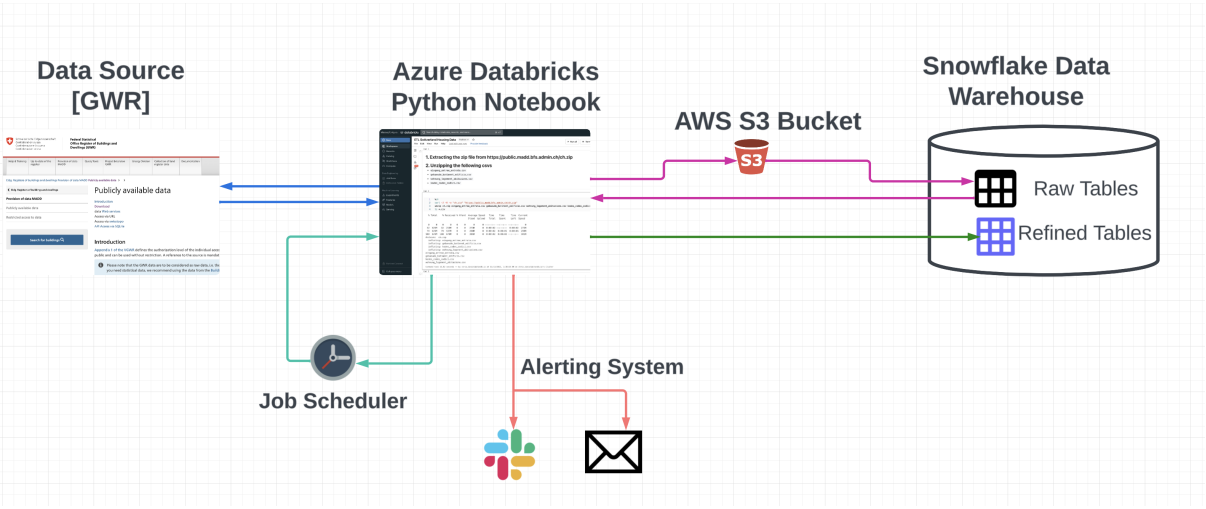
Data Engineering Code Test

Data Pipeline Architecture



Execution Mechanism

- The Python notebook sends a request to the GWR API

```
curl -J -O -o "ch.zip" "https://public.madd.bfs.admin.ch/ch.zip"
```

- The GWR API sends a zip file that contains the CSVs and additional files that are not needed for ETL
- Within the Python notebook, we extract only the necessary files and they are
  - `Kodes_codes_codici.csv`
  - `eingang_entree_entrata.csv`
  - `gebaeude_batiment_edificio.csv`
  - `Wohnung_logement_abitazione.csv`
- For accountability purposes, we append today's date to all the CSV files

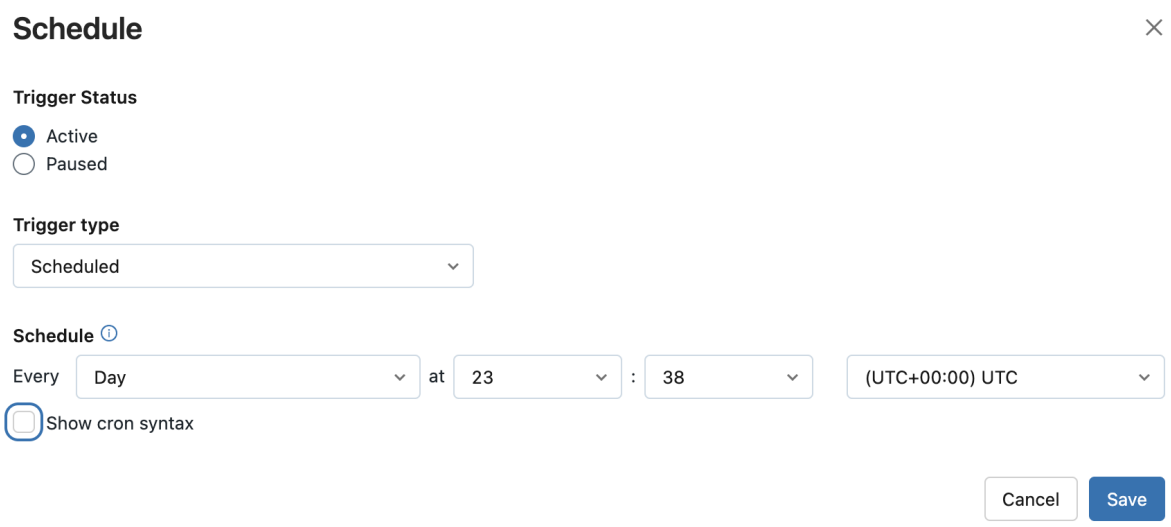
> 💡 E.g Kodes_codes_codici.csv is transformed to Kodes_codes_codici_*2023_11_13*.csv

- After CSV name transformation the files are loaded into two folders in Aws S3. They are
  - *Latest folder* - this contains only the recent pull
  - *Archive folder* - this contains all historical CSVs
- Files dropped on AWS S3 triggers a SQS notification. Snowpipe reads this notification and loads all the data into an external Snowflake stage.
- From the external Snowflake stage data is moved into raw tables.
- From the raw tables data is transformed and moved into refined tables

---

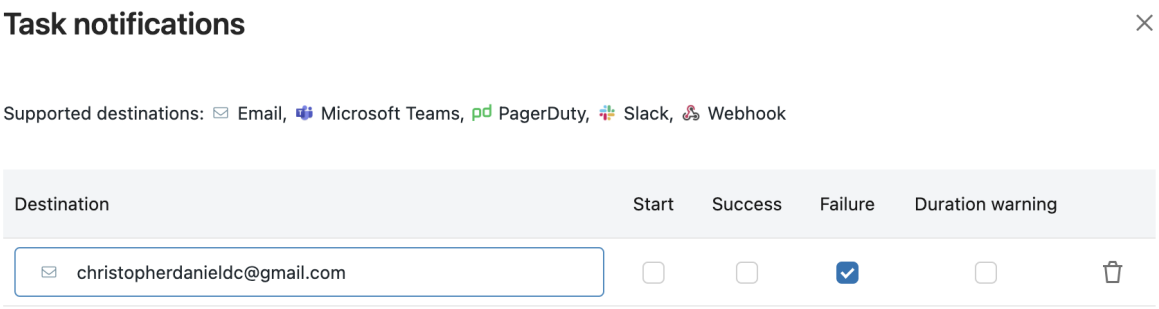
Automating the ETL

- The job scheduler can start the ETL process at a set time every day / hour etc
- The ETL process can also be triggered sporadically based on custom triggers.

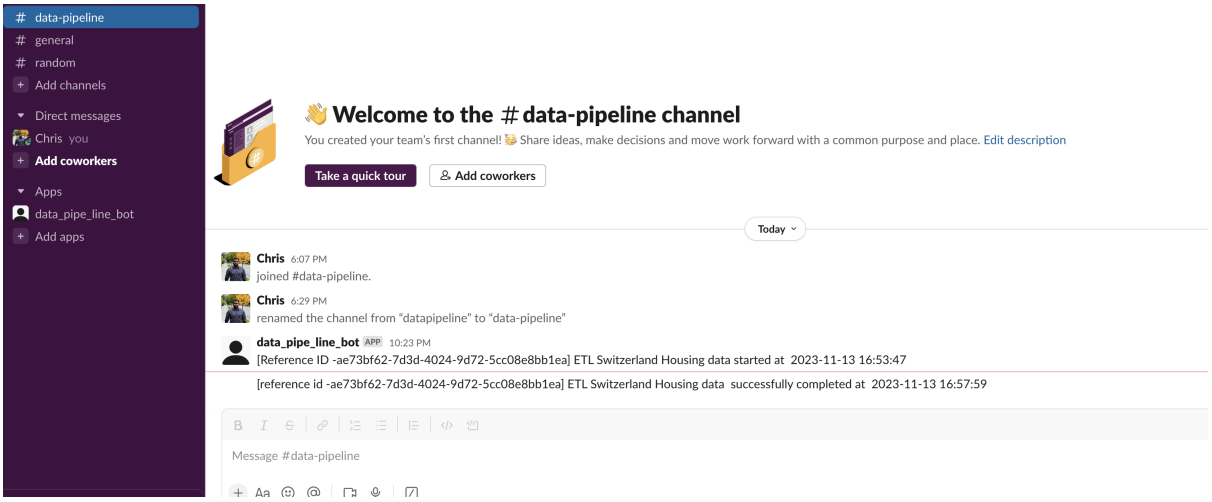


Reporting and Status Updates

- If the Python notebook fails to start, the job scheduler will send an email



- Once the Python notebook starts all notifications can be received via Slack



AWS S3 File Structure

```
.
└── chris-properti-ftp[S3 Bucket]/
    ├── latest/
    │   ├── Kodes_codes_codici_2023_11_13.csv
    │   ├── eingang_entree_entrata_2023_11_13.csv
    │   ├── gebaeude_batiment_edificio_2023_11_13.csv
    │   └── Wohnung_logement_abitazione_2023_11_13.csv
    └── archive/
        ├── Kodes_codes_codici/
        │   ├── -Kodes_codes_codici_2023_11_13.csv
        │   └── -Kodes_codes_codici_2023_11_12.csv
        ├── eingang_entree_entrata/
        │   ├── eingang_entree_entrata_2023_11_13.csv
        │   └── eingang_entree_entrata_2023_11_12.csv
        ├── gebaeude_batiment_edificio/
        │   ├── gebaeude_batiment_edificio_2023_11_13.csv
        │   └── gebaeude_batiment_edificio_2023_11_13.csv
        └── Wohnung_logement_abitazione/
            ├── Wohnung_logement_abitazione_2023_11_13.csv
            └── Wohnung_logement_abitazione_2023_11_13.csv
```

- *Latest folder* is used for processing the current pull
- *Archive folder* is used for storing historical pull

Snowflake Audit Table

- Every step of the run is recorded on a Snowflake table called `data_pipeline_audit_table`

	UNIQUE_ID	START_TIME	STEP	STATUS	END_TIME
1	931661b1-a4a3-4474-96	2023-11-13 05:51:47.098	Extracting CSVs from https://public.madd.bfs.admin.ch/ch.zip	ETL Started	2023-11-13 05:51:47.098
2	931661b1-a4a3-4474-96	2023-11-13 05:51:58.626	CSVs extracted to Azure Databricks	ETL Running	2023-11-13 05:51:58.626
3	931661b1-a4a3-4474-96	2023-11-13 05:51:59.230	CSVs are transformed	ETL Running	2023-11-13 05:51:59.230
4	931661b1-a4a3-4474-96	2023-11-13 05:52:01.130	Cleaning AWS S3 latest folder	ETL Running	2023-11-13 05:52:01.130
5	931661b1-a4a3-4474-96	2023-11-13 05:52:03.975	Truncating existing tables in Snowflake	ETL Running	2023-11-13 05:52:03.975
6	931661b1-a4a3-4474-96	2023-11-13 05:52:11.278	Loading wohnung_logement_abitazione.csv to S3	ETL Running	2023-11-13 05:52:11.278
7	931661b1-a4a3-4474-96	2023-11-13 05:52:19.017	Loading eingang_entree_entrata.csv to S3	ETL Running	2023-11-13 05:52:19.017
8	931661b1-a4a3-4474-96	2023-11-13 05:52:29.263	Loading gebaeude_batiment_edificio.csv to S3	ETL Running	2023-11-13 05:52:29.263
9	931661b1-a4a3-4474-96	2023-11-13 05:52:30.177	Loading kodes_codes_codic.csv to S3	ETL Running	2023-11-13 05:52:30.177
10	931661b1-a4a3-4474-96	2023-11-13 05:54:59.717	Loading data into Raw tables	ETL Running	2023-11-13 05:54:59.717
11	931661b1-a4a3-4474-96	2023-11-13 05:55:22.451	Loading data into Refined tables	ETL Running	2023-11-13 05:55:22.451
12	931661b1-a4a3-4474-96	2023-11-13 05:55:23.278	Cleaning directories	ETL Successfully Completed	2023-11-13 05:55:23.278

Reasoning Behind Choice of Tools

- Databricks Python Notebook
  - One of the best in the game
  - Comes with a job scheduler
  - I have personally migrated close to a million Asana tasks using this notebook
  - It is pay-per-use and costs $0.07 / DBU
- AWS S3
  - Best in class
  - A data pipe can be easily built between Snowflake & S3 in no time
  - Storage costs are much less compared to other major players
- Snowflake
  - Easy to commission [less than 5 minutes]
  - Can query billions of data at ease
  - Apart from SQL, supports other languages such as Python and Scala
  - All popular BI tools can be connected with Snowflake with ease