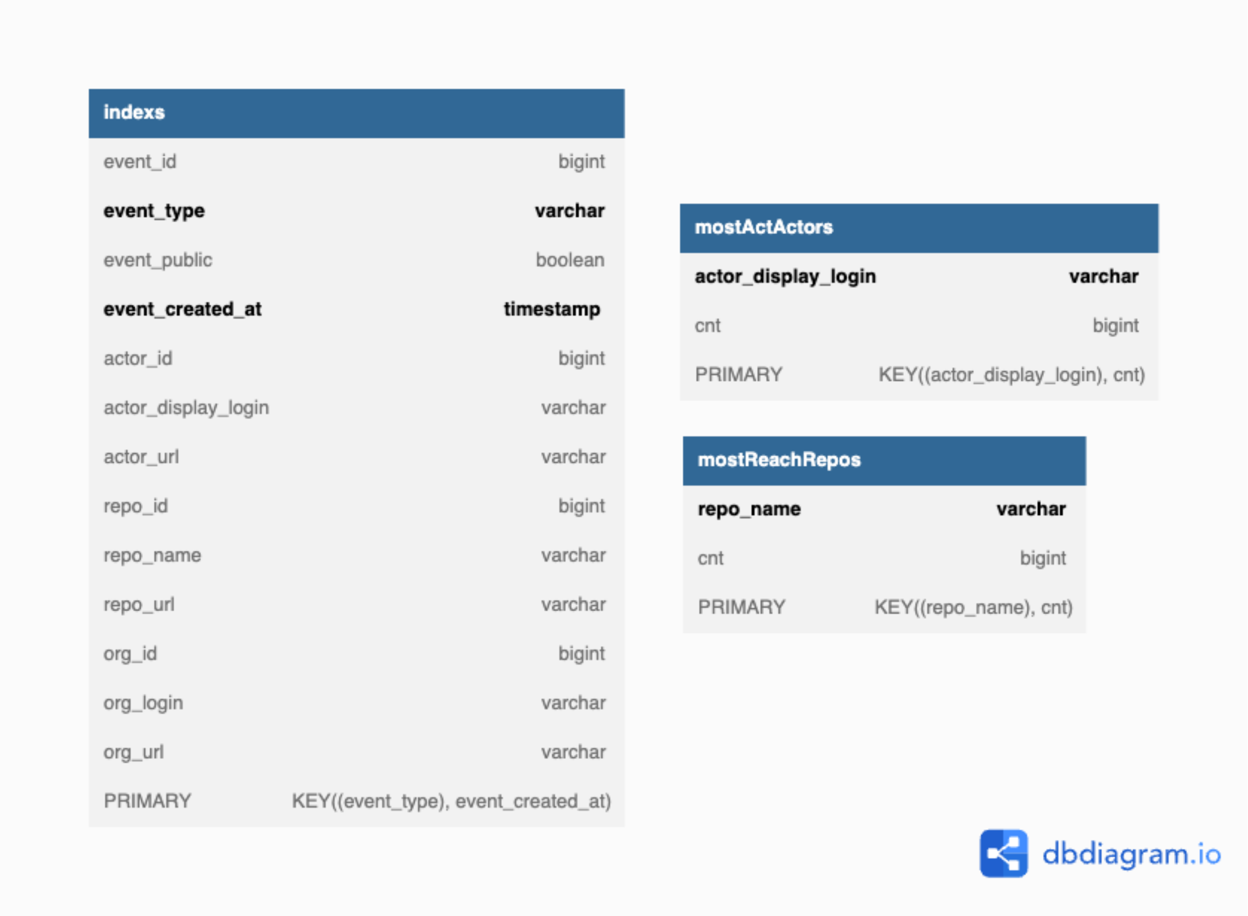
Week 2 – Project Data model ii

Data model



indexs	
event_id	bigint
event_type	**varchar**
event_public	boolean
event_created_at	**timestamp**
actor_id	bigint
actor_display_login	varchar
actor_url	varchar
repo_id	bigint
repo_name	varchar
repo_url	varchar
org_id	bigint
org_login	varchar
org_url	varchar
PRIMARY	KEY((event_type), event_created_at)

mostActActors	
actor_display_login	**varchar**
cnt	bigint
PRIMARY	KEY((actor_display_login), cnt)

mostReachRepos	
repo_name	**varchar**
cnt	bigint
PRIMARY	KEY((repo_name), cnt)

DS525 - Chin Lertvipada - 64199130039

Tables

indexs:

This table can serve the data for a main page of logging administration. It includes the data about events, repos, actors and orgs.

The partition is ‘event_type’, so the data are separately kept by type of event. This case, we can manipulate the data based on the type of event.

The clustering column is ‘event_create_at’, so the data will be sorted based on the event creation date-time.

mostActActors:

This table allows us to see the frequency of event creation by each user. So, we can see the participation of each user and we can find the top contributory user.

mostReachRepo:

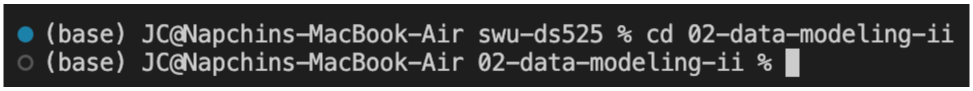
This table allows us to see the frequency of reaching the repository for all event type. So, we can see the activity on each repository and we can find the top engaged repository.

DS525 - Chin Lertvipada - 64199130039

Project implementation instruction

1. Reach the project repository ‘swu-ds525/02-data-modeling-ii’

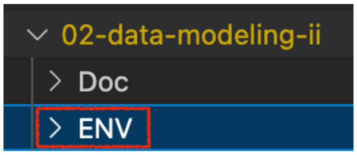
   : *$ cd 02-data-modeling-ii*



```
● (base) JC@Napchins-MacBook-Air swu-ds525 % cd 02-data-modeling-ii
○ (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % █
```

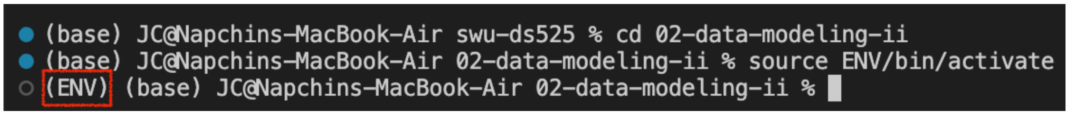
2. Create visual environment for the project’s resources, named ‘ENV’ (only 1st time)

   : *$ python -m venv ENV*



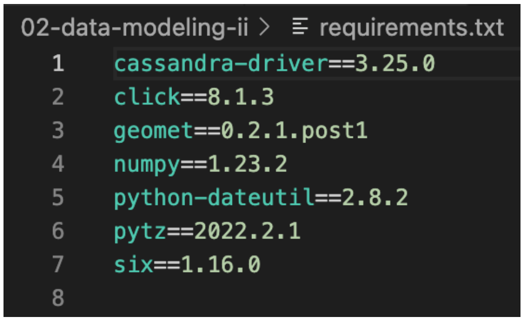
3. Activate visual environment ‘ENV’ to be used

   : *$ source ENV/bin/activate*



```
● (base) JC@Napchins-MacBook-Air swu-ds525 % cd 02-data-modeling-ii
● (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % source ENV/bin/activate
○ (ENV) (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % █
```

4. Install mandatory libraries from configuration file, named ‘requirements.txt’ (only 1st time)



02-data-modeling-ii > requirements.txt

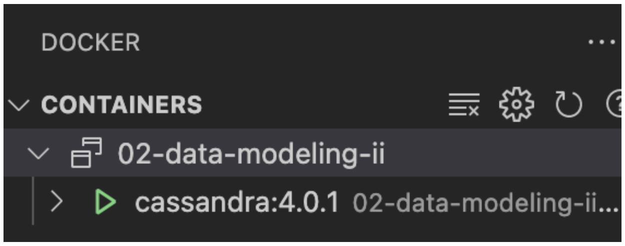
```
cassandra-driver==3.25.0
click==8.1.3
geomet==0.2.1.post1
numpy==1.23.2
python-dateutil==2.8.2
pytz==2022.2.1
six==1.16.0

```

   : *$ pip install -r requirements.txt*

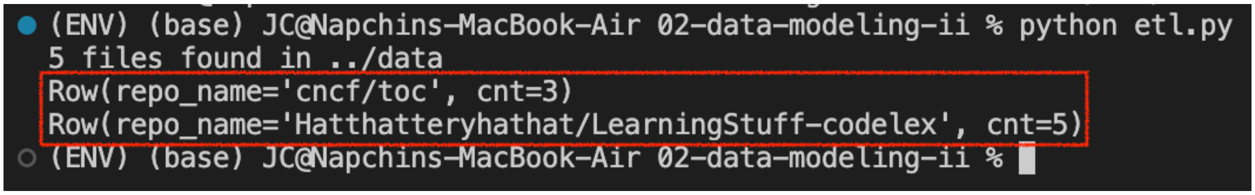
5. Start Docker with services ‘Cassandra’ from YMAL file, named ‘docker-compose.yml’

   : *$ docker-compose up*



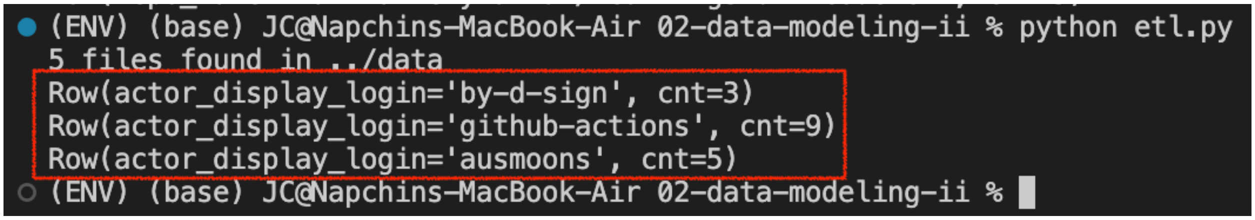
6. Create tables, Insert data, Query data thru python script, named ‘etl.py’

   : *$ python etl.py*



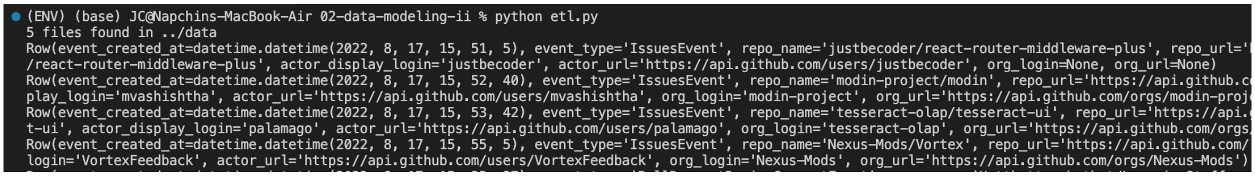
```
● (ENV) (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % python etl.py
  5 files found in ../data
  Row(repo_name='cncf/toc', cnt=3)
  Row(repo_name='Hatthatteryhathat/LearningStuff-codelex', cnt=5)
○ (ENV) (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % █
```

** the query to show the repository with engagement is more than 2 events



```
● (ENV) (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % python etl.py
  5 files found in ../data
  Row(actor_display_login='by-d-sign', cnt=3)
  Row(actor_display_login='github-actions', cnt=9)
  Row(actor_display_login='ausmoons', cnt=5)
○ (ENV) (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % █
```

** the query to show the actor with participation is more than 2 events



```
● (ENV) (base) JC@Napchins-MacBook-Air 02-data-modeling-ii % python etl.py
  5 files found in ../data
  Row(event_created_at=datetime.datetime(2022, 8, 17, 15, 51, 5), event_type='IssuesEvent', repo_name='justbecoder/react-router-middleware-plus', repo_url='
  /react-router-middleware-plus', actor_display_login='justbecoder', actor_url='https://api.github.com/users/justbecoder', org_login=None, org_url=None)
  Row(event_created_at=datetime.datetime(2022, 8, 17, 15, 52, 40), event_type='IssuesEvent', repo_name='modin-project/modin', repo_url='https://api.github.c
  play_login='mvashishtha', actor_url='https://api.github.com/users/mvashishtha', org_login='modin-project', org_url='https://api.github.com/orgs/modin-proj
  Row(event_created_at=datetime.datetime(2022, 8, 17, 15, 53, 42), event_type='IssuesEvent', repo_name='tesseract-olap/tesseract-ui', repo_url='https://api.
  t-ui', actor_display_login='palamago', actor_url='https://api.github.com/users/palamago', org_login='tesseract-olap', org_url='https://api.github.com/orgs
  Row(event_created_at=datetime.datetime(2022, 8, 17, 15, 55, 5), event_type='IssuesEvent', repo_name='Nexus-Mods/Vortex', repo_url='https://api.github.com/
  login='VortexFeedback', actor_url='https://api.github.com/users/VortexFeedback', org_login='Nexus-Mods', org_url='https://api.github.com/orgs/Nexus-Mods')
```

** the query to show data for main page of logging administration

7. [optional] Shutdown the environment

   a. Stop ‘Cassandra’ by shutdown Docker

      : *$ docker-compose down*

   b. Deactivate the visual environment ‘ENV’

      : *$ deactivate*