Senior Data Engineer Test

1. Design and code a Spark Job that ingests 1 or multiple CSV files into DeltaLake

- The Job must be able to ingest files with and without header
- The Job must add 2 extra columns to the output DataFrame :
    - ingestion_tms : ingestion timestamp format : YYYY-MM-DD HH:mm:SS
    - batch_id : UUID v4
- The job must use APPEND write mode to atomically add new data to Delta table

2. Produce a Docker Compose YAML file to run the job from a Container

- Produce the Spark Job Dockerfile
- Add SparkHistoryServer container to the service:
    - Image : gcr.io/spark-operator/spark:v2.4.0
    - Command :
        - /sbin/tini
        - -s
        - --
        - - /opt/spark/bin/spark-class
        - -Dspark.history.fs.logDirectory=/[PATH TO LOG DIR]/
        - org.apache.spark.deploy.history.HistoryServer

3. Produce a production ready system diagram of your solution deployed to either Public Cloud provider (AWS/GCP/Azure) or Kubernetes.

- The system design must include job orchestration

Implicit requirements:

1. Development language – Python3 – PySpark 3.3.1
2. DeltaLake 1.2.1 (feel free to use any storage provider “local or cloud provided”)
3. The code produced by you is expected to be of high quality
4. The Spark job Logs/Traces must be persistent and accessible from SparkHistoryServer
5. The solution must have tests, runnable locally
6. Use common sense

Please put your work on GitHub or Bitbucket.