Algolia - Data Engineer Assignment

The Integration team has deployed a cron job to dump a CSV file containing all the new Shopify configurations daily at 2 AM UTC.

The task will be to build a daily pipeline 3 steps that will :

- Extract :
  - Download the CSV file from the s3 bucket `alg-data-public`, the name of the files `[YYYY-MM-DD].csv` (replace `[YYYY-MM-DD]` by the date),
- Transform :
  - Filter out each row with empty `application_id`
  - Add a `has_specific_prefix` column set to `true` if the value of `index_prefix` differs from `shopify_` else to `false`
- Load :
  - Load the valid rows to a PostgreSQL instance

The pipeline should process files from 2019-04-01 to 2019-04-07.

This pipeline **should be runnable easily using** `docker` and `docker-compose`.

Note that you can leverage/extend/improve the following docker-compose and Dockerfile if that may be helpful: Algolia Data Engineer Assignment . Those are heavily inspired by the official codebase which we recommend you to review.

This pipeline is relatively simple on purpose because we want you to concentrate on delivering an assignment as close as possible to something we could put in production. Even if it seems overkill in this context, your architecture choices should make it possible to scale in higher data volumes.

You will be evaluated on the implementation of the solution and whether it works. But also strongly on whether the code is production-ready.

Hence why you should keep in mind the following :

- The code must be available on a Github repo
- Using Python is mandatory
- Writing Python unit tests is mandatory
- Using Airflow as the orchestration tool is mandatory
- Writing documentation and a detailed readme is mandatory
- The quality of the code will be assessed

Finally, we prefer the candidate to take a few more days to polish the code instead of rushing it to ship fast.