

# ROAD TO DATA ENGINEER 2.0

# WORKBOOK



CHAPTER 1

# *Data Collection & Basic SQL*

# ยินดีต้อนรับ สู่บทที่ 1

ในบทนี้ เราจะไปเรียนรู้พร้อมกันกับน้องໂຮດดีว่า:

- SQL คืออะไร ทำไมต้องเรียน
- คำสั่งพื้นฐานของ SQL
- การ Optimize SQL ให้เร็วขึ้น
- การสร้าง Data Pipeline และ ETL / ELT
- ประเภทของการประมวลผลข้อมูล
- การทำ Data Integration และวิธีแก้ปัญหา

แล้วเจอกันในบทเรียนครับ / ค่ะ



แอดเพิร์ด



แอดฟบ

# สิ่งที่ต้องทำก่อน ทำ Workbook นี้

- ดูวีดีโอ Class 2 ซึ่งจะสอน
  - Basic SQL
  - Chapter 1 Data Collection
  - Workshop 1
- (ไม่บังคับ) ถ้าคุณอยากรู้ SQL มากขึ้น
  - ดูวีดีโอบอกเรียน "โจทย์ SQL เพิ่มเติม" และบกเรียน "Walkthrough เฉลยโจทย์ SQL เพิ่มเติม"
- (ไม่บังคับ) ถ้าคุณอยากรู้ Pandas มากขึ้น
  - อ่านบกเรียน "โจทย์ Pandas เพิ่มเติม" และ "เฉลยโจทย์ Pandas เพิ่มเติม"



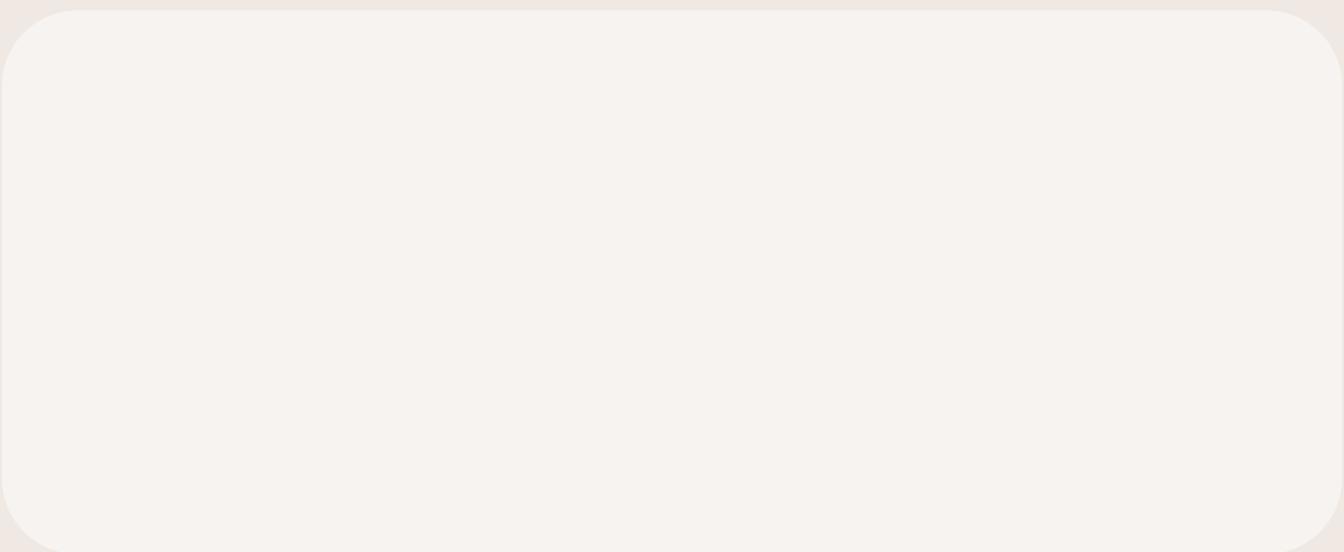
ถ้าเสร็จด้านบนแล้ว  
มาทำ Workbook นี้กันเลย >>



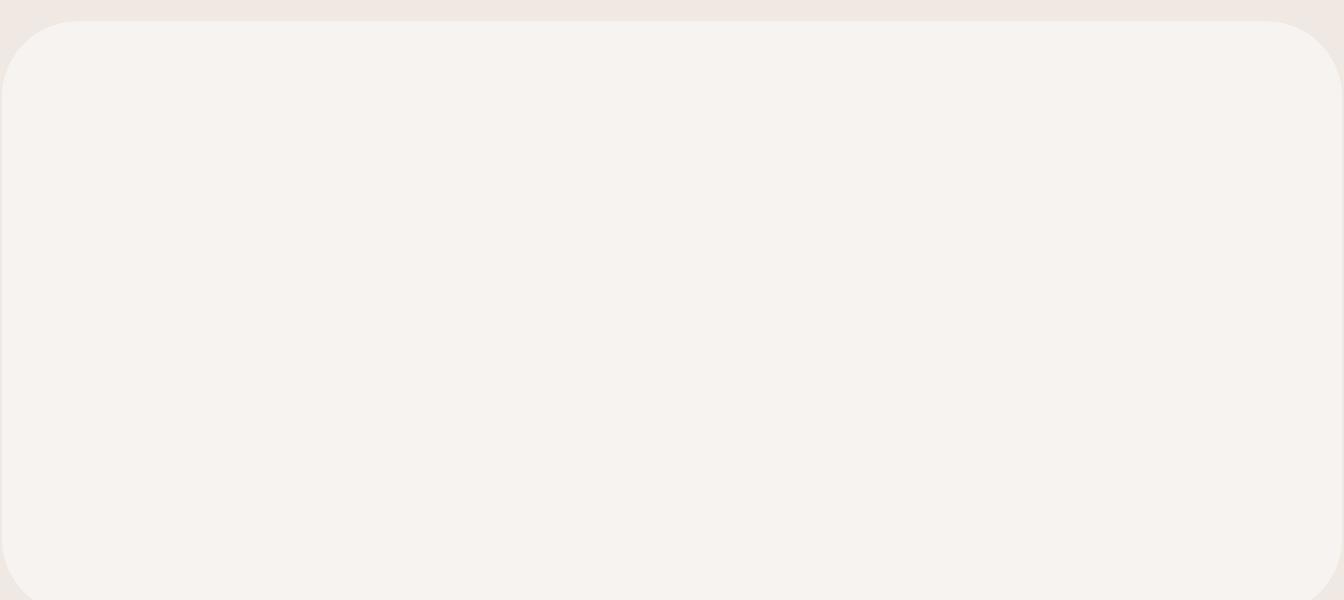


# Basic SQL (1)

SQL คือ



สาเหตุที่คนทำงานสาย Data ต้องเรียน SQL คือ



# Basic SQL (2)

ແດວແນວນອນ ເຮັດວຽກ

student_id	name	birthday	weight	height
1	David	2/1/1990	75	175
2	John	4/30/1989	67	169
3	Mary	6/22/1993	58	171
4	Jane	8/10/1990	60	153

ແດວແນວຕັ້ງ ເຮັດວຽກ

ຈາກຕາරາງດ້ານບັນ

ດ້າຕ້ອງການເລື່ອກນັກເຮັດວຽນທີ່ສ່ວນສູງເກີນ 170 ຕ້ອງເຂີຍນ SQL ວ່າ

ຈາກຕາරາງດ້ານບັນ

ດ້າຕ້ອງການກຣາບຈຳນວນນັກເຮັດວຽນທີ່ສ່ວນສູງເກີນ 170 ຕ້ອງເຂີຍນ SQL ວ່າ

# Basic SQL (3)

Table 1: students

student_id	name	birthday	weight	height
1	David	2/1/1990	75	175
2	John	4/30/1989	67	169
3	Mary	6/22/1993	58	171
4	Jane	8/10/1990	60	153

Table 2: scores

score_id	student_id	subject	score
1	1	Maths	70
2	1	Computer	85
3	1	Science	74

จากตารางกั้งสองด้านบน

ถ้าต้องการแสดง ชื่อนักเรียน, ชื่อวิชา, คะแนนวิชานั้น  
ต้องเขียน SQL ว่า

Exclusive Left Join กับ Inclusive Left Join ต่างกันที่

# Basic SQL (4)

เราสามารถโหลดข้อมูลจากไฟล์ CSV เข้าไปในฐานข้อมูลได้มั้ย



ได้



ไม่ได้



DDL ต่างกับ DML ที่

DDL

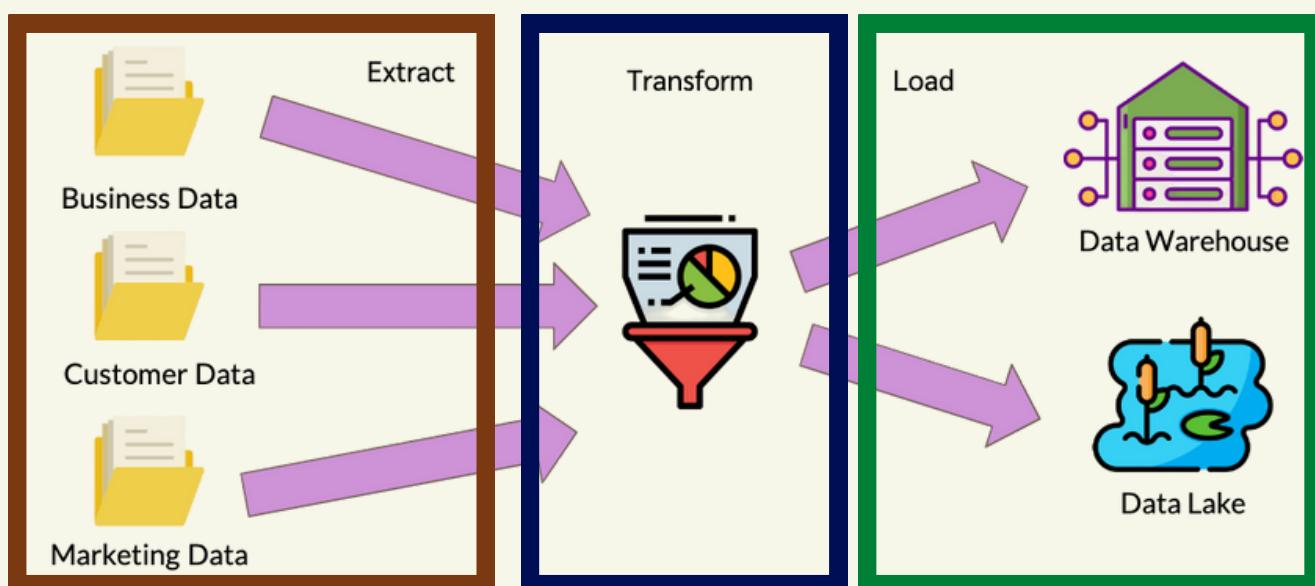
DML

วิธีการทำงาน และประโยชน์ของการทำ Index คือ

# Data Pipeline (1)

Data Pipeline คือ

ขั้นตอน Extract มีหน้าที่ \_\_\_\_\_



ขั้นตอน Transform มีหน้าที่ \_\_\_\_\_

ขั้นตอน Load มีหน้าที่ \_\_\_\_\_

# Data Pipeline (2)



การสร้าง Data Pipeline ที่ดี ควรคำนึงถึง

Initial Load กับ Incremental Load ต่างกันที่

*Initial Load*

*Incremental Load*

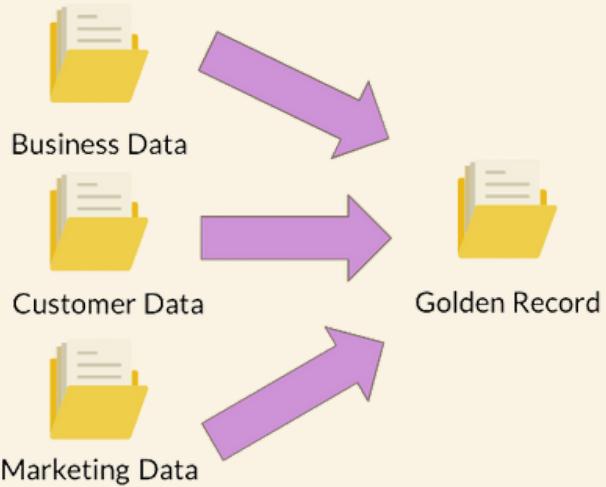
Batch กับ Streaming ต่างกันที่

*Batch*

*Streaming*

# Data Integration (1)

Data Integration คือ



Customer 360 คือ

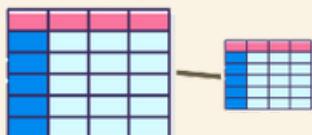


ข้อมูลที่เราใช้ในการวิเคราะห์ สามารถมาจาก...



# Data Integration (2)

ปัญหาในการทำ Schema Integration คือ



Client ID  
=  
Customer ID

1.6 Kilometer  
=  
1 Miles

ปัญหาในการทำ Data Integration คือ

Mister Jonathan  
=  
Mr. Jon

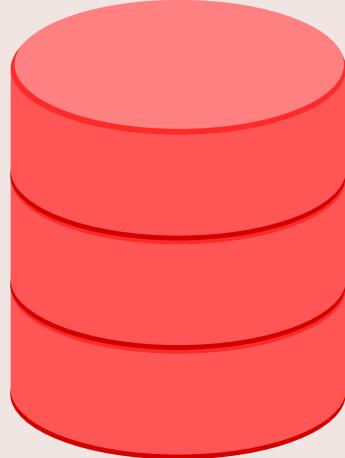


ไปลุย Workshop 1 กันเลย!



# Workshop 1 (1)

แพ็คเกจ PyMySQL uu Python ใช้สำหรับ

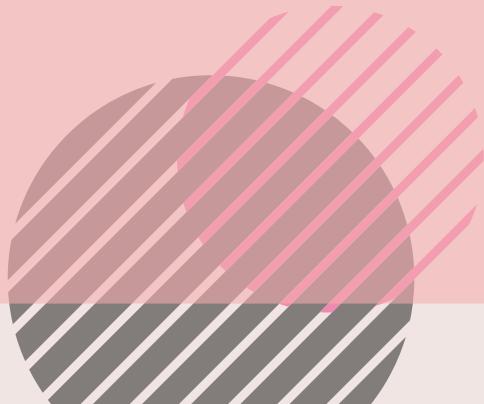


เราไม่ควรใส่พาสเวิร์ดตรง ๆ ในโค้ด เพราะ



\* \* \* \*

แพ็คเกจ Pandas ใช้สำหรับ





# Workshop 1 (2)

Pandas รองรับการใช้ภาษา SQL ดึงข้อมูลมั้ย

- รองรับ
- ไม่รองรับ

คำสั่ง Pandas ต่อไปนี้ ใช้สำหรับ

`read_sql`

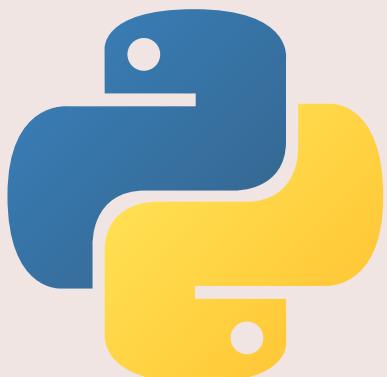
`merge`

`rename`

`apply`

`to_csv`

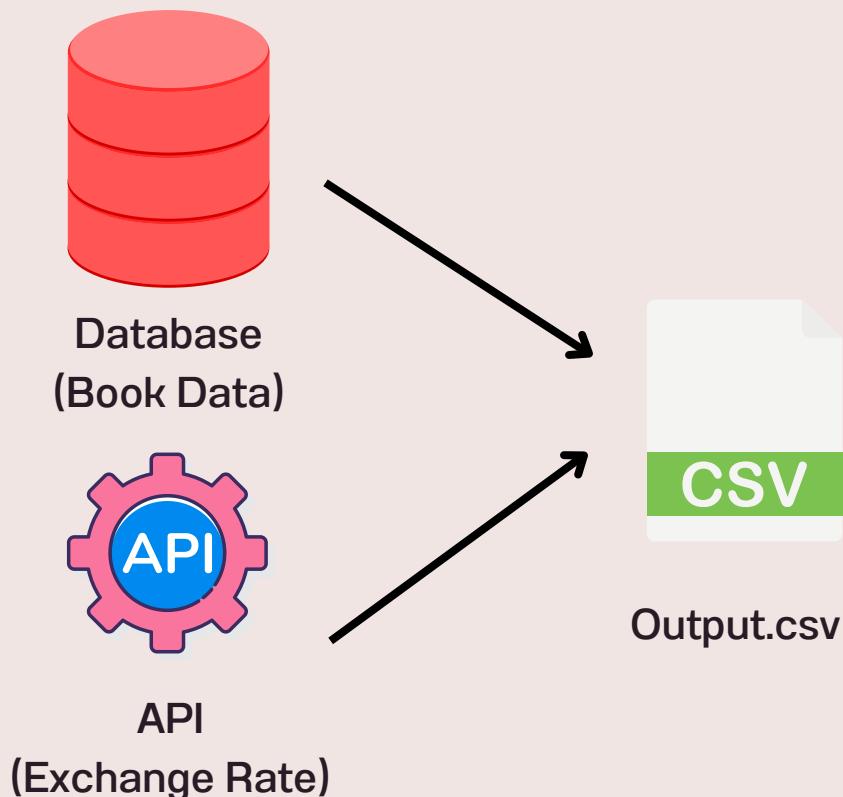
แพ็คเกจ Requests ใช้สำหรับ



# Workshop 1 (3)

< หน้านี้สำหรับจด Note ได้อิสระ >

สรุปสิ่งที่เราทำใน Workshop 1 คือ



# ยินดีด้วย คุณจบบทที่ 1 แล้ว!



คุณพร้อมไปต่อบทที่ 2 เรื่อง Data Cleansing และ<sup>+</sup>  
ในบทต่อไปเราจะเรียนรู้เรื่อง Data Quality และ Distributed  
Processing (Hadoop / Spark) กัน สนุกแน่นอนนะ



มาพ邪乍ນໄປຄ້ວຍກັນນະ!

Recap สิ่งที่เราได้เรียนรู้ จากบทที่ 1:

- SQL คืออะไร ทำไมต้องเรียน
- คำสั่งพื้นฐานของ SQL
- การ Optimize SQL ให้เร็วขึ้น
- การสร้าง Data Pipeline และ ETL / ELT
- ประเภทของการประมวลผลข้อมูล
- การทำ Data Integration และวิธีแก้ปัญหา