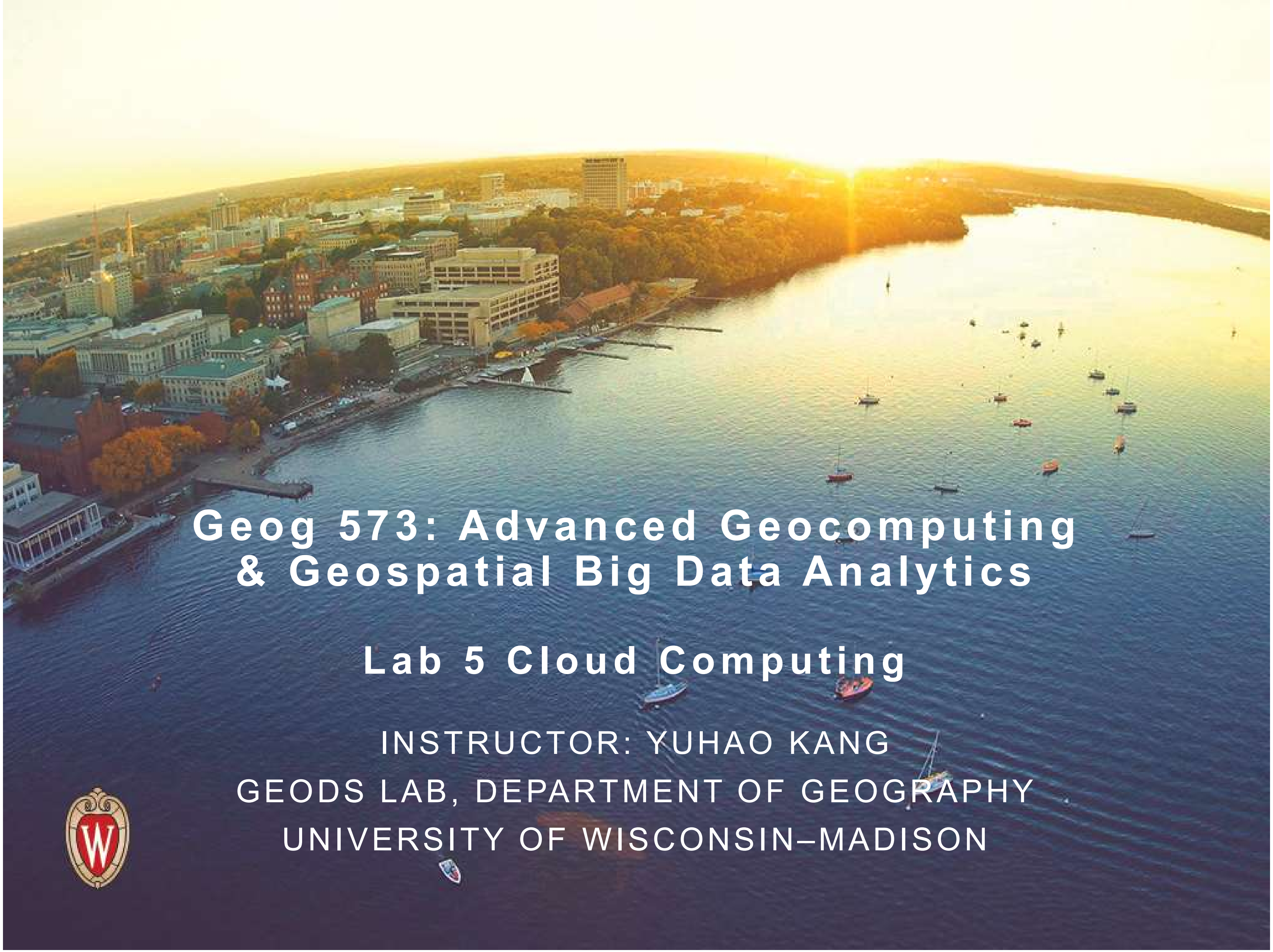


Geog 573: Advanced Geocomputing & Geospatial Big Data Analytics

Lab 5 Cloud Computing

INSTRUCTOR: YUHAO KANG

GEODS LAB, DEPARTMENT OF GEOGRAPHY

UNIVERSITY OF WISCONSIN–MADISON

Review and Notes

Web Scraping & Geocoding

- Understand HttpRequests (GET/POST).
- Geocoding.
- Collect house information from REDFIN.

How to Collect Photos?

- <div style=“url”>url</div>
- soup.find('div', class_=“")["style"]
- soup.find('div', class_=“").string

Shp (Shapefile)

The shp format is a geospatial vector data format for geographic information system (GIS) software.

It is developed and regulated by Esri as a mostly open specification for data interoperability among Esri and other GIS software products.

Shapefile shape format (.shp): the main file (.shp) contains the geometry data.

Shapefile shape index format (.shx): the index file.

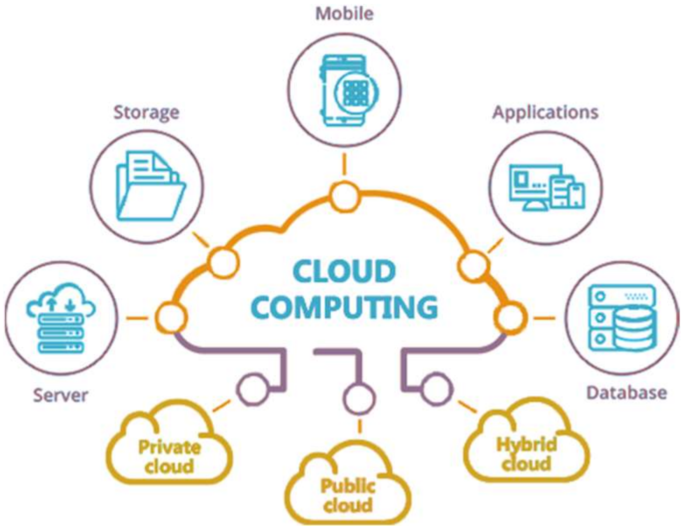
Shapefile attribute format (.dbf): this file stores the attributes for each shape.

Objectives

- **Cloud Computing**

- Launch a virtual machine and a cluster.
- Use AWS S3 to store data.
- Learn basic Linux operations.
- Run Hadoop on the virtual machine.

Why Cloud Computing?



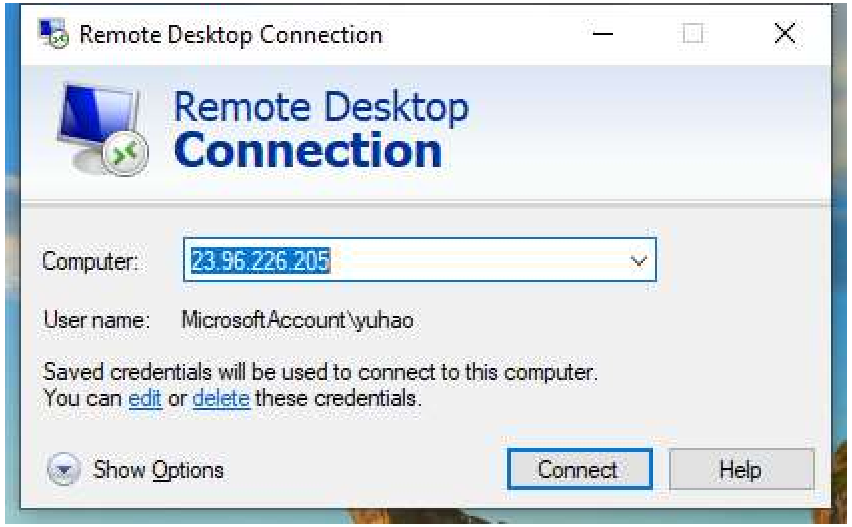
- With Cloud Computing, users can access database resources via the Internet from anywhere.
- Cloud Computing allows us to create, configure, and customize applications with high-performance servers.
- There is no need to worry about any maintenance or management of actual resources.

Cloud Server Examples

esri
GEOSPATIAL INTELLIGENCE
LANGUAGES
DATA PLATFORMS
GeoAI Data Science VM
ML & AI TOOLS
DATA INGESTION TOOLS
DATA EXPLORATION & VISUALIZATION

- GeoAI Data Science VM
- IP: 23.96.226.205
- Account: yuhao
- Password: Spring2020!!!

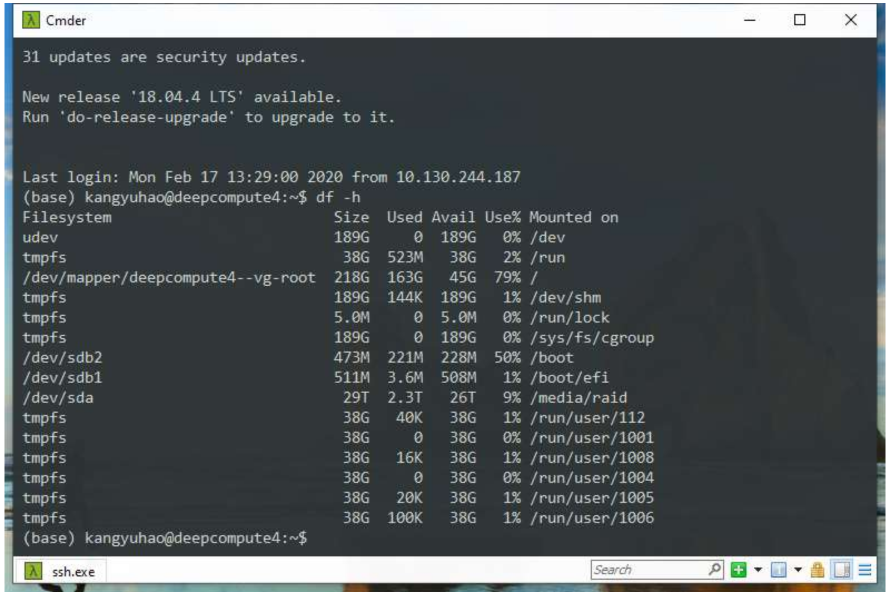
- Cloud Server of GeoDS Lab



Computer: 23.96.226.205

User name: MicrosoftAccount\yuhao

Saved credentials will be used to connect to this computer.
You can edit or delete these credentials.



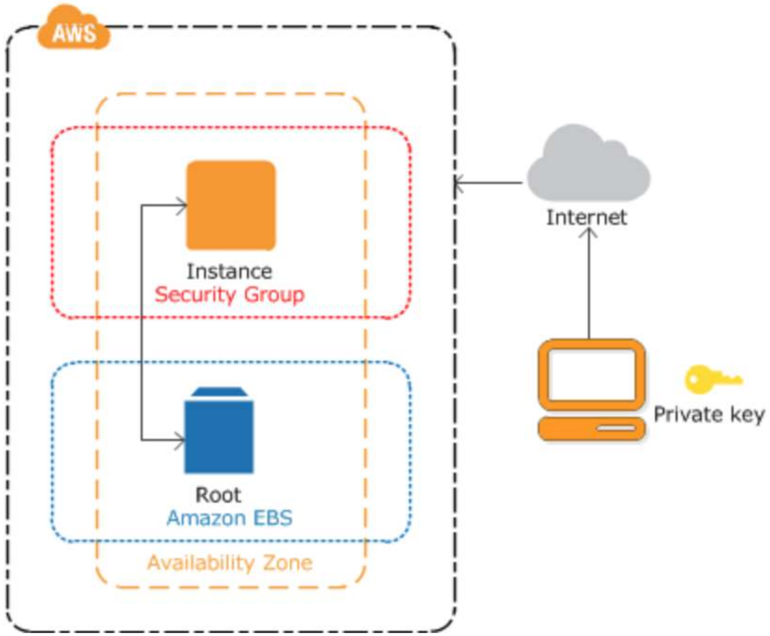
```
31 updates are security updates.

New release '18.04.4 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Feb 17 13:29:00 2020 from 10.130.244.187
(base) kangyuhao@deepcompute4:~$ df -h
Filesystem                         Size  Used Avail Use% Mounted on
udev                               189G     0  189G   0% /dev
tmpfs                               38G  523M   38G   2% /run
/dev/mapper/deepcompute4--vg-root  218G  163G   45G  79% /
tmpfs                              189G  144K  189G   1% /dev/shm
tmpfs                              5.0M     0  5.0M   0% /run/lock
tmpfs                              189G     0  189G   0% /sys/fs/cgroup
/dev/sdb2                          473M  221M  228M  50% /boot
/dev/sdb1                          511M  3.6M  508M   1% /boot/efi
/dev/sda                            29T  2.3T   26T   9% /media/raid
tmpfs                               38G   40K   38G   1% /run/user/112
tmpfs                               38G     0   38G   0% /run/user/1001
tmpfs                               38G   16K   38G   1% /run/user/1008
tmpfs                               38G     0   38G   0% /run/user/1004
tmpfs                               38G   20K   38G   1% /run/user/1005
tmpfs                               38G  100K   38G   1% /run/user/1006
(base) kangyuhao@deepcompute4:~$
```

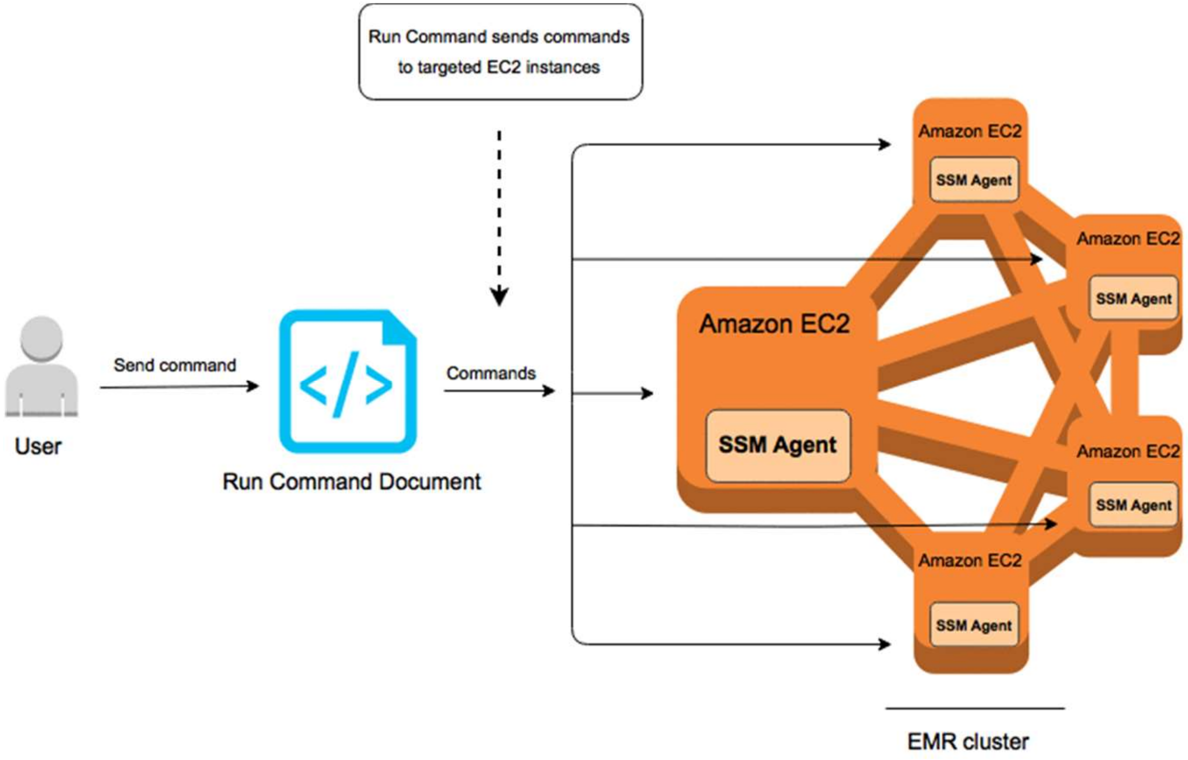
Virtual Machine

- ❑ AWS Educate Account with $50 free credits. https://aws.amazon.com/education/awseducate/?nc1=h_ls
- ❑ Amazon Elastic Compute Cloud (EC2) is the Amazon Web Service you use to create and run virtual machines in the cloud.



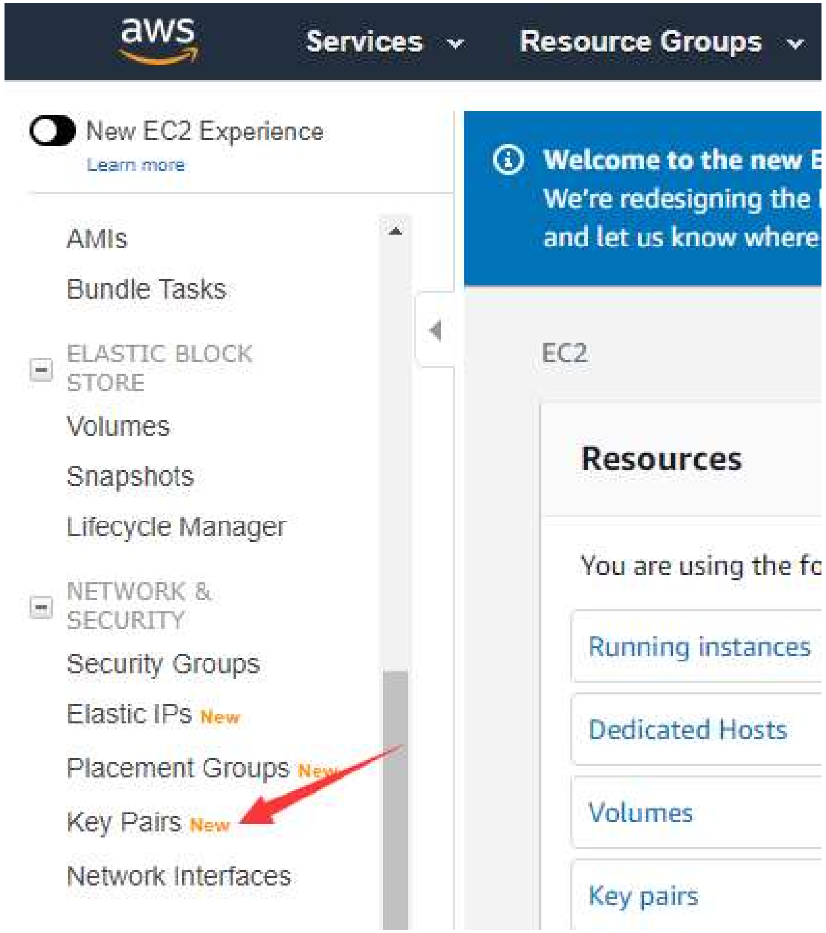
AWS EMR

- Amazon EMR is the industry leading cloud-native big data platform for processing vast amounts of data quickly and cost-effectively at scale. https://aws.amazon.com/emr/
- One master node + n slave nodes running on n+1 instances.



Launch an AWS EMR Cluster

- Create a new key pair
- EC2 -> Key Pairs -> Create keys.
  - For Linux/Mac: pem.
  - For Windows: ppk. (Download Putty: https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html)



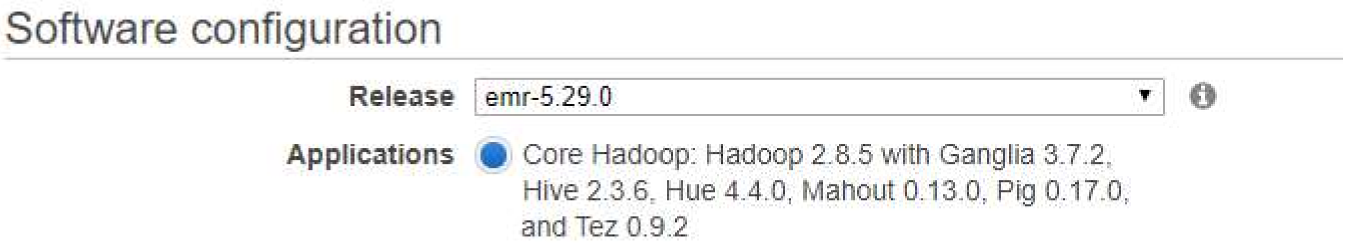
Welcome to the new E
We're redesigning the
and let us know where

You are using the fo

Launch an AWS EMR Cluster

- Launch a cluster: EMR -> Core Hadoop -> EC2 key pair.

- Choose Core Hadoop.

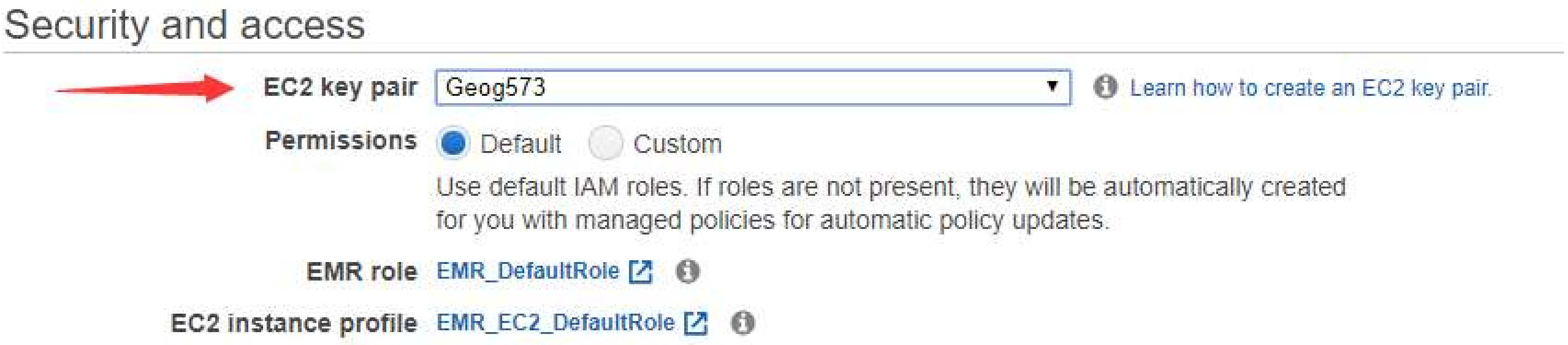


Software configuration

Release: emr-5.29.0

Applications: ● Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.6, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

- Use the key pair you just created.



Security and access

EC2 key pair: Geog573

Permissions: ● Default ○ Custom

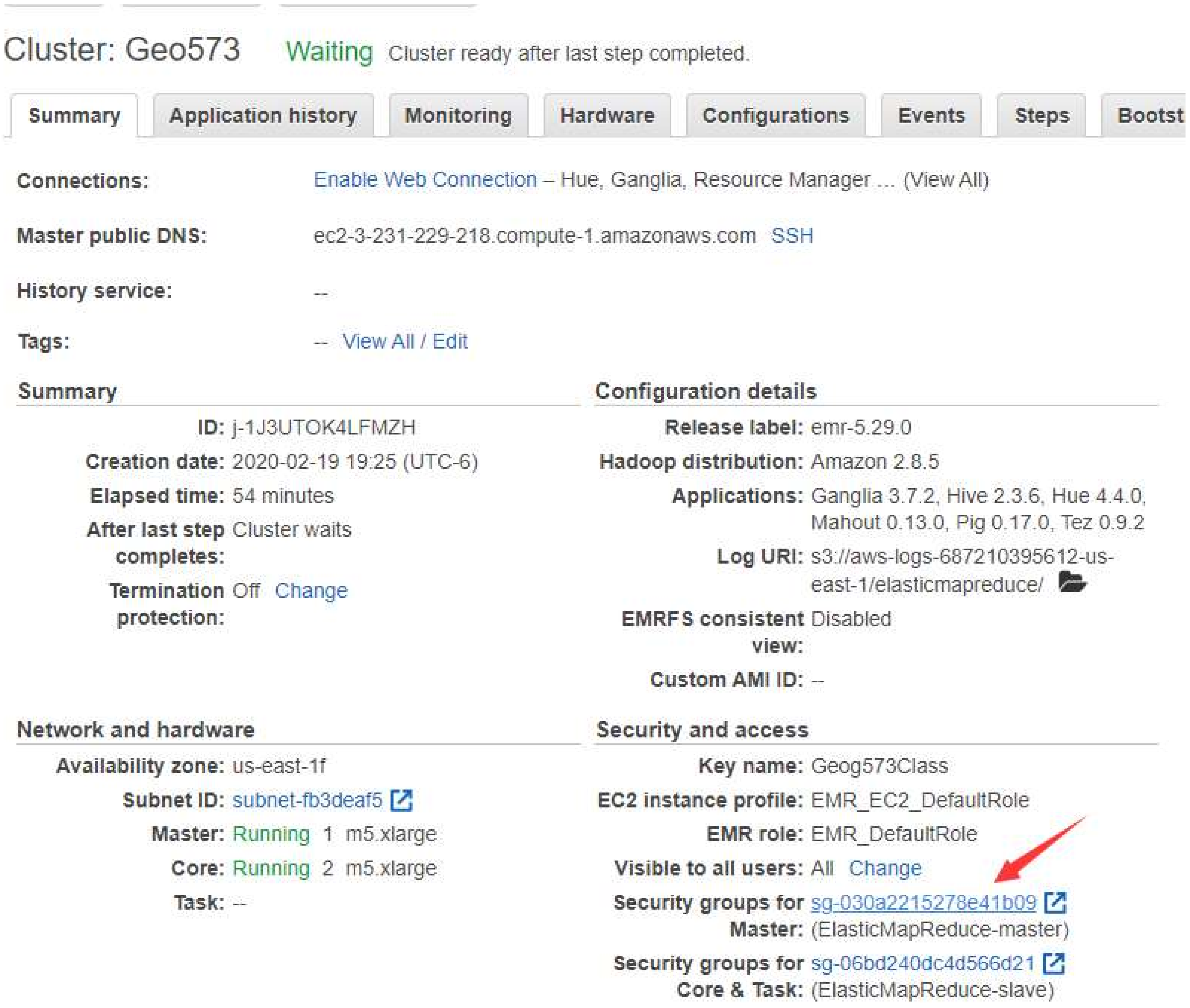
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole

EC2 instance profile: EMR_EC2_DefaultRole

Launch an AWS EMR Cluster

- Edit the Security Group



Cluster: Geo573 Waiting Cluster ready after last step completed.

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

Master public DNS: ec2-3-231-229-218.compute-1.amazonaws.com SSH

History service: --

Tags: -- View All / Edit

Summary

ID: j-1J3UTOK4LFMZH
Creation date: 2020-02-19 19:25 (UTC-6)
Elapsed time: 54 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-687210395612-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

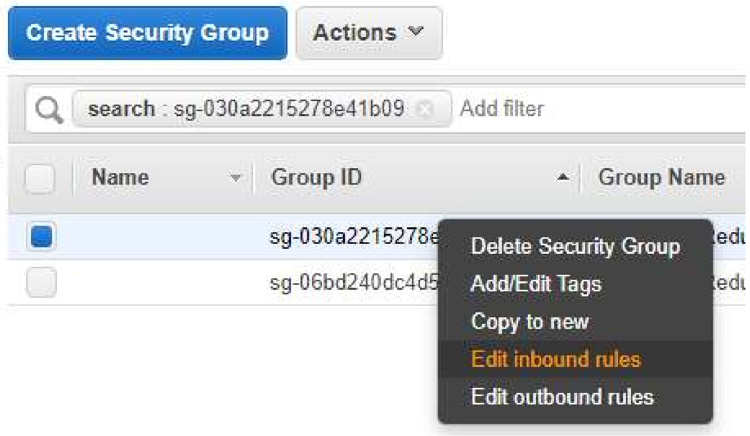
Availability zone: us-east-1f
Subnet ID: subnet-fb3deaf5
Master: Running 1 m5.xlarge
Core: Running 2 m5.xlarge
Task: --

Security and access

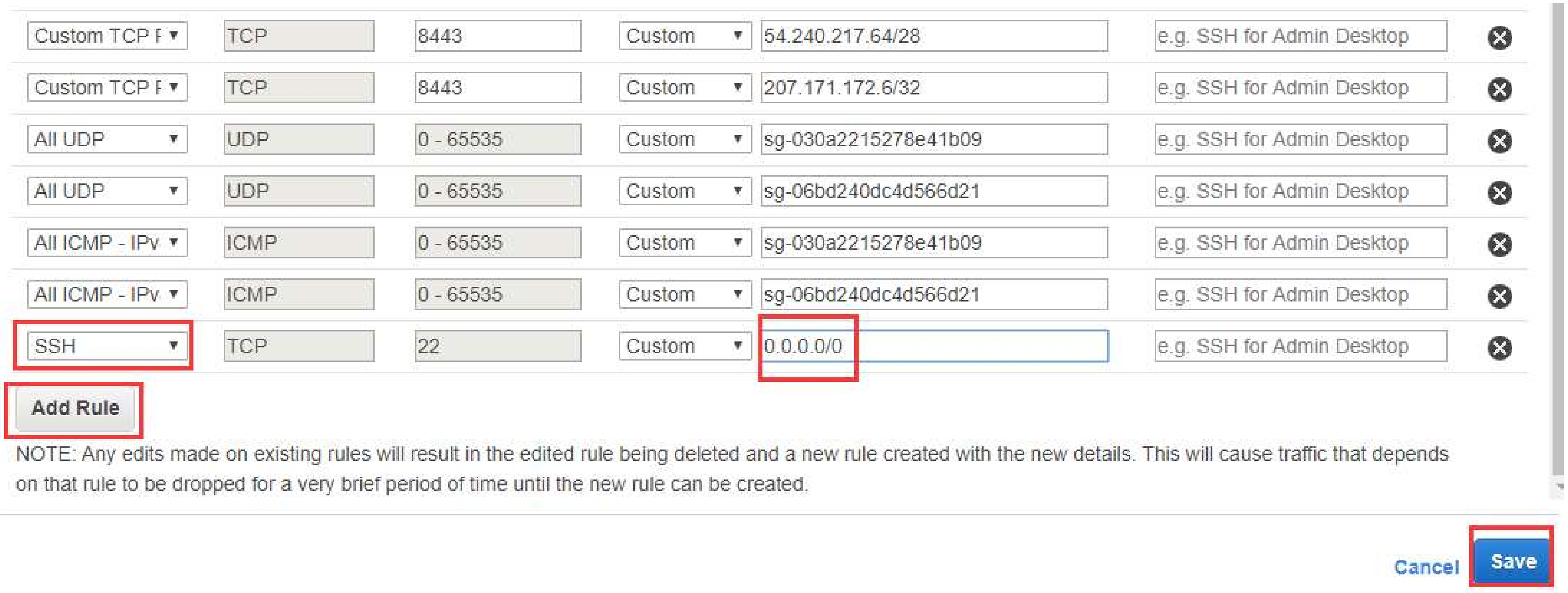
Key name: Geog573Class
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-030a2215278e41b09 (ElasticMapReduce-master)
Security groups for Core & Task: sg-06bd240dc4d566d21 (ElasticMapReduce-slave)

Launch an AWS EMR Cluster

- Edit inbound rules for **both** groups.
- If there is no SSH rule as follows, add new rule and save:
- Type: SSH, Source: 0.0.0.0/0



	Name	Group ID	Group Name
■		sg-030a2215278e	edu
□		sg-06bd240dc4d5	edu

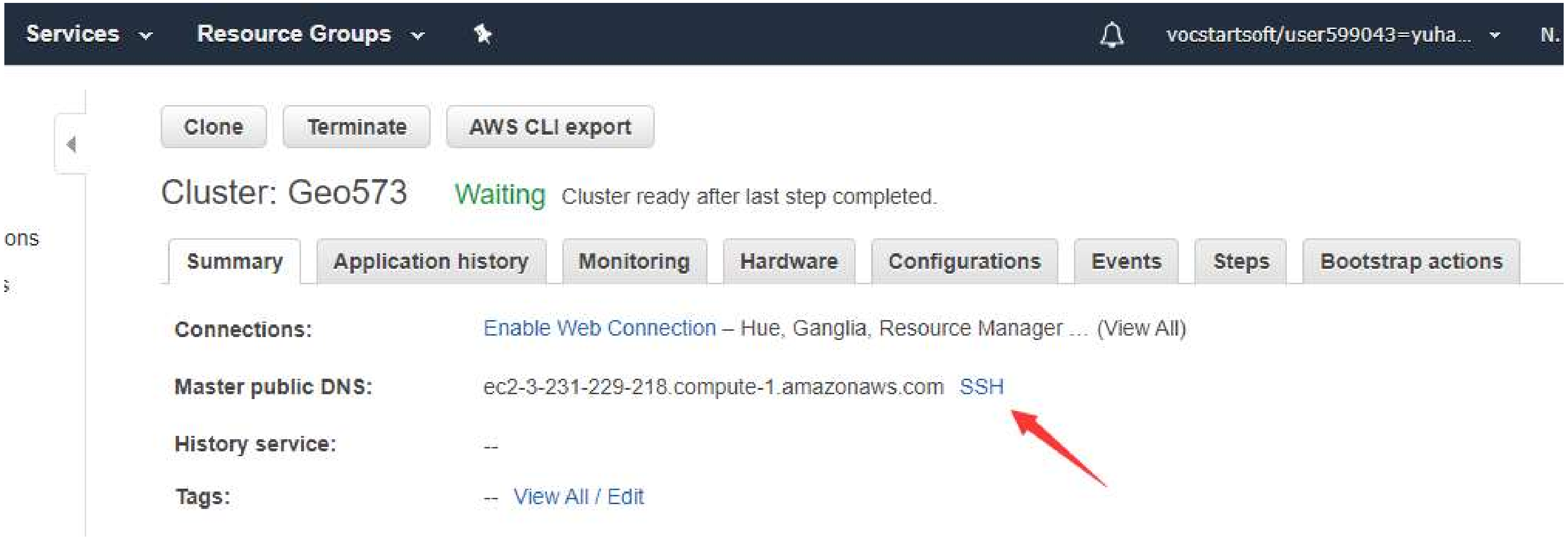


Custom TCP F	TCP	8443	Custom	54.240.217.64/28	e.g. SSH for Admin Desktop	⊗
Custom TCP F	TCP	8443	Custom	207.171.172.6/32	e.g. SSH for Admin Desktop	⊗
All UDP	UDP	0 - 65535	Custom	sg-030a2215278e41b09	e.g. SSH for Admin Desktop	⊗
All UDP	UDP	0 - 65535	Custom	sg-06bd240dc4d566d21	e.g. SSH for Admin Desktop	⊗
All ICMP - IPv	ICMP	0 - 65535	Custom	sg-030a2215278e41b09	e.g. SSH for Admin Desktop	⊗
All ICMP - IPv	ICMP	0 - 65535	Custom	sg-06bd240dc4d566d21	e.g. SSH for Admin Desktop	⊗
SSH	TCP	22	Custom	0.0.0.0/0	e.g. SSH for Admin Desktop	⊗

NOTE: Any edits made on existing rules will result in the edited rule being deleted and a new rule created with the new details. This will cause traffic that depends on that rule to be dropped for a very brief period of time until the new rule can be created.

Launch an AWS EMR Cluster

- Connect to the EMR cluster.



Cluster: Geo573 Waiting Cluster ready after last step completed.

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

Master public DNS: ec2-3-231-229-218.compute-1.amazonaws.com SSH

History service: --

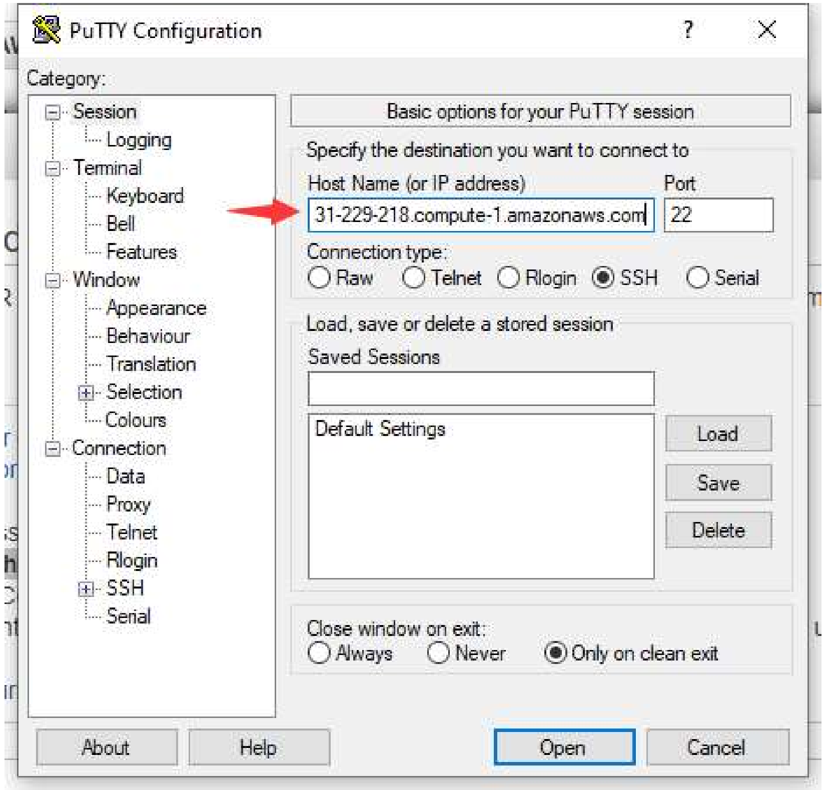
Tags: -- View All / Edit

- **Take care: Please Remember to Terminate the Cluster when You do not Use it!**

Launch an AWS EMR Cluster

Connect to the EMR cluster.

- Input Host Name field.



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): 31-229-218.compute-1.amazonaws.com

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

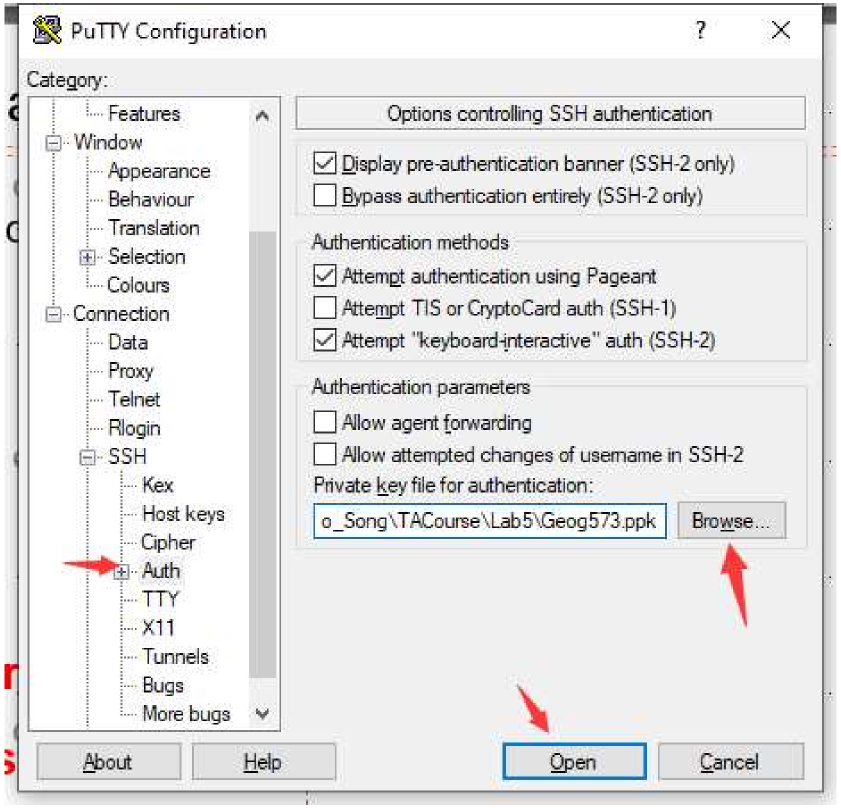
Load, save or delete a stored session

Saved Sessions

Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

- Input SSH Auth file.



PuTTY Configuration

Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)
☐ Bypass authentication entirely (SSH-2 only)

Authentication methods

☑ Attempt authentication using Pageant
☐ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

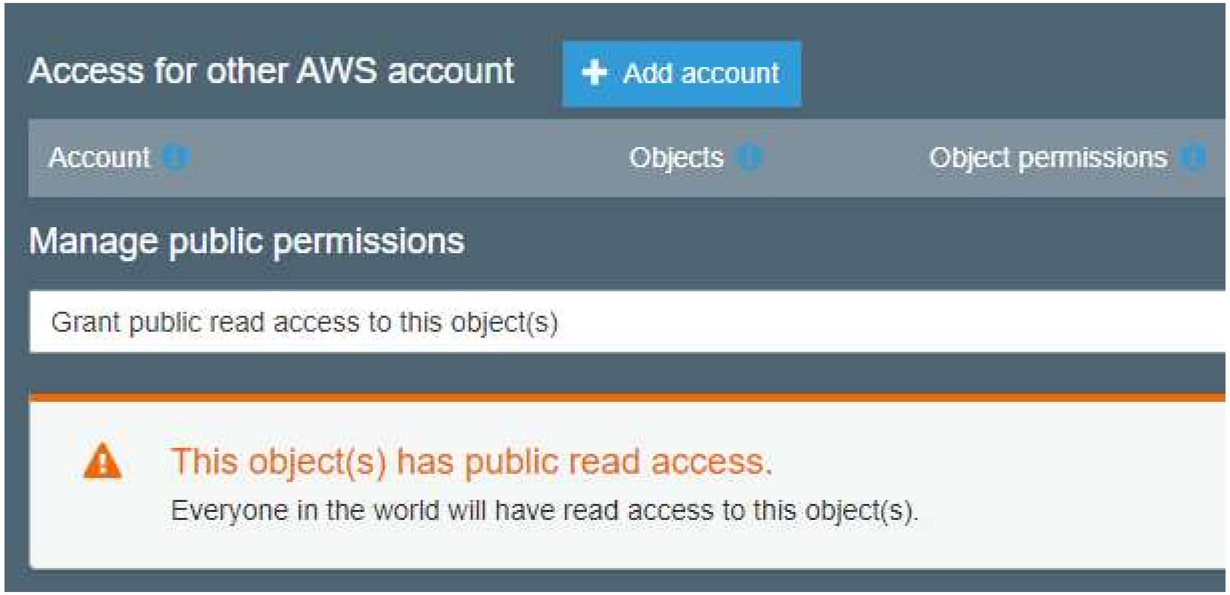
Authentication parameters

☐ Allow agent forwarding
☐ Allow attempted changes of username in SSH-2

Private key file for authentication: o_Song\TACourse\Lab5\Geog573.ppk

AWS S3

- Amazon's S3 is an object-based Web scale NAS and provides highly scalable, reliable, fast data storage infrastructure with a single logical namespace across an entire region.
- Grant public read access to the objects.



Access for other AWS account

Manage public permissions

Grant public read access to this object(s)

This object(s) has public read access.

Everyone in the world will have read access to this object(s).

Linux Basic Commands

- **Try on Windows Powershell**
- pwd: Display current folder path.
- ls: List files in the current folder.
  - ls –a: List all files including hidden files
  - ls –l: list all files with details

- **File/Folder Path**
- cd: Change folder
  - cd folder
  - cd **..** (parent directory)
  - cd ~ (home directory)
- Absolute path: contains the full path to the file.
- Relative Path: only contains a portion of the full path.
  - file
  - **./** (current directory)
  - **../** (parent directory)

Linux Basic Commands

- cp: Copy files
  - cp file new_file
- mv: Move files (Rename)
  - mv old_path new_path
- mkdir: Make a directory
  - mkdir folder
- rm: Remove files
  - rm file
- rmdir: Remove directory
  - rmdir folder

Linux Basic Commands

- wget: Download the webpage.
  - wget url
- unzip: Unzip .zip files.
  - unzip *.zip
- cat: Show the content of the file.
  - cat file
- tac: Show the content of the file reversely.
  - tac file
- grep: Search for the specific strings.
- piplines: Connect different commands.
  - cat | grep

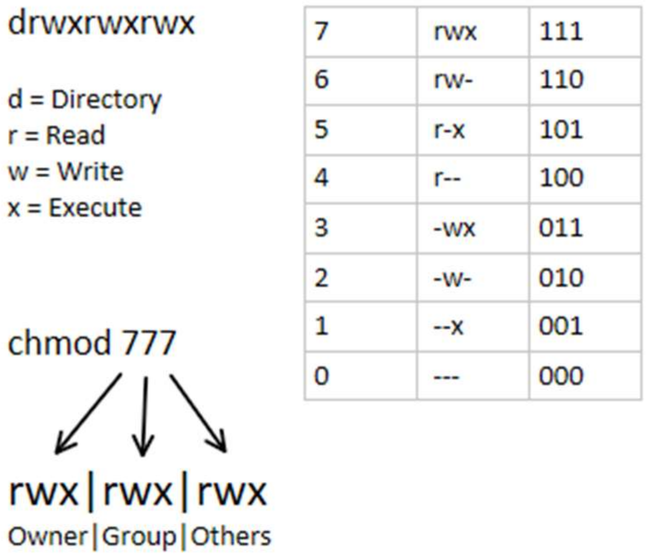
User Permission

- Three groups of users:
- Owner, group, others

- sudo: execute as a root user.
- chmod: change file readable, writable, executable mode.

  chmod digits file

- Meaning of digits:
  - 4 stands for "read",
  - 2 stands for "write",
  - 1 stands for "execute", and
  - 0 stands for "no permission.“
  - 7=4+2+1(read, write, and execute)
  - 5=4+0+1 (read, no write, and execute)
  - …



drwxrwxrwx

d = Directory
r = Read
w = Write
x = Execute

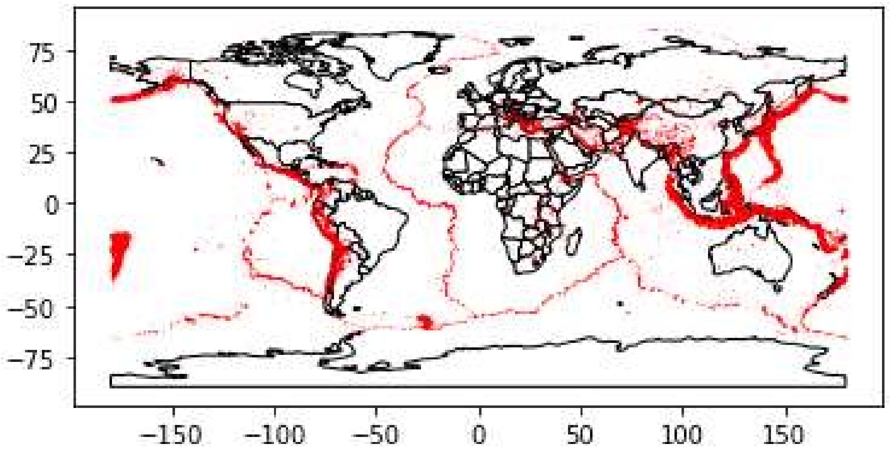
7	rwx	111
6	rw-	110
5	r-x	101
4	r--	100
3	-wx	011
2	-w-	010
1	--x	001
0	---	000

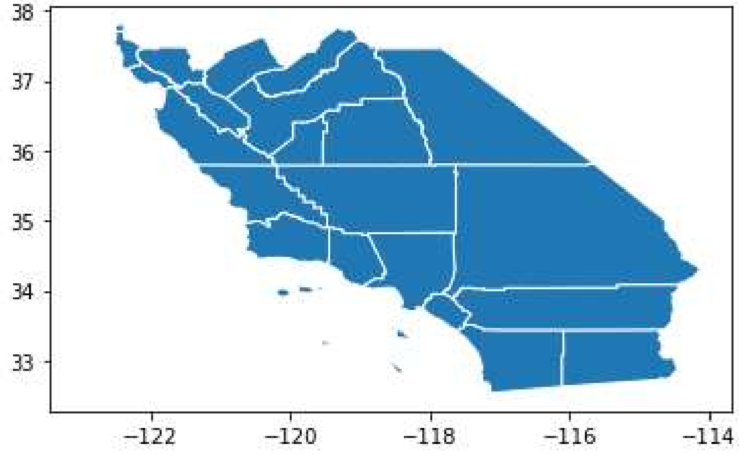
chmod 777

rwx|rwx|rwx
Owner|Group|Others

Hadoop on Cluster

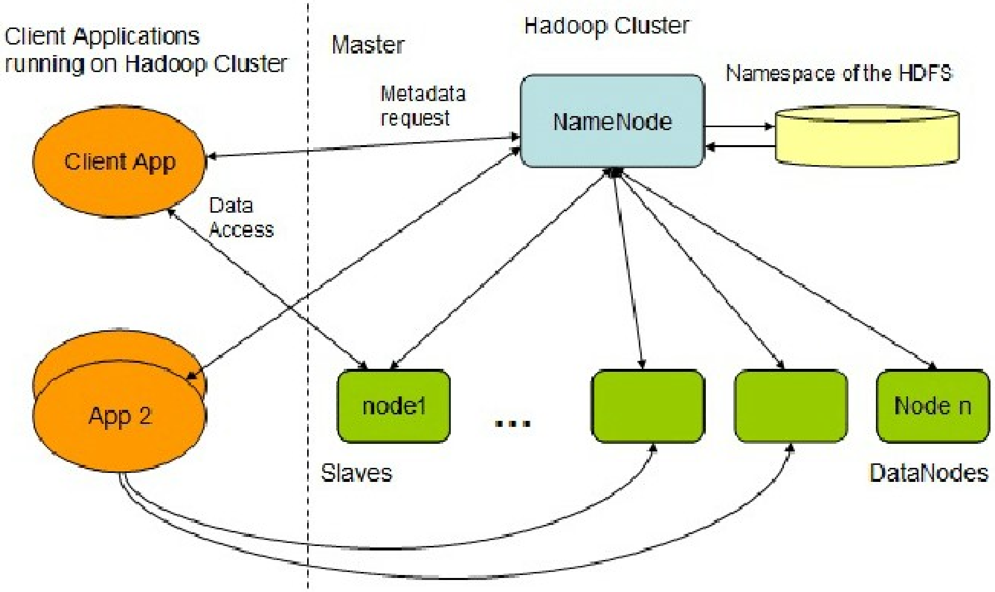
- Aggregate earthquake data to counties.
- Task: Count earthquake events in each county.
- Example: https://github.com/Esri/gis-tools-for-hadoop/tree/master/samples/point-in-polygon-aggregation-hive
- Hadoop and Hive have been installed by default.



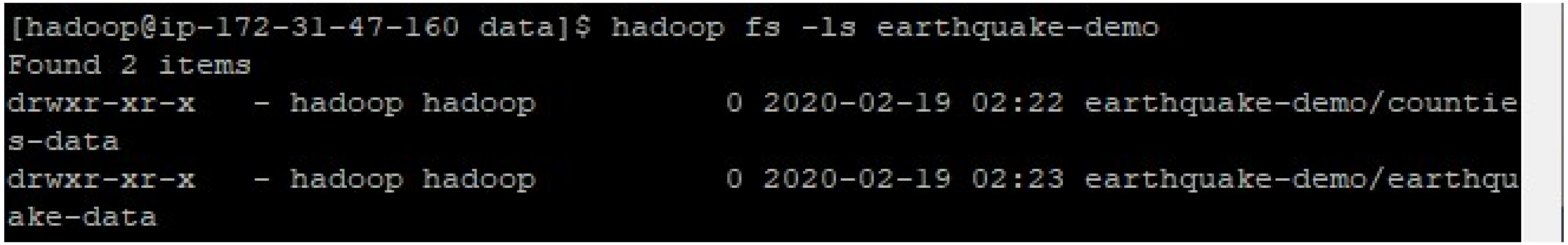


Aggregate Points to Polygons

- Upload data to Hadoop HDFS (**Be careful of the path**):
  - hadoop fs -mkdir earthquake-demo
  - hadoop fs -put DATA_PATH/counties-data earthquake-demo
  - hadoop fs -put DATA_PATH/earthquake-data earthquake-demo



- Check whether you have put them on Hadoop:



```
[hadoop@ip-172-31-47-160 data]$ hadoop fs -ls earthquake-demo
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2020-02-19 02:22 earthquake-demo/countie
s-data
drwxr-xr-x   - hadoop hadoop          0 2020-02-19 02:23 earthquake-demo/earthqu
ake-data
```

Aggregate Points to Polygons

- Open Hive command line and add external libraries:
  - add jar ${env:HOME}/DATA_PATH/lib/esri-geometry-api-2.0.0.jar;
  - add jar ${env:HOME}/DATA_PATH/lib/spatial-sdk-hive-2.0.0.jar;
  - add jar ${env:HOME}/DATA_PATH/lib/spatial-sdk-json-2.0.0.jar;
  - create temporary function ST_Point as 'com.esri.hadoop.hive.ST_Point';
  - create temporary function ST_Contains as 'com.esri.hadoop.hive.ST_Contains’;

- Drop tables if exist.
  - drop table earthquakes;
  - drop table counties;

Aggregate Points to Polygons

- Create new tables.
  - CREATE TABLE earthquakes (earthquake_date STRING, latitude DOUBLE, longitude DOUBLE, depth DOUBLE, magnitude DOUBLE, magtype string, mbstations string, gap string, distance string, rms string, source string, eventid string)
  - ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
  - STORED AS TEXTFILE;

  - CREATE TABLE counties (Area string, Perimeter string, State string, County string, Name string, BoundaryShape binary)
  - ROW FORMAT SERDE 'com.esri.hadoop.hive.serde.EsriJsonSerDe'
  - STORED AS INPUTFORMAT 'com.esri.json.hadoop.EnclosedEsriJsonInputFormat'
  - OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat';
- Load data from file.
  - LOAD DATA INPATH 'earthquake-demo/earthquake-data/earthquakes.csv' OVERWRITE INTO TABLE earthquakes;
  - LOAD DATA INPATH 'earthquake-demo/counties-data/california-counties.json' OVERWRITE INTO TABLE counties;

Aggregate Points to Polygons

- Run the demo: Count earthquakes in each county.
  - SELECT counties.name, count(*) cnt FROM counties
  - JOIN earthquakes
  - WHERE ST_Contains(counties.boundaryshape, ST_Point(earthquakes.longitude, earthquakes.latitude))
  - GROUP BY counties.name
  - ORDER BY cnt desc;
- Exit Hive: Ctrl-C.

Potential Bugs

- Solve the problem that Hive reports FAILED:
  - SemanticException Cartesian products are disabled for safety reasons. If you know what you are doing, please sethive.strict.checks.cartesian.product to false and that hive.mapred.mode is not set to 'strict' to proceed. Note that if you may get errors or incorrect results if you make a mistake while using some of the unsafe features.
- Solution:
  - set hive.mapred.mode=nonstrict;

Lab Assignment (Due Feb.26th)

Task 1*:*

- Upload your Lab 4 code to an S3 bucket and allow public access. Submit the url in the text entry.

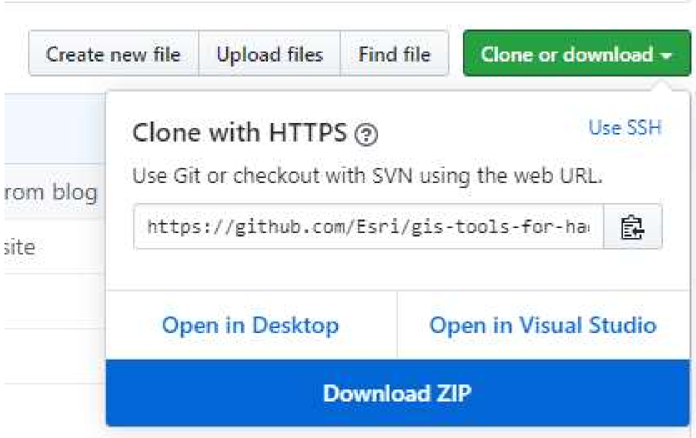
Task 2*:*

- Launch a cluster and practice all Linux operations taught today. Download your Lab 4 code to the instance. Show the user permission information of the code. Change its mode so that only the owner can read, write, and execute, while others can only read. Then execute the code. Submit a screenshot to show all the commands and outputs.
- At least three commands should be contained: wget, chmod, python/python3.

Lab Assignment (Due Feb.26th)

Task 3:

- Follow the Hadoop example on GitHub to aggregate earthquake data to counties. Run the demo analysis and submit the screenshots of the results.
- To download the GitHub repository, please visit: https://github.com/Esri/gis-tools-for-hadoop
- Right click Download ZIP and copy link address.



Clone with HTTPS

Use Git or checkout with SVN using the web URL.

https://github.com/Esri/gis-tools-for-ha

Note:

- You only need to submit the Lab 4 code S3 url, and two (or more if necessary) screenshots.
- **Please Remember to Terminate the Cluster when You do not Use it!**

THANK YOU