MSDS 498 CAPSTONE

Dan Kuratko

MY PROJECT

PROJECT 3 CLOUD MAPREDUCE

- **Project Objectives**
  - Implement Chapter 16 in *Python for Programmers* on MS Azure
  - Port the same example to run on Google Cloud Platform
- **Specific Objectives**
  - Manipulate SQL relational database
  - Store tweets in a MongoDB NoSQL JSON database, visualize via Folium map
  - Execute Hadoop MapReduce using MS Azure HDInsight and GCP
  - Use Apache Spark to process data in mini-batches with Azure and GCP
  - Visualize IoT data via publish/subscribe model

SQL and NoSQL

SQL

- Review of SQL fundamentals
- Book database example
- SQLite library within Python

`titles` Table

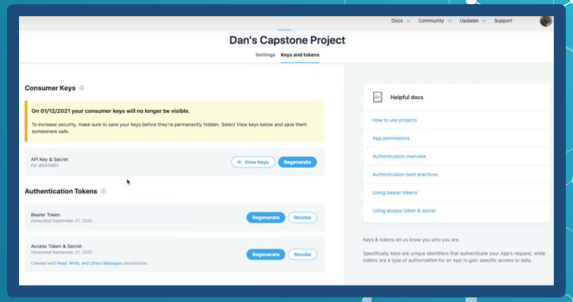
```
[6]: pd.read_sql('SELECT * FROM titles', connection)
```

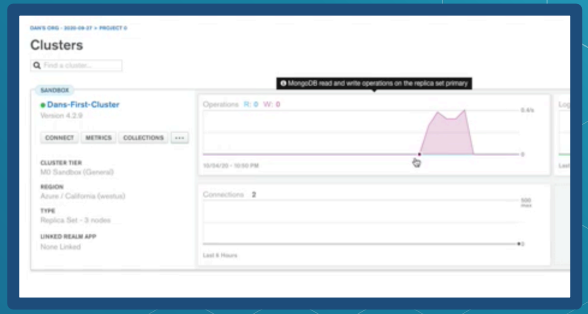
[6]:

	isbn	title	edition	copyright
0	0135404673	Intro to Python for CS and DS	1	2020
1	0132151006	Internet & WWW How to Program	5	2012
2	0134743350	Java How to Program	11	2018
3	0133976890	C How to Program	8	2016
4	0133406954	Visual Basic 2012 How to Program	6	2014
5	0134601548	Visual C# How to Program	6	2017
6	0136151574	Visual C++ How to Program	2	2008
7	0134448235	C++ How to Program	10	2017
8	0134444302	Android How to Program	3	2017
9	0134289366	Android 6 for Programmers	3	2016

NoSQL

- Create a MongoDB Atlas Cluster
- Tweet Streaming with Tweepy
- Use Folium maps to visualize tweet density



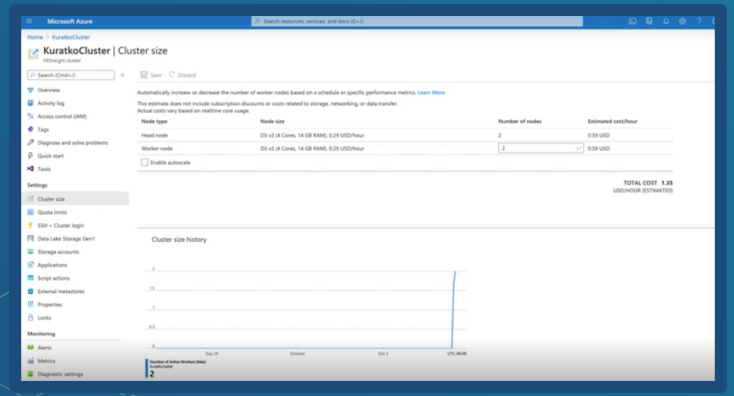


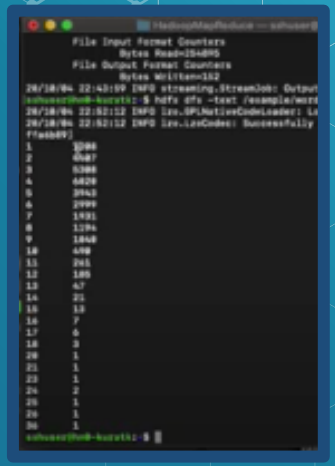
Hadoop MapReduce

With MS Azure HDInsight

MapReduce Word Count

- Create Azure HDInsight multi-node cluster
- Run Hadoop MapReduce on this cluster
- Case study: Romeo and Juliet word count



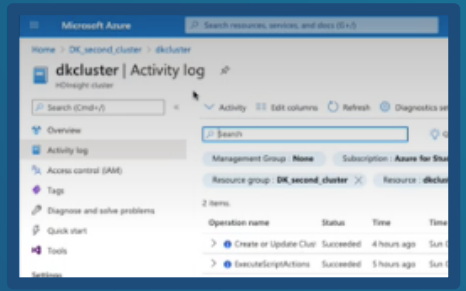


Apache Spark

Employing a Docker stack

MapReduce with Spark

- Execute a Docker stack with Spark and PySpark
- Run Hadoop MapReduce on this cluster
- Case study: Romeo and Juliet word count

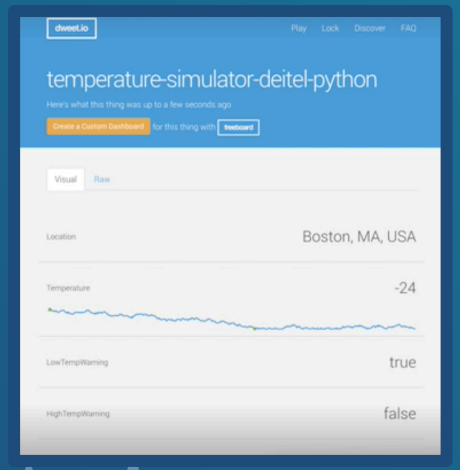


IoT and Dashboards

Visualizing live data streams

Live Data Visualization

- Visualize simulated PubNub data with Freeboard.io
- Simulate IoT thermostat using JSON "dweets"

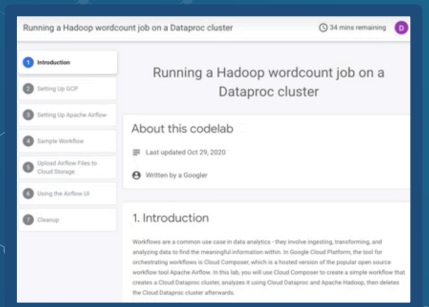


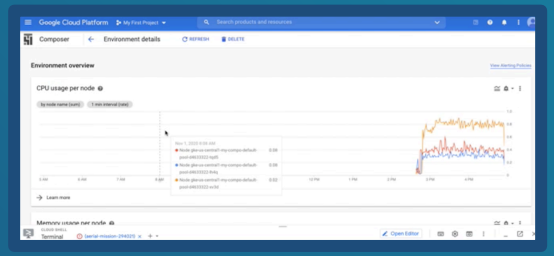
Hadoop with Google

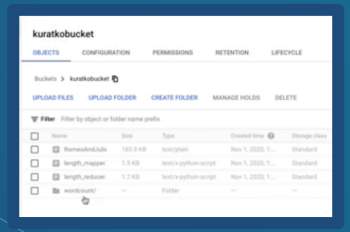
Replicating a MapReduce in GCP

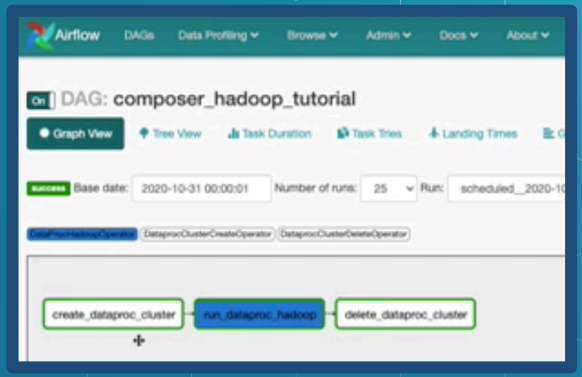
MapReduce with GCP

- GCP Dataproc cluster, composer environment, storage bucket
- Word count harnessing GCP and Apache Airflow







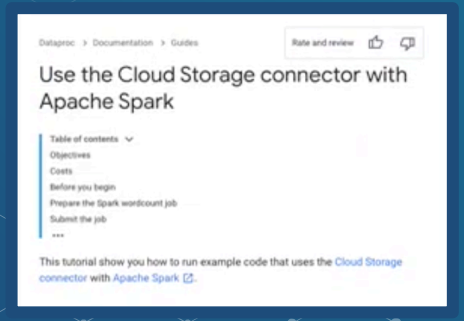


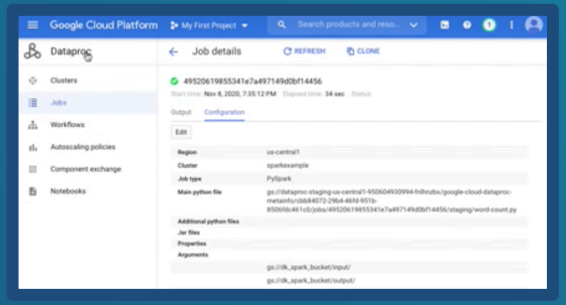
Spark and GCP

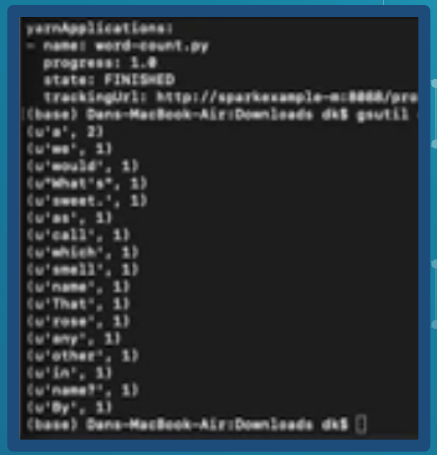
Apache Spark word count in GCP

Spark MapReduce with GCP

- Same word count example using Apache Spark







THANK YOU!

Dan Kuratko
dankuratko@gmail.com
(619)-618-5093