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

Front end technologies – Angular, Bootstrap, PrimeNG

Backend technologies- Java, Spring boot, Maven

Database – mysql

File storage – Azure storage account – blog storage.

Data cleaning & Profileing – Python

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.

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 he 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

Cloud Deployment

A diagram of a computer

Description automatically generated

GIT flow

CI/CD