A Kibe Creation

I. Setting up a Data Engineering Infrastructure: A step-by-step guide

Installations and configurations on windows 10

a. **Apache Nifi**
b. **Apache Airflow**
c. **Elasticsearch**
d. **Kibana**
e. **PostgreSQL**
f. **pgAdmin 4**

1. Installing and configuring Apache Nifi

Curl https://mirrors.estointernet.in/apache/nifi/1.12.1/nifi-1.12.1-bin.tar.gz #cmd

https://www.clearpeaks.com/installing-apache-nifi-on-windows/ #for window

Install a valid Java Development Kit

Download JDK from https://www.oracle.com/ae/java/technologies/javase/javase8u211-later-archive-downloads.html

Installation

Open ~ **edit the system variable environment** ~ from the search bar ~ click on **environment Variable ~** select Path **c://windows \system32 .. .. wwbem** then click **edit ~ new ~ then add java path then okay – open cmd and type javac to confirm installation**

Download: JDK Development Kit ~
https://www.oracle.com/ke/java/technologies/downloads/#java22

Starting Nifi

Download Nifi from https://nifi.apache.org/download/

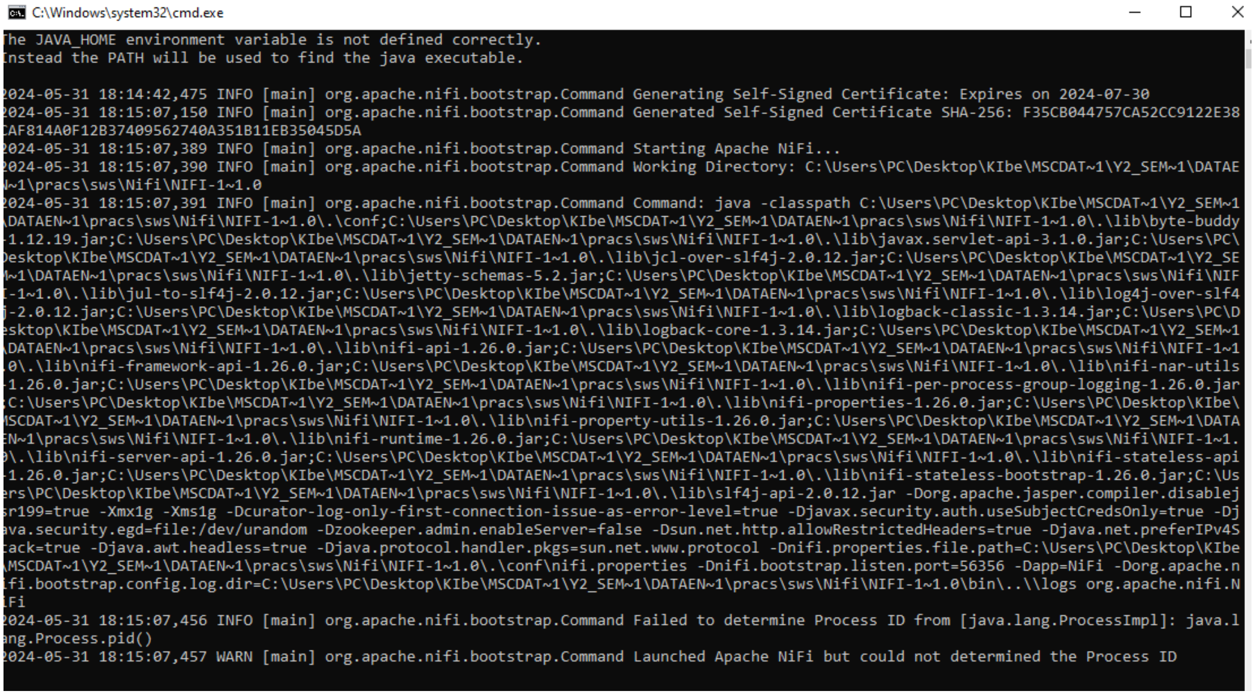
Extract the compressed Nifi file, then go to bin, then click or run nifi.bat file it should start nifi environment ... then access the **Nifi UI** from ***localhost:8443/nifi/***

Successful after many trials

Generated Username [2dc3de96-af49-46b8-ab1d-b2c52483e711]

Generated Password [t9IzC9H3BESmG7Pl6IyvvqBEsKHg+n8K]

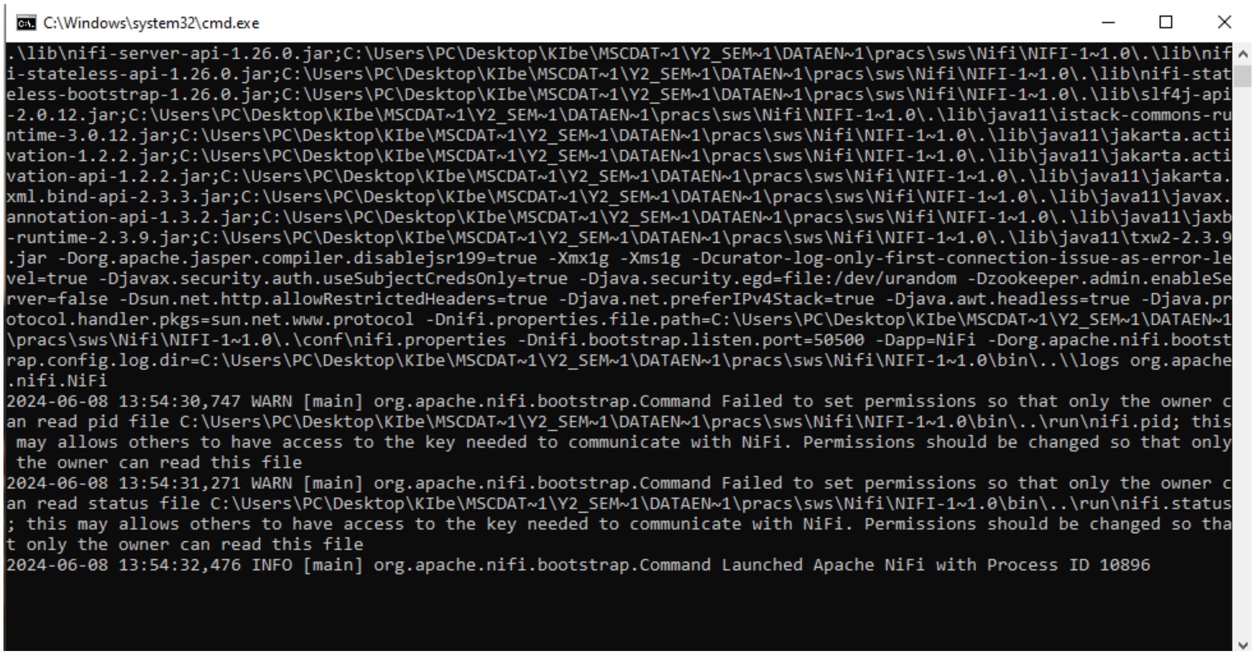
Problem at a glance ~ latest version did not process an ID



```
he JAVA_HOME environment variable is not defined correctly.
nstead the PATH will be used to find the java executable.

024-05-31 18:14:42,475 INFO [main] org.apache.nifi.bootstrap.Command Generating Self-Signed Certificate: Expires on 2024-07-30
024-05-31 18:15:07,150 INFO [main] org.apache.nifi.bootstrap.Command Generated Self-Signed Certificate SHA-256: F35CB044757CA52CC9122E38
AF814A0F12B37409562740A351B11EB35045D5A
024-05-31 18:15:07,389 INFO [main] org.apache.nifi.bootstrap.Command Starting Apache NiFi...
024-05-31 18:15:07,390 INFO [main] org.apache.nifi.bootstrap.Command Working Directory: C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAE
~1\pracs\sws\Nifi\NIFI-1~1.0
024-05-31 18:15:07,391 INFO [main] org.apache.nifi.bootstrap.Command Command: java -classpath C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1
DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\conf;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\byte-buddy
1.12.19.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\javax.servlet-api-3.1.0.jar;C:\Users\PC\
esktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\jcl-over-slf4j-2.0.12.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SE
~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\jetty-schemas-5.2.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIF
-1~1.0\.\lib\jul-to-slf4j-2.0.12.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\log4j-over-slf4
-2.0.12.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\logback-classic-1.3.14.jar;C:\Users\PC\D
sktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\logback-core-1.3.14.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1
DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-api-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1
0\.\lib\nifi-framework-api-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-nar-utils
1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-per-process-group-logging-1.26.0.jar
C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-properties-1.26.0.jar;C:\Users\PC\Desktop\KIbe\
SCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-property-utils-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATA
N~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-runtime-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.
\.\lib\nifi-server-api-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-stateless-api
1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-stateless-bootstrap-1.26.0.jar;C:\Us
rs\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\slf4j-api-2.0.12.jar -Dorg.apache.jasper.compiler.disablej
r199=true -Xmx1g -Xms1g -Dcurator-log-only-first-connection-issue-as-error-level=true -Djavax.security.auth.useSubjectCredsOnly=true -Dj
va.security.egd=file:/dev/urandom -Dzookeeper.admin.enableServer=false -Dsun.net.http.allowRestrictedHeaders=true -Djava.net.preferIPv4S
ack=true -Djava.awt.headless=true -Djava.protocol.handler.pkgs=sun.net.www.protocol -Dnifi.properties.file.path=C:\Users\PC\Desktop\KIbe
MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\conf\nifi.properties -Dnifi.bootstrap.listen.port=56356 -Dapp=NiFi -Dorg.apache.n
fi.bootstrap.config.log.dir=C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\bin\..\\logs org.apache.nifi.N
Fi
024-05-31 18:15:07,456 INFO [main] org.apache.nifi.bootstrap.Command Failed to determine Process ID from [java.lang.ProcessImpl]: java.l
ng.Process.pid()
024-05-31 18:15:07,457 WARN [main] org.apache.nifi.bootstrap.Command Launched Apache NiFi but could not determined the Process ID
```

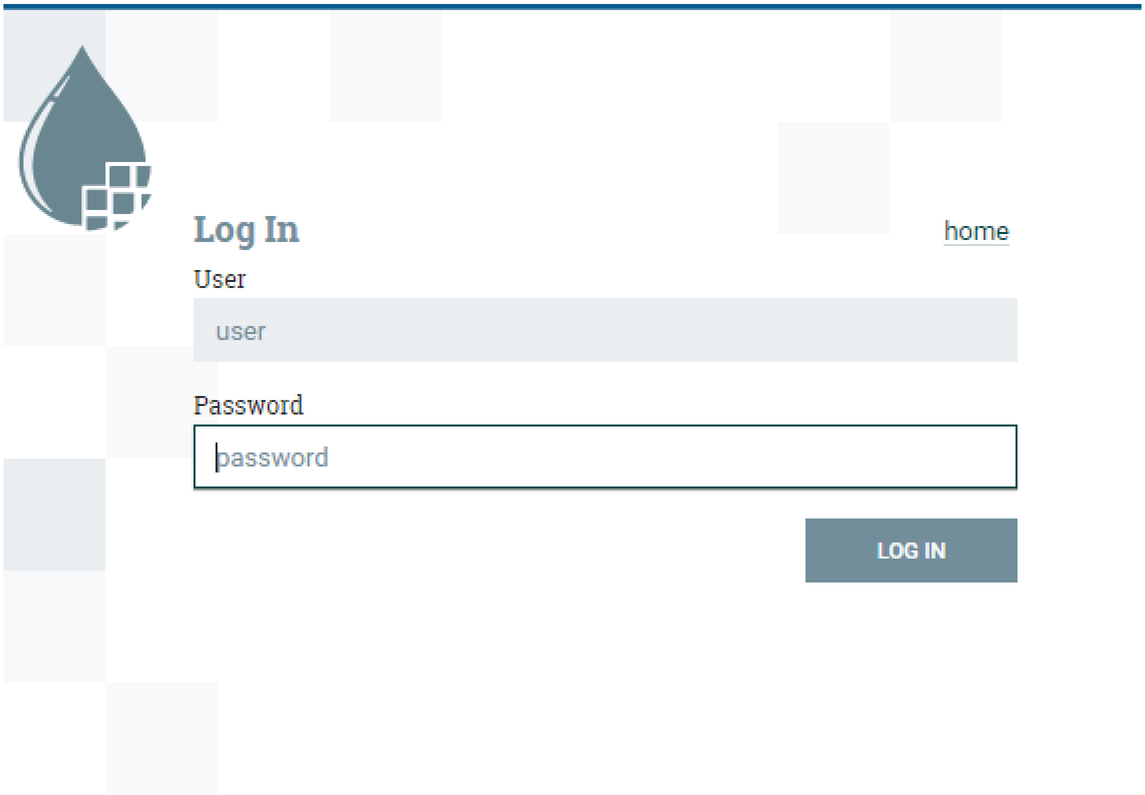
Figure 1: Nifi Issues

Processed ID



```
.\lib\nifi-server-api-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nif
i-stateless-api-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\nifi-stat
eless-bootstrap-1.26.0.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\slf4j-api
-2.0.12.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\java11\istack-commons-ru
ntime-3.0.12.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\java11\jakarta.acti
vation-1.2.2.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\java11\jakarta.acti
vation-api-1.2.2.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\java11\jakarta.
xml.bind-api-2.3.3.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\java11\javax.
annotation-api-1.3.2.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\java11\jaxb
-runtime-2.3.9.jar;C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\.\lib\java11\txw2-2.3.9
.jar -Dorg.apache.jasper.compiler.disablejsr199=true -Xmx1g -Xms1g -Dcurator-log-only-first-connection-issue-as-error-le
vel=true -Djavax.security.auth.useSubjectCredsOnly=true -Djava.security.egd=file:/dev/urandom -Dzookeeper.admin.enableSe
rver=false -Dsun.net.http.allowRestrictedHeaders=true -Djava.net.preferIPv4Stack=true -Djava.awt.headless=true -Djava.pr
otocol.handler.pkgs=sun.net.www.protocol -Dnifi.properties.file.path=C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1
\pracs\sws\Nifi\NIFI-1~1.0\.\conf\nifi.properties -Dnifi.bootstrap.listen.port=50500 -Dapp=NiFi -Dorg.apache.nifi.bootst
rap.config.log.dir=C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\bin\..\\logs org.apache
.nifi.NiFi
2024-06-08 13:54:30,747 WARN [main] org.apache.nifi.bootstrap.Command Failed to set permissions so that only the owner c
an read pid file C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\bin\..\run\nifi.pid; this
 may allows others to have access to the key needed to communicate with NiFi. Permissions should be changed so that only
 the owner can read this file
2024-06-08 13:54:31,271 WARN [main] org.apache.nifi.bootstrap.Command Failed to set permissions so that only the owner c
an read status file C:\Users\PC\Desktop\KIbe\MSCDAT~1\Y2_SEM~1\DATAEN~1\pracs\sws\Nifi\NIFI-1~1.0\bin\..\run\nifi.status
; this may allows others to have access to the key needed to communicate with NiFi. Permissions should be changed so tha
t only the owner can read this file
2024-06-08 13:54:32,476 INFO [main] org.apache.nifi.bootstrap.Command Launched Apache NiFi with Process ID 10896
```

Successful: https://localhost:8443/nifi/



Log In

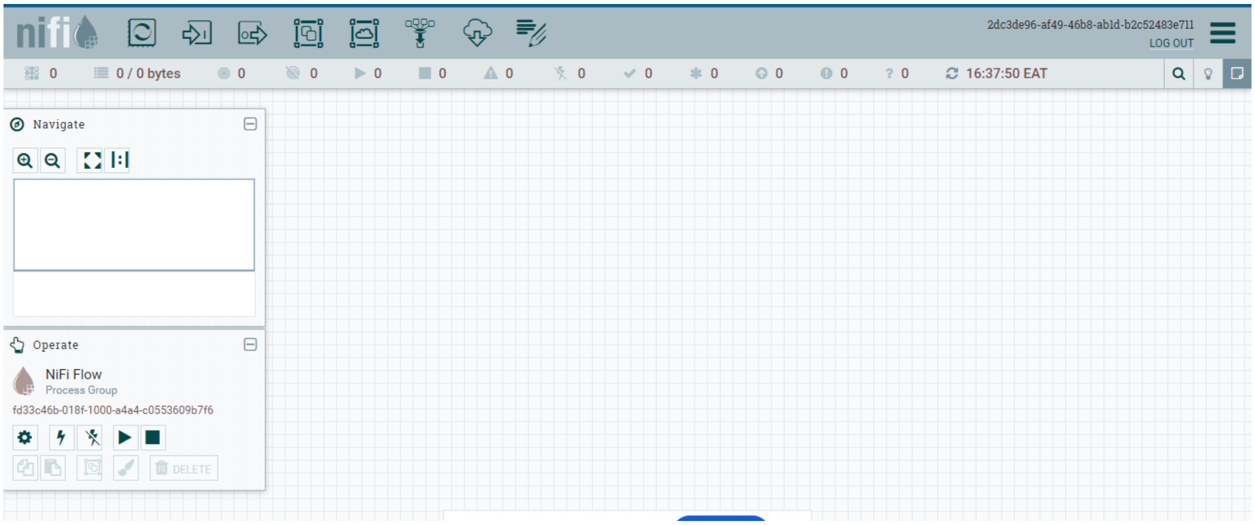
home

User

user

Password

password



2dc3de96-af49-46b8-ab1d-b2c52483e711

NiFi Flow
Process Group
fd33c46b-018f-1000-a4a4-c0553609b7f6

2. Installing and configuring Apache Airflow

Tools needed

- Docker
- Visual studio Code

-Download and install docker – follow the serious SQL procedure (very hectic and stubborn this one)

- Set up Apache Airflow

Save file; https://airflow.apache.org/docs/apache-airflow/2.5.1/docker-compose.yaml as a .yaml file otherwise it will not work. You can use VS code to edit. Save file in a folder then open said folder from VS code environment.

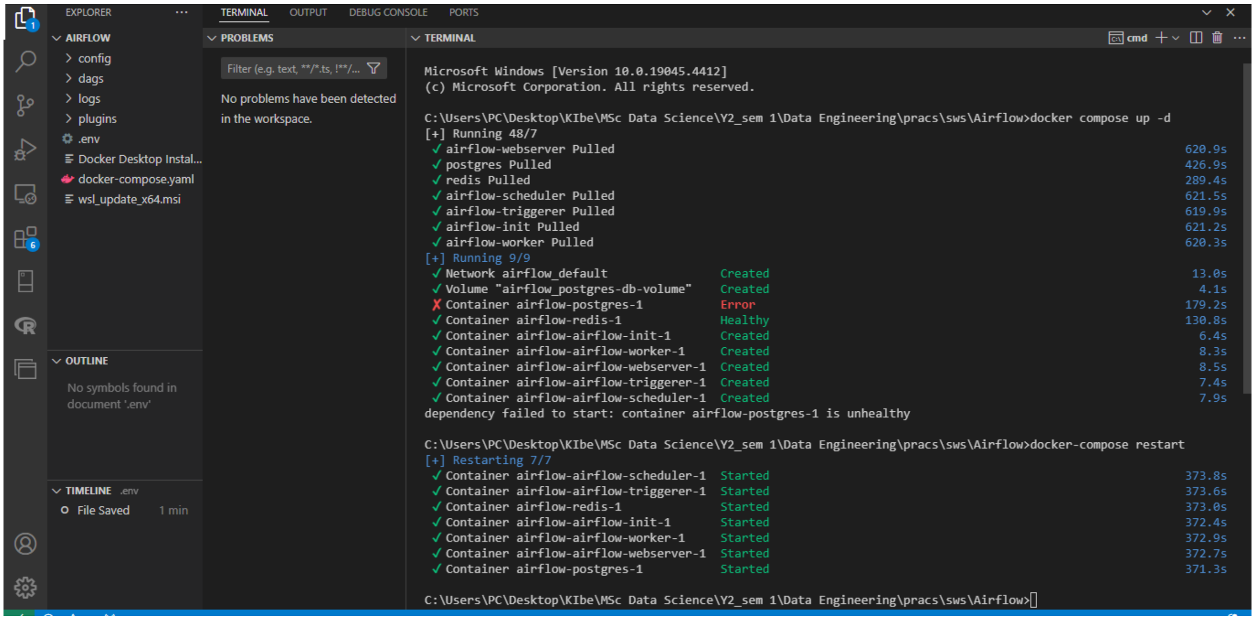
Create a .env file then paste below code and save.

AIRFLOW_IMAGE_NAME= apache/airflow:2.4.2

AIRFLOW_UID=50000

Open a terminal and run ***docker compose up -d***, these should pull all relevant resources and create an airflow container on docker, executable through ***Localhost:8080/Login/***

Problem



No problems have been detected in the workspace.

```
Microsoft Windows [Version 10.0.19045.4412]
(c) Microsoft Corporation. All rights reserved.

C:\Users\PC\Desktop\KIbe\MSc Data Science\Y2_sem 1\Data Engineering\pracs\sws\Airflow>docker compose up -d
[+] Running 48/7
 ✓ airflow-webserver Pulled                                                                     620.9s
 ✓ postgres Pulled                                                                              426.9s
 ✓ redis Pulled                                                                                 289.4s
 ✓ airflow-scheduler Pulled                                                                     621.5s
 ✓ airflow-triggerer Pulled                                                                     619.9s
 ✓ airflow-init Pulled                                                                          621.2s
 ✓ airflow-worker Pulled                                                                        620.3s
[+] Running 9/9
 ✓ Network airflow_default               Created                                                 13.0s
 ✓ Volume "airflow_postgres-db-volume"   Created                                                  4.1s
 ✗ Container airflow-postgres-1          Error                                                  179.2s
 ✓ Container airflow-redis-1             Healthy                                                130.8s
 ✓ Container airflow-airflow-init-1      Created                                                  6.4s
 ✓ Container airflow-airflow-worker-1    Created                                                  8.3s
 ✓ Container airflow-airflow-webserver-1 Created                                                  8.5s
 ✓ Container airflow-airflow-triggerer-1 Created                                                  7.4s
 ✓ Container airflow-airflow-scheduler-1 Created                                                  7.9s
dependency failed to start: container airflow-postgres-1 is unhealthy

C:\Users\PC\Desktop\KIbe\MSc Data Science\Y2_sem 1\Data Engineering\pracs\sws\Airflow>docker-compose restart
[+] Restarting 7/7
 ✓ Container airflow-airflow-scheduler-1 Started                                                373.8s
 ✓ Container airflow-airflow-triggerer-1 Started                                                373.6s
 ✓ Container airflow-redis-1             Started                                                373.0s
 ✓ Container airflow-airflow-init-1      Started                                                372.4s
 ✓ Container airflow-airflow-worker-1    Started                                                372.9s
 ✓ Container airflow-airflow-webserver-1 Started                                                372.7s
 ✓ Container airflow-postgres-1          Started                                                371.3s

C:\Users\PC\Desktop\KIbe\MSc Data Science\Y2_sem 1\Data Engineering\pracs\sws\Airflow>
```

Figure 2: Airflow Issues

3. Installing and configuring Elasticsearch

curl https://artifacts.elastic.co/downloads/ Elasticsearch/elasticsearch-7.6.0-darwin-x86_64.tar.gz --output elasticsearch.tar.gz **#cmd**

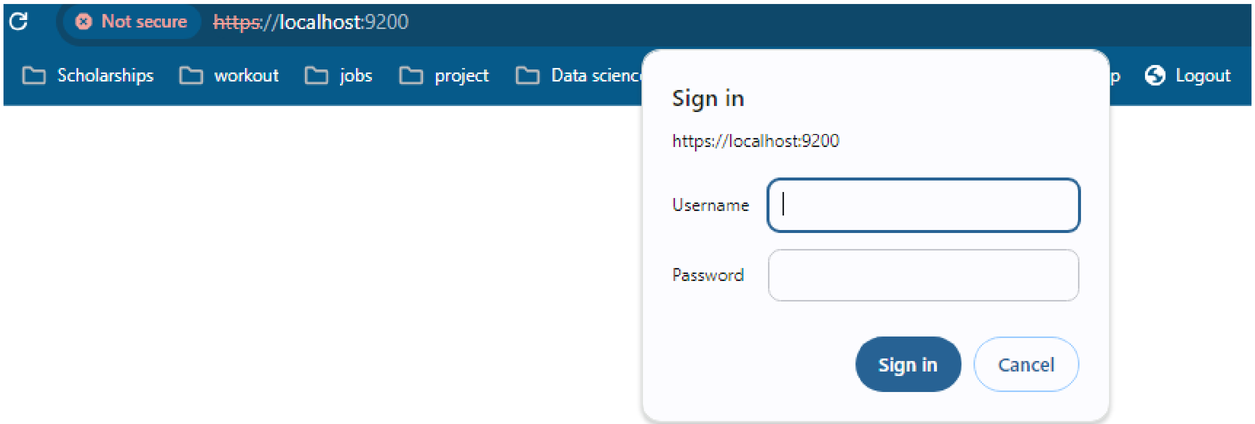
Download Elasticsearch zip file from: https://www.elastic.co/downloads/elasticsearch then extract the file or https://www.elastic.co/downloads/past-releases#elasticsearch

Then go bin in the folder look for Elasticsearch type window batch file, then run a command line from there with (**elasticsearch.bat**), it will now create our application, accessible on a URL on port 9200 at https://localhost:9200 or 9300

Play around with ***Elasticsearch-service.bat remove*** to try different versions if a process fails

sc delete elasticsearch-service-x86_64 **or** Control Panel\System and Security\Administrative Tools **or** services.msc ~ on win + r **or** on #cmd

Successful



Setting up a password

Run CMD as an ADMIN, then move to the bin with Elasticsearch batch file, copy the path then paste on cmd e.g. cd C:\Users\ ... ... \Elasticsearch\elasticsearch-8.11.4\bin

Then, below command to reset a new password with username: elastic

```
-Elasticsearch-reset-password -u elastic --interactive
```

- Elasticsearch-reset-password -u elastic --interactive –verbose with debug

Do not close the previous cmd running Elasticsearch application.

ERROR: Failed to determine the health of the cluster. , with exit code 69

4. Installing and configuring Kibana

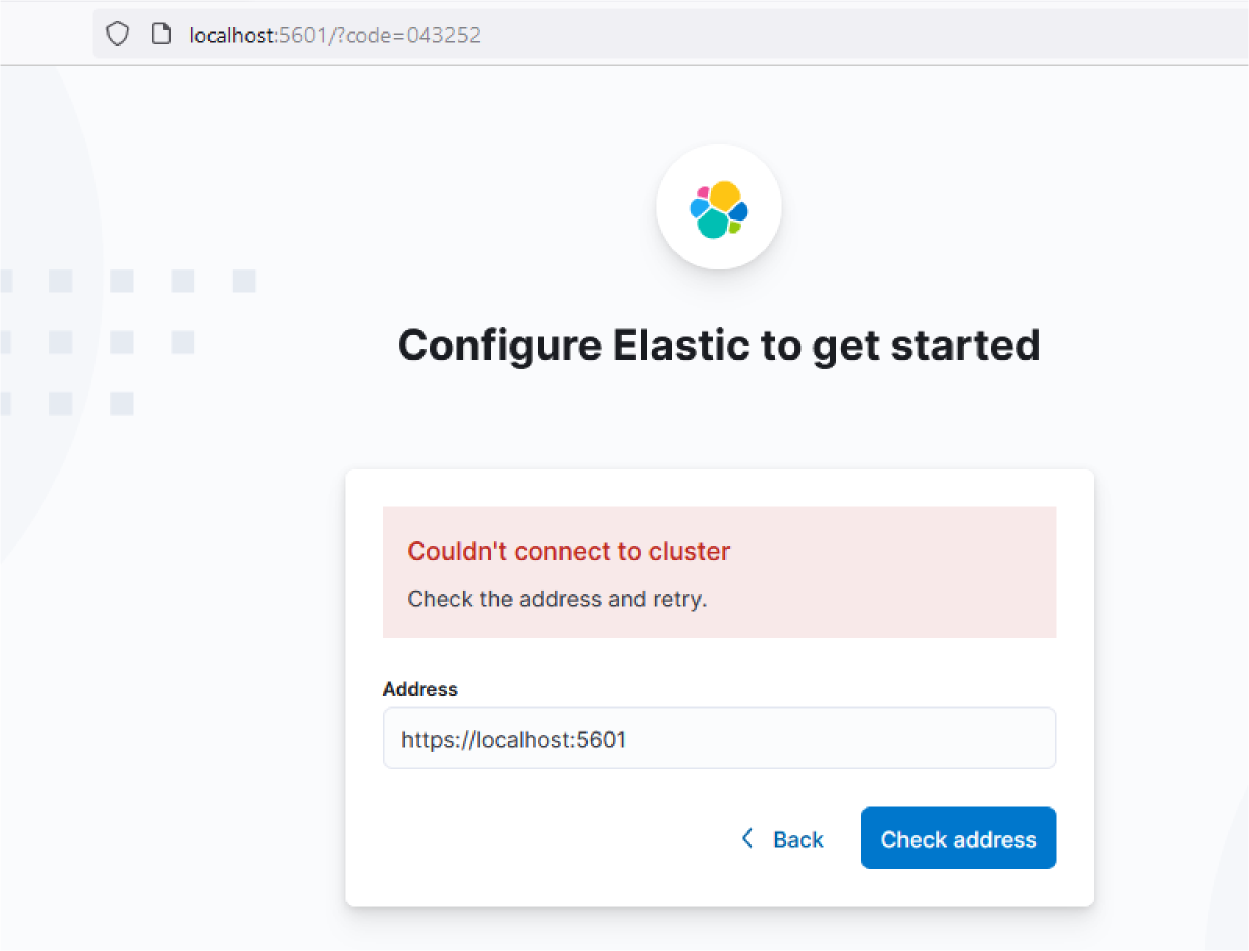
Download Kibana: https://www.elastic.co/downloads/kibana

Basically, our Kibana run on port 5601 at default.

Then follow below steps to launch

Run a CMD from the Kibana folder, basically, select the whole path on the search environment then type CMD, this should launch a working shell automatically, with path C:\Users\PC\ ‘your directory’ ... \kibana-8.14.0> .. **On CMD** environment Run **.\bin\kibana.ba,** this should install Kibana. On Firefox (Recommended) open **http://localhost:5601/?code=043252**

Successful but with conflicting address issues as below



localhost:5601/?code=043252

Configure Elastic to get started

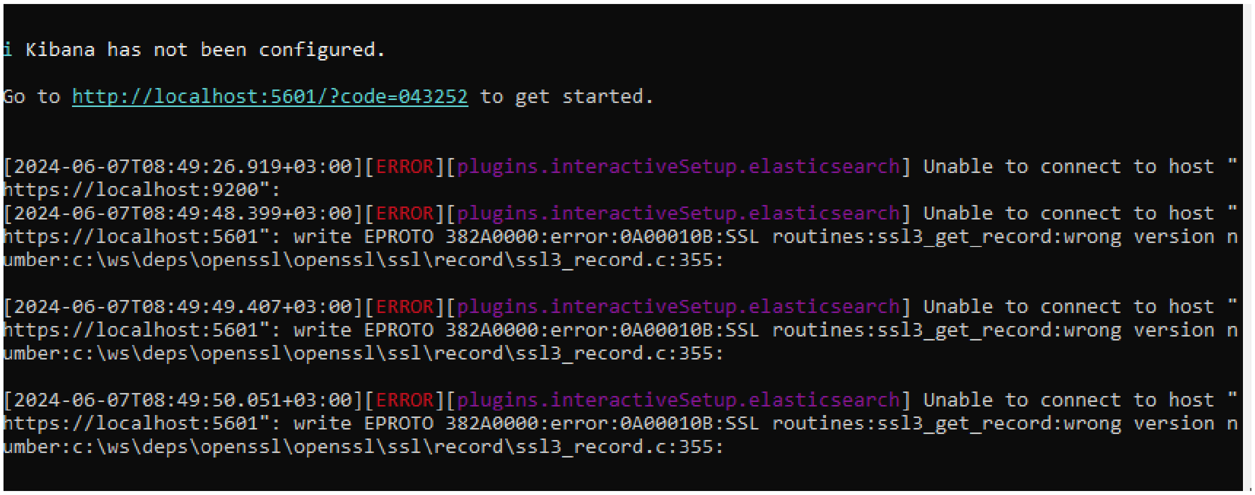
Couldn't connect to cluster

Check the address and retry.

Address

https://localhost:5601

Figure 4a: Kibana Issues



```
i Kibana has not been configured.

Go to http://localhost:5601/?code=043252 to get started.


[2024-06-07T08:49:26.919+03:00][ERROR][plugins.interactiveSetup.elasticsearch] Unable to connect to host "
https://localhost:9200":
[2024-06-07T08:49:48.399+03:00][ERROR][plugins.interactiveSetup.elasticsearch] Unable to connect to host "
https://localhost:5601": write EPROTO 382A0000:error:0A00010B:SSL routines:ssl3_get_record:wrong version n
umber:c:\ws\deps\openssl\openssl\ssl\record\ssl3_record.c:355:

[2024-06-07T08:49:49.407+03:00][ERROR][plugins.interactiveSetup.elasticsearch] Unable to connect to host "
https://localhost:5601": write EPROTO 382A0000:error:0A00010B:SSL routines:ssl3_get_record:wrong version n
umber:c:\ws\deps\openssl\openssl\ssl\record\ssl3_record.c:355:

[2024-06-07T08:49:50.051+03:00][ERROR][plugins.interactiveSetup.elasticsearch] Unable to connect to host "
https://localhost:5601": write EPROTO 382A0000:error:0A00010B:SSL routines:ssl3_get_record:wrong version n
umber:c:\ws\deps\openssl\openssl\ssl\record\ssl3_record.c:355:
```

Figure 4b: Kibana Issues

5. Installing and configuring PostgreSQL

Download PostgreSQL: https://www.enterprisedb.com/downloads/postgres-postgresql-downloads

Then install normally, this will take a few minutes Port = 5432, remember password = ***k#$e***

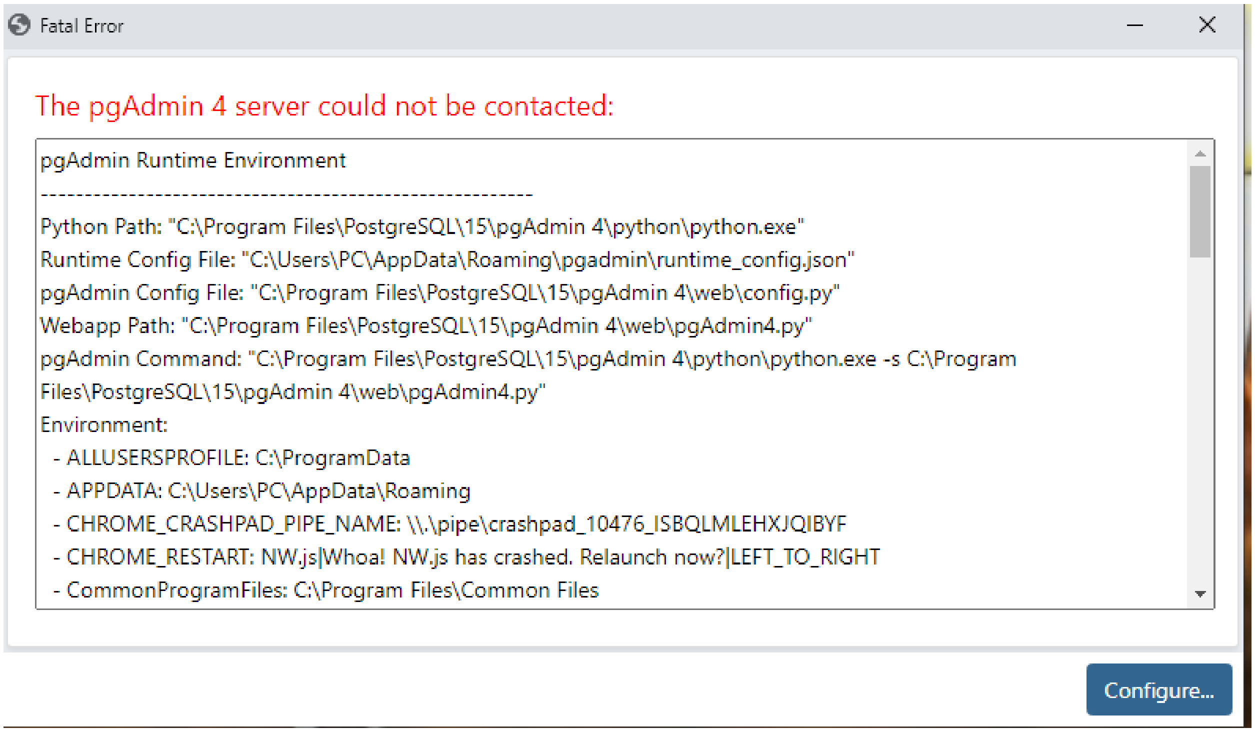
Launches successfully as below

6. Installing pgAdmin 4

pgAdmin 4 has been installed as a PostgreSQL component above.

Problem

Asks configure a static port in one case, in another case it asks to return to default, then now launches successfully.



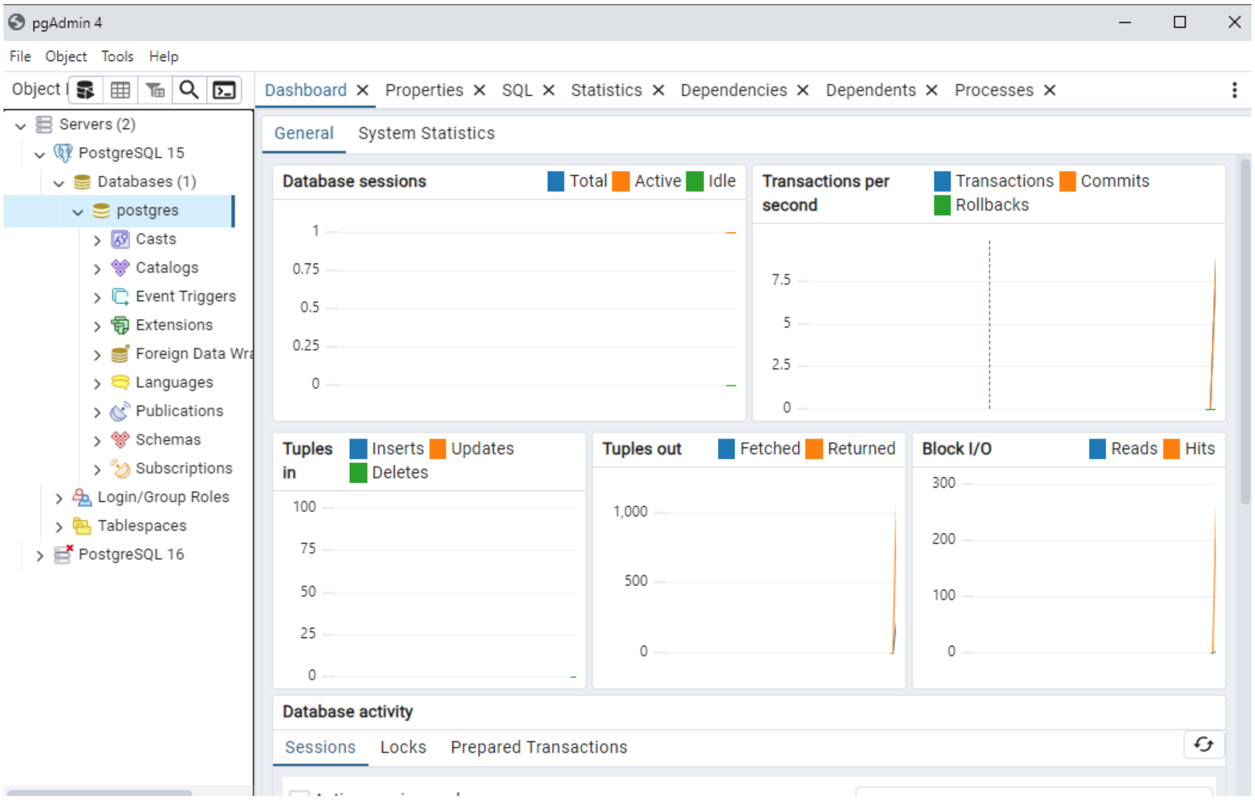
Fatal Error

The pgAdmin 4 server could not be contacted:

```
pgAdmin Runtime Environment
--------------------------------------------------------
Python Path: "C:\Program Files\PostgreSQL\15\pgAdmin 4\python\python.exe"
Runtime Config File: "C:\Users\PC\AppData\Roaming\pgadmin\runtime_config.json"
pgAdmin Config File: "C:\Program Files\PostgreSQL\15\pgAdmin 4\web\config.py"
Webapp Path: "C:\Program Files\PostgreSQL\15\pgAdmin 4\web\pgAdmin4.py"
pgAdmin Command: "C:\Program Files\PostgreSQL\15\pgAdmin 4\python\python.exe -s C:\Program
Files\PostgreSQL\15\pgAdmin 4\web\pgAdmin4.py"
Environment:
  - ALLUSERSPROFILE: C:\ProgramData
  - APPDATA: C:\Users\PC\AppData\Roaming
  - CHROME_CRASHPAD_PIPE_NAME: \\.\pipe\crashpad_10476_ISBQLMLEHXJQIBYF
  - CHROME_RESTART: NW.js|Whoa! NW.js has crashed. Relaunch now?|LEFT_TO_RIGHT
  - CommonProgramFiles: C:\Program Files\Common Files
```

Figure 6: Postgres issue

Successful

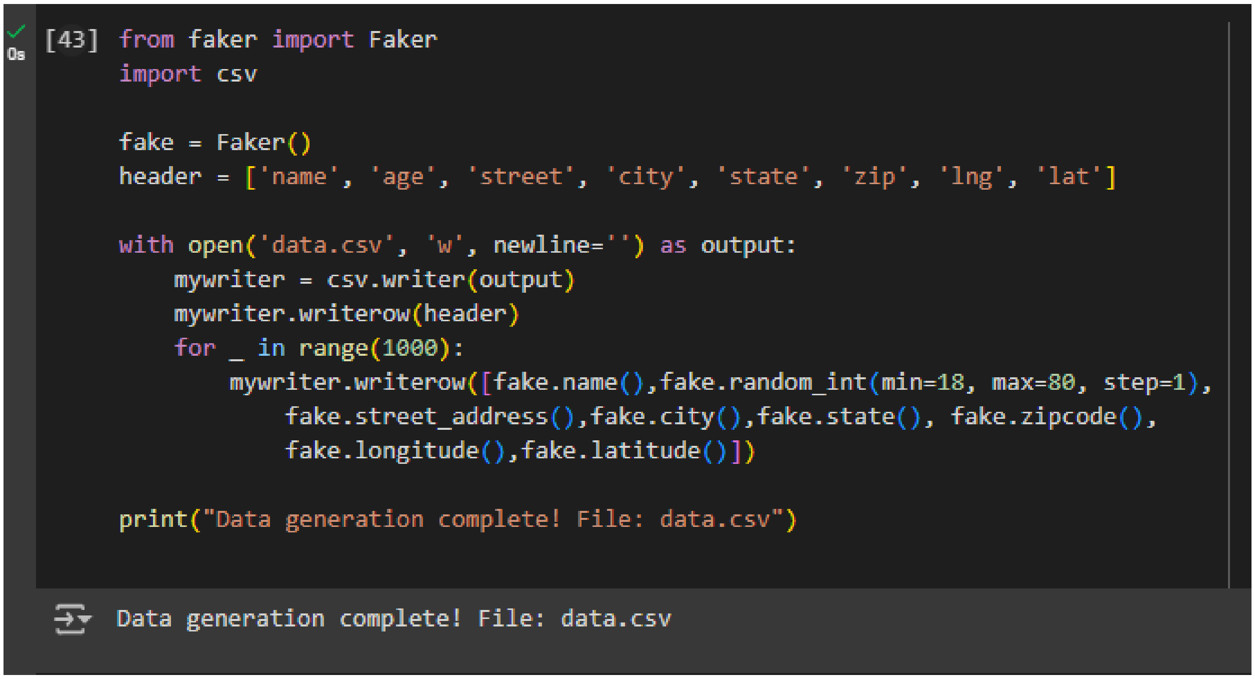


II. Reading and Writing Files

Explore the process of reading and writing CSV files using Python’s built-in CSV library and pandas Dataframes. Compare the two approaches in terms of performance, ease of use, and handling of large datasets. Provide a scenario where one approach might be preferred over the other.

a. How to read and write CSV files using Python’s built-in CSV library

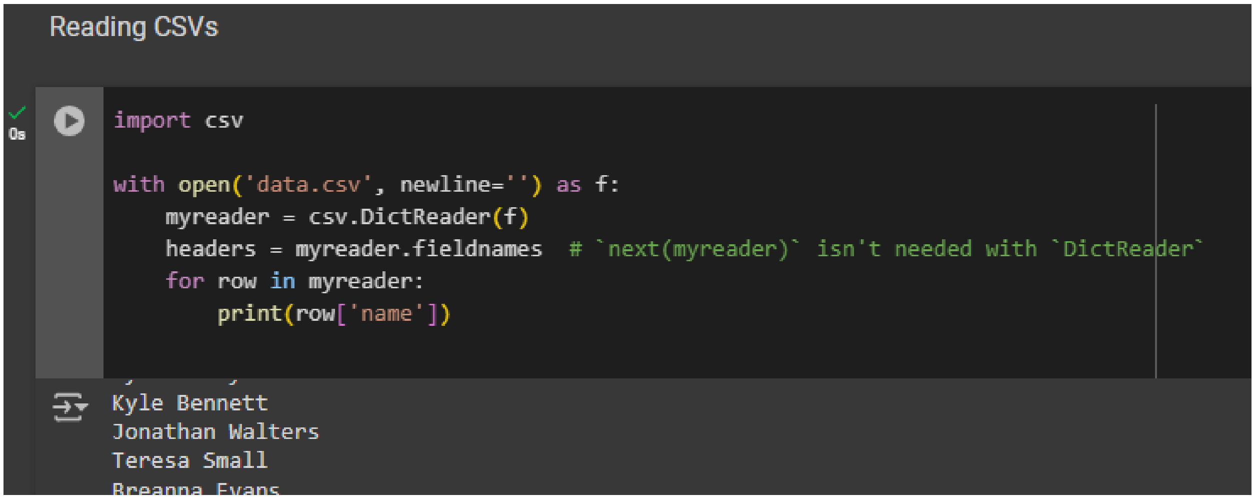
Sample code to **write** CSV file s in python’s built in CSV library



```
from faker import Faker
import csv

fake = Faker()
header = ['name', 'age', 'street', 'city', 'state', 'zip', 'lng', 'lat']

with open('data.csv', 'w', newline='') as output:
    mywriter = csv.writer(output)
    mywriter.writerow(header)
    for _ in range(1000):
        mywriter.writerow([fake.name(),fake.random_int(min=18, max=80, step=1),
            fake.street_address(),fake.city(),fake.state(), fake.zipcode(),
            fake.longitude(),fake.latitude()])

print("Data generation complete! File: data.csv")
```

```
Data generation complete! File: data.csv
```

Sample codes to **read** CSV file s in python’s built in CSV library



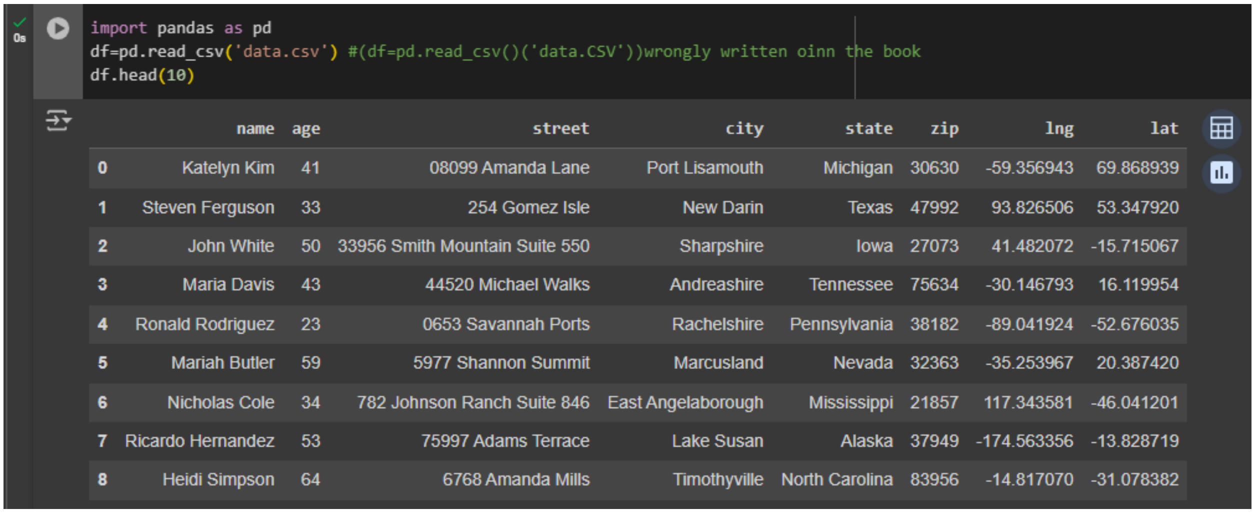
Reading CSVs

```
import csv

with open('data.csv', newline='') as f:
    myreader = csv.DictReader(f)
    headers = myreader.fieldnames  # `next(myreader)` isn't needed with `DictReader`
    for row in myreader:
        print(row['name'])
```

```
Kyle Bennett
Jonathan Walters
Teresa Small
Breanna Evans
```

b. How to Read and write CSVs using pandas Dataframes

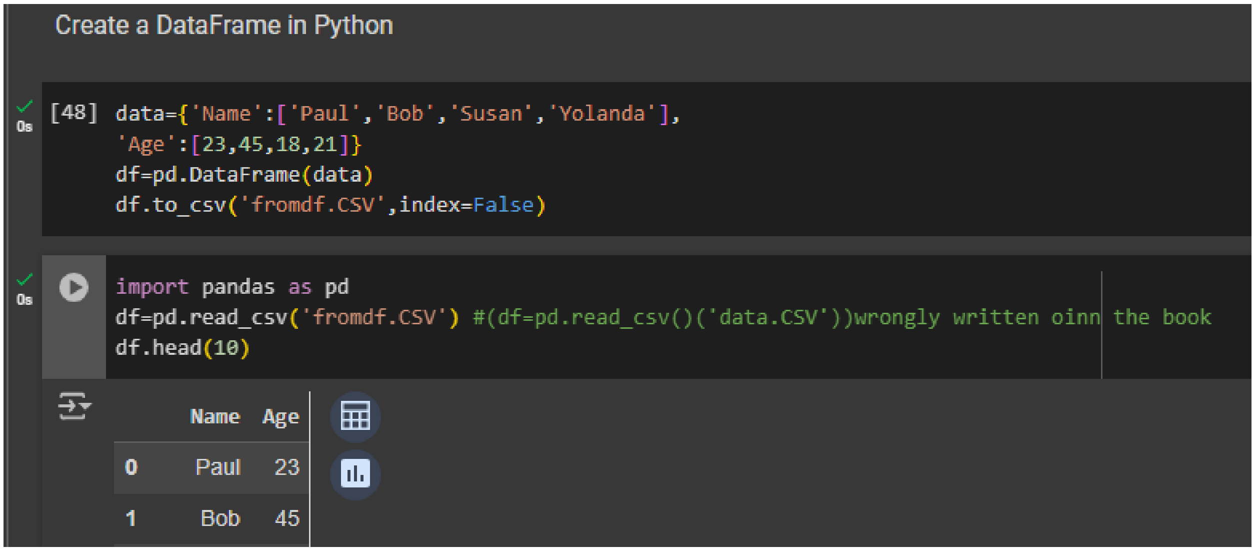
Sample code to **read** CSVs using pandas Dataframes



```
import pandas as pd
df=pd.read_csv('data.csv') #(df=pd.read_csv()('data.CSV'))wrongly written oinn the book
df.head(10)
```

	name	age	street	city	state	zip	lng	lat
0	Katelyn Kim	41	08099 Amanda Lane	Port Lisamouth	Michigan	30630	-59.356943	69.868939
1	Steven Ferguson	33	254 Gomez Isle	New Darin	Texas	47992	93.826506	53.347920
2	John White	50	33956 Smith Mountain Suite 550	Sharpshire	Iowa	27073	41.482072	-15.715067
3	Maria Davis	43	44520 Michael Walks	Andreashire	Tennessee	75634	-30.146793	16.119954
4	Ronald Rodriguez	23	0653 Savannah Ports	Rachelshire	Pennsylvania	38182	-89.041924	-52.676035
5	Mariah Butler	59	5977 Shannon Summit	Marcusland	Nevada	32363	-35.253967	20.387420
6	Nicholas Cole	34	782 Johnson Ranch Suite 846	East Angelaborough	Mississippi	21857	117.343581	-46.041201
7	Ricardo Hernandez	53	75997 Adams Terrace	Lake Susan	Alaska	37949	-174.563356	-13.828719
8	Heidi Simpson	64	6768 Amanda Mills	Timothyville	North Carolina	83956	-14.817070	-31.078382

Sample code to **write** CSVs using pandas Dataframes



Create a DataFrame in Python

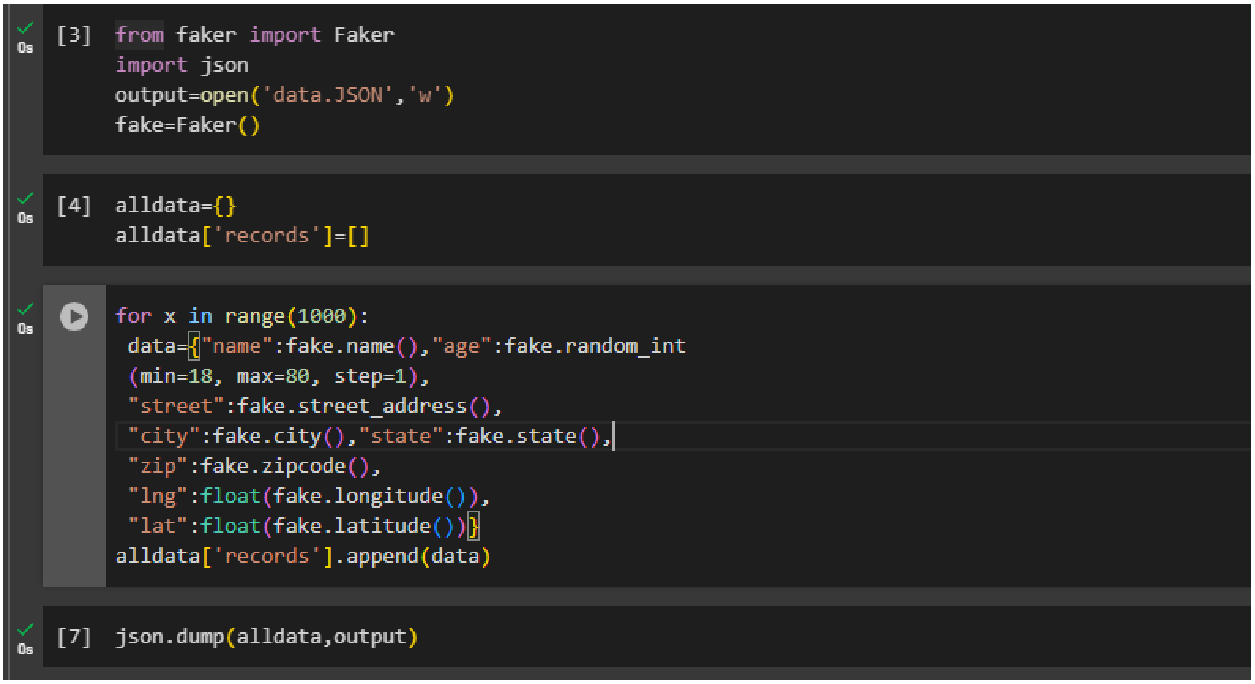
```
[48] data={'Name':['Paul','Bob','Susan','Yolanda'],
     'Age':[23,45,18,21]}
     df=pd.DataFrame(data)
     df.to_csv('fromdf.CSV',index=False)
```

```
import pandas as pd
df=pd.read_csv('fromdf.CSV') #(df=pd.read_csv()('data.CSV'))wrongly written oinn the book
df.head(10)
```

	Name	Age
0	Paul	23
1	Bob	45

c. Writing JSON with Python

Sample code to write JSON file in Python



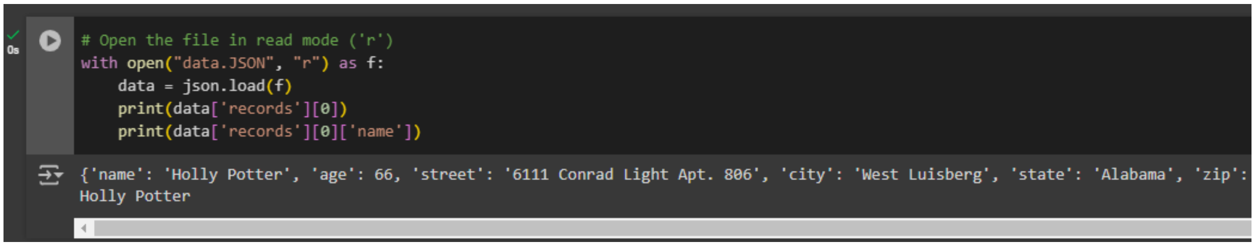
```python
from faker import Faker
import json
output=open('data.JSON','w')
fake=Faker()
```

```python
alldata={}
alldata['records']=[]
```

```python
for x in range(1000):
 data={"name":fake.name(),"age":fake.random_int
 (min=18, max=80, step=1),
 "street":fake.street_address(),
 "city":fake.city(),"state":fake.state(),
 "zip":fake.zipcode(),
 "lng":float(fake.longitude()),
 "lat":float(fake.latitude())}
alldata['records'].append(data)
```

```python
json.dump(alldata,output)
```

Sample code to read JSON files in Python

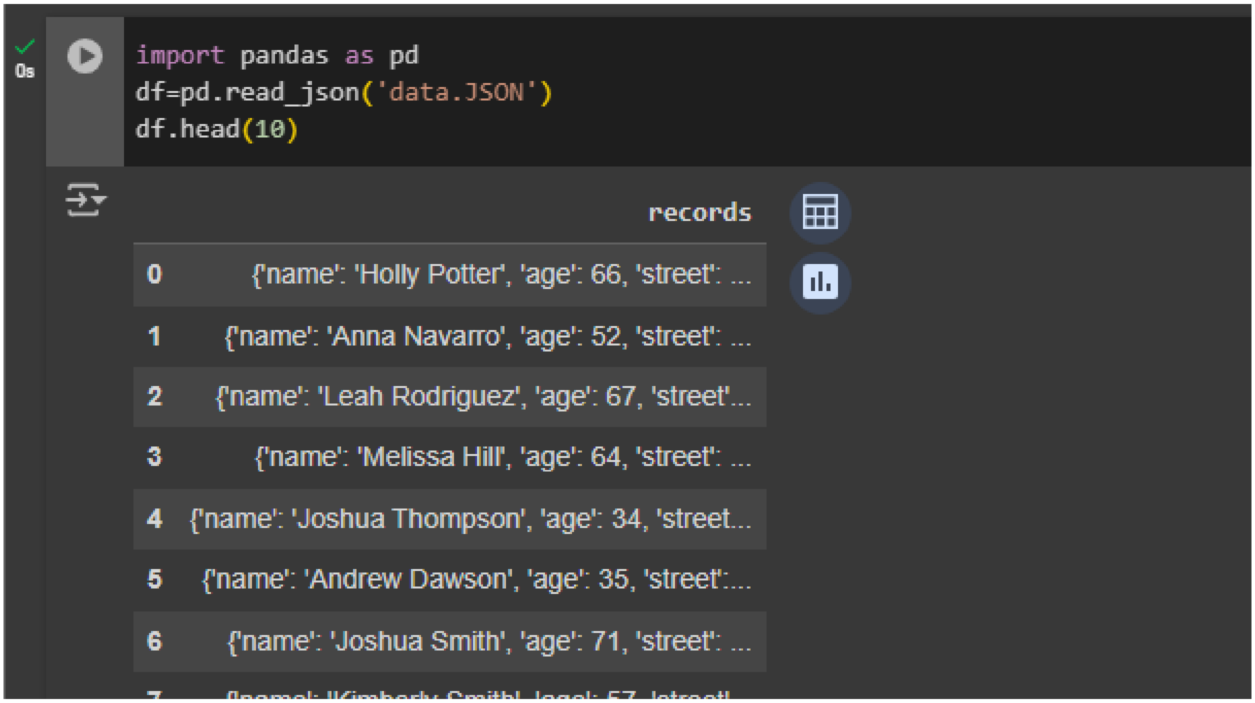


```python
# Open the file in read mode ('r')
with open("data.JSON", "r") as f:
    data = json.load(f)
    print(data['records'][0])
    print(data['records'][0]['name'])
```

```
{'name': 'Holly Potter', 'age': 66, 'street': '6111 Conrad Light Apt. 806', 'city': 'West Luisberg', 'state': 'Alabama', 'zip':
Holly Potter
```

d. Reading and writing JSON with Dataframes

Sample code read Json files

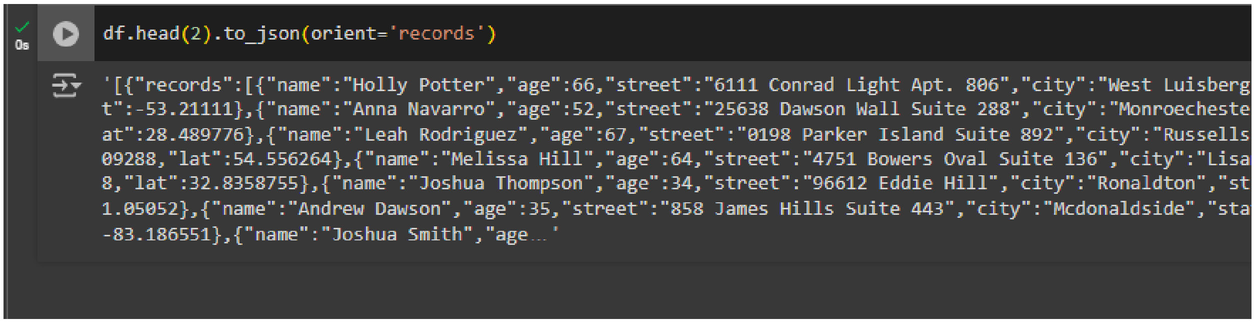


```python
import pandas as pd
df=pd.read_json('data.JSON')
df.head(10)
```

	records
0	{'name': 'Holly Potter', 'age': 66, 'street': ...
1	{'name': 'Anna Navarro', 'age': 52, 'street': ...
2	{'name': 'Leah Rodriguez', 'age': 67, 'street'...
3	{'name': 'Melissa Hill', 'age': 64, 'street': ...
4	{'name': 'Joshua Thompson', 'age': 34, 'street...
5	{'name': 'Andrew Dawson', 'age': 35, 'street':...
6	{'name': 'Joshua Smith', 'age': 71, 'street': ...
7	{'name': 'Kimberly Smith', 'age': 57, 'street':

Pass the orient parameter, which determines the format of the JSON that is returned

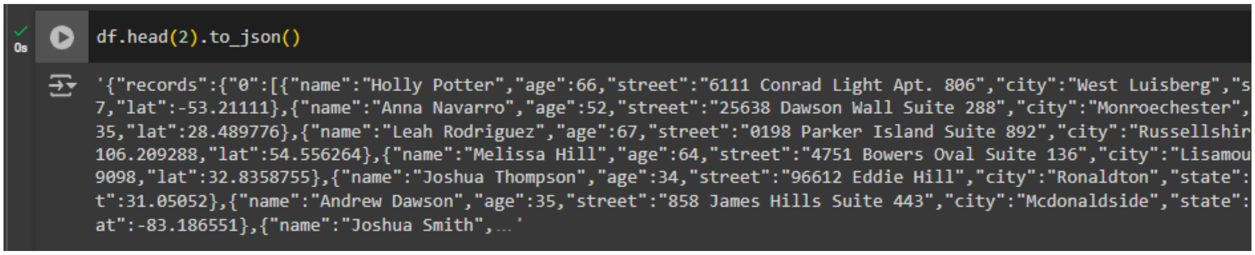
Sample results as below



```python
df.head(2).to_json(orient='records')
```

```
'[{"records":[{"name":"Holly Potter","age":66,"street":"6111 Conrad Light Apt. 806","city":"West Luisberg
t":-53.21111},{"name":"Anna Navarro","age":52,"street":"25638 Dawson Wall Suite 288","city":"Monroecheste
at":28.489776},{"name":"Leah Rodriguez","age":67,"street":"0198 Parker Island Suite 892","city":"Russells
09288,"lat":54.556264},{"name":"Melissa Hill","age":64,"street":"4751 Bowers Oval Suite 136","city":"Lisa
8,"lat":32.8358755},{"name":"Joshua Thompson","age":34,"street":"96612 Eddie Hill","city":"Ronaldton","st
1.05052},{"name":"Andrew Dawson","age":35,"street":"858 James Hills Suite 443","city":"Mcdonaldside","sta
-83.186551},{"name":"Joshua Smith","age...'
```

Default Json file looks like below

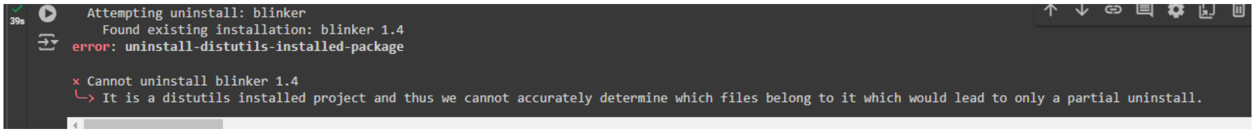


```python
df.head(2).to_json()
```

```
'{"records":{"0":[{"name":"Holly Potter","age":66,"street":"6111 Conrad Light Apt. 806","city":"West Luisberg","s
7,"lat":-53.21111},{"name":"Anna Navarro","age":52,"street":"25638 Dawson Wall Suite 288","city":"Monroechester",
35,"lat":28.489776},{"name":"Leah Rodriguez","age":67,"street":"0198 Parker Island Suite 892","city":"Russellshir
106.209288,"lat":54.556264},{"name":"Melissa Hill","age":64,"street":"4751 Bowers Oval Suite 136","city":"Lisamou
9098,"lat":32.8358755},{"name":"Joshua Thompson","age":34,"street":"96612 Eddie Hill","city":"Ronaldton","state":
t":31.05052},{"name":"Andrew Dawson","age":35,"street":"858 James Hills Suite 443","city":"Mcdonaldside","state":
at":-83.186551},{"name":"Joshua Smith",...'
```

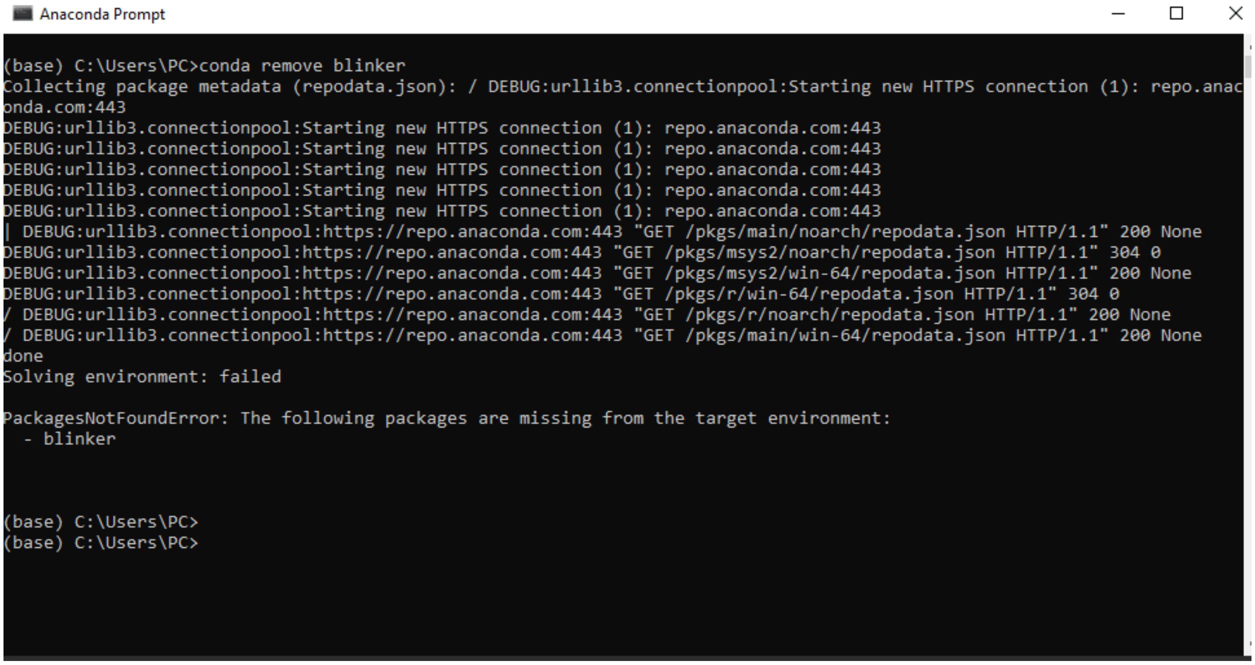
Building a CSV to a JSON data pipeline

Wanted to install apache-airflow on my IDE but ran into below error



```
  Attempting uninstall: blinker
    Found existing installation: blinker 1.4
error: uninstall-distutils-installed-package

× Cannot uninstall blinker 1.4
╰─> It is a distutils installed project and thus we cannot accurately determine which files belong to it which would lead to only a partial uninstall.
```

The system was configuring the environment in preparation for apache-airflow installation. I tried to uninstall it through conda remove blinker and still ran into an error as below.



```
(base) C:\Users\PC>conda remove blinker
Collecting package metadata (repodata.json): / DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1): repo.anac
onda.com:443
DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1): repo.anaconda.com:443
DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1): repo.anaconda.com:443
DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1): repo.anaconda.com:443
DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1): repo.anaconda.com:443
DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1): repo.anaconda.com:443
| DEBUG:urllib3.connectionpool:https://repo.anaconda.com:443 "GET /pkgs/main/noarch/repodata.json HTTP/1.1" 200 None
DEBUG:urllib3.connectionpool:https://repo.anaconda.com:443 "GET /pkgs/msys2/noarch/repodata.json HTTP/1.1" 304 0
DEBUG:urllib3.connectionpool:https://repo.anaconda.com:443 "GET /pkgs/msys2/win-64/repodata.json HTTP/1.1" 200 None
DEBUG:urllib3.connectionpool:https://repo.anaconda.com:443 "GET /pkgs/r/win-64/repodata.json HTTP/1.1" 304 0
/ DEBUG:urllib3.connectionpool:https://repo.anaconda.com:443 "GET /pkgs/r/noarch/repodata.json HTTP/1.1" 200 None
/ DEBUG:urllib3.connectionpool:https://repo.anaconda.com:443 "GET /pkgs/main/win-64/repodata.json HTTP/1.1" 200 None
done
Solving environment: failed

PackagesNotFoundError: The following packages are missing from the target environment:
  - blinker


(base) C:\Users\PC>
(base) C:\Users\PC>
```

Conclusion

Basically, this project requires a very robust system, for example core i7 and above, unfortunately I was using core i5. So, as I was approaching towards using the whole Data engineering infrastructure at once to perform ETL, my system was overwhelmed. However, the project is perfectly doable (Chapter 4: Working with Databases, Chapter 5: Cleaning and Transforming Data, Chapter 6: Building a 311 Data Pipeline) if and when the data engineering infrastructure is configured correctly. Continuing to struggle to make progress with my weak system is an unsatisfactory endeavor.