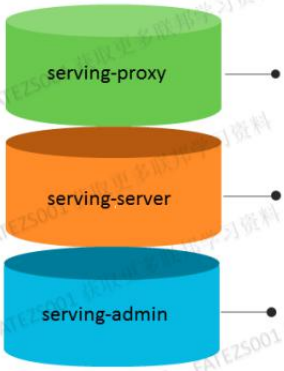
8 月 27 日，FATE 开源社区第三次圆桌会完美落幕。这次的圆桌会，我们与老朋友们分享了更多在实战案例的应用。会上，我们首先对 FATE-Serving 2.0 的新版进行了全面的介绍，在其架构、流程、组件以及业内同类产品进行了横比，然后再回来我们熟系的环节，帮助解决社区朋友们在日常使用过程中遇到的问题。

以下为分享 PPT 的部分内容：

版面有限，有兴趣的朋友欢迎联系小助手（FATEZS001）获取完整内容。

	FATE Serving	Tensorflow Serving	Paddle Serving	TorchServe
联合推理	√	×	×	×
通信框架	GRPC	GRPC	BRPC	HTTP
集群能力	√	×	×	×
服务治理	√	×	×	×
服务保护	√	×	×	×
监控、鉴权	√	×	×	×
插件化	√	×	×	×
多协议	√	√	√	×

FATE-Serving2.0 的组件

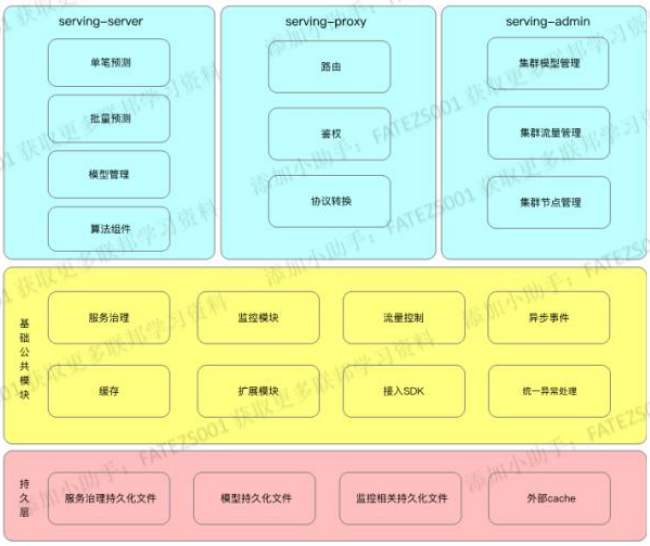


在多方交互时提供路由功能、协议转换功能

预测逻辑的执行模块、包括各种算法组件

提供集群管理功能，可视化功能

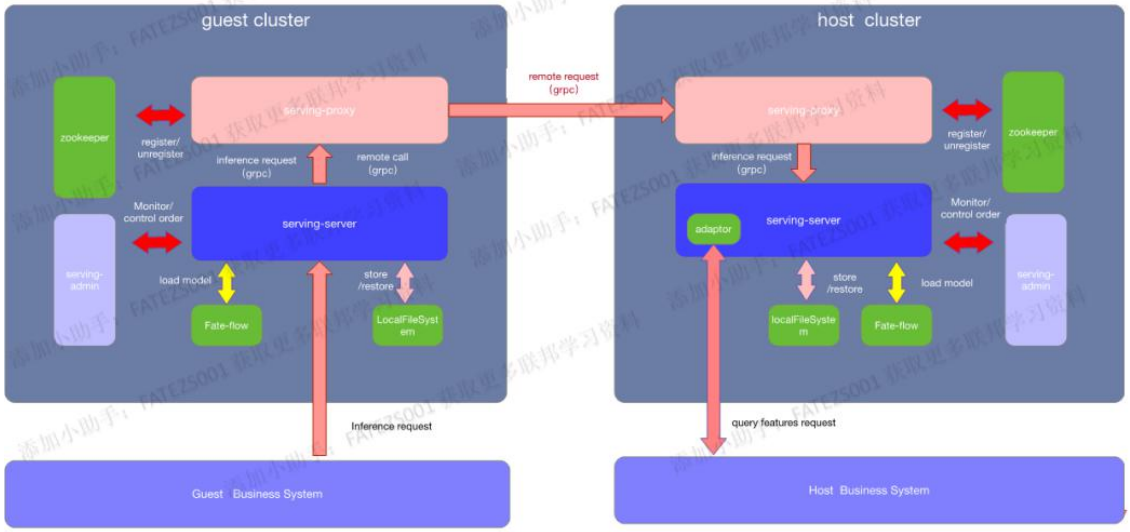
FATE-Serving2.0架构图



FATE-Serving2.0工作流程

- 打包部署，从github下载源码打包或者直接下载腾讯云上已打好包的版本
- 从Fatelow推送模型
- 模型初始化，Serving-server收到推送命令后拉取模型数据并初始化进内存
- Serving-server将模型信息转换为接口信息注册进入注册中心
- 调用方从注册中心拉取该模型的接口地址并发起调用

FATE-Serving2.0部署架构



以下为圆桌会上的精选问答

1、serving 是必须要用 ZK 吗？

推荐使用 zk，因为不使用 zk 配置起来会比较麻烦，目前是支持不使用 zk 的。

2、集群各方都需要部署一个 guest 的 host 吗？

Guest 和 host 至少是有一个，之后的版本有可能会支持一个 guest 对应多个 host。

但是目前 guest 和 host 至少得是一对一，只有一个 guest，或者只有一个 host，那都是不成立的。

3、模型绑定设备，是不是只能在 guest 发起之后？预测的请求也是只能发给 guest？

目前是，只能在 GUEST 发起。

因为 HOST 方发起的话，host 侧是只有一个半模型，算出来结果实际上是无效的。

4、各方都部署了，只是各方的集群是什么意思？

各方部署并不是只有一个进程，由多个组件多个进程组成了一个集群。

5、serving 和 1.3 版本一样是通过配置而决定使用 zk 还是不使用 zk 吗

是的。这个是可以配置不使用 zk，但是不推荐这样使用，因为不使用 zk 的话，很多同学会把 IP 给配错，然后流程走不通。

所以最好是能使用 ZK 就使用。

6、有更为详细的文档吗。

有，现在 FATE-SERVING readme 里面，留了一个链接叫 document，指向了一个腾讯云文档，这个文档目前还在持续更新中。

7、横向的预测还用 FATE-SERVING 吗？

目前 FATE-SERVING 只是纵向的。

8、有一个最大 QPS 延时多少毫秒？

用 16 核机器，如果是批量预测的话，吞吐量可以到 4000~5000qps 。

每个请求是在延时在 100~200 毫秒之内。

9、SERVING 默认使用 ZK 作为注册中心，之前 redis 方式还支持吗？

支持的，不过之前 redis 不作为注册中心，而是作为一个预测结果的缓存而已。

10、模型 unbind 接口是否有开放出来？

这个是开放出来的，并且可以调用，但目前的方式是通过页面调用，如果想结合到自己的系统里面也是可以直接调用的。

我们有接口的 protobuf 文件，如果有这个需要，可以自行开发。

11、sendToRemoteFeatureData 如果传 id 的话，还要知道对方有哪些 ID 么？

guest 和 host 进行联通的话，首先得知道对方系统是需要什么 ID 的，在了解这个之后，然后再进行 guest 和 host 双方的联调。

12、横向建模的流程是怎样的？

横向建模流程目前不支持。

13、大批量预测如何实现？例如预测数据都存 CSV 在文件中的情形。

预测数据都存在 CSv 是指特征数据都存在 CSv 吗？还是说预测请求存在 csv？

14、如果所有的预测请求都存在 cSv 文件里面的话，现在我没有办法做预测吗？

预测请求使用 CSV 可否在预测之前用其他系统去把从 CSV 里面读出来？

读出来也行，但如果组成 url 的形式的话，因为 url 长度也有限，我们一般如果存在 CSd 文件里面，它都可能就是大批量的预测？

你说的大批量会是上千笔上万笔这种吗？

上万笔目前是不支持的。

目前的批量大概 500， 1000 以下。

因为这是同步预测，就是 guest 的 host 的双方同步预测他的超时时间是有限的，比如说是 2~3 秒，这时候如果是大批量预测，hos 那边就可能计算不过来，几秒的超时时间可能不够，可以把那种大批量切分成小的批量，比如说一批只放一两百笔去发送会更好。

后面会支持像那种 csv 格式吗？

如果是指支持很大批量的那种预测，这个有可能会对流程的改造非常大，因为肯定不能适用现在这种同步预测，而是做一个离线异步预测之类的场景。

离线预测目前是没有支持的，所以看之后有没有这种需求，可以规划起来。

15、在 1.3 版本，score 出现负值是啥情况？

这个问题得看一下模型具体训练的什么，以及最后模型训练成什么样了。

没有一个准确的答案能够解释这个问题，得看当时是怎么训练的，然后采用了哪些组件。

然后 guest host 那边是不是都网络打通了。

16、刚提到一个 guest 和多个 host 的情况，可以详细说下吗？

这个会在接下来的小版本中支持，目前还不支持。

17、目前 FATE 是直接向 serving 推送模型的，有时候训练环境和推理环境不在同一网络下，这样推送模型的时候需要 serving 具备公网。2.0 架构还需要训练环境配置全部的分设为节点的地址吗？

现在 fate-flow 已经可以支持模型迁移功能，因此 serving 就没有提供专项支持。由 fate-flow 去做模型的迁移，能从一个网络环境 copy 出来，然后放到另外一个环境下，由它来做这个事情，所以 serving 目前没有关注这个点。

18、目前可以部署三方两个 guest，一个 host 或者两个 host，一个 guest 吗？

第一个情况部署是支持的，这种也是常态，一个 host 对应很多很多的 guest ，但是这些 guest 之间是没有关联的，互相不知道对方的存在。第二种情况目前不支持，但是接下来会支持。

19、新的用户是指不是共同的用户，是有一方有该用户？

在另一方会出现没有特征，新的用户就是有一方没有。它这边的错码会非零的，同时也会返回一个分数。这个错码会提示你在 host 侧是没有特征的，你就知道这个分数有可能是不可靠的，因为 host 方没有查到用户的特征。

20、unbind 的接口开放出来了吗？

Http 接口我们是没有开放的，开放的是 grpc 接口，HTTP 接口只开放了预测接口。

21、预测这个过程超时时间大概会到 2~3 秒，因为整个过程会从 host 拉取特征，数据对批量预测的话，2~3 秒这个时间会够吗？

批量预测在之前的很那个图里面会有一个 adaptor，实际上会有两种 adaptor，一个是单笔的，批量是要求对方业务系统是要有实现批量接口，如果没有实现批量接口的话，这个问题肯定是存在的，他肯定会超时，一笔笔去轮循环去拉这个特征的话，他肯定已经超时了。

所以你对外如果说采用批量，这肯定是要求对方的 host 的业务系统是支持能够支持批量拉取特征的。

22、请求入口只能由 guest 发起，一般应用方或者业务方他是作为 host，比如说我们如果要配置的话，host 方是我们自己。

比如说用户发了一个请求，是要先把请求链接发到 guest 上集群上，然后再开始进行一个预测的过程？

一般情况下，host 方数据相对来说会比较多，是它持有大量数据，而 guest 方可能并没有 host 方所需要的数据。

这个实验也有好处，因为 guest 的和 host 双方保存了共同的交集，只能对交集部分打会比较可靠，这样相比较从后期的作为，这样的时延会大大降低，还是说没有关系，其实就是这么设计的？

打个比方，按你说的这种从 host 方发起，然后在 host 上预测，然后请求会往 guest 那边再去算一次吗？

但是还是在数据量上会不一样，我理解的 host 也就是我们自己业务方有自己的用户标签，但是想要用到 guest 的那些特征？

可以这样理解，并不是说一方他只能做 guest 或只能做 host，他其实 guest/host 他都可以做的。在自己本方发起了一个预测，然后又跑到另外一方去算了一下，其实那方就是 host 的角色，这个过程其实是互不冲突的，host 其实可以随时切换。

只不过训练的时候要指定谁作为 guest 谁作为 host，训练出来的模型中有自己这方属于哪个角色的信息 。总结一下：预测请求一定是从 guest 发给 host，但是如果有需要可以在建模的时候调换一下双方的角色，这个有可能需要修改双方 FATE 的一些配置。

在预测的时候，其实看请求是先发给谁入口，就是作为 guest?

训练好的模型里面会告诉你这方模型是 guest 的模型还是 host 的模型，在线集群是可以 guest 模型和 host 模型同时存在在一个实例中。谁作为 guest 谁作为 host 在模型训练的时候决定。

23、假设我既作为一个 guest 又作为一个 host，那么 host 跟 guest 是部署在同一套系统中，还是说这边可以部署两套系统?

部署在同一套系统也是可以支持的，只不过会比较复杂，必须对这里面的配置相当的熟悉，需要解决的问题是请求出去以后，怎么能回到自己的这套系统。

流程的话也是从 guest 到 host，只不过 host 就是自己。不建议用同一套系统来做，最好两套。

24、预测的话它是从 guest 方发起的话，它只能通过 url 去发送，这样的话就不可以发送那种语音数据或图片数据?

目前是不支持图片和语音的。

25、预测结果中的 inputdatahitrate 和 modelingfeaturehitrate 表示什么?

2.0 版本这两个字段去掉了。

26、横向联邦预测还用 fateserving 吗?

横向预测没有提供服务，因为业界这样的方案很多。横向我们后续会提供一些模型转换工具，比如适配 tensorflow-serving。

原文链接：https://mp.weixin.qq.com/s/F19sh70KU_U7UJkGaL6FQw