Apache Spark at Home

A Case Study with Twitter Sentiment Analysis

Frazier Baker

Background Derived from Photo by Amador Loureiro on Unsplash

Overview

- Building a Spark Cluster
  - Apache Spark
  - Installation / Setup
- Running Twitter Sentiment Analysis
  - Sentiment Analysis
  - Stanford CoreNLP
  - Twitter
  - Preliminary Results
  - Performance Analysis
- Future Directions
  - Future of the Cluster
  - Future of my NLP Work

Building an Apache Spark Cluster

What is Apache Spark?

- Distributed Computing
- Versatile & General
- Built for Big Data

Why I Used Spark

- Existing Resources
- Pyspark is Gorgeously Simple
- Interest in Distributed/Cloud Computing

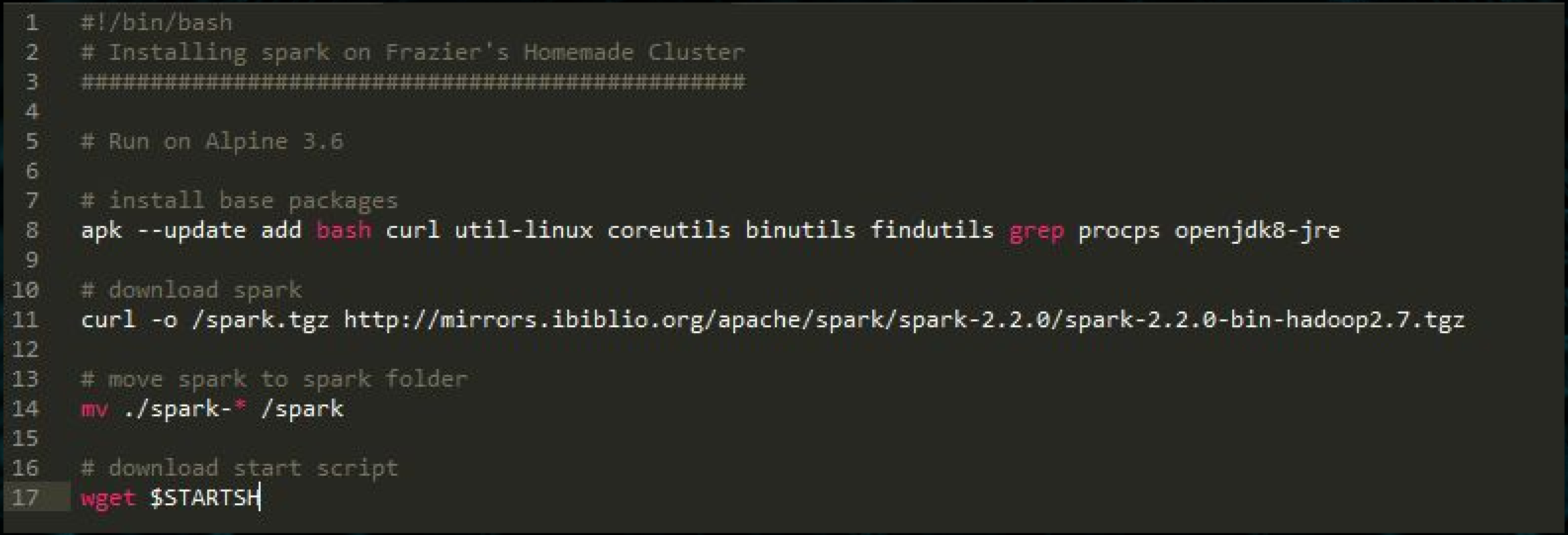
Python Logo Source: Wikimedia Commons,
https://upload.wikimedia.org/wikipedia/commons/thumb/0/0a/Python.svg/2000px-Python.svg.png

Photo by Pero Kalimero on Unsplash

Installation / Setup

Running Alpine Linux 3.6

Connected to LAN with Static IP; Connected to Internet via Proxy



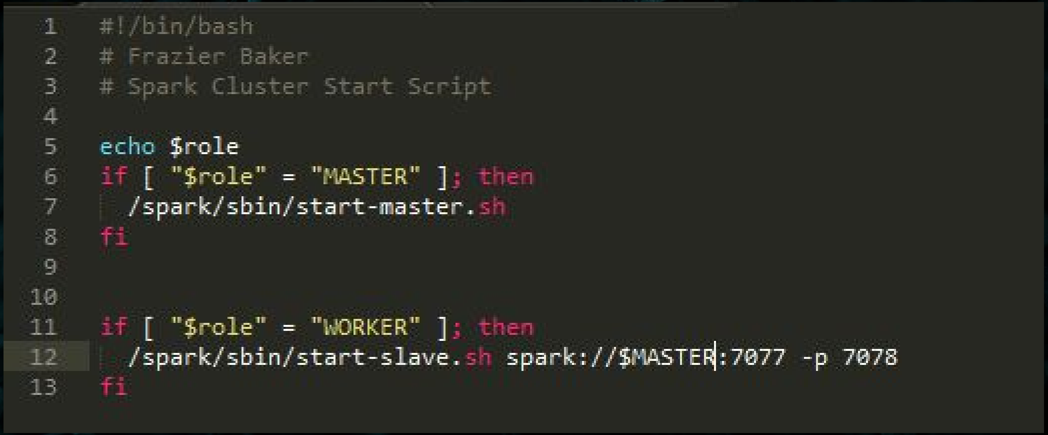
```
#!/bin/bash
# Installing spark on Frazier's Homemade Cluster
################################################

# Run on Alpine 3.6

# install base packages
apk --update add bash curl util-linux coreutils binutils findutils grep procps openjdk8-jre

# download spark
curl -o /spark.tgz http://mirrors.ibiblio.org/apache/spark/spark-2.2.0/spark-2.2.0-bin-hadoop2.7.tgz

# move spark to spark folder
mv ./spark-* /spark

# download start script
wget $STARTSH
```

Installation / Setup

Start Script simply runs Spark’s Built-in Start Scripts

Run `role=MASTER ./start.sh` on machine you want to be master

Run `MASTER=hostname_of_master role=WORKER ./start.sh`
on all other machines

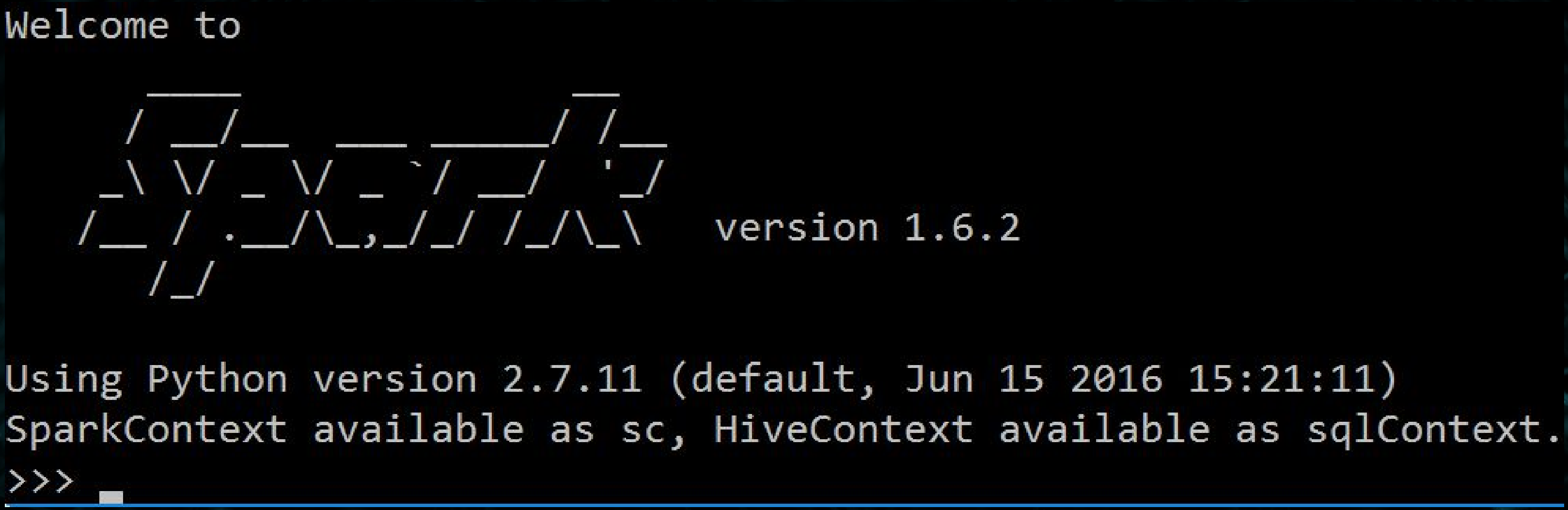


```bash
#!/bin/bash
# Frazier Baker
# Spark Cluster Start Script

echo $role
if [ "$role" = "MASTER" ]; then
  /spark/sbin/start-master.sh
fi


if [ "$role" = "WORKER" ]; then
  /spark/sbin/start-slave.sh spark://$MASTER:7077 -p 7078
fi
```

Installation / Setup

Run `pyspark`

Do pythony stuff



```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.6.2
      /_/

Using Python version 2.7.11 (default, Jun 15 2016 15:21:11)
SparkContext available as sc, HiveContext available as sqlContext.
>>>
```

Image Source: https://yokekeong.com/running-apache-spark-with-sparklyr-and-r-in-windows/

What about Docker?

Docker is wonderful for controlling environment

But Docker has overhead

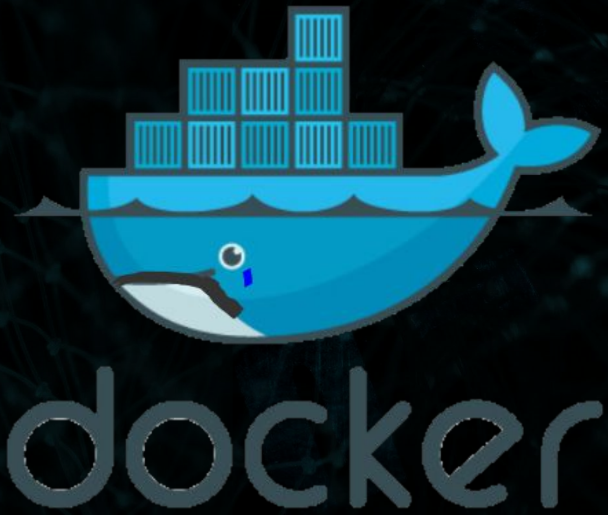


Image modified from photo found at:
https://www.flickr.com/photos/xmodulo/14098888813

Sentiment Analysis on Twitter

Sentiment Analysis

“I hate you.”

“You’re really great!”

“I like puppies.”

“I couldn’t imagine not liking puppies.”

Stanford CoreNLP

Built in Java

Run as Java Server on each node

Access through Python API Wrapper

0=VeryNegative 1=Negative 2=Neutral 3=Positive 4=VeryPositive

Twitter Data

Publicly available data

Simple search API

?keywords=@username

Limitations of the Twitter API

- Only over last 7 days

- Max 3200 tweets, 100 at a time

- Bigger data transactions require a bigger wallet

Preliminary Results

Preliminary Results

uofcincy,CarnegieMellon,miamiuniversity, Harvard,MIT,OhioState,Stanford,UCBerkeley, XavierUniv

11696 Tweets
~1.5MB of data

Negative Sentiment [-1, -0.6]

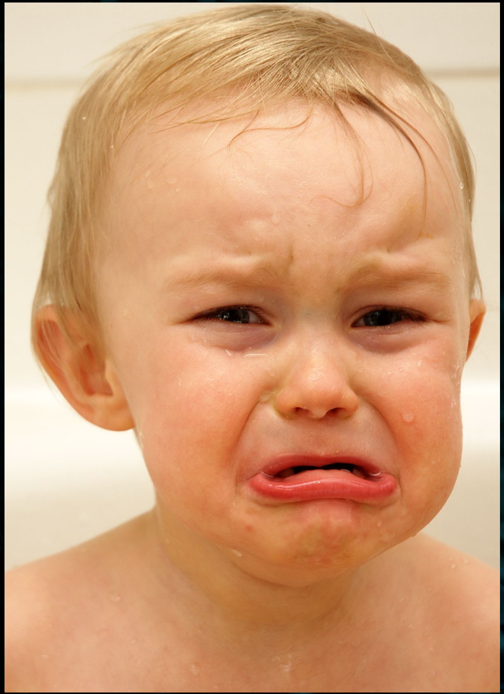
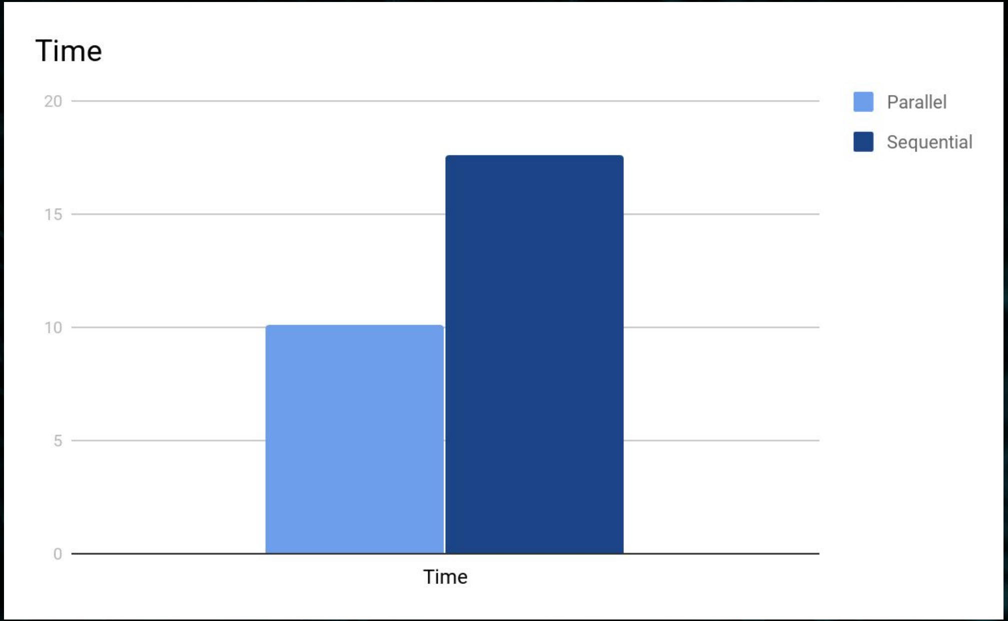


Image Source: Kyle Flood https://commons.wikimedia.org/wiki/File:Waaah!.jpg

Performance Analysis

Parallel	Sequential
10min 7.796s	17min 34.547s



Future Directions

NLP Experiments

- Get better data/more data

- Focus on measuring stress and more complex emotions

Cluster

- Likely use in future projects
    - Potentially in my Senior Design Project

- I have more old computers to add
    - And will probably continue to collect them as years go on

Acknowledgements

- Professor Fred Annexstein
  - For teaching the course that required this project, gave me an excuse to build the cluster I've been wanting to build for years, and offering feedback during office hours.

- Professor Paul Talaga
  - For getting me interested in Cloud Computing and teaching me some spark basics in his CS 6065 class before he left UC.

- Professor Shomir Wilson
  - For advice during his office hours on NLP and pointing me towards some cool resources

- Grace Gamstetter
  - Originally my partner on this project, we discussed ideas early on in the semester and settled the idea of doing NLP with Twitter data.

Questions?

References

Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: cluster computing with working sets. In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing (HotCloud'10). USENIX Association, Berkeley, CA, USA, 10-10.

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.[pdf][bib]

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).

Dirk Merkel. 2014. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014, 239, pages.