

ปัญหา ปัจจัยใดที่ส่งผลกระทบต่อการสมัครผลิตภัณฑ์เงินฝากประจำทางโทรศัพท์ของลูกค้า

เมื่อยุคสมัยเปลี่ยนไป แต่มีพฤติกรรมบางอย่างของลูกค้า อาจจะยังไม่เปลี่ยนตาม แม้จะเป็น ยุคดิจิทัล ที่มีการใช้งาน Internet กันแทบทุกวัน แต่สำหรับการเสนอขายผลิตภัณฑ์ของธนาคาร การสนทนาก็โดยตรงทางโทรศัพท์กับเจ้าหน้าที่ ยังเป็นวิธีการที่ทำให้ลูกค้าตัดสินใจสมัครมากที่สุด เพราะด้วยพฤติกรรมของมนุษย์ที่ยังคงเน้นการสื่อสารระหว่างบุคคล และ 2 ฝั่ง มากกว่าที่จะสื่อสารผ่านเดียว รวมไปถึงการเสนอผลิตภัณฑ์ให้ลูกค้าโดยตรงแบบเห็นหน้า ก็มีโอกาสสูงที่ลูกค้าจะให้ความสนใจ แต่นั้น หมายความว่า ลูกค้าจะต้องเดินทางมาที่สาขาของธนาคาร และด้วยความที่ ผลิตภัณฑ์ ด้าน ธนาคารหรือการเงิน เป็นผลิตภัณฑ์ที่มีรายละเอียดเยอะ และทำความเข้าใจยาก การยิง ads โฆษณา การส่ง e-mail การส่ง sms จึงทำให้ลด ความสนใจของลูกค้าได้ แต่ในการติดต่อทางโทรศัพท์นั้นก็มีโอกาสที่จะทำให้ลูกค้ารู้สึกถึงความไม่เป็นส่วนตัว หรือ การเจ้าหน้าที่ ติดต่อไปที่ลูกค้า บ่อยจนเกินไป อาจทำให้ลูกค้า มีหัศจรรย์ ที่ไม่ต่อ ธนาคาร และผลิตภัณฑ์ได้

จากที่กล่าวมาข้างต้น การนำข้อมูล การตัดสินใจในการสมัครซื้อผลิตภัณฑ์ ของลูกค้าธนาคารมีวิเคราะห์ ตามปัจจัย หรือ Insight พฤติกรรมของลูกค้า และพฤติกรรมการติดต่อของพนักงาน จะทำให้ธนาคารหรือองค์กร สามารถ นำมาปรับเปลี่ยนวิธีการในการนำเสนอผลิตภัณฑ์มีวิเคราะห์ และ ออกแบบ แผนการดำเนินงาน เพื่อเพิ่มโอกาสให้ลูกค้ายินยอมและตกลงซื้อ ผลิตภัณฑ์เพิ่มมากขึ้น

- Dataset ที่นำมาใช้

ข้อมูลผลการนำเสนอ ผลิตภัณฑ์ เงินฝากประจำทางโทรศัพท์ Bank Marketing Dataset -> จำนวน 41,118 records

โดยข้อมูลด้านบน เป็น public dataset ซึ่งเป็น ข้อมูลของสถาบันการเงินแห่งหนึ่งในประเทศโปรตุเกส เกี่ยวข้องกับ การติดต่อลูกค้าทางโทรศัพท์เพื่อเสนอขาย ผลิตภัณฑ์ เงินฝากประจำ

ชี้ง ที่มีการเลือกใช้ Dataset ดังกล่าวเนื่องจากถือเป็นข้อมูล ที่แต่ละสถาบันการเงิน มีการ record ไว้ และเพื่อหาตัวอย่างอย่างง่ายในการท้าความเข้าใจกับข้อมูล

- Data model

การออกแบบ Data model เน้นไปที่ข้อมูลที่จะมีการดึงมาใช้งานในการแก้ปัญหาด้านบน โดยจะเลือกเฉพาะข้อมูล ที่สำคัญ และ คาดว่า จะเป็นสำหรับการนำไปวิเคราะห์ โดย ออกแบบเป็น 1 ตารางหลัก โดย จะไม่มีข้อมูลที่ระบุถึงตัวตน หรือ Id ของ ลูกค้า เนื่องจาก ไม่ได้มีการใช้งานในข้อมูลส่วนนั้น

และ 2 ตารางย่อย ได้แก่

- ตารางข้อมูลการเสนอขาย

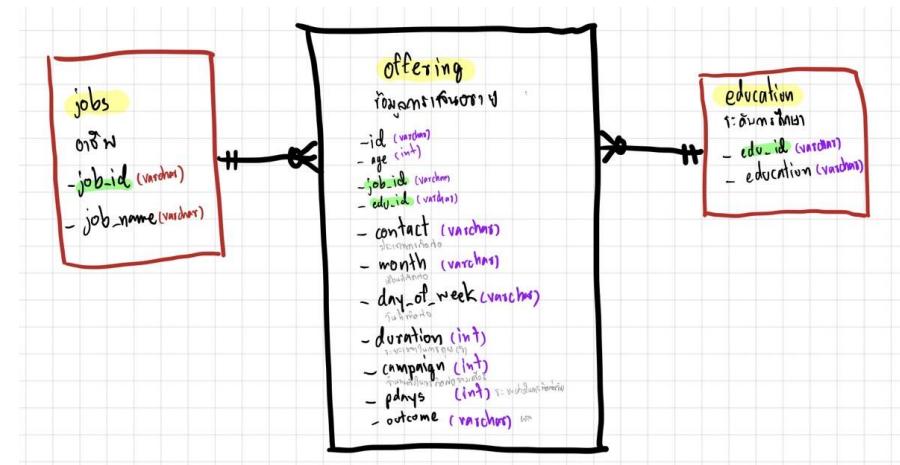
อายุ , รหัสอาชีพ, รหัสการศึกษา, ประเภทการติดต่อ , เดือนที่ ติดต่อ, วันในเดือนที่ติดต่อ, ระยะเวลา สนทนาระยะห่างในการติดต่อ, จำนวนครั้งในการติดต่อ, ผลกระทบในการขาย

- ตาราง อาชีพ

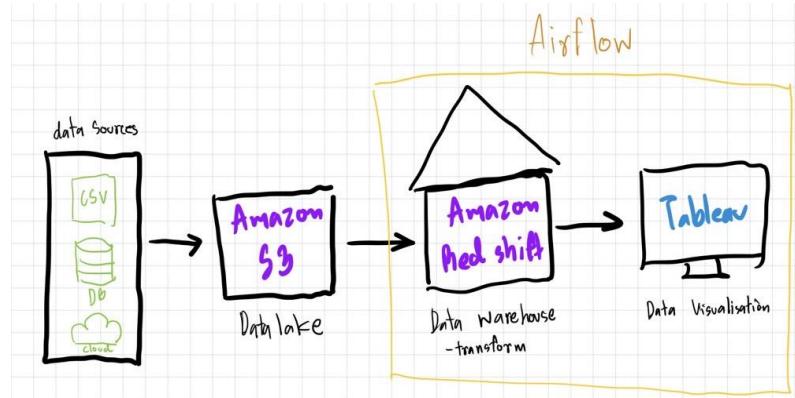
รหัสอาชีพ, อาชีพ

- ตารางระดับการศึกษา

รหัสระดับการศึกษา, ระดับการศึกษา



Data Pipeline



* ในความเป็นจริงควรใช้ AirFlow ในการทำงานตั้งแต่ต้น เพื่อควบคุมการทำงานและการไหลของข้อมูลได้สอดคล้อง แต่ด้วยข้อจำกัดของเครื่องมือในการทดลองทำให้ไม่สามารถตั้งค่าให้ AirFlow ทำงานได้

Step การทำงานมีดังนี้

1. Data source

นำ Raw data วางไว้ที่ s3 เพื่อเป็น data source จำลองในการดึงข้อมูลมาใช้งาน

- สร้าง Bucket ใน s3

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose [Provide feedback](#).

Amazon S3 > Buckets > datalake-bank-pleng

Properties

Bucket overview

AWS Region US East (N. Virginia) us-east-1	Amazon Resource Name (ARN) arn:aws:s3:::datalake-bank-pleng	Creation date December 10, 2022, 10:48:03 (UTC+07:00)
---	--	--

Bucket Versioning

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. [Learn more](#)

Bucket Versioning
Disabled

- ทำการสร้าง Folder landing เก็บ Raw_data

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with links for Buckets, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, and Access analyzer for S3. Below that are sections for Block Public Access settings, Storage Lens, Dashboards, AWS Organizations settings, and a Feature spotlight. At the bottom of the sidebar is a link to AWS Marketplace for S3. The main content area shows the 'datalake-bank-pleng' bucket. It has a 'Publicly accessible' label. Below it are tabs for Objects, Properties, Permissions, Metrics, Management, and Access Points. The 'Objects' tab is selected, showing a list of objects with 2 items. The table columns are Name, Type, Last modified, Size, and Storage class. The objects listed are 'export/' (Folder) and 'landing/' (Folder). At the top of the main content area, there's a message: 'We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose Provide feedback.' There are also links for 'Provide feedback', 'Global', and 'voclabs/user1591037=nattakan.towpunwong@g.swu.ac.th @ 2830-10...'. The URL in the address bar is https://s3.console.aws.amazon.com/s3/buckets/datalake-bank-pleng?region=ap-southeast-1.

2. Datalake Zone

เป็นส่วนที่ใช้จัดเก็บไฟล์ต่างๆ ที่ได้ทำการดึงข้อมูลมา และนำมาจัดเก็บไว้ ก่อนที่จะนำไป ETL หรือ Transform ใน Step ถัดไป โดยจะทำการสร้างไฟล์ datalake_s3.py เพื่อดึงข้อมูลจาก Folder ในส่วนของ Landing Zone มาจัดเก็บ ในส่วนนี้จะเก็บไว้ใน s3 โดยใช้คำสั่ง Spark ให้การดึงข้อมูลและสร้างตารางในการจัดเก็บ

- ทำการสร้าง Folder Export เก็บ clean data
- ตั้งค่าการเชื่อมต่อ AWS กับไฟล์ py
ใช้คำสั่ง \$ cat ~/.aws/credentials
 - aws_access_key_id
 - aws_secret_access_key
 - aws_session_token

The screenshot shows the AWS Learner Lab interface. On the left is a sidebar with navigation links: Home, Modules, Discussions, Calendar, Inbox, History, and Help. The main area has tabs for Home, Modules, and Discussions. A terminal window is open with the following command and output:

```
dd0_v1_u_Dm_1383602@unweb69255:~$ cat ~/.aws/credentials
[default]
aws_access_key_id = ASTAUOZPGRZAS30TSI
aws_secret_access_key = 123Kf0k0lts3MPaej15X883YoiSl1ts5xS0In0
aws_session_token = Fw0GDXIVXzGEfUk0GD05eSx0Bggg4M1LQAbJw4IDb/XBvAIFcA/g4y/r/mxHsd0p163wCICLojIVTmJTyW0iaZxHndrDUL0fFn1fTUestUsaR
RL0dC0D9b156e602ekhNt5t9AFaKS07vYH2el05SGhPOG+eXQd452H+rT0dRpwpVkmw45RVEt1pJf+j3n92fj8z3wPs68yPhUlg223gnYMPtU06wVUQceIX
cnuSuAPh12p...p.../nwd...088t98y17efcsmkae2/z291q51pgB1lo3IHqAVy...0x00CXB88tP0a8c35yc5070ze1n8eb/nvvKvVzyhQ6Rs2f2f1x10N...
dd0_v1_u_Dm_1383602@unweb69255:~$ < C
dd0_v1_u_Dm_1383602@unweb69255:~$ < C
dd0_v1_u_Dm_1383602@unweb69255:~$ < C
```

To the right of the terminal is a sidebar titled "Learner Lab" containing links to "Environment Overview", "Environment Navigation", "Access the AWS Management Console", "Region restriction", "Service usage and other restrictions", "Using the terminal in the browser", "Running AWS CLI commands", "Using the AWS SDK for Python", "Preserving your budget", "Accessing EC2 instances", "SSH Access to EC2 instances", "SSH Access from Windows", and "SSH Access from a Mac".

- ทำการรันผ่าน docker compose และเข้าไปรันไฟล์ที่ port 8888
- ```
cd capstone-project
python -m venv ENV
source ENV/bin/activate
pip install -r requirements.txt
docker-compose up
port 8888
```
- สร้างไฟล์ datalake\_s3.py บน Jupyter

The screenshot shows a Jupyter Notebook interface with several code cells. The first cell contains:

```
data = spark.read.option("header","true").option("multiline", "true").option("sep",",").csv(landing_zone)
```

The second cell contains:

```
data.printSchema()
```

The third cell contains:

```
data.show()
```

The output of the third cell shows the schema of the data:

```
+--+
| age | job | marital | education | default | balance | housing | loan | contact | day | month | duration | campaign | pdays | previous | poutcome | y |
+--+
| 58 | management | married | tertiary | no | 2143 | yes | no|unknown| 5 | may | 261 | 1 | -1 | 0 | unknown | no |
```

At the bottom, it says "Jupyter Server Remote Cell 1 of 20 Layout: US Ports: 4040 4041 8080 8888 39991".

### 3. Data Warehouse zone

เป็นส่วนที่ใช้จัดเก็บไฟล์ที่ผ่านการ Transform หรือ ETL ให้อยู่ในโครงสร้างที่ต้องการและสามารถนำไปใช้ต่อได้อย่างสะดวก โดยเก็บไว้ที่ Amazon Redshift และใช้ Airflow control ทำหน้าที่เป็น Data Pipeline ควบคุมการโหลดของข้อมูล และสามารถ Monitor การทำงานของ Workflow ได้

- ทำการ Create Cluster ใน Amazon Redshift

- ทำการสร้างไฟล์ `datawarehouse.py` เก็บไว้ใน Folder Dags และทำการสร้างตารางที่ต้องการนำไปใช้งานตามที่ออกแบบไว้ (`create_tables[สร้างตาราง] >> truncate_tables[ลบตารางที่เคยสร้าง] >> load_staging_tables >> insert_tables`)

```

 1 import os
 2 import glob
 3 from sqlite3 import Timestamp
 4 from typing import List
 5 import json
 6 from datetime import datetime
 7 import psycopg2
 8
 9 from airflow import DAG
 10 from airflow.utils import timezone
 11 from airflow.operators.python import PythonOperator
 12 # from airflow.operators.bash_operator import BashOperator
 13 # from airflow.hooks.postgres_hook import PostgresHook
 14
 15 curr_date = datetime.today().strftime('%Y-%m-%d')
 16
 17 create_table_queries = [
 18 """
 19 CREATE TABLE IF NOT EXISTS edu (
 20 edu_id bigint,
 21 education text
 22);
 23 """,
 24 """
 25 CREATE TABLE IF NOT EXISTS job (
 26 job_id bigint,
 27 job text
 28);
 29 """,
 30 """
 31 CREATE TABLE IF NOT EXISTS offering (
 32 age int,
 33 job_id bigint,
 34 edu_id bigint,
 35 contact text,
 36 month text,
 37 day int,
 38 duration int,
 39 campaign int,
 40 pdays int,
 41 outcome text
 42);
 43 """

```

Lin 158, Col 73 Spaces: 4 UTF-8 LF Python 3.8.13 64-bit ('shims': pyenv) Layout: US Ports: 4040:4041:8080:8888:39591

- ทำการ docker compose และเข้าไปที่ port 8080

cd capstone-project

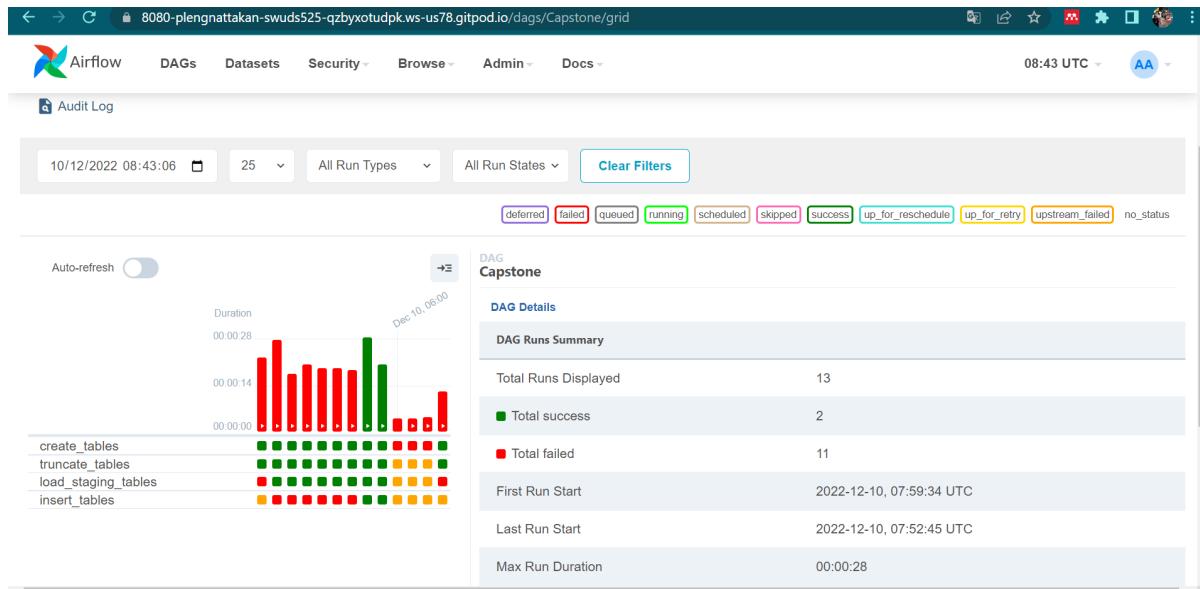
mkdir -p ./dags ./logs ./plugins

echo -e "AIRFLOW\_UID=\$(id -u)" > .env

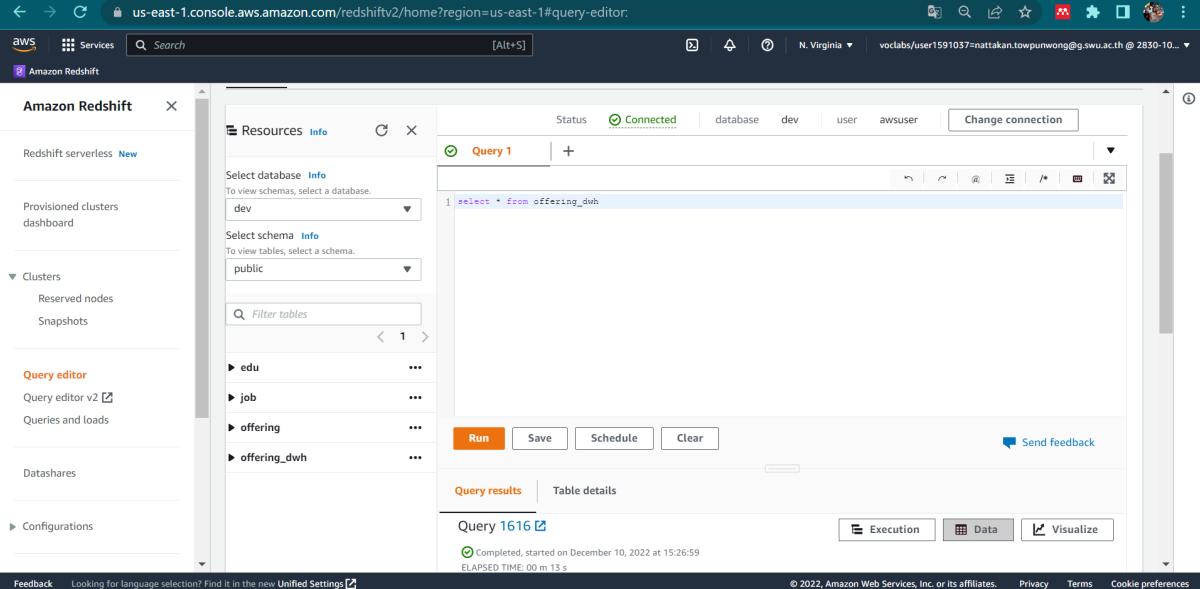
docker-compose up

port 8080

- กดรัน task บน Airflow



- ทดสอบ query ข้อมูล บน Amazon Redshift และ Download เป็น CSV



The screenshot shows the Amazon Redshift Query Editor interface. On the left is a sidebar with navigation links like 'Amazon Redshift', 'Redshift serverless', 'Provisioned clusters dashboard', 'Clusters', 'Query editor v2', 'Data and loads', 'Dashboards', and 'Configurations'. The main area has tabs for 'Resources' and 'Info'. Under 'Info', there are dropdowns for 'Select database' (set to 'dev') and 'Select schema' (set to 'public'). A search bar 'Filter tables' is present. Below these are tree views for 'edu', 'job', 'offering', and 'offering\_dwh'. A central panel displays a query editor with the SQL command 'select \* from offering\_dwh'. Below the editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. To the right of the editor is a 'Query results' section showing 'Query 1616' completed successfully. It includes tabs for 'Execution', 'Data', and 'Visualize'. The 'Data' tab shows a table with 45195 rows, with columns including age, job\_id, job, edu\_id, education, contact, month, and day. The table content is as follows:

| age | job_id | job           | edu_id | education | contact   | month | day |
|-----|--------|---------------|--------|-----------|-----------|-------|-----|
| 47  | 2      | blue-collar   | 2      | secondary | telephone | jul   | 28  |
| 58  | 2      | blue-collar   | 2      | secondary | cellular  | jul   | 7   |
| 25  | 2      | blue-collar   | 2      | secondary | cellular  | jul   | 7   |
| 34  | 8      | services      | 2      | secondary | cellular  | jul   | 15  |
| 48  | 10     | technician    | 2      | secondary | cellular  | jul   | 16  |
| 25  | 7      | self-employed | 3      | tertiary  | cellular  | jul   | 17  |
| 39  | 5      | management    | 2      | secondary | cellular  | jul   | 24  |

## 4. Data Visualization

เมื่อได้ข้อมูลที่พร้อมนำมาใช้งานตามโครงสร้างที่ได้ออกแบบแล้ว ได้ทำงานการนำข้อมูลดังกล่าวไปใช้งานต่อ ในส่วนของ Data Visualization เพื่อให้สามารถนำไปทำกราฟที่ทำให้เข้าใจข้อมูลได้ง่ายขึ้น โดยทำการสร้าง Dashboard จาก Power Bi มีการนำเข้าไฟล์ csv ที่ได้จาก Amazon Redshift

<https://app.powerbi.com/view?r=eyJrljoiYjlMOTIwNmEtOTM2My00ZjJiLWJiZGMtOTVINmM3M2MxMmNhliwidCI6ImY5MGM0NjQ3LTg4NmYtNGI0Yy1iMmViLTU1NWRmOWVjNGU4MSIsImMiOjEwfQ%3D%3D&pageName=ReportSection>



จากการ Dashboard สรุปได้ว่า จำนวนลูกค้าที่สนใจสมัครมีจำนวนทั้งสิ้น 5290 คน ใช้เวลาเฉลี่ยในการติดต่อขายประมาณ 4 นาที สายงานที่มีการสมัครสูงที่สุด คือ Management จำนวน 1,301 คน รองลงมาเป็น Technician จำนวน 840 คน โดยอายุของผู้สมัครสูงสุดอยู่ระหว่าง 30-39 ปี โดยสูงที่สุดคือ อายุ 32 ปี จำนวน 221 คน และระดับการศึกษาที่สมัครสูงที่สุดคือ Secondary จำนวน 2,450 คน โดยหากวิเคราะห์ตามรายเดือน เดือน พฤษภาคม จะมีผู้สมัครในจำนวนที่สูง จำนวน 925 คน รองลงมาเป็นสิงหาคม จำนวน 688 คน และสมัครสูงสุดในวันที่ 30 ของแต่ละเดือน จำนวน 271 ชั่วโมงระหว่างเดือน ปริมาณการสมัครใกล้เคียงกัน แต่จะมีช่วงก่อนวันที่ 30 ปริมาณการสมัครจะต่ำที่สุด