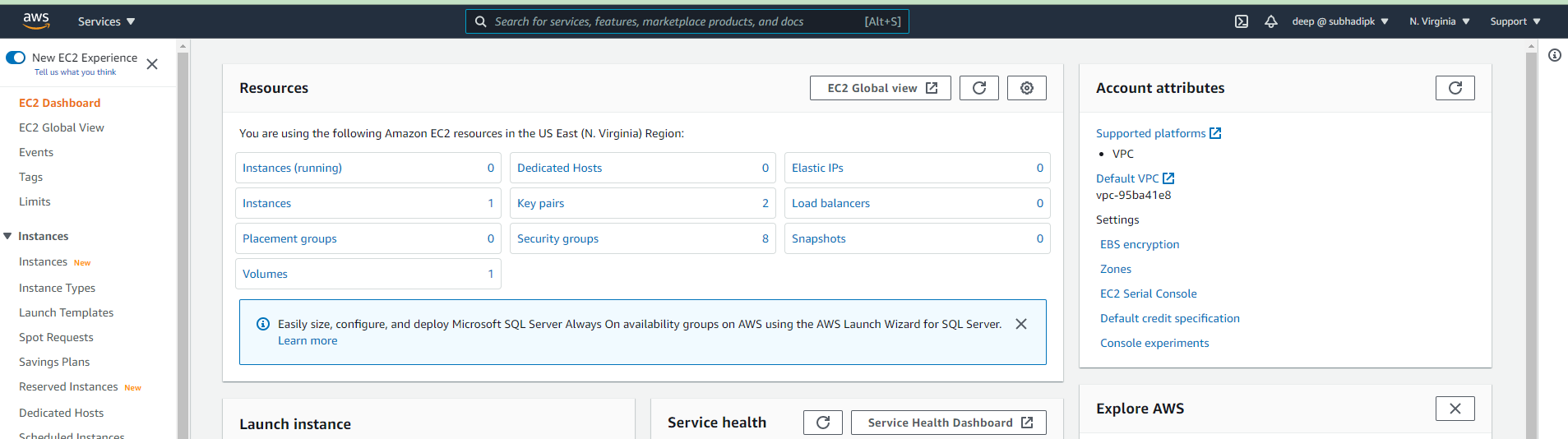
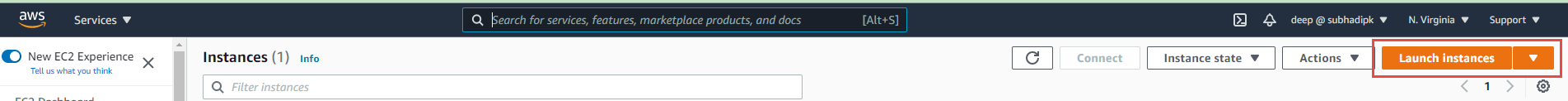
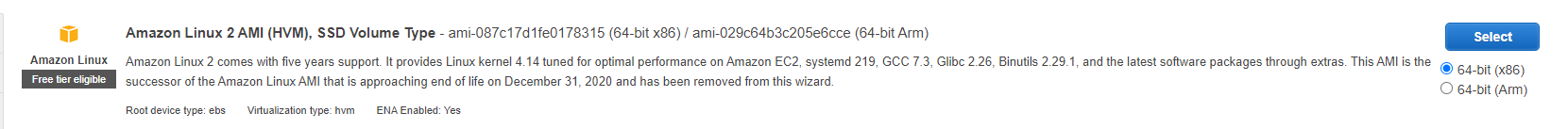
# Install and Setup EC2 instance

Go to AWS Management Console -> Create a t2.micro instance on AWS with default settings

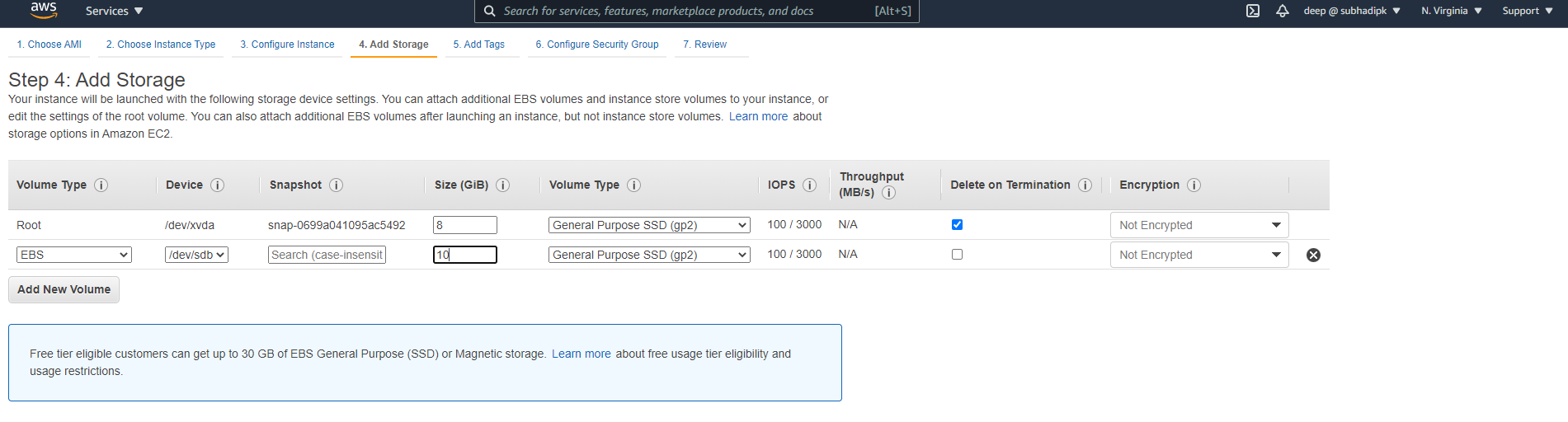




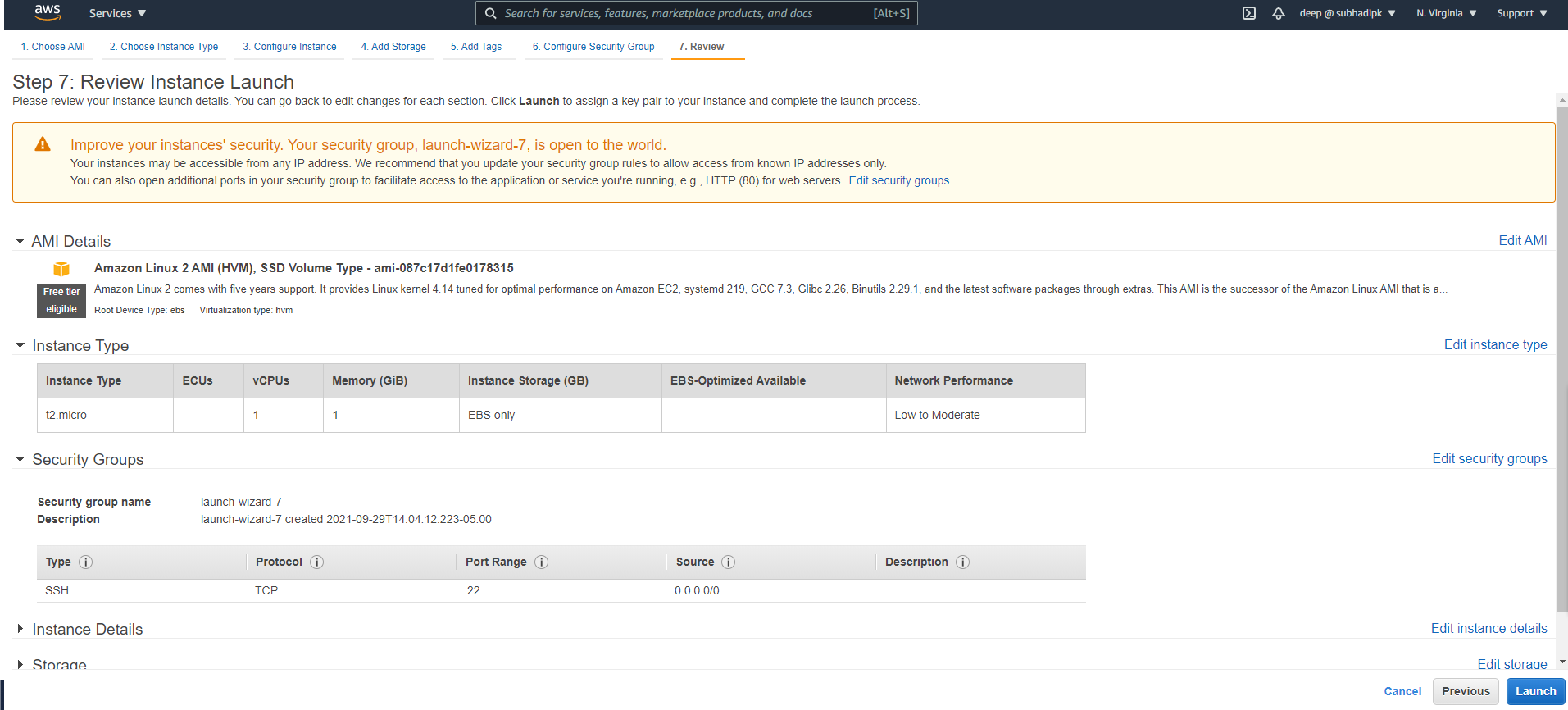
Select a t2.micro instance Amazon Linux2



I selected default settings except I added extra 10GB of space

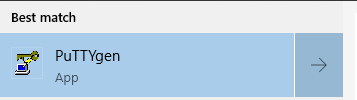


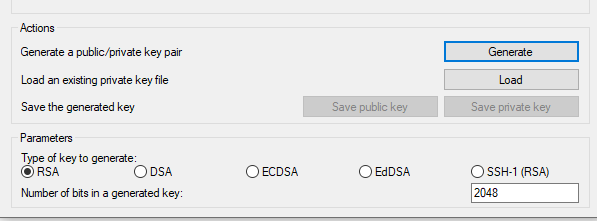
Review and Launch

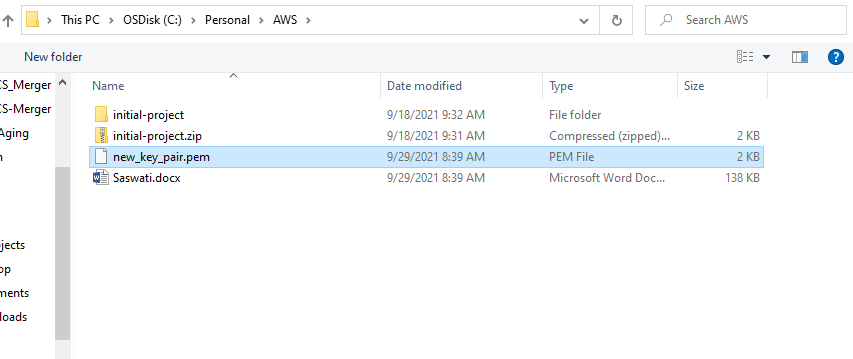


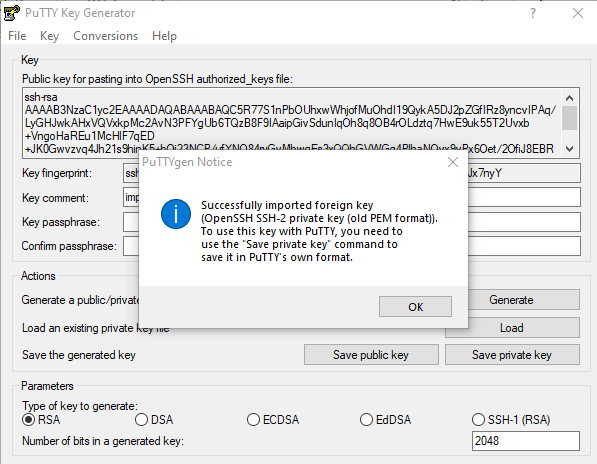
Download the private key and save as .pem file

Convert .pem file to .ppk file using PuttyGen, connect to putty using .ppk file



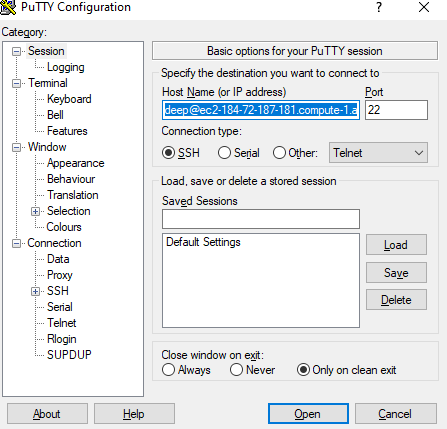


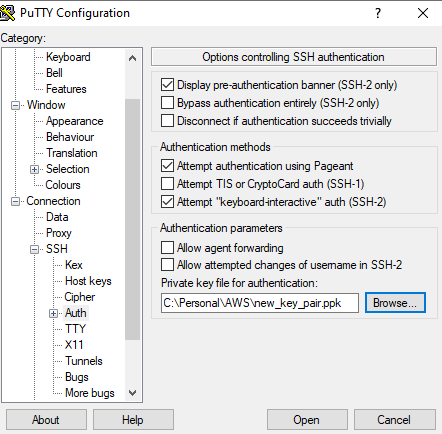






new\_key\_pair





Once connected go to the EC2 instance and perform below:

# Install Google-Chrome

*sudo curl https://intoli.com/install-google-chrome.sh | bash*

*sudo mv /usr/bin/google-chrome-stable /usr/bin/google-chrome*

*cd /usr/bin/*

*google-chrome – version*

*94.0.4606.61*

# Install chromedriver

*cd /tmp/*

*sudo wget https://chromedriver.storage.googleapis.com/94.0.4606.61/chromedriver\_linux64.zip*

*sudo unzip chromedriver\_linux64.zip*

*sudo mv chromedriver /usr/bin/chromedriver*

*chromedriver – version*

# Install python3, pip and other bindings (selenium, boto3, requests, time etc)

*Install python3*

*sudo amazon-linux-extras install pythonpip 3*

*curl -O* [*https://bootstrap.pypa.io/get-pip.py*](https://bootstrap.pypa.io/get-pip.py)

*python3 get-pip.py –user*

*pip –version*

*pip install awsebcli --upgrade –user*

*pip install awsebcli --upgrade –user*

*pip install selenium*

*pip install boto3 requests pandas numpy*

# Create directories and adjust the code

*sudo mkdir saswati*

*chmod –R 775 saswati*

*chown –R ec2-user saswati*

*cd saswati*

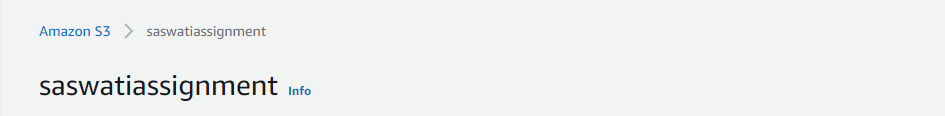
*mkdir log code tmp*

*cd code*

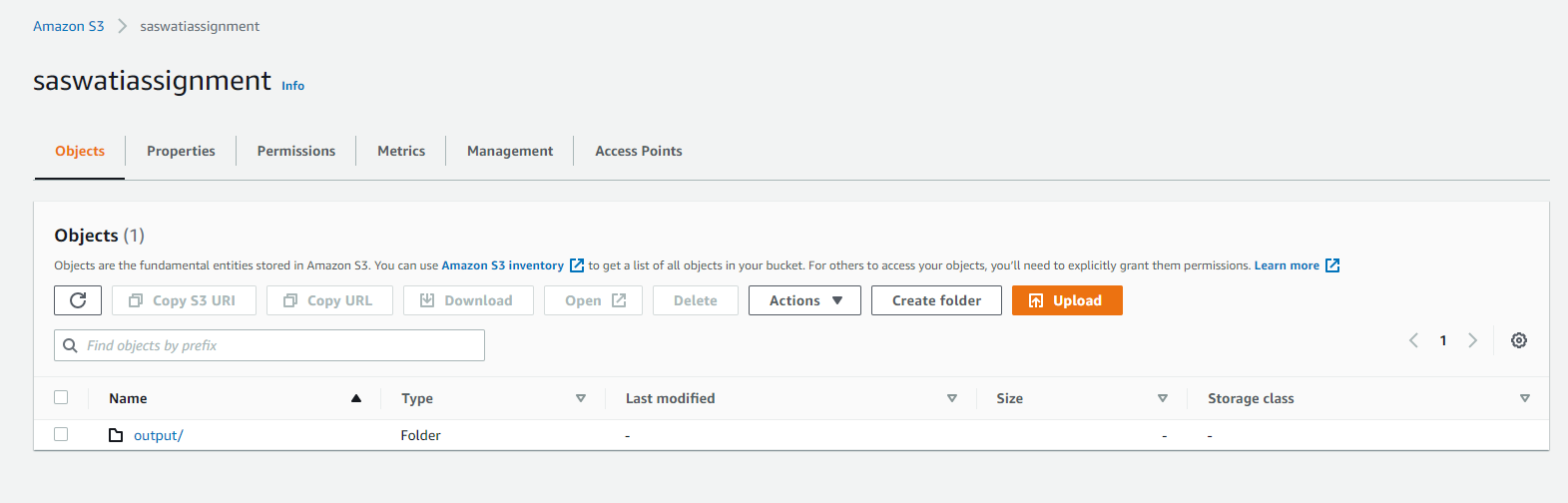
Copy the code that you have created in your local machine

# Setup S3 bucket

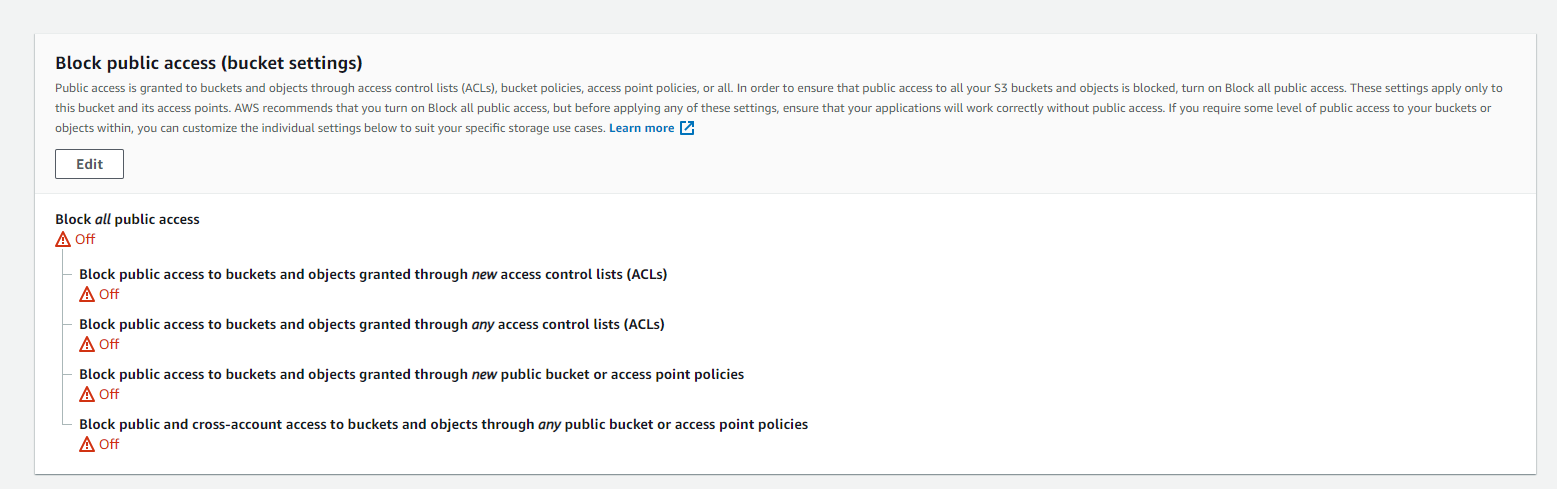
Create a public bucket saswatiassignment



Inside that create a folder output



Permission = public



# Adjust the code

I made below adjustments to the code for Linux.

1. Changed selenium chrome webdriver in silent mode

*chrome\_options.add\_argument("--headless")*

*chrome\_options.add\_argument("--silent")*

*chrome\_options.add\_argument('--no-sandbox')*

1. Added some sleep value as few button click code was failing –with this error message “stale element reference: element is not attached to page document”

*time.sleep(2)*

1. Adjusted chromedriver path, file output path pertaining to Linux
2. Finally added boto3 code to push the output to S3 bucket

*import boto3*

*REGION = 'us-east-1'*

*ACCESS\_KEY\_ID = 'xxxxxxxxxxxxxxxxxx'*

*SECRET\_ACCESS\_KEY = 'xxxxxxxxxxxxxxxxxxxx'*

*PATH\_IN\_COMPUTER = '/saswati/log/idaho\_license\_details.csv'*

*BUCKET\_NAME = 'saswatiassignment'*

*KEY = 'output/idaho\_license\_details.csv' # file path in S3*

*s3\_resource = boto3.resource(*

*'s3',*

*region\_name = REGION,*

*aws\_access\_key\_id = ACCESS\_KEY\_ID,*

*aws\_secret\_access\_key = SECRET\_ACCESS\_KEY*

*)*

*s3\_resource.Bucket(BUCKET\_NAME).put\_object(*

*Key = KEY,*

*Body = open(PATH\_IN\_COMPUTER, 'rb')*

*)*

# Setup Crontab

First to check if there is any crontab job running

sudo crontab -l

<no job found>

Add a new crontab

sudo crontab –e

<Add below lines for daily job at midnight>

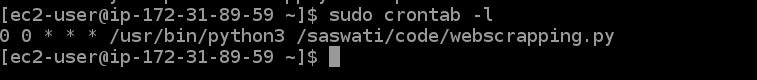
0 0 \* \* \* /usr/bin/python3 /saswati/code/webscrapping.py

Save the file



Verify

sudo crontab -l



# Output

Change file permission to public read

<https://saswatiassignment.s3.amazonaws.com/output/idaho_license_details.csv>