



DataStax Enterprise

Big Data Management for the Enterprise

WHITE PAPER

By DataStax Corporation

March 2012

Contents

Introduction.....	3
A Look at Key Data Management Challenges	3
Highly Available, High-Velocity Workloads	3
Mixed Workloads	4
Scale, Big Data, and Architecture Limitations.....	5
Widespread Data Distribution	7
Miscellaneous Challenges	8
What Is the Answer?	9
What Is DataStax Enterprise?	9
DataStax Enterprise – Powered by Apache Cassandra	9
DataStax Enterprise – Certified Cassandra for Production Applications	11
Smartly Combining Real-Time Data With Analytics and Search	12
Hadoop Analytics.....	12
Search With Solr.....	12
A Complete Big Data Platform	13
Visual Database Management.....	14
Enterprise Production Support and Services	16
Application Use Cases.....	16
Conclusion.....	16
About DataStax	17

Introduction

Successful businesses recognize the value of capturing the massive volume of daily customer interactions – from purchased transactions to what products customers viewed – and analyzing that data to discover insights about their customers that help make smart business decisions. Achieving success in this area requires a union between distinct types of technology: a real-time database infrastructure that supports the operational data needs of the business, an analytical framework capable of handling the massively parallel analysis of that data, and an enterprise search ability to quickly search and mine that data.

The real-time side of modern data management has experienced a shift in direction in recent years. Traditional legacy relational database management system (RDBMS) technologies have not been able to keep up with the explosive growth of new data types requiring millisecond-time performance access on a scale that involves both large numbers of concurrent users and high data volumes. Whether the application is serving high-volume web session and user data, reacting to a high-speed financial market feed, aggregating distributed sensor grid events, processing social network messages and connections, or providing real-time intelligence and entity classification, it all comes down to being able to process, store, and respond to large data volumes as fast as possible.

Once such real-time data is stored, it is only natural for decision-makers to use it for analysis. But challenges such as mixed workload management (for instance, separating real-time and analytic operations on the same data), a distributed business and workforce, and the need to store and process extremely large sets of data have stymied even the best IT professionals trying to use legacy RDBMS software to squeeze the proverbial square peg into the round hole.

This paper examines these and other key data management challenges facing modern businesses. It also explains how DataStax Enterprise provides the first post-relational database solution to handle real-time, analytic, and search data in a way that solves these problems without the major compromises and costs associated with using RDBMS solutions.

A Look at Key Data Management Challenges

What are the primary challenges that successful businesses face when managing their growing data infrastructure? Although every company is different, a distinct pattern emerges with respect to data management problems.

Highly Available, High-Velocity Workloads

For the majority of applications at the core of modern businesses, gone are the days when systems only needed to be available between 9 a.m. and 5 p.m. Instead, around-the-clock availability has become the standard for databases that serve as the system of record (i.e., real-time data) for key production applications.

In addition, service level agreements (SLAs) for today's applications extend beyond just uptime and also concern themselves with response times. As web-based businesses know, the competition is just a click away, so the e-commerce experience must be one where product searches are accomplished very quickly, and the buying process supports a positive transactional experience.

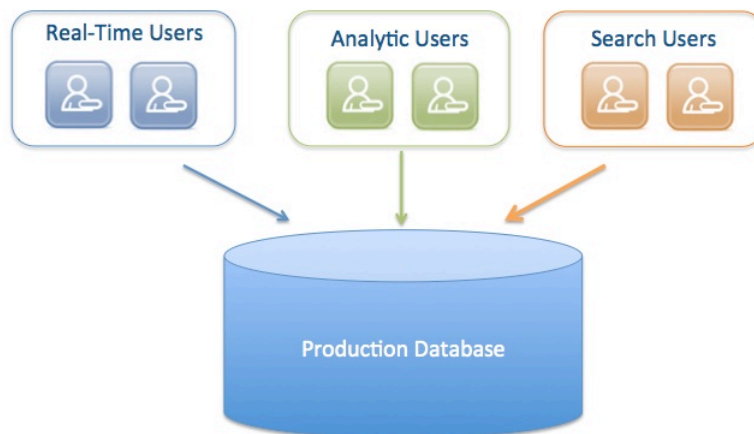
These benchmarks may be easily accomplished when there are not many concurrent users on a system and the underlying database is centrally located and not large. But that's not the situation for businesses experiencing tremendous growth in user traffic and data volume, whose services are extended to multiple geographic locations, and whose input involves device or sensor data (sometimes referred to as "data exhaust").

For these "high-velocity" workloads, one challenge is expanding capacity when needed to handle user and data growth. Moreover, the type of data needed both externally (by customers) and internally (by corporate staff) is real-time, analytic, and enterprise search in nature, with all data needing to be online 24x7.

This need produces the next major challenge for modern businesses: mixed workloads.

Mixed Workloads

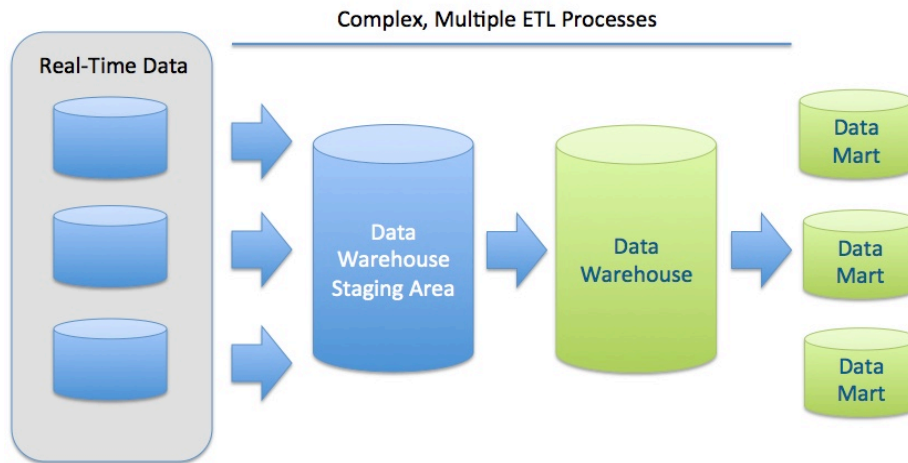
Industry analyst Gartner, Inc. identifies mixed-workload management (e.g., online transaction processing (OLTP) and analytics, batch/real-time analytics) among the top challenges faced by data management professionals. In addition, Gartner identifies mixed-workload management as a continuing issue for 2012.¹



¹ "Gartner Identifies Nine Key Data Warehousing Trends for the CIO in 2011 and 2012," Gartner, Inc., media release, February 2011: <http://www.gartner.com/it/page.jsp?id=1542914>.

Some vendors such as Oracle propose that businesses solve the mixed-workload challenge by purchasing a solution like Oracle Exadata. Exadata is a very powerful combined set of hardware and software; however, it also comes with a very high price tag. Such a purchase is either impossible or undesirable for many businesses.

Instead, many companies physically separate data such as real-time and analytic into different databases (often many different databases) to separate the distinct workloads.

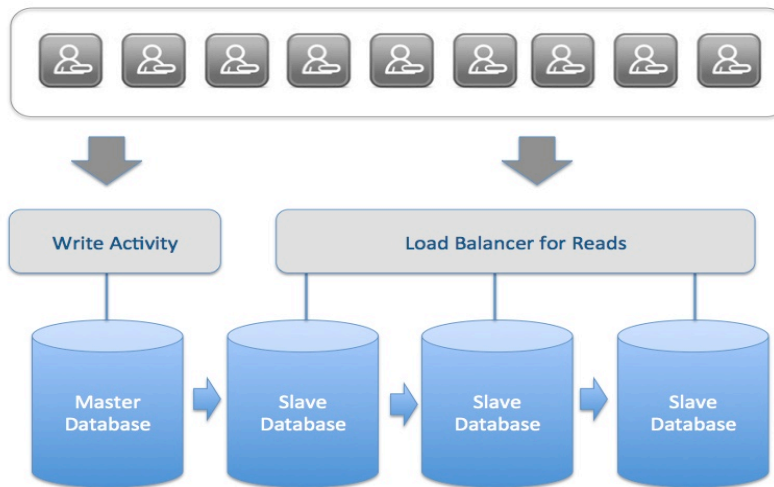


The glue that holds these infrastructures together are multiple, complex, extract, transform, and load (ETL) processes and sets of software designed to move data from real-time datastores into various databases designated for business intelligence and search operations. In many organizations, such implementations grow to the point where entire IT departments are devoted to ETL work, with ETL specialists pulled from the developer and database administrator ranks to staff these teams.

Scale, Big Data, and Architecture Limitations

It's no secret that a major challenge facing growing businesses is scaling their systems and managing the "big data" problem for real-time, analytic, and search data. It must be understood that big data does not simply equate to data warehousing. Rather, big data is defined by (1) the high-speed velocity at which data comes in; (2) the variety of different types of data (i.e., structured, semi-structured, and unstructured); (3) the volume of data being managed; and (4) the complexity of managing data across data centers, multiple geographies, and so on.

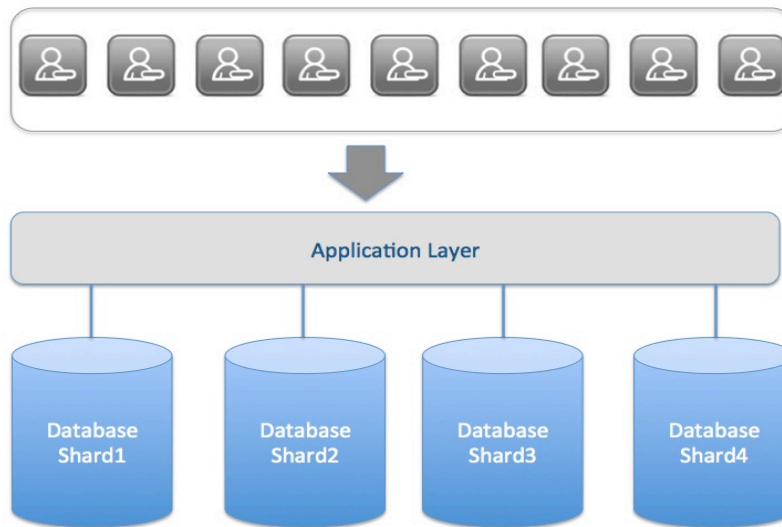
Legacy RDBMS systems are insufficient for managing big data, although there have been a number of attempts at making them work. The three most common methods involve master-slave replication, sharding, and intermediate caching layers put in front of RDBMSs.



While replication scale-out can work for moderately trafficked databases, the master-slave architecture suffers from well-known limitations. The master database quickly becomes a bottleneck for write-intensive workloads; latency between the master and slave servers presents problems (especially when many slave servers are added to the mix); and breakages in the replication chain (i.e., one or more slave servers stopped being replicated to) can frequently occur.

Perhaps the thorniest problem in master-slave architectures is when the master database server itself fails. While various techniques to failover to a slave server can be implemented, there is no easy way to switch back to the original master once it comes back online. Additionally, there is uncertainty as to whether all data from the master made it to the failed-over slave before the master went down.

To overcome traditional issues with master-slave architectures, IT professionals turned to sharding. With sharding, the front-end application programmatically implements a form of data partitioning to spread portions of a database across multiple servers manually. The idea is to split what would otherwise be an unmanageable database for one server among many machines in hopes of getting performance where it needs to be.



As many have discovered, manual sharding can become a data management nightmare. Reasons include the inability to easily expand existing shards or add new ones; the change control issues that exist with spreading a single schema across many databases (i.e., easily managing schema changes); constant updates to the application layer that controls the data partitioning; and the introduction of multiple points of failure for a logically partitioned database.

In addition to replication and sharding, some businesses have tried caching layers to better scale their systems, and there are some use cases that show such caching tiers can provide demonstrable performance benefits.

However, caching layers are not right for every situation and they bring management burdens of their own. As caching tiers have no data permanence, all write I/O must eventually find its way to the underlying RDBMS, so they become ineffective in heavy write situations. Furthermore, they become yet another set of software and layers for the IT staff to manage – as well as another cost added to the underlying application.

Widespread Data Distribution

Modern businesses increasingly need highly distributed databases that often span multiple data centers and geographic regions. As companies expand operations across the globe, they need to service all corporate locations from a data perspective and ensure all have fast access to data. This requirement is not confined to big data systems alone, but also can apply to much smaller databases.



Although replication has been a main feature in literally every legacy RDBMS, none offer a simple method for distributing data between different data centers (widely dispersed or otherwise) where performance isn't an issue. Furthermore, the issues of managing write activity among the various locations and reconciling it all together can become quite complex.

Lastly, simply moving to a cloud architecture does not overcome the above problems. Many cloud databases use some form of master-slave architecture and can deliver wildly different performance metrics if data is not smartly distributed among the geographical locations in the cloud provider's infrastructure.

Miscellaneous Challenges

Other data management challenges that modern businesses encounter include:

- **Fixed vs. dynamic schemas** – The legacy RDBMS vendor's fixed schema paradigm often does not accommodate an organization's need to easily manage structured, semi-structured, and unstructured data. In addition, making alterations to fixed schema designs can result in database objects being offline for the change (e.g., MySQL), or require considerable resources when large data volumes are present. Instead, organizations want dynamic schemas that offer the flexibility to manage all forms of data and provide ways to modify designs that don't involve downtime or impact overall performance.
- **Multiple vendors** – In trying to force legacy RDBMSs to conform to modern data management needs, companies are forced to engage with many different vendors to provide all the pieces to the infrastructure puzzle (e.g., a different RDBMS, ETL tools, load balancers, caching software, and management tools) that they try to put together.
- **High Cost** – Products like Oracle Exadata carry very high initial licensing costs and yearly maintenance fees. Even general purpose RDBMS software can be quite

expensive, and when combined with other components needed to manage the overall growing data infrastructure, the total price tag can be prohibitively high. Lastly, the management overhead of taking care of a complex data management framework, with respect to employee headcount and the specific expertise needed, should not be forgotten.

What Is the Answer?

For these reasons and others, experts agree that the concept of big data management (that involves real-time, analytic, and search data) equates to more than simply trying to retrofit legacy RDBMSs to tackle modern data-driven systems. This is why IDC defines “big data” as follows:

Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.²

IDC’s definition of big data incorporates all types of data (e.g., real-time, analytic) managed by systems that must scale to handle constantly increasing user workloads and data volume. Such capabilities are found in the primary offering from DataStax: DataStax Enterprise.

What Is DataStax Enterprise?

DataStax is the leading provider of enterprise NoSQL software products and services based on Apache Cassandra™. Through its offerings, DataStax supports businesses that need a progressive data management system that can serve as a primary system of record/real-time datastore for critical production applications, and delivers built-in analytic and search capabilities for analyzing and searching that data once it is in Cassandra.

DataStax Enterprise inherits all of Cassandra’s powerful feature set (described below) for servicing modern real-time applications, and uses it to merge in a fault-tolerant, analytics, and enterprise search platform that provides Hadoop MapReduce, Hive, and Pig support for analytics and uses Apache Solr for fast enterprise search.

A key differentiator of DataStax Enterprise is that real-time, analytic, and search operations are smartly separated across a DataStax Enterprise cluster so that no competition for underlying compute resources or data is encountered.

² *Extracting Value from Chaos*, by John Gantz and David Reinsel, IDC, June 2011, <http://idcdocserv.com/1142>.

DataStax Enterprise – Powered by Apache Cassandra

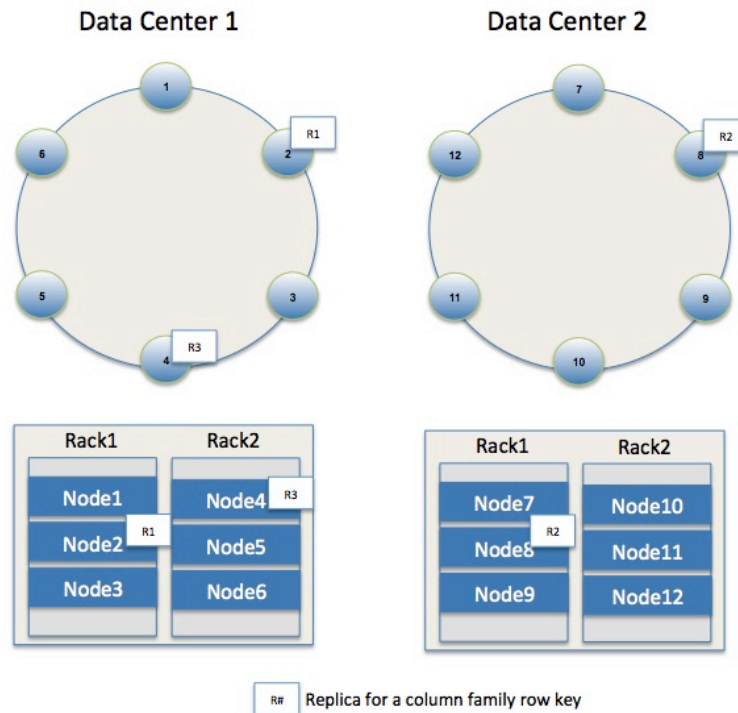
The foundation of DataStax Enterprise is Apache Cassandra. Cassandra is a highly scalable and high-performance distributed database management system that can serve as both a real-time database (the “system of record”) for online/transactional applications, and as a read-intensive datastore for business intelligence systems.

Key technical differentiators of Cassandra over its RDBMS predecessors, as well as other NoSQL offerings, include the following:

- A built-for-scale architecture that can handle petabytes of information and thousands of concurrent users/operations per second as easily as it can handle much smaller amounts of data and user traffic
- Peer-to-peer design that offers no single point of failure for any database process or function
- Online capacity additions that deliver linear performance gains for both read and write operations
- Read/write anywhere capabilities that equate to a true network-independent method of storing and accessing data
- Tunable data consistency that allows Cassandra to offer comparable data durability and protection like an RDBMS, but with the flexible choice of relaxing that consistency when application use cases allow
- Flexible/dynamic schema design that accommodates structured, semi-structured, and unstructured data; data is represented in Cassandra via column families that are dynamic in nature and accommodate all modifications online
- Simplified replication that provides data redundancy and is capable of being multi-data center and cloud in nature
- Data compression that reduces the footprint of raw data by over 80 percent in some use cases
- A SQL-like language (CQL) that lessens the learning curve for developers and administrators coming from the RDBMS world
- Support for key developer languages (e.g., Java) and operating systems
- No requirement for any special equipment; Cassandra runs on commodity hardware

Cassandra is built with the assumption that failures can and will occur in a database infrastructure. Therefore, data redundancy to protect against hardware failure and other data loss scenarios is built into and managed transparently by Cassandra. Furthermore, this capability can be configured to be quite sophisticated so that data in a single cluster can be distributed across multiple, geographically dispersed data centers, between different physical racks in a data center, and between public cloud providers and on-premise managed data centers.

These and other capabilities make Cassandra and DataStax Enterprise the smart choice for modern businesses that have outgrown their RDBMS software, and are looking for a better way to store and access their real-time data.



DataStax Enterprise – Certified Cassandra for Production Applications

Cassandra is a top open source project for the Apache foundation and enjoys strong community support and developer involvement. New community releases and patches are produced very quickly, with the understanding that community builds are not put through any enterprise-styled quality assurance process, and often contain a mixture of enhancements plus bug fixes.

By contrast, DataStax Enterprise only contains selected Cassandra releases that are chosen by the expert staff and committers at DataStax. Each selected release is then put through a rigorous certification process designed by DataStax engineers and quality assurance (QA) staff to ensure it is stable and ready for enterprise production systems. Any found issues are immediately fixed and applied to the DataStax Enterprise server.

In addition, DataStax also provides enterprises with predictable, certified quarterly service pack updates as well as other software benefits such as emergency hot fixes (for production outages) and bug escalation privileges that prioritize customers' issues over community-submitted bugs.

Smartly Combining Real-Time Data With Analytics and Search

A primary benefit that DataStax Enterprise provides to enterprises needing smart big data management capabilities is its ability to service real-time, analytic, and enterprise search data operations in the same database cluster without any of the loads impacting the other. The key to making this possible is the underlying architecture of Cassandra.

Hadoop Analytics

Built into DataStax Enterprise is an enhanced Hadoop distribution that utilizes Cassandra for many of its core services. DataStax Enterprise provides integrated Hadoop MapReduce, Hive, Pig, and job/task tracking capabilities, replacing Hadoop's HDFS storage layer with Cassandra (CassandraFS). The end product is a single integrated solution that provides increased reliability, simpler deployment, and lower total cost of ownership (TCO) than a traditional Hadoop solution. DataStax Enterprise also is fully compatible with existing HDFS, Hadoop, and Hive tools and utilities.

Another benefit of using Hadoop in DataStax Enterprise is that it eliminates the complexity and single points of failure of the typical Hadoop HDFS layer. From an operational standpoint, there is no need to set up a Hadoop name node, secondary name node, Zookeeper, and so on.

Instead, DataStax Enterprise provides a single layer in which every node is a peer of the others and automatically knows its position in the cluster. On startup, all DataStax Enterprise nodes automatically start a Hadoop task tracker, and one of the nodes is elected to be the job tracker. If the job tracker node fails, the job tracker is automatically restarted on a different node. DataStax Enterprise utilizes full data locality awareness for Hadoop task assignment.

Search With Solr

DataStax Enterprise includes strong enterprise search support via Lucene and Apache Solr. Coming from the Apache Lucene project, Solr is the most popular open source enterprise search platform in use today.

Solr's primary features include robust full-text search, hit highlighting, faceted search, rich document (e.g., PDF, Microsoft Word) handling, and geospatial search.

By integrating Solr into the DataStax Enterprise big data platform, DataStax extends Solr's capabilities and overcomes a number of shortcomings that native Solr has by delivering the following:

- An easily scalable search platform
- 100% data durability
- No single point of failure
- No write bottleneck as in community Solr

- Automatic data sharding
- Multi-data center capabilities
- Easy, ad-hoc index rebuilds
- The ability to query search data with Cassandra's CQL

In essence, in the same way that DataStax Enterprise takes Hadoop and delivers a fault-tolerant, no single point of failure, and dynamically scalable Hadoop/analytics system, it automatically does the same thing for Solr and enterprise search operations. Using Cassandra as the underlying foundation, DataStax Enterprise allows search data to be written to any participating search node in a DataStax Enterprise cluster. New search nodes can be added online to increase both fault tolerance and performance, with gains being near linear in nature.

Those currently using Solr will be at home with DataStax Enterprise, as the solution is 100 percent Solr compatible, with all Solr utilities, APIs, and so on, included.

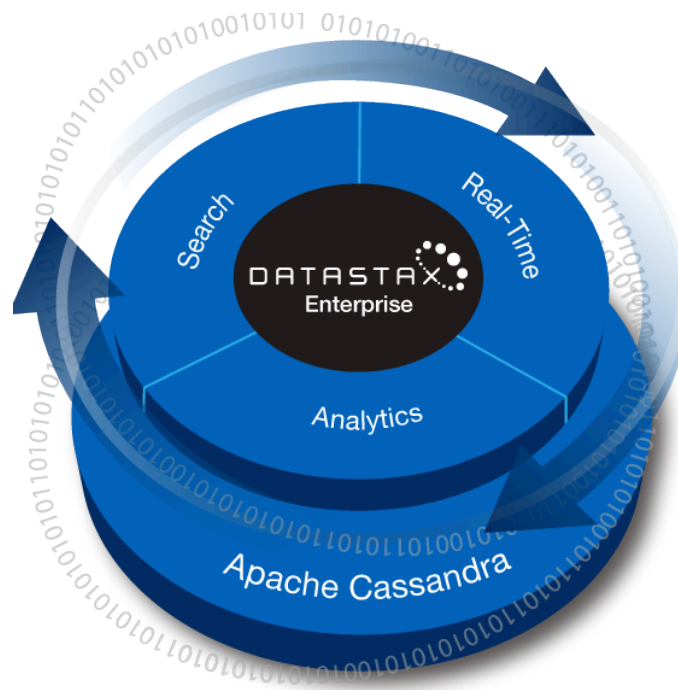
A Complete Big Data Platform

A key benefit of DataStax Enterprise is the tight feedback loop it has between real-time applications and the analytics and search operations that naturally follow. Traditionally, users would be forced to move data between systems via complex ETL processes, or perform both functions on the same system with the risk of one impacting the other. In big data environments, this process can be time-consuming and burdensome.

With DataStax Enterprise, real-time, analytic, and search big data operations take place in the same distributed system, but users have the ability to dedicate certain nodes solely for analytics or search so their workloads don't slow down real-time processing.

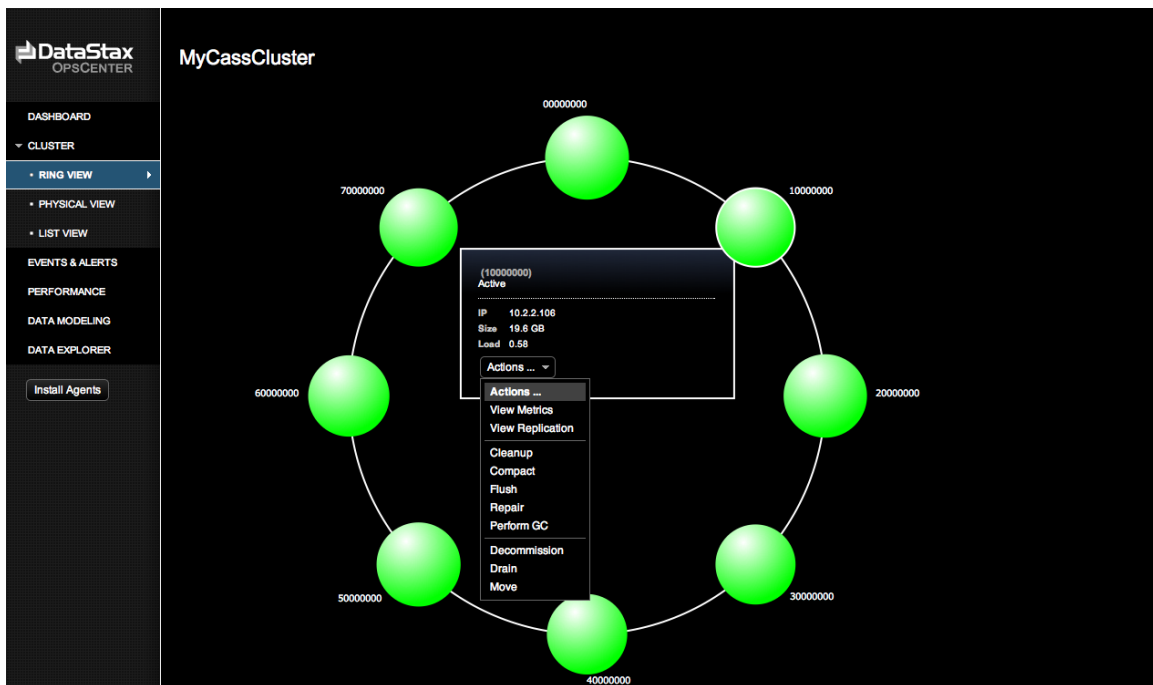
Users simply define one or more replica groups, and configure the role of each – one or more Cassandra, Hadoop, or HDFS (i.e., HDFS without job/task tracker), and search/Solr nodes. Writes are instantly replicated between all nodes.

With DataStax Enterprise, users truly have the best of all worlds for big data management. They have all the power of Cassandra serving their highest-volume and high-velocity, real-time applications; the power of Hadoop, Hive, and Pig working directly against the same data for analytics; and Solr for enterprise search in the same distributed database. The result is smart workload isolation for big data application, which is much simpler to manage and more reliable than any of the alternatives.

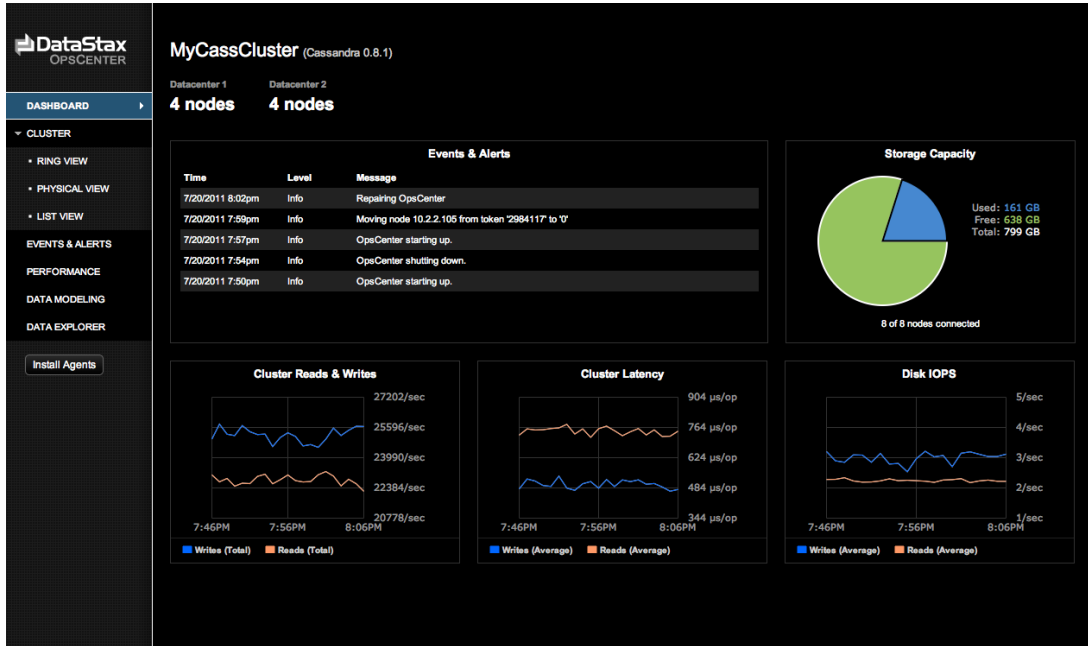


Visual Database Management

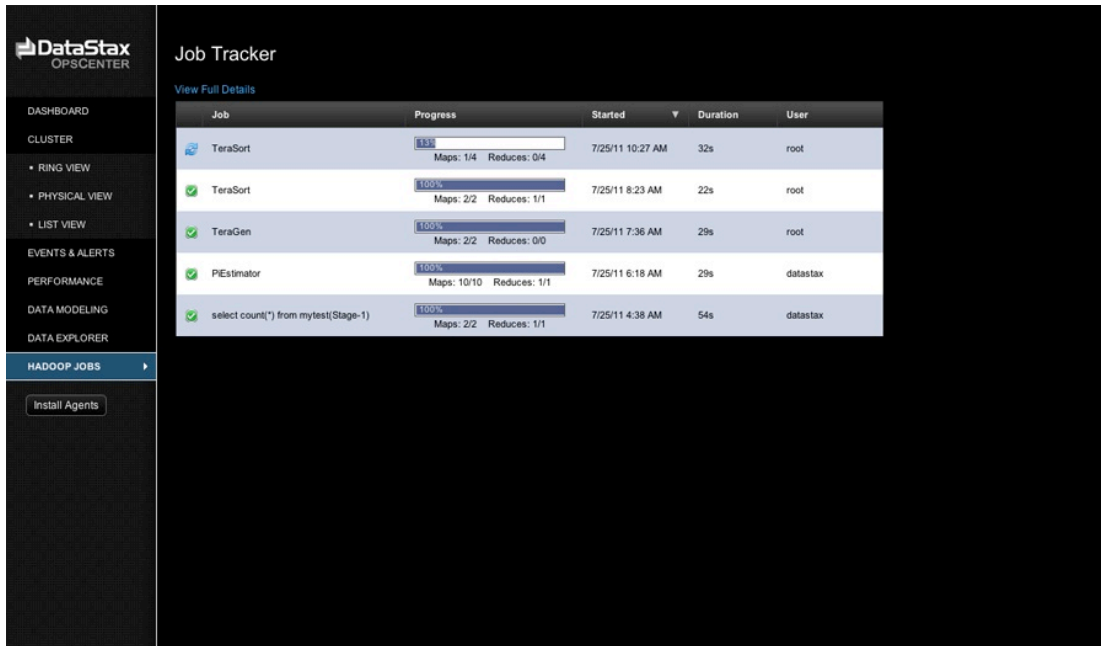
DataStax Enterprise includes a visual, browser-based management solution called OpsCenter Enterprise. OpsCenter Enterprise allows a developer or administrator to manage and monitor the health of an entire DataStax Enterprise cluster from a centralized web console.



OpsCenter Enterprise uses an agent-based architecture to monitor and carry out tasks on each node in a DataStax Enterprise cluster. Through an intuitive point-and-click interface, a user can understand the state of a cluster, which nodes are up and down, and what type of performance users are experiencing. Key events are reported into a centralized dashboard displayed along with other vital statistics.



Hadoop analytic operations also can be monitored and controlled from within OpsCenter Enterprise:



Enterprise Production Support and Services

DataStax Enterprise includes professional production support and services from the Cassandra experts. Customers can choose the right production support package for their business needs, including rapid response SLAs and consultative help.

DataStax also provides certified quarterly service pack updates for DataStax Enterprise as well as other benefits such as emergency hot fixes (for production outages) and bug escalation privileges for customers.

Additionally, DataStax offers professional training on Cassandra and Hadoop, with classes offered in many major cities and on-site for corporations that need many staff members trained at once.

Application Use Cases

Because of its progressive architecture and built-in capabilities, DataStax Enterprise excels in the following application use cases:

- Time series data management
- High-velocity device data ingestion and analysis
- Media streaming (e.g., music, movies)
- Social media input and analysis
- Online web retail (e.g., shopping carts, user transactions)
- Web log management
- Web click-stream analysis
- Real-time data analytics
- Online gaming (e.g., real-time messaging)
- Write-intensive transaction systems
- Buyer event analytics
- Fraud detection and analysis
- Risk analysis and management
- Web product searches

Conclusion

DataStax Enterprise supplies a progressive and modern data management framework that overcomes the key data management challenges documented in the beginning of this paper.

For organizations needing a highly available database that supports high-velocity workloads (i.e., many users and large data volumes), DataStax Enterprise's architecture was designed from the ground up to offer the highest possible availability and meet the most stringent performance requirements of modern systems.

As for handling the mixed-workload problem, DataStax Enterprise supports real-time, analytic, and search workloads within the same database cluster and isolates all workloads so none competes with the others for compute resources or data.

When it comes to scaling a successful system and accommodating big data, DataStax Enterprise's scale-out architecture comfortably scales from gigabytes to petabytes, while offering high performance no matter the data volume. Moreover, DataStax Enterprise does not rely on outdated master-slave architectures and manually shared frameworks, which are difficult to manage and maintain; or intermediate caching software. Instead, it uses a smart peer-to-peer design tailor-made for today's data-driven applications.

With respect to handling tough data distribution requirements, DataStax Enterprise's underlying Cassandra foundation makes it easy to replicate and distribute data across multiple data centers, multiple geographies, and different cloud provider zones. It also handles hybrid on-premise and cloud implementations. Additionally, DataStax Enterprise's dynamic schema design makes it easy to change a database's underlying schema and have those changes replicate to all nodes in a cluster in an online fashion.

Lastly, rather than dealing with many different IT vendors and high cost, DataStax Enterprise supplies all the functionality a growing business needs for its data infrastructure at a cost that is a fraction of what traditional RDBMS vendors charge.

To find out more about DataStax Enterprise and to download software, please visit www.datastax.com or email info@datastax.com.

About DataStax

DataStax offers products and services based on the popular open-source database, Apache Cassandra™ that solve today's most challenging big data problems. DataStax Enterprise combines the performance of Cassandra with analytics powered by Apache Hadoop and enterprise search with Apache Solr, creating a smartly integrated, big data platform. With DataStax Enterprise, real-time, analytic, and search workloads never conflict, giving you maximum performance with the added benefit of only managing a single database.

The company has over 140 customers, including leaders such as Netflix, Disney, Cisco, Rackspace and Constant Contact, and spans verticals including web, financial services, telecommunications, logistics and government. DataStax is backed by industry leading investors, including Lightspeed Venture Partners and Crosslink Capital and is based in San Mateo, CA.

For more information, visit www.datastax.com.