Security Level:

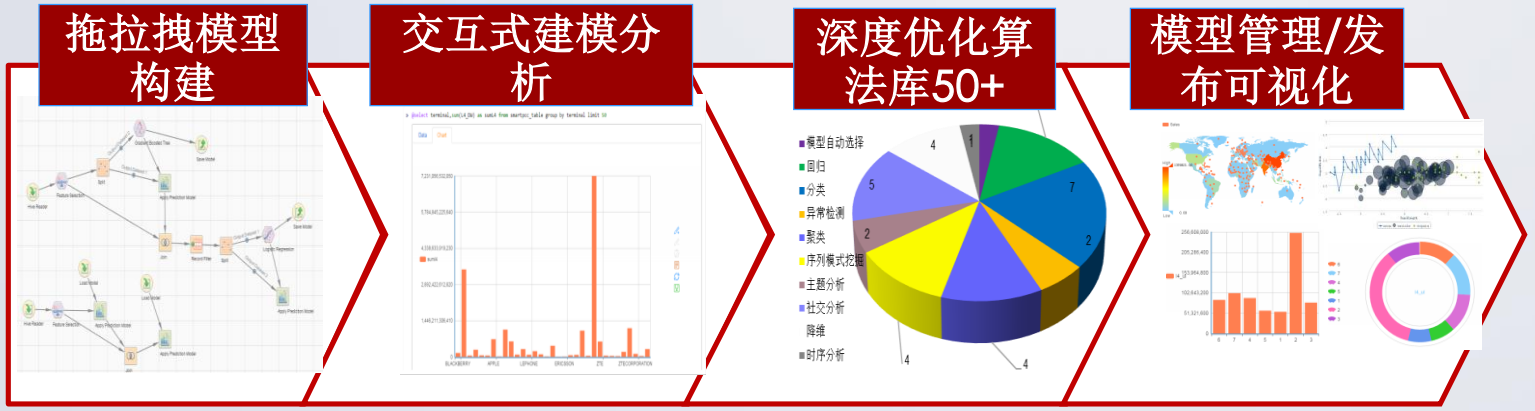
华为FusionInsight Miner
产品介绍

www.huawei.com

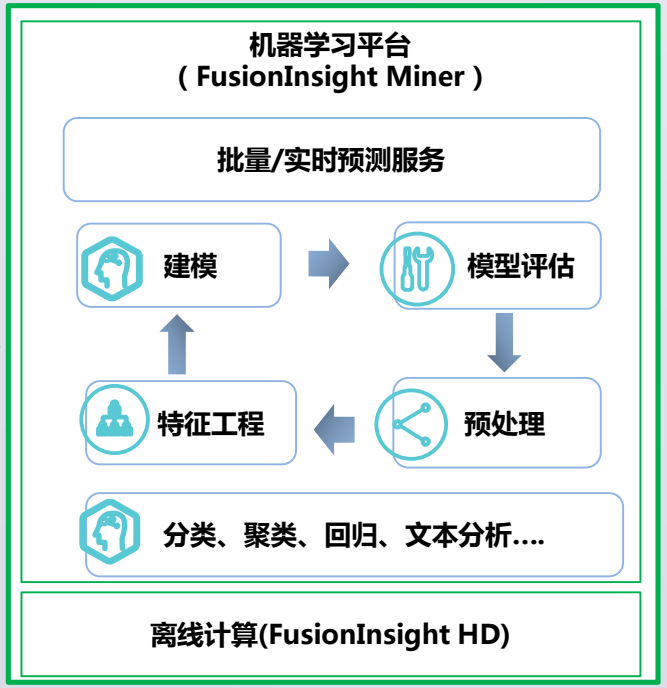
HUAWEI

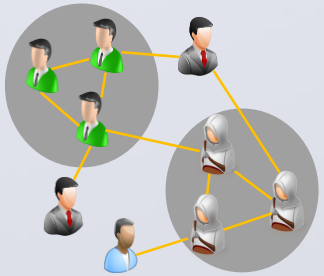
目录

FusionInsight Miner 概述



用户数据：年龄、收入等





用户画像

机器学习平台

业务场景

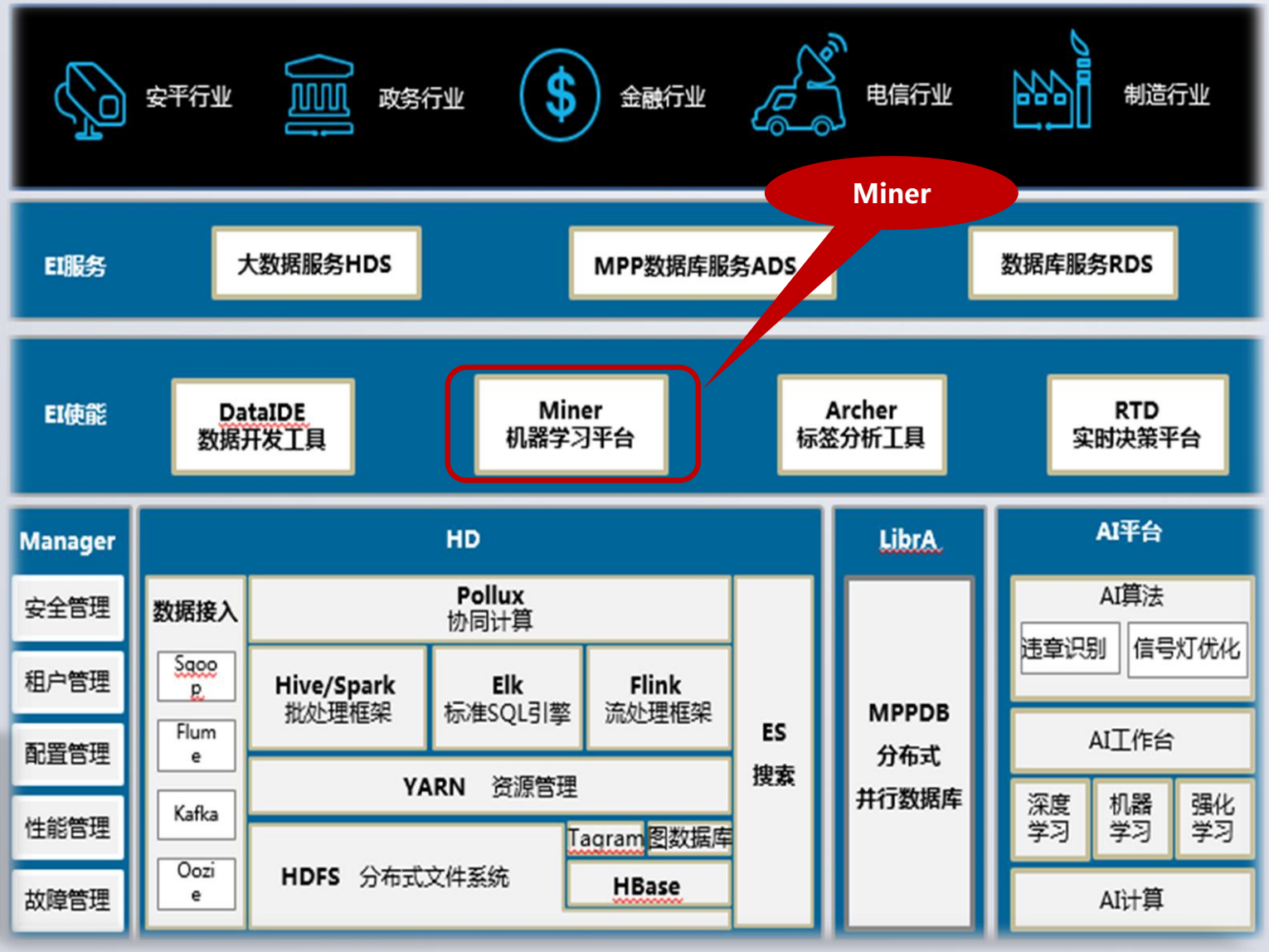
- **极大团分析**
- **反电信诈骗、拟合分析**
- **金融交易实时反欺诈**

	特性
	支持R语言
	支持Python2、Python3
	工作流支持自定义算法
	支持对接推理平台
	支持Elk、Sparksql数据源
	内置50+多种算法

业务价值

- **易用：可视化界面**降低AI模型开发门槛，提升开发效率
- **预置模型：**预置常用模型，快速实现业务建模，并支持增量训练
- **开放：**支持主流开发语言、对接多种数据源

Miner在FusionInsight大数据解决方案中的位置



FusionInsight

- 华为企业级大数据存储、查询、分析的统一平台，能够帮助企业快速构建海量数据信息处理系统，通过对海量信息数据实时与非实时的分析挖掘，发现全新价值点和企业商机；
- 由FusionInsight HD、FusionInsight LibrA、FusionInsight Miner、FusionInsight Farmer和1个操作运维系统FusionInsight Manager构成

FusionInsight Miner

- 企业级的机器学习平台，基于华为FusionInsight HD的分布式存储和并行计算技术，提供从海量数据中挖掘出有价值信息的平台；
- 更易用，更开放，更丰富，提供特征工程、算法、建模、预测、模型管理的一站式机器学习应用

FusionInsight Miner 业务架构

数据探索

提供Web化的自由探索界面，用户可以使用自己熟悉的脚本语言交互式探索各类数据，并将探索结果通过可视化的图形展示出来，大大提高数据分析师的探索效率。

特征工程

提供了大量提取特征的算子和功能，满足绝大部分特征工程的需要，尤其适合高维稀疏大数据的特征处理。

建模分析

建模分析模块集成常用分类、聚类、推荐算法，可以支持个性化营销、个性化推荐场景的预测模型。

目录

部署流程

Miner基于HD集群之上搭建

Miner安装流程方便快捷

具体只需要两步：
1.调用mangaer进行注册组件
2.在manager界面添加Miner服务

部署方案

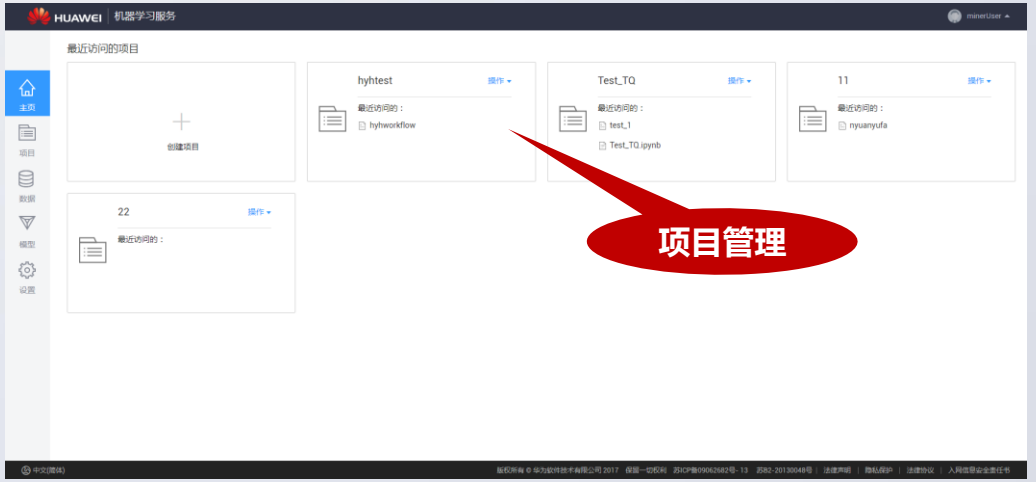
Miner部署需求：

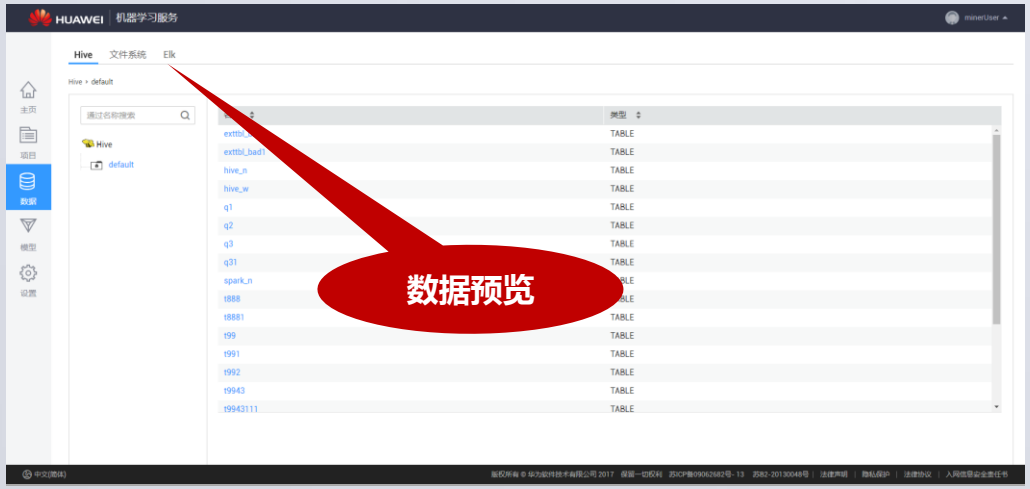
属性	描述
角色	MinerServer
最小内存要求	2GB
部署原则	主备部署在两个**控制节点**上
依赖组件	Hive、Spark2x、Yarn、HDFS、DBService、KrbServer、LdapServer

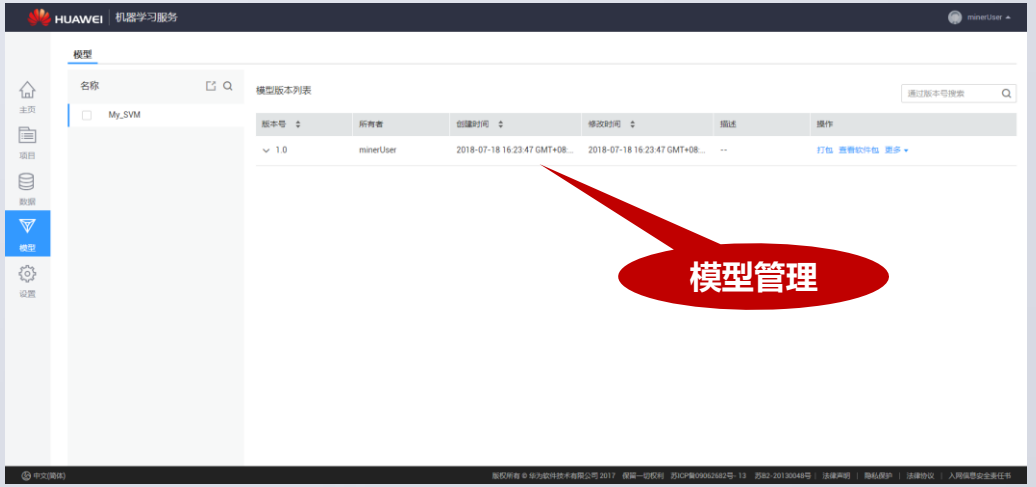
目录

Web管理界面，提供项目、模型、数据管理，让管理更方便

- 提供了大量提取特征的算子和功能，满足绝大部分特征工程的需要，尤其适合高维稀疏大数据的特征处理。

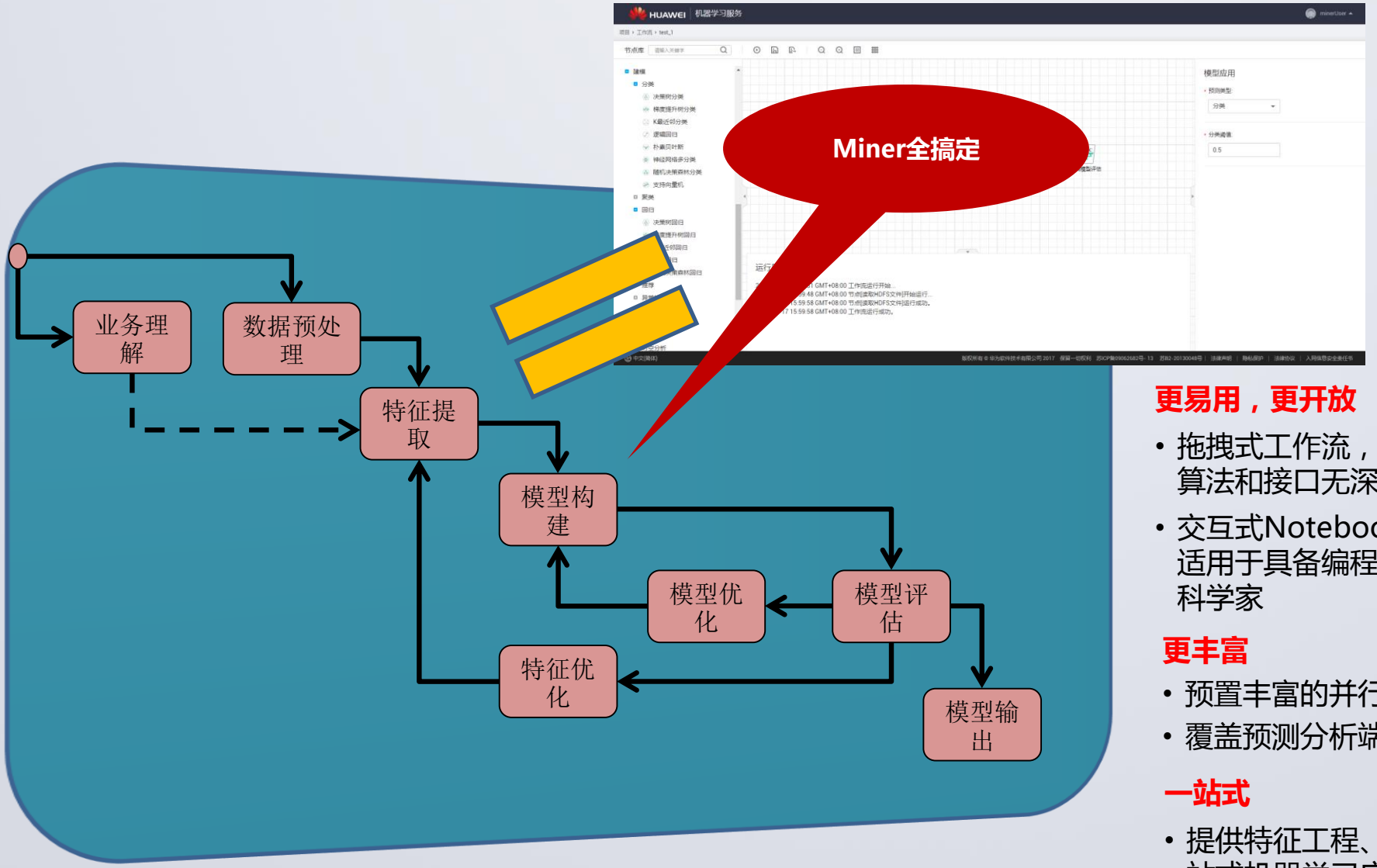






HUAWEI

一站式机器学习应用，提供特征工程、算法、建模、预测、模型管理



更易用，更开放

- 拖拽式工作流，可以直观的展示处理流程，适合对算法和接口无深入了解的数据分析业务人员
- 交互式Notebook，能够灵活编写代码，交互性强，适用于具备编程能力，了解数据分析及算法的数据科学家

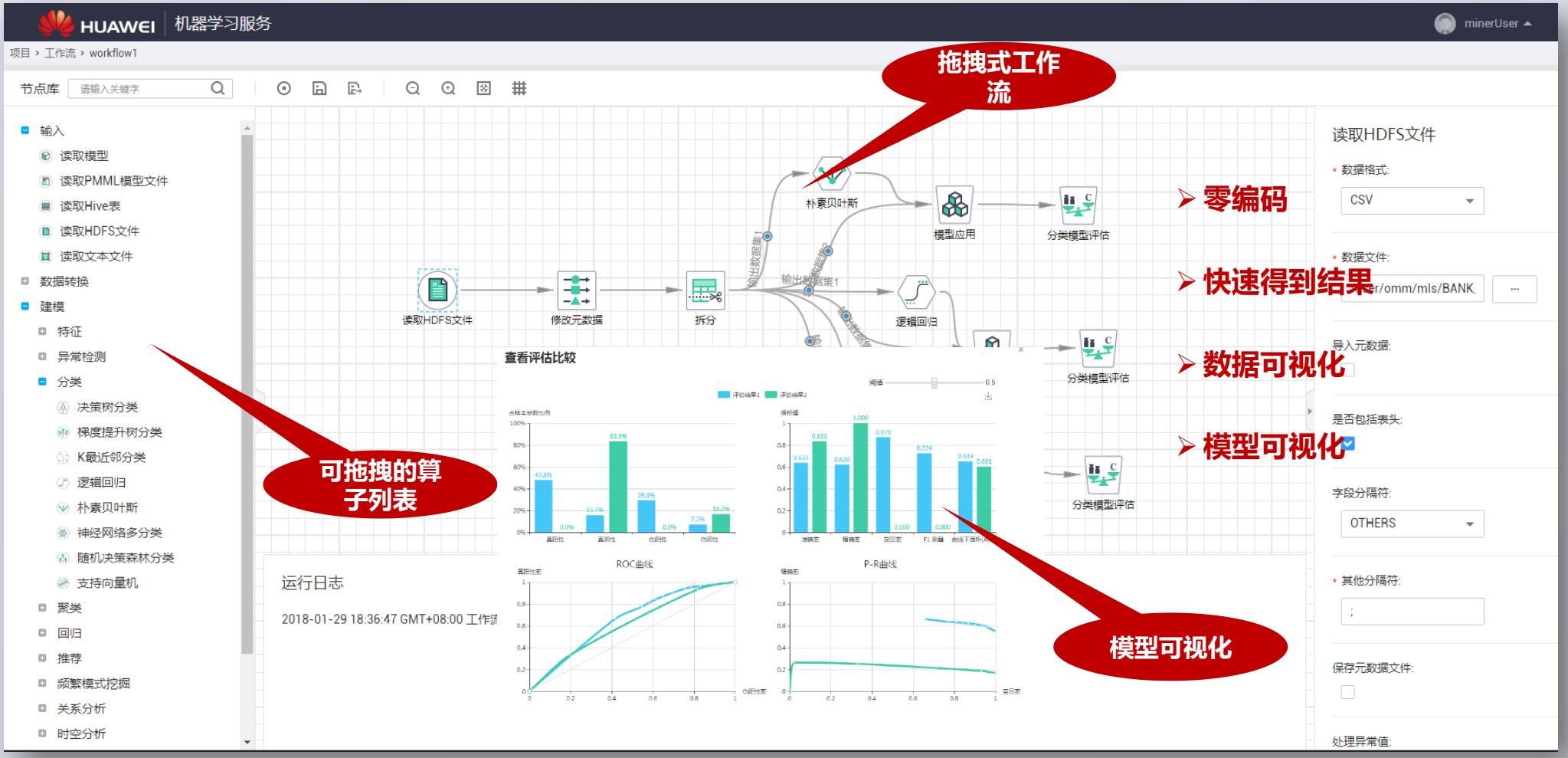
更丰富

- 预置丰富的并行化机器学习算法库，适合各种场景
- 覆盖预测分析端到端业务

一站式

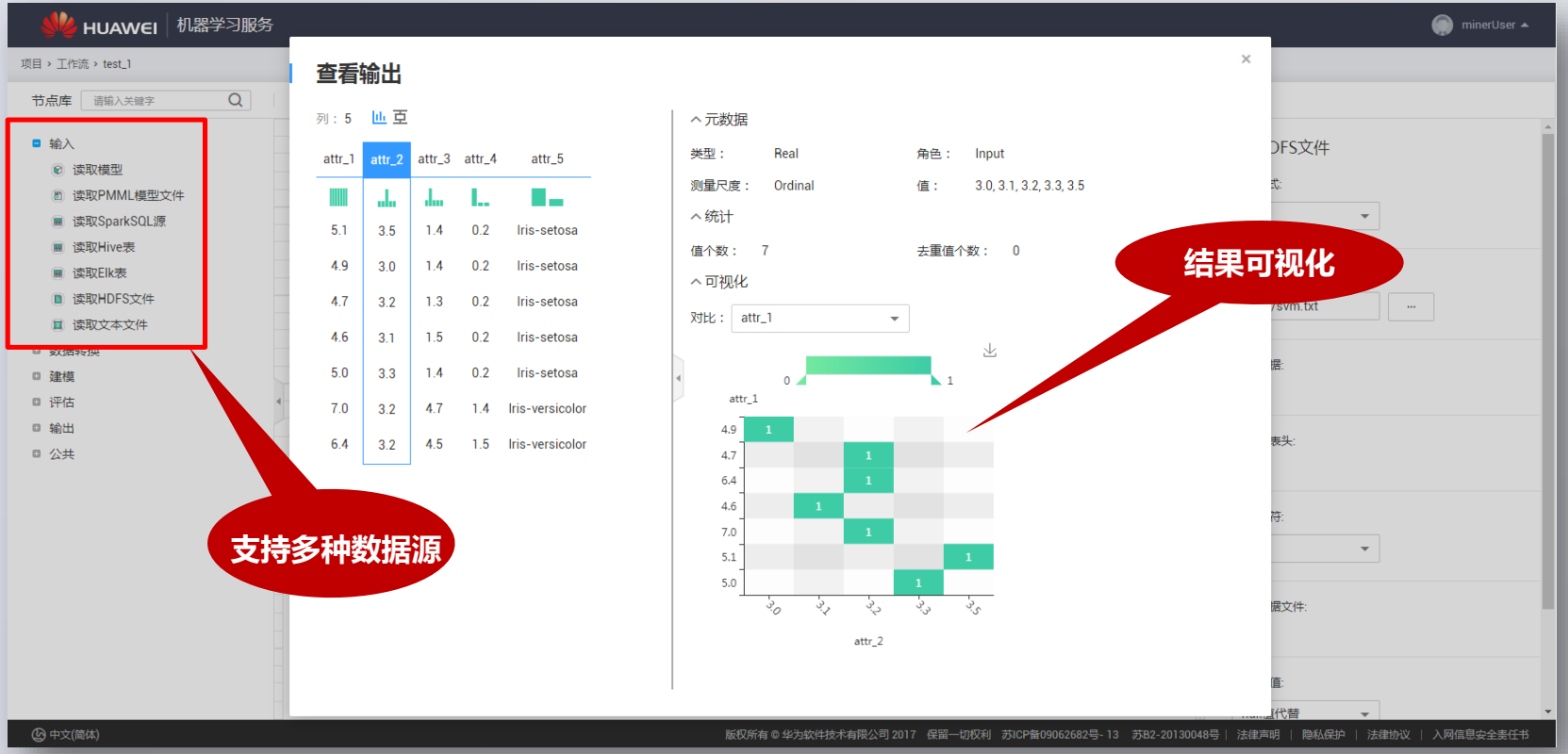
- 提供特征工程、算法、建模、预测、模型管理的一站式机器学习应用

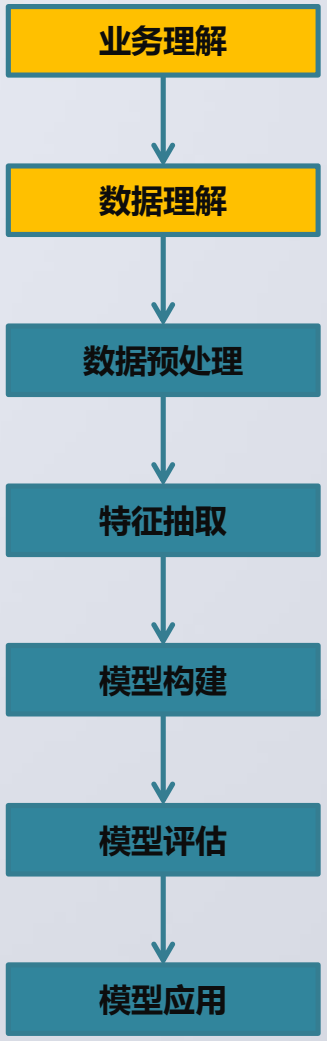
拖拽式工作流，实现数据建模、分析、可视化，更易用



对接多种数据源，提供可视化功能，让数据探索更直接

- 支持HDFS、Hive数据源，用户可动态添加华为MPPDB（Elk，LibrA）类型数据源；
- 提供Web化的自由探索界面，将探索结果通过可视化的图形展示出来，大大提高数据分析师的探索效率。

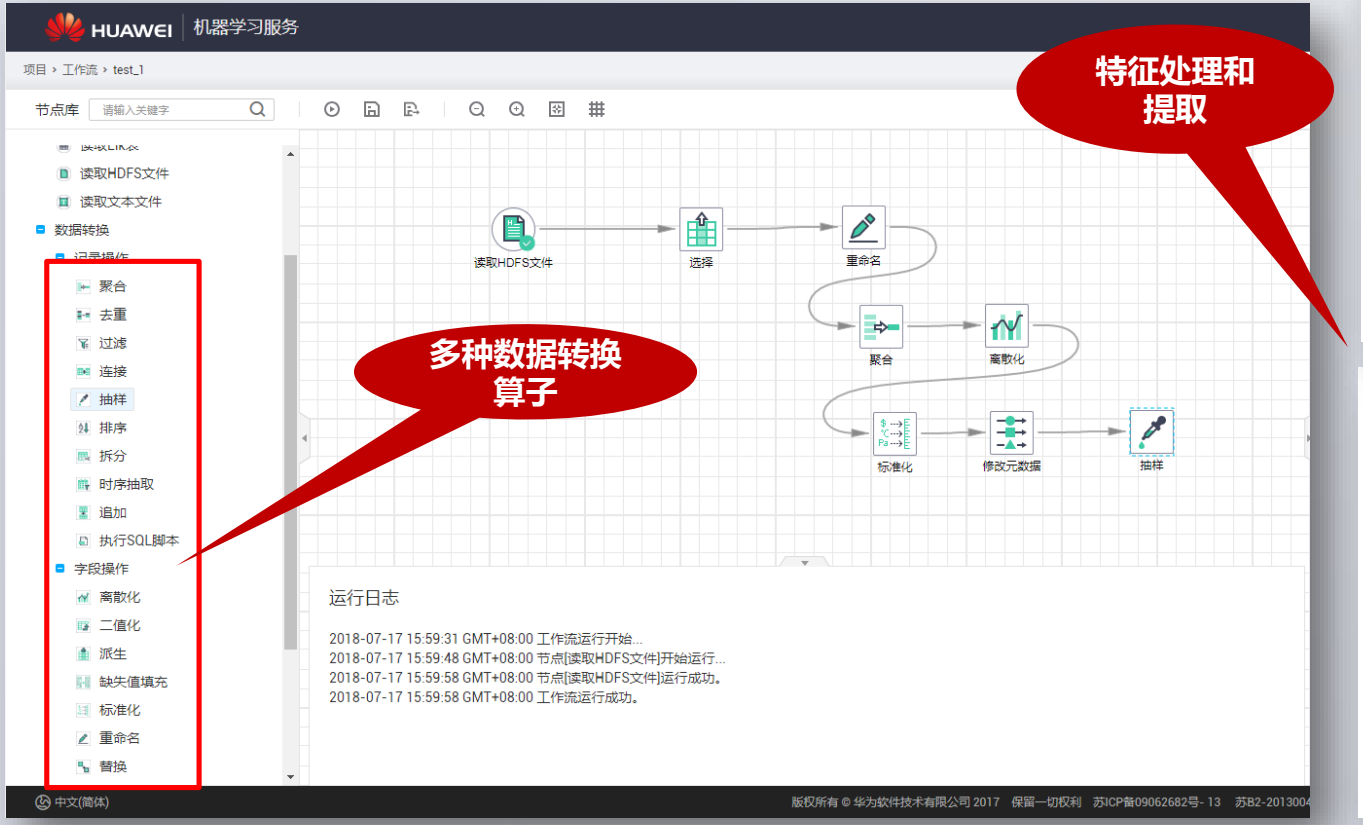


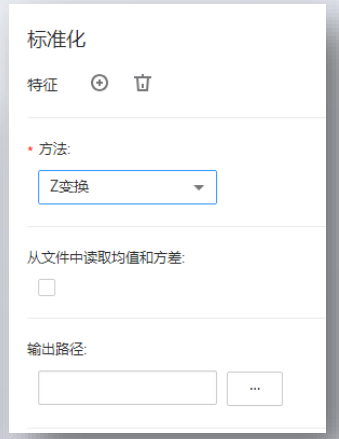


HUAWEI

多种算子交互使用，让构建特征工程变的得心应手

- 提供了大量提取特征的算子和功能，满足绝大部分特征工程的需要，尤其适合高维稀疏大数据的特征处理。

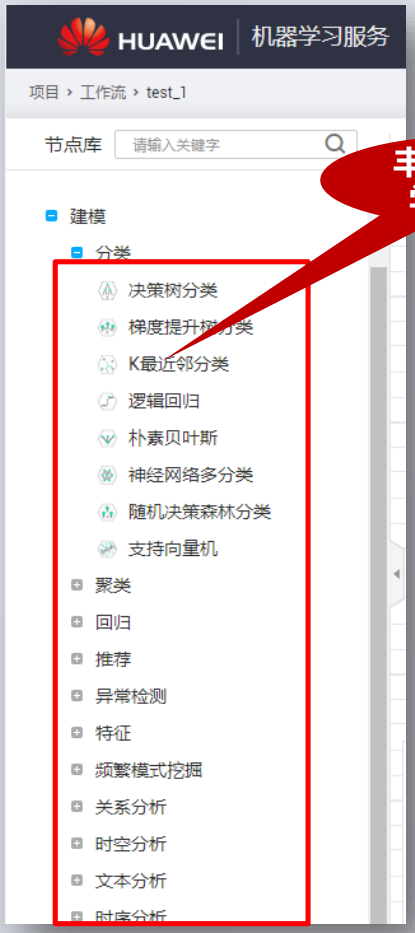




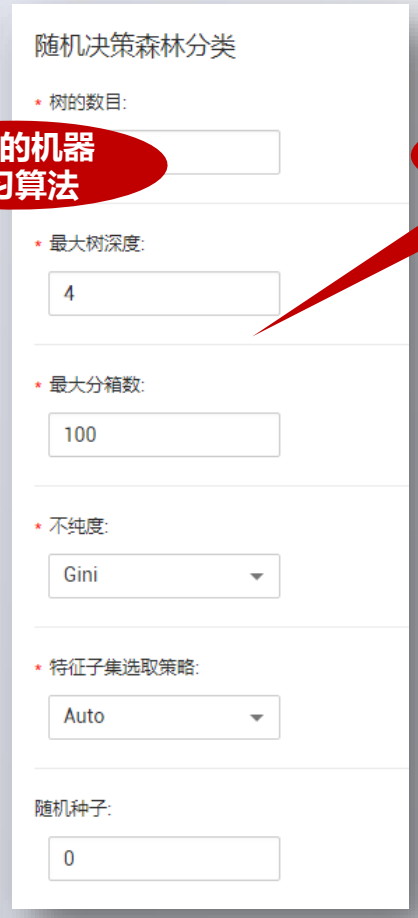
HUAWEI

覆盖各种场景的分布式机器学习算法库，让建模分析更轻松

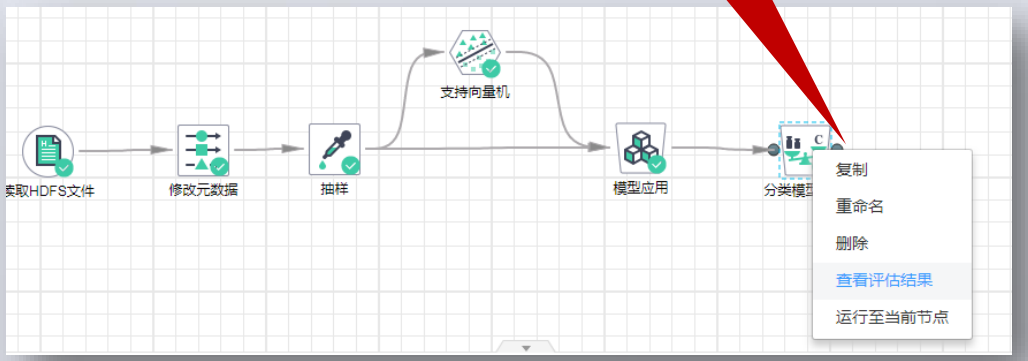
- 预置丰富的分布式机器学习算法库，满足各种建模场景需求；
- 一键查看模型评估，随时调参，让模型性能更加。



丰富的机器学习算法



多个参数可调节

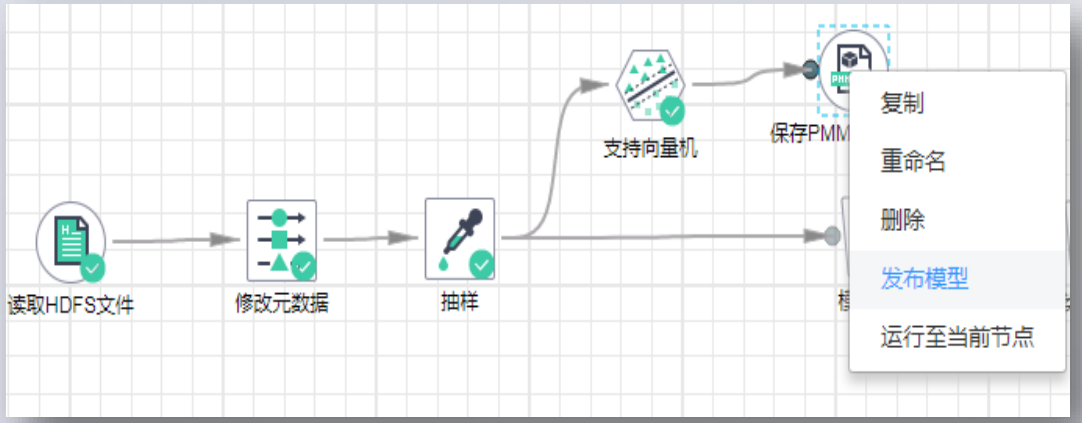


一键查看模型评估结果

HUAWEI

支持模型构建、发布及部署，让模型跨平台复用

- 支持模型构建及发布模型，让模型得到复用；
- 提供打包功能，让模型跨平台复用

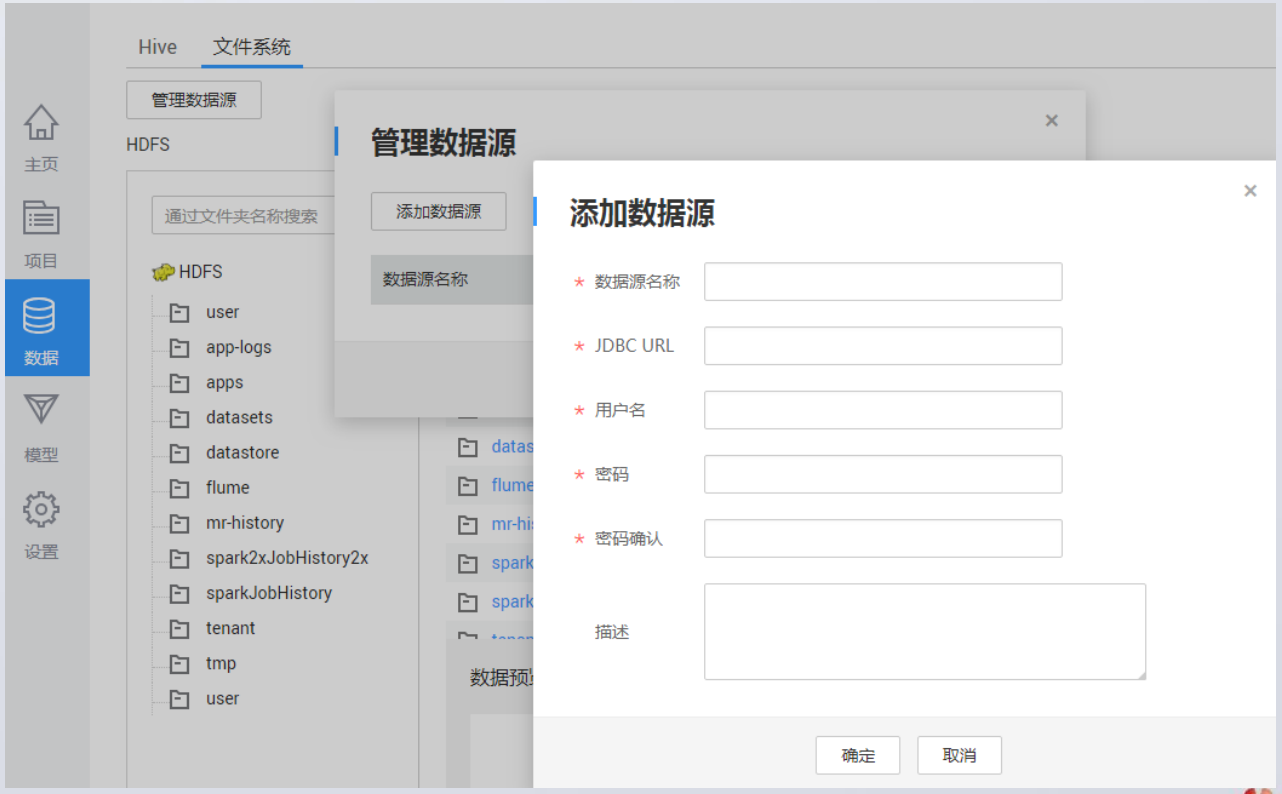


HUAWEI

支持多租户

1.按用户所属的用户组与角色控制数据访问权限

2.用户动态扩展的数据源专属于该用户



HUAWEI

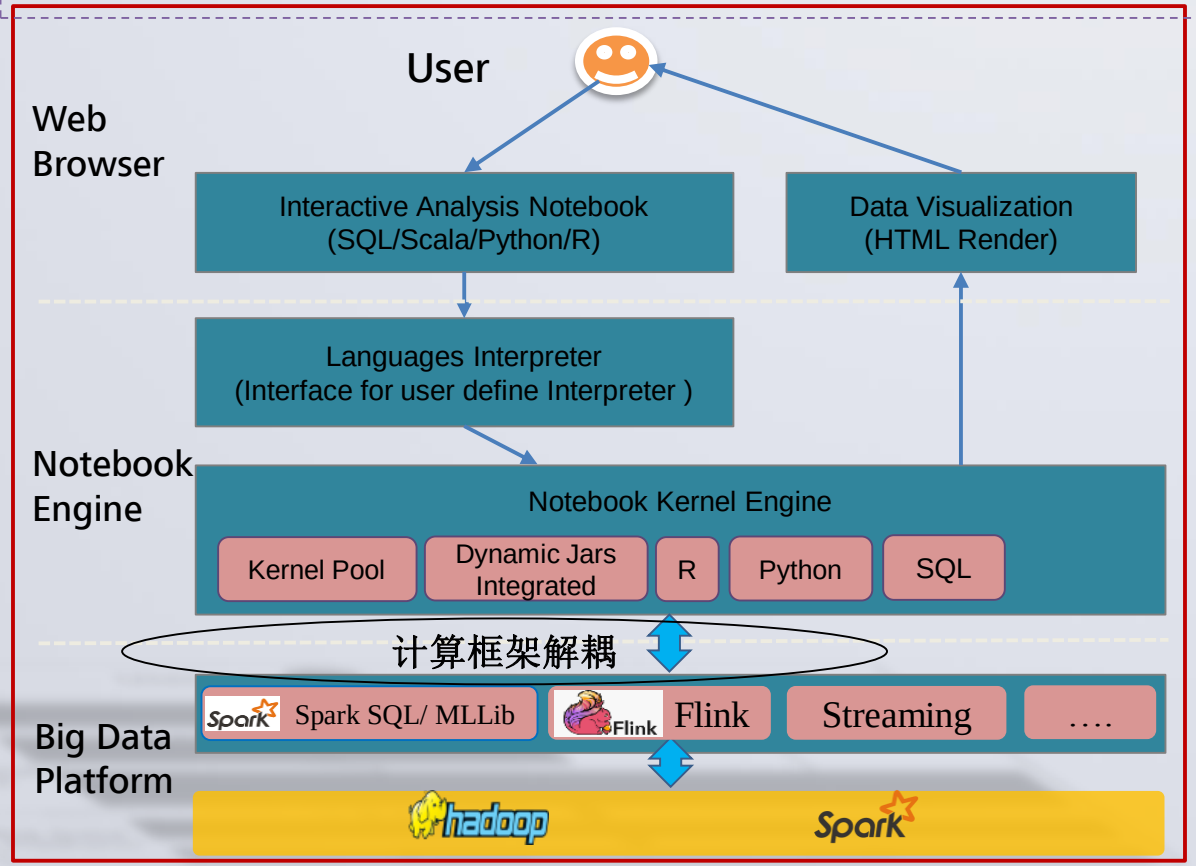
可操作算子达到100+，机器学习算法库达到70+

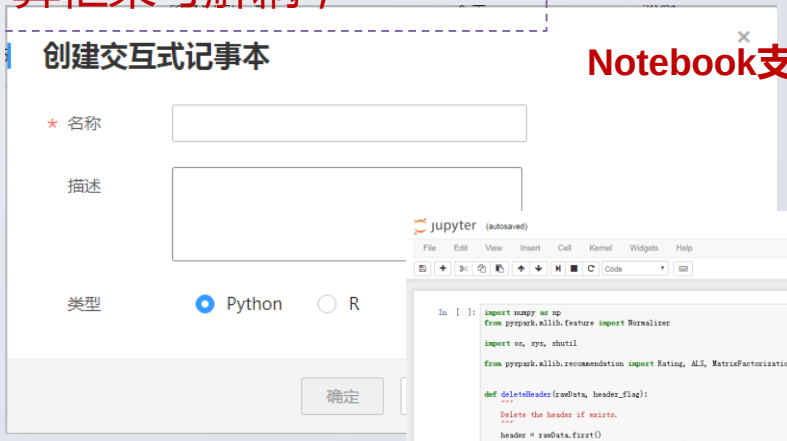
算子类别	算子名称（中英）	算子分类	合计
输入输出	读取模型、读取PMML模型文件、读取SparkSQL源、读取Hive表、读取Elk表、读取HDFS文件、读取文本文件； 保存HDFS文件、保存模型、保存PMML模型文件、保存SparkSQL数据、保存Hive表、保存Elk表	数据源	13
数据转换（字段、记录）	聚合、去重、过滤、连接、抽样、排序、拆分、时序抽取、时间窗抽取、追加、执行SQL脚本； 相关性选择、转换、离散化、二值化、派生、缺失值填充、标准化、重命名、替换、选择、设置元数据、重建、修改元数据	特征工程	24
特征选择	相关性分析、卡方检验、PCA、N元语法	特征工程	4
分类	Random Forest、Logistic Regression、KNN、SVM、Naive Bayes、神经网络、Desion Tree、GBDT	有监督学习	8
回归	Linear Regression、GBDT、KNN、Random Forest、Desion Tree	有监督学习	5
推荐	关联规则、域分解机、ALS	有监督学习、无监督学习	3
聚类	K-均值	无监督学习	1
关联规则（频繁项集挖掘）	FP-Growth、Apriori、PrefixSpan	无监督学习	3
时序分析	ARMA	时序分析	1
评估	模型应用、分类模型评估比较、分类模型评估、回归模型评估比较、回归模型评估	模型评估	5
异常检测	孤立森林、基于PCA的异常检测	异常检测	2
轨迹分析	轨迹点清洗、轨迹异常检测、轨迹切分、停留点聚合、停留点发现、相似轨迹分析	时空分析、轨迹分析	6
文本分析	去除停用词、分词、繁简转换、关键词提取、命名实体识别、Word2Vec、TF-IDF、文本推荐、自动摘要、HTML标签过滤、字符串相似度、词频统计、文本断句、LDA、短文本相似度、文本规范化	文本分析	16
关系分析	多源最短路径、PageRank、图直径与半径、三角形检测、Simrank相似度、共同邻居、度统计、社团趋势分析、社团发现、三角形统计、LabelPropagation、Simrank相似度模型、相似项分析、聚集系数、离心率、连通区域	图计算	16
公共	重分区		1

合计： 108

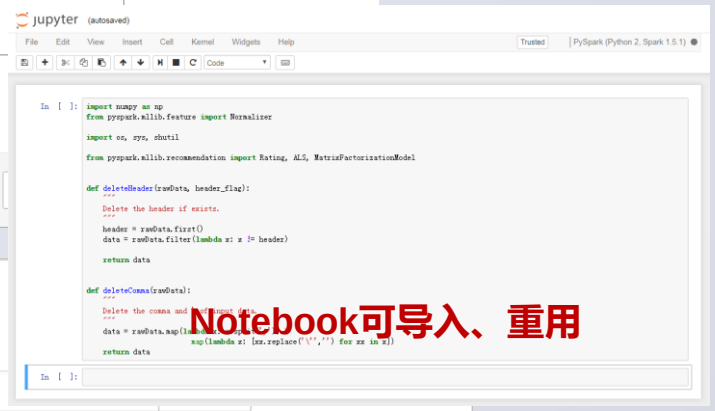
交互式Notebook，支持多种建模语言（Python、R），更开放

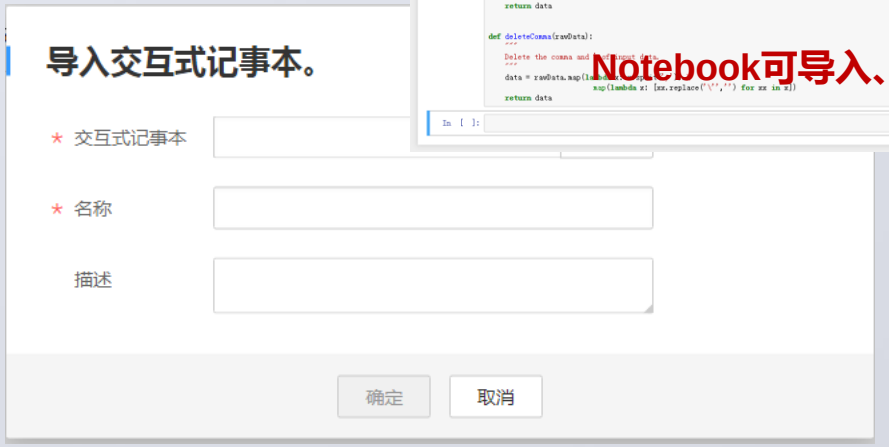
- 让科学家用最习惯的语言编程(Python 2.7/Python 3.6/R 3.4)，支持跨语言的使用。
- 数据分析 的“浏览器”，统一入口，与底层大数据平台计算框架可解耦；





Notebook支持多语言





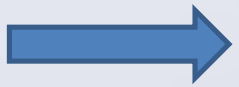
Notebook可导入、重用

HUAWEI

目录

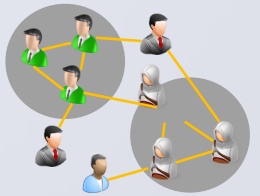
预置覆盖各种场景的分布式机器学习算法库，更丰富

算法分类	算法	场景
特征工程	相关性分析、卡方检验、PCA	特征选择
	N元语法	文本挖掘
有监督学习	Random Forest、Logistic Regression、KNN、SVM、Native Bayes、神经网络	分类
	Linear Regression、GBDT、KNN、Random Forest、Desion Tree	回归
	FTRL、域分解机、ALS	推荐
无监督学习	K-Means	聚类
	FP-Growth、Apriori、PrefixSpan	关联分析
时间序列分析	ARMA	时间序列分析
异常检测	孤立森林、基于PCA的异常检测	异常检测
时空分析	轨迹点清洗、轨迹异常检测、轨迹切分	轨迹分析
文本分析	词频、文本断句、去除停用词、隐含狄利克雷分布、分词、、繁简转换、关键词提取、命名实体识别、文本词向量、词频-逆文档频率、短文本相似的、文本规范化、文本推荐、自动摘要、HTML标签过滤、字符串相似度	文本挖掘
图计算	PageRank、LabelPropagation	关系分析



产品推荐

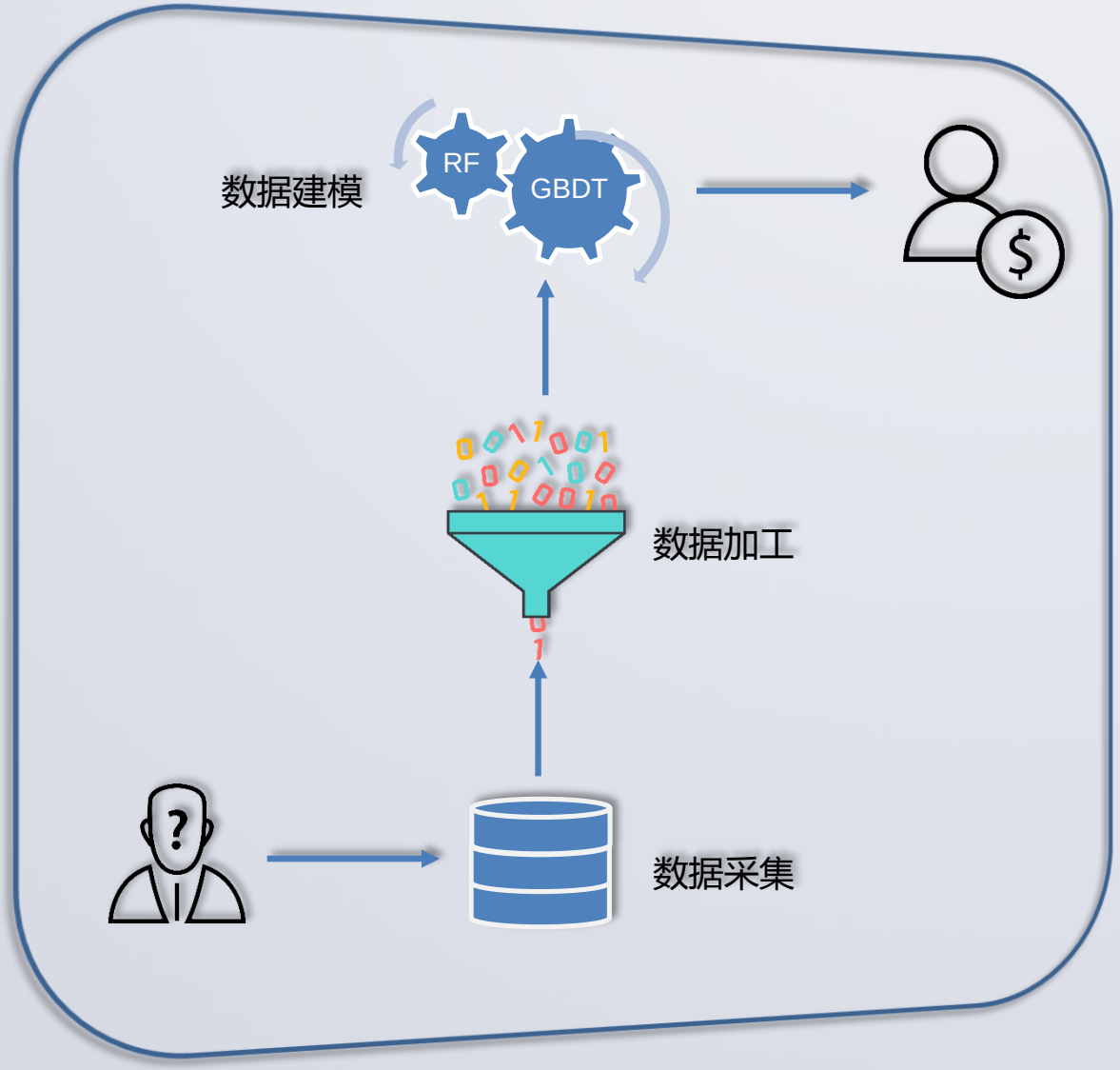
客户分群



异常检测

预测性维护

FusionInsight Miner典型应用场景——精准营销



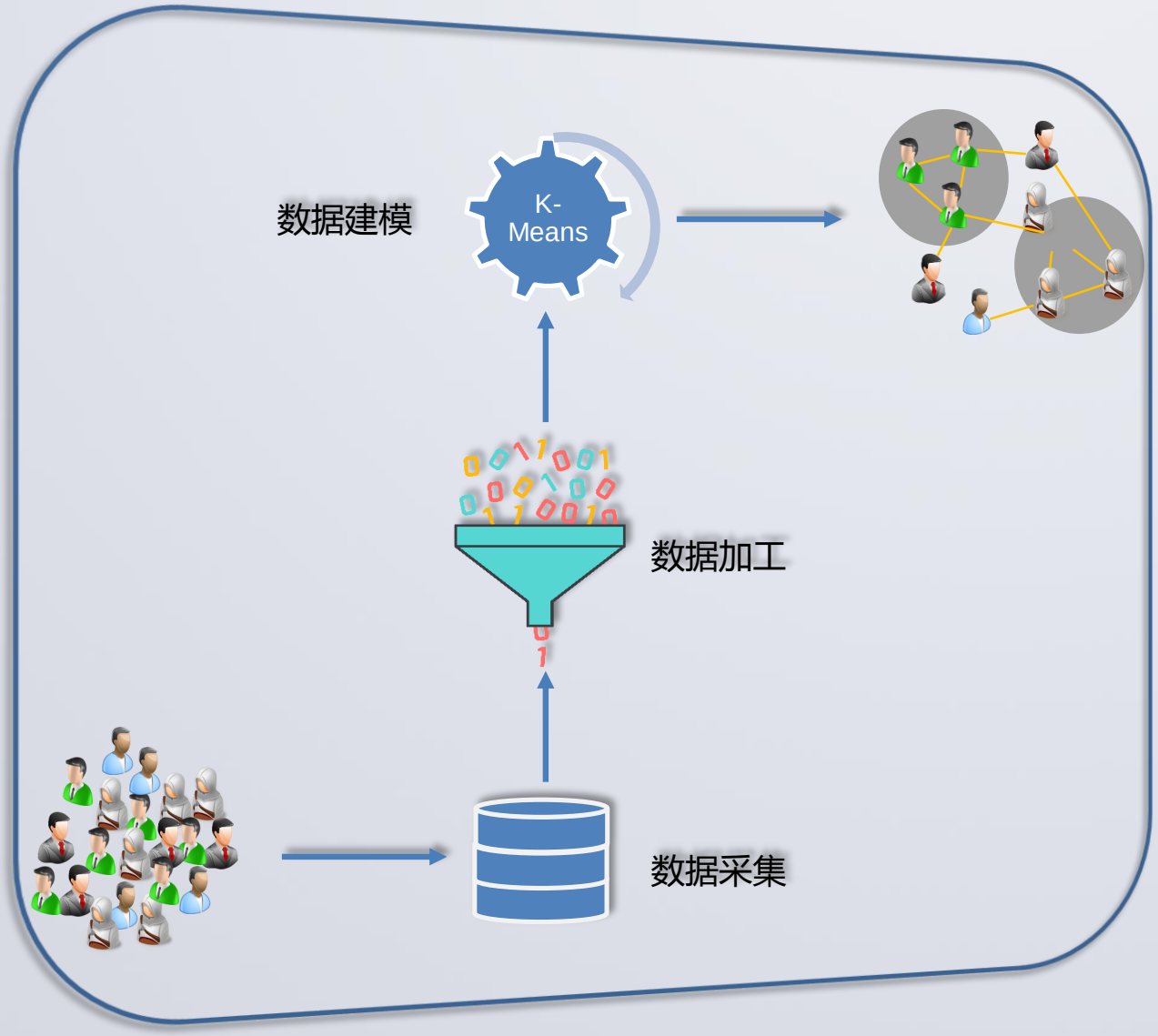
精准营销

- 根据客户本身属性和行为特征等（年龄、性别、工作类型、婚姻状况、文化程度、个人贷款、收入情况），预测客户是否愿意办理相关业务，为客户提供个性化的业务推荐，从而达到精准营销的目的。

相关内容

- 推荐算法：随机森林、梯度提升决策树
- 案例：银行理财推荐、车辆市场定价预测

FusionInsight Miner典型应用场景——客户分群



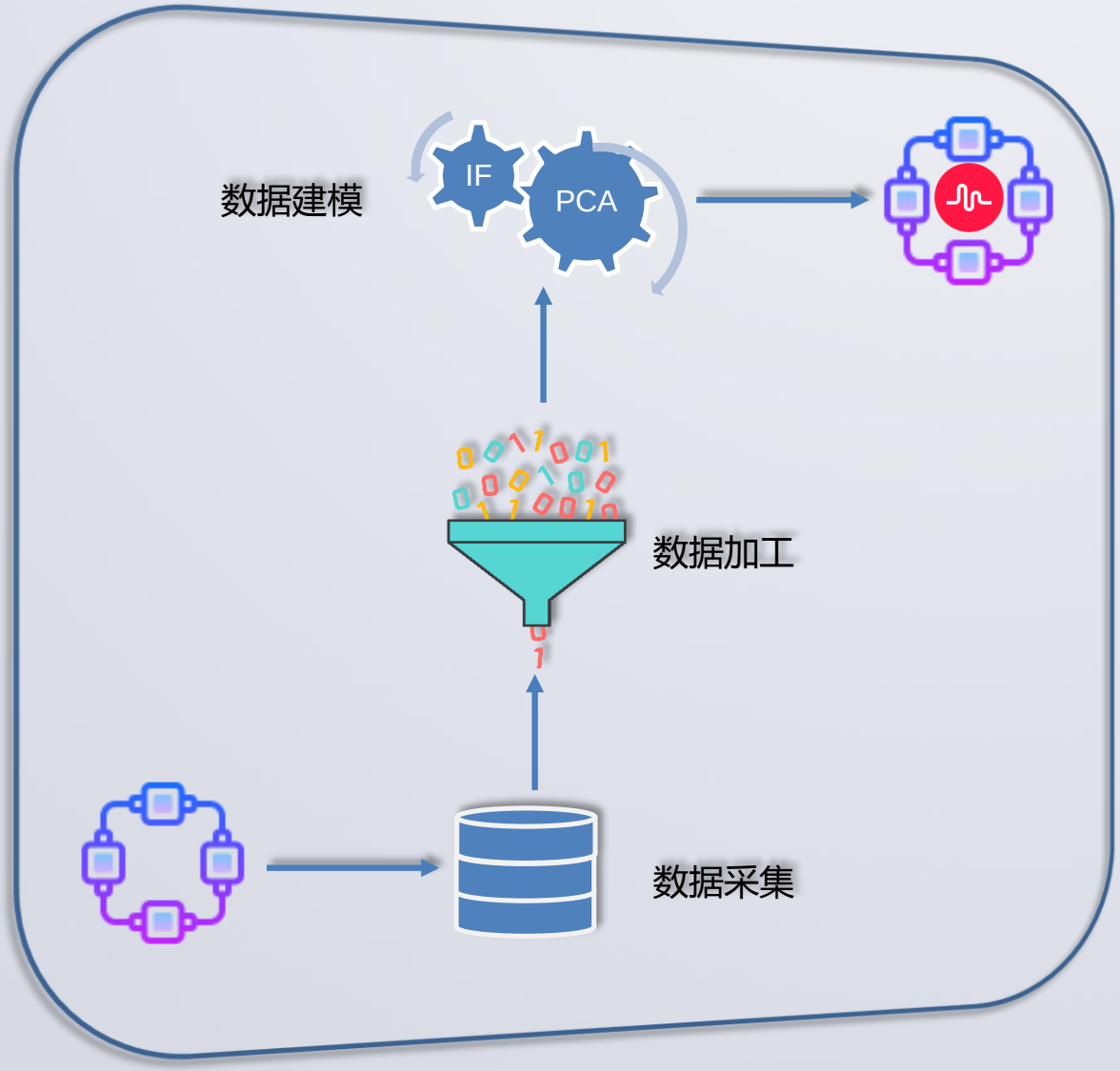
客户分群

- 通过数据挖掘来给客户做科学的分群，依据不同分群的特点制定相应的策略，从而为客户提供适配的产品、制定针对性的营销活动和管理用户，最终提升产品的客户满意度，实现商业价值。

相关内容

- 推荐算法：K-均值
- 案例：零售商客户分群

FusionInsight Miner典型应用场景——异常检测



异常检测

- 在网络设备运行中，用自动化的网络检测系统，根据流量情况实时分析，预测可疑流量或可能发生故障的设备

相关内容

- 推荐算法：基于PCA的异常检测、孤立森林
- 案例：网络入侵检测

FusionInsight Miner 思考题

问题1：客户分群可以用下列哪个算法？

A.线性回过 B.逻辑回归 C. K-mean D.SVM

问题2：下列哪几项是机器学习模型训练步骤？

A.数据预处理 B.特征抽取 C.模型构建 D.模型评估

问题3：线性回归属于分类算法？

A.对 B.不对

Thank You

HUAWEI