**设计文档**

**1.需求分析**

**1.核心需求为**全球城市空气质量预测，细化为未来 1 天的 AQI 预测，24 小时制”。对于美国城市，需遵循美国环境保护署 (以下简称 EPA) 设定的美国 AQI 定义。对于美国城市，使用 2 个公开的科学数据集作为来源：1）美国国家海洋和大气管理局：每日全球表面摘要，2）OpenAQ。模型的训练应使用AutoML进行训练和评估。

2.最终用户有两种类型：企业和个人。面向个人终端用户的内容为体现城市特色、当日主题、AQI信息、健康信息的图片。其中相关图片应使用GenAI生成。面向企业的需求待进一步沟通。

3.技术需求：产品应满足扩展性需求以满足数据的不断增加。

总体原则是利用AWS平台满足他们的业务和技术需求。

**EPA关于AQI的定义概要如下：**

AQI是基于几种主要空气污染物的浓度来计算的，这些污染物主要包括：

1.颗粒物 (PM2.5 和 PM10)：直径小于或等于 2.5 微米（PM2.5）和 10微米（PM10）的细小颗粒物。

2.臭氧 (O₃)：地面级别的臭氧，通常在夏季形成，特别是在阳光强烈、气温较高的情况下。

3.二氧化氮 (NO₂)：主要来自机动车尾气和工业排放。

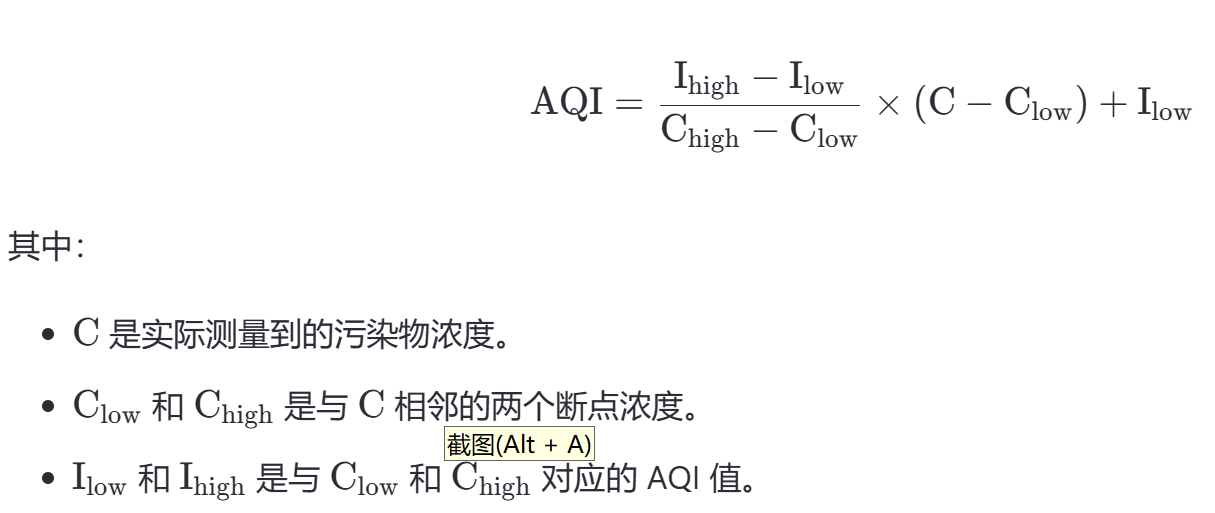
4.一氧化碳 (CO)：主要来自不完全燃烧的燃料，如汽车尾气和工业过程。

5.二氧化硫 (SO₂)：主要来自燃煤电厂和其他工业活动。

**AQI 的计算方法**

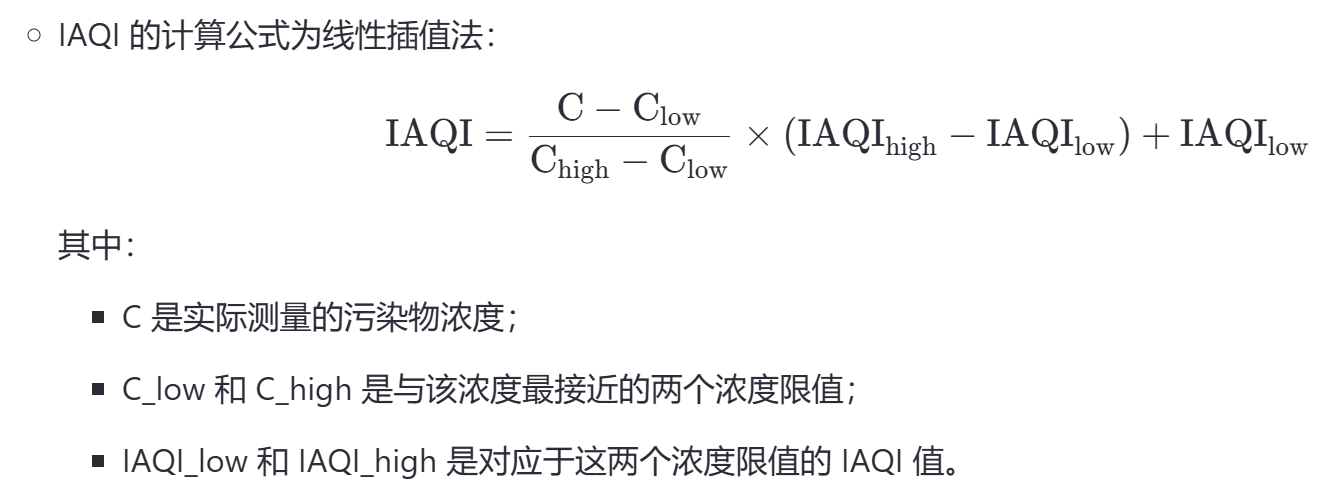
AQI 的值是根据每种污染物的浓度转换而来的。EPA 为每种污染物设定了不同的“断点”（breakpoints），这些断点对应于特定的健康标准。AQI 的值是从 0 到 500 不等，数值越高，表示空气质量越差，健康风险越大。

对于每种污染物，EPA 都有一个公式将污染物的浓度转换为 AQI 值。具体公式如下：



最终的 AQI 值是所有污染物中最高的那个值，即最影响空气质量的那个污染物的 AQI 值。

除美国EPA之外，其他国家对于AQI的定义和计算方法大同小异，也都是基于PM2.5，PM10，O₃，NO₂，CO，SO₂六种污染物浓度计算而来，比如中国对于AQI的计算方法如下：

****

**有了以上公式，需求就转变为：预测未来一天的**PM2.5，PM10，O₃，NO₂，CO，SO₂的浓度，并根据所处地区计算出AQI值。

据查，以上污染物主要的影响因素如下：

1. 气象条件：太阳辐射，气温，风速，降水等
2. 地理位置：城市，工业区，山谷，盆地，沿海等
3. 季节和时间变化：具备比较明显的季节性影响，部分具有日变化影响
4. 汽车尾气：尾气排放也是一个影响因素
5. 工业/农业活动：工业生产过程中会排放污染物或者重要前体物

**根据以上影响条件，结合以下的数据分析，可以进一步确定目前可以使用的污染物影响因素的数据，通过机器学习预测污染物浓度。**

**2.数据分析**

天气会对空气质量产生重大影响，因为天气的不同方面会影响特定区域中存在的臭氧和颗粒物的数量。阳光、雨水、高温、风速、空气湍流和混合深度都会影响污染物浓度。

1.阳光会导致某些污染物发生化学反应，从而形成烟雾

2.较高的空气温度会加速空气中的化学反应

3.雨水通常会减少污染，因为它会冲走颗粒物，可以冲走可溶解的污染物

4.风速、空气湍流和混合深度都会影响污染物如何扩散或从一个区域扩散

**依据以上信息，我们需要关注对空气质量有影响的主要因素，总结如下：**

**1.时间和周期：污染物浓度具有明显的时间依赖性，比如中国北方冬季的pm25指数通常比夏季高，月份的影响因素最大**

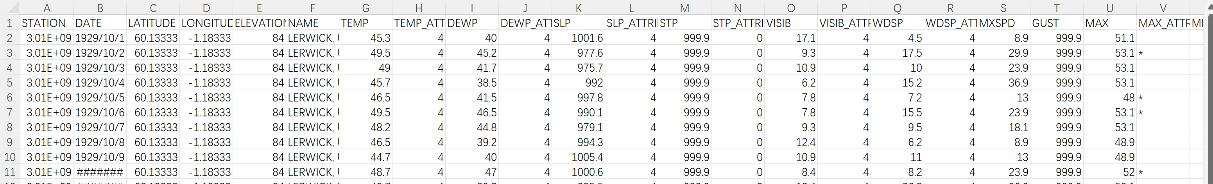
**2.地理位置：通常中国北方的pm25指数比中国南方高，沿海地区的pm25指数更低**

**3.气温：高气温加速空气中的化学反应**

**4.雨水：冲走和溶解污染物**

**5.风速：影响污染物的扩散**

**NOAA数据格式如下：**

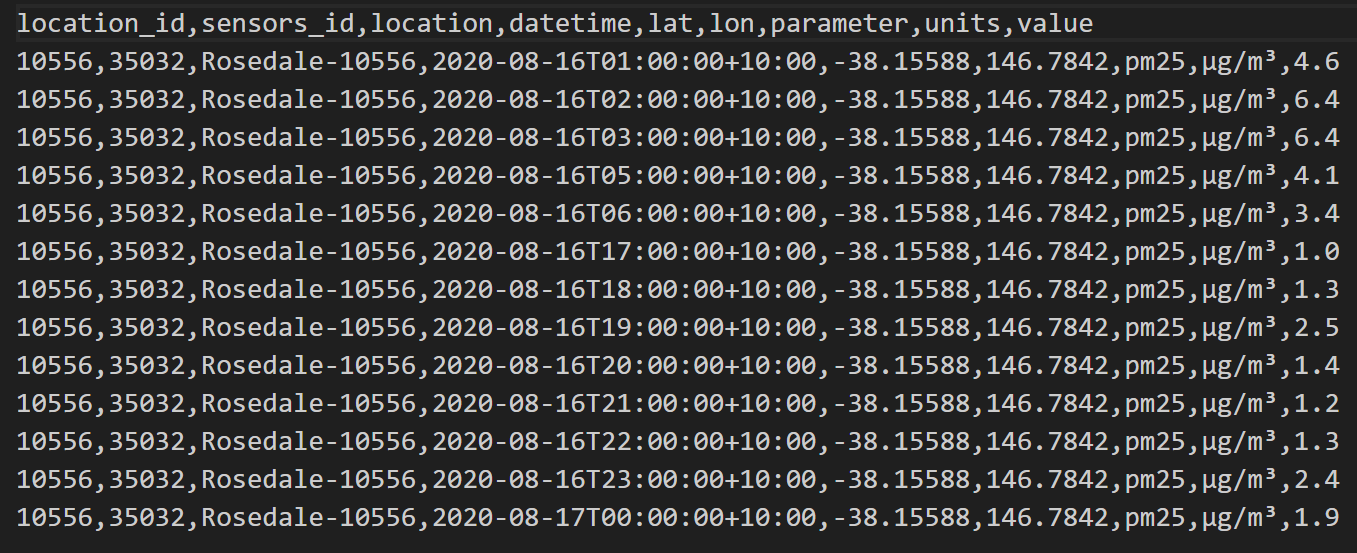
****

**相关的指标解释如下：**

* 1. **STATION：监测点编码，训练采用此字段表示地理位置**
  2. **DATE：观测日期，抽取月份作为主要条件**
  3. **LATITUDE/** **LONGITUDE：经纬度，用于NOAA和OPENAQ合并的依据**
  4. **TEMP：平均气温**
  5. **MAX：最高气温（暂不使用）**
  6. **MIN：最低气温（暂不使用）**
  7. **DEWP：平均露点温度（预测时暂不能提供未来24小时数据，暂忽略）**
  8. **WDSP：平均风速**
  9. **MXSPD：最大风速（预测时暂不能提供未来24小时数据，暂忽略）**
  10. **GUST：阵风速度（预测时暂不能提供未来24小时数据，暂忽略）**
  11. **PRCP：降水量**

**NOAA数据总量为582784条。**

**OPENAQ的数据格式如下：**

****

**通过经纬度与NOAA数据合并，可以得到监测点ID，月份，气温，风速，降水等与污染物相关联的数据。**

**合并策略如下：**

**获取某年的NOAA数据，得到经纬度，从OPENAQ数据中搜索与此经纬度不超过10km（可配）的监测点数据。**

**Openaq的数据量起码在800万以上，无法全部下载，最终采用调用openaq的api方式进行数据准备。由于openaq有限流措施，数据准备因为时间关系并不完整。**

**3.数据预处理**

1. **NOAA的数据是天气数据，OPENAQ的数据为空气质量数据，两者需经过经纬度合并为一份数据，形成6份可训练数据**
2. **同时可以去除我们暂时不关注的数据项。数据拆分为训练数据和测试数据，比例按8：2进行**
3. **考虑到污染指数与月份强相关，新增月份列，数据来源为日期**
4. **训练数据处理：过滤负值，同时数据分布呈现长尾特性（即大部分值较小，少数值非常大），应用对数转换减少大值的影响，保持小值的差异**

**4.模型设计**

**AutoGluon 支持多种机器学习任务：**

**1.分类：分类任务的目标是预测一个样本属于多个类别中的哪一个。AutoGluon 可以处理二分类和多分类问题。例如，预测一封电子邮件是否为垃圾邮件（二分类），或者识别一张图片中的人物身份（多分类）。**

2.回归：回归任务的目的是预测一个连续值。例如，预测房价、股票价格等

3.时间序列预测：时间序列预测涉及到根据历史数据来预测未来的值。AutoGluon 提供了专门的时间序列预测模块，可以处理单变量或多变量的时间序列数据。

**4.物体检测：物体检测任务旨在识别图像中不同物体的位置和类别。虽然这个功能在较新的版本中有所提及，但请注意它可能不是所有版本都支持的功能，并且需要特定的数据格式和配置。**

5.图像分类：图像分类任务与常规分类任务类似，但输入数据为图像。AutoGluon 的 MultiModalPredictor 模块也支持图像分类任务。

6.文本分类：文本分类任务涉及对文本内容进行分类。这可以通过 MultiModalPredictor 模块实现，它可以处理文本数据作为输入。

7.多模态学习：多模态学习指的是结合不同类型的数据（如文本、图像、表格数据）来进行预测。AutoGluon 的 MultiModalPredictor 模块能够处理这种复杂的数据组合。

**根据前述分析，污染物浓度具有明显的时间依赖性，主要是月份，同时和地理位置相关，与天气（气温，风速，降水）关系密切。基于以上特性，采用**回归任务预测，监测点ID，月份，气温，风速，降水作为特征输入模型。

总共训练六个模型，预测目标分别为PM2.5，PM10，O₃，NO₂，CO，SO₂。

推理部分，输入为城市，通过城市获取到经纬度，在NOAA数据集中搜索最近的监测点，得到监测点ID，并获取城市对应的预计**气温，风速，降水，与未来一天所对应的月份一起传入模型进行预测，得到**PM2.5，PM10，O₃，NO₂，CO，SO₂的预测值以后，通过城市对应国家的AQI算法最终获取预测AQI值。

训练可以采用Amazon SageMaker服务，可以大大加快训练进程。

**5.模型评测**

分别对PM2.5，PM10，O₃，NO₂，CO，SO₂六种主要污染物进行回归训练，六个模型的相关评测数据如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **评测指标** | PM10 | PM2.5 | O₃ | CO | NO₂ | SO₂ |
| **均方根误差RMSE** | **0.45** | **0.45** | **0.48** | **0.87** | **0.5** | **0.48** |
| **均方误差MSE** | **0.21** | **0.20** | **0.23** | **0.75** | **0.26** | **0.23** |
| **平均绝对误差MAE** | **0.30** | **0.31** | **0.27** | **0.45** | **0.37** | **0.34** |
| **中位数绝对误差MedAE** | **0.21** | **0.22** | **0.15** | **0.22** | **0.27** | **0.24** |
| **有效数据量** | **2724629** | **2904370** | **1664952** | **986575** | **2426410** | **1211543** |

**由于对数据进行了对数转换，误差会比实际误差更小，比如平均绝对误差0.30，经对数转换后实际误差应为1.35左右。**

**分析上表评测数据，以及与实际数据量级对比，其中**PM2.5，PM10，O₃，NO₂，SO₂误差较小，CO的误差较大，原因是CO的数据跨越较大，且没有很好的体现出与温度，降水，风速的相关性，考虑到模型预测值水平对AQI影响面较小，仍然纳入AQI计算，未来增加更多影响因素，如**工业/农业活动, 尾气排放等，增加准确性**。

**6.应用设计**

**应用设计为基于Web的架构，提供高可用，高可扩展性服务，并在aws构建的基础设施上实施和部署。**

**应用功能设计为：**

**1.核心功能：提供全球城市的AQI预测**

**2.用户管理：用户分为企业用户和个人用户，不同的用户的功能设计不同**

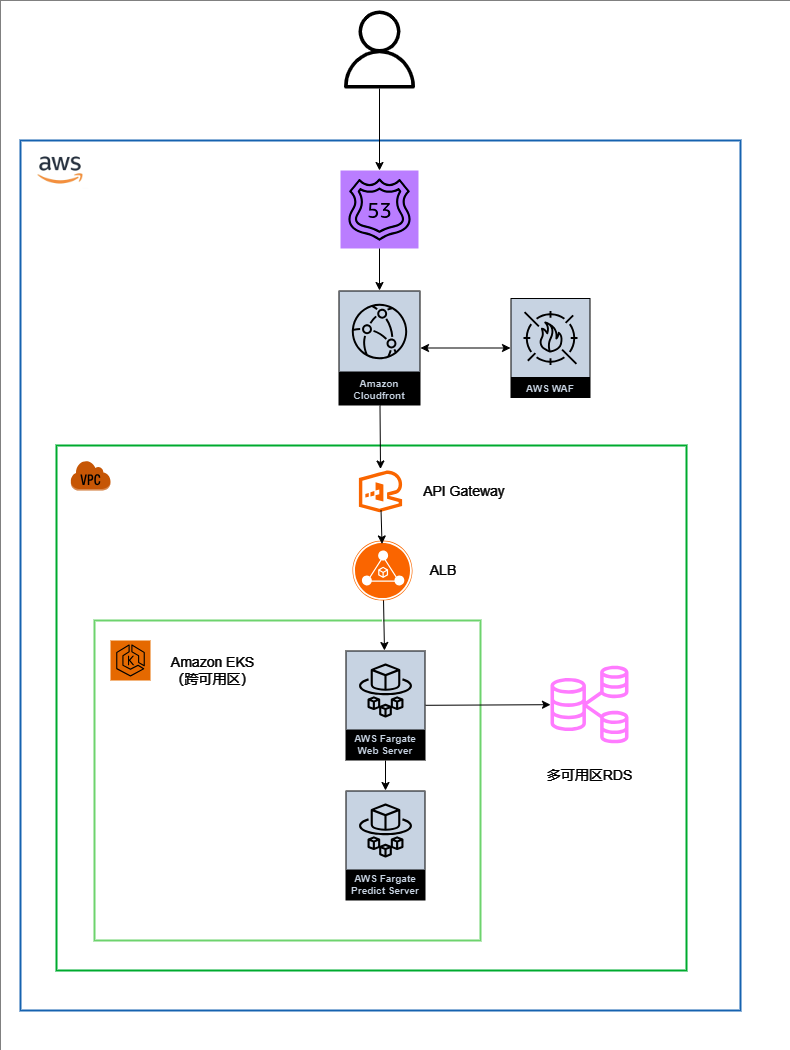
**3.全球城市管理：包含基础管理和地理位置**

**其中AQI预测主要由Predict Server提供，其他服务由Web Server提供。Predict Server需要大量的GPU计算资源，可考虑数据库记录已预测的数据。**

**aws为我们提供了强大的基础设施，借助aws可以很方便的实施我们的应用。**

1. **借助amazon route53服务完成域名的管理工作，以及DNS服务**
2. **借助amazon cloudfront服务，加速全球客户访问web的速度**
3. **借助amazon waf服务，有效提升web的安全性**
4. **借助amazon api gateway服务，提供统一的接口，完成认证，限流等**
5. **借助alb，完成后端服务的负载均衡能力**
6. **采用**amazon elastic kubernetes service (EKS)，创建托管的 Kubernetes，将web server和predict server容器化部署在EKS中。EKS为应用提供了高可用，自动扩展的能力。
7. Web Server和Predict Server将自主研发，借助amazon fargate服务，可以更方便的部署在EKS中，并实现基于负载的自动扩展。Fargate支持GPU，将大大提升AQI的预测速度。
8. 采用amazon multi-az RDS作为应用的数据库，实现跨区高可用，保证应用数据安全。在应对数据的不断增长上，RDS可以通过提升规格，分片到不同实例的方式为数据扩容。如果数据增长的需求非常强劲，我们可以采用amazon aurora，它能提供更高的性能，以及每个实例最高可扩展到128TB的。目前架构设计仍然采用RDS。

**应用整体架构图如下:**



**7.改进建议**

1. **数据部分：如前所述，目前模型采用的因素字段主要为监测点ID，月份，温度，风速，降水，可以增加更详细的气象数据进行训练，经纬度也可以作为训练数据进行尝试。同时，污染物浓度还与工业/农业活动, 尾气排放相关，尝试收集更多相关的数据有利于模型学习这些关系。**
2. **准备更完整的训练数据，收集全球更多的数据集，更大的数据规模和更完整的数据通常有助于模型的泛化效果。**
3. **尝试使用时间序列模型，捕捉污染物和除月份以外更多的周期关系，比如不同时间段、不同的季节或节假日，污染物浓度可能不同。**