BASIC	DSE STANDARD	DSE MAX
Database			
Enterprise/Production Certified Apache Cassadra™ Release	⊘	⊘	⊘
DSE Advanced Security Option		⊘	⊘
DSE Automatic Management Services		⊘	⊘
DSE Advanced Replication		⊘	⊘
DSE Tiered Storage		⊘	⊘
DSE Multi-Instance		⊘	⊘
DSE In-Memory			⊘
DSE Search			⊘
DSE Analytics			⊘
DSE Graph		Option	Option
Management and Development Solutions			
DataStax OpsCenter		⊘	⊘
DataStax Studio		⊘	⊘
Support and Professional Services			
Expert 24 x 7 x 365 Support with 1 Hour SLA	⊘	⊘	⊘
Platform Certification	⊘	⊘	⊘
Certified Service Packs	⊘	⊘	⊘
Hot Fixes and Bug Escalation	⊘	⊘	⊘
Free Software Deployments for Non-Production Environments	⊘	⊘	⊘
Health Checks/Performance Reviews	⊘	⊘	⊘

DATASTAX

ABOUT DATASTAX

DataStax delivers the always-on, Active Everywhere distributed hybrid cloud database built on Apache Cassandra™. The foundation for personalized, real-time applications at massive scale, DataStax Enterprise makes it easy for enterprises to exploit hybrid and multi-cloud environments via a seamless data layer that eliminates the issues that typically come with deploying applications across multiple on-premises data centers and/or multiple public clouds.

Our product also gives businesses full data visibility, portability, and control, allowing them to retain strategic ownership of their most valuable asset in a hybrid/multi-cloud world. We help more than 400 of the world's leading brands across industries transform their businesses through an enterprise data layer that eliminates data silos and cloud vendor lock-in while powering modern, mission-critical applications. For more information, visit DataStax.com and follow us on @DataStax.

Why You Need the Best Distribution of Apache Cassandra™, and So Much More