OPEN SOURCE BUSINESS CONFERENCE

Building Your Big Data Future with Open Source

COMPUTERWORLD
OSBC
SAN FRANCISCO

COMPUTERWORLD OSBC SAN FRANCISCO

Building Your Big Data Future with Open Source

DataStax

DataStax' Brisk – Celebrity Open-Source Super Couple.

Hadoop Powered by Cassandra

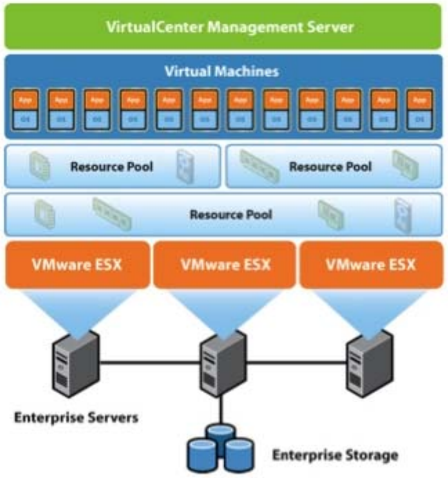
Ben Werther

VP of Products

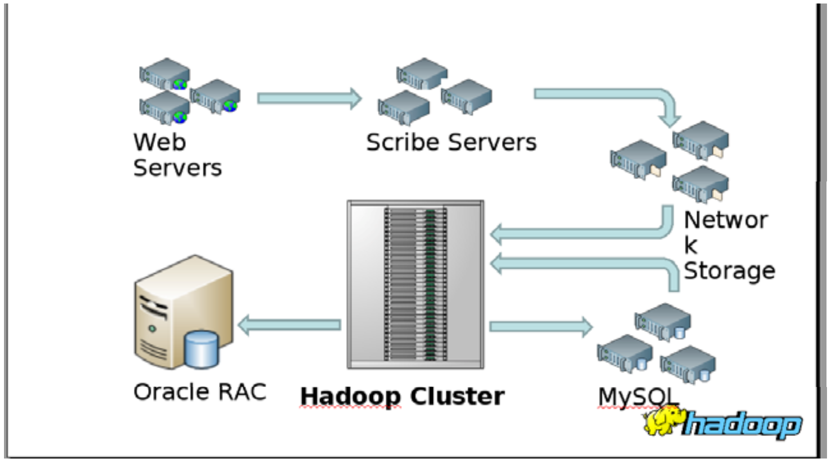
DataStax

The Shift to Data-Centricity

- Before… app- and server-centric infrastructure
- But look around – it is a data-centric world



App-Centric – VMware Virtualization



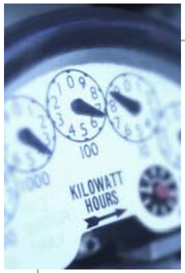
Data-Centric – Facebook’s Dataflow

A Few Examples

High-Volume Websites

Finance and Capital Markets

Retail



Smart Grid Sensors

State of Play

- Batch Analytics: Hadoop and Hive
  - Strong ecosystem, very scalable, not highly tuned
  - Complex to run in production, SPOF (HDFS)
- Low Latency: Cassandra
  - Very scalable and extremely high performance
  - No SPOF, but no batch analytics capabilities
- Customers – We Need These Unified!
  - Goals: Simpler stack, no manual ETL, batch analytics and low-latency in one system, resource isolation

Apache Cassandra™

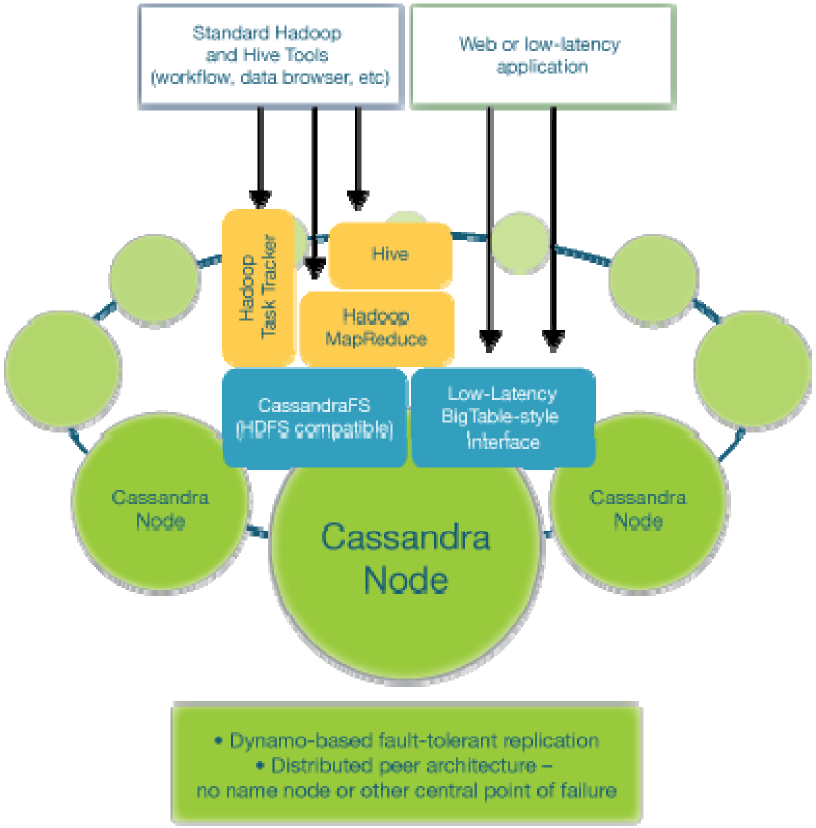
- Was incubated at Facebook by Avinash Lakshman
  - Incorporated the best of Google’s BigTable and Amazon’s Dynamo models in one system
- Was open-sourced by Facebook in 2008
  - Became an Apache top-level project under the leadership of Jonathan Ellis (DataStax)
- The ‘best-of-breed’ big-data low-latency infrastructure
  - In use at 1000s of organizations worldwide, including Twitter, Netflix, Cisco, Rackspace, as well as in government/intelligence, financial services, telecommunications and logistics

Cassandra – Technical Differentiators

- Massively scalable ring architecture
- Flexible schema-less data modeling
- Extreme write performance with durability
- Gossip-based fault detection and recovery
- Incremental and elastic expansion
- Multi-datacenter replication
- Cache-like performance

Introducing Brisk

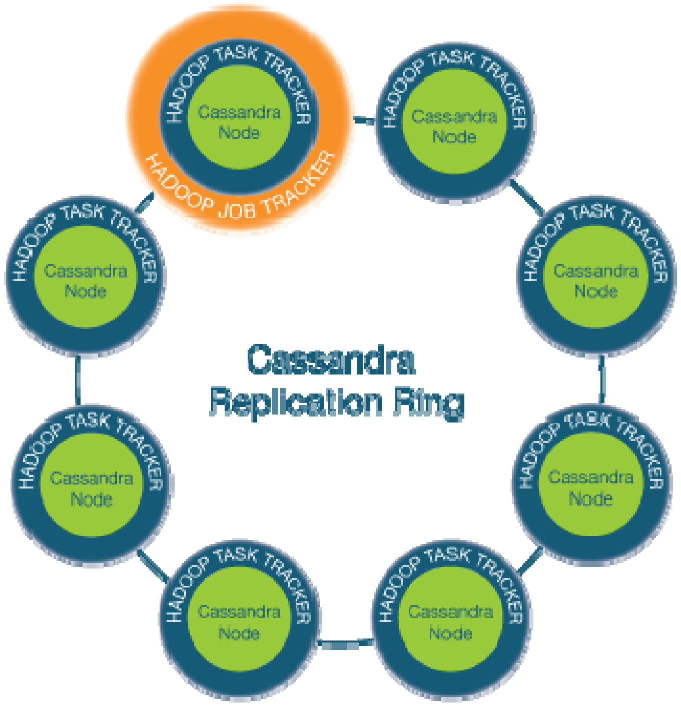
- A New Hadoop Distribution powered by Cassandra
  - Best-of-Breed combination of Low-Latency Database and Batch Analytics
  - Dramatically simplifies the Hadoop stack, while retaining full compatibility
- Open-source Apache 2.0 license
  - Downloadable now at datastax.com/brisk


Standard Hadoop and Hive Tools (workflow, data browser, etc)
Web or low-latency application
Hadoop Task Tracker
Hive
Hadoop MapReduce
CassandraFS (HDFS compatible)
Low-Latency BigTable-style Interface
Cassandra Node
Cassandra Node
Cassandra Node
• Dynamo-based fault-tolerant replication
• Distributed peer architecture – no name node or other central point of failure

Hadoop - Radically Simplified

- Fully Integrated Stack
  - Hadoop 0.20.2, Hive 0.7, Cassandra 0.7.4
  - Everything is started automatically
    - Hadoop job trackers and task trackers managed by Cassandra nodes
  - All nodes are peers, with no single point of failure
    - No Hadoop name nodes, Zookeeper, Region servers, etc.

"Hadoop Powered by Cassandra" Deployed



Brisk Performance

[slide to be inserted]

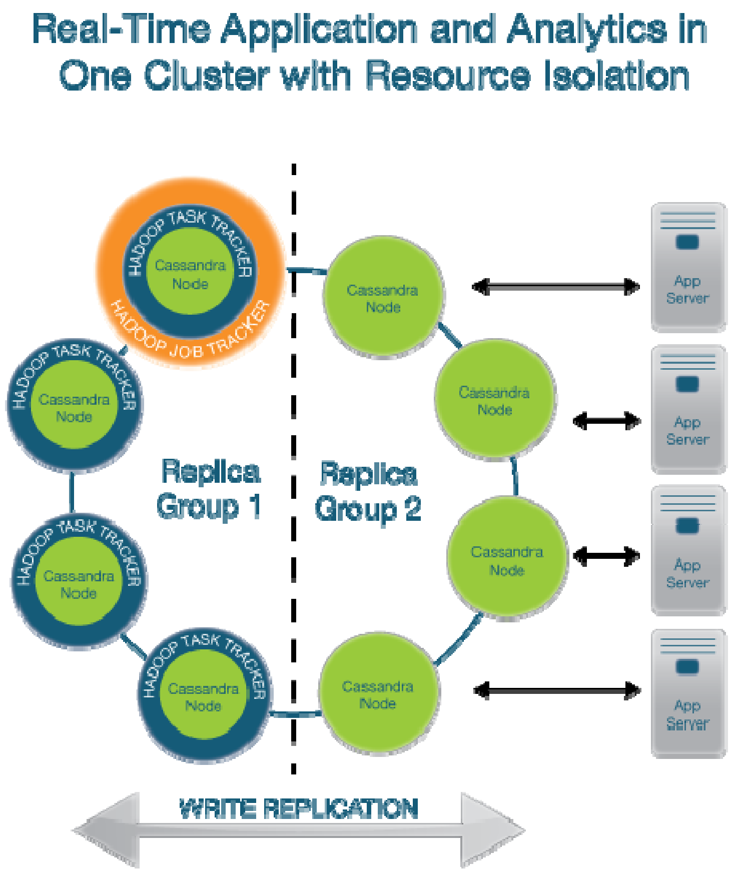
Brisk Internals

- HDFS Compatible Layer (CassandraFS)
  - 2 Column Families (inode, block)
  - No Namenode, Secondary Namenode. No SPOF.
  - hadoop distcp hdfs:///mydata cassandra:///mydata
- JobTracker and TaskTracker management
  - 1 Seed node is elected JobTracker
  - No config for this
- BriskSnitch splits cluster for OLAP and OLTP workloads

More on Hive

- Hadoop and Hive Drivers for accessing Cassandra data
  - Access both low-latency Cassandra data and HDFS-style data
  - High performance – equal or faster than other distributions
- Two types of access
  - Fixed column access (rowid, firstname, lastname, zip)
  - Dynamic column access (rowid,column,value)
- Hive MetaStore in Cassandra
  - Unified schema view from any node. No SPOF

Isolation w/ Zero-Delay Feedback Loop

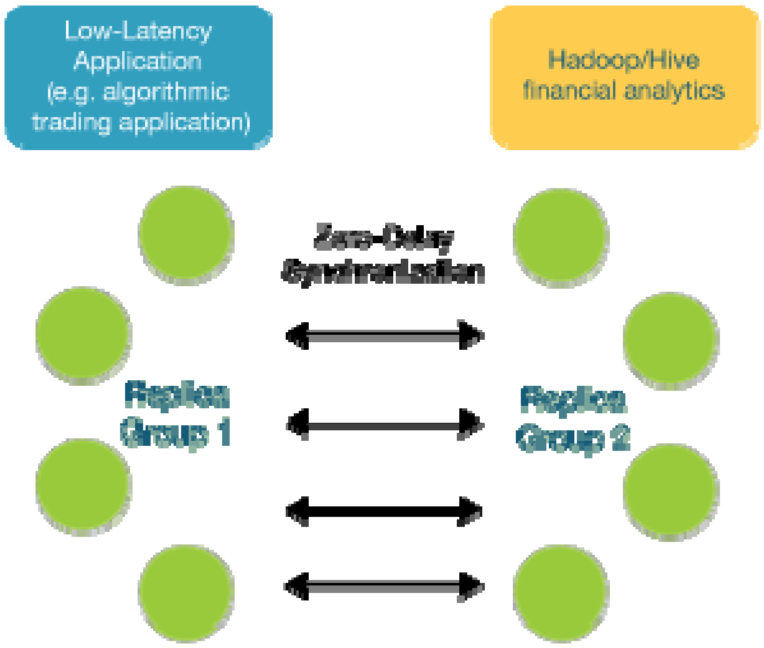


Real-Time Application and Analytics in One Cluster with Resource Isolation

- Built-in Support for Role-Based Replica Groups
  - Assign replica to do low-latency, analytics or both
- Zero-Delay Loop Between App and Analysis
  - Application can do millions of fine-grained reads/writes per second
  - Analysis always sees latest data
  - Analytical results instantly available to the application

Trading Example In Action

1. Trading app receives a stream of market events that it stores and responds to in real-time based on a predictive model


Low-Latency Application (e.g. algorithmic trading application)
Hadoop/Hive financial analytics
Zero-Delay Synchronization
Replica Group 1
Replica Group 2

2. Every few minutes a Hive query runs to update the predictive model based on the very latest data.

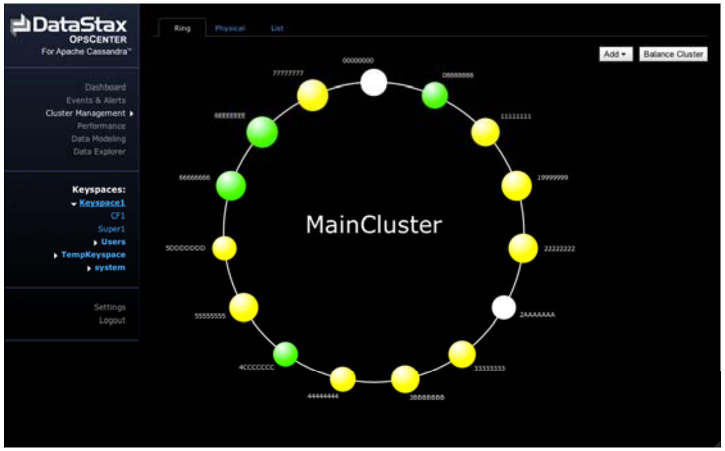
This is written back into the system with Hive.

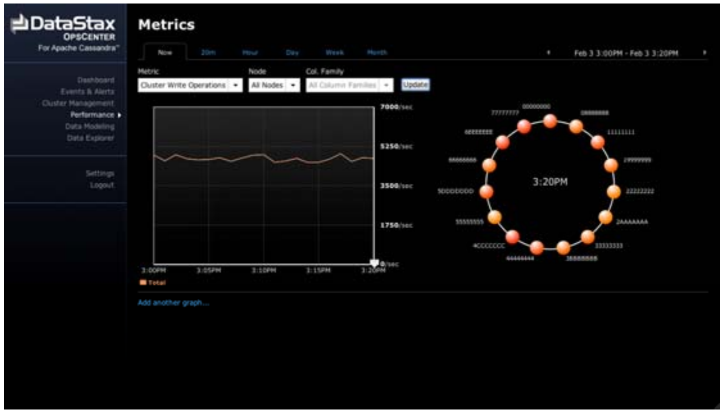
3. The updated predictive model is immediately available for low-latency processing.

DataStax OpsCenter

for Apache Cassandra & Brisk

- DataStax OpsCenter is the first platform for managing, monitoring and operating Brisk and Cassandra applications.
  - Sophisticated visualizations of a Brisk or Cassandra cluster
  - Real-time Hadoop job tracking
  - Comprehensive management and configuration
  - Health and performance monitoring.
- Freely downloadable for non-production use





Metric: Cluster Write Operations

Node: All Nodes

Col. Family: All Column Families

About DataStax

- DataStax is the commercial leader in Apache Cassandra™ and the developer of Brisk
  - Build products and services 'For' or 'Powered by' Apache Cassandra™
- Founded in early 2010 by Matt Pfeil and Jonathan Ellis
  - Jonathan is the leader and project chair of Apache Cassandra
- More than 80 customers including:
  - Netflix, Cisco, Openwave, Ooyala, Constant Contact, RealNetworks, Rackspace
- Based in Burlingame, CA
  - With offices in Austin, TX and Stamford, CT
- More than 30 employees
  - Most of the core Cassandra project developers, plus superb pool of enterprise distributed systems talent

Investors Include:

LIGHTSPEED™ VENTURE PARTNERS

SEQUOIA CAPITAL
THE ENTREPRENEURS BEHIND THE ENTREPRENEURS

rackspace
IT HOSTING

Questions?