# iCloudMiner

# 产品白皮书

成都数之联科技有限公司

**2018 年 11 月**

# 版权声明



# 免责声明



# 技术支持

感谢您选择数之联科技有限公司的商务智能产品，在系统的使用过程中如果遇到问题，请通过如下的方式与我们联系，我们将为您提供竭诚的服务。

主页：http://www.unionbigdata.com

电话：028-86661321

地址：成都市高新区吉泰五路 88 号香年广场 T3 栋 5 层

邮箱：shuzhilian@unionbigdata.com

# 目录

# 产品定位

iCloudMiner 是一个可视化云端数据挖掘平台。

iCloudMiner 分为企业版和个人版，其中企业版主要面向拥有海量数据沉淀能力，并期望分析，挖掘和利用数据的企事业单位、政府服务部门、科研院所；而个人版则主要面向高校师生、数据分析师、数据科学从业者，为其提供一个有限空间和计算能力的数据分析挖掘平台。iCloudMiner 解决方案覆盖了各个领域，包括制造、通信、金融、以及政务、公共事业等各个行业。

用户通过浏览器，无需编程 （code-free），采用拖拽算子的可视化操作方式，就可实现多人协作，完成海量数据的价值发现工作。它能够帮助用户从海量数据中发现商业最优决策、成本漏洞，生产隐患 等各类数据价值，并提供对业务预测模型或预测结果的可视化呈现。

# 产品概述

基于分布式技术的 iCloudMiner，能够完成传统数据分析工具无力胜任的海量数据挖掘任务，它以分析挖掘为核心，还提供用户管理、数据管理和可 视化、建模可视化、多用户协作几大功能模块。。我们提供多源异构数据的处理方案，让数据挖掘工程师能够在集成了上百种数据探索和机器学习算法的界面上，以可视化的方式对客户运作中沉淀下来的数据进行探索和透视。

在 iCloudMiner 平台上，一个数据挖掘团队中的多位合作伙伴能够轻松进行分工合作，通过共享和互动的方式，协力完成每一项数据挖掘工作。如果需要执行耗时较长的数据探索或模型训练过程，而计算资源又有限，数据挖掘工程师还可以设置自动任务，让 iCloudMiner 在合适的时间，按照指定顺序来执行批量作业，省去人工触发和长时间值机的消耗。

它可以用于提高市场回应率、降低客户流失、检测机械故障、计划预测性维护以及检测错误等。

生态介绍

业界使用最为广泛的数据挖掘流程是 CRISP-DM，它是由 NCR、OHRA、SPSS、Daimler-Benz

等全球企业一起开发出来的数据挖掘方法论。CRISP-DM 没有特定的工具限制，也没有特定领域局限，是适用于所有行业的标准方法论，下图给出 CRISP-DM 所倡议的通用数据挖掘流程。

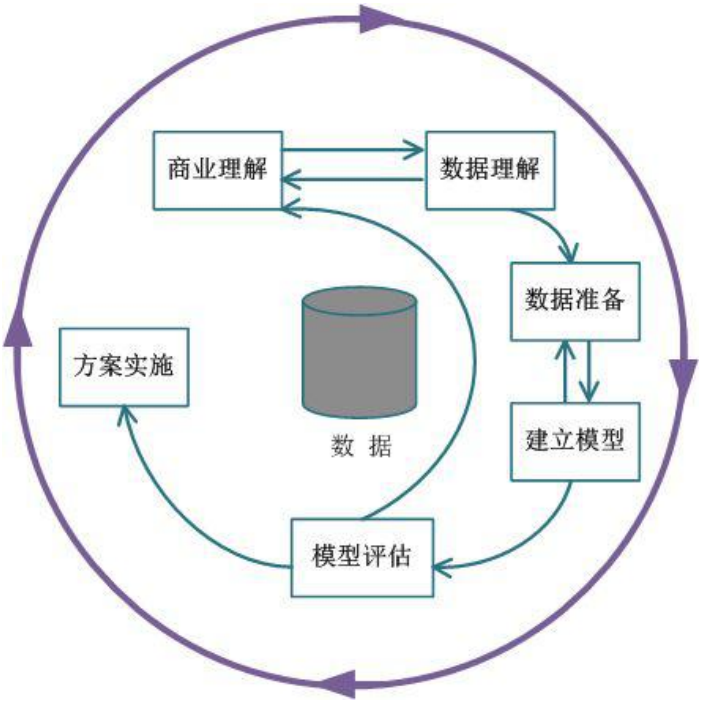


图 3 数据挖掘通用流程（CRISP-DM）

在数之联的 iCloudMiner 数据挖掘平台中，数据挖掘工程师在浏览器中，通过 iCloudMiner 的界面进行可视化业务建模，业务建模流程如下图所示：

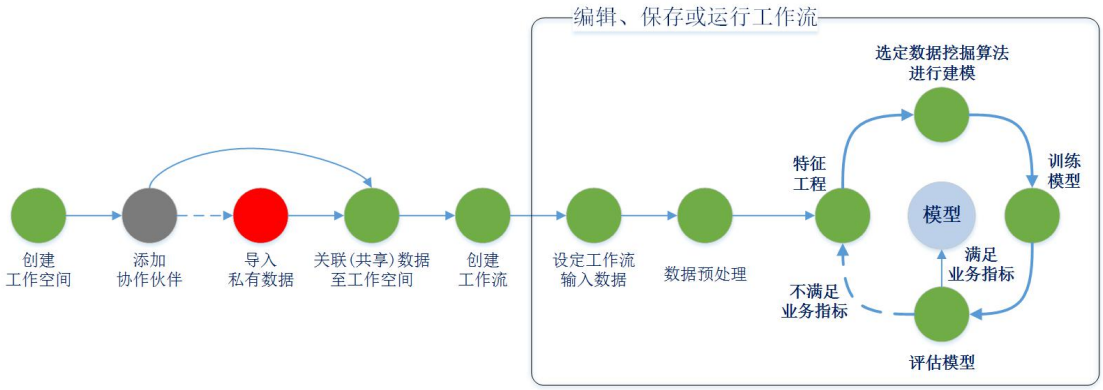


图 4 iCloudMiner 数据挖掘流程

如果训练模型环节得到的模型尚不满足事先指定的性能指标(如分类准确率、召回率等)，则需要重复特征工程、选定数据挖掘算法和训练模型的过程，直至模型评估的结果满足事先指定的性能指标。

在模型训练工作流中，可通过模型持久化算子将模型以指定名称持久化到模型仓库中；当训练出的模型满足事先指定的性能指标时，该模型便可以供上层系统调用。

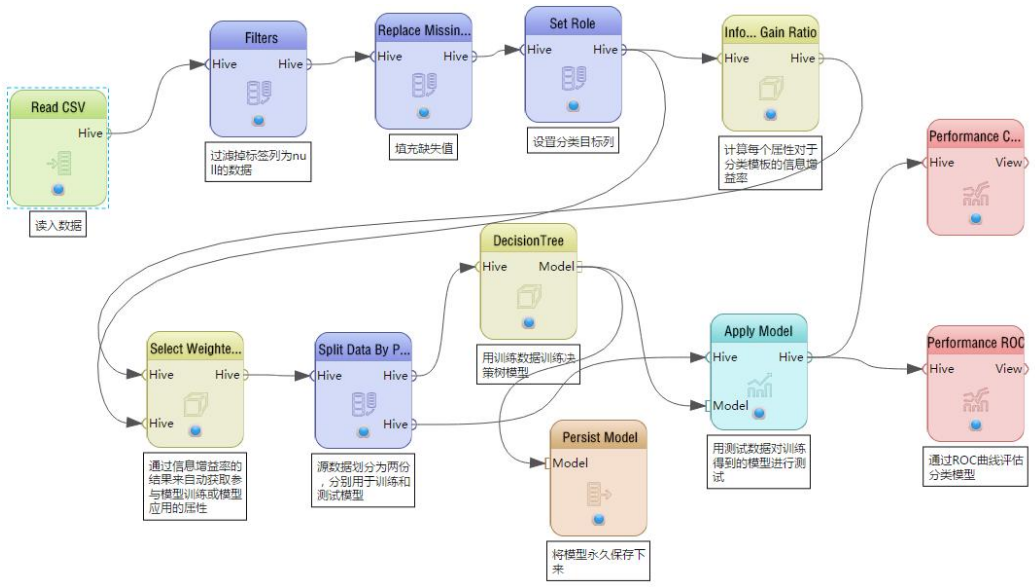


图 5 训练模型所用的工作流示例

一旦训练得到满足事先指定性能指标的业务模型，便可以将训练过程所做的数据预处理、特征工程和应用模型的操作（模型应用算子）组成一个新的可用于生产系统调度的工作流

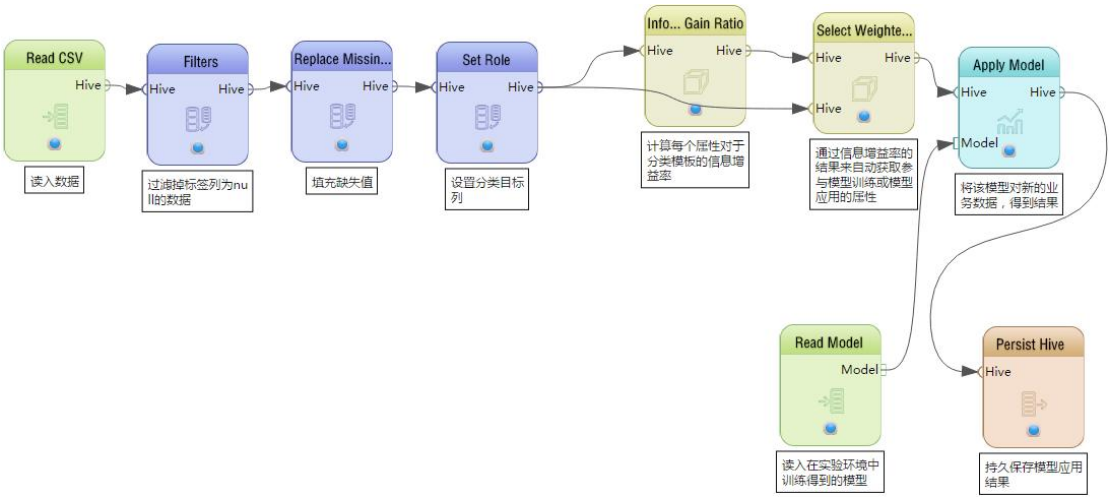


图 6 应用模型所用的工作流示例

# 产品价值

## 商业价值

从短期效益来看，充分利用生产业务数据，开展生产大数据创新应用，能够帮助客户降低生产或运营成本、提高生产或运营效率，从而增加政企单位收益率，提高服务水平。

从长期效益来看，对争气单位的采购、生产、销售、售后、运输、排放等各环节有效地监测预警，能够有效保证生产环节、环保环节的长期安全可控，为客户的管理、数据或业务储备、名声带来诸多有利影响，促进白酒行业互联网转型提速方面将产生积极的引领作用。

## 社会价值

大数据技术的创新应用促使企业长期创造社会价值，促进经济发展。数据驱动的决策和业务创新，能给予企业信誉、可持续发展、社会声誉、投资者关系、商业模式、科技创新、产品与技术和经营管理等各方面积极正面的影响，拓宽企业的业务增长空间。

# 特点与优势

iCloudMiner 是一个分布式的、无需编码的可视化大数据分析与挖掘平台。平台提供简单高效的可视化模型构建能力，用户通过可视化操作即可完成对数据的分布式分析挖掘任务，平台自动将可视化工作流转化为分布式计算框架下的执行任务，并实现工作流的任务调度机制，完成对数据的分析挖掘，平台集成各种数据挖掘和机器学习算法，通过普适化服务层以简单高效的方式为各行业提供大数据价值发现服务支撑。iCloudMiner 平台支持流行的 Hadoop 发行版 CDH 和 HDP，支持在云端进行海量数据挖掘，与传统的单服务器数据分析平台相比，它具备如下特点和优势：

- ✓ 简单：可视化的操作界面，让用户无需编程，通过直观易懂的图形化拖拽方式即可完成数据挖掘所需的数据探索和业务建模的一系列流程，可大大降低企业发现数据价值的门槛，使得企业更轻松地驾驭大数据。
- ✓ 高效：随时随地可开始、保存和继续进行数据价值发现的工作，大大提升数据价值发现的效率。
- ✓ 协作：多人协作、互动地进行数据挖掘，合作伙伴之间能够随时了解数据挖掘项目的动态和进展，轻松共享建模数据和业务数据探索成果。
- ✓ 安全：支持私有云，客户不必将数据转移至第三方分析机构。内建的数据访问控制机制保证业务和数据的私密性。
- ✓ 开放：对有数据挖掘和业务建模技术的客户，iCloudMiner 对其开放数据价值发现

的过程，让客户充分参与和优化业务探索的过程；对缺乏数据挖掘和业务建模技术的客户，iCloudMiner 团队可对其开放数据价值发现的结果，让客户聚焦于数据规律或价值本身，无需花费人力去了解繁琐的技术实现细节。

## 安全的业务数据访问控制

通过内建的数据访问控制机制，iCloudMiner 能够保证每个用户导入平台或在平台生成的数据的私密性，除非用户主动将数据共享给协作伙伴。另外，用户对已经共享给别人的数据还可以取消共享。

## 支持多源异构数据

iCloudMiner 支持从本地文件系统、关系型数据库（如 MySQL 和 Oracle）或平台底层的 HDFS 导入普通文件、CSV 文件或数据库表数据，从而支撑分析挖掘过程的数据需求。未来还将根据客户需求，不断完善支持更多的数据源和数据格式。

## 丰富的数据挖掘算法

iCloudMiner 集成了数近百种数据挖掘过程所需的函数或算法，涵盖特征选择和提取、分类、聚类、回归、关联规则分析、文本挖掘和深度学习算法类别，并将数据读入(Import)、预处理或转换(Attribution Selection, Transform)、写出（Export）操作、模型性能评估(Performance)、模型应用操作（Apply Model, Predicate）统一封装为算子，以保持工作流构建过程的简便和一致性。

## 可视化业务建模过程

iCloudMiner 始终追求数据挖掘过程成本的最小化，可视化的工作流构建界面、典型的建 模样例和准确到位的帮助信息，使得用户无需编程，仅通过拖拽、连线的形式，加上直观 的参数配置，便可以集中精力于创造性的数据挖掘工作中。

## 多人协作和共享

内建的用户权限管理、工作空间和共享机制，使得多个用户可以通过工作空间进行协作，用户能将数据一键分享给工作空间内的其他协作伙伴，也能轻松共享协作伙伴精心积累的数据建模逻辑（工作流），还能将数据和预测模型持久化到 HDFS，或将数据和建模逻辑 下载到本地文件系统。

## 自动任务管理

数据挖掘有时候是一个漫长的过程，可能几小时、几天甚至更长的时间，iCloudMiner 的 作业容器和作业调度功能，使得用户在构建好工作流作业后，只需配置作业容器（Job）和作业列表（Task），并设置起止时间和调度周期，就能轻松实现不同周期粒度的批量 自动任务调度。批量任务一旦执行完成，iCloudMiner 将自动根据用户设定，将任务执行 状态通过邮件发送给自己或用户指定的协作伙伴，使得用户无需值守平台，就能让作业自 动开始、完成，并让用户及时了解作业完成情况。

## 可视化数据和模型预览

在模型逻辑建立好以后，可以将建模逻辑以工作流的形式保存在 iCloudMiner 持久化存储 系统中，并可以随时手动或自动运行自己或协作伙伴的工作流以观察建模效果，我们提供了文本、表格、直方图、ROC 曲线图、树、森林等视图，以满足建模过程中的不同阶 段、不同输出类型数据的预览需求。

# 组成及功能概述

## 产品架构

目前 Hadoop 是大数据的既定事实标准，它是大数据生态系统的核心；而 Spark 是为大规模机器学习设计的高性能分布式计算框架，Spark 与 Hadoop 可形成一种良好的互补的关系。Hadoop 提供了 Spark 所没有的功能特性，比如分布式文件系统 HDFS，而 Spark 为需要

它的那些数据集提供了高性能内存处理。

iCloudMiner 团队以公司多年的技术沉淀为基础，结合开源的大数据生态技术，建立一个大数据挖掘和智能价值发现的平台，把由大数据挖掘或分析的结果作为智能知识的基础，探讨智能知识和传统知识结构的逻辑关系。

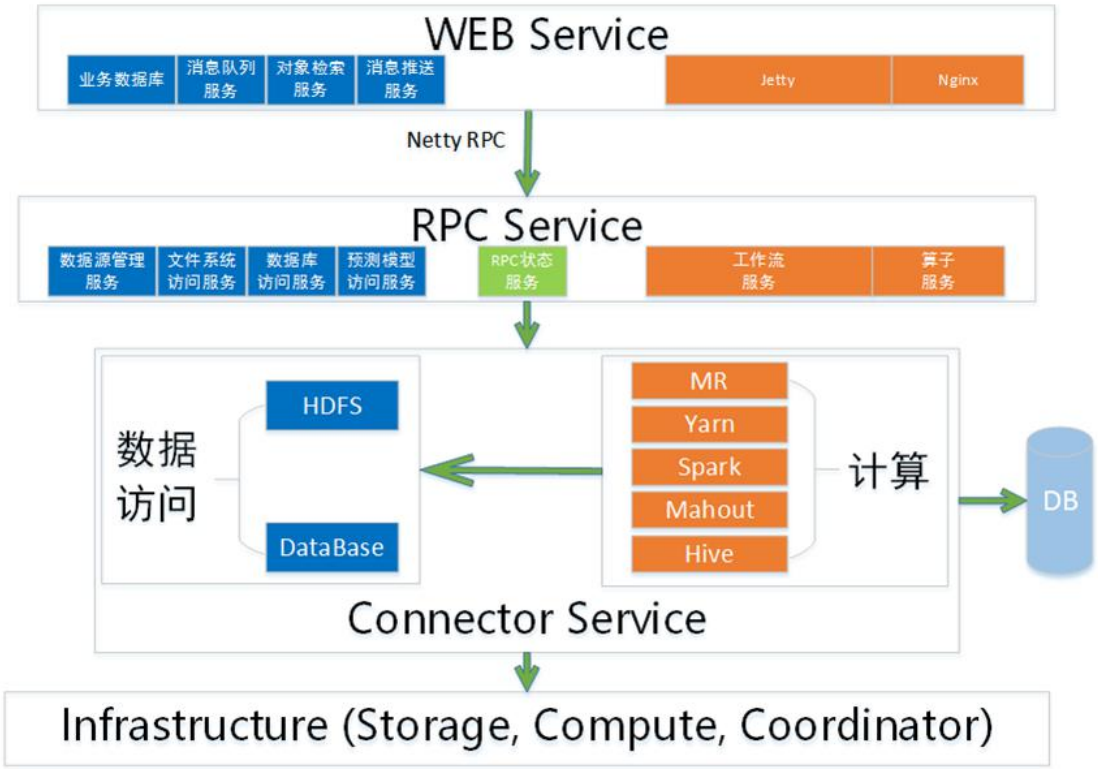


图 1 iCloudMiner 技术架构

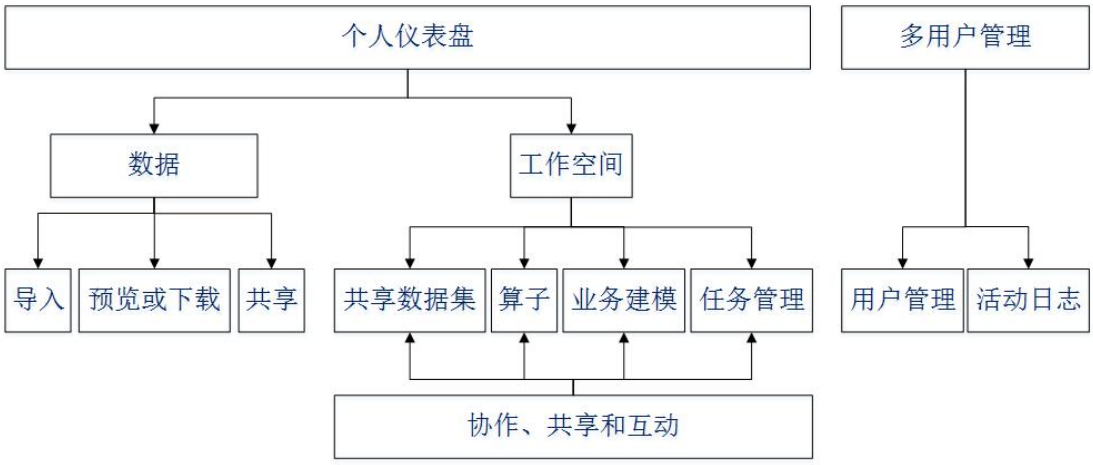


图 2 iCloudMiner 功能架构

# 功能概述

# 管理功能

管理员用户可以增减用户，对已有用户进行信息更改

管理员用户可以将平台活动日志下载到本地

管理员用户可以查看平台 License 信息、重新导入 License

# 用户界面管理

导航式的主功能界面

可进行模块定制的用户个人主页

面包屑风格的 HDFS 数据目录视图

自由的标签分类风格

通过可拖拽和连线的方式构建有序工作流

# 数据管理

在 HDFS 上创建文件夹

上传文件到 HDFS

删除 HDFS 上的文件或文件夹

导入 HDFS 上指定的文件到 iCloudMiner 存储系统

从关系型数据库导入表数据到 iCloudMiner 存储系统

# 数据预处理

丰富的聚合函数

选择合理的特征类型

对特征数据进行各种转换

## 数据挖掘

将诸多算法抽象为输入输出风格一致的算子，易于构成建模逻辑工作流

多人协作以共享数据和挖掘逻辑

随时提交工作流，或终止运行中的工作流

实时地查看运行中的工作流执行状态和结果报告

查看工作流最近的运行历史

## 批量作业管理

将多个工作流设置为顺序执行的批量作业，并能轻松调整每个任务内的作业执行顺序。自动调度批量作业

## 活动和通知

与用户相关的几乎所有活动都会得到记录并呈现给用户

后台主动推送事件通知

作业完成状态邮件通知

用户还可以对一些活动加以评论或添加见解

## 标签和搜索

为平台内各种元素，如数据集、工作空间和工作流设定标签，方便归类和定位

全局搜索平台内各种元素

# 产品环境配置

下表列出了已经过 iCloudMiner 测试验证过的典型系统要求，如果客户的集群环境与此表有差异，需另行调研和测试验证。

表格 1 iCloudMiner 典型基础大数据平台要求

| 组件名称 | 组件版本 | |
|---|---|---|
| **Hadoop** | **CDH** | **HDP** |

| 发行版 | 5.6.0 | 2.3.2 |
|---|---|---|
| Hadoop | hadoop-2.6.0+cdh5.6.0+1025 | 2.7.1 |
| Spark | spark-1.5.0+cdh5.6.0+115 | 1.4.1 |
| Hive | hive-1.1.0+cdh5.6.0+377 | 1.2.1 |
| Sqoop | sqoop-1.4.6+cdh5.6.0+33 | 1.4.6 |

表格 2 iCloudMiner 典型部署和客户端软硬件要求

| 资源分类 | 资源名称 | 规格要求 |
|---|---|---|
| **部署服务器硬件环境** | CPU | 主频 3GHz+ |
| | | 核心数目 8+ |
| | 内存容量 | 32GB+ |
| | 硬盘容量 | 1TB+ |
| **部署服务器软件环境** | 操作系统 | CentOS 5.5~6.7 |
| | JDK | 1.7 |
| | PostgreSQL | 9.35 |
| **客户端软件环境** | 浏览器 | Chrome 32+ |
| | | Firefox 32+ |
| | | Internet Explorer 9+ |