WordCount

amazon

EMR

Word count example

To run word count in AWS you have two different ways; either use the already exist WordCount program, or to write your own file.

We will be using our own word count program

Prerequisites :

a. S3 bucket

b. EC2 key pair

Steps:

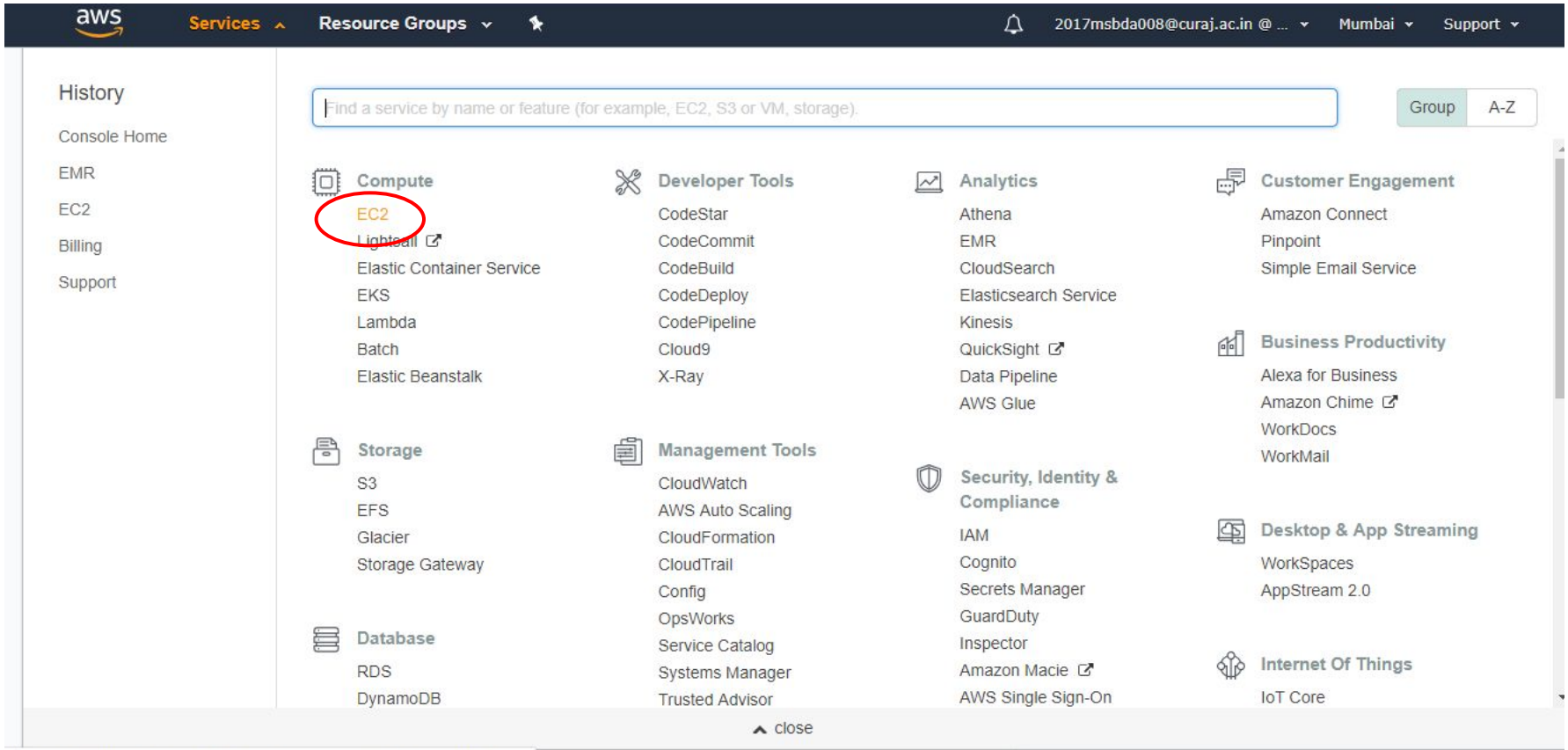
1. Login to your account in AWS

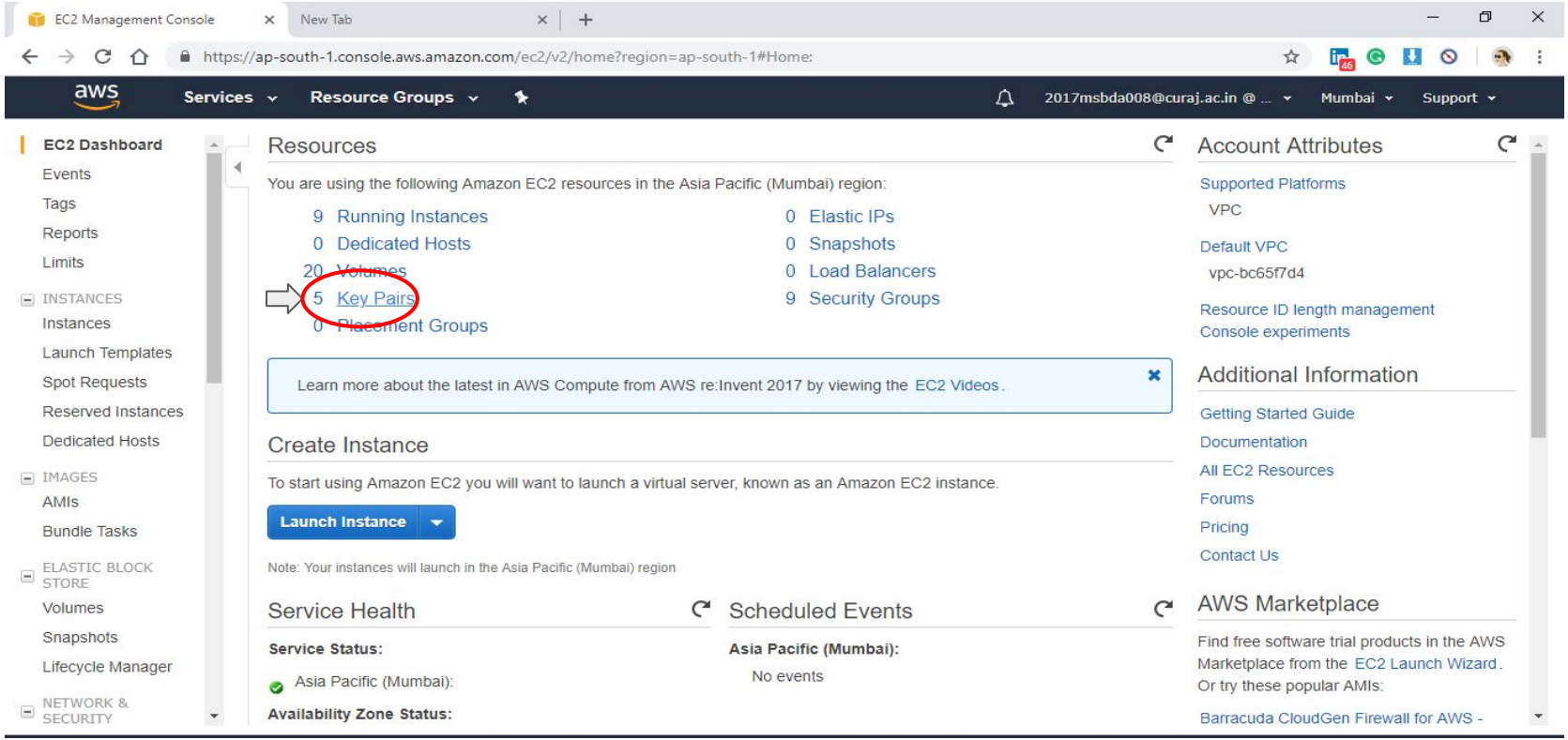
2. If you did not create Key-Pair and S3 (see below i explained how to do).

3. From Amazon Web Services, go to EMR

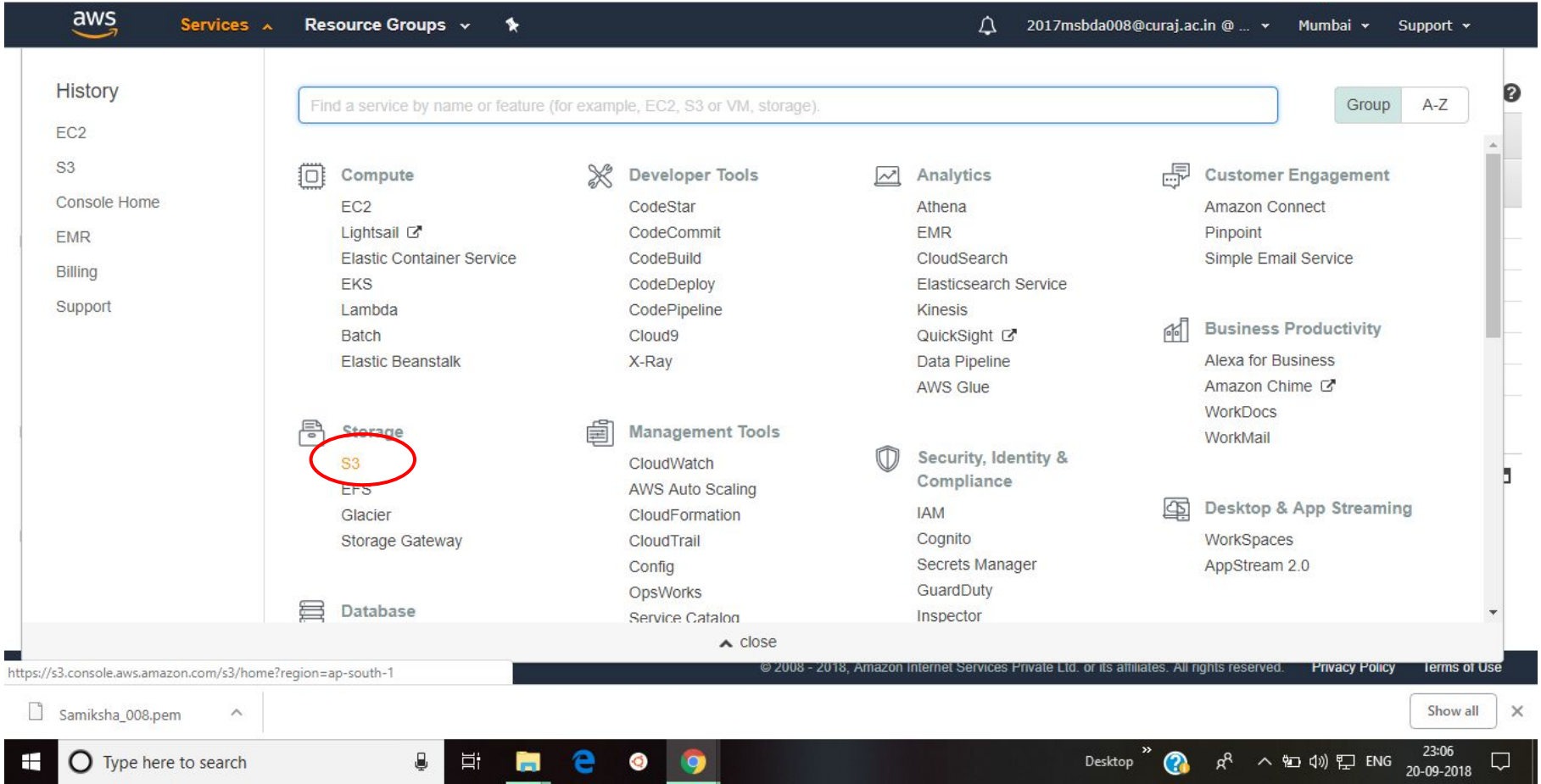
Create Key Pair

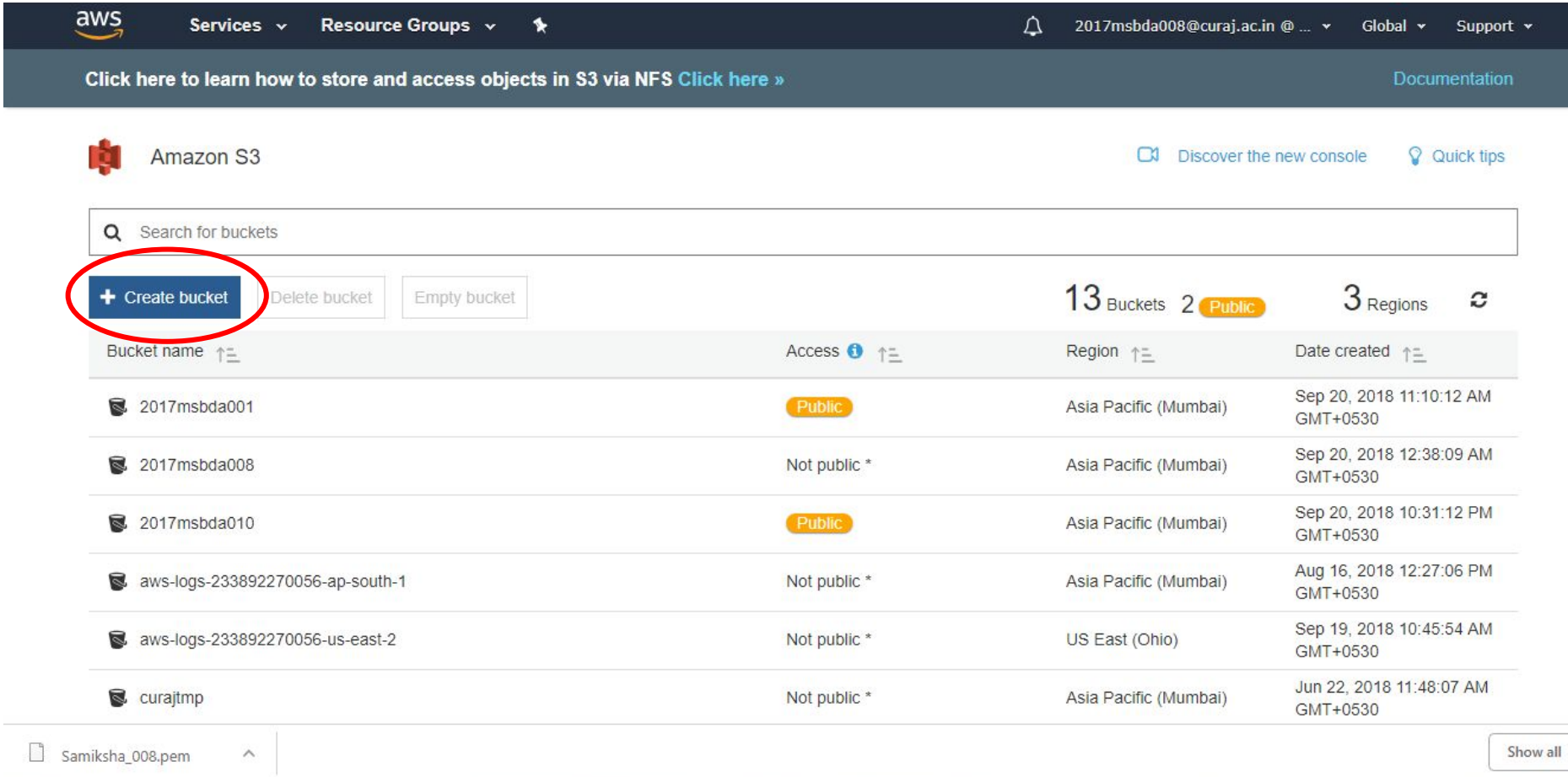
Step1- Create key-Pair -> Go to the EC2



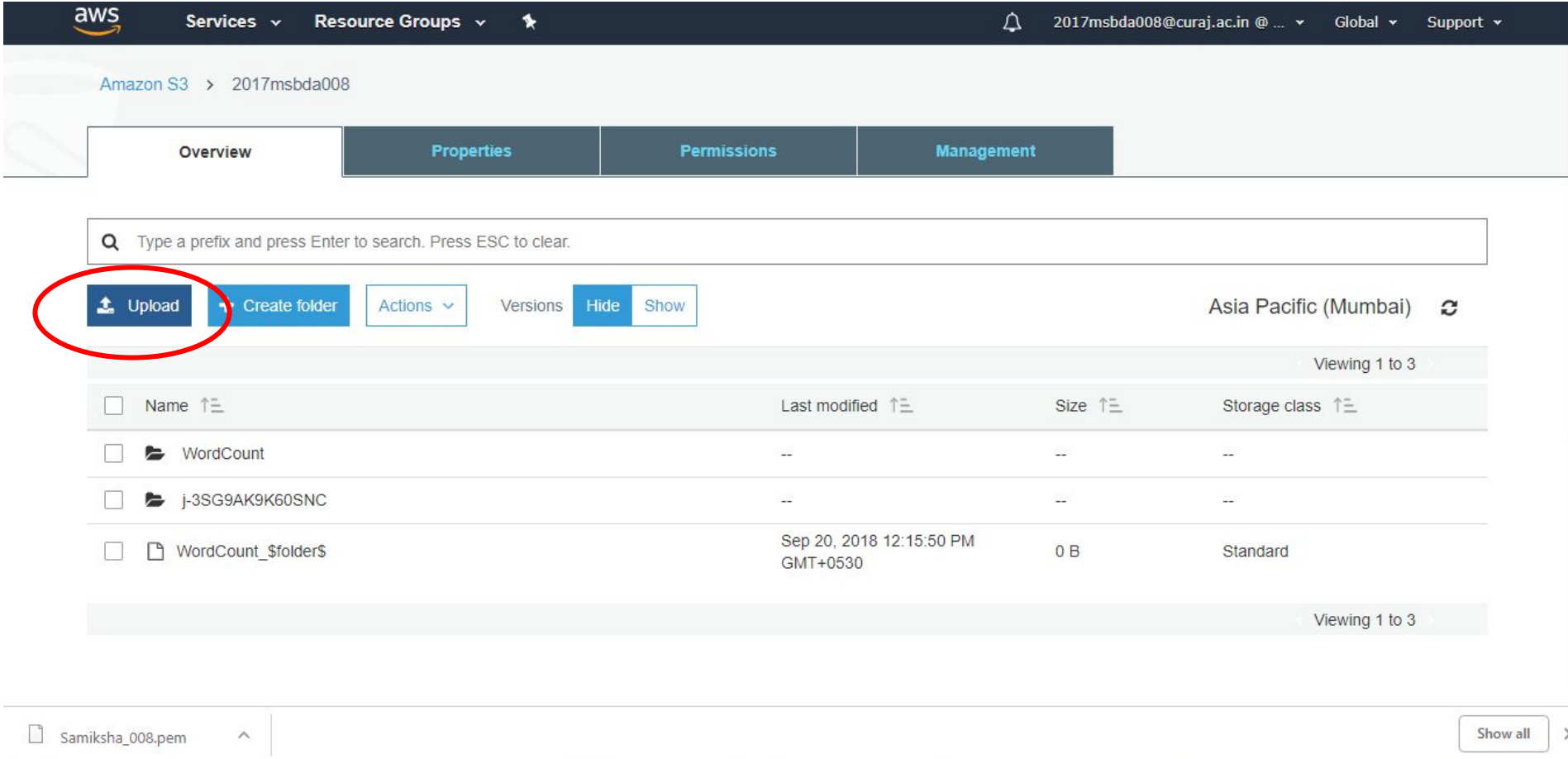

EC2 Management Console
New Tab
https://ap-south-1.console.aws.amazon.com/ec2/v2/home?region=ap-south-1#Home:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
NETWORK & SECURITY
Resources
You are using the following Amazon EC2 resources in the Asia Pacific (Mumbai) region:
9 Running Instances
0 Dedicated Hosts
20 Volumes
5 Key Pairs
0 Placement Groups
0 Elastic IPs
0 Snapshots
0 Load Balancers
9 Security Groups
Learn more about the latest in AWS Compute from AWS re:Invent 2017 by viewing the EC2 Videos.
Create Instance
To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.
Launch Instance
Note: Your instances will launch in the Asia Pacific (Mumbai) region
Service Health
Service Status:
Asia Pacific (Mumbai):
Availability Zone Status:
Scheduled Events
Asia Pacific (Mumbai):
No events
Account Attributes
Supported Platforms
VPC
Default VPC
vpc-bc65f7d4
Resource ID length management
Console experiments
Additional Information
Getting Started Guide
Documentation
All EC2 Resources
Forums
Pricing
Contact Us
AWS Marketplace
Find free software trial products in the AWS Marketplace from the EC2 Launch Wizard. Or try these popular AMIs:
Barracuda CloudGen Firewall for AWS -

Create S3 bucket..

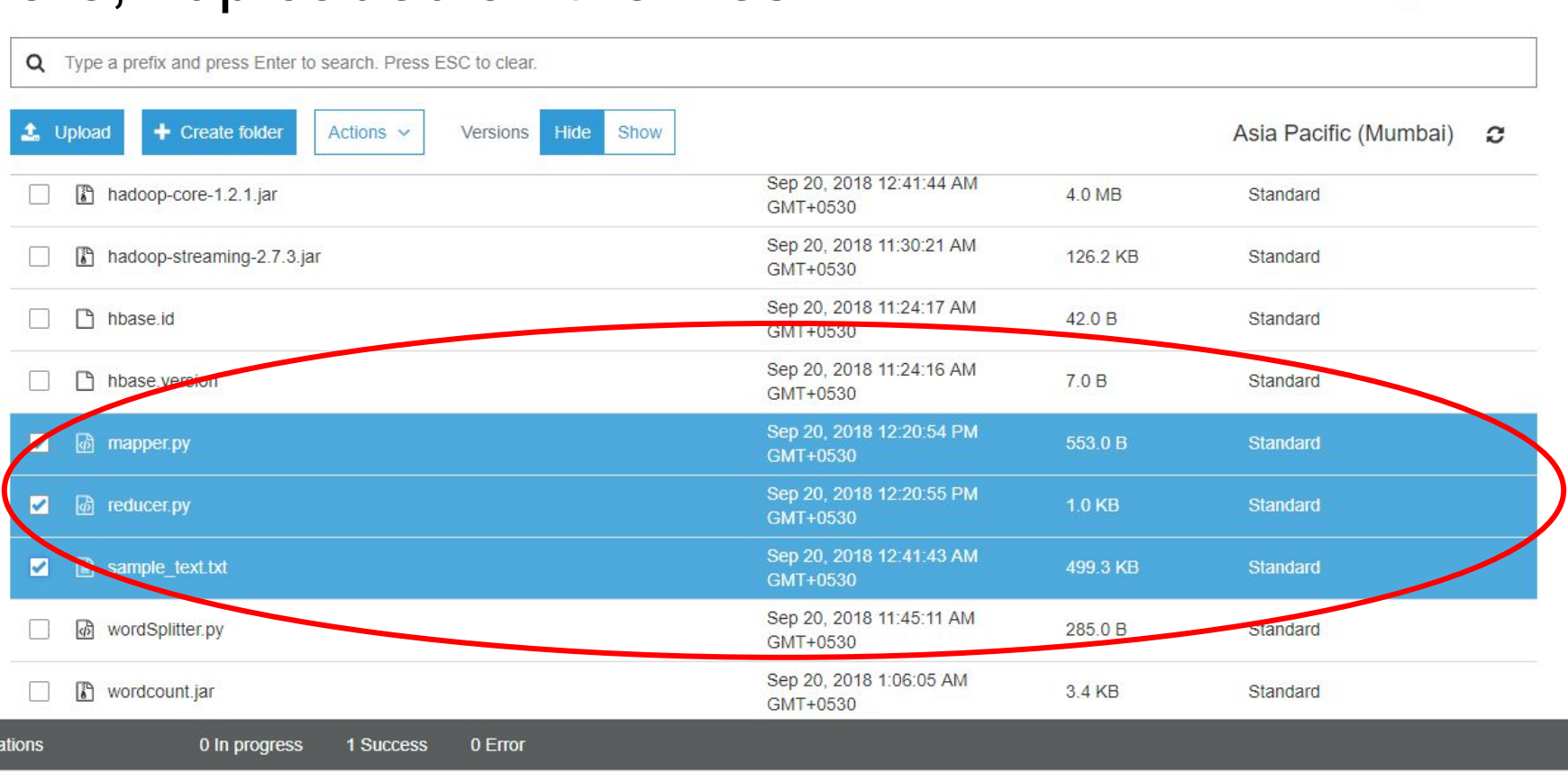



aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Global
Support
Click here to learn how to store and access objects in S3 via NFS Click here »
Documentation
Amazon S3
Discover the new console
Quick tips
Search for buckets
Create bucket
Delete bucket
Empty bucket
13 Buckets 2 Public
3 Regions
Bucket name
Access
Region
Date created
2017msbda001
Public
Asia Pacific (Mumbai)
Sep 20, 2018 11:10:12 AM GMT+0530
2017msbda008
Not public *
Asia Pacific (Mumbai)
Sep 20, 2018 12:38:09 AM GMT+0530
2017msbda010
Public
Asia Pacific (Mumbai)
Sep 20, 2018 10:31:12 PM GMT+0530
aws-logs-233892270056-ap-south-1
Not public *
Asia Pacific (Mumbai)
Aug 16, 2018 12:27:06 PM GMT+0530
aws-logs-233892270056-us-east-2
Not public *
US East (Ohio)
Sep 19, 2018 10:45:54 AM GMT+0530
curajtmp
Not public *
Asia Pacific (Mumbai)
Jun 22, 2018 11:48:07 AM GMT+0530
Samiksha_008.pem
Show all

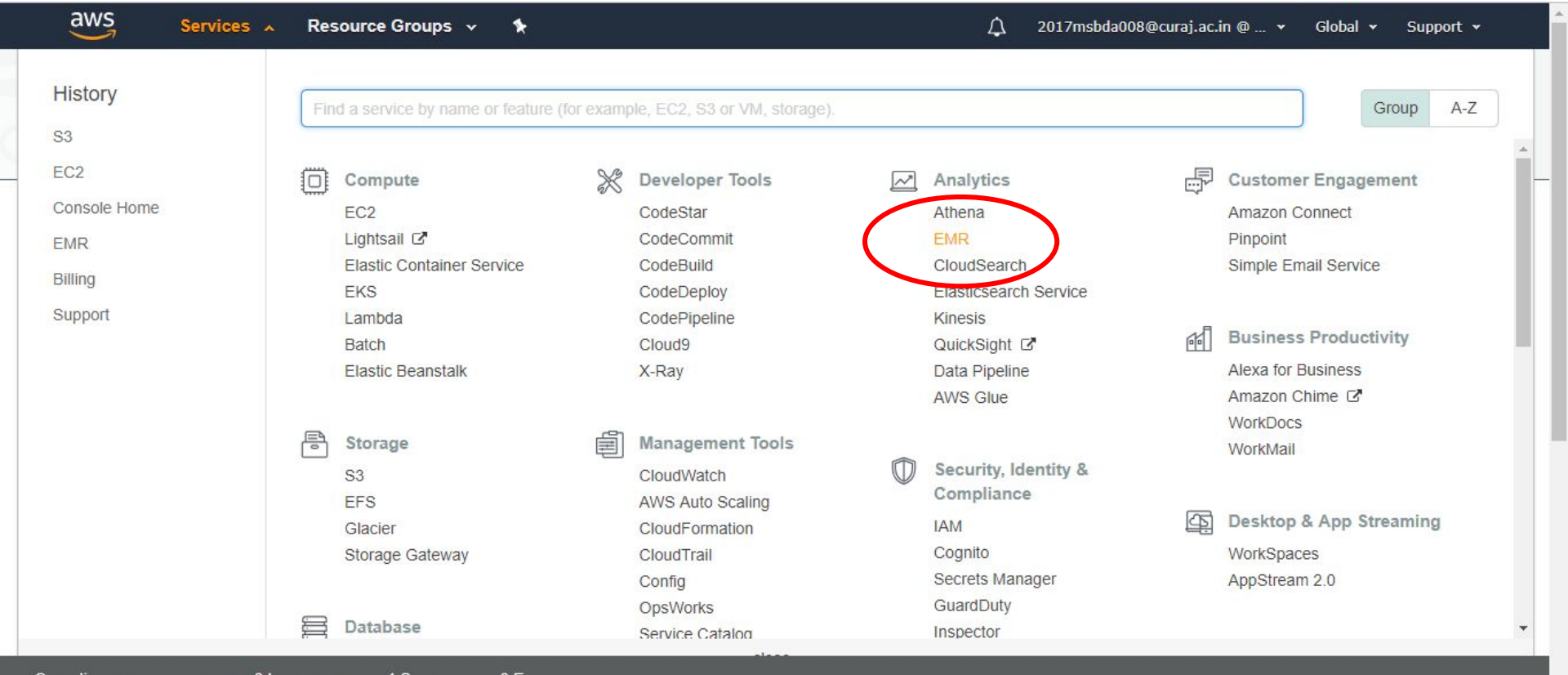
Now, create a folder WordCount and upload your Mapper and Reducer program and your Input file in this folder..



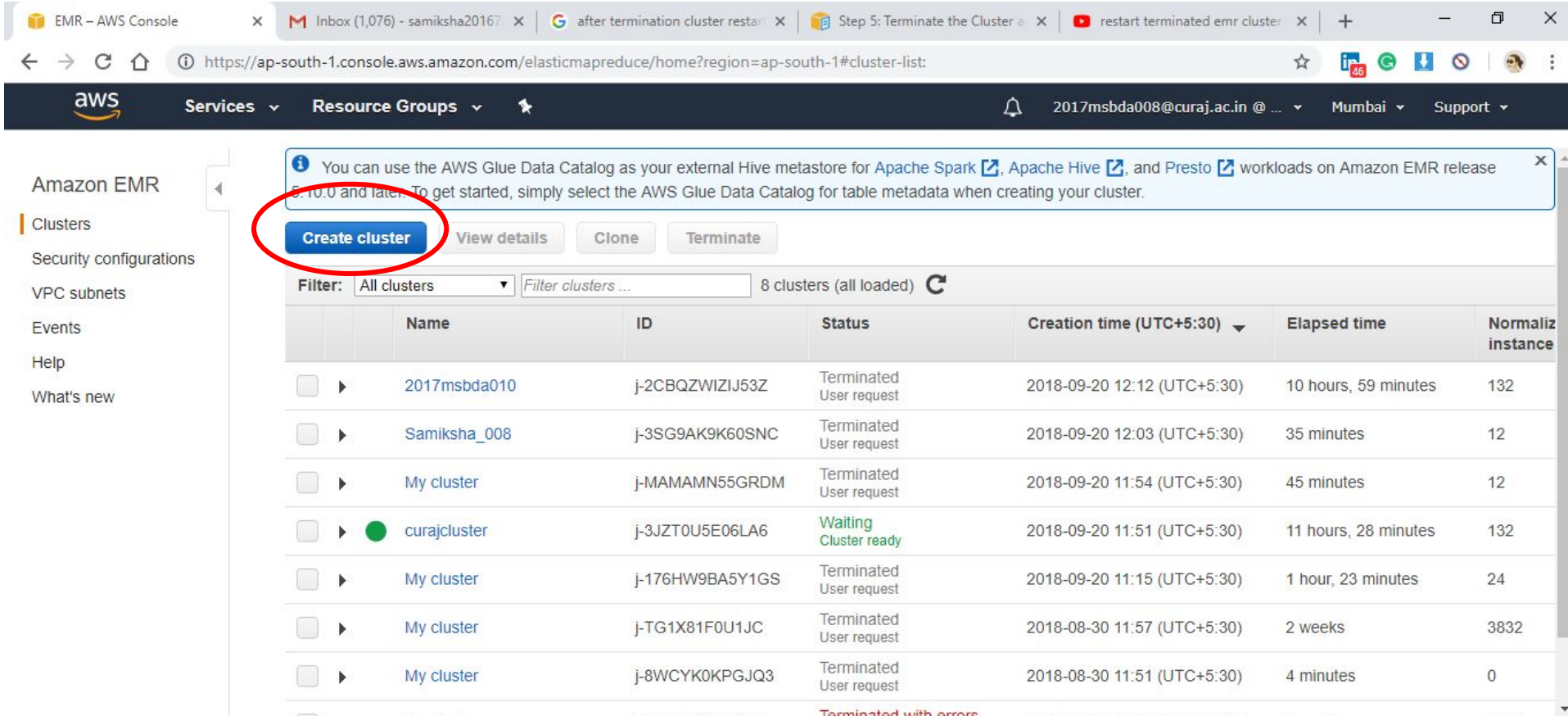
Here, i uploaded all the files..

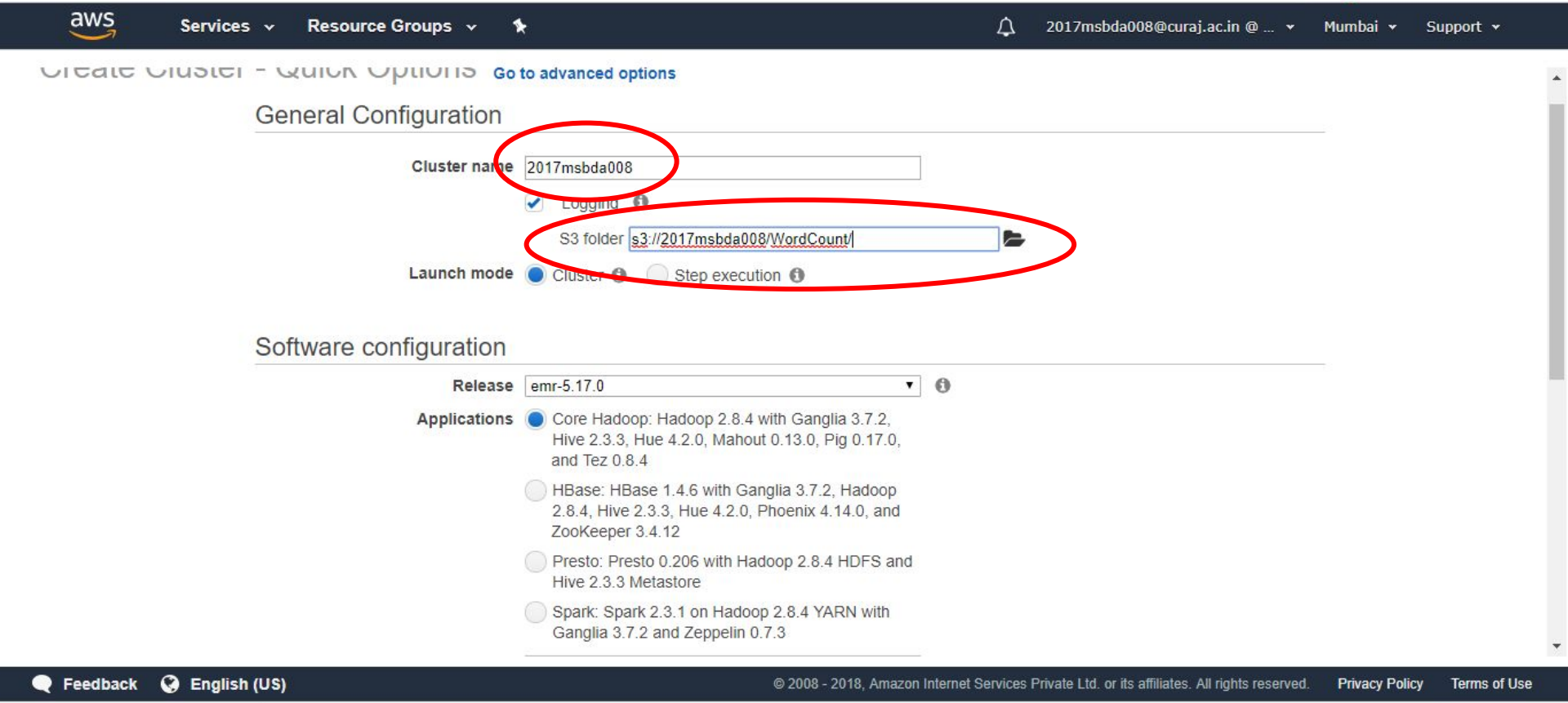


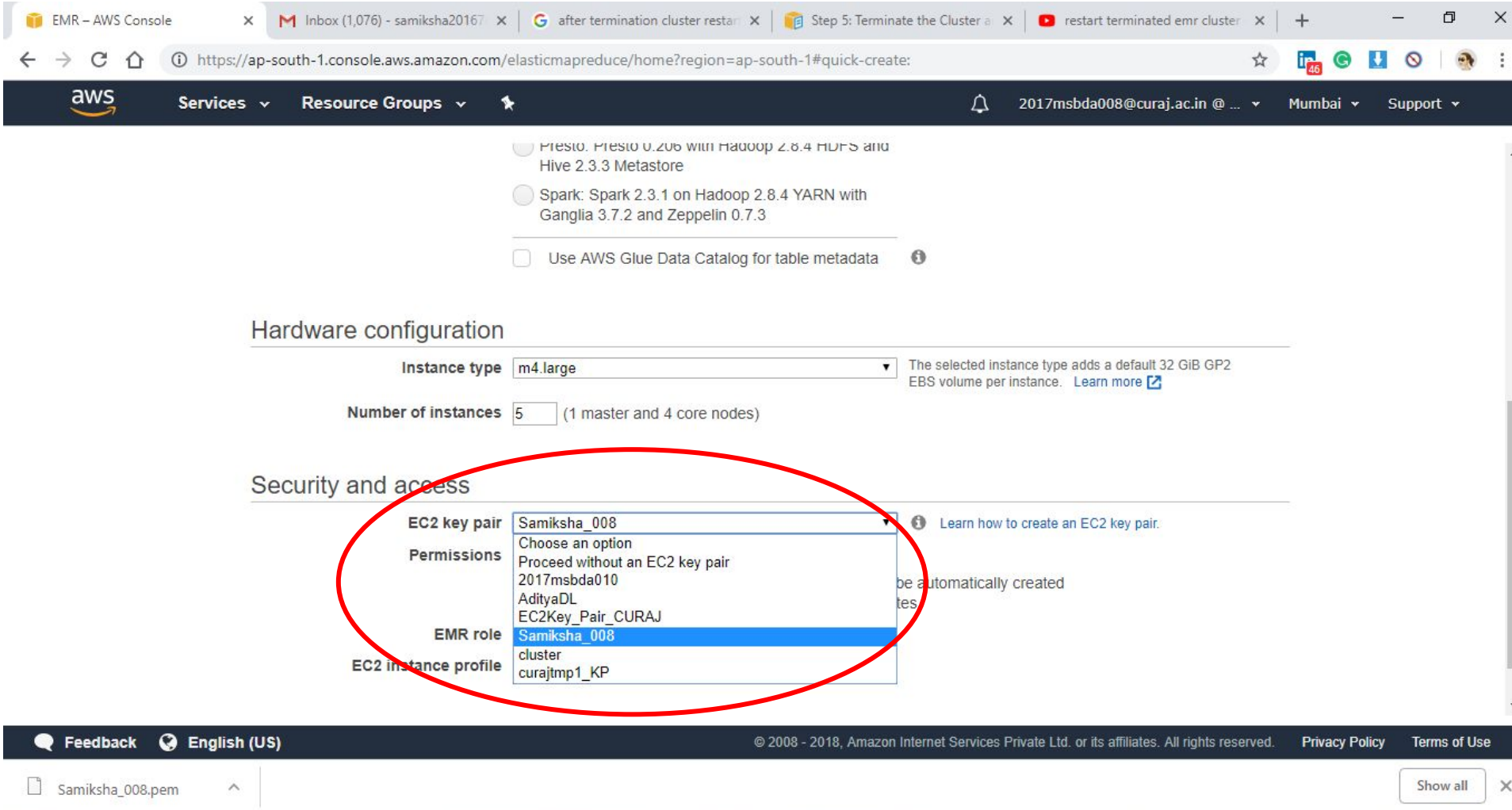
Go to EMR

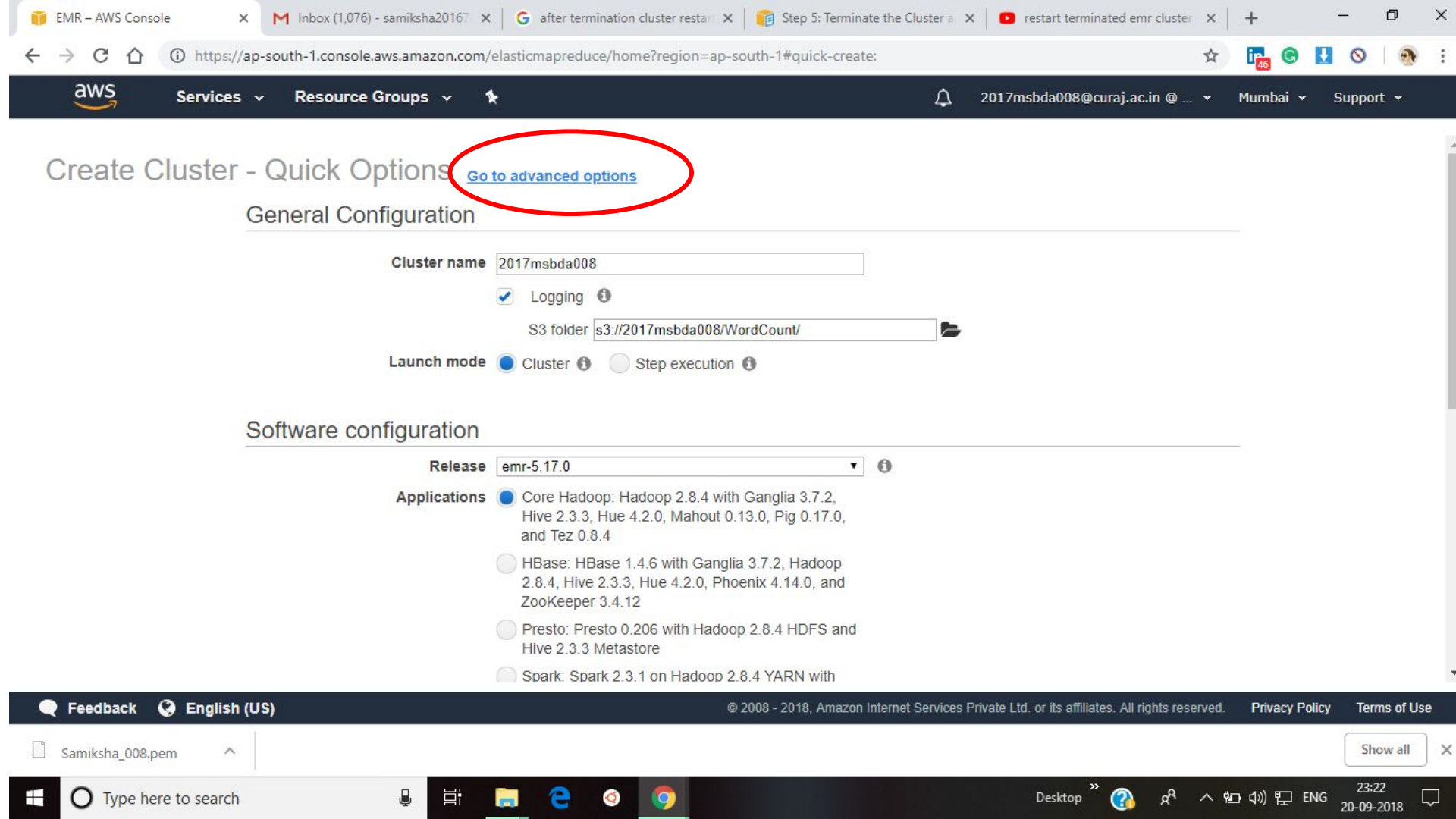

aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Global
Support
History
S3
EC2
Console Home
EMR
Billing
Support
Find a service by name or feature (for example, EC2, S3 or VM, storage).
Group
A-Z
Compute
EC2
Lightsail
Elastic Container Service
EKS
Lambda
Batch
Elastic Beanstalk
Developer Tools
CodeStar
CodeCommit
CodeBuild
CodeDeploy
CodePipeline
Cloud9
X-Ray
Analytics
Athena
EMR
CloudSearch
Elasticsearch Service
Kinesis
QuickSight
Data Pipeline
AWS Glue
Customer Engagement
Amazon Connect
Pinpoint
Simple Email Service
Business Productivity
Alexa for Business
Amazon Chime
WorkDocs
WorkMail
Storage
S3
EFS
Glacier
Storage Gateway
Management Tools
CloudWatch
AWS Auto Scaling
CloudFormation
CloudTrail
Config
OpsWorks
Service Catalog
Security, Identity & Compliance
IAM
Cognito
Secrets Manager
GuardDuty
Inspector
Desktop & App Streaming
WorkSpaces
AppStream 2.0
Database

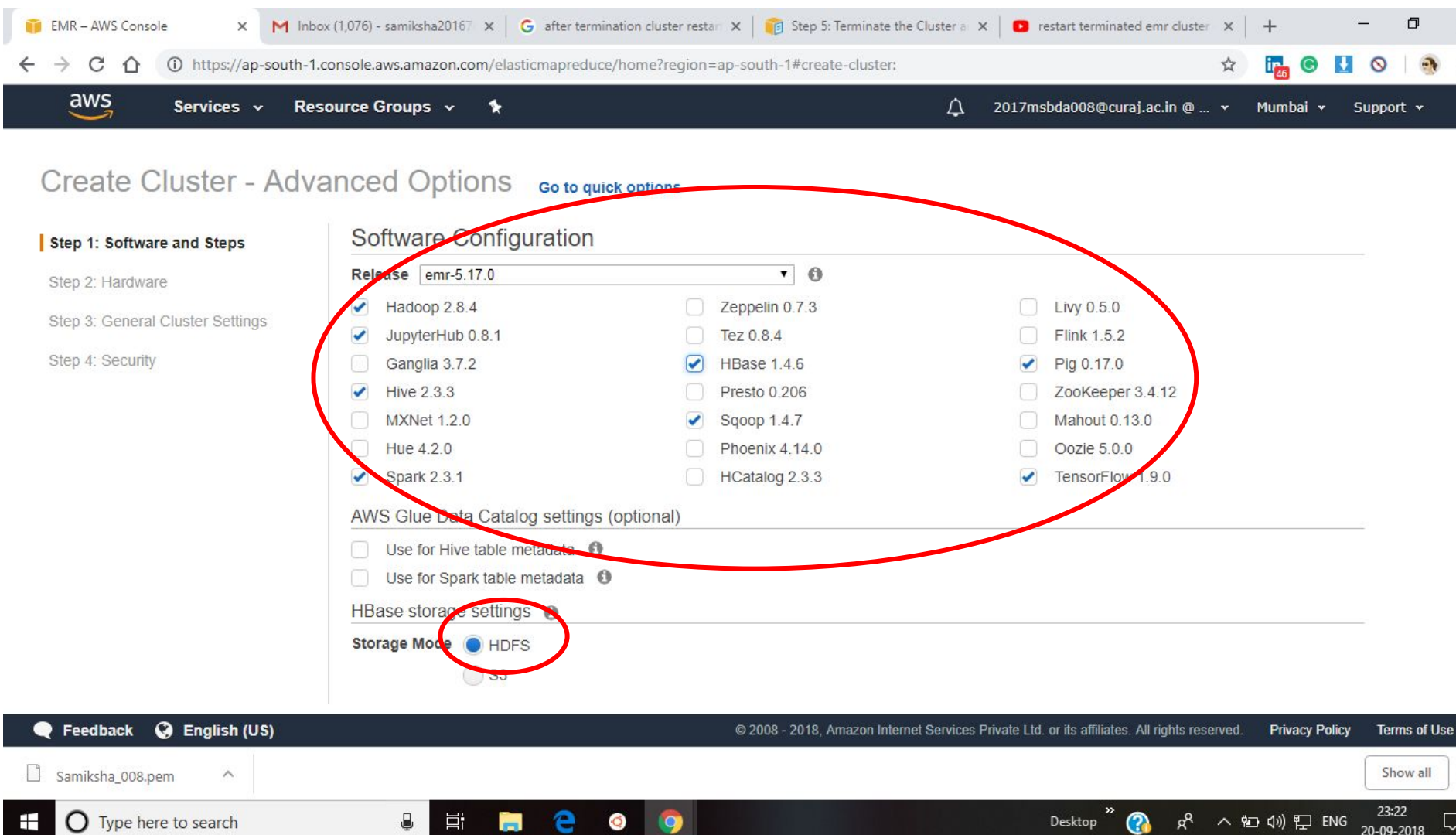
Create cluster..

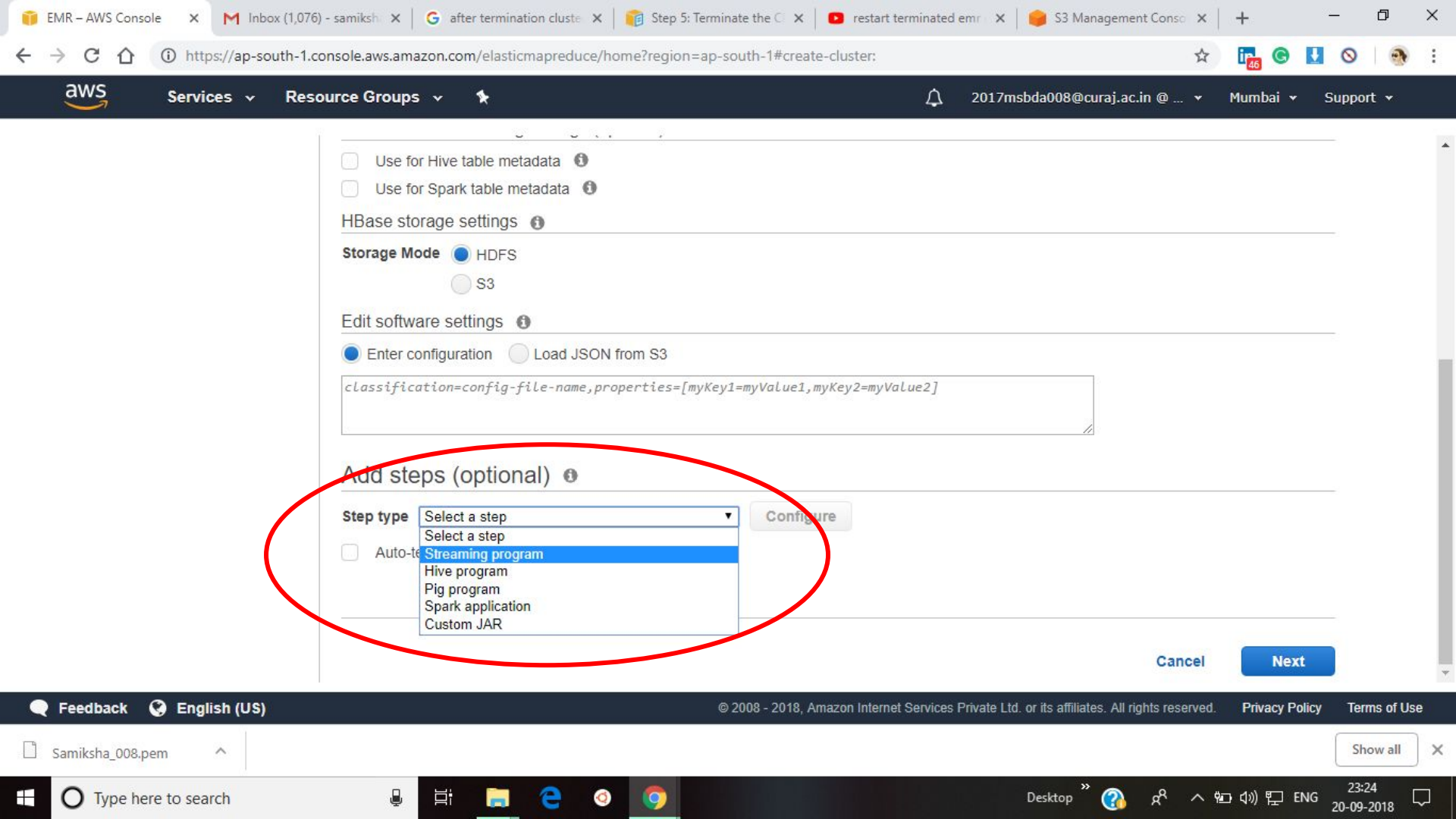


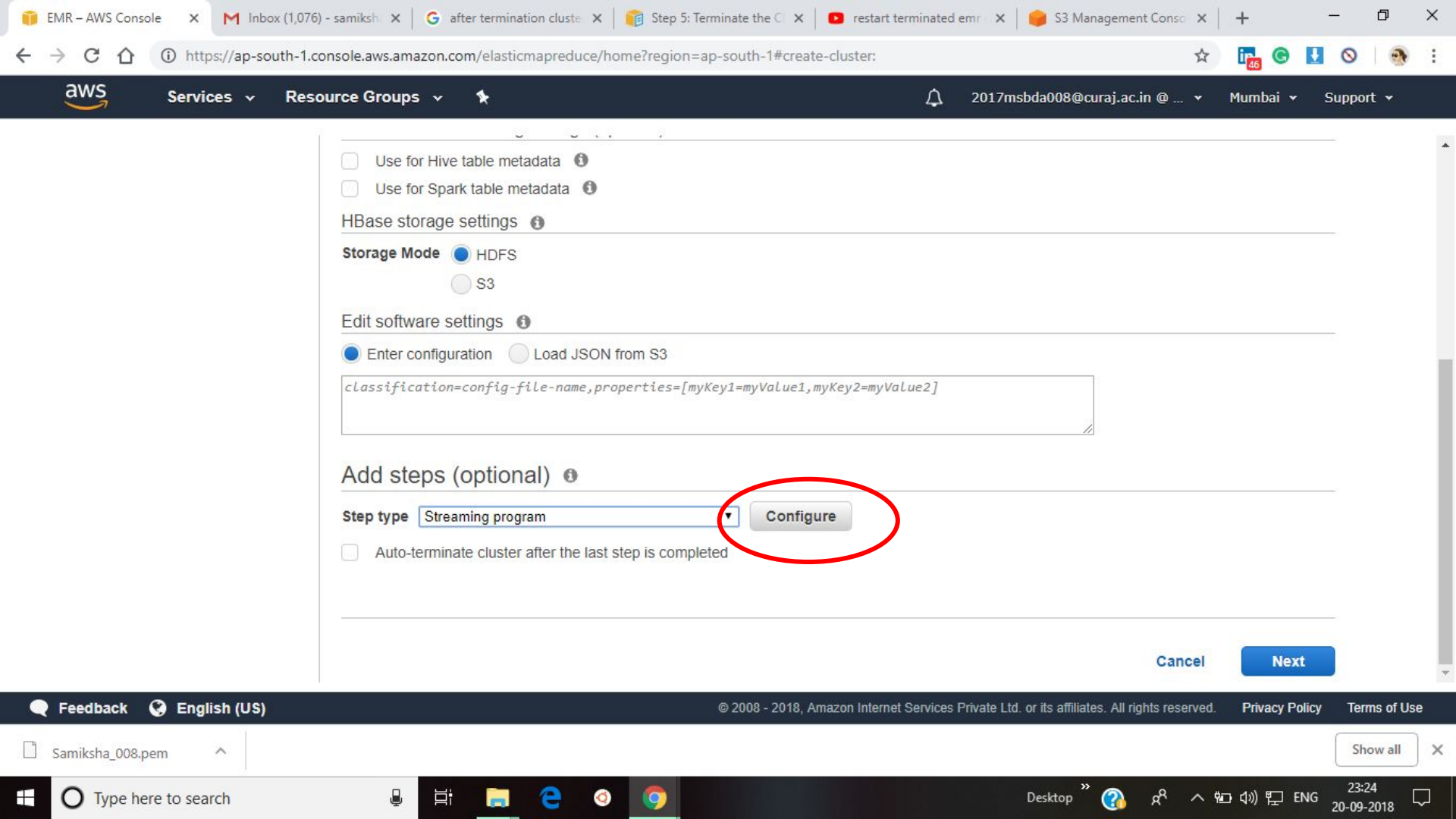

aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Create Cluster - Quick Options
Go to advanced options
General Configuration
Cluster name
2017msbda008
Logging
S3 folder
s3://2017msbda008/WordCount/
Launch mode
Cluster
Step execution
Software configuration
Release
emr-5.17.0
Applications
Core Hadoop: Hadoop 2.8.4 with Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4
HBase: HBase 1.4.6 with Ganglia 3.7.2, Hadoop 2.8.4, Hive 2.3.3, Hue 4.2.0, Phoenix 4.14.0, and ZooKeeper 3.4.12
Presto: Presto 0.206 with Hadoop 2.8.4 HDFS and Hive 2.3.3 Metastore
Spark: Spark 2.3.1 on Hadoop 2.8.4 YARN with Ganglia 3.7.2 and Zeppelin 0.7.3
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all

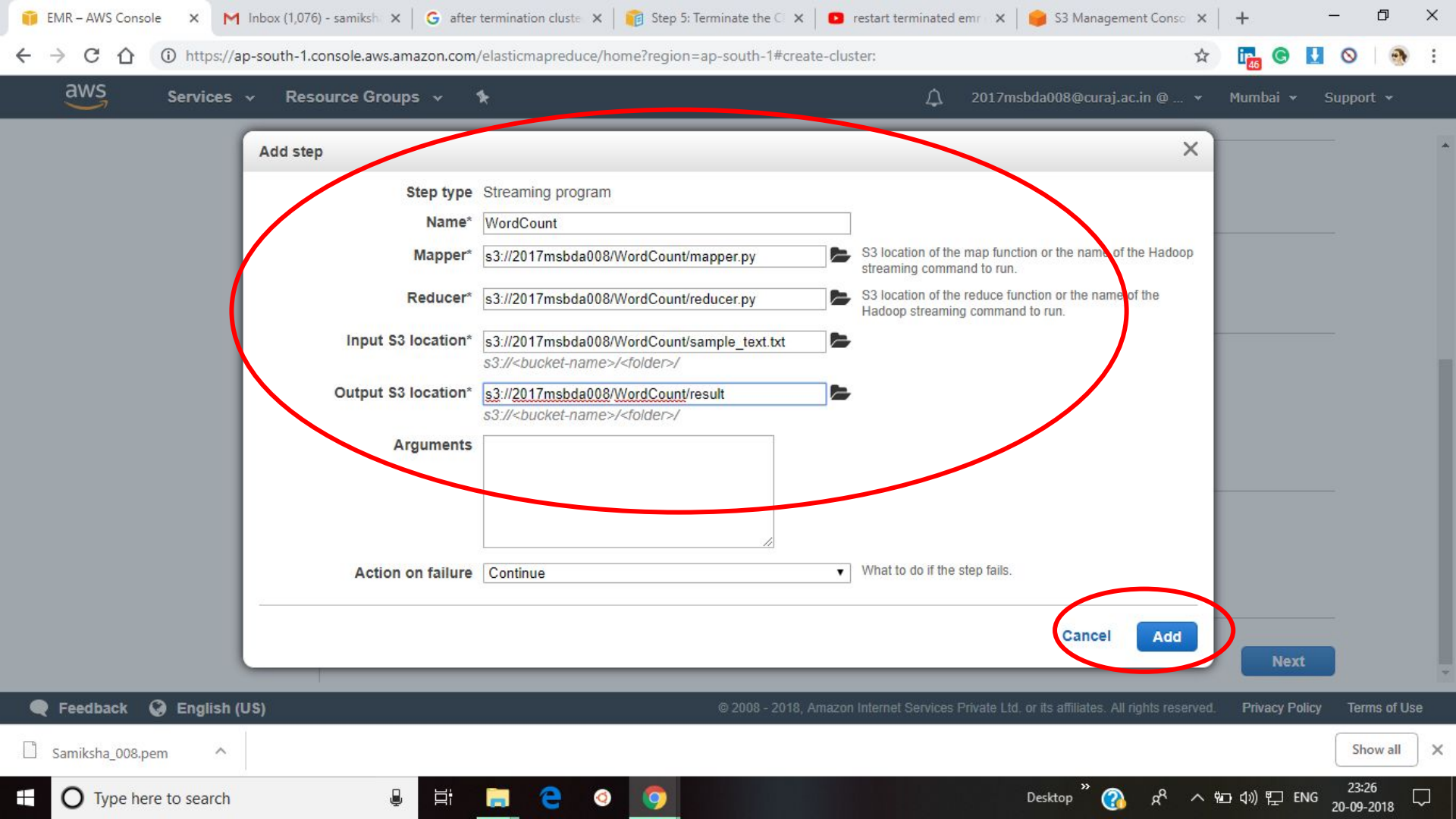

EMR – AWS Console
Inbox (1,076) - samiksha20167
after termination cluster restar
Step 5: Terminate the Cluster a
restart terminated emr cluster
https://ap-south-1.console.aws.amazon.com/elasticmapreduce/home?region=ap-south-1#quick-create:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Hive 2.3.3 Metastore
Spark: Spark 2.3.1 on Hadoop 2.8.4 YARN with Ganglia 3.7.2 and Zeppelin 0.7.3
Use AWS Glue Data Catalog for table metadata
Hardware configuration
Instance type
m4.large
The selected instance type adds a default 32 GiB GP2 EBS volume per instance. Learn more
Number of instances
5
(1 master and 4 core nodes)
Security and access
EC2 key pair
Samiksha_008
Learn how to create an EC2 key pair.
Permissions
Choose an option
Proceed without an EC2 key pair
2017msbda010
AdityaDL
EC2Key_Pair_CURAJ
Samiksha_008
cluster
curajtmp1_KP
be automatically created
tes
EMR role
EC2 instance profile
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all

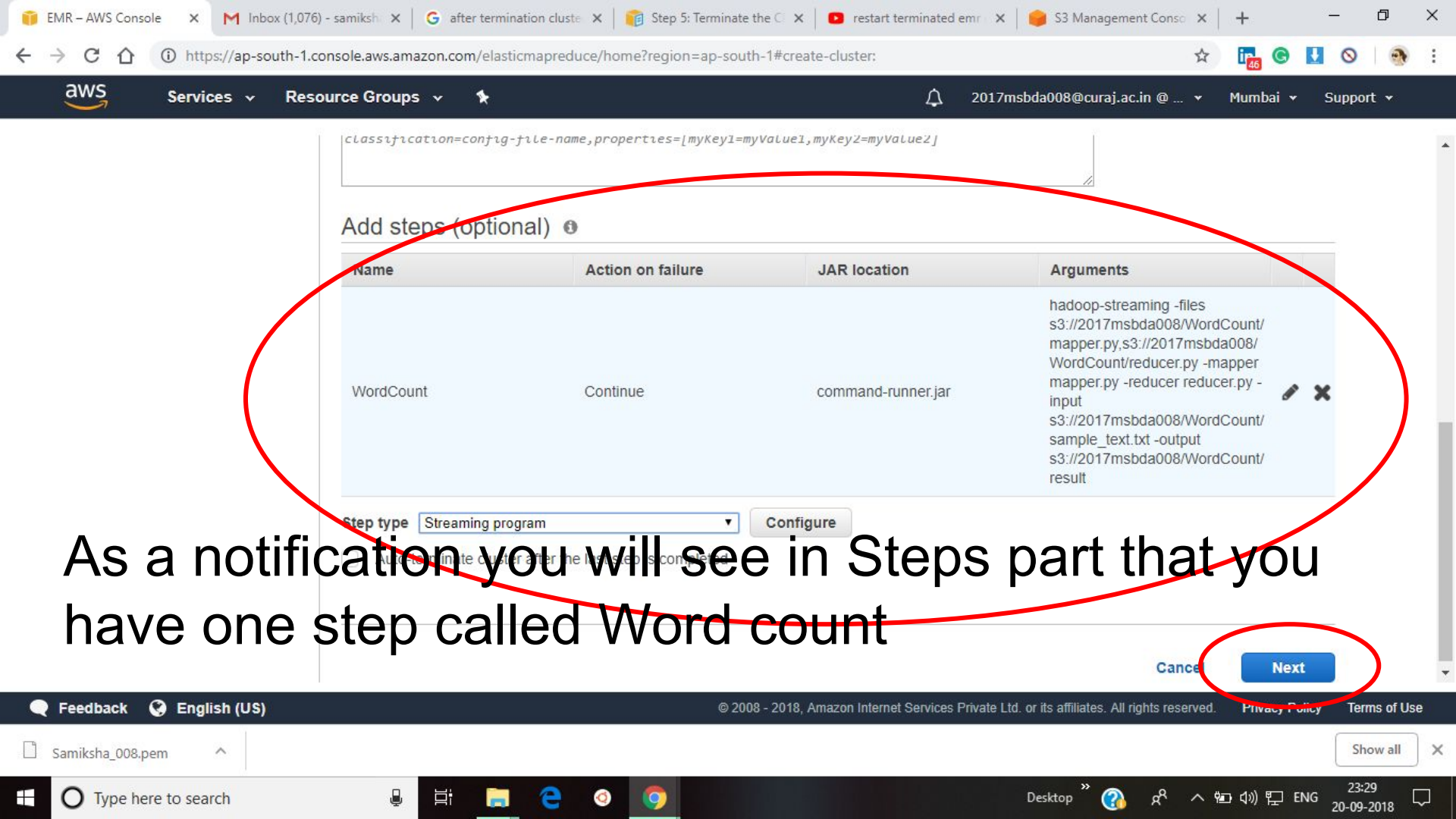

EMR – AWS Console
Inbox (1,076) - samiksha20167
after termination cluster restar
Step 5: Terminate the Cluster a
restart terminated emr cluster
https://ap-south-1.console.aws.amazon.com/elasticmapreduce/home?region=ap-south-1#quick-create:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Create Cluster - Quick Options
Go to advanced options
General Configuration
Cluster name
2017msbda008
Logging
S3 folder
s3://2017msbda008/WordCount/
Launch mode
Cluster
Step execution
Software configuration
Release
emr-5.17.0
Applications
Core Hadoop: Hadoop 2.8.4 with Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4
HBase: HBase 1.4.6 with Ganglia 3.7.2, Hadoop 2.8.4, Hive 2.3.3, Hue 4.2.0, Phoenix 4.14.0, and ZooKeeper 3.4.12
Presto: Presto 0.206 with Hadoop 2.8.4 HDFS and Hive 2.3.3 Metastore
Spark: Spark 2.3.1 on Hadoop 2.8.4 YARN with
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all
Type here to search
Desktop
ENG
23:22
20-09-2018


EMR – AWS Console
Inbox (1,076) - samiksha20167
after termination cluster restar
Step 5: Terminate the Cluster a
restart terminated emr cluster
https://ap-south-1.console.aws.amazon.com/elasticmapreduce/home?region=ap-south-1#create-cluster:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Software Configuration
Release
emr-5.17.0
Hadoop 2.8.4
JupyterHub 0.8.1
Ganglia 3.7.2
Hive 2.3.3
MXNet 1.2.0
Hue 4.2.0
Spark 2.3.1
Zeppelin 0.7.3
Tez 0.8.4
HBase 1.4.6
Presto 0.206
Sqoop 1.4.7
Phoenix 4.14.0
HCatalog 2.3.3
Livy 0.5.0
Flink 1.5.2
Pig 0.17.0
ZooKeeper 3.4.12
Mahout 0.13.0
Oozie 5.0.0
TensorFlow 1.9.0
AWS Glue Data Catalog settings (optional)
Use for Hive table metadata
Use for Spark table metadata
HBase storage settings
Storage Mode
HDFS
S3
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all
Type here to search
Desktop
ENG
23:22
20-09-2018

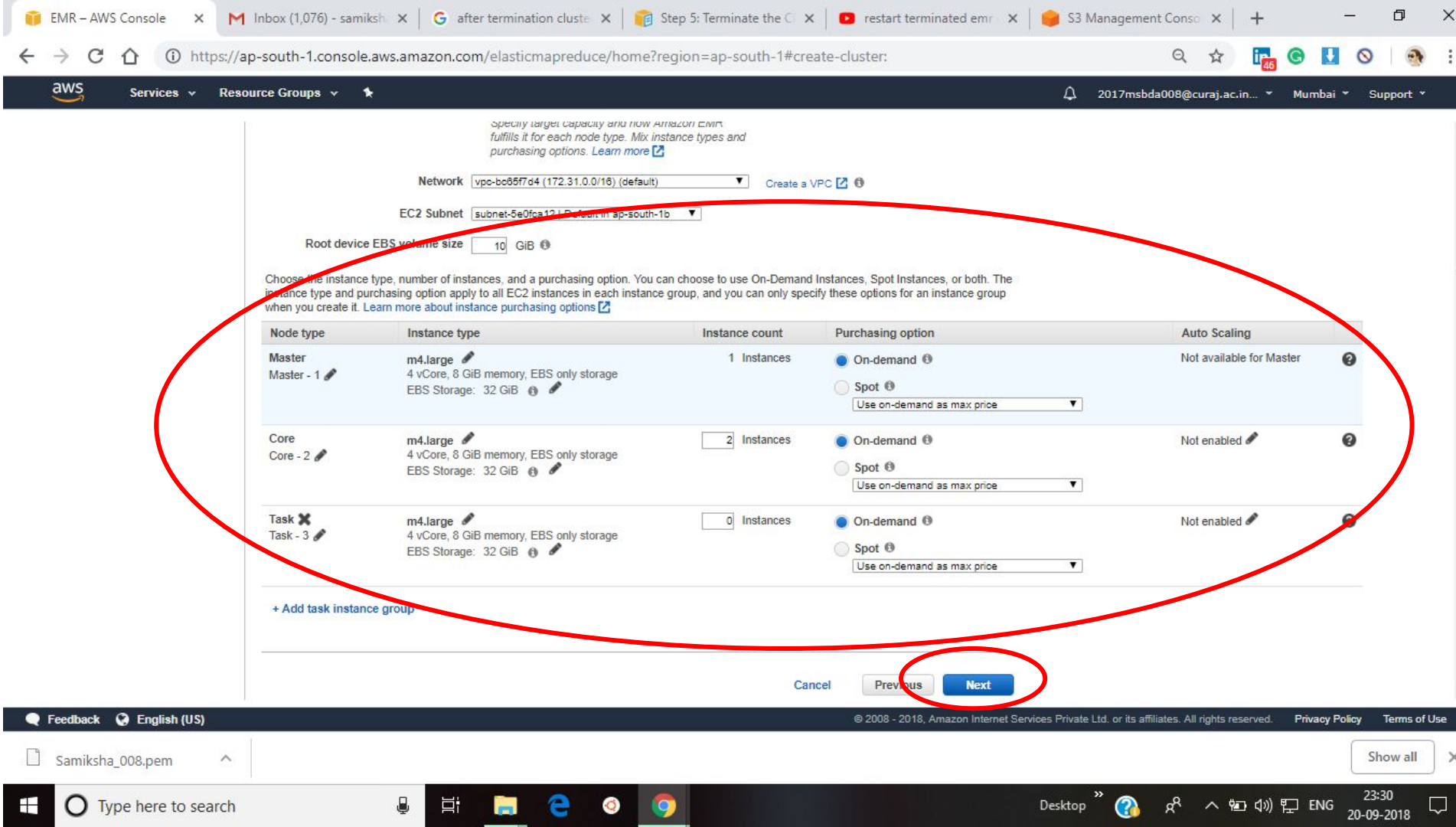

EMR – AWS Console
Inbox (1,076) - samiksh
after termination cluste
Step 5: Terminate the C
restart terminated emr
S3 Management Conso
https://ap-south-1.console.aws.amazon.com/elasticmapreduce/home?region=ap-south-1#create-cluster:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Use for Hive table metadata
Use for Spark table metadata
HBase storage settings
Storage Mode
HDFS
S3
Edit software settings
Enter configuration
Load JSON from S3
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
Add steps (optional)
Step type
Select a step
Configure
Auto-te
Select a step
Streaming program
Hive program
Pig program
Spark application
Custom JAR
Cancel
Next
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all
Type here to search
Desktop
ENG
23:24
20-09-2018

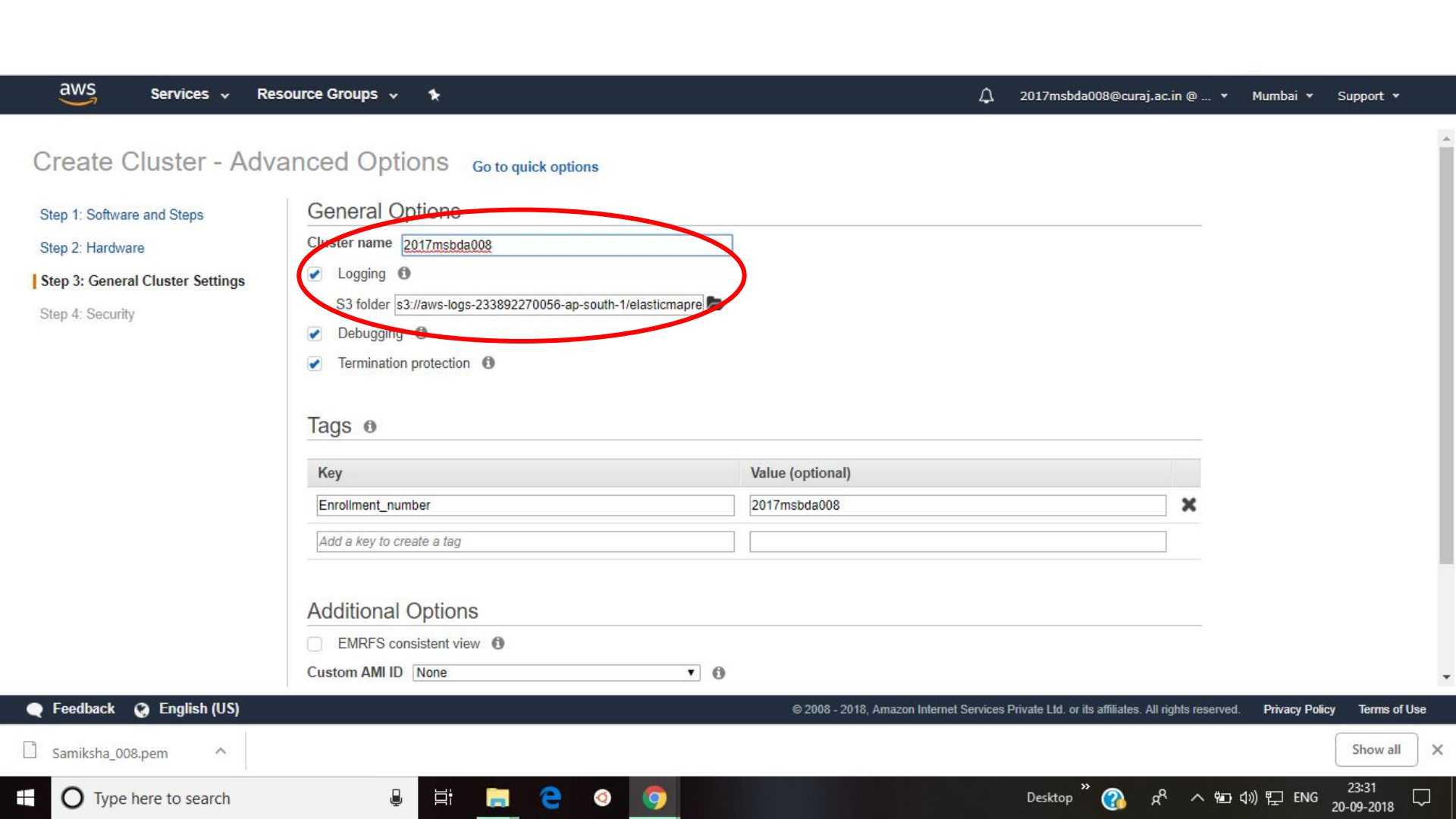

EMR – AWS Console
Inbox (1,076) - samiksh
after termination cluste
Step 5: Terminate the C
restart terminated emr
S3 Management Conso
https://ap-south-1.console.aws.amazon.com/elasticmapreduce/home?region=ap-south-1#create-cluster:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Use for Hive table metadata
Use for Spark table metadata
HBase storage settings
Storage Mode
HDFS
S3
Edit software settings
Enter configuration
Load JSON from S3
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
Add steps (optional)
Step type
Streaming program
Configure
Auto-terminate cluster after the last step is completed
Cancel
Next
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all
Type here to search
Desktop
ENG
23:24
20-09-2018


EMR – AWS Console
Inbox (1,076) - samiksh
after termination cluste
Step 5: Terminate the C
restart terminated emr
S3 Management Conso
https://ap-south-1.console.aws.amazon.com/elasticmapreduce/home?region=ap-south-1#create-cluster:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Add step
Step type
Streaming program
Name*
WordCount
Mapper*
s3://2017msbda008/WordCount/mapper.py
S3 location of the map function or the name of the Hadoop streaming command to run.
Reducer*
s3://2017msbda008/WordCount/reducer.py
S3 location of the reduce function or the name of the Hadoop streaming command to run.
Input S3 location*
s3://2017msbda008/WordCount/sample_text.txt
s3://<bucket-name>/<folder>/
Output S3 location*
s3://2017msbda008/WordCount/result
s3://<bucket-name>/<folder>/
Arguments
Action on failure
Continue
What to do if the step fails.
Cancel
Add
Next
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all
Type here to search
Desktop
ENG
23:26
20-09-2018



As a notification you will see in Steps part that you have one step called Word count


EMR – AWS Console
Inbox (1,076) - samiksh
after termination cluste
Step 5: Terminate the C
restart terminated emr
S3 Management Conso
https://ap-south-1.console.aws.amazon.com/elasticmapreduce/home?region=ap-south-1#create-cluster:
aws
Services
Resource Groups
2017msbda008@curaj.ac.in...
Mumbai
Support
fulfills it for each node type. Mix instance types and purchasing options. Learn more
Network
vpc-bc65f7d4 (172.31.0.0/16) (default)
Create a VPC
EC2 Subnet
subnet-5e0fca12 | Default in ap-south-1b
Root device EBS volume size
10
GiB
Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. Learn more about instance purchasing options
Node type
Instance type
Instance count
Purchasing option
Auto Scaling
Master
Master - 1
m4.large
4 vCore, 8 GiB memory, EBS only storage
EBS Storage: 32 GiB
1 Instances
On-demand
Spot
Use on-demand as max price
Not available for Master
Core
Core - 2
m4.large
4 vCore, 8 GiB memory, EBS only storage
EBS Storage: 32 GiB
2 Instances
On-demand
Spot
Use on-demand as max price
Not enabled
Task
Task - 3
m4.large
4 vCore, 8 GiB memory, EBS only storage
EBS Storage: 32 GiB
0 Instances
On-demand
Spot
Use on-demand as max price
Not enabled
+ Add task instance group
Cancel
Previous
Next
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all
Type here to search
Desktop
ENG
23:30
20-09-2018


aws
Services
Resource Groups
2017msbda008@curaj.ac.in @ ...
Mumbai
Support
Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
General Options
Cluster name
2017msbda008
Logging
S3 folder
s3://aws-logs-233892270056-ap-south-1/elasticmapre
Debugging
Termination protection
Tags
Key
Value (optional)
Enrollment_number
2017msbda008
Add a key to create a tag
Additional Options
EMRFS consistent view
Custom AMI ID
None
Feedback
English (US)
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Samiksha_008.pem
Show all
Type here to search
Desktop
ENG
23:31
20-09-2018

Step 3: General Cluster Settings
Step 4: Security

Logging
S3 folder s3://aws-logs-233892270056-ap-south-1/elasticmapre
Debugging
Termination protection

Tags

Key	Value (optional)	
Enrollment_number	2017msbda008	✖
Add a key to create a tag		

Additional Options

EMRFS consistent view
Custom AMI ID None
Bootstrap Actions

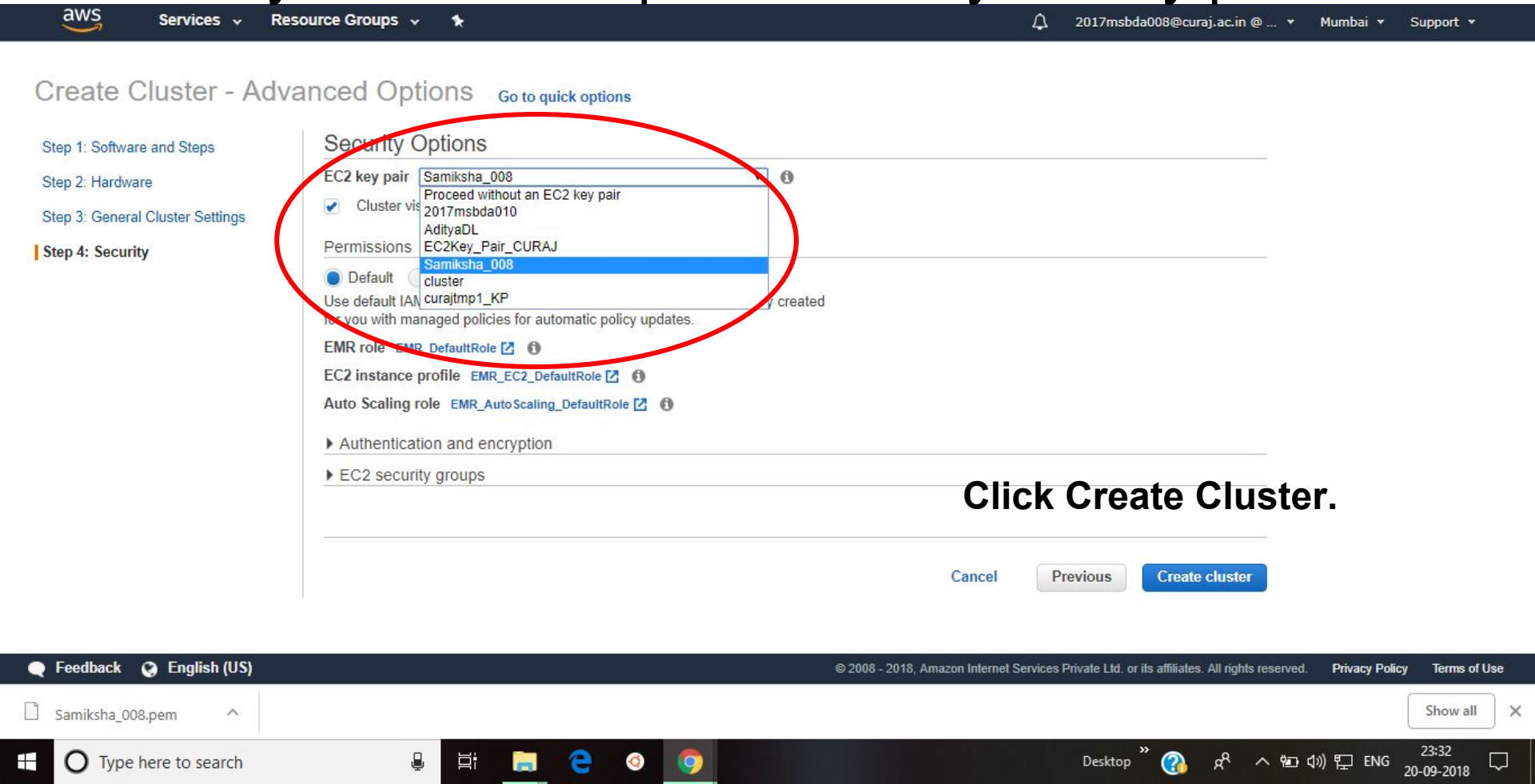
Cancel Previous Next

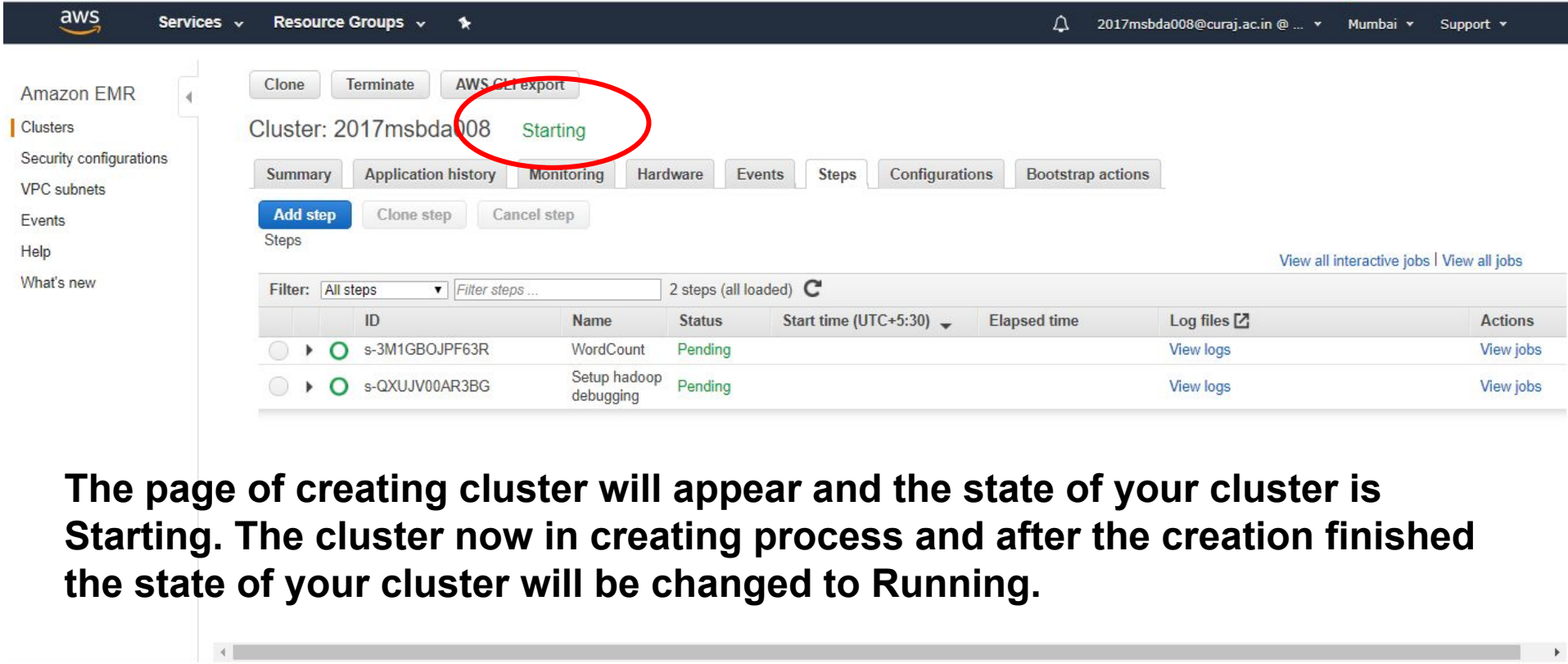
© 2008 - 2018, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved.

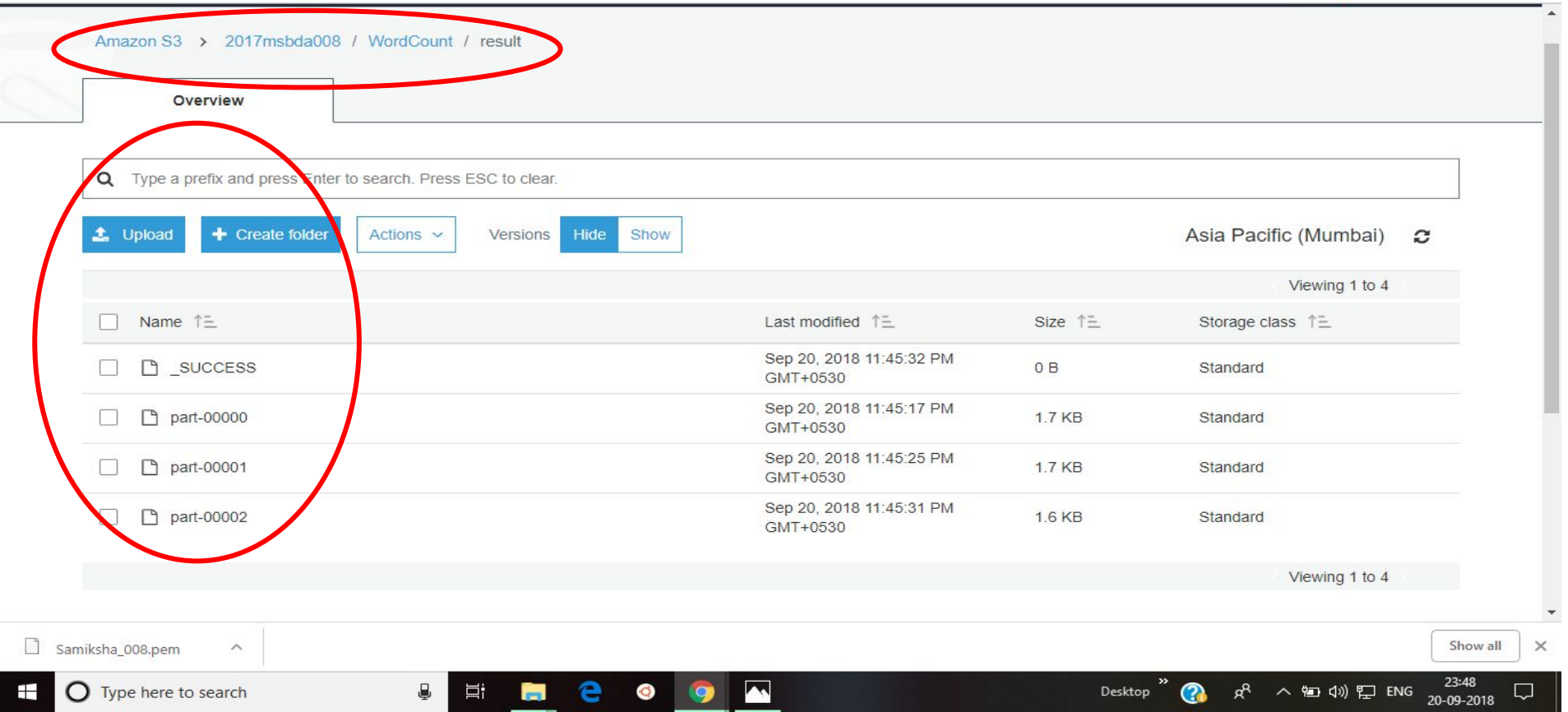
Samiksha_008.pem
Show all

Type here to search
Desktop ENG 23:31 20-09-2018

In Security and Access part choose your key pair





The page of creating cluster will appear and the state of your cluster is Starting. The cluster now in creating process and after the creation finished the state of your cluster will be changed to Running.

Go to S3 console and choose the bucket ,you will find the result folder that contains the output, it's something like in figure below.