# Week 4:04-building-a-data-lake

### Data model

## staging\_events

id: VARCHAR NOT NULL [ PK ]

type: VARCHAR NOT NULL
actor\_id: BIGINT NOT NULL
actor\_name: VARCHAR NOT NULL
actor\_url: VARCHAR NOT NULL
repo\_id: BIGINT NOT NULL
repo\_name: VARCHAR NOT NULL
repo\_url: VARCHAR NOT NULL
public: BOOLEAN NOT NULL
created\_at: VARCHAR NOT NULL
org\_id: VARCHAR NOT NULL
org\_avatar\_url: VARCHAR NOT NULL
org\_gravatar\_id: VARCHAR NOT NULL
org\_login: VARCHAR NOT NULL
org\_url: VARCHAR NOT NULL

#### events

id: VARCHAR NOT NULL [ PK ]

created\_at: VARCHAR NOT NULL public: VARCHAR NOT NULL type: VARCHAR NOT NULL id\_actor: VARCHAR NOT NULL repo: VARCHAR NOT NULL date)oprt: VARCHAR NOT NULL

### repos

id: BIGINT NOT NULL [ PK ]

name: VARCHAR NOT NULL url: VARCHAR NOT NULL date oprt: VARCHAR NOT NULL

#### actors

id: BIGINT NOT NULL [ PK ]

avatar\_url: VARCHAR NOT NULL display\_login: VARCHAR NOT NULL gravatar\_id: VARCHAR NOT NULL login: VARCHAR NOT NULL url: VARCHAR NOT NULL date\_oprt: VARCHAR NOT NULL

# **Project Processing**

- 1. change directory to project 04-building-a-data-lake: \$ cd 04-building-a-data-lake
- 2. Applying code for saving jupyter lab (Any update on coding) sudo chmod 777.

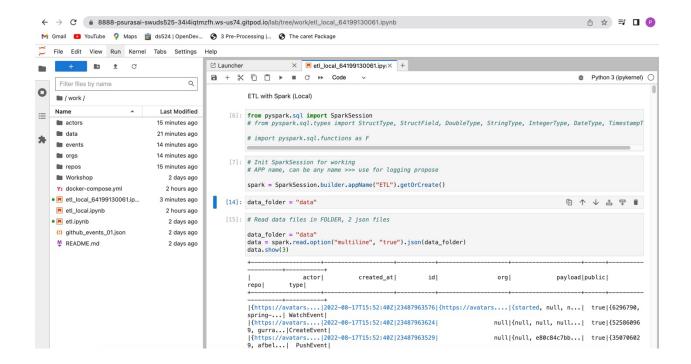
3. prepare environment workspace by Docker:

\$ docker-compose up

4. Open JupyterLab URL:



5. Execute the Notebook 'etl local 64199130061.ipynb' step by step:



6. Chcek the cleaned output data in folders, partition by 'date oprt':

### actors : [actors]( https://github.com/psurasai/SWU-DS525/tree/main/04building-a-data-lake/actors)

### repos : [repos] (https://github.com/psurasai/SWU-DS525/tree/main/04building-a-data-lake/repos)

### orgs : [orgs] ( https://github.com/psurasai/SWU-DS525/tree/main/04building-a-data-lake/orgs)

### events : [events] ( https://github.com/psurasai/SWU-DS525/tree/main/04building-a-data-lake/events)

<br>

7. Shutdown environment workspace:

\$ docker-compose down