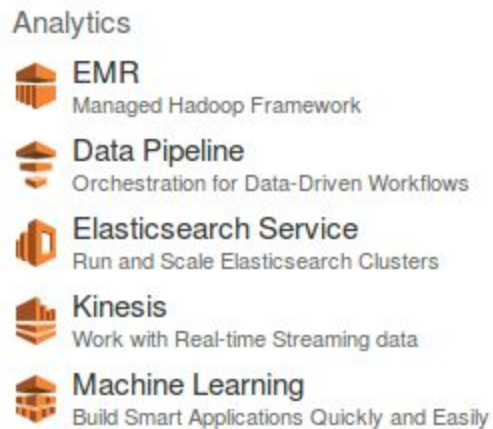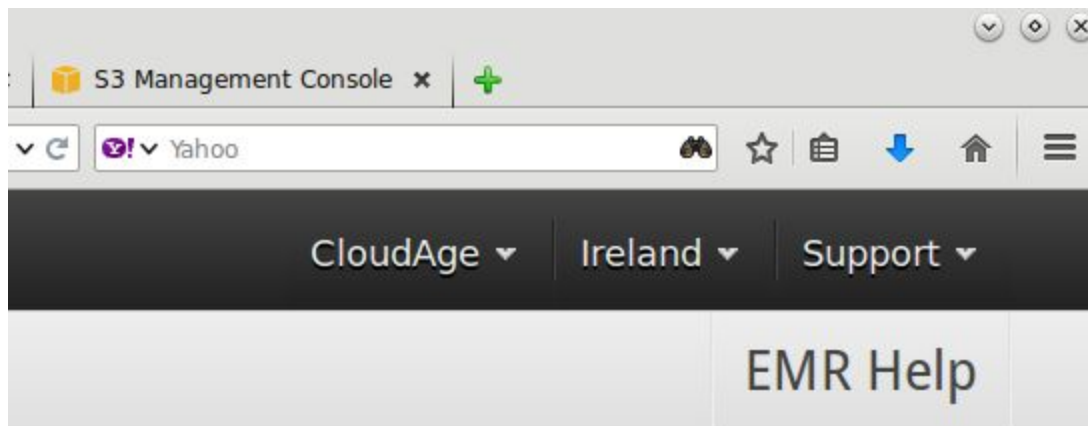# EMR (ELASTIC MAP REDUCE) Streaming Hadoop Program
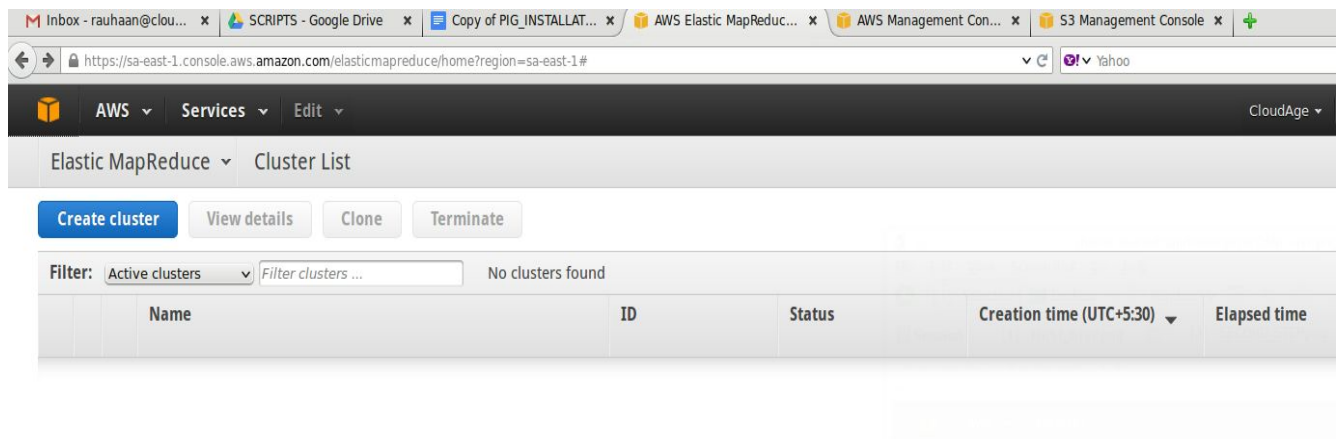
1> Sign in to AWS Console

2> Click on EMR on Analytics Section

3> Select Data Center

3> Click on Create Cluster

4> Give the Cluster Name Enable Logging and give S3 folder location for Logs

5> Click Step execution as Launch mode

6> Select Step Type as Streaming Program

Add steps

ⓘ A step is a unit of work submitted to an application running on your EMR cluster. EMR programn execute the added steps. Learn more

Name	Action on failure	JAR location

Step type Streaming program ⌄

Configure

7> Click on Configure

Paste this as Mapper : s3://elasticmapreduce/samples/wordcount/wordSplitter.py

reducer : aggregate

Give Input S3 location (Location of dataset residing on S3)

Give Output S3 location (Just give Bucket name and write name of output folder as hadoop will create output Folder. Note: if you select Folder Cluster will terminate with error "Output folder already exists")

Select Terminate as Action on failure

Add Step

Step type	Streaming program
Name*	Streaming program
Mapper*	lasticmapreduce/samples/wordcount/wordSplitter.py — S3 location of the map function or the name of the Hadoop streaming command to run.
Reducer*	aggregate — S3 location of the reduce function or the name of the Hadoop streaming command to run.
Input S3 location*	s3://datasetss/Datasets/in — s3://<bucket-name>/<folder>/
Output S3 location*	s3://baltee/output — s3://<bucket-name>/<folder>/
Arguments	
Action on failure	Terminate cluster — What to do if the step fails.

8> Select Vendor in Software configuration option

Software configuration

Vendor	● Amazon ○ MapR
Release	emr-4.1.0
Applications	Hadoop 2.6.0

9> Select Instance type and Number of instances in Hardware configuration

Hardware configuration

Instance type	m1.medium
Number of instances	3 (1 master and 2 core nodes)

10> Click on Create Instance