

CSE 435/535 Information Retrieval

COMPUTING ENVIRONMENT SETUP GUIDEBOOK

Contents

1. Introduction	3
2. Twitter Developer Account	3
3. AWS EC2.....	3
3.1 AWS account.....	4
3.2 EC2 Setup	4
4. Apache Lucene Solr.....	6
5. FAQs.....	7

1. Introduction

This guidebook is designed for students to help setup Twitter developer account, AWS EC2 instance and Apache Lucene Solr. For smooth execution of all projects in this course, students are expected to complete setups as soon as possible and let TAs know if they are facing any problems within first week of the class.

2. Twitter Developer Account

Since July, 2018, Twitter requires you to have a developer account to access their APIs. To create developer account, you need to have a Twitter account. Please sign up for it if you don't already have one and then follow these steps to apply for developer account and obtain access tokens:

- Go to <https://developer.twitter.com/>
- Click "Apply" link on the top right corner of the page. And then click "Apply for a developer account"
- If you do not have an account, Twitter will ask you to create one.
- After logging into Twitter account, specify your country and account name. You can use meaningful account name like cse535IR.
- On the next page, select "Student" as the type of user and enter following description in the usage description box:

"I am a student taking the Information Retrieval course (CSE 435/535) in the computer science department at the University at Buffalo. Our project involves ingesting tweet data in different languages. We are required to collect ~100,000 tweets over a period of three to four weeks. We require access to the twitter developer API.

We will be analyzing COVID related tweets from public officials and health organizations based on multilingual search and topic analysis which are taught during the course. This is for academic requirements only and the data and findings will not be distributed."

- Uncheck **all** other usage boxes and proceed to the next page.
- After accepting the agreement, you will receive an email for verification. Click the "Confirm your email" box and you should get your developer account approved.

To apply for access tokens, follow these steps:

- Create an app by going on "Apps" page
- Enter the App details. The actual values do not really matter but filling some meaningful values is recommended.
- Review developer policy and create!
- At the end of this step, you should have values for the following four fields under the "Keys and Access Tokens" tab : *Consumer Key, Consumer Secret, Access Token and Access Token Secret*. You will need these four values to be able to connect to Twitter using a client and querying for data.

3. AWS EC2

The first project (and possibly others) will involve the use of an EC2 instance on Amazon AWS. This section will guide you in setting up an AWS account and EC2 instances which will be used

for projects in this course. You are required to complete this setup before class on September 7th so that you are prepared for hands-on session in the class.

3.1 AWS account

Amazon Web Services (AWS) is Amazon's cloud web hosting platform that offers flexible, reliable, scalable and easy-to-use solutions. Before we begin, you would need to sign up for an AWS account if you don't already have one: <https://aws.amazon.com> Although the sign-up requires a credit card for verification purposes, you can use a gift card instead. Note that you can even share a gift card amongst a group if you desire.

UB is a part of the AWS educate program. It gives you \$100 in annual credit. Follow instructions at <http://www.buffalo.edu/ubit/service-guides/teaching-technology/aws.html> to setup your AWS student account with free credits. We do not anticipate students using more than their free tier allocation.

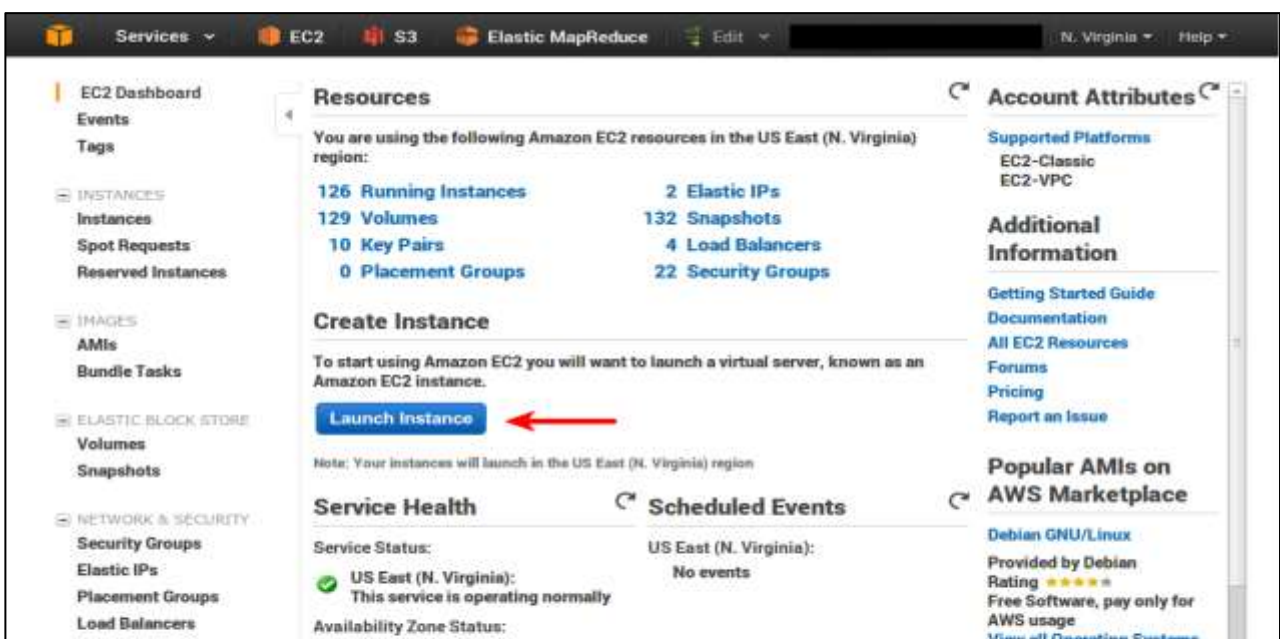
If you are struggling in understanding UB's steps for setup, you can also follow these streamlined steps a student came up with last year to set up quickly:

1. Open this link [Select 'Apply for AWS Educate'](#)
2. Click on "Apply for AWS Educate for Students"
3. Confirm your account via email. Make sure you are using UB email.
4. Create a new [aws account](#) with your buffalo email id. Select the **BASIC** plan. Enter your debit/credit card details.
5. Login. Go to My account. Copy your **account id**.
6. Login to [aws educate](#) with your buffalo email id. If you are not able to log in, click forgot password and try again.
7. Click **back to login**. Click '**Sign in to personal AWS account**' (visible on your right). Enter the account id you copied in step 4. You will receive the promotional code.
8. Copy the promotional code.
9. Go to My account in aws. Go to Credits and enter the promotional code with the captcha.

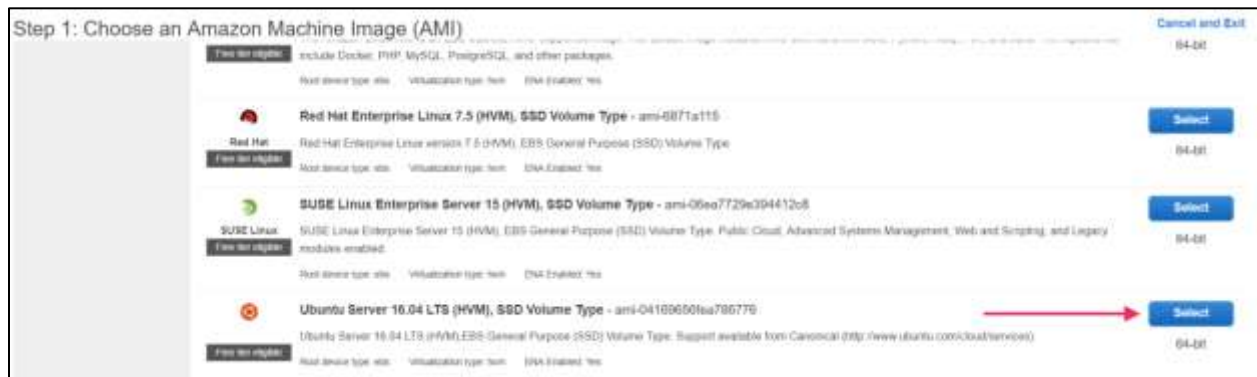
3.2 EC2 Setup

Although there are several guides available, instructions here are adapted from Solr's EC2 guide here: <https://wiki.apache.org/solr/SolrOnAmazonEC2>

1. Login to your AWS account and navigate to the EC2 dashboard.
2. Create an instance
 - a. Click on "Launch Instance"



- b. Select an AMI type. For this demo, we are using Ubuntu 16.04 LTS

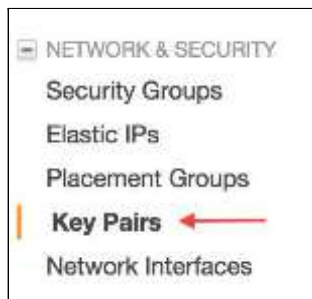


- c. Choose an instance type. We keep the default option (General purpose, t2.micro). You may need to change this to t2.small for your project later.
- d. Keep the default options for steps 3,4 and 5 (Configure Instance, Add Storage and Tag Instance)
- e. In step 6 (Configure Security Group), create a new security group. Provide access to SSH for your IP and global access for port 8983 (see screenshot below). We will later provide a more restricted IP list. You could restrict it to your IP for the time being.



- f. Review and launch!

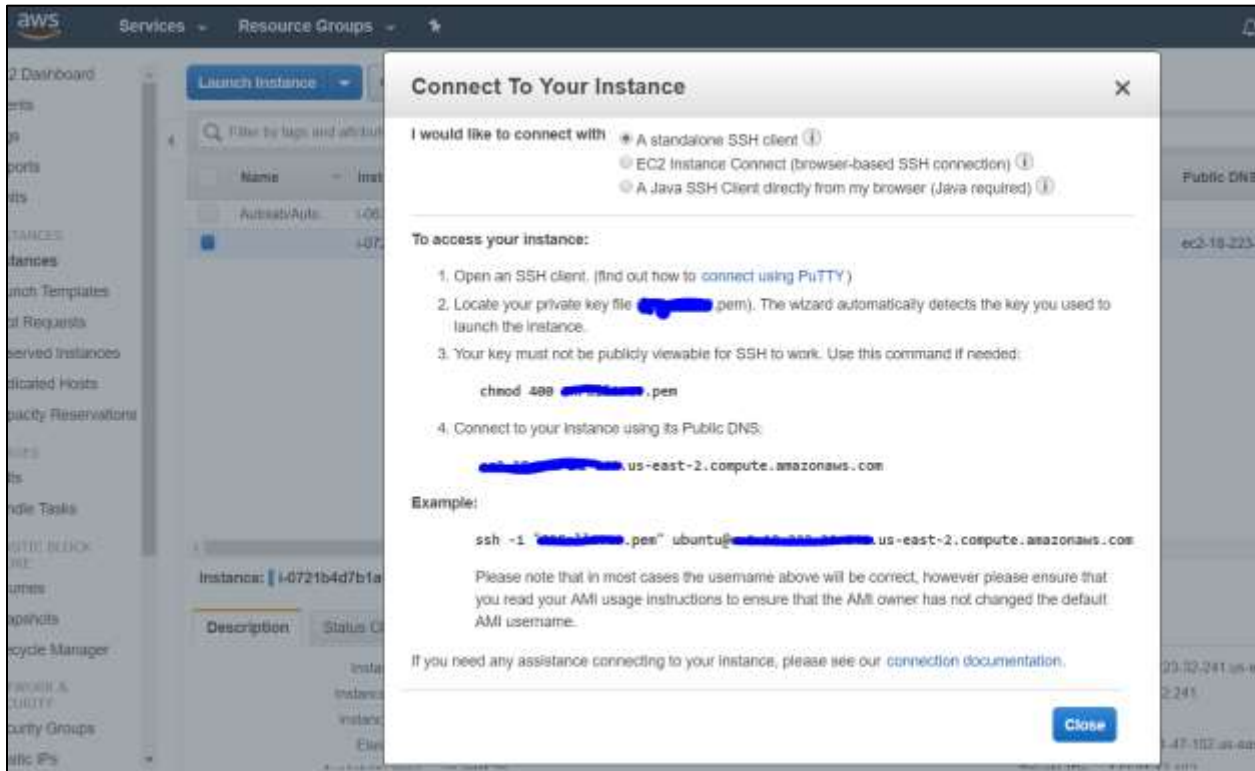
3. Create a keypair to log in to your instance.
- a. Click on “Key Pairs” under the “Network and Security” group in the left pane



- b. Create a new key pair by giving it some meaningful name. Download and save the file. For most Unix/Linux based systems `~/.ssh` is a good place. However, make sure that this key is only with read permission: `chmod 400 <path to key>`

4. Connect to your instance.

- a. By now your instance must be up and running. Find its hostname or ip address in the instance description.
- b. Click on “Connect” button at the top panel and follow the instructions:



- c. **Windows users** can use PuTTY to login to their EC2 Ubuntu instance. You can download PuTTY software from <https://www.putty.org/>. It is an SSH client and is open source. Please refer [this document](#) to connect your Ubuntu instance from Windows using Putty.

4. Apache Lucene Solr

Solr is an open source enterprise search platform, written in Java, from the Apache Lucene project. In this project, we will be using Solr for indexing and querying purposes. You will play around with Solr schema, log files and various indexing operations.

It is fairly easy to have a Solr instance up and running, at least for sanity checks. We will be using Solr 8.2 for all the projects in this course. Refer to [Solr Reference Guide 8.1](#) which contains detailed explanation of operations you will be using. If you have used Solr before with different version, please [read the major changes](#) implemented in this release.

SETUP ON EC2

1. Connect to your EC2 instance.
2. Install Java 8

```
sudo apt-get update
```

```
sudo apt-get install default-jre
sudo apt-get install default-jdk
```

3. Download Solr:

```
wget https://archive.apache.org/dist/lucene/solr/8.2.0/solr-8.2.0.tgz
```

4. Untar Solr and enter the directory "solr-8.2.0":

```
tar xf solr-8.2.0.tgz
```

5. Start a schemaless server: `bin/solr start -p 8983 -e schemaless`

6. Verify the instance is working on [http://\[host-name\]:8983/solr](http://[host-name]:8983/solr)

7. Index/post data using : `bin/post -c gettingstarted example/exampledocs/*.xml`

8. Verify if the data is indexed by going to your solr instance.

9. Stop solr by `bin/solr stop -p 8983`

Note: During project implementation, Windows users will have to transfer crawled data from their OS to EC2 instance. You can use FileZilla or other file transferring tools to do this.

5. FAQs

Q: I am not able to login using ssh. I think there is something wrong with EC2 security group.

```
$ ssh -i ~/.ssh/new.pem ubuntu@[REDACTED].us-west-2.compute.amazonaws.com
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@ WARNING: UNPROTECTED PRIVATE KEY FILE! @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```

Permissions 0644 for '/Users/[REDACTED]/.ssh/new.pem' are too open.

It is required that your **private key files are NOT accessible by others**.

This private key will be ignored.

Load key "/Users/[REDACTED]/.ssh/new.pem": **bad permissions**

Permission denied (publickey).

A: Try all of the solutions below and see what works for you:

Sol-1. You need to change the permissions for your key. Try '`chmod 400 <your_key_file>`' and then execute. That will fix it since by changing the file permissions to 400 you are making it private which is what is required in this case. After giving this permission, You might need to reset permissions for the directory '`sudo chmod 755 <your_key_file>`'

Sol-2. Instead of using Public DNS, try using the public IP that u can see in instances

Sol-3. Apart from the solutions mentioned above, following are the reasons that can also cause the same error:

<http://stackoverflow.com/questions/18551556/permission-denied-publickey-when-ssh-access-to-amazon-ec2-instance>

This error message means you failed to authenticate.

These are common reasons that can cause that:

1. Trying to connect with the wrong key. Are you sure this instance is using this keypair?
2. Trying to connect with the wrong username. ubuntu is the username for the ubuntu based AWS distribution, but on some others it's ec2-user (or admin on some Debians, according to Bogdan Kulbida's answer)(can also be root, fedora, see below)
3. Trying to connect the wrong host. Is that the right host you are trying to log in to?

Note that 1. will also happen if you have messed up the /home/<username>/.ssh/authorized_keys file on your EC2 instance.

About 2., the information about which username you should use is often lacking from the AMI Image description. But you can find some in AWS EC2 documentation, bullet point 4.

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AccessingInstancesLinux.html>

Q: While creating the EC2 instance, I did not download the Key pair instance. Can you tell me how to proceed further?

A: You will have to delete the old EC2 instance you created and create a new. This time don't forget to download the key pair. Other option is to generate a key pair yourself and import it in the key pairs tab and then restart your instance using the new keypair.

Q: I am unable to connect to EC2 suddenly. It was working when I was at home but not working on campus (or working at campus but not working at home.)

A: Every time you move to different network, you need to give access to your IP address. Go to "Security Groups" -> Set "My IP" wherever necessary.