CS643 - Programming Assignment 2 Instructions

Kendy Colon (krc9)
4/22/2022

Step 1: Parallel training on 4 ec2 Instances

This project was built with the Spark data frames API and the MLib libraries, and the application is automatically parallelized and distributed natively.

The wine prediction program is built with Spark DataFrames and MLlib. When you run it on an AWS EMR cluster, it automatically parallelizes and distributes task execution. For locating dataset files and storing trained models, the Hadoop Distributed Files system is employed.

How to create an EMR Cluster ?

1. After logging into AWS console,
2. Go to EMR Service & Create Cluster
    a. Enter cluster name
    b. Launch Mode cluster
    c. Vendor Amazon
    d. Release emr-5.3.10
    e. Select spark application – version 2.4.5
    f. Hardware Configurations
        i. select the instance type
        ii. number of instances to 4 (1 master 3 slaves)
    g. select the ec2 key pair or generate one to access the master node.
    h. click on create cluster.

Upload files to EMR Cluster Master node

1. Once Cluster goes into waiting state, copy master node dns address and open command prompt on local machine.
2. Open sftp connection to master node
    ○ sftp -i cluster-keypair.pem hadoop@ec2-3-128-26-180.us-east-2.compute.amazonaws.com
3. Upload TrainingDataSet.csv, ValidationDataset.csv and winequality-1.0.jar to master node.

SSH Master node :

ssh -i ~/cluster-keypair.pem hadoop@ec2-3-128-26-180.us-east-2.compute.amazonaws.com

Copy files to HDFS :

Now all files are on our master node we wanna move them to HDFS so that all slave nodes can also access them and we don’t have to manually copy them to all ec2 nodes.

1. Use this command to copy files from Master node to HDFS.
    a. hadoop fs -put TrainingDataset.csv /user/hadoop/TrainingDataset.csv
    b. hadoop fs -put ValidationDataset.csv /user/hadoop/ValidationDataset.csv
2. Use this command to verify if files are successfully copied to HDFS
    a. hdfs dfs -ls -t -R

Launch TrainingModel application :

Now everything is done, we want to launch an apache-spark application on the EMR cluster. Execute following command to run application

1. spark-submit winequality-1.0.jar

We can confirm job execution by going to the monitor tab then spark dashboard.

1. This will create a TrainingModel folder and store trained models to it.
2. Verify model is created by executing following :
    a. hdfs dfs -ls -t -R
3. Now copy This folder back to our master node using following
    a. hdfs dfs -copyToLocal TrainingModel /home/hadoop/wine

Training part is done here, so we need training files to be transferred to our local environment so that we can use them to predict on single ec2 with or without docker.

1. Make a tar.gz zip of folder so that we can download it
    a. tar czf model.tar.gz TrainingModel
2. In our sftp session execute following to download mode.tar.gz on local machine

    a. get wine/model.tar.gz

Step 2: Predict wine quality on single ec2 instance

At this stage we are interested in executing prediction code on a single ec2 instance. For that we need TestDataset.csv, wine-quality-predict.jar, model.tar.gz (from task1)

Ec2 instance Create:

- After logging into AWS console,
- Go to EC2 -> launch instance, select AMI : **ami-0053f34b22df259f2**
- Select keypair and launch it.

Ec2 instance pre configuration:

- Do ssh to ec2 instance public dns,
    - ssh -i ec2-A.pem ec2-user@ec2-54-158-81-112.compute-1.amazonaws.com
- **Install SCALA:**
    - wget http://downloads.typesafe.com/scala/2.11.6/scala-2.11.6.tgz
    - tar -xzvf scala-2.11.6.tgz
    - Update PATH environment variable:
        - vim ~/.bashrc
        - copy following lines into file and then save it

            § export SCALA_HOME=/home/ec2-user/scala-2.11.6

            § export PATH=$PATH:/home/ec2-user/scala-2.11.6/bin

        § source ~/.bashrc

- **Install SPARK:**

    ○ wget https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz

    ○ sudo tar xvf spark-2.4.5-bin-hadoop2.7.tgz -C /opt

    ○ sudo chown -R ec2-user:ec2-user /opt/spark-2.4.5-bin-hadoop2.7

    ○ sudo ln -fs spark-2.4.5-bin-hadoop2.7 /opt/spark

    - Update PATH Environment
        - vim ~/.bash_profile

        - copy following lines into file and then save it

            § export SPARK_HOME=/opt/spark

            § PATH=$PATH:$SPARK_HOME/bin

            § export PATH

        § source ~/.bash_profile

- **Upload trained model and jar files :**
    - Login to ec2 instance using sftp
    - sftp -i ec2-A.pem ec2-user@ec2-54-158-81-112.compute-1.amazonaws.com
    - put wine-quality-predict-1.0.jar
    - put TestDataset.csv
    - put model.tar.gz
- **Extract model.tar.gz :**
    - tar -xzvf model.tar.gz
- **Disable unnecessary log4j :**
    - cp $SPARK_HOME/conf/log4j.properties.template $SPARK_HOME/conf/log4j.properties
    - vi $SPARK_HOME/conf/log4j.properties

        - (on line 19 of the file, change the log level from INFO to ERROR)

        - log4j.rootCategory=ERROR, console

        - Save the file and exit the text editor.

Run wine-predict application :

- spark-submit wine-quality-predict-1.0.jar

Step 3 : Predict wine quality using docker:

For Predicting a wine quality on TestDataset.csv using docker. We need to have a full local file path of TestDataset.csv and provide it as an input argument while running a docker. So that TestDataset.csv can be copied to the local docker environment before running.

Test filename has to be **TestDataset.csv** and file has to be placed under the data/ directory of the container. To do this use -v parameter to map volumes.

Execute the following command.

1. docker pull krc993/krc9_cs643_assignment2
2. docker run -v ..data/TestDataset.csv

General use following format :

1. docker run -v [local_testfile_directory:/data] krc993/krc9_cs643_assignment2/TesDataset.csv