

机器学习平台

产品白皮书

海云数据

产品技术中心-焦华琛

# 产品概述

## 产品背景

随着网络和信息技术的不断普及，人类产生的数据量正在呈指数级增长。现如今，政府和企业越来越关注数据资产的价值，大量新数据源的出现则导致了非结构化、半结构化数据爆发式的增长。但这些由我们创造的信息背后产生的这些数据早已经远远超越了目前人力所能处理的范畴，政府和企业通常缺乏高效的处理手段，难以挖掘大数据真实的价值。

“大数据”需要新的处理模式才能具有更强的决策力、洞察发现力和流程优化能力，然而，现如今传统的方法都是通过编码的方式来实现大数据分析，通常工期长不具备通用型，每次都是针对具体的任务定制化的开发，传统IT从业者和企业业务人员因大数据分析技术门槛高，无法掌握大数据分析的能力，数据分析人才紧缺，导致政府和企业无法通过大数据分析的方式挖掘数据价值。

## 产品定位

机器学习平台是一个企业级大数据可视化数据分析建模平台，提供数据分析建模的全流程可视化设计管理能力，包括交互式模型配置、数据分析、机器学习、业务算法、可视化组件等多种类型的算子组件和模型工具，将数据分析算法步骤封装成算子，通过前端交互式模型配置页面来操作大数据分析流程，降低大数据分析的上手门槛，为用户提供一站式的业务数据分析和数据建模快速开发服务，让数据分析模型从传统的依赖于专家建模的方式，向在线、自助式、可视化的方式转变。

**产品主要服务于以下场景：**

* + 预测类场景：隐性吸毒人员预测、医学病人患病预测
  + 分类：垃圾邮件检测、有害信息识别营销类场景：商品推荐、用户群体画像、广告精准投放
  + 金融类场景：贷款发放预测、金融风险控制、股票走势预测、黄金价格预测

## 产品优势

### 可视化数据建模

对底层的数据分析算法封装，用户无需具备专业数据分析技能，通过可视化建模工具即可完成业务建模和数据分析，降低技术门槛，无需编写代码，大大缩短开发周期。

### 知识分享

在平台中建立的模型可以选择上传至模型超市共享给其他用户，通过模型超市业务知识将不断积累并分享给企业中的其他用户，将模型生成的结果通过api服务的方式灵活接入其他系统中。

### 打通数据孤岛

统一数据管理平台，支持用户可视化配置异源异构数据，包括传统关系型数据mysql、oracle等，以及非关系型数据库hbase、hive等。对接入的数据源进行统一管理，实现数据表跨库关联。

### 灵活可扩展

* + - * 算子组件化，支持灵活扩展算子
      * 支持用户自定义组合算子生成新算子
      * 支持用户灵活配置模型

## 客户价值

### 降低技术门槛

通过对底层的数据分析算法封装，用户无需具备专业变成技能，通过可视化建模工具即可完成业务建模和数据分析，摆脱传统的专家建模方式，降低技术门槛。

### 缩短项目开发周期

通过可视化建模工具完成模型构建和数据分析，无需编写代码，开发人员可以专注于优化业务逻辑，大大缩短开发周期。

### 提高工作效率

用户可以动态调整模型的参数，修改模型调度策略，提升工作效率。

### 降低开发成本

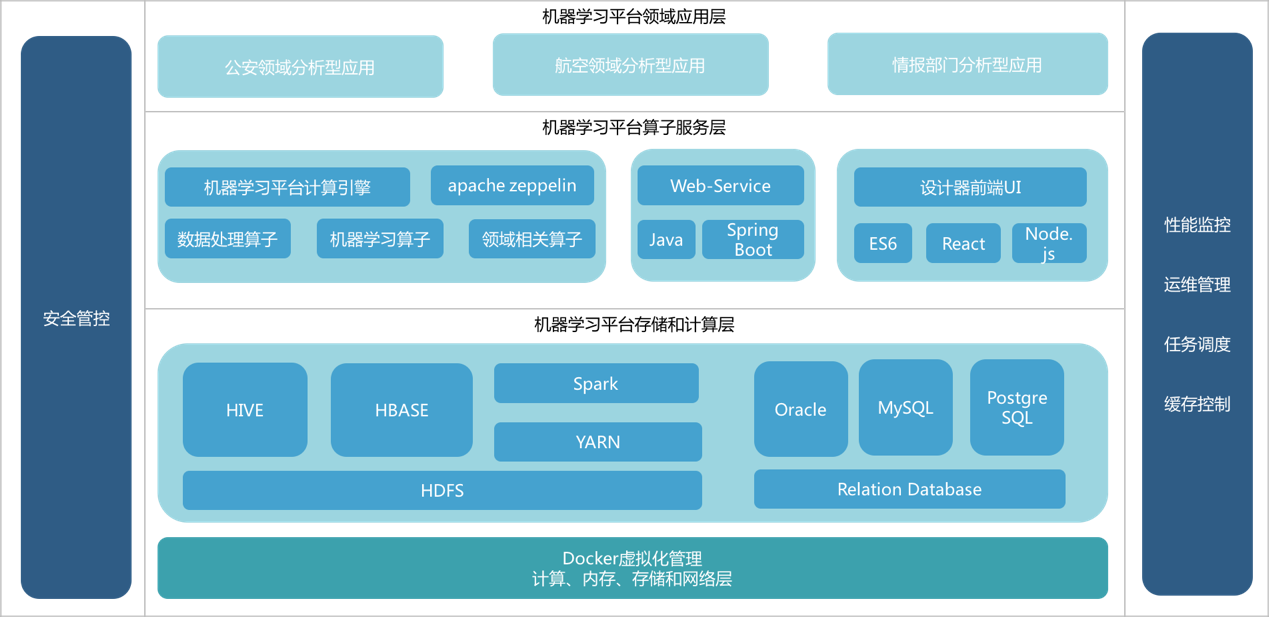
提供机器学习和数据挖掘的基础框架，和公安技战法业务算子，降低对第三方技术开发公司的依赖，无需重复建设底层框架，减少项目开发成本。

### 业务知识积累、分享

提供模型超市，将用户制作的模型保存并发布到模型超市，发布到模型超市的模型可供其他用户调用、下载，实现业务知识的积累和分享。

### 挖掘数据价值

# 产品架构



《机器学习平台》主张利用数据分析技术，将企业系统中海量的异构数据进行整合，根据行业业务需求，形成通用性的业务算子，并通过对底层的分布式数据分析算法进行封装，提供交互式数据分析建模工具，降低数据分析建模的上手门槛、提供工作效率。并提供模型调度功能，使模型训练的结果可以和业务实现无缝对接，用户还可以结合可视化功能，一站式解决从整合数据、分析数据、建立模型、成果可视化的问题。帮助1人员有效认知、了解、提取有价值、高质量的数据，探索数据价值，实现大数据落地业务场景，实现数据价值。

# 产品功能

## 高性能分布式数据资源管理

基于Hadoop和Spark生态来组建整个框架，因此天然支持大数据存储和和分布式计算的能力。Hadoop作为分布式数据存储的基础平台组建，能够实现大数据量的分布式存储；Spark作为分布式计算框架，在某种程度承担着连接不同数据源和计算过程，能够实现快速的数据处理和数据分析操作能力。整个平台支持的数据源类型如下：

1. 支撑结构化、半结构化、非结构化数据源的连接
2. 支持关联数据库Oracle，SQLServer，MySQL等主流数据源
3. 支持Excel、CSV等文本类型文件直接作为数据源导入；
4. 支持Hive和HBase的连接；
5. 支持接入kafaka和stream等流数据源。

## 机器学习引擎

机器学习是数据分析的核心部分，主要是从数据中挖掘规律与特征，学习相应的模型，将分析的结果应用于决策等过程中。但由于机器学习的门槛相对较高，普通的开发者难以开发和应用恰当的模型，因此模型工程也是将机器模型的不同功能块划分为算子，然后通过组合算子的方式来形成最终的模型。从而加快模型开发的速度，降低开发成本，同时也让更多普通开发人员也能够快速的实现特定的模型。

spark生态系统提供了分布式计算的能力，因为机器学习平台是基于hadoop和spark生态开发，形成独有的技术解决方案。同时spark具有很强的粘合能力，能够将多种数据源快速的接入整个系统中，并且在此基础上执行分布式计算等任务；

## 交互式模型设计

提供交互式模型设计器，通过对底层的分布式算法封装，将复杂的数据分析模型建立过程变成数据和算子的拖拉组合，让数据挖掘的创建过程像搭积木一样简单，降低用户的上手难度，加快模型开发的速度，降低开发成本。

### 丰富的算法

提供丰富的算法，支持传统的机器学习算法，也支持强大的深度学习算法，能够实现包括特征工程、数据预处理、统计分析、机器学习、神经网络、预测与评估等算子组建等算子，共 100 余种，同时也能够通过前端的设计器界面快速实现模型构建和算法验证等。

### 可视化数据分析

强大的图表引擎提供了超过 30 种不同的可视化选项

* 气泡图和三维散点图
* 网络和树状图，以及更多

图表类型包括柱形图、条形图、折线图、饼图、面积图、组合图、仪表盘和矢量地图等40多个图表。

## 系统管理

### 权限管理

如今，为了保密和业务的需要，各种应用系统都有权限方面的控制。同样，为保证政府、企业数据的安全性，机器学习平台提供权限方面的控制

1. 配置各个用户相应的BI访问权限
2. 用户第一次访问受保护的资源时，会发出访问请求，服务器接收到请求后会验证用户权限，如果没有通过验证则返回登录页面
3. 如果通过验证用户获取一个唯一标识自身的令牉（ticket），并将携带这个令牌进行BI的访问操作

#### 权限管理的层次

**业务数据的权限：**支持不同用户对于数据业务的数据权限仅限于自己权限范围内；

**数据的明细权限**：支持不同的用户对于同一个模型，只能够访问权限范围内的数据结果；

**模型的使用权限**：支持不同的用户对于模型管理使用（查看、删除、修改）权限进行控制。

### 调度管理

调度管理作为系统的核心组件，主要是任务的调度和资源的调度，实现在有限的资源上，高效的利用存储和计算资源，且能够根据用户指定和修改任务的优先级，动态的调整任务执行的先后顺序。

### 缓存管理

模型在运行过程中，存在大量的中间结果需要缓存到分布式存储系统中，为了避免大量的重复计算，提高计算的效率。因此需要对缓存进行控制和管理，实现缓存的创建、删除以及更新等操作。

# 产品部署

暂无

# 结束语

现如今，越来越多的行业客户依赖于大数据可视分析来获得有价值的数据关系，重要的业务指标，帮助行业用户做出正确的预测和决策。图易使用户以全新的方式分析数据，数据可视化技术完美呈现工作成果。贴合业务的数据模型与简洁直观的操作方式，让行业业务全员都能迅速掌握，为您挖掘数据价值提供强劲的驱动程序。