**Documentation**

This document is designed to give overall steps and details of the code repository, instructions to run the spark application created as part of the HelloFresh Data Engineering Test.

**Folder Structure for the code:**

Graphical user interface, text, application

Description automatically generated

**etl\_job.py:**

This Python module contains an example Apache Spark ETL job definition

It can be submitted to a Spark cluster (or locally) using the 'spark-submit' command found in the '/bin' directory of all Spark distributions (necessary for running any Spark job, locally or otherwise).

**Spark\_submit.txt:**

This file contains all the steps on how to submit the spark application.

My spark application consists of:

1. Modularized code for extract, transform and load operation
2. Exception Handing
3. Logging Level
4. Defined abstractions
5. Comments on each functions hence Readable

Config management:

CI/CD pipeline can be configured using below Applications:

Graphical user interface, application

Description automatically generated

**Orchestration:**

Apache Airflow or Oozie can be used to run the application periodically