## Elastic MapReduce Tutorial

You will mapReduce jobs on the Amazon Web Services (AWS) Elastic MapReduce Clusters.

**Account**

You will need to create an AWS account. The account can be linked to your existing AWS account, or created new. You will most likely need to enter a credit card information. Note that AWS only bills you for services used.

For this class, you will get a $100 AWS credit. This is more than enough for the project. The remaining credit expires after one year. **The code to redeem your AWS credit is in your Sakai dropbox.**

**How to redeem and view AWS credits** – Sign into your account. In the upper right corner, click on the arrow next to your name and go to Billing & Cost Management. Next, in your Dashboard menu on the left, click on [Credits](https://console.aws.amazon.com/billing/home?#/credits) and once you are there, you will be able to see all the relevant info such as the remaining balance, applicable products and services, and expiration date.

**Pricing**

Amazon awarded an AWS in Education Coursework Grant Award for this project. You each can spend $100 running instances and clusters on Amazon using your account. After you exceed your quota, the account will be charged, and your credit card billed.

To avoid overages, make sure to terminate your clusters after you use them. One possibility is to set them as terminate on exit/fail. You can decide to leave your cluster running to perform several jobs on it (since AWS charges by the hour, regardless of how long you used the cluster for). **In this case make sure to terminate your clusters!**

S3 storage will also be billed on a usage basis. It is unlikely you will run more than a few cents a month with this project, but if you forget to clean your s3 buckets after your $100 credit expires in a year you may get charged then.

Information on MapReduce pricing can be found here:

<http://aws.amazon.com/elasticmapreduce/pricing/>

**Storage**

You each can create buckets on AWS S3. Since the results of your executions will be on S3, you will need to give me ([cyramel@gmail.com](mailto:cyramel@gmail.com)) and Valia (TBA) access to the bucket that holds your results.   
  
**Because Elastic MapReduce needs to access jars from a public repository, you will have to create a folder for your jar files under on of your buckets. You will store your jar files, to run on Elastic MapReduce there. Please note that the folder is PUBLIC.**

**Getting started**

Connect **the AWS console. For this project, you will need to access AWS S3 and Elastic MapReduce.**

Follow the Elastic Mapreduce tutorial at <http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-get-started-count-words.html> to launch a test WordCount MapReduce job.

* Send the output and the log to the bucket that you will use you’re your assignment results. Make sure to give us permission to access that bucket AND to give us the name of the bucket in your submission.
* The number of instances on your account are limited (you can play with the settings).

For the assignment, you will have to create jar files with the MapReduce code described in the assignment. You can either launch clusters using the console, or using the command line interface (see local setup below).

Details on how to create MapReduce jar files can be found on <http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html>

**Local Setup**

This will allow you to use the command line interface, and not the console, to run your MapReduce jobs.

Install the AWS Command Line interface

<http://aws.amazon.com/cli/>

This will allow you to perform AWS operations from your command line.

You will need to access the AWS console to create an Access Key and a secret Key.

$ aws configure

AWS Access Key ID [None]: <YOUR AWS ACCESS KEY>

AWS Secret Access Key [None]: <YOUR AWS SECRET KEY>

Default region name [None]: us-east-1

Default output format [None]: json

Install the Elastic MapReduce (EMR) CLI. This will provide you with a build environment to compile and test your java sources**. The clusters you start are on AWS so do not forget to terminate them.**

http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-cli-install.html

In the credentials.json file

{

"access\_id": "**Your AWS Access Key ID**",

"private\_key": "**Your AWS Secret Access Key**",

"keypair": "**Your key pair name**",

"key-pair-file": "**The path and name of your PEM file**",

"log\_uri": "**A path to a bucket you own on Amazon S3, such as, s3n://mylog-uri/**",

"region": "**The region of your cluster, either us-east-1, us-west-2, us-west-1, eu-west-1, ap-northeast-1, ap-southeast-1, ap-southeast-2, or sa-east-1**"

}

The AWS access key and secret key are the same as before.

The keypair is the <your\_aws\_user\_id>.pem file we provided you.

The key pair name is <your\_aws\_user\_id>

The log\_uri is your s3n://437fall2013/home/<your\_aws\_user\_id>/ directory

The region is us-east-1

Once you have installed the EMR CLI, you can create a local interactive cluster on which you can test your MapReduce jobs:

<http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-build-binaries.html>