DataStax

Big Data: Beyond the Hype

Why Big Data Matters to You

WHITE PAPER

Big Data:
Beyond the Hype

Why Big Data Matters to You

By DataStax Corporation

October 2011

DataStax

Table of Contents

Introduction

"Big data" is a big buzz phrase in the IT and business world right now – and there are a dizzying array of opinions on just what these two simple words really mean.

Technology vendors in the legacy database or data warehouse spaces say "big data" simply refers to a traditional data warehousing scenario involving data volumes in either the single or multi-terabyte range. Others disagree: They say "big data" isn't limited to traditional data warehouse situations, but includes real-time or operational data stores used as the primary data foundation for online applications that power key external or internal business systems.

It used to be that these transactional/real-time databases were typically "pruned" so they could be manageable from a data volume standpoint. Their most recent or "hot" data stayed in the database, and older information was archived to a data warehouse via extract-transform-load (ETL) routines.

But big data has changed dramatically. The evolution of the web has redefined:

- The speed at which information flows into these primary online systems
- The number of customers a company must deal with
- The acceptable interval between the time that data first enters a system, and its transformation into information that can be analyzed to make key business decisions

Some analysts such as Forrester Research and others have attempted to categorize these changes by what is being called "the four V's" of big data:

1. Volume – TB's to PB's of data
2. Velocity – how fast the data is coming in
3. Variety – all types are now being captured (structured, semi-structured, unstructured)
4. Value – mining the valuable pieces of data from among data that does not matter

Because of these changes, new definitions for big data have been proposed, with a focus on technologies to handle such data. Analyst firms such as IDC say legacy RDBMSs designed to run moderately sized data volumes on single machines do not offer sufficiently powerful engines for the big data scenarios with which modern businesses are wrestling.

> **Here's how IDC defines "big data":**
>
> *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.*[1]

This definition incorporates all types of data (e.g., real-time, analytic) managed by next-generation systems that must scale to handle constantly increasing user workloads and data volume.

David Kellogg, meanwhile, simply defines big data as being "too big to be reasonably handled by current/traditional technologies."[2] Consulting and research firm McKinsey & Company agrees with Kellogg's concept of big data and defines it as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."[3]

This paper examines the growing prevalence of big data across nearly every industry; explains why being good at using and understanding big data is critical for firms that want to compete in their chosen market; and details how businesses can use DataStax Enterprise—a solution specifically designed to manage big data easily and effectively—to exploit the benefits derived from handling big data smartly.

---

[1] *Extracting Value from Chaos*, IDC: http://idcdocserv.com/1142

[2] "Big data has jumped the shark," DBMS 2: www.dbms2.com/2011/09/11/big-data-has-jumped-the-shark/

[3] *Big Data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011, p. 11: www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf
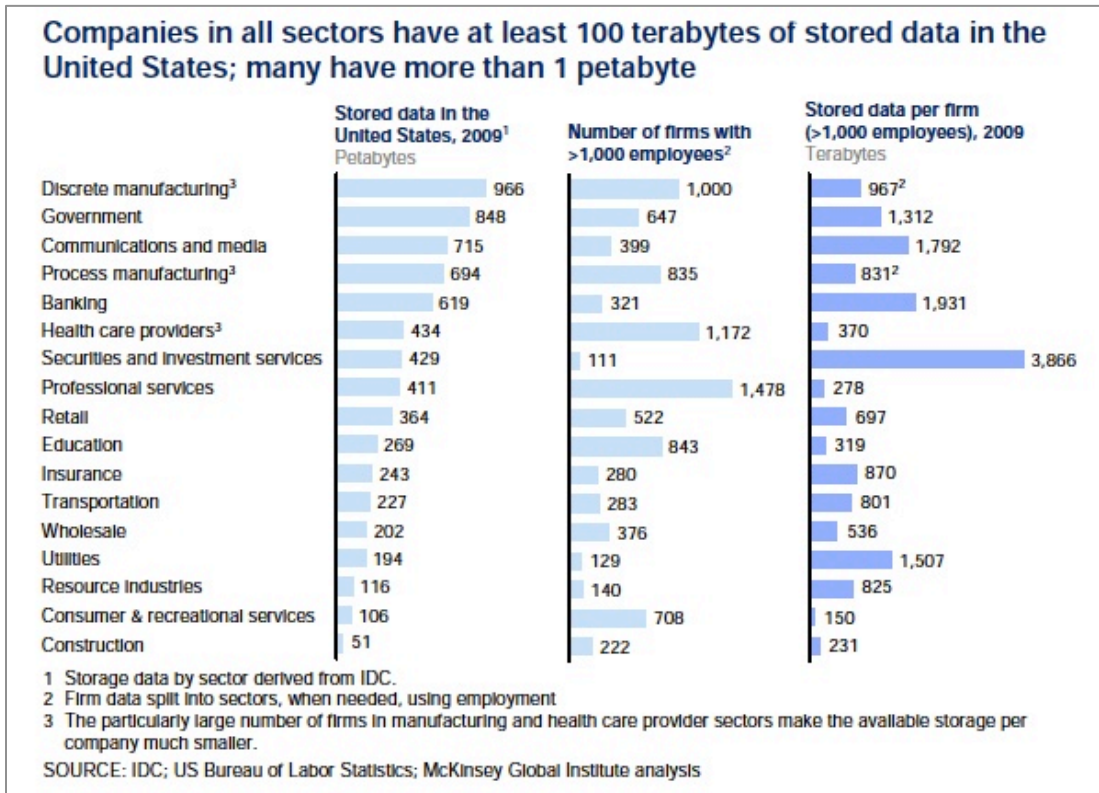
Big Data and You

How much should you care about effectively managing big data? A lot – in fact you're likely already dealing with big data without even knowing it.

Big Data Is More Prevalent than You Think

Many businesses believe big data is something only companies like Facebook and Google deal with. However, a 2011 McKinsey Global Institute study says otherwise.

For instance, the McKinsey report found that the average investment firm with fewer than 1,000 employees has 3.8 petabytes of data stored, experiences a data growth rate of 40 percent per year, and stores structured, semi-structured, and unstructured data.[4] Overall, McKinsey found that 15 out of 17 industry sectors in the United States have more data stored per company than the U.S. Library of Congress (which had 235 terabytes of information at the time of McKinsey's study)[5] and that companies in all sectors have at least 100 terabytes stored[6], as Figure 1 shows:

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte

	Stored data in the United States, 2009[1] Petabytes	Number of firms with >1,000 employees[2]	Stored data per firm (>1,000 employees), 2009 Terabytes
Discrete manufacturing[3]	966	1,000	967[2]
Government	848	647	1,312
Communications and media	715	399	1,792
Process manufacturing[3]	694	835	831[2]
Banking	619	321	1,931
Health care providers[3]	434	1,172	370
Securities and investment services	429	111	3,866
Professional services	411	1,478	278
Retail	364	522	697
Education	269	843	319
Insurance	243	280	870
Transportation	227	283	801
Wholesale	202	376	536
Utilities	194	129	1,507
Resource industries	116	140	825
Consumer & recreational services	106	708	150
Construction	51	222	231

1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

---

[4] McKinsey, p. 19.

[5] McKinsey, p. vi.

Figure 1: Stored data by industry sector

These data volumes are not confined to enterprise data warehouses that assist only internal decision makers, but instead exist in the real-time database systems that serve external-facing customers. And that data continues to grow and expand as the underlying business becomes more successful. In 2010, the United States alone stored more than 3,500 petabytes of new information,[7] as shown in Figure 2:
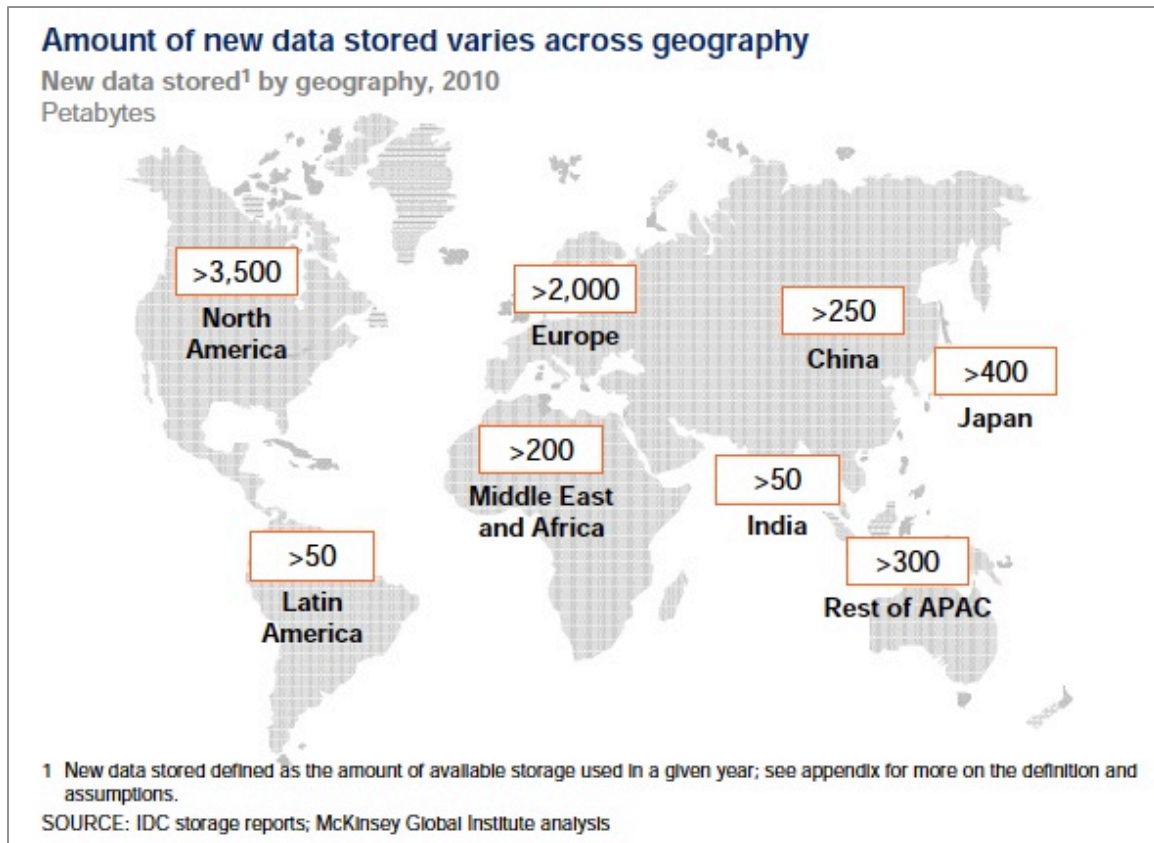

Figure 2: Data stores by geography

---

[6] McKinsey, p. 19.

[7] McKinsey, p. 103.

Big Data Formats

Part of the need for new technologies for big data (versus older, legacy RDBMSs) has to do with the format of the data coming in from online applications. A more dynamic, flexible database schema format is needed to handle the structured, semi-structured, and unstructured data that comprises today's big data[8] (see Figure 3):
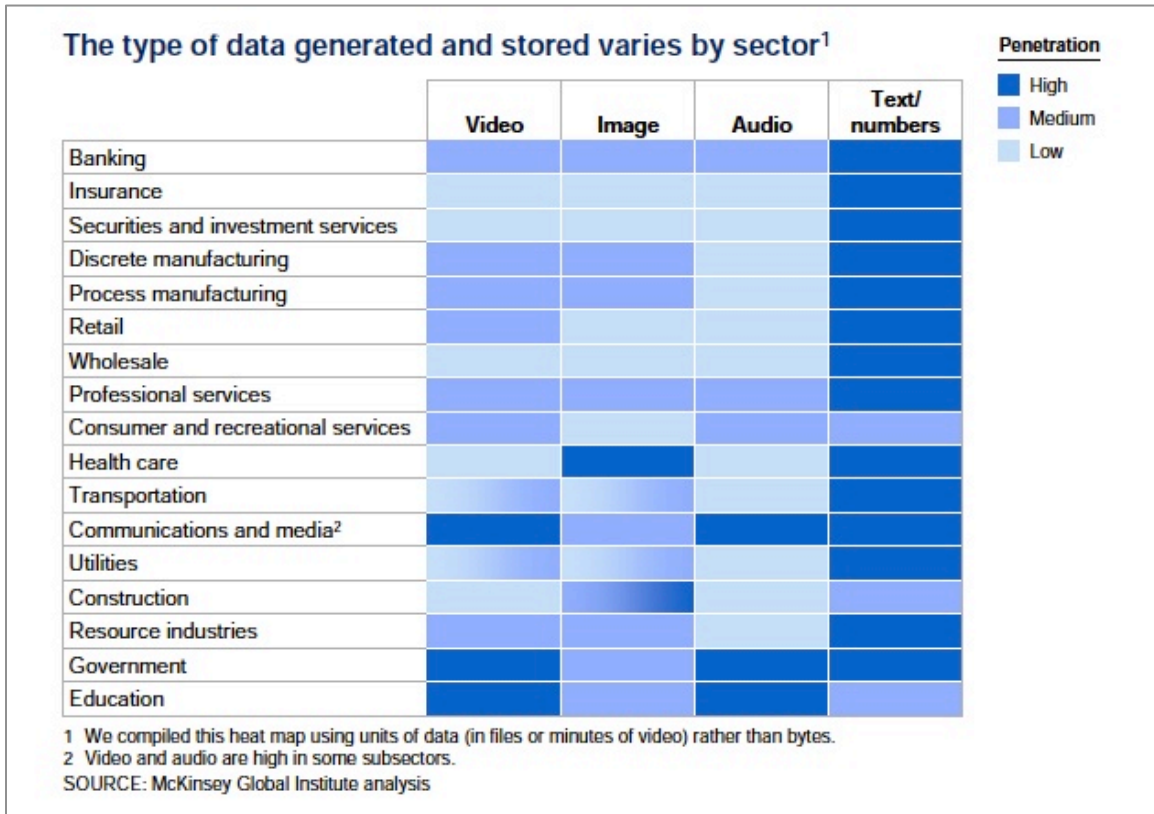
The type of data generated and stored varies by sector[1]

	Video	Image	Audio	Text/ numbers
Banking	Medium	Medium	Low	High
Insurance	Low	Low	Low	High
Securities and investment services	Low	Low	Low	High
Discrete manufacturing	Medium	Medium	Low	High
Process manufacturing	Medium	Medium	Low	High
Retail	Medium	Low	Low	High
Wholesale	Medium	Medium	Low	High
Professional services	Medium	Medium	Medium	High
Consumer and recreational services	Medium	Low	Medium	Medium
Health care	Low	High	Low	High
Transportation	Medium	Low	Low	High
Communications and media[2]	High	Medium	High	Medium
Utilities	Low	Low	Low	High
Construction	Low	Medium	Low	High
Resource industries	Medium	Medium	Low	High
Government	High	Medium	High	High
Education	High	Medium	High	Medium

Penetration
- High
- Medium
- Low

1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

Figure 3: Data stores by sector

---

[8] McKinsey, p. 20.

Competitive Advantages Gained Through Big Data

Today, no one today argues with the fact that a company's data is its most strategic impersonal asset. If you don't use it as a competitive weapon against your market competitors, it's guaranteed that you will be at a disadvantage. In a presentation given at the Strata New York conference in September 2011, McKinsey & Company showed the eye opening, 10-year category growth rate differences (see Figure 4, below) between businesses that smartly use their big data and those that do not.
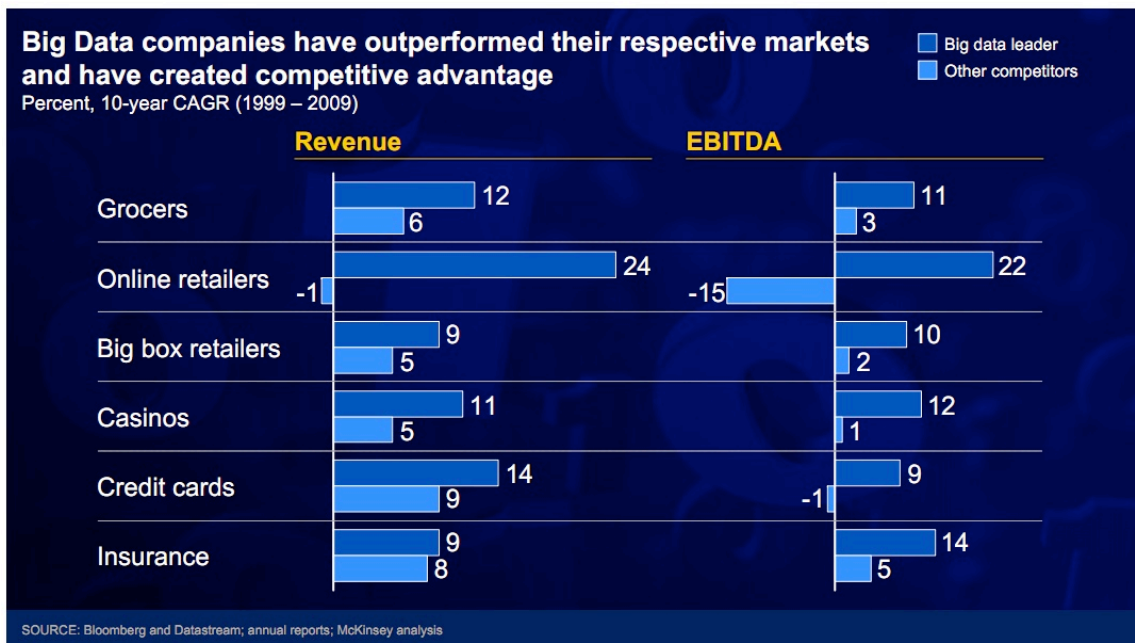
Figure 4: Big data companies have a very real competitive advantage

Clearly, the biggest differences exist between online retailers who don't use big data well, and those that do. And today, online retailing is a business every company is in, whether they're 100 percent web-based or not. Forrester Research predicts that by 2013, over half of all U.S. sales will be online in nature[9] (see Figure 5):

---

[9] McKinsey, p. 66.
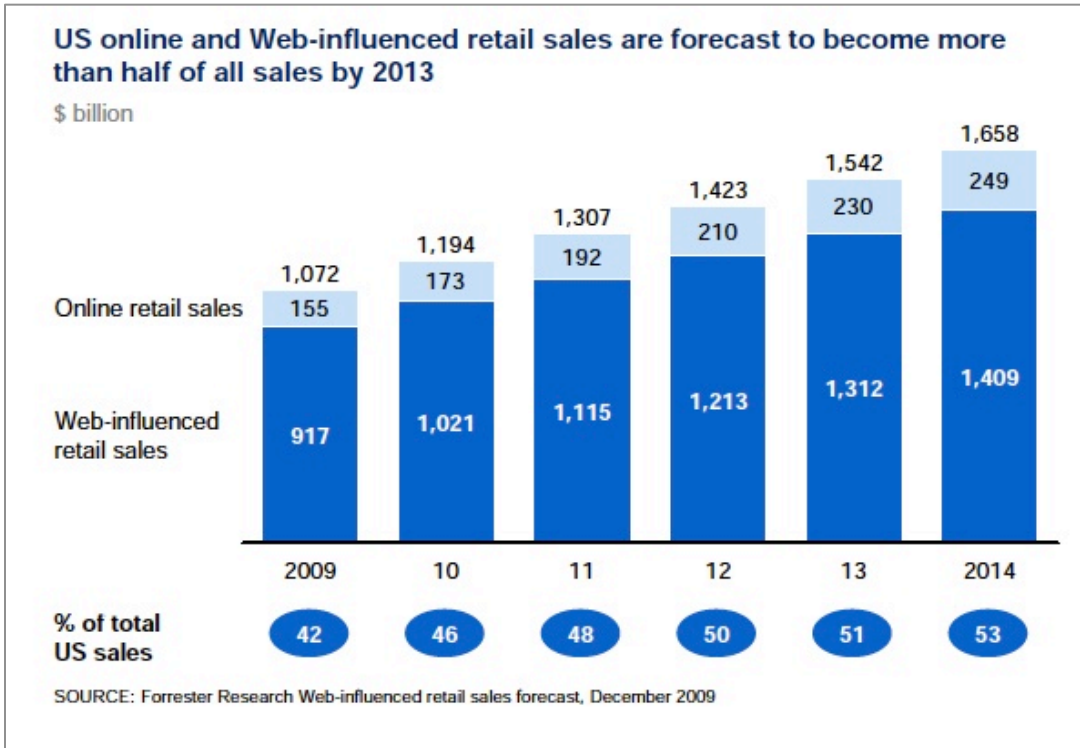
Figure 5: Online retail sales growth

The recognition of these market realities has led smart businesses to concentrate on effectively utilizing their big data. This is reflected in a strong jobs growth trend, as illustrated in the chart below from Indeed.com:
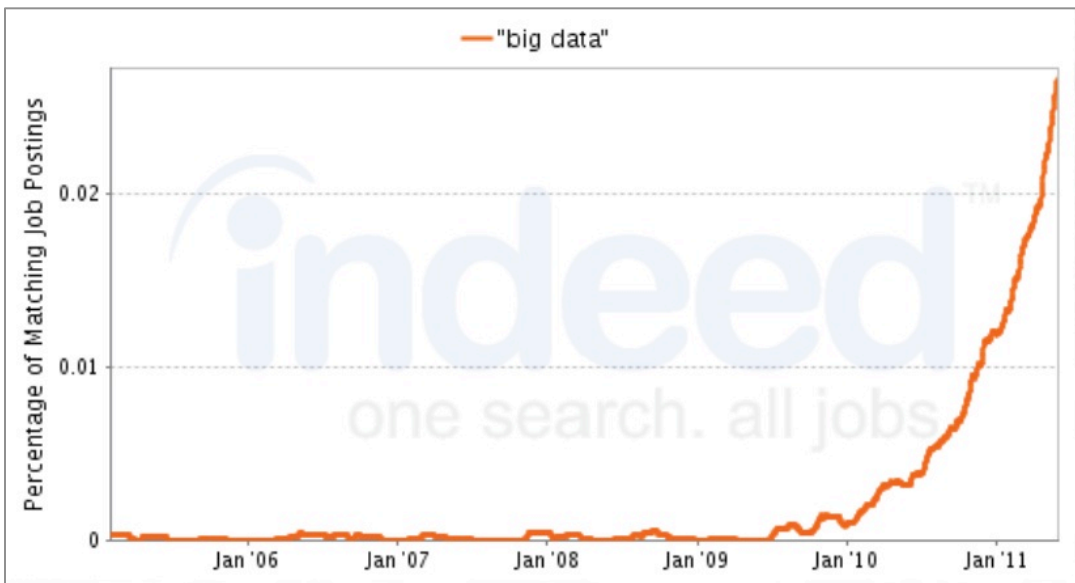

Figure: 6: Big data job postings

Once data professionals are hired and put to work, there are many different data-driven projects they can be assigned to. As to the types of processes that can benefit from big data efforts, McKinsey found many different activities across all core internal corporate functions can provide value to a modern business through the use of big data[10] (see in Figure 7):

Big data retail levers can be grouped by function

Function	Big data lever
Marketing	• Cross-selling • Location based marketing • In-store behavior analysis • Customer micro-segmentation • Sentiment analysis • Enhancing the multichannel consumer experience
Merchandising	• Assortment optimization • Pricing optimization • Placement and design optimization
Operations	• Performance transparency • Labor inputs optimization
Supply chain	• Inventory management • Distribution and logistics optimization • Informing supplier negotiations
New business models	• Price comparison services • Web-based markets

SOURCE: McKinsey Global Institute analysis

Figure 7: Big data retail levers

Now What?

The facts about big data really do speak for themselves. The nonstop growth of data, new data formats that must be managed, and the competitive advantages that come from managing big data well all underscore why big data should matter to you.

---

[10] McKinsey, p. 67.

But what should you do about it? If you're an IT professional, you know how difficult it can be to find a solution capable of handling a task like big data management that combines the following benefits:

- Scalability
- Performance
- Ease of use
- Low total cost of ownership

Some database offerings may have two of these four features, but it's rare to find one with all four – and that delivers on them well. The good news is there *is* a solution available that confidently provides checkmarks for the four criteria above.

DataStax Enterprise:
The Best Solution for Managing Big Data

DataStax is the leading provider of modern enterprise database software products and services based on Apache Cassandra™. It 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 also deliver built-in analytic capabilities for analyzing that data once it's in Cassandra.

DataStax Enterprise is tailor-made to manage big data effectively. The solution inherits all of Cassandra's powerful feature set for servicing modern real-time applications, and merges in a fault-tolerant, analytics platform that provides Hadoop MapReduce, Hive, and Pig support for business intelligence systems.

A key differentiator of DataStax Enterprise over other big data providers is that real-time and analytic workloads are smartly separated across a distributed DataStax Enterprise database cluster, so that no competition for underlying compute resources or data occurs.

Powered by Apache Cassandra

The foundation that enables DataStax Enterprise to tackle big data is Apache Cassandra™. Cassandra enjoys an industry reputation for being the only NoSQL database solution able to truly handle big data requirements. It's a highly scalable and high-performance distributed database management system that can handle real-time big data applications that drive key systems for modern and successful businesses.

Key technical differentiators of Cassandra versus its legacy RDBMS predecessors, as well as other NoSQL offerings, include:

- A built-for-scale architecture that can handle petabytes of information and thousands of concurrent users/operations per second as easily as it can manage 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 the data durability and protection like an RDBMS, but with the flexible choice of relaxing data consistency when application use cases allow
- Flexible/dynamic schema design that accommodates all formats of big data applications, including 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 big data by over 80 percent in some use cases
- An 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 big data 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 so that big data applications can use a single large database that is 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.
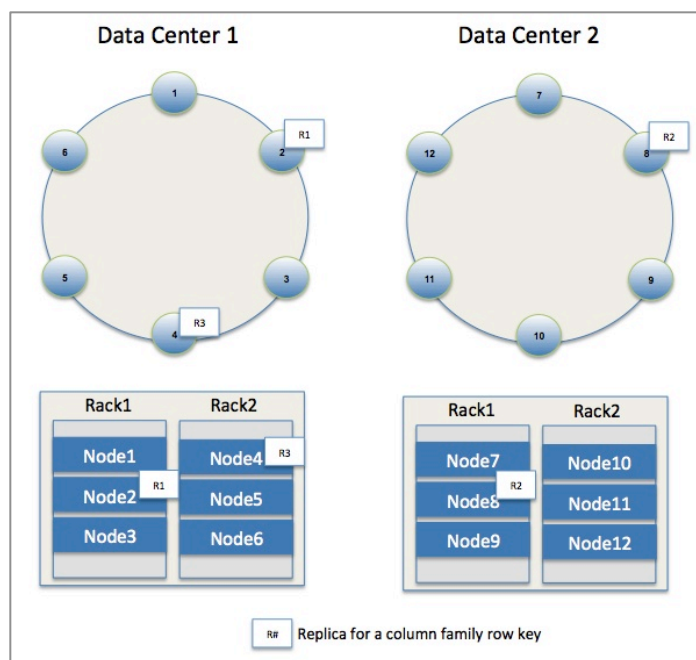
Figure 8: Cassandra multi-data center capabilities

These and other capabilities make Cassandra and DataStax Enterprise the smart choice for modern businesses whose big data management needs have outgrown their traditional RDBMS software.

Big Data Analytics

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

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 TCO than a traditional Hadoop solution. DataStax Enterprise also is fully compatible with existing HDFS, Hadoop, and Hive tools and utilities.

Another benefit of 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.
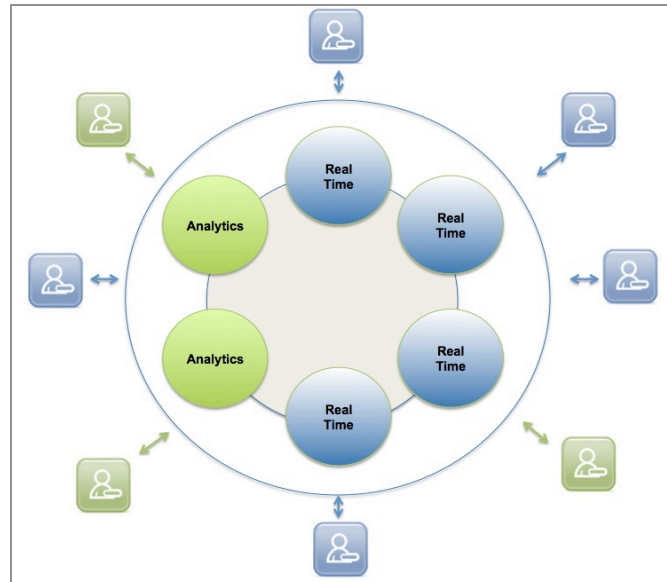

Figure 9: DataStax Enterprise – real-time and analytic data in one database

Another key benefit of DataStax Enterprise is the tight feedback loop it has between real-time application and the analytics 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, both real-time and analytic big data operations take place in the same distributed system, but users have the ability to dedicate certain nodes solely for analytics so that workload doesn'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) nodes. Writes are instantly replicated between both real-time and analytic nodes.

With DataStax Enterprise, users truly have the best of both worlds for big data management. They have all the power of Cassandra serving their highest-volume and high-velocity real-time applications, and the power of Hadoop, Hive, and Pig working directly against the same data for analytics. The result is smart workload management 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 named OpsCenter Enterprise. OpsCenter Enterprise allows a developer or administrator to manage and monitor the health of big data clusters from a centralized web console.
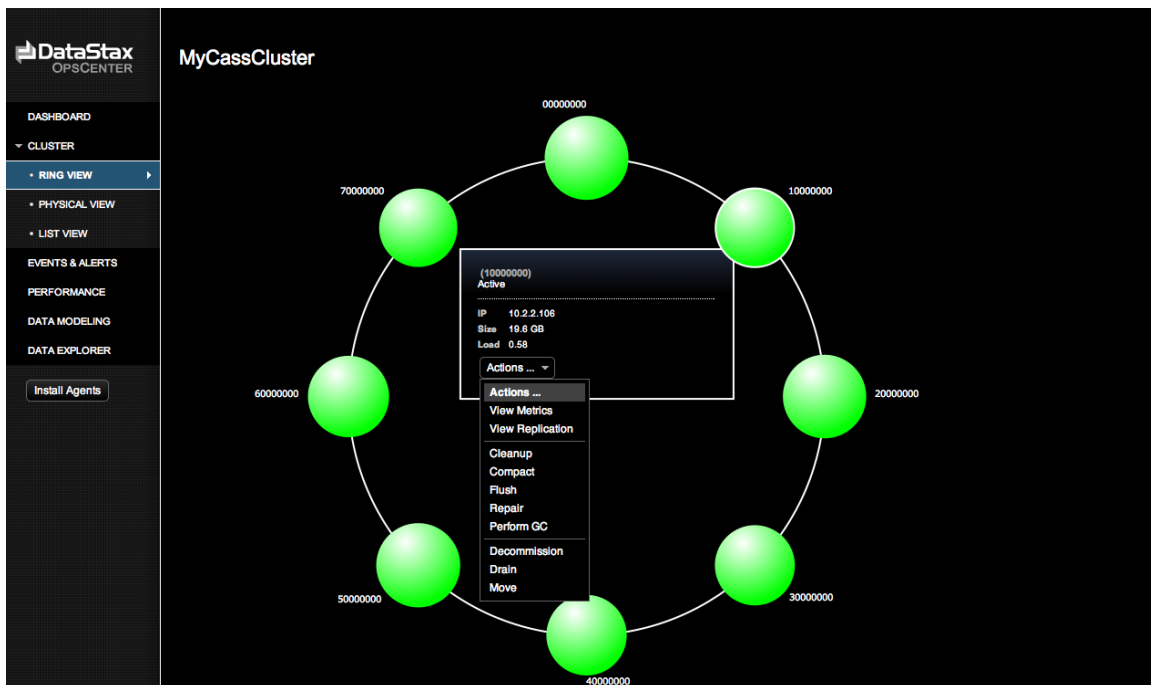

Figure 10: OpsCenter Enterprise database cluster ring view

OpsCenter Enterprise uses an agent-based architecture to monitor and carry out tasks on each node in a DataStax Enterprise cluster. Through a graphical and 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.
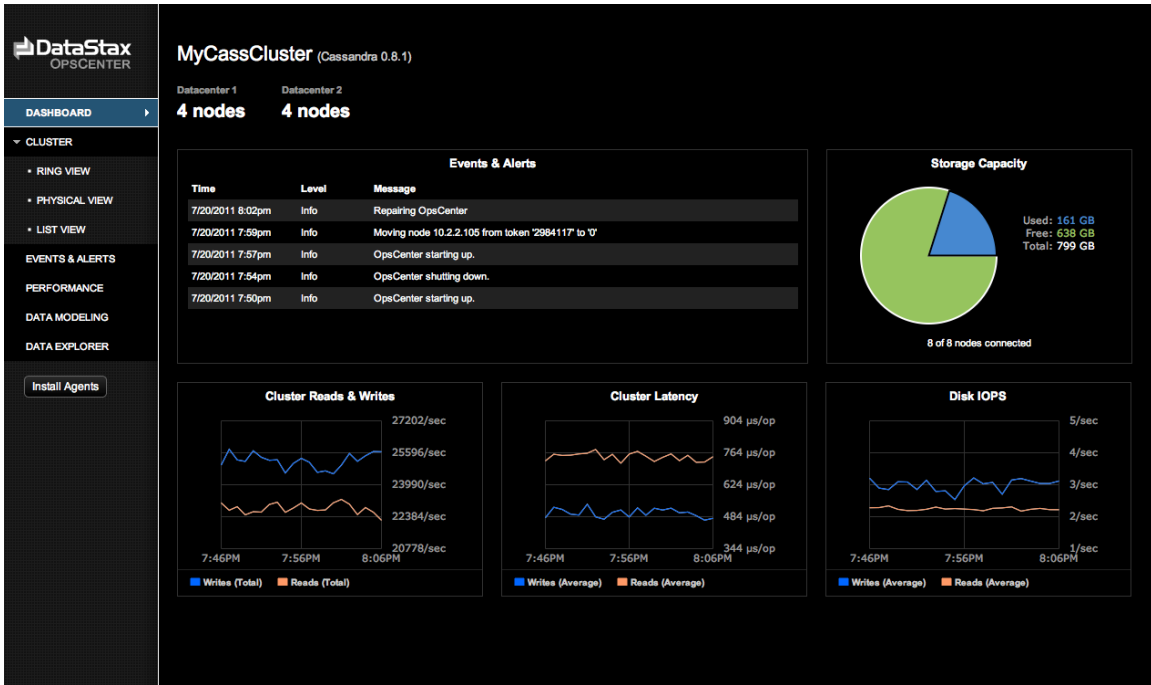
Figure 11: OpsCenter dashboard

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

Figure 12: OpsCenter analytic operations monitoring

Enterprise Production Support and Services

Big data situations often require fast access to skilled expertise. DataStax Enterprise includes experienced production support and consultative services from Cassandra experts. You can

choose the right production support package for your 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.

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

Conclusion

Big data isn't just hype—and it's much more than a buzz phrase. Today, companies across industries are finding they not only need to manage increasingly large data volumes in their real-time systems, but also analyze that information so they can make the right decisions, fast, to compete effectively in the market.

Modern businesses looking for a solution to handle their big data easily and effectively should consider DataStax Enterprise. Its scale-out architecture comfortably scales into the petabytes data range, while offering high performance for reads and writes no matter the data volume size.

DataStax Enterprise also differs from competitors by providing a single integrated database platform that smartly manages both real-time and analytic data. Additionally, DataStax Enterprise does all of this at a fraction of the cost charged by traditional RDBMS vendors.

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

About DataStax

DataStax is the developer of DataStax Enterprise, a distributed, scalable, and highly available database platform that delivers optimal performance either on premise or in the cloud for modern enterprise applications that manage both real-time and analytic workloads. The company has over 100 customers, including leaders such as Netflix, Cisco, Rackspace and Constant Contact, and spanning 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 Burlingame, CA with offices in Austin, TX and Stamford, CT.

For more information, visit www.datastax.com.