ABC 公司空气预测系统设计

1. 引言

1.1 项目背景

数百万人生活在空气污染可能导致严重健康问题的地区，个人和企业用户对于空气质量越来越关注，需用通过某种方式获取空气质量预测，提前采取措施来预防对身体的伤害。

1.2 项目目标

构建一个高精度、可扩展的全球城市空气质量预测系统，为企业用户和个人用户提供准确、及时的 AQI 预测信息，利用生成式 AI 技术为个人用户提供个性化的空气质量信息展示。

1.3 范围

- 数据源（NOAA GSOD, OpenAQ）
- 预测的时间（未来 24 小时）
- 用户群体（企业用户，个人用户）
- GenAI 图片生成

2. 需求分析

2.1 功能需求

2.1.1 数据采集与处理

国家海洋和大气管理局（NOAA）的全球每日表面总结，数据集 URL: https://registry.opendata.aws/noaa-gsod/ 本数据是开放数据。对该数据的使用没有限制。数据文档地址：http://www.ncdc.noaa.gov/。该数据在 AWS S3 中有存储，可以直接从 S3 中获取。

全球聚合物理空气质量数据来自政府、研究级和其他公共数据来源。数据集 URL: https://registry.opendata.aws/openaq/ 。数据许可遵守 CC BY 4.0 ，文档地址: https://openaq.org。该数据需要注册并获取 apikey 从官方网站获取。

原始数据有很多问题，有些数据偏差很大，有些数据缺失严重，在 ML 之前需要对数据进行 EDA，根据 EDA 结果，进行聚合、清洗、筛选等操作，需要通过代码来完成。

2.1.2 模型开发与训练

模型开发的目标是尽可能根据天气状况，对 AQI 进行准确的预测，GOSD 的天气数据和 OpenAQ 的 AQI 数据是作为输入数据源。

根据 EDA 的结果和特征工程，只保留对 AQI 有影响的气候因素，采用 AutoGluon 的文本模型进行训练，AutoGluon 可以自动进行数据的归一化，标准化等操作，模型优化也会自动进行，最后采用均方误差 (MSE)、均方根误差 (RMSE)、R 平方 (R^2)等方法对训练结果进行评估。

2.1.3 AQI 预测服务

使用训练的 AutoGluon 模型，部署到后端服务，对每日获取的天气数据进行 AQI 预测，并提供 API 服务给 web 端调用。

个人用户的健康提示图片，使用 Stable Diffusion 提供的 GenAI 服务，提供文生图功能，使用 prompt 工程，预制 prompt 来提示 GenAI 生成图片，把生成的图片和 AQI 结果一起存储，并提供 API 给 web 端调用。

提供 AQI 获取和健康图片的 API，根据用户类型不同返回不同的结果，展示在前端页面。

提供用户登录 API，来区分是企业用户还是个人用户来返回不同的内容。

2.2 非功能需求

2.2.1 性能需求

API 的平均响应时间小于 100ms。系统支持每秒 1000 个并发请求。

2.2.2 可靠性需求

系统可用性达到 99.99%，数据备份频率为每周一次全量备份，每天增量备份。

2.2.3 安全性需求

系统采用登录体系，拒绝直接访问系统 api；实现严格的访问控制，防止未授权访问。

2.2.4 可扩展性需求

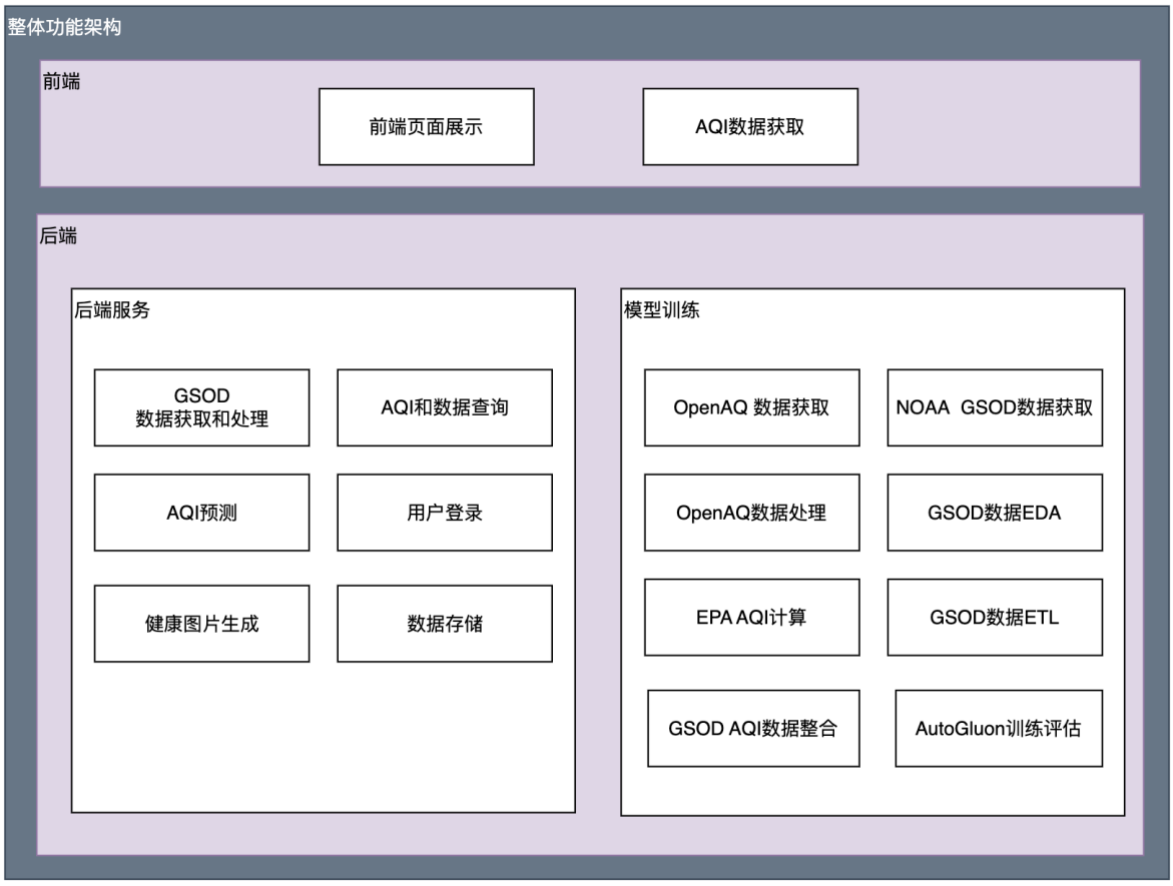
系统支持横向扩展，使用 Nginx 等负载均衡软 Nginx 件可以根据用户量和功能增加提供更大的并发量；并提供 CDN，Nginx，Redis，本地缓存等多级缓存体系来增加并发量。

2.2.5 成本效益

系统采用 AWS 相关的服务或者组件，可以让公司专注业务，其他的由亚马逊来负责管理维护，降低整体的复杂度和开发维护成本

3. 解决方案设计

3.1 总体功能架构



该系统整体上分为三个部分，模型训练，后端服务和前端页面。模型训练分为数据获取、数据规范化和模型的训练评估三个主要阶段。后端服务包括 GSOD 数据的获取模块；根据模型训练生层的模型对 AQI 进行预测，根据预测结果和具体的城市等情况通过 GenAI 生成健康图片模块；用户的登录认证模块；数据存储模块和 AQI 数据等 API 响应模块等。前端主要分为静态页面、登录认证和 AQI 获取处理等。

3.2 数据采集与存储设计

3.2.1 数据源分析

NOAA GSOD 数据在 AWS 的 S3 存储中获取，aws s3 ls --no-sign-request s3://noaa-gsod-pds/。OpenAQ 的数据从 https://registry.opendata.aws/openaq/获取，使用 http 协议。GSOD 数据包含很多天气数据如温度、气压、风力、降水等，需要根据需要进行筛选。OpenAQ 数据包含六种主要污染物的数据，有些站点只有部分污染物的数据，需要根据需要进行聚合。

3.2.2 数据存储

训练数据存储：由于是 demo 项目，采集数据的量不大，使用本地 csv 文件的格式存储。EDA 后的数据处理结果和 OpenAQ 数据聚合处理结果也是以本地 csv 文件格式存储。

实际预测数据存储：实际预测数据采用关系数据库 mysql 储存，方便 API 查询等操作。

3.2.3 数据处理流程

先进行 EDA，通过 pandas 来对数据进行整体的了解，使用 pandas 的直方图和箱线图对数据的分布进行了解。使用特征工程的方法进行数据的具体分析。根据 EDA 和特征工程的结果对数据进行聚合、清洗、筛选等操作，得到整合后的用于训练和测试推理的数据。

3.2.4 模型开发与训练设计

经过分析对比，AutoML 选用 aws 的 AutoGluon，由于要预测 AQI 值，并不是一个二元模型，所以选择 AutoGluon 的回归模型。

AutoGluon 有很多优秀的特点，仅用少量的代码就可以快速进行训练，利用云预测器和预构建容器从实验转向生产，可以方便的通过自定义功能处理、模型和指标进行扩展。

3.2.5 模型训练

使用 AutoGluon 的 TabularPredictor 进行训练，使用准备好的训练数据，使用 TabularPredictor 的 fit 方法，presets 参数选择 best_quality 进行训练，AutoGluon 会自己进行比较调优，训练出性能最好的模型。

训练完模型之后，使用训练出的模型进行预测，并根据预测值和实际值进行模型的预测能力评估，使用 MAE、RMSE 和 R2 的方法进行推理评估。

3.2.6 模型部署

AutoGluon 的模型在训练完成后，默认保存为一个 Predictor 对象，你可以直接加载，也可以封装成 API、打包成 Docker 镜像、或者部署到云平台，非常灵活方便。

3.3 AQI 预测服务设计

系统从 GSOD 定时获取各个城市的天气数据，使用训练生成的 AutoGluon 模型进行 AQI 预测，预测生成的结果会放入数据库中，通过 API 提供前端查询展示。

3.4 个人健康图片生成

个人健康图片是一个 GenAI 文生图的功能，需要根据城市和天气情况来构造 prompt，通过 API 调用文生图的大模型来生成跟城市和天气相关的图片，我们可以是用 AWS 部署的 Stable Diffusion 也可以使用 Replicate 的 Stabl Diffusion 来生成，生成完之后可以放到数据库中存储，通过 API 随 AQI 信息一起返回给前端用户展示。

3.5 API 服务设计

3.5.1 API 接口

系统采用 Restful 风格的接口，主要有如下几个 API：获取 AQI 预测结果和个人健康图片信息的接口，获取城市信息的接口，仅获取 AQI 预测信息的接口，用户登录认证接口。

3.5.2 认证与授权

系统内置两个默认用户，分别是 enterprise 和 individual 用户，使用 jwt 技术进行登录认证，用户和用户类型进行关联，根据企业用户和个人用户分别返回不同的信息给前端展示。

4. 设计假设与约束

4.1 数据假设

- GSOD 和 OpenAQ 的数据的质量可以满足需求；
- GSOD 数据源的更新频率能够满足预测的实时性要求；
- GSOD 和 OpenAQ 的历史数据满足训练模型的要求。

4.2 模型假设

- 有合适的机器学习模型，能够准确预测 AQI；
- 有合适的 GenAI 模型，能够生成符合需求的图片；
- 模型可以在合理的资源和时间内训练完成。

4.3 用户假设

- 用户能够理解 AQI 的含义和使用方式；
- 用户对个性化的空气质量信息展示有需求；
- 用户能够接受系统的性能和可用性。

5. 风险评估与应对

5.1 技术风险

- 数据源不稳定，譬如获取不到 GSOD 的天气数据；
- 模型预测准确率不高；
- GenAI 模型生成图片质量不达标；

5.2 数据风险

- 存储的数据丢失或损坏；
- 数据泄露或被篡改；
- 数据质量问题；
- 数据合规性问题。

5.3 应对策略

- 数据源不稳定属于第三方的问题，可以给出明确提示预警；
- 模型预测性不高的问题，可以通过不断优化，更多数据训练的方法解决；
- 可以使用不同的 GenAI 来对比生成的图片质量；
- 采用备份策略来防止数据丢失损坏；
- 加强安全措施，防止数据泄露或被篡改；
- 加强数据审查，对有问题的数据进行提示预警；
- 密切关注数据使用协议，有合规性问题及时找到解决方法

6. 参考文献

- NOAA GSOD 数据集文档
- OpenAQ 数据集文档
- EPA AQI 标准文档
- AutoGluon 文档
- AWS 官方文档
- GitHub 提供的 AQI 示例项目