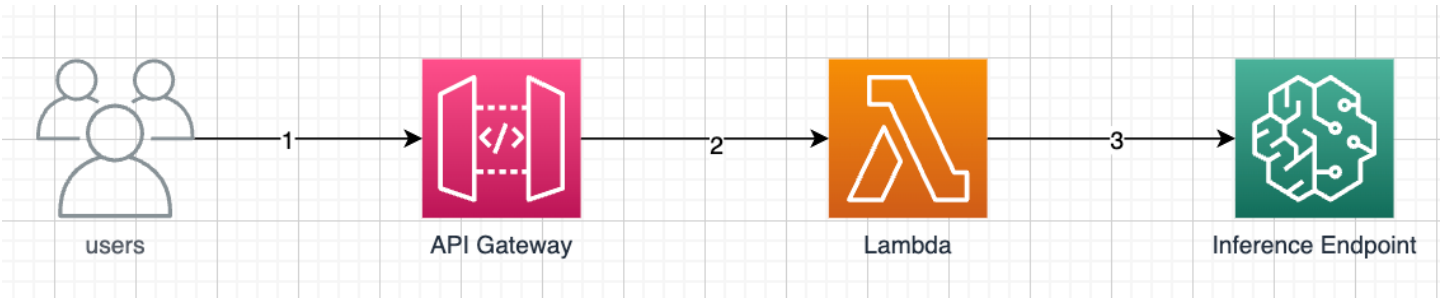
Serverless & SageMaker实现托管推理和微调平台

方案概述

基于AWS Serverless相关服务和SageMaker构建托管的推理和微调平台，使得平台用户可以以类SaaS的形式使用LLM。

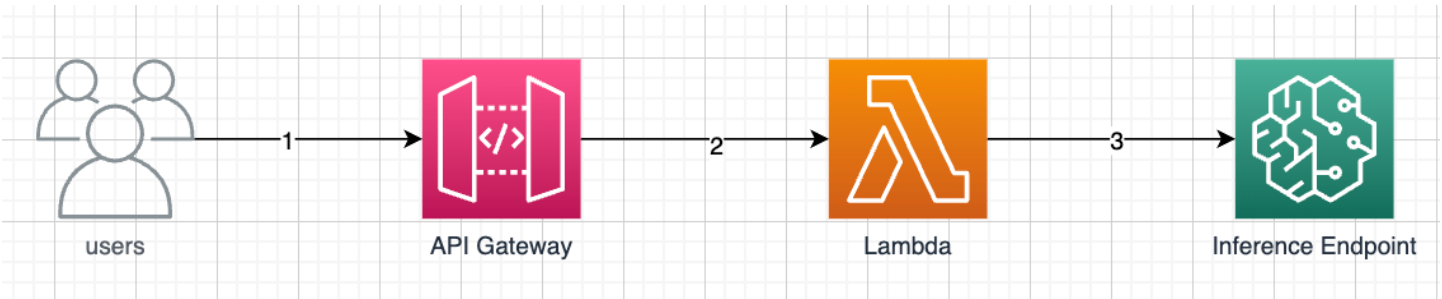
架构图

1. 创建推理端点



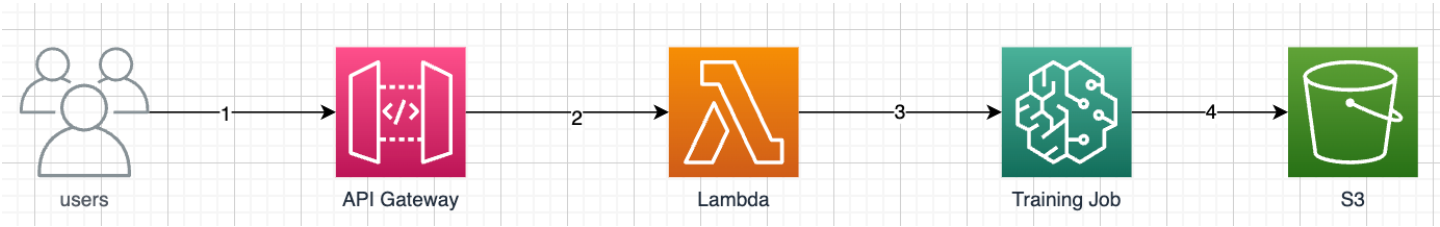
1. 用户（client）请求通过API Gateway暴露的API，提交相关信息（endpoint_name，模型S3路径等）；
2. API Gateway触发集成好的Lambda函数；
3. Lambda函数通过SDK创建推理端点；

2. 推理



1. 用户（client）请求通过API Gateway暴露的API，提交推理请求（endpoint_name，输入参数等）；
2. API Gateway触发集成好的Lambda函数；
3. Lambda函数通过SDK调用推理端点进行推理，并同步返回推理结果；

3. 训练（微调）



1. 用户（client）请求通过API Gateway暴露的API，提交训练任务请求（基础模型，数据集，输出路径，其它参数等）；
2. API Gateway触发集成好的Lambda函数；
3. Lambda函数通过SDK调用SageMaker Training Job，创建训练任务，并返回相关信息；
4. Training Job完成后会自动将产出物上传至指定的S3存储桶；

API

1. 创建推理端点

Api: `/create_inference_endpoint`

Method: `POST`

请求参数:

参数名	类型	是否必须	示例	备注
model_name	string	是	cpm-bee-10b	模型名称，用于标识部署哪个模型
endpoint_name	string	是	cpm-bee-2023082201	推理端点名称，全局唯一，用于后续调用推理服务使用
delta_path	string	否	s3://bucket/test.ckpt	delta增量微调结果文件的S3 uri

响应结果:

```
// 示例
{
    "error_no": 0,
    "error_msg": "",
    "data": {
        "endpoint_name":"cpm-bee-2023082201",
```

```
        "endpoint_status": "creating"
    }
}
```

2. 推理端点调用

Api: `/invoke_inference_endpoint`

Method: `POST`

请求参数:

参数名	类型	是否必须	示例	备注
endpoint_name	string	是	cpm-bee-2023082201	推理端点名称，全局唯一
input	string	是	{"input":"我想回家","prompt":"汉译英","<ans>":""}	推理输入
max_new_tokens	int	否	100	

响应结果:

```
// 示例
{
    "error_no": 0,
    "error_msg": "",
    "data": {
        "result":"{'input': '我想回家', 'prompt': '汉译英', '<ans>': 'I want to go
    }
}
```

3. 删除推理端点

Api: `/delete_inference_endpoint`

Method: `POST`

请求参数:

参数名	类型	是否必须	示例	备注

endpoint_name	string	是	cpm-bee-2023082201	推理端点名称，全局唯一

响应结果：

```
// 示例
{
    "error_no": 0,
    "error_msg": "",
    "data": {}
}
```

4. 查询推理端点状态

Api: `/describe_inference_endpoint`

Method: `GET`

请求参数：

参数名	类型	是否必须	示例	备注
endpoint_name	string	是	cpm-bee-2023082201	推理端点名称，全局唯一

响应结果：

```
// 示例
{
    "error_no": 0,
    "error_msg": "",
    "data": {
        "endpoint_name":"cpm-bee-2023082201",
        // 'OutOfService'|'Creating'|'Updating'|'SystemUpdating'|'RollingBack'|'
        // 只有 InService 状态时才能提供推理服务
        // 只有 非 Creating/Deleting 状态时才能提供删除操作
        "endpoint_status":"deleted/creating",
        ....待补充
    }
}
```

5. 创建微调任务

Api: `/crete_training_job`

Method: `POST`

请求参数:

参数名	类型	是否必须	示例	备注
model_name	string	是	cpm-bee-10b	模型名称，用于标识基于哪个模型进行微调
dataset_path	string	是	s3://bucket/dataset.json	微调数据集，具体格式参照官方推荐
eval_dataset_path	string	是	s3://bucket/eval_dataset.json	eval数据集
epoch	int	否	3	训练轮次
batch_size	int	否	8	数据批次大小
max_length	int	否	2048	最大长度
instance_type	string	否	ml.g5.12xlarge	实例类型

响应结果:

```
// 示例
{
    "error_no": 0,
    "error_msg": "",
    "data": {
        "job_name":"cpm-bee-finetune-2023-08-14-08-17-16-651",
        "output_path":"s3://bucket/output/a.ckpt"
    }
}
```

6. 查询微调任务

Api: `/describe_training_job`

Method: `GET`

请求参数:

参数名	类型	是否必须	示例	备注
job_name	string	是	cpm-bee-10b	job名称

响应结果:

```
// 示例
{
    "error_no": 0,
    "error_msg": "",
    "data": {
        "job_name":"cpm-bee-finetune-2023-08-14-08-17-16-651",
        // 'InProgress'|'Completed'|'Failed'|'Stopping'|'Stopped'
        // 只有 Completed 状态时才能进行部署操作
        // 只有 InProgress 状态时才能执行stop操作
        "job_status":"InProgress",
        ...  待定
    }
}
```

7. 停止微调任务

Api: `/stop_training_job`

Method: `POST`

请求参数:

参数名	类型	是否必须	示例	备注
job_name	string	是	cpm-bee-10b	job名称

响应结果:

```
// 示例
{
    "error_no": 0,
    "error_msg": "",
    "data": {
```

```
        "job_name":"cpm-bee-finetune-2023-08-14-08-17-16-651",
        "job_status":"InProgress"
    }
}
```