**AWS**

**AWS S3** (load data to data lake by creating buckets -> upload files) -> **Glue** (manage structured data) / (**EMR spark cluster with EC2-Create key pairs -> EMR create cluster to apply py code-> S3)** -> **Redshift** (data warehouse by creating buckets -> connect to dataset -> use query to create empty table -> load data to tables from S3) -> **SageMaker** (Launch app –> import dataset)

1. Nodes and cluster: A set of nodes called an Amazon Redshift cluster, which consists of a leader node and one or more compute nodes.
2. Snapshots: Snapshots are point-in-time backups of a cluster.
3. IAM role, credential: AWS associate cluster with IAM role, use that role’s credential to access to S3 data
4. Database name: dev, Database user: awsuser
5. EMR Elastic MapReduce: run tasks in parallel

Q: what is IAM role, what is Labrole

Q: Do I be familiar with how to write query to load data in Redshift

Q: What is the difference of studio and Canvas in SageMaker’s Launch app

**Google's BigQuery**

SQL Fundamentals - Common Table Expressions (CTEs)

* A temporary named resulted
* Ex: With CTE\_NAME as (Select customer\_id….)

-> Select …. where customer\_id in CTE\_NAME

Raw data ->(DBT)-> Transformed

Modern data pipeline: DBT, Looker, BigQuery, Snowflake

DBT/Fivetran Certificate

Window functions

1. 從fivetran裡面讀取公開的s3檔案( <- flightviews raw data)存到我的googlesheet的Google's BigQuery裡面
2. Google sheet 裡面建一個 credential ，把這個credential的key下載成json檔讓dbt有權限讀取
3. 在DBT開一個brench後，用staging file清乾淨 ( .sql用來讀取資料後用 .yml清理資料)後
4. 再用 .sql把清理好的表整合到marts裡面，輸出 (cleaning and transforming should be separated by structure design in dbt)
5. Deploy -> Environment -> create environment -> add job to automize what dbt shall run on schedule
6. Now we can use looker to visualize data in Google BigQuery

E(Extract)L(Loading)T(Transformation)

Q1: why dbt test/run the sql file

Q2: what is “version two” in yml file