**Read Me**

**Run etl.py**

1. The data files are hosted on G-Drive. The code fetching data in chunks and saving locally to multiple csv files in case if it’s a big csv file.
2. Three DDLs are created for three tables.

**Postgres DB**

1. Connection to Postgres database is made and tables are created with the method create\_table() using the DDLs.
2. SQL-alchemy engine with Postgres connection credentials is created for loading.
3. The data from the files are read and transformed before loading to the Postgres db.

**Note: LocalStack, an AWS cloud emulator has been used instead of AWS.**

**AWS S3**

1. Batch transfer to AWS S3 is made in the bucket named travel-bucket.

**AWS Lambda**

1. A zip file name lambda.zip with the script lambda.py has been created.
2. In terminal, we’d execute the following command using aws-cli in the same directory of zip file.
3. We’d use AWS EventBridge to schedule Lambda invocation.

aws lambda create-function --function-name myLambdaFunction \

--runtime python3.10 --handler lambda\_function.lambda\_handler \

--role arn:aws:iam::xxxxxxxxxxxx:role/service-role/my-lambda-role \

--zip-file fileb://lambda.zip

**CIA (Confidentiality, Integrity and Availability)**

## **Data Confidentiality**

### Implement Data Encryption- To safeguard data during transit and at base, data encryption using strong encryption algorithms could be implemented. AWS offers built-in encryption features that can be leveraged for enhanced security. We could use on sensitive fields like passwords, mobile, address, e-mail.

### Role-Based Access Control (RBAC)- RBAC is a method of restricting network access based on the roles of individual users within an enterprise. AWS IAM service provides RBAC model to assign roles to grant access to its resources.

### Data Masking - We can replace sensitive data with realistic but fictitious data, to ensure confidentiality of the original data while sharing data between testing and development.

### Secure Credentials Management – AWS provides credentials in form of database passwords, API keys and Tokens. We’d use secure storage mechanisms, like environment variables or credential management systems instead of hard-coding credentials in ETL scripts.

## **Data Integrity**

### Physical Integrity - Physical integrity is jeopardized in case of natural calamities or power outage disrupting database functionality. AWS provides a solution to by creating a copy of our database across multiple Availability Zones.

1. Logical Integrity - Logical integrity protects data from human mistakes and hackers.
2. **Entity Integrity -** It ensures no duplicates and that no field in a database is null. We have used SQL keywords like ‘NOT NULL’, ‘UNIQUE’ to avoid null values and duplicates, wherever needed.
3. **Domain Integrity -** It ensures that each piece of data in a domain is accurate. We have used constraints while creating the DDL for the tables by limiting the format, kind, and amount of data to enter. Keywords like ‘INT’,’TIMESTAMP’,’FLOAT’,’BOOLEAN’ can be used. Data validations could also be done before it is allowed into any data storage system.

## **Data Availability**

1. Cloud Solutions – Cloud providers like AWS comes with multiple availability zones within a region itself providing preventive measures such as redundancy for high availability.
2. Workflow management tool – Apache Airflow is tool to manage ETL pipelines. In case of database usage throughout the day for dashboarding and reporting purposes, one can schedule the pipeline to run overnight to avoid overloading and ensuring faster data processing. Logs and errors are automatically saved and made available for debugging and tracking purposes in Airflow.
3. Optimization – One could use few optimization tricks to ensure faster data retrieval and easy accessibility. Tricks like breaking down complex queries into smaller ones to find a good trade-off between speed and maintenance. Another trick would be to use the correct joins to minimize table movements, along with fetching the required columns and avoiding unnecessary fields for faster retrieval.