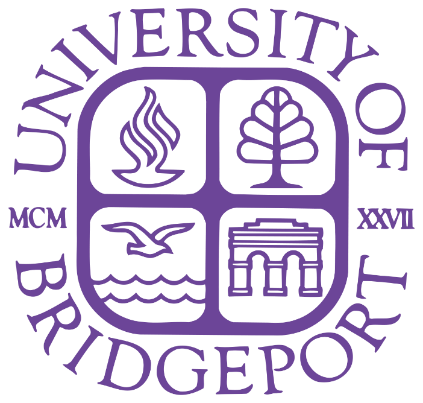
PROJECT REPORT

**Analyzing Big Data Using Amazon Elastic MapReduce**



**Srivatsan Ananthakrishnan (1021022)**

TABLE OF CONTENTS

[**Abstract** 3](#_Toc512798046)

[**Introduction** 4](#_Toc512798047)

[**Hadoop MapReduce 4**](#_Toc512798050)

[**Benifits of using EMR 5**](#_Toc512798051)

[**EMR Architecture 5**](#_Toc512798052)

[**Implementation 6**](#_Toc512798053)

[**Code for Mapper 13**](#_Toc512798054)

[**Code for Reducer 14**](#_Toc512798055)

[**Result 15**](#_Toc512798056)

[**Resetting the Environment 15**](#_Toc512798057)

[**Conclusion 15**](#_Toc512798058)

ABSTRACT

The topic of our project is Analyzing Big Data with Amazon EMR. We are using AWS services for Analyzing a massive dataset and coming up with solutions for various problems. The primarily used Amazon Web Service in this project is Amazon Elastic MapReduce, which is a service to create a cluster to work with the dataset. S3 is used for uploading our dataset, and in the cluster, the input path and the output path of the code will be given which triggers the EMR service to pick the input and code from S3 and produces the desired output to the S3 path specified in the cluster description.

The input path will hold the dataset and the code for mapper/reducer or hive whereas the Output path will hold the result of the solution. The number of EC2 instances required to perform EMR is based on the size of the data in which we are analyzing. The master and the core nodes will be allocated based on the instances.

The dataset which is used to perform this analysis is the Crime data set of LA city from 2010-present which is a public dataset took from government website. Analysis of various problems has been implemented with this dataset using the EMR service and python is used as a streaming language for implementing mapper and reducer.

INTRODUCTION

[Apache Hadoop](https://aws.amazon.com/elasticmapreduce/details/hadoop/) is an open-source Java software framework that supports massive data processing across a cluster of instances. It can run on a single instance, or thousands of instances. Hadoop uses a programming model called MapReduce to distribute processing across multiple instances. It also implements a distributed file system called HDFS that stores data across multiple instances. Hadoop monitors the health of instances in the cluster, and can recover from the failure of one or more nodes. In this way, Hadoop provides increased processing and storage capacity, as well as high availability.

Amazon Elastic MapReduce (EMR) is an Amazon Web Services tool for big data processing and analysis. Amazon Elastic MapReduce offers the expandable low-configuration service as an easier alternative to running in-house Cluster Computing. Amazon EMR is a managed service that makes it fast, easy, and cost-effective to run Apache Hadoop and Spark to process vast amounts of data.

Amazon EMR also supports powerful and proven Hadoop tools such as Presto, Hive, Pig, HBase, and more. In this project, you will deploy a fully functional Hadoop cluster, ready to analyze log data in just a few minutes. You will start by launching an Amazon EMR cluster and then use a Streaming Program script to process sample log data stored in an Amazon S3 bucket. Python, is a scripting language for data warehousing and analysis. We use a Mapper code and a Reducer code for taking the input from the file, which then gives us the desired output. You can then use a similar setup to analyze your own log files.

The dataset is taken from the following US Government Website which has the data of crimes that taken place in Los Angeles since 2010- Present.

<https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq>

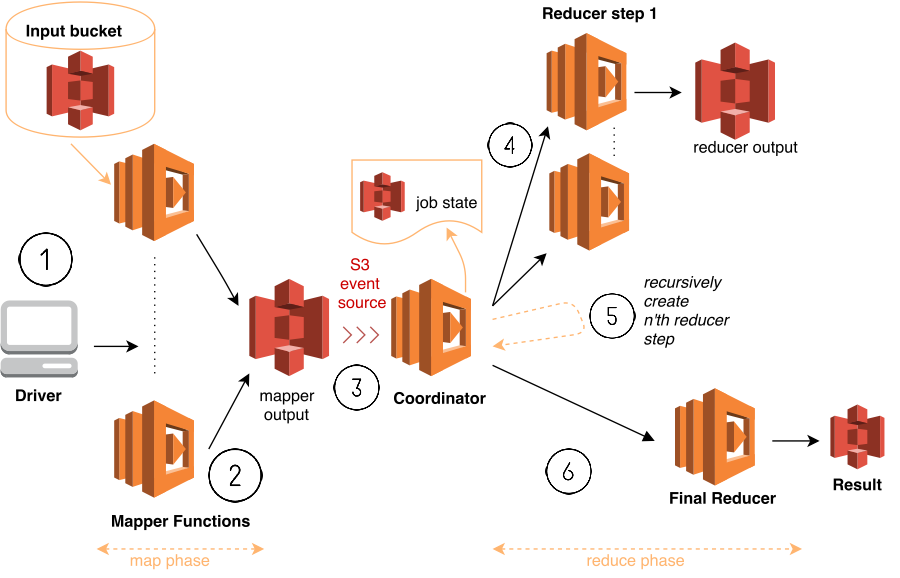
HADOOP MAPREDUCE

Hadoop MapReduce is an open-source programming model for distributed computing. It simplifies the process of writing parallel distributed applications by handling all of the logic, while you provide the Map and Reduce functions. The Map function maps data to sets of key-value pairs called intermediate results. The Reduce function combines the intermediate results, applies additional algorithms, and produces the final output. There are multiple frameworks available for MapReduce, such as Hive, which automatically generates Map and Reduce programs.

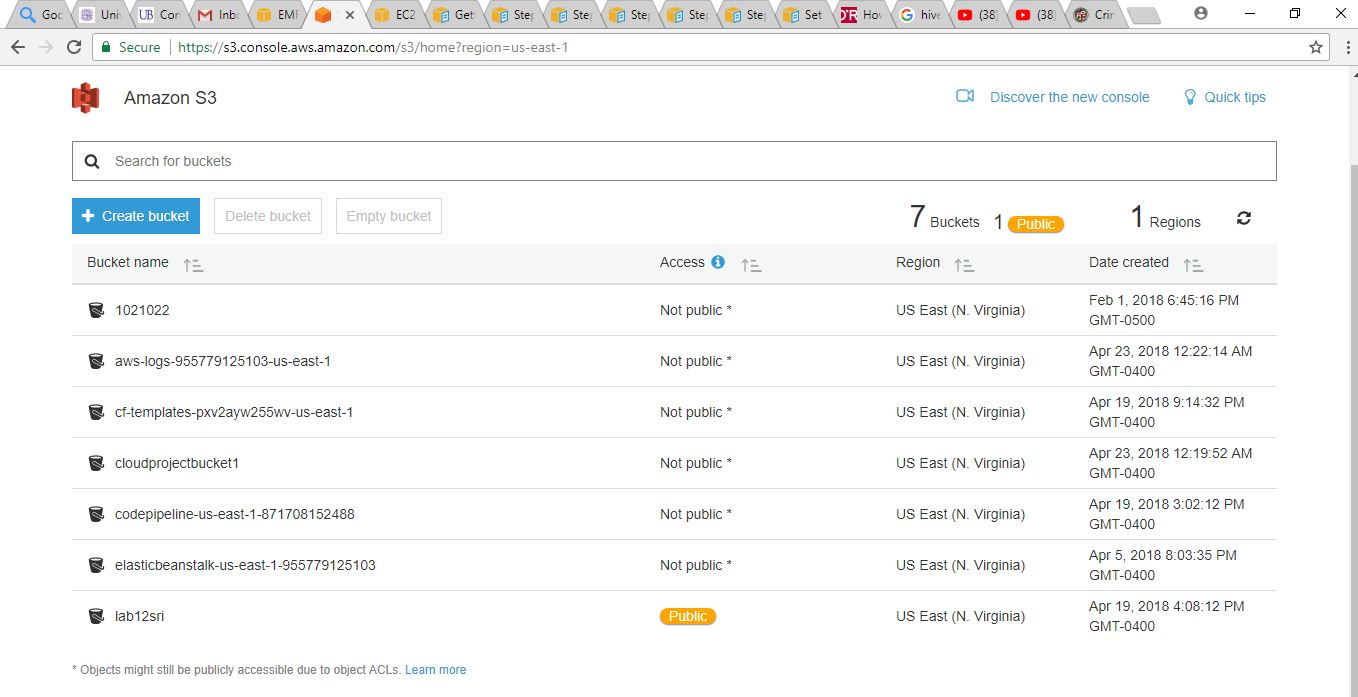
BENEFITS OF USING EMR

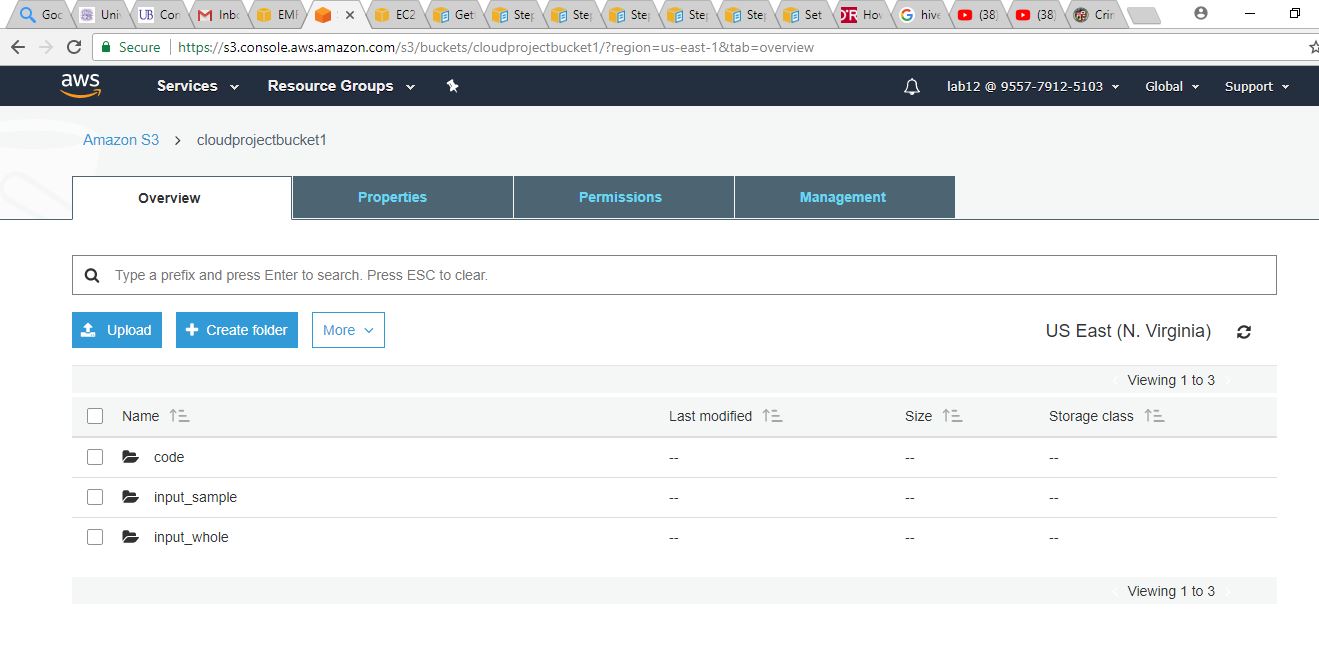
* Cost Savings
* AWS Integration
* Deployment
* Scalability and Flexibility
* Reliability
* Security
* Monitoring
* Management Interfaces

EMR ARCHITECTURE

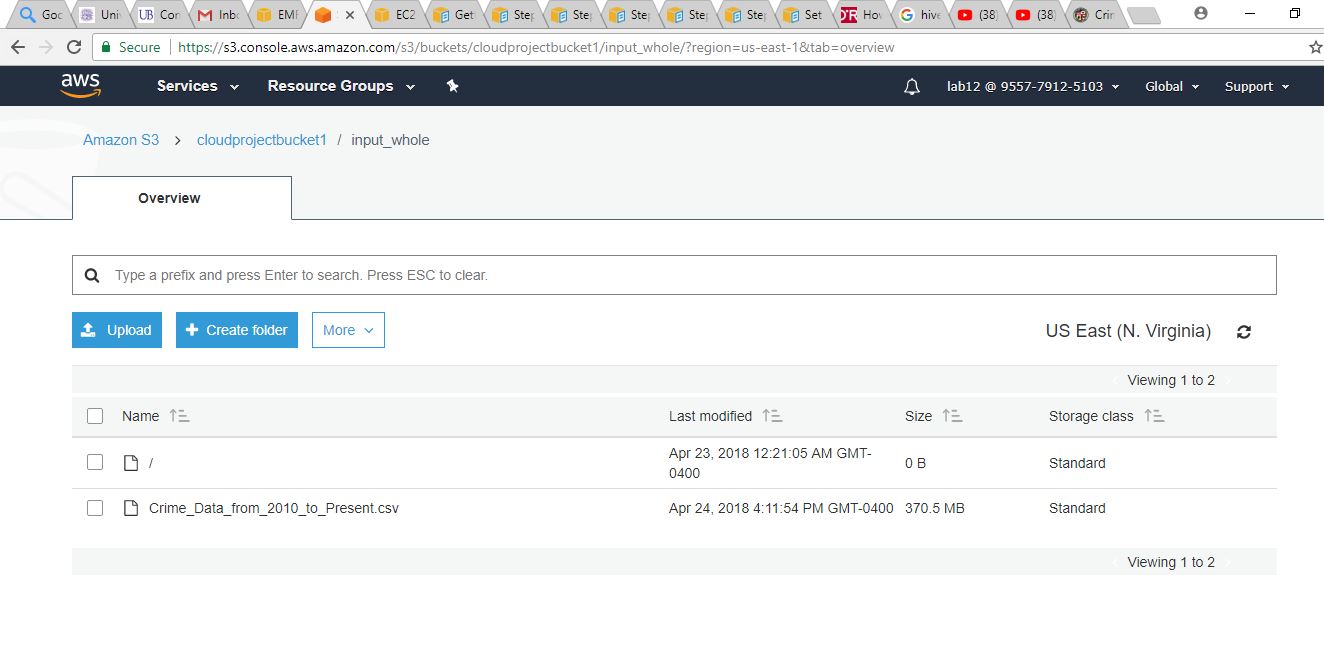


IMPLEMENTATION

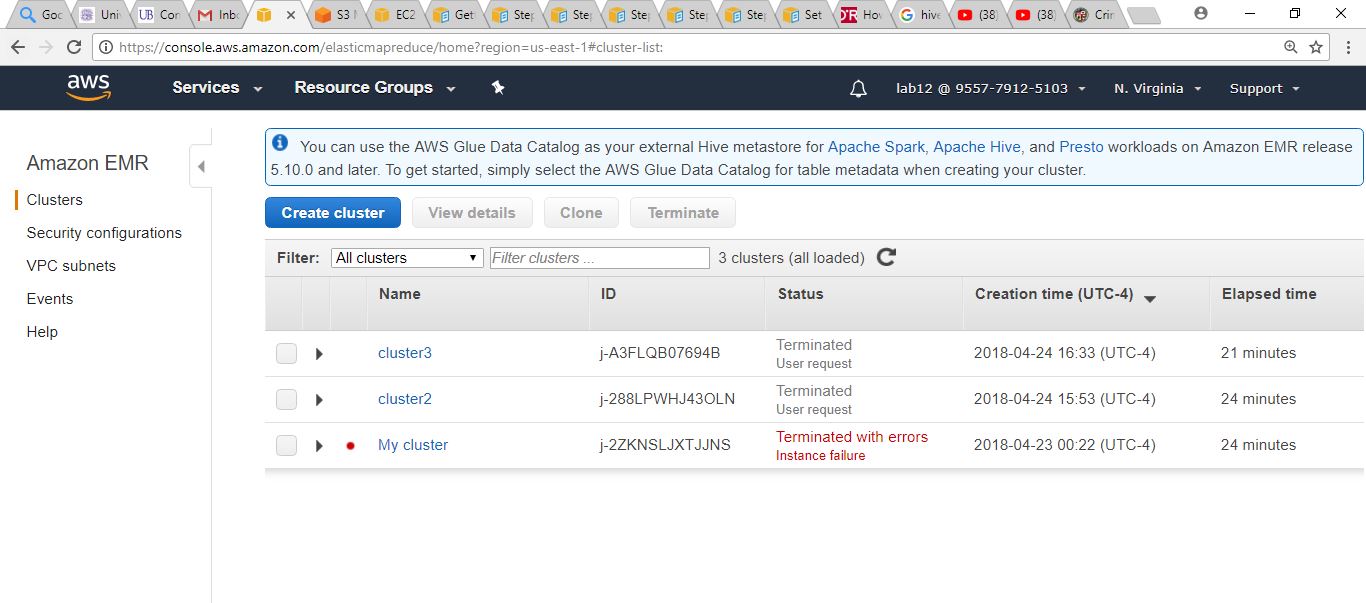
* **Create S3 Bucket and EC2 Key Pair.**
* **Save the dataset in the S3 Bucket and also save the code for mapper and reducer**



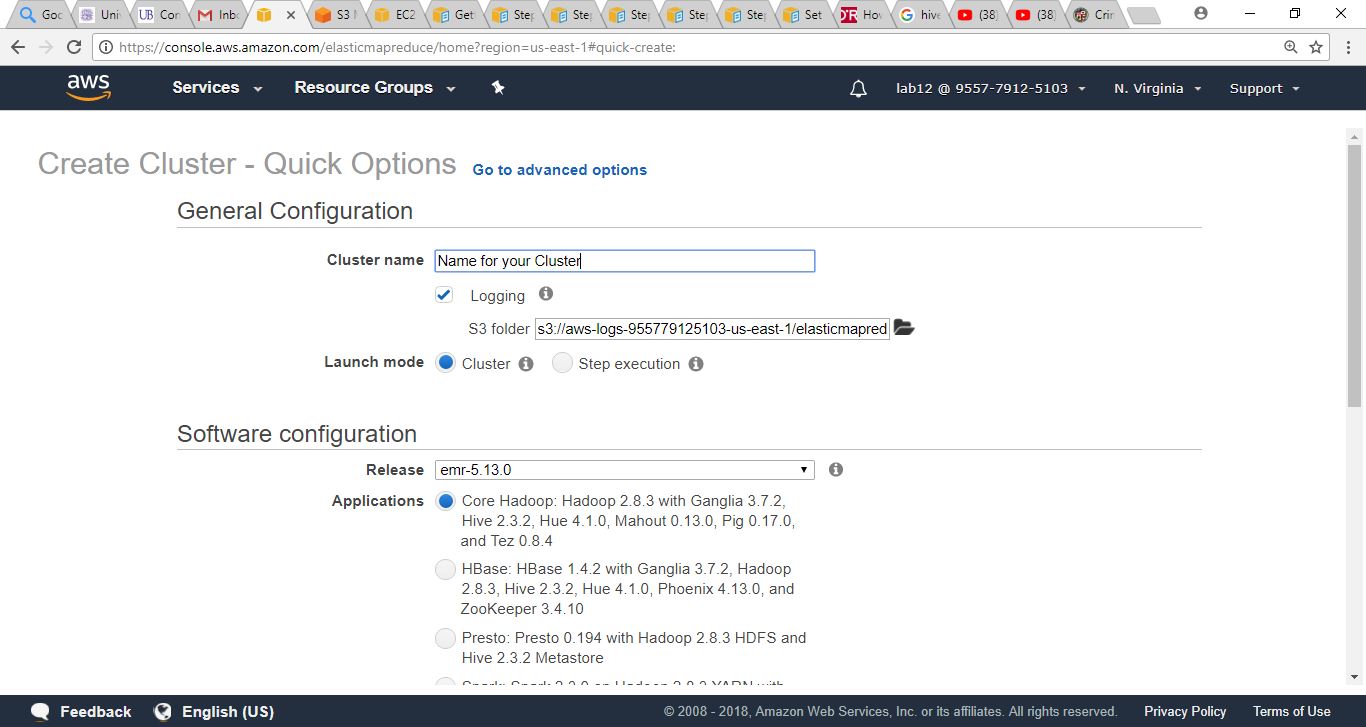
* **The dataset is saved in the S3 bucket.**

****

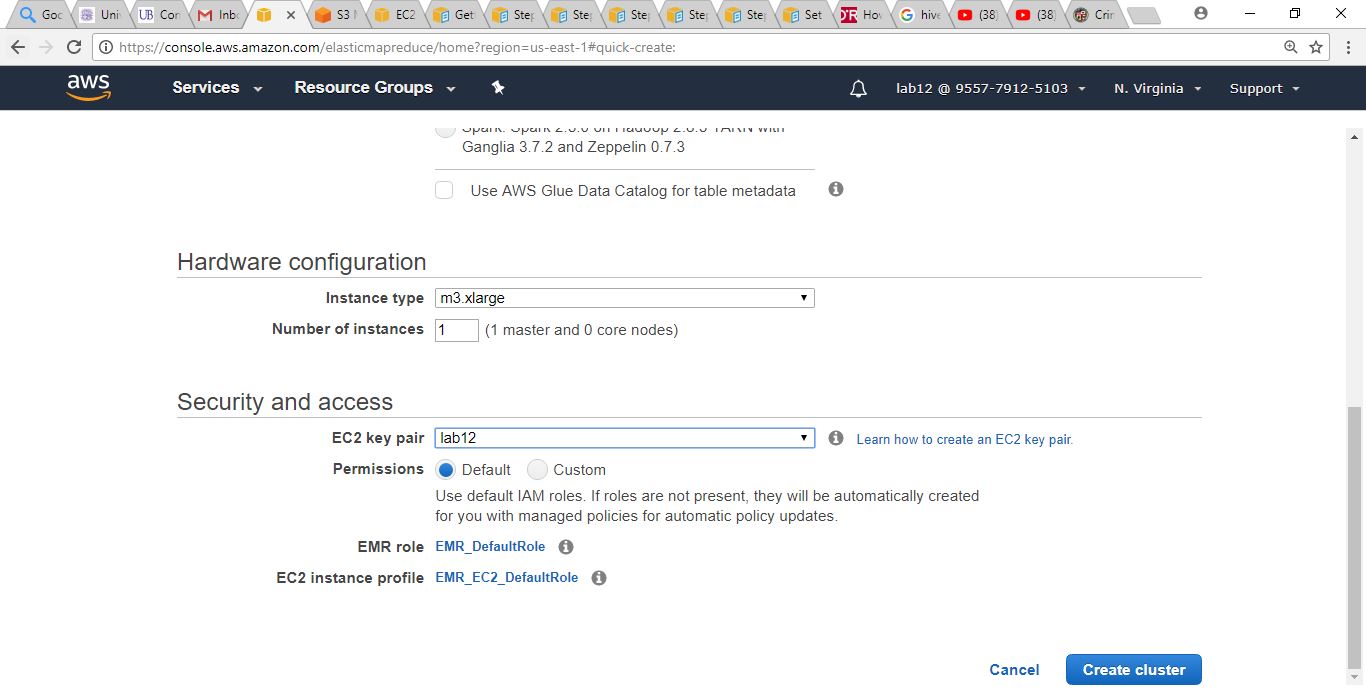
* **Creating Hadoop Cluster using Elastic MapReduce.**

****

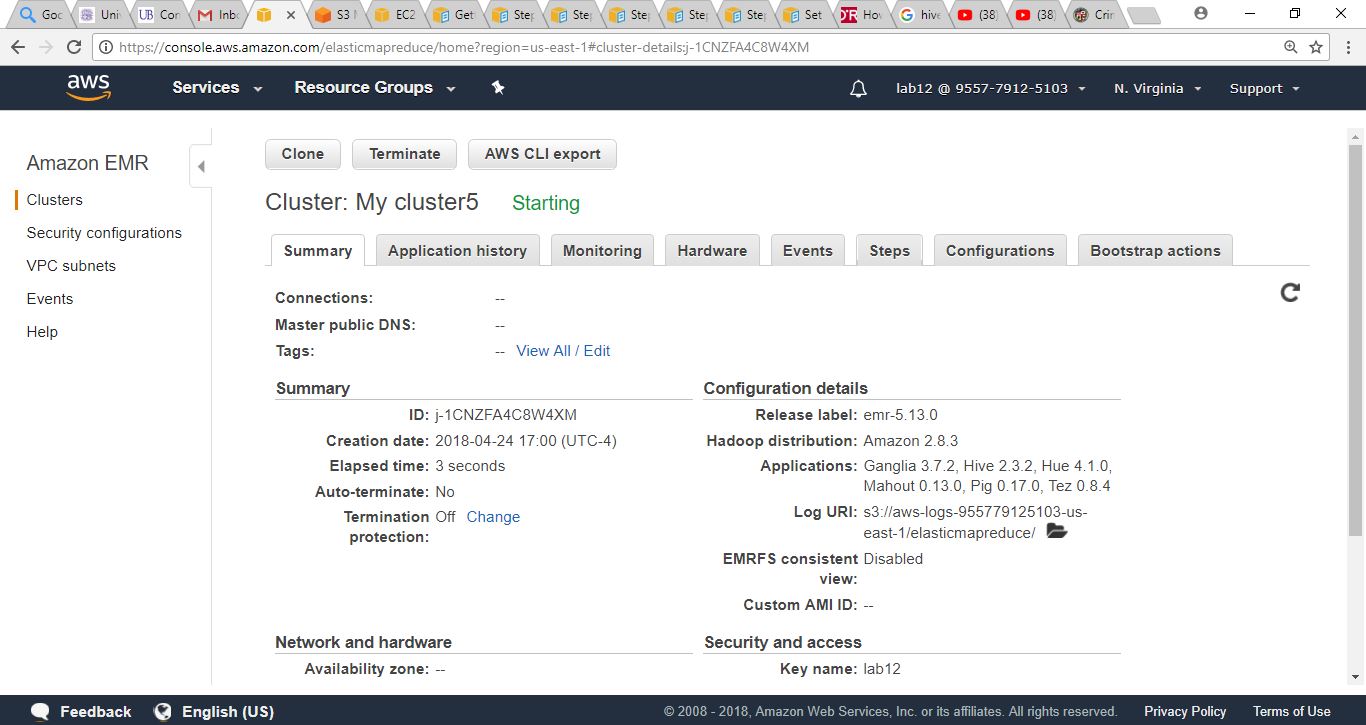
* **Configuration of the Cluster should be done. Software Configuration should be chosen as Core Hadoop as we use Streaming Program.**



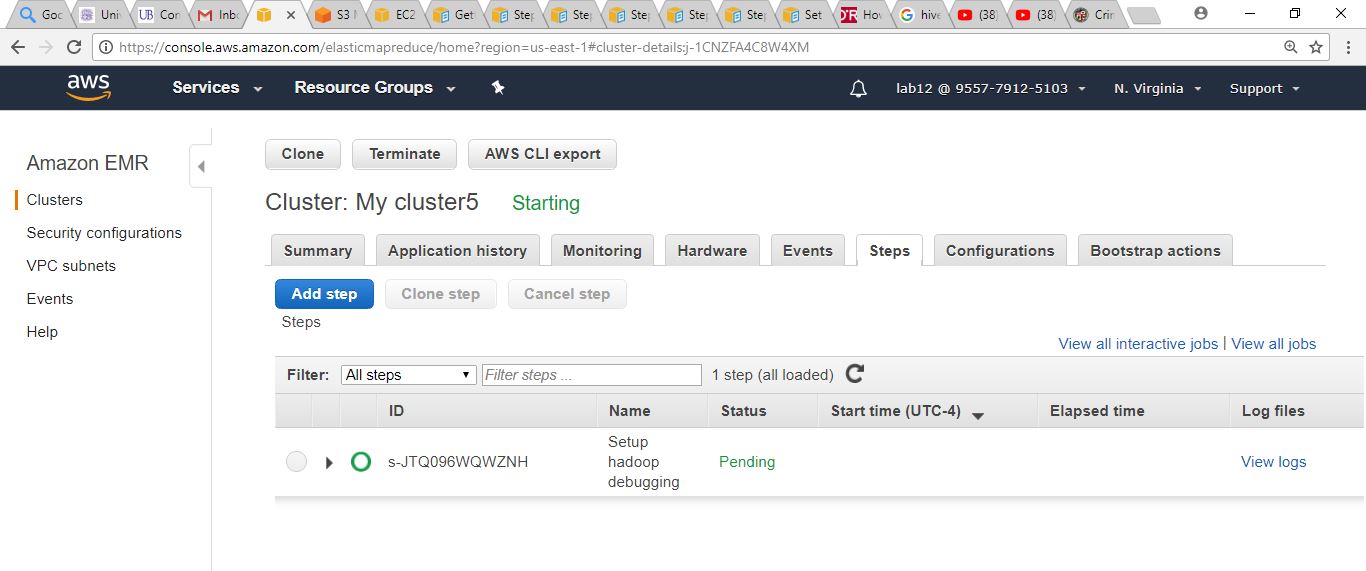
* **Select m3.xlarge (default). The number of instances launch can be selected by choice.**



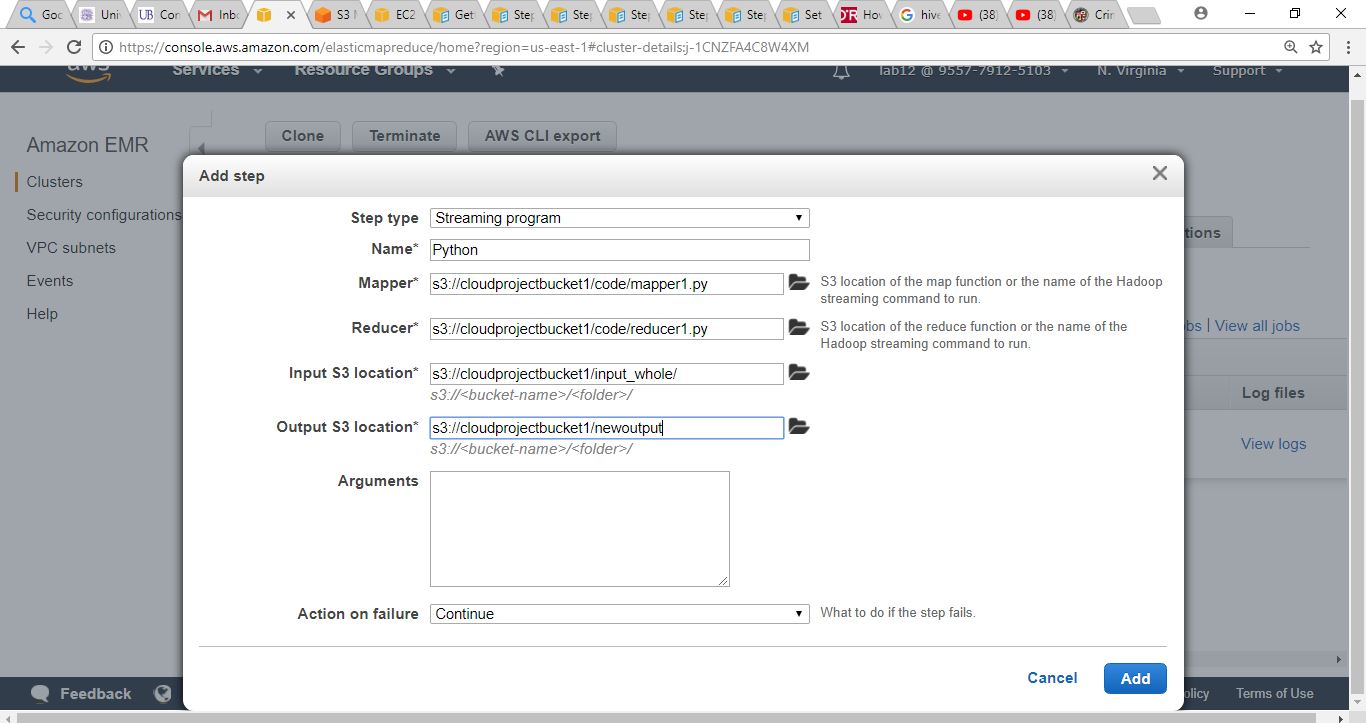
* **The cluster has been initiated.**



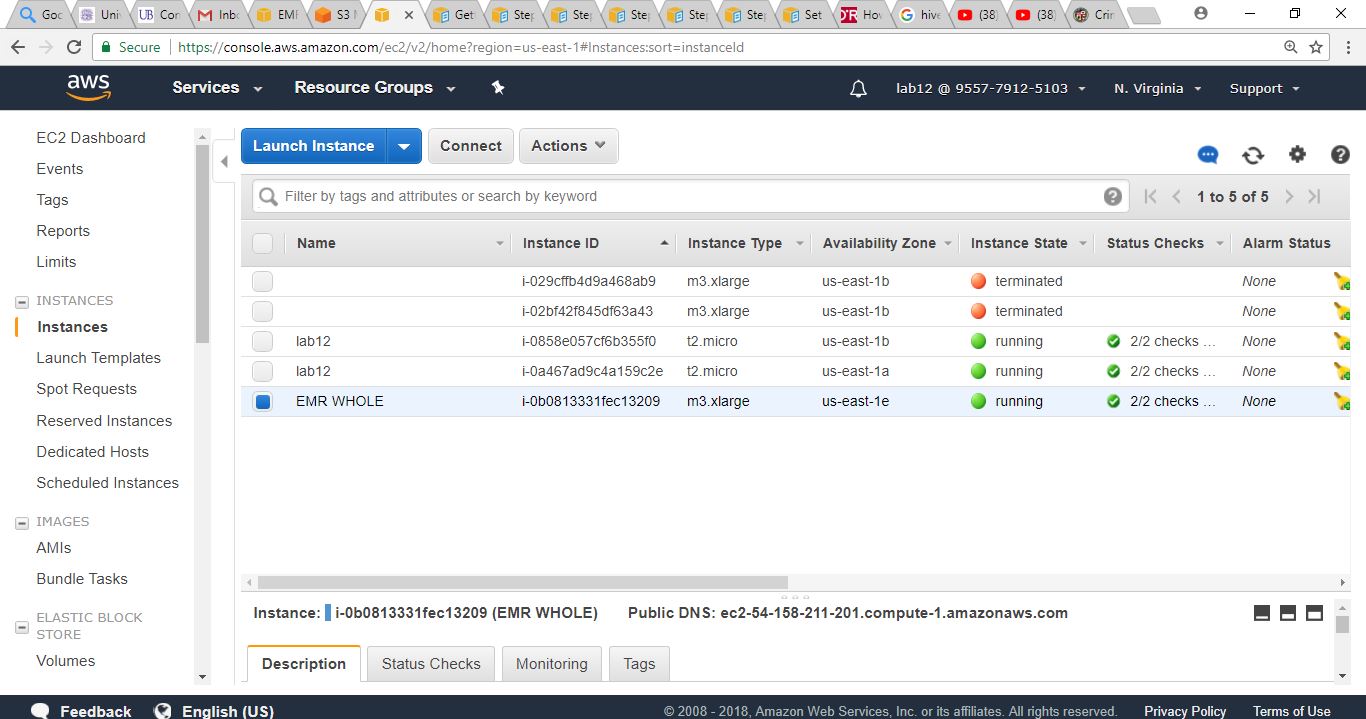
* **The cluster has been started. In steps the Hadoop Environment has started initiating.**



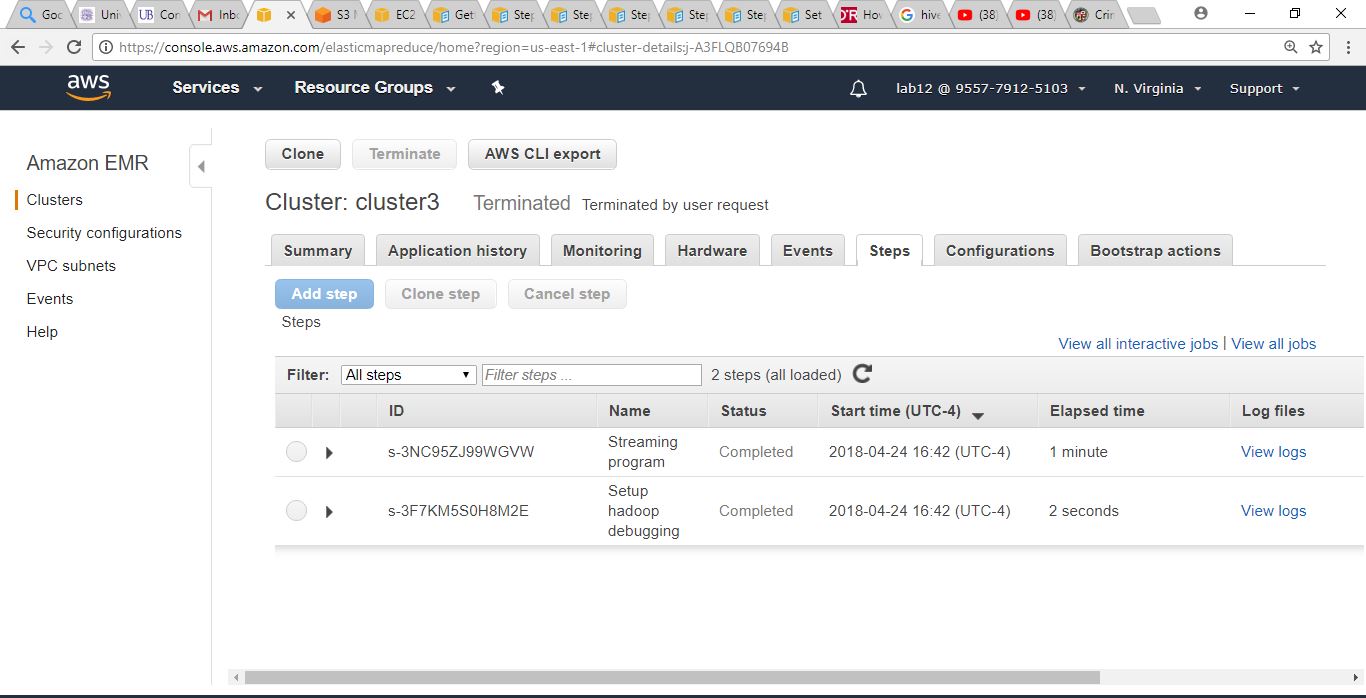
* **Another step is added, we will choose step type as Streaming Program. We give the path of Mapper code and Reducer code that is implemented through this step.**



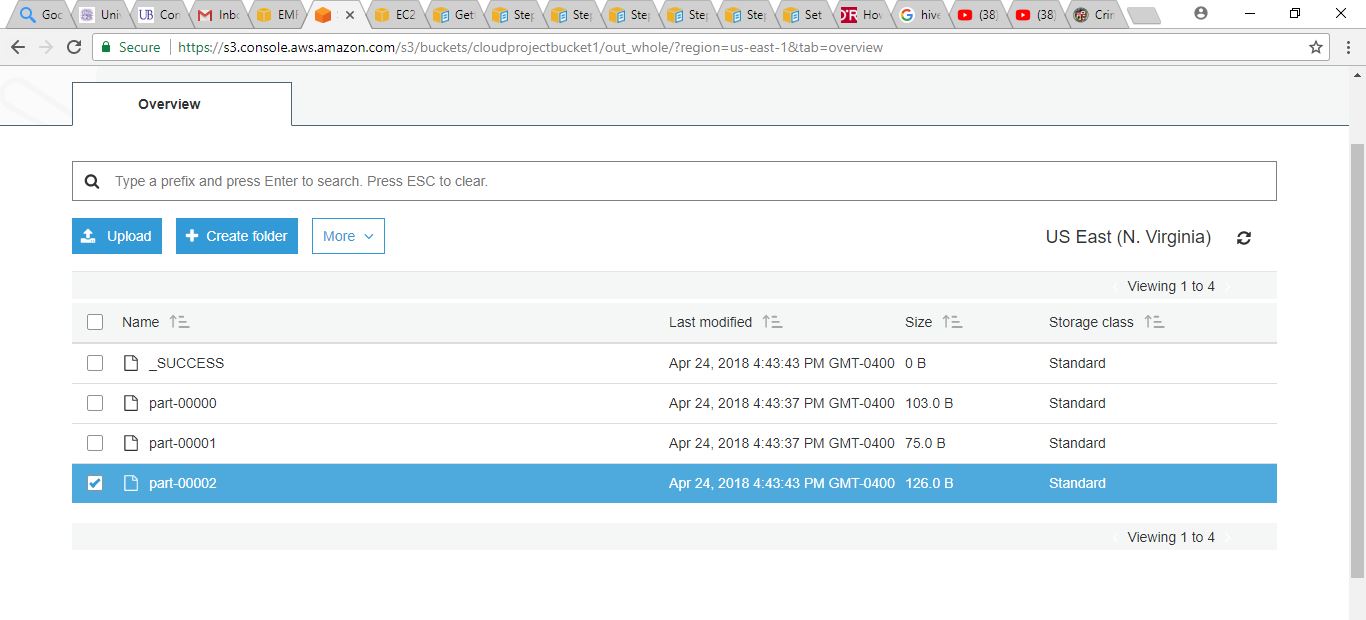
* **The instance has been launched.**



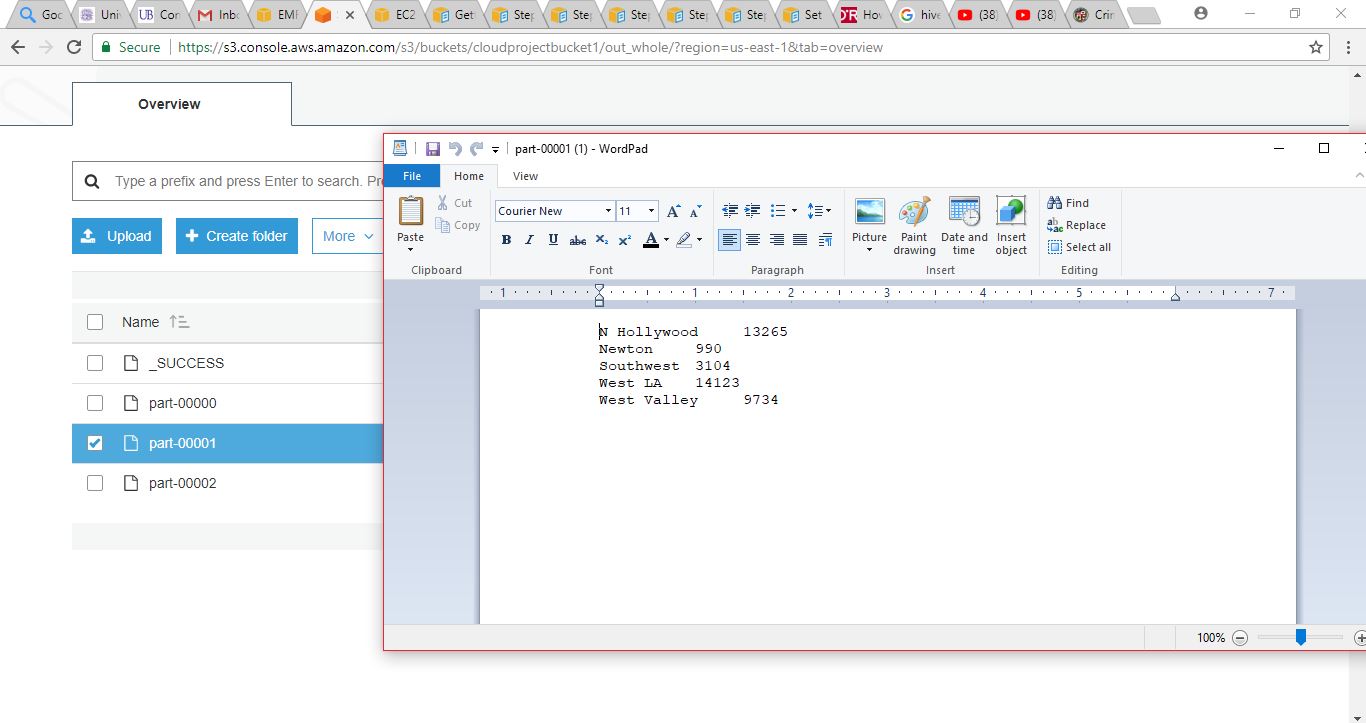
* **The Hadoop environment has been created along with the Streaming Program Step.**



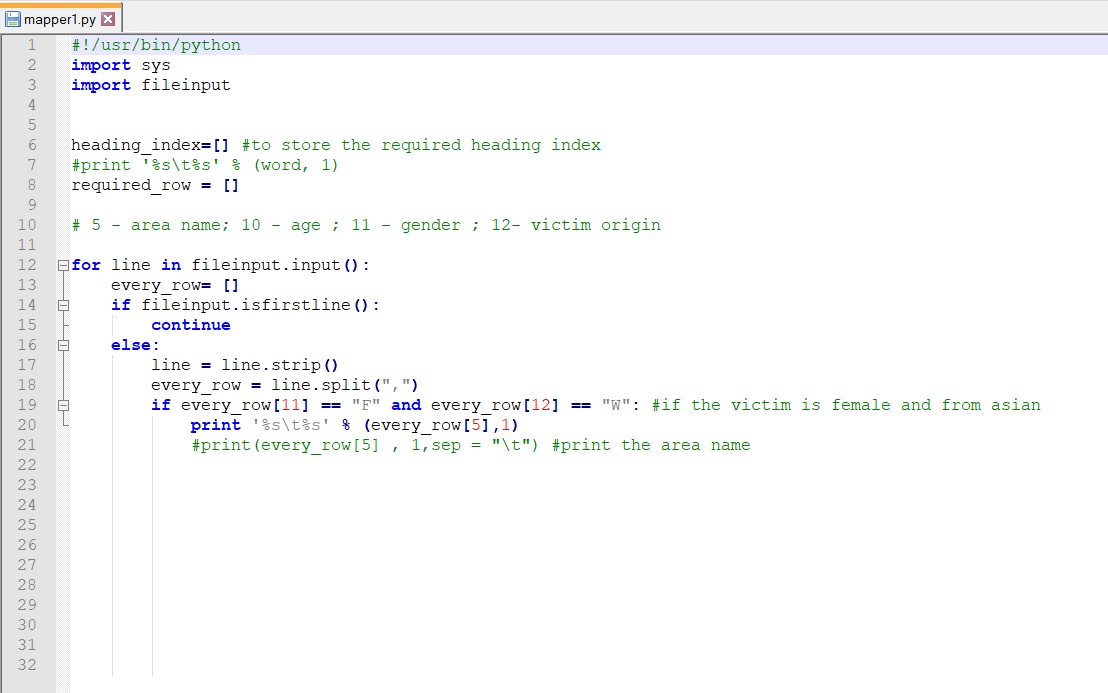
* **The output can be seen in the S3 bucket, under the folder which can be named dynamically.**



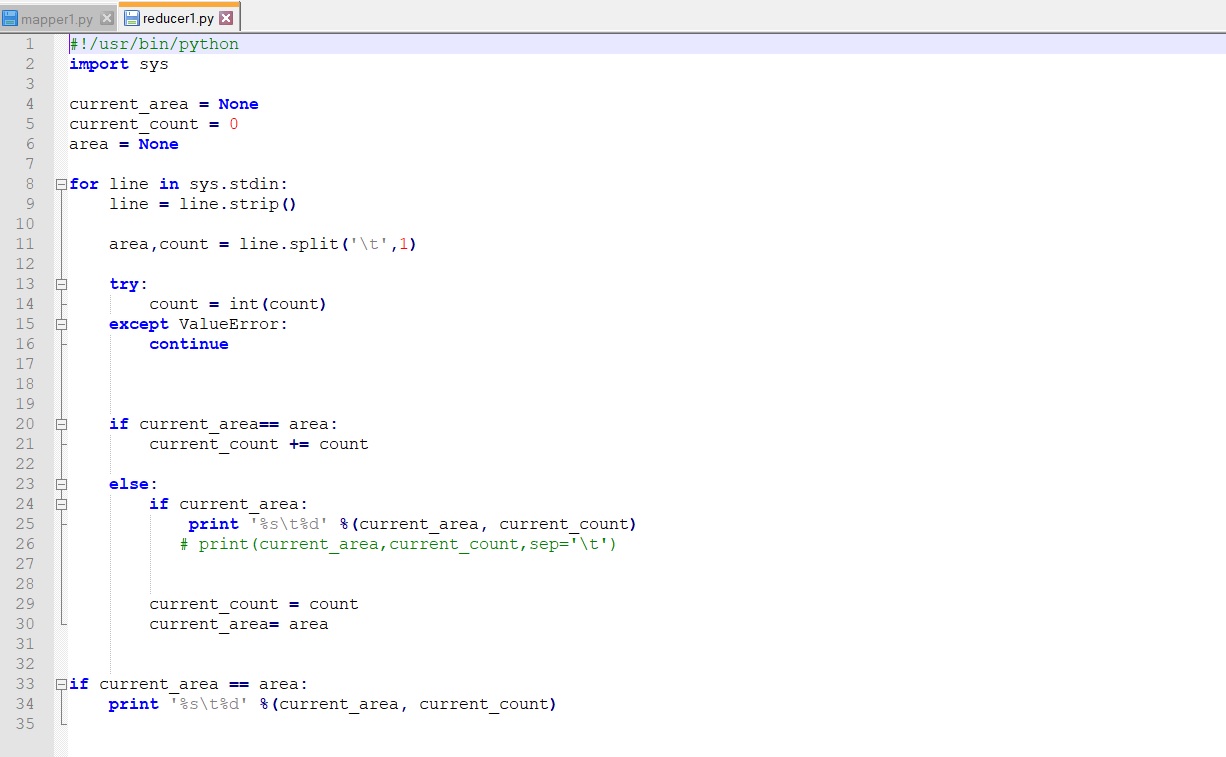
* **The output has been obtained.**



CODE FOR MAPPER (Python)

****

CODE FOR REDUCER (Python)

****

RESULT

After the step completes successfully, the output produced by the mapreduce is stored in the Amazon S3 output folder that you specified when you submitted the step.

To view the output of the mapreduce job

Open the Amazon S3 console at [https://console.aws.amazon.com/s3/](https://console.aws.amazon.com/s3/" \t "_blank)

In the Amazon S3 console, select the bucket that you used for the output data; for example, s3://mybucket/

Select the output folder

The mapreduce results are stored in a text file. To download the file, right-click it, choose Download, open the context (right-click) menu for Download, choose Save Link As, and then save the file to a suitable location.

Open the file using a text editor such as WordPad (Windows), TextEdit (Mac OS), or gEdit (Linux). In the output file, you should see the number of access requests by operating system.

RESETTING THE ENVIRONMENT

* Deleting Amazon S3 bucket
* Terminating the cluster

CONCLUSION

In this study, we obtained a statistical report on the safety measures and crime preventions in LA City which is rated high amongst other US major cities. I expect to gain knowledge in setting up cluster in Elastic MapReduce and uploading the bi dataset to S3 bucket, solving the computations over the data using the MapReduce python programming. As a future work, we would implement the same cluster with other application such as spark, hive, presto which are available in the Amazon EMR. Other than the technical skills, this will improve the analytical skills too.