



**Silesian
University
of Technology**

Analyze Big Data with Hadoop
Hadoop Ecosystem Lab 3

Authors:

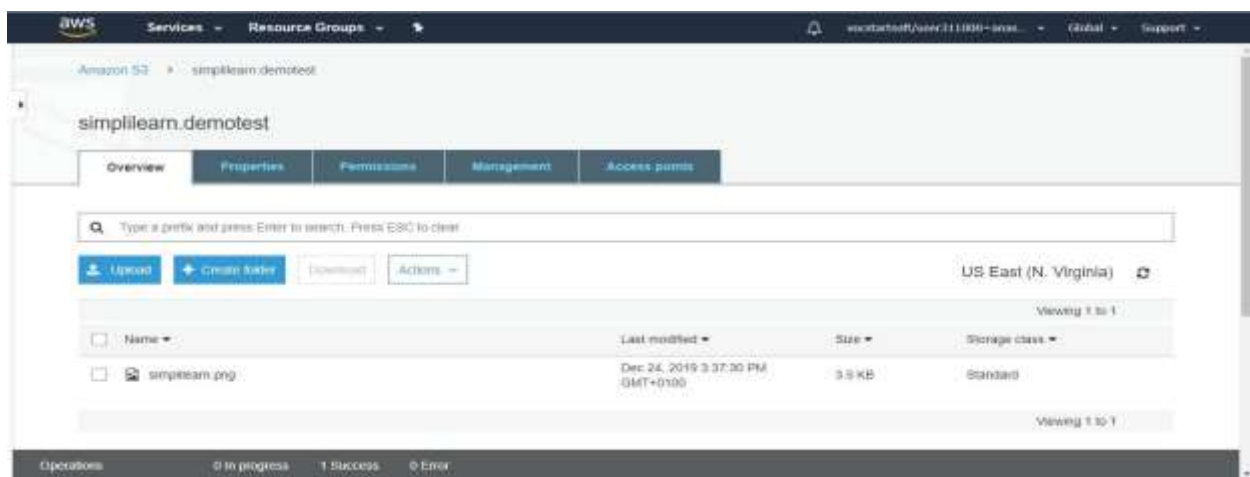
Ankit Rathi

Jean Remy Uwacu

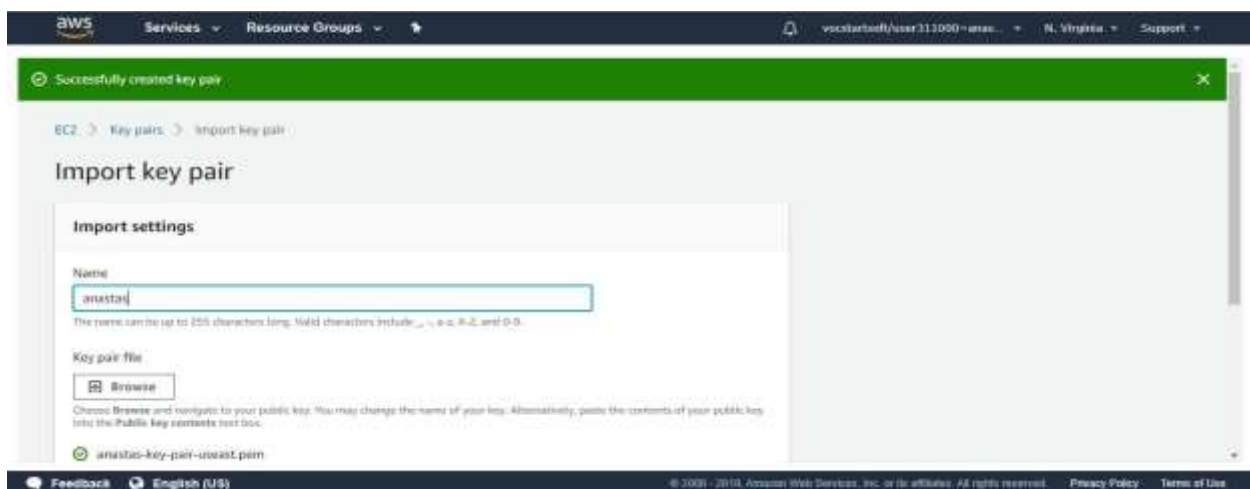
Step 1: Set Up Prerequisites for Sample Cluster

Create an Amazon S3 Bucket

We need to create a Amazon s3 bucket and a folder inside it –



The next step that we need to perform is to create a key pair.



Step 2: Launch Your Sample Amazon EMR Cluster

In this step we need to create a cluster and for this we will need the key pair that we created before, here are the steps –

In this step, we launch our sample cluster by using Quick Options in the Amazon EMR console and leaving most options to their default values.

The screenshot shows the 'Create cluster' wizard in the Amazon EMR console. The 'Hardware configuration' section shows the instance type set to 'm5.xlarge' and the number of instances set to 3 (1 master and 2 core nodes). The 'Security and access' section shows the EC2 key pair set to 'anastas-key-pair-ec2', permissions set to 'Default', and the EMR role set to 'EMR_DefaultRole'. The 'EC2 instance profile' is set to 'EMR_EC2_DefaultRole'. The 'Create cluster' button is visible at the bottom right.

Step 3: Allow SSH Connections to the Cluster From Your Client

In this step we will have to modify the access rights in order to secure the cluster. And then after this we will ssh connection only from our computer.

The screenshot shows the 'Create Security Group' wizard in the Amazon EC2 console. The 'Create Security Group' button is visible at the top. Below it, a table lists the existing security groups:




Name	Group ID	Group Name	VPC ID	Owner	Description
sg-041315dc2d4401996	sg-041315dc2d4401996	ElasticMapReduce-master	vpc-4a7e1930	956468776634	Master group for Elastic M
sg-094c62cd5aa5b45	sg-094c62cd5aa5b45	ElasticMapReduce-slave	vpc-4a7e1930	956468776634	Slave group for Elastic Ma

Below the table, the 'Security Group: sg-041315dc2d4401996' is shown. The 'Inbound' tab is selected, and a table lists the inbound rules:

Type	Protocol	Port Range	Source	Description
All TCP	TCP	0 - 65535	sg-041315dc2d4401996 (Elastic	
All TCP	TCP	0 - 65535	sg-094c62cd5aa5b45 (Elastic	

Step 4: Process Data by Running the Hive Script as a Step

We will Now use Hive script to add data into our cluster, and to perform this we need to add new step in our cluster with given parameter—

Step type	<input type="text" value="Hive program"/>	
Name	<input type="text" value="Hive program"/>	
Script S3 location*	<input type="text" value="s3://us-west-2.elasticmapreduce.samples/cloudfront/o"/> <small>s3://<bucket-name>/<path-to-file></small>	 S3 location of your Hive script.
Input S3 location	<input type="text" value="s3://us-west-2.elasticmapreduce.samples"/> <small>s3://<bucket-name>/<folder>/</small>	 S3 location of your Hive input files.
Output S3 location	<input type="text" value="s3://mybucketbenlab2/MyHiveQueryResults/"/> <small>s3://<bucket-name>/<folder>/</small>	 S3 location of your Hive output files.
Arguments	<div></div>	Specify optional arguments for your script.
Action on failure	<input type="text" value="Continue"/>	What happens if the step fails

We will run HiveScript and after that we will download myhivequery, than when we come back to our bucket we can find the following data –

Android 855

Linux 833

Macos 852

Osx 799

Windows 883

Ios 794

So the previous operation works now we have added the data.

Step 5: Terminate the Cluster and Delete the Bucket

Now that we have done the demonstration, we can close the bucket and cluster.

