**Project youtube**

**1. Create AWS Account**

**2. Create IAM for admin**

- grant AdministratorAccess.

**3. Create S3 Bucket and upload raw data into S3**

- Create S3 bucket: de-on-youtube-raw-apsoutheast1-alivebook.

- Upload data with format "\*.json" into raw\_statistics.

A screenshot of a computer

Description automatically generated

- Upload data with format "\*.csv" into raw\_statistics\_reference\_data and dividing by region.

A screenshot of a computer

Description automatically generated

**4. Build Glue Crawler and catalog file json**

- From s3 (de-on-youtube-raw-apsoutheast1-alivebook/youtube/raw\_statistics) to db: de-on-youtube-clean

- Create s3 (de-on-youtube-raw-apsoutheast1-alivebook-athena)

- In athena, manage location (choosing s3: de-on-youtube-raw-apsoutheast1-alivebook-athena) and query.

- Query => Error -> convert json to parquet file.

**5. Create lambda function to convert json file to parquet file and automatically crawler**

- Code.

- Set environtment variable.

"s3\_cleansed\_layer":"s3://de-on-youtube-clean-apsoutheast1-alivebook/youtube",

"glue\_catalog\_db\_name":"cleand\_statistics\_reference\_data",

"glue\_catalog\_table\_name":"de-on-youtube-clean",

"write\_data\_operation":"append"

- Test.

- Add trigger to when upload data into S3, it will automatically convert json to parquet and craw data into db de-on-youtube-clean (table: cleand\_statistics\_reference\_data).

(from S3: de-on-youtube-raw-apsoutheast1-alivebook, Prefix: youtube/raw\_statistics\_reference\_data/, Suffix: .json)

A picture containing text, screenshot, software, computer icon

Description automatically generated

A screenshot of a computer

Description automatically generated

A picture containing text, software, computer icon, web page

Description automatically generated

**6. Build ETL job to convert csv file to parquet file**

- From S3: de-on-youtube-raw-apsoutheast1-alivebook /youtube/raw\_statistics/

- Transform datatype.

- To S3: de-on-youtube-clean-apsoutheast1-alivebook/youtube/raw\_statistics/

- PartitionKeys: region.

- Run ETL.

**7. Build Glue Crawler and catalog file**

- From s3 (de-on-youtube-clean-apsoutheast1-alivebook/youtube/raw\_statistics/) to db: de-on-youtube-clean.

A screenshot of a computer

Description automatically generated

**8. Build ETL job**

- From db: de-on-youtube-clean/ table: raw\_statistics and db: de-on-youtube-clean/ table: statistics\_reference\_data.

- Transfrom datatype bigint for category\_id and id.

- Inner join with condition: category\_id=id.

- To S3: de-on-youtube-finish-apsoutheast1-alivebook with format Parquet.

- Run ETL.

A screenshot of a computer

Description automatically generated  
A picture containing text, software, computer icon, web page

Description automatically generated