Jeremy Beard – MSDS 610 – 20220508 – Week 1 Assignment

What is Linux, and why is it important for data engineers and data scientists?

Linux is an operating system, as compared to the Windows 10 operating system or the OSX operating system. There are many different variations of the Linux operating system. The two (2) main varieties of Linux I believe are Red Hat Linux and Ubuntu Linux. I personally use Red Hat Linux as it is the equivalent of what my company uses, Lockheed Martin.

Who are the main cloud services providers, and what sort of things do they provide?

The main cloud service providers are Amazon Web Services (AWS), Azure, and Google Cloud. Each provides a cloud service that can assist the software developer in working with cloud servers and platforms which are utilized in the cloud.

What can we do with the cloud providers related to data engineering?

Cloud providers give users a platform which may be utilized in order to store data in the cloud, allow for platform-agnostic capabilities, and allow for high-availability applications that can be accessed from anywhere. For data engineering, this means data can move around a lot easier and doesn't need to be confined to one machine. Data Engineers can use cloud platforms to host an application or data analysis which can be accessed on multiple other mediums and is much more flexible as a result.

What is Docker, how is it useful, and why is it important to know?

Docker is a service that can take a working application and package it into a unit which has all dependencies and computing resources in it already. This makes it much easier to take an application on one platform and get it working on another platform without much effort. This is primarily what makes Docker useful (and containerization applications in general, in my case Podman). The main difference between containers and virtual machines is containers utilize the same kernel as the host, where in virtual machines the kernel is independent to the virtual machine.

Docker is also important to know because of both how prevalent it is in the software world today, and how easy it is to understand and use. Docker containers are lightweight and are quite simple to execute and work with which makes them an ideal candidate for the fast-paced software development world we live in today.

Technical Section

It is stated that if the assignment is done locally, steps should be shown for how the environment was set up. Historically, I have ran containers via Podman on a Virtual Machine using VirtualBox. The same is true with this assignment. For this assignment, I opened my Red Hat Linux Virtual Machine in VirtualBox and pulled the Hadoop container using Podman, running it interactively and mounting a local volume to the container. Podman is much better suited for Red Hat Linux machines and this assignment worked seamlessly with my environment.

First, I ran the command to pull the Hadoop container:

```
docker pull docker pull ngeorge/ubuntu-hadoop-quickstart
```

Next, I cloned the Word Count Hadoop repository using the following command:

```
git clone https://github.com/Regis-University-Data-
    Science/simple_Hadoop_MapReduce_example.git
```

I then mapped the simple_Hadoop_MapReduce_example repo to the Hadoop-quickstart container using the following options in the "docker run" (in this case, podman run) command:

```
docker run -it --rm --name hadoop1 -v
/home/jeremy/Documents/GitStuff/Regis/simple_Hadoop_MapReduce_example:/hadoop:Z
3b127e41e697
```

This command runs the Hadoop container and mounts the local Hadoop_MapReduce_Example repository to the container at the /hadoop directory.

From there, I was able to work with Hadoop and the MapReduce example data. I ran the following commands, according to the instructions in the assignment:

```
wget http://norvig.com/ngrams/shakespeare.txt
```

This 'wget' command downloaded the Shakespeare data from the Norvig.com website. Then I needed to work with Hadoop to create a couple directories and also copy the Shakespeare data to that directory.

```
hdfs dfs -mkdir /shakespeare
ls
hdfs dfs -mkdir /shakespeare/input
hdfs dfs -copyFromLocal shakespeare.txt /shakespeare/input
hdfs dfs -ls /shakespeare/input
```

After getting the directories set up and the data copied over, I was able to use the mapred command to initiate the map reducer function. I used the following command:

mapred streaming -mapper mapper.py -reducer reducer.py -input /shakespeare/input -output /shakespeare/output

I then copied the results to a file (in the mapped drive) using the following command:

```
hdfs dfs -copyToLocal /shakespeare/output/part-00000 result
```

This creates the 'result' file which contains the word count information. Upon sorting the information using the following command:

```
sort -g -r -k 2 result | head
```

The data was sorted from the greatest word count to the least. The data needs cleaned still as punctuation was appearing on the list of most frequently used words.

```
root@90bdb81508bd:/hadoop# sort -g -r -k 2 result | head
,       81827
.       36514
the     23272
I       20041
;       17274
and     16817
to      15506
of      15037
you     12361
a       12155
```

This was an interesting assignment! I think word counts can be used in numerous ways to show people how the general sentiment of news and articles is, as well as showing people what the primary topics talked about are.

References

1) freeCodeCamp.org. (2016, March 4). A Beginner-Friendly Introduction to Containers, VMs and Docker. Retrieved May 8, 2022, from https://www.freecodecamp.org/news/a-beginner-friendly-introduction-to-containers-vms-and-docker-79a9e3e119b
2) Compare AWS and Azure services to Google Cloud | Google Cloud Free Program. (2022, February 16). Google Cloud. Retrieved May 8, 2022, from https://cloud.google.com/free/docs/aws-azure-gcp-service-comparison
3) MSDS 610 Week 1 Lab Presentation Document, MSDS 610 Week 1.pdf
4) Hadoop – Apache Hadoop 3.3.2. (2021, February 21). Apache Hadoop. Retrieved May 8, 2022, from https://hadoop.apache.org/docs/stable/