**机器学习平台简要说明文档**

# 产品背景

随着大数据的发展,在公安和航空等领域累积越来越多的数据，如何应用机器学习技术快速的挖掘数据背后的规律成为最紧迫的需求；特别是天津模型工厂的开发过程中，我们对公安的业务和数据有了更多的了解，同时客户可希望应用机器学习和数据挖掘的技术来分析他们的数据，以及可配置的分析工具提供给他们使用。因此需要开发机器学习平台来应对快速变化的分析业务，主要是为了解决以下问题：

1. 提高分析型业务的代码复用和业务；
2. 提供机器学习和数据挖掘的基础框架；
3. 降低机器学习和数据挖掘的门槛；
4. 为能力赋能平台提供基础平台支撑；
5. 提高项目的开发速度，降低项目开发成本；

# 产品定位

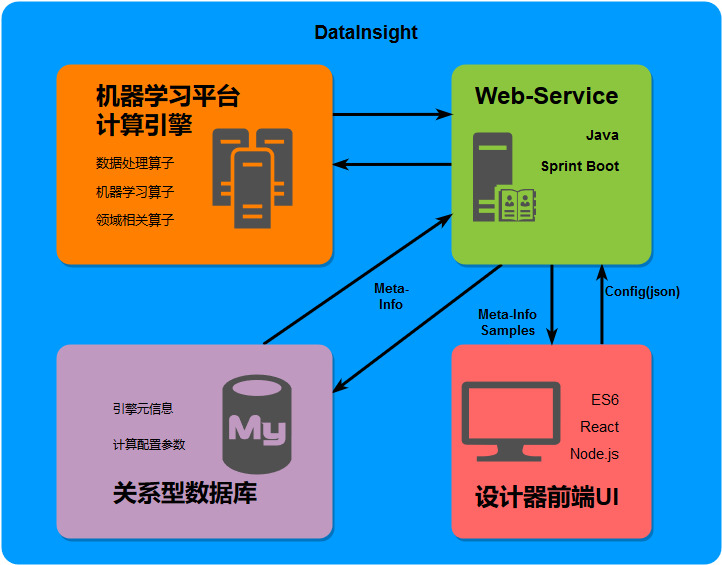
机器学习平台（DataInsight）是一款一站式的机器学习平台，包含数据处理、特征工程、常规机器学习算法、模型的评估以及预测这一整套机器学习相关服务。并且，DataInsight提供本地单机多线程、深度学习、分布式计算三大类框架，以应对不同的业务场景。另外，DataInsight将算法组件进行封装，并增添了丰富的可视化工具，让用户可以低门槛上手，通过简单的拖拽就可以实现整套机器学习工作流程。

# 产品架构

机器学习平台主要由3个子系统构成：

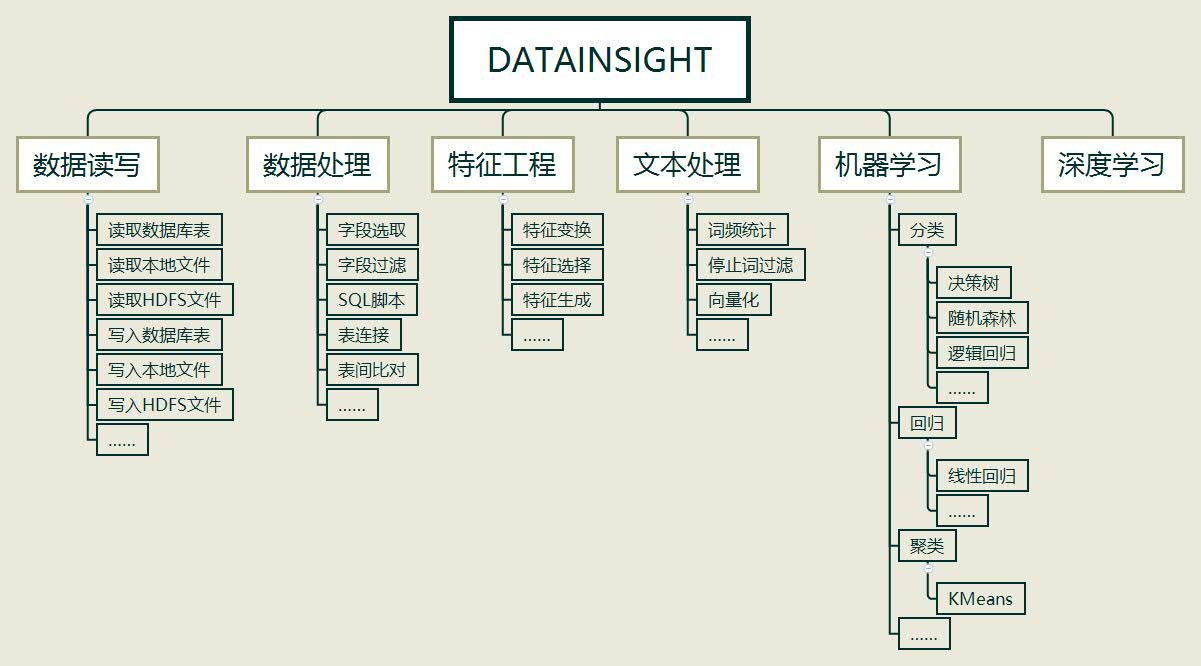
1. 计算引擎(Engine)：封装Operator来做数据处理、挖掘和分析，以此来提供强大、灵活的数据处理和分析的能力；
2. Web前端配置系统(Web-FrontEnd)：前端配置界面从后端服务上取得各种元信息(比如算子、参数和数据源配置信息等)，以及提交算子和应用的配置参数并保存到数据库中；
3. Web服务系统 (Web-Service)：Web服务有两个主要功能：
   1. 响应前端配置界面的请求，从数据库中获取元信息并发送给前端，将前端提交的配置信息保存到数据库中；
   2. 启动计算引擎中的计算任务，提交到计算集群上进行运算，保存日志和错误信息等，以及计算任务的分配和调度等功能；

机器学习平台各子系统之间的交互关系：

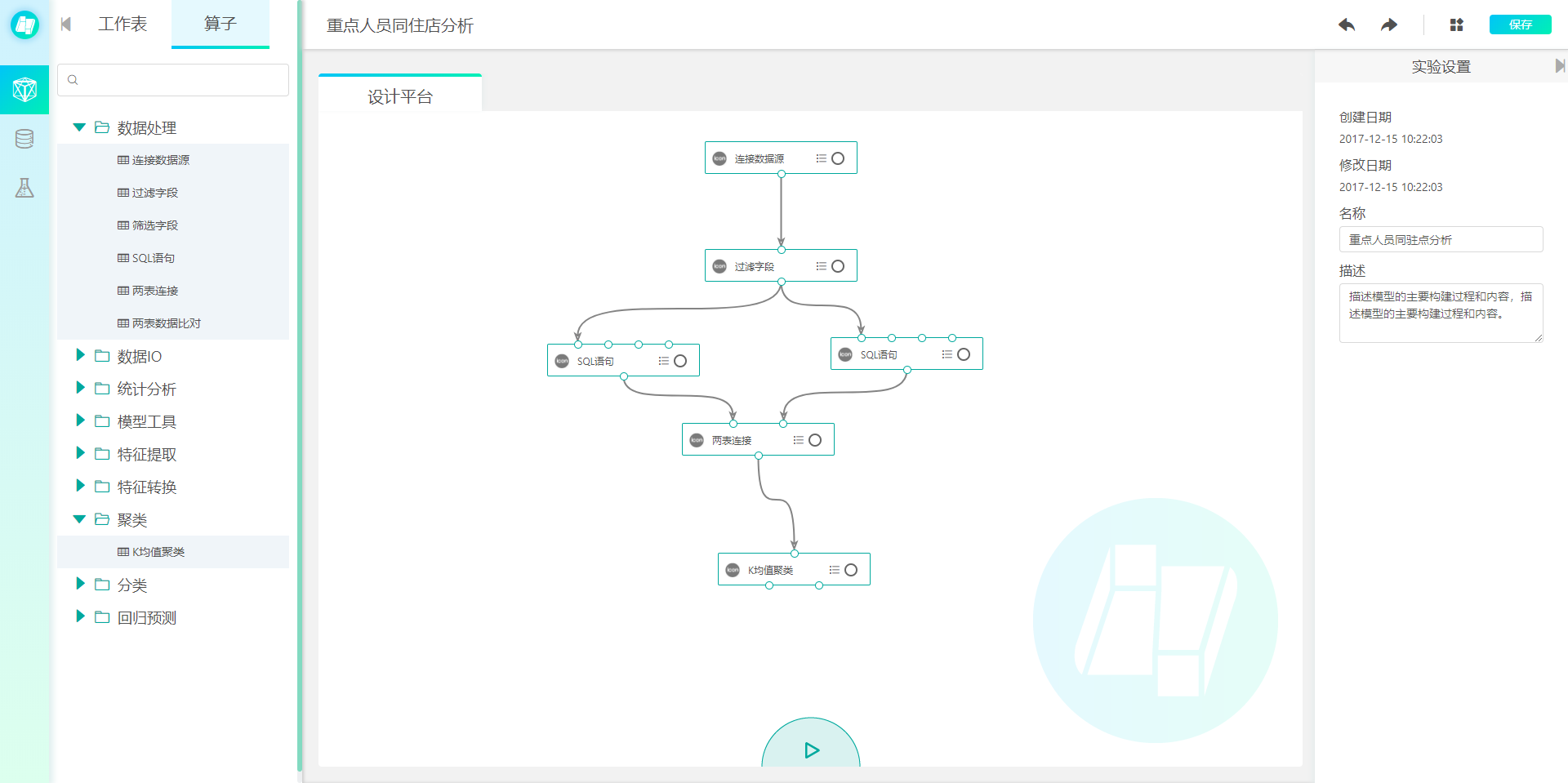


# 产品特性

1. 支持多种主流数据源接入
   1. 分布式数据仓库：Hive、Impala
   2. 关系型数据库：Mysql、Oracle、Postgresql
2. 支持三类计算框架
   1. 单机多线程：适用于小规模数据场景，无需接入分布式集群，帮助用户快速处理数据
   2. 分布式计算框架：适用于大规模数据场景，得益于Spark分布式快速计算引擎，可为用户提供PB级别数据的高效计算保证
   3. 深度学习：适用于需要大规模数据量迭代计算的场景，充分运用Deeplearning4J深度学习库，为用户提供高效的计算能力
3. 具有丰富的算法组件
   1. 数据预处理组件
   2. 数据读写组件
   3. 特征工程组件
   4. 机器学习算法组件
   5. 模型应用组件
   6. 模型评估与预测组件



1. 简洁的可视化操作环境



1. 一站式机器学习体验

提供模型训练、模型评估以及模型持久化功能，模型可以重复使用、重新编辑、多次迭代，模型既可以用于在线预测，也可以离线调度，让机器学习训练结果和业务无缝衔接