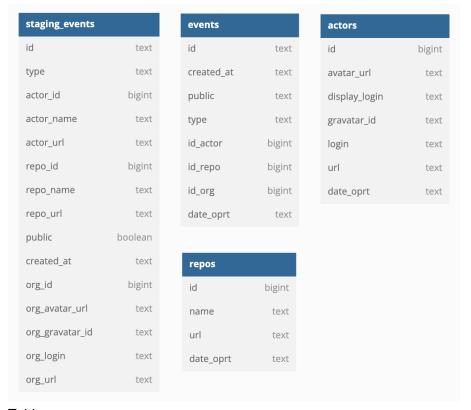
Lab4 – Project Building Datalake

Data model



<u>Tables</u>

staging events:

This is a temporary table using for stacking the data loaded from json files.

events:

This table collects the information related to event, appear the id to connect with table 'actors', 'repos' and 'orgs'.

There is a partition 'date_oprt' to identify the operation date.

actors:

This table collects the information related to actor.

There is a partition 'date_oprt' to identify the operation date.

repos:

This table collects the information related to repository.

There is a partition 'date_oprt' to identify the operation date.

Orgs:

This table collects the information related to organization.

There is a partition 'date_oprt' to identify the operation date.

Project implementation instruction

1. Reach the project repository 'swu-ds525/04-building-a-data-lake'

: \$ cd 04-building-a-data-lake

- (base) JC@Napchins-MacBook-Air swu-ds525 % cd 04-building-a-data-lake
 (base) JC@Napchins-MacBook-Air 04-building-a-data-lake %
- 2. Prepare the environment workspace thru 'docker-compose.yml'
 - : \$ docker-compose up
 - (base) JC@Napchins-MacBook-Air swu-ds525 % cd 04-building-a-data-lake(base) JC@Napchins-MacBook-Air 04-building-a-data-lake % docker-compose up
- 3. Open the JupyterLab URL

```
● (base) JC@Napchins-MacBook-Air swu-ds525 % cd 04-building-a-data-lake
○ (base) JC@Napchins-MacBook-Air 04-building-a-data-lake % docker-compose up
    [+] Running Z/2

# Network 04-building-a-data-lake_default Created

# Container 04-building-a-data-lake-pyspark-notebook-1

Attaching to 04-building-a-data-lake-pyspark-notebook-1

Attaching to 04-building-a-data-lake-pyspark-notebook-1

Fintered start.sh w
                                                                                                                                                                                                          notebook-1
Entered start.sh with args: jupyter lab
/usr/local/bin/start.sh: running hooks in /usr/local/bin/before-notebook.d as uid / gid: 1000 / 100
/usr/local/bin/start.sh: running script /usr/local/bin/before-notebook.d/spark-config.sh
/usr/local/bin/start.sh: done running hooks in /usr/local/bin/before-notebook.d
Executing the command: jupyter lab
[II 2022-10-04 12:19:11.505 ServerApp] jupyterlab | extension was successfully linked.
[I 2022-10-04 12:19:11.515 ServerApp] mbclassic | extension was successfully linked.
[I 2022-10-04 12:19:11.517 ServerApp] Writing Jupyter server cookie secret to /home/jovyan/.local/shar
       04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
      e/jupyter/runtime/jupyter_cookie_secret
        04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
                                                                                                                                                                                                             [I 2022-10-04 12:19:11.705 ServerApp] notebook_shim | extension was successfully linked.
[I 2022-10-04 12:19:11.723 ServerApp] notebook_shim | extension was successfully loaded.
[I 2022-10-04 12:19:11.725 LabApp] JupyterLab extension loaded from /opt/conda/lib/python3.10/site-pac
                                                                                                                                                                                                            [I 2022-10-04 12:19:11.725 LabApp] JupyterLab application directory is /opt/conda/share/jupyter/lab [I 2022-10-04 12:19:11.728 ServerApp] jupyterlab | extension was successfully loaded. [I 2022-10-04 12:19:11.732 ServerApp] nbclassic | extension was successfully loaded. [I 2022-10-04 12:19:11.733 ServerApp] Serving notebooks from local directory: /home/jovyan [I 2022-10-04 12:19:11.733 ServerApp] Jupyter Server 1.18.1 is running at: [I 2022-10-04 12:19:11.733 ServerApp] http://8d2b05828758:8888/lab?token=d4d5e05447577a0bda0fa8c02c258
      kages/Jupyrerlab
04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
     04-building-a-data-lake-pyspark-notebook-1 | [I 2022-10-04 12:19:11.733 ServerApp] Jupyter Server 1.18.1 is running at: 04-building-a-data-lake-pyspark-notebook-1 | [I 2022-10-04 12:19:11.733 ServerApp] http://8d2b05828758:8888/lab?token=d4d5e05447577a0bda0fa8c02c258 ce983331ae9ff9e8827 | [I 2022-10-04 12:19:11.733 ServerApp] or http://127.0.0.1:8888/lab?token=d4d5e05447577a0bda0fa8c02c258 | [I 2022-10-04 12:19:11.733 ServerApp] or http://127.0.0.1:8888/lab?token=d4d5e05447577a0bda0fa8c02c258 | [I 2022-10-04 12:19:11.733 ServerApp] | Or http:
                                                                                               ake-pyspark-notebook-1 | [I 2022-10-04 12:19:11.733 ServerApp] Use Control-C to stop this server and shut down all kernels (twi
       04-building-a-data-lake-p
ce to skip confirmation).
      de to skip confirmation).

de-building-a-data-lake-pyspark-notebook-1

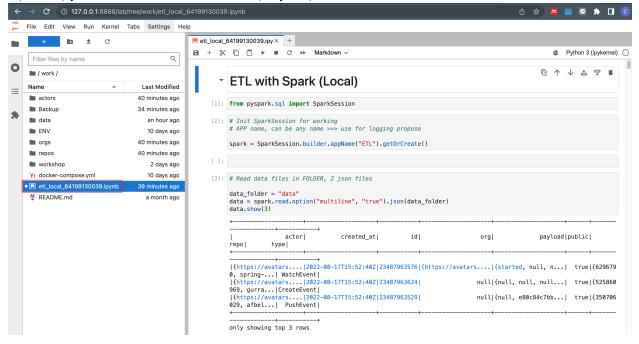
04-building-a-data-lake-pyspark-notebook-1

04-building-a-data-lake-pyspark-notebook-1

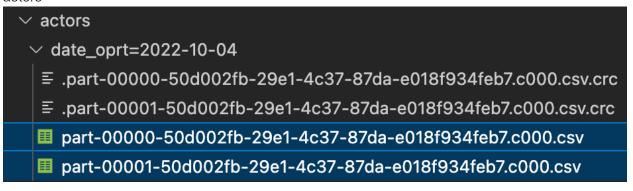
04-building-a-data-lake-pyspark-notebook-1

04-building-a-data-lake-pyspark-notebook-1
                                                                                                                                                                                                             [C 2022-10-04 12:19:11.737 ServerApp]
                                                                                                                                                                                                                             To access the server, open this file in a browser:
    file:///home/jovyan/.local/share/jupyter/runtime/jpserver-7-open.html
Or copy and paste one of these URLs:
    http://8d2b05828758:8888/lab?token=d4d5e05447577a0ba067a8c02c258ce883331ae9ff9e8827
       04-building-a-data-lake-pyspark-notebook-1
04-building-a-data-lake-pyspark-notebook-1
                                                                                                                                                                                                                                  or http://127.0.0.1:8888/lab?token=d4d5e05447577a0bda0fa8c02c258ce983331ae9ff9e8827
```

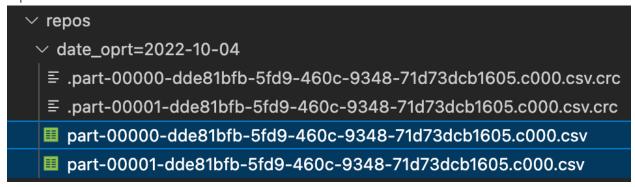
4. Open JupyterNotebook and execute step by step



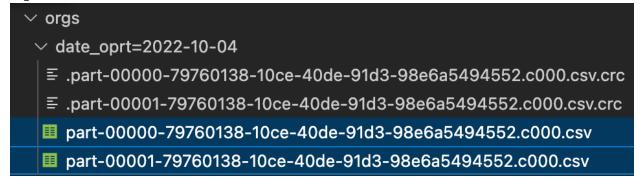
- 5. Check the cleaned output data in folders which partition by 'date_oprt'
 - actors



- repos



- orgs



events

