Data Engineering - Programming Exercise

This exercise is an exemplary scenario designed to test your data engineering skills. Solve it using the Yelp data set (https://www.kaggle.com/yelp-dataset/yelp-dataset) and PySpark/Spark. You may add other libraries, that you find suitable for the task.

As a data engineer, you should create a (small) data lake. Its requirements are:

- It should consist of raw, cleaned and aggregated Yelp data.
- All JSON files should be part of the raw data.
- Consider all raw files for cleaning.
- The aggregation should at least include the stars per business on a weekly basis and the number of checkins of a business compared to the overall star rating.
- Write a README including a short explanation of your design decisions and a section on how to run your code.
- Package your Spark application as a Docker container (or multiple containers).

Create a git bundle of your repository and send it back to the same email-address that send you this exercise once you are finished.

Good luck!