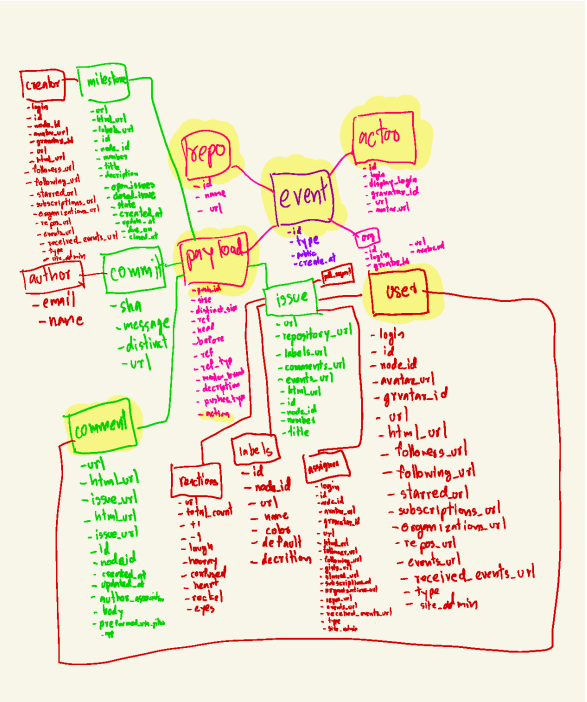
Building an ETL Pipeline to Transform Data in Data Lake (From Landing Zone to Cleaned Zone)

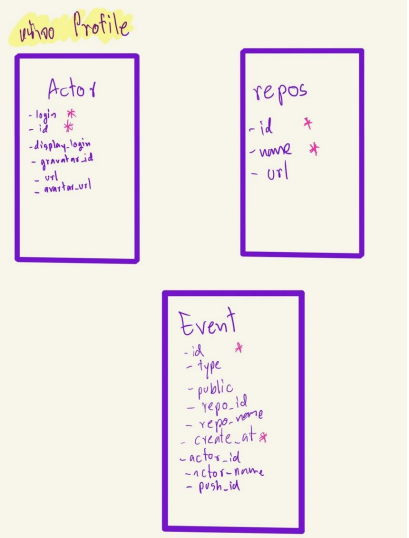
วิธีการดำเนินการ

1. วิเคราะห์โครงสร้างข้อมูลของ JSON ไฟล์ ได้ดังรูป



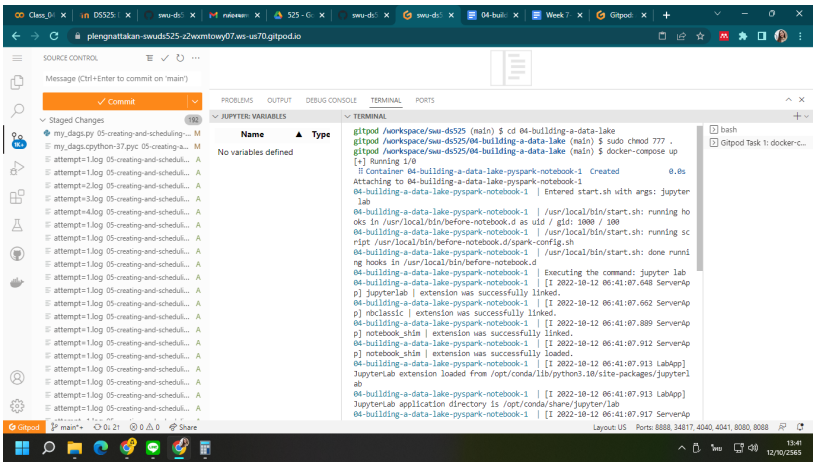
จากภาพ ข้อมูลใน Json เป็นข้อมูล เหตุการณ์ หรือ Event ของ Github โดยมีโครงสร้างที่ยืดหยุ่น โดยแต่ละ record จะมีข้อมูลที่ไม่เท่ากัน และไม่เหมือนกัน

2. ดำเนินการออกแบบตาราง ได้ตารางดังนี้ (โดยทุกตารางไม่มีการเชื่อมโยงกัน และทุกตารางมีการกำหนด Partition Key เพื่อการจัดกลุ่มข้อมูล และ Clustering Key เพื่อการจัดเรียงข้อมูลที่อยู่ใกล้เคียงกัน รวมเป็น Primary Key เพื่อทำให้การ Query เข้าถึงได้อย่างรวดเร็ว)

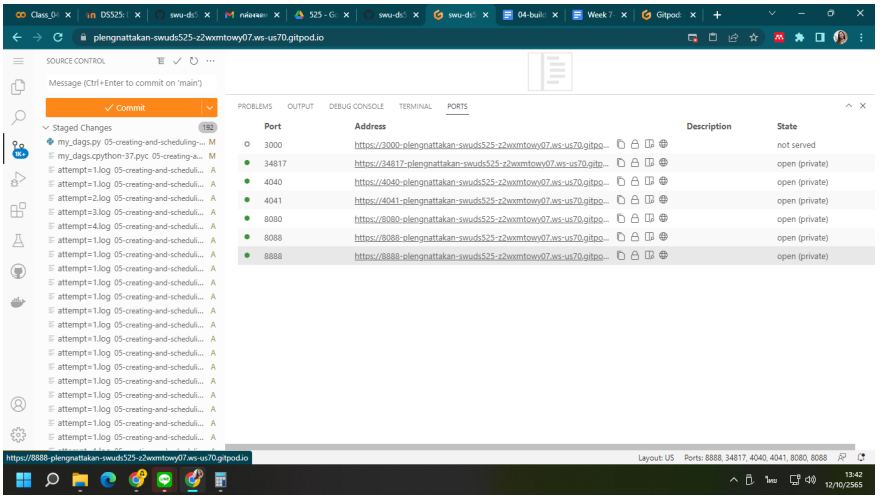


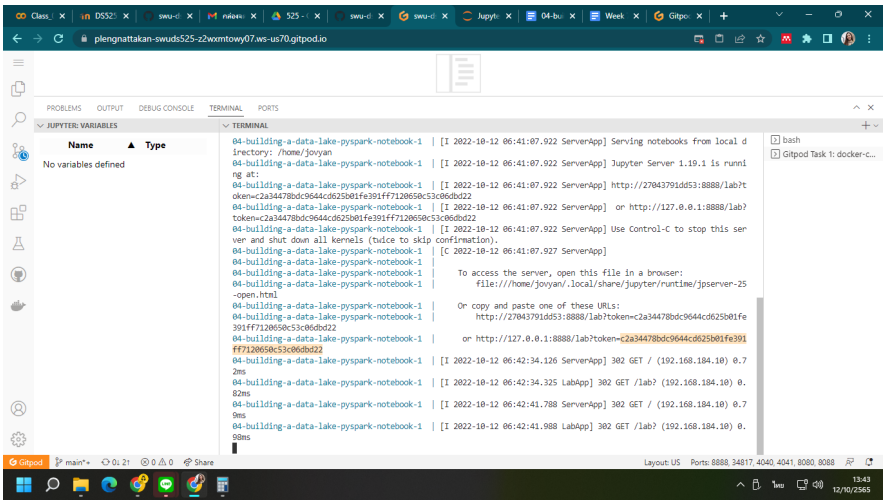
โดย **ตาราง event** คือ ตารางที่เก็บข้อมูลการดำเนินการหรือ Activity ต่างๆ ของผู้ใช้งาน เช่น มีการ push สิ่งใด จำนวนกี่ครั้ง โดย Partition Key คือ type เนื่องจาก Type เป็นลักษณะของหมวดหมู่อยู่แล้ว และ Clustering Key คือ Create_at เป็นลักษณะของวันและเวลา เหมาะสำหรับการนำมา sort ข้อมูล
ตาราง repo คือ ตารางที่เก็บข้อมูลในส่วนของพื้นที่ workspace ของแต่ละคน โดย Partition Key คือ id เนื่องจากตารางนี้เป็นตารางเก็บข้อมูลพื้นที่การทำงานของแต่ละ 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 จากในตารางไม่มีข้อมูล Datetime จึงใช้ Login ควบคู่ไปกับ id

3. ทำการ เข้าไปที่ Folder 04 ด้วยคำสั่ง cd 04-building-a-data-lake และกำหนดให้สามารถเขียนไฟล์ได้ด้วยคำสั่ง sudo chmod 777 . หลังจากนั้นใช้คำสั่ง docker-compose up เพื่อรัน Docker

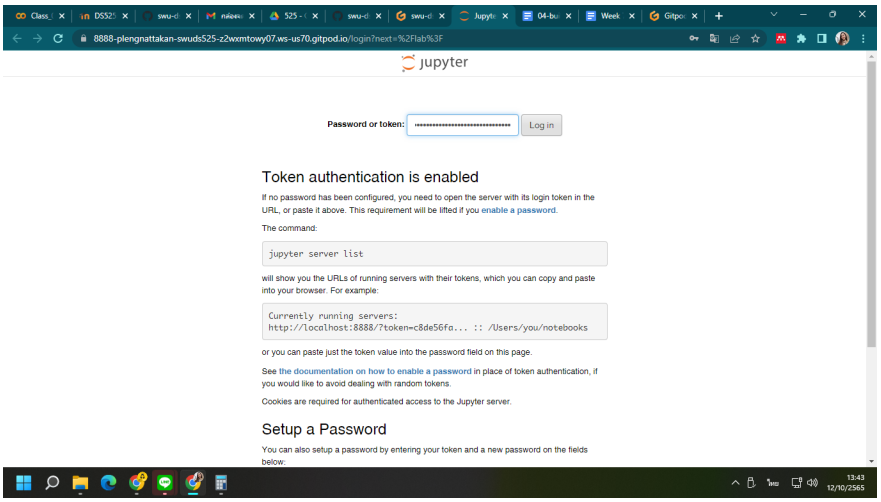


4. เข้าไปที่ tab port เลือก port 8888 และนำ token ไปกรอกในช่อง





c2a34478bdc9644cd625b01fe391ff7120650c53c06dbd22



5. จากนั้นเข้าไปที่ไฟล์ etl_local_0043.ipynb เพื่อทำการสร้างตารางตามที่ออกแบบโดยใช้คำสั่ง spark และ sql