Deloitte SFL Presentation

By

Emmanuel Oyekanlu

Presentation Outline

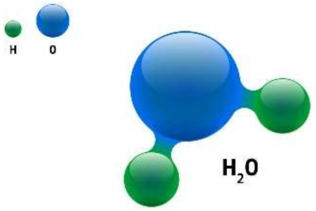
- Work Plan for the Implementation of ChemBERTa at BPC (slide 2 – slide 9)
- PostgreSQL Based Data ETL Solution on Docker with pgAdmin and Streamlit Frontend (slide 10 – slide 15)
- Deep Learning Model Deployment on Docker with Flask API (slide 16 – slide 19)
- Questions/Suggestions

Work Plan for the Implementation of ChemBERTa at BPC

Background

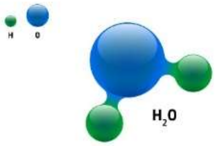
- Big Pharma Company (BPC) seeks to deploy ChemBERTa machine learning algorithm for its downstream applications.
- Deploying ChemBERTa will enable BPC to do chemical fingerprinting , molecules representation and property prediction of BPC’s materials and products.
- BPC have IT/Cloud team with no knowledge of AI/ML workloads
- BPC have approached Deloitte-SFL team to implement the ChemBERTa algorithm and deploy it as part of BPCs downstream applications.
- BPC needs a work plan from the Deloitte-SFL team

Work Plan for the Implementation of ChemBERTa at BPC



Background

- Molecules exists with 3D geometries (biological, chemical & physical properties)
- Big Data
- Simplified Molecular Input Line Entry System (SMILES) data set.
- World's largest open source collections of chemical & molecular structures.
- ChemBERTa algorithm based on RoBERTa algorithm was trained on the SMILES data set.



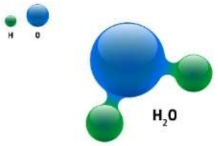
- HuggingFace Open Source Machine Learning Libraries

Work Plan for the Implementation of ChemBERTa at BPC

Deliverables by Deloitte SFL Team

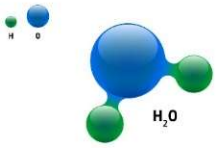
- A ChemBERTa based ML model that can accomplish molecules property prediction using BPC's data
- Integration of the developed NLP model for downstream applications at BPC
- Deployments through cloud APIs. Possibly desktop apps deployments. Based on needs.
- Reliability of the deployed model will be accomplished by containerization and Kubernetes technologies.

Work Plan for the Implementation of ChemBERTa at BPC

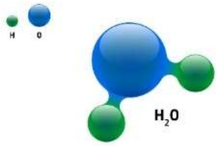


Team & Resources

- A ChemBERTa based ML model that can accomplish molecules property prediction using BPC's data
- Integration of the developed NLP model for downstream applications at BPC
- Deployments through cloud APIs. Possibly desktop apps deployments. Based on needs.
- Reliability of the deployed model will be accomplished by containerization and Kubernetes technologies.

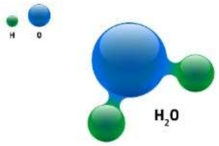


Team & Resources Needed



- DeloitteSFL Project Manager (Data Engineering / AI Applications)
- Two (2) Data Engineers with knowledge of GPU, CUDA, RAPIDS (cuDF, Dask, cuxFilter), Vaex and/or OpenCL
- Two (2) DeloitteSFL NLP Engineers with knowledge of HuggingFace suite of libraries.
- DeloitteSFL Data Visualization Engineer with knowledge of rendering 2D and 3D interactive graphs. Plotly, Chemplot, Mayavi, PyVista, etc.
- Two (2) molecules/chemical/materials engineers from BPC
- Two (2) BPC Engineers. Data/ML engineers. BPC engineers that can be quickly trained can also fit in.
- One (1) Software Engineer from DeloitteSFL
- One (1) DevOps Engineer from DeloitteSFL

Time Estimates

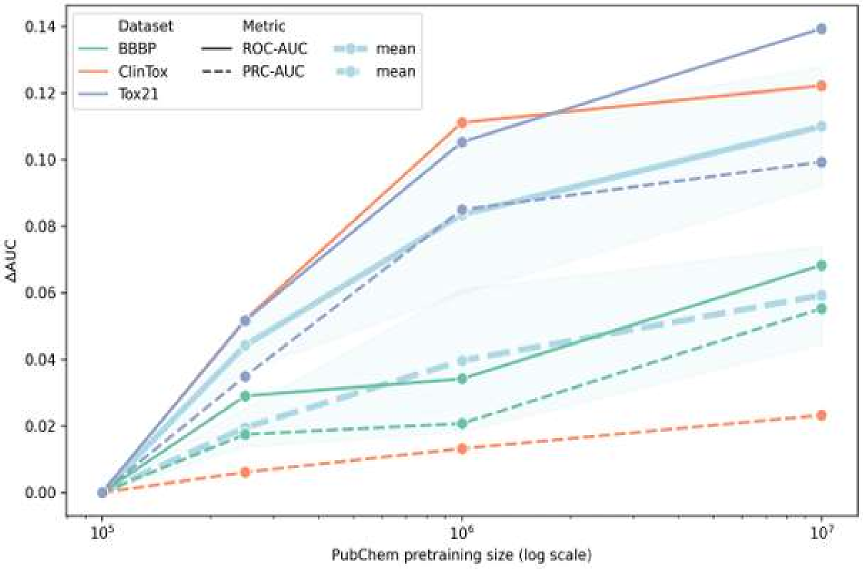


• Initial ETL/ELT activities	60 hours
• ChemBERTa model transfer learning using BPC’s data. Possible retraining of RoBERTa from ground up to achieve a better ChemBERTa model	80 hours
• Software & DevOps Engineer, dockerization, Kubernetes, cloud and possibly desktop app deployments	55 hours
• Usage manual design, training of BPC material engineers with regards to app usage	50 hours

Time estimate is approximately 245 – 400 hours. Time estimates can be reviewed after further discussion with the entire team at Deloitte SFL and BPC

Success Criteria

- Authors of ChemBERTa suggested using ROC/AUC curves to measure the performance of the ChemBERTa algorithm.
- Apart from test data sets, performance can also be tested with known molecules to see if the model generalizes well.
- Performance data can also be harvested over time from field reports by BPC users.



PostgreSQL Based Data ETL Solution on Docker with pgAdmin and Streamlit Frontend

Background

An Extraction, Transform and Loading (ETL) project that automates the ingestion of data from a provided CSV/xlsx file, transform the data and load the data into a PostgreSQL database.

Result

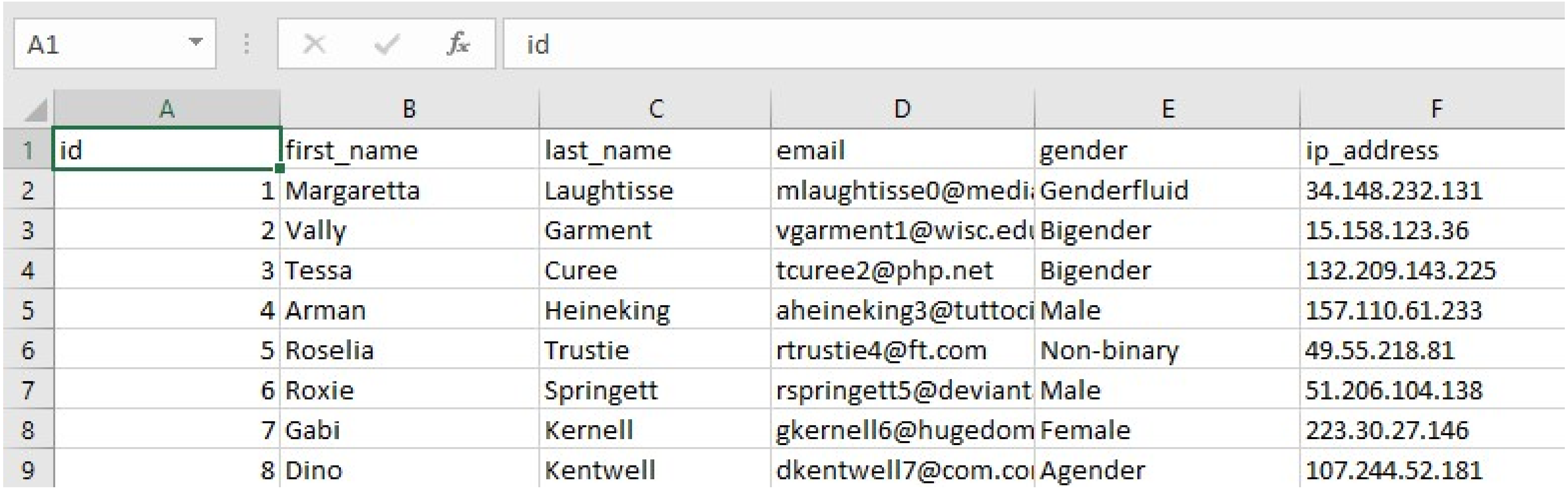
Loaded data can be viewed through using PostgreSQL’s pgAdmin user interface (UI).

Streamlit (front end) for user’s convenience.

Docker & virtual environment for reproducibility

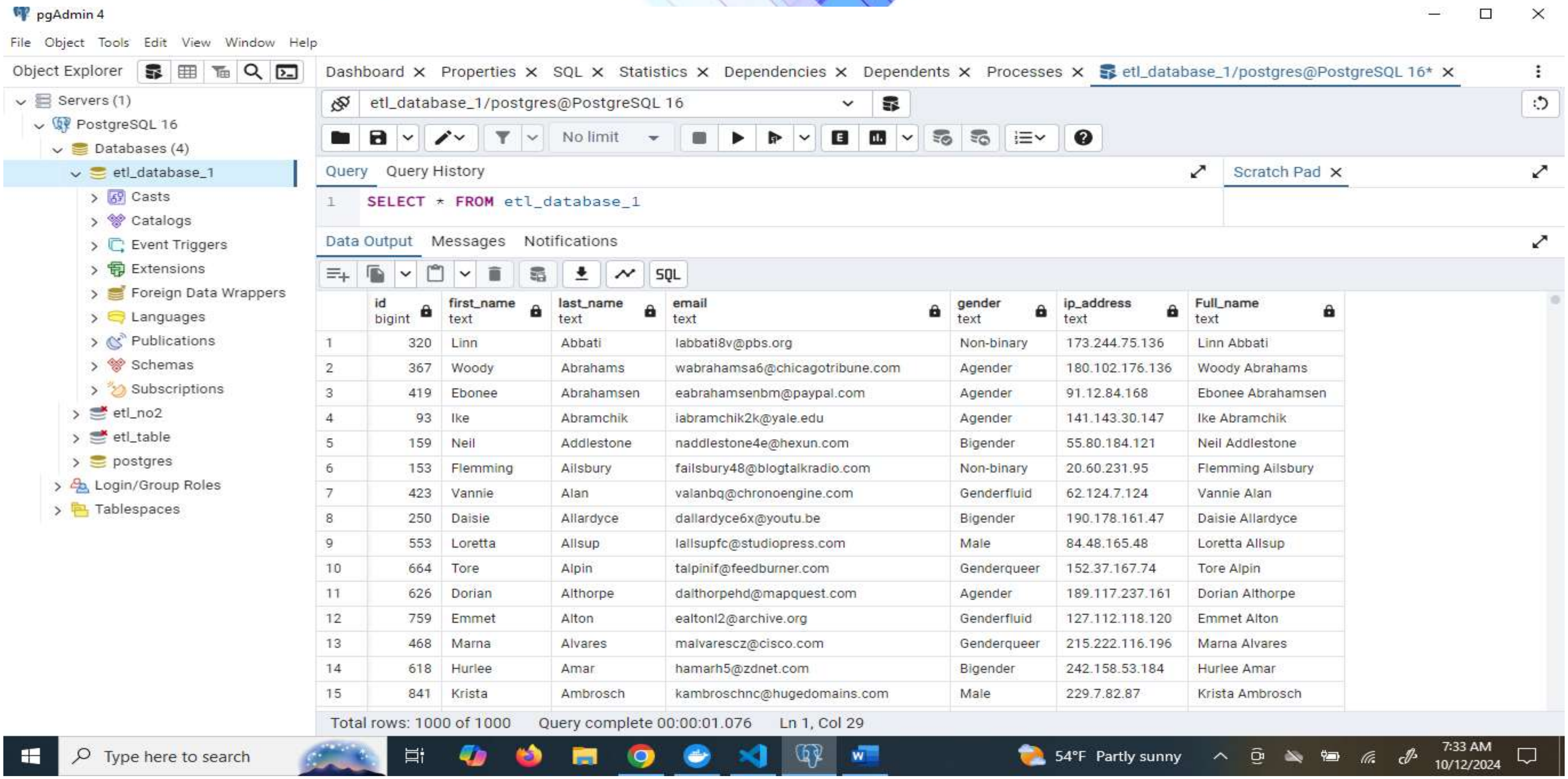
Snapshot of input data

DATA INTEGRATION



	A	B	C	D	E	F
1	id	first_name	last_name	email	gender	ip_address
2	1	Margaretta	Laughtisse	mlaughtisse0@medi	Genderfluid	34.148.232.131
3	2	Vally	Garment	vgarment1@wisc.edu	Bigender	15.158.123.36
4	3	Tessa	Curee	tcuree2@php.net	Bigender	132.209.143.225
5	4	Arman	Heineking	aheineking3@tuttoci	Male	157.110.61.233
6	5	Roselia	Trustie	rtrustie4@ft.com	Non-binary	49.55.218.81
7	6	Roxie	Springett	rspringett5@deviant	Male	51.206.104.138
8	7	Gabi	Kernell	gkernell6@hugedom	Female	223.30.27.146
9	8	Dino	Kentwell	dkentwell7@com.co	Agender	107.244.52.181

After ETL : pgAdmin UI

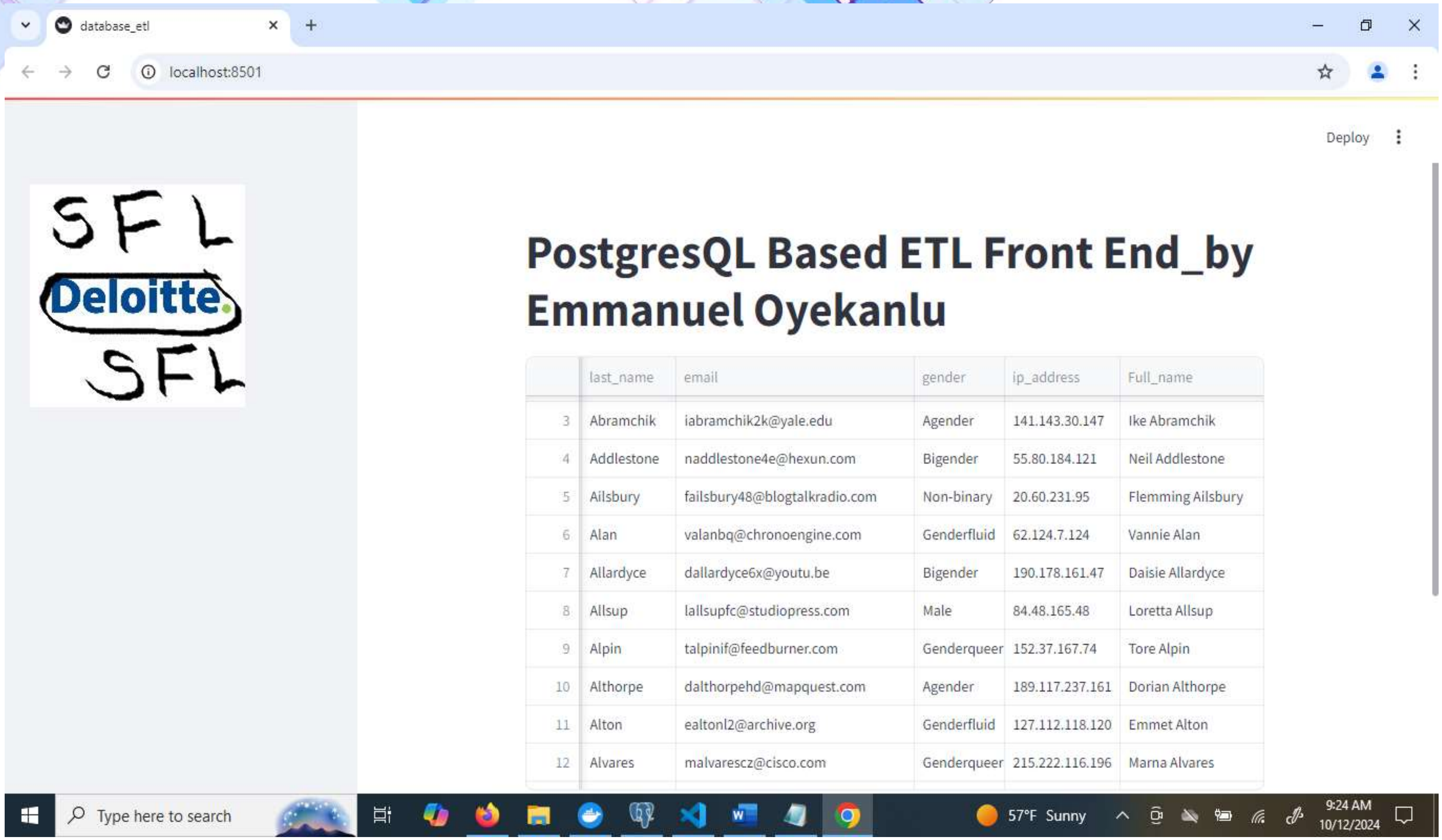


```
SELECT * FROM etl_database_1
```

	id bigint	first_name text	last_name text	email text	gender text	ip_address text	Full_name text
1	320	Linn	Abbati	labbati8v@pbs.org	Non-binary	173.244.75.136	Linn Abbati
2	367	Woody	Abrahams	wabrahamsa6@chicagotribune.com	Agender	180.102.176.136	Woody Abrahams
3	419	Ebonee	Abrahamsen	eabrahamsenbm@paypal.com	Agender	91.12.84.168	Ebonee Abrahamsen
4	93	Ike	Abramchik	iabramchik2k@yale.edu	Agender	141.143.30.147	Ike Abramchik
5	159	Neil	Addlestone	naddlestone4e@hexun.com	Bigender	55.80.184.121	Neil Addlestone
6	153	Flemming	Ailsbury	failsbury48@blogtalkradio.com	Non-binary	20.60.231.95	Flemming Ailsbury
7	423	Vannie	Alan	valanbq@chronoengine.com	Genderfluid	62.124.7.124	Vannie Alan
8	250	Daisie	Allardyce	dallardyce6x@youtu.be	Bigender	190.178.161.47	Daisie Allardyce
9	553	Loretta	Allsup	lallsupfc@studiopress.com	Male	84.48.165.48	Loretta Allsup
10	664	Tore	Alpin	talpinif@feedburner.com	Genderqueer	152.37.167.74	Tore Alpin
11	626	Dorian	Althorpe	dalthorpehd@mapquest.com	Agender	189.117.237.161	Dorian Althorpe
12	759	Emmet	Alton	ealtonl2@archive.org	Genderfluid	127.112.118.120	Emmet Alton
13	468	Marna	Alvares	malvarescz@cisco.com	Genderqueer	215.222.116.196	Marna Alvares
14	618	Hurlee	Amar	hamarh5@zdnet.com	Bigender	242.158.53.184	Hurlee Amar
15	841	Krista	Ambrosch	kambroschnc@hugedomains.com	Male	229.7.82.87	Krista Ambrosch

Total rows: 1000 of 1000 Query complete 00:00:01.076 Ln 1, Col 29

After ETL : Streamlit API



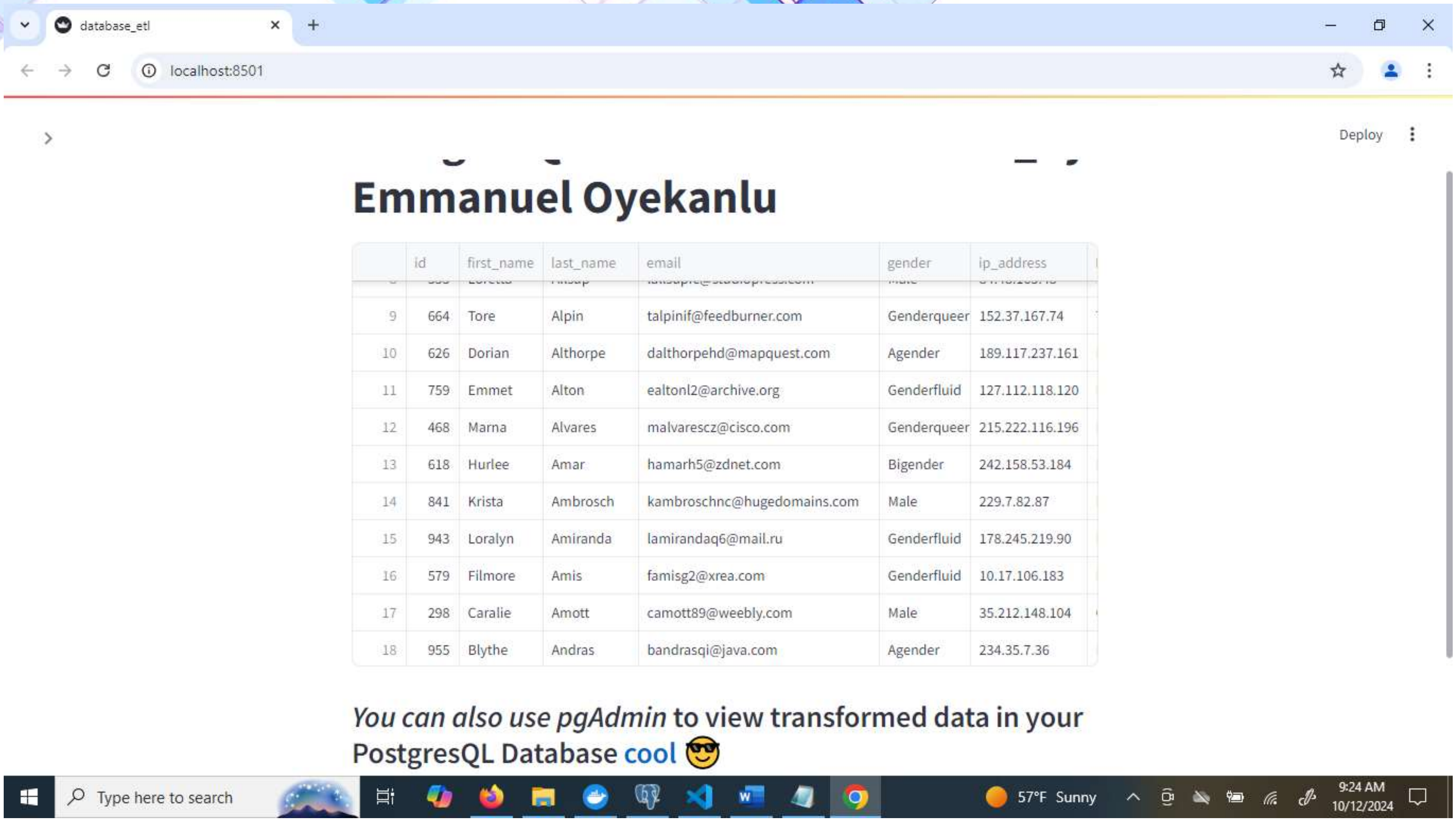
PostgresQL Based ETL Front End_by Emmanuel Oyekanlu

	last_name	email	gender	ip_address	Full_name
3	Abramchik	iabramchik2k@yale.edu	Agender	141.143.30.147	Ike Abramchik
4	Addlestone	naddlestone4e@hexun.com	Bigender	55.80.184.121	Neil Addlestone
5	Ailsbury	failsbury48@blogtalkradio.com	Non-binary	20.60.231.95	Flemming Ailsbury
6	Alan	valanbq@chronoengine.com	Genderfluid	62.124.7.124	Vannie Alan
7	Allardyce	dallardyce6x@youtu.be	Bigender	190.178.161.47	Daisie Allardyce
8	Allsup	lallsupfc@studiopress.com	Male	84.48.165.48	Loretta Allsup
9	Alpin	talpinif@feedburner.com	Genderqueer	152.37.167.74	Tore Alpin
10	Althorpe	dalthorpehd@mapquest.com	Agender	189.117.237.161	Dorian Althorpe
11	Alton	ealtonl2@archive.org	Genderfluid	127.112.118.120	Emmet Alton
12	Alvares	malvarescz@cisco.com	Genderqueer	215.222.116.196	Marna Alvares

Apologies for the crudely designed DeloitteSFL logo

☻

After ETL : Streamlit API



Emmanuel Oyekanlu

	id	first_name	last_name	email	gender	ip_address
9	664	Tore	Alpin	talpinif@feedburner.com	Genderqueer	152.37.167.74
10	626	Dorian	Althorpe	dalthorpehd@mapquest.com	Agender	189.117.237.161
11	759	Emmet	Alton	ealtonl2@archive.org	Genderfluid	127.112.118.120
12	468	Marna	Alvares	malvarescz@cisco.com	Genderqueer	215.222.116.196
13	618	Hurlee	Amar	hamarh5@zdnet.com	Bigender	242.158.53.184
14	841	Krista	Ambrosch	kambroschnc@hugedomains.com	Male	229.7.82.87
15	943	Loralyn	Amiranda	lamirandaq6@mail.ru	Genderfluid	178.245.219.90
16	579	Filmore	Amis	famisg2@xrea.com	Genderfluid	10.17.106.183
17	298	Caralie	Amott	camott89@weebly.com	Male	35.212.148.104
18	955	Blythe	Andras	bandrasqi@java.com	Agender	234.35.7.36

You can also use pgAdmin to view transformed data in your PostgresQL Database cool 😎

PostgreSQL Based Data ETL Solution on Docker with pgAdmin and Streamlit Frontend

Background

An Extraction, Transform and Loading (ETL) project that automates the ingestion of data from a provided CSV/xlsx file, transform the data and load the data into a PostgreSQL database.

Result

Loaded data can be viewed through using PostgreSQL’s pgAdmin user interface (UI).

Streamlit (front end) for user’s convenience.

Docker & virtual environment for reproducibility

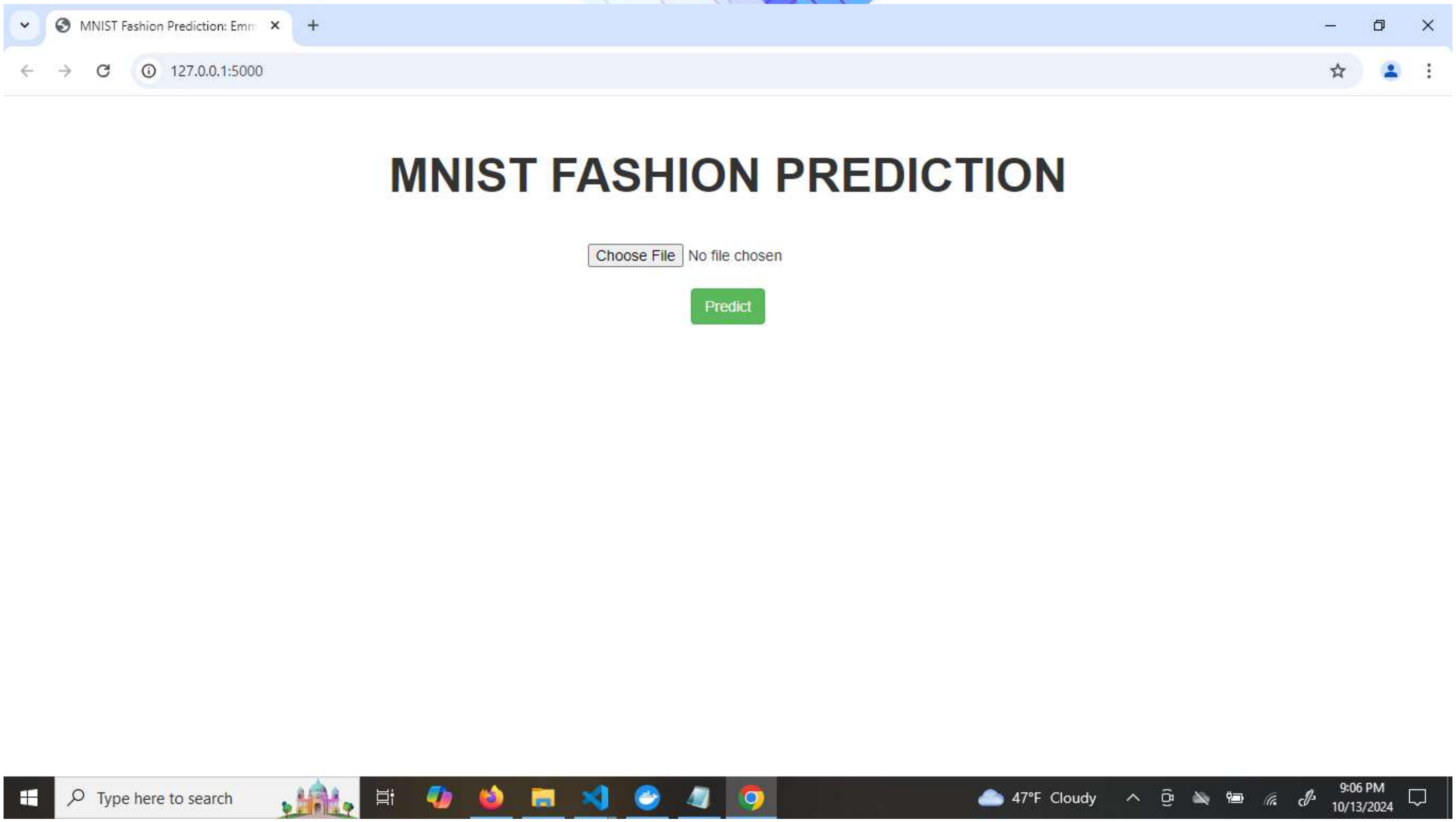
Deep Learning Model Deployment on Docker with Flask API

Background

- One of the major goals of training machine learning models is to solve real world problems, and this goal can only be accomplished when a trained model is deployed in productions and being actively used by consumers.

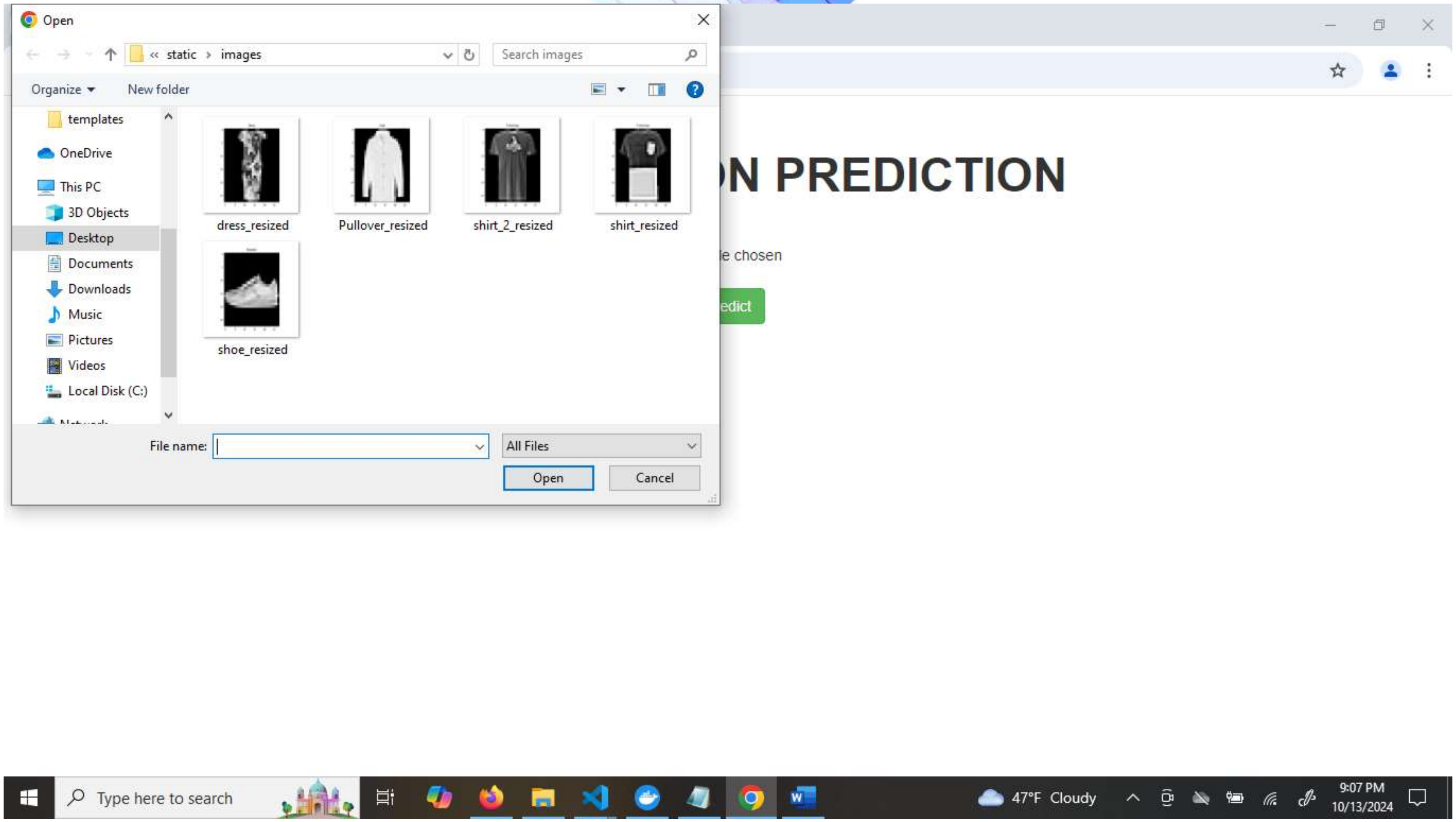
- This projects shows how a deep learning model that has been trained and saved in a desired format (e.g. .keras, .h5, .hdf5 etc) can be deployed as an API endpoint, using the lightweight Flask API on Docker.

Some Results

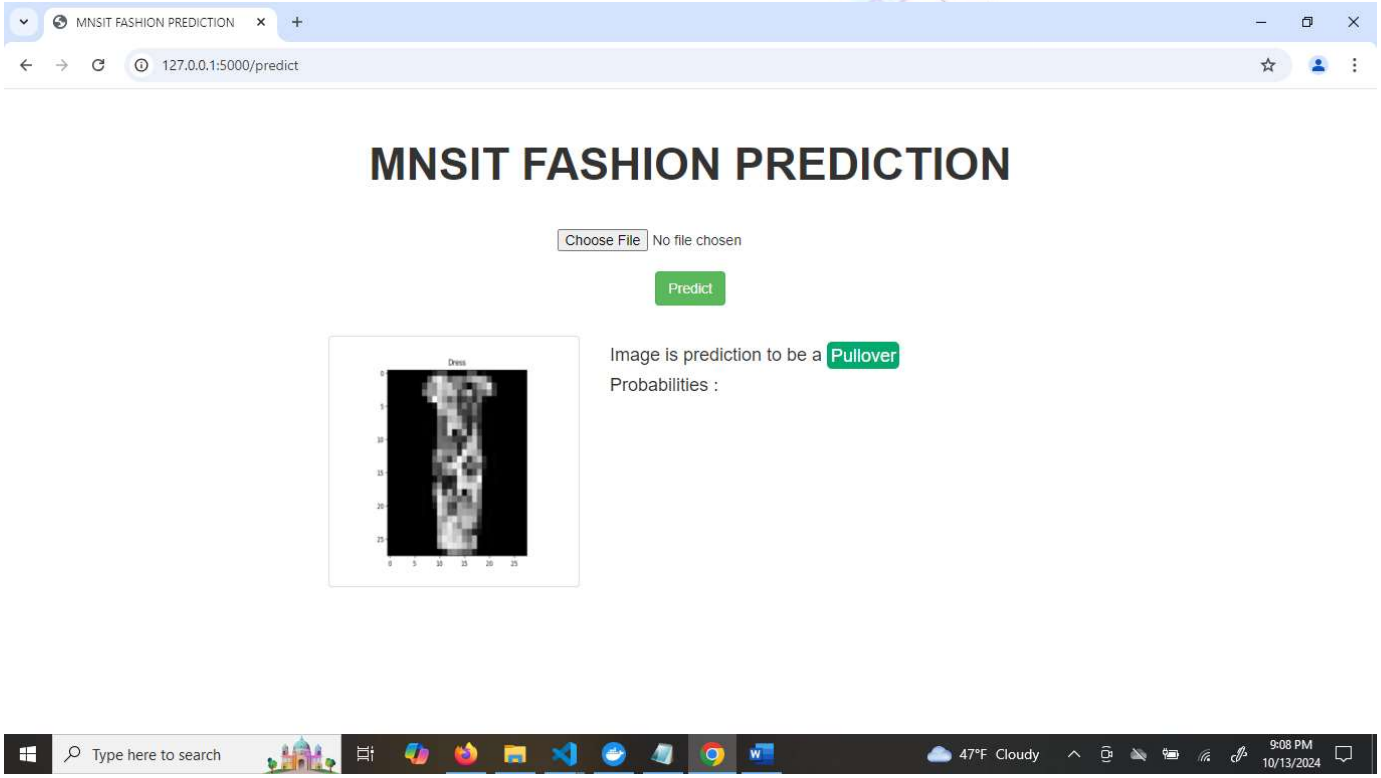


MNIST FASHION PREDICTION

Some Results



Some Results



MNSIT FASHION PREDICTION

Image is prediction to be a Pullover

Probabilities :

Focus has been on deployment of the model as an end point API using Flask.

Prediction accuracy shall be improved in future iterations of the project.

Github Links

Problem 1

https://github.com/manuelbomi/Data-Engineering-ETL-Using-PostgreSQL-Docker-and-Streamlit-Front-End.git

Problem 2

https://github.com/manuelbomi/Deep-Learning-Image-Prediction-with-Flask-API-End-Point-on-Docker.git

Problem 3

https://github.com/manuelbomi/WorkPlan-for-ChemBERTa-NLP-Algorithm-Implementation-at-BPC.git

Presentation Slides

https://github.com/manuelbomi/Presentation.git

Questions/Suggestions

Thanks to the Deloitte SFL team
for the opportunity