

## Interview Challenge (Data Engineer) - 4 days

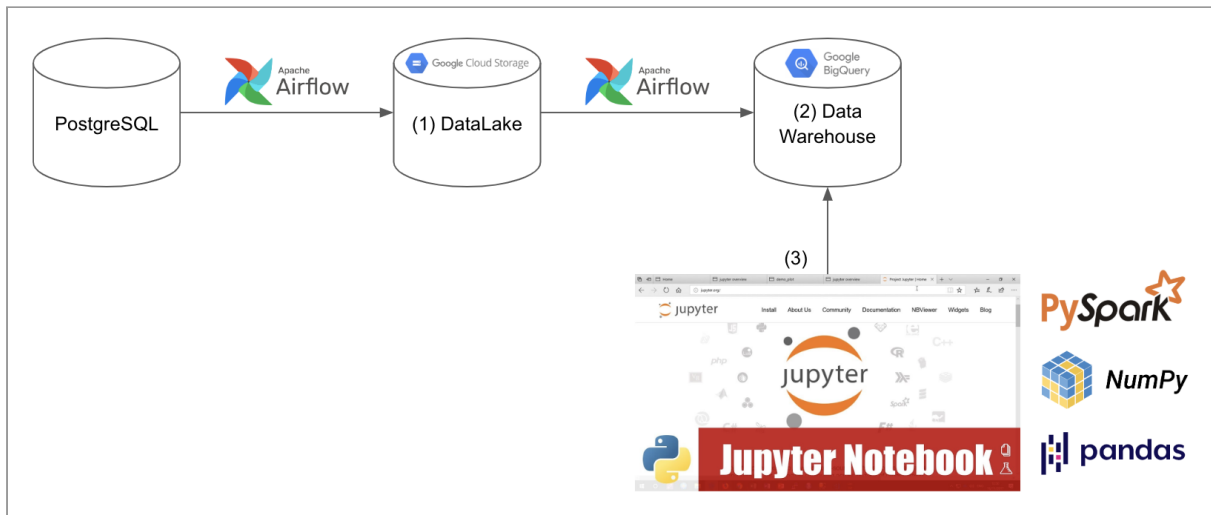
### เกณฑ์การประเมิน

- การทำงานได้ครบถ้วนตาม requirement
- ความง่ายและยืดหยุ่นในการปรับเปลี่ยนค่าต่างๆ (config file, environment variable)
- ความง่ายในการแก้ไขเปลี่ยนแปลง และเพิ่มลด feature ในภายหลัง
- การทำงานได้โดยไม่เกิดความผิดพลาด
- การรักษาความปลอดภัยของข้อมูล
- คุณภาพของ Architecture ที่ออกแบบมา
- คุณภาพของ source code (project structure, design pattern, solid principle)
- การเขียน document
- การเขียน readme
- ความสามารถในการเรียนรู้เครื่องมือและเทคนิคใหม่ในเวลาจำกัด
- ความเข้าใจในเครื่องมือและเทคนิคที่เลือกมาใช้งาน

### extra point

- Docker
- Kubernetes
- Monitor tools for Airflow Job, Jupyter healthcheck
- Apache Kafka or RabbitMQ

### โจทย์



### ฐานข้อมูลตั้งต้น (PostgreSQL 11)

host: 35.247.174.171:5432

user: exam

pass: bluePiExam

db: postgres

## ทำบน Google cloud platform ใน Project ที่เตรียมไว้ให้เท่านั้น

1. สร้าง Data Lake ด้วย Google Storage  
ติดตั้ง Apache Airflow เพื่อทำ scheduled job โดยให้ทำงานทุก 1 ชั่วโมง  
สำหรับ Clone ข้อมูลจาก ฐานข้อมูลตั้งต้น (users table, user\_log table)  
ไปเก็บแบบ Raw data ที่ Data Lake
2. สร้าง Data Warehouse แล้วเชื่อมต่อเข้ากับ Google Big Query  
เพิ่ม job ใน Apache Airflow เพื่อทำการ Transform ข้อมูลจาก Data Lake แล้วเก็บข้อมูลที่  
แปลงแล้วที่ Data Warehouse โดยให้งานต่อหลังจาก job load ข้อมูลเข้า data lake เสร็จสิ้น  
  
โดยสิ่งที่ต้อง Transform คือ Table ของ user\_log  
ให้เปลี่ยนชื่อ column จาก status เป็น success และเก็บข้อมูลเป็นชนิด Boolean  
ทำการแปลงข้อมูล 0 เป็นค่า FALSE และแปลงข้อมูล 1 เป็นค่า TRUE
3. ติดตั้ง Apache Spark  
ติดตั้ง Jupyter Notebook สำหรับเขียนโค้ด Python3  
ติดตั้ง library PySpark, Pandas, Numpy
4. เขียนโค้ด Python3 ที่ Jupyter Notebook โดยใช้ PySpark เรียกใช้งาน BigQuery เพื่ออ่าน  
ข้อมูลจาก Data Warehouse สำหรับเป็นตัวอย่างการใช้งานให้แก่ Data Scientist
5. ส่งมอบ
  - file code ทั้งหมดที่เขียนผ่าน Git
  - document อธิบายว่าใช้ server ตัวไหนบ้าง มีการติดตั้งหรือเก็บไฟล์ไว้ที่ใดบ้าง
  - ลิงค์สำหรับเข้าใช้งาน Jupyter Notebook