

Capstone Project

Employee Turnover

Problem

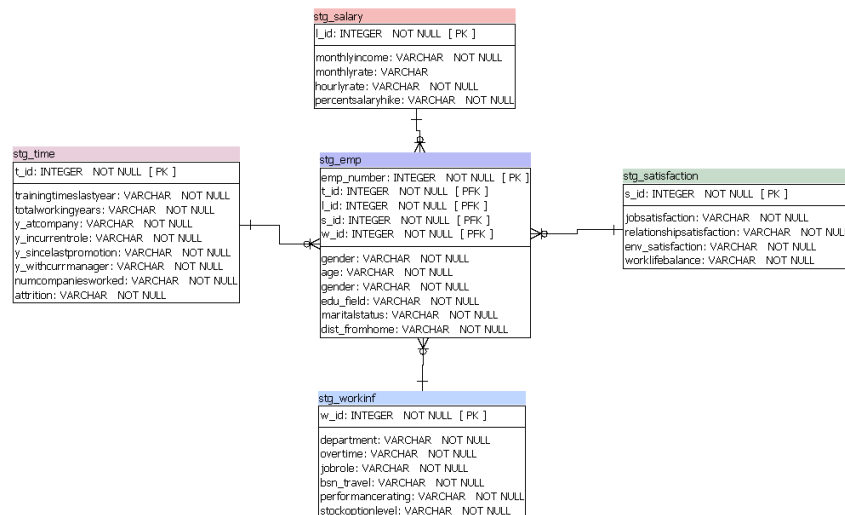
เนื่องจากพนักงานในองค์กรมีจำนวนของการลาออกเพิ่มมากขึ้น ทำให้พนักงานในแต่ละตำแหน่งมีจำนวนไม่เพียงพอกับงานในองค์กร ส่งผลให้งานที่ออกมาคุณภาพลดลง และการส่งงานมีความล่าช้ามากขึ้น ไม่เป็นไปตามเวลาที่ได้ตกลงกันไว้ และอาจส่งผลให้ธุรกิจหยุดชะงักได้ หากมีพนักงานไม่เพียงพอต่อความต้องการของงาน โดยการทำโปรเจกต์นี้เพื่อการหาสาเหตุหลักที่ทำให้พนักงานลาออกจากองค์กร 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 ซึ่งบางส่วนจะไม่มี ความสำคัญ จึงแบ่ง table ออกตามกลุ่มข้อมูลที่เราสนใจ เพื่อให้ง่ายต่อการ query ซึ่งไม่ซับซ้อนและใช้เวลานาน และตัดข้อมูล column ที่ไม่จำเป็นออก ให้แต่ละ table เก็บเฉพาะข้อมูลที่สนใจเท่านั้น

Process

1. สร้าง Bucket ที่ S3 สำหรับเก็บไฟล์ raw data

Amazon S3 > Buckets > Create bucket

Create bucket [Info](#)

Buckets are containers for data stored in S3. [Learn more](#)

General configuration

Bucket name

Bucket name must be globally unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#)

AWS Region

Copy settings from existing bucket - optional
Only the bucket settings in the following configuration are copied.

Unblock all public เพื่อให้สามารถเข้าถึงข้อมูลใน bucket ได้

Block Public Access settings for this bucket

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to this bucket and its objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to this bucket or objects within, you can customize the individual settings below to suit your specific storage use cases. [Learn more](#)

☐ **Block all public access**
Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.

☐ **Block public access to buckets and objects granted through new access control lists (ACLs)**
S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.

☐ **Block public access to buckets and objects granted through any access control lists (ACLs)**
S3 will ignore all ACLs that grant public access to buckets and objects.

☐ **Block public access to buckets and objects granted through new public bucket or access point policies**
S3 will block new bucket and access point policies that grant public access to buckets and objects. This setting doesn't change any existing policies that allow public access to S3 resources.

☐ **Block public and cross-account access to buckets and objects through any public bucket or access point policies**
S3 will ignore public and cross-account access for buckets or access points with policies that grant public access to buckets and objects.

Warning Turning off block all public access might result in this bucket and the objects within becoming public. AWS recommends that you turn on block all public access, unless public access is required for specific and verified use cases such as static website hosting.

☐ I acknowledge that the current settings might result in this bucket and the objects within becoming public.

[Feedback](#) Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

2. สร้าง Cluster ใน Redshift สำหรับการสร้าง Data Warehouse

us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#create-cluster

Amazon Redshift > Clusters > Create cluster

Create cluster [info](#)

Cluster configuration

Cluster identifier
This is the unique key that identifies a cluster.

redshift-cluster-1

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

What are you planning to use this cluster for?

☒ **Production**
Configure for fast and consistent performance at the best price.

☐ **Free trial**
Configure for learning about Amazon Redshift. This configuration is free for a limited time if your organization has never created an Amazon Redshift cluster.

Choose the size of the cluster

☒ **I'll choose**

☐ **Help me choose**

Node type [info](#)
Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

Feedback Looking for language selection? Find it in the new Unified Settings.

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

เลือก Node type หากข้อมูลมีจำนวนไม่มาก แนะนำให้เลือกเพียง 1 node

us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#create-cluster

Choose the size of the cluster

☒ **I'll choose**

☐ **Help me choose**

Node type [info](#)
Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

ra3.xlplus

Number of nodes
Enter the number of nodes that you need.

1

Range (1-32)

Configuration summary [info](#)
ra3.xlplus | 1 node

\$781.92/month Estimated on-demand compute price Save more than 60% of your costs by purchasing reserved nodes. Learn more	32 TB Max compressed storage RA3 stores data in Redshift managed storage. Each RA3 node gets up to 64 TB of compressed data capacity in managed storage to ensure optimal query performance.	\$0.024/GB/month Estimated storage price Pay only for the amount of data you store in managed storage when running an RA3 cluster.
--	---	---

Feedback Looking for language selection? Find it in the new Unified Settings.

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

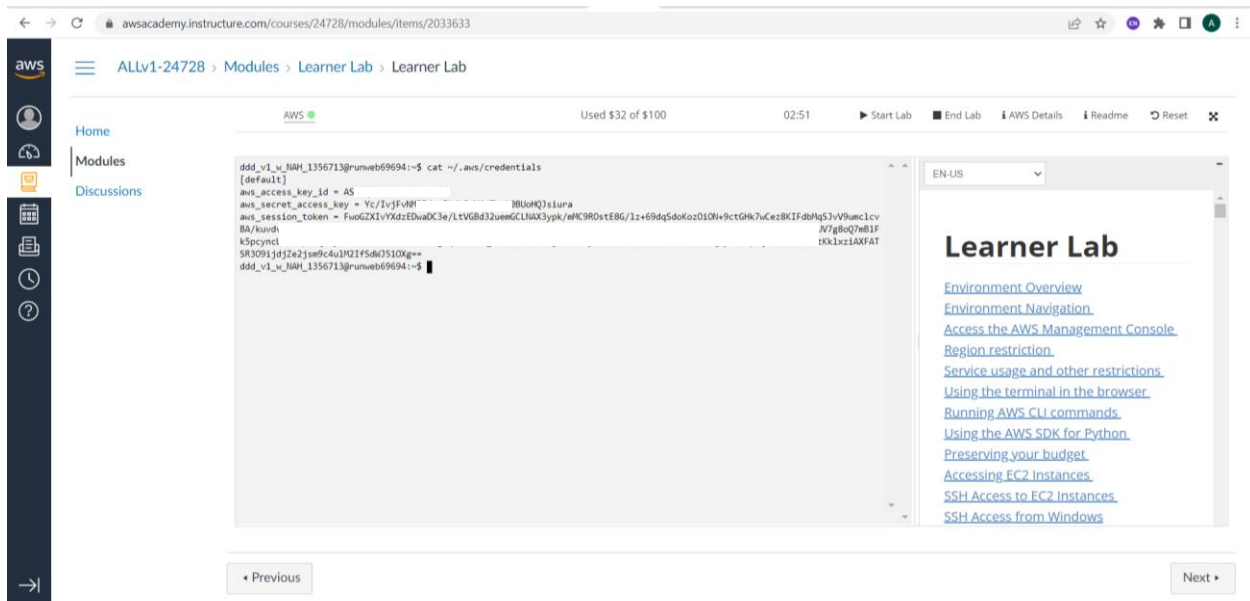
กำหนด username และ password สำหรับการเข้าถึง cluster

The screenshot shows the AWS Redshift console interface. At the top, there's a navigation bar with the AWS logo, 'Services' link, a search bar, and a user profile dropdown. The main content area is divided into sections: 'Sample data' with a 'Load sample data' checkbox and description; 'Database configurations' with fields for 'Admin user name' (set to 'awsuser') and 'Admin user password' (masked with dots); and 'Cluster permissions' at the bottom. A footer bar contains a feedback link, copyright notice, and privacy/terms links.

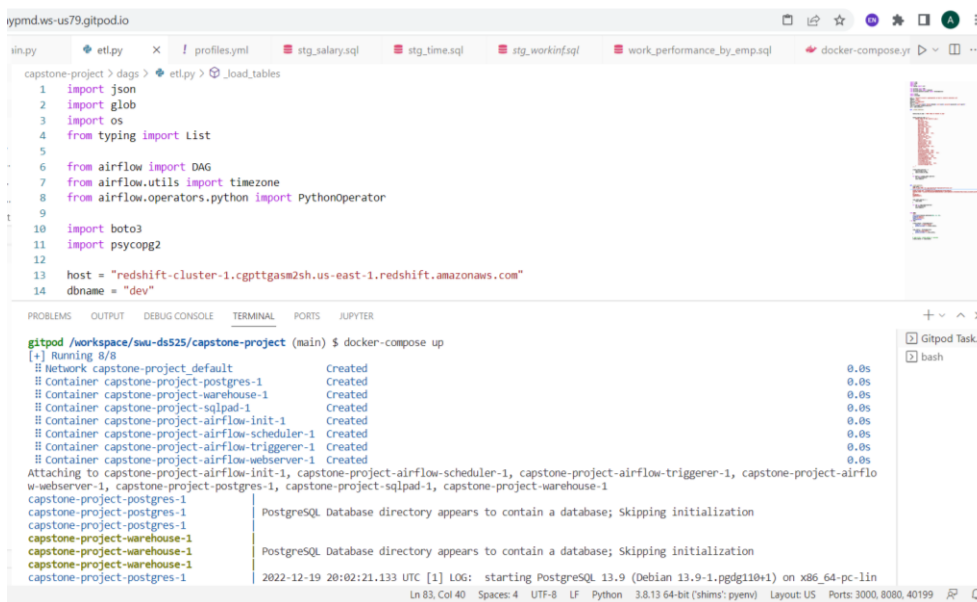
โดยต้องกำหนดสิทธิ์ของ Cluster ให้เป็นแบบ public เพื่อให้สามารถเข้าถึงได้

This screenshot shows the 'redshift-cluster-1' details page in the AWS console. A modal dialog titled 'Edit publicly accessible' is open in the center. The dialog has a 'Publicly accessible' section with a checked checkbox 'Turn on Publicly accessible' and a warning icon with text: 'When you turn on this feature, clients can connect to the database from outside the VPC.' Below this is an 'Elastic IP address' dropdown menu currently set to 'None'. At the bottom of the dialog are 'Cancel' and 'Save changes' buttons. The background shows the cluster's general information and a navigation bar at the bottom.

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



โดยใช้งาน Airflow ที่ port 8080

Port	Address	Description	State
3000	https://3000-amornjie-swuds525-01api@nypmd.ws-us79.gitpod.io		open (private)
8080	https://8080-amornjie-swuds525-01api@nypmd.ws-us79.gitpod.io		open (private)
40199	https://40199-amornjie-swuds525-01api@nypmd.ws-us79.gitpod.io		open (private)

สร้าง connection เพื่อเชื่อมต่อกับ Redshift

8080-amongie-swuds525-01api0nypmd.ws-us79.gitpod.io/connection/edit/1

Airflow DAGs Datasets Security Browse Admin Docs 17:13 UTC AA

Connection Id * redshift

Connection Type * Postgres
Connection Type missing? Make sure you've installed the corresponding Airflow Provider Package.

Description

Host cluster-name.cgptlgasm2sh.us-east-1.redshift.amazonaws.com

Schema database name

Login awsuser

Password

Port 5439

Extra

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

s3.console.aws.amazon.com/s3/buckets/gg-capstone?region=us-east-1&tab=objects

Amazon S3 Buckets gg-capstone

gg-capstone info

Objects Properties Permissions Metrics Management Access Points

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	hr-employee-attrition.csv	csv	December 19, 2022, 01:48:18 (UTC+07:00)	222.6 KB	Standard

Feedback Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

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

The top screenshot shows the Airflow web interface for the 'etl' DAG. The status is 'success'. The DAG Run Details table shows the following information:

Field	Value
Status	success
Run ID	manual__2022-12-19T17:08:18.770928+00:00
Run type	manual
Run duration	00:00:08
Last scheduling decision	2022-12-19, 17:08:27 UTC
Started	2022-12-19, 17:08:19 UTC
Ended	2022-12-19, 17:08:27 UTC

The bottom screenshot shows the DAG graph with two tasks: 'create_tables' and 'load_tables'. Both tasks are in a 'success' state.

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

```
gitpod /workspace/svu-ds525 (main) $ cd capstone-project
gitpod /workspace/svu-ds525/capstone-project (main) $ dbt init
17:18:36 Running with dbt=1.3.1
Enter a name for your project (letters, digits, underscore): dbt_empdata
Which database would you like to use?
[1] postgres
[2] redshift
(Don't see the one you want? https://docs.getdbt.com/docs/available-adapters)

Enter a number: 2
host (hostname.region.redshift.amazonaws.com): redshift-cluster-1.cgptgasm2sh.us-east-1.redshift.amazonaws.com

Problems OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER VARIABLES
[2] Gitpod Task...
[2] python3 ca...
```

```
Problems OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER VARIABLES
(Don't see the one you want? https://docs.getdbt.com/docs/available-adapters)

Enter a number: 2
host (hostname.region.redshift.amazonaws.com): redshift-cluster-1.cgptgasm2sh.us-east-1.redshift.amazonaws.com
port [5439]:
user (dev username): awsuser
[1] password
[2] iam
Desired authentication method option (enter a number): 1
password (dev password):
dbname (default database that dbt will build objects in): dev
schema (default schema that dbt will build objects in): public
threads (1 or more) [1]:
```


Run คำสั่ง dbt debug เพื่อตรวจสอบว่าสามารถ connection ได้หรือไม่

```
gitpod /workspace/swu-ds525/capstone-project/dbt_empdata (main) $ dbt debug
17:24:02 Running with dbt=1.3.1
dbt version: 1.3.1
python version: 3.8.13
python path: /home/gitpod/.pyenv/versions/3.8.13/bin/python3
os info: Linux-5.15.0-47-generic-x86_64-with-glibc2.29
using profiles.yml file at /home/gitpod/.dbt/profiles.yml
using dbt_project.yml file at /workspace/swu-ds525/capstone-project/dbt_empdata/dbt_project.yml

Configuration:
  profiles.yml file [OK found and valid]
  dbt_project.yml file [OK found and valid]

Required dependencies:
  - git [OK found]

Connection:
  host: redshift-cluster-1.cgptgasm2sh.us-east-1.redshift.amazonaws.com
  port: 5439
  user: awsuser
  database: dev
  schema: public
  search_path: None
  keepalives_idle: 240
  sslmode: None
  method: database
  cluster_id: None
  iam_profile: None
  iam_duration_seconds: 900
  connection test: [OK connection ok]

All checks passed!
gitpod /workspace/swu-ds525/capstone-project/dbt_empdata (main) $
```

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

```
capstone-project > dbt_empdata > models > example > time_performance_by_emp.sql

1  select
2      e.emp_number,
3      e.gender,
4      e.edu_field,
5      e.age,
6      t.y_atcompany,
7      t.totalworkingyears,
8      t.numcompaniesworked,

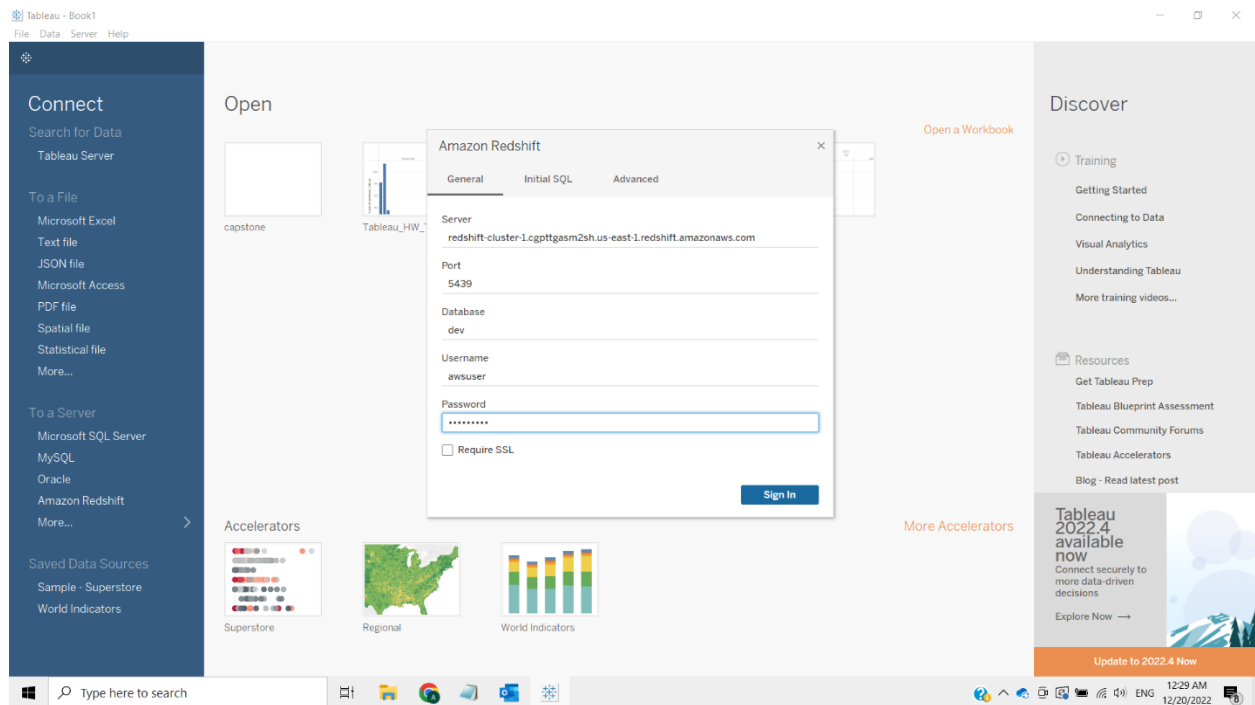
gitpod /workspace/swu-ds525/capstone-project/dbt_empdata (main) $ dbt run
17:25:38 Running with dbt=1.3.1
17:25:38 Partial parse save file not found. Starting full parse.
17:25:40 Found 8 models, 4 tests, 0 snapshots, 0 analyses, 327 macros, 0 operations, 0 seed files, 0 sources, 0 exposures, 0 metrics
17:25:40 Concurrency: 1 threads (target='dev')
17:25:42
17:25:42 1 of 8 START sql table model public.my_first_dbt_model ..... [RUN]
17:25:45 1 of 8 OK created sql table model public.my_first_dbt_model ..... [SELECT in 2.63s]
17:25:45 2 of 8 START sql view model public.stg_emp ..... [RUN]
17:25:46 2 of 8 OK created sql view model public.stg_emp ..... [CREATE VIEW in 1.91s]
17:25:46 3 of 8 START sql view model public.stg_salary ..... [RUN]
17:25:48 3 of 8 OK created sql view model public.stg_salary ..... [CREATE VIEW in 1.78s]
17:25:48 4 of 8 START sql view model public.stg_satisfaction ..... [RUN]
17:25:50 4 of 8 OK created sql view model public.stg_satisfaction ..... [CREATE VIEW in 1.82s]
17:25:50 5 of 8 START sql view model public.stg_time ..... [RUN]
17:25:52 5 of 8 OK created sql view model public.stg_time ..... [CREATE VIEW in 2.38s]
17:25:52 6 of 8 START sql view model public.stg_workinf ..... [RUN]
17:25:54 6 of 8 OK created sql view model public.stg_workinf ..... [CREATE VIEW in 1.80s]
17:25:54 7 of 8 START sql view model public.my_second_dbt_model ..... [RUN]
17:25:56 7 of 8 OK created sql view model public.my_second_dbt_model ..... [CREATE VIEW in 1.98s]
17:25:56 8 of 8 START sql view model public.time_performance_by_emp ..... [RUN]
17:25:58 8 of 8 OK created sql view model public.time_performance_by_emp ..... [CREATE VIEW in 1.73s]
17:25:59
17:25:59 Finished running 1 table model, 7 view models in 0 hours 0 minutes and 19.04 seconds (19.04s).
17:25:59 Completed successfully
17:25:59 Done. PASS=8 WARN=0 ERROR=0 SKIP=0 TOTAL=8
gitpod /workspace/swu-ds525/capstone-project/dbt_empdata (main) $
```


ตรวจสอบข้อมูลใน Redshift หากดำเนินการสำเร็จจะสามารถ query ข้อมูลออกมาได้

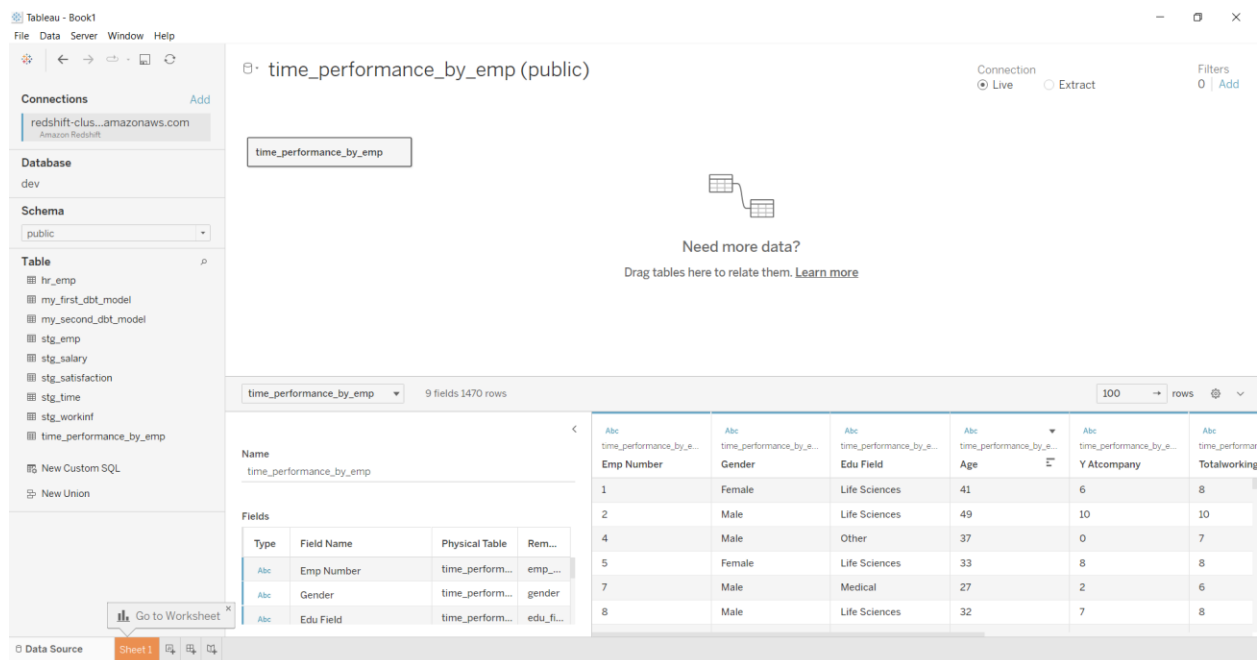
The screenshot displays the Amazon Redshift Query Editor interface. The left sidebar contains navigation options: Amazon Redshift, Redshift serverless, Provisioned clusters dashboard, Clusters, Query editor (selected), Queries and loads, Datashares, Configurations, and AWS Partner Integration. The main panel shows the 'Query editor' view with a SQL query: `select * from time_performance_by_emp`. The query has been executed successfully, as indicated by the 'Query results' tab and the status 'Completed, started on December 20, 2022 at 00:26:42'. The results show 1470 rows returned. The table structure includes columns: emp_number, gender, edu_field, age, y_atcompan, and totalworkingyear. The first few rows of data are visible.

emp_number	gender	edu_field	age	y_atcompan	totalworkingyear
1	Female	Life Sciences	41	6	8
2	Male	Life Sciences	49	10	10
4	Male	Other	37	0	7
5	Female	Life Sciences	33	8	8
7	Male	Medical	27	2	6
8	Male	Life Sciences	32	7	8
10	Female	Medical	59	1	12

8. เชื่อมต่อ Redshift กับ Tableau สำหรับการทำ Data Visualization

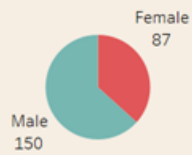


เลือก table ที่ต้องการใช้ในการทำ Dashboard ตามต้องการ

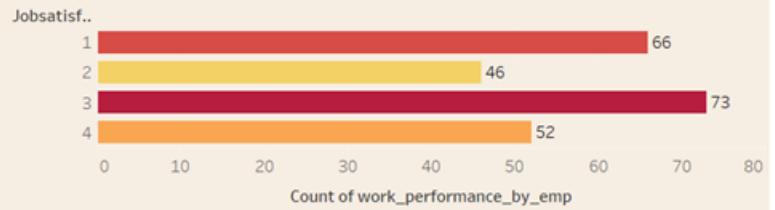


Employee Turnover

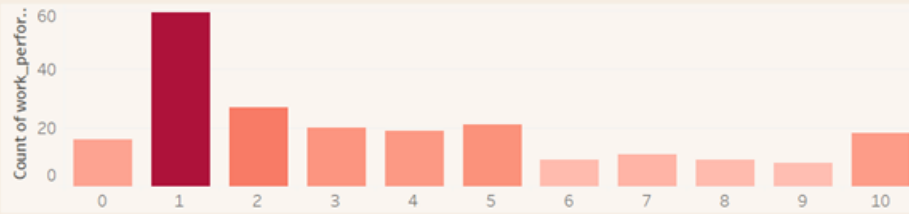
Gender



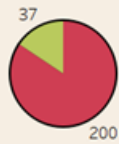
Job Satisfaction



Years at company

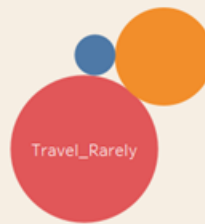


Performance Rating



Performance Rating

Business Travel



Business Travel

Non-Travel
Travel_Frequently
Travel_Rarely