CSE 435/535 Information Retrieval

SETUP GUIDEBOOK

Contents

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.
- Verify your account against your US phone number, select the country of residence and give meaningful name to your account (eg: 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 ~200,000 tweets over a period of three to four weeks. We require access to the twitter developer API.

  We will be analyzing rhetoric from prominent public leaders based on multilingual search, sentiment analysis 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 August 28th so that you are prepared for hands-on session in the class.

3.1 AWS account

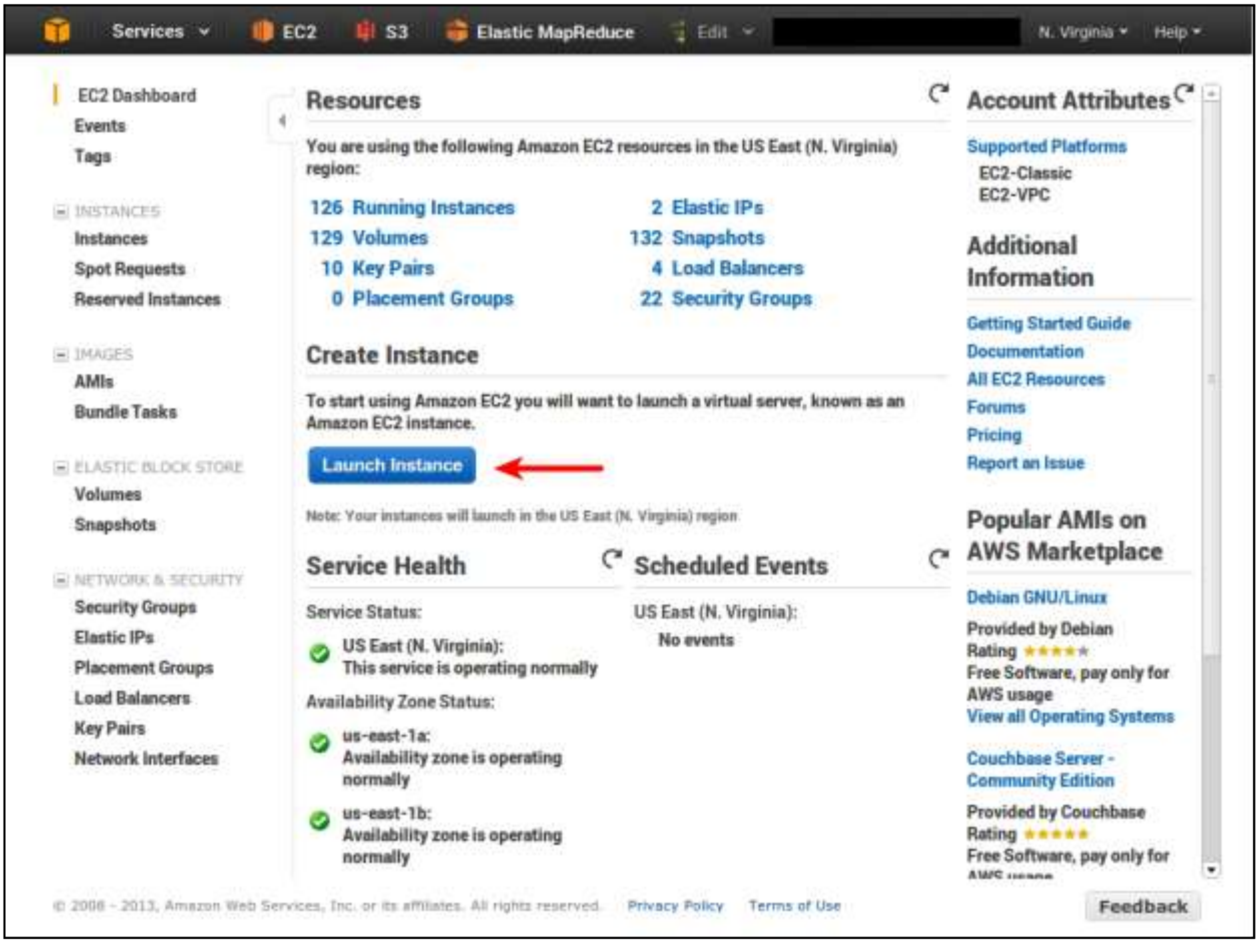
Amazon Web Services (AWS) is Amazon's cloud web hosting platform that offers flexible, reliable, scalable and easy-to-use solutions. Before we begin, you would need to sign up for an AWS account if you don't already have one: https://aws.amazon.com Although the sign-up requires a credit card for verification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

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"



Resources

You are using the following Amazon EC2 resources in the US East (N. Virginia) region:

126 Running Instances	2 Elastic IPs
129 Volumes	132 Snapshots
10 Key Pairs	4 Load Balancers
0 Placement Groups	22 Security Groups

Create Instance

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Note: Your instances will launch in the US East (N. Virginia) region

Service Health

Service Status:

US East (N. Virginia):
This service is operating normally

Availability Zone Status:

us-east-1a:
Availability zone is operating normally

us-east-1b:
Availability zone is operating normally

Scheduled Events

US East (N. Virginia):
No events

Account Attributes

Supported Platforms
EC2-Classic
EC2-VPC

Additional Information

Getting Started Guide
Documentation
All EC2 Resources
Forums
Pricing
Report an Issue

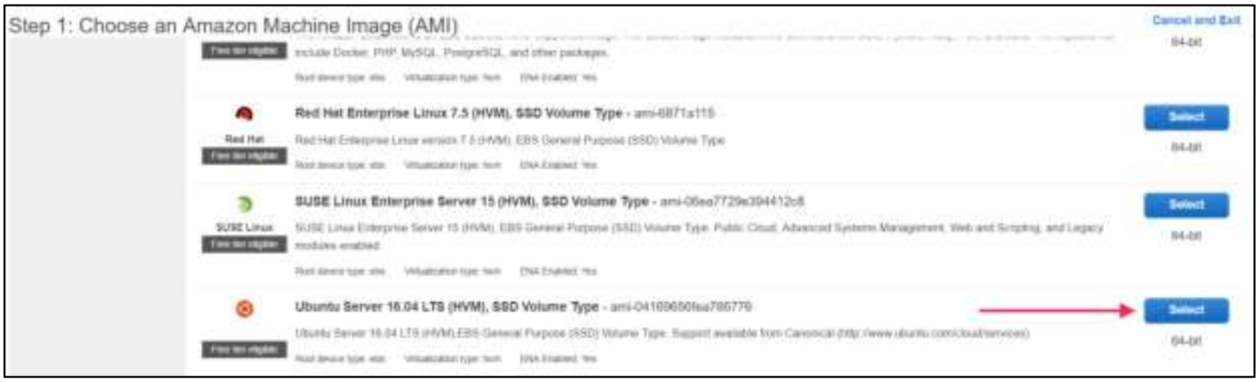
Popular AMIs on AWS Marketplace

Debian GNU/Linux
Provided by Debian
Rating ★★★★★
Free Software, pay only for AWS usage
View all Operating Systems

Couchbase Server - Community Edition
Provided by Couchbase
Rating ★★★★★
Free Software, pay only for AWS usage

© 2008 - 2013, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

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



Step 1: Choose an Amazon Machine Image (AMI)

Cancel and Exit

Free tier eligible

include Docker, PHP, MySQL, PostgreSQL, and other packages.

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

64-bit

Red Hat Enterprise Linux 7.5 (HVM), SSD Volume Type - ami-6871a115

Red Hat Enterprise Linux version 7.5 (HVM), EBS General Purpose (SSD) Volume Type

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

Red Hat · Free tier eligible · Select · 64-bit

SUSE Linux Enterprise Server 15 (HVM), SSD Volume Type - ami-06ea7729e394412c8

SUSE Linux Enterprise Server 15 (HVM), EBS General Purpose (SSD) Volume Type. Public Cloud, Advanced Systems Management, Web and Scripting, and Legacy modules enabled.

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

SUSE Linux · Free tier eligible · Select · 64-bit

Ubuntu Server 16.04 LTS (HVM), SSD Volume Type - ami-04169656fea786776

Ubuntu Server 16.04 LTS (HVM),EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services).

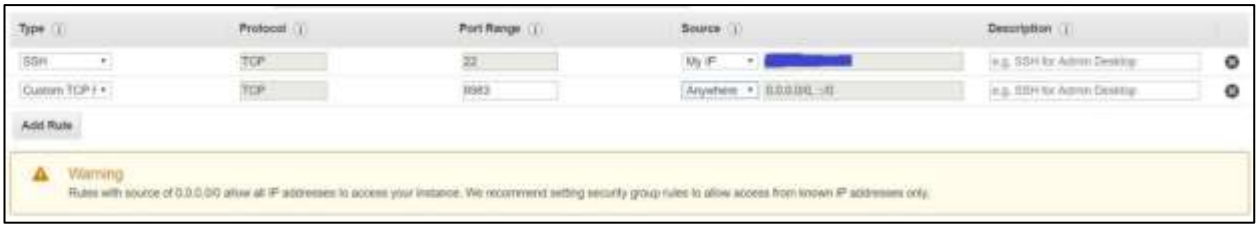
Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

Free tier eligible · Select · 64-bit

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. Create a new security group. Provide access to SSH for your IP and global access for port 8983. We will later provide a more restricted IP list. You could restrict it to your IP for the time being.



Type	Protocol	Port Range	Source	Description
SSH	TCP	22	My IP [redacted]	e.g. SSH for Admin Desktop
Custom TCP I	TCP	8983	Anywhere 0.0.0.0/0, ::/0	e.g. SSH for Admin Desktop

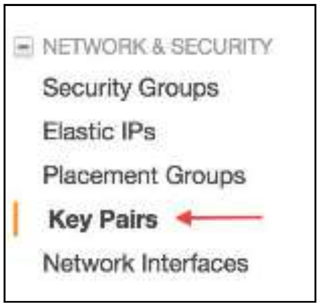
Add Rule

Warning
Rules with source of 0.0.0.0/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

f. Review and launch!

3. Create a keypair to log in to your instance.

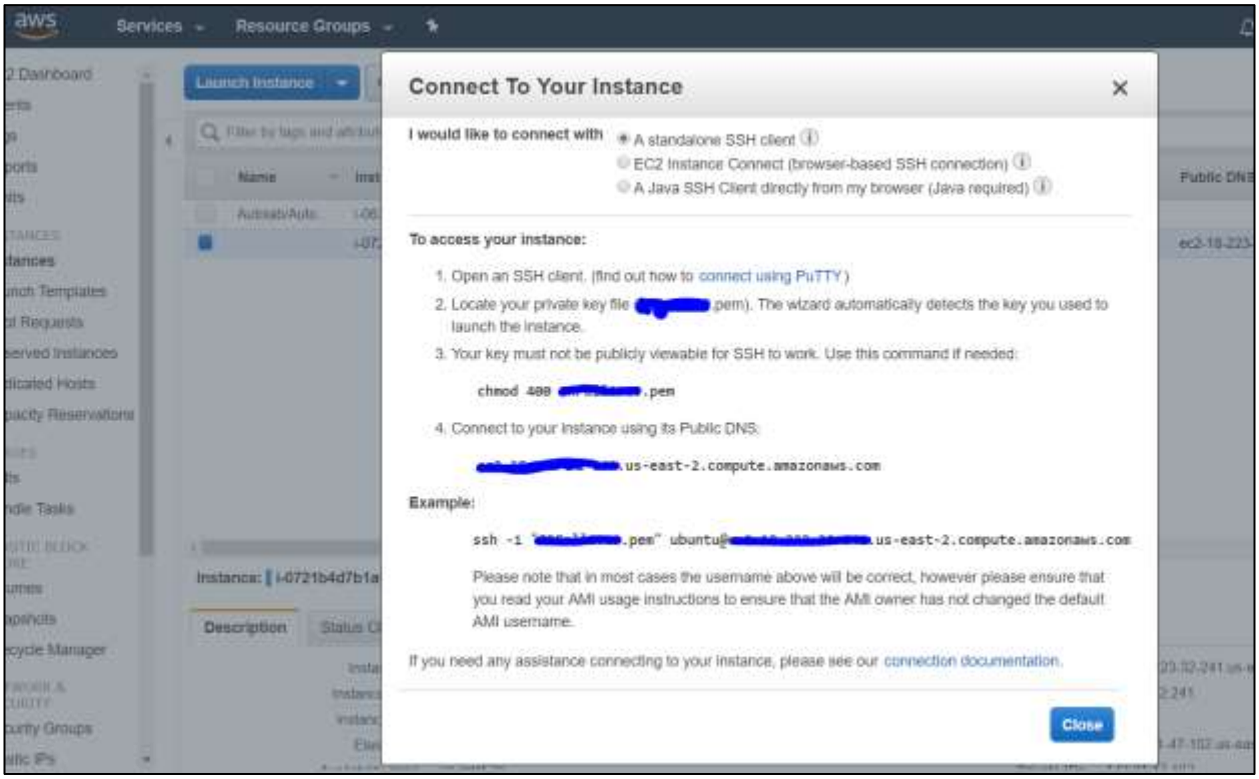
   a. Click on “Key Pairs” under the “Network and Security” group in the left pane



   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:



Connect To Your Instance

I would like to connect with
● A standalone SSH client
○ EC2 Instance Connect (browser-based SSH connection)
○ A Java SSH Client directly from my browser (Java required)

To access your instance:

1. Open an SSH client. (find out how to connect using PuTTY)
2. Locate your private key file ([redacted].pem). The wizard automatically detects the key you used to launch the instance.
3. Your key must not be publicly viewable for SSH to work. Use this command if needed:

```
chmod 400 [redacted].pem
```

4. Connect to your instance using its Public DNS:

```
[redacted].us-east-2.compute.amazonaws.com
```

Example:

```
ssh -i "[redacted].pem" ubuntu@[redacted].us-east-2.compute.amazonaws.com
```

Please note that in most cases the username above will be correct, however please ensure that you read your AMI usage instructions to ensure that the AMI owner has not changed the default AMI username.

If you need any assistance connecting to your instance, please see our connection documentation.

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 http://www.gtlib.gatech.edu/pub/apache/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 standalone 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.