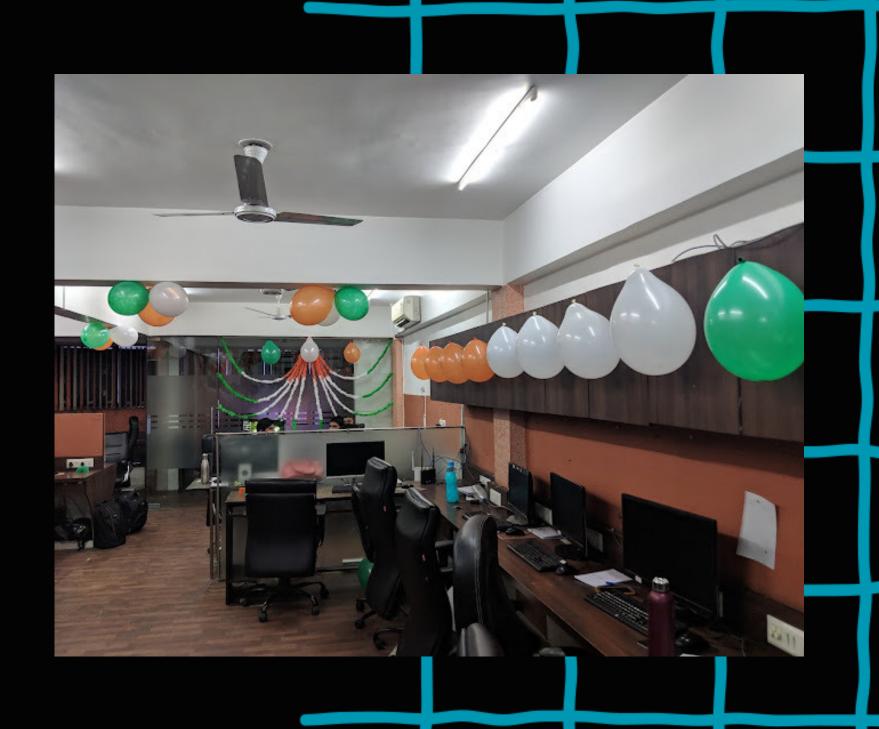
MODI SHRESHTHA(190130111081)

Anonymization pipeline



Eternal is a AWS partner web development based consulting company based in Ahmedabad and UK that provides cloud based product design and development solutions for businesses and startups





401 Satyam Mall Nr Mansi Circle, Satellite, Ahmedabad, Gujarat 380051 Page 3 of 24

Training









Taining







Week 7

Project Introduction and setting up the project enviournment

Week 8

Configuring the arx servers and anonymizing the dummy arx data

Weet



Installing python on Linux, virtual environments in python and the code skeleton of python, introduction to vscode intellisense ide and shortcuts for vscode



basic and nested loops, switch statements, basic calculator and a mad libs game



Data types, variables, typecasting

Desk Setup





String, list, dictionary and array methods such as slicing, indexing and appending



Functions and modules, global and local scope, basics of understanding the errors and debugging them in python

Taking user input and then building a calculator with the same

```
var1= input("Please enter the first variable")
var2=input("Please Enter the second variable")
operation=input("Please Enter the operation that you would like to perform with the variable")
def calculator (var1, var2):
     if (type(var1)and type(var2)==int):
        var1, var2, operation=var1, var2, operation
        if (operation != ['*' or '+' or '/' or '-' or '%']):
            print("Enter a valid function")
        else:
            if(operator== '+'):
                print(var1+var2)
            elif(operator == '-'):
                print(var1-var2)
            elif(operator== '*'):
                print(var1*var2)
            elif(operator == '/'):
                print(var1/var2)
            else:
               (varl%var2)
            return
```

Please enter the first variable2
Please Enter the second variable3
Please Enter the operation that you would like to perform with the variable'*'

01

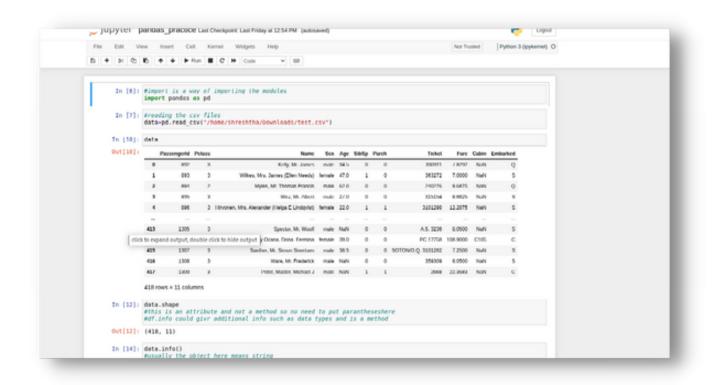
Classes and inheritance and understanding of scope

03

Global and local variables and iterators



Modules and various types of modules, introduction to the pandas and numpy modules



Pandas (importing the csv data, understanding the csv data, getting the columns and other methods), working with csv data

05

API and apis using python, pip and python file handling

```
script_name = sys.argv[0]
res = {
    "total_lines":"",
    "total_characters":"",
    "total_words":"",
    "unique_words":"",
    "special_characters":""
try:
    textfile = sys.argv[1]
    with open(textfile, "r", encoding = "utf_8") as f:
        data = f.read()
        res["total_lines"] = data.count(os.linesep)
        res["total_characters"] = len(data.replace(" ","")) - res["total_lines"]
        counter = collections.Counter(data.split())
        d = counter.most_common()
        res["total_words"] = sum([i[1] for i in d])
        res["unique_words"] = len([i[0] for i in d])
        special_chars = string.punctuation
```



Introduction to web scraping using beautiful soup using selenium



Python revision and breif reading for further resources



What is cloud computing, advantages of cloud as compared to on prem and major cloud providers, models of cloud computing, aws free tier, storage and networking



Introduction to aws and the aws free tier, advantages of aws over other servies



AWS storage and compute services such as ec2, s3 and elastic file sharing

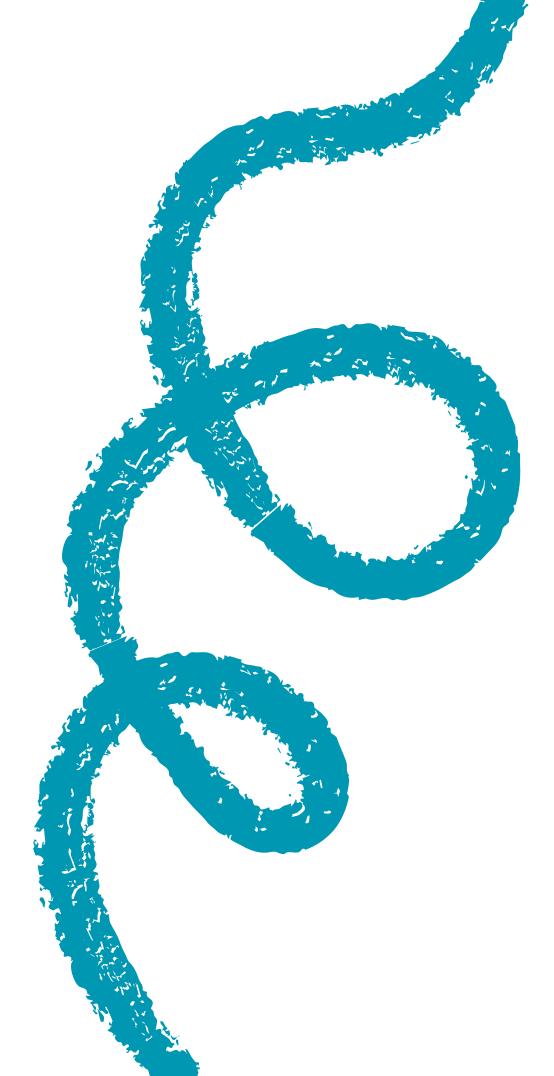
```
from selenium import webdriver
import csv
import time
items=[]
driver=webdriver.Chrome(r"C:/Users/hp/Anaconda3/chromedriver.exe")
driver.get('https://www.youtube.com/watch?v=iFPMz36std4')
driver.execute script('window.scrollTo(1, 500);')
#now wait let load the comments
time.sleep(5)
driver.execute_script('window.scrollTo(1, 3000);')
username_elems = driver.find_elements_by_xpath('//*[@id="author-text"]')
comment_elems = driver.find_elements_by_xpath('//*[@id="content-text"]')
for username, comment in zip(username_elems, comment_elems):
    item = {}
    item['Author'] = username.text
    item['Comment'] = comment.text
    items.append(item)
filename = 'C:/Users/hp/Desktop/commentlist.csv'
with open(filename, 'w', newline='', encoding='utf-8') as f:
    w = csv.DictWriter(f,['Author','Comment'])
   w.writeheader()
    for item in items:
        w writerow/item\
```





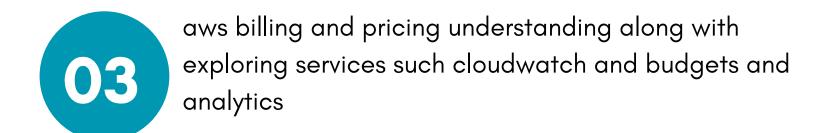
- Ol Asking the users for input in python
- O2 How the python interpreters work
- O3 Looping and branching to give structure to your code
- Making code reusable with functions and modules
- O5 Components of clean code and how to write clean code
- O6 Understanding complex open source code and how to write code that others will understand

- O7 Getting data from the web and analyzing it
- one of modules in python and how modules make a developers life easier
- O9 How to ask for help and talk in technical terms
- Opening Analyzing data





Understanding the databases in aws and understanding how to deploy a static website on aws+ advanced s3





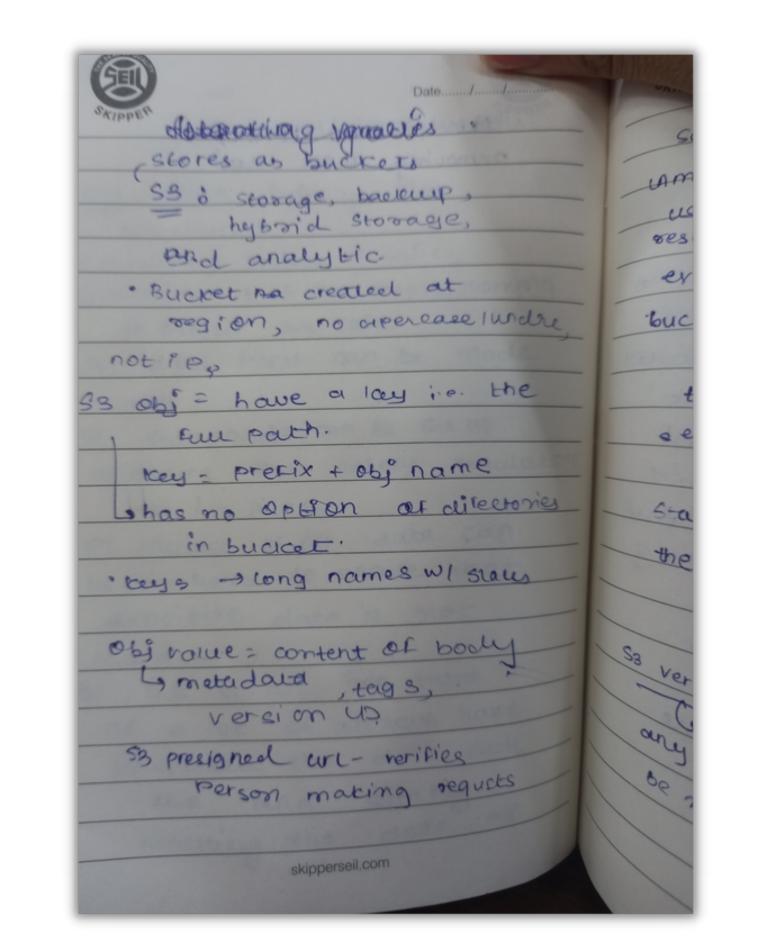
Introduction to fully managed services in aws such as lambda and fargate



AWS free tier introduction and introduction to free tier of various services and how to stay in free tier



Introduction to the shared responsibility model of the aws, aws IAM and roles



AE AE
Opening the second seco
de la cale
Scaration and holy
resoure
scerity rosoure weer band
user par can access
user permiss or access it it
- Plicit denu
Explicit dery bucket policies
- Point
bucket policies - Principal- account or user to apply
the to copply
the policy to . Actions.
eet at Alp to
set- af Alpl to alwro 1 clery
Ciari Clery
Static website - " "
DPGOLL
the index on the
Lowproad
THE THE PARTY OF T
S8 versionie
Signal some succest level
Constitute of the second
The state of the s
File hosen
be nuev. version will
acc.



Project breif and aws cdn services and relaiblity



Misc aws services such as sns, inspector and macie for threat detection, Introduction to docker



Setting up docker on the local instance, docker components, architecture and docker basics



Frequently used docker commands and building a local docker instance from an image



What is version control, various types of version control systems and the difference between git and github, setting up your own github account

sudo apt-get install docker-ce docker-ce-cli containerd.io docker-buildx-plugin docker-compose-plugin

The docker is now installed. Can verify the install using hello world. Issues faced:

Dependencies installation error

(had to reinstall curl and sudo)

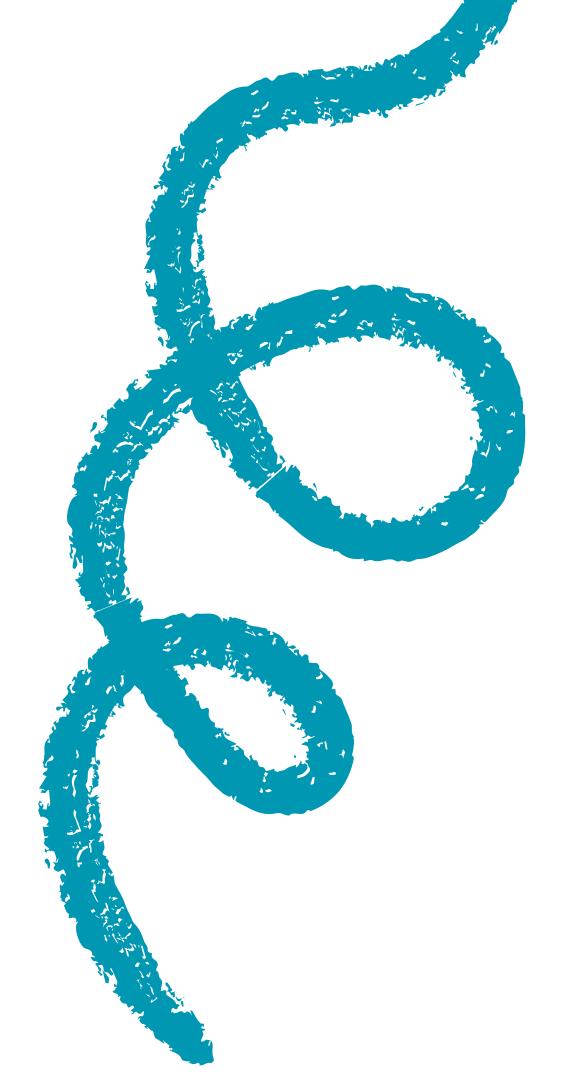
My system had <u>virtualiation</u> and kvm not installed hence had to reinstall it

After getting the docker set up and running, we can use:

https://github.com/navikt/arxaas to install the docker image on the local server:

Before running the instance, we have to login as root first and that can be done using:

Sudo -s



Page 12 of 24

Learnings so far

- Learning about where our datagoes from the website and howit is all stored
- O2 Advantages of cloud as compared to on prem
- 7 Types of cloud providers and the cloud race
- Amazon Wbe services and why they are dominating the industry

- Various aspects and services of aws
- O8 How cloud can stack up bills quickly and how to save money on them
- O9 Some use cases are better suited for cloud and someon prem
- integrating various parts of the project with cloud

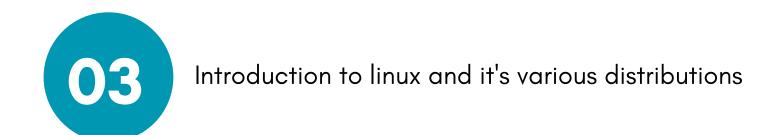
01

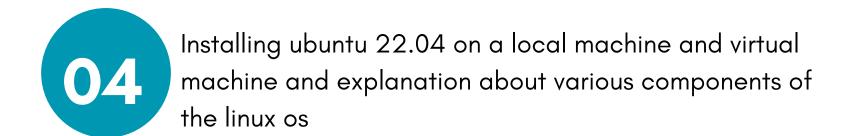
Frequently used github commands to clone a repo, push and pull

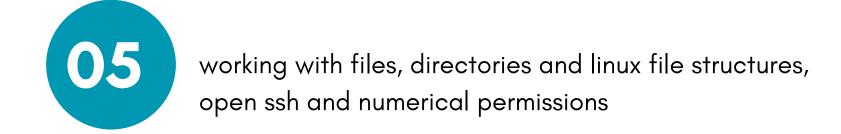


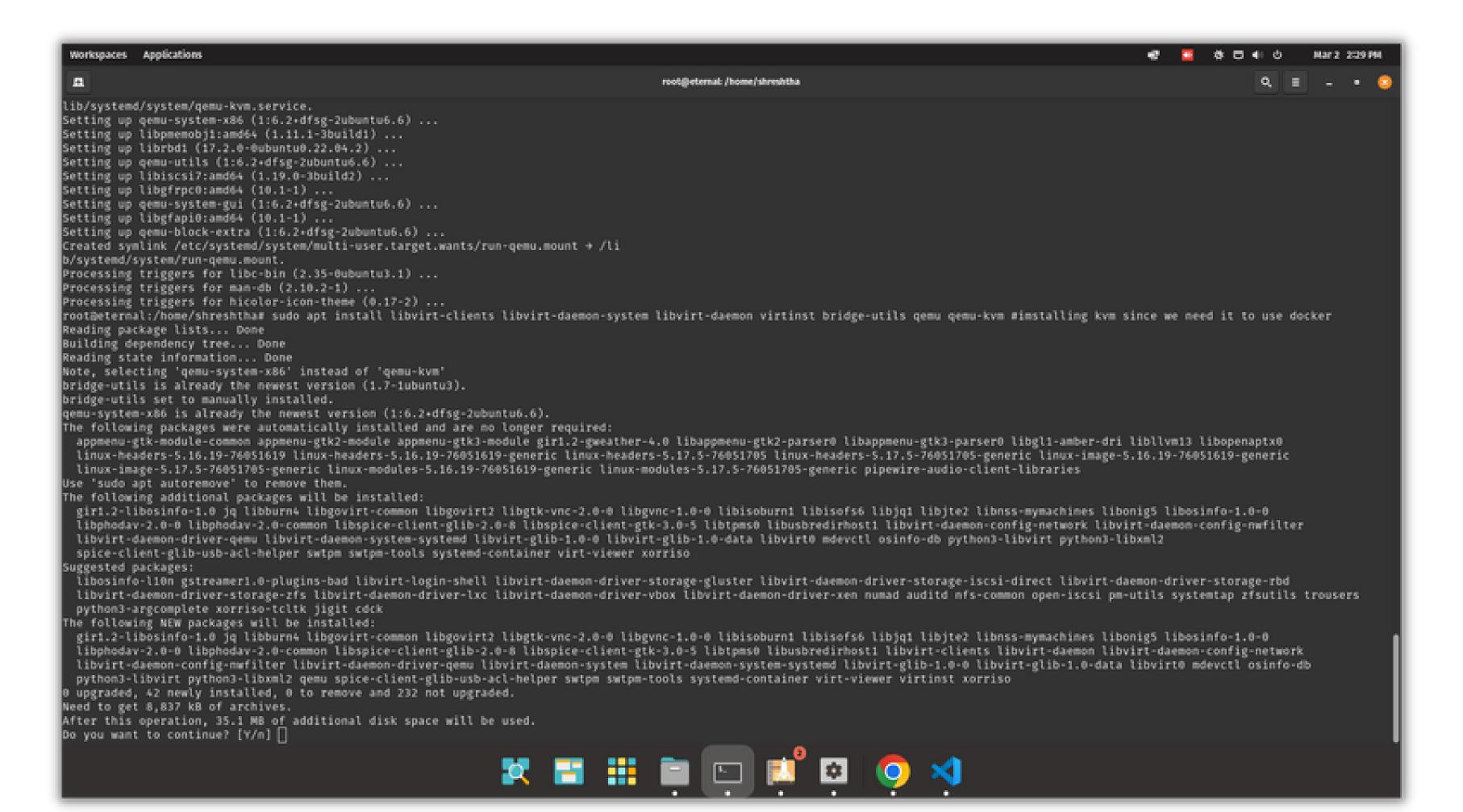
Introduction to Operating Systems and kernels













shell training to learn more about installing and updating the packages, debugging dependency errors and using regex in shell to automate things, package manager



IManaging system units and logs, and managing users, scp and rsync



Project breif and project setup with the practical use case

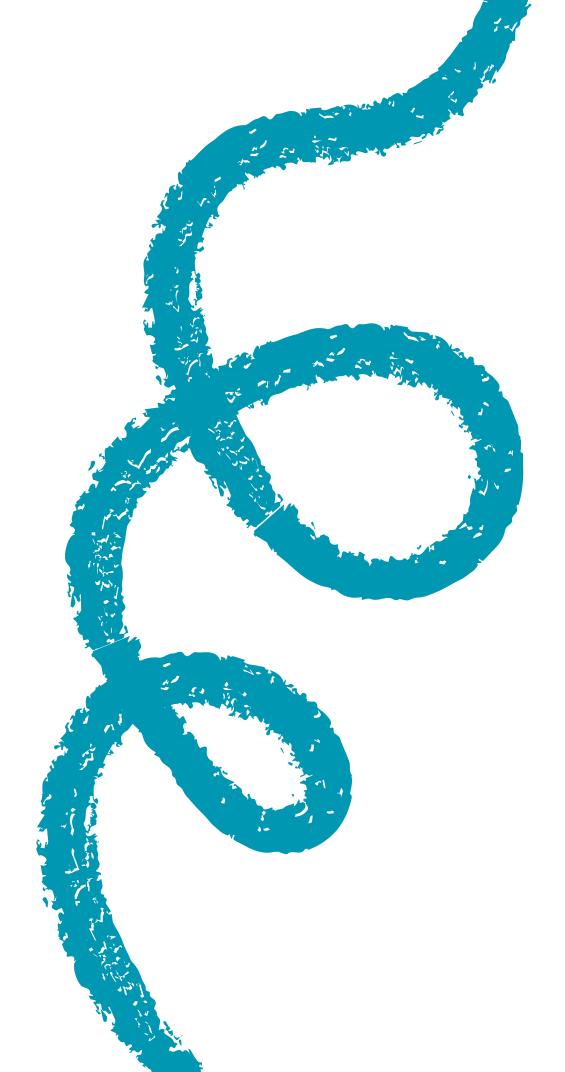


Understanding the scope of the project and reading about various types of anonymization techniques in aws



Installing arx

```
#importing the neccesary modules and functions
import numpy as np
import pandas as pd
import pyarxaas
from pyarxaas import ARXaaS
from pyarxaas.hierarchy import IntervalHierarchyBuilder, OrderHierarchyBuilder
from pyarxaas import AttributeType
from pyarxaas.privacy_models import KAnonymity
from pyarxaas import Dataset
#creating a local arx instance, can be created using running the docker image locally
arxaas = ARXaaS("http://localhost:8080/")
#print(data)
#loading the data which contains the zipcodes
data=pd.read_csv("/home/shreshtha/Documents/zipcode.csv", header=None, usecols=[0], names=['zipcode'])
dataset = Dataset.from_pandas(data)
#loading the zipcode hierarchies present in the csv files, if dont want to define manually then can be done automatically with arx
zipcode_hierarchy=pd.read_csv("/home/shreshtha/Documents/zipcode.csv", header=None, usecols=[1,2,3,4,5])
#print(zipcode_hierarchy.head)
#setting the attribute types for the dataset column
dataset.set_attribute_type(AttributeType.QUASIIDENTIFYING)
#creating a risk profile
risk_profile = arxaas.risk_profile(dataset)
#reidentification risk states how much the data is at the risk of reidentifying
#print(risk_profile.re_identification_risk)
#attacker success rate tells you how much the data can be attacked by which model of attacking
#print(risk_profile.attacker_success_rate)
print(risk_profile.population_model)
```



Page 16 of 24

Learnings so far

V	arious	types	of o	perating
---	--------	-------	------	----------

- 01 system and computer architecture
- 02 What is linux
- 75 The beauty of open source
- $\bigcap A$ How to learn tough concepts
- 05 Moving away from windows and getting used to the cli

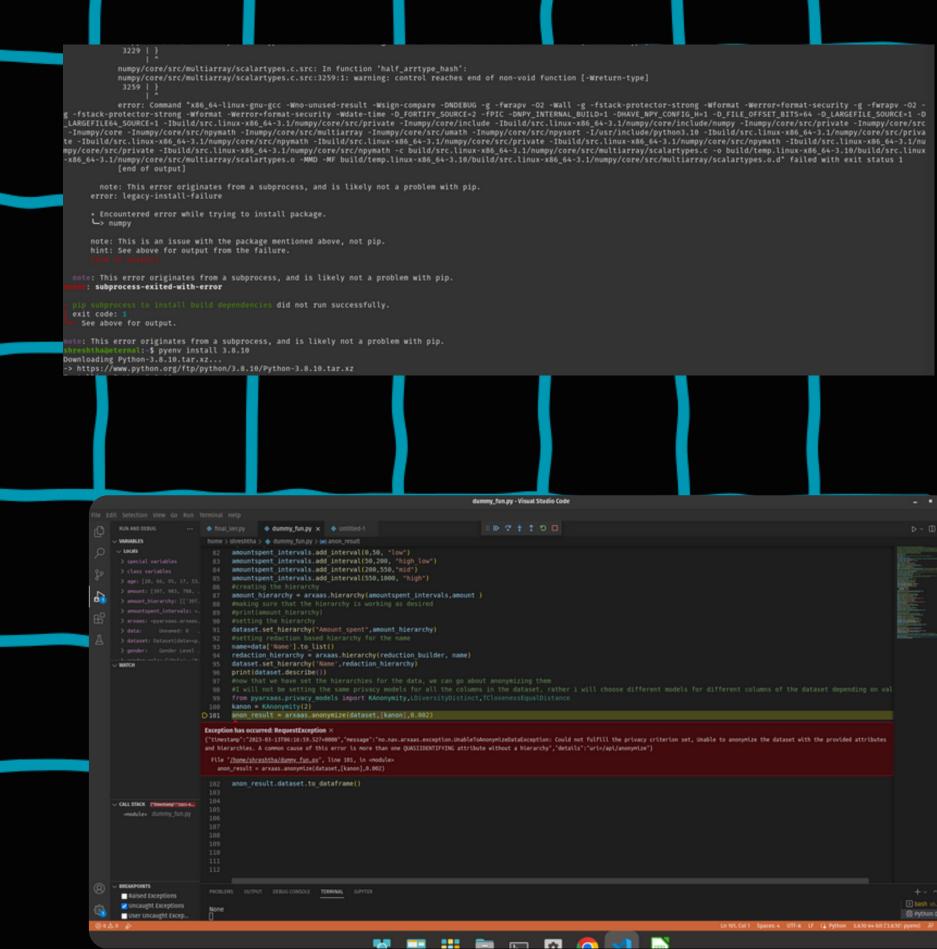
- Understanding how linux works and the flexiblity it provides
- O7 IDocker and the importance of container technology
- 1 Importance of version control in the life of a developer
- 09 If it works, don't touch it

Project Abstract

Internet has an approximate of 4.6 billion users worldwide and they are growing at an astronomical rate. With the rise in users, privacy concerns also pop up. Data privacy is a complex area which happens to be a vital part of the internet ecosystem. This project aims at protecting the privacy of the user data by anonymizing the data using various statistical tools and techniques and proposing an end to end solution of the anonymized data on the AWS. The proposed solution's schema can be modified according to the need of the user and the type of the data with ease

Challanges

Although facing challanges is a common part of programming and it more often than not indicates that you are going in the right direction, sometimes the challanges can often be a roadblock to the further progress. One of the biggest challanges i faced while working for internship till date was trying to configure the pyarxaas module. The module has been outdated with no support for the latest python version or the bugs. Having a system which runs on python 3.10 i could not downgrade the python version globally as that would mean dependency error for a lot of the system default modules, i had to figure out how to have multiple versions of python in the same memory limited system along with making sure that the code and the dependency are not broken



Thank You So Much!