Developer Day
DSE Overview and when to use

Your Instructors

Amanda Moran, *Developer Advocate*

Programming languages & Areas of Interest: Python, Bash. Analytics, cloud, Linux, databases, Spark, Cassandra, Tensorflow, Big Data, DevOps.
Interesting Projects: Character recognition with Hadoop, Tensorflow, Teradata, Aster, and Android app.
Github Handle: @amandamoran
Twitter Handle: @AmandaDataStax
LinkedIn: in/Amanda-Kay-Moran

Cedrick Lunven, *Developer Advocate*

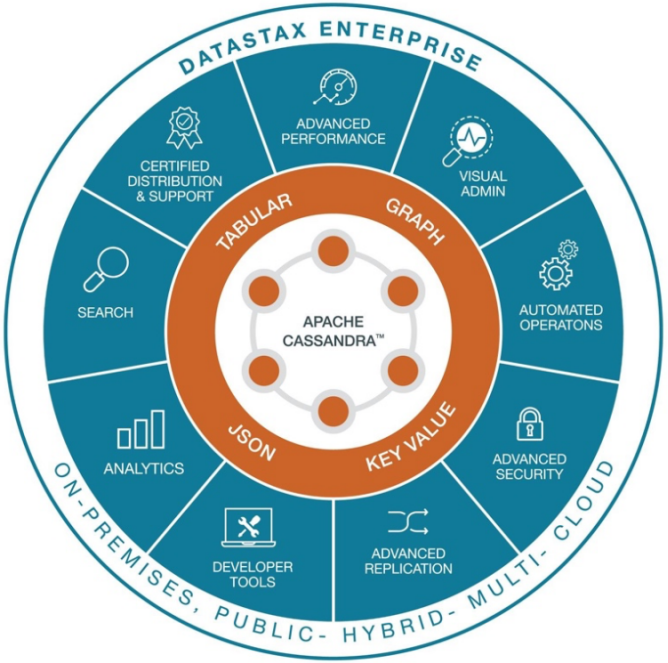
Programming languages & Areas of Interest: Java, Javascript, Python. Distributed architectures and integration.
Interesting Projects: Creator and main commit of FF4j, Feature Flipping for Java.
Github Handle: @clun
Twitter Handle: @clunven
LinkedIn: in/clunven

Goals for this Session

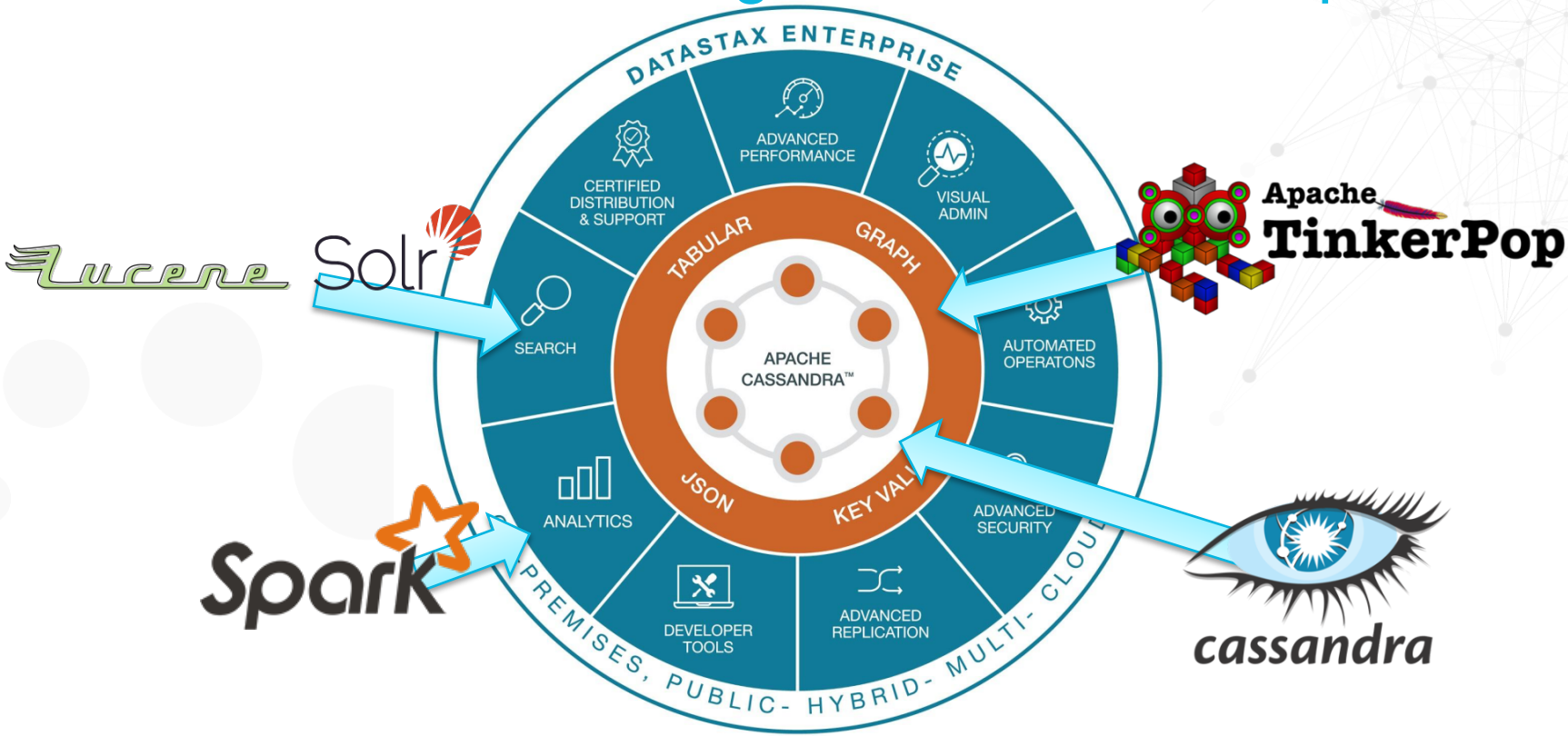
- Know what makes DataStax Enterprise different from OSS Apache Cassandra
- Gain a high-level understanding of each of the DSE products and tools
- Learn about the business use-cases for each of the DSE products
- Watch some pretty *cool* demos
- Learn something new, have fun, ask questions!

Integrated Data Platform for Cloud Applications

- DataStax Enterprise
  - Core
  - Search
  - Analytics
  - Graph
- DataStax Studio
- DataStax Drivers
- DataStax OpsCenter



OSS Foundations, Integrated for the Enterprise



Apache Cassandra, Spark, Lucene, Solr, TinkerPop ® Apache Software Foundation

DataStax ENTERPRISE

> 2X Writes

Advanced Performance

~ 2X Reads

> “This is a ‘once every seven years’ rewrite.”
>
> **JONATHAN ELLIS**

3X Query

DSE Spark Connectivity

~ 4X Load

DataStax Bulk Loader

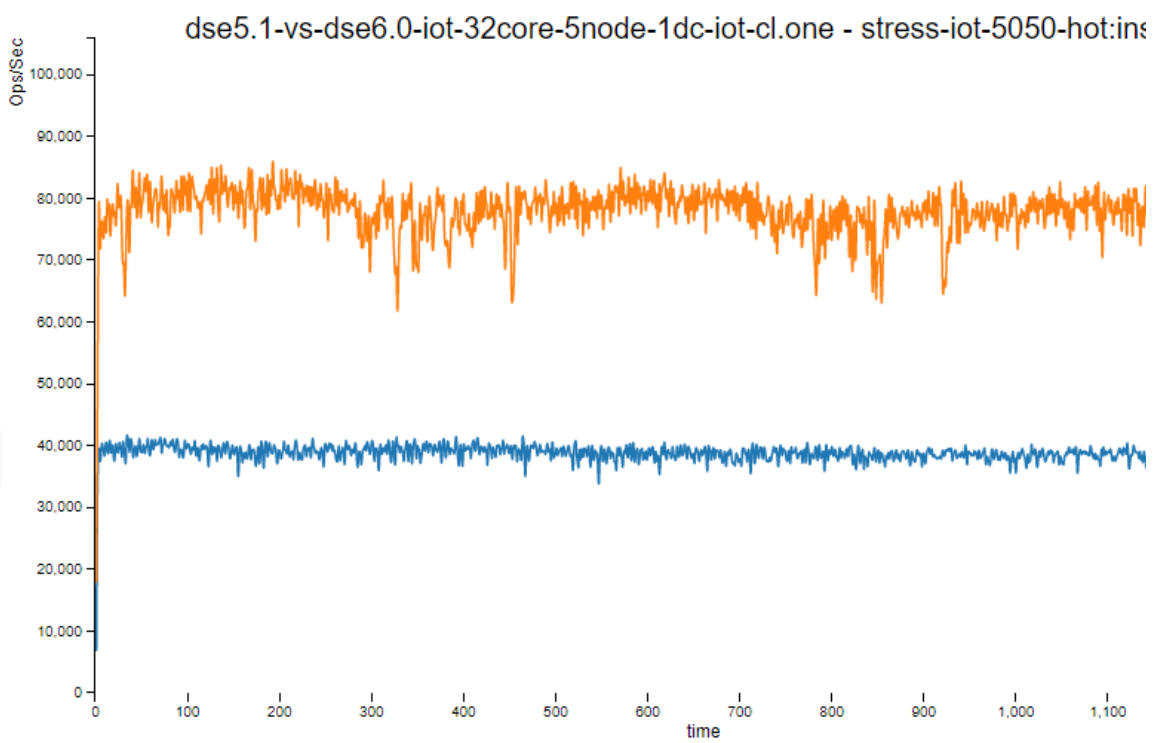
> 4X Unload

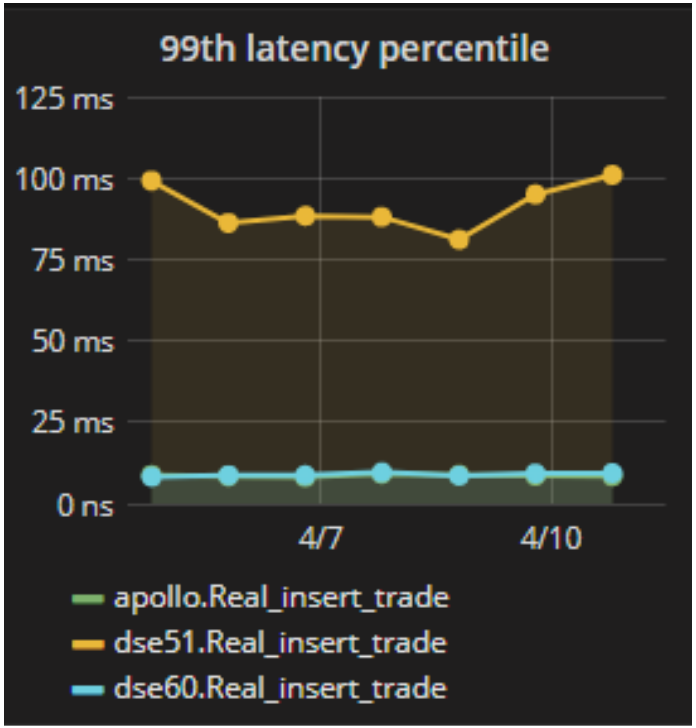
> “Please give me a faster and easier way to load data.”
>
> **DATASTAX CUSTOMER**

What's in DataStax Enterprise 6?

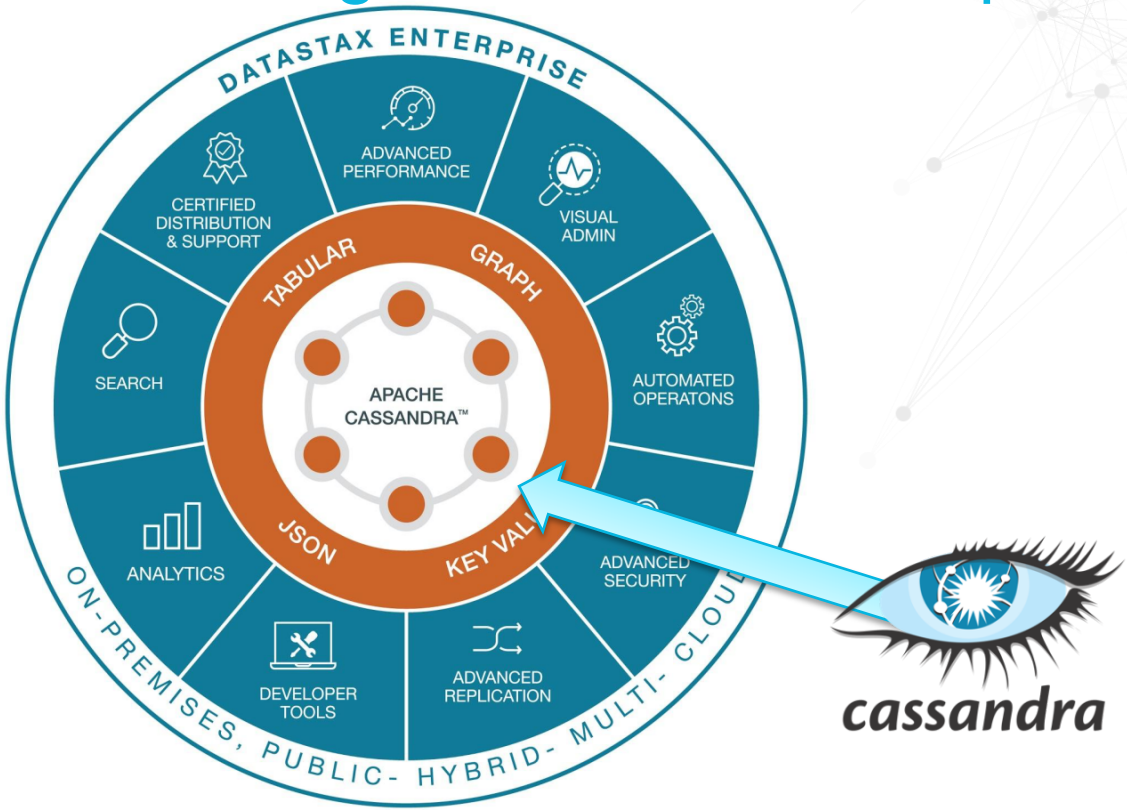
- Double the performance
  - Less hardware required for your workload
  - Improved p99 read/write latency
- Self-driving operational simplicity
  - Public / Private / Hybrid clouds
  - Reduced repairs with NodeSync
- Robust transactional analytics
  - Support for Spark SQL in DataStax Studio
  - AlwaysOn SQL
  - Streaming from Apache Kafka, file systems, other sources

2x Increased Throughput, Decreased Latency





OSS Foundations, Integrated for the Enterprise



Apache Cassandra and DataStax Enterprise

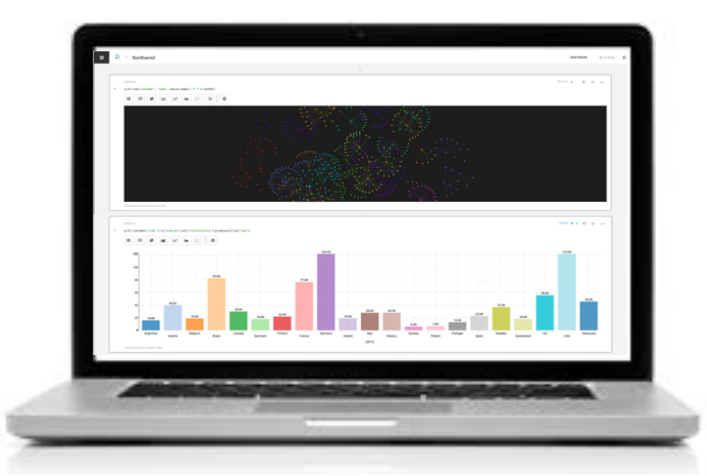
- First developed by Facebook
- Top-level Apache project since 2010
- Distributed, decentralized
- Elastic scalability / high performance
- High availability / fault tolerant
- Tunable consistency
- Partitioned row store

- The best distribution of Apache Cassandra™
- Production certified Cassandra
- Advanced Performance
- Advanced Security
- Advanced Replication

Apache Cassandra ® Apache Software Foundation

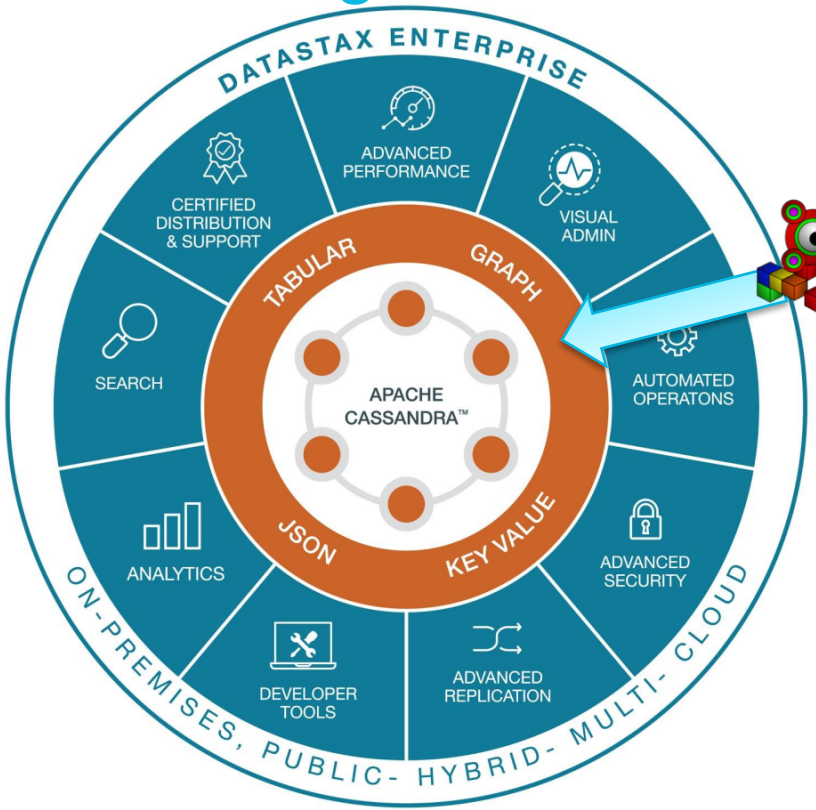
DataStax Studio

Explore, query, and analyze CQL, DSE Graph, and Gremlin Query Language



- Explore and query CQL, Gremlin Query Language, and Spark SQL
- Auto-completion, result set visualization, execution management, and much more
- Friendly fluent API Graph and data visualization

OSS Foundations, Integrated for the Enterprise



Demo: Fraud Detection with DSE Graph

Gremlin Query Language (GQL)

```
g.V().has("name","gremlin").
      repeat(in("manages")).until(has("title","ceo")).
      path().by("name")
>> The management chain from Gremlin to the CEO
```

Comparing SQL and Gremlin

```
SELECT TOP (5) [t14].[ProductName]
  FROM (SELECT COUNT(*) AS [value],
        [t13].[ProductName]
    FROM [customers] AS [t0]
  CROSS APPLY (SELECT [t9].[ProductName]
        FROM [orders] AS [t1]
      CROSS JOIN [order details] AS [t2]
      INNER JOIN [products] AS [t3]
          ON [t3].[ProductID] = [t2].[ProductID]
      CROSS JOIN [order details] AS [t4]
      INNER JOIN [orders] AS [t5]
          ON [t5].[OrderID] = [t4].[OrderID]
       LEFT JOIN [customers] AS [t6]
          ON [t6].[CustomerID] = [t5].[CustomerID]
      CROSS JOIN ([orders] AS [t7]
            CROSS JOIN [order details] AS [t8]
            INNER JOIN [products] AS [t9]
              ON [t9].[ProductID] = [t8].[ProductID])
        WHERE NOT EXISTS(SELECT NULL AS [EMPTY]
                FROM [orders] AS [t10]
              CROSS JOIN [order details] AS [t11]
              INNER JOIN [products] AS [t12]
                  ON [t12].[ProductID] = [t11].[ProductID]
                WHERE [t9].[ProductID] = [t12].[ProductID]
                  AND [t10].[CustomerID] = [t0].[CustomerID]
                  AND [t11].[OrderID] = [t10].[OrderID])
          AND [t6].[CustomerID] <> [t0].[CustomerID]
          AND [t1].[CustomerID] = [t0].[CustomerID]
          AND [t2].[OrderID] = [t1].[OrderID]
          AND [t4].[ProductID] = [t3].[ProductID]
          AND [t7].[CustomerID] = [t6].[CustomerID]
          AND [t8].[OrderID] = [t7].[OrderID]) AS [t13]
      WHERE [t0].[CustomerID] = N'ALFKI'
    GROUP BY [t13].[ProductName]) AS [t14]
ORDER BY [t14].[value] DESC
```

vs

```
g.V().has("customerId", x).as("customer").
      out("ordered").out("contains").
      out("is").aggregate("products").
      in("is").in("contains").in("ordered").
      where(neq("customer")).out("ordered").
      out("contains").out("is").
      where(without("products")).
      groupCount().order(local).by(valueDecr)
```

DSE Graph

- A scalable, distributed graph database optimized for storing, traversing, and querying complex graph data in real- time
- Value data between relationships
- DSE Analytics and DSE Search integrated
- Perfect for use cases:
  - Customer 360
  - Recommendations
  - Fraud Detection

DSE Graph Enhancements

- Performance increases due to thread-per-core and other optimizations.
- Increased unification with DSE Analytics; improved graph analytics engine.

DataStax ENTERPRISE 6

When to Use DSE Graph

Problem Characteristics

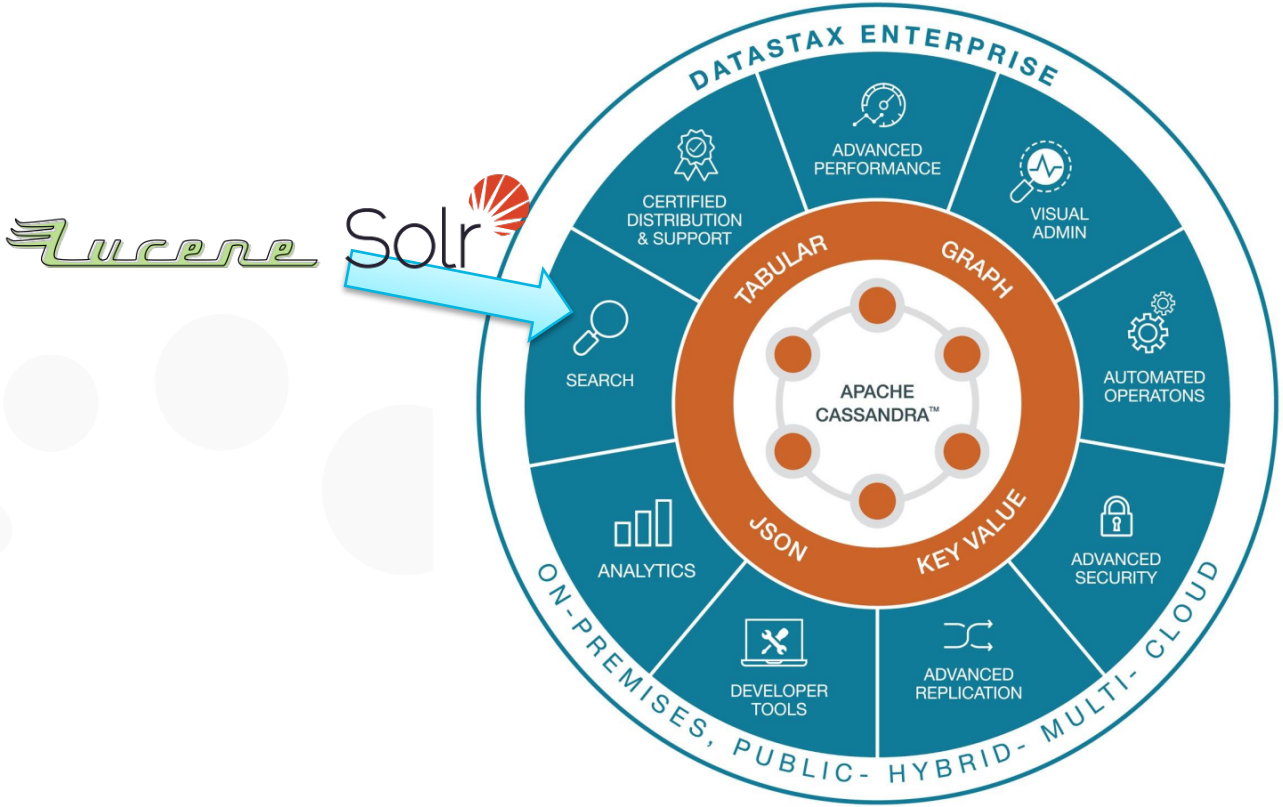
- Relationships between elements are as important as the elements themselves
- Relationships between elements can be many-to-many
- Have very large graph/s
- Need fast traversals by many users

Use Cases

- Customer 360--understand customer fully in real time
- Social Media analysis
- Fraud detection
- HealthCare
- Recommendation Engine

Demo: Recommendations with KillRVideo

OSS Foundations, Integrated for the Enterprise



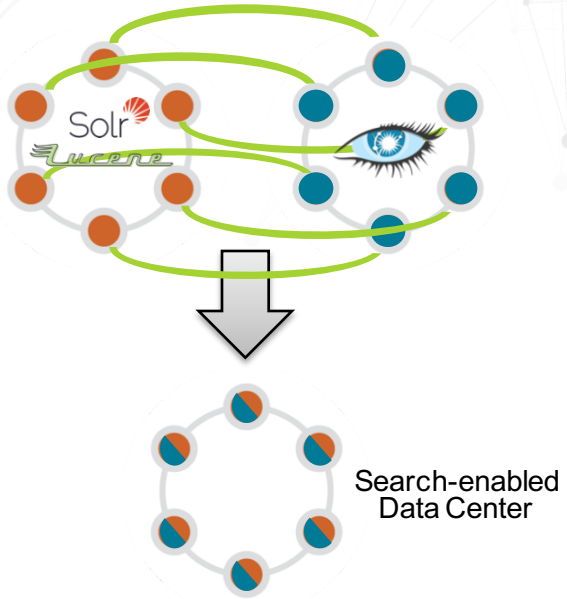
Apache Cassandra, Spark, Lucene, Solr, TinkerPop ® Apache Software Foundation

Demo: Geospatial-enabled queries with DSE Search

DSE Search: Apache Solr / Lucene + Cassandra

- Supports ad-hoc queries not supported by Cassandra
  - Full text search, Faceting, Suggester
  - Geospatial Search
- Live indexing engine
  - Automatic indexing on insert
  - Higher ingestion throughput
  - Distributed query optimization
- Compared to self-managed:
  - No separate search cluster to manage
  - No ETL or sync to build and maintain

- Search indexes co-located with Cassandra



DSE Search Queries

- SOLR Based Queries (releases through DSE 5.1)

```
SELECT title, release_year FROM killrvideo.videos WHERE solr_query =
'title:inception';
```

- CQL Search Queries (new in DSE 6)

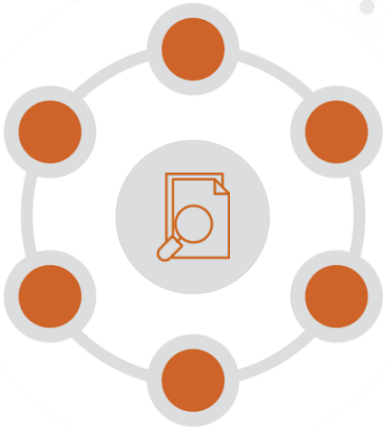
```
SELECT title, release_year FROM killrvideo.videos WHERE title="inception";
```

- New additional keywords IN, LIKE supported in CQL

```
SELECT * FROM killrvideo.videos WHERE title LIKE "incept%";
```

DSE Search Enhancements

- CQL-enabled search
- Like queries
- Eliminate Solr syntax in CQL
- Support for satisfying standard CQL index queries with DSE Search
  - Predicates
  - Sorting
  - Limits



DataStax ENTERPRISE 6

When to Use DSE Search

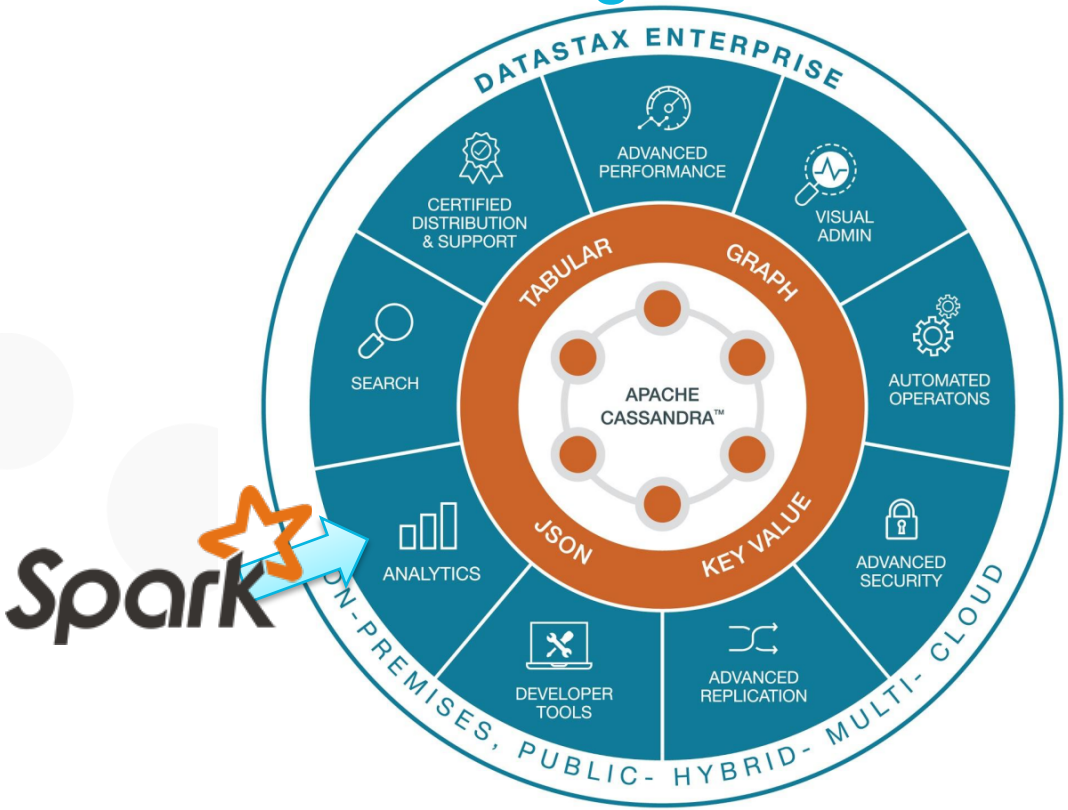
Problem Characteristics

- Need to perform queries not supported by Cassandra
- Need to search by a column not in the Primary Key
- Do not want a separate cluster for Search
- Large text fields

Use Cases

- Searching product catalogs
- Ad-Hoc Reporting
- Document Repositories
- Type-ahead search, autocompletion
- Geospatial search -- nearest Starbucks to me!
- Log analysis

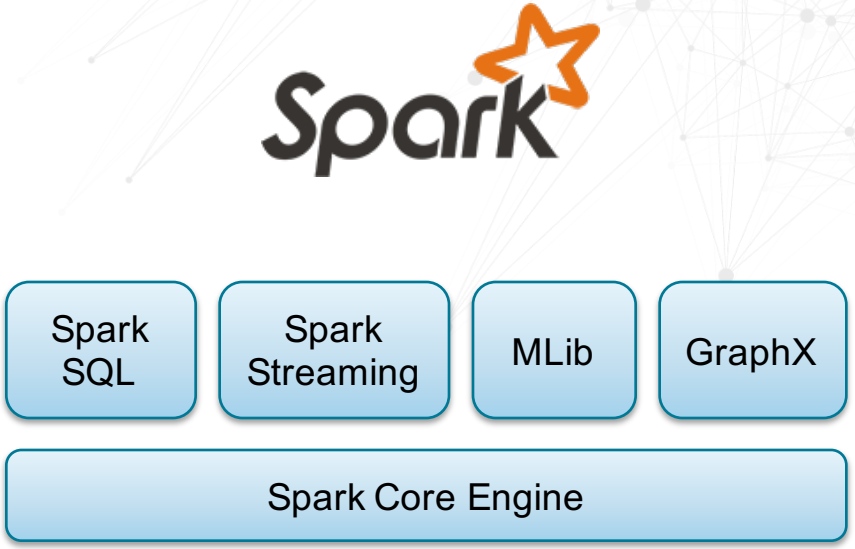
OSS Foundations, Integrated for the Enterprise



Demo: Working with SparkSQL

Apache Spark at a Glance

- Distributed computing framework
  - Similar cluster structure
- Generalized DAG (Directed Acyclic Graph) execution
- Easy Abstraction for Datasets
- Integrated SQL Queries
- Streaming
- Machine Learning Library



Apache Spark ® Apache Software Foundation

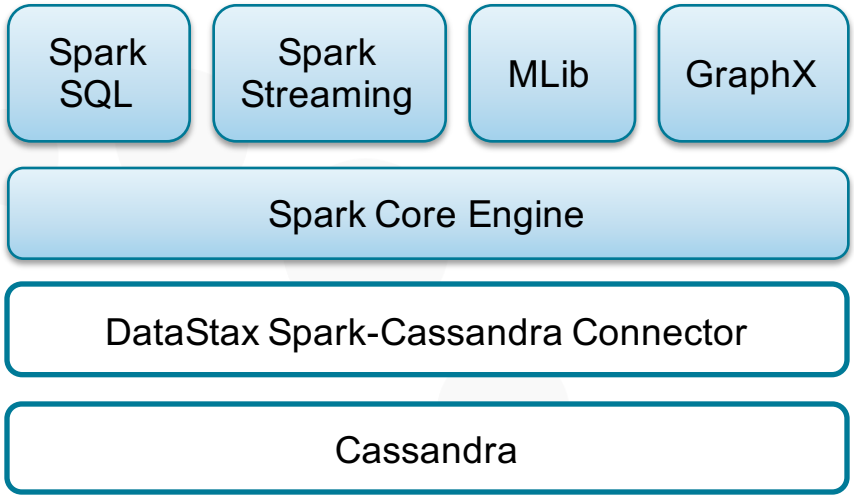
DSE Analytics: SparkSQL

- SQL query engine on top of Spark
  - Hive compatible (JDBC, UDFs, types, metadata, etc.)
  - Support for in-memory processing
  - Pushdown of predicates to Cassandra when possible
- Provide a single interface for working with structured data including Apache Hive, Parquet, JSON files and any DataSource
- Friendly abstraction layer for Spark Batch

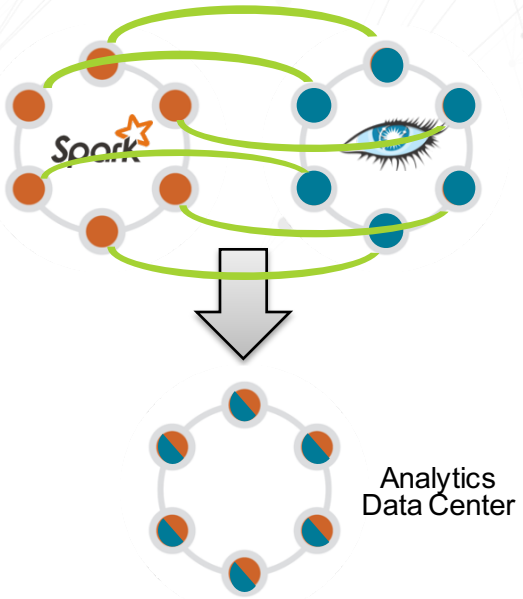
```
SELECT merchant,
  sum(amount) AS total_amount,
  count(amount) AS total_trans
  FROM dsbank.transactions
  GROUP BY merchant
  ORDER BY total_amount desc
  limit 10;
```

DSE Analytics: Spark + Cassandra

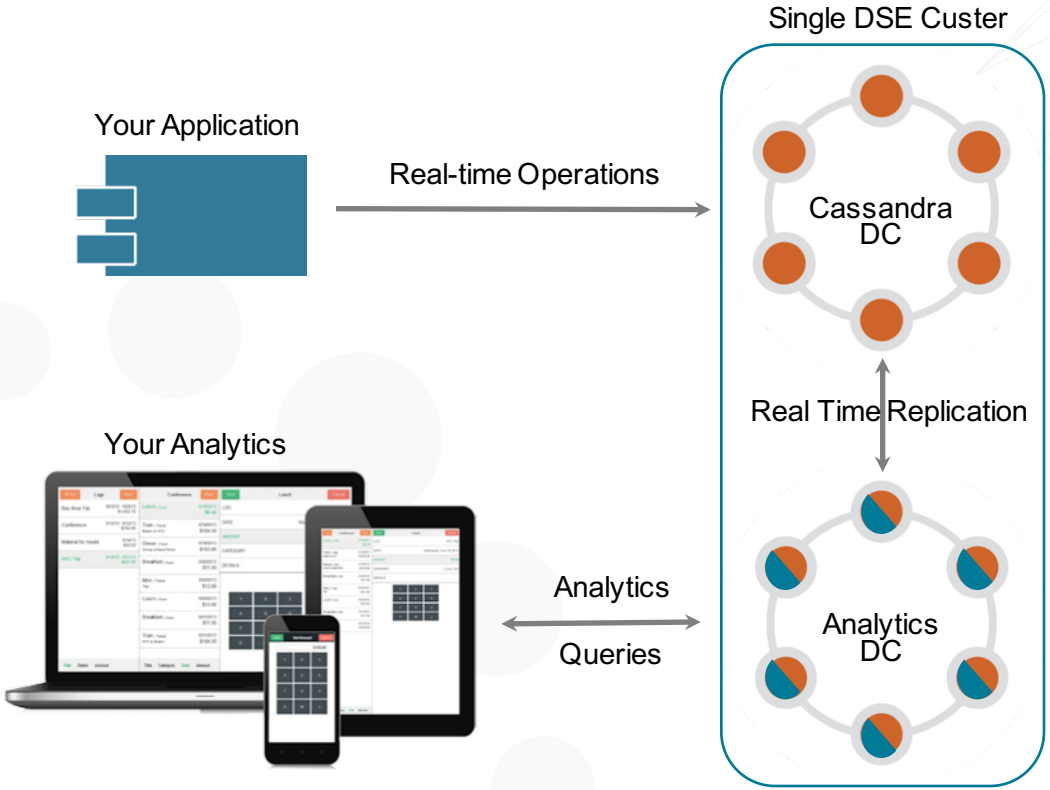
Read/write Cassandra data from Spark via DataStax Connector



Co-located Spark with Cassandra



Deploying DSE Analytics for Real-time Results



Streaming, ad-hoc, and batch

- High-performance
- High availability
- Workload management
- SQL reporting

Compared to self-managed:

- No ETL
- True HA without Zookeeper

DSE Analytics Enhancements

- DSE Spark Connector with:
  - Increased performance and optimizer enhancements
  - Custom join strategies
  - Implicit use of DSE Search
  - Structured Streaming to Cassandra
- AlwaysOn SQL support for Spark SQL via ODBC/JDBC
  - Supports external BI/ETL/other tool connectivity
  - Unified Authentication with DSE
  - Security improvements via proxy execution
- Apache Spark 2.2 including Structured Streaming

DataStax ENTERPRISE 6

When to Use DSE Analytics

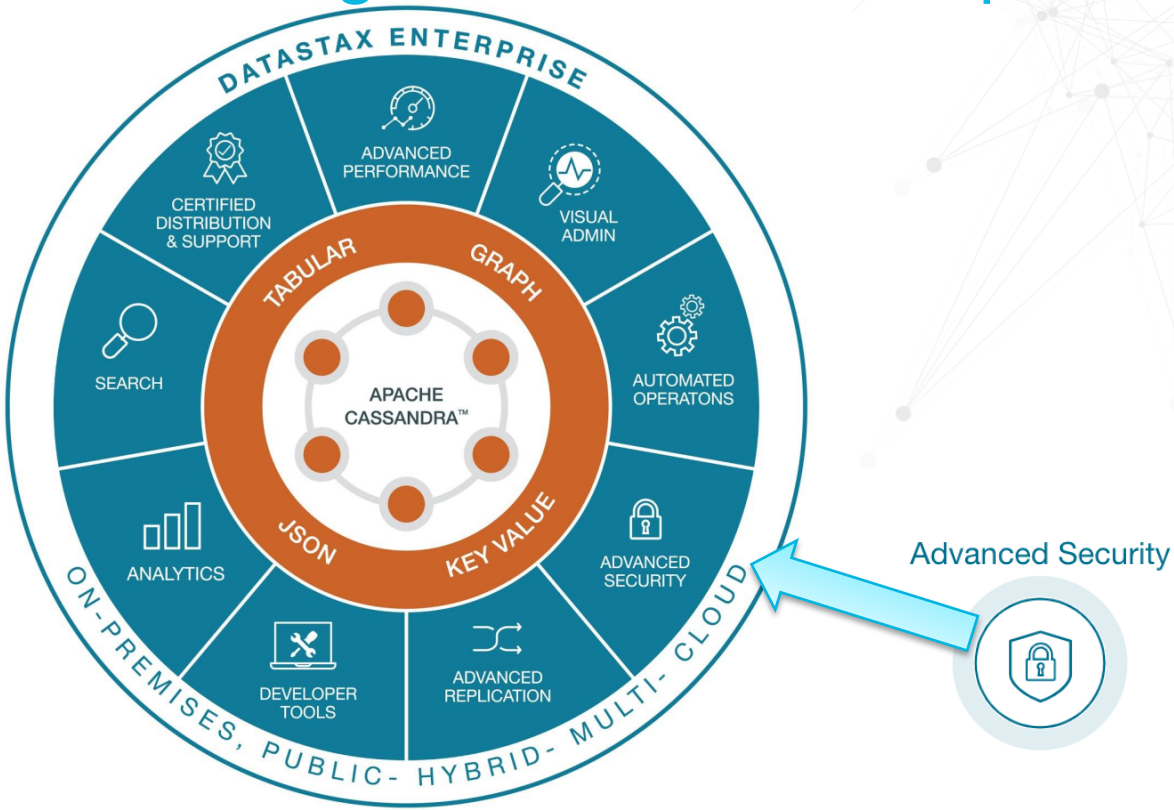
Problem Characteristics

- Don’t know queries in advance
- Need to be able to transform, ingest, and perform analytics on data NOW
- Don’t want to move my data to perform analytics --no ETL
- Have both historical and streaming data
- I need to migrate my DSE tables

Use Cases

- Personalization
- Customer 360
- Recommendation Engines
- Complex Event Processing
- Model building with Spark MLlib
- Online Learning

OSS Foundations, Integrated for the Enterprise



Apache Cassandra, Spark, Lucene, Solr, TinkerPop ® Apache Software Foundation

DSE Advanced Security

External Authentication

- External validation of authorized users
- Leverages Kerberos & LDAP/AD
- Single sign-on to all data domains

Transparent Data Encryption

- Protects sensitive data at rest via encryption
- No changes needed at app level

Data Auditing

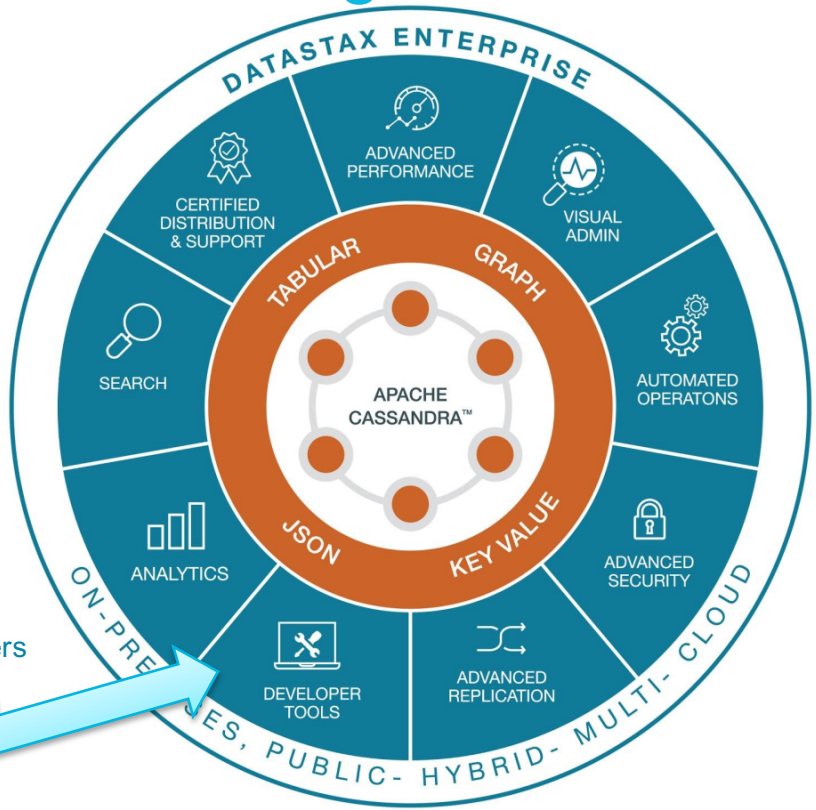
- Audit trail of all accesses and changes
- Control to audit only what's needed
- Uses log4j interface to ensure performance & efficient audit

DSE Advanced Security Enhancements

- Privatized schemas – protects against unauthorized viewing of database schemas
- Secured administrator – allows administration work without data access

DataStax ENTERPRISE 6

OSS Foundations, Integrated for the Enterprise

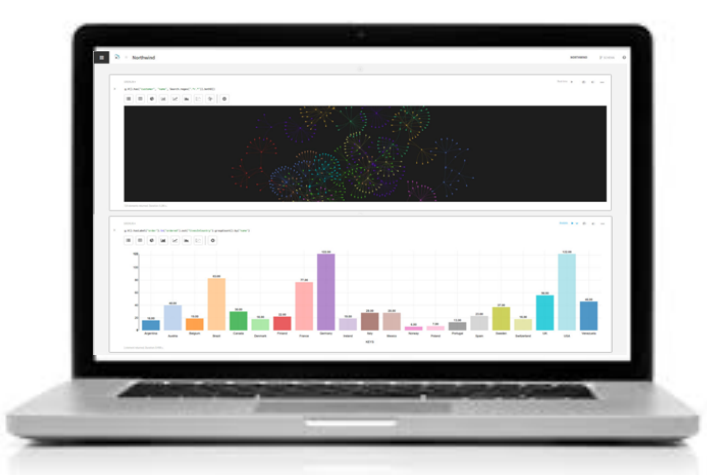


Tools for Developers

Apache Cassandra, Spark, Lucene, Solr, TinkerPop ® Apache Software Foundation

DataStax Studio

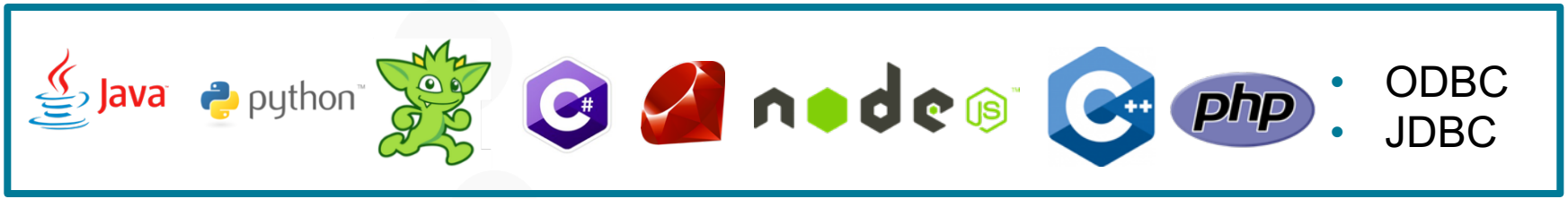
Explore, query, and analyze CQL, DSE Graph, and Gremlin Query Language



- Explore and query CQL, Gremlin Query Language, and Spark SQL
- Auto-completion, result set visualization, execution management, and much more
- Friendly fluent API Graph and data visualization

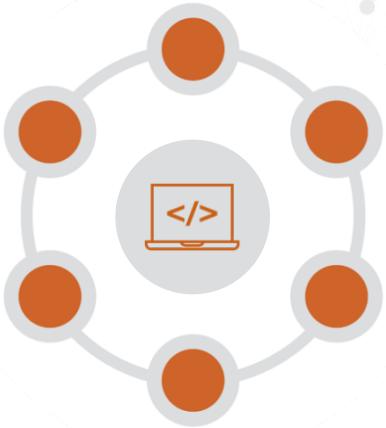
DataStax Drivers

- DataStax Cassandra Drivers (OSS)
  - CQL Support
  - Sync / Async API
  - Address Translation
  - Load Balancing Policies
  - Retry Policies
  - Reconnection Policies
  - Connection Pooling
  - Auto Node Discovery
  - SSL
  - Compression
  - Query Builder
  - Object Mapper
- DataStax Enterprise Drivers
  - OSS Drivers capabilities plus Enterprise improvements for
    - Performance, Usability, Scalability, Ecosystem
  - DSE Advanced Security, Unified Authentication
  - DSE Graph Fluent API
  - DSE Geometric Types



DataStax Studio and Drivers Enhancements

- DSE Analytics/Spark support in DataStax Studio
- Export/Import of DataStax Studio notebooks
- Fluent API support in drivers for DSE Graph



DataStax ENTERPRISE 6

DataStax Bulk Loader

Move data in and out of DSE

- Simplifies data loading and unloading
- Supports CSV and JSON formats
- Command-line interface
- Available as part of DSE and standalone
- Up to 4x faster than community utilities

DSE OpsCenter

Visual management tool for DSE

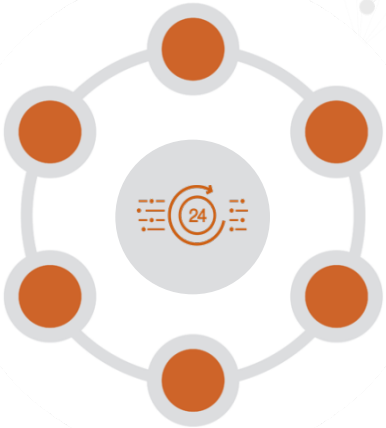
"... OpsCenter allows us to act on things before they become a problem."

SAMSUNG SmartThings

- Control automatic management services including transparent repair
- Manage and schedule backup and restore operations
- Perform capacity planning with historical trend analysis and forecasting capabilities
- Proactively manage all clusters with threshold and timing-based alerts
- Visually create new clusters or upgrade existing clusters with a few mouse clicks either on premise or in the cloud
- Integrate with the enterprise landscape
- Built-in Automatic Failover

DSE OpsCenter 6.5 Enhancements

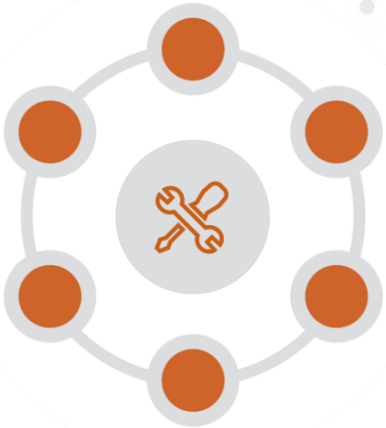
- Backup Service improvements
- DSE 6 feature support
  - Advanced Performance metrics
  - DSE Analytics AlwaysOn Service status
  - Support for NodeSync
  - Upgrade Service



DataStax ENTERPRISE 6

DSE NodeSync

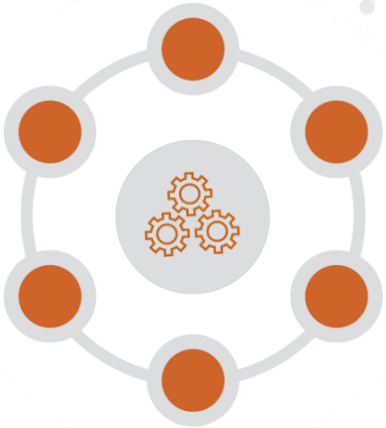
- Automatically resolves data inconsistencies among nodes
- Little to no need for manual repair operations intervention
- Runs continuously and efficiently on the cluster
- Part of DSE Server foundation



DataStax ENTERPRISE 6

DSE Upgrade Service

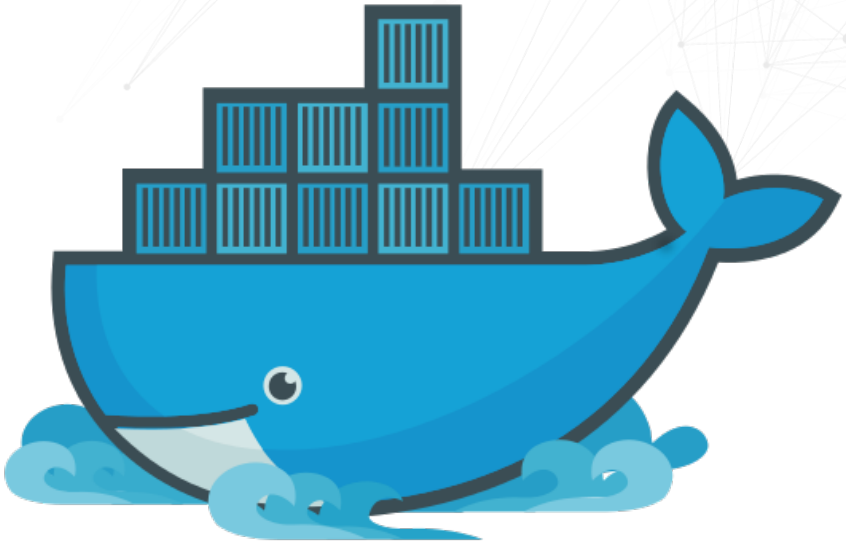
- Automatically keeps DSE clusters up to date with latest patch releases
- Downloads software updates onto servers or staging area
- Applies updates in rolling restart fashion so no downtime is experienced



DataStax ENTERPRISE 6

DataStax Enterprise Docker Images

- DataStax Docker Images for Non-Production available on Docker Hub
  - DSE Server
  - OpsCenter
  - Studio
- https://hub.docker.com/u/datastax/



Management and hosting options

OpsCenter

Any Cloud

Self-Managed
And Hosted

Cloud Hosted

What Should I Do Now?

Some resources to learn more

- DataStax Academy
  - DS201 - DataStax Enterprise 6 Foundations of Apache Cassandra
    - https://academy.datastax.com/resources/ds201-datastax-enterprise-6-foundations-of-apache-cassandra
- Follow Us
  - Twitter @AmandaDataStax and @clunven
  - LinkedIn https://www.linkedin.com/in/amanda-kay-moran/ and https://www.linkedin.com/in/clunven/
  - Subscribe to our YouTube channel: https://www.youtube.com/channel/UCAIQY251avaMv7bBv5PCo-A

Questions?