SOFT8026 – Data-driven Microservices – Assignment 1

**Due:** Sunday 21st March at 23:59

**Worth:** 50%

# Overview

Thus far, we have covered the concept of microservices, bounded contexts and various architectural and integration concepts. We have also covered tools and technologies that can help us build and deploy microservices (see lab documents 1 to 4). This assignment requires that you analyse a requirement and devise an appropriate architecture and implementation.

# Problem Description

You are to analyse a stream of data from Reddit. You will simulate the stream of data by reading posts from a data file and serving them through a stream channel to an analytics client. The client should analyse each post as it arrives and calculate 4 different types of metric / result:

1. an aggregate metric, such as a total or average

2. a rolling 3-minute metric, e.g. sentiment within the last 3 minutes

3. a single post which is the most or least of some criteria

4. anything else of your choice that is not based on the post title

There should be a single web page that displays those metrics / results whenever the page is loaded or refreshed. It is up to you how and where the web page gets the data.

It is recommended that on average 2 posts per second should be streamed, though there should be some random variability.

# The Posts

The file r\_dataisbeautiful\_posts.csv has more than 190,000 posts in CSV format. All the posts are from the “Data is Beautiful” subreddit. The file was downloaded from <https://www.kaggle.com/unanimad/dataisbeautiful> and has the following comma separated fields:

*Id, title, score, author, author\_flair\_text, removed\_by, total\_awards\_received, awarders, created\_utc, full\_link, num\_comments, over\_18*

There is no post text or comments associated with the posts, just the post title.

# The Microservices

There are likely to be at least 3 microservices in your architecture, e.g. one that reads the posts and streams them to a client connection, the client microservice that does the analytics, and a web server microservice that serves out that single web page. There may also be a data storage microservice.

# Technical Considerations

Use gRPC as the communication mechanism and use a stream channel to send the posts to the client.

Flask is recommended for the web server and web page. A simple table is all that is required. You could consider using a charting library or you could improvise by using hash symbols to build up a bar chart, for example. You have a lot of freedom here.

Because a sleep is placed between each tweet read, it should be possible to keep refreshing the web page (either automatically or manually) to see updated metrics / results.

Use Docker (with Dockerfiles) for each microservice and use Docker Compose to orchestrate your system. It should be possible to simply enter “docker-compose up” to bring your system up.

# What to Submit

* There will be a form to fill in (Word document) where you will explain your architecture and the metrics / results you chose to calculate. You will include sample output and complete a checklist to indicate which of the requirements you have completed. *This form will be uploaded in the next few days.*
* You will submit your project in zip format – exclude the dataset, but indicate in the form where I should place the CSV file in your directory structure.