Tech Challenge - Senior Data Engineer

Welcome to your Tech Challenge!

Hey Data Engineer. Welcome. Your mission is to implement a data processing strategy which is relevant at orderbird.

The Data (see repository)

You are given sample data files consisting of invoice data of various orderbird customers. A row consists of

- venue_id (string)
- invoice_id (string)
- datetime_created (string)
- total (double)

Where:

- venue_id is an unique outlet identifier
- invoice_id is a UUID of an invoice
- datetime_created denotes the timestamp when the invoice was created
- total is the invoice amount in EURO

The date in a filename indicates that the invoice data was synced at that specific date to the backend system. It is important to notice, however, that an invoice could be created at any time in the past before the sync date!

The Task:

Your task is to implement a strategy to efficiently update a table holding aggregated venues' totals.

Technical requirements:

- The ETL-job is implemented in the Apache Airflow framework running inside a Docker container.
- The ETL-job awaits a csv-file (in the format described above) in a mounted folder.

orderbird GmbH
Ritterstraße 12
10969 Berlin
Deutschland

www.orderbird.com
E-Mail: hello@orderbird.com
Telefon: +49 30 208983099
Fax: +49 32121468189

Geschäftsführer: Mark Schoen
und Jakob Schreyer

- when a new data file is detected then its content is inserted into a RDS (running inside the Docker environment as well).

Here comes the challenging part:

- After successfully inserting the data then a second table holding aggregated total amounts per venue and per date of creation is updated.
- suppose that the invoice table was huuuuuge. Can you think of a non brute-force approach to efficiently update that aggregation table?

Please send us a zipped file containing your docker project together with instructions on how to use it. Please do not create a pull request or fork this repository as your solution should not end up being public afterwards.

Link: https://github.com/orderbird/data-engineer-challenge

Good luck and happy coding!