5 月 26 日，FATE 开源社区第九期圆桌会圆满落幕。本次圆桌会，微众陈伟敬为我们分享了 SecureBoost：挑战千万级别样本训练之性能提升篇。

接下来带大家回顾一下经典**问答环节**，为新老朋友答疑解惑。

#问答环节

想问下 fast-sbt 在不同数据上和 hetero-sbt 对比?

以 fast sbt 的 mix 模式为例，mix 模式下，一代用 guest 的特征建树，一代用 host 的特征建树，以此往复，这样很像是跑 hetero-sbt 时对特征做一个列采样。我们在几份样例数据上测试 fast-sbt，最终达到的效果是差不多的，但是 fast-sbt 它可能会需要多跑一些代数，最终才能达到 hetero-sbt 的效果。

为什么没有使用 lightGBM 的优化呢，Xgboost 只是 lifewise，leafwise？

Xgboost是层优先的，layer-wise，是一层一层的构建的，lightgbm的是leafwise，lightgbm用到的goss，直方图相减FATE-1.6都是用到了的。

1.7 的升级会在架构上和 1.6 有很大差别吗？还是侧重于训练过程优化?

4 月圆桌我们有对 1.7 进行展望，有兴趣可以去回顾（文末有链接），1.7fate-flow 等会拆分，架构会有比较大的变动。

想问下 sbt 安全性问题，目前像 mix 模式, host 方的完整树结构是需要发给 guest，这会不会有一定安全性问题。

mix 模式，host 树结构是不会发给 guest 的。

有没有考虑实现密文下比大小的操作？这样 host 就不需要回传分裂点给 guest。

是在 host 做分裂点收益比较吗？目前在同态加密的情况下，在 host 没有办法计算出分裂点收益，所以也就没办法进行比较了。

用 eggroll 来做的计算框架，用单机 standalone 的时候，按照我的 CPU 的核数来分配的，就是单机的，后来我变成集群模式的时候，也是相当于核数有几台就扩充了几台，总核数都利用上，但是在效果上发现训练时长甚至比原来单机的时间还要长，不知道有没有这方面的一个排查问题的思路?

配置里面会有一些并发参数，有两个参数，一个是 computing_partition: 数据的分块，分了几个 partition；还有一个是 task_cores：并行的时候，用上多少个核，如果没有配上的话，确实有可能是跑得比较慢的，另外一个情况是集群模式下有调度和网络传输开销。

（接上一个问题）这两个参数都是有配置的，而且配置了 48，和我们的核数符合。

在运行 Hetero-SBT 的时候，我们这边也碰到一种情况，partition 越大的时候，写出的时候，每个 partition 就会涉及一个加密直方图写出的过程，这个过程每个 partition 根据 key 将结果分发到不同的 nodemanager 上，这是一个 shuffle 的过程。那么这种情况下涉及 IO 开销和调度开销。所以当你数据量不是很大，计算已经很快了，那增大 partition 可能会导致其他方面有出现瓶颈。

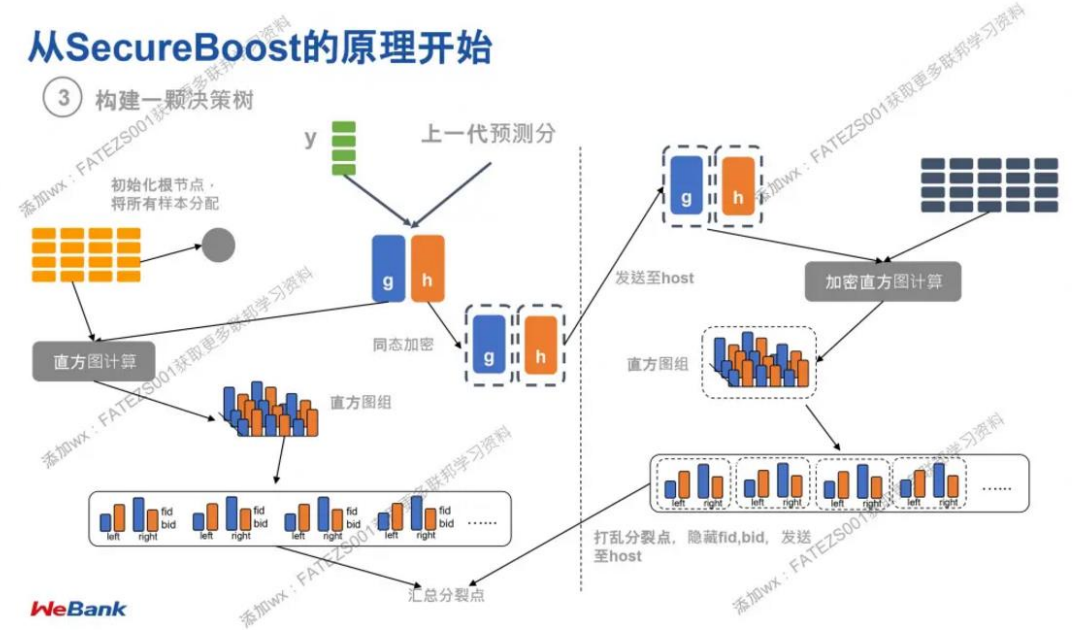
（接上一个问题）所以其实有的时候是需要去减少分区数量?

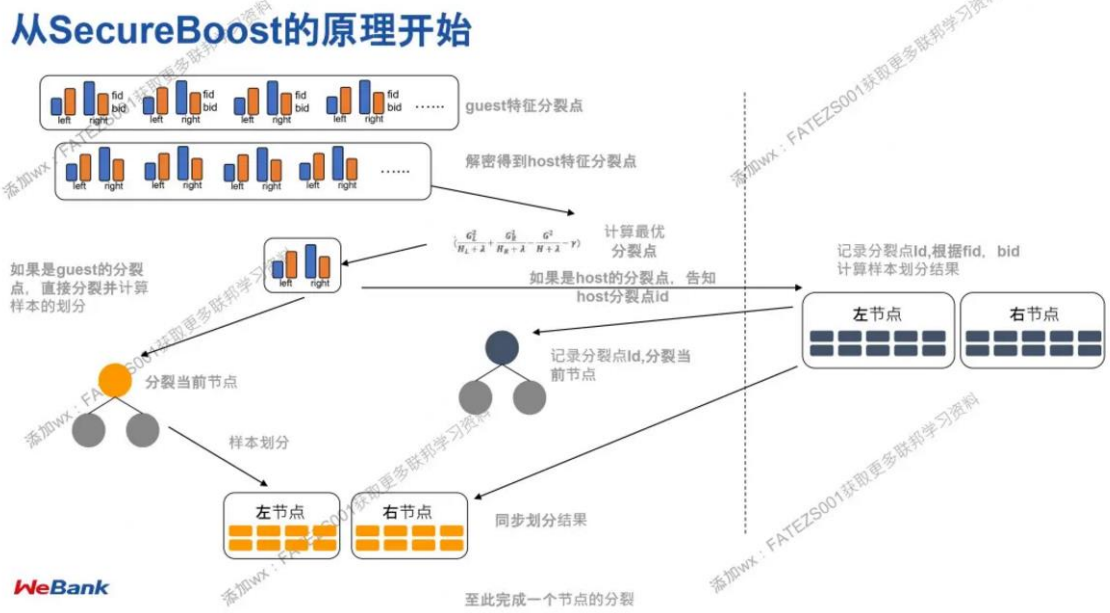
是的，数据量比较小，或者计算性能提升到极致后，会导致有其他瓶颈的情况出现。

（接上一个问题）所以你建议的一个优化的方案就是说适当去减小分区数吗?

理论上来说计算和 IO 调度等消耗是有一个平衡点的，并不是无限的增大资源，就一定会提升速度。

以下为本次圆桌会的部分内容介绍，添加小助手（FATEZS001）可获取详细资料：





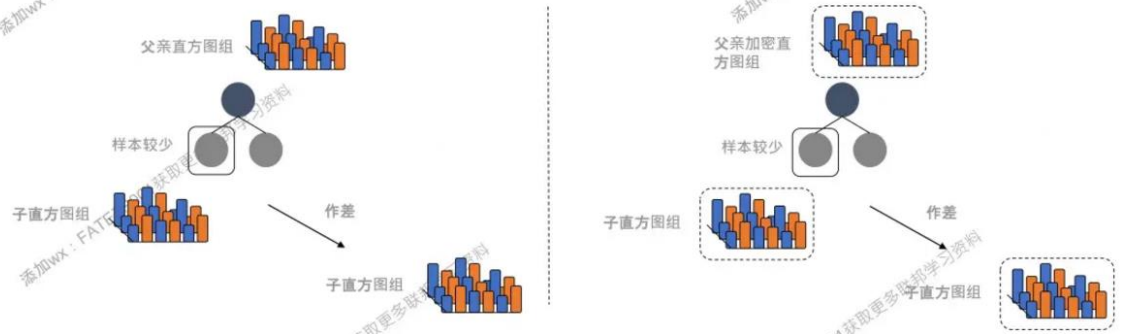
优化的方向

从最基础的SecureBoost开始，我们可以做什么来加快训练/预测的速度，容纳更大的训练样本？

WeBank

直方图相减

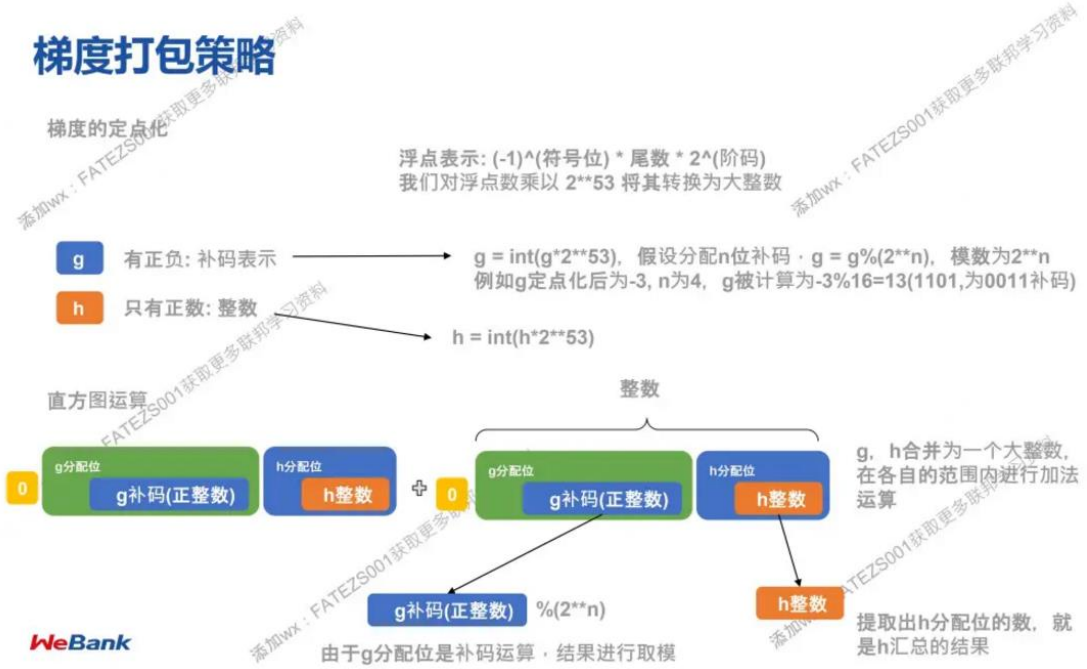
在构建每一棵树的时候，各方都会自动缓存父层的直方图，
选取样本量少的节点计算，执行直方图相减

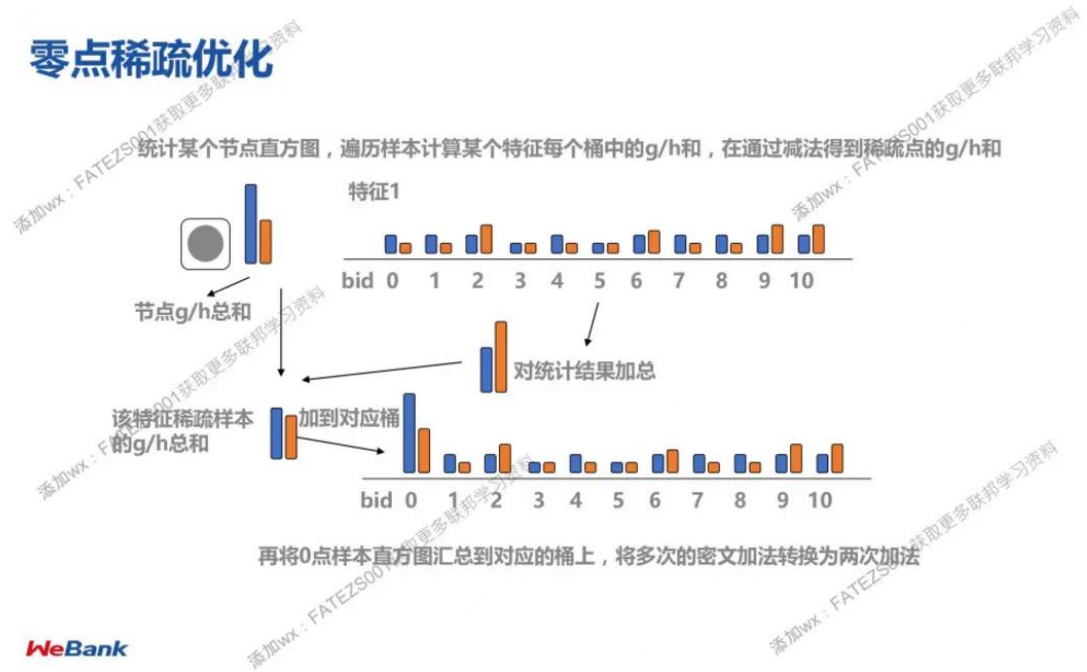


即使在最差的情况下，涉及计算的样本只会有原来的一半，实际情况更优

n_sample = n_sample / 2

WeBank





获取会议 PPT，或对圆桌会还有别的疑问？欢迎联系 FATE 开源社区助手获得帮助。

原文链接： https://mp.weixin.qq.com/s/41TVzaG6oBLhNYbGy6qtzw