面向机器学习的数据平台设计与搭建

袁凯
个推 大数据架构师

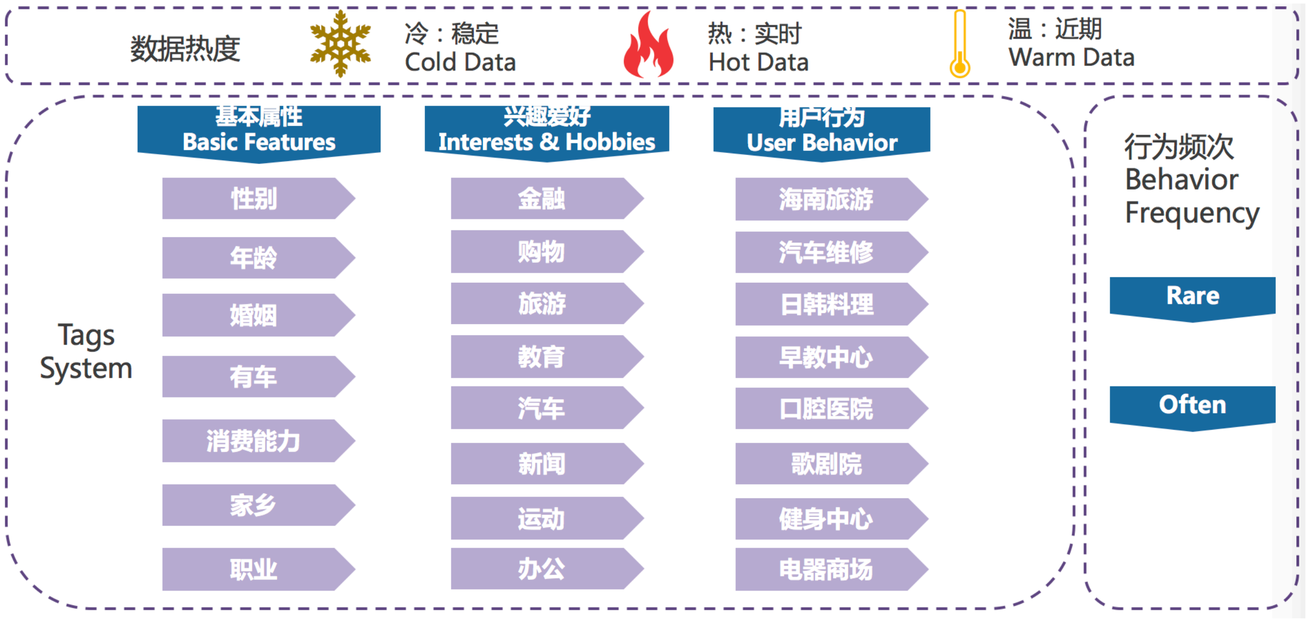
TABLE OF

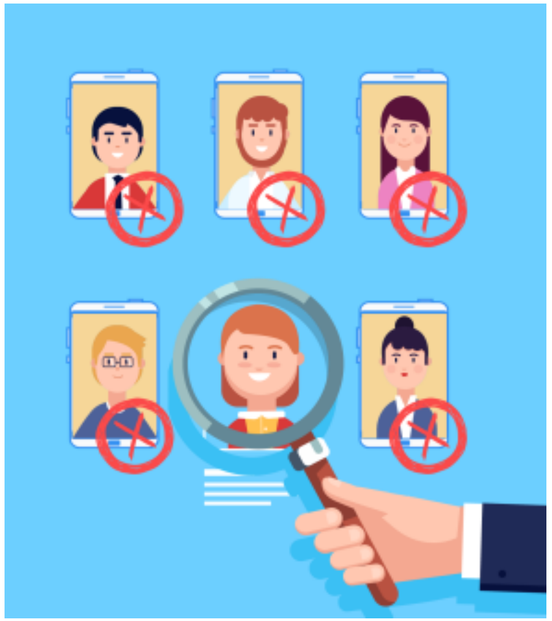
CONTENTS 大纲

- 背景简介
- 机器学习的落地问题
- 机器学习平台建设
- 总结与未来

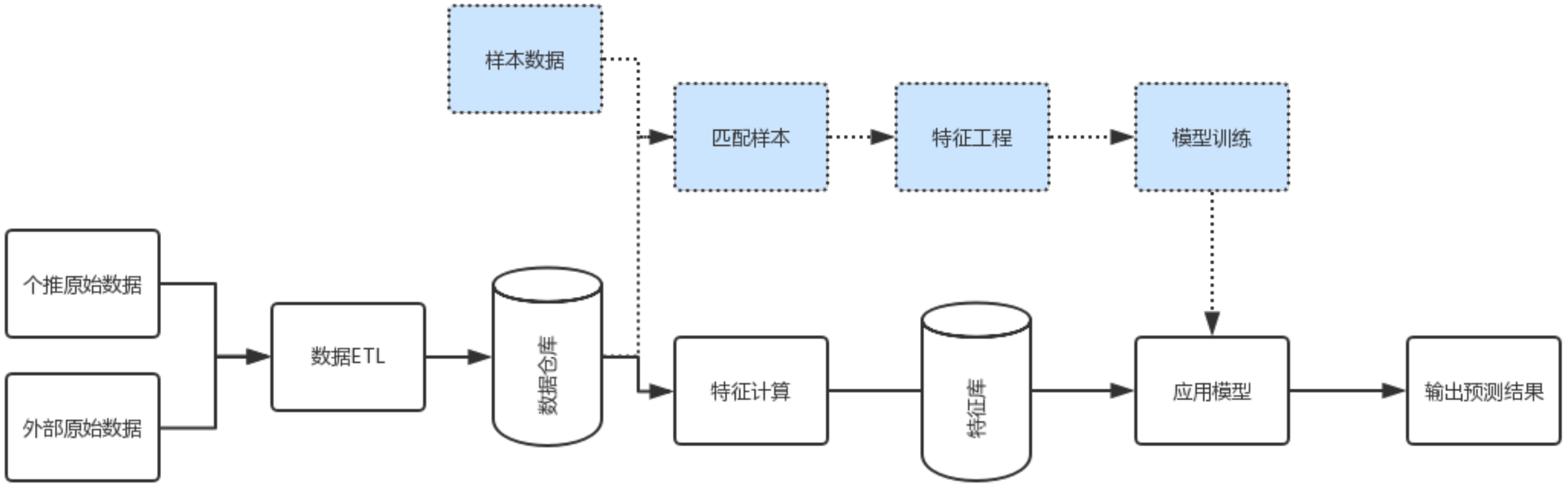
机器学习 in 个推

- 用户精准画像
- 智能推送+精准营销
- 商圈景区人流预测
- 虚假设备识别
- 个性化推荐
- APP用户流失预测
- ……





机器学习 in 个推


样本数据
匹配样本
特征工程
模型训练
个推原始数据
外部原始数据
数据ETL
数据仓库
特征计算
特征库
应用模型
输出预测结果

机器学习标准工作流

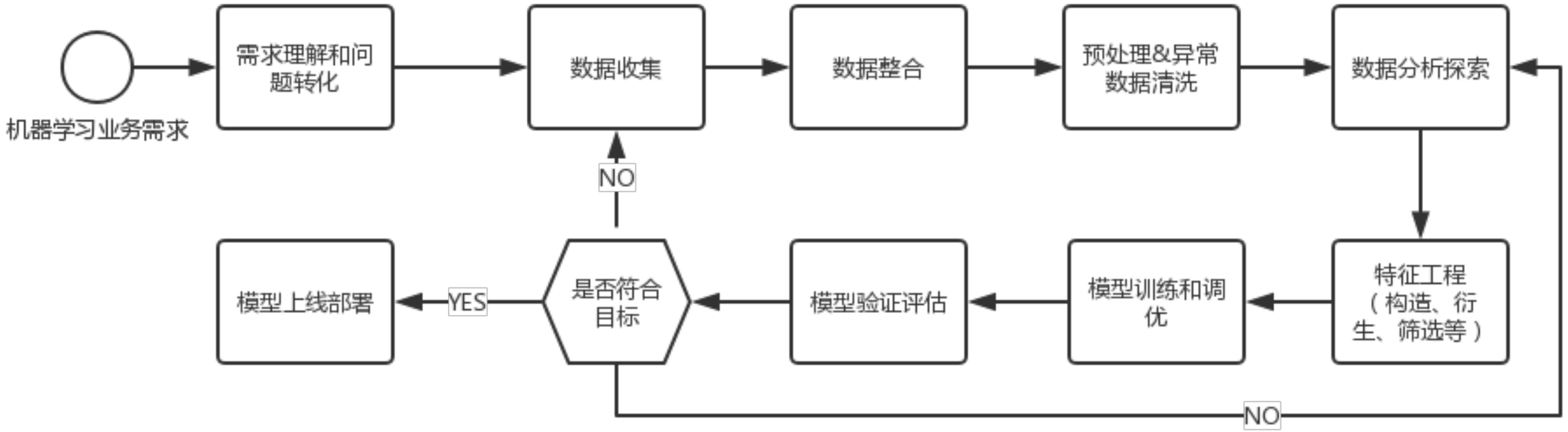

机器学习业务需求
需求理解和问题转化
数据收集
数据整合
预处理&异常数据清洗
数据分析探索
特征工程（构造、衍生、筛选等）
模型训练和调优
模型验证评估
是否符合目标
NO
YES
模型上线部署
NO

TABLE OF

CONTENTS 大纲

- 背景简介
- 机器学习的落地问题
- 机器学习平台建设
- 总结与未来

落地机器学习时的问题（一）

- 大数据时代建模人员需要熟悉很多大数据组件使用。

  例如：Spark，Hive， Hadoop，Yarn等。

- 海量数据下业务相关数据（用户行为数据、属性等）匹配抽取效率低下。

- 建模工具不统一，机器学习pipeline 不统一，代码重复率高，建模过程不能很好沉淀，缺少标准化监控。

- 建模人员大多聚焦在模型实验阶段，不擅长工程实现，依赖工程开发人员翻译集成最终的pipeline。

落地机器学习时的问题（二）

- 数据探索时数据种类繁多，使用数据成本高。
- 相同特征存在重复构建，比较好用特征未得到广泛应用。
- 机器学习快速推广困难。

TABLE OF

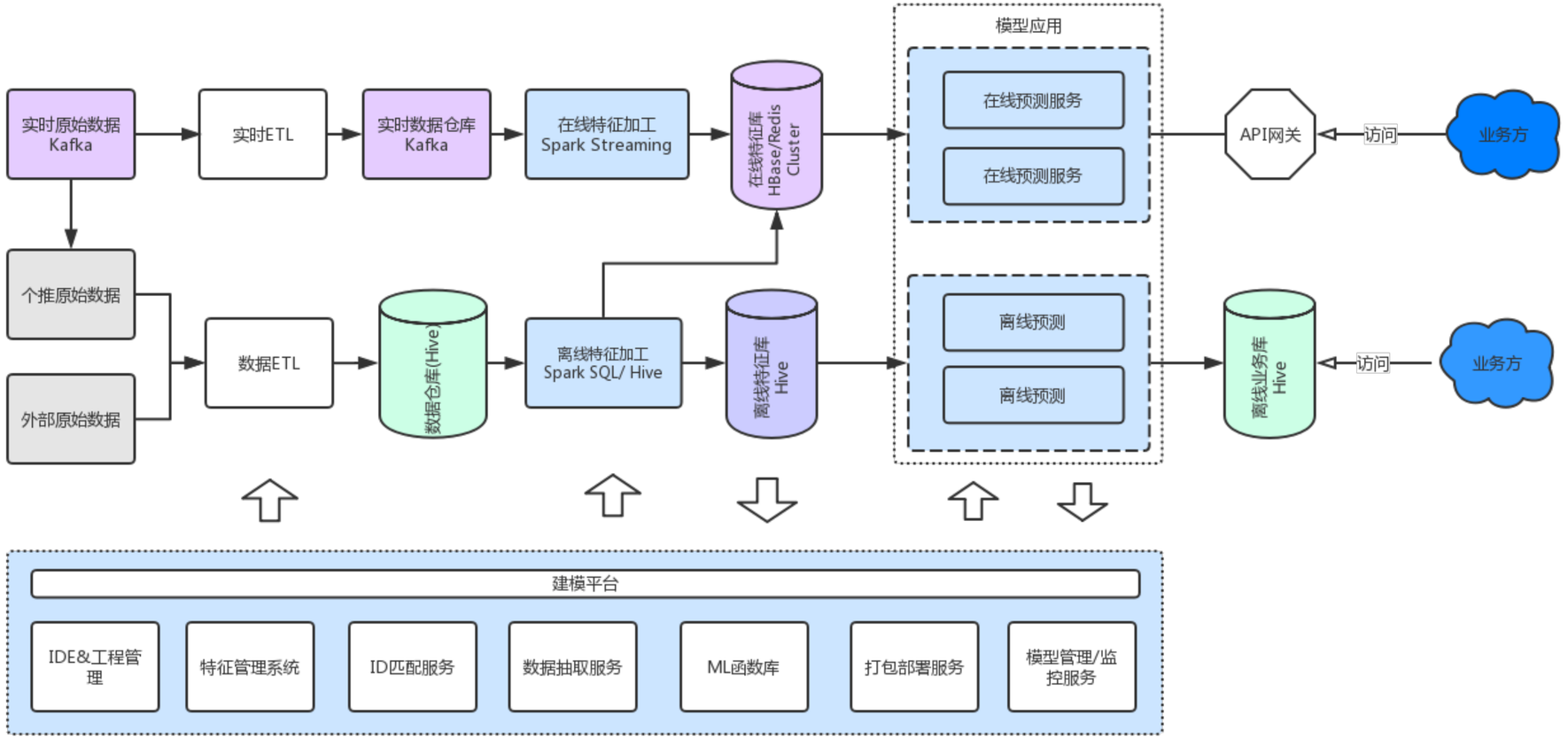
CONTENTS 大纲

- 背景简介
- 机器学习的落地问题
- 机器学习平台建设
- 总结与未来

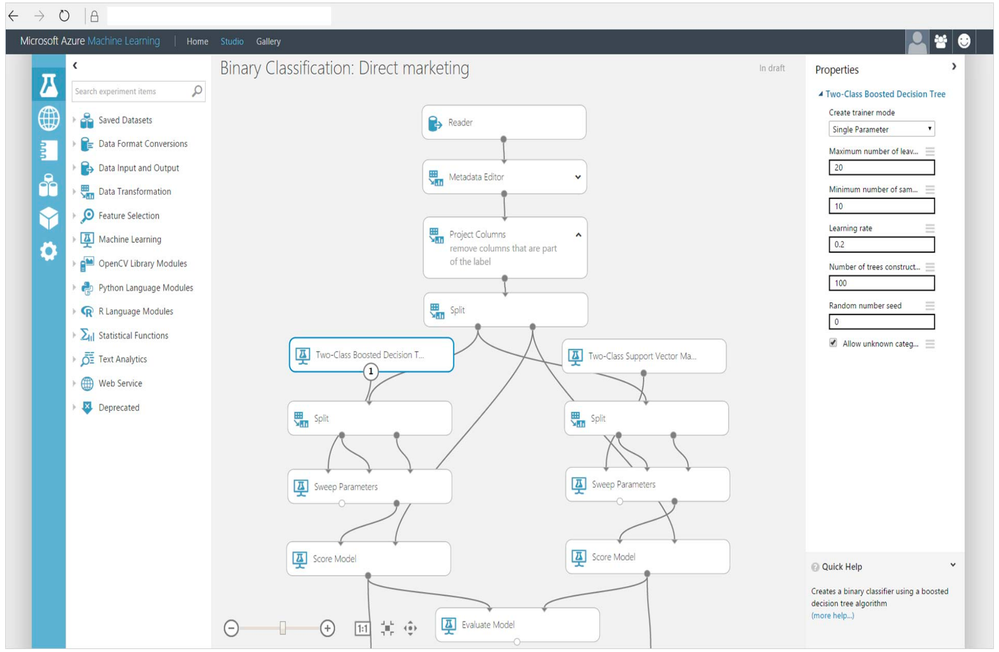
个推机器学习平台的目标

- 内部建模流程规范化。
- 模型开发到上线应用的全流程支持。
- 特征数据可运营、可共享。
- 面向专家和半专家，提高建模效率。
- 支持多租户。
- 数据安全。

个推机器学习平台方案


模型应用
实时原始数据 Kafka
实时ETL
实时数据仓库 Kafka
在线特征加工 Spark Streaming
在线特征库 HBase/Redis Cluster
在线预测服务
在线预测服务
API网关
访问
业务方
个推原始数据
外部原始数据
数据ETL
数据仓库(Hive)
离线特征加工 Spark SQL/ Hive
离线特征库 Hive
离线预测
离线预测
离线业务库 Hive
访问
业务方
建模平台
IDE&工程管理
特征管理系统
ID匹配服务
数据抽取服务
ML函数库
打包部署服务
模型管理/监控服务

统一分析建模交互工具 - Jupyter

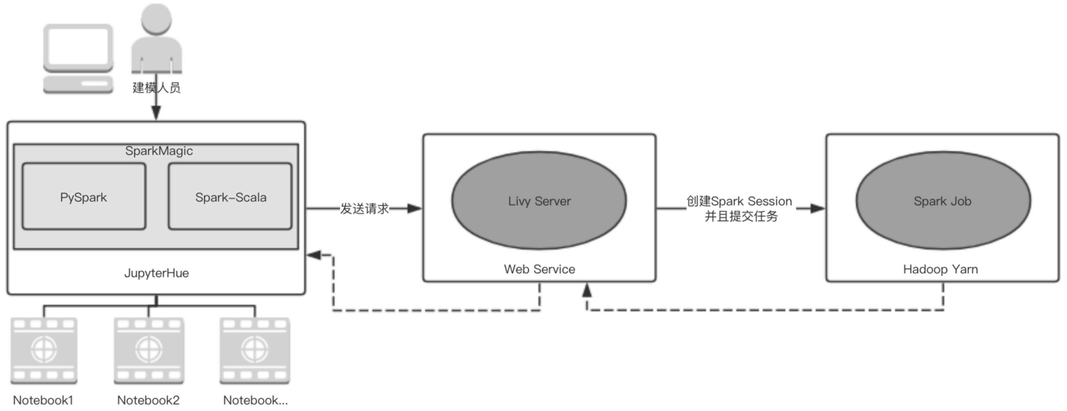


图片来自:https://azure.microsoft.com/zh-cn/services/machine-learning-studio

- 选择Jupyter：简单、高效、易扩展、文档代码一体。

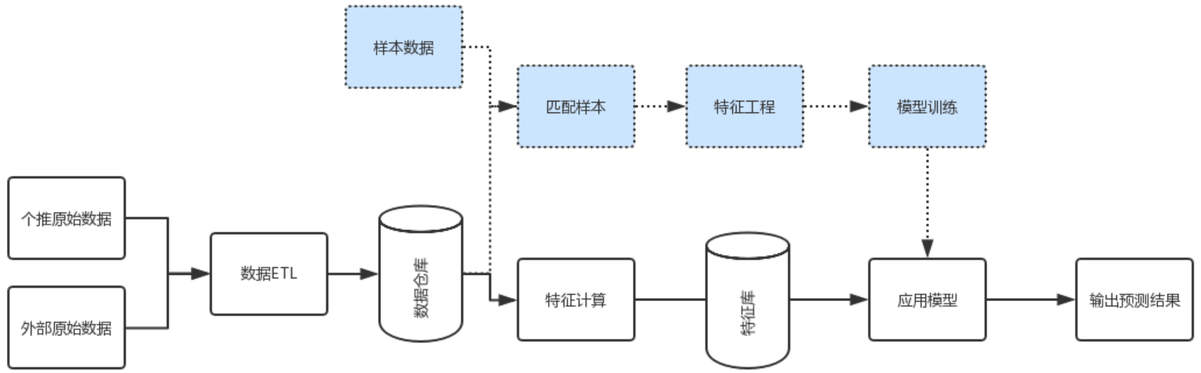
统一分析建模交互工具 - Jupyter

- 多租户：Jupyterhub。
- 机器学习库: Tensorflow、Pyspark、sk-learn等。
- 交互式Spark：SparkMagic+Livy。
- Notebook管理与共享：Git。
- 插件：扩展模板化，tensorboad插件，自定义插件等。
- ……



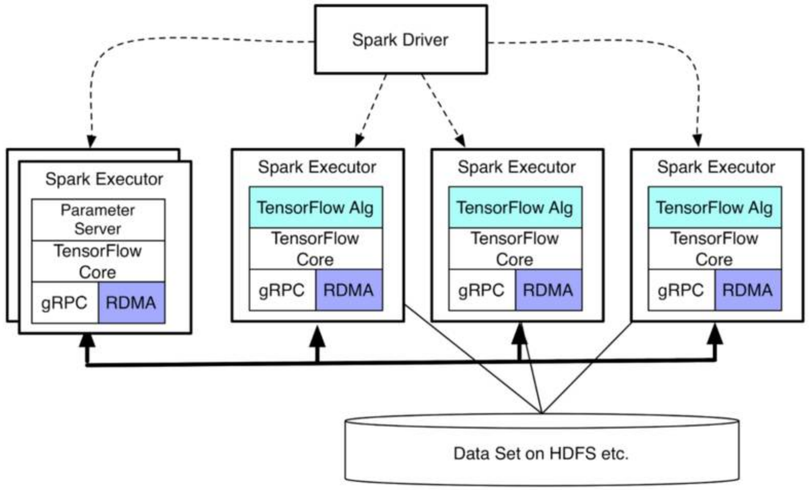
内部统一ML函数库

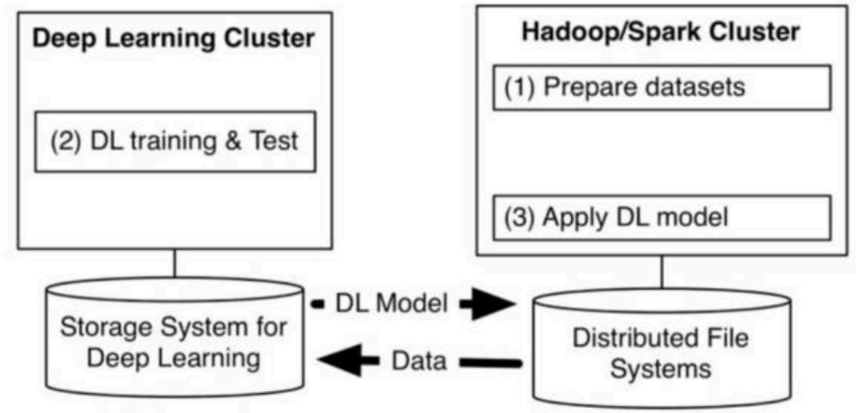
- 标准化IDMapping
- 数据抽取
- 特征分析与可视化
- 可视化工具
- Notebook转AzkabanFlow
- ……



TensorflowOnSpark

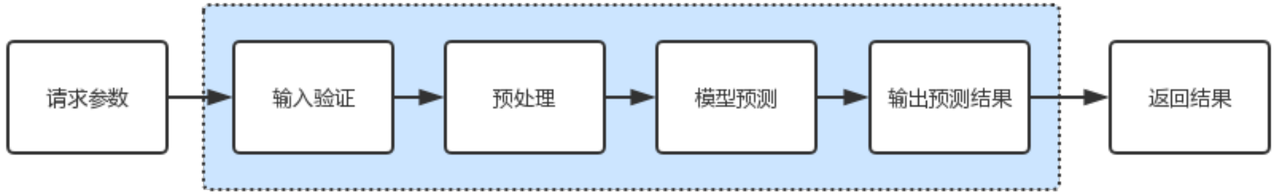
- 资源调度透明，解决原生Tensorflow Cluster分布式问题。
- 迁移TensorflowOnSpark简单，代码基本不用改。
- 支持GPU和CPU混部集群，资源易复用。
- 团队对hadoop、spark熟悉。



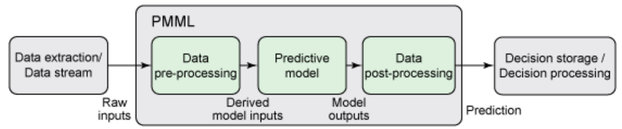


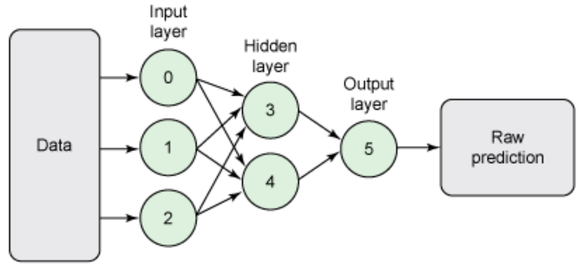
模型部署应用

- 交付部署方便
- 流程规范化



1. 输入参数结构规范化。
2. 反馈格式规范化。
3. 预测代码框架化（输入、验证、特征处理、预测、返回）。
4. 支持标准化模型文件（pmml、tensorflow pb等）。





模型部署应用

- 具体方式

1. Docker + Spring Boot+ PMML
2. Docker + Flask + Tensorflow Serving
3. Spark SQL/Hive + Java/Python UDF + PMML

TABLE OF

CONTENTS 大纲

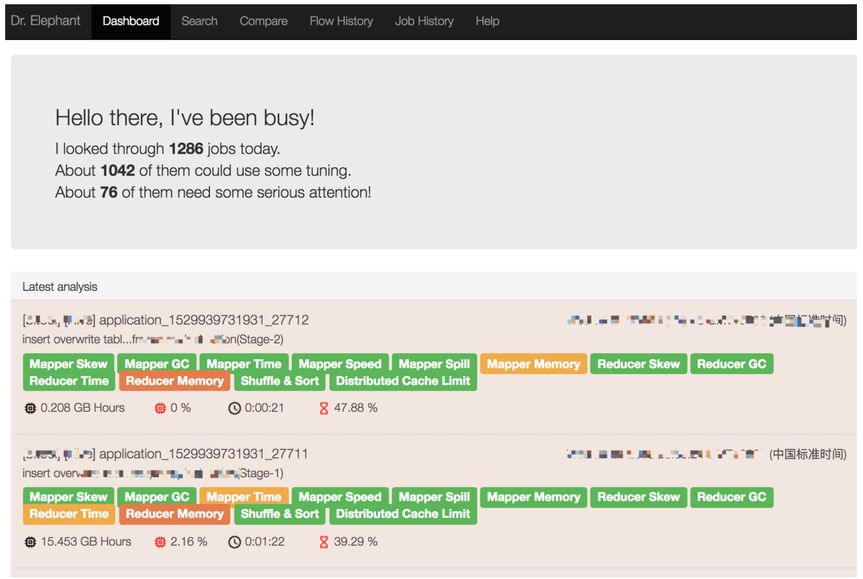
- 背景简介
- 机器学习的落地问题
- 机器学习平台建设
- 总结与未来

经验与总结

- TensorflowOnSpark : PS数量、资源使用优化、模型导出。
- Jupyter使用: Sparkmagic、本地函数库、远程函数库。
- PMML：性能、扩展性。

经验与总结

- Spark/Hive 性能 (Dr.elephant 诊断）。
- 亿级数据下的Spark SQL构建特征（避免shuffle操作）。
- 特征工程代码复用（Pipeline框架化开发）。
- 特征库和模型运营（可见共享、价值评估、稳定性）。



经验与总结

- 硬件选型注意事项（带宽、内存、GPU）。
- HBase运维（监控、region合并）。
- 平衡需求与技术栈分裂

未来

- TensorFlow on Kubernetes。
- 系统内各个组件融合完善。
- 机器学习相关周边工具补齐和增强。

THANKS