DATASTAX ACADEMY

Classification and Clustering

Paired with Wine and Chocolate

Amanda K Moran
Developer Advocate for DataStax

But first… A Little About Amanda

- Lived in Seattle for 10 years! University of Washington Alumni -- BS Biology
- Graduated with MS in Computer Science and Engineering from Santa Clara University in 2012
- Worked as a Software Engineer for 6 years at Lockheed Martin, HP, Teradata, Esgyn and now DataStax as 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

- What Problem Are We Trying To Solve?
  - Can I use clustering ML algorithms to find which wine comes from which vineyard?
  - Can I use classification ML algorithms to find which country a candy bar comes from?
- Introduction to Apache Cassandra
- Introduction To Apache Spark
- What is KMeans?
  - Demo: KMeans and Wine
- What is Naive Bayes?
  - Demo: Naive Bayes and Chocolate

What Problem Are We *Really* Trying to Solve?

Can I use Machine Learning with Apache Spark with Wine and Chocolate?

- Can I use clustering ML algorithms to find which wine comes from which vineyard?
- Can I use classification ML algorithms to find which country a candy bar comes from?
- Data Analytics doesn’t have to be complicated!
- We are going to utilize the power of Big Data using
  - Apache Cassandra
  - Apache Spark
  - Spark Machine Learning library
    - Kmeans
    - Naive Bayes
  - Jupyter notebooks
  - Python

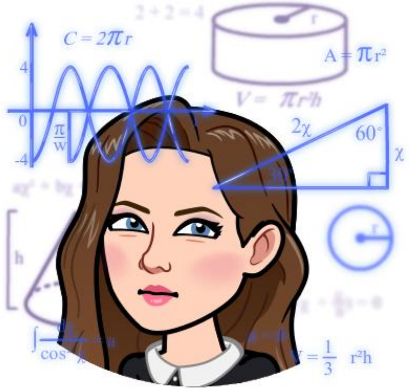
Introduction to Apache Cassandra

What is Apache Cassandra?

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

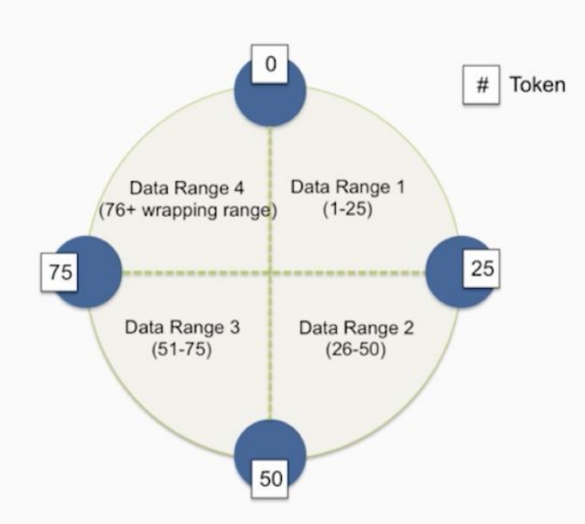
- Let’s talk about 4 Big Topics:
  - Distributed
  - Replication
  - Elastically Scalable
  - High Availability



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

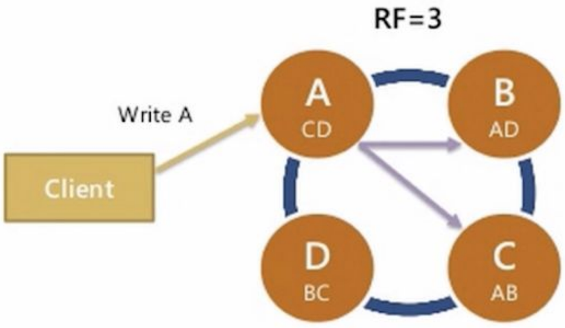
Distributed

- 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



Replication

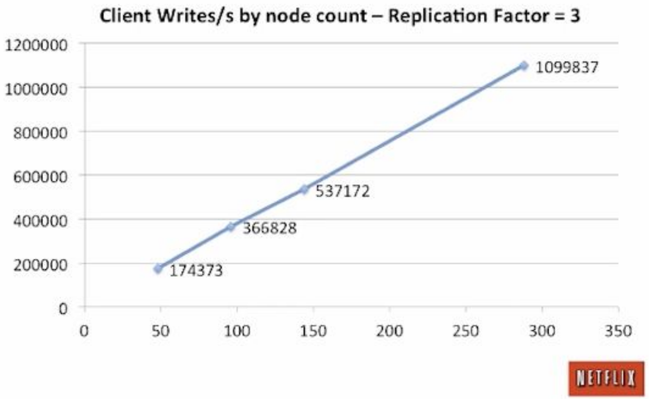
- To be able to survive a node going down data must be copied to other nodes
- The Replication Factor (RF) is set by the user
  - 1-Number of nodes in the Cluster (not recommended)
- The data is asynchronously replicated
  - Automatic
  - Peer-to-peer communication



Elastically Scalable

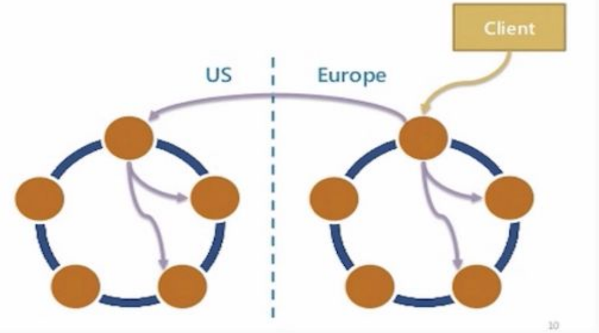
- As more nodes are added, performance increases linearly
- You can scale up or down with no downtime
  - Not even a restart!
- Reads and Writes both scale

Scale-Up Linearity



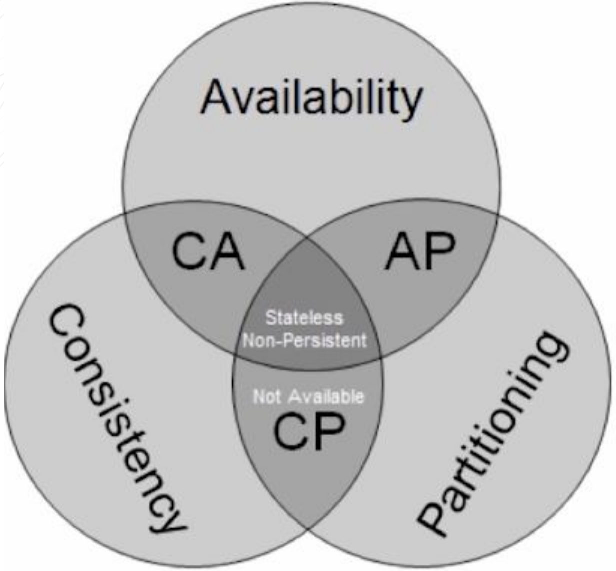
High Availability

- 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
- Multiple Data Center support is out of the box



One Small Trade Off

- The CAP Theorem
  - Availability
  - Consistency
  - Partitioning
- You can’t have it all --Impossible
  - Cassandra chooses to have eventual consistency as the default
- But you can prioritize consistency over availability
  - Consistency levels are configurable!



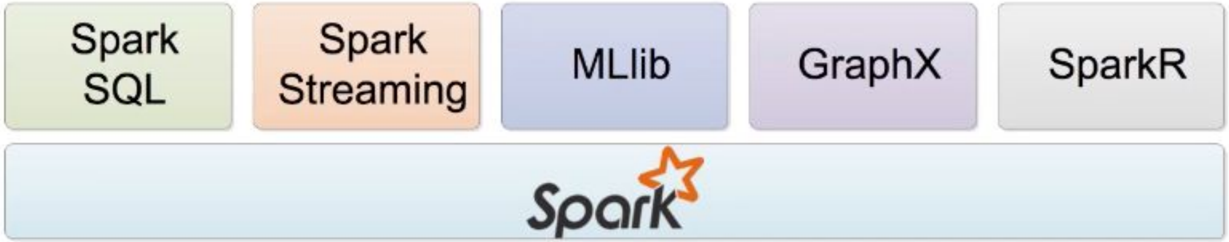
Why Do I Need Cassandra?

- Think about your application:
  - Do you have Big Data (a lot of it!)?
  - Do you need to be able to read/write fast?
  - Do you need to be able to scale up/down easily?
  - Do you need High Availability?
  - Do you need multiple data center support?
    - Multi-cloud/hybrid cloud support?

(Small) Introduction to Apache Spark

What is Apache Spark?

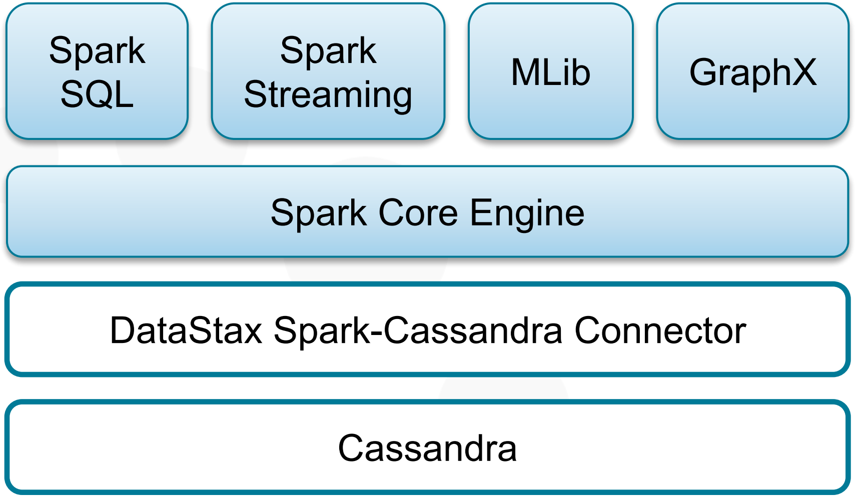
- “Apache Spark is a unified analytics engine for large-scale data processing” --https://spark.apache.org
- 100 times faster than Hadoop for analytics
  - Utilizes in-memory processing
  - Amazing parallelism
- Machine Learning library -- Spark MLlib



What is DataStax Analytics?

DSE Analytics: Spark + Cassandra

Read/write Cassandra data from Spark via DataStax Connector



Spark SQL

Spark Streaming

MLib

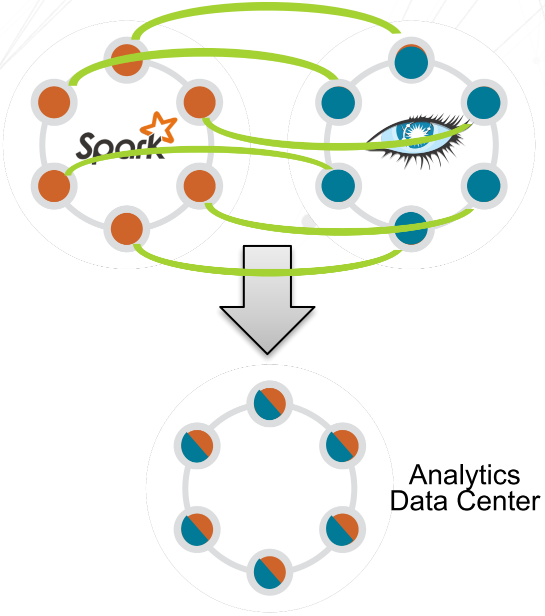
GraphX

Spark Core Engine

DataStax Spark-Cassandra Connector

Cassandra

Co-located Spark with Cassandra



Analytics
Data Center

What is Clustering?

What are Clustering Algorithms ?

- **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

What is KMeans ?

- K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems.
- It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand.

What is KMeans ?

- The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

When to use KMeans?

- The *K*-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.
  - Behavior Segmentation of customers
    - Segment by purchase history
  - Finding anomalies

Demo: KMeans and Wine

Can I use clustering ML algorithms to find which wine comes from which vineyard?

What is Classification?

What are Classification Algorithms ?

- In machine learning and statistics, **classification** is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

What are Classification Algorithms ?

- Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.

What is Naive Bayes ?

- Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.
- There are many algorithms implemented for Naive Bayes classifiers but all assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Demo: Naive Bayes and Chocolate

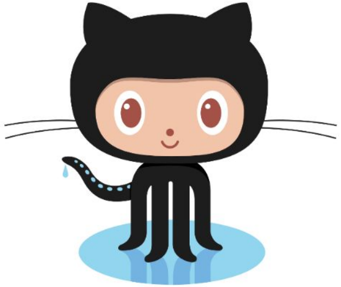
Can I use classification ML algorithms to find which country a candy bar comes from?

Okay, this was awesome! What now?

Information and Links

- Learn more about Cassandra: https://academy.datastax.com/
- Learn more about Spark: https://spark.apache.org/
- Learn more about DataStax: https://www.datastax.com/
- Follow me on Twitter: @AmandaDataStax
- Get this demo on Github: https://github.com/amandamoran/wineAndChocolate

DATASTAX ACADEMY



Thank you