Capstone Project

Employee Turnover

Problem

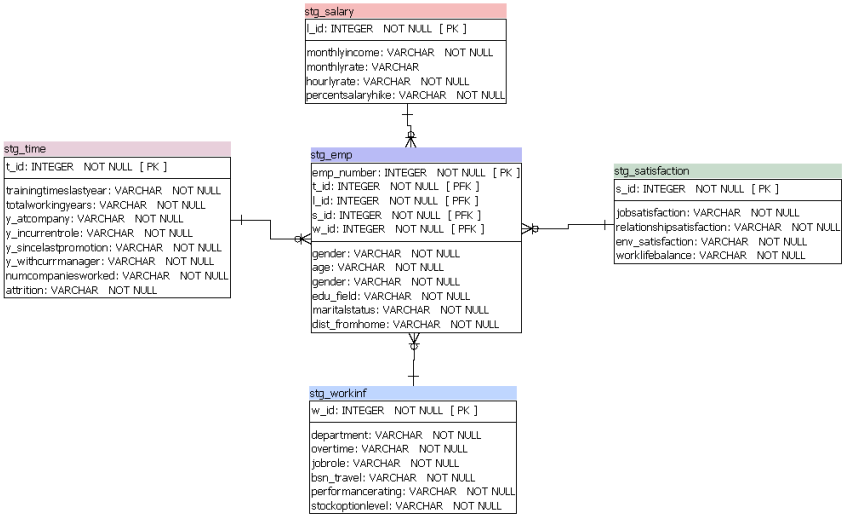
เนื่องจากพนักงานในองค์กรมีจำนวนของการลาออกเพิ่มมากขึ้น ทำให้พนักงานในแต่ละตำแหน่งมีจำนวนไม่เพียงพอกับงานในองค์กร ส่งผลให้งานที่ออกมาคุณภาพลดลง และการส่งงานมีความล่าช้ามากขึ้น ไม่เป็นไปตามเวลาที่ได้ตกลงกันไว้ และอาจส่งผลให้ธุรกิจหยุดชะงักได้ หากมีพนักงานไม่เพียงพอต่อความต้องการของงาน โดยการทำโปรเจคนี้เพื่อการหาสาเหตุหลักที่ทำให้พนักงานลาออกจากองค์กร และเสนอแนวทางที่จะลดการลาออกของพนักงานภายในองค์กรได้

Dataset

แบบสำรวจของพนักงานภายในองค์กร ที่ได้มาจากการสัมภาษณ์พนักงาน จำนวน 1,470 คน

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?resource=download

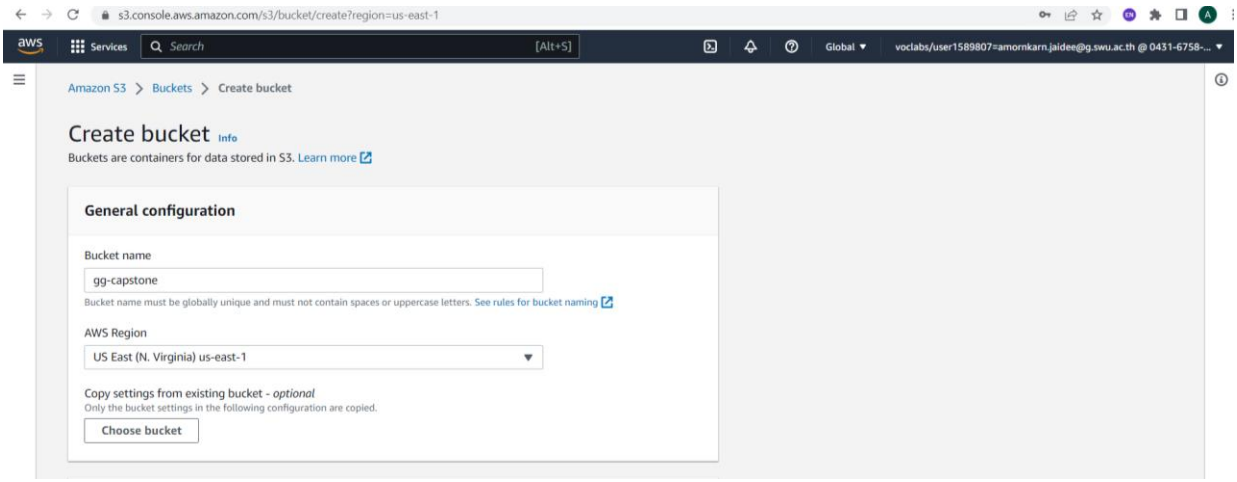
Data Model



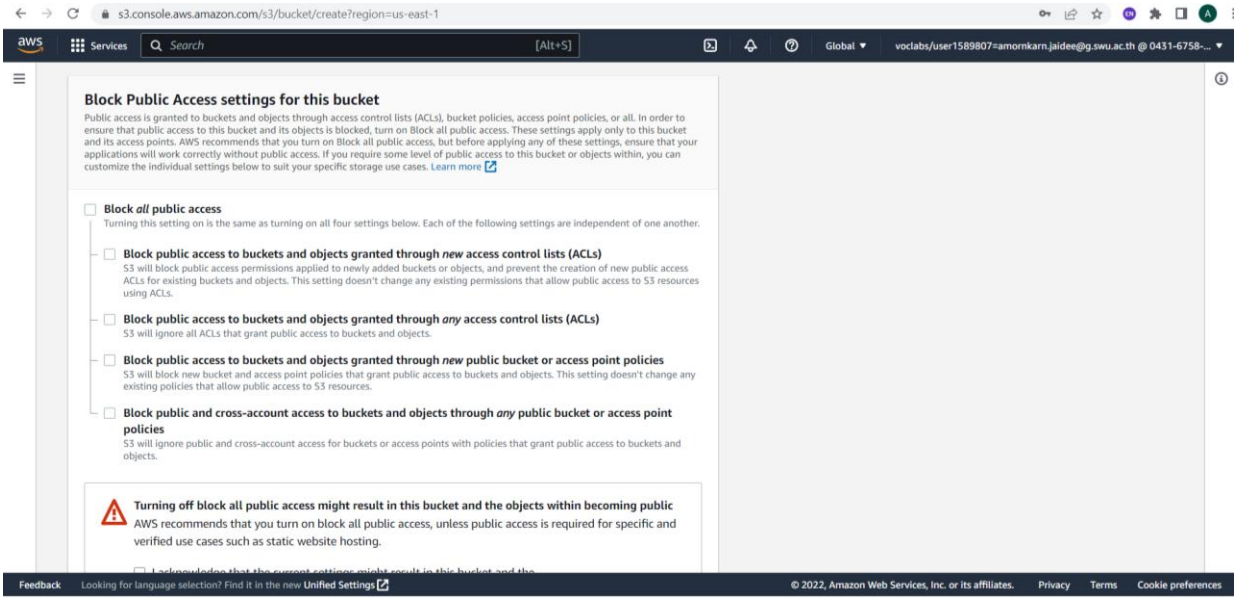
ออกแบบ data model ในรูปแบบนี้เนื่องจากใน table หลัก มี column จำนวนมาก ซึ่งบางส่วนจะไม่มีความสำคัญ จึงแบ่ง table ออกตามกลุ่มข้อมูลที่เราสนใจ เพื่อให้ง่ายต่อการ query ซึ่งไม่ซับซ้อนและใช้เวลานาน และตัดข้อมูล column ที่ไม่จำเป็นออก ให้แต่ละ table เก็บเฉพาะข้อมูลที่สนใจเท่านั้น

Process

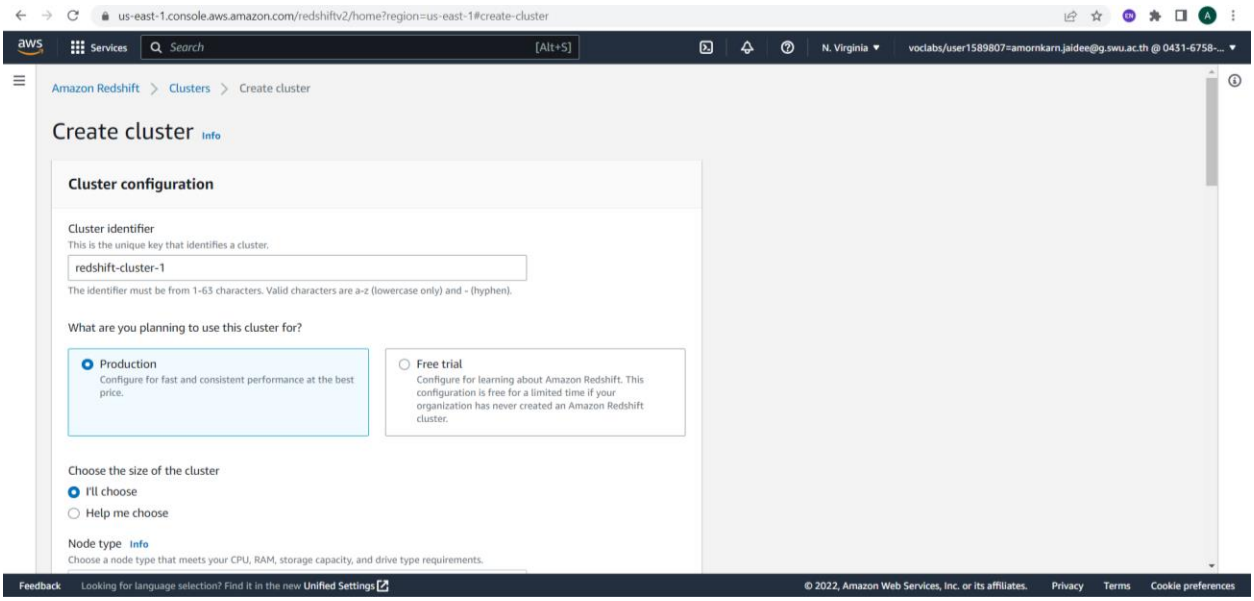
1. สร้าง Bucket ที่ S3 สำหรับเก็บไฟล์ raw data



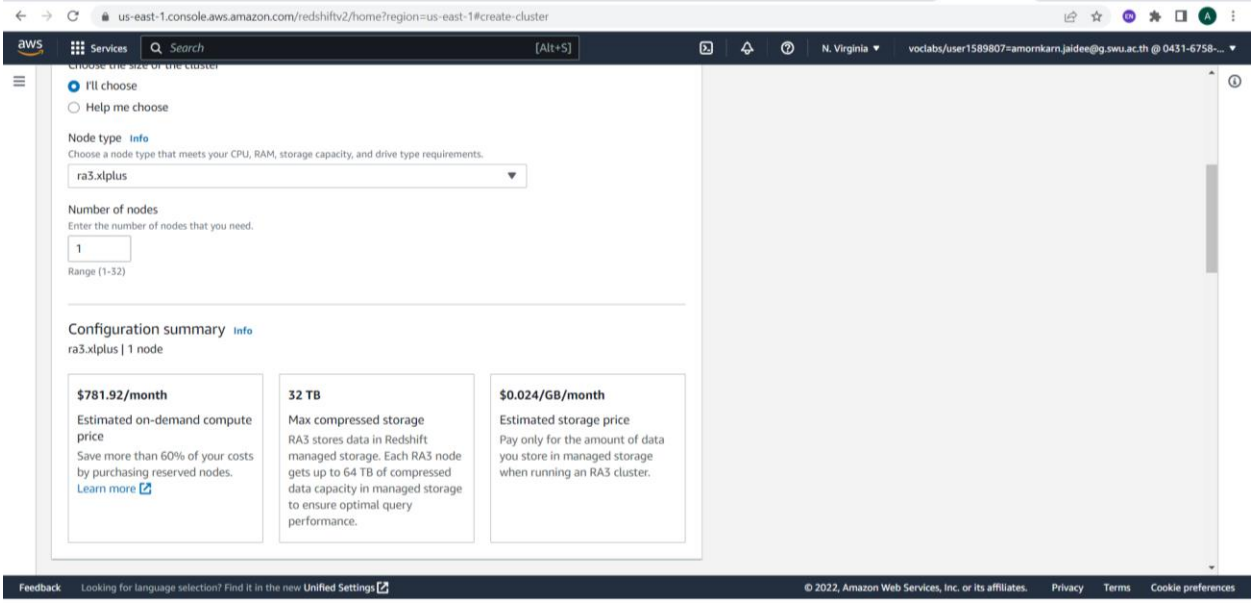
Unblock all public เพื่อให้สามารถเข้าถึงข้อมูลใน bucket ได้



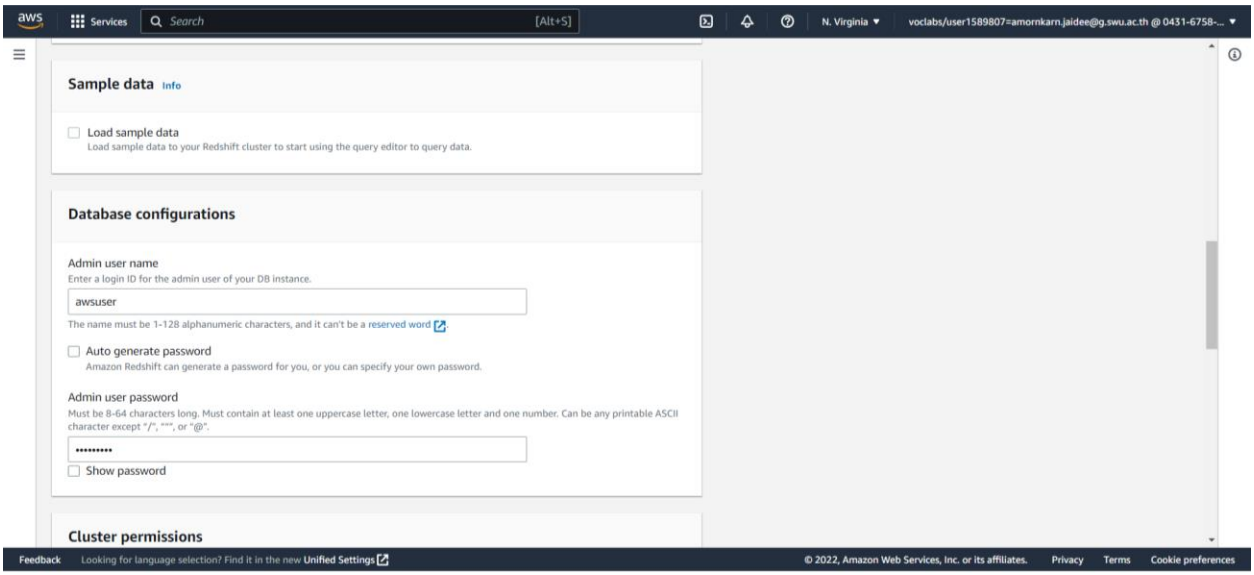
2. สร้าง Cluster ใน Redshift สำหรับการสร้าง Data Warehouse



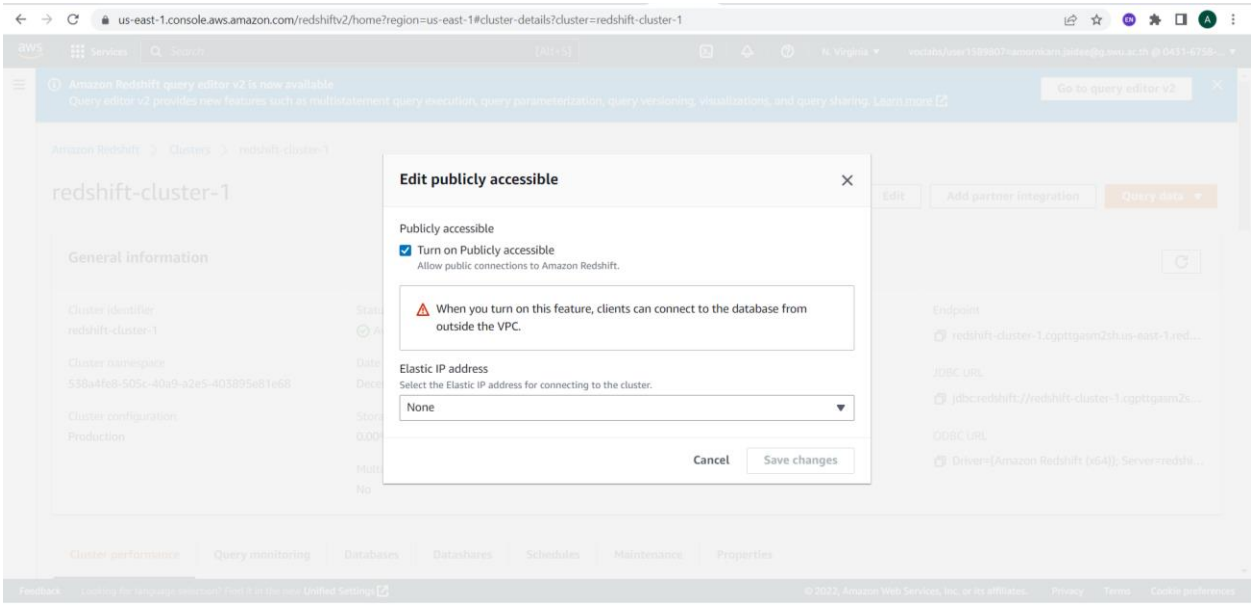
เลือก Node type หากข้อมูลมีจำนวนไม่มาก แนะนำให้เลือกเพียง 1 node



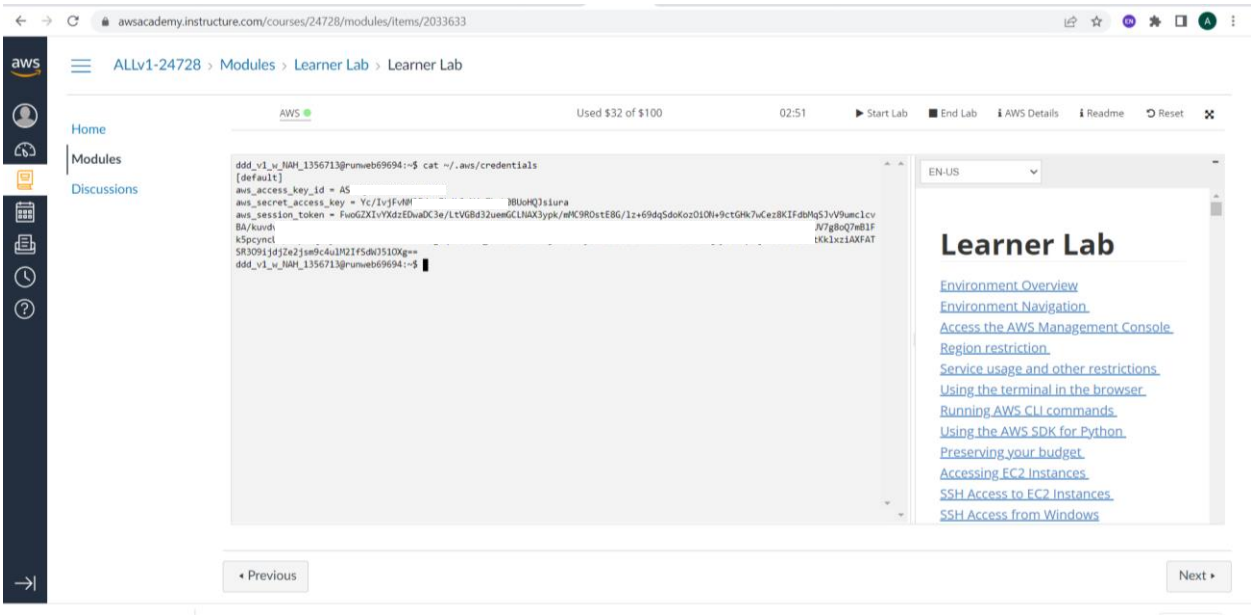
กำหนด username และ password สำหรับการเข้าถึง cluster



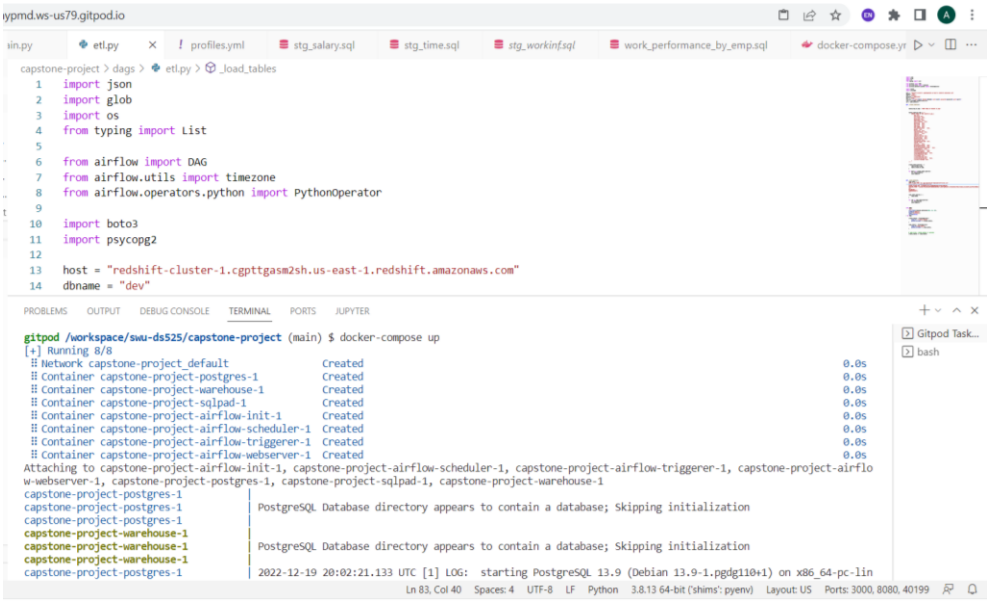
โดยต้องกำหนดสิทธิ์ของ Cluster ให้เป็นแบบ public เพื่อให้สามารถเข้าถึงได้



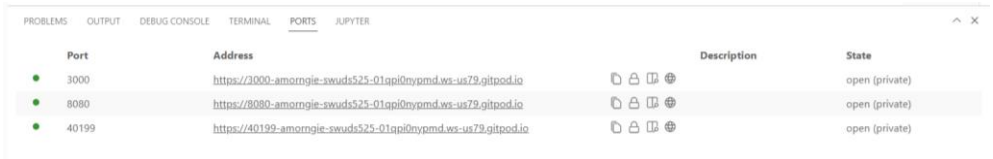
3. ดู secret keys ใน AWS console เพื่อใช้ในการเข้าถึง S3 โดยใช้คำสั่ง cat ~/.aws/credentials



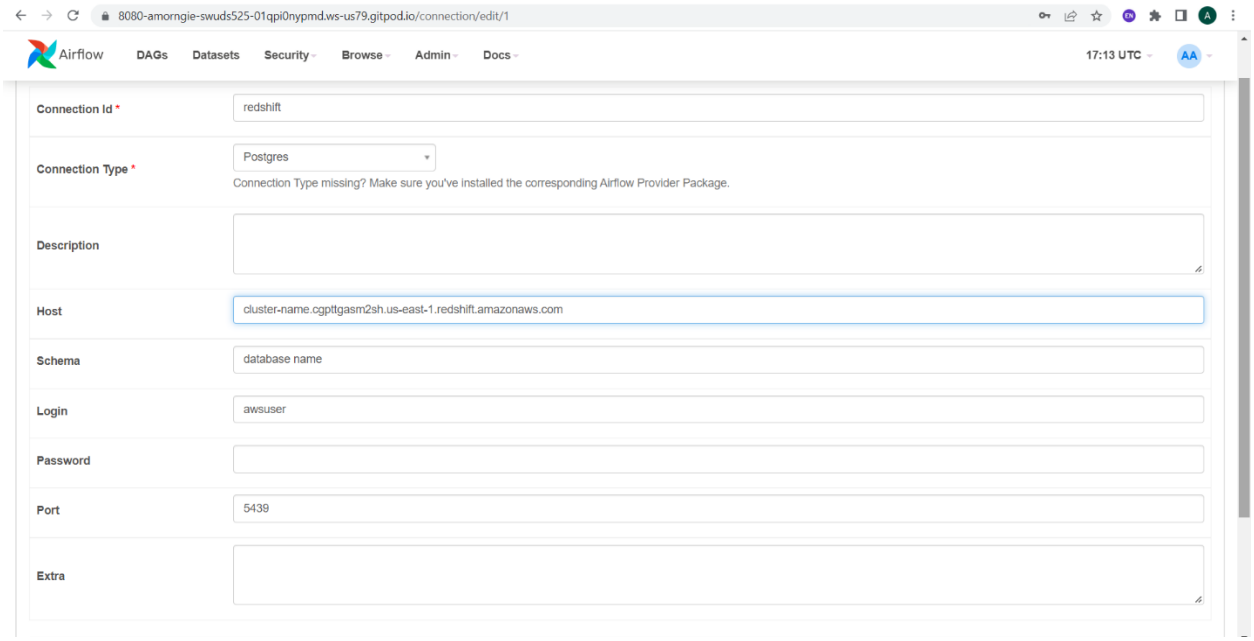
4. run คำสั่ง docker-compose up เพื่อเปิดใช้งาน Apache Airflow เพื่อสร้าง Automating data pipelines



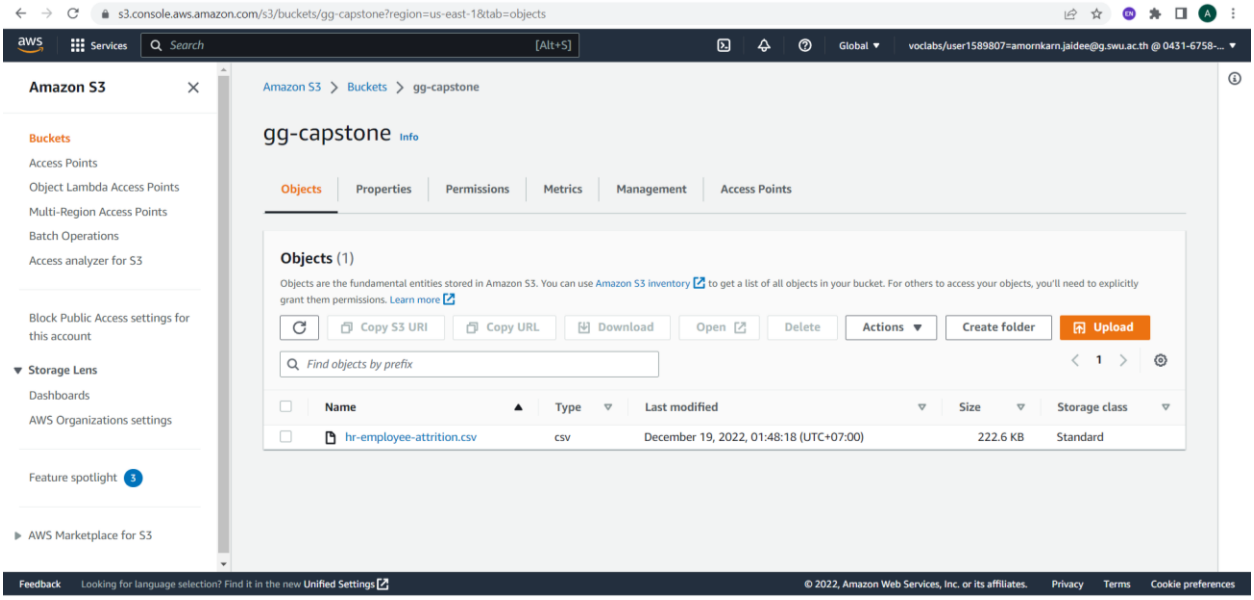
โดยเข้าใช้งาน Airflow ที่ port 8080



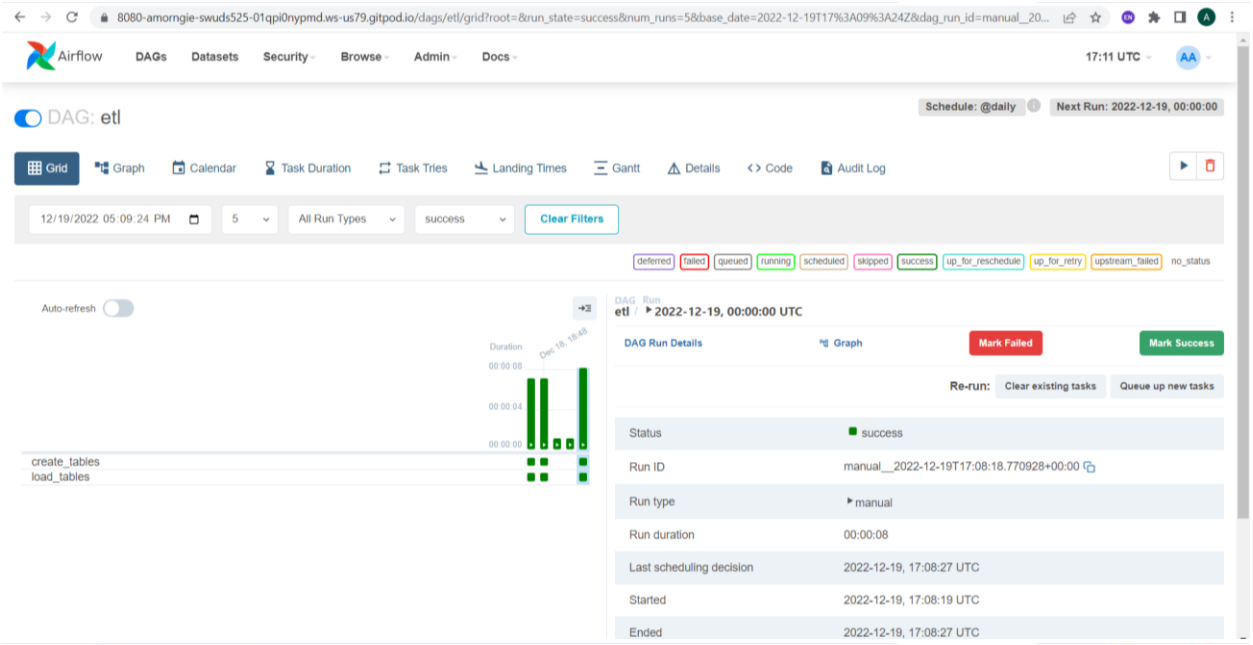
สร้าง connection เพื่อเชื่อมต่อกับ Redshift

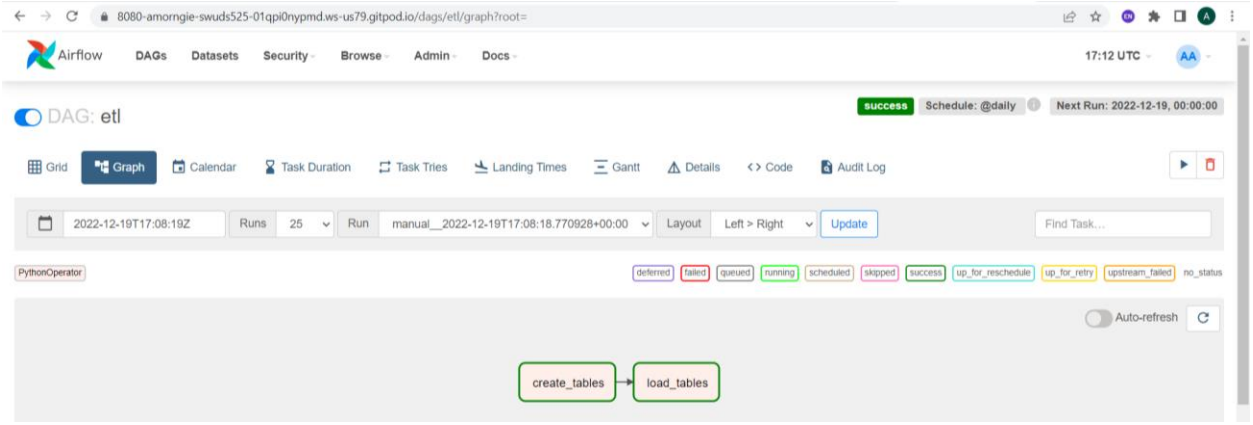


5. Run คำสั่ง python main.py เพื่อ upload raw data ไว้ที่ S3



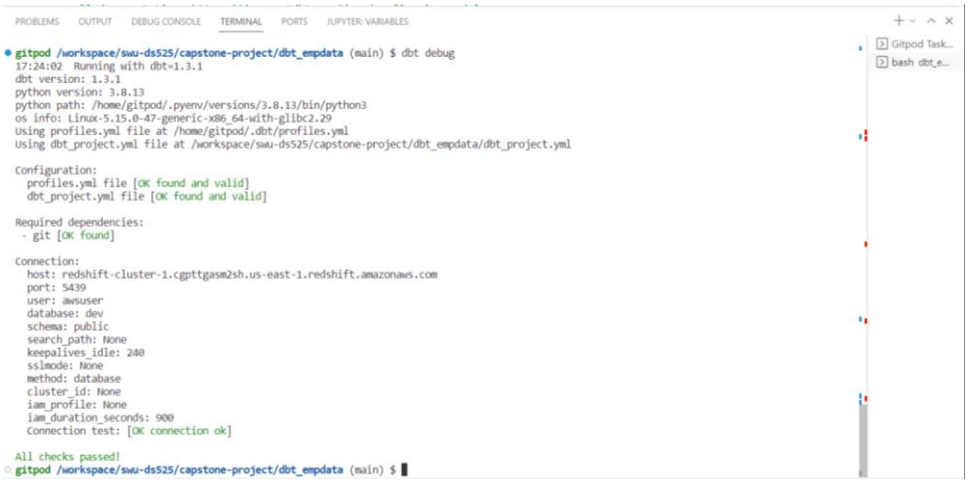
6. Run คำสั่ง python etl.py เพื่อสร้าง table และ load data จาก S3 ไปยัง Redshift หากดำเนินการสำเร็จ ใน Airflow จะแสดงสถานะ success ทุก task



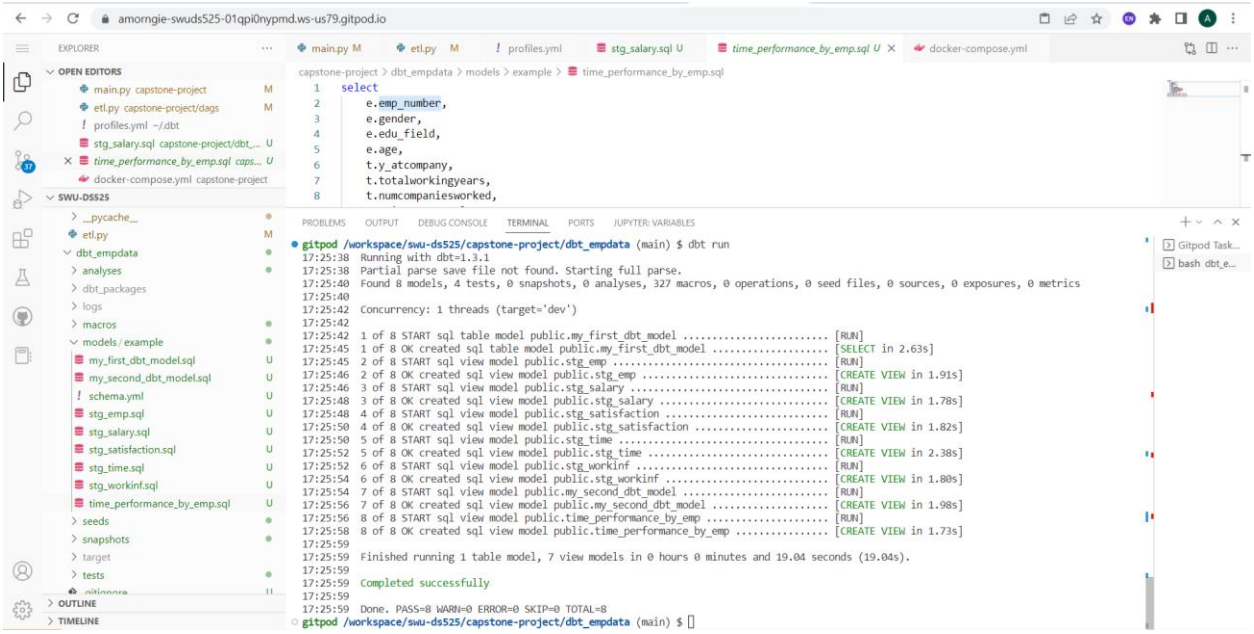


7. ใช้ dbt ในการสร้าง data model และ transform data โดย run คำสั่ง dbt init และใส่ข้อมูลที่กำหนด

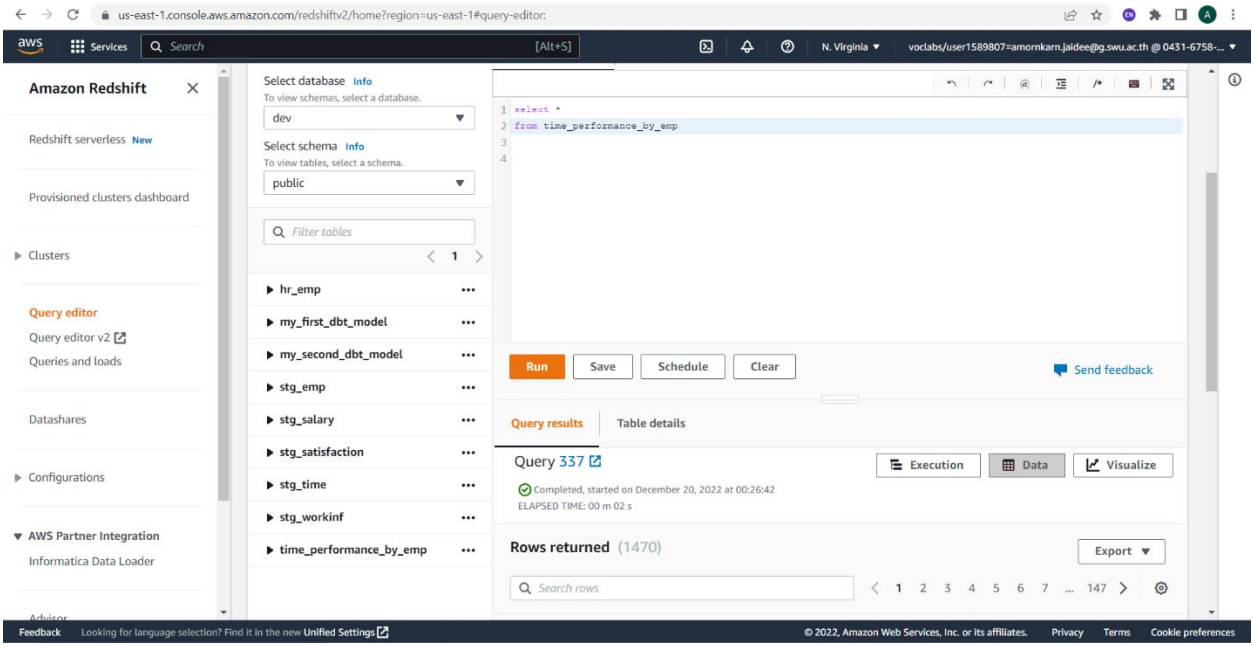
Run คำสั่ง dbt debug เพื่อตรวจว่าสามารถ connection ได้หรือไม่

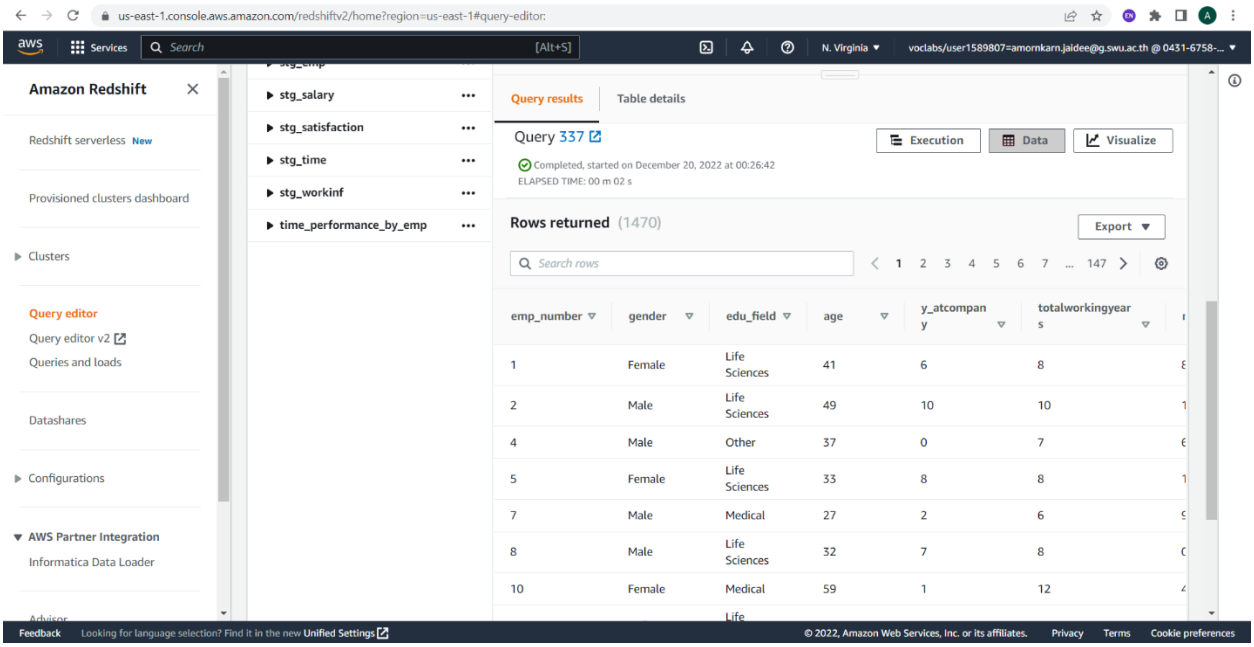


เมื่อสร้าง sql สำหรับ data model เรียบร้อยแล้ว ให้ run คำสั่ง dbt run เพื่อสร้าง data model และ transform data

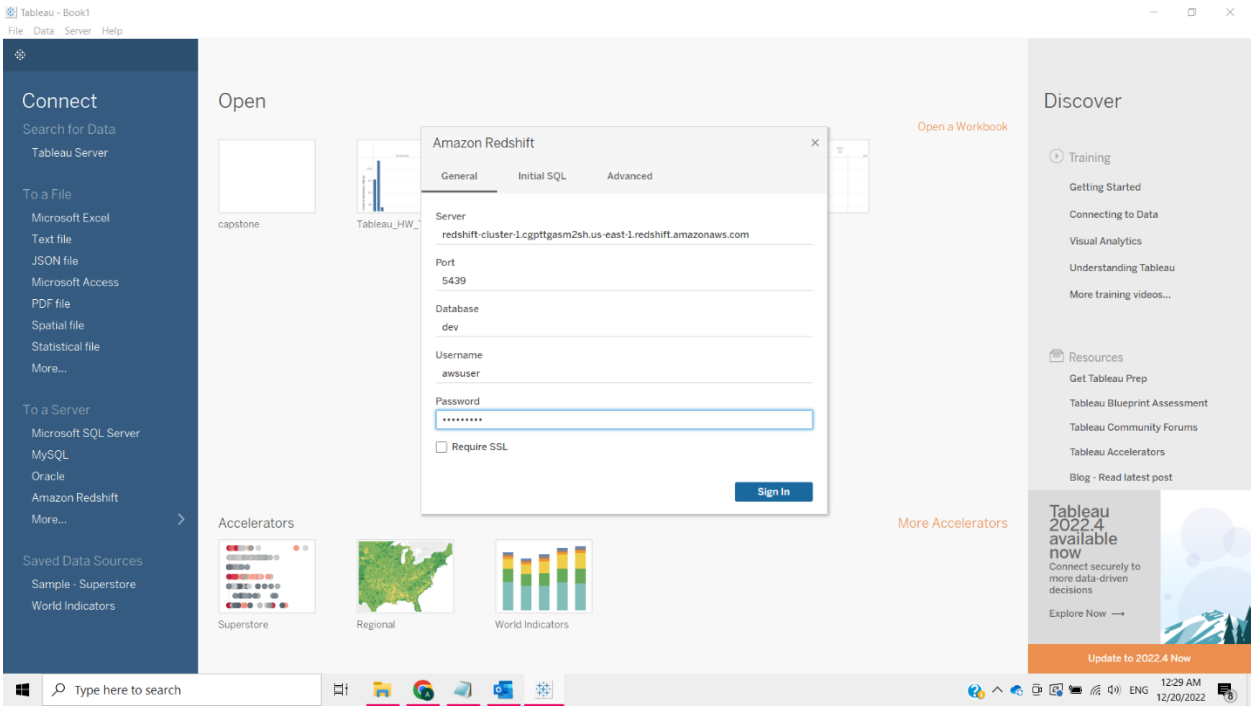


ตรวจสอบข้อมูลใน Redshift หากดำเนินการสำเร็จจะสามารถ query ข้อมูลออกมาดูได้

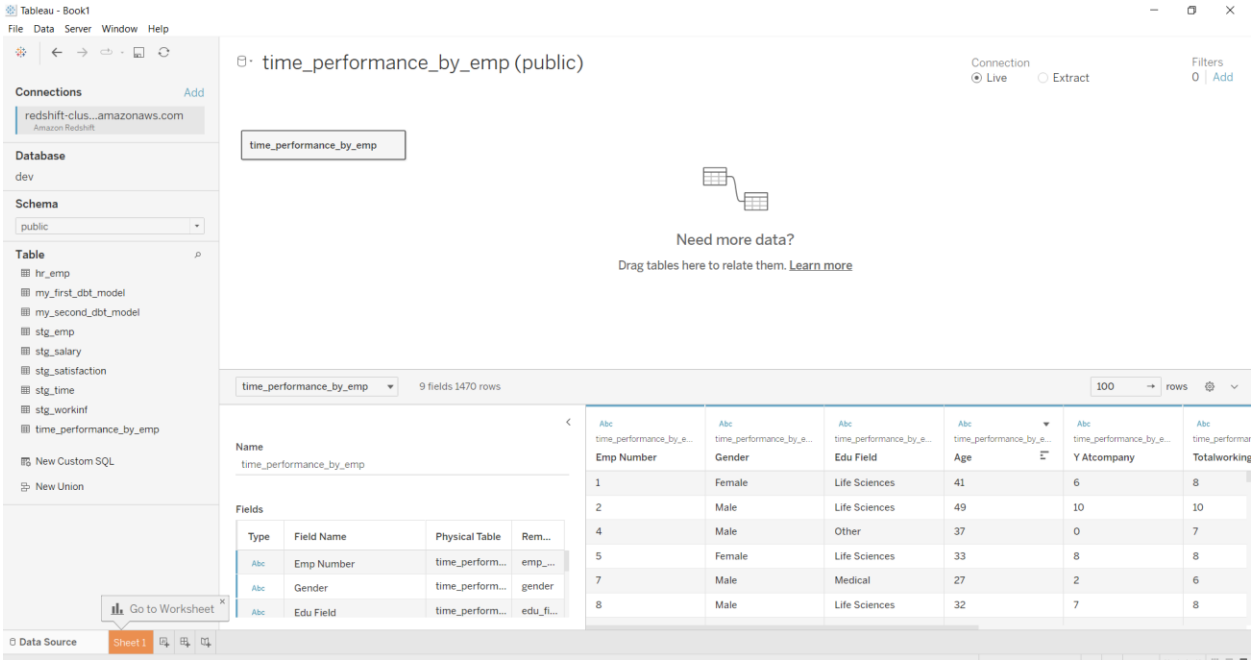




8. เชื่อมต่อ Redshift กับ Tableau สำหรับการทำ Data Visualization

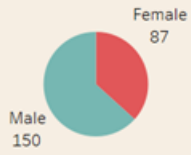


เลือก table ที่ต้องการใช้ในการทำ Dashboard ตามต้องการ

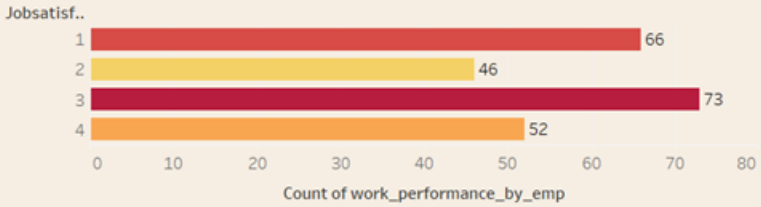


Employee Turnover

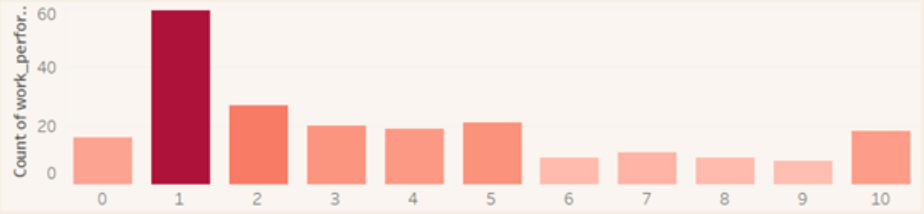
Gender



Job Satisfaction



Years at company



Performance Rating

Performancerating
3
4

Business Travel

Bsn Travel
Non-Travel
Travel_Freque..
Travel_Rarely