Big Data Infrastructure and Technologies

Practice 02 – Getting started with Hadoop

Creating and accessing AWS EMR Cluster

Installing and running AWS EMR cluster

- Prerequisite
- EMR cluster deployment steps

- Accessing Hadoop cluster with web browser
  - Configuring tunnel and Hadoop services

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html
- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html

Prerequisites

- Create an Amazon S3 Bucket
  - To be referred as *s3://mybucket/MyHiveQueryResults*
  - Can be also created automatically during EMR creation
- Create EC2 keypair
  - Create and store during VM instance creation
  - Store created keypair with remembered name
    - Example: keypair2020bdit4da.pem => for Windows convert to .ppk with PuTTYgen
    - It will be prompted during EMR cluster creation

- For more information - https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html

What to consider before starting

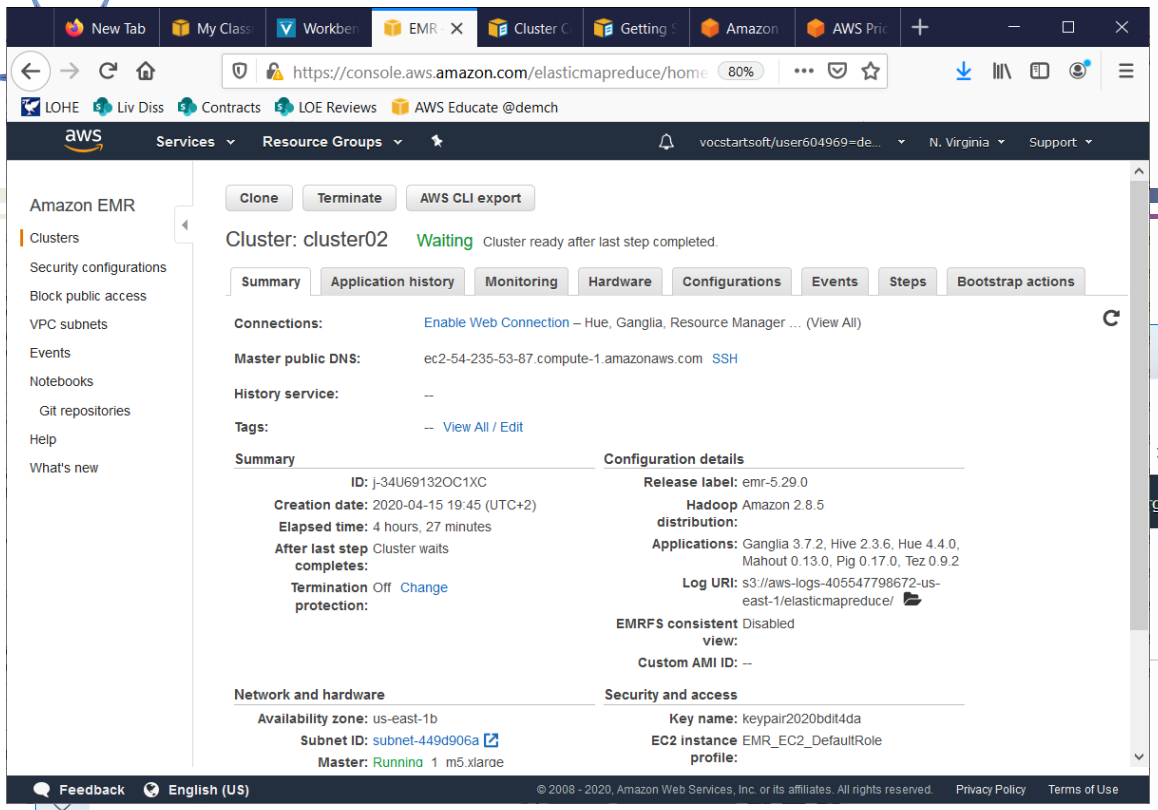
- Estimate Pricing for Amazon EMR and Amazon EC2 (On-Demand) - https://aws.amazon.com/emr/pricing/
  - The Amazon EMR price is in addition to the Amazon EC2 price (the price for the underlying servers) and Amazon EBS price (if attaching Amazon EBS volumes - billed per-second, with a one-minute minimum).
- **Save money with Reserved and Spot Instances**

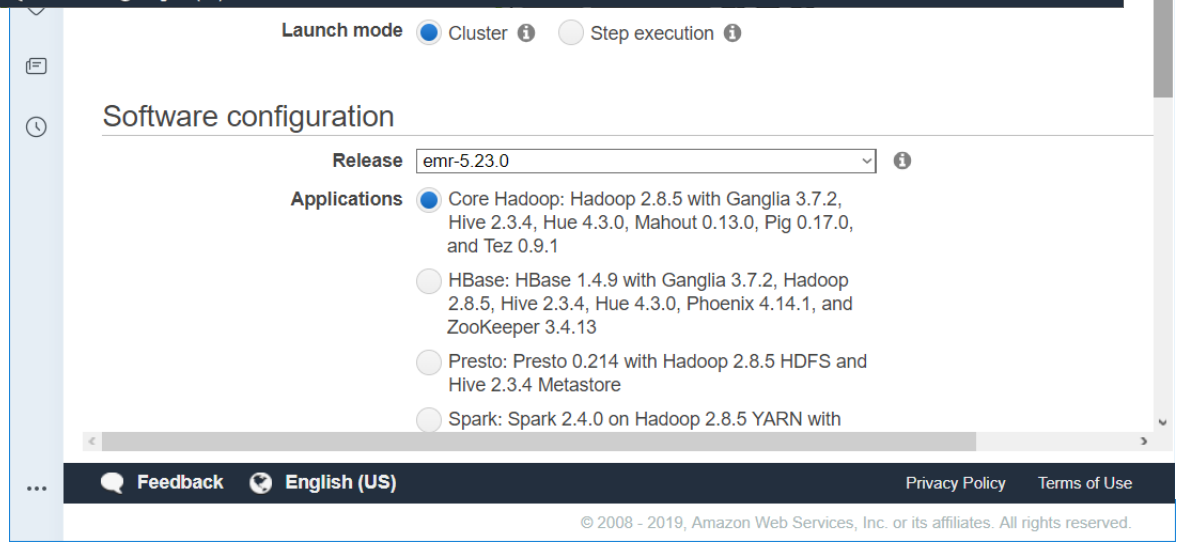
- Estimate your monthly bill using the AWS Pricing Calculator https://calculator.aws/
  - Helps only with EC2/S3/RDS services, some Management services
  - Old pricing calculator https://calculator.s3.amazonaws.com/index.html#s=EMR (available until June 2020)

EMR Price Comparison table (2020-04)

Instance type	Amazon EC2 Price	Amazon EMR Price
m5.xlarge	$0.192 per Hour	$0.048 per Hour
m5.2xlarge	$0.384 per Hour	$0.096 per Hour
m5a.2xlarge	$0.344 per Hour	$0.086 per Hour
m4.large	$0.10 per Hour	$0.03 per Hour
m4.xlarge	$0.20 per Hour	$0.06 per Hour
m4.2xlarge	$0.40 per Hour	$0.12 per Hour
c5.xlarge	$0.17 per Hour	$0.043 per Hour
c4.large	$0.10 per Hour	$0.026 per Hour
c4.xlarge	$0.199 per Hour	$0.052 per Hour

EMR cluster deployment steps

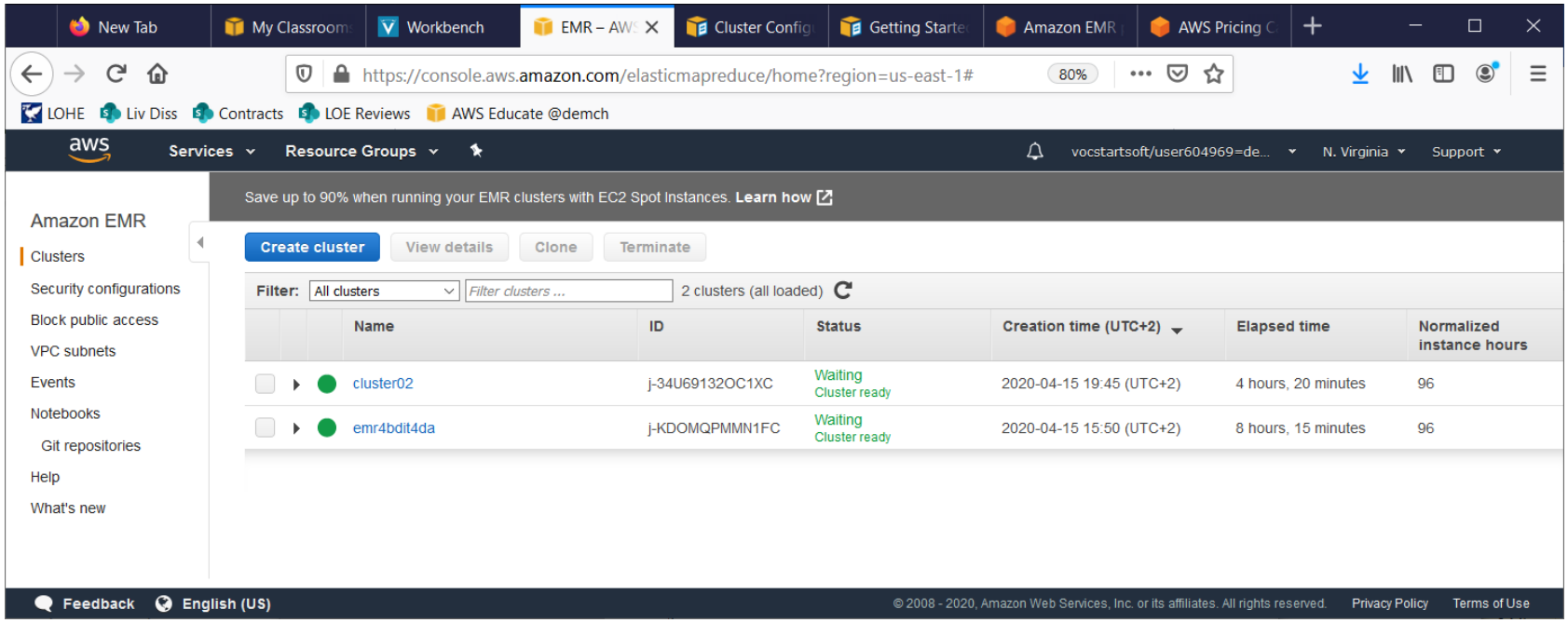




- Follow steps “Create Cluster” on Amazon EMR console
- Creation takes time of 10+ min
- Depends on the size of VM instance
- Check View Details for access details

- Terminate cluster after finishing practice tasks
- Use Clone to restore previously used configurations
  - Note: All your tunnels and ports configuration will remain

Actual Cluster compute cost by total cores



- Note: Your charge time is calculated based on total number of cores used – See Normalised instance hours

Cluster configuration and Connection URL

New Tab | My Class | Workben | EMR | Cluster C | Getting S | Amazon | AWS Pric

https://console.aws.amazon.com/elasticmapreduce/home 80%

LOHE | Liv Diss | Contracts | LOE Reviews | AWS Educate @demch

aws | Services | Resource Groups | vocstartsoft/user604969=de... | N. Virginia | Support

Amazon EMR

- Clusters
- Security configurations
- Block public access
- VPC subnets
- Events
- Notebooks
- Git repositories
- Help
- What's new

Clone | Terminate | AWS CLI export

Cluster: cluster02 Waiting Cluster ready after last step completed.

Summary | Application history | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Connections: Enable Web Connection – Hue, Ganglia, Resource Manager ... (View All)

Master public DNS: ec2-54-235-53-87.compute-1.amazonaws.com SSH

History service: --

Tags: -- View All / Edit

Summary

ID: j-34U69132OC1XC
Creation date: 2020-04-15 19:45 (UTC+2)
Elapsed time: 4 hours, 27 minutes
After last step completes: Cluster waits
Termination protection: Off Change

Configuration details

Release label: emr-5.29.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-405547798672-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-east-1b
Subnet ID: subnet-449d906a
Master: Running 1 m5.xlarge

Security and access

Key name: keypair2020bdit4da
EC2 instance profile: EMR_EC2_DefaultRole

Feedback | English (US) | © 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved. | Privacy Policy | Terms of Use

View web interfaces to EMR cluster

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-web-interfaces.html

- Option 1: Set Up an SSH Tunnel to the Master Node Using Local Port Forwarding
  - Use tunneled port http://localhost:8157/

- Option 2, Part 1: Set Up an SSH Tunnel to the Master node using Static and Dynamic Port Forwarding

- Option 2, Part 2: Configure Proxy Settings to view web interface hosted on the Master node

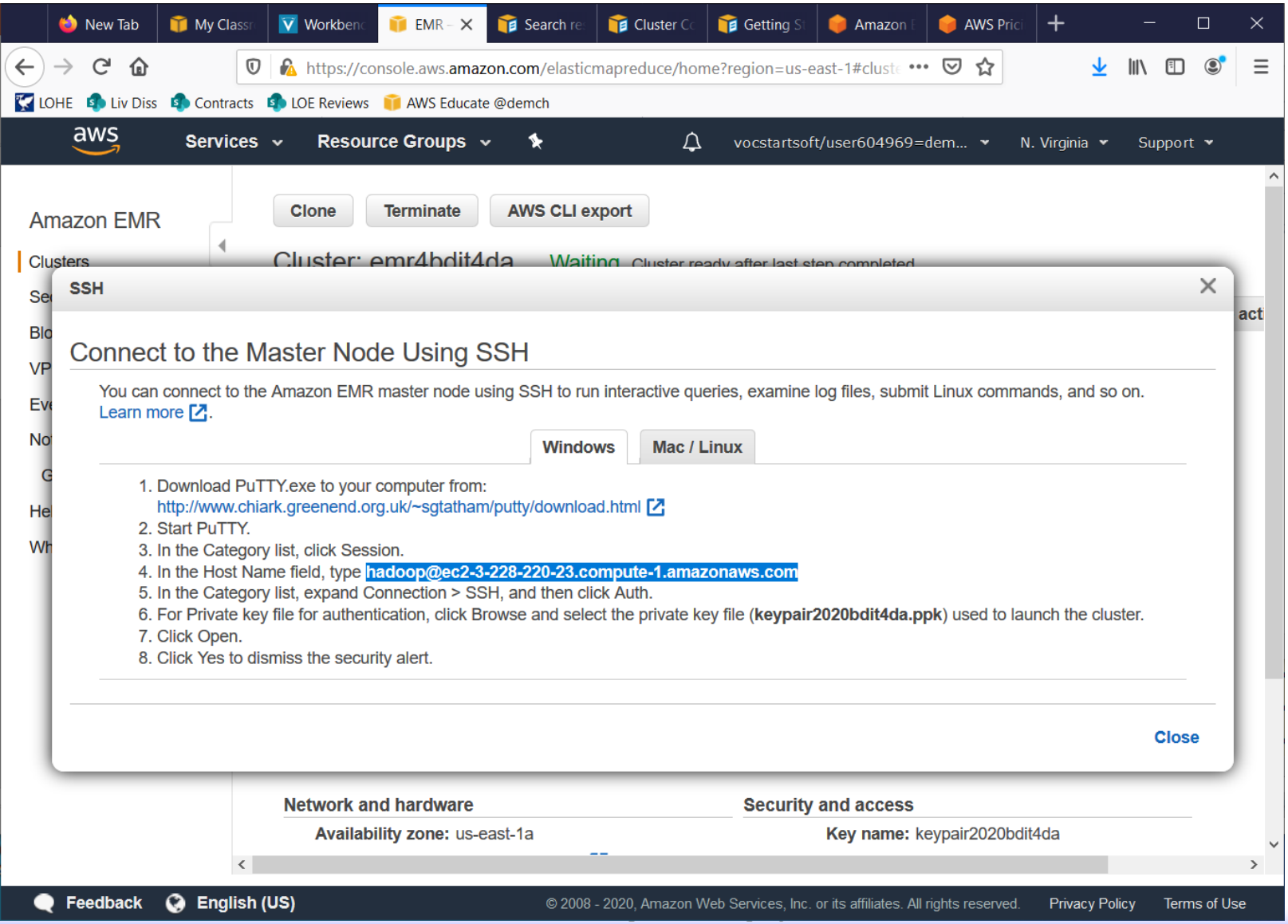
- Access the Web Interfaces on the Master Node using web browser

- NOTE: Check Master Node Security Group has open 22 port via Security group > Modify Rules

Important EMR ports and services

Name of interface	URI
Ganglia	http://master-public-dns-name/ganglia/
Hadoop HDFS NameNode	https://master-public-dns-name:50470/
Hadoop HDFS DataNode	https://coretask-public-dns-name:50475/
HBase	http://master-public-dns-name:16010/
Hue	http://master-public-dns-name:8888/
JupyterHub	https://master-public-dns-name:9443/
Livy	http://master-public-dns-name:8998/
Spark HistoryServer	http://master-public-dns-name:18080/
Tez	http://master-public-dns-name:8080/tez-ui
YARN NodeManager	http://coretask-public-dns-name:8042/
YARN ResourceManager	http://master-public-dns-name:8088/
Zeppelin	http://master-public-dns-name:8890/

Connect to the Master Node Using SSH

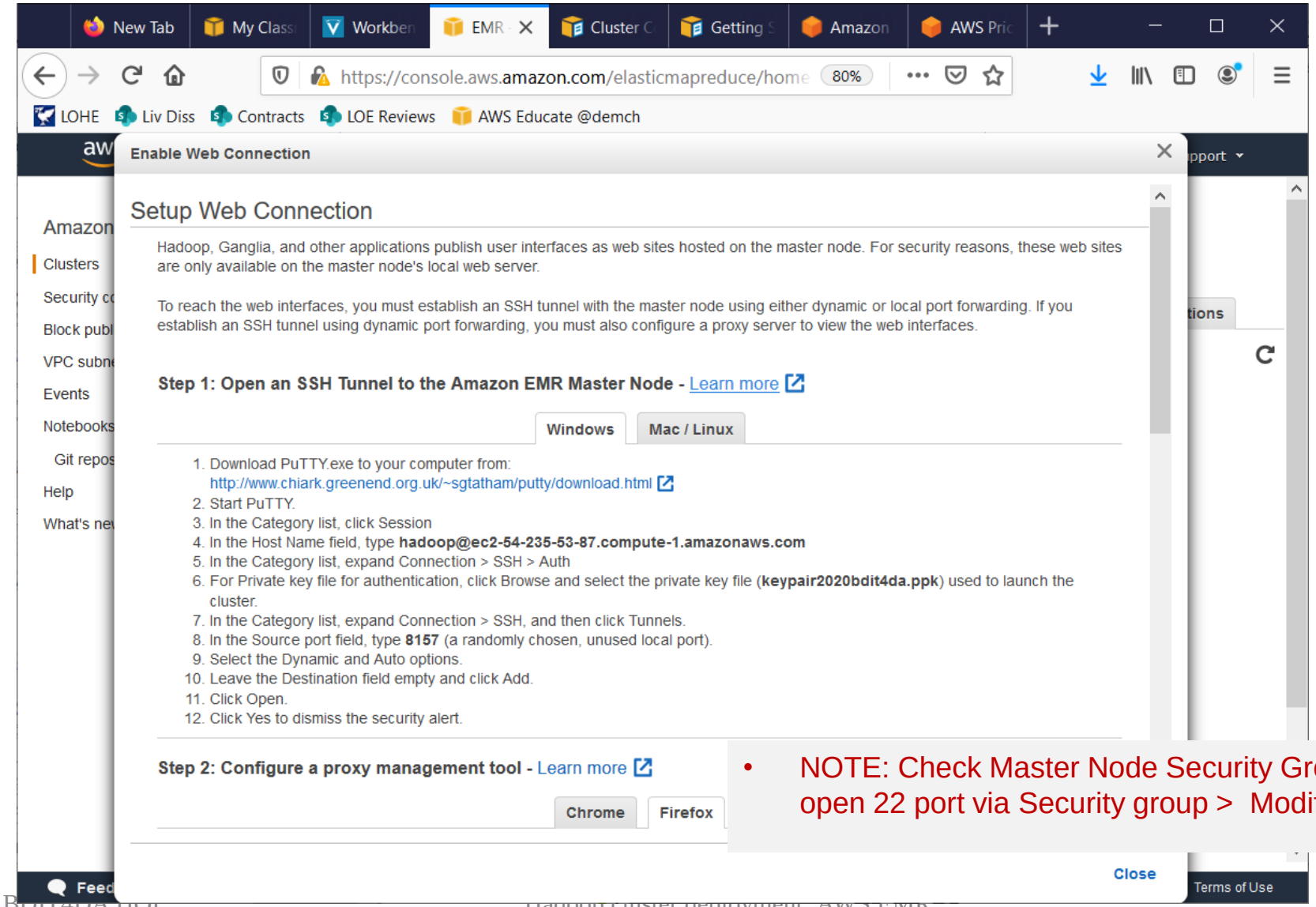


Connect to the Master Node Using SSH
(Instruction pops up when starting ER cluster)

1. Connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on.
2. Download PuTTY.exe to your computer from: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
3. Start PuTTY.
4. In the Category list, click Session.
5. In the Host Name field, type **hadoop@ec2-3-228-220-23.compute-1.amazonaws.com**
6. In the Category list, expand Connection > SSH, and then click Auth.
7. For Private key file for authentication, click Browse and select the private key file (**keypair2020bdit4da.ppk**) used to launch the cluster.
8. Click Open.
9. Click Yes to dismiss the security alert.

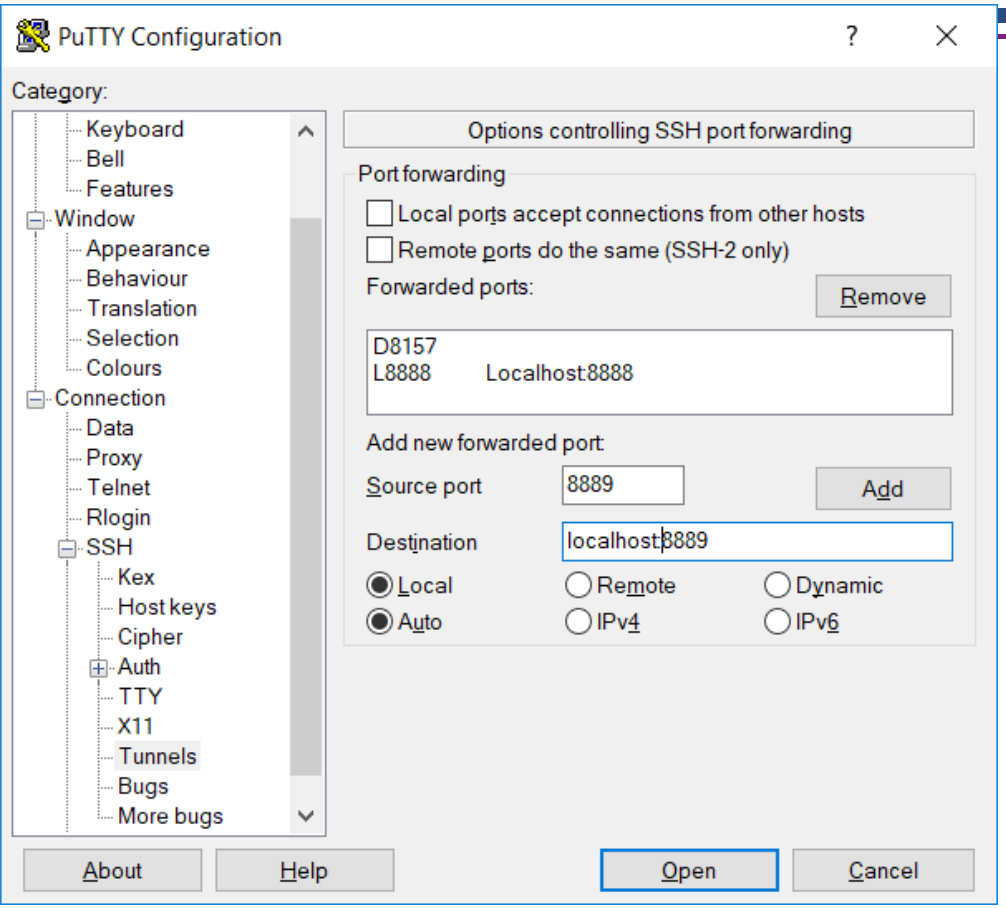
NOTE:
Check Master Node Security Group has open 22 port via Security group
If not, add port 22 forwarding (0.0.0.0/0) via Modify Rules

Enabling Web Connection (via Tunnel configuration)



- NOTE: Check Master Node Security Group has open 22 port via Security group > Modify Rules

Configuring Tunnels



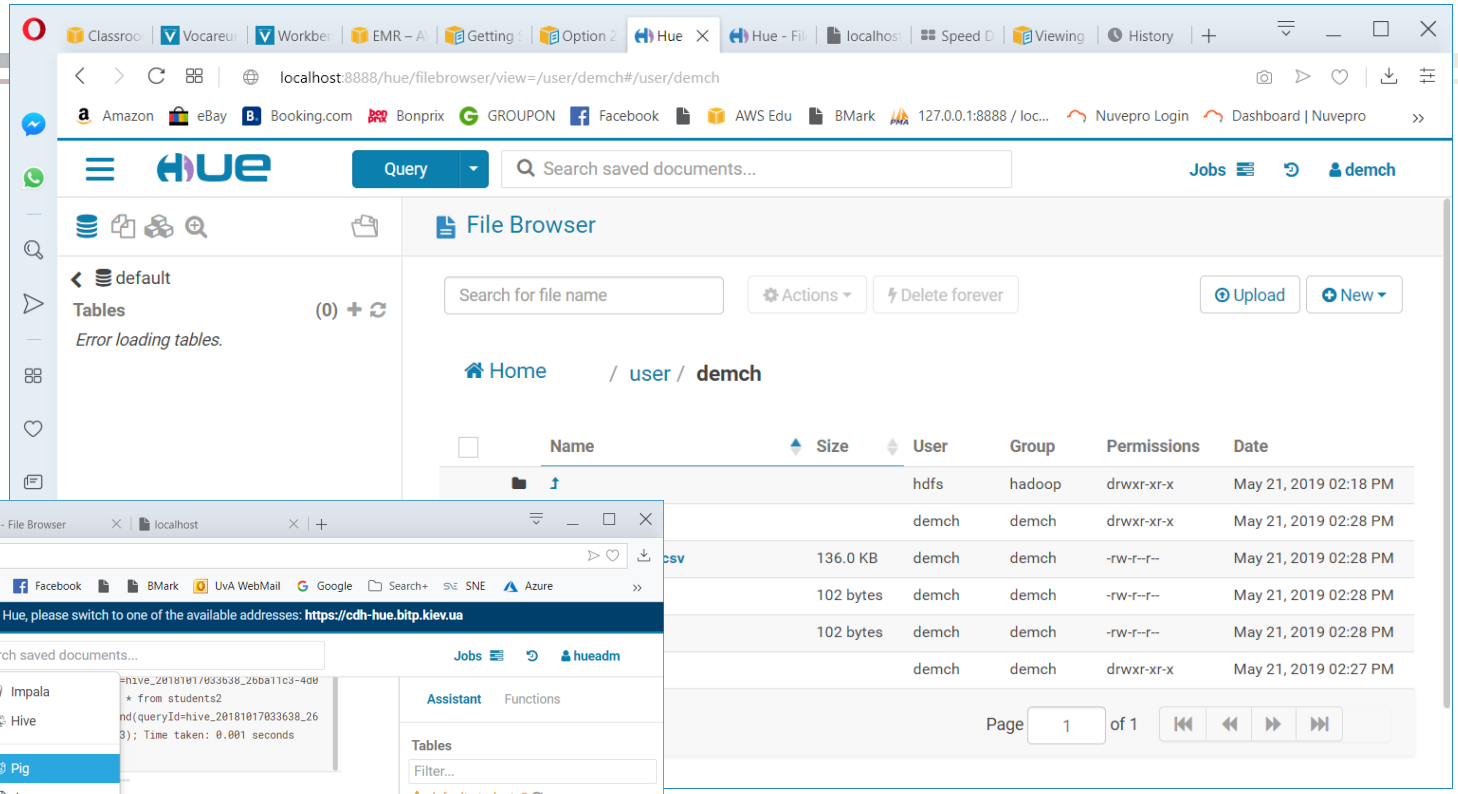
- Using PuTTY SSH client
- First, configure dynamic tunnel to EMR cluster
  - Port 8157 (or any random)
  - Using Source port and Dynamic radio button
- Configure other local ports
  - See EMR ports table

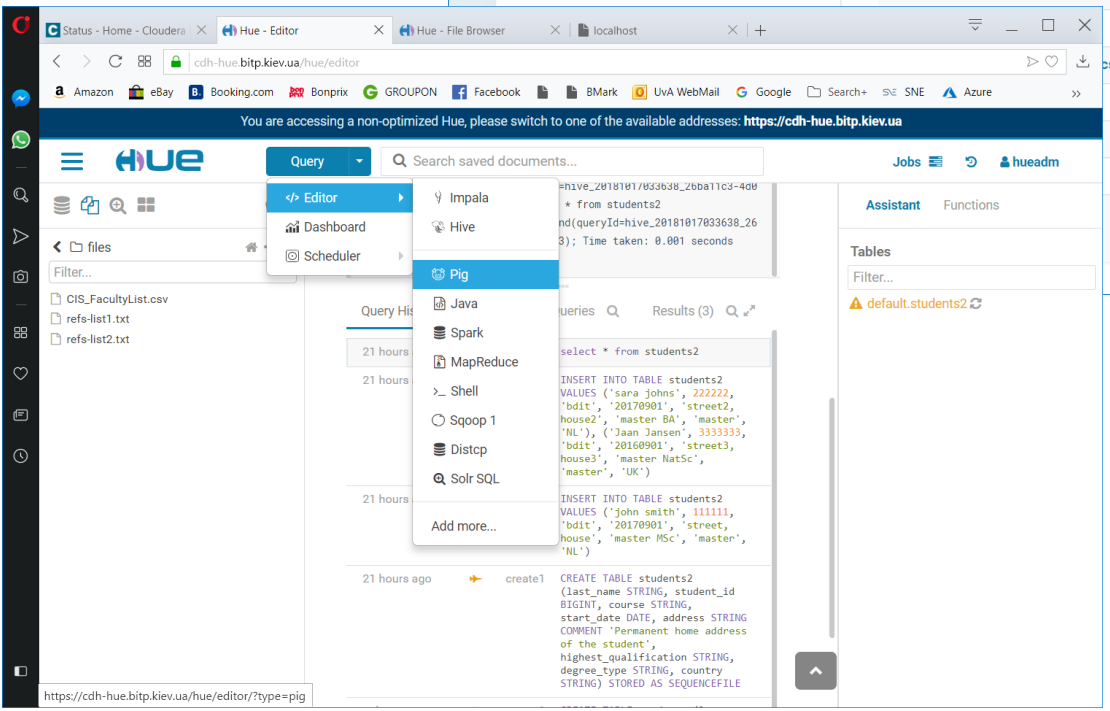
- This will create a secure tunnel by forwarding a port (the “destination port”) on the remote server to a port (the “source port”) on the local host (127.0.0.1 or localhost).

Connecting to Hue web interface http://*localhost*:8888/

- Use http://*localhost*:8888/ - or other port you assigned to browser access
  - Note: Because of security reasons using http://*master-public-dns-name*:8888/ is not allowed
- At first login assign username and password
  - It is recommended to use username “hadoop” (the same as SSH login) to have access to HDFS space, otherwise you will get message of type:
    - Cannot access: /user/cluster02/. Note: you are a Hue admin but not a HDFS superuser, "hdfs" or part of HDFS supergroup, "hadoop".
- Hue web interface invocation takes 2-3 min
  - Started with a simple guide
- Use Query dropdown menu to select Query>Editor Pig, Hive, Java, Scala, etc
- Use top-left dropdown menu to select:
  - Apps: Editor, Scheduler
  - Browser: Documents, Files, Tables, Indexes, Jobs
- Use Browser > Files to upload your datasets or script (works for S3 directrories)

Hue: File Browser/Directory and Editor view





- Normally, you can upload file/data via File Browser in Hue
- However the main way of upload data is via S3 bucket and File Browser in Hue

Supported features in EMR

https://docs.amazonaws.cn/en_us/emr/latest/ReleaseGuide/emr-hue-supported-features.html

- Amazon S3 and Hadoop File System (HDFS) Browser
  - With the appropriate permissions, you can browse and move data between the ephemeral HDFS storage and S3 buckets belonging to your account.
  - By default, superusers in Hue can access all files that Amazon EMR IAM roles are allowed to access. Newly created users do not automatically have permissions to access the Amazon S3 filebrowser and must have the filebrowser.s3_access permissions enabled for their group.
- Hive - Run interactive queries on your data. This is also a useful way to prototype programmatic or batched querying.
- Pig - Run scripts on your data or issue interactive commands.
- Oozie - Create and monitor Oozie workflows.
- Metastore Manager - View and manipulate the contents of the Hive metastore (import/create, drop, and so on).
- Job browser - See the status of your submitted Hadoop jobs.
- User management - Manage Hue user accounts and integrate LDAP users with Hue.
- AWS Samples - There are several "ready-to-run" examples that process sample data from various AWS services using applications in Hue. When you log in to Hue, you are taken to the
- Hue Home application where the samples are pre-installed.
- Livy Server is supported only in Amazon EMR version 5.9.0 and later.
- To use the Hue Notebook for Spark, you must install Hue with Livy and Spark.
- The Hue Dashboard is not supported.
- Files upload from browser is recently not supported. You need to do this via S3 bucket.

How to Get Data Into Amazon EMR

Amazon EMR provides several ways to get data onto a cluster.

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-get-data-in.html

- The most common way is to upload the data to Amazon S3 and use the built-in features of Amazon EMR to load the data onto your cluster. https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/hue_use_s3_source_sink.html
- Use the Distributed Cache feature of Hadoop to transfer files from a distributed file system to the local file system.
- The implementation of Hive provided by Amazon EMR (Hive version 0.7.1.1 and later) includes functionality that you can use to import and export data between DynamoDB and an Amazon EMR cluster.

Error messages and problems

- HUE File Browser interface:
  Cannot access: /user/cluster02/. Note: you are a Hue admin but not a HDFS superuser, "hdfs" or part of HDFS supergroup, "hadoop".
  - Hue superusers are different from hdfs superusers. With this error message, it complains that the admin and hdfs users you created through hue are not part to hdfs superuser.
    - You should be able to fix the issue by fixing the permission of the directories to the user/group on the cluster.
    - Simple solution assign/use username “hadoop” when accessing Hue
- **SSH connection: SSH connection timed out**
  - Check if cluster allows external access with port 22 (optionally HTTPS) and Master Node security group has open the same ports 22 and optionally 80 or 453
- Cannot connect to Hue via browser, site is not reachable
  - This is disable due to security reasons. Use SSH tunnel for this
- Hue interface is not responsive, doesn’t browse files, tables, etc.
  - Logout and login to refresh Hue state
-

Additional Materials