



微博机器学习平台实践



微博机器学习平台负责人
黄波, @黄波_WB
2017.11.23



➤ 微博

➤ 微博机器学习平台

- 大规模机器学习
- 大规模深度学习
- 机器学习 workflow
- 平台效果

➤ 业务实践

- Feed机器学习排序

微博 中国领先的社交媒体平台



1.65亿

日活跃用户DAU



3.76亿

月活跃用户MAU



92%

移动月活占比

微博 中国领先的社交媒体平台



6.2亿

视频发布总量



8700万

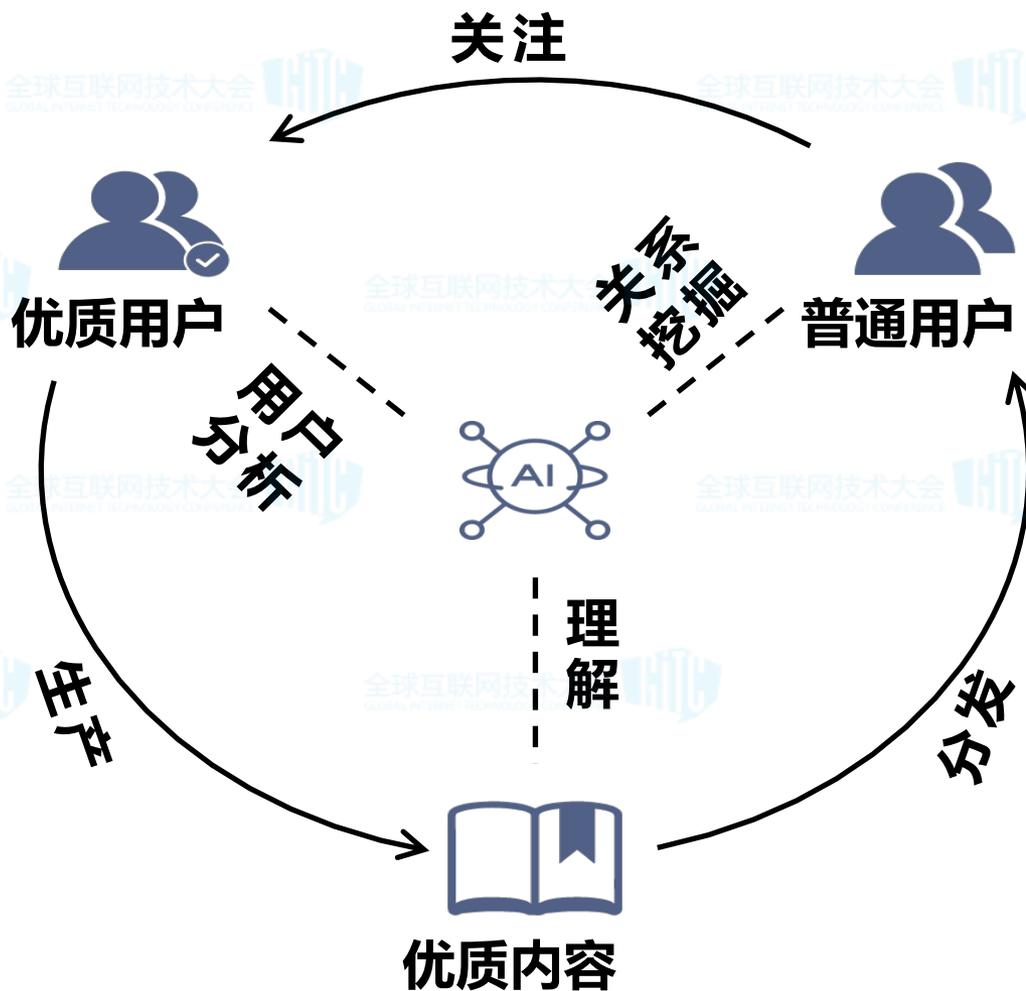
直播总场次数



2600亿

博文发布总量

微博 中国领先的社交媒体平台



大数据

大规模

高频访问

用户体量大

用户间关系纷杂

微博内容体量巨大

微博内容数据多样
(文本、图片、音频、视频, 等)

特征类别多

特征维度巨大

近百亿级别特征维度

近万亿级别样本量

机器学习

算法模型多样化

深度学习

LR,SVM,GBDT,CNN,...

业务场景多样性

业务场景复杂

Feed,热门,用户增长,反垃圾,...

流程无标准, 沟通效率低

开发流程冗长

长期迭代调优

计算框架多样性

hive,Hadoop,spark,tensorflow,storm,...

系统运行门槛高

执行性能差

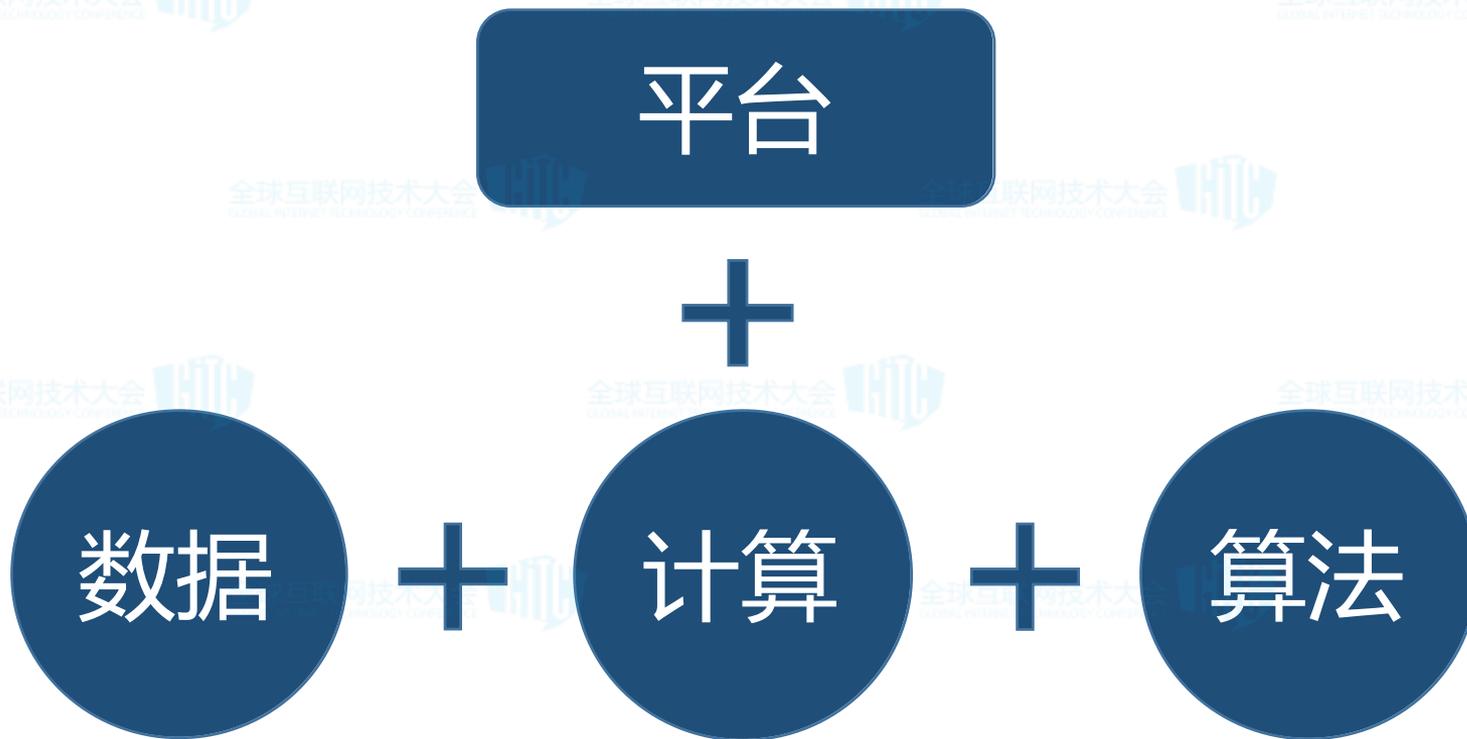
业务程序依赖多

重复建设成本

人力成本

标准化

平台化



➤ 大规模机器学习

➤ 大规模深度学习

➤ 机器学习工作流

➤ 大规模机器学习

➤ 微博实践证明：机器学习规模越大，效果越好

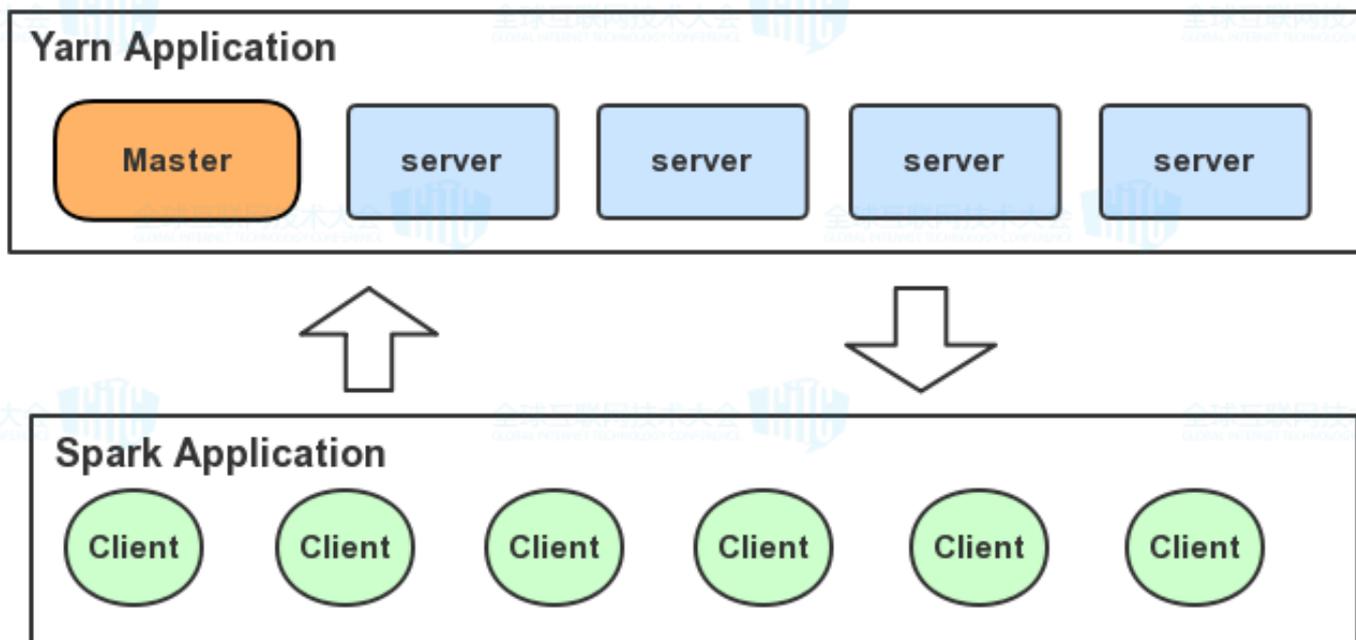


➤ 规模：几千亿样本、几十亿特征

- Hadoop : MapReduce
- Spark : RDD、MLLib
- 参数服务器WeiPS : 解决样本和特征规模化问题

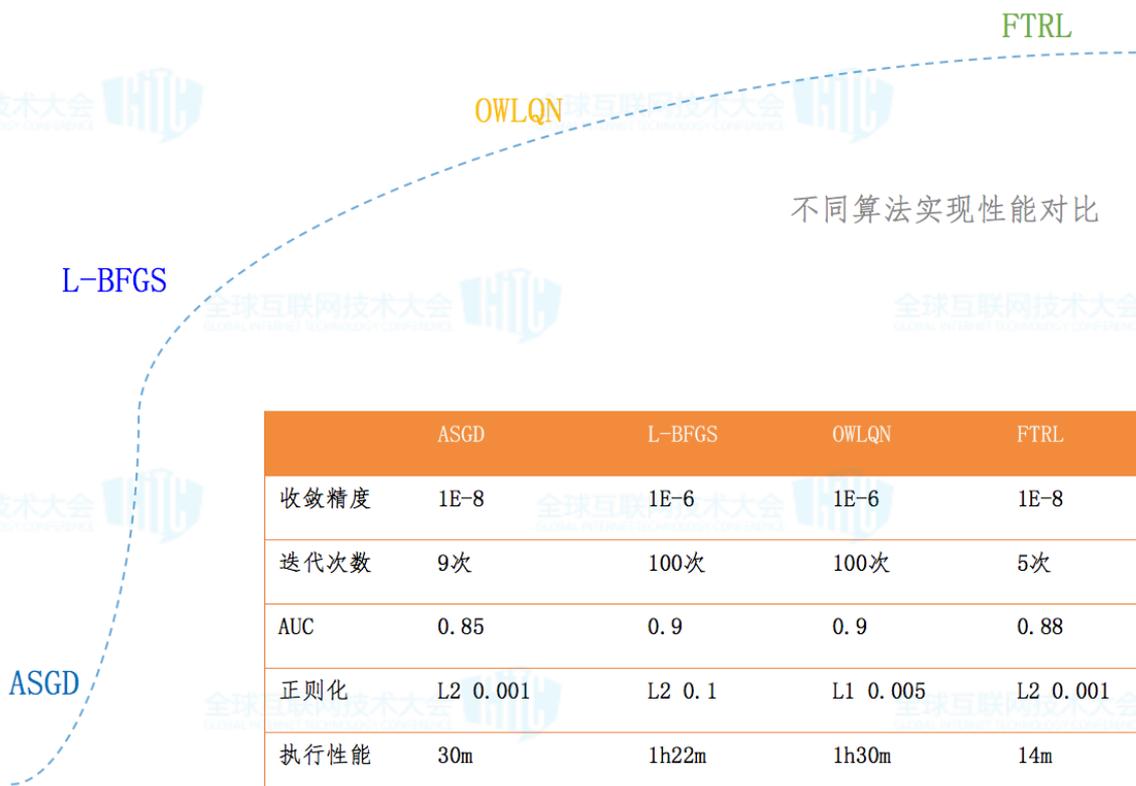
➤ 参数服务器WeiPS-架构

- 参数存储：分布式
- 同步控制：ASP/BSP/SSP
- 容灾机制：Checkpoint/多副本
- 功能扩展：psFunction支持简单运算和分布式向量运算



➤ 参数服务器WeiPS-算法

- ASGD
- L-BFGS
- OWL-QN
- FTRL



➤ 参数服务器WeiPS-优化

➤ PS sever count

➤ Batch size

➤ Msg 序列化方式

➤ 参数同步比例

➤ 大规模深度学习

➤ 深度学习平台分层架构

应用(人脸识别/CTR/...)

模型(DNN/CNN/RNN/...)

框架(Tensorflow/Caffe/Kaldi/...)

调度(K8s/Mesos/Yarn/...)

基础库(CUDA/CuDNN/NCCL/...)

硬件(GPU/FPGA/...)

➤ 大规模深度学习-调度

➤ Tensorflow on K8S

相对
成熟

灵活
配置

MPI

➤ 大规模深度学习-框架

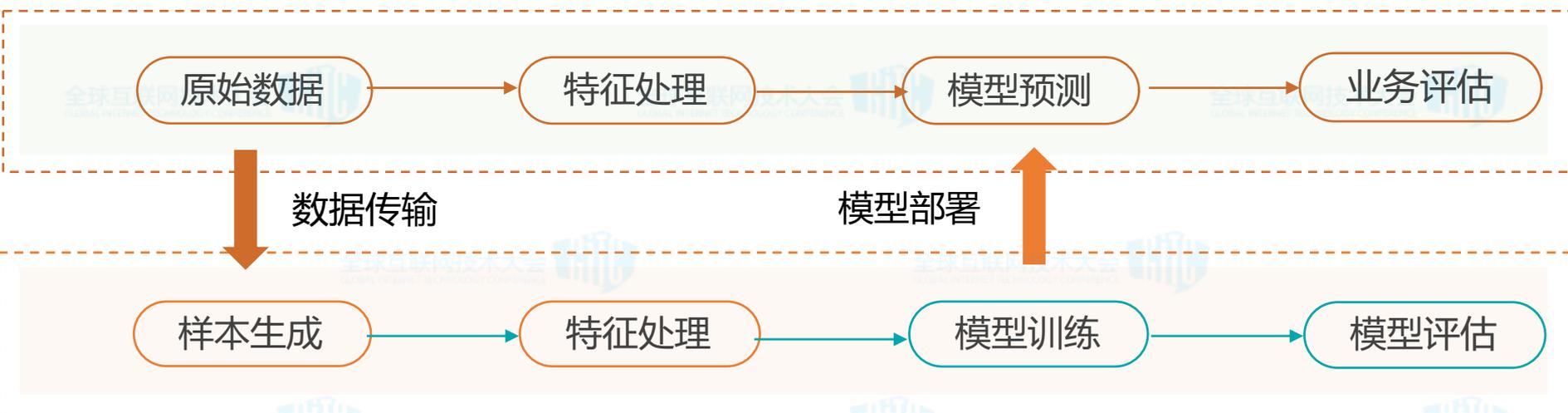
➤ Tensorflow on MPI



版本	TF1.1	TF1.1	TF1.4	TF on MPI
优化方向	IO优化	IO优化	通信优化	通信优化
主要内容	引入pydoop	多进程替换多线程	grpc版本升级	MPI替换gprc 引入NCCL2支持多GPU通信

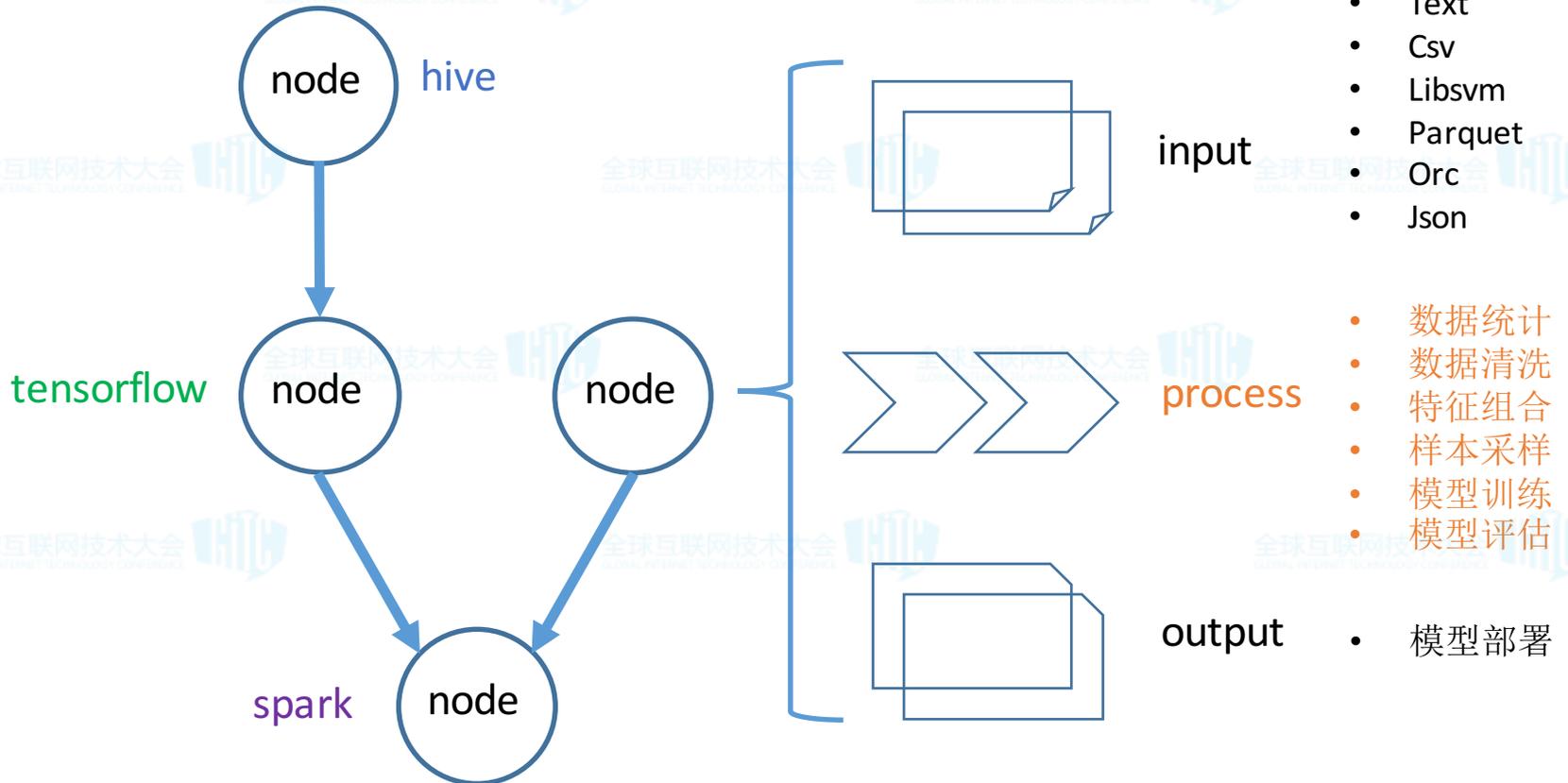
➤ 机器学习 workflow

- 标准化
- 机器学习 workflow 框架 WeiFlow



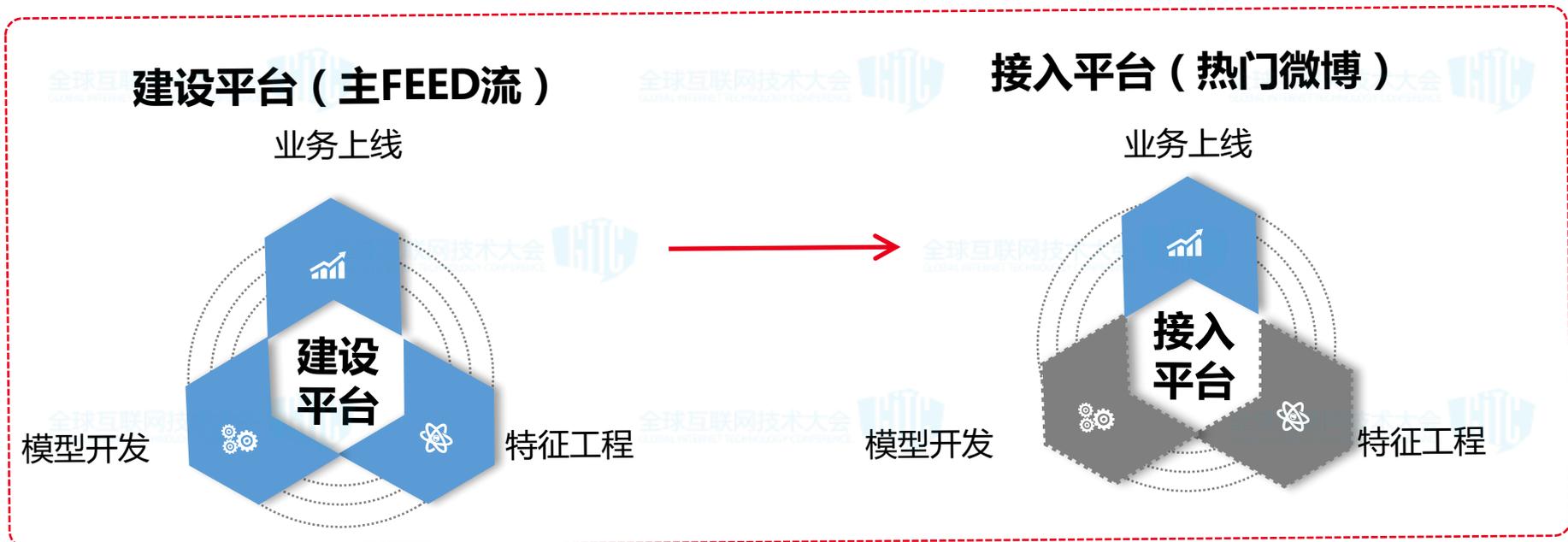
机器学习 workflow 框架 WeiFlow

- 兼容异构环境
- 统一数据计算框架



全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE





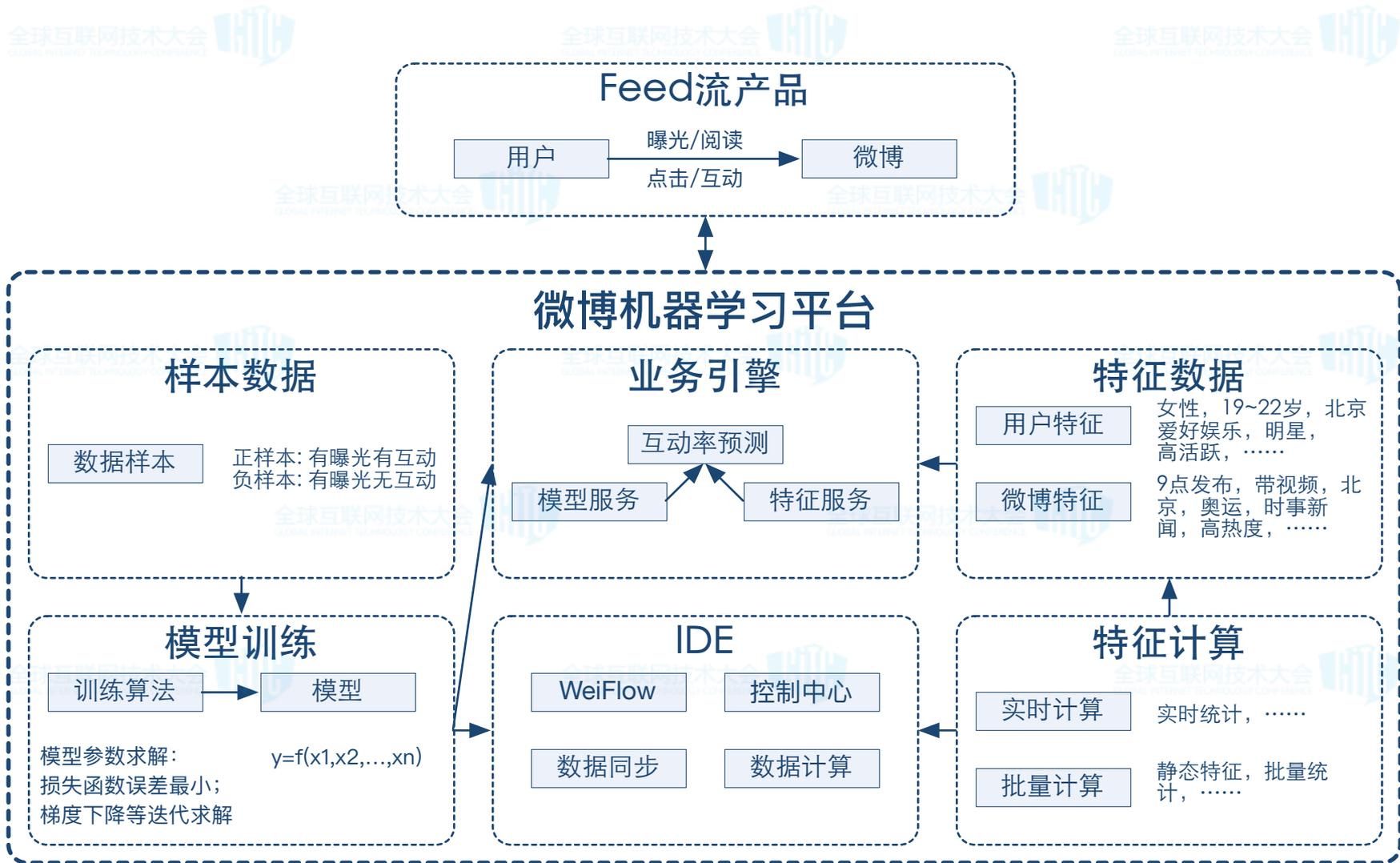
微博Feed

➤ Feed流-主信息流

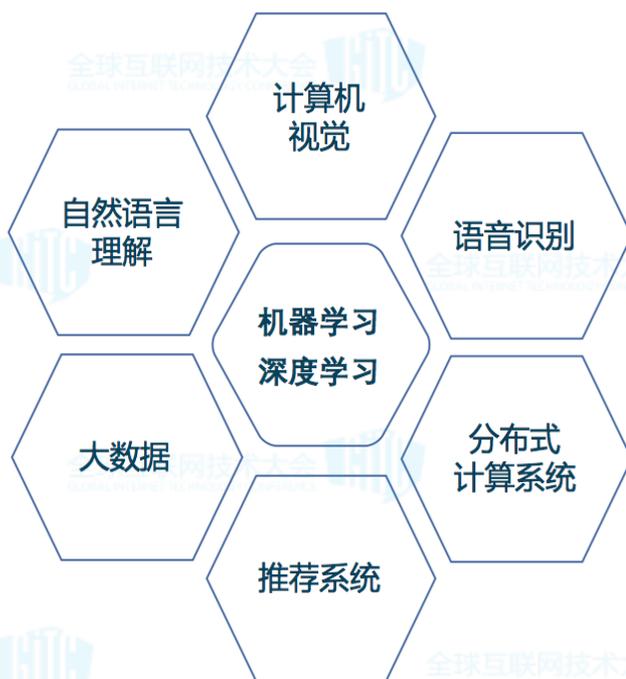
- 文本
- 短视频
- 图片
- 长文
- 问答



业务实践 - Feed机器学习排序



➤ 微博机器学习团队 & AILab 诚聘英才



扫描二维码



简历投递

✉ ailab@weibo.com

全球互联网技术大会
GLOBAL INTERNET TECHNOLOGY CONFERENCE

谢谢!