DATASTAX

Spark and Cassandra

Martin Van Ryswyk, EVP Engineering

Partnership

DATASTAX + databricks

Jonathan Ellis
Chairman, Apache Cassandra

Matei Zaharia
Creator, Apache Spark

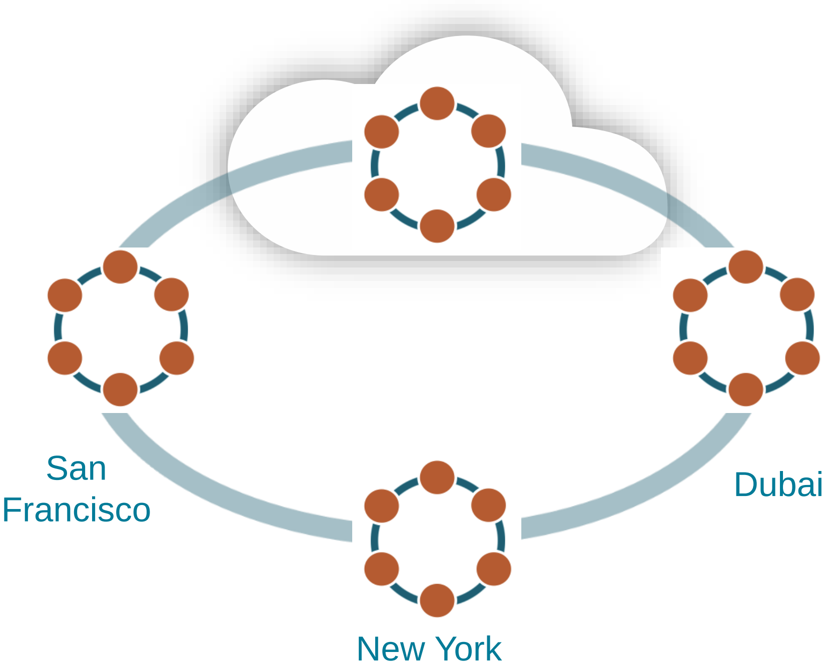
Announced in May 2014

Based on clear feedback from both communities

Committed to an integration between Spark and Cassandra

What is Apache Cassandra?

Apache Cassandra



Massively scalable database

Always on

Fully distributed

Linear scale performance

Flexible NoSQL data model

Operationally simple

SQL-Like language

Free tools and drivers

1 9 2 .com People, Businesses & Places
Acunu
adform
adku
Adobe
Aeris COMMUNICATIONS
AGENTIS
aiMatch ad intelligence
airpim
ALTOROS SYSTEMS
ampush
apigee
APIXIO

AppDynamics
applifier
AppScale
appssavvy
Apptegic
Arise
ARKIVUM ASSURED ARCHIVING
ARRAYENT
AURELIUS
AZUL SYSTEMS
backupify Secure your online life
b:

Burt.
CAKESOLUTIONS the vital ingredient to your business
CallFire
CANONICAL
CARTERA commerce
CERN
chaordic
CharityUSA.com Greater Good Network
Chegg The Student Hub
Chill
citypath BETA YOUR CITY, YOUR WAY
clearleap
*clearspring

Cloudian
cloudkick
CLOUDTALK
cloze
CODEMETTLE
comcast.
COMPUSCAN
CONCORD
Constant Contact Connect. Inform. Grow.
CREATIVE MOBILE GAMES
Crowd Riff
CSR Push every boundary.
DachisGroup

DATADIO
DATADOG
DELL
Demandforce Growth. On Demand.
despegar.com
digg
Digital River
DISQUS
easou 宜搜
ebay
ebuddy
egnyte Cloud File Server

Σi
elasticinbox
embed.ly
ENGINE
Engine Yard
enkata
ERICSSON
EUROTECH Imagine. Build. Succeed.
FanBridge
Fight Monster
FINN
formspring
Friend Facts
Frugal Mechanic
F-Secure

FUSION-IO
GF GAMEFLY Games Delivered
Gavagai
GIGASPACES
Globant we are ready
GNIP
GoDaddy.com
goingon
GOTIME.com
gotosolr
GRIDPOINT
HAIL
Hangtime

HEALTH MARKET SCIENCE
HealthCare ANYTIME
Healthline Connect to Better Health
hobsons
hp
HubSpot
IBM
IGNITIONONE
IMPETUS
Infinite Convergence
inkling
instaclustr
Intelligent Systems Group

inteluma ENERGY SYSTEMS
IHG InterContinental Hotels Group
IID
InWorldz
iovation
ISIDOREY CLOUD SOLUTIONS
ITHAKA
jetpac
Jumptap
justSignal
KISSmetrics
krux
lacuna SYSTEMS

LeaseEagle
librato
Life360
lightt
LiveOffice Your Email Experts
LIVEPERSON
looplogic
Mahalo
MARKED UP
martinimedia
Mass Relevance
mate1.com find someone

media6degrees
MEDIALETS
medidata Medidata Solutions Worldwide
medio
MedProcure
meta broadcast
Acquired By Microsoft
MolloM
monitis Monitor everything
moodyo
MORNINGSTAR
movile
napshare
NASA

NAVTEQ
nearbuy
NETFLIX
news360
NEXGATE
NEXT BIG SOUND
Nextag
ngmoco:)
NICE TO WEB YOU
NREL National Renewable Energy Laboratory
nSCALED
nToklo
одноклассники
ONEsite
one spot

onsip
OOYALA
opensource connections
OpenAdExchange
openNMS
opentracker event tracking and analytics
OPENWAVE
OpenX
Other inbox
outbrain
PACE ANTI PIRACY
page2rss

Palantir
paper li
PAPERKUT MAKING PAPER RECEIPTS, PAPERLESS
parseix
PBS KIDS
PCPhase
PDF SOLUTIONS
PeopleBrowsr
PERFECT WORLD
Ping Identity
ping4alerts!
PitneyBowes
platform46
plaxo

POLYVORE
PRESS+
EPIC Colorado
proofpoint
PULSAR Big Social Data for Research
quantiFind
QUOVA a Neustar service
Rakuten
RALLY SOFTWARE
RAZORlabs
real
Recruiting.com
reddit

referentia
Reltio
resultly
retickr
Rhapsody
RIGHTSCALE CLOUD MANAGEMENT
riptide io
RockMelt
rockyou
Ruckus WIRELESS
simbiose
SCALE UNLIMITED
Scandit

Scoop.it!
SeaChange
see | me
SEOmoz
SEVEN
sevocity ELECTRONIC HEALTH RECORDS MADE EASY
ShareThis
shazam
SHODAN Computer Search Engine
shopsavvy
SimpleGeo

simplereach
skillpages
smeet
SocialArtisan
SOCIALFLOW
SOCIALPARENT
Software Projects
Sonico
SourceNinja
spidertracks
Spotify
spotright

spotXchange
Spredfast
springpad
SQUIDOO
Startapp
stayfilm
Stormpath
Strmur
Symantec.
SYNAPSE
Synchronoss
Taboola
Tapulous

TEADS
TENDRIL
the echonest
thefind
THIRD IRIS
THOMSON REUTERS
trap!t
travelocity
trueffect First-Party Targeting and Media Measurement Platform
Tutao
TWIMPACT
twitter
URBAN AIRSHIP

Veezio
verticalscope
VigLink
Visual DNA
UNG embracing challenges
VTEX Remarkable Technology
W2O Group

Netflix Ensures Constant Uptime with DataStax Enterprise

World’s leading streaming media provider with digital revenue $1.5BN+

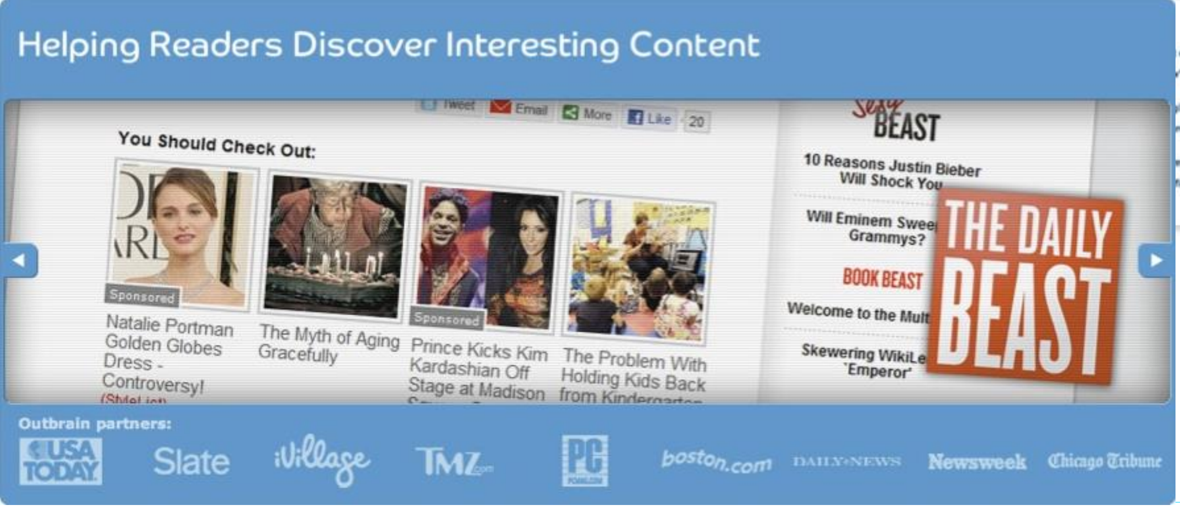
95% of all Netflix data stored in DSE

Introduction of ‘Profiles’ drove throughput to over 10M transactions per second

Does 1 trillion transactions/day with DSE

Replaced Oracle in six data centers, worldwide, 100% in the cloud

NETFLIX



Helping Readers Discover Interesting Content

You Should Check Out:

Natalie Portman Golden Globes Dress - Controversy! (StyleList)

The Myth of Aging Gracefully

Prince Kicks Kim Kardashian Off Stage at Madison S

The Problem With Holding Kids Back from Kindergarten

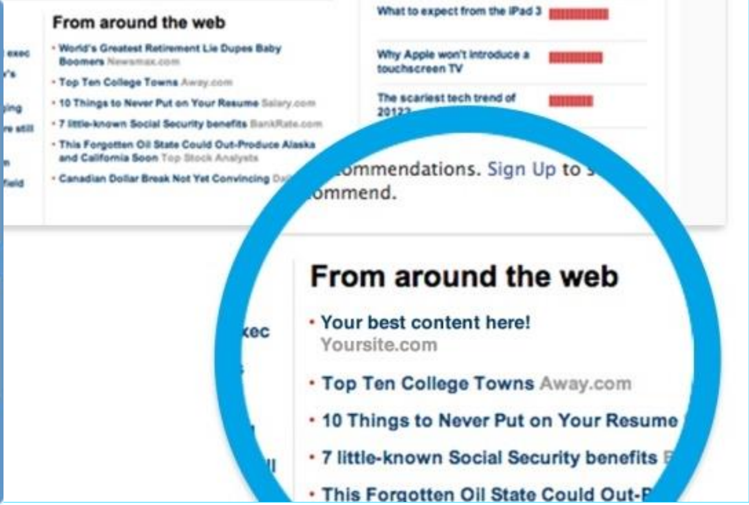
10 Reasons Justin Bieber Will Shock You

Will Eminem Swee Grammys?

Welcome to the Mult

Skewering WikiLe 'Emperor'

Outbrain partners:



From around the web

- World's Greatest Retirement Lie Dupes Baby Boomers Newsmax.com
- Top Ten College Towns Away.com
- 10 Things to Never Put on Your Resume Salary.com
- 7 little-known Social Security benefits BankRate.com
- This Forgotten Oil State Could Out-Produce Alaska and California Soon Top Stock Analysts
- Canadian Dollar Break Not Yet Convincing Da

What to expect from the iPad 3

Why Apple won't introduce a touchscreen TV

The scariest tech trend of 2012?

ommendations. Sign Up to s
ommend.

From around the web

- Your best content here! Yoursite.com
- Top Ten College Towns Away.com
- 10 Things to Never Put on Your Resume
- 7 little-known Social Security benefits B
- This Forgotten Oil State Could Out-P

Delivers 150+ Billion Content Recommendations Per Month

Serves content for largest media brands in the world: Reuters, Wall St Journal, USA Today

Needed a massively scalable data store

High velocity of data with 58,000 links to content per second

Always-on data architecture

Lost a data center during Hurricane Sandy but never went offline

Outbrain

Mostly Cloudy

FEELS LIKE 76°

Dry conditions will continue

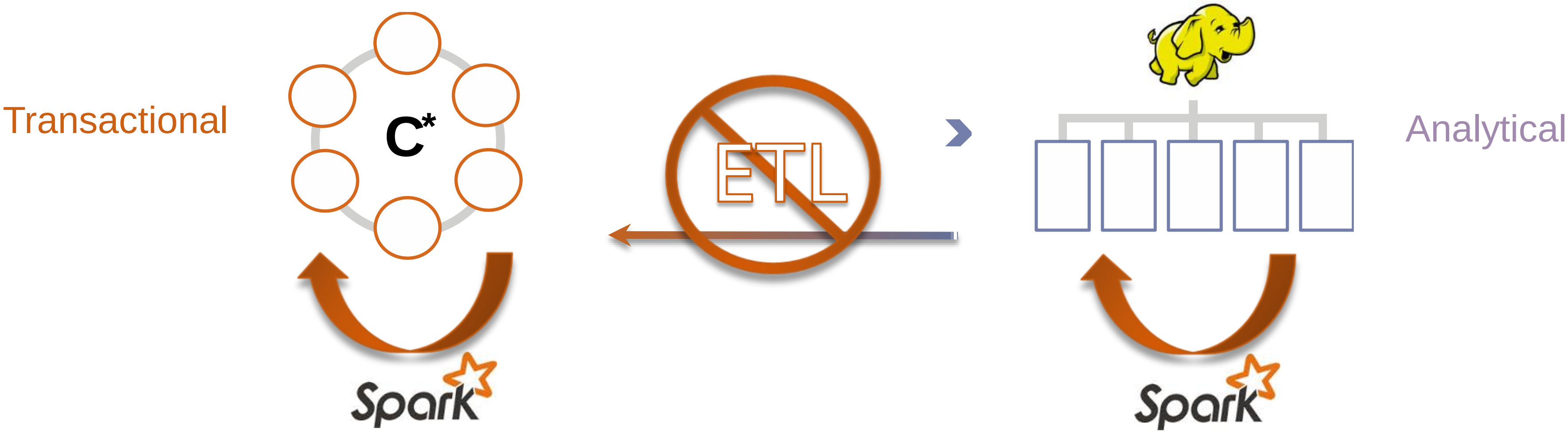
Updated about 20 seconds ago

The Weather Channel

While I have years of experience using Cassandra, my team was mostly new to it; ***CQL made their transition essentially painless****. But where Cassandra really shines is in* ***speed and operational simplicity****, and I would say those two points were critical."*

ROBBIE STRICKLAND *Software Dev Manager*

Faster Feedback Loops



Announcing cassandra-driver-spark

cassandra-driver-spark v1.0

NEW

- Developed by DataStax with support and review by Databricks
- Free with Apache 2.0 license from DataStax

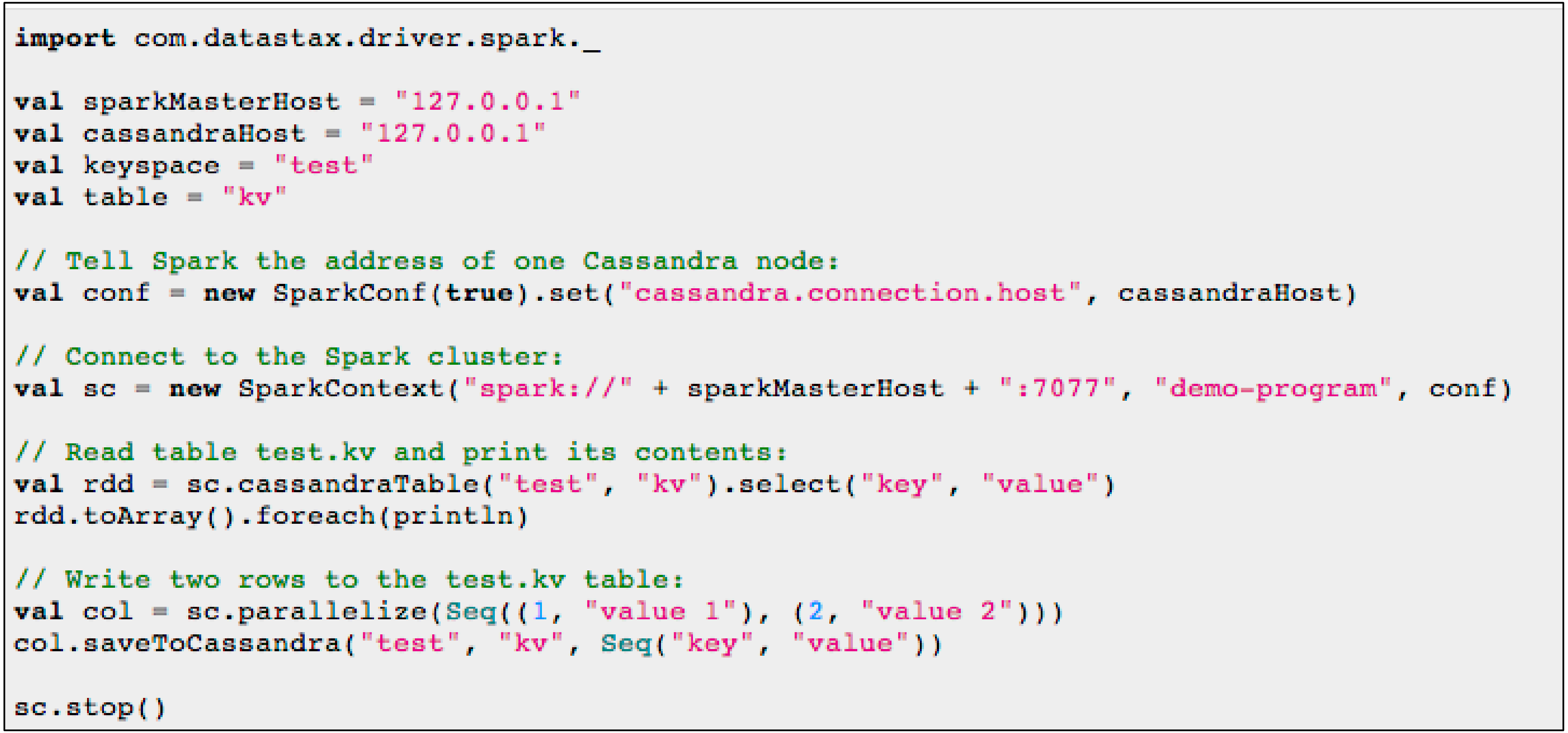
  https://github.com/datastax/cassandra-driver-spark
- Question to Apache Spark User List
  - user@spark.apache.org
- Offering driver code to Apache Spark community

cassandra-driver-spark v1.0

NEW

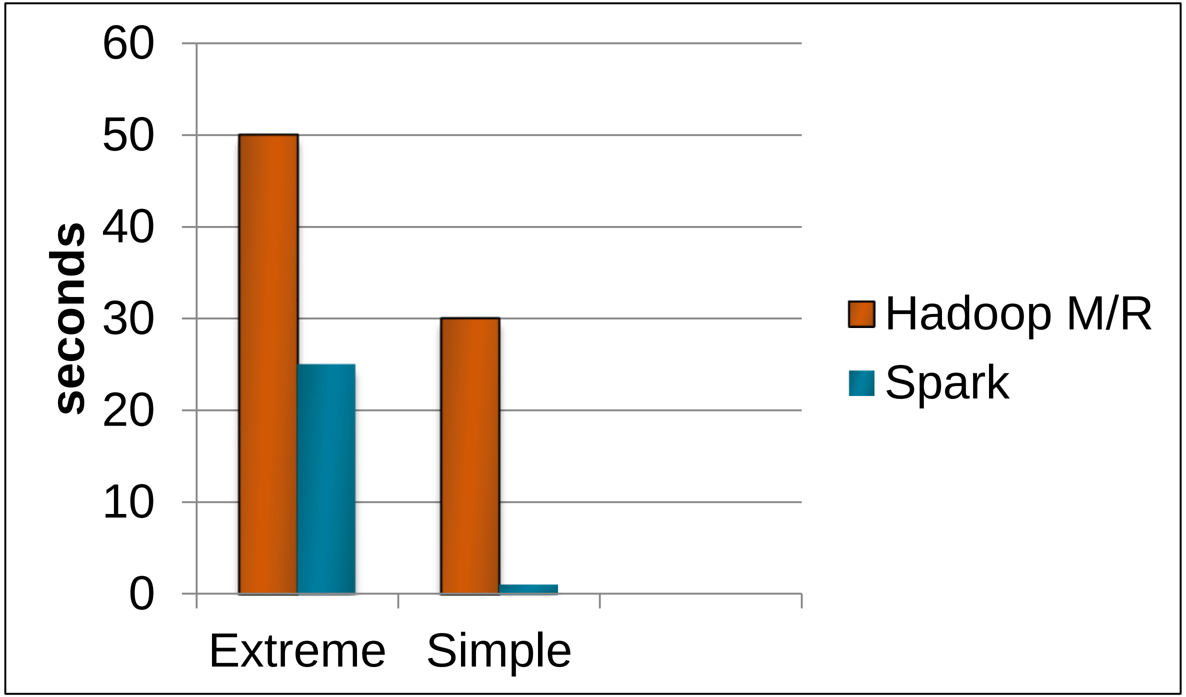
- Exposes Cassandra tables as RDD
- Map table rows to CassandraRow objects
- Data type conversions between Cassandra and Scala
- Save RDDs back to Cassandra by implicit saveToCassandra call
- Filter rows on server via CQL WHERE clause (CassandraRDD#where method)
- Select subset of columns (CassandraRDD#select method)
- Optimizations for Cassandra vnodes

Example



```
import com.datastax.driver.spark._

val sparkMasterHost = "127.0.0.1"
val cassandraHost = "127.0.0.1"
val keyspace = "test"
val table = "kv"

// Tell Spark the address of one Cassandra node:
val conf = new SparkConf(true).set("cassandra.connection.host", cassandraHost)

// Connect to the Spark cluster:
val sc = new SparkContext("spark://" + sparkMasterHost + ":7077", "demo-program", conf)

// Read table test.kv and print its contents:
val rdd = sc.cassandraTable("test", "kv").select("key", "value")
rdd.toArray().foreach(println)

// Write two rows to the test.kv table:
val col = sc.parallelize(Seq((1, "value 1"), (2, "value 2")))
col.saveToCassandra("test", "kv", Seq("key", "value"))

sc.stop()
```

Performance



2-30x faster than highly optimized "Hadoop on C*" implementation in DataStax Enterprise

Extreme: data not in memory and fetched from multiple nodes
Simple: data set fits in memory

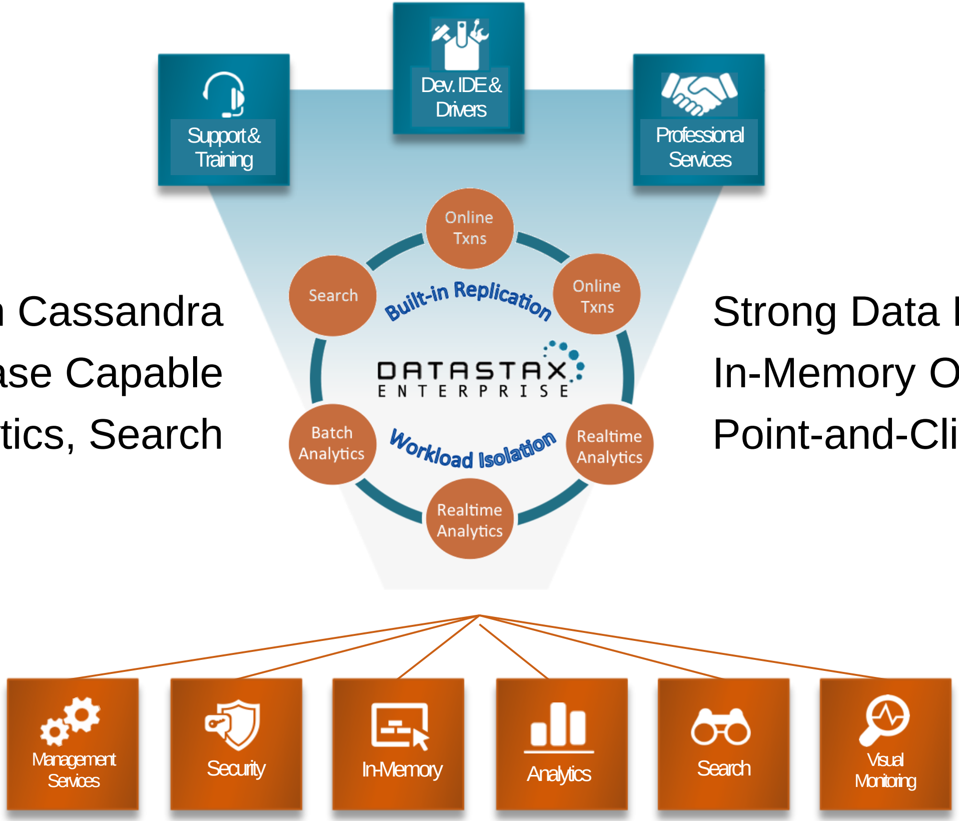
What is DataStax Enterprise?

Delivering Apache Cassandra to the Enterprise

Certified Production Cassandra

Multi-Workload/Use Case Capable

Integrated OLTP, Analytics, Search



Strong Data Protection

In-Memory OLTP/Analytics

Point-and-Click/Automated Mgmt

Announcing

NEW

INTRODUCING

DATASTAX ENTERPRISE 4.5

The World's Fastest, Most Scalable Distributed Database Technology

Certified Spark Distribution

cassandra-driver-spark Roadmap

- Integrate into Apache Spark
- Update for Spark 1.0
- Support for Additional Spark components
  - Spark SQL
  - Spark Streaming
  - GraphX
- Listen to both communities…

Next Steps

Today:

Attend Tupshin and Al’s talk at 1pm

Visit DataStax booth – meet experts

Tonight:

Download the integration	github.com/datastax/cassandra-driver-spark
Learn about Cassandra	planetcassandra.org
Download DataStax Enterprise	www.datastax.com/downloads

September:

datastax.com/cassandrasummit14