**HCMC UNIVERSITY OF TECHNOLOGY AND EDUCATION**

**FACULTY OF INFORMATION TECHNOLOGY**

----🙦🙦🕮🙤🙤----

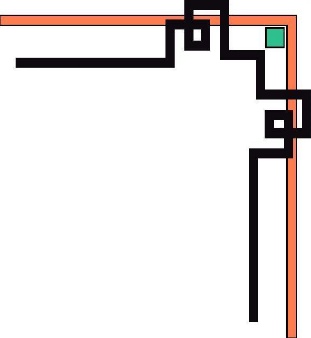
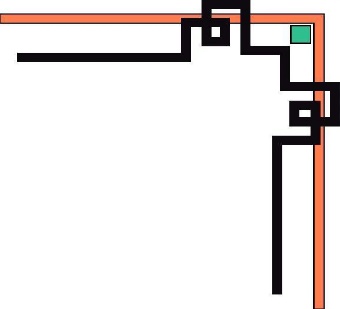
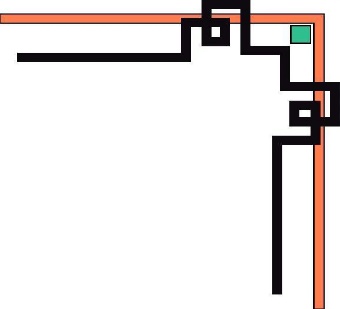
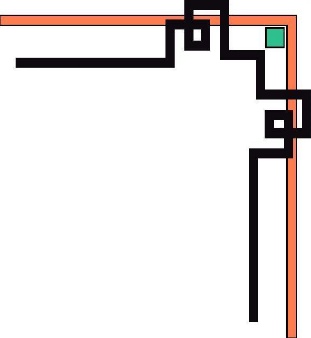


Instructor : Mrs . Truong Thi Ngoc Phuong

**Find out how to use Amazon EMR**

Bùi Tuấn Anh 17110001

Ho Chi Minh City, June 2020



Contents

[I. Overview and pricing 2](#_Toc42676848)

[II. Cluster’s state 4](#_Toc42676849)

[III. Step’s states 6](#_Toc42676850)

[IV. S3 service and amazon EMR 6](#_Toc42676851)

[V. Amazon EMR and development environment 8](#_Toc42676852)

[VI. Monitoring and managing the cluster in Amazon EMR with the console 9](#_Toc42676853)

[VII. Amazon EMR notebook overview. 10](#_Toc42676854)

[VIII.](#_Toc42676855) Security………………………………………………………………………………..11

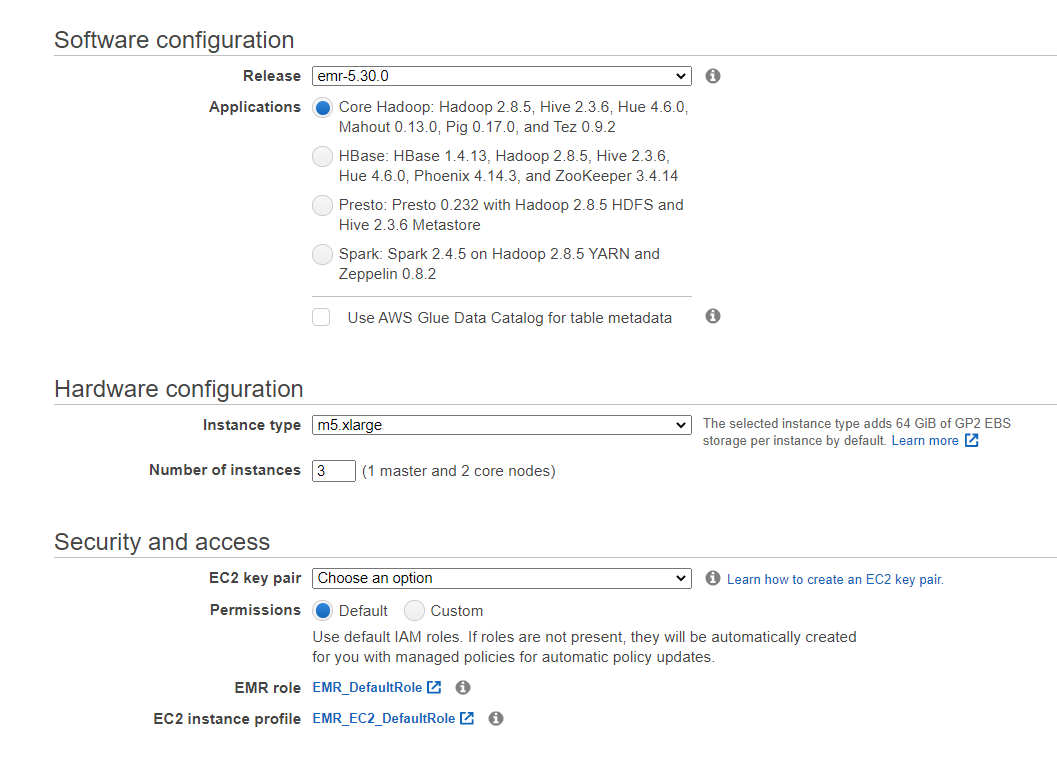
IX. Work assignment…………………………………………………………………………11

X. Reference…………………………………………………………………………………11

# 

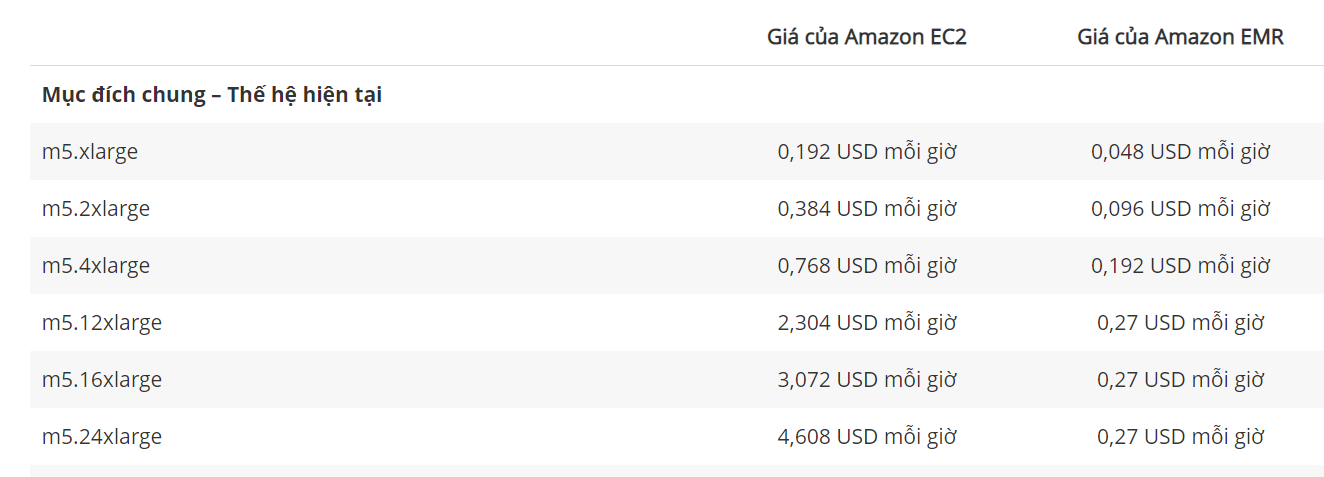
# Overview and pricing

* Amazon EMR is a service helps you to process big data using frameworks like Hadoop and hive, ….
* Amazon EMR is classified as PaaS.
* Amazon EMR will use instances in EC2 to run programs (as steps in amazon EMR)



picture 1

* The instances run based on EC2 services.
* In Amazon EMR, the Apache Hadoop map-reduce and some big data processing frameworks (Hive, Spark, Pig, …) are installed on instances.
* The price is based on what instances you choose but the price in EMR is different from the price of instance in EC2.



picture 2

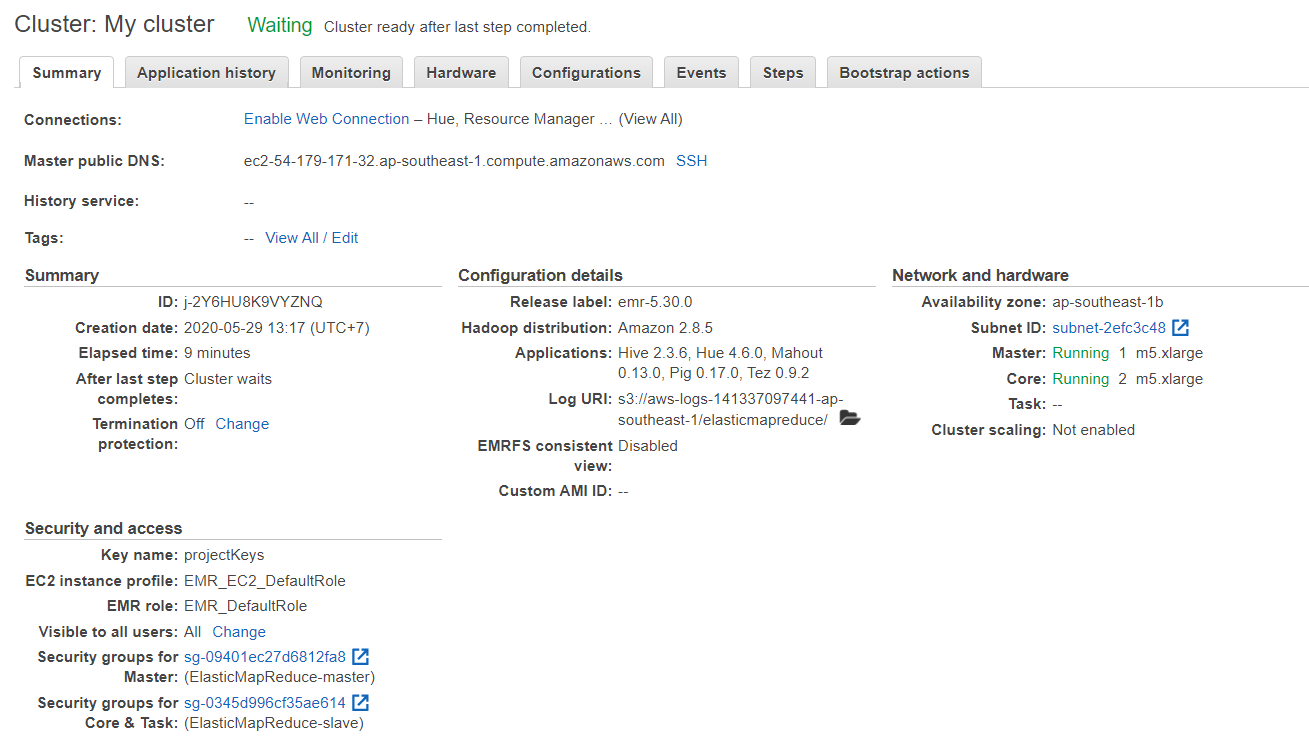
* In this service, you create cluster with as many instances as you want and choose the correct version of EMR (you can see in the picture 1 that the versions of some big data frameworks).

# Cluster’s state

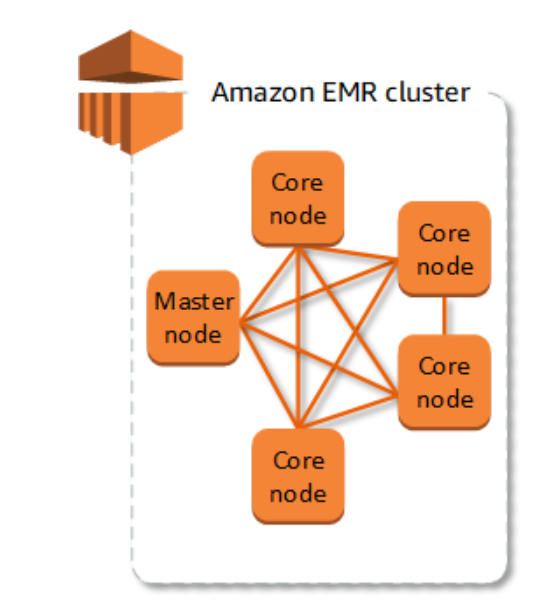
* When you created a cluster in amazon EMR, you can see that there will be three instances running in EC2.
* The cluster will go through 4 stages

Diagram 1 a normal state transition of a cluster

* Provisioning means that the EC2 is provisioning the instances for the cluster.
* Boostrapping means that your cluster is running bootstrap action to install more packages (maybe you want install tensorflow for example)
* Running means that your cluster is running the steps (the programs to process big data). Because there will be a default step will be running to prepare your cluster, so your cluster will be at running state when you first create your cluster.
* Waiting means that you can submit your steps to your cluster and it will execute the step you just submitted immediately.
* Some time your cluster will be automatically terminated due to some errors (bootstrap action failure, …)
* It means that you will have 1 master and 2 core nodes in amazon EMR.
* When you create a cluster and submit your steps, it will not execute immediately. You will have to wait until the EC2 instances running then your cluster will be ready to execute your steps. Just like the picture below



picture 3 a ready cluster



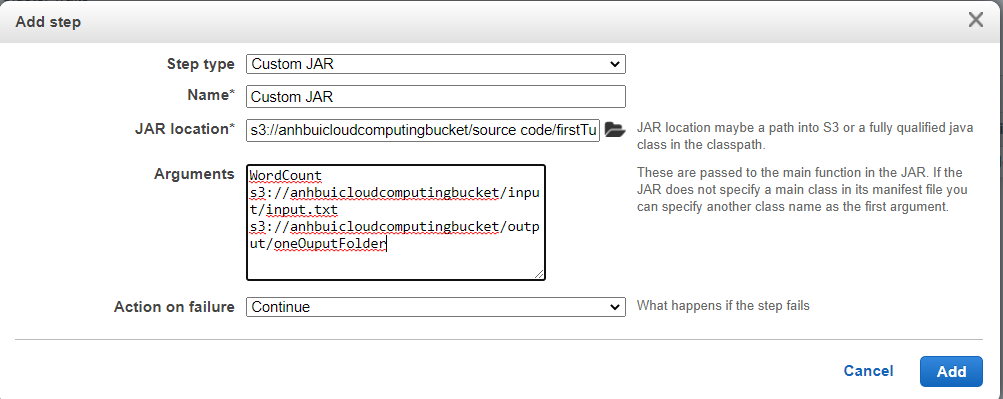
picture 4

# Step’s states

* Pending: The step is waiting to be processed
* Running: The step is being processed
* Completed: Amazon EMR has finished processing the step
* Failed: Amazon EMR failed to process the step due to something.

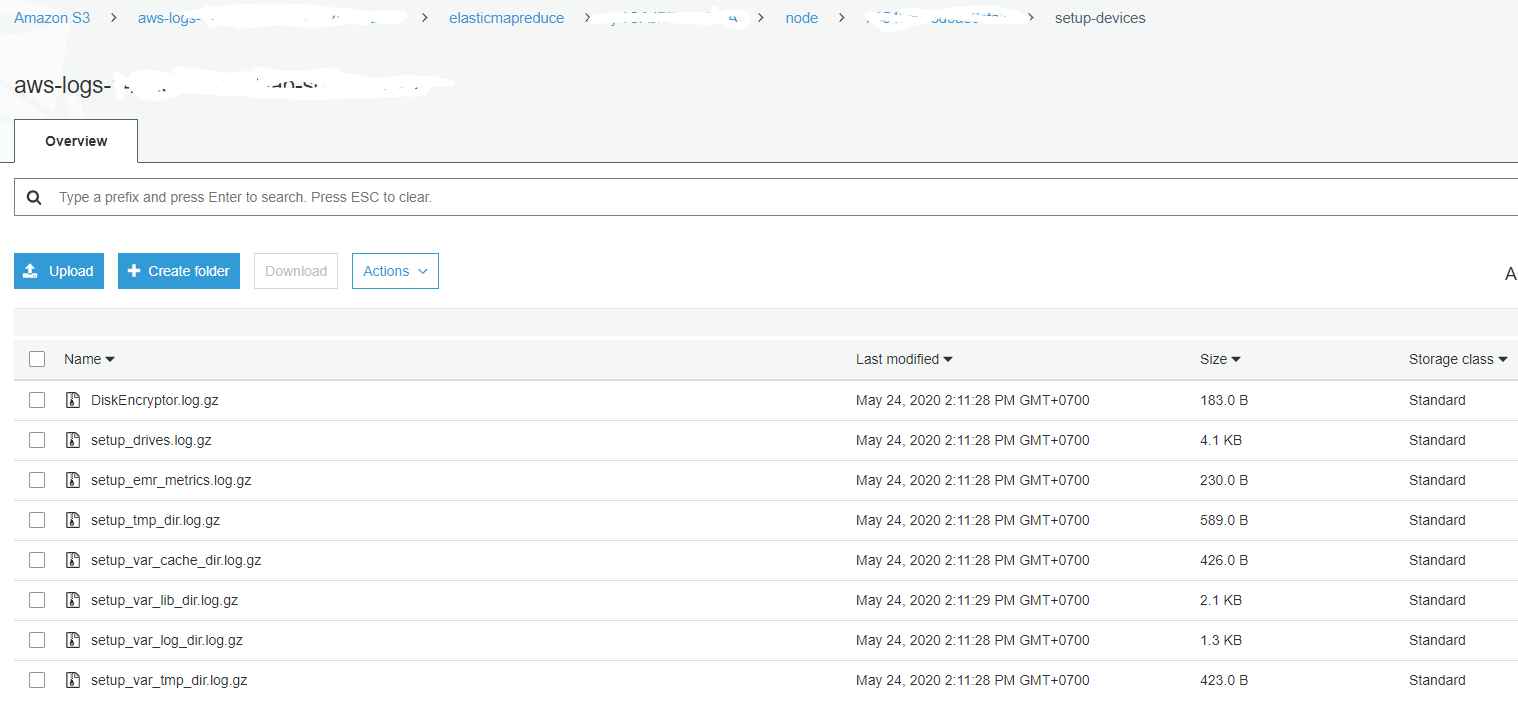
# S3 service and amazon EMR

* The files that you use in amazon EMR service is stored in your S3 buckets.
* Amazon EMR and other services of AWS only use the S3 of AWS.
* S3 service like a HDFS of Hadoop apache.
* The results that you get when the steps in amazon EMR completed will be in S3.
* Every files that you want to use to submit to steps of the cluster you can only interact with the files that are submitted to S3 service.



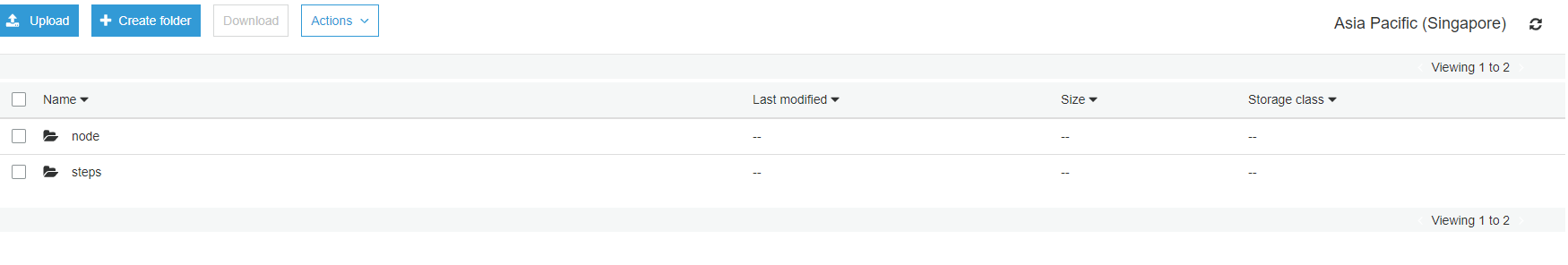
picture 5 input and output files are stored in S3

* When you create a cluster , amazon EMR will create a bucket for the cluster.
* In this bucket of your cluster, there will be log files that record the status or everything related to the cluster.



picture 6 log files location

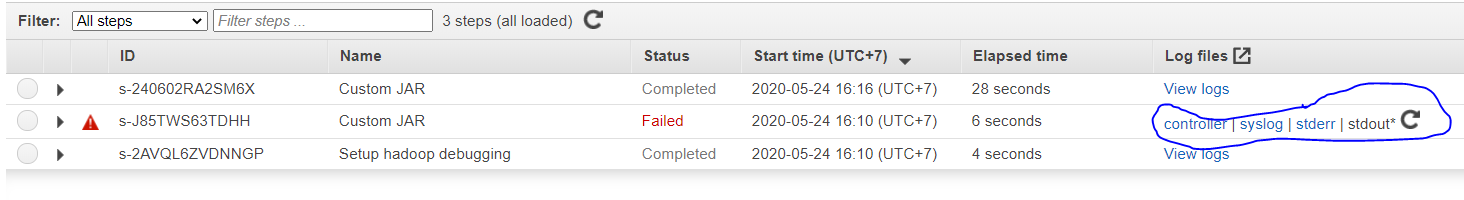
* The location of log files will have the following format.
* Amazon s3 > aws-logs - id of cluster bucket – your region > elasticmapreduce > cluster id.
* When you create a cluster, a bucket will created and the name of the bucket will be aws-logs - id of cluster bucket – your region. The ID of the bucket will be randomly generated.
* The folder cluster id will contain all the log files.



picture 7 folder contains log files

# Amazon EMR and development environment

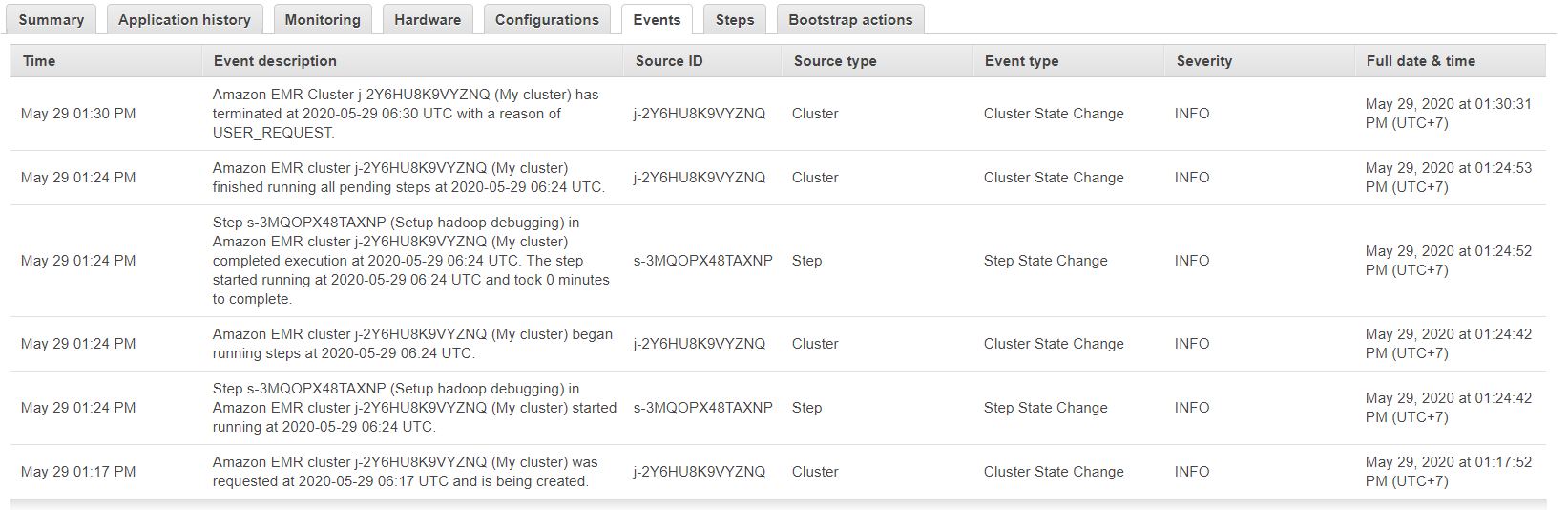
* The amazon EMR runtime uses java 8. Beware of this when you want to run a step with a custom JAR file. If your file is compiled in a different java version then maybe when you submit your step the step will be failed.
* Even if your step is failed due to something else, you still can find out the reasons why it is failed since amazon EMR provide log files for debugging.
* But still the log files are not enough to debug your program. Sometimes, you should run locally on your machine and use different tools to debug.



picture 8 The log files will tell the error why step failed are highlighted in the picture

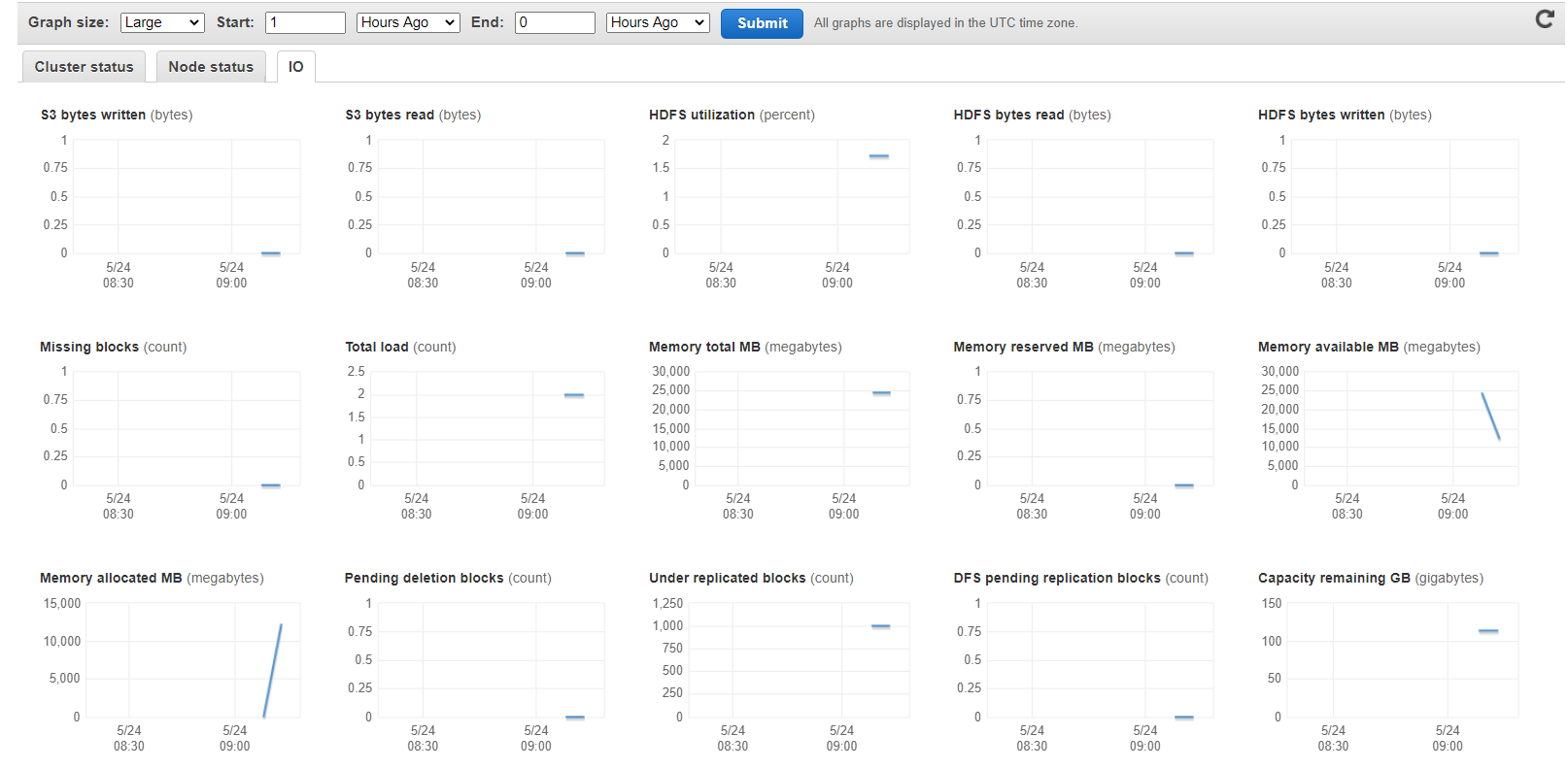
# Monitoring and managing the cluster in Amazon EMR with the console

* The AWS console provide a friendly UI for you to manage your clusters and programs. You will know when cluster’s state change or when your step has finished.
* This is the UI to see what events have happened. You can see what events just happened in your cluster very easily with the UI that AWS provided.



picture 9

* To manage your program or steps, the amazon EMR provide a UI that looks like picture 8. As you can see from the picture, you will know the time need to execute one step. Also you can know the status (failed, completed…) and type of step (Custom JAR, Hive program…).
* The summary information of a cluster you can refer to picture 3.
* You can monitor also how the cluster is performing, the current status of cluster.



picture 10

# Amazon EMR notebook overview.

* Amazon EMR notebook is jupyter notebook that uses amazon emr clusters to run your applications. Note that the clusters must run apache spark.
* You can create and store your notebook in amazon s3.
* It has some pre-installed libraries and packages that are necessary for big data analyzing (packages related to spark)
* For more detail information about the versions of notebook , python and spark . Please refer to this link <https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-managed-notebooks-considerations.html>
* If you need additional software or packages (like tensorflow...), you still can install them using bootstrap actions.
* Note that you should use EMR 5.29 because this version is more stable

# References

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html>