

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:

```
PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL PORTS
04-building-a-data-lake-pyspark-notebook-1 | [I 2022-11-07 15:26:19.638 ServerApp] jupyterlab | extension was successfully loaded.
04-building-a-data-lake-pyspark-notebook-1 | [I 2022-11-07 15:26:19.642 ServerApp] nbclassic | extension was successfully loaded.
04-building-a-data-lake-pyspark-notebook-1 | [I 2022-11-07 15:26:19.643 ServerApp] Serving notebooks from local directory: /home/jovyan
04-building-a-data-lake-pyspark-notebook-1 | [I 2022-11-07 15:26:19.643 ServerApp] Jupyter Server 1.21.0 is running at:
04-building-a-data-lake-pyspark-notebook-1 | [I 2022-11-07 15:26:19.643 ServerApp] http://c0fa2f98c374:8888/lab?token=cf11487fa0f6bb1de344d6dafb41bc3ca9700f141d1596c9
04-building-a-data-lake-pyspark-notebook-1 | [I 2022-11-07 15:26:19.643 ServerApp] or http://127.0.0.1:8888/lab?token=cf11487fa0f6bb1de344d6dafb41bc3ca9700f141d1596c9
04-building-a-data-lake-pyspark-notebook-1 | [I 2022-11-07 15:26:19.643 ServerApp] Use Control-C to stop this server and shutdown all kernels (twice to skip confirmation).
04-building-a-data-lake-pyspark-notebook-1 | [C 2022-11-07 15:26:19.647 ServerApp]
04-building-a-data-lake-pyspark-notebook-1 |
04-building-a-data-lake-pyspark-notebook-1 | To access the server, open this file in a browser:
04-building-a-data-lake-pyspark-notebook-1 | file:///home/jovyan/.local/share/jupyter/runtime/jpserver-25-open.html
04-building-a-data-lake-pyspark-notebook-1 | Or copy and paste one of these URLs:
04-building-a-data-lake-pyspark-notebook-1 | http://c0fa2f98c374:8888/lab?token=cf11487fa0f6bb1de344d6dafb41bc3ca9700f141d1596c9
04-building-a-data-lake-pyspark-notebook-1 | or http://127.0.0.1:8888/lab?token=cf11487fa0f6bb1de344d6dafb41bc3ca9700f141d1596c9
```

5. Execute the Notebook 'etl_local_64199130061.ipynb' step by step:

```
ETL with Spark (Local)

[6]: from pyspark.sql import SparkSession
# from pyspark.sql.types import StructType, StructField, DoubleType, StringType, IntegerType, DateType, TimestampType
# import pyspark.sql.functions as F

[7]: # Init SparkSession for working
# APP name, can be any name >>> use for logging propose

spark = SparkSession.builder.appName("ETL").getOrCreate()

[14]: data_folder = "data"

[15]: # Read data files in FOLDER, 2 json files

data_folder = "data"
data = spark.read.option("multiline", "true").json(data_folder)
data.show(3)

+-----+-----+-----+-----+-----+-----+
| repo | actor | created_at | id | org | payload|public |
+-----+-----+-----+-----+-----+-----+
| {https://avatars... | [2022-08-17T15:52:40Z | 23487963576 | {https://avatars... | [started, null, n... | true | {6296790, spring-... | WatchEvent |
| {https://avatars... | [2022-08-17T15:52:40Z | 23487963624 | null | null, null, null... | true | {52586096 9, gurra... | CreateEvent |
| {https://avatars... | [2022-08-17T15:52:40Z | 23487963529 | null | null, e80c84c7bb... | true | {35070602 9, afbe1... | PushEvent |
```

6. Check the cleaned output data in folders, partition by 'date_oprt':

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

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

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

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


7. Shutdown environment workspace:

```
$ docker-compose down
```