Technical Solution

This section addresses solution for problems discussed in previous section

System overview

Data polish is a web based application primarily for Master’s and PhD student to clean datasets before performing any machine learning or deep learning techniques on their dataset using simple user interface.

Data import

In this page end use will be able to import their dataset for cleaning

It will be saved in cloud and they will get unique ID to track the progress.

A computer screen shot of a cloud

Description automatically generated

Data Profile

In this page end user will the data quality index and full profile of the dataset

A blue background with a square with black text

Description automatically generated with medium confidence

A screen shot of a graph

Description automatically generated

Data cleaning

In this page end user will be able to click the button and data cleaning will be

A blue and white screen

Description automatically generated

Export Data

In this page user will be able to download the cleaned data file

A computer screen shot of a computer screen

Description automatically generated

Basic Architecture

A diagram of a computer program

Description automatically generated

We're developing a modern web application using a stack that includes front-end technologies like Angular, Bootstrap, and Prime NG(framework) for the user interface and on the backend we're utilizing Java, Spring Boot, and Maven to build our API. Our data is stored in a MySQL database, and for file storage, we use Azure Storage Account's blob storage. For data quality and profiling, we've integrated Python into our application.

User interact with application with Angular frontend application. Angular app send HTTP Request to spring boot API which can query database for any required information, and put message into queue stating what it needs and python application takes the message from there and it can also store data files in blob storage.

Cloud Deployment

Once base architecture was fixed the basic application is also developed to be hosted. While hosting we faced several issues and we have to change the architecture multiple times to ensure robust integrations between services.

In order to keep up with learning curve, need for complex deployments, need of multiple environment to test code and limited cloud resources we decided to use AWS EC2 – Ubuntu server running tomcat server for spring boot application and angular applications and other python services are deployed as daemon service in the same server.

For DB we use AWS RDS service with MySQL engine.

For messaging system, we use Microsoft Azure service bus and for storing binary files we are using Microsoft Azure Storage Account Blob Storage

A diagram of a computer

Description automatically generated

Flow of our application

1. Data Upload

A diagram of a software development

Description automatically generated

This is the starting point for the user to our application. This flows starts by uploading dataset by end user in the angular application. From the angular app file is sent to spring boot API where file is saved in blob storage and name of the file will be stored in MySQL DB.

When file name is inserted into DB unique GUID is generated and returned to API. Unique GUID is Job ID and published to azure service bus in the form of message and same ID is returned to user.

Python app running on backend will pick up the message from the queue and start processing it once data profile is done it save the output back into blob storage and puts URI in DB.

Data ProfileA diagram of a software development process

Description automatically generated

All the necessary operations for data profiles will be finished when the user uploaded the data. So when user wants to see the results of data profile they can just give the job id they have into frontend. It will be passed to API from there DB will be queried with the job id and gets back the URI of data profile output

Once the API gets URI it will fetch file from the blob storage and return it front end application where the data will be transformed into visualization.

Data Cleaning

A diagram of a software application

Description automatically generated

When user is stastified with the data profile output they can start cleaning data. They can set the rules

Export DataA diagram of a software development process

Description automatically generated

GIT flow

CI/CD