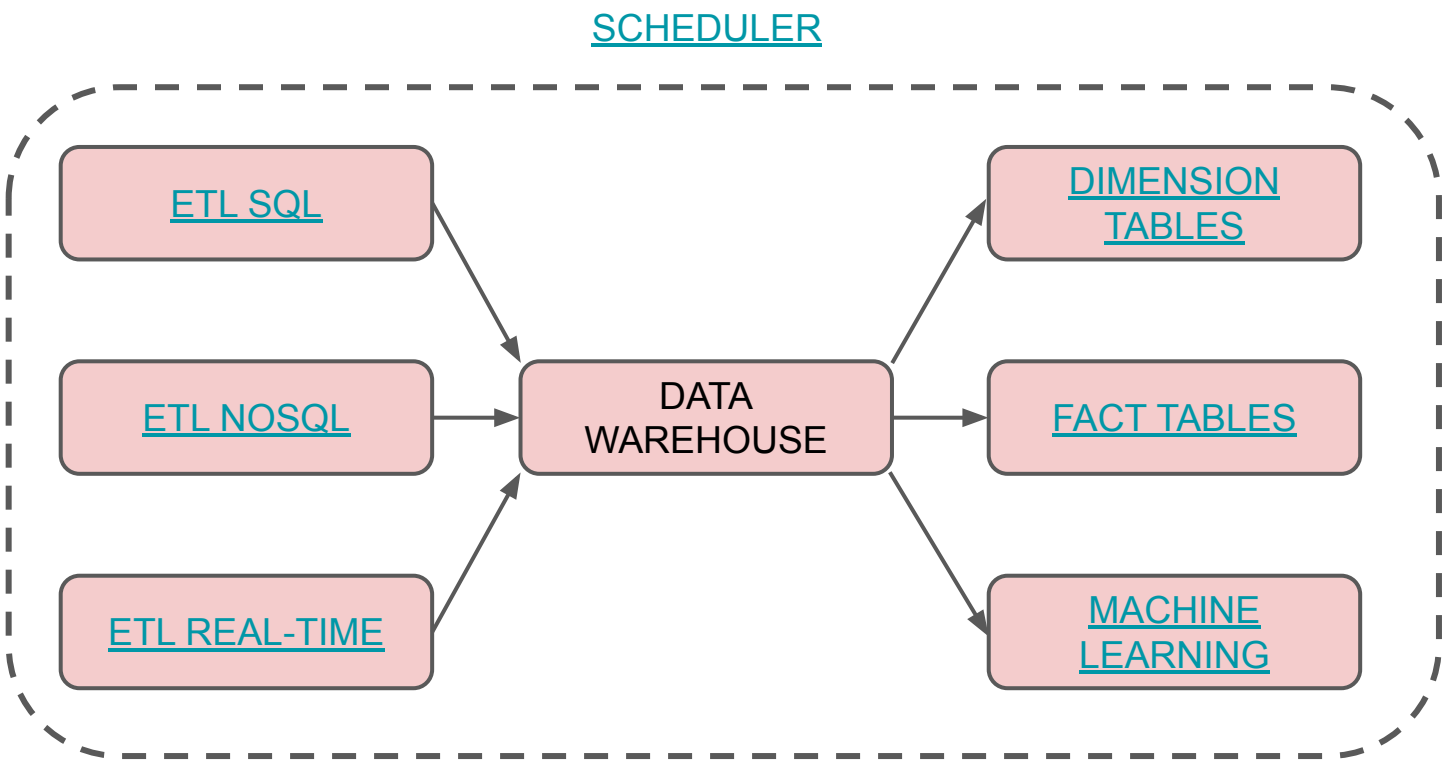
FINAL PROJECT PRESENTATION

DATA ENGINEER BATCH 9

AGNES SEPTILIA

WORKFLOW


SCHEDULER
ETL SQL
ETL NOSQL
ETL REAL-TIME
DATA WAREHOUSE
DIMENSION TABLES
FACT TABLES
MACHINE LEARNING

ETL SQL

Using Spark, extract data from CSV and Load to MySQL through JDBC

Using Spark, extract data from MySQL and Load to Postgres through JDBC

Tools :

- Spark
- Python (VS Code)

Libraries :

- Findspark
- PySpark

ETL SQL

Home Credit Default Risk Application Train

postgres Databases postgres Schemas final_project Tables home_credit_default_risk_application_train

home_credit_default_risk_applica | Enter a SQL expression to filter results (use Ctrl+Space)

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY
1	100,002	1	Cash loans	M	N	Y
2	282,780	0	Cash loans	F	Y	Y
3	100,003	0	Cash loans	F	N	N
4	100,004	0	Revolving loans	M	Y	Y
5	100,006	0	Cash loans	F	N	Y
6	100,007	0	Cash loans	M	N	Y
7	100,008	0	Cash loans	M	N	Y

Home Credit Default Risk Application Test

postgres Databases postgres Schemas final_project Tables home_credit_default_risk_application_test

home_credit_default_risk_applicatior | Enter a SQL expression to filter results (use Ctrl+Space)

	SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDI
1	306,030	Cash loans	F	N	N	
2	100,001	Cash loans	F	N	Y	
3	306,047	Cash loans	F	Y	N	
4	306,051	Cash loans	F	N	N	
5	306,063	Cash loans	F	Y	Y	
6	100,005	Cash loans	M	N	Y	
7	306,066	Cash loans	M	Y	Y	

ETL NoSQL

Data from MongoDB Atlas: **Sample Training Zips** and **Companies**

For **Zips**, flatten the **loc** fields into Latitude and Longitude

For **Companies**, take non-array fields and flatten first value of **offices** field

PostgreSQL

Load to Data Warehouse

Tools :

- MongoDB
- Python (VS Code)

Libraries :

- PyMongo
- Pandas
- SQL Alchemy

mongoDB®

ETL NoSQL

Sample Training Zips

sample_training_zips | Enter a SQL expression to filter results (use Ctrl+Space)

	id	city	zip	latitude	longitude	pop	state
1	5c8eccc1caa187d17ca6ed19	BAILEYTON	35019	34.268298	86.621299	1,781	AL
2	5c8eccc1caa187d17ca6ed16	ALPINE	35014	33.331165	86.208934	3,062	AL
3	5c8eccc1caa187d17ca6ed18	ACMAR	35004	33.584132	86.51557	6,055	AL
4	5c8eccc1caa187d17ca6ed1b	BLOUNTSVILLE	35031	34.092937	86.568628	9,058	AL
5	5c8eccc1caa187d17ca6ed1c	BRIERFIELD	35035	33.042747	86.951672	1,282	AL
6	5c8eccc1caa187d17ca6ed25	NEW SITE	35010	32.941445	85.951086	19,942	AL
7	5c8eccc1caa187d17ca6ed2f	EMPIRE	35063	33.825589	87.016139	2,429	AL

Sample Training Companies

Properties | Data | ER Diagram | postgres | Databases | postgres | Schemas | final_project | Tables | sample_training_companie

sample_training_companies | Enter a SQL expression to filter results (use Ctrl+Space)

		overview	total_money_raised	offices_description	offices_address1	offices_addre
1	TC 2013	<p>Wetpaint is a technology platform company that	$39.8M		710 - 2nd Avenue	Suite 1100
2	TC 2013	<p>StumbleUpon is the easiest way to discover new a	$18.5M			
3	C 2013	<p>Currently in public beta, Omnidrive makes it easy	$800k		Suite 200	654 High Street
4	TC 2013	<p>Slacker is the most complete music service on Ear	$73.1M		16935 W. Bernardo Dr. Suite 101	
5	TC 2013	<p>Babelgum is an integrated web and mobile video	$13.2M			
6	2013	<p>Cisco designs and sells hardware, software, netw	$2.5M	Headquarters	170 West Tasman Dr.	
7	C 2013	<p>Powerset is a search engine focused on natural la	$22.5M	[NULL]	475 Brannan St	

ETL Real Time

Extract the data from **freeforexapi.com** using **Kafka Producer**

→

Transform the data received in json → to information table

→

Load the data to Postgres data warehouse using **Kafka Topic Consumer**

→

PostgreSQL

Tools :

- Kafka
- Python (VS Code)

Libraries :

- Kafka
- json
- SQL Alchemy

kafka

ETL Real Time

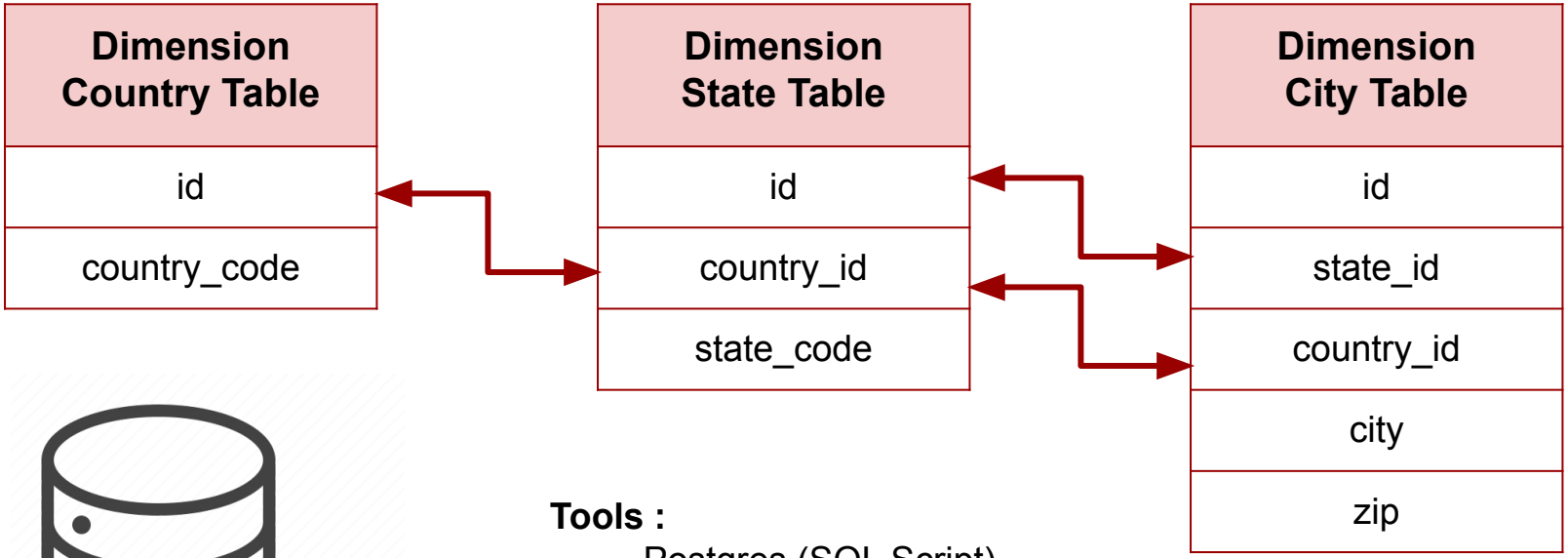
Topic Currency

topic_currency | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	currency_id	currency_name	rate	timestamp
1	EURUSD	US Dollar	1.052255	2022-12-05 22:59:03
2	EURGBP	Pound Sterling	0.861733	2022-12-05 22:59:03
3	USDEUR	Euro	0.95034	2022-12-05 22:59:03
4	EURUSD	US Dollar	1.052776	2022-12-05 23:01:03
5	EURGBP	Pound Sterling	0.861686	2022-12-05 23:01:03
6	USDEUR	Euro	0.94987	2022-12-05 23:01:03
7	EURUSD	US Dollar	1.05296	2022-12-05 23:02:03
8	EURGBP	Pound Sterling	0.861827	2022-12-05 23:02:03

Query Dimensions Table



Tools :

- Postgres (SQL Script)

Data is taken from **sample_training_companies** + **sample_training_zips** table in Postgres Data Warehouse

Query Dimensions Table

Dimension Currency Table		
id	currency_name	currency_code

Tools :

- Postgres (SQL Script)

Data is taken from **topic_currency** table in Postgres Data Warehouse

Query Dimensions Table

Dim Country Table

dim_country | Enter a SQL expression to filter results (use Ctrl+Space)

	id	country_code
1	6c9d653b-6e0b-4bde-961d-1297f53d31d0	others
2	9181446b-c853-4291-bbc2-bddb032a0d9f	IND
3	09af3ad4-3679-46e1-b067-e98f248319e0	EGY
4	68e1fee5-17de-44f1-8e7e-6d0378e918bb	CHL
5	77accdc2-f06e-4a1e-b416-b61b123b8747	BGR
6	4188a1d9-67c9-458e-bcc6-a0cf2d3f131e	NLD

Dim State Table

dim_state | Enter a SQL expression to filter results (use Ctrl+Space)

	id	country_id	state_code
1	db08f0d2-0438-4dc7-9037-90aa1be0fac6	e7cfcd56-a550-4a08-8bcd-c82d90fdba7a	others
2	809968fb-f9a6-4614-8196-fec3d15e5137	f554df66-eaa3-4148-b1e4-8e649e811a7a	DC
3	d154f3fc-bc7c-49e7-9c00-34bfb1e36e2c	6c9d653b-6e0b-4bde-961d-1297f53d31d0	CA
4	6c0a6c2b-d3bc-41f8-8e30-f48607b8f154	f554df66-eaa3-4148-b1e4-8e649e811a7a	IA
5	a763163a-378a-4372-bb11-e1f3892832f3	6c9d653b-6e0b-4bde-961d-1297f53d31d0	MS
6	72b666e4-71b6-4c0a-b795-e9954855a4b1	f554df66-eaa3-4148-b1e4-8e649e811a7a	NV

Query Dimensions Table

Dim City Table

dim_city | Enter a SQL expression to filter results (use Ctrl+Space)

	id	country_id	state_id	city_name	zip_code
1	8bfec459-f59d-48c7-bf03-35320a7bf9fc	c01258e2-5643-49b4-9539-4b4b7ca7f259	f5e22cb1-2666-4de8-a44b-044e34ead5e0	Salt Lake City	84117
2	81521735-7181-4ba2-97d2-b2fecac0f86b	d1fc86d0-b39b-47bd-8f2e-ab4090ef7f04	447b82d4-bc9e-41aa-9ebf-1e9b8ce88575	Maroochydore	4558
3	8ddfe3b0-4124-4e0d-aed4-4a641c639519	d1fc86d0-b39b-47bd-8f2e-ab4090ef7f04	447b82d4-bc9e-41aa-9ebf-1e9b8ce88575	sydney	2010
4	604ee4f8-0b37-4088-9800-c7a52d00fac5	d1fc86d0-b39b-47bd-8f2e-ab4090ef7f04	447b82d4-bc9e-41aa-9ebf-1e9b8ce88575	Sydney	2000
5	a02eff30-e2fb-405a-abde-84411b44f76a	d1fc86d0-b39b-47bd-8f2e-ab4090ef7f04	447b82d4-bc9e-41aa-9ebf-1e9b8ce88575	Sydney	2154
6	0f606868-3a89-4288-8054-fd6b197d2801	a5b4ad13-edf3-46b9-9b31-881b1f95df21	1c89220c-a8a1-4a30-ba12-51c9df82339f	Linz	4020
7	60001b7a-6979-41b1-9e2b-3bbe3f153b95	44d58f47-7687-41a7-80b3-05a5be54fc51	1f490b4c-9f8f-43f6-b0bd-dded724345d3	Brussel	1170
8	8963e5e4-27f6-4ff6-a6fe-375edad2a68f	77accdc2-f06e-4a1e-b416-b61b123b8747	fc6ae122-f518-4ad0-8ef4-87b47fc64598	Sofia	1505

Dim Currency Table

dim_currency | Enter a SQL expression to filter results (use Ctrl+Space)

	id	currency_name	currency_code
1	d4c210ef-7f9d-4f16-b4e2-eec63a892a9b	US Dollar	EURUSD
2	d0b6f9f3-5bad-4d36-bd20-bd084d053652	Pound Sterling	EURGBP
3	cb18b3b1-a5a6-4e7c-bf41-8518ac2aa8f1	Euro	USDEUR

Query Facts Table

Fact Total City and Office per State Table
state
total_city
total_office

Fact Daily Average Currency Table
currency_id
currency_name
day
avg_rate

Fact Monthly Average Currency Table
currency_id
currency_name
end_of_month
avg_rate

Tools :

- Postgres (SQL Script)

Function :

- Airflow macros for scheduler

Query Facts Table

Fact Total City and Office per State Table

fact_total_per_state | Enter a SQL expression to filter resu

Grid | Text

	state	total_city	total_office
1	AL	7	13
2	AR	4	5
3	AZ	9	51
4	CA	204	1670
5	CO	24	88
6	CT	18	31
7	DC	7	39
8	DE	6	9
9	FL	66	156
10	GA	21	87
11	HI	3	5
12	IA	7	13
13	ID	7	8
14	IL	32	122

Query Facts Table

Fact Daily Average Currency Table

fact_daily_avg_currency | Enter a SQL expression to filter results (use Ctrl+Space)

	currency_id	currency_name	day	avg_rate
1	EURGBP	Pound Sterling	2022-12-12	0.8596201667
2	EURUSD	US Dollar	2022-12-12	1.0559505
3	USDEUR	Euro	2022-12-12	0.9470141667

Fact Monthly Average Currency Table

fact_monthly_avg_currency | Enter a SQL expression to filter results (use Ctrl+Space)

	currency_id	currency_name	end_of_month	avg_rate
1	EURGBP	Pound Sterling	2022-12-31	0.8620127463
2	EURUSD	US Dollar	2022-12-31	1.0521541045
3	USDEUR	Euro	2022-12-31	0.9504351642

Machine Learning

Extract csv data from **kaggle : Home Credit Default Risk: Train and Test** dataset

Data Cleaning: remove column with >60% null

For remained columns, **impute** null values

Fitting Model **Logistic Regression**, check the **prediction** and **probability**

PostgreSQL

Load to Data Warehouse

Tools :

- Python (Jupyter Notebook)

Libraries :

- Pandas
- Scikit-Learn
- Imblearn
- SQL Alchemy & Psycopg2

scikit learn

Machine Learning

Home Credit Default Risk Application Train Clean Table

postgres Databases postgres Schemas final_project Tables home_credit_default_risk_application_train_clean

home_credit_default_risk_application_trai | Enter a SQL expression to filter results (use Ctrl+Space)

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDRE
1	100,002	1	Cash loans	M	N	Y	
2	282,780	0	Cash loans	F	Y	Y	
3	100,003	0	Cash loans	F	N	N	
4	100,004	0	Revolving loans	M	Y	Y	
5	100,006	0	Cash loans	F	N	Y	
6	100,007	0	Cash loans	M	N	Y	

Home Credit Default Risk Application Test Clean Table

postgres Databases postgres Schemas final_project Tables home_credit_default_risk_application_test_clean

home_credit_default_risk_application_tes | Enter a SQL expression to filter results (use Ctrl+Space)

	SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_
1	306,030	Cash loans	F	N	N	0	
2	100,001	Cash loans	F	N	Y	0	
3	306,047	Cash loans	F	Y	N	0	
4	306,051	Cash loans	F	N	N	0	
5	306,063	Cash loans	F	Y	Y	1	
6	100,005	Cash loans	M	N	Y	0	

Machine Learning

Home Credit Default Risk Application Machine Learning Result Table

home_credit_default_risk_application_ml_ | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	123 SK_ID_CURR	123 prediction_target	ABC probability
1	306,030	0	{0.9486590082237689,0.05134099177623112}
2	100,001	0	{0.9163565538072872,0.08364344619271283}
3	306,047	0	{0.9557436731528267,0.04425632684717326}
4	306,051	0	{0.9199971008724842,0.0800028991275158}
5	306,063	0	{0.916849669313629,0.08315033068637095}
6	100,005	0	{0.8891978888380373,0.11080211116196272}
7	306,066	0	{0.9353502730477948,0.0646497269522052}

SCHEDULER

DAG Spark
Unscheduled

DAG MongoDB
Unscheduled

DAG Machine Learning
Unscheduled
Dependency: DAG Spark

DAG Dims Tables
Unscheduled
Dependency: DAG Spark and DAG MongoDB

DAG Fact Table Daily
Daily Schedule
Dependency: DAG Dims Table

DAG Fact Table Monthly
Monthly Schedule
Dependency: DAG Dims Table

Tools :

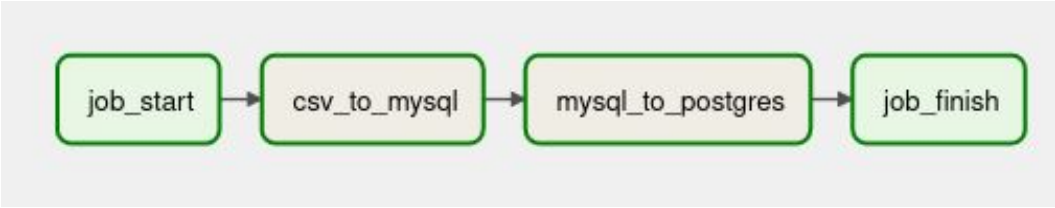
- Airflow
- Python (VS Code)

Libraries :

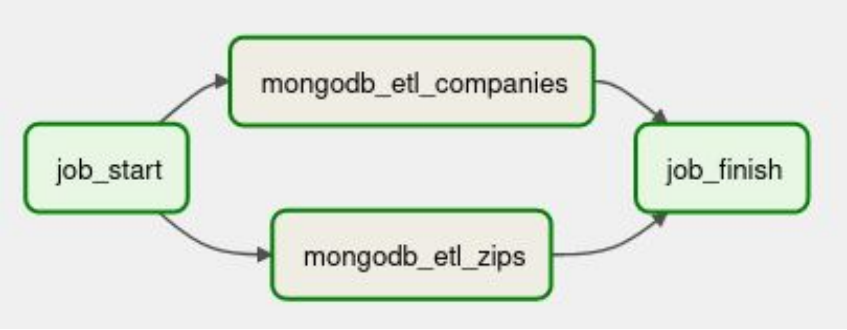
- Airflow
- Datetime

SCHEDULER

Dag: etl_spark



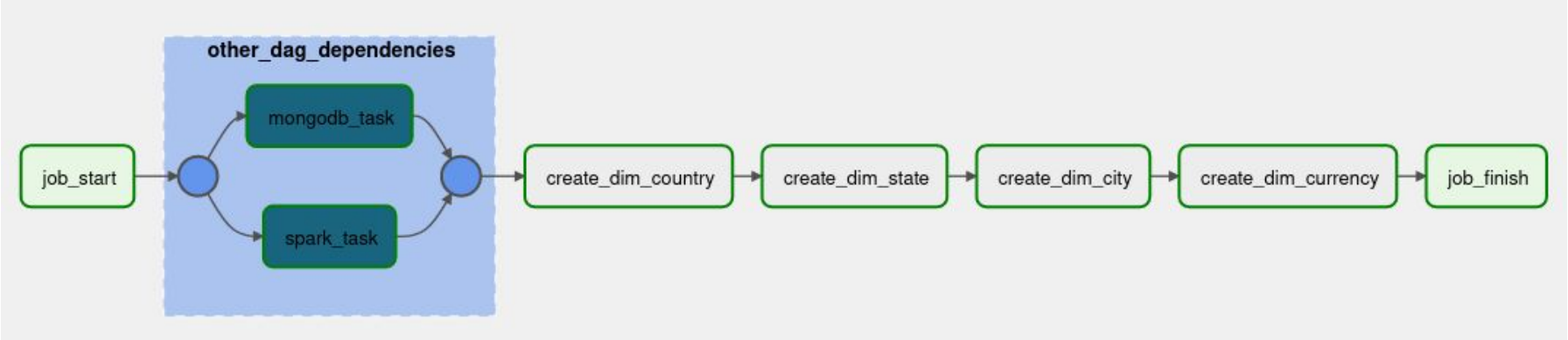
Dag: etl_mongodb



SCHEDULER

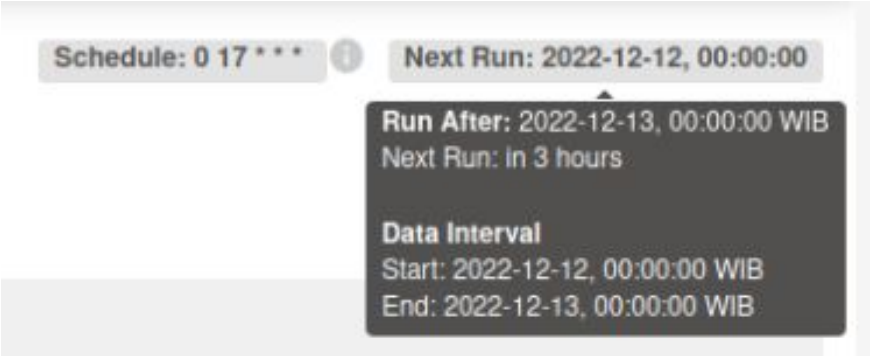
Dag: machine_learning

Dag: dim_tables



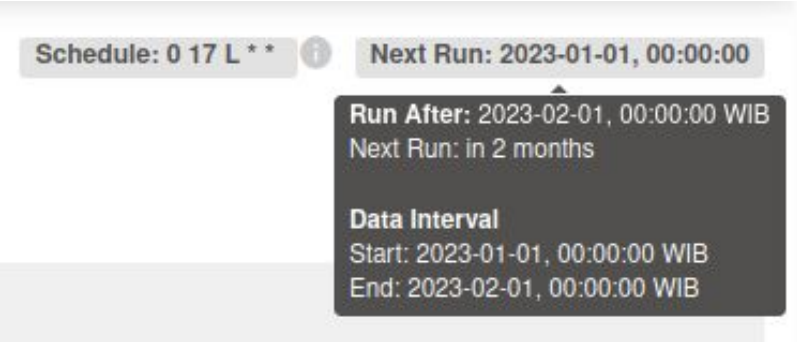
SCHEDULER

Dag: fact_daily_table



SCHEDULER

Dag: fact_monthly_table



GITHUB LINK

Code Source of the Final Project:

https://github.com/agnes-septilia/final_project_agnes_septilia.git

THANK YOU!