Extended Abstract: cloud-based document analysis application

Florian Bauer
School of Computer Science
University of Bristol
Bristol, UK
ya18048@bristol.ac.uk

Nathalie Pett
School of Computer Science
University of Bristol
Bristol, UK
aq18034@bristol.ac.uk

***Abstract*—This document describes the idea, choice of technology, archtitecture, scalability and security measure we intend to implement as part of our *COMSM0010 Cloud Computing* coursework project.**

I. Idea

The application that we intend to develop and deploy in the cloud as part of the coursework for the module *COMSM0010 Cloud Computing* aims to support the analysis of PDF text documents.

As university students we are often faced with an abundance of resources regarding specific units or even certain topics within a unit. These range from lecture notes or slides to personal notes and additional scientific papers as well as e-books or extracts thereof. The first step in the exploration process in this case is to familiarise ourselves with these materials, that is to find key aspects and how different sources are possibly related. This is what we are hoping to facilitate with our application.

More specifically the functionalities of our application include keyword search and tagging, suggestions of related documents based on textual analysis and eventually the automated generation of short summaries from related documents. The functionality could be extended by the possibility to share insights with other users of the application, as they might have added additional sources for the same class or topic to their account. However, in the scope of this project the main focus is on the deployment of our application in the cloud. Therefore we might not implement all of the aforementioned features and will not focus on optimizing the results of the analysis, but rather develop a prototype application with the potential for further development and extensions.

The very rough proposed process for using the application is as follows: This document describes the idea, choice of technology, archtitecture, scalability and security measure we intend to implement as part of our *COMSM0010 Cloud Computing* coursework project. After logging into the application the user can upload one or more documents. These documents are then processed and stored by our application. By applying OCR techniques the application ensures that the documents are machine readable. The documents will then be analysed and the user will be presented with the outcome of this analysis.

Lastly, the general idea of this application is not limited to the use case of university students. Any scenario where people are confronted with a large number of different and possibly complex text sources could make use of the application, e.g. in the context of information gathering for management decisions in the industry.

II. Choice of Technology Stack

In this section a preliminary technology stack is introduced. There might be changes to these current choices, when the development of the application progresses.

The backend of the web application will be written in Python, while React will be used for the frontend. For storing the uploaded PDF documents an object storage will be used. Other data, such as account and session information or the results of the analysis will be stored and managed through a Mongo DB. The application will run in Docker containers orchestrated by Kubernetes.

The necessary machine learning tasks for the analysis will be implemented using Keras on top of Tensorflow.

III. Architecture

The application uses a microservice architecture with the following four microservices: Frontend, Backend, Analysis and Login / Account.
Each microservice is containerised and all containers are integrated into a deployment on a Kubernetes cluster. All of the data associated with the application will be stored outside of the cluster for fault tolerance reasons.

IV. Scaling

Kubernetes is managing and monitoring the services and is responsible for scaling. Each service can be scaled individually by adding or removing instances.

Load balancing is used to distribute incoming traffic and requests over the cluster. As for the storage (database and object) scalable PaaS are used, it can easily be scaled out (more instances) and up (more resources assigned to an instance).

V. Security

Running an application in the cloud requires certain security measures, which are described in this section. Software Defined Networking is applied to the application and all its components. More specifically, firewalls are used to block ports, only allow specific traffic and prevent DDOS attacks on the cluster. Role based access control is applied to protect all resources of unwanted access.

Encryption is applied to the database and object storage (in rest and in transit). SSL and TLS is applied to encrypt network traffic.