# Indeed Analysis for Data Science Job Postings

**Description:**

The main goal of the project is to implement an end to end pipeline for the analysis of job postings on Indeed. The pipeline consists of:

1. Data Scraping
2. Data Cleaning
3. Data analysis

The input data is acquired using Web Scraping, and the below analysis are performed on data with help of the pandas library:

1. Number of Data Scientist job openings per state in the United States.
2. Distribution of mean salary across different states.
3. Number of job openings for different specializations /categories in the Data science field, across different states.
4. Salary visualization across different job specializations for each state.

Workflow:

A close up of a map

Description automatically generated

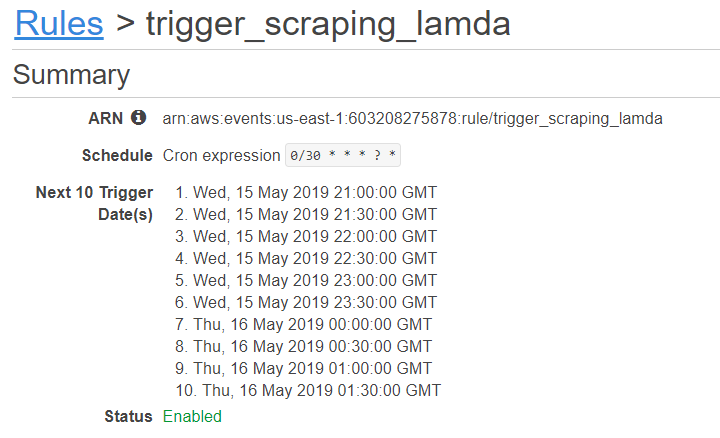
1. The user places the code and the virtual environment for the lambda function in a S3 bucket called ‘indeed-scraping-function’ . Please find the code attached below:



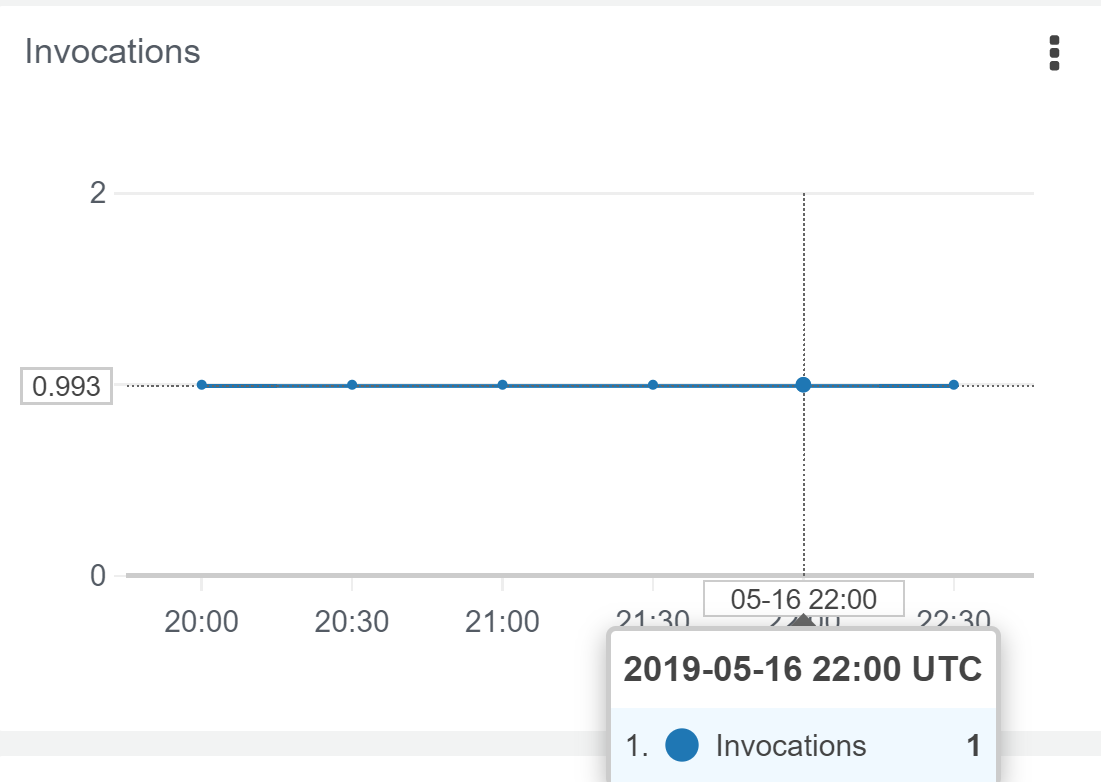
The same source code is placed in the directory ‘Analysis on AWS’ as part of the project deliverable (We have excluded the virtual environment from the project deliverable due to the huge file size)

1. The AWS CloudWatch is configured to trigger the lambda function every 30 mins

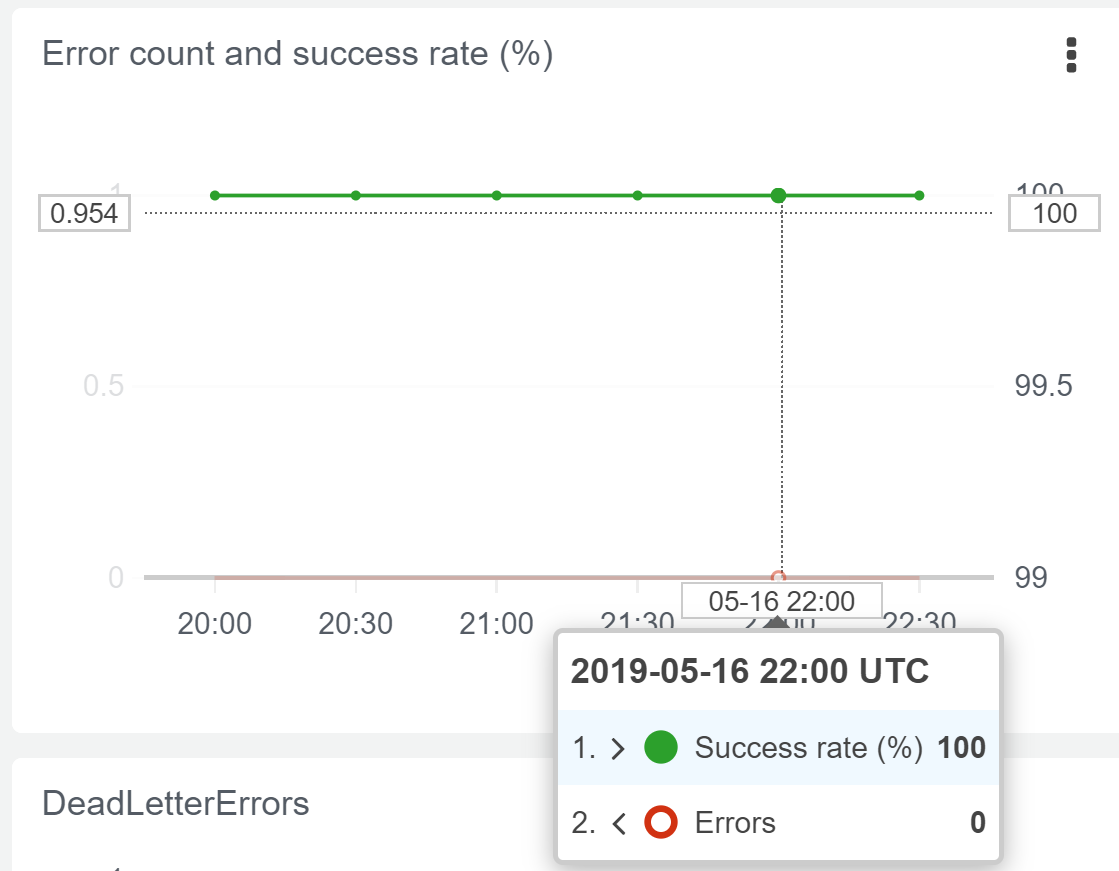
Cloudwatch trigger schedule:



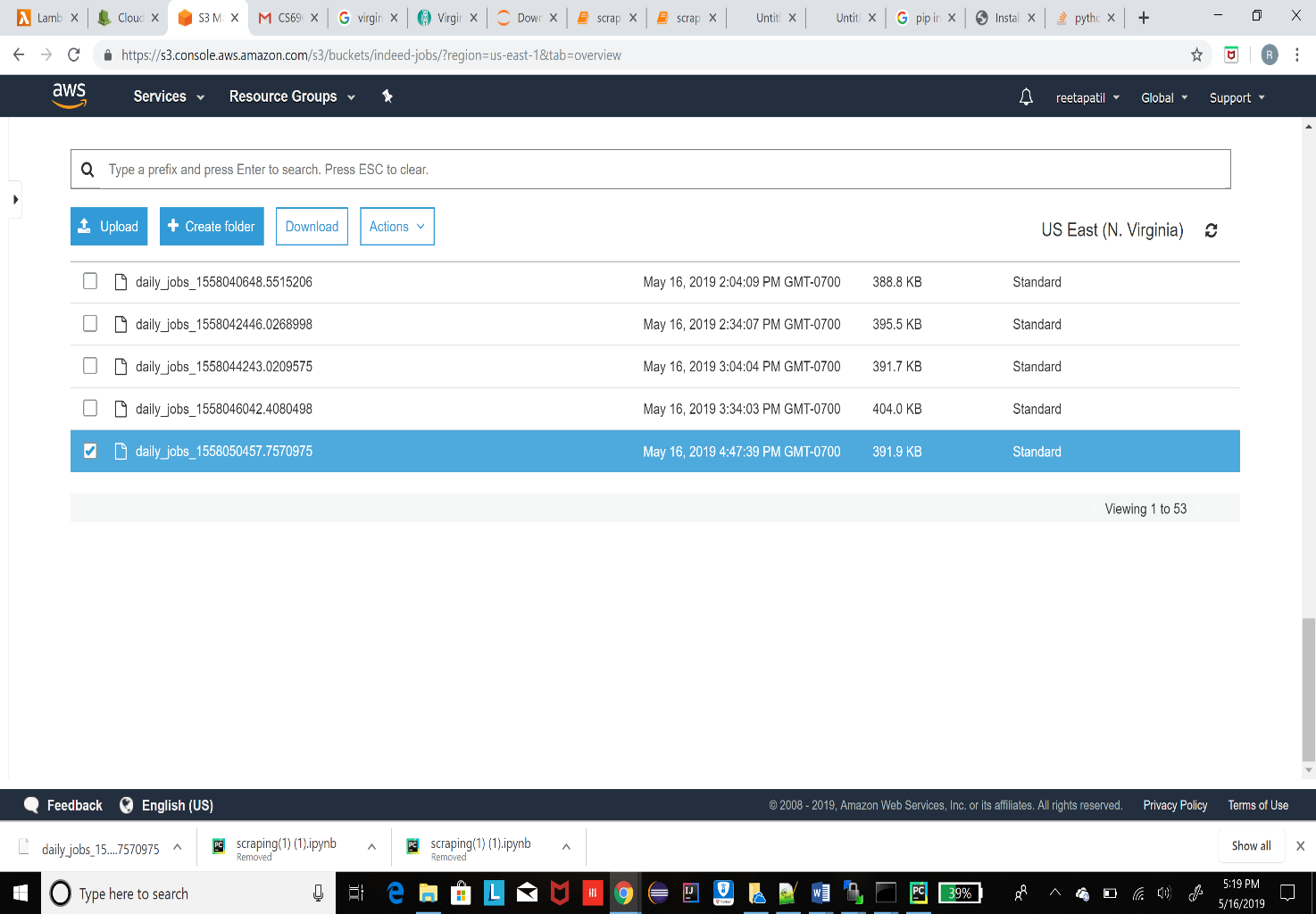
Cloudwatch invocations for the Lambda Function:



Cloudwatch success metrics for the Lambda invocation:



1. Once the CloudWatch is triggered the Lambda function scrapes the indeed data, and places the result in a file on S3 bucket (‘indeed-jobs’)





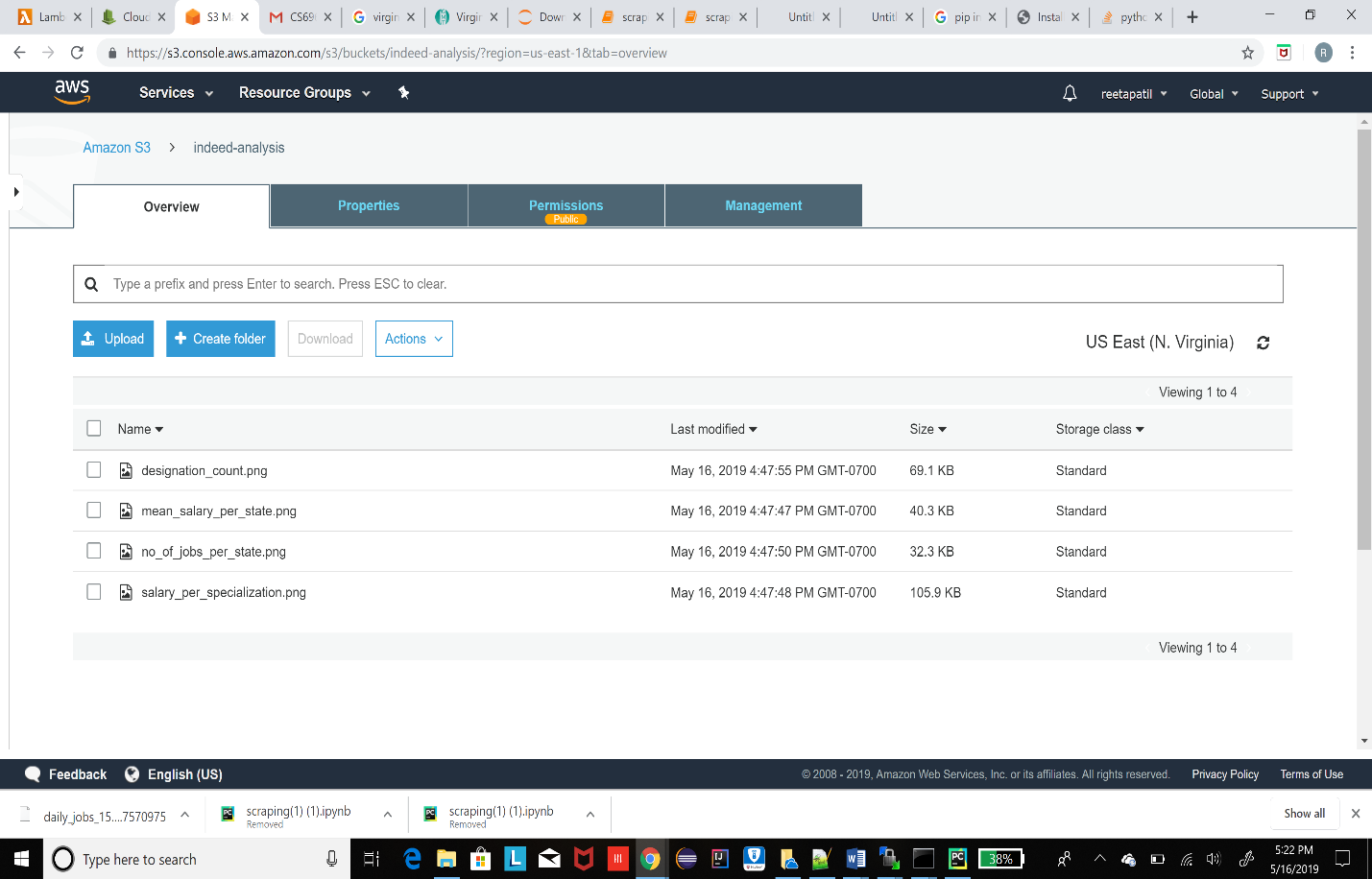
1. The above data is then preprocessed and analyzed in the same lambda function to generate graphs. The below are the link to the graphs of analysis on AWS:

<https://s3.amazonaws.com/indeed-analysis/no_of_jobs_per_state.png>

<https://s3.amazonaws.com/indeed-analysis/mean_salary_per_state.png>

<https://s3.amazonaws.com/indeed-analysis/designation_count.png>

<https://s3.amazonaws.com/indeed-analysis/salary_per_specialization.png>



**Instructions :**

Jupyter Notebook has been implemented to review the functionality of Indeed scraping and analysis without AWS. (Please refer the 'Analysis on Jupyter Notebook' directory for the same). The same code has been tweaked for AWS Implementation using Cloudwatch, Lambda and S3 Bucket (Please refer the 'Analysis on AWS' directory for the same).

To run the notebook , following modules are required:

1. Requests : $ **pip install requests**.
2. BeautifulSoup : **$ pip install bs4**

For running the lambda function on AWS, we have used the following additional modules:

1. Boto3
2. StringIO, BytesIO