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

目录

- 微博
- 微博机器学习平台
  - 大规模机器学习
  - 大规模深度学习
  - 机器学习工作流
  - 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

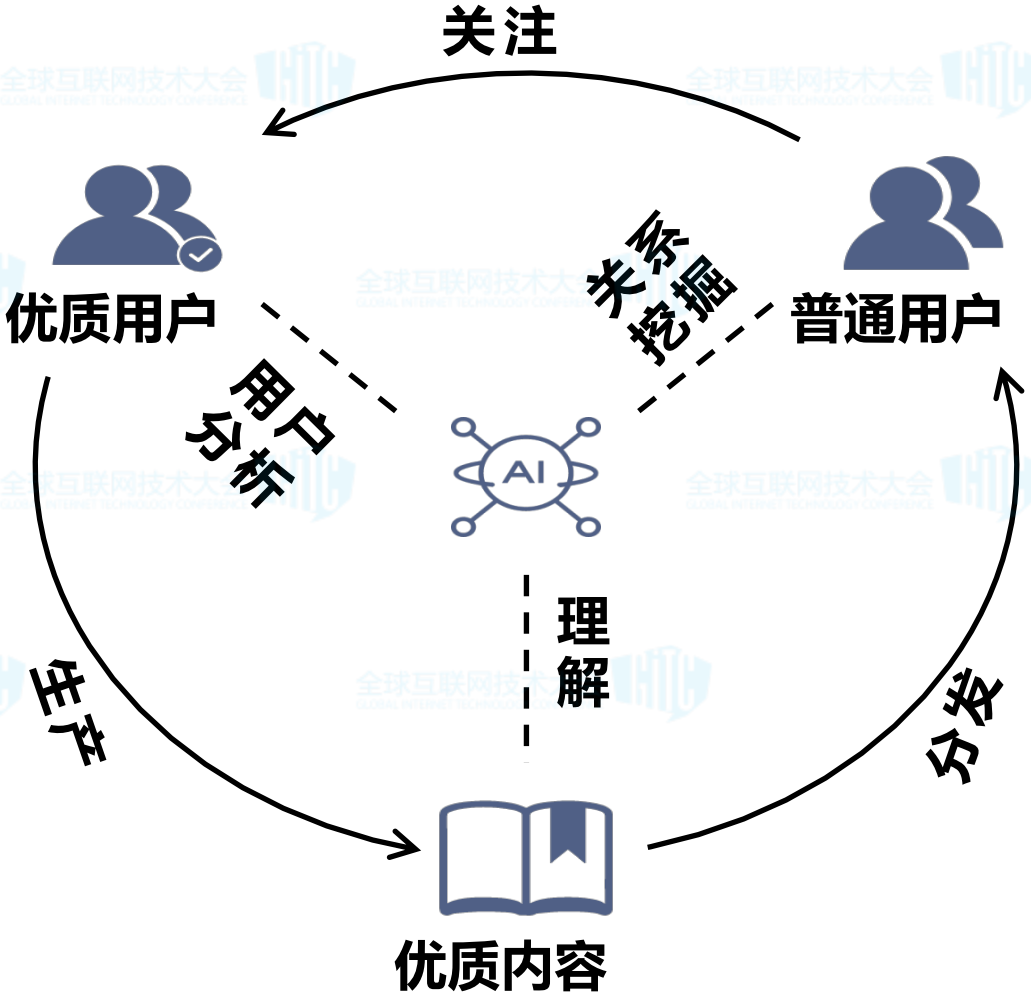
8700万

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

高频访问

用户体量大

用户间关系纷杂

微博内容体量大

微博内容数据多样
(文本、图片、音频、视频，等)

大规模

特征维度巨大

特征类别多

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,...

标准化

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

平台化

计算框架多样性

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

执行性能差

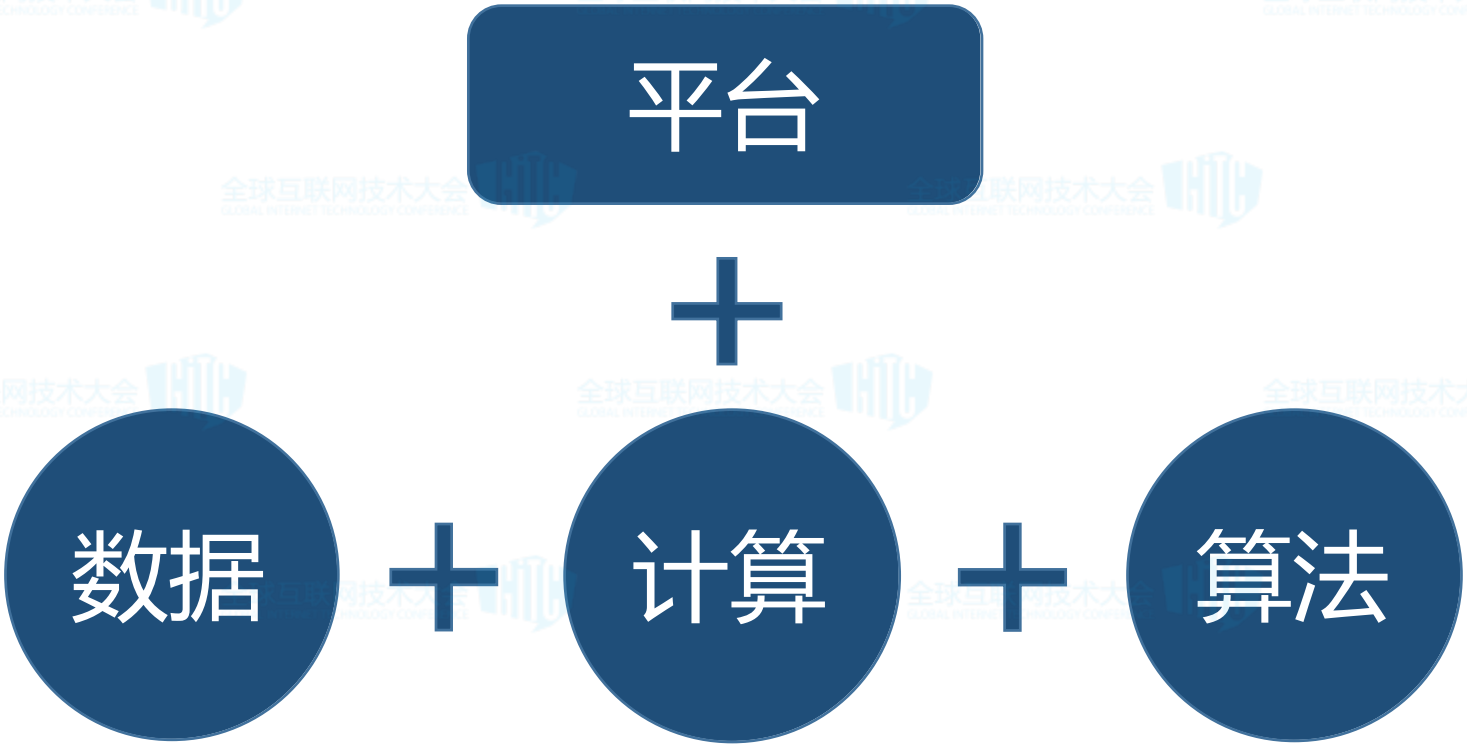
系统运行门槛高

业务程序依赖多

重复建设成本

人力成本

微博机器学习平台

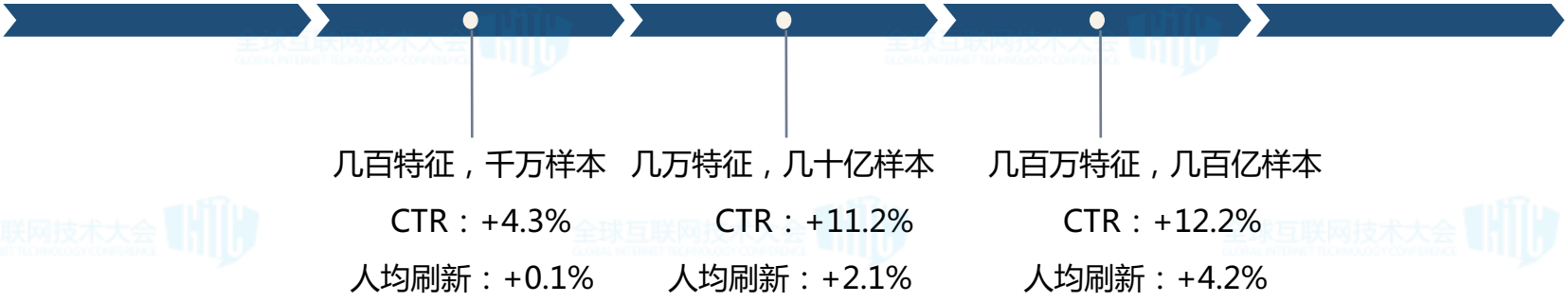


微博机器学习平台 – 算法

- 大规模机器学习
- 大规模深度学习
- 机器学习工作流

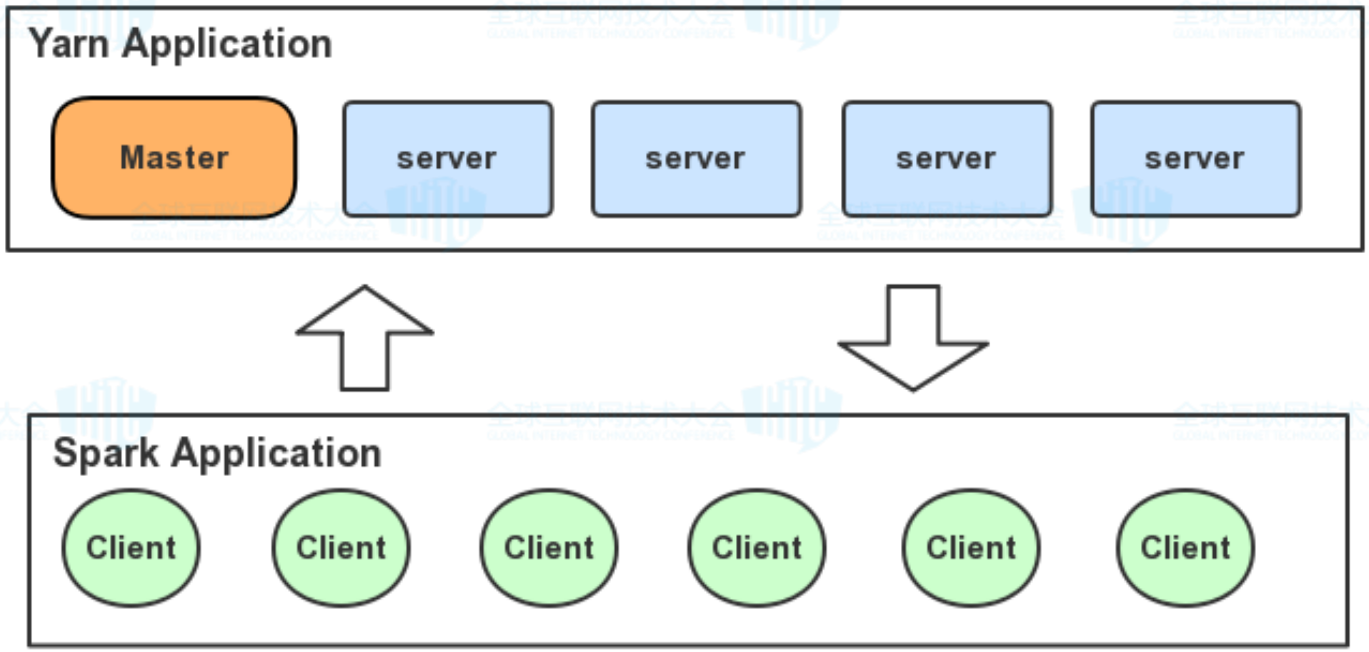
微博机器学习平台 - 大规模机器学习

- 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



微博机器学习平台 - 大规模机器学习

- 参数服务器WeiPS-算法
  - ASGD
  - L-BFGS
  - OWL-QN
  - FTRL

	ASGD	L-BFGS	OWLQN	FTRL
收敛精度	1E-8	1E-6	1E-6	1E-8
迭代次数	9次	100次	100次	5次
AUC	0.85	0.9	0.9	0.88
正则化	L2 0.001	L2 0.1	L1 0.005	L2 0.001
执行性能	30m	1h22m	1h30m	14m

微博机器学习平台 - 大规模机器学习

- 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

微博机器学习平台 - 大规模深度学习

- 大规模深度学习
  - 深度学习平台分层架构

微博机器学习平台 - 大规模深度学习

- 大规模深度学习-调度
  - Tensorflow on K8S

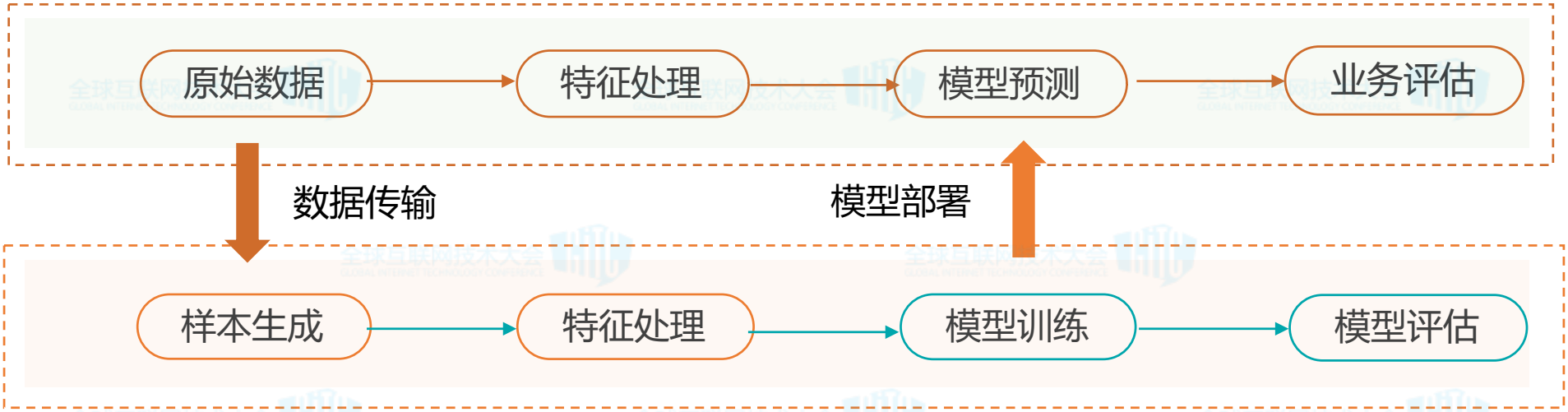
微博机器学习平台 - 大规模深度学习

- 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

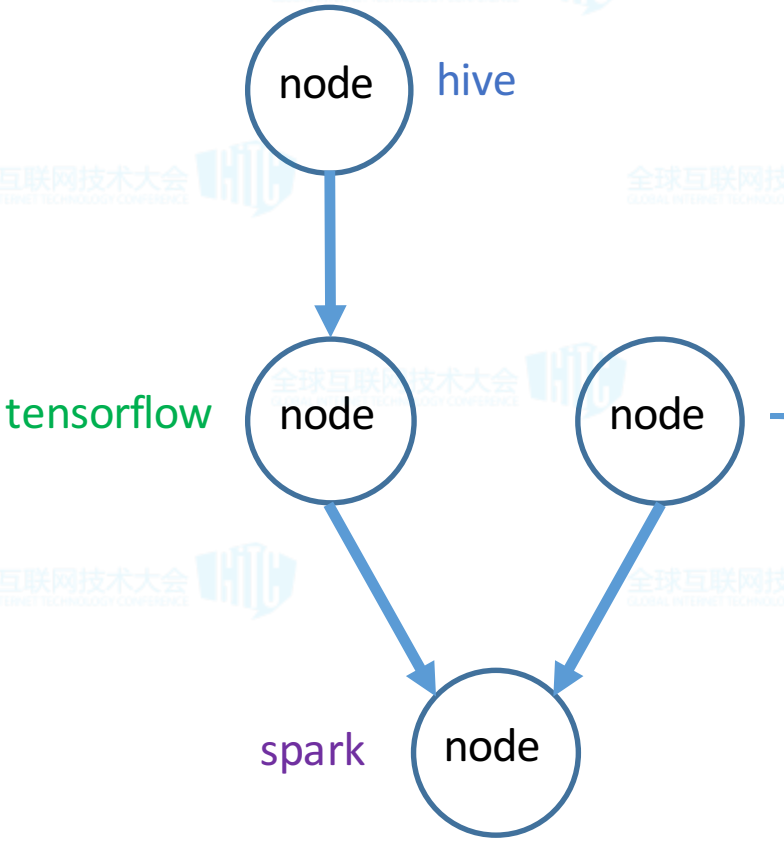
微博机器学习平台 - 机器学习工作流

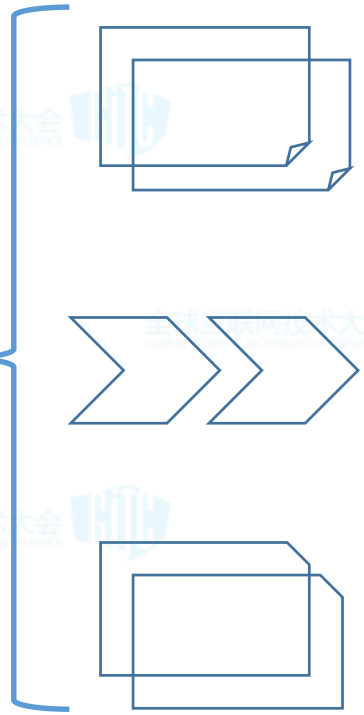
- 机器学习工作流
  - 标准化
  - 机器学习工作流框架WeiFlow



微博机器学习平台 - 机器学习工作流

- 机器学习工作流框架WeiFlow
  - 兼容异构环境
  - 统一数据计算框架





input

- Text
- Csv
- Libsvm
- Parquet
- Orc
- Json

process

- 数据统计
- 数据清洗
- 特征组合
- 样本采样
- 模型训练
- 模型评估

output

- 模型部署

微博机器学习平台 – 平台效果

效果

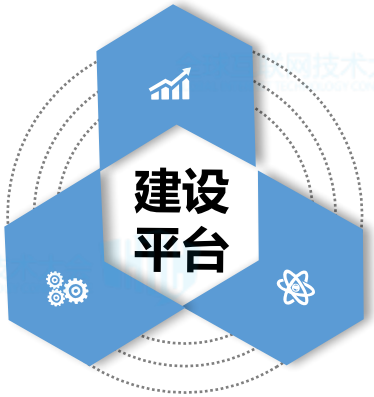
效率

成本

微博机器学习平台 – 平台效果

建设平台（主FEED流）

业务上线

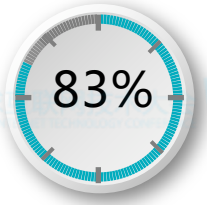


模型开发　　特征工程

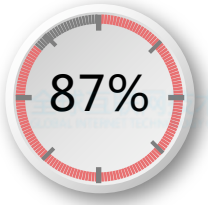
接入平台（热门微博）

业务上线

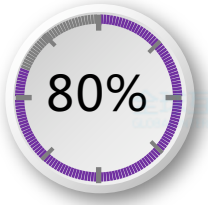
模型开发　　特征工程



人力成本



时间成本



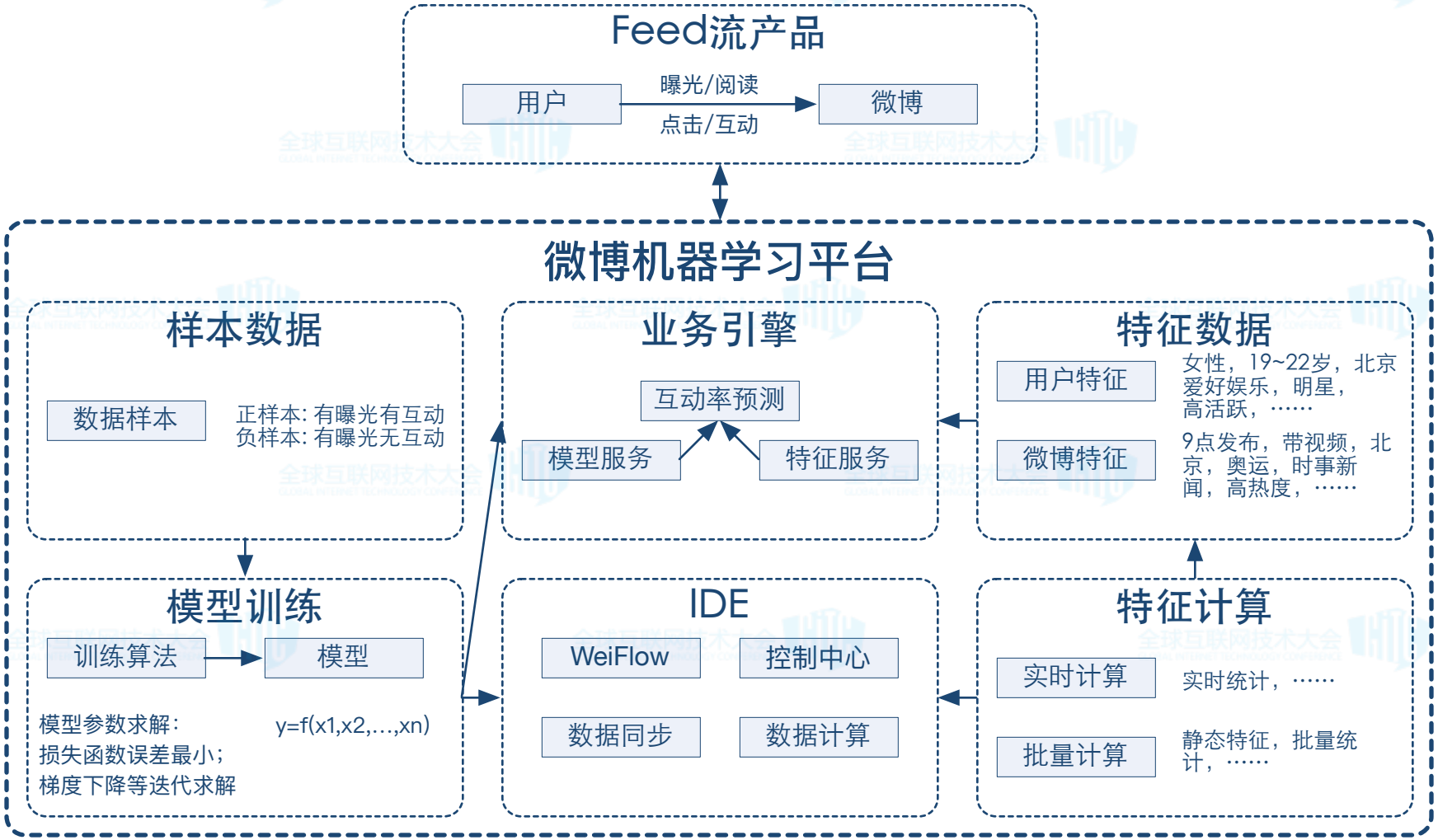
机器成本

业务实践 – Feed机器学习排序

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



人工智能在社交媒体领域大有可为

➢ 微博机器学习团队 & AI Lab 诚聘英才

扫描二维码

简历投递

☛ ailab@weibo.com

谢谢！