**Mounika Gampa BigData Project**

**Running IMDB datasets for sentiment analysis using Spark R**

1. There are multiple cloud providers available to choose from in order to analyze spark R services so a few are AWS, Azure, Cloudera, databricks, google cloud, etc. Here we choose AWS EMR using spark to run.
2. Create an EMR Cluster on your AWS Account.

**Graphical user interface, table

Description automatically generated**

1. Go to advanced options

**Graphical user interface, text, application, email

Description automatically generated**

1. In first step, Configure Software and Steps,please choose the latest emr release in Software Configuration and select the below options

**Graphical user interface

Description automatically generated with medium confidence**

1. Maximize resource allocation for your current emr cluster using below configuration

{

"Classification": "spark",

"Properties": {

"maximizeResourceAllocation": "true"

}

}

]

**Graphical user interface, text, application, email

Description automatically generated**

1. In second step, configure Hardware as below

**Graphical user interface, text, application, email

Description automatically generated**

1. Go to third step, to configure general cluster settings so here we configure bootstrap action.

Next go to amazon blog - <https://aws.amazon.com/blogs/big-data/running-sparklyr-rstudios-r-interface-to-spark-on-amazon-emr/> and edit the shell script as below:

<https://aws-bigdata-blog.s3.amazonaws.com/artifacts/aws-blog-emr-rstudio-sparklyr/rstudio_sparklyr_emr5.sh> and save it as sh file in R.

1. Go to aws and open S3, Create a bucket - [sparkbigdataproj](https://s3.console.aws.amazon.com/s3/buckets/sparkbigdataproj?region=us-east-1)

Graphical user interface, application

Description automatically generated

1. Upload the sh file into it

Graphical user interface, application, email

Description automatically generated

1. Open the object and Copy the S3 URI Path

Graphical user interface, text, application, email

Description automatically generated

1. Choose Custom Action then Configure and add the URI Path in emr cluster.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

1. Go to last step, Security unselect the cluster visible option and create a cluster

Graphical user interface, text, application, email

Description automatically generated

1. We will wait cluster to start and then configure the security groups for Master and add the edit inbound rules as below:

Text

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

1. Then, create a cluster, after sometime the cluster status changes to bootstrapping.

Graphical user interface, text, application, email

Description automatically generated

1. Now, copy the Master public DNS Graphical user interface, text, application, email

   Description automatically generated
2. Add the 8787 port to the url - <http://ec2-35-170-249-215.compute-1.amazonaws.com:8787/> and open it in a new tab- you will be able to see the login page of R studio.

Graphical user interface, application

Description automatically generated

1. Installed the required packages

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

1. Read csv files and display data(training testing validation)

Training Data Image

A close-up of a document

Description automatically generated with medium confidence

Validation Data glimpse(df\_val)

A picture containing text, newspaper

Description automatically generated

A close-up of a document

Description automatically generated with medium confidence

train %>%

select(Class) %>%

group\_by(Class) %>%

summarise(count = n()) %>%

glimpse

Text

Description automatically generated

1. Implement logistic regression

Logistic Regression op1

Table

Description automatically generated

Logical Regression op2

Table

Description automatically generated

1. Implement random forest