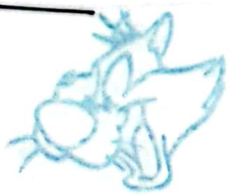
Week 2 : Data Ingestion

- Data Lake.
- workflow orchestration.
- Airflow locally.
- Ingesting data to local postgres.
- Transfer service (AWS > GCP)

Data lake:

- Central repository that hold big data from many sources.
- Data may be structured or not.
- Main goal of DL to ingest data as quickly as possible making it available to others.
- DL should be secure, scalable, Non-expensive

	Data Lake (DL)	Data warehouse (DW)
processing	unstructured data, raw undergone minimal processing	structured data, refined data has cleaned, preprocessed
size	Large, petabytes(1024)TB Data transformed in need only data stored indefinitely للأبد	small compared, TB, alway processed before ingestion, data may purged تنظف periodically.
users	Data scientist, analysist	Business analysist
steps	~~ETL "Export Trasfom load"~~ Extract ELT Transformation last layer AWS, GCP, AZure	~~EIL~~ ETL

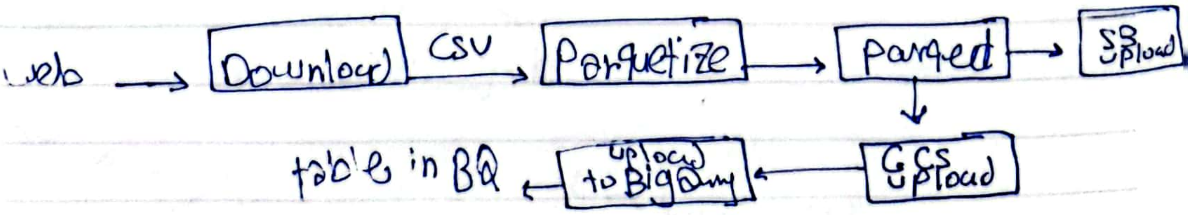


- Data lakes Come into existence because as companies started to realize importance of data, they soon Founded they couldn't ingest data to DW, but they didn't want to waste uncollected data when their devs hadn't finished developing necessary relatiaship for DW, so Data lake was born to collect any potentially useful data till structure and organise this data.

- Data swamp (مستنقع)
  (Data lakes gone wrong)
  - Data lakes are only useful if data can be processed from it that happen when:
    - No versioning of data.
    - Incompatible schemes for same data.
    - No metadata associated.
    - Joins between different dataset not possible.

Orchestration (تناغم) with Airflow:

For massive pipeline we call it work flow in previus work we downloaded & process csv file at same block, we should splet that to differant steps, separating help to avoid problems like internet lose and repeat download file.

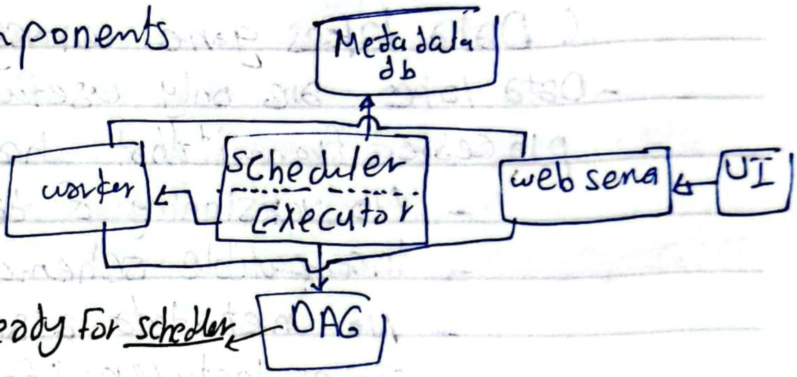


- parquet (باركت)
  Columnar storage datafile format, more efficient then csv.

- The previous graph are called Directed Acyclic Graph (DAG)
  it miss for loops
  and Flow is well defined.

Airflow Architecture:

Graph for airflow components



File with DAG Files ready for scheduler

- setting up docker (Airflow)
  - full version - lite version
  - we use docker compose for this task.

all done on your pc

Creating DAG "Ingest Data to GCP with Airflow"

. DAG is created as <u>python script</u> importing series of libiraries from airflow.

```
> with DAG (dag-id = "my-dag-name") as dag :
      op1 = Dummy operator (task-id = "task1")
      op2 . . . .
       op1 >> op2
executing op1 then op2 . . .
```

. we will focus on <u>operator tasks</u> :
- Bashoperator : simple bash command.
- Python operator : calls python method rather bash.

```
> download-dataset-task = Bashoperator (
    task-id = "download-dataset-task",
    bash-command = f"curl -sS {dataset-url} > {path . . . .")
                instead of wget
```

```
7 format-to-parquet-task = PythonOperator (
    task-id = "format-to-parquet-task",
    python-callable = format-to-parquet,
    op-kwargs = {
            "src-file" : f" {path . . }"
```

. we can schedule intervals to run code at a certin time.
using <u>Cron Job</u>

space فيه

```
              * * * * * → day of week
minute(0-59)  Hour  day     month    (0-6)         7(Sunday) also
              (0-23) month  (1-12)   (Sunday : Sat)
                     (1-31)
```

0 6 2 * *
means each month second day of this month
at 6 am.

for example:
~~Dobbn~~ load this file each file and ingest it
to dataset.

- schedule_interval = "0 6 2 * * "

Tricks

- URL = 'yellow-trip data-{{ execution-data .strftime(\'%Y-%m
  -%m\')}}.csv
  to download current data

```
with my-workflow as dag:
    download-task = BashOperator(
        task-id = 'download'
        bash-command = f curl -sS L {URL-Templet} > {output}'
```

to pass argument

scilent
show error

- Airflow offer series of predefined variable and macros.

- A DAG is idempotency(كسال ن) , if end result is identical regardless multiples times.

- we have also emailoperator send emails and even more, like GCP tools

2.3.3 ingest data to local postgres

- Using Airflow we can run certin code at certin time like
  - running python code to do something.
  - running bash.

- we are downloading NY- taxi data set using bash in airflow and want to repeat that every month using cron.
- After that we would ingest this data to postgress so we will use python operator and pass python file to it.

Hint to write in multible line press ctrl + D

- we can creat Network between 2 docker compose

- URL_TEMPLATE = URL_PREFIX + '/yellow-tripdata_{{ execution-data. strftime('%Y-%m') }}.csv'

get current time when we execute code.

2.4.1 GCP Transfer Service

- GCP Enable us to trasfer files from different resources like (AWS, Azure, URL..)

  speed of transfem : 790 MB/S

  Charge : .012 $/GB

. we can do trasfer using

- UI
- Teraform

  "Just sheet which you edit variables"