

การสร้างแบบจำลองข้อมูลด้วย Cassandra (NoSQL)

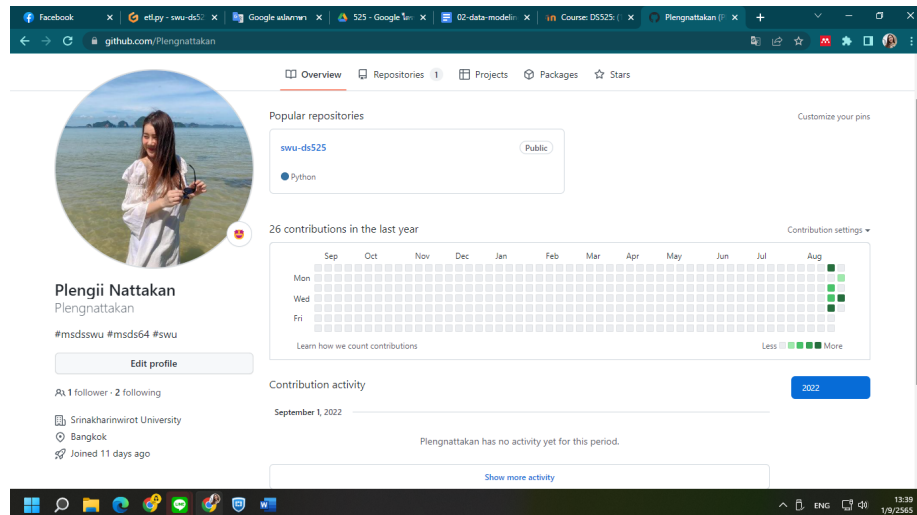
วิธีการดำเนินการ

1. วิเคราะห์โครงสร้างข้อมูลของ JSON ไฟล์ ได้ดังรูป

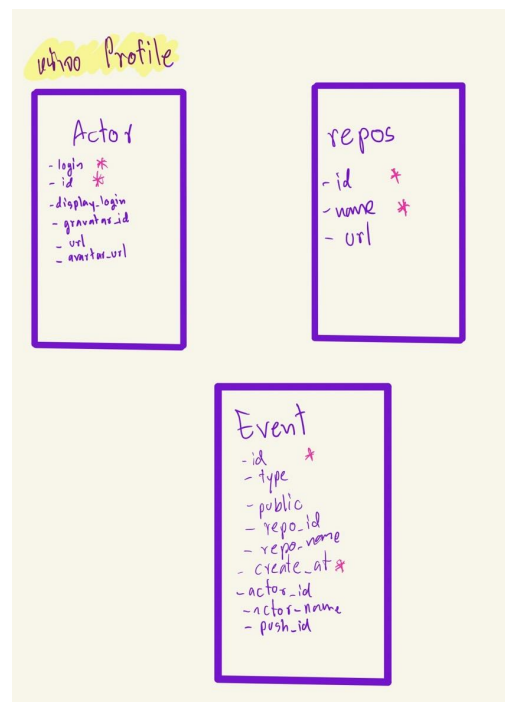


จากภาพ ข้อมูลใน Json เป็นข้อมูล เหตุการณ์ หรือ Event ของ Github โดยมีโครงสร้างที่ยืดหยุ่น โดยแต่ละ record จะมีข้อมูลที่ไม่เท่ากัน และไม่เหมือนกัน

2. เลือกหน้าจอกที่ต้องการแสดงผลข้อมูล โดยได้เลือกหน้าจอ Profile ของผู้ใช้งาน Github ตามหน้าจอดังนี้



3. ดำเนินการออกแบบตาราง ตามหน้าจอที่ได้เลือก ได้ตารางดังนี้ (โดยทุกตารางไม่มีการเชื่อมโยงกัน และทุกตารางมีการกำหนด Partition Key เพื่อการจัดกลุ่มข้อมูล และ Clustering Key เพื่อการจัดเรียงข้อมูลที่อยู่ใกล้เคียงกัน รวมเป็น Primary Key เพื่อให้การ Query เข้าถึงได้อย่างรวดเร็ว)

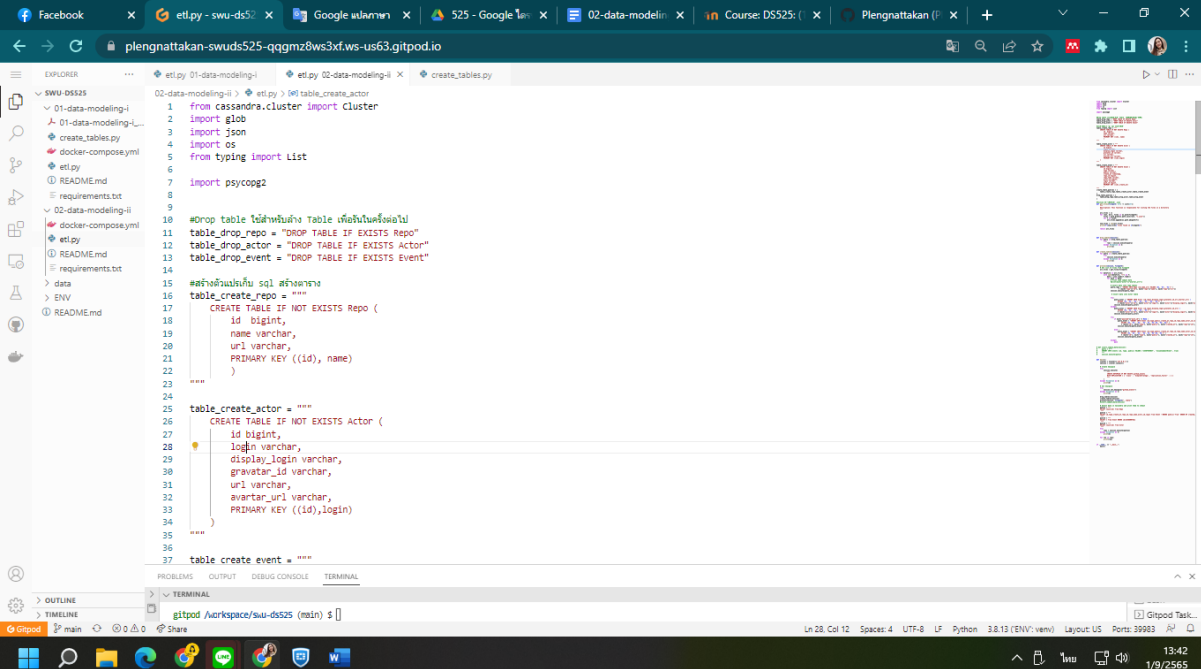


โดย ตาราง event คือ ตารางที่เก็บข้อมูลการดำเนินการหรือ Activity ต่างๆ ของผู้ใช้งาน เช่น มีการ push สิ่งใด จำนวนกี่ครั้ง โดย Partition Key คือ type เนื่องจาก Type เป็นลักษณะของ

หมวดหมู่อยู่แล้ว และ Clustering Key คือ Create_at เป็น
ลักษณะของวันและเวลา เหมาะสำหรับการนำมา sort ข้อมูล
ตาราง repo คือ ตารางที่เก็บข้อมูลในส่วนของพื้นที่ workspace
ของแต่ละคน โดย Partition Key คือ id เนื่องจากตารางนี้เป็น
ตารางเก็บข้อมูลพื้นที่การทำงานของแต่ละ actor 1 actor
สามารถมีได้หลาย repo 1 repo สามารถแชร์การทำงานร่วมกันได้
และ Clustering Key คือ name จากในตารางไม่มีข้อมูล
Datetime จึงใช้ name ควบคู่ไปกับ id

ตาราง actor คือ ตารางที่เก็บข้อมูลเจ้าของ Profile โดย
Partition Key คือ id เนื่องจากตารางนี้เป็นตารางเก็บข้อมูลต่างๆ
ของแต่ละ actor 1 และ Clustering Key คือ login จากในตาราง
ไม่มีข้อมูล Datetime จึงใช้ Login ควบคู่ไปกับ id

4. ทำการเตรียม Gitpod สำหรับการเขียนโปรแกรม โดยทำการ
Sync Github account และ Gitpod เข้าด้วยกัน ติดตั้ง
Postgresql เชื่อมกับ Gitpod เพื่อสามารถสร้างตาราง SQL บน
Gitpod ได้ และทำการเตรียม environment
5. แก้ไขไฟล์ etl.py เพื่อทำการสร้างตารางและดึงข้อมูลจากลิงค์
json เข้าไปยังตารางที่สร้างไว้

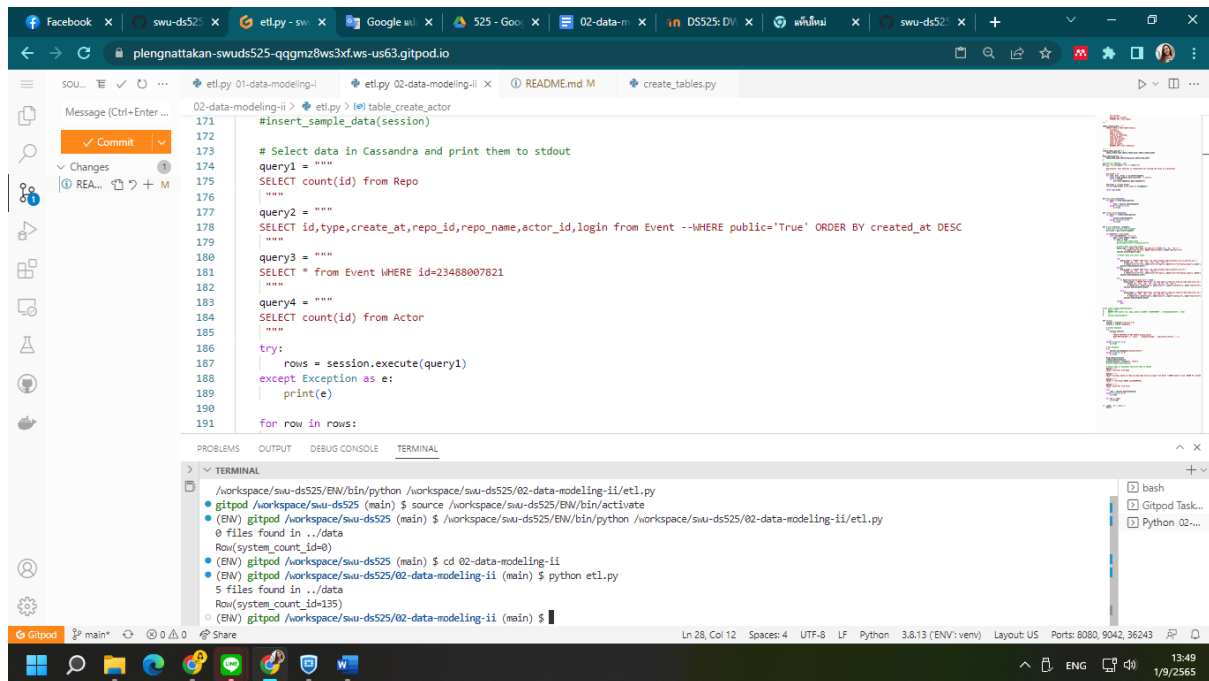


The screenshot shows a Gitpod workspace environment. The main editor displays a Python script named `etl.py` with the following content:

```
1 from cassandra.cluster import Cluster
2 import glob
3 import json
4 import os
5 from typing import List
6
7 import psycopg2
8
9 #Drop table ถ้าพบว่ามี table เหมือนในไฟล์แล้ว
10 table_drop_repo = "DROP TABLE IF EXISTS Repo"
11 table_drop_actor = "DROP TABLE IF EXISTS Actor"
12 table_drop_event = "DROP TABLE IF EXISTS Event"
13
14 #สร้างและเก็บ sql สร้างจาก
15 table_create_repo = """
16 CREATE TABLE IF NOT EXISTS Repo (
17     id bigint,
18     name varchar,
19     url varchar,
20     PRIMARY KEY ((id), name)
21 )
22 """
23
24
25 table_create_actor = """
26 CREATE TABLE IF NOT EXISTS Actor (
27     id bigint,
28     login varchar,
29     display_login varchar,
30     gravitar_id varchar,
31     url varchar,
32     avantar_url varchar,
33     PRIMARY KEY ((id), login)
34 )
35 """
36
37 table_create_event = """
```

The terminal at the bottom shows the command `gitpod /workspace/swu-ds525 (main) $` and the system status bar at the bottom right indicates the time is 13:42 on 1/9/2565.

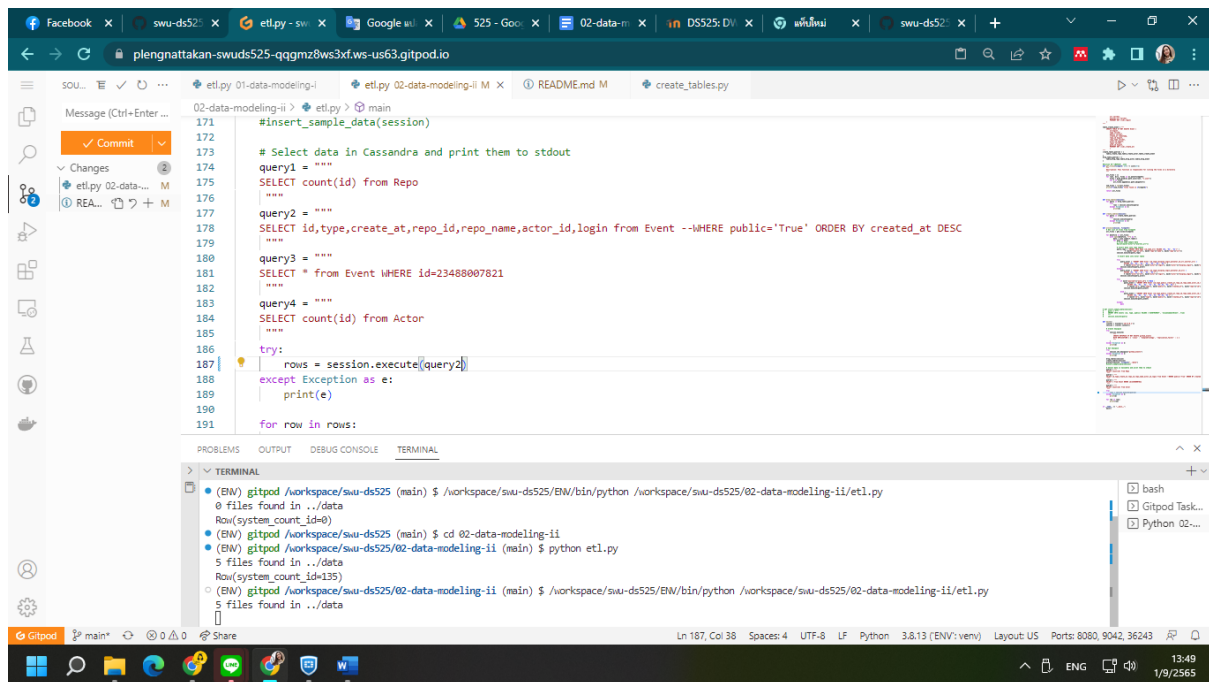
6. ทดสอบโปรแกรม



```
171 #insert_sample_data(session)
172
173 # Select data in Cassandra and print them to stdout
174 query1 = """
175 SELECT count(id) from Repo
176 """
177
178 query2 = """
179 SELECT id,type,create_at,repo_id,repo_name,actor_id,login from Event --WHERE public='True' ORDER BY created_at DESC
180 """
181
182 query3 = """
183 SELECT * from Event WHERE id=23488007821
184 """
185
186 query4 = """
187 SELECT count(id) from Actor
188 """
189
190 try:
191     rows = session.execute(query1)
192 except Exception as e:
193     print(e)
194
195 for row in rows:
```

TERMINAL

```
/workspace/swu-ds525/BW/bin/python /workspace/swu-ds525/02-data-modeling-ii/etl.py
gitpod /workspace/swu-ds525 (main) $ source /workspace/swu-ds525/BW/bin/activate
(BW) gitpod /workspace/swu-ds525 (main) $ /workspace/swu-ds525/BW/bin/python /workspace/swu-ds525/02-data-modeling-ii/etl.py
0 files found in ../data
Row(system_count_id=0)
(BW) gitpod /workspace/swu-ds525 (main) $ cd 02-data-modeling-ii
(BW) gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ python etl.py
5 files found in ../data
Row(system_count_id=135)
(BW) gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $
```



```
171 #insert_sample_data(session)
172
173 # Select data in Cassandra and print them to stdout
174 query1 = """
175 SELECT count(id) from Repo
176 """
177
178 query2 = """
179 SELECT id,type,create_at,repo_id,repo_name,actor_id,login from Event --WHERE public='True' ORDER BY created_at DESC
180 """
181
182 query3 = """
183 SELECT * from Event WHERE id=23488007821
184 """
185
186 query4 = """
187 SELECT count(id) from Actor
188 """
189
190 try:
191     rows = session.execute(query2)
192 except Exception as e:
193     print(e)
194
195 for row in rows:
```

TERMINAL

```
(BW) gitpod /workspace/swu-ds525 (main) $ /workspace/swu-ds525/BW/bin/python /workspace/swu-ds525/02-data-modeling-ii/etl.py
0 files found in ../data
Row(system_count_id=0)
(BW) gitpod /workspace/swu-ds525 (main) $ cd 02-data-modeling-ii
(BW) gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ python etl.py
5 files found in ../data
Row(system_count_id=135)
(BW) gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ /workspace/swu-ds525/BW/bin/python /workspace/swu-ds525/02-data-modeling-ii/etl.py
5 files found in ../data
```