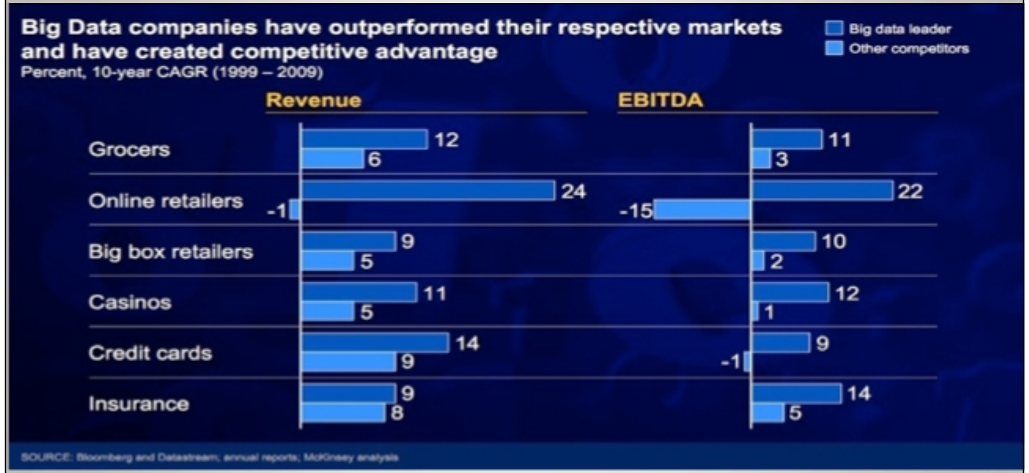
DATASTAX

Data Velocity *and* Continuous Availability

Michael Shaler
Senior Director, Business Development

What is Big Data's payoff?

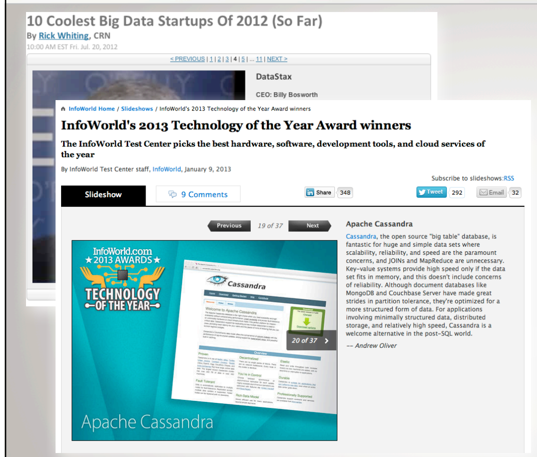


Big Data companies have outperformed their respective markets and have created competitive advantage

Percent, 10-year CAGR (1999 – 2009)

SOURCE: Bloomberg and Datastream; annual reports; McKinsey analysis

DataStax: CRN's "10 Coolest Big Data Startups of 2012 (So Far)"
Cassandra: InfoWorld's 2013 Technology of the Year Award Winner



10 Coolest Big Data Startups Of 2012 (So Far)
By Rick Whiting, CRN
10:00 AM EST Fri. Jul. 20, 2012

< PREVIOUS | 1 | 2 | 3 | 4 | 5 | ... 11 | NEXT >

DataStax

CEO: Billy Bosworth

InfoWorld Home / Slideshows / InfoWorld's 2013 Technology of the Year Award winners

InfoWorld's 2013 Technology of the Year Award winners

The InfoWorld Test Center picks the best hardware, software, development tools, and cloud services of the year

By InfoWorld Test Center staff, InfoWorld, January 9, 2013

Subscribe to slideshows:RSS

Apache Cassandra

Cassandra, the open source "big table" database, is fantastic for huge and simple data sets where scalability, reliability, and speed are the paramount concerns, and JOINs and MapReduce are unnecessary. Key-value systems provide high speed only if the data set fits in memory, and this doesn't include concerns of reliability. Although document databases like MongoDB and Couchbase Server have made great strides in partition tolerance, they're optimized for a more structured form of data. For applications involving minimally structured data, distributed storage, and relatively high speed, Cassandra is a welcome alternative in the post-SQL world.

-- Andrew Oliver

LIGHTSPEED
VENTURE PARTNERS

CROSSLINK CAPITAL

MERITECH
CAPITAL PARTNERS

1,000+ production deployments and 300 customers, including
~~23~~ 24 of the Fortune 100

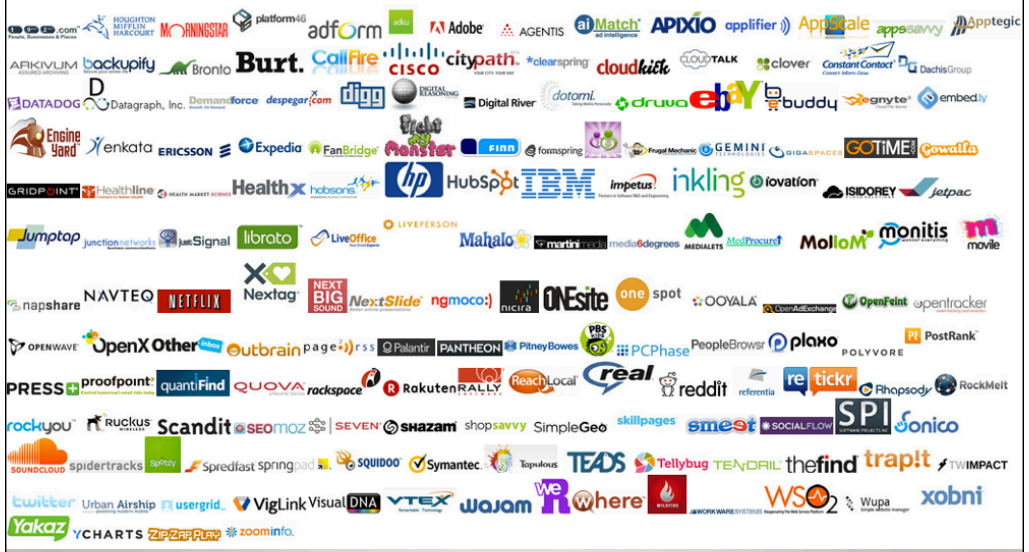
DATASTAX

Documented Use Cases

accenture
Constant Contact
NASA
ebay
NETFLIX
HealthCare ANYTIME
Adobe
BARRACUDA NETWORKS
OOYALA
comcast

DATASTAX

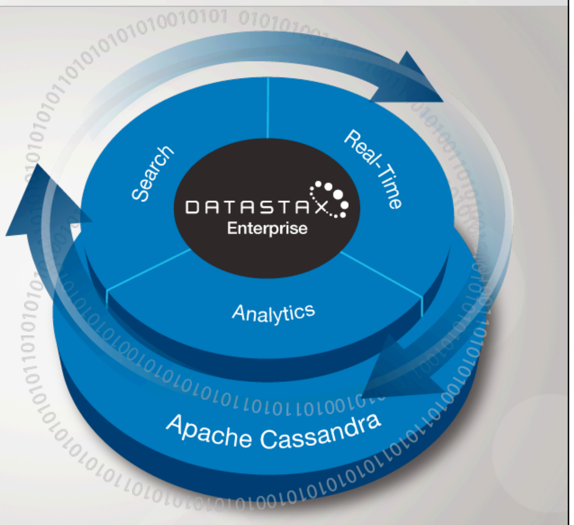
Real Growth In Production

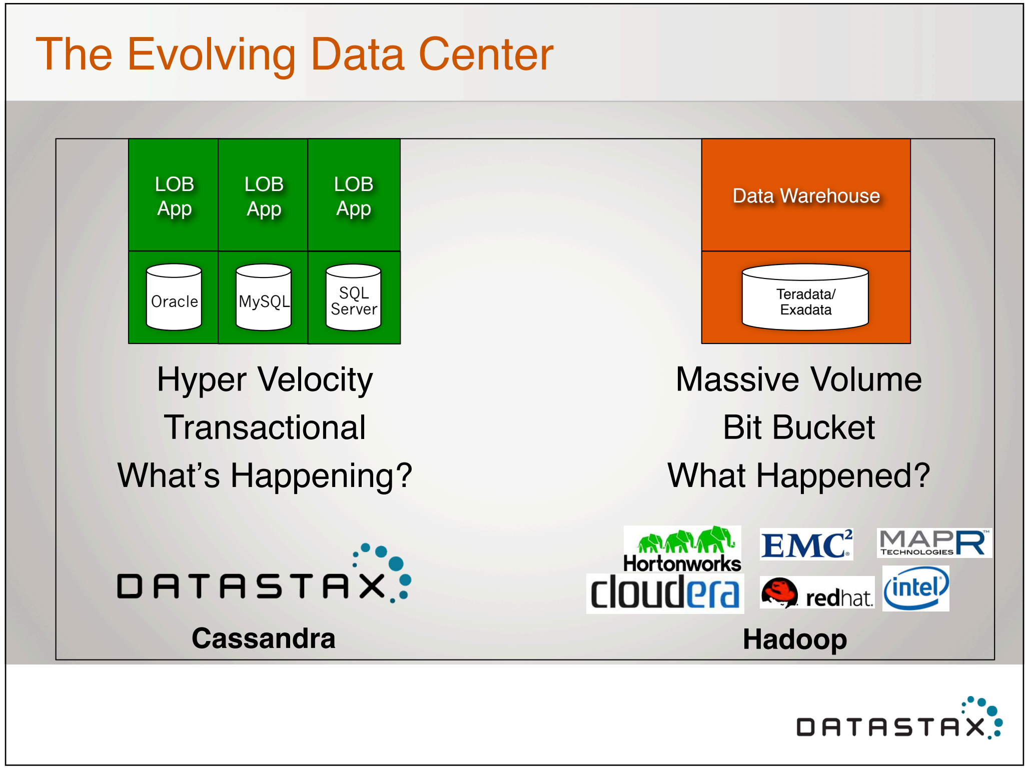


Our Solution

- DataStax Enterprise powers the big data apps that transform business.

- Velocity at Scale
- Continuous Availability





The Evolving Data Center

LOB App | LOB App | LOB App

Oracle | MySQL | SQL Server

Hyper Velocity

Transactional

What's Happening?

DATASTAX

Cassandra

Data Warehouse

Teradata/ Exadata

Massive Volume

Bit Bucket

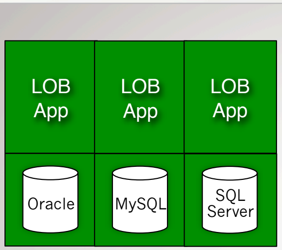
What Happened?

Hadoop

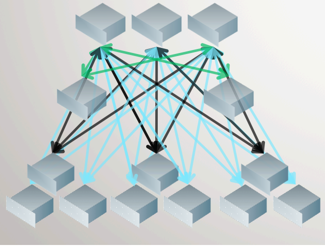
Put the Cloudera logo on the bottom of the their stack and us on ours. Put the technology names under it.

Follow it up with the case studies.

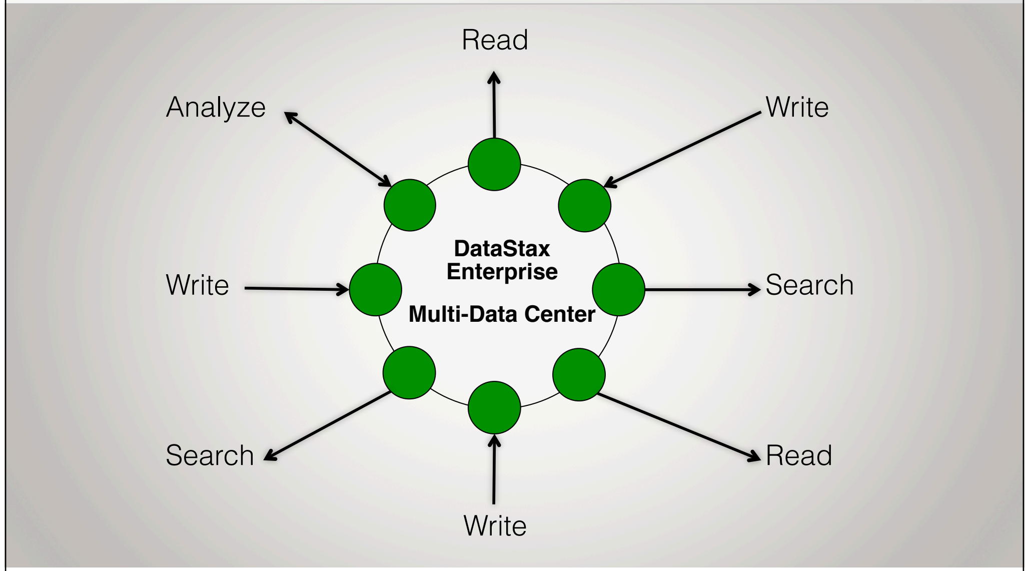
We Tried… Oh, How We Tried.



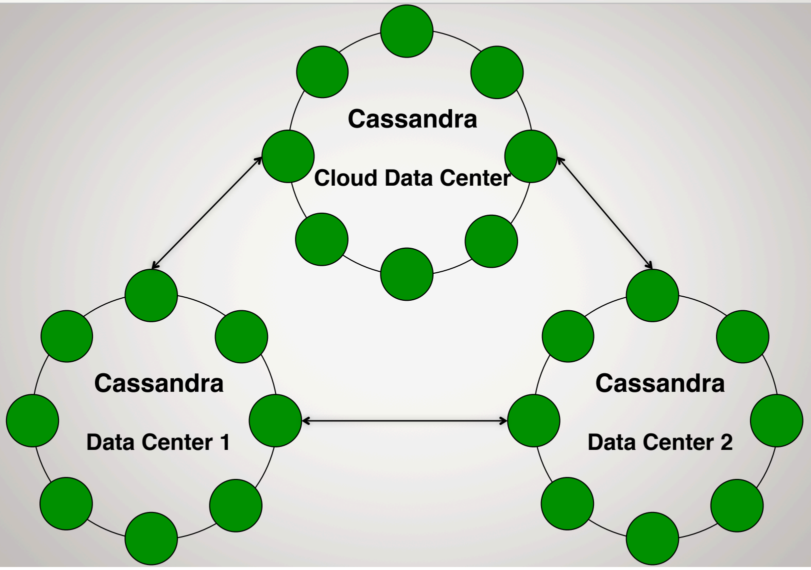
RDBMS Sharding



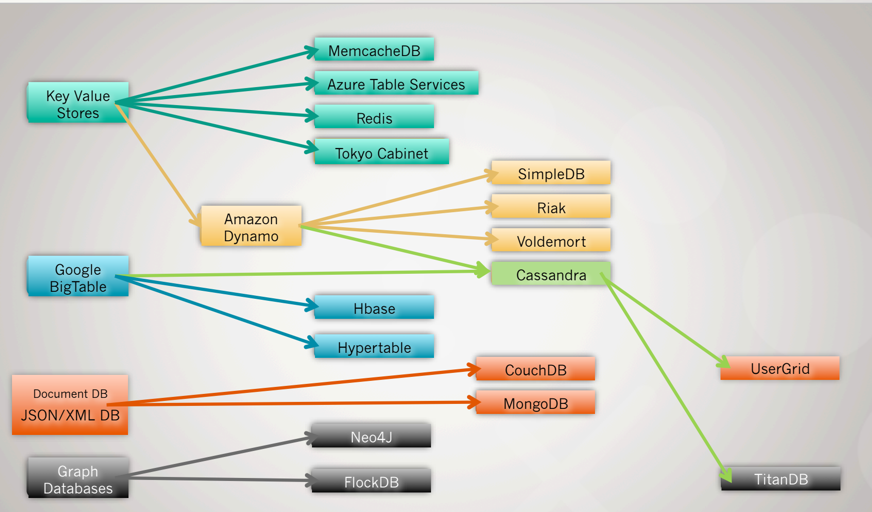
LOB Application Demands



The New Online Architecture



Open Source Database Pedigrees

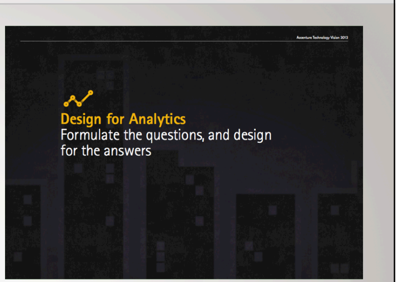


* Courtesy of @GuyHarrison

Hadoop Use Cases

Variety

Data factory



Accenture Technology Vision 2013

Design for Analytics
Formulate the questions, and design
for the answers

Volume

Data reservoir

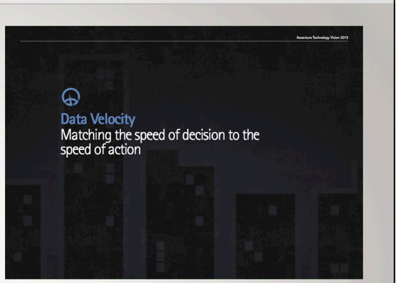
Accenture Technology Vision 2013

Relationships at Scale
Moving beyond transactions
to digital relationships

Cassandra Use Cases

Velocity

Data Velocity at Scale



Accenture Technology Vision 2013

Data Velocity
Matching the speed of decision to the speed of action

Complexity

Continuously available

Accenture Technology Vision 2013

Relationships at Scale
Moving beyond transactions to digital relationships

Paul Maritz: “We need to bring consumer-grade back to the enterprise.”

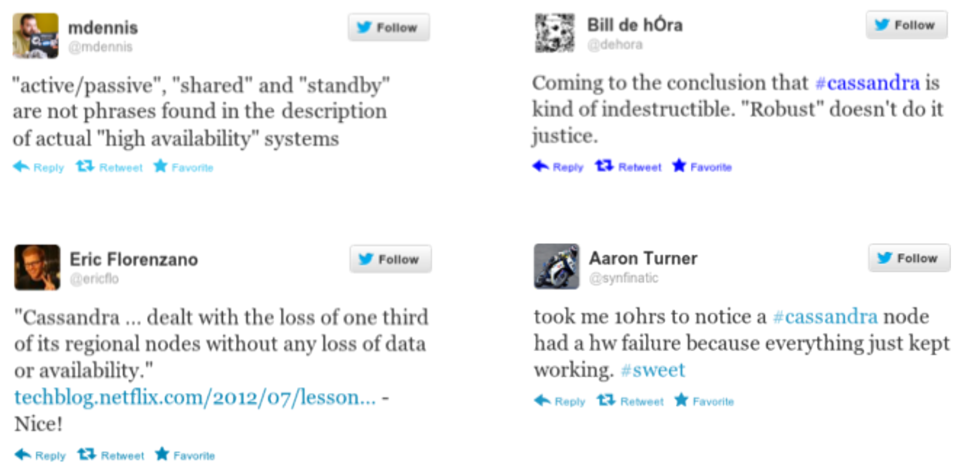
Common Use Cases

- **Cassandra**
  - Big data OLTP and write intensive systems
  - Time series data management
  - High velocity device data consumption and analysis
  - Healthcare systems input and analysis
  - Media streaming (music, movies, etc.)
  - Online Web retail (shopping carts, user transactions, etc.)
  - Online gaming (real-time messaging, etc.)
  - Real time data analytics
- **Hadoop**
  - Social media input and analysis
  - Web click-stream analysis
  - Buyer event and behavior analytics
  - Fraud detection and analysis
  - Risk analysis and management
  - Supply chain analytics
- **Solr**
  - Web product searches
  - Internal document search (law firms, etc.)
  - Real estate/property searches
  - Social media match ups
  - Web & application log management / analysis

Cassandra as Foundation

Benefit	Feature
Fully Distributed: no SPOF	Peer-to-peer architecture for continuous availability and operational simplicity
Multi-Datacenter	Node-, rack- and DC-aware with tunable consistency
Massively Scalable	Multiple customers > 10M writes/second
SSD and Cloud optimized	All writes are linear, and all files are immutable
Rich Application Data Model	CQL (no joins or 2PC), integration with ODBC/JDBC et al

Continuous Availability Commentary



mdennis
@mdennis

"active/passive", "shared" and "standby" are not phrases found in the description of actual "high availability" systems

Bill de hÓra
@dehora

Coming to the conclusion that #cassandra is kind of indestructible. "Robust" doesn't do it justice.

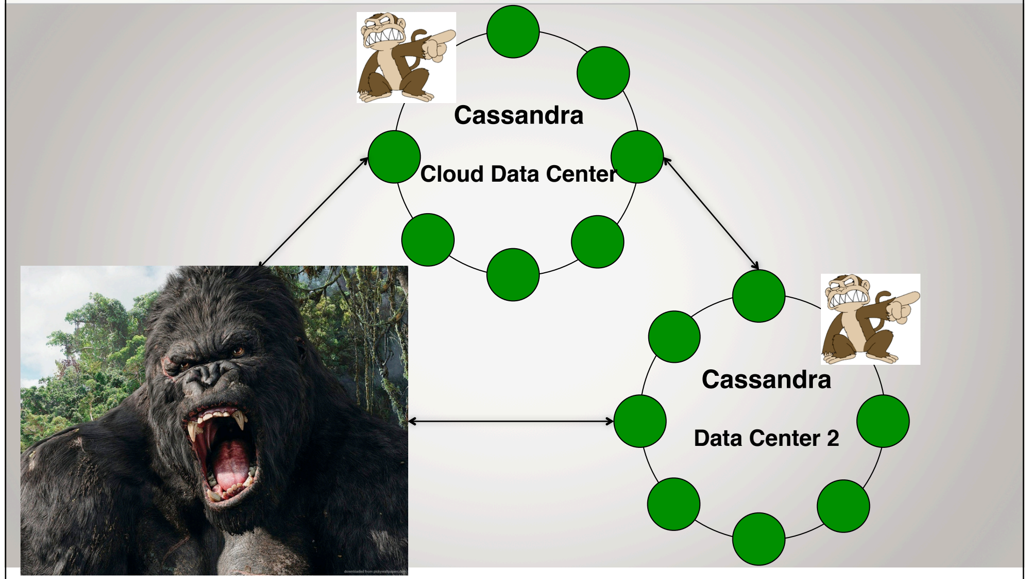
Eric Florenzano
@ericflo

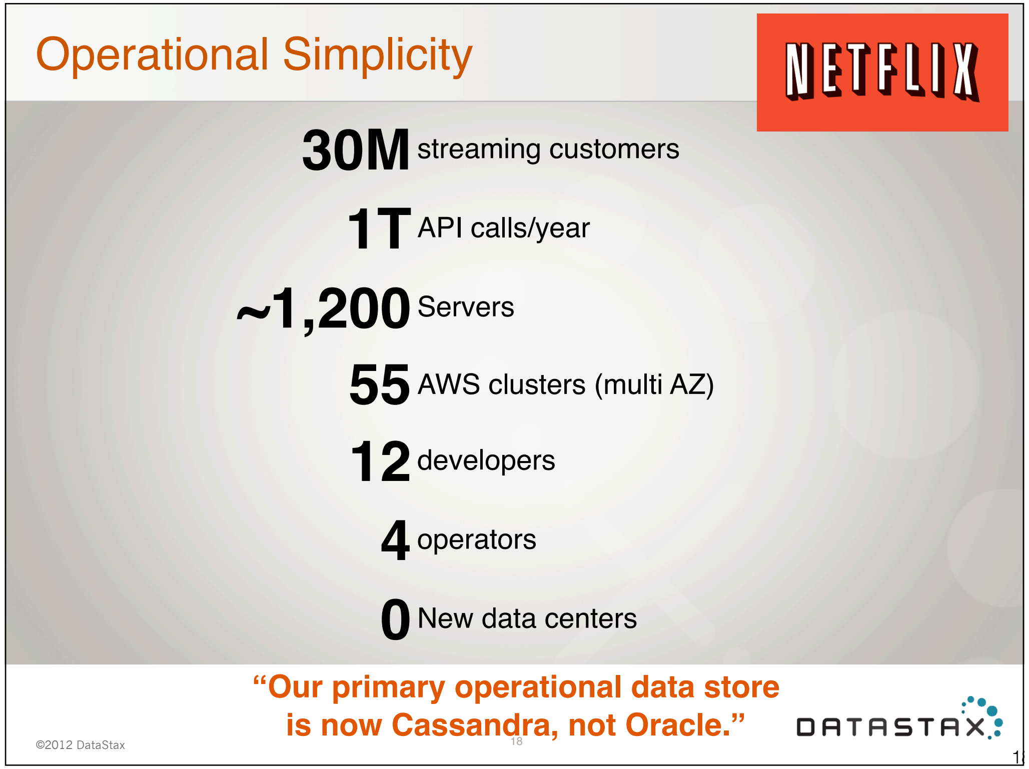
"Cassandra ... dealt with the loss of one third of its regional nodes without any loss of data or availability." techblog.netflix.com/2012/07/lesson... - Nice!

Aaron Turner
@synfinatic

took me 10hrs to notice a #cassandra node had a hw failure because everything just kept working. #sweet

The New DR: Simian Army “Dystopia as a Service”





Operational Simplicity

30M streaming customers

1T API calls/year

~1,200 Servers

55 AWS clusters (multi AZ)

12 developers

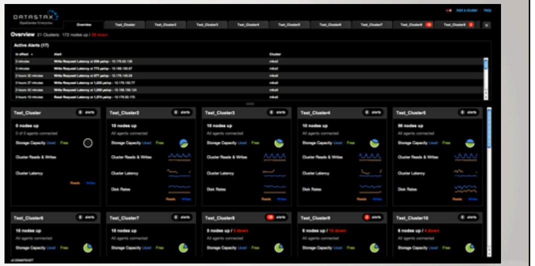
4 operators

0 New data centers

“Our primary operational data store is now Cassandra, not Oracle.”

Time Series Analytics: 70B readings

Smart Grid Proof of Concept: Analyze 2 years of Smart Meter data for 1M households
Improvements in demand forecasting could yield EBITDA > $100M per GW saved

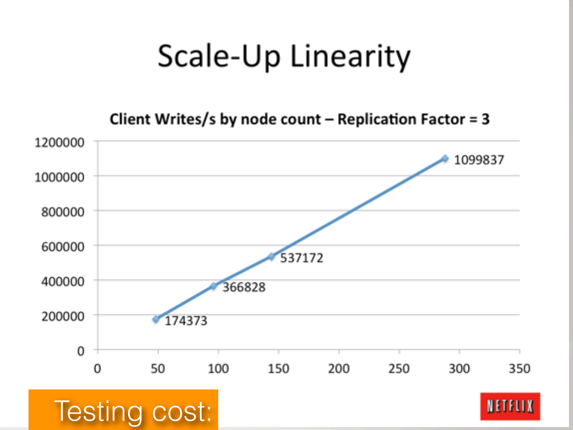


- $5M CAPEX
- 10 man/months delivery (Deploy, DevOps, Tuning)
- Ongoing OPEX of > $1M

- $450K OPEX
- 2 DevOps running 15 AWS nodes
- Faster performance in 2 weeks
- …All in the cloud

Linear Scalability on Commodity Hardware

- Yale University: “The Oracle system that is able to achieve 500,000 transactions per second costs a prohibitive $30,000,000!”



Testing cost: $561.68

DATASTAX

Performance: NoSQL Leadership

Cassandra vs. HBase:

- *10x more read throughput*
- *100x faster read latency*
- *8x more write throughput*
- *8x faster scan latency*
- *4x more scan throughput*

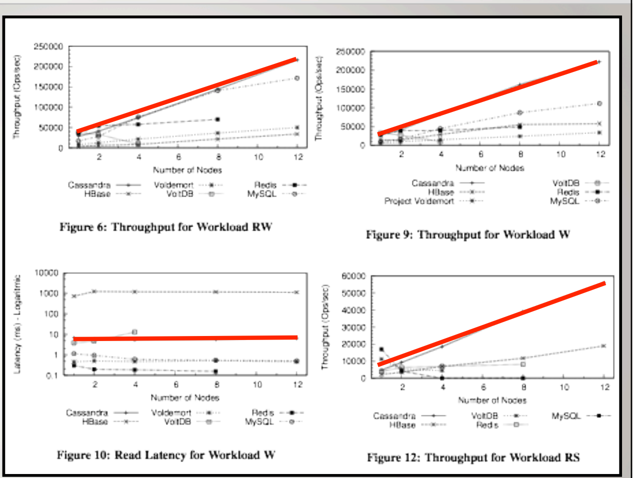


Figure 6: Throughput for Workload RW

Figure 9: Throughput for Workload W

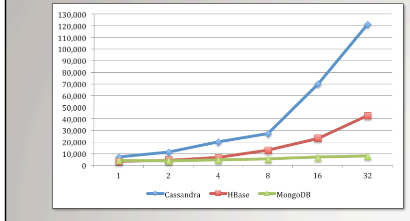
Figure 10: Read Latency for Workload W

Figure 12: Throughput for Workload RS

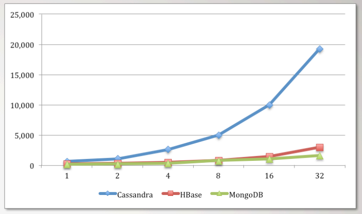
Source: *Solving Big Data Challenges for Enterprise Application Performance Management*
Tillman Rabl, University of Toronto et al VLDB 2012 (August 2012, Istanbul)

Performance: NoSQL Leadership

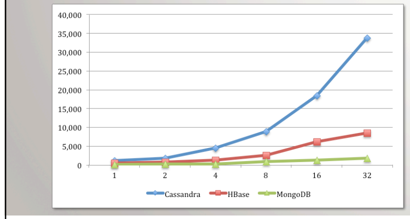
YCSB Load Process



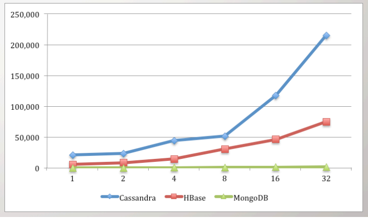
YCSB Read-mostly



YCSB Read-write mix



YCSB Write-mostly



Community: Cassandra Summit 2012

NETFLIX
accenture
High performance. Delivered.
ebay
SAMSUNG
Instagram
@WalmartLabs
Disney
Expedia
media6degrees
SourceNinja
apigee
github SOCIAL CODING
simple reach

MVPs

Ed Anuff (@edanuff), VP, Mobile Platform at Apigee

Ed Anuff is the VP of App Services at Apigee, which he joined after they acquired Usergrid, a mobile back-end platform. Prior to starting Usergrid, he was the EVP of Platform Services at Six Apart. He co-founded Widgetbox, the widget marketplace, and Epicentric, a leading provider of enterprise portal software, which was bought by Vignette. He was also responsible for the launch of the HotBot search engine at Wired.

Tim Berglund (@tlberglund), Trainer and Coder at Github

Tim is a full-stack generalist and passionate teacher who loves working with people as much as he loves to code. He is a GitHubberwhose mission is to make it easy for everybody in the world to use Git. He is a speaker internationally and on the No Fluff Just Stuff tour in the United States, who loves to speak on Git, Cassandra, and other topics. He is co-president of the Denver Open Source User Group, co-presenter of the best-selling O'Reilly Git Master Class, co-author of Building and Testing with Gradle, a member of the O'Reilly Expert Network, and a member of the GigOM Pro Analyst Network. He occasionally blogs at http://timberglund.com. He lives in Littleton, CO, USA with the wife of his youth and their three children.

Russell Bradberry (@DevDazed), Principal Architect at SimpleReach

Russell Bradberry is the Principal Architect at SimpleReach where he is responsible for architecting and building out highly scalable data solutions. He has put into market a wide range of products including a real time bidding ad server, a rich media ad management tool, a content recommendation system and most recently a real time social intelligence platform. He is also the author of the NodeJS Cassandra driver Helenus.

Rick Branson (@rbranson), Infrastructure Engineer at Instagram

Rick is an Infrastructure Engineer at Instagram where he spends his days building scalable systems to deal with their meteoric growth rates. Rick is a totally impatient dude, so he's spent the last 10 years obsessed with making complex production systems faster and more reliable. Previously he was a Solutions Architect at DataStax, and has held various engineering roles at Coroutine, Mosaik Solutions, and FedEx.

Edward Capriolo (@edwardcapriolo), Hadoop Systems Admin at m6d.com

Edward Capriolo is a Hadoop System Admin at m6d.com. Always interested clustered, fault tolerant, and grid computing systems he became dazzled by Apache Cassandra. He is organizer for the NYC Cassandra User Group, NoSQL NYC, and the New York Cloud and Grid Computing Meetup Group. In his spare time, he authored Cassandra High Performance Cookbook for both fun and homage.

Adrian Cockcroft (@adrianco), Cloud Architect at Netflix

Adrian Cockcroft is the director of architecture for the Cloud Systems team at Netflix. He is focused on availability, resilience, performance, and measurement of the Netflix cloud platform, and has presented at many conferences, including the Cassandra Summit, QCon/GOTO, Cloud Connect, Velocity, Gluecon and Structure.

Adrian is also well known as the author of several books while a Distinguished Engineer at Sun Microsystems: Sun Performance and Tuning; Resource Management; and Capacity Planning for Web Services. From 2004-2007 he was a founding member of eBay Research Labs. He graduated with a BSc in Applied Physics from The City University, London.

Eric Lubow (@elubow), CTO at SimpleReach

Eric Lubow, CTO of SimpleReach, began his career building secure systems at Guardian Digital before joining Conductor.com, where he developed and ran their ad systems. After graduating from Rutgers University with a degree in Information Systems, Eric did multiple tours with the U.S. Army including serving in Iraq as a Lieutenant and Albania as a Captain. He is a board advisor for the IronMatt.org charity for pediatric brain tumors, a mixed martial artist, motorcyclist, and seasoned skydiver.

Patrick McFadin (@PatrickMcFadin), Chief Architect at Hobsons

Patrick is the Chief Architect at Hobsons, an education services company. His responsibilities include ensuring product availability and scaling for all higher education products. Prior to this position, he was the Director of Engineering at Hobsons which he came to after they acquired his company, Link-11 Systems, a software services company. While at Link-11 Systems, he built the first widely popular CRM system for universities, Connect. He obtained a BS in Computer Engineering from Cal Poly, San Luis Obispo and holds the distinction of being the only recipient of a medal (as anyone can find out) for hacking while serving in the US Navy.

Aaron Morton (@aaronmorton), Freelance Developer, Apache Cassandra Committer

Aaron Morton is a Freelance Developer based in New Zealand, and a Committer on the Apache Cassandra project. In 2010 he gave up the RDBMS world for the scale and reliability of Cassandra. He now spends his time advancing the Cassandra project and helping others get the best out of it.

Brian O'Neill (@boneill42), Lead Architect at Health Market Science

Brian O'Neill is Lead Architect at Health Market Science, where he heads design and development of their Master Data Management (MDM) solution and Big Data platform, powered by Cassandra. He also leads the Virgil project, which is a services layer providing a REST interface to Cassandra. In the past, Brian participated on numerous Java Community Process expert groups and has patents in artificial intelligence and dynamic application data routing.

Stephen O'Sullivan (@steveos), Senior Director, Platform and Systems, @WalmartLabs

Stephen O'Sullivan works on data architecture and infrastructure to Lead Walmart's next generation of e-commerce. He is an expert in Hadoop usage and architecture and the future of open source solutions for Big Data. He brings more than 20 years of experience creating enterprise applications and data management solutions for high availability and scale to his current position.

Prior to @WalmartLabs, Stephen spent over three years as director of system engineering at LiveOps running the company's data, reporting and API teams. His tenure covers three years in data management for Yahoo's search group and eleven years as a senior engineer at Sun Microsystems. Stephen has a National Diploma in Computer Science from Buckinghamshire Chilterns University College.

Vijay Parthasarathy (@vijay2win), Principal Engineer at Samsung Electronics

As a Principal Engineer at Samsung Vijay is responsible for building consumer facing, distributed, high performance realtime big data solutions.

Prior to samsung he worked at Netflix, he was tasked to implement distributed, highly available datastore running on commodity hardware for the next generation Netflix app's in the cloud.

Vijay is a also a Apache-Cassandra Committer and a Technical Adviser for Trap.it. Before Netflix Vijay worked at Cisco Webex on several distributed backed services spanning 8 DC's, Projects including sharding RDBMS which led to his initial work on Cassandra.

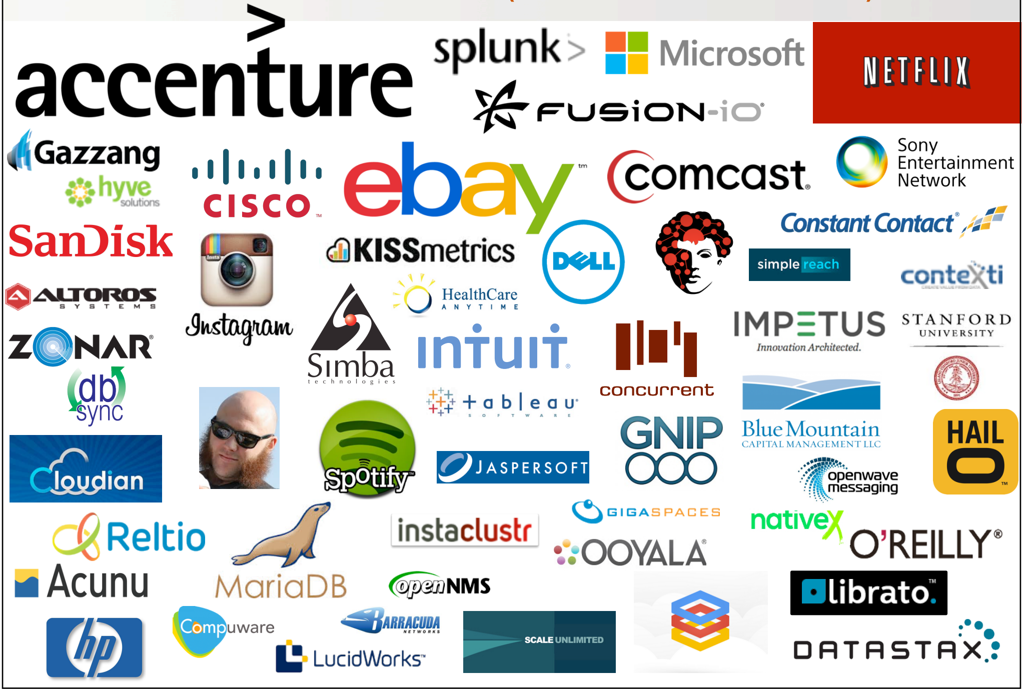
Peter Sanford (@psanford), Software Engineer at Nearbuy Systems

Peter is a distinguished member of the technical staff and first engineering hire at Nearbuy Systems. Prior to joining Nearbuy he ran a software development group in Aruba Networks' AirWave division. Outside of work Peter enjoys drinking coffee and tinkering with embedded systems. Peter holds a BS in Computer Science from Wheaton College (IL).

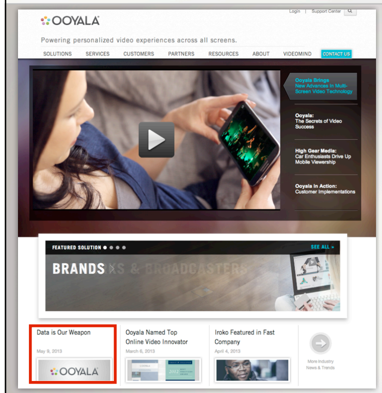
Eddie Satterly (@eddie_satterly), Chief Big Data Evangelist, Office of the CTO at Splunk

Mr. Satterly has served in a variety of roles including developer, engineer, architect and CTO over his 23 year career for Fortune 500 companies. He has been a long time Splunk fan and user. More recently he was able to revolutionize the way Expedia delivers their core web applications that resulted in improved user experience. At Expedia, he created a highly scalable and flexible Splunk environment and as a result was able to retire 35 other systems. With Splunk, he saved $14M in the first 6 months Splunk was installed and avoided many more $MM in outages. He is a recognized expert in the field of big data and has presented at many conference globally on the topic. Eddie has a BS Comp Sci from Indiana University

Cassandra Summit SF (June 11-12, 2013)



Ooyala



Powering personalized video experiences across all screens.

Ooyala Brings New Advances In Multi-Screen Video Technology

Ooyala: The Secrets of Video Success

High Gear Media: Car Enthusiasts Drive Up Mobile Viewership

Ooyala In Action: Customer Implementations

Data is Our Weapon
May 9, 2013

Ooyala Named Top Online Video Innovator
March 6, 2013

Iroko Featured in Fast Company
April 4, 2013

Application/Use Case

- Enable clients, such as ESPN/Disney, to monetize video streaming properties
- Analytics for all end-user interactions with videos

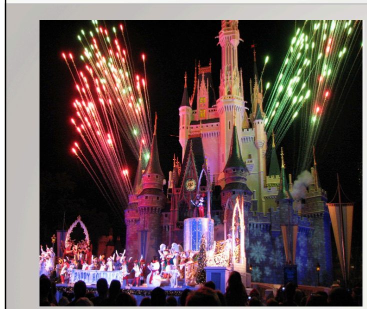
Why DataStax?

- Prior MySQL solution would not scale
- Amount of data coming in too fast and was too large for typical RDBMS
- Online data elasticity for increasing capacity with no downtime a major requirement

“With a conventional database, we’d have to be really in the trenches, or completely re-architecting how we absorb that data. But because we had Apache Cassandra, we knew we’d have to add a few additional nodes to the cluster, at most, and without having to fundamentally re-architect our solution. It gives us tremendous competitive advantage.”

- Harry Robertson, Tech Lead

Disney



Application/Use Case

- Unified data management for 300+ websites → single platform for internal functional sites
- Real-time, analytic, and search functions

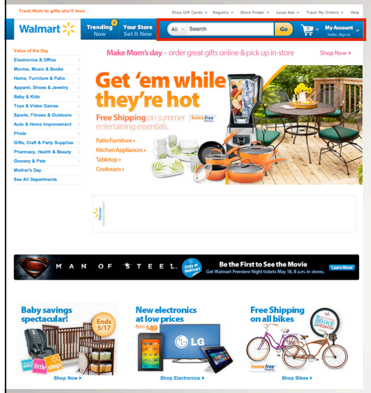
Why DataStax/Cassandra?

- DataStax Enterprise chosen because of its ability to support real-time and search
- Needed scalability to serve data needs of hundreds of websites
- Continuous availability for 24x7 uptime
- Search support crucial

"We use MongoDB, but Cassandra is getting a lot more usage here. We also looked at trying to make HBase work across multiple data centers and couldn't figure out how to do it."
- Arun Jacob, Director of Data Services

DATASTAX

Walmart


Walmart
Trending Now
Your Store
Search
Go
My Account
Make Mom's day – order great gifts online & pick up in-store
Shop Now
Get 'em while they're hot
Free Shipping on summer entertaining essentials.
Patio Furniture
Kitchen Appliances
Tabletop
Cookware
MAN OF STEEL
Be the First to See the Movie
Baby savings spectacular!
Ends 5/17
New electronics at low prices
LG
Free Shipping on all bikes
Shop Now
Shop Electronics
Shop Bikes

Application/Use Case

- Need to maintain global product catalog
- Hundreds of millions of items to manage
- Need to search for products on web site
- Need to analyze product/customer interaction
- Need new product on-line shopping cart

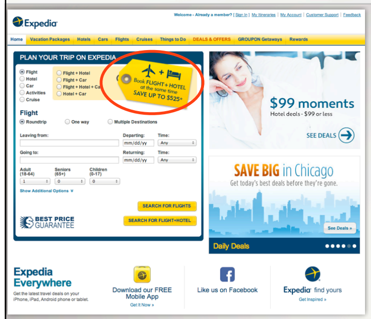
Why DataStax?

- Cassandra for fast data input and change (40% of items change per day)
- Hadoop for batch analysis of operations
- Solr for accelerated web site search
- DataStax Enterprise integrates all the above in one package

"We need a system with low latency, high throughput, and highly available for us to be able to add to our catalog 24 x 7 with staggering updates occurring in that stream of data."
-Rajkumar Venkat, Wal-Mart Labs

DATASTAX

Expedia



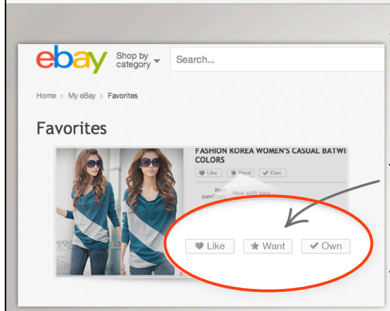
Application/Use Case

- Real-time hotel pricing support
- Database requirements exceeded Splunk capacity
- Need to run analytics on log data
- Faster search for Website (500ms SLA = $2.2M revenue)

Why DataStax?

- Able to process big data log information much more efficiently than other competitors
- DataStax Enterprise provides built in analytics for inputted log data
- Searches on Expedia.com being implemented for faster search response times with DataStax Enterprise
- Avoid ETL data movement between systems
- Saved $1.5 million over old SQL Server system

eBay


Shop by category
Search...
Home › My eBay › Favorites
Favorites
FASHION KOREA WOMEN'S CASUAL BATWI COLORS
Like
Want
Own

Application/Use Case

•Variety of use cases supported including write heavy logging and tracking as well as mixed workloads.

•“Social Signal” project, which enables like/own/want features on eBay product pages.

Why DataStax?

•Needed database to handle heavy write workloads with continuous availability.

•Multi-data center support essential.

•Easily split out functional area workloads across multiple clusters.

•Required expert Cassandra production support.

GNIP

Gnip is the Largest Provider of Social Media Data to the Enterprise - Never Miss a Tweet, Post, Comment or Like

FULL HISTORICAL ACCESS TO THE TWITTER FIREHOSE AVAILABLE EXCLUSIVELY FROM GNIP!

When It's Mission Critical

Gnip serves 8 of the 9 largest social media monitoring firms. When reliable access to social media data is critical to your business, you can count on Gnip.

Full Twitter Firehose

In 2010, Twitter partnered with Gnip to be the first provider of enterprise-grade access to the full firehose of Tweets. Looking for Twitter data? This is the place.

90% of the Fortune 500

Gnip's customers serve over 90% of the companies in the Fortune 500. They demand a robust, scalable enterprise-grade service. Gnip delivers.

3 Billion Activities Per Day

With the most sources, the most customers, and the most robust infrastructure, Gnip reliably delivers an enormous stream of social media data.

A few of our social media sources...

Application/Use Case

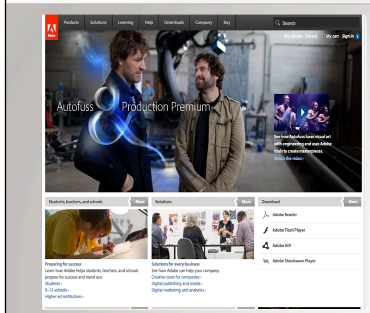
- Social media provider, ingesting all data from Twitter, Facebook, Tumblr, Wordpress, and others.
- 90% of the Fortune 500 get their social media data from Gnip

Why DataStax/Cassandra?

- Sheer data velocity. Twitter alone can generate 20,000 tweets per second
- Massive write load and need for real-time access
- Multi-data center and cloud capable
- Serves system-of-record, compliance, and time series use cases

"We need a real-time, massively scalable architecture, where no one node is specific, that can easily span multiple data centers and cloud availability zones, and that's Cassandra."
- Greg Greenstreet, VP Engineering

Adobe



See how Autofuss fuses visual art with engineering and uses Adobe tools to create masterpieces.
Watch the video ›

Preparing for success
Learn how Adobe helps students, teachers, and schools prepare for success and stand out.
Students ›
K–12 schools ›
Higher ed institutions ›

Solutions for every business
See how Adobe can help your company.
Creative tools for companies ›
Digital publishing and media ›
Digital marketing and analytics ›

Application/Use Case

- Adobe AudienceManager service, an audience optimization solution that manages customers' digital data assets across the enterprise, helping them to perform web analytics, web content management and online advertising

Why DataStax?

- Needed a distributed cache for the ever-changing user profile data Adobe AudienceManager stores for customers
- Looked at HBase, and Membase but chose Cassandra because of linear scale for reads/writes and availability across multiple data centers

"I can tell you we've been very pleasantly surprised by just how well Cassandra performs ."
- Dave Weinstein, Adobe

Call to action

More Great Resources

Events

- 6/11-12: Cassandra Summit (SF)
- 6/17: Bloomberg Next Big Thing (Half Moon Bay, CA)

- Case Studies and Interviews
  datastax.com/casestudies
- Cassandra Conference Presentations and Videos
  datastax.com/events/cassandrasummit2012/presentations
- White Papers
  datastax.com/resources/whitepapers
- Webinars
  datastax.com/resources/webinars
- Training
  datastax.com/services/training

Thank You

We power the big data apps
that transform business.