

DataStax Meetups



Make your data speak



Hands-on Codelab



Apache Cassandra

3 hours, intermediate level + dinner break

An initiative by:



Powered by:



Agenda

A faint, abstract network graph with numerous small, light-blue circular nodes connected by thin white lines, forming a complex web-like structure across the slide's background.

A – What is Apache Cassandra and why do you care ?

1. Getting starting with Apache Cassandra™ and use Cases
2. CodeLab : *Getting Started with Apache Cassandra*
3. Apache Spark™ and DataStax Enterprise Analytics

B – Machine Learning with DataStax Enterprise

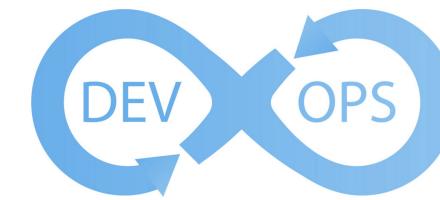
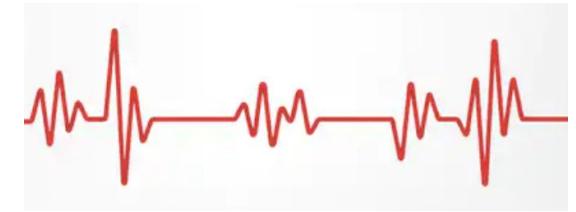
1. CodeLab : *Clustering with K-means*
2. CodeLab : *Classification with Naïve Bayes*
3. CodeLab : *Regression and Classific. with RandomForest*
4. CodeLab : *Recommendation with FP-Growth*
5. CodeLab : *Recommendation with Collaborating Filtering*

C – Resources and next steps

Your Instructors



Cedrick Lunven



Aleks
Volochnev



Before Starting

Hands-on Codelab

 **cassandra** Apache Cassandra
3 hours, intermediate level + dinner break

An initiative by:
 ITALIAN ASSOCIATION FOR MACHINE LEARNING

Powered by:
 **SOURCESENSE** Open Solutions for your Value 

Dear guest,
Welcome to this event organized in collaboration with DataStax and SourceSense. As a codelab we expect you to have your laptops to do the exercises.
The session has been designed for intermediate level software engineers and data scientists. We have a lot to cover and unfortunately not a lot of time to help you installing. Don't worry we made things as simpler as we can, still if you are beginner try to team up !

Prerequisites: To run the exercises you will simply need : Docker (cf link below)

Agenda

PART I – What is Apache Cassandra and why do you care ?

- Getting starting with Apache Cassandra™ and use Cases
- CodeLab : Getting Started with Apache Cassandra
- Apache Spark™ and DataStax Enterprise Analytics

PART II – Machine Learning with DataStax Enterprise

- CodeLab : Clustering with K-means
- CodeLab : Classification with Naïve Bayes
- CodeLab : Regression and Classific. with RandomForest
- CodeLab : Recommendation with FP-Growth
- CodeLab : Recommendation with Collaborating Filtering

Installation

```
git clone https://github.com/HadesArchitect/CaSpark.git
cd CaSpark
docker-compose up -d
docker-compose logs -f jupyter
http://localhost:8888
```

Download Docker	http://download.docker.com/
Github Repository	https://github.com/HadesArchitect/CaSpark.git
DataStax Studio	http://localhost:9091
DataStax Academy	http://academy.datastax.com
DataStax Community	http://community.datastax.com



- You should have one of those sheet.
- Please execute the Installation steps **as soon as possible**. This will download a few docker images that can take some time !

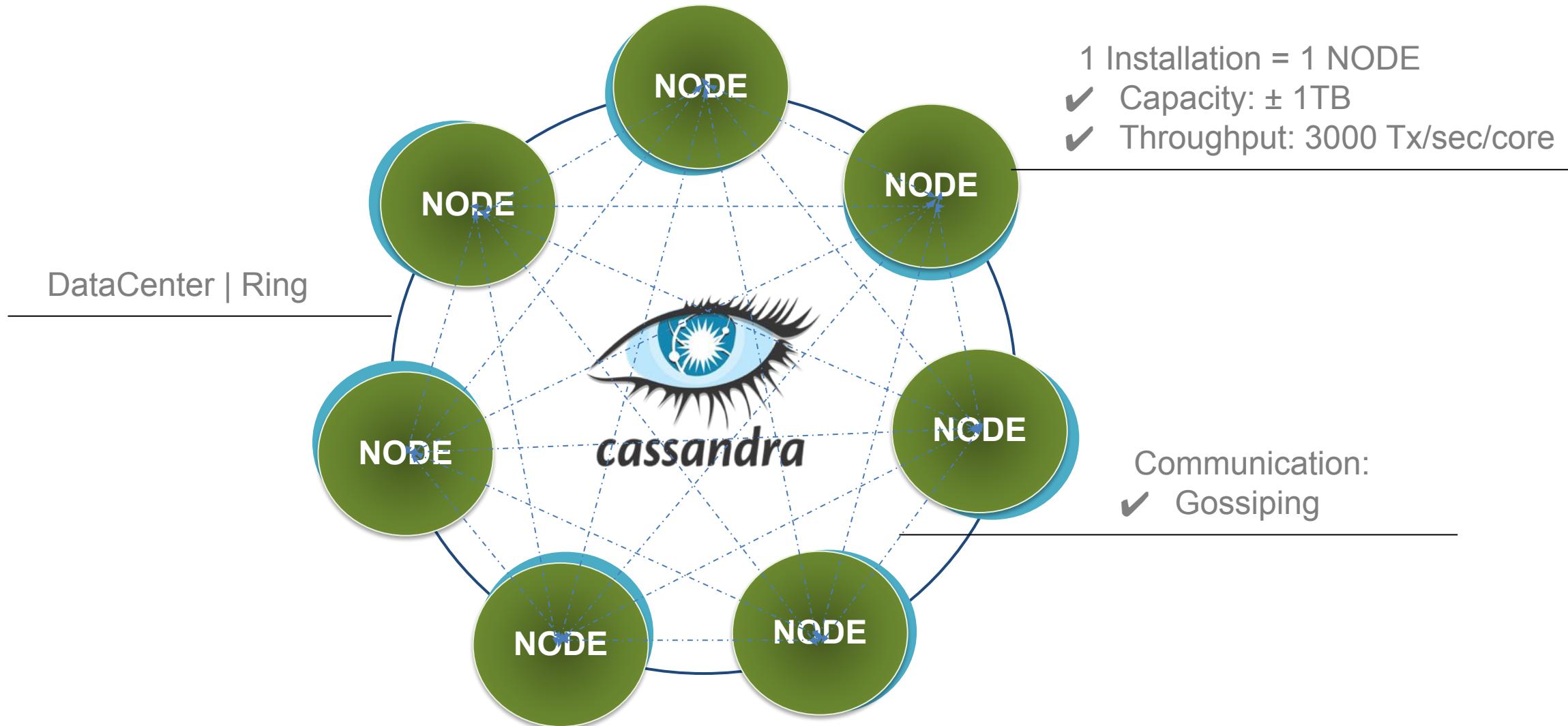
DataStax Meetup



Getting Started with Apache Cassandra

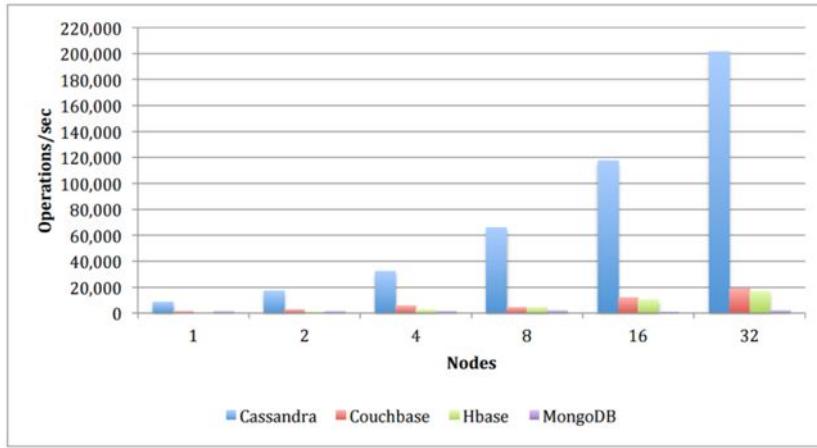


Apache Cassandra™ = Distributed NoSQL Database

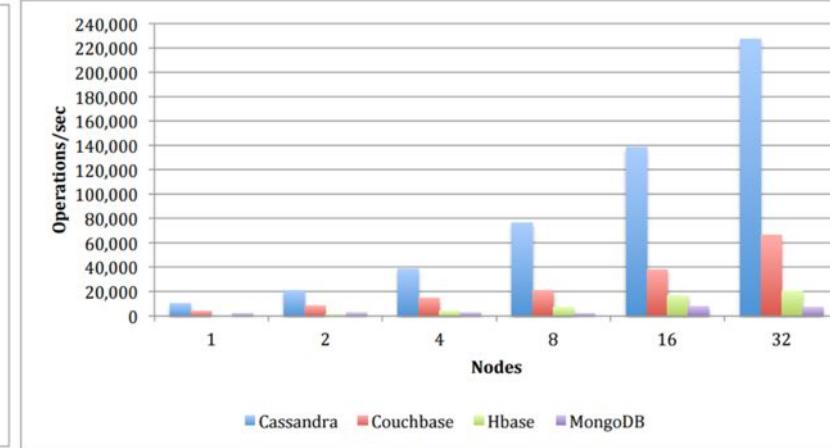


Linear Scalability

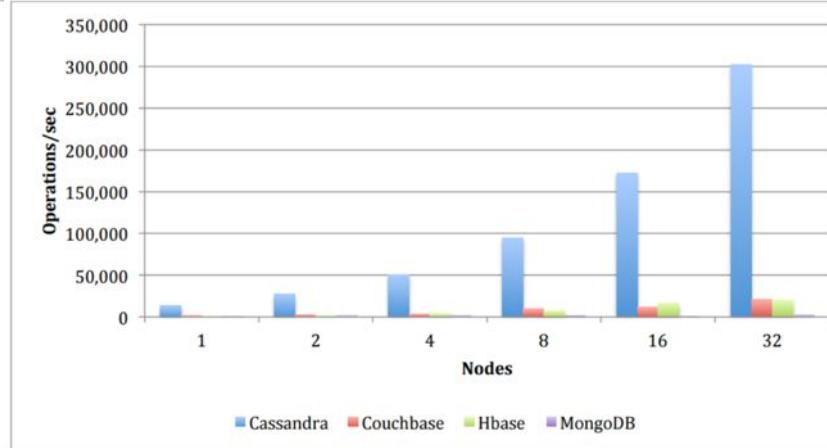
Read-Modify-Write Workload



Read-mostly Workload



Balanced Read/Write Mix



- Need More Capacity ? Add new nodes
- Need more Throughput ? Add new nodes

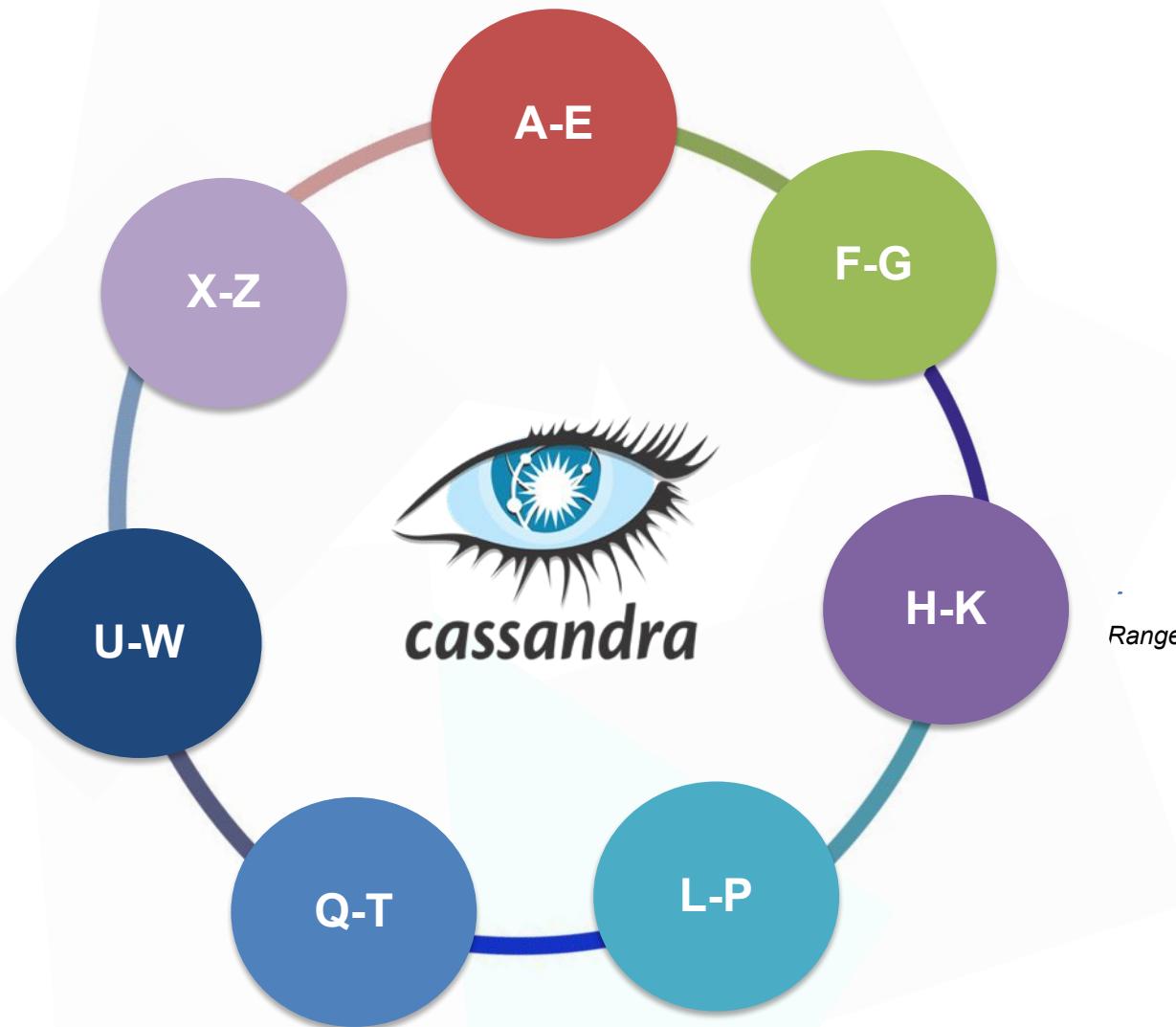
Data is Distributed



Country	City	Habitant
USA	New York	8.000.000
USA	Los Angeles	4.000.000
FR	Paris	2.230.000
DE	Berlin	3.350.000
UK	London	9.200.000
AU	Sydney	4.900.000
DE	Nuremberg	500.000
CA	Toronto	6.200.000
CA	Montreal	4.200.000
FR	Toulouse	1.100.000
JP	Tokyo	37.430.000
IN	Mumbai	20.200.000

Partition Key

Data is Distributed



Data is *Evenly-distributed*

CO	City	Habitant
AU	Sydney	4.900.000
CA	Toronto	6.200.000
CA	Montreal	4.200.000
DE	Berlin	3.350.000
DE	Nuremberg	500.000

Partitioner
Hashing Function

CO	City	Habitant
59	Sydney	4.900.000
12	Toronto	6.200.000
12	Montreal	4.200.000
45	Berlin	3.350.000
45	Nuremberg	500.000

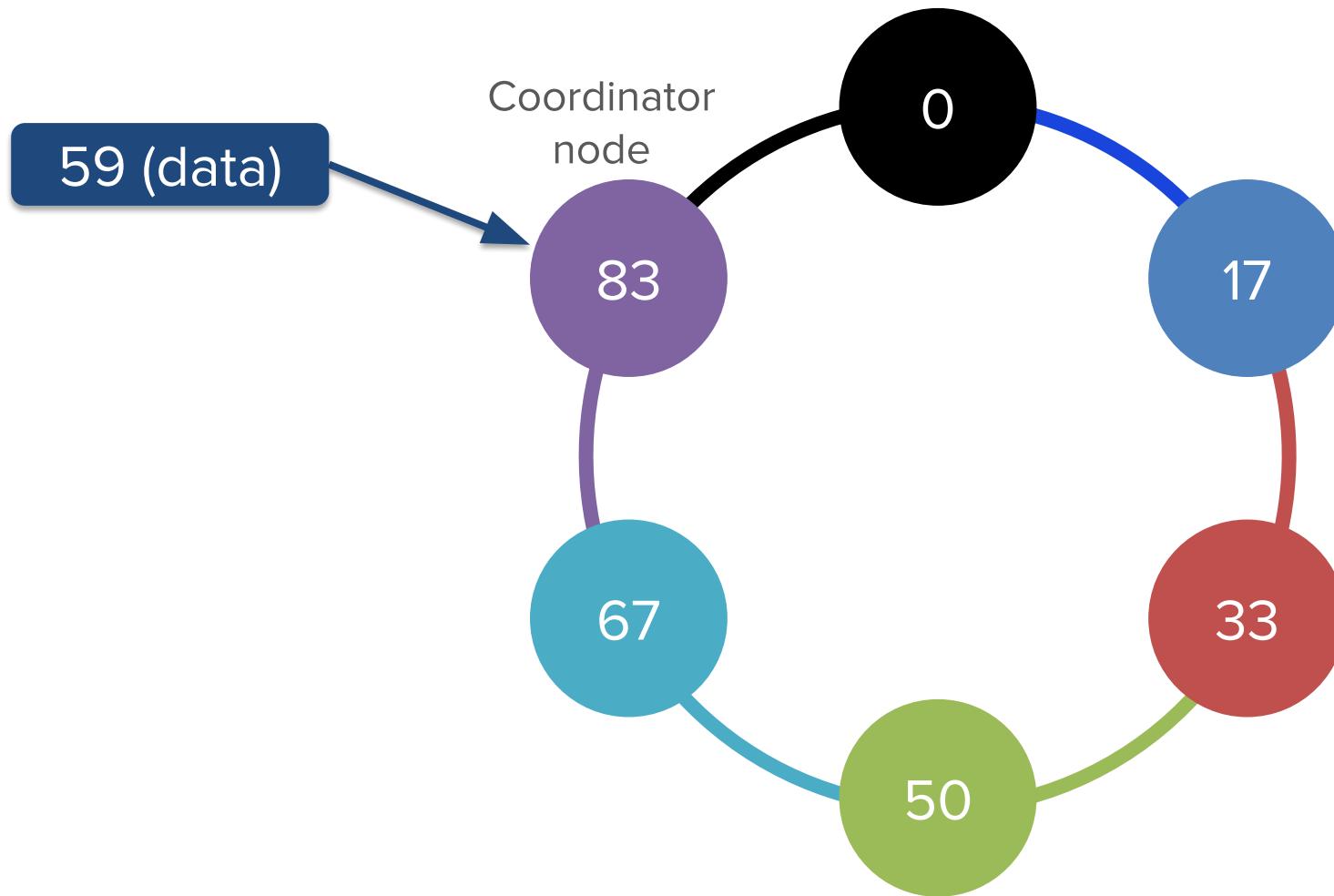


Partition Key

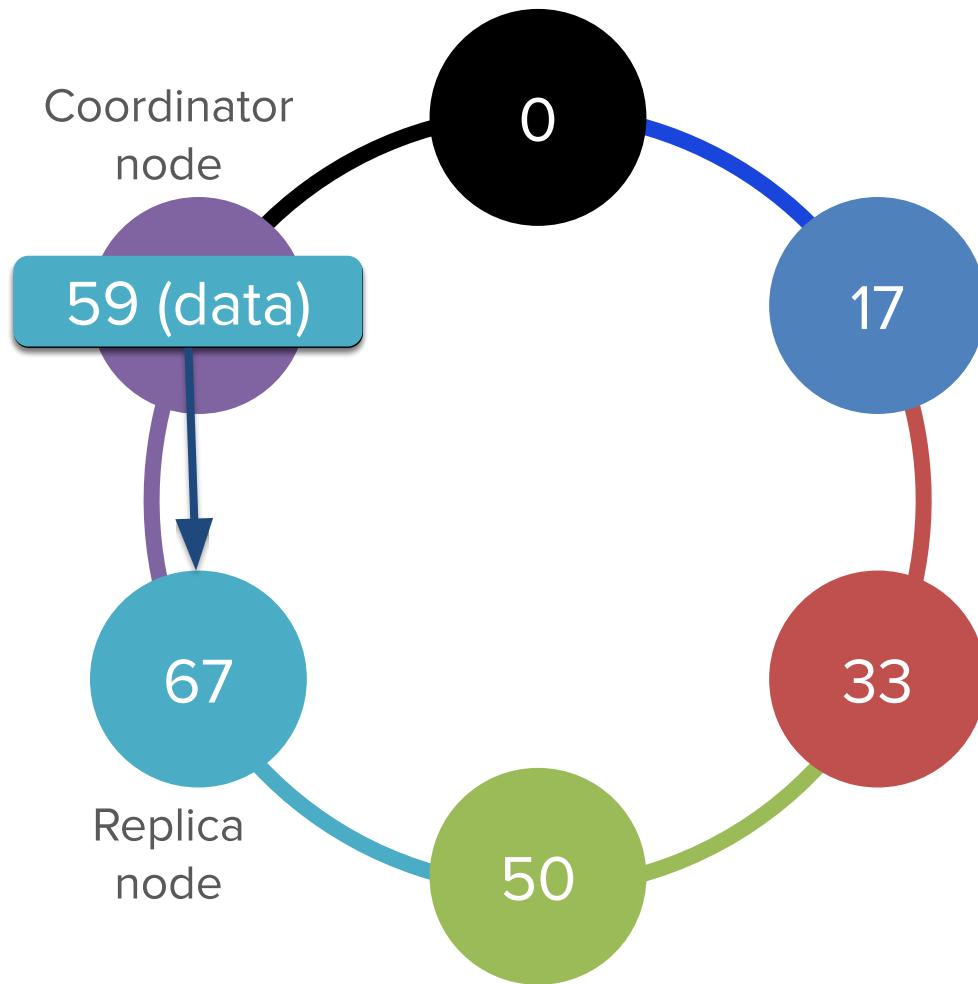


Tokens

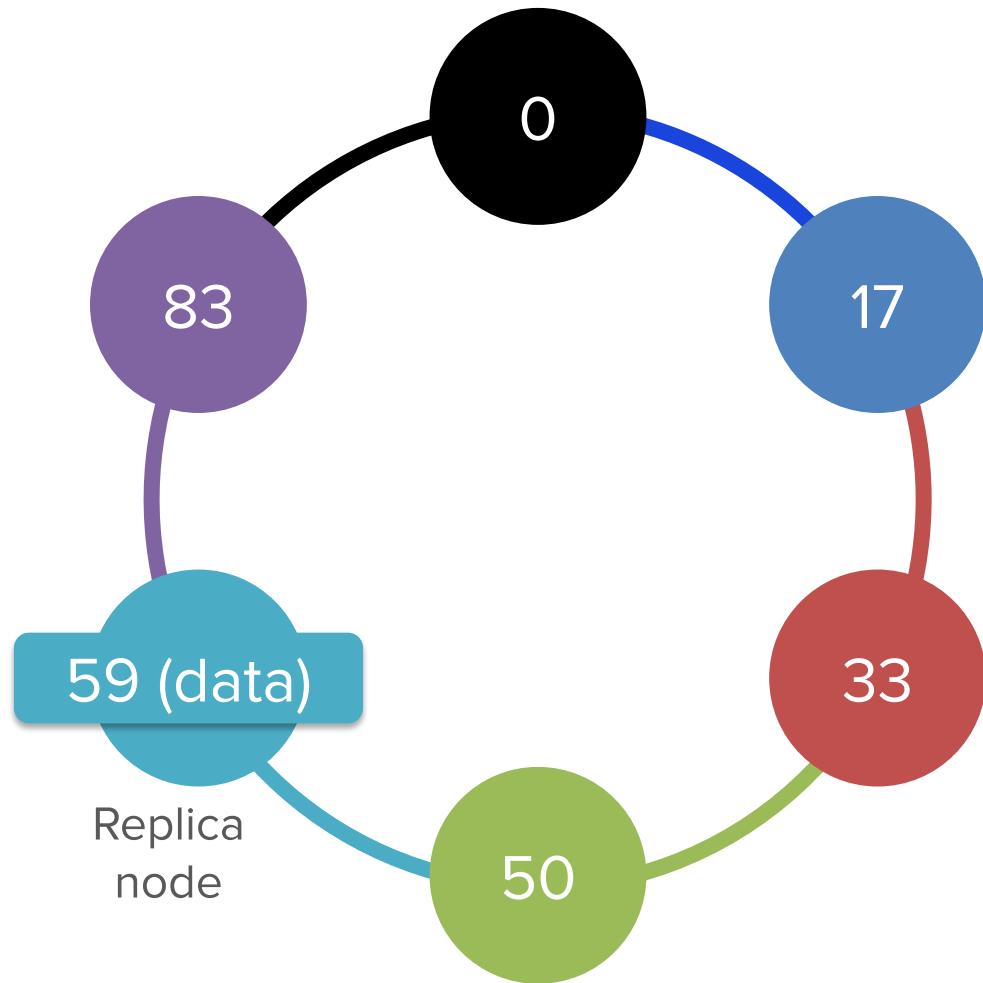
How the Ring Works



How the Ring Works

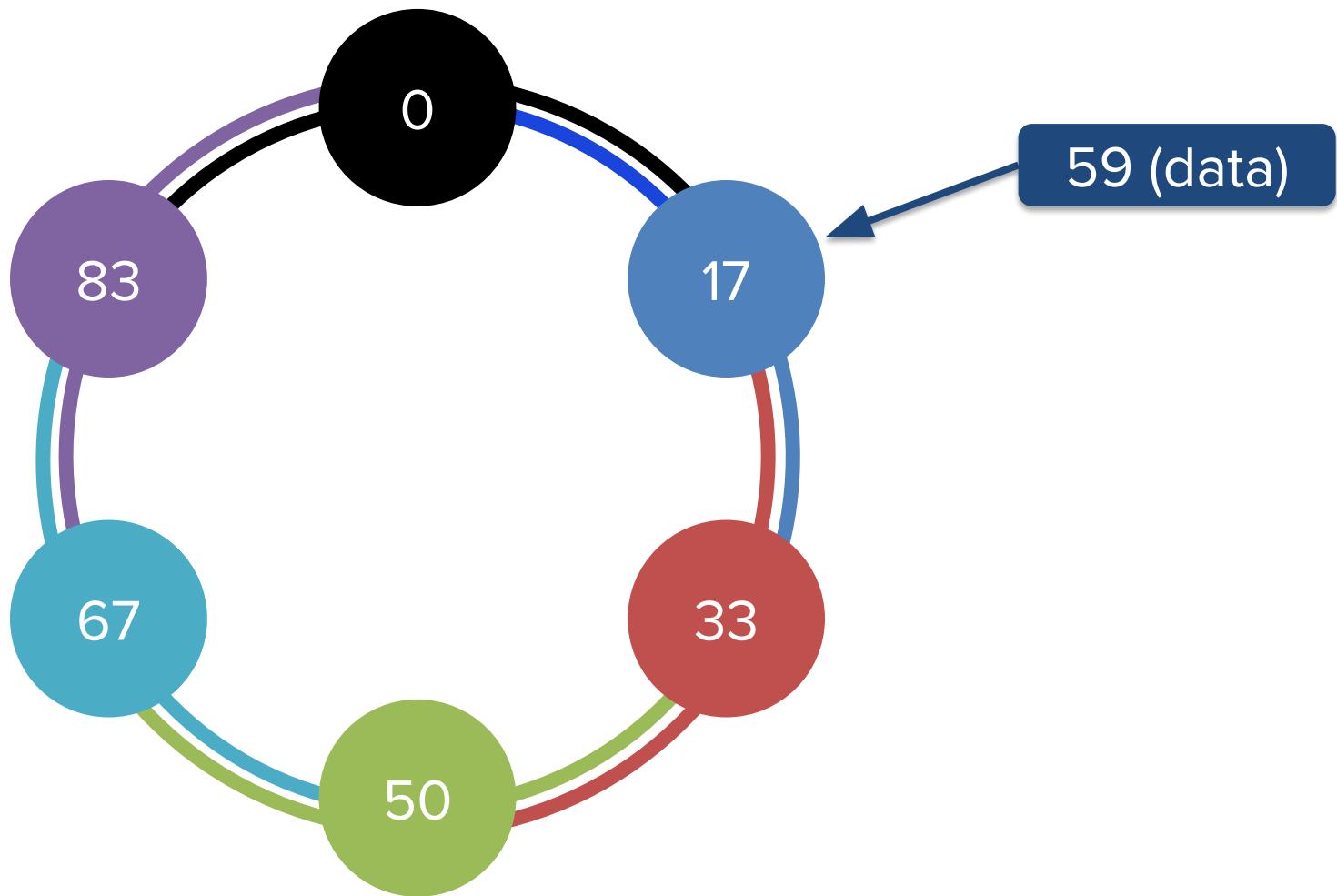


How the Ring Works



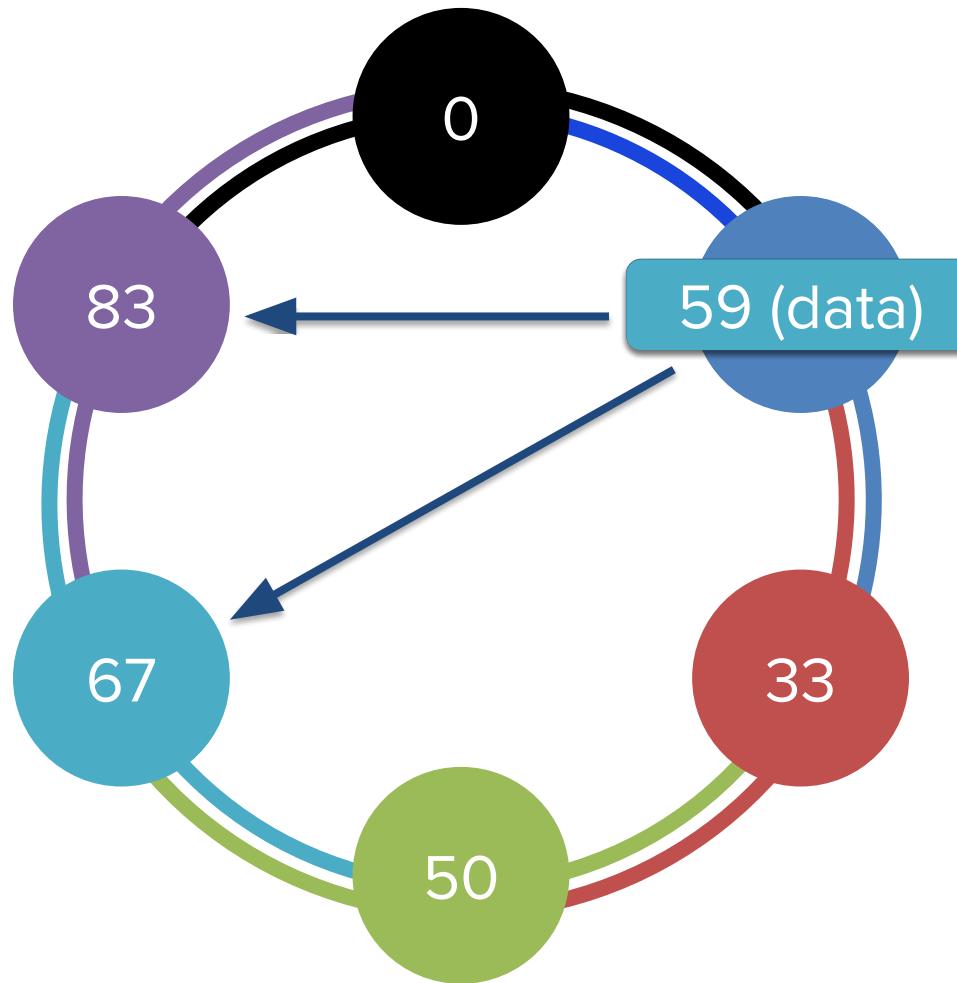
Replication within the Ring

RF = 2



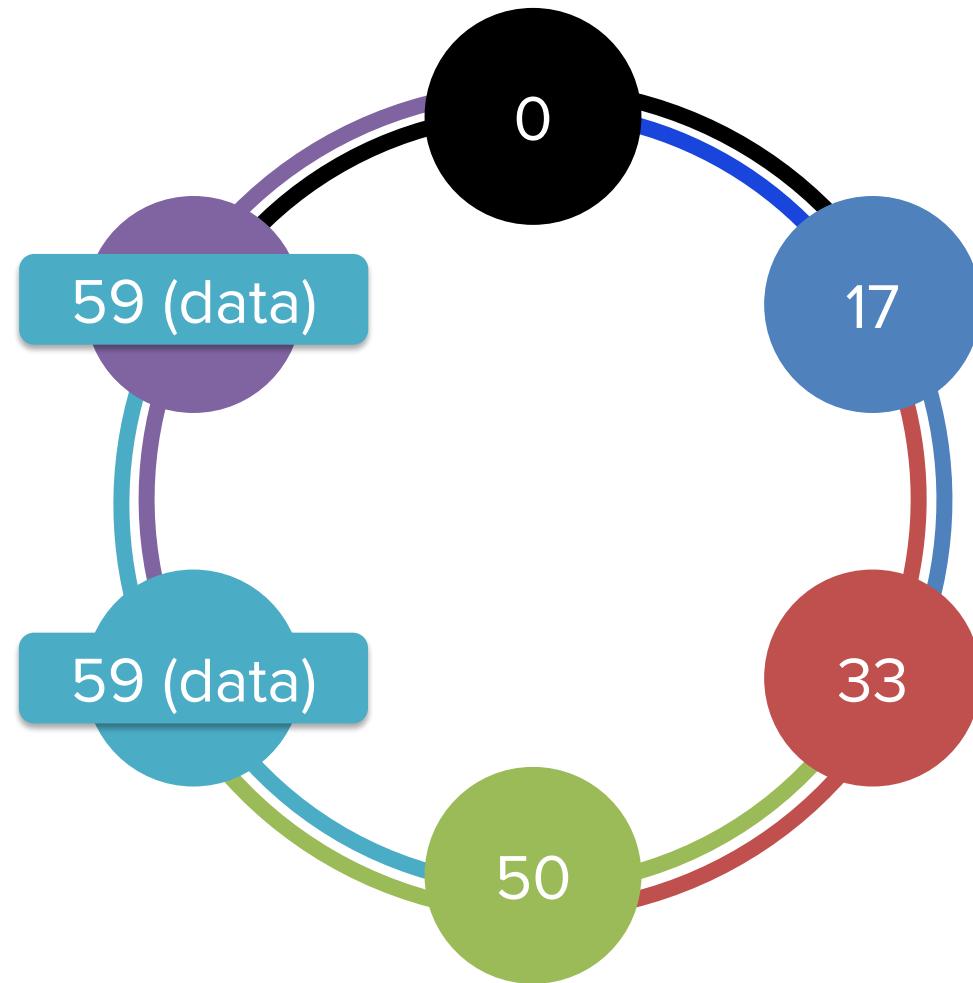
Replication within the Ring

RF = 2



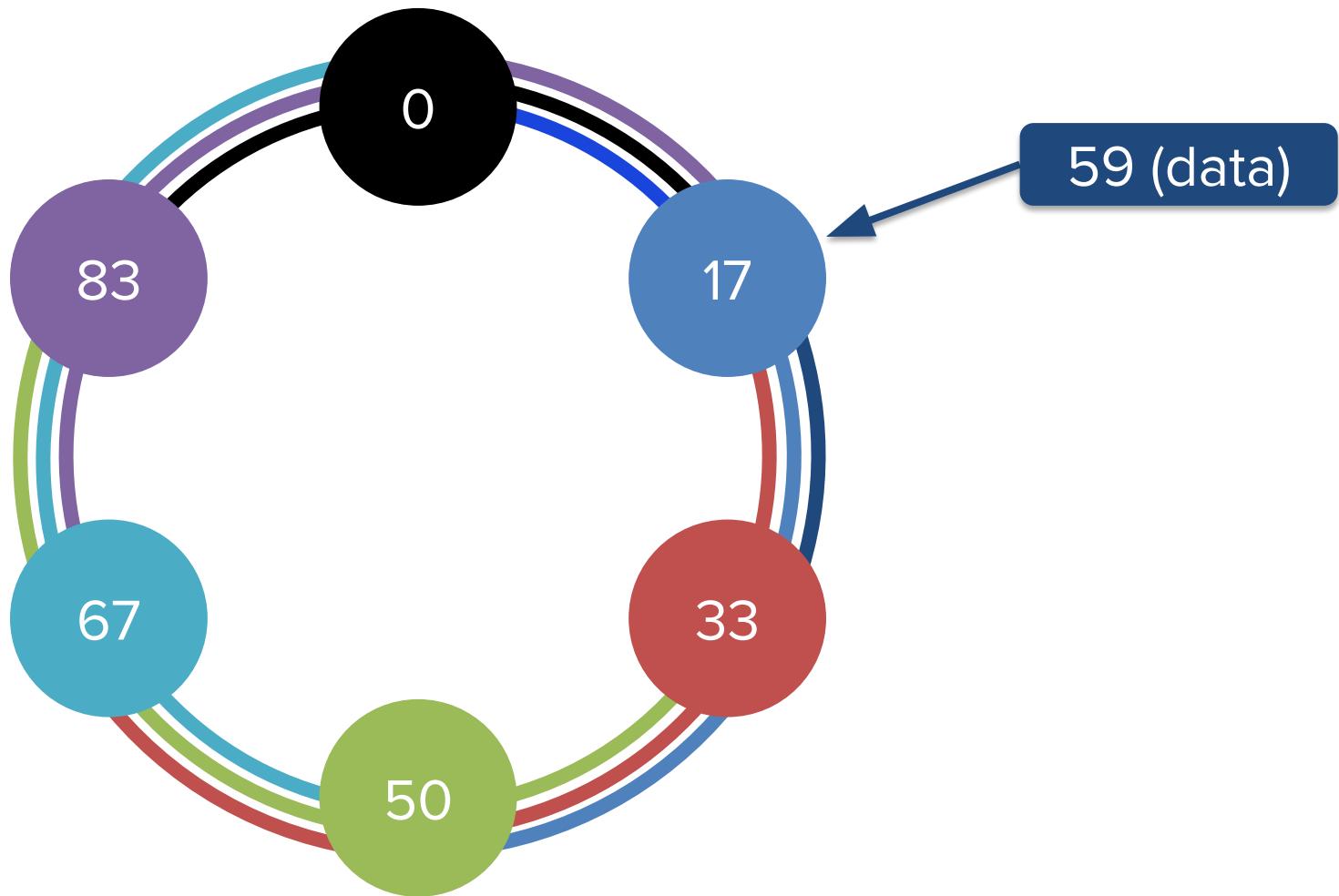
Replication within the Ring

RF = 2



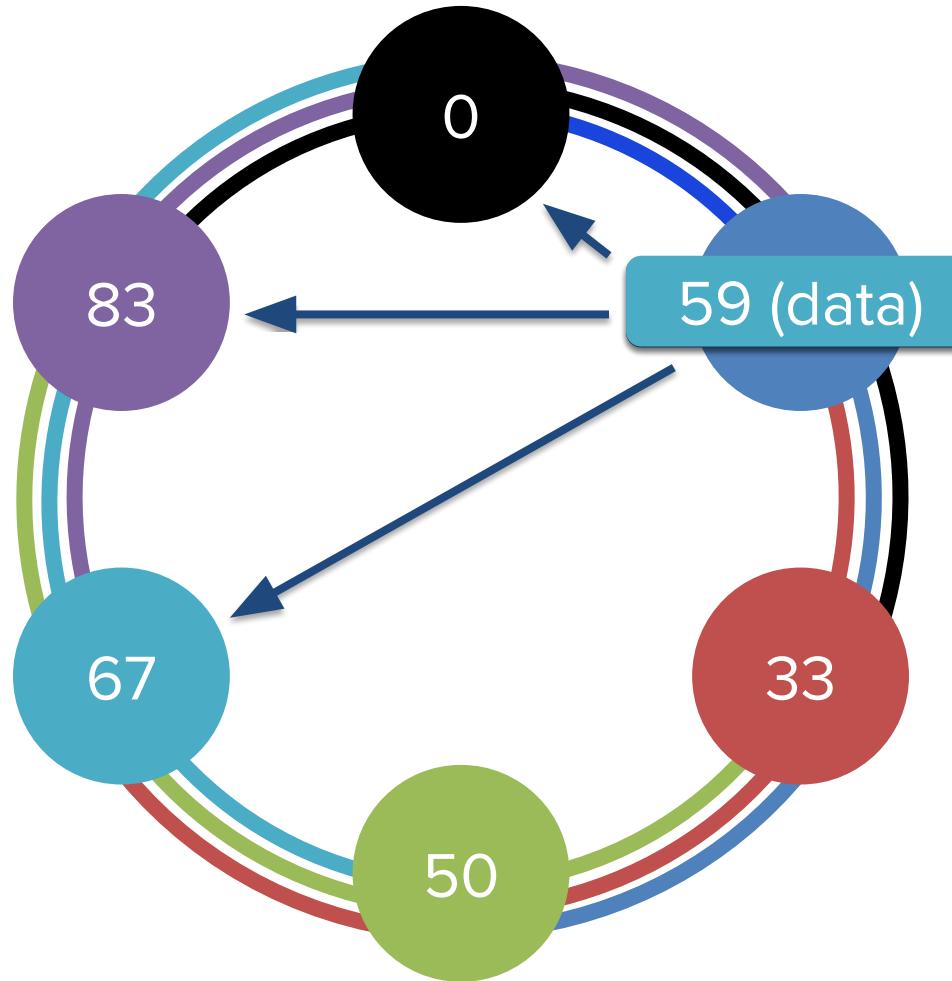
Replication within the Ring

RF = 3



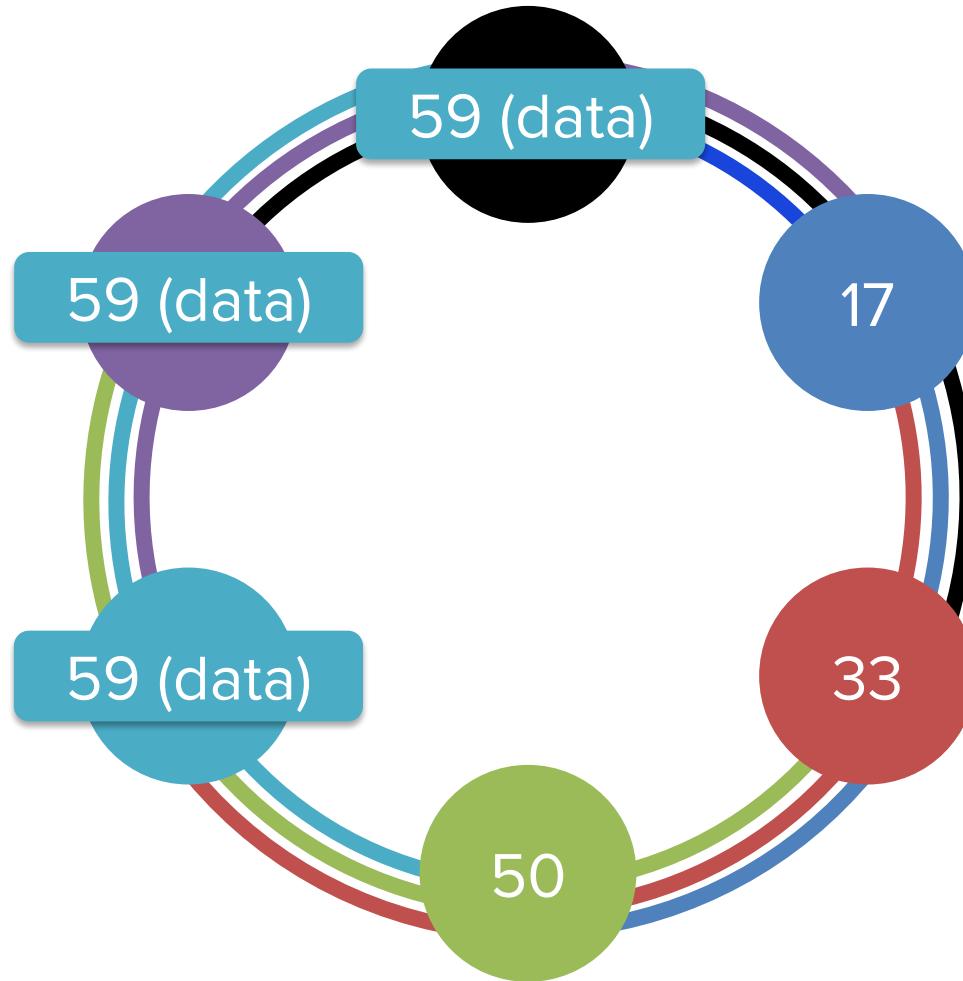
Replication within the Ring

RF = 3



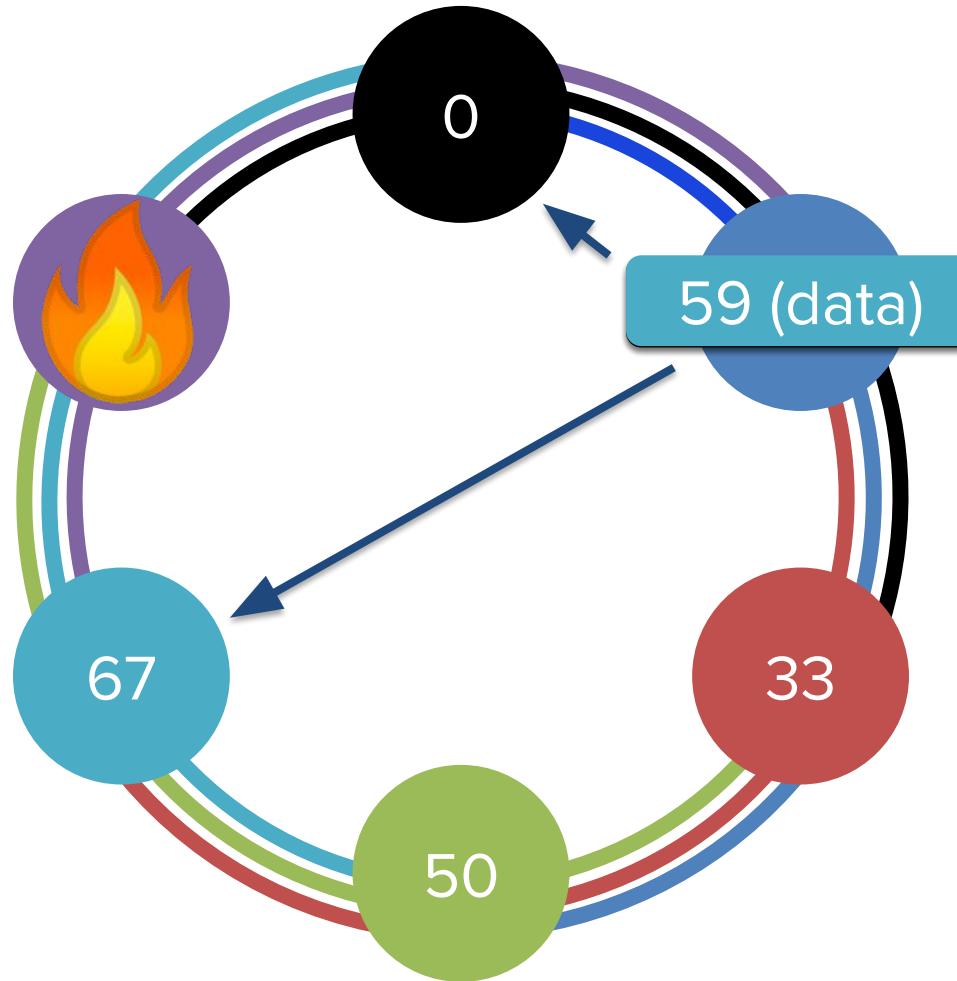
Replication within the Ring

RF = 3



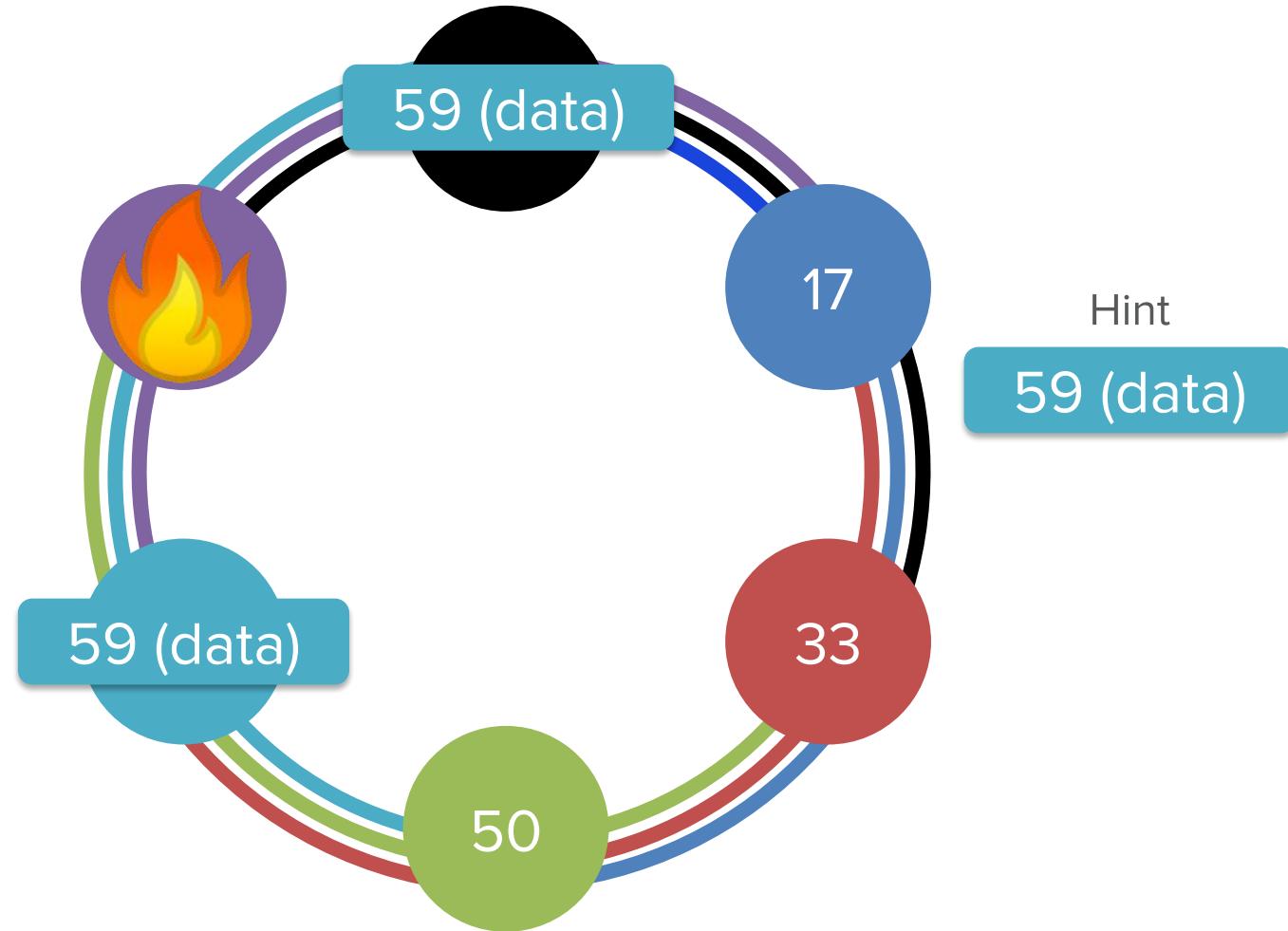
Node Failure

RF = 3



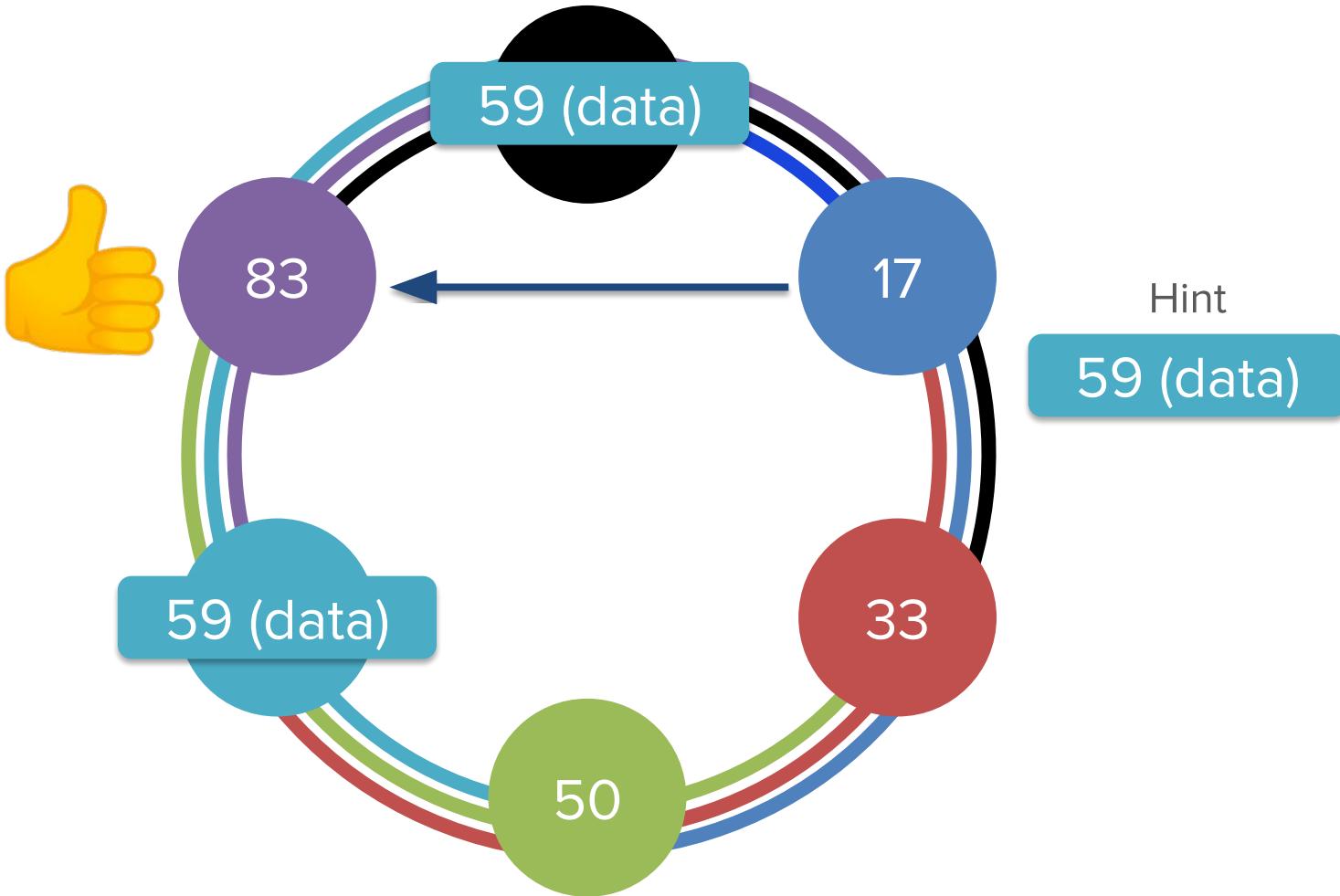
Node Failure

RF = 3



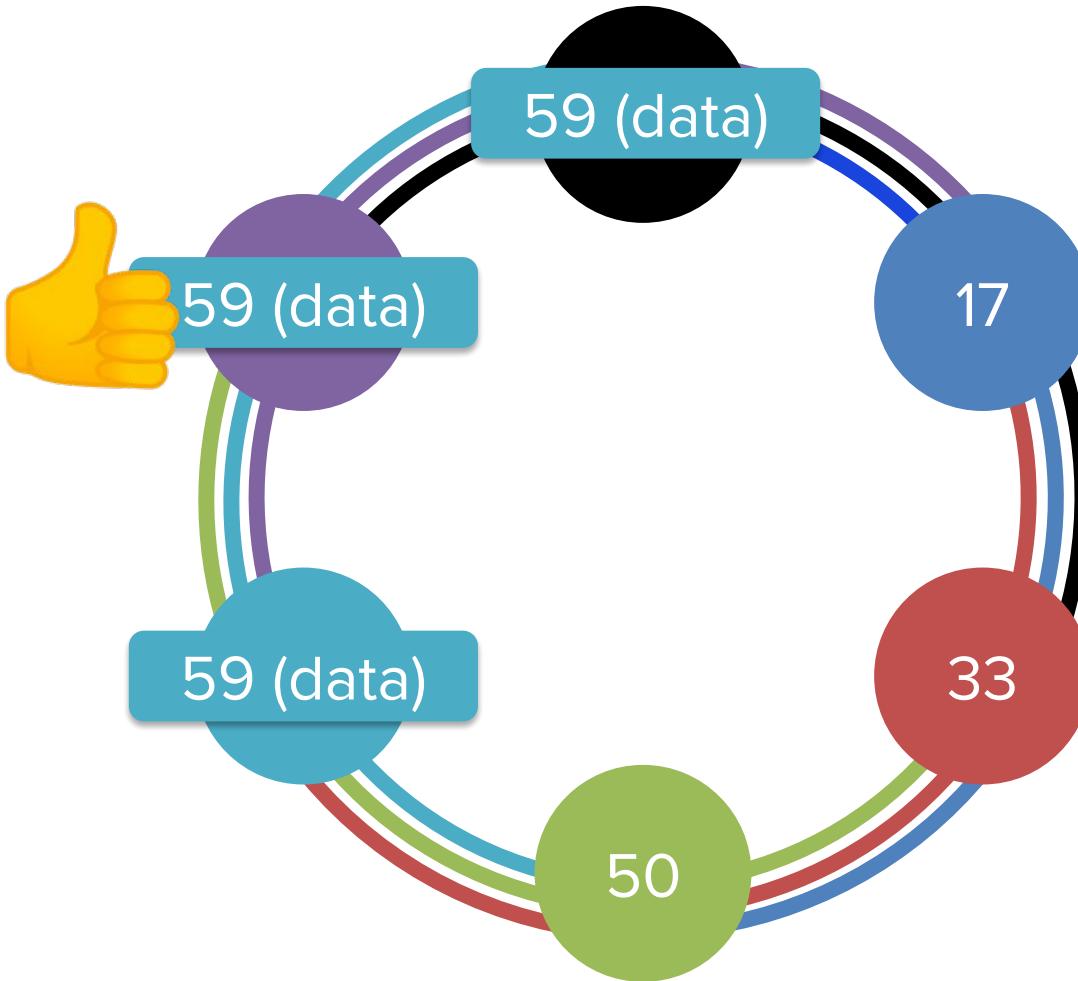
Node Failure

RF = 3

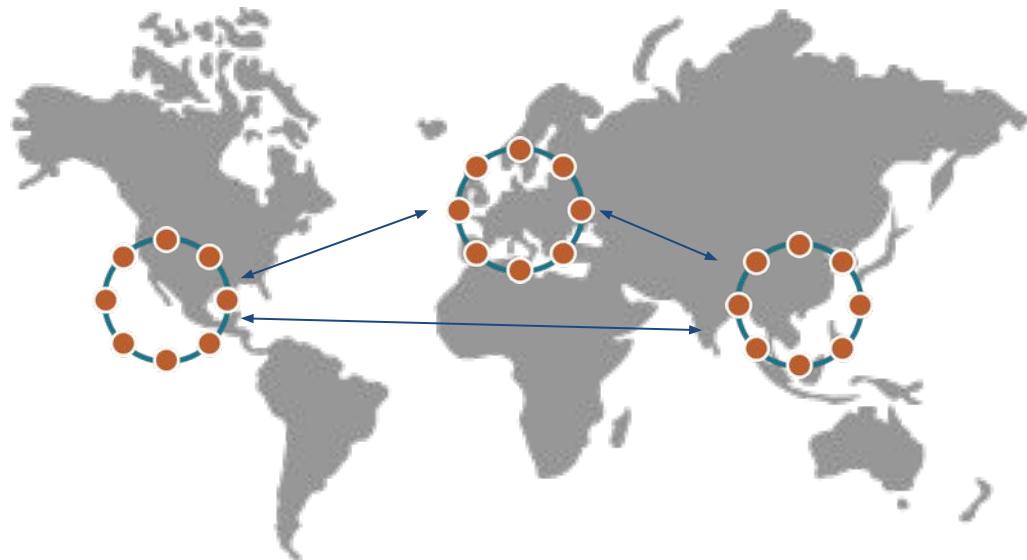


Node Failure – Recovered!

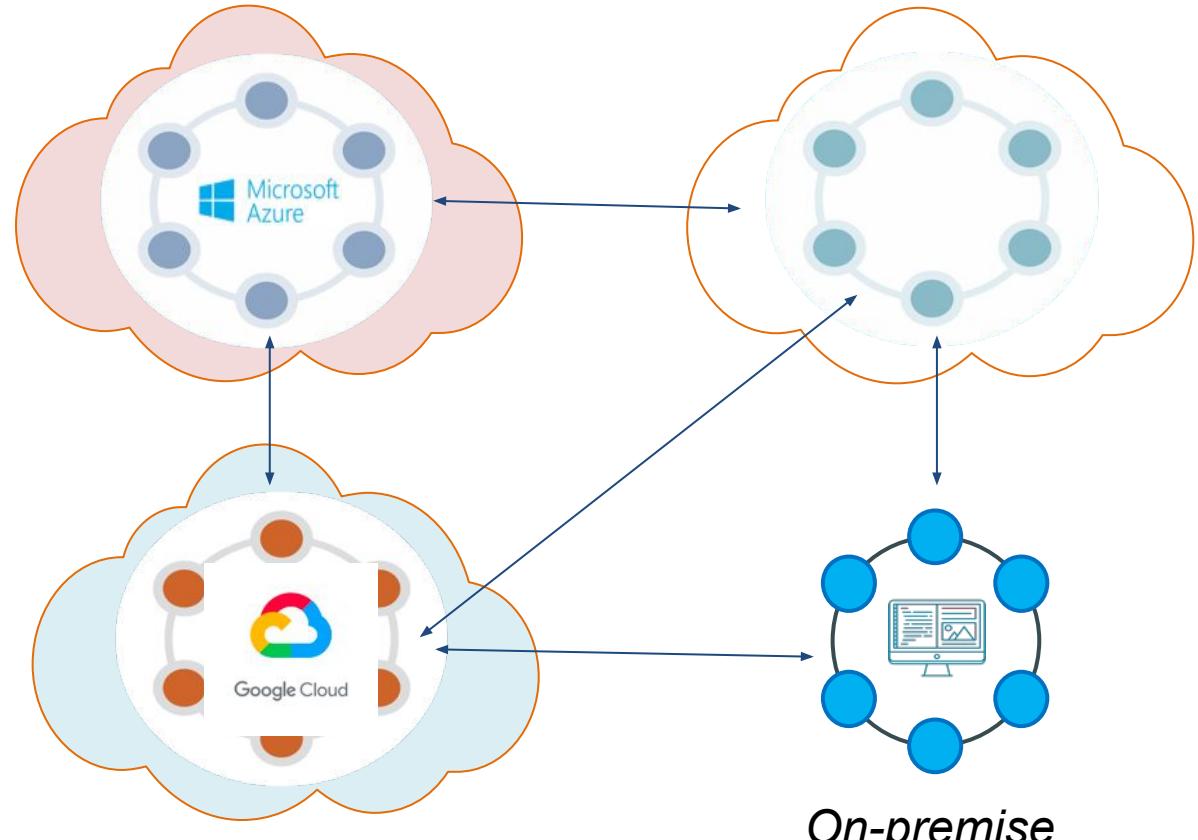
RF = 3



Data Distributed Everywhere



GEOGRAPHICALLY



HYBRID- MULTI CLOUD

Understanding Use Cases

S

High Throughput
High Volume

Heavy Writes

Heavy Reads

Event Streaming
Internet of Things
Log Analytics
Any TimeSeries

Caching

Market Data

Prices

A

Mission Critical Availability

No Data Loss
Responsive System

Banking
Track and Trace
Customer Apps

R

Realtime

Any CRUD
API Layer

Global Company

Enterprise Data Layer
Applications

D

Distributed

Geographically Deployments

Retailers

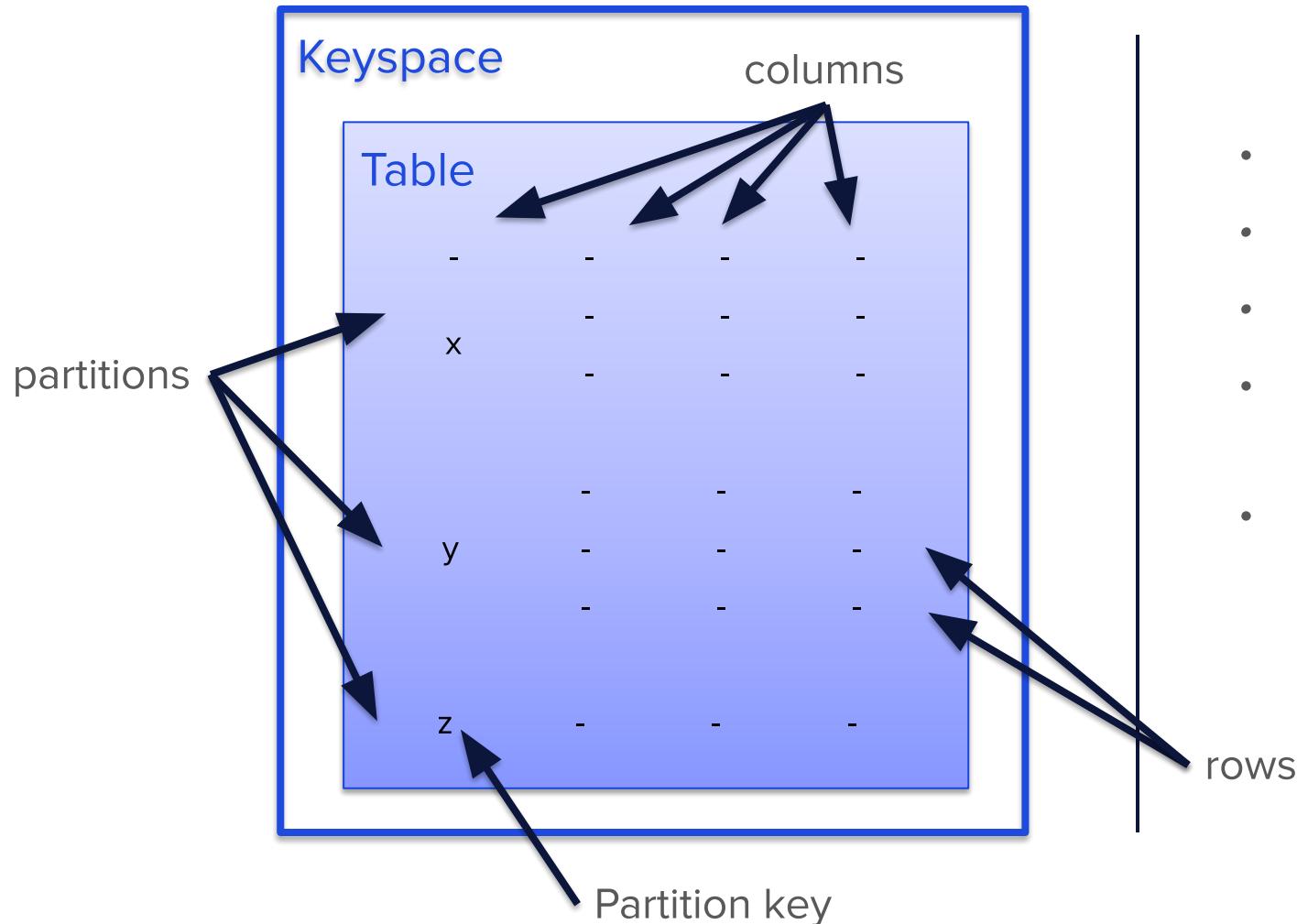
Hybrid Cloud
MultiCloud

DataStax Meetup

Cassandra's Data Model



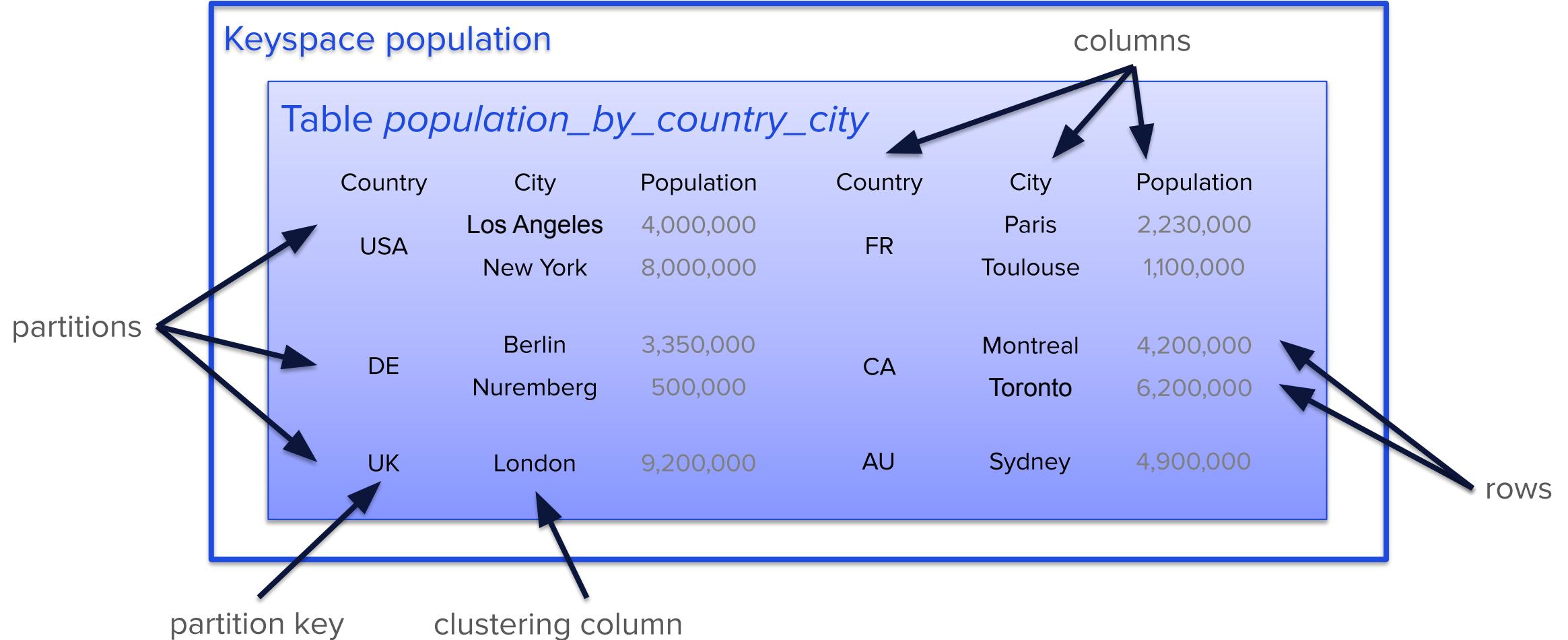
How does Cassandra structure data?



- Tabular data model, with one twist
- *Keyspaces* contain *tables*
- *Tables* are organized in *rows* and *columns*
- Groups of related rows, called *partitions*, are stored together on the same node (or nodes)
- Each row contains a *partition key*
 - One or more columns that are hashed to determine which node(s) store that data



Example Data – City populations organized by country



Example Data – City populations organized by country

Keyspace population

Table *population_by_country_city*

Country	City	Population
USA	Los Angeles	4,000,000
	New York	8,000,000
DE	Berlin	3,350,000
	Nuremberg	500,000
UK	London	9,200,000

CQL Equivalent:

```
CREATE TABLE population_by_country_city (
    country text,
    city text,
    population int,
    PRIMARY KEY((country), city)
);
```

partition key

clustering column



@DataStaxDevs #DataStaxDeveloperDay

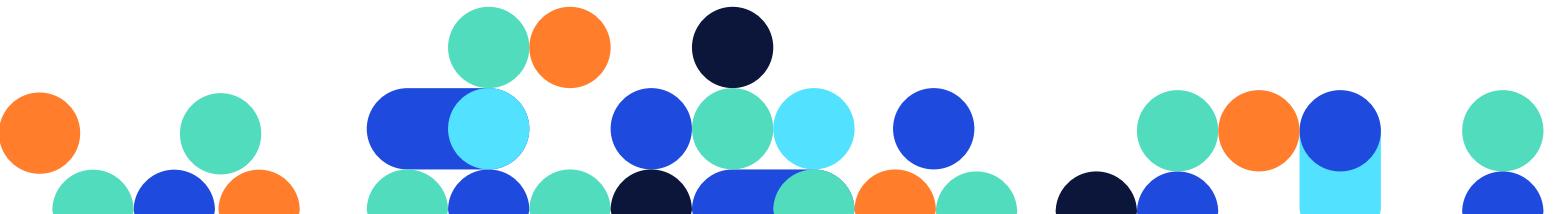
<https://community.datastax.com>



Time for an exercise!



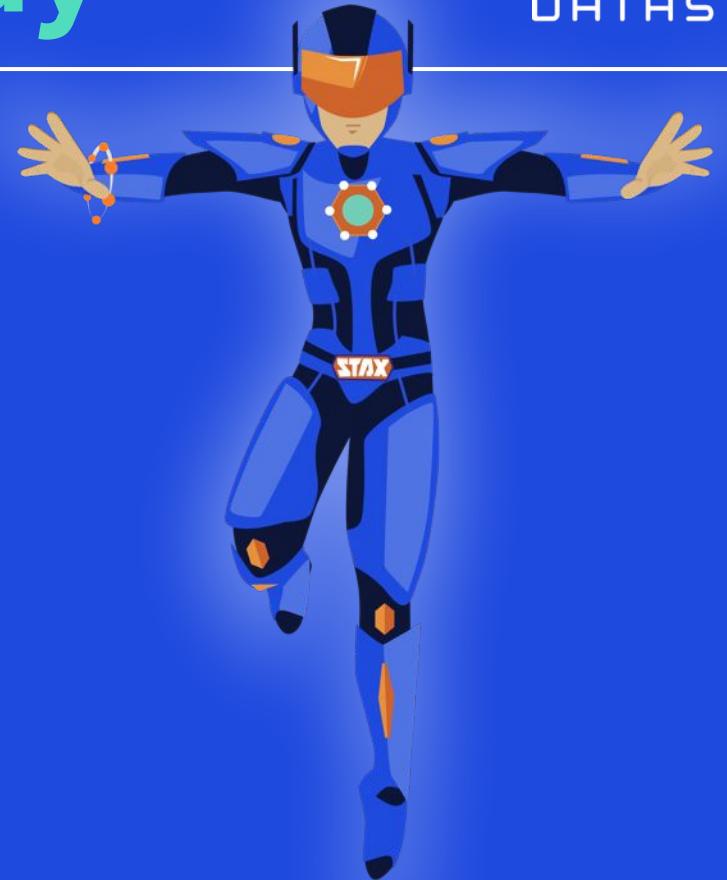
“Getting Started with Apache Cassandra™” Notebook



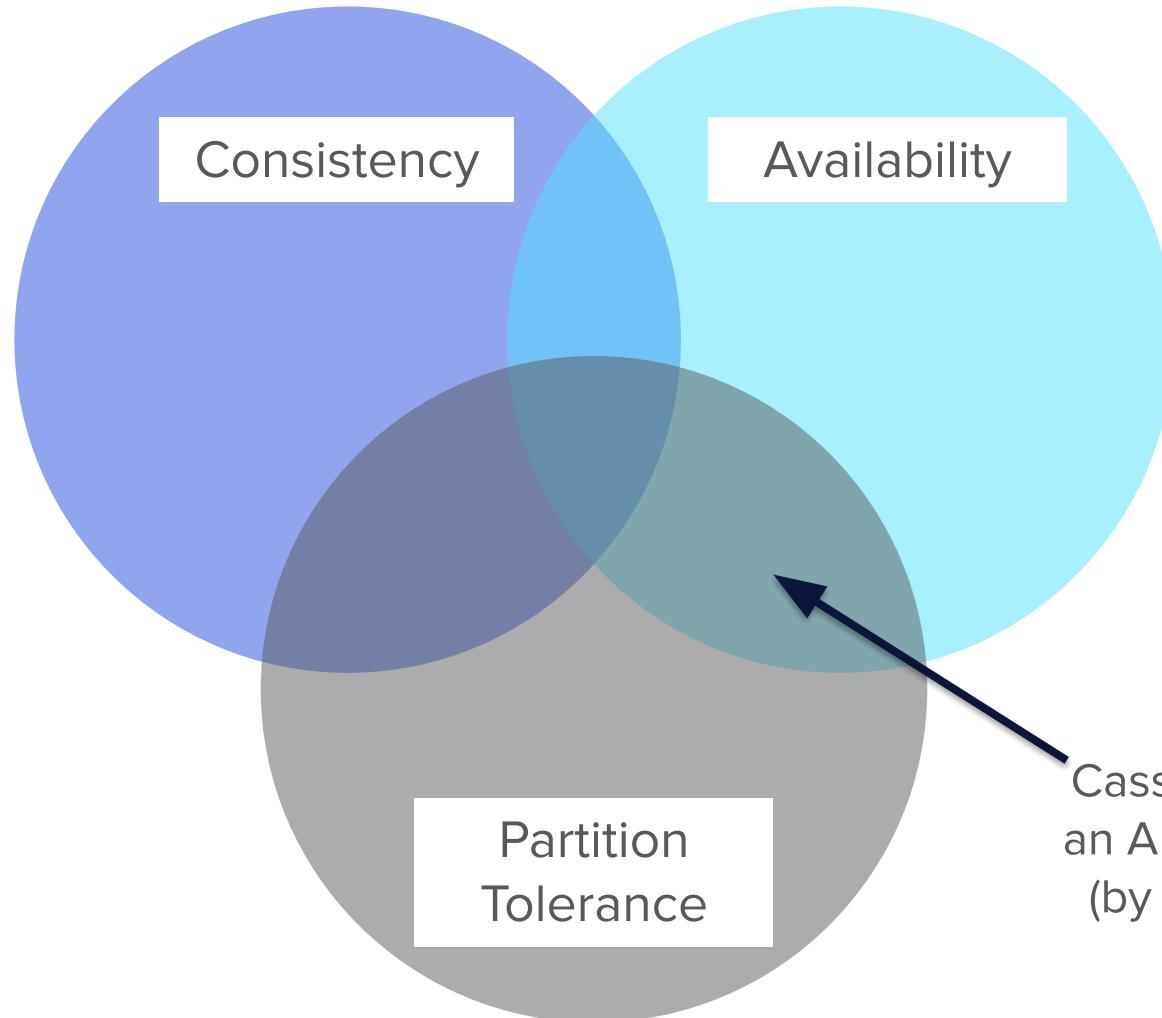
DataStax Developer Day



Cassandra's Consistency



CAP Theorem



Cassandra is
an AP system
(by default)

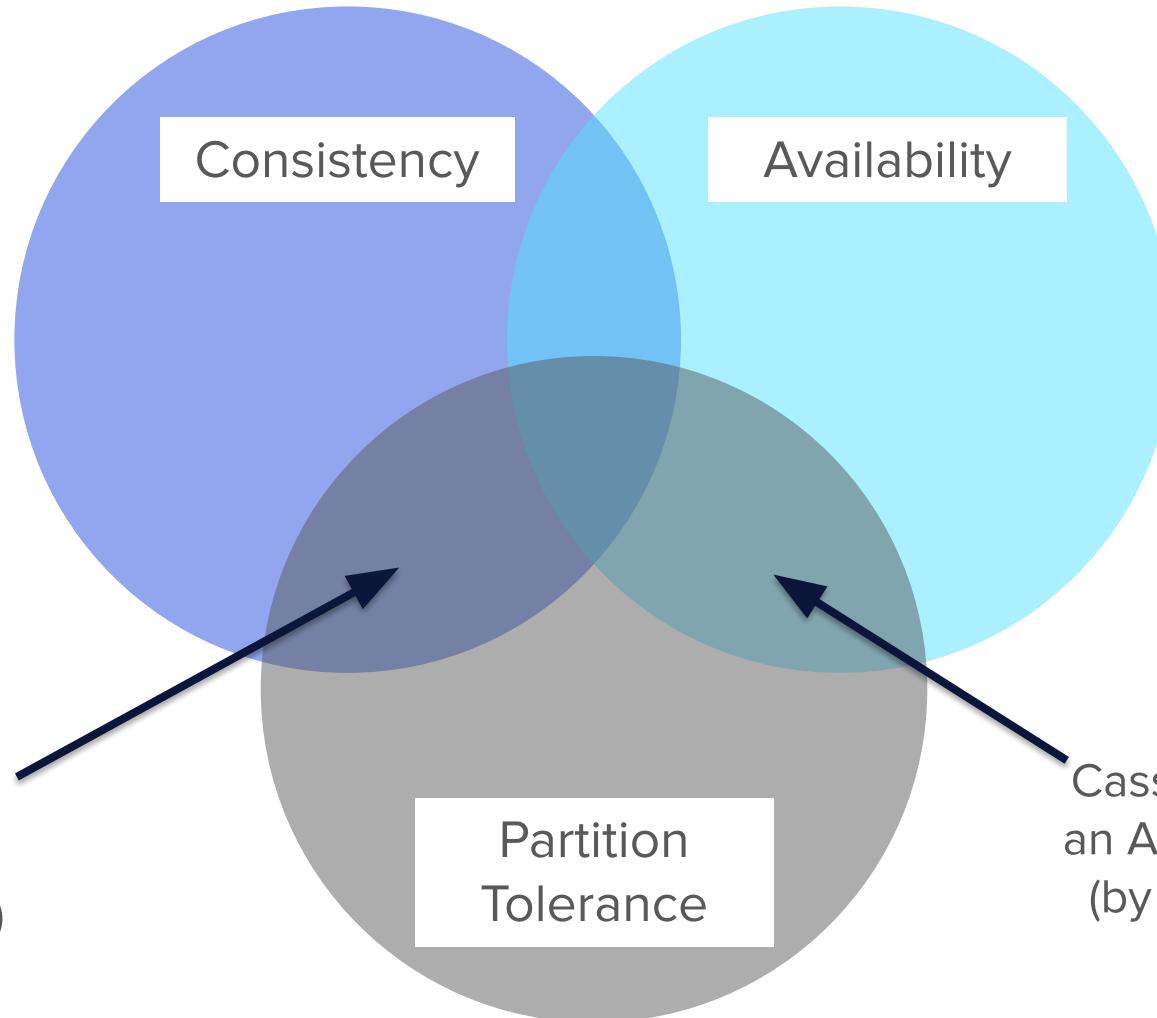


@DataStaxDevs #DataStaxDeveloperDay

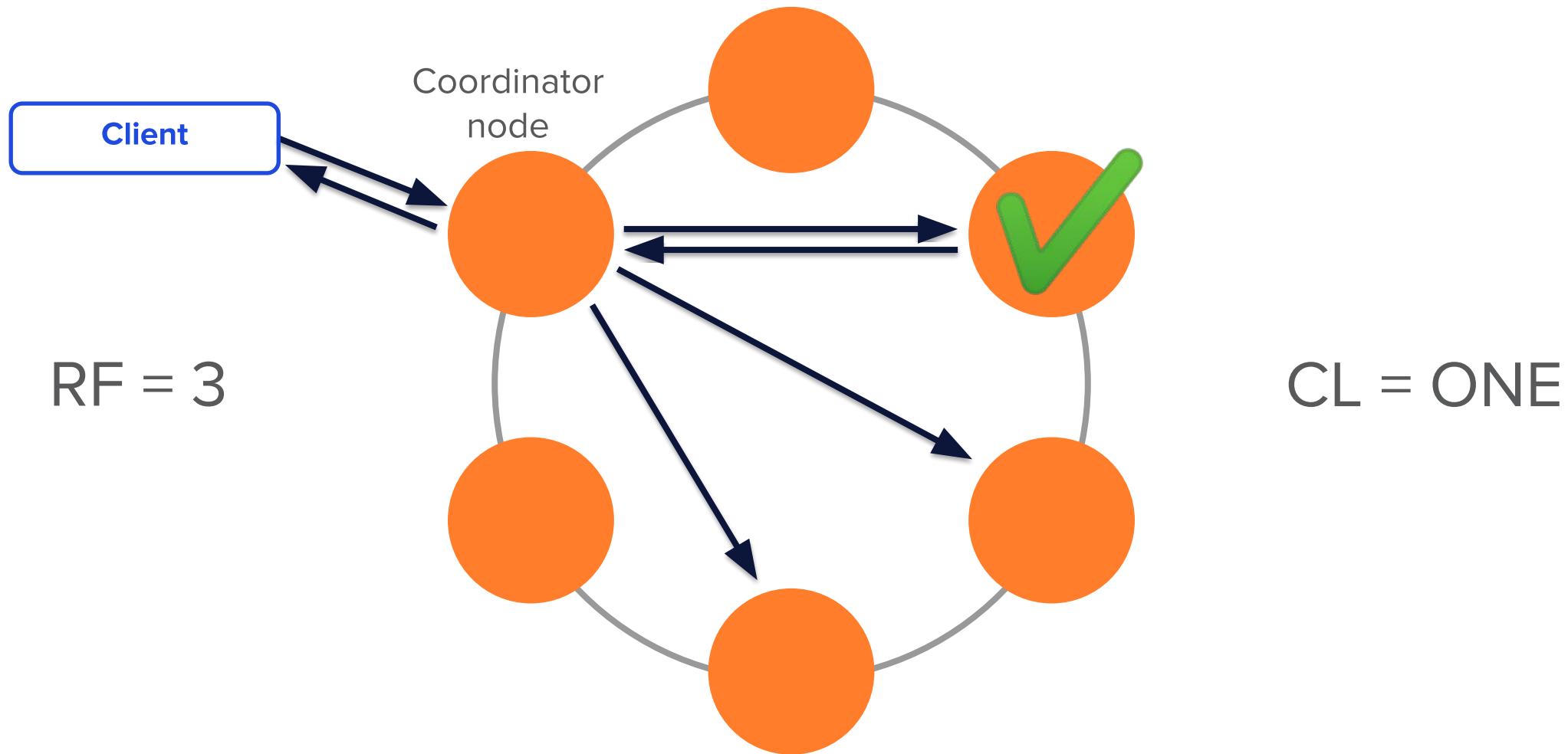
<https://community.datastax.com>



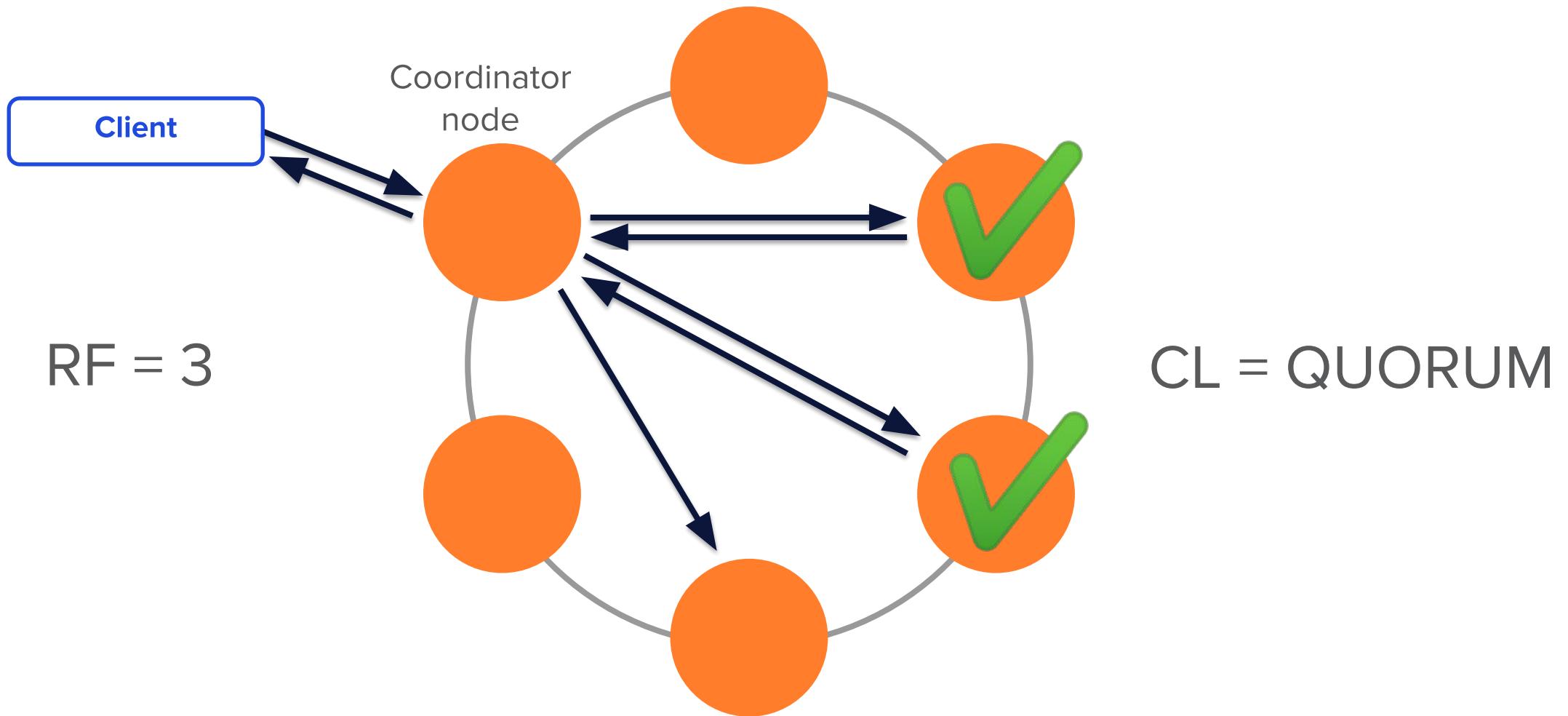
CAP Theorem



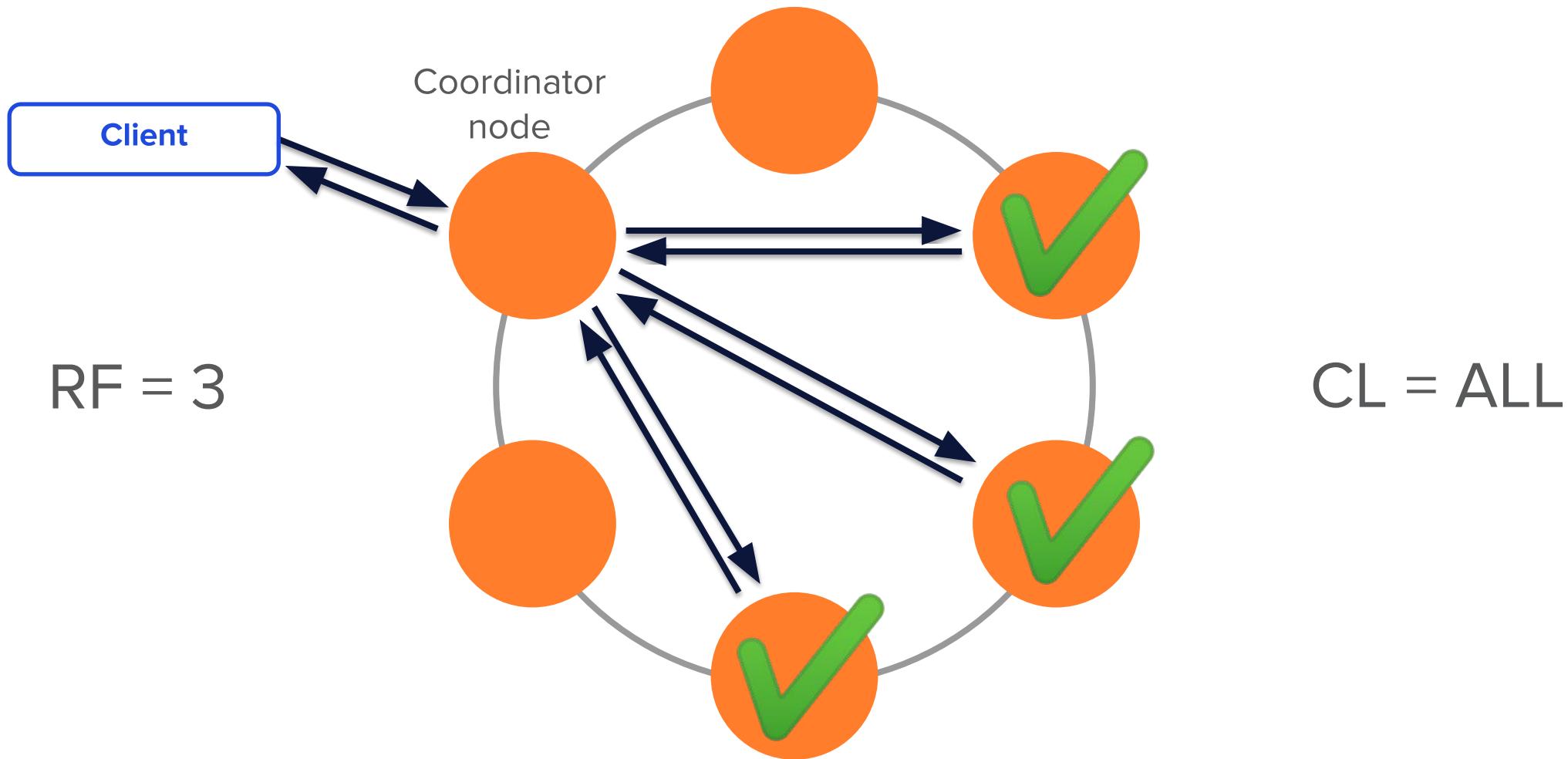
Consistency Levels



Consistency Levels



Consistency Levels



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

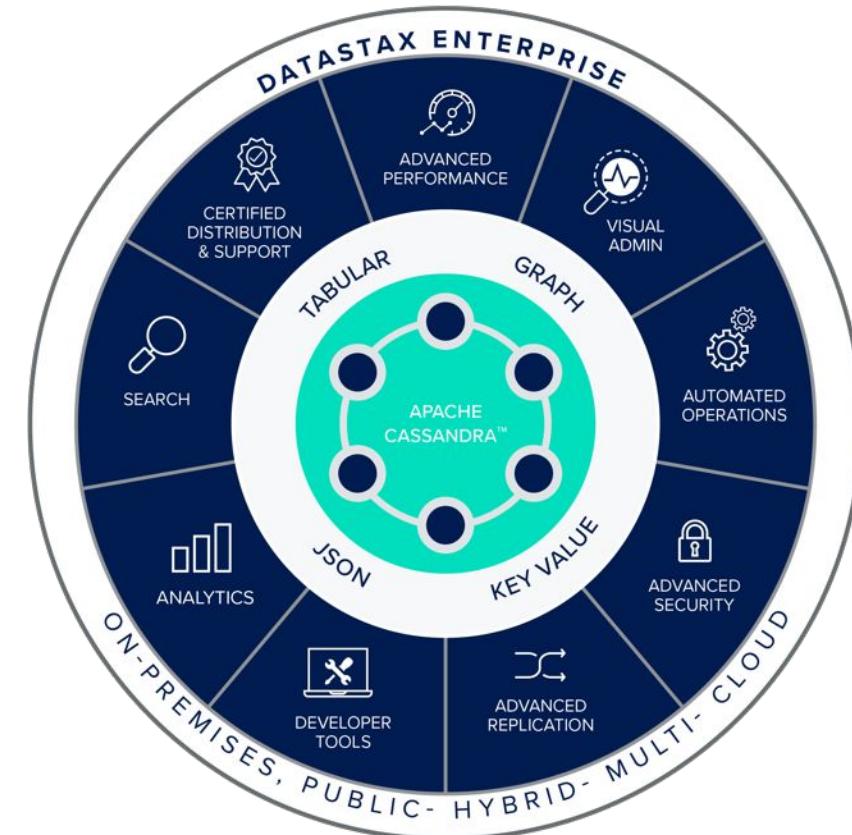


DataStax Meetup

Getting Started with Apache Spark

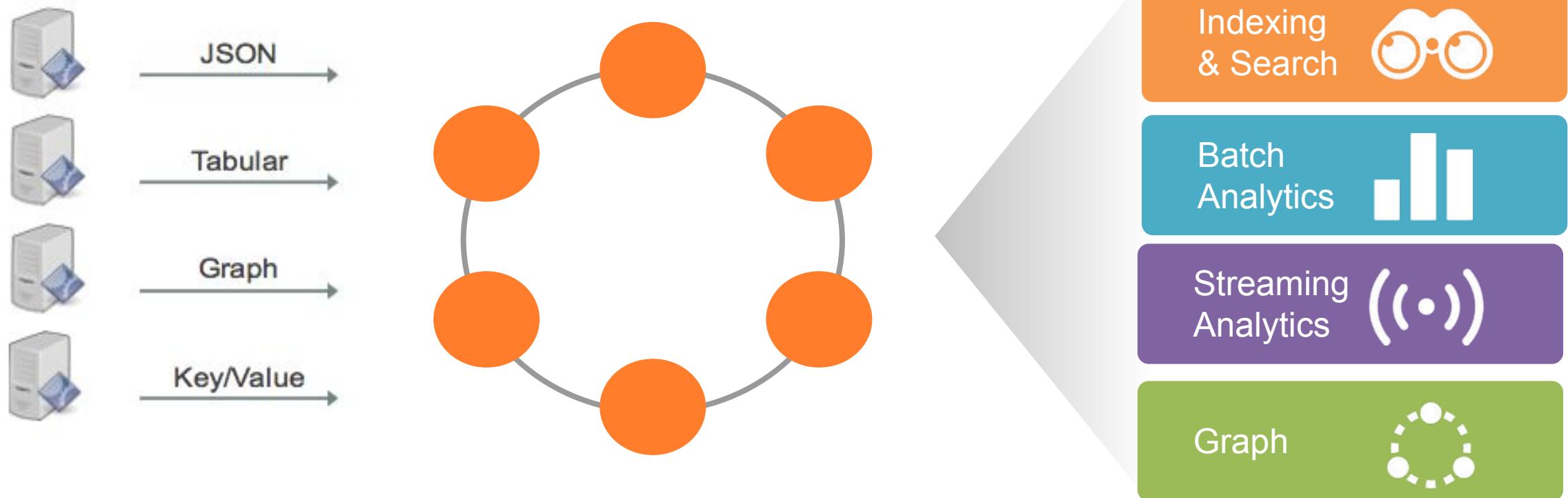


Our solution DataStax Enterprise (DSE)



- A **unified data layer** of database, search, and analytics, all independent of the public cloud provider and portable
- Consistent data management built for on-prem, hybrid-, and/or multi-cloud
- Consistent security **model** across entire data layer
 - Row and columnar level control of your data helps you achieve data governance and compliance

Integrated multi-model/mixed workload platform



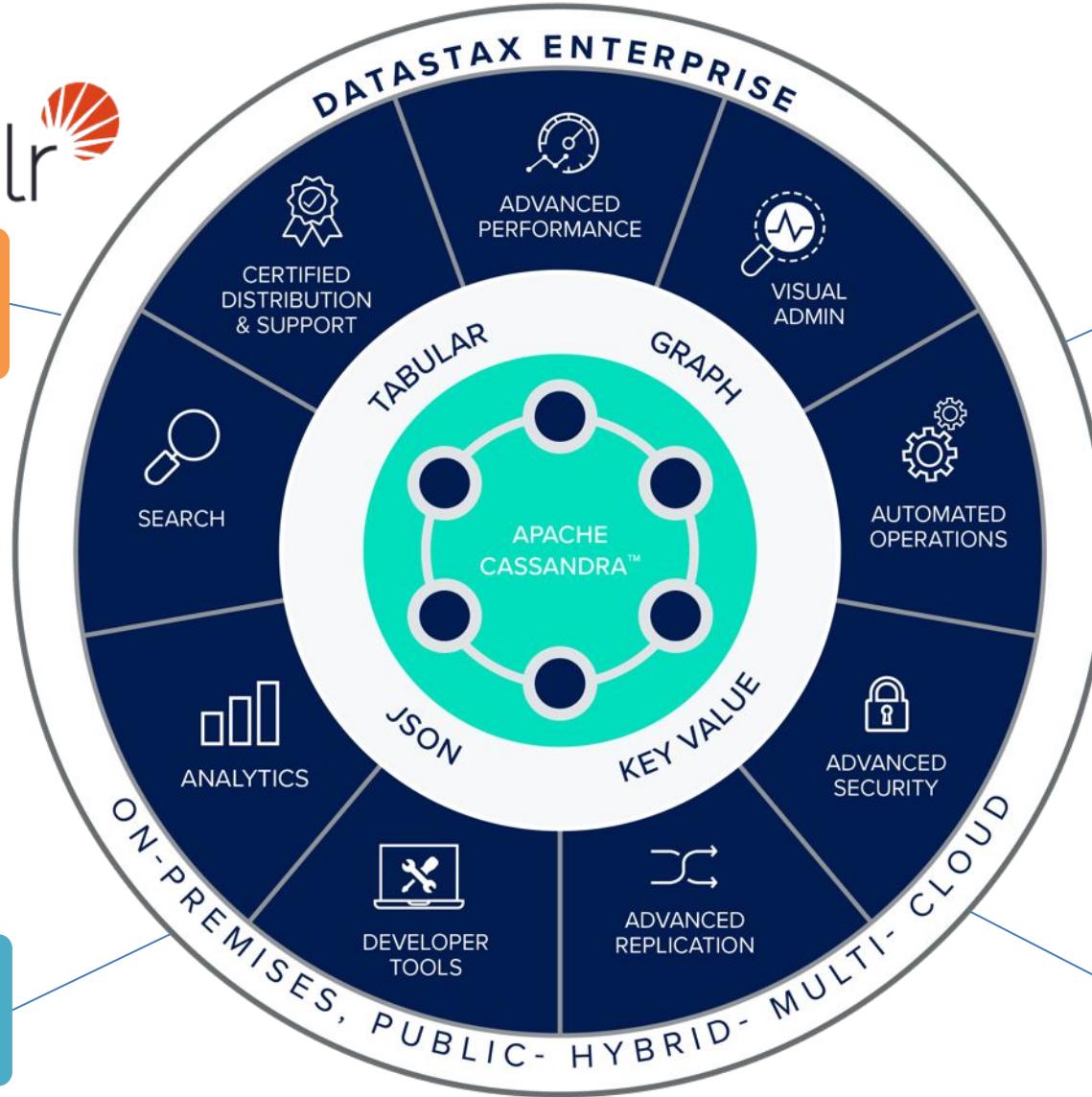
DataStax Enterprise



Indexing & Search



Batch Analytics



Graph



OLTP



Data Analytics

- Definition

Science and craft of building applications from data analysis steps to discover useful information and support data-driven decision making

- Use cases

- Recommendations
- Fraud detection
- Social networks and Web link analysis
- Marketing and advertising decisions
- Customer 360
- Sales and stock market analytics
- IoT analytics

- Analysing Steps

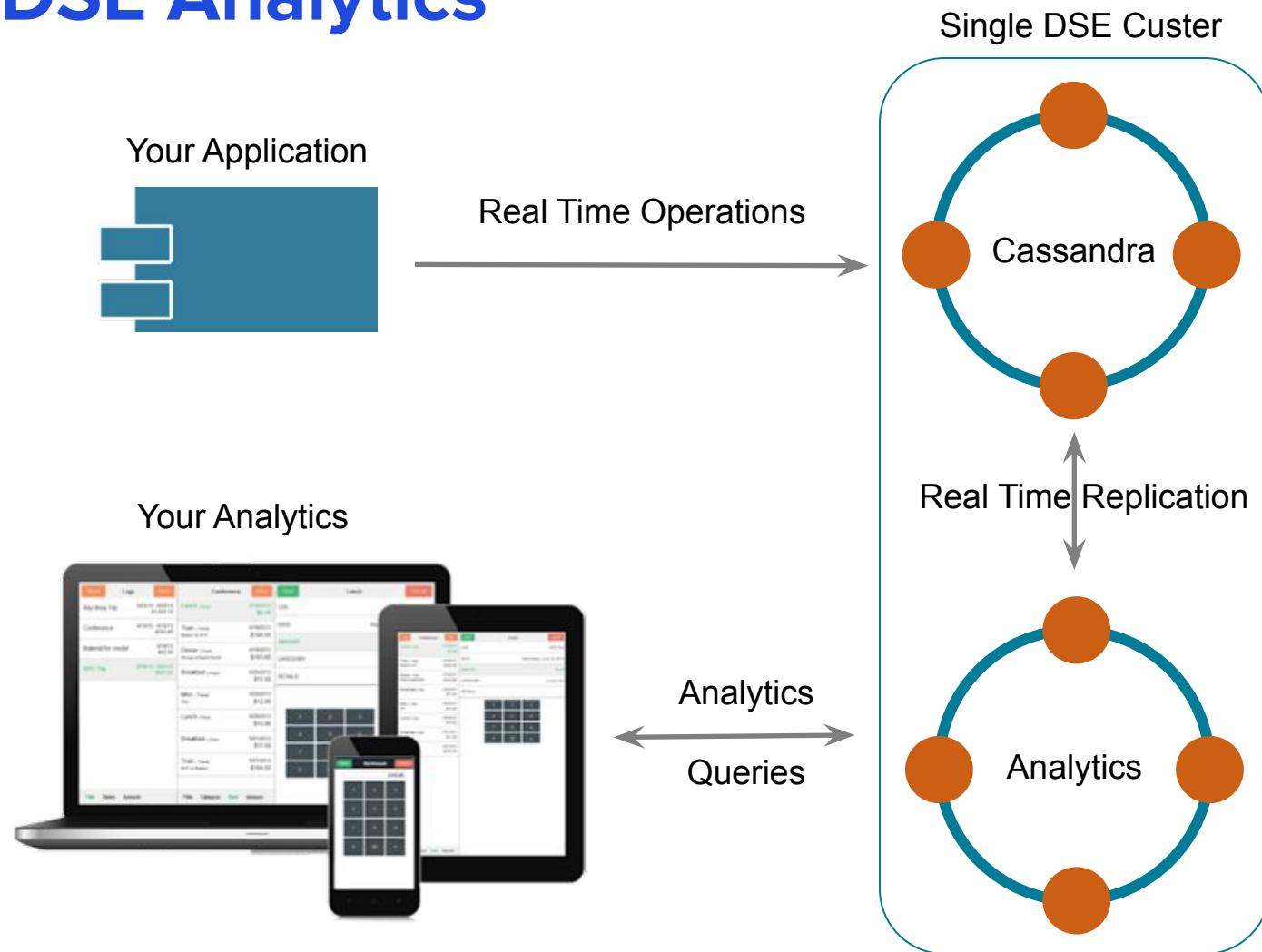
- Statistical analysis
- Classification
- Clustering
- Regression
- Similarity matching
- Collaborative filtering
- Profiling
- Dimensionality reduction
- Feature extraction

Distributed computation engine designed for big data and in-memory processing

- Interactive and batch analytics
- Up to 100x faster than Hadoop
- 5-10x less code than Hadoop
- Efficiency and scalability
- Fault-tolerance



DSE Analytics



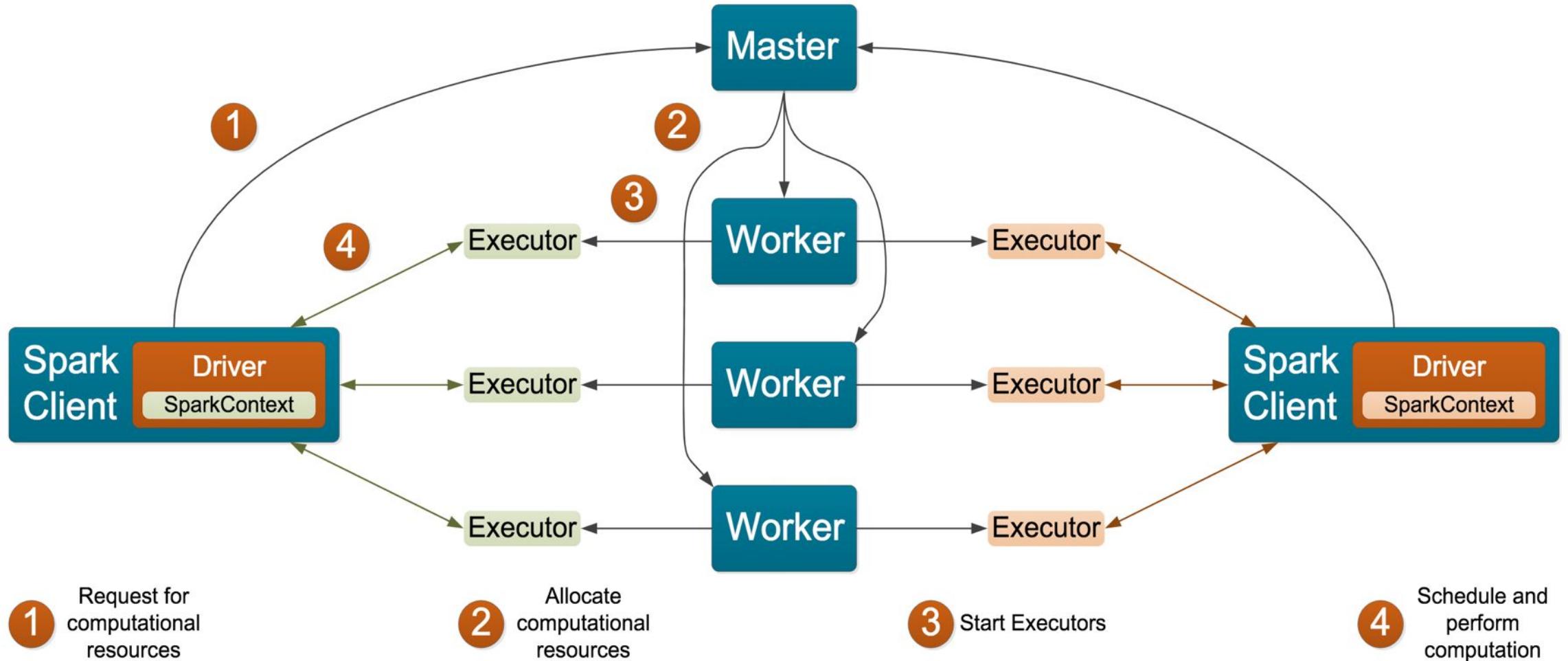
Streaming, ad-hoc, and batch

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

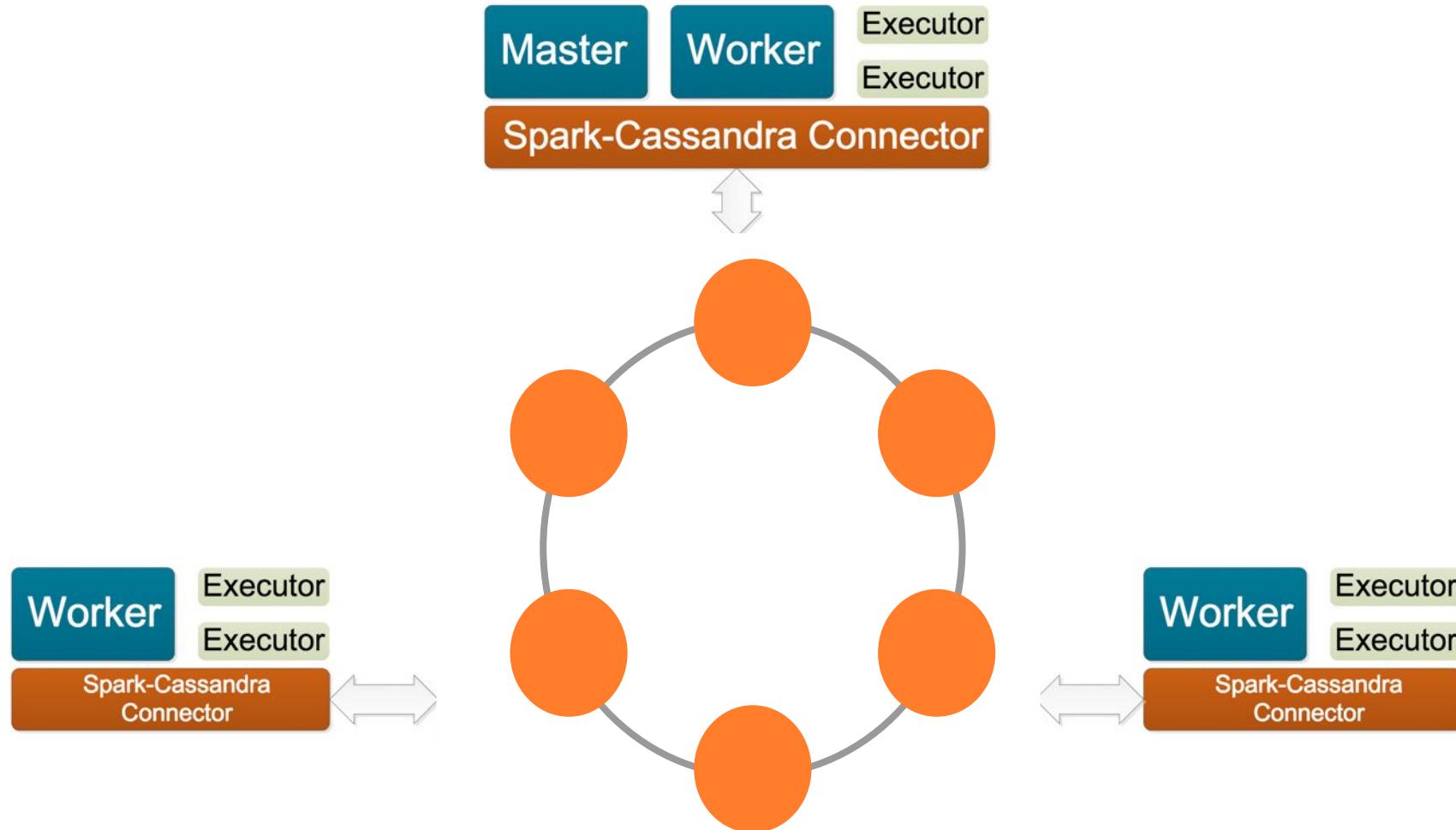
Compared to self-managed:

- No ETL
- True HA without Zookeeper

Spark Architecture



Architecture with Spark DSE Driver



Database Access with DataStax Driver

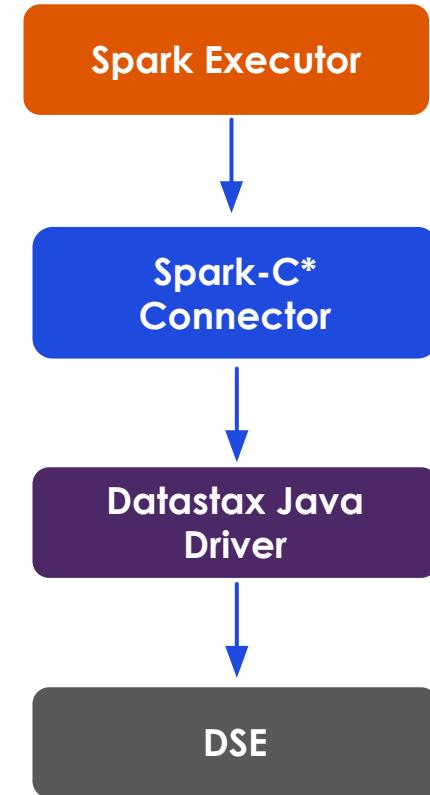
- DataStax Cassandra Spark driver
 - Implemented mostly in Scala
 - Scala + Java APIs
 - Does automatic type conversions

```
// Spark connection options
val conf = new SparkConf(true)
.set("spark.cassandra.connection.host", "127.0.0.1")
.set("spark.cassandra.auth.username", "cassandra")
.set("spark.cassandra.auth.password", "cassandra")

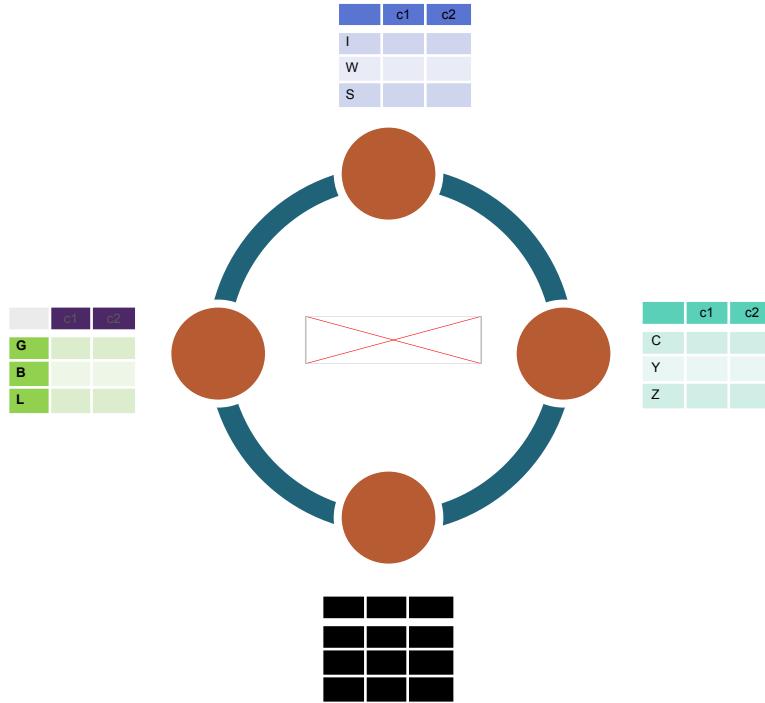
val sc = new SparkContext("spark://127.0.0.1:7077", "myapp", conf)

// Read from DSE and add partitioner with primary key
val rdd = sc.cassandraTable("my_keyspace", "my_table").byKey("pk", "cc")

// Save to DSE
rdd.saveToCassandra("my_keyspace", "my_table", SomeColumns("key",
"value"))
```



Data Locality

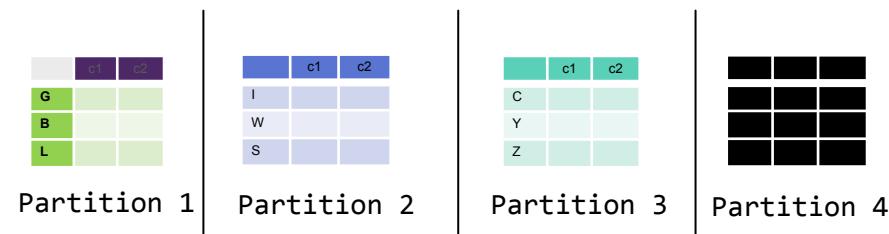


- DSE Analytics respects data locality
- No need for ETL between separated clusters
- Spark Master HA

Every Spark task uses a CQL-like query to fetch data for a given token range:

```
SELECT "key", "value" FROM "keyspace"."table"  
WHERE  
    token("key") > 384023840238403 AND  
    token("key") <= 38402992849280  
ALLOW FILTERING
```

In Memory: Distributed on all available nodes



Integrating Analytics into your application

```
dse cassandra -k
```

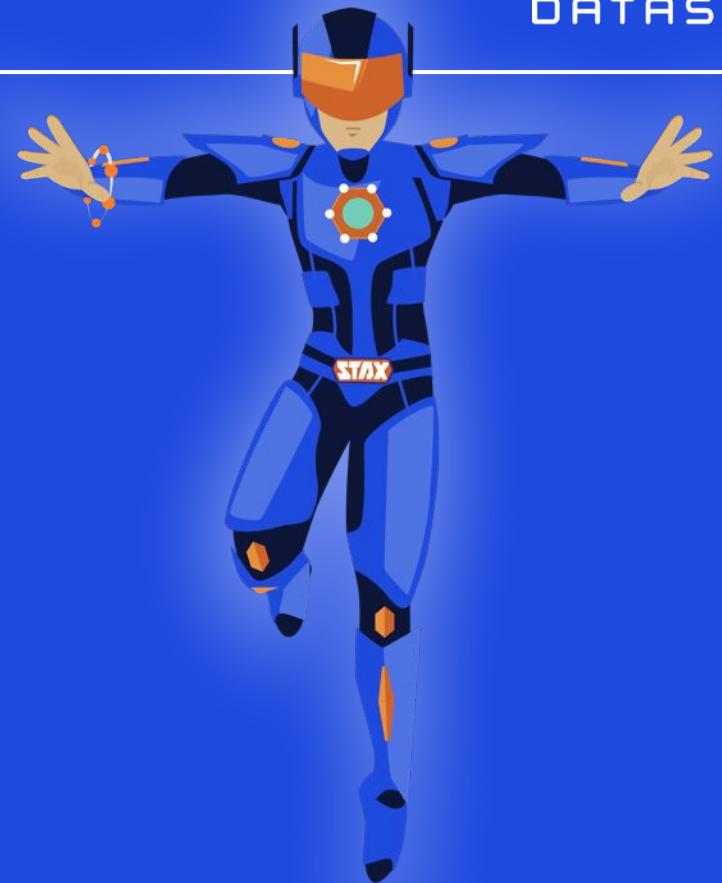
- Aaaaand you're done



DataStax Meetup



Exercises Bootstrap



Before Starting

Hands-on Codelab

 **cassandra** Apache Cassandra 3 hours, intermediate level + dinner break

An initiative by:  ITALIAN ASSOCIATION FOR MACHINE LEARNING Powered by:  

Dear guest,
Welcome to this event organized in collaboration with DataStax and SourceSense. As a codelab we expect you to have your laptops to do the exercises.
The session has been designed for intermediate level software engineers and data scientists. We have a lot to cover and unfortunately not a lot of time to help you installing. Don't worry we made things as simpler as we can, still if you are beginner try to team up !
Prerequisites : To run the exercises you will simply need : Docker (cf link below)

Agenda

PART I – What is Apache Cassandra and why do you care ?

- Getting starting with Apache Cassandra™ and use Cases
- CodeLab : Getting Started with Apache Cassandra
- Apache Spark™ and DataStax Enterprise Analytics

PART II – Machine Learning with DataStax Enterprise

- CodeLab : Clustering with K-means
- CodeLab : Classification with Naïve Bayes
- CodeLab : Regression and Classific. with RandomForest
- CodeLab : Recommendation with FP-Growth
- CodeLab : Recommendation with Collaborating Filtering

Installation

```
git clone https://github.com/HadesArchitect/CaSpark.git  
cd CaSpark  
docker-compose up -d  
docker-compose logs -f jupyter  
http://localhost:8888
```

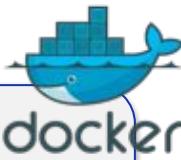
Download Docker	http://download.docker.com/
Github Repository	https://github.com/HadesArchitect/CaSpark.git
DataStax Studio	http://localhost:9091
DataStax Academy	http://academy.datastax.com
DataStax Community	http://community.datastax.com



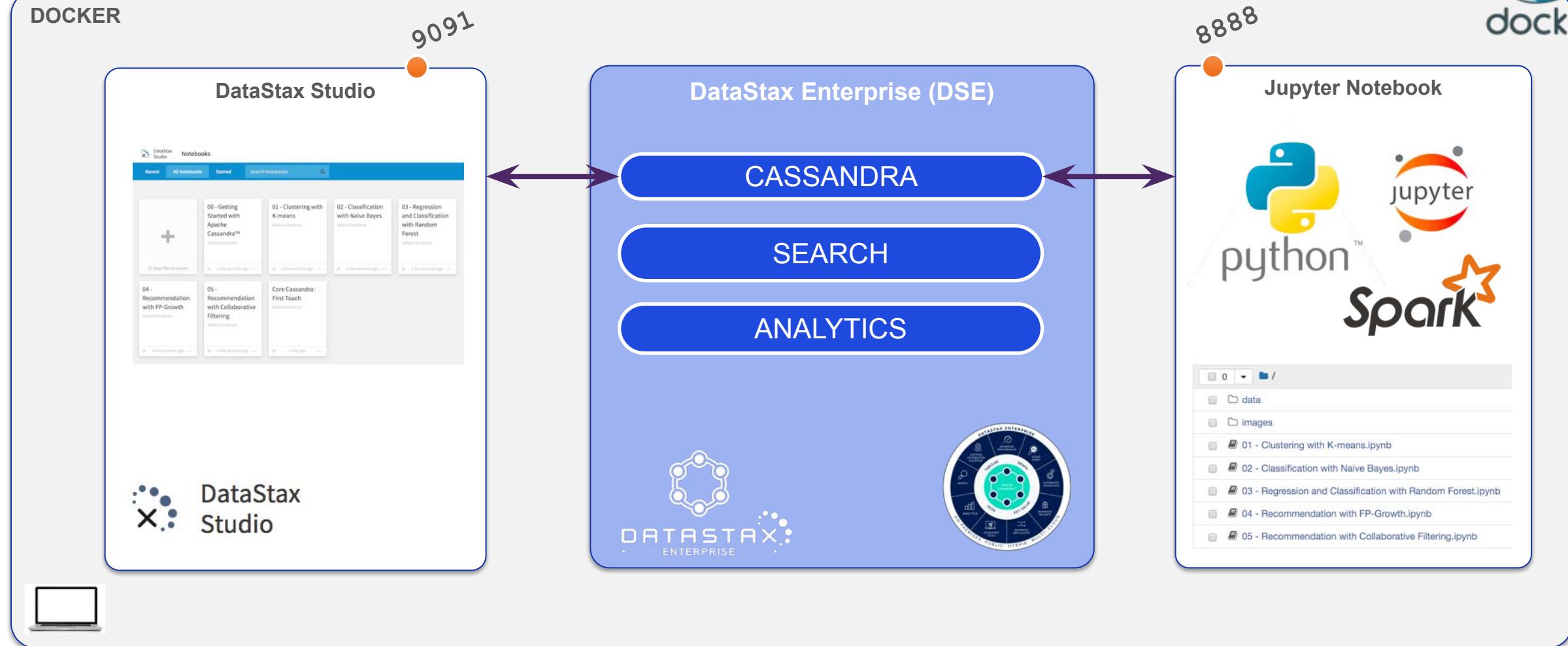
- You should have one of those sheet.
- Please execute the Installation steps **as soon as possible**. This will download a few docker images that can take some time !

Your environment



docker-compose up -d

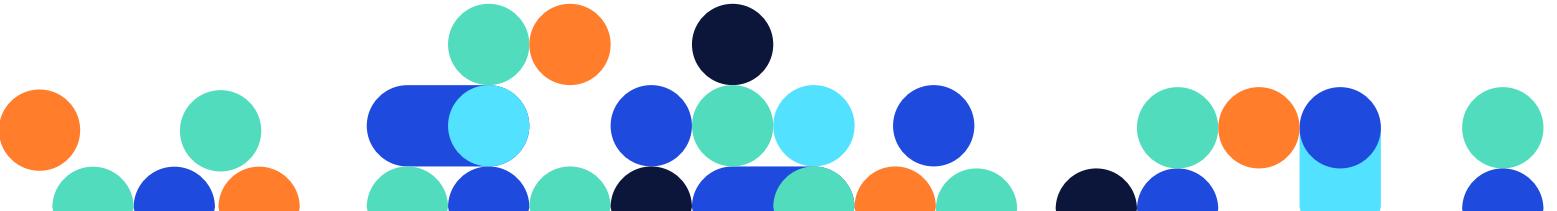
DOCKER



Time for an exercise!



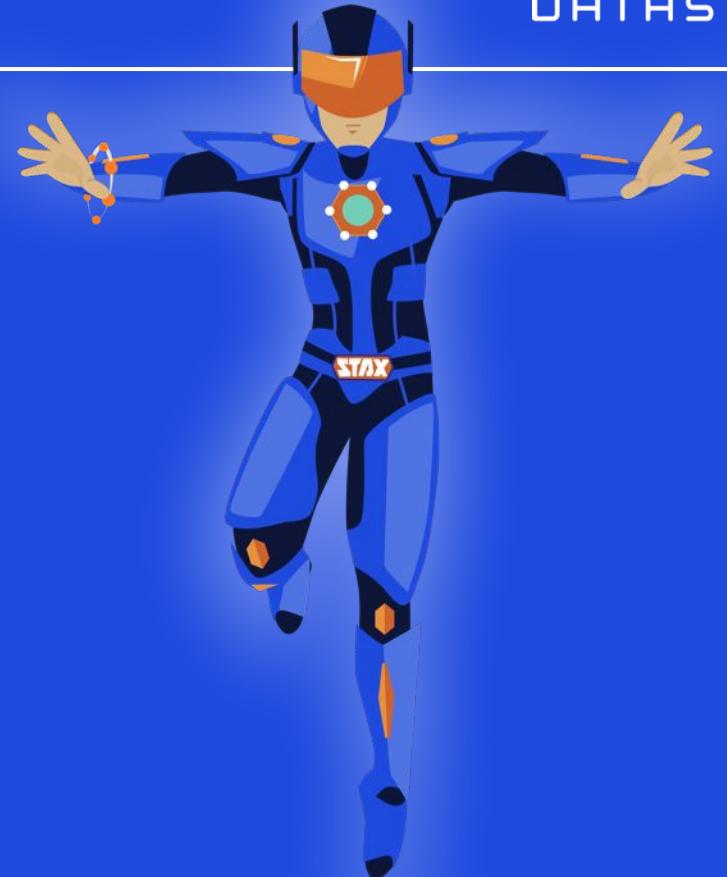
“Getting Started with Apache Spark™” Notebook



DataStax Meetup



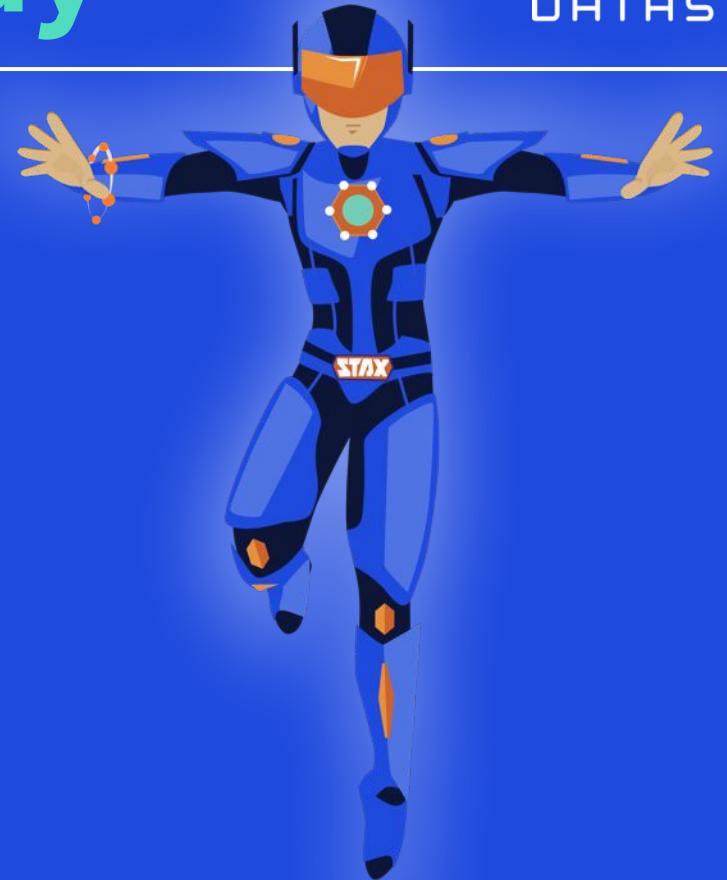
Machine Learning



ALEKS SLIDES

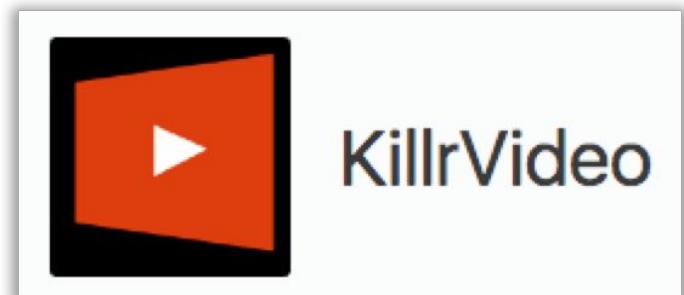
DataStax Developer Day

Developer Resources



Resources for Developers

- DataStax Academy
 - [Training Courses](#) and [Certifications](#)
 - [Developer Blog](#)
 - [Distributed Data Show podcast](#)
- Live Events
 - [Developer Day](#)
 - [Meetups](#)
- More Content
 - [YouTube Channel](#)
 - [Live coding on Twitch](#)
 - [KillrVideo reference application](#)
 - [DataStax Academy Slack](#)



Developer Day

A day of hands-on learning about DataStax Enterprise and Apache Cassandra™

- Use Cases
- Core Cassandra
- Cassandra Data Modeling
- Application Development
- Search, Analytics and Graph
- Operations and Security

Network with experts, Developer Advocates and peers

Open to the public, or schedule a private event



Training Courses at DataStax Academy

- Free self-paced DSE 6 courses
 - [DS201: DataStax Enterprise 6 Foundations of Apache Cassandra™](#)
 - [DS210: DataStax Enterprise 6 Operations with Apache Cassandra™](#)
 - [DS220: DataStax Enterprise 6 Practical Application Data Modeling with Apache Cassandra™](#)
 - [DS330: DataStax Enterprise 6 Graph](#)
 - [DS332: DataStax Enterprise 6 Graph Analytics \(NEW\)](#)



Learning Paths on DataStax Academy

- Unsure where to start?
- Follow a learning path to learn about topics related to your role.
 - Administrator
 - Analytics Specialist
 - Architect
 - Developer
 - Graph Specialist
 - Search Specialist

<https://academy.datastax.com/paths>

community.datastax.com

DATASTAX COMMUNITY

Find posts and topics... 

Events Live Coding About Login or Sign Up

ESPACES ▾ 

Bringing together the Apache Cassandra experts from the community and DataStax.

Want to learn? Have a question? Want to share your expertise? You are in the right place!

Not sure where to begin? [Getting Started](#)

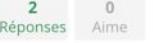
Tous les messages 

 nagasam17_177126 répondu · il y a 19 heures · General forum
[Course status not updated to completion](#) 

 tim.mason_177848 demandé · il y a 23 heures · General forum
[How to get free book after completion of academy courses](#) 

 Erick Ramirez répondu · il y a 3 jours · General forum
[multi node cluster installation on ubuntu](#) 

 Erick Ramirez répondu · il y a 5 jours · General forum
[Will changing data partitioning help with the load ?](#) 

 pmcfadin répondu · il y a 6 jours · General forum
[Data Modeling : One to Many and Many to Many](#) 

 Erick Ramirez répondu · il y a 6 jours · General forum
[Datastax Exam Issue](#) 

 Erick Ramirez répondu · il y a 6 jours · General forum
[How can I troubleshoot OpsCenter showing agent is unreachable?](#) 

SUJETS POPULAIRES

cassandra spark spark-connector driver performance
dse search search opcenter dse certification graph
data modelling docker datastax cassandra upgrade
compaction query solr datastax community dse 6.7.3
java delete dse-java-driver-core partition restore
datastax enterprise server 6.7.4 timeout opscenter-backup
cpp cassandra-connection

VOIR TOUT

BADGES RÉCENTS

 nagasam17_177126
 Beck
 amitosh
 csplinter
 joao.reis_163533
 muzimilbasha_178465
 muzimilbasha_178465
 \$500 Erick Ramirez

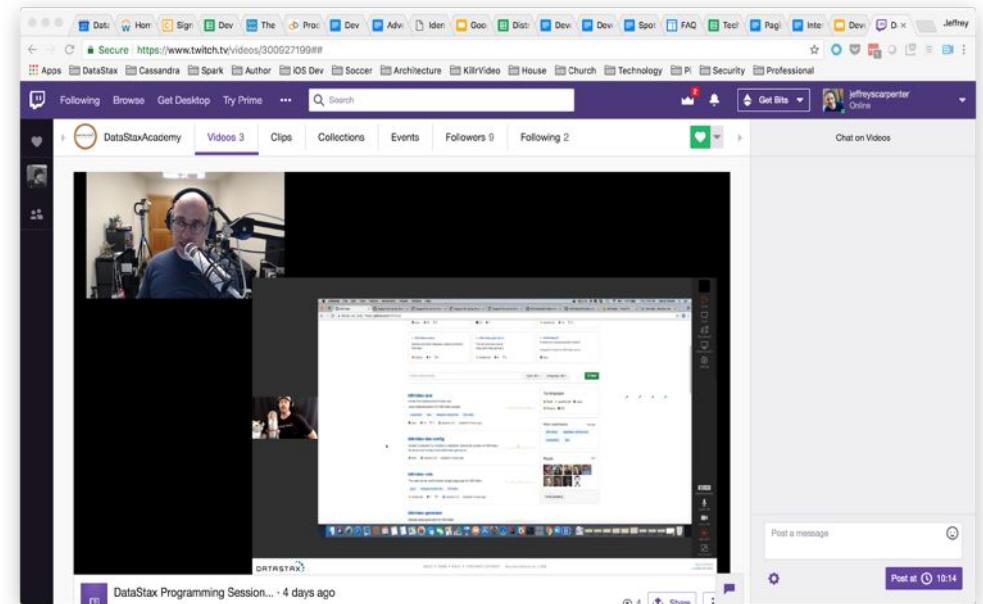
Distributed Data Show

- Interview-style show featuring a mix of DataStax and industry guests
- We go in-depth on the technology and challenges of data in large-scale distributed systems
- Released weekly on DataStax Academy [YouTube channel](#) and as a podcast
- Send us your suggestions for topics and guests – we love customer use cases



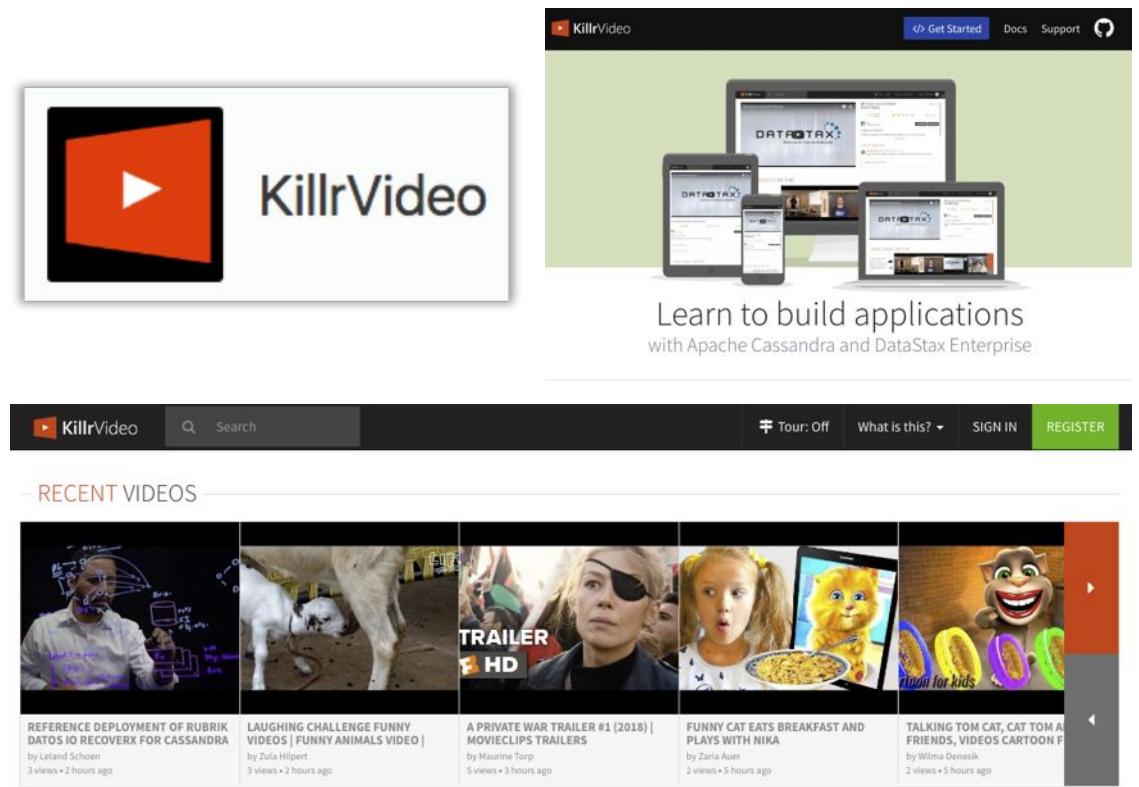
Live Coding on Twitch

- Live coding sessions with advocates and guests each Thursday
 - <https://www.twitch.tv/datastaxacademy>
- Working through the challenges of building distributed systems
- Join the conversation and ask questions
- Some advocates also do streaming on personal channels



KillrVideo Reference Application

- Reference application for learning how to use Apache Cassandra and DataStax Enterprise
 - DataStax Drivers
 - Docker images
- Source code freely available
 - <https://github.com/killrvideo>
- Live version
 - <http://killrvideo.com>
- Download, test, modify, contribute!



DataStax Meetup

We need you



Bureau of Diplomatic Projects
Antimatter Tracking ▾

DATA MODEL ^

- Antimatter
- Official
- Purchase
- Trader
- TraderPhoneNumber

DATA CHANGE EVENTS ^

- ↑ OfficialUpdate
- ↑ PurchaseRecord

QUERIES ^

- Bookmark AntimatterById
- Bookmark DangerousAntimatter
- Bookmark TradersByNameWithPurchases

DEPLOYMENTS ^

- Development Deployment

Build your custom data layer in 4 simple steps:

- ← 1. Add entities and relations to define your data model.
- ← 2. Add events to define changes to your data, data ingests, and data flows.
- ← 3. Add queries to define the read APIs for your data layer.
- ← 4. Access the development deployment to test your data layer or launch production deployments.

VISIONIZATION X

```

graph TD
    Antimatter[Antimatter] -- "atomicWeight Int" --> Purchase[Purchase]
    Antimatter -- "container String" --> Official[Official]
    Purchase -- "buyer Trader" --> Trader[Trader]
    Purchase -- "seller Trader" --> Trader
    Official -- "notarizedPurchases Purchase" --> Purchase
    Trader -- "phone PhoneNumber" --> TraderPhoneNumber[TraderPhoneNumber]
    Trader -- "sold Purchase" --> Purchase
    Trader -- "bought Purchase" --> Purchase
    Purchase -- "PurchaseRecordEffect" --> Purchase
    Official -- "OfficialUpdateEffect" --> Official
  
```

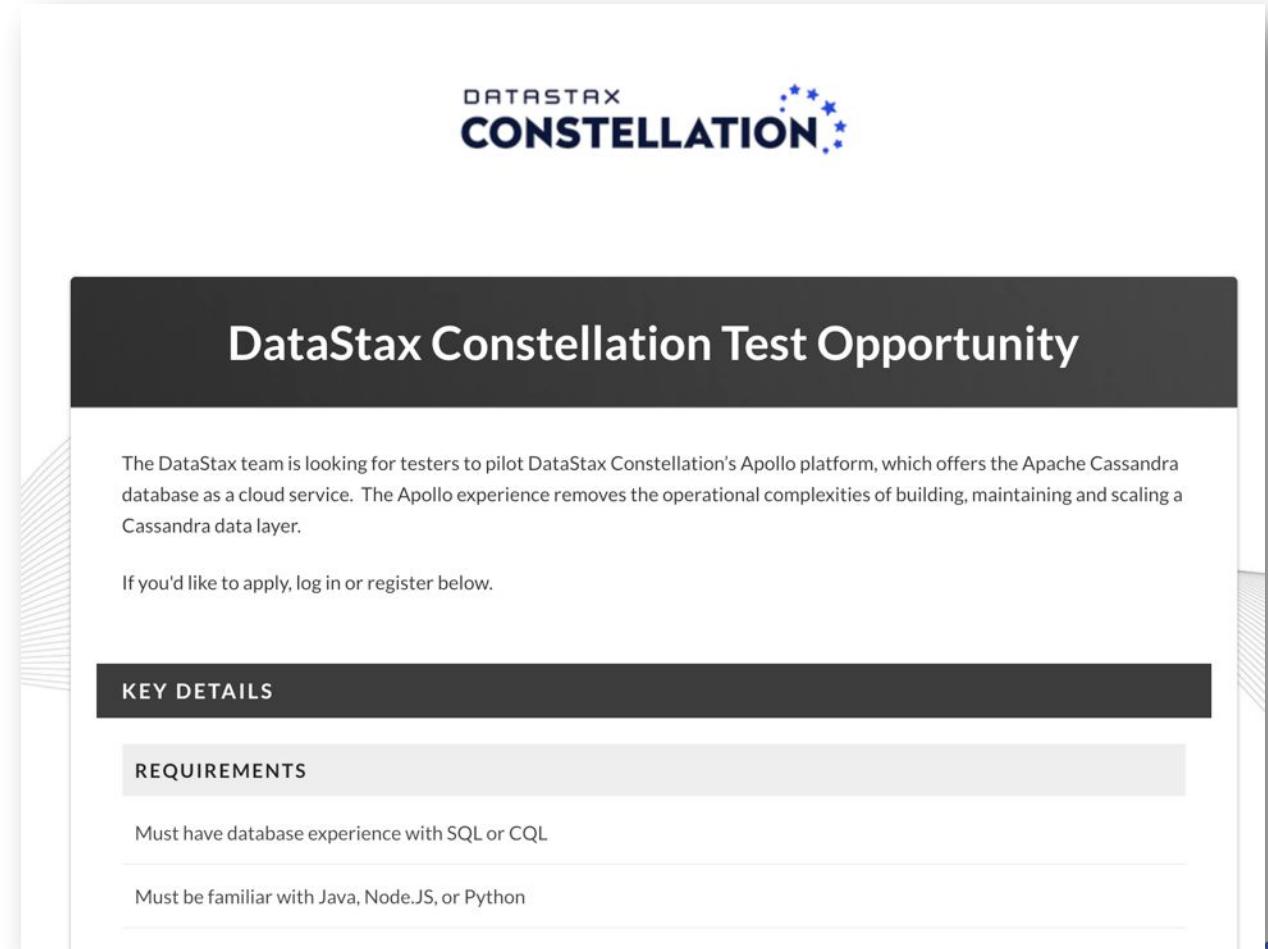
Constellation

<http://constellation.datastax.com>



The landing page for DataStax Constellation features a dark blue background with a network graph pattern. At the top left, it says "COMING SOON..." above the DataStax Constellation logo, which includes the word "Cloud Data Platform". Below the logo is a brief description: "A cloud-native platform with smart services that radically simplify and accelerate application development while eliminating the complex overhead of database operations." To the right is a sign-up form for the early access program. It includes fields for First Name, Last Name, Email Address, Company Name, Job Title, and Country, each with an asterisk indicating it is required. There is also a checkbox for participating in the early access program and a link for terms of use. A yellow "SUBMIT" button is at the bottom.

<https://datastax.centercode.com/key/opportunity>



The page for the DataStax Constellation Test Opportunity has a white background. At the top is the DataStax Constellation logo. Below it is a large dark header with the text "DataStax Constellation Test Opportunity". The main content area starts with a paragraph about the team looking for testers to pilot the Apollo platform. It then asks if you'd like to apply, log in or register below. Below this is a "KEY DETAILS" section, followed by a "REQUIREMENTS" section. The requirements listed are "Must have database experience with SQL or CQL" and "Must be familiar with Java, Node.JS, or Python". The DataStax logo is at the bottom right.

DataStax Labs (<https://downloads.datastax.com/#labs>)

Download DataStax

DataStax Desktop
With DataStax Desktop you're a few clicks away from a working DSE and DataStax Studio launched in a local or remote Kubernetes cluster! More to come!

Tools Drivers **Labs**

DataStax Labs

DataStax Labs provides the Apache Cassandra™ and DataStax communities with non-supported previews of enhancements that may or may not be included in future DataStax production software well as tools, aids, and partner software designed to increase productivity.

As a guest, have fun with DataStax Labs previews, and try it out. And note our disclaimer that these features are not supported, and so should not be put into production.

You try out some of our new Labs technologies, tools, and experimental features we would like to hear your feedback. Good or bad, let us know!

connect with us through the [DataStax Community](#).

"Who better to help us shape our software than our developers every day. We have a lot of ideas, but we want to make sure we're sharing them with the community and getting feedback from the cutting-edge builds and let us know."
- PATRICK MCFADIN, VP DEVELOPER RELATIONS, DATASTAX

DataStax CDC for Apache Kafka

DataStax CDC for Apache Kafka extends the existing Sink Connector with Source functionality. Now changes may be pushed from a source DataStax Enterprise cluster to Kafka topics. Change Data Capture events include inserts, updates, and deletes.

DataStax Proxy for DynamoDB™ and Apache Cassandra™

Preview version of an open source tool that enables developers to run their AWS DynamoDB™ workloads on Apache Cassandra™. With the proxy, developers can run DynamoDB workloads outside of AWS (including on-premises, other clouds, and in hybrid configurations).

DataStax Spring Boot Starter

The DataStax Spring Boot Starter streamlines the development of Spring applications with Cassandra and DataStax.



Insights (<https://www.datastax.com/products/datastax-insights>)



<https://youtu.be/iZ47rrKENuc>



Thank You

@CLUNVEN
@HadesArchitect

