

Week2 : 02-data-modelling-ii

Data Model

events
id: BIGINT NOT NULL [PK]
type: VARCHAR NOT NULL
actor_id: BIGINT NOT NULL
actor: VARCHAR NOT NULL
public: BOOLEAN NOT NULL
created_at: TIMESTAMP NOT NULL

actors
actor_id: BIGINT NOT NULL [PK]
actor: VARCHAR NOT NULL
number_events: INTEGER NOT NULL

Project Methodology

1. Access project repository on gitpod “gitpod /workspace/SWU-DS525/02-data-modelling-ii”

```
● gitpod /workspace/SWU-DS525 (main) $ cd 02-data-modelling-ii/
```

2. Create virtual environment named “ENV” (only first time)

```
● gitpod /workspace/SWU-DS525/02-data-modelling-ii (main) $ python -m venv ENV
```

3. Activate virtual environment

```
● gitpod /workspace/SWU-DS525/02-data-modelling-ii (main) $ source ENV/bin/activate
```

4. Install required libraries

```
● (ENV) gitpod /workspace/SWU-DS525/02-data-modelling-ii (main) $ pip install -r requirements.txt
Collecting cassandra-driver==3.25.0
  Downloading cassandra_driver-3.25.0-cp38-cp38-manylinux1_x86_64.whl (3.6 MB)
  3.6/3.6 MB 68.5 MB/s eta 0:00:00
Collecting click==8.1.3
  Downloading click-8.1.3-py3-none-any.whl (96 kB)
  96.6/96.6 KB 39.5 MB/s eta 0:00:00
Collecting geomet==0.2.1.post1
  Downloading geomet-0.2.1.post1-py3-none-any.whl (18 kB)
Collecting numpy==1.23.2
  Using cached numpy-1.23.2-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (17.1 MB)
Collecting python-dateutil==2.8.2
  Using cached python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
Collecting pytz==2022.2.1
  Using cached pytz-2022.2.1-py2.py3-none-any.whl (500 kB)
Collecting six==1.16.0
  Using cached six-1.16.0-py2.py3-none-any.whl (11 kB)
```

```
(ENV) gitpod /workspace/SWU-DS525/02-data-modelling-ii (main) $ pip freeze
cassandra-driver==3.25.0
click==8.1.3
geomet==0.2.1.post1
numpy==1.23.2
python-dateutil==2.8.2
pytz==2022.2.1
six==1.16.0
```

5. Start Docker with Cassandra services

```
(ENV) gitpod /workspace/SWU-DS525/02-data-modelling-ii (main) $ docker-compose up
[+] Running 10/10
  :: cassandra Pulled                                8.9s
  :: 08c01a0ec47e Pull complete                      1.9s
  :: cd22f8f9007b Pull complete                      4.1s
  :: 1a67ef3809c8 Pull complete                      6.0s
  :: 86878d46e3d7 Pull complete                      6.0s
  :: 80d5daecc561 Pull complete                      6.1s
  :: 21d84a377e44 Pull complete                      6.2s
  :: b91dabf4c4b8 Pull complete                      6.3s
  :: bba635112678 Pull complete                      7.4s
  :: eb0cf34d7562 Pull complete                      7.4s
[+] Running 3/3
  :: Network 02-data-modelling-ii_default             Created                                0.0s
  :: Volume "02-data-modelling-ii_cassandra-data-volume" Created                0.0s
  :: Container 02-data-modelling-ii-cassandra-1       Created                                0.0s
Attaching to 02-data-modelling-ii-cassandra-1
02-data-modelling-ii-cassandra-1 | OpenJDK 64-Bit Server VM warning: Option UseConcMarkSweepGC was deprecated in version 9.0 and will likely be removed in a future release.
```

6. Create tables, insert data, and query data through python command by file “etl.py”

Python etl.py

```
(ENV) gitpod /workspace/SWU-DS525/02-data-modelling-ii (main) $ python etl.py
5 files found in ../data
query PushEvent
Row(actor_id='106313266', created_at=datetime.datetime(2022, 8, 17, 15, 55, 5), actor='kenobiwins', id='23488014712', public=True, type='PushEvent')
Row(actor_id='94136524', created_at=datetime.datetime(2022, 8, 17, 15, 52, 40), actor='Sidalvik', id='23487963448', public=True, type='PushEvent')
Row(actor_id='106591964', created_at=datetime.datetime(2022, 8, 17, 15, 54, 45), actor='kunede', id='23488007824', public=True, type='PushEvent')
Row(actor_id='37953029', created_at=datetime.datetime(2022, 8, 17, 15, 52, 40), actor='BadProfessor', id='23487963492', public=True, type='PushEvent')
Row(actor_id='47776594', created_at=datetime.datetime(2022, 8, 17, 15, 55, 5), actor='baileys-li', id='23488014647', public=True, type='PushEvent')
Row(actor_id='46447321', created_at=datetime.datetime(2022, 8, 17, 15, 51, 5), actor='allcontributors[bot]', id='23487929588', public=True, type='PushEvent')
Row(actor_id='5146167', created_at=datetime.datetime(2022, 8, 17, 15, 51, 5), actor='markbush', id='23487929536', public=True, type='PushEvent')

.

.

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Row(actor_id='100631798', created_at=datetime.datetime(2022, 8, 17, 15, 55, 5), actor='simonkernel', id='23488014775', public=True, type='PushEvent')
Row(actor_id='74971347', created_at=datetime.datetime(2022, 8, 17, 15, 51, 5), actor='MathisGD', id='23487929674', public=True, type='PushEvent')
Row(actor_id='97834402', created_at=datetime.datetime(2022, 8, 17, 15, 54, 45), actor='Master-Stroke', id='23488007810', public=True, type='PushEvent')
Row(actor_id='17869732', created_at=datetime.datetime(2022, 8, 17, 15, 51, 5), actor='moiify', id='23487929591', public=True, type='PushEvent')
Row(actor_id='66924041', created_at=datetime.datetime(2022, 8, 17, 15, 51, 5), actor='yousabu', id='23487929676', public=True, type='PushEvent')
Row(actor_id='18267059', created_at=datetime.datetime(2022, 8, 17, 15, 55, 5), actor='ewquon', id='23488014738', public=True, type='PushEvent')
Row(actor_id='90624848', created_at=datetime.datetime(2022, 8, 17, 15, 51, 4), actor='ofzlo', id='23487929349', public=True, type='PushEvent')
Row(actor_id='91143053', created_at=datetime.datetime(2022, 8, 17, 15, 51, 5), actor='BR-Junior', id='23487929661', public=True, type='PushEvent')
Row(actor_id='17073422', created_at=datetime.datetime(2022, 8, 17, 15, 54, 45), actor='alexpassss', id='23488007745', public=True, type='PushEvent')
query number of events by each actor if number of event more than 1
(ENV) gitpod /workspace/SWU-DS525/02-data-modelling-ii (main) $
```

7. Environment Deactivation by; Docker-compose down and deactivate