

Project3 Final

Project: Building an ETL Pipeline for a Cloud Data Warehouse (Redshift/Google BigQuery)

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

ในโปรเจกต์นี้เราจะใช้ข้อมูล GitHub event data จาก [API](#) หรือสามารถดาวน์โหลดไฟล์ JSON ได้ที่ [URL](#) นี้
สิ่งที่คาดหวังในโปรเจกต์นี้

ในโปรเจกต์นี้เราจะตั้งต้นว่าเรามีไฟล์อยู่บน S3 อยู่แล้ว ดังนั้นทุกคนสามารถที่จะอัปโหลดไฟล์ขึ้นไปไว้บน S3 ได้เลย

1. มีโค้ดที่ทำ ETL จากข้อมูล JSON โหลดเข้าไปใน Redshift ตาม data model ที่เราออกแบบไว้
2. มีการเขียน documentation อธิบายสิ่งที่ตัวเองทำลงไป รวมไปถึงการออกแบบ data model
3. มี instruction ในการรันโค้ดของตัวเอง

อ่านวิธีโหลด JSON เข้า Redshift ได้ที่ [COPY from JSON format](#)

Last modified: Saturday, 3 September 2022, 8:05 AM

1.สร้าง Redshift ใน aws

2.สร้าง S3 เก็บ JSON

3.สร้างตาราง และ import file จาก s3 มา redshift ทดสอบflow

4. เขียนpython ใน gitpod เพื่อ สร้างไฟล์ python สำหรับการ สร้างตาราง และ import file จาก s3 มา redshift โดย script