When Rotten Tomatoes Isn’t Enough

Analyzing Twitter Movie Reviews using DataStax Enterprise

PyData

Amanda K Moran
Developer Advocate for DataStax

Agenda

- Introduction to DataStax, Apache Cassandra, Apache Spark, and Sentiment Analysis
- Overview of the Demo
- Walk through the notebook
- Installation of Docker and DataStax with Jupyter --Walk through together
- Free time to explore the notebooks on your own

But first… A Little About Amanda

- California Native born in Redlands #siliconbeach
- 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?
- Introduction to Apache Cassandra
  - What is it?
  - Why do I need it?
- Introduction To Apache Spark
  - What is it?
  - Why do I need it?
- What is Sentiment Analysis?
- How does DataStax Enterprise Analytics Help Us?
- Overview of Demo
- Setup and play with Demo on your laptop with Docker!

What Problem Are We *Really* Trying to Solve?

What Movie Should I See?

- Wouldn't it be great if I could ask 1 million people this question?
- Wouldn't it be great if I could automate this process?
- 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
  - Jupyter notebooks
  - Python
  - Twitter Tweets and API
  - Pattern for Sentiment Analysis

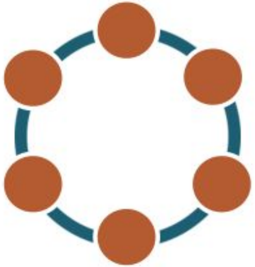
(Brief) Introduction to Apache Cassandra

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 is DataStax Enterprise?

- DataStax has been some of the key contributors to the Cassandra project
  - DataStax Enterprise is a commercial product that provides
    - More cool features
    - More QA
    - More support



The power behind the moment.

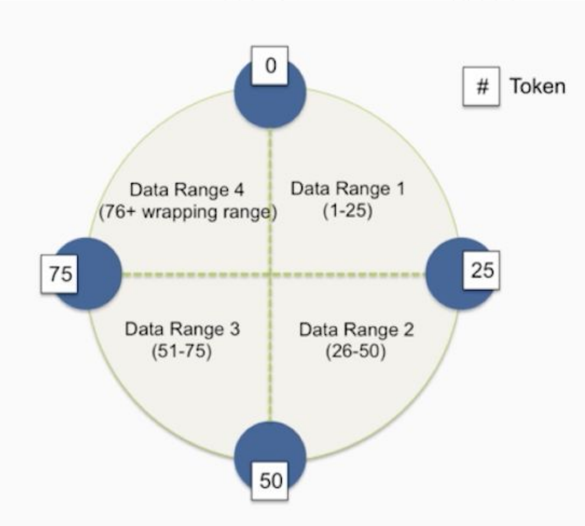
What Does All This Mean?

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

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

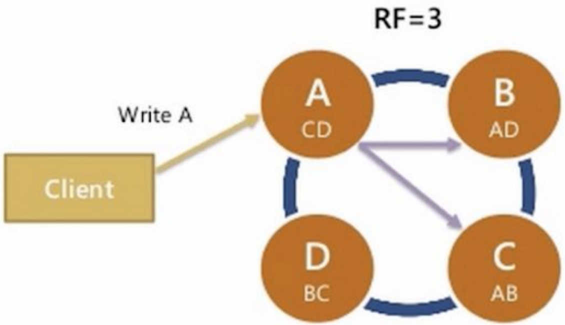
Distributed

- 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

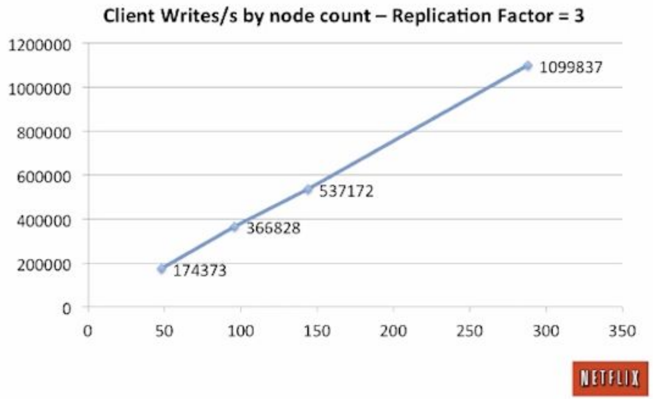
- To be able to survive a node going down data must be copied to other nodes
- The Replication Factor (RF) is set by the user
  - 1-Number of nodes in the Cluster (not recommended)
- The data is asynchronously replicated
  - Automatic
  - Peer-to-peer communication



Elastically Scalable

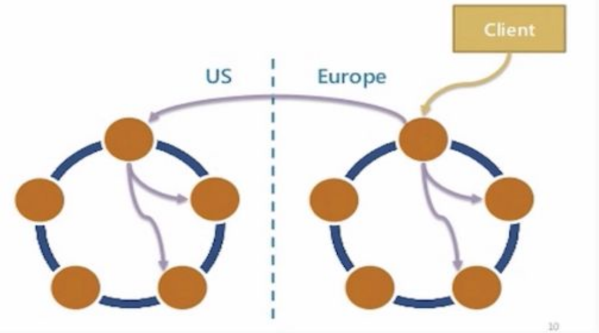
- As more nodes are added, performance increases linearly
- You can scale up or down with no downtime
  - Not even a restart!
- Reads and Writes both scale

Scale-Up Linearity



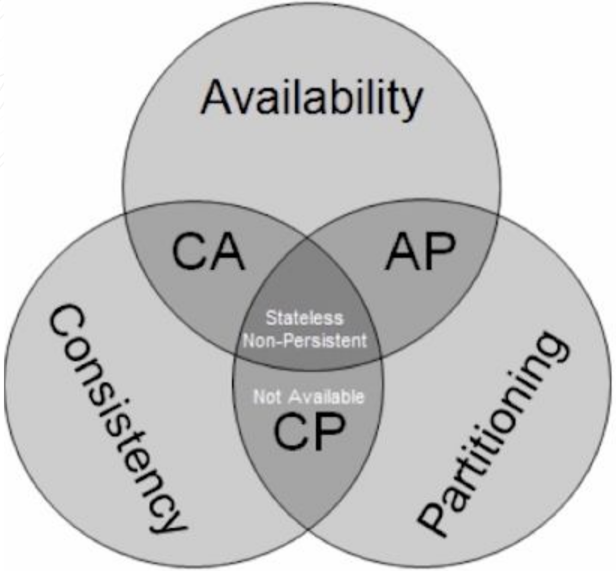
High Availability

- 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?

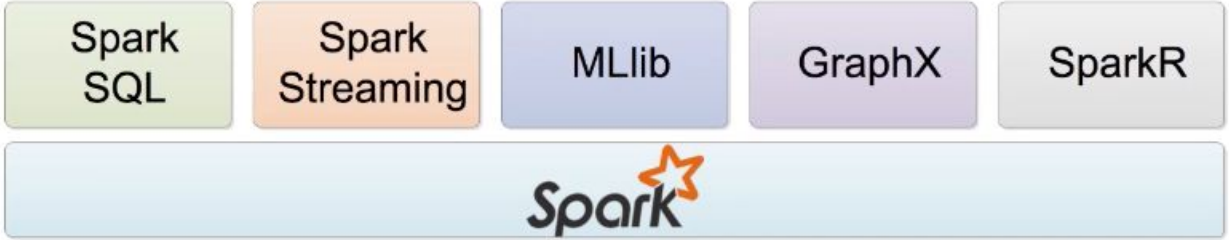
- Think about your application:
  - Do you have Big Data (a lot of it!)?
  - Do you need to be able to read/write fast?
  - Do you need to be able to scale up/down easily?
  - Do you need High Availability?
  - Do you need multiple data center support?
    - Multi-cloud/hybrid cloud support?

If you application needs any of these things, you want to consider Cassandra!

(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



Why Do I Need Spark?

- Think about your application:
  - Do you have Big Data?
  - Do you need High Availability?
  - Do you need analytics at lightning speed?
  - Do you need a simple way to get insights into your data?

If you application needs any of these things, you want to consider Spark!

What is Sentiment Analysis?

What is Sentiment Analysis?

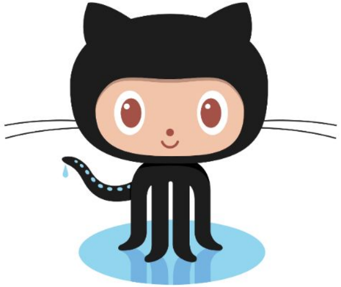
- Sentiment Analysis at a high level is very simple
- Natural Language processing and text analytics to determine if a word or sentence
  - Positive
  - Negative
  - Neutral
- This is easy to understand, but difficult for machines to learn how to do!

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
- Check us out on Twitch: https://www.twitch.tv/datastaxacademy

DATASTAX ACADEMY

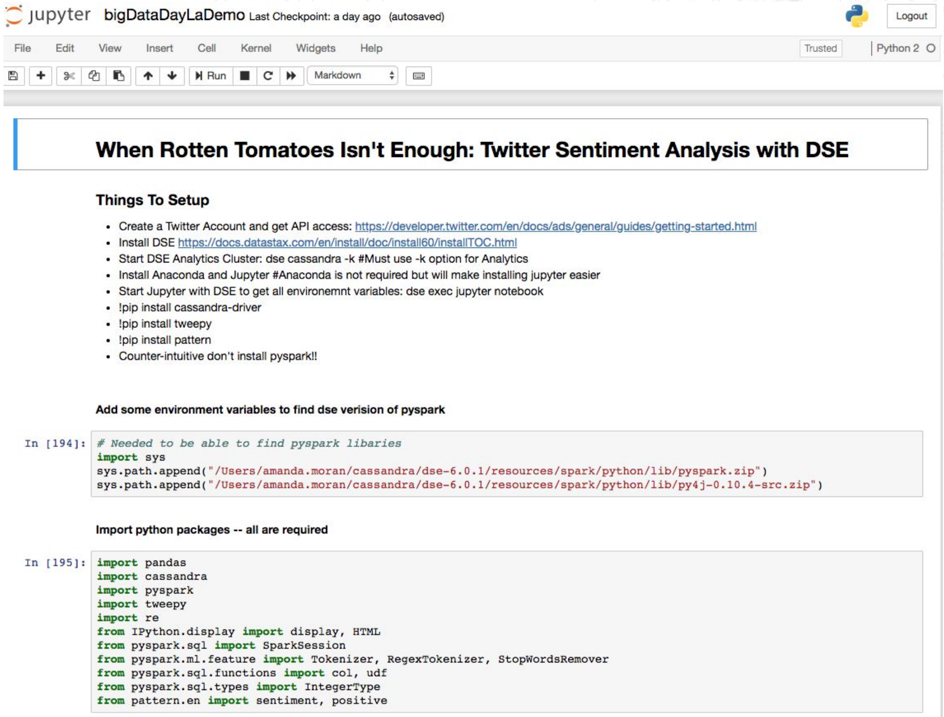


DATASTAX
The power behind the moment.
Amanda
@AmandaDataStax

Demo Overview

Analyzing Twitter with DSE Analytics and Jupyter

- Install Docker
- Install/Run DSE/Jupyter Image
- Create keyspace and tables
- Insert into Cassandra
- Create Spark Dataframe
- Use Spark ML
- Sentiment Analysis with Pattern
  - Gives positive/negative
- Take average of these scores
- Should I see this movie??



When Rotten Tomatoes Isn't Enough: Twitter Sentiment Analysis with DSE

Things To Setup

- Create a Twitter Account and get API access: https://developer.twitter.com/en/docs/ads/general/guides/getting-started.html
- Install DSE https://docs.datastax.com/en/install/doc/install60/installTOC.html
- Start DSE Analytics Cluster: dse cassandra -k #Must use -k option for Analytics
- Install Anaconda and Jupyter #Anaconda is not required but will make installing jupyter easier
- Start Jupyter with DSE to get all environemnt variables: dse exec jupyter notebook
- !pip install cassandra-driver
- !pip install tweepy
- !pip install pattern
- Counter-intuitive don't install pyspark!!

Add some environment variables to find dse verision of pyspark

In [194]:
```
# Needed to be able to find pyspark libaries
import sys
sys.path.append("/Users/amanda.moran/cassandra/dse-6.0.1/resources/spark/python/lib/pyspark.zip")
sys.path.append("/Users/amanda.moran/cassandra/dse-6.0.1/resources/spark/python/lib/py4j-0.10.4-src.zip")
```

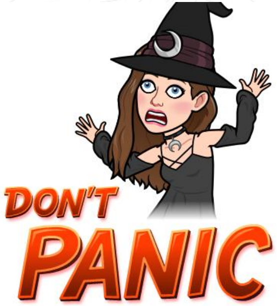
Import python packages -- all are required

In [195]:
```
import pandas
import cassandra
import pyspark
import tweepy
import re
from IPython.display import display, HTML
from pyspark.sql import SparkSession
from pyspark.ml.feature import Tokenizer, RegexTokenizer, StopWordsRemover
from pyspark.sql.functions import col, udf
from pyspark.sql.types import IntegerType
from pattern.en import sentiment, positive
```

Hands On Time!

Cassandra, Spark, Python, Jupyter Notebooks, Cassandra Python driver, Tweets, and Pattern

Instructions all on Github



- Follow along with me
  - OR go it alone… if you dare!
  - https://github.com/amandamoran/pydata
- Download or do a git clone
- Double-click on Getting Started.html to view a notebook in your browser.
  - Install Docker, then pull down Docker DSE/Jupyter Image. (about 6 minutes)

Thank you

WE ARE HIRING!