DATASTAX ACADEMY

Cassandra and the Multi-Cloud

Amanda K Moran
Developer Advocate for DataStax

But first… A Little About Amanda

- Graduated with MS in Computer Science and Engineering from Santa Clara University in 2012
- Worked as a Software Engineer for 6 years and now is a Developer Advocate
- Apache Committer, PMC Member, and initial contributor to all installation and deployment work for Apache Trafodion
- Keywords: Disney, Cloud, Dogs, Veggies, Linux, Databases, Big Data, Analytics, Testing, and Running

What Are We Talking About Today

- Introduction to Apache Cassandra
- What are Multiple DataCenters?
- Why all this talk about MultiCloud?!
- Apache Cassandra and the MultiCloud
- Demo!

Introduction to Apache Cassandra

What is Apache Cassandra?

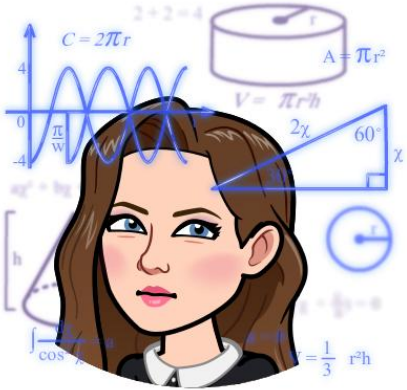
- First developed by Facebook
- Became a top-level Apache Foundation project in 2010
- NoSQL database
- ***Distributed***, decentralized database
- Elastic scalability -- add/remove nodes with no downtime

What is Apache Cassandra?

- High performance
  - Very fast -- low latency
- High availability / fault tolerant
  - No single point of failure
- Solves many of the problems faced with a traditional DB for certain workloads

What Does All This Mean?

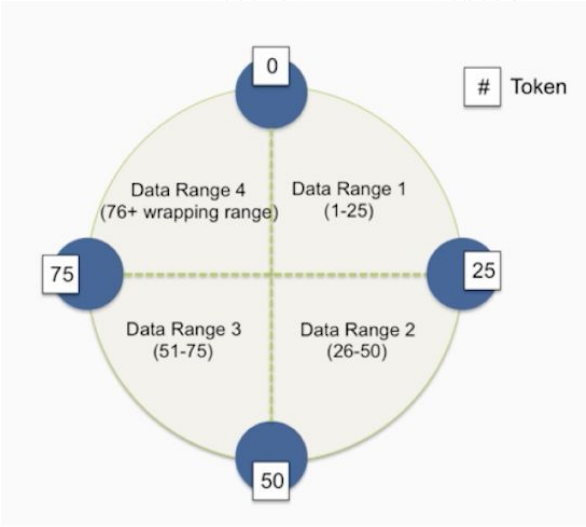
- Let's talk about the Big Topics:
  - Distributed Systems
  - Replication
  - Elastically Scalable
  - High Availability
  - Latency
    - Read path
    - Write path



Note: Don't forget this is just a brief intro!

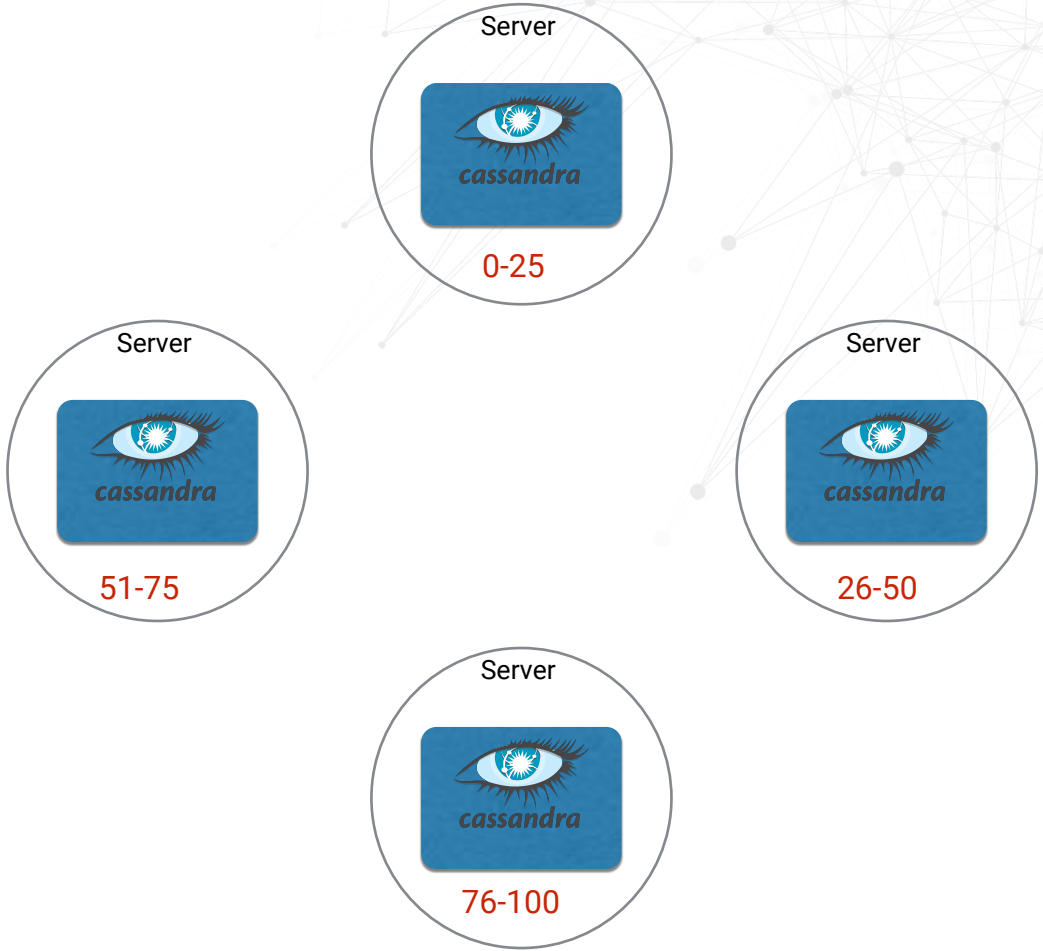
Distributed System

- Every node in the cluster has the same role
  - Really!
  - Cassandra does not have a Master-Worker Architecture
- Any client can connect to any node
  - All nodes are Read and Write ready
- But this is not to say that all nodes contain all data



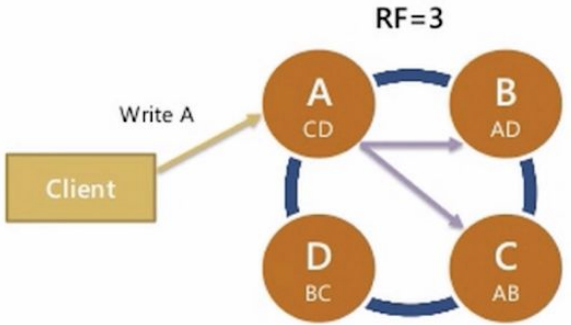
The cluster

Token	Range
0	0-25
26	26-50
51	51-75
76	76-100

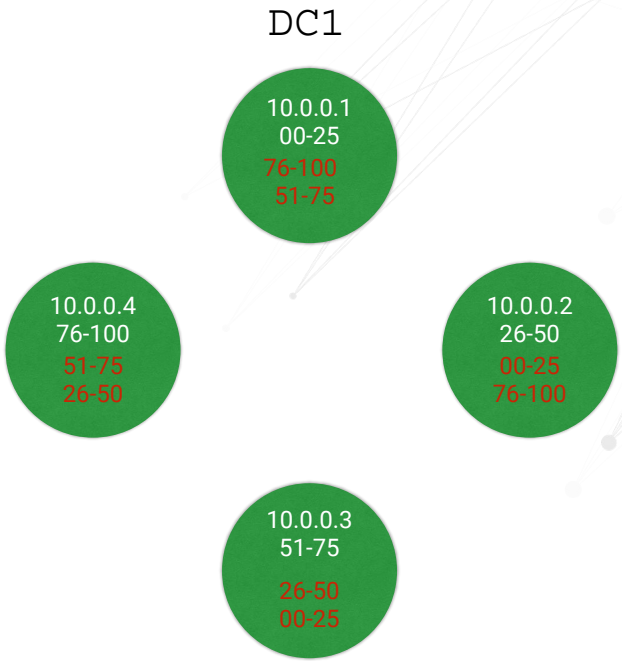


Replication

- To be able to survive a node going down data must be copied to other nodes
- The Replication Factor (RF) is set by the user
  - 1-Number of nodes in the Cluster (not recommended)
- The data is asynchronously replicated
  - Automatic
  - Peer-to-peer communication



Replication

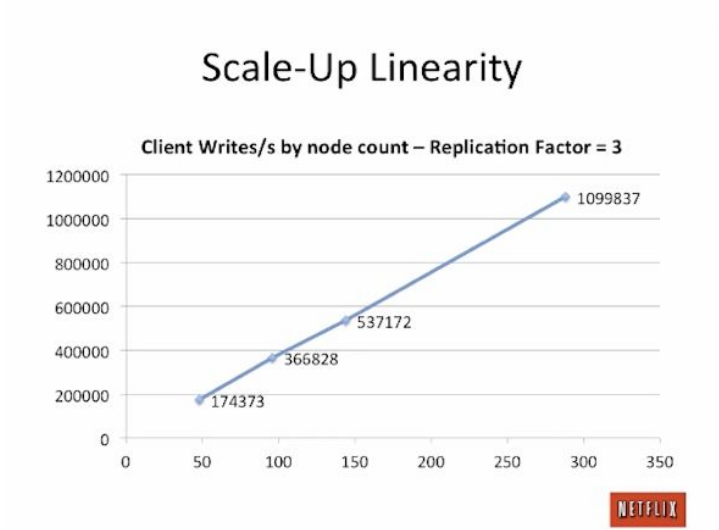


DC1: RF=3

Node	Primary	Replica	Replica
10.0.0.1	00-25	76-100	51-75
10.0.0.2	26-50	00-25	76-100
10.0.0.3	51-75	26-50	00-25
10.0.0.4	76-100	51-75	26-50

Elastically Scalable

- As more nodes are added, performance increases linearly
- You can scale up or down with no downtime
  - Not even a restart!
- Reads and Writes both scale



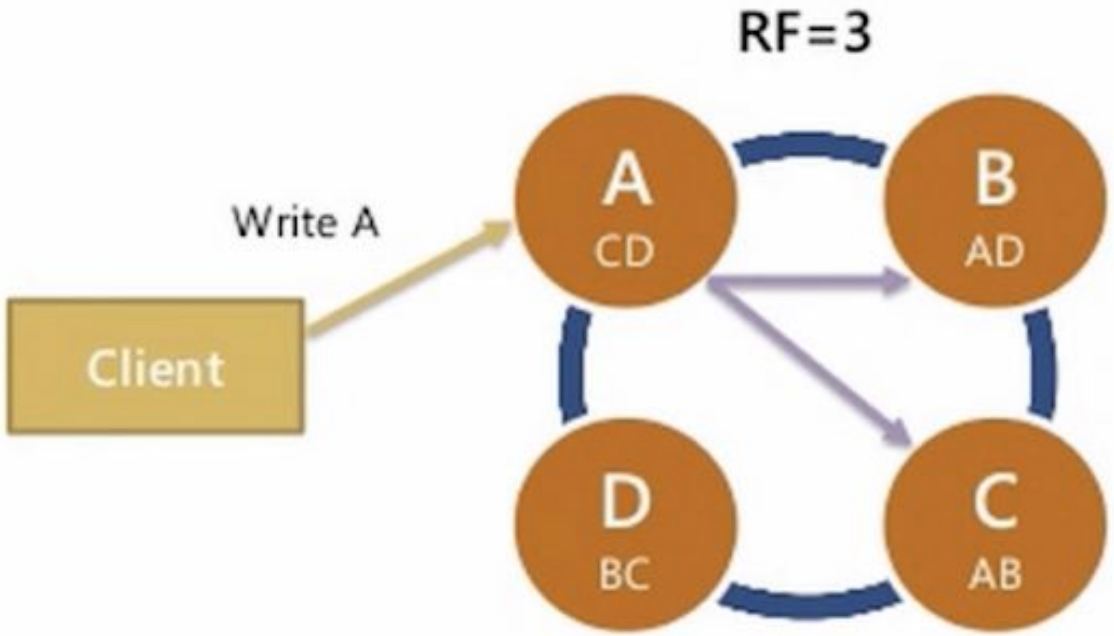
High Availability

- The lack of a Master node allows for high availability
  - No single point of failure
- Replication allows nodes to fail and data to still be available
  - Cassandra expects nodes to fail and doesn't panic

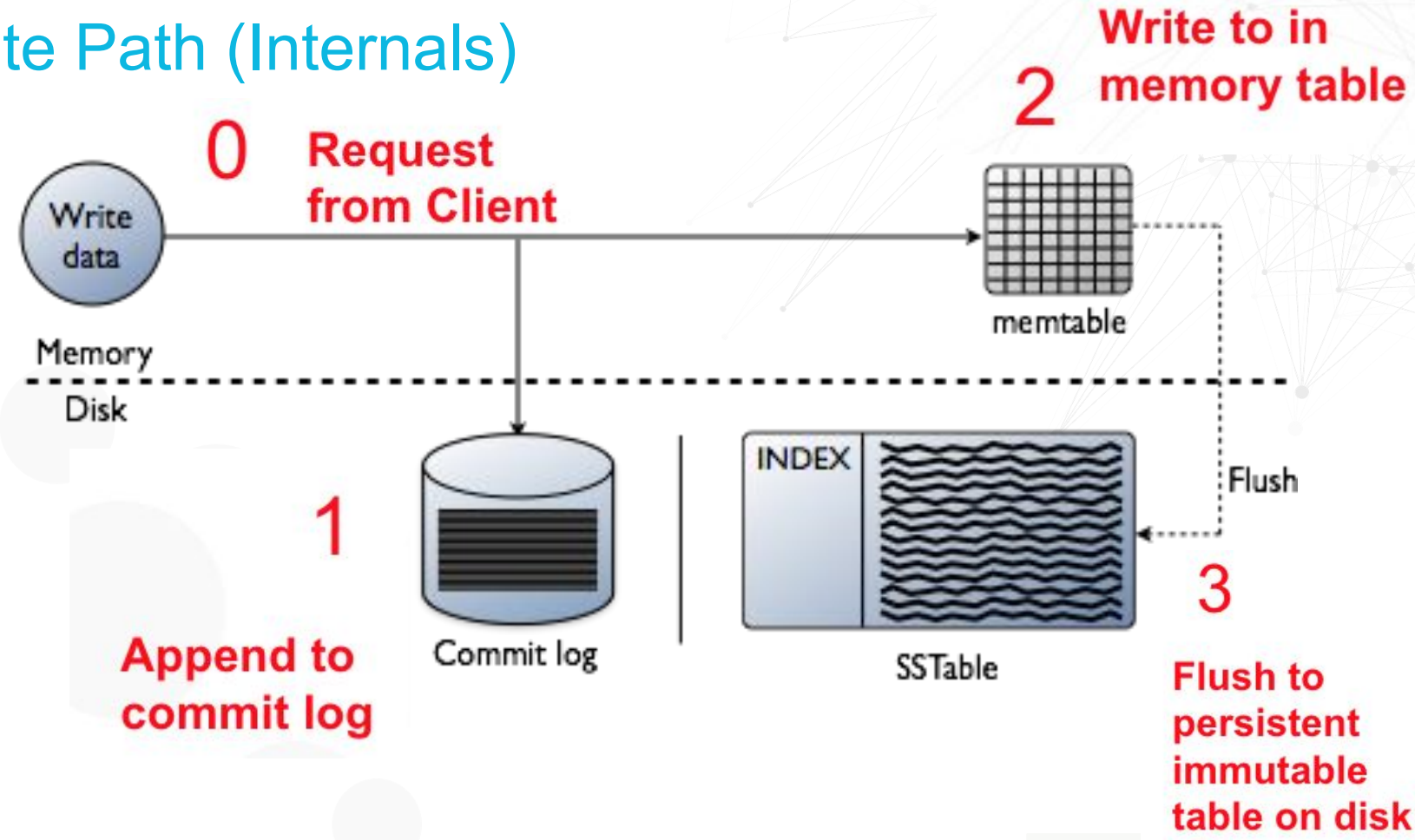
Latency

- How is Apache Cassandra able to achieve such low latency?
- It's all about the read and write path!
  - The write path is truly beautiful in its simplicity!
  - High throughput with quick responses times are easy to achieve

Write Path (Client to Cluster)



Write Path (Internals)



Read Path (Client to Cluster)

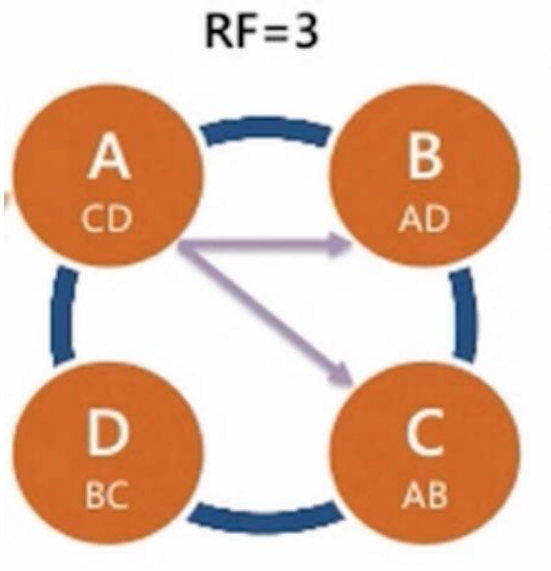
- Data modeling comes in to play here!
  - This is the one “simple trick about Cassandra/Nosql”
- Partition data by nodes
- Query will essentially query one node and return the data
  - Constant time READ access

```
select * from myTable where state =`CA`
```

Looks like SQL but it IS NOT! It’s Cassandra Query Language

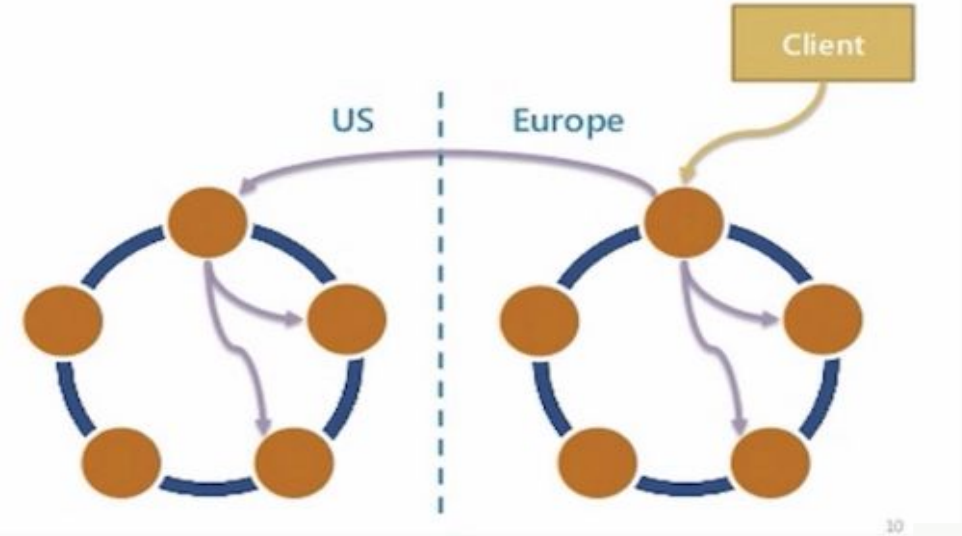
Multiple Data Centers

- Cassandra Cluster represented as a ring
- Can support multiple rings chained together
  - Separated by region
  - Separated by workload



Multiple Data Centers

- Multiple Data Center support is out of the box
- Replication happens between data centers automatically
  - No need to sync data



What is DataStax?

- DataStax is the enterprise version of Apache Cassandra
- 70% of the the commits to the open source project
- 2x the Write performance of Apache Cassandra
- 2x the Read performance
- Add in the ability to do Search, Analytics, and Grpah
- Cool tools!

WHY Multi-Cloud?

What is Multi-Cloud?

- Two or more public cloud providers at the same time
  - Data moving between two++ providers

Google Cloud

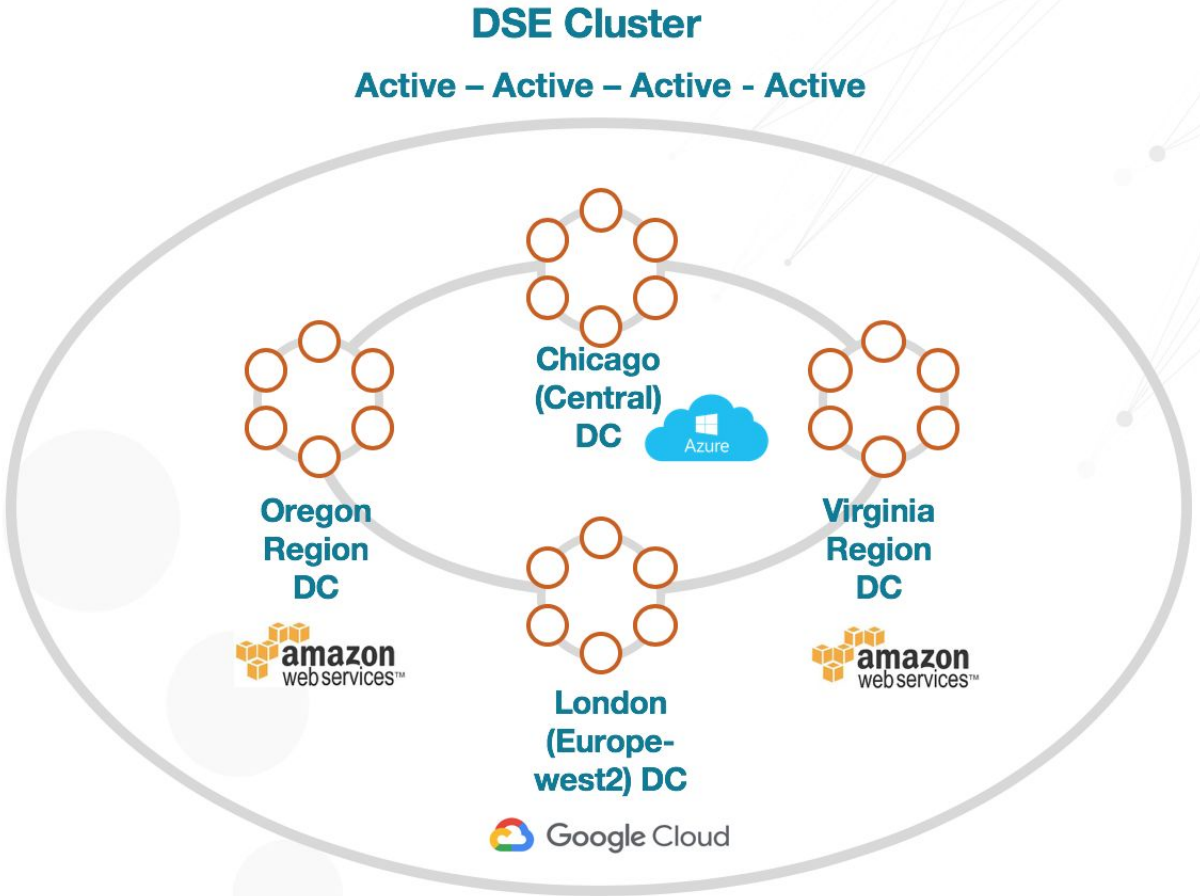
Why Multi-Cloud?

- Data Center Locality
  - Not all zones are in each provider
- Provider Specific Services
- Cloud provider competition
  - Can shop around cheap compute! -- Maybe
  - More likely -- afraid of lock in to a competitor
- Cost

Why can Apache Cassandra do MultiCloud?

- Multi Data Center support -- out of the box
- Cloud Native database
  - Built for the cloud
    - Multi region support
  - Expanded to Hybrid cloud
  - Easy expansion to Multi-Cloud
- Every node has the same job

Only Database that supports Multi-Cloud



DATASTAX

Only Database that supports Multi-Cloud

Demo

Why NOT Multi Cloud?

Issues with Multi-Cloud

- Complexity
- Networking!
  - Latency
- Security
  - Boundary protection
- Legal
- Scaling at the Application layer

Okay, this was awesome! What now?

Information and Links

- Learn more about Cassandra: https://academy.datastax.com/
- Learn more about DataStax: https://www.datastax.com/
- Follow me on Twitter: @AmandaK_Data
- Github: https://github.com/amandamoran

DATASTAX ACADEMY

Join us at Accelerate!!

www.datastax.com/accelerate

Discount Code: ADVOCATE20