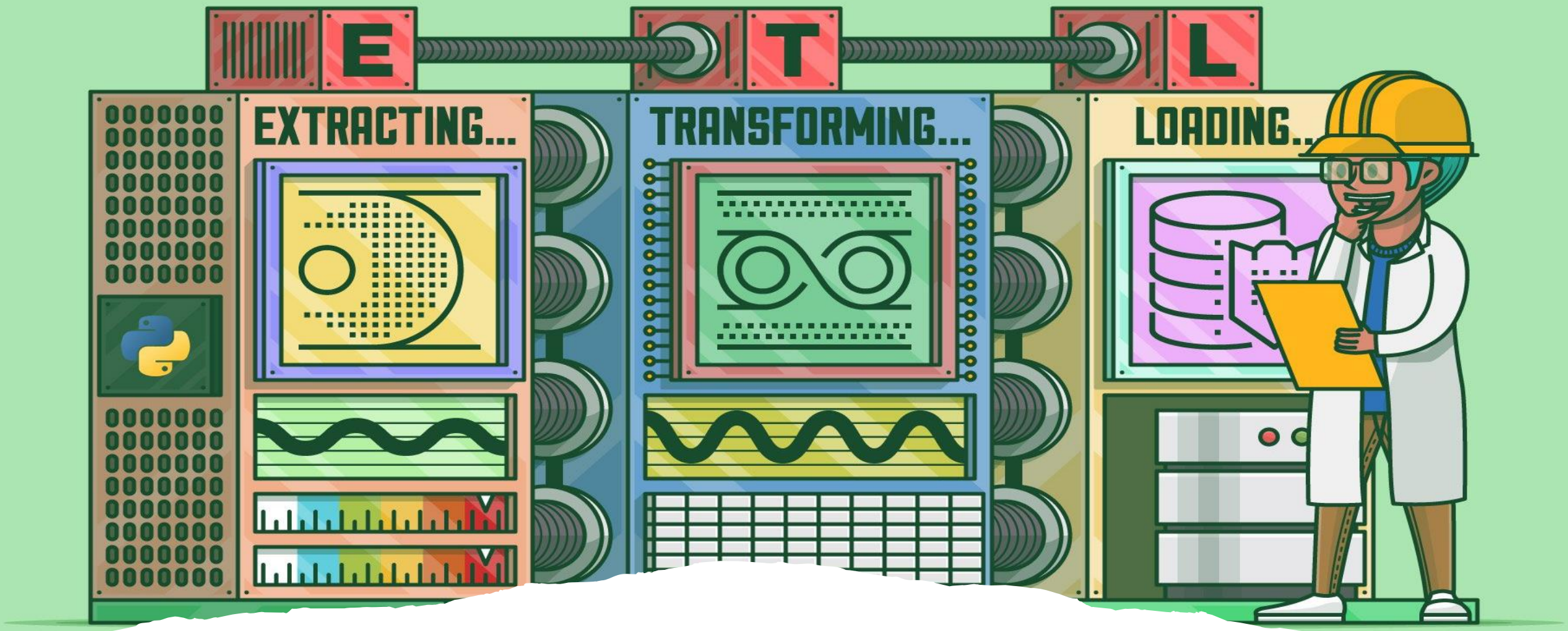


THERATTIL ANITTA SAJU

DATA ENGINEER (4.5 years)

CONTENTS

01 ABOUT ME

02 TECH STACK

03 PROJECT 1

04 PROJECT 2

05 PROJECT 3

06 PROJECT 4

07 PROJECT 5

08 CERTIFICATIONS

09 ACHEIVEMENTS

10 CONTACT DETAILS

About Me

I have been consistently appreciated for my accountability, technical acumen, and out-of-box thinking by my peers and leaders throughout my 4.5 years of professional experience in the field of Big Data and Data Engineering. I am a quick learner, a proactive problem solver, and open to learning new platforms. I look forward to being a part of an organization that would help me grow and in turn help me use my skills to contribute to the company's vision.

2017.09 — **Graduation**

2017.12 — **System Engineer** — Infosys — *Client* — Apple

2019.12 — **Senior Analyst – Product Development** — American Express

2022.03 — **Immigrated to Korea**

2022.04

Tech Stack - I

Programming Languages	python, UNIX shell Scripting
Database	SQL
Big Data	hadoop HDFS; PySpark; Spark SQL; HIVE, TEZ; APACHE SQOOP
Data Presentation tools	tableau; matplotlib; X Pivot Tables

Tech Stack - II

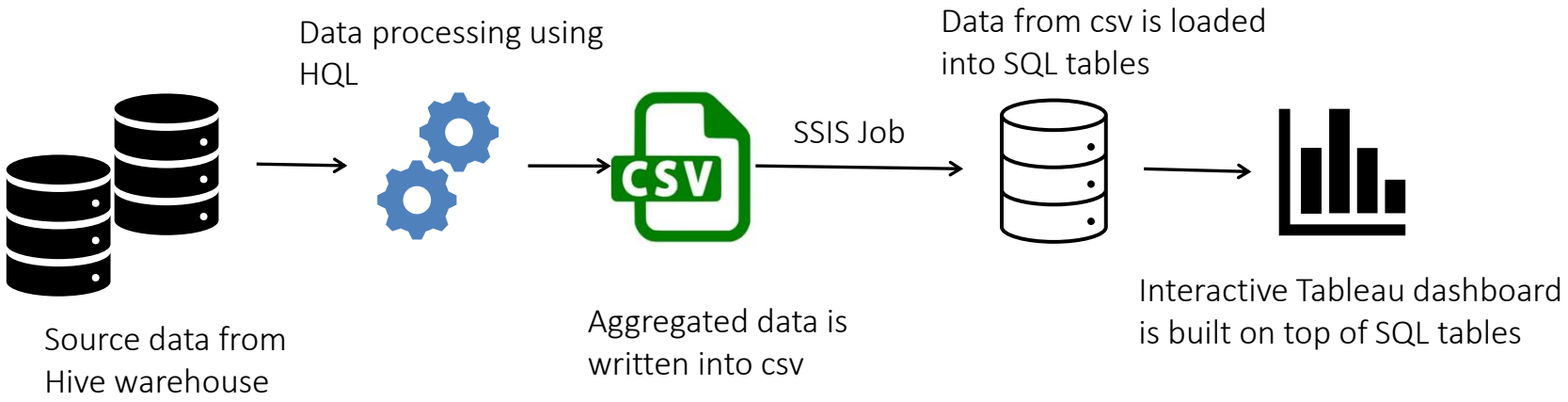
BI Tools	Microsoft SQL Server SSIS Microsoft SQL Server Analysis Services
Other skills/Tools	jupyter teradata. R jethro git

Project 1 : Fraud Dashboard

Why ? :
To report key fraud metrics to senior leadership and strategic partners for daily insights into fraud trends

What's new ? :
This is a daily interactive dashboard with monthly and YoY components which is also capable of interactively visualizing trends in different categories like markets and types of fraud across years.

Flow diagram :



Technology Used

HIVE TEZ
+ableau
Microsoft SQL Server SSIS
SQL

Project 2 : Authorization Reporting

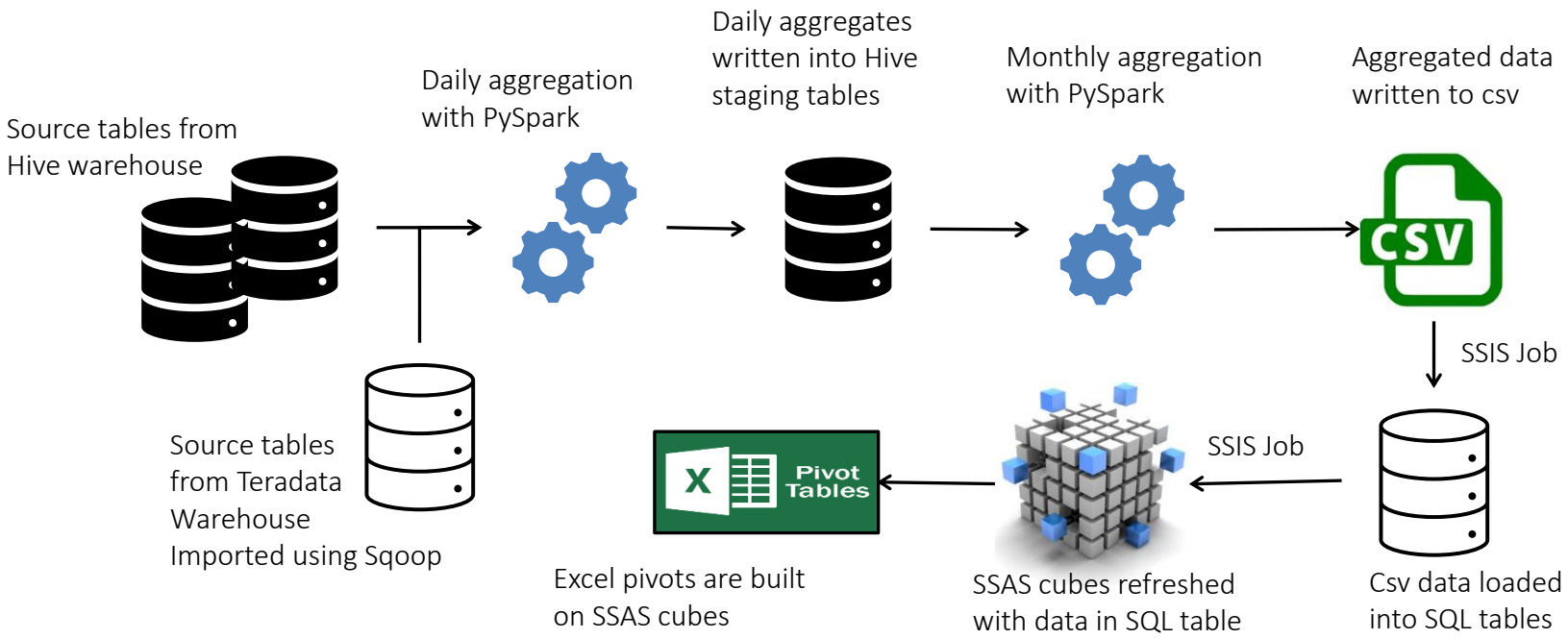
Why ? :

To report key Authorization metrics to senior leadership and strategic partners for monthly insights into transaction authorization trends.

What's new ? :

Input data has a high volume (~5 million records per day) with 30+ columns being used, and the size of the output file is ~80 GB per month after aggregations. Nevertheless, using multiple optimization techniques, monthly processing is completed within 3 hrs.

Flow diagram :



Technology Used

APACHE SQOOP

SQL

Microsoft SQL Server SSIS

Spark SQL

Microsoft SQL Server Analysis Services

Project 3 : Event Engine Revamped

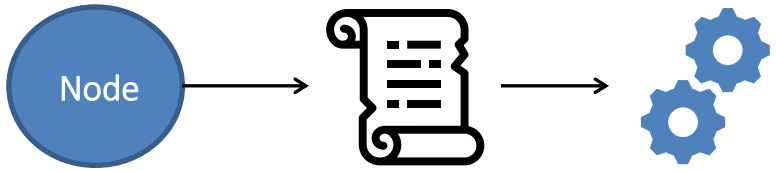
Why ? :

A workflow scheduler was created to tackle the problems of rigidity and redundancy in the existing infrastructure.

What's new ? :

Is customizable to any process with minor alterations and all the dynamic parameters can be passed via the GUI.

Flow diagram :



Event Engine Node Wrapper script

If all conditions are met,
then the script to be executed
is submitted.

Wrapper behaves differently based on node type variable passed from the node. Node types are :

- ➢ data validation
- ➢ processing
- ➢ history processing
- ➢ trigger node
- ➢ exit check nodes

Wrapper is a shell script, and it can :

- ✓ Pass extra arguments
- ✓ Execute .py, .hql, .sh scripts

Example of checks done by wrapper before exe cuting the script passed :

- ✓ Process sequence checks
- ✓ Whether another instance is running
- ✓ Resuming or restarting process check
- ✓ Process exit checks

Technology Used

UNIX shell Scripting

Note:
Event Engine is an American Express home-grown scheduler, and this project is built on top of it.

Project 4 : Regression Testing

Why ? :

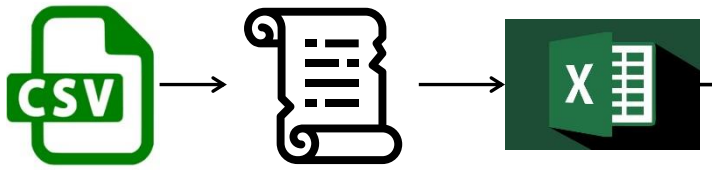
An automation utility that can be used for an extensive comparison of metrics and dimensions between two hive tables or between baseline data in excel to a hive table. Useful during UAT or migrations.

What's new ? :

This utility can compare data for all the selected metrics across all the selected dimensions and as it is built using PySpark and python, it has good performance on large datasets as well.

Flow diagram :

Python script uses pandas and PySpark for aggregations



Configs are provided in csv, like :
- ✓ Hive table names
- ✓ Metrics and Dimensions
- ✓ Group by columns if any,
- ✓ filters if any.

Table1	db_dev.customer_details
Table2	db_prod.customer_details
Filter	year=2021, month=10

Country	Total Amount 1	Total Amount 2	Diff Total Amount	Diff % Total Amount
US	980	1000	20	2.0408
IN	50	50	0	0
AU	67	68	1	1.4925
BR	0	20	20	100

Table1	db_dev.customer_details
Table2	db_prod.customer_details
Filter	year=2021, month=10

Metric Name	Table 1	Table 2	Diff	Diff %
Total Amount	1097	1138	41	3.7375
#Customers	1000000	1000000	0	0
Total Default	50	48	-2	-4

Test results are written as two sheets in excel:
Sheet1: Comparison of metric aggregates along dimension categories
Sheet2: Comparison of overall metric aggregates

Technology Used

Pandas

HIVE TEZ

UNIX shell Scripting

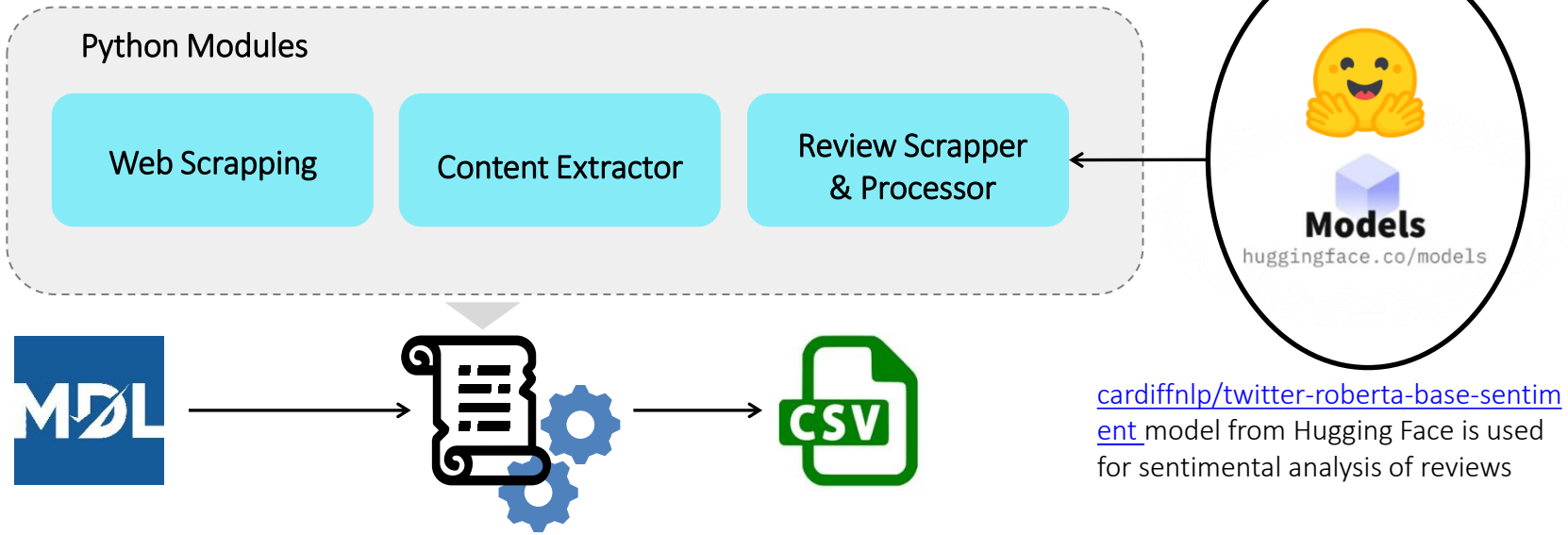
PySpark

Project 5 : MyDramaList Scrapper 2022.08

Why ? :
To generate a comprehensive dataset of the top 500 dramas, that contains general details along with review and reviewer details. Has been uploaded on Kaggle for public use.

What's new ? :
This data has been scrapped from MyDramaList website and sentimental analysis has been performed using the Hugging Face model. Multithreading and Multi-processing approaches were used to de-crease the time of execution and this script can also run for fetching all the dramas on MyDramaList.

Flow diagram :



Drama details are scrapped from MDL page

Scrapped data is processed using python.
Code is available here.

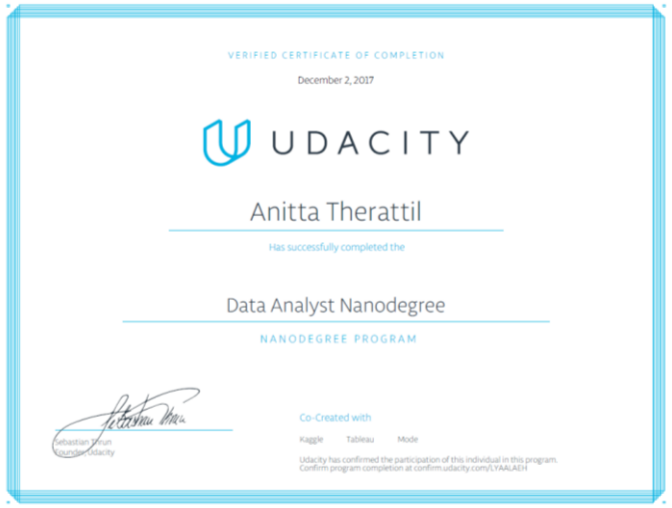
Processed data is written in csv. Dataset can be found on Kaggle.

cardiffnlp/twitter-roberta-base-sentiment model from Hugging Face is used for sentimental analysis of reviews

Technology Used

jupyter
Beautiful Soup
python
git

CERTIFICATIONS



Achievements

Leadership in Action Award

Global Education Center Stars Award

Bronze Medalist in Engineering (EC)

ANITTA SAJU THERATTIL

Data Engineer

+82-10-9640-7835

anittasaju1996@naver.com

https://github.com/anittasaju1996

www.linkedin.com/in/anitta-therattil

경기도 성남시 분당구