# README

# Pollution Data API

## About

This repository contains code for accessing and analysing live pollution data from the Indian government's pollution API. The API provides information about the pollution levels in majorly polluted states/counties where receiving stations are located. The data is updated every hour, providing users with fresh information about the current state of pollution in India.

## Usage

To use this code, you will need to have access to the API provided by the Indian government. Once you have access to the API, you can use the code in this repository to access and analyse the data. The code is written in Python and is designed to be easily modifiable to suit your needs.

To get started, you will need to clone this repository to your local machine. Once you have done that, you can open the code in your favourite Python editor and begin exploring the pollution data. The code is well-commented and should be easy to follow.

API Used:

[https://api.data.gov.in/resource/3b01bcb8-0b14-4abf-b6f2-c1bfd384ba69?api-key=579b464db66ec23bdd0000017789d1ca79674d856cd0d050548fe535&format=json&offset=0&filters[country]=India&limit=3000](https://api.data.gov.in/resource/3b01bcb8-0b14-4abf-b6f2-c1bfd384ba69?api-key=579b464db66ec23bdd0000017789d1ca79674d856cd0d050548fe535&format=json&offset=0&filters%5bcountry%5d=India&limit=3000)

## Pre-requisites

Visual Studio- 1.77.3

Python version- 3.11

Setting up the python path: C:\Users\Manu\AppData\Local\Programs\Python\Python311

Dagster Version- 1.3.1

Protobuff- 3.20.0

MySql server in openstack- 8.0.32

PostgresSql server in openstack- 11.19

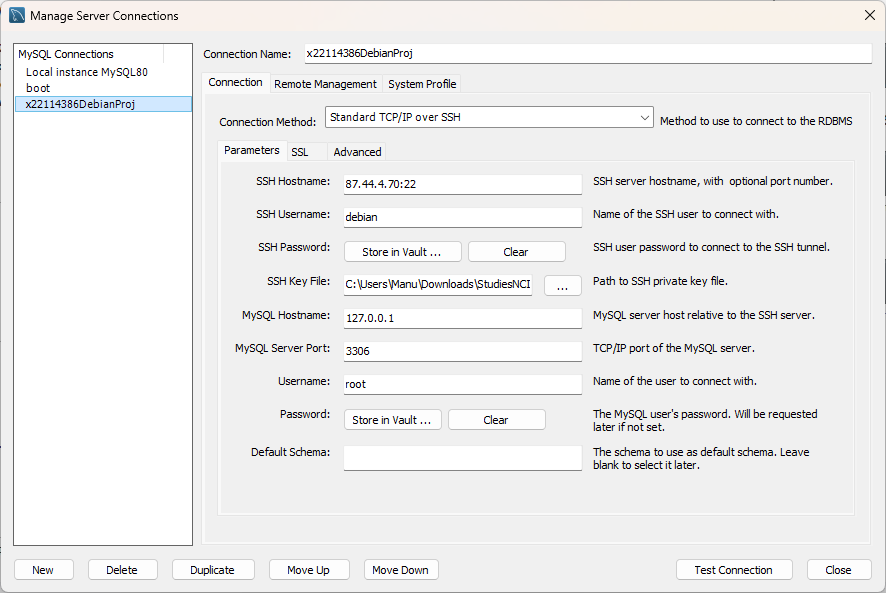
PEM file in the link – new.pem

PPK file in the link – new.ppk

Debian server in openstack- 10.0

Debian Instance- [x22114386Project](https://openstack.cloudenci.ie/project/instances/b22b1dab-19d0-494f-b5c0-9616dfb7ea1b/) - 87.44.4.70

* Mysql server details and database details:



* Postgress Server and database details:

user = "dap",

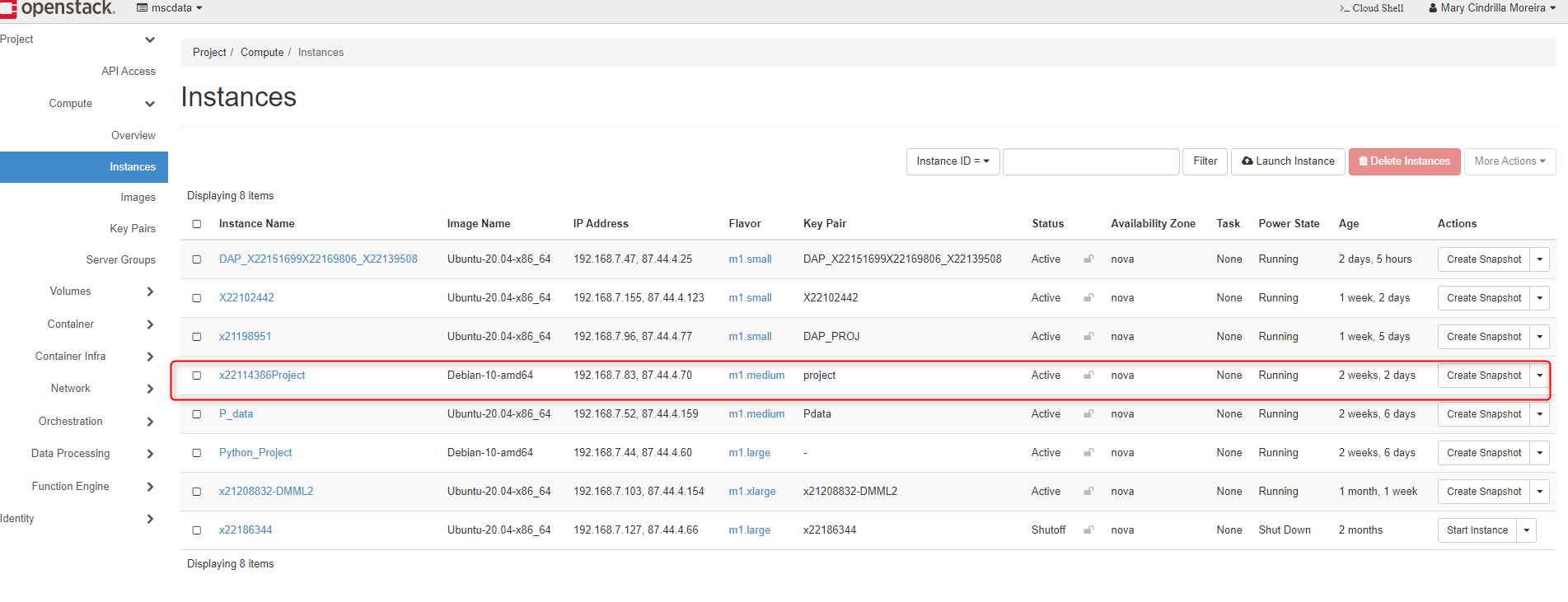
password = "dap",

host = "87.44.4.70",

port = "5432",

database = "postgres"

* Openstack cloud



More insights on the DB connections are attached in file “MySQL Server DATA BASE CONNECTION DETAILS.DOCX” in the file structure

## Sequence of execution of Program

1. **1\_X22114386Project\_API. ipynb:**

This is the first file to execute this file consist of

a. Triggering the open API

b. Extracting the Json Output to a CSV file in the local path .The output file is competation.csv

1. **2\_X22114386DataBase. ipynb:**

The code is related to Data base connection

a. Connection string to create a connection pipe line to Database

b. Creation of Database Schema

c. Dump the entire competation.csv file to the table ""

d. Nomalization of the data

e. Insertion of data with respect to tables such as "CountryTable","StateTable","CityTable","PollutionDataTable"

1. **X22114386SettingUpPostgresDB.ipynb**

This code gives the connectivity to postgress pipe line in case we are unable to create a pipeline in dagster just an intermediate file for reference

1. **3\_dagsterPipeLine\_final.py**

The path of this file is in the dagster virtual environment:

\DataBase\_And\_Analytics\_Project\airPollution

To set the Dagster pipeline and create the environment for dagster use the below commands

Creating a Dagster environment

- pip install virtualenv for creating a virtualenv

- python -m venv env

- pip install dagster dagit- installs the library we need to work with

Creation of project in dagster

- dagster project scaffold --name airPollution

Setting up a dagster home

-dagit

Make sure you set the path

DAGSTER\_HOME (for Logs)

C:\Users\Manu\Downloads\StudiesNCI\MichealBrad\_DB\projects\DataBase\_And\_Analytics\_Project\

Scripts that set in ENV file of DAGSTER C:\Users\Manu\AppData\Local\Programs\Python\Python39\Scripts

**Dbcon.py**

Supports all the DB connection files that is necessary for the Dagster pipeline connectivity

After the Dagster pipeline we will be able to see the cleaned dataset into postgreSql

1. **4\_CleaningAndAnalysisOfdata.ipynb**

The CSV file is taken and analysis of the file is been done

Post this all the 4 dataset is merged, analysed on a common ground.

MergedDataset.csv- merged all the 4datasets in git hub master branch

X2215392MergingDatasets- analysed dataset python code in git hub master branch

## Other files in the folder structure in Cindrilla branch github

1. competitions.csv- sample CSV data
2. airPollution- this is the dagster virtual folder
3. DagsterLogs- dagster logs
4. ArcheitectureDiagramOfFirstSemProject- architecture of the project
5. get-pip- pip file
6. FewLinesAboutTheDataSet.txt- about the dataset
7. Env- this folder has supporting libraries for dagster
8. MySQL Server DATA BASE CONNECTION DETAILS- for the data base connectivity using ssh tunnel
9. New.pem- private key for ssh tunnelling
10. New.ppk- for putty connection