Test task for Data Platform Engineer

Background

This task should not take more than few hours to complete by someone who has good working knowledge of any data processing system. It may take slightly more if used system is new and some learning is required. In any case you should not spend more than 8 hours on this task.

Writing solid production-grade solution that will work in various conditions can be very demanding task. It's OK to cut corners in test task and point out shortcomings of offered solutions or assumptions made.

Data and questions

For this task we are using pre-generated Deals dataset from Amazon S3 (s3://pdw-export.alpha/test_tasks/deals.csv.gz).

S3 access credentials:

- **Access key ID** - AKIAI7LDULX7AFF2VNYQ
- **Secret access key** - 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

Do the following:

- Register trial Pipedrive account (https://www.pipedrive.com)
- Write a small program to download dataset from S3 and create/update deals via Pipedrive API in Pipedrive account (https://developers.pipedrive.com/docs/api/v1)
- Apply simple transformation logic on each row (like multiply "value" field by 2).
- Think of it as it is regular daily job to keep these two endpoints in sync.

Requirements:

- Use Python

Additional questions:

- Think about the performance and be prepared to discuss on it.
    - What happens if there will be 100K, 1M deals to synchronize
- How easy is it to change source (for example, to URL) and destination (for example, to Postgres) and how to make this data loading more generic.