Data Science
Exercise X: Building a streaming pipeline with MQTT

Prof. Dr. Thomas Kopinski

November 3, 2022

Abstract

In this series of exercises you will become familiar with using feeds, Docker, MQTT and InfluxDB to build a streaming pipeline that grabs data from the web, streams it to a local broker based framework, calculates values on the incoming data to stream it to a database that is optimized for timeseries data. This Pipeline can be adapted for similar needs in an IoT environment.

Step 0: Setting up Docker

Docker is a containerization setup that lets you run packaged software on various systems (mostly) independent of the underlying operating system. It is a framework commonly used in production settings where reliability and transferability play a crucial role.

- go to the getting started page of Docker and click on the installation guide suitable for your system
- Install Docker and test the installation with the suggested methods

Step 1: Setting up the MQTT Broker

MQTT is a communication protocol designed for communication in networks. The main focus is streaming data with high possible latencies and unstable connections. It is based around a broker that handles the published messages from a client in so called topics. Other clients can subscribe to topics and receive messages whenever an event comes in.

- For the broker we will use the Eclipse Mosquitto MQTT broker in a dockerized environment for easy setup. The installation instructions can be found here
- Install and test your MQTT broker publishing and subscribing to a test topic. In a terminal type:
- mosquitto_sub -v -t 'test/topic'
- mosquitto_pub -t 'test/topic' -m 'helloWorld'

Step 2: Setting up InfluxDB

- install and test InfluxDB with the tutorial provided here

Step 3: Using python to stream data

- install the necessary python packages for this pipeline in a conda or virtual python environment. You can use the .env file provided in the git repository of the course.
- Download the weather data with the python api

- test the paho mqtt python client by connecting to the locally running broker
- stream the weather data in a json serialized payload format to the weather topic
- open another notebook or python script for a subscriber
- subscribe to the weather topic to receive messages
- Extract the data from the json payload to fit into a pandas dataframe
- use the forecast times as the timestamp and index for the dataframe
- write the transformed data into the influxdb to the weather bucket with the _measurement tag set to the time when the forecast was requested

Step 4: Read data from the influxDB and plot it

- use a Flux query to access all the data from your weather bucket
- use a pandas multi-index dataframe to store the data properly
- plot the weather forecasts for the next 2 weeks