Project: Building a Data Modeling with Cassandra (NoSQL)

สามารถเริ่มต้นจากโค้ดที่ https://github.com/zkan/swu-ds525 ได้

ในโปรเจคนี้เราจะใช้ข้อมูล GitHub event data จาก API หรือสามารถดาวน์โหลดไฟล์ JSON ได้ที่ URL นี้

สิ่งที่คาดหวังในโปรเจคนี้

1. มีโค้ดที่ทำ ETL จากข้อมูล JSON โหลดเข้าไปใน Cassandra
2. มีการเขียน documentation อธิบายสิ่งที่ตัวเองทำลงไป รวมไปถึงการออกแบบ data model
3. อธิบายการออกแบบ primary key และ clustering columns
4. มี instruction ในการรันโค้ดของตัวเอง

หมายเหตุ โปรเจคนี้จะคล้าย ๆ กับโปรเจคที่ 1 แต่จะเปลี่ยนจาก relational database ไปเป็น NoSQL database อย่าง Apache Cassandra แทน

1. File JSON ที่จะเป็นข้อมูลเข้า DB


```
[
    {
        "id":"23487929637",
        "type":"IssueCommentEvent",
        "actor":{
            "id":1696078,
            "login":"sukhada",
            "display_login":"sukhada",
            "gravatar_id":"",
            "url":"https://api.github.com/users/sukhada",
            "avatar_url":"https://avatars.githubusercontent.com/u/1696078?"
        },
        "repo":{…},
        "payload":{…},
        "public":true,
        "created_at":"2022-08-17T15:51:05Z",
        "org":{
            "id":24305026,
            "login":"350org",
            "gravatar_id":"",
            "url":"https://api.github.com/orgs/350org",
            "avatar_url":"https://avatars.githubusercontent.com/u/24305026?"
        }
    },
    {
        "id":"23487929676",
        "type":"PushEvent",
        "actor":{
            "id":66924041,
            "login":"yousabu",
            "display_login":"yousabu",
            "gravatar_id":"",
            "url":"https://api.github.com/users/yousabu",
            "avatar_url":"https://avatars.githubusercontent.com/u/66924041?"
        },
        "repo":{…},
        "payload":{…},
        "public":true,
        "created_at":"2022-08-17T15:51:05Z"
    },
```

2. มีการเขียน documentation อธิบายสิ่งที่ตัวเองทำลงไป รวมไปถึงการออกแบบ data model

คำสั่งในการ รัน

docker
gitpod /workspace/swu-ds525 (main) $ cd 02-data-modeling-ii/
gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ source ENV/bin/activate
(ENV) gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ docker-compose up

bash
gitpod /workspace/swu-ds525 (main) $ cd 02-data-modeling-ii/
gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ source ENV/bin/activate
(ENV) gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ pip install -r requirements.txt
(ENV) gitpod /workspace/swu-ds525/02-data-modeling-ii (main) $ python etl.py

ตาราง event

- id คือ partition key
- type คือ clustering column

Code ส่วนที่เป็นการสร้างตาราง

```
table_create = """
  CREATE TABLE IF NOT EXISTS events
  (
     id text,
     type text,
     public boolean,
     created_at text,
     PRIMARY KEY (
        id,
        type
     )
        )
"""
```

```python
# 2. เพิ่ม ค่าใน created_at เข้ามา "2022-08-17T15:51:05Z"

def insert_sample_data(session):
    query = f"""
    INSERT INTO events (id, type, public, created_at) VALUES ('23487929661', 'PushEvent', true, '2022-08-17T15:51:05Z' )
    """
    session.execute(query)


#5. Read JSON แล้วprintout ข้อมูลที่เซตที่ 3มาดู
def process(session, filepath):
    # Get list of files from filepath
    all_files = get_files(filepath)

    for datafile in all_files:
        with open(datafile, "r") as f:
            #load file มาอ่าน
            data = json.loads(f.read())
            for each in data:
                # Print some sample data
                #5.จากเดิม print(each["id"], each["type"], each["actor"]["login"])
                #  ล้อตาม2.ที่ insert_sample_data เปลี่ยนเป็น
                # Print มาดู ยังไม่ได้ใส่ตาราง
                print(each["id"], each["type"], each["public"], each["created_at"])

                # Insert data into tables here
                #6. ใส่ข้อมูลในตาราง
                query = f"""
                INSERT INTO events (id, type, public, created_at) VALUES ('{each["id"]}', '{each["type"]}', {each["public"]}, '{each["created_at"]}' )
                """
                session.execute(query)
```

Output ที่แสดงผล

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
23488014740 PushEvent True 2022-08-17T15:55:05Z
23488014725 IssueCommentEvent True 2022-08-17T15:55:05Z
23488014708 IssuesEvent True 2022-08-17T15:55:05Z
23488014752 PullRequestReviewEvent True 2022-08-17T15:55:05Z
23488014635 PushEvent True 2022-08-17T15:55:05Z
23488014633 PushEvent True 2022-08-17T15:55:05Z
23488014680 PullRequestReviewCommentEvent True 2022-08-17T15:54:51Z
23488014638 PullRequestReviewEvent True 2022-08-17T15:55:05Z
23488014647 PushEvent True 2022-08-17T15:55:05Z
23488014662 PushEvent True 2022-08-17T15:55:05Z
23488014629 PushEvent True 2022-08-17T15:55:05Z
Row(id='23487929661', type='PushEvent', created_at='2022-08-17T15:51:05Z', public=True)
```