**Data Analytics with Spark and EMR**

**Introduction**

You work for an organization that has a wide variety of users. You have been tasked with running some data analytics for an upcoming marketing campaign. The end goal is to determine the most common users, grouped by their gender and age. You have all the data you need stored in an S3 bucket. In this lab, you will be in charge of running data analytics on hundreds/thousands of files containing CSV data about the users who interact with the application. To accomplish this, you will first need to create an EMR cluster and copy user data into HDFS. Next, you will run a PySpark Apache Spark script to count the number of users, grouping them by their age and gender. Finally, you will need to load the results into S3 for further analysis.

**Helpful Documentation**

* [Getting Started with Amazon EMR](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html)
* [Apache Spark on EMR](https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark.html)
* [Using s3-dist-cp on EMR](https://docs.aws.amazon.com/emr/latest/ReleaseGuide/UsingEMR_s3distcp.html)
* [Using the command-runner.jar on EMR](https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-commandrunner.html)

**Lab Resources**

* [User Data CSV Files](https://das-c01-data-analytics-specialty.s3.amazonaws.com/Labs/user-data-acg.zip)
* [PySpark Script for gender and age groupings](https://das-c01-data-analytics-specialty.s3.amazonaws.com/Labs/emr-pyspark-code.py)

**Solution**

Log into the AWS Management Console using the credentials provided for the lab. Ensure you are using the us-east-1 region.

**Create the EMR Cluster**

1. Navigate to EMR using the *Services* menu or the unified search bar.
2. Click **Create cluster**, then select **Go to advanced options** to ensure the correct software is installed on the EMR cluster.
3. In the *Software Configuration* section, ensure **Hadoop 2.10.1** and **Spark 2.4.7** are selected.
4. Leave all other default software selections, then click **Next**.
5. In the *Cluster Nodes and Instances* section, update the *Master* and *Core* instance types to m4.large:
   * Select the pencil icon to the right of the *Master* instance type.
   * In the *Instance types* popup, select **m4.large**, then click **Save**.
   * Select the pencil icon to the right of the *Core* instance type.
   * In the *Instance types* popup, select **m4.large**, then click **Save**.
6. Update the *Instance count* for the *Core* node to **1**.
7. Leave all other default settings and click **Next**.
8. In the *Cluster name* field, enter a name for the cluster (e.g. age-and-gender-analytics-cluster), then click **Next**.
9. Leave the default security options and click **Create cluster**. We do not need to create a key pair for this hands-on lab. The cluster takes 10-15 minutes to create.

**Open the HDFS Port**

1. After the EMR cluster is created, select the **Application user interfaces** tab.
2. In the *On-cluster application user interfaces* section, copy the port number from the *HDFS Name Node* URL.
3. Select the **Summary** tab, then open the *Security groups for Master* URL in a new browser tab.
4. Select the *ElasticMapReduce-master* security group ID link.
5. On the right, click **Edit inbound rules**.
6. Scroll down to the bottom of the page and click **Add rule**. A new rule line is added to the bottom of the rule list.
7. Configure the new rule:
   * Leave the *Type* as *Custom TCP*.
   * In the *Port range* field, paste the copied HDFS port number.
   * Click into the IP field and select **0.0.0.0/0**.
8. In the bottom right corner of the page, click **Save rules**.
9. Close out of the *Security Groups* tab and navigate back to EMR.

**Copy Data from S3 to HDFS Using s3-dist-cp**

1. From the EMR cluster, select the **Application user interfaces** tab.
2. In the *On-cluster application user interfaces* section, copy the *HDFS Name Node* URL and open it in a new browser tab. This provides an overview and the details of your Hadoop cluster.
3. Along the top of the page, use the *Utilities* dropdown to select **Browse the file system**. You will load data onto the HDFS cluster and monitor it through this web interface.
4. Navigate back to EMR and select the **Steps** tab. You will create a few steps that will run on the EMR cluster.
5. Click **Add step** and fill in the step details:
   * Use the *Step type* dropdown to select **Custom JAR**.
   * In the *Name* field, enter a description (e.g. Copy data and script to HDFS).
   * In the *JAR location* field, enter command-runner.jar. The JAR application is pre-installed onto the EMR cluster and allows you to run a particular set of scripts within the EMR cluster.
   * In the *Arguments* field, enter the s3-dist-cp command followed by the source (the public S3 bucket) and the destination (the HDFS cluster):

s3-dist-cp --src=s3://das-c01-data-analytics-specialty/Data\_Analytics\_With\_Spark\_and\_EMR/ --dest=hdfs:///

* + Leave the *Action on failure* field set to *Continue*.

1. After the step details are complete, click **Add**.

The step copies data from the S3 bucket provided for the lab onto the Hadoop cluster, which takes about 5 minutes. You may need to refresh the step to see an updated status.

**Run a PySpark Script Using spark-submit**

1. After the first step is created, navigate back to the HDFS tab.
2. Refresh the page to ensure your files and folders are now available on HDFS. You should see a user-data-acg folder and a pyspark-script folder.
3. View the *user-data-acg* and *pyspark-script* folders:
   * Select the **user-data-acg** folder. The folder contains hundreds of CVS files which house user data. You will run data analytics to find the most common age and gender.
   * Navigate back to the *Steps* tab and select the **pyspark-script** folder. The folder contains the PySpark script, which creates a spark session, reads in all the CSV files in the HDFS cluster, creates a data frame that groups the data by users' age and gender, shows these results to the standard output, and writes the results to a CSV file stored on the HDFS cluster.
4. Navigate back to the ERM tab and click **Add step**.
5. Fill in the step details:
   * Use the *Step type* dropdown to select **Custom JAR**.
   * In the *Name* field, enter a description (e.g. Run PySpark script).
   * In the *JAR location* field, enter command-runner.jar.
   * In the *Arguments* field, enter the spark-submit command followed by the Pyspark source code link.

spark-submit hdfs:///pyspark-script/emr-pyspark-code.py

* + Leave the *Action on failure* field set to *Continue*.

1. After the step details are complete, click **Add**. The step takes some time to run the PySpark script. You may need to refresh the step to see an updated status.
2. When the step's status is *Completed*, click **View logs** on the right.
3. Select **stdout** to view the script's standard output results. The list shows 20 results, and the top of the list shows the most common age and gender.
4. Navigate back to the HDFS tab and view the root directory by entering / into the search bar. You should now see a results folder in the directory.
5. Select the **results** folder and confirm it contains a CSV file. This file contains the full list of user data.

**Copy Data from HDFS to S3 Using s3-dist-cp**

1. Open S3 in a new browser tab, then click **Create bucket**.
2. In the *Bucket name* field, enter a unique name for the bucket (e.g. gender-age-analytics-bucket).
3. Copy the bucket name to your clipboard, then click **Create bucket**.
4. Close the S3 tab and navigate back to the EMR tab.
5. Ensure the *Steps* tab is selected, then click **Add step**.
6. Fill in the step details:
   * Use the *Step type* dropdown to select **Custom JAR**.
   * In the *Name* field, enter a description (e.g. Load results to S3).
   * In the *JAR location* field, enter command-runner.jar.
   * In the *Arguments* field, enter the s3-dist-cp command followed by the source (the *Results* folder on HDFS) and the destination (the new S3 bucket).

s3-dist-cp --src=hdfs:///results --dest=s3://<YOUR\_BUCKET\_NAME>/

* + Leave the *Action on failure* field set to *Continue*.

1. After the step details are complete, click **Add**. The step takes some time to copy the data from HDFS into the S3 bucket. You may need to refresh the step to see an updated status.
2. After the step's status is *Completed*, navigate to S3 in a new tab and select the bucket you created for the lab. You should see a CSV file with the data that was loaded from HDFS into S3.
3. Select the CSV file and use the *Actions* dropdown to select **Download**.
4. Review the file to verify that it shows all the data, aggregated by age and gender.

**Conclusion**

Congratulations — you've completed this hands-on lab!