机器学习平台-DataInsght

1. 平台背景

随着大数据以及人工智能的发展，众多领域都积累越来越多的数据，而市场都强烈要求企业能够运营客户数据，结合实际需求快速迭代分析产品以及分析方案，因此机器学习技术如何快速落地成为紧迫的需求。

为满足客户日益旺盛的分析需求，需要开发更加简易以及复用的分析算法模块，同时降低使用门槛，让更多普通开发人员或者分析人员能够在AI时代更好的发挥价值，为企业创造更多的利润。

机器学习平台(DataInsight)主要是为解决上述提出的问题，主要表现在以下方面：

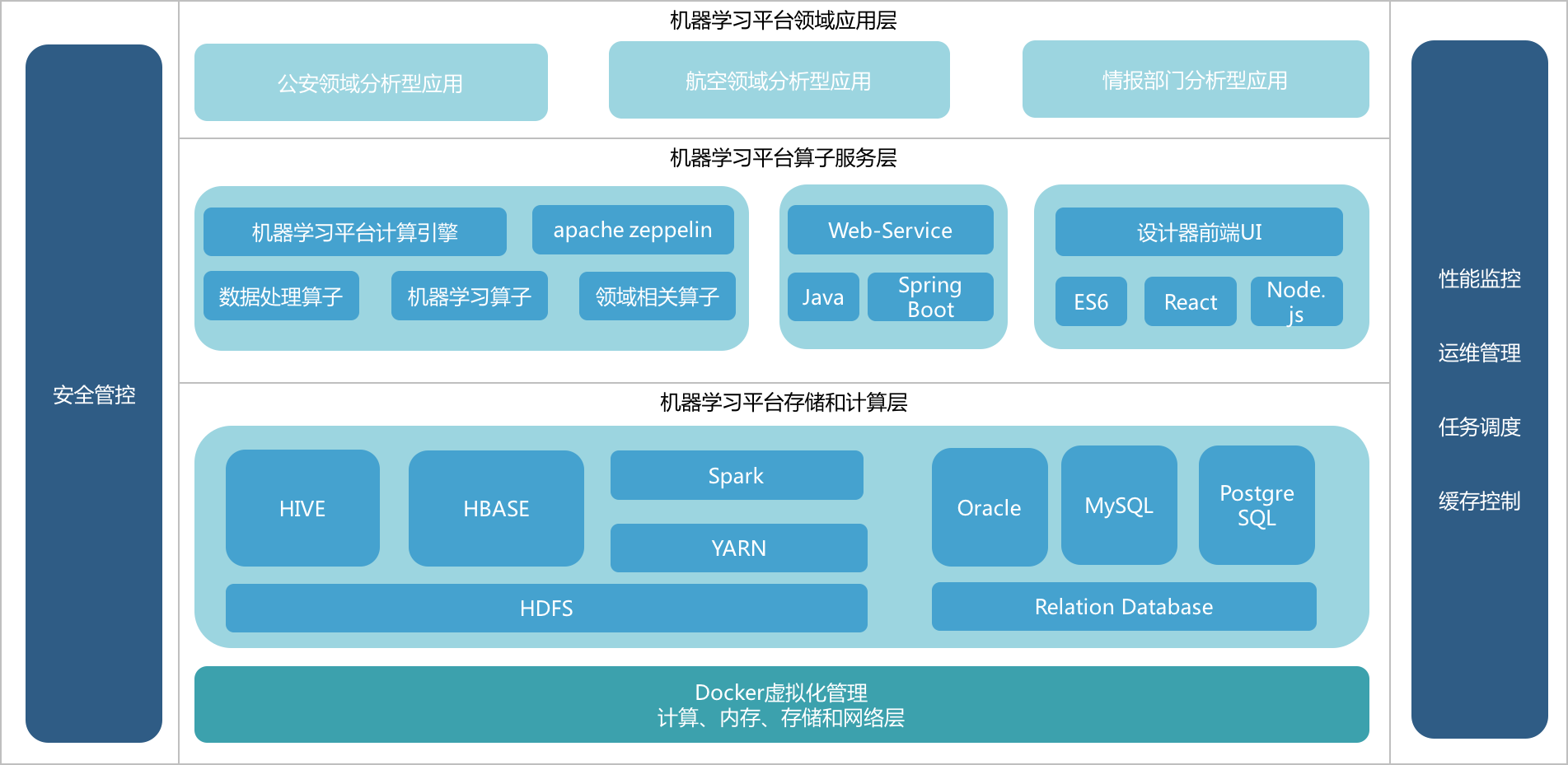
1. 提高分析性业务的代码复用以及领域知识的留存；
2. 提供机器学习和数据挖掘的基础框架；
3. 降低机器学习和数据挖掘的门槛；
4. 为算法分析与模型构建提供底层的支撑；
5. 提高模型构建速度，降低算法的开发成本；
6. 平台定位

DataInsight是算法分析以及模型构建的平台，包含数处理、特征工程以及机器学习算法的算法模块，并且能够以本地单机多线程环境、集群分布式计算集群以及GPU环境下运行，为客户和开发人员提供多种选择；同时集成可视化工具以及交互式数据探索环境，便于分析人员快速构建分析模型；然后结合图易视化平台，将分析结果以大屏或者应用页面输出，为客户提供智能辅助决策。

由于分析领域的技术特殊性，需要使用的人员同时具备计算机算法知识以及领域模型，从而根据现有的数据以及实际的需求，构架出有效的分析模型。因此平台的使用主要还是数据挖掘领域的分析建模人员。

1. 架构设计

DataInsight是基于大数据存储以及分布式计算技术为基础开发的，支持大数据量的分析以及计算。该平台主的架构设计图如下：



机器学习平台算子服务层主要由3个子系统以及交互式数据分析环境构成：

1. 计算引擎(Engine)：封装算子(Operator)来做数据处理、挖掘以及算法分析，通过配置参数代替编码的方式来构建整个分析应用，而算子通过有向无环图的方式组合起来形成最终的模型。
2. 前端设计器(Designer)：前端模型设计界面从后端服务商获取各种元信息(比如算子、参数以及数据源配置信息等)，提供拖拽式的可视化交互，配置计算引擎所需的json配置文件，并将这些模型配置信息保存到数据库中。除此之外，前端也提供数据源编辑以及实验模型管理以及运行的界面。
3. Java微服务(Micro-Service)：其主要由两个功能
   1. 为前端设计器系统提供restful接口，从数据库中获取元信息并发送给前端，同时将前端构建模型配置参数保存到数据库中，提供分析工具到图易的关联；
   2. 根据配置文件启动计算引擎，获取相应的计算硬件资源，提交并运行，保存模型的日志以及错误信息。
4. 交互式建模环境：在构建分析模型的过程中，我们经常需要实时与数据进行交互，以及快速迭代分析模型，因此交互式数据分析工具为这个需求提供极大的便利，能够让分析人员在该平台下完成整个模型原型设计到模型部署。

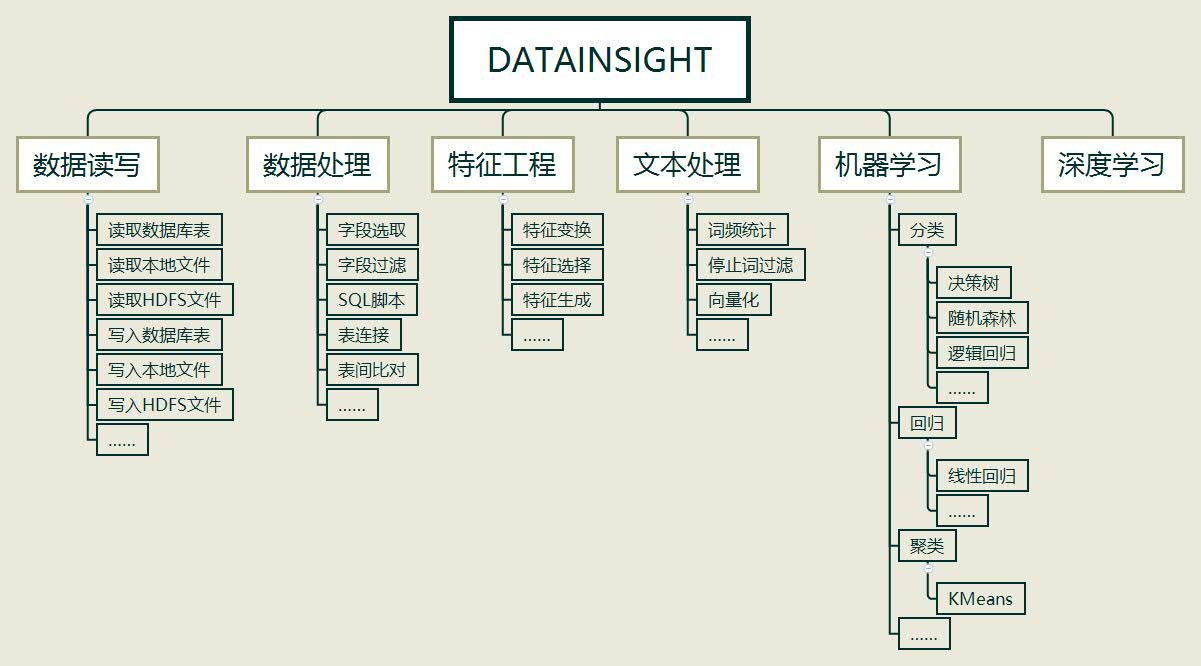
提供Scala、Python、R、SQL以及Markdown等多种语言的支持，同时也集成了Spark，Hive以及其他众多数据，除此之外也能够快速的实现统计可视化等。以下是支持的语言后端的完整列表：



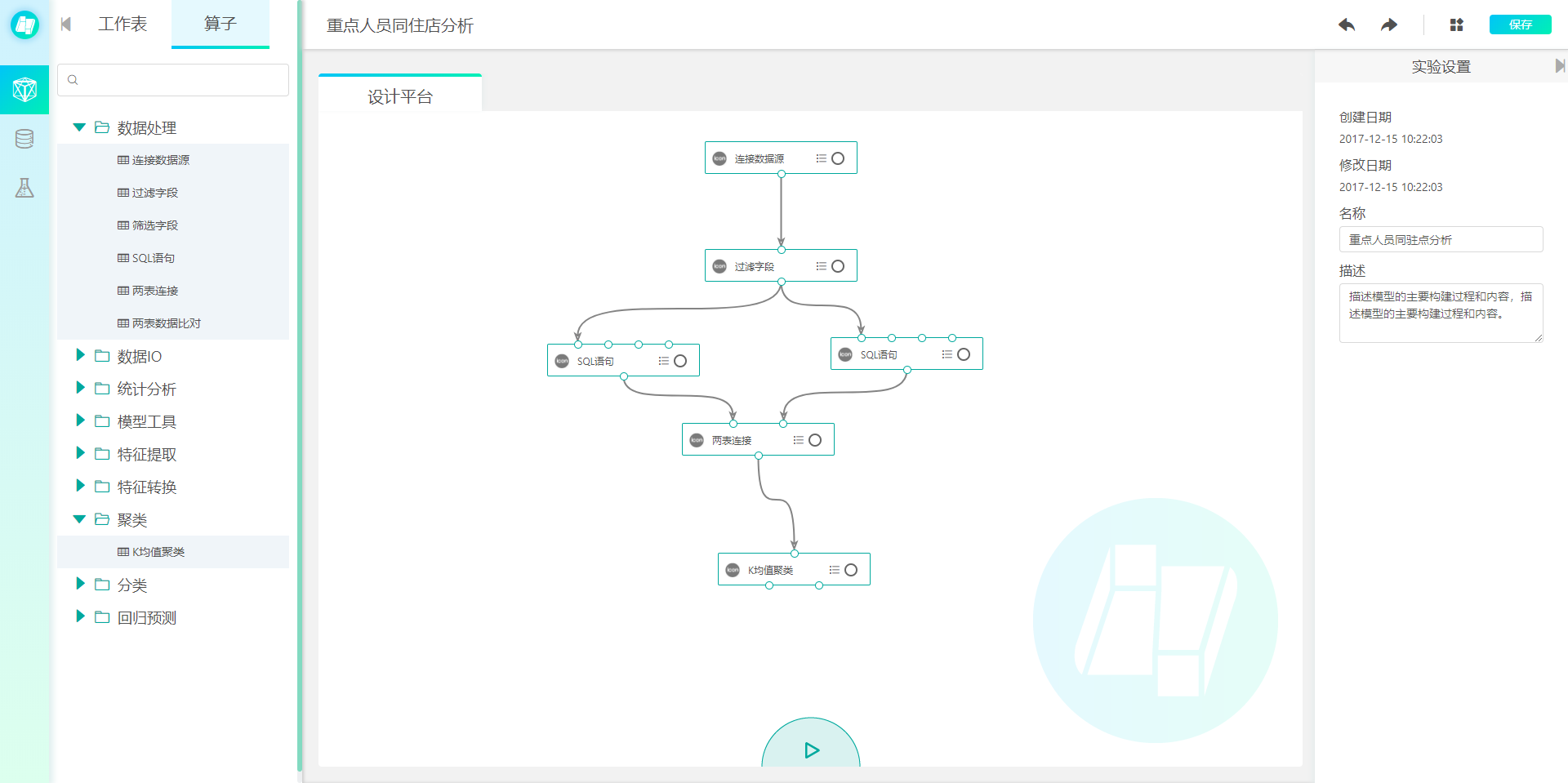
1. 平台特性
2. 支持多种主流数据源
   1. 分布式数据仓库： Hive，Impala
   2. 关系型数据库：MySQL，Orcale，PostgreSQL
3. 支持多种计算环境
   1. 单机多线程环境：适用于小规模数据场景，无需分布式集群环境，帮助用户快速处理小批量的数据；
   2. 分布式集群计算环境：使用大数据场景，得益于hadoop以及spark等开源大数据存储以及计算生态，可为客户体统PB级数据的高效计算；
   3. GPU环境：深度学习是当下最热门的人工智能技术，在图像以及自然语言领域取得突出的战绩，我们的分析平台也融合深度学习的技术，为客户提供深度学习的研发以及部署环境。
4. 丰富的算法组件

提供算法组件共50多个，涵盖建模以及机器学习的方方面面，为客户提供众多的模块选择，以便提供强大的分析和建模能力，主要包括以下几大类：

* 1. 数据预处理
  2. 数据源读写
  3. 特征抽取与特征转换
  4. 监督学习与非监督学习
  5. 模型应用
  6. 模型评估与预测
  7. 图分析与计算



1. 交互式界面
   1. 前端模型设计器：



* 1. 交互式建模环境

