Backend Test Project

Objective

Create a streaming data processing pipeline which can test for anomalous data in real-time.

Details

Anomaly detection is an important part of data processing, bad quality data will lead to bad quality insights. As such the first part of our data processing does some basic anomaly detection.

As part of this test project you will create a simple anomaly detection pipeline in Apache Flink using the API it provides. The pipeline will read data from the provided files, do stream processing to allocate an anomalous score, and then write the data into InfluxDB. Both the original values and the anomalous score should be written to InfluxDB for each sensor reading.

The following dataset can be used for this project:
https://www.dropbox.com/s/3ww0xoitwkzaate/TestFile.zip?dl=0

Anomaly Detection Method

There are libraries which provide anomaly detection functionality, however many don’t work well for streaming data. The following algorithm can be used to give a score:

1. For a sliding window of values (100 values should give ok results)
2. Calculate the interquartile range (IQR) for the array
3. Based on the IQR, score the value being processed with the following:
    a. If the value is < 1.5 * IQR, assign 0
    b. If it is >= 1.5 * IQR and < 3 * IQR, assign 0.5
    c. If it is >= 3 * IQR, assign 1

Constraints

The project should be provided in a Git repository such as on gitlab.com
It is expected it will be in Java using Maven to build the project
InfluxDB is available as a Docker image
Instructions on how to run the project should be provided