Spark Project I

November,2, 2018

1. Instructions

You can start working on your first Spark project by installing and configuring the Spark virtual machine on your computer (not the lab computr!) following the instructions in section 2.

Once you have succesfully configured the Spark VM, you can download and extract the `SENG501Sparklabf6.zip` file from D2L.

Next, open your jupyter notebook using the following URL:

`http://localhost:8888`

Once the `Jupyter notebook` is open, upload the files that you downloaded (i.e., `deefoot.csv` and `SENG501_lab1.ipynb`) to the notebook.

Open the `SENG501_lab1.ipynb` notebook to start working on your first Spark Project. Make sure you change the notebook to point to the correct location of deerfoot.csv.

2. Installing and Con iguring the Apache Spark Virtual Machine

This section contains instructions on how to download and install a virtual machine that runs Apache Spark.

2.1. Install VirtualBox

First, you need to install `VirtualBox` on your computer. If you have already installed `virtualbox` on your system, you can skip this step. Download and install the `Virtualbox` using the following link:

`https://www.virtualbox.org/wiki/Downloads`

2.2. Install Vagrant

Once `Virtualbox` is installed on your system, download and install `vagrant` using the link below:

`https://www.vagrantup.com/downloads.html`

`Vagrant` is a tool for creating and configuring virtual development environments.

2.3. Download Spark Box

Once `Vagrant` is installed, you have to download a `Vagrant box` that contains Apache Spark. The box is uploaded to the cluster and you can download it using either scp or winscp . The box can be found at the following location on the cluster:

```
/home/instructor/spark_vagrant_box/spark_package.box
```

2.4. Configuring the Spark VM

Now use `Vagrant` and the `spark_package.box` to create the virtual machine. First, go to the folder where you downloaded `spark_package.box` . Open your terminal (or the command line for Windows users) and import the box into Vagrant:

```
vagrant box add my_spark_vm spark_package.box
```

Then create a folder called `spark_vm` somewhere on your computer. Off the `spark_vm` directory that you just created run the following command:

```
vagrant init my_spark_vm
```

This creates a `Vagrantfile` under the `spark_vm` directory.

2.5. Launch the Spark VM

In order to launch the VM go to the directory where you created your Vagrantfile (*e.g.*, `/home/user/Desktop/spark_vm`) and start your virtual machine by running the following command:

```
vagrant up
```

It takes approximately one minute for your virtual machine to start. Once started, open your browser and go to the following address to access the `jupyter notebook`:

```
http://localhost:8888
```

You can turn off your VM by issuing the following command:

```
vagrant halt
```

To turn on the virtual machine next time, follow the same instructions listed in this section.