Setting up a Spark Cluster

Re-submit Assignment

Due No Due Date **Points** 5 **Submitting** a file upload

AWS Setup: Spark Cluster via Elastic MapReduce

Note: This assignment will use the EC2 key pair you created in **Getting Started with AWS** assignment.

Starting a Cluster with Spark

To run a Spark program on AWS, you need to start up an cluster using the Elastic MapReduce Management (EMR) Console.

To set up and connect to a cluster, perform the following steps:

1. Go to **https://us-east-1.console.aws.amazon.com/elasticmapreduce/home** **(https://us-east-1.console.aws.amazon.com/elasticmapreduce/home)** and sign in. Make sure that the N. Virginia region is selected on the panel at the top right.
2. Click the "Create Cluster" button.
3. Click "Go to advanced options". You will have four steps of options to fill out.

Step 1: Software and Steps

In the "Software Configuration" section, set "Vendor" to be "Amazon" and select "Release" as "emr-4.2.0". Uncheck "Hive 1.0.0" and "Pig 0.14.0". Check "Zeppelin-Sandbox 0.5.5" and "Spark 1.5.2".
In the software settings section, copy and paste the following configuration which allows Spark to use all of the memory of the cluster:

```
[{
  "Classification": "spark",
  "Properties": {
    "maximizeResourceAllocation": "true"
  }
}]
```

then click "Next" at the bottom of the page.

Step 2: Hardware

In the "Hardware Configuration" section, don't change the default "Network" and "EC2 Subnet", and don't create a VPC. Change the count of core instances to be 5 for the homework, but for playing around and experimenting, change the count to be 1. If you find queries are too slow, you can resize the cluster and increase the count of core instances. Keep the count of task instances as 0. You can change the instance type of the instances if you want, but the larger the instance, the more expensive. To start out, keep it as m3.xlarge. For more information, go to **https://aws.amazon.com/ec2/instance-types/ (https://aws.amazon.com/ec2/instance-types/)** and **https://aws.amazon.com/ec2/pricing/ (https://aws.amazon.com/ec2/pricing/)** . Lastly, you do not need to check the "Request spot" option, but if you want to experiment with bidding for a machine (like in an action) rather than getting the set price, see **https://aws.amazon.com/ec2/spot/ (https://aws.amazon.com/ec2/spot/)** .

Step 3: General Options

In the "General Options" section, in the "Cluster name" field, type a name such as "Spark Cluster" or "Spark Homework 8". Uncheck "Logging" and "Termination protection" and keep the rest unchanged. Click "Next" at the bottom of the page.

Step 4: Security Options

In the "Security Options" section, select the EC2 key pair you created above. **IMPORTANT:** *make sure to select a key pair, otherwise you won't be able to ssh to the cluster.* Leave the rest unchanged. Then click "Create cluster" at the bottom of the page.

1. Go back to the cluster list and should see the cluster you just created. It may take some minutes for the cluster to launch (up to 45 min). If your cluster fails or takes an extraordinarily long time, Amazon may be near capacity. Try again later. If it still doesn't work, contact the TA.
2. On the cluster details page for your newly created cluster, make note of the Master Public DNS, listed on the top of the page. We will all this <master.public-dns-name.amazonaws.com>.

Prepare to Use Amazon's Public Dataset

For the homework, you are going to connect to one of Amazon's public datasets at **https://aws.amazon.com/public-data-sets/ (https://aws.amazon.com/datasets/)** . The following steps will work with any dataset stored on Amazon **EBS (https://aws.amazon.com/ebs/)** .

- For the in-class demo or to test Spark: use the **Freebase Simple Topic Dump (https://aws.amazon.com/datasets/freebase-simple-topic-dump/)**
- **(https://aws.amazon.com/datasets/freebase-simple-topic-dump/)** For the assignment: use the **Freebase Quad Dump (https://aws.amazon.com/datasets/freebase-quad-dump/)** for the homework.

To prepare the data, perform the following steps:

1. Find the snapshot id of the dataset you want to use. This is listed on the public dataset's page. We will refer to it as <snapshotID>. For the Freebase Quad Dump, the id is snap-b2ca9bdc.
2. Go to **Amazon EC2 Console (https://us-west-2.console.aws.amazon.com/ec2/v2/home)** , and click on "Instances" under "Instances". You should see the cluster instances you just created in the prior step.
3. Find and take note of the instance ID and availability zone of the master node. For example, the ID of i-b0ead669 and availability zone of us-east-1b. To do this, find the instance that has the Public DNS matching <master.public-dns-name.amazonaws.com>. You will need these in the next steps. You may simplify finding your master instance by setting its name. Click the pencil icon of the name field and enter a name (e.g. 'Spark Master').
4. Click on "Volume" under "Elastic Block Storage" on the left.

5. Click on "Create Volume"
6. Keep the volume type unchanged and make the size large enough to fit the data (100 GiB should be fine). Select the availability zone to be the same one of the master node you found above. Under snapshot ID, enter <snapshotID>. Select the data from the drop down that appears. Click "Create Volume".
7. Once loaded, check the volume you just created, and under "Actions", select "Attach Volume". In the instance field, select the instance ID of the master node. Keep the device field as it is. Note, a warning should come up about newer Linux kernels being renamed. That is okay. Click "Attach".

Running First Query on Spark

You are now ready to connect to the data you just attached using Spark and run some queries.

Perform the following:

1. Go back to the **EMR Portal** **(https://us-west-2.console.aws.amazon.com/elasticmapreduce/home)** and select your newly created cluster.
2. On the top, under the cluster name, click on the "Enable Web Connection" link and follow the instructions. You will set up FoxyProxy and establish a SSH tunnel. Make sure to set up FoxyProxy **BEFORE** establishing an SSH tunnel, even though the instructions are in the other order on AWS. For the installation of FoxyProxy, choose FoxyProxy Standard With the SSH tunnel, it looks as if the command is hanging, but that is okay. You are supposed to leave the terminal window open and running.
3. Back on the cluster details page, next to the Master Public DNS name, click on "SSH" and follow those instructions in a new terminal window (keep the previous one running). This will connect you to the master node.
4. Once connected, run

   ```
   cd /
   ```

   This will take you to the root directory where we will mount the data to the device. For more information, go to **http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-using-volumes.html** **(http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-using-volumes.html)** .
5. To make sure things are correct, run

   ```
   lsblk
   ```

   This should so you the disk devices available. You should see xvdf and xvdf1 where xvdf1 has the same size as the public data we are using.
6. Run

   ```
   sudo mkdir /data
   ```

   This will make a folder, called /data, where we will mount the public dataset.
7. Run

   ```
   sudo mount /dev/xvdf1/ /data/
   ```

Then navigate to /data/. You should find a file of the public dataset, ready for use.

8. Run

   ```
   sudo chmod 664 /data/freebase-datadump-quadruples.tsv
   ```

   to make the file readable for the next step.

9. Then run

   ```
   hadoop fs -mkdir /data/
   ```

   to make a folder on HDFS for Spark to use. This may take a few seconds.

10. To put the public dataset file in HDFS, run

    ```
    hadoop fs -put /data/freebase-datadump-quadruples.tsv /data/spark_data.tsv
    ```

    (assuming you are using the Freebase Quad Dump). It may take a while (up to 90 min) for the data to be transferred and you must stay connected to the server during the transfer (see the note about screen below). You can SSH to your master node using a separate terminal window and run

    ```
    hadoop fs -ls -h /data
    ```

    to monitor the current size of the file (the end size will be around 30GB). You may use a command such as **screen (https://www.gnu.org/software/screen/manual/screen.html)** or **nohup (http://linux.die.net/man/1/nohup)** to let the transfer run even if you log off.

11. Go back to the cluster detail page, and on the top of the page, under the cluster name, click on the "Zeppelin" link. This will open up **Zeppelin (https://zeppelin.incubator.apache.org/)** .

12. Click on "Create new note", and give your new notebook a name like "Homework 8 Spark". Click on that notebook. This will open up an interpreter to let you run Scala code to load the data into Spark and run SQL to actually run queries.

13. Copy and paste the following:

    ```
    import org.apache.commons.io.IOUtils
    import java.net.URL
    import java.nio.charset.Charset

    //open file in hdfs
    val FBText = sc.textFile("/data/spark_data.tsv")
    //create the schema of the table for Freebase data
    case class RDFRow(subject: String, predicate: String, obj: String, context: String)
    //loop through each row of data, split it on a tab, and make a RDFRow object
    //to toDF makes it a DataFrame, which is equivalent to a relational table
    val fbRow = FBText.map(s => s.split("\t")).map(s =>
      RDFRow(
        if (s.length >= 1) s(0) else "",
        if (s.length >= 2) s(1) else "",
        if (s.length >= 3) s(2) else "",
        if (s.length >= 4) s(3) else "")).toDF()
    //makes a table of the data called dbFacts
    fbRow.registerTempTable("fbFacts")
    ```

14. Click "Run" on the top right of the white paragraph box. This is going to run the code on your cluster without you having to do it yourself. Pretty nice!
15. In a new paragraph, enter (without the semi-colon):

    ```
    %sql
    SELECT *
    FROM fbFacts
    LIMIT 1
    ```

    This sets the language to SQL and lets you run SQL on your fbFacts table. This sample query just returns a single row of data. You are now ready to run more queries, which you will be doing in a later assignment.

What to turn in: A screenshot of the response to the query above.

Shutting Down Your Cluster

You MUST shut down your cluster when you are done. Closing the browser will not work. Amazon charges you per instance hour, which means you could spend a lot of money if you forget to shut down your cluster.

To shut down your cluster, do the following:

1. Go back to the EMR portal by clicking **https://us-west-2.console.aws.amazon.com/elasticmapreduce/** **(https://us-west-2.console.aws.amazon.com/elasticmapreduce/)** .
2. You should see a list of your clusters. Click on the name of your running cluster.
3. At the top, click "Terminate". You may have to turn off termination protection. It may take a few minutes for everything to shut down.
4. If you are totally done with the homework, delete your Volumes. As Volumes are cheap, it's okay to keep the running while you are working on your homework. To delete them, go back to the EC2 dashboard by clicking on **https://us-west-2.console.aws.amazon.com/ec2/v2/home** **(https://us-west-2.console.aws.amazon.com/ec2/v2/home)** . Click on "Volumes" from the left hand side under "Elastic Block Storage". Delete your Volumes.