<<TASK 1>> create EMR cluster

Make sure no custom inbound rules exist in ElasticMapReduce-master security group

1. go to EMR
2. click create cluster
3. s/w config: release – **emr-5.33.1**; applications – **Spark**
4. h/w config: **m4.xlarge**
5. add your EC2 key-pair
6. click create cluster
7. wait for it to start up – approx. takes up to 15 mins
   [when status is 'waiting'; you are ready to use the cluster]

<<TASK 2>> connect to the zeppelin notebook

1. go to EC2 and find the master Security group
   [name would be ElasticMapReduce-master]
2. edit inbound rules to allow port 8890 for anywhere IPv4
   [8890 is where zeppelin resides]
3. go to Summary
   [in EMR – for created cluster]
4. use the MasterPublicDNS:8890 to access the zeppelin notebook from any browser
   example URL – "ec2-37-142-218-13.compute-1.amazonaws.com:8890"

<<TASK 3>> Implement the map-reduce task

1. create new notebook in zeppelin

   [default interpreter let it be as spark]

   [At this point, json log files should already exist in your bucket from Sparkify 3;

   if not, you can upload manually - for this assignment only]
2. use %spark.pyspark magic function in the beginning of each cell
3. Fetch top 10 artists and songs using map-reduce functionality
   - Map: (artistA, 1); (artistB, 1); (artistA, 1) ...
   - Reduce: (artistA, totalCount); (artistB, totalCount) ...
   - use sortBy to sort according to totalCount

These resources will come in handy:

Read JSON/log files — Spark RDD Guide

Spark Map Transformation — Spark Reduce Transformation

Spark sortBy — Python Lambda functions

Extra resources [if you want to better understand all spark RDD functionality]:

General RDD operations — Pyspark Programming