GMS Hadoop 1 Day Workshop

Welcome back!

Running Hadoop on AWS

Running Hadoop on AWS

Amazon EMR (Elastic MapReduce)

Managed Hadoop framework using AWS instances and storage (S3).

Pay-as-you-go

Running Hadoop on AWS

Pre-requisite:

AWS user account

Running Hadoop on AWS

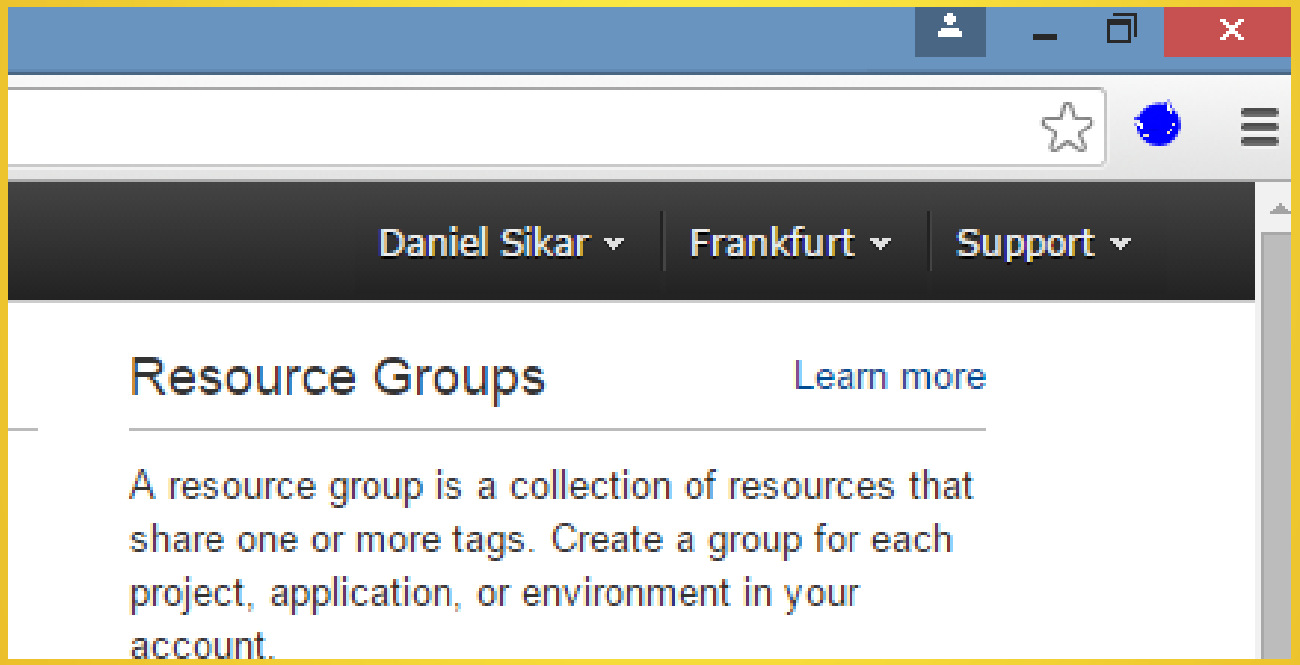
PRACTICE 1 - Sign in

https://aws.amazon.com/console/

Sign in to the Console

Running Hadoop on AWS

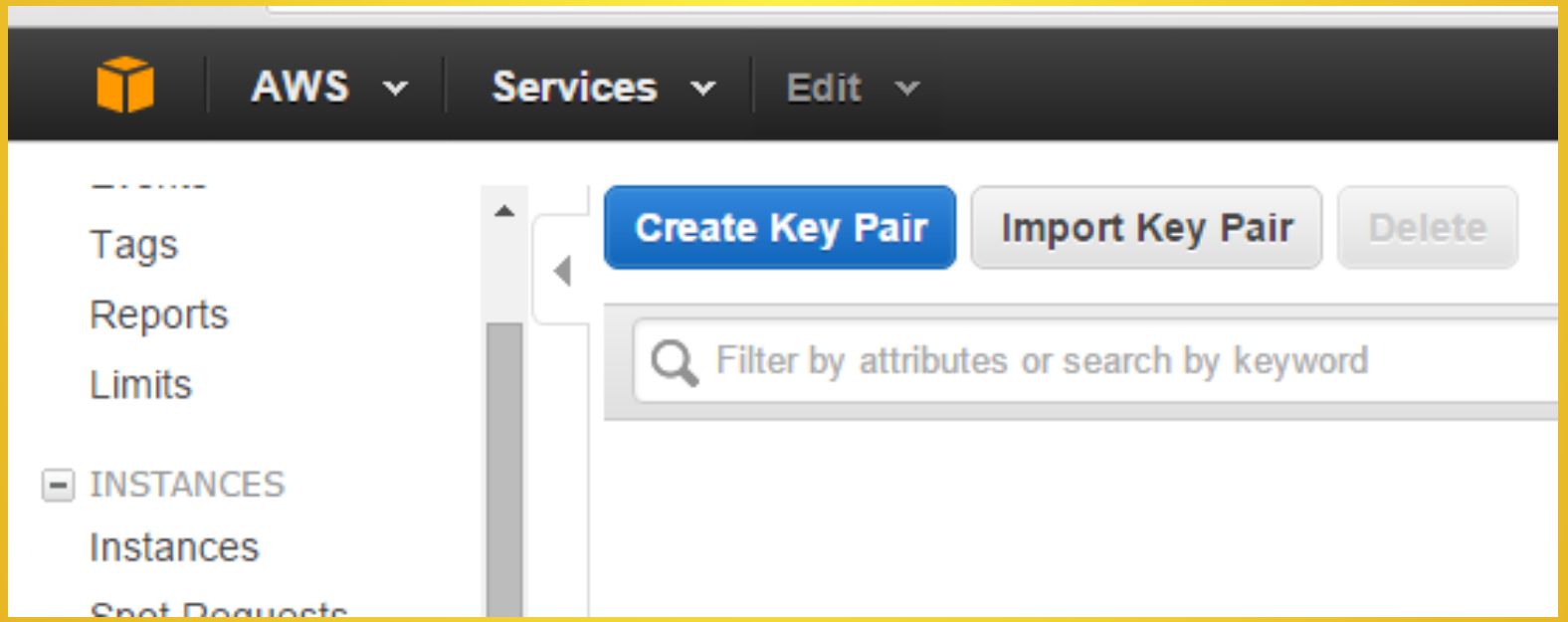
PRACTICE 2 - Change to nearest Region



Running Hadoop on AWS

PRACTICE 3 - Create Key Pair

Services > EC2 > Key Pairs



Running Hadoop on AWS

A Note About AWS Regions:

Resources are contained within Regions.

Change Regions and your recently created Key Pair will not be there.

PRACTICE 4 – Try it

Running Hadoop on AWS

A Note About Key Pairs:

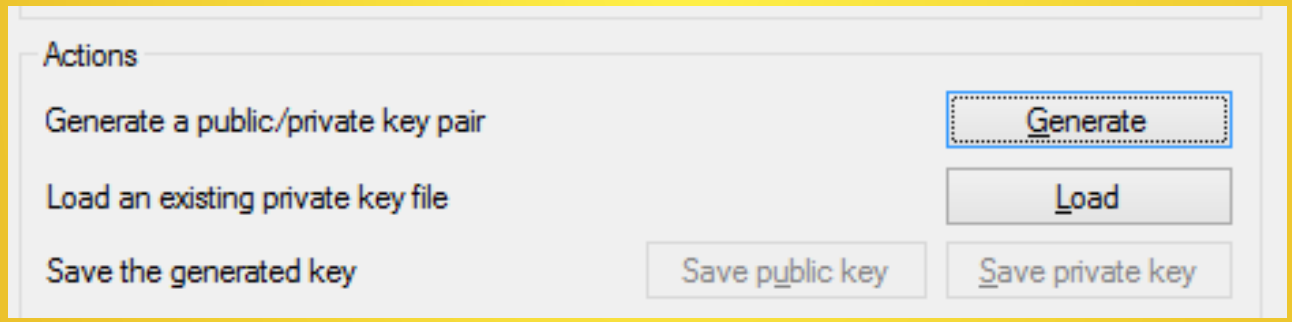
If accessing AWS Instances (Virtual Machines) through Linux SSH, GMSHadoopWorkshop.pem will do, otherwise create .ppk file with puttygen and access with putty.

More to follow...

Running Hadoop on AWS

PRACTICE 5 – Converting Key Pair

Using puttygen.exe, load the private key (Key Pair) file and “Save private key”



Running Hadoop on AWS

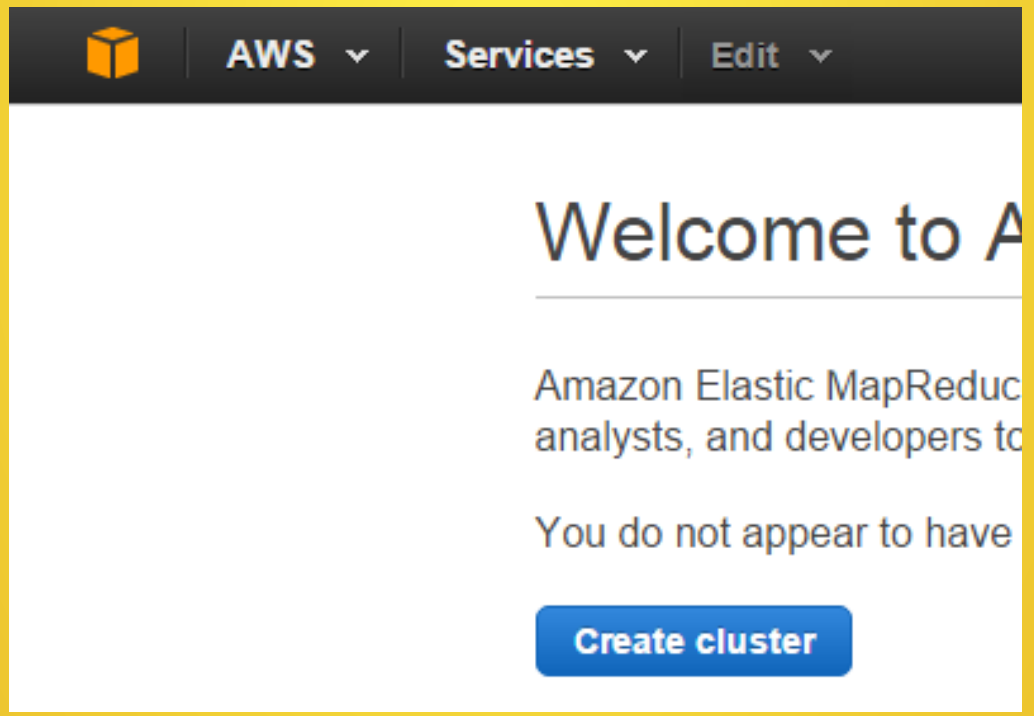
PRACTICE 6 – Save both .pem and ppk file to:

(…)\Documents\GSMHadoopWorkshop\.ssh

Running Hadoop on AWS

PRACTICE 7 – Create Hadoop Cluster

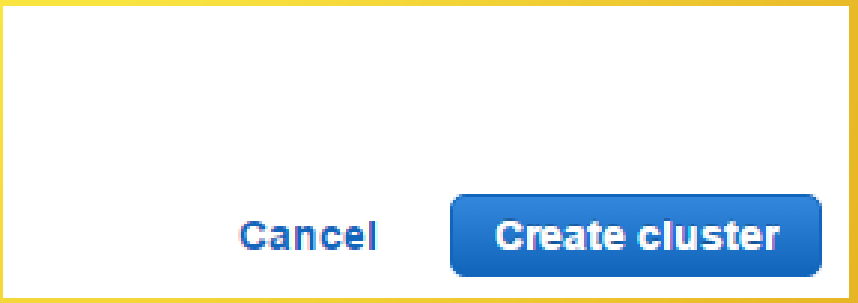
> Services > EMR > Create Cluster



Running Hadoop on AWS

PRACTICE 7 cont. – Configure Cluster

Cluster Name: <My Name> GMS Hadoop Cluster
Instance Type: m3.xlarge ($0.315 instance/hour)
EC2 key pair: GMSHadoopWorkshop



Running Hadoop on AWS

A Note About Pricing:
https://aws.amazon.com/ec2/pricing/

Cluster with 3 x m3.xlarge instances:
Aproximate hourly cost 3 x $0.315 ~ 1 USD
Annual Cost (1 USD X 24 X 365) ~ 8.7k USD

DON'T FORGET TO TERMINATE YOUR CLUSTER

Running Hadoop on AWS

PRACTICE 8 – SSH into Master Node

with putty (Auth ~ .ppk)

ssh

>ssh -i .ssh\<my key pair>.pem hadoop@<public dns>

Running Hadoop on AWS

PRACTICE 9 – SSH into Slave Node

Add SSH rule then SSH as per PRACTICE 8.

PRACTICE 10 – assert HDFS works
as per HDFS Practice

Running Hadoop on AWS

PRACTICE 11 – Terminate Cluster and delete S3 storage (bucket)

Services > S3 > Actions > Delete Bucket

Tip: Copy Bucket Name from Properties.