



MININGLAMP  
明 略 数 据

# DataInsight: 企业级的大数据挖掘平台

---

## 产品白皮书

明略数据

2015/10/1

明略 DataInsight 是一款企业级的大数据挖掘平台产品。明略 DataInsight 应用先进的大数据技术，帮助企业实现在海量数据上的数据挖掘，获取隐藏在大数据下的知识，为企业创造新的业务价值。

明略 DataInsight 专注于企业大数据挖掘的全过程。通过提供一体化、并行化的高效数据挖掘工具和模型应用平台，帮助企业提高大数据挖掘落地速度，降低大数据挖掘落地成本。

## 一体化大数据挖掘应用平台

明略 DataInsight 提供高效的建模工具帮助企业在大数据上进行数据挖掘，同时提供模型应用系统帮助企业整合从模型开发到模型上线的大数据挖掘落地的全过程，真正使得大数据挖掘能在企业轻松落地。

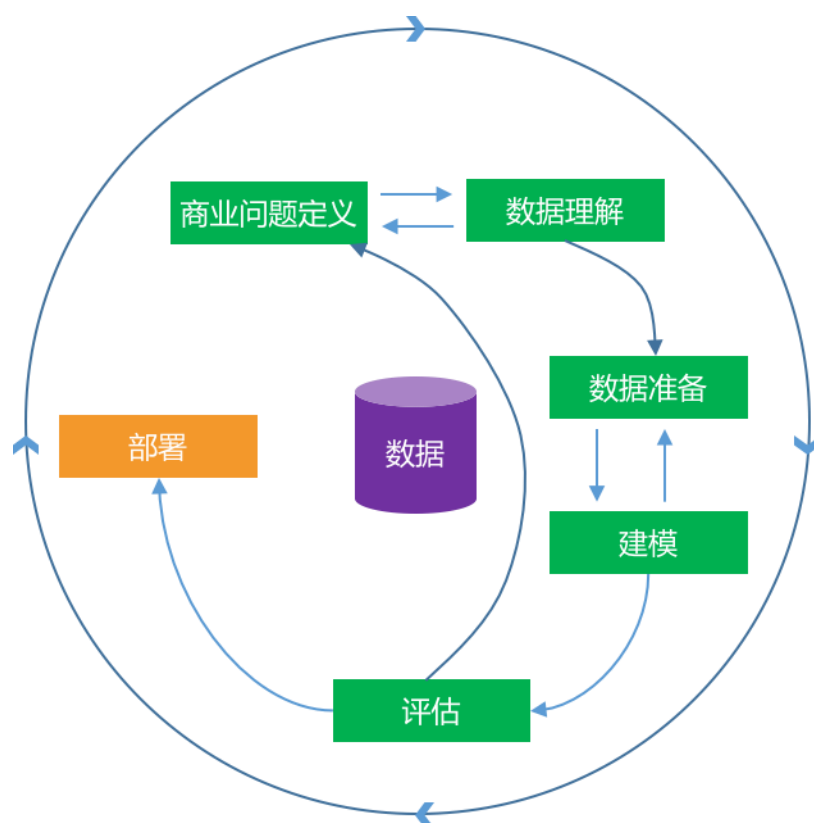


图 1 跨行业数据挖掘标准流程

1999 年的《跨行业数据挖掘标准流程》定义了数据挖掘的 6 个步骤。但是，传统的数据挖

掘软件往往只涉及 6 个步骤中的前 5 个步骤，即只关心模型如何建立，模型建立完成后，如何部署、应用该模型，则很少有软件涉及。但是，从企业数据挖掘实践来看，建立模型只是企业数据挖掘工作的一小部分，后续还有大量的模型部署、更新、维护的工作。目前这些工作缺乏良好的系统来进行管理，导致数据挖掘落地时间长、效率低。

明略 DataInsight 从企业业务落地出发，帮助企业有效的管理大数据挖掘的各个阶段，不单单为企业提供了高效的建模工具，同时也提供模型应用管理系统帮助企业管理已经建立好的模型，降低了模型部署、更新等维护的成本。明略 DataInsight 是基于大数据平台的产品，这就意味着明略 DataInsight 的实验环境和生产环境融为一体。在实验环境中建立的模型可以很方便的在生产环境中应用，降低了模型的部署和迁移成本。

## 并行化大数据挖掘运行平台

大数据时代的特点是数据量规模巨大，传统的单机算法已经无法满足大数据时代的需要，我们需要使用并行算法来处理和计算大数据。明略 DataInsight 是一款基于 Apache Spark 架构的真正意义上的大数据挖掘平台，其通过 Spark 提供的高性能内存迭代计算引擎在多个节点上并行挖掘，解决了单机节点无法挖掘海量数据的问题，同时提高了挖掘速度。



图 2 DataInsight 并行架构

明略 DataInsight 自主研发了大量并行运行在 Spark 之上的数据挖掘算法。这些算法覆盖了数据挖掘工作中最常用的算法种类。用户可以非常方便的在明略 DataInsight 中使用这些算法来创建模型。

在大数据挖掘实践中，使用并行化的数据挖掘算法只是整个挖掘工作中的一部分，更多的工作集中在数据预处理方面。在大数据环境下，单机处理海量数据显然是不现实的。明略 DataInsight 提供了多种数据预处理的并行化算法，帮助用户高效的对数据进行处理。

明略 DataInsight 是完全基于 Hadoop 和 Spark 的并行化的平台，其计算能力受限于整个大数据平台的整体计算能力。当大数据平台的节点得到扩展时，明略 DataInsight 的计算能力也随之扩展。

## 高效的大数据挖掘落地平台

明略 DataInsight 专注于提高企业大数据挖掘的效率。在大数据挖掘实践中，往往 70%-80% 的时间和精力耗费在数据探索和数据处理上面。因此，提高数据探索和数据处理的效率会大大加速整个大数据挖掘的落地速度。

明略 DataInsight 提供了交互式可视化的数据探索工具。明略 DataInsight 帮助用户实时对数据数据进行探索，来指引其寻找更好的解决方法。同时，明略 DataInsight 提供了大量的图表形式供用户更加直观的感受数据，寻找数据中的规律。

在数据处理方面，明略 DataInsight 除了提供大量的并行化数据处理算法之外，还对 SQL 进行了良好的支持。用户可以通过 SQL 语句对建模的中间结果进行各种操作，这样，具有 SQL 技能的用户可以零学习成本的使用 SQL 对数据进行各种探索和处理。

明略 DataInsight 帮助用户真正的处理大数据挖掘问题，加速了大数据挖掘效率，降低了大数据挖掘的成本。明略 DataInsight 提供了大量的富有创造性的特性加速企业大数据挖掘落地。

## 模型 workflow

企业大数据挖掘是一项系统性的工程，其涉及到从数据整合、数据探索、数据处理、模型建立、模型评估、模型调优等一系列的过程。因此，单一的步骤是不足以表示大数据挖掘全过程的。明略 DataInsight 中使用 workflow 的概念来表示整个建模过程。

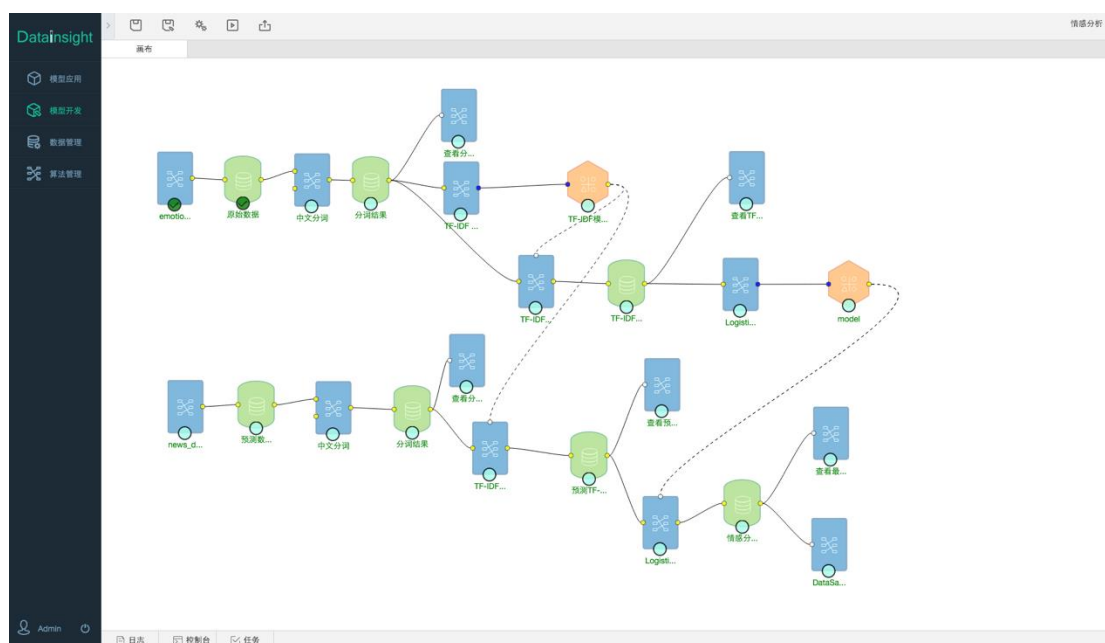


图 3 DataInsight 模型 workflow

在明略 DataInsight 中，每个建模步骤都看做一个算子，每个算子接受若干输入，并且产生若干输出。每个算子的输出都可以作为其他算子的输入，这样，整个建模过程就可以形成一幅有向无环图。建模过程的目的就变成了将原始的输入通过一系列算子组合得到最终的业务结果。

建模过程是一个不断尝试不断探索的过程。用户从原始数据出发，进过对数据的探索和处理，应用合适的算法，最终形成业务上可用的模型。在建模的过程中，会有很多尝试性的步骤，用户可以通过在模型 workflow 中添加分支的方法来进行不同方法的尝试。

当模型开发完毕后，用户可以从模型中生成应用，这样就可以在生产环境中使用应用来产生业务结果。应用从本质上来看依然是 workflow，记录了原始数据如何一步一步的变为最终的业务结果的过程，所不同的是，应用中的 workflow 只保留了产生业务结果的必要步骤，不再保留用户在建模时所做的尝试性的工作。

## 交互式数据探索

对数据的理解是数据挖掘过程中的重中之重。只有理解数据，才能知道如何从数据中挖掘出有价值的信息。数据探索的过程同时也是数据预处理的过程，用户在探索数据时，需要将多种数据进行关联，并且对数据进行不同形式的转换，甚至使用多种算法来对数据进行尝试性的挖掘，来探索数据的意义。

在形成最终模型之前，用户需要通过大量的实验来找到一条切实可行的挖掘方法。在用户进行实验时，对实验的时效性要求非常高。用户希望能够尽快试错，排除那些不可靠的方法，快速的找到可行的方法。

明略 DataInsight 提供了交互式数据探索工具供用户对数据进行实验性的探索工作。明略 DataInsight 通过先进的大数据技术，缩短了数据探索的时间，帮助用户实时的对数据进行探索和实验。

同时，明略 DataInsight 通过可视化的方法，提供了常用的数据统计和分析的图表，供用户能够直观的从图形中发掘数据背后的意义。

## 并行化挖掘算法

明略 DataInsight 是基于 Spark 架构的并行化数据挖掘平台。明略 DataInsight 自主研发了大量的并行化数据挖掘算法，这些算法解决了单机算法不能挖掘大数据的问题，极大的方便了用

户在大数据上的数据挖掘。

此外，明略 DataInsight 是一款覆盖整个建模过程的产品。除了数据挖掘的算法之外，明略 DataInsight 还提供了若干数据处理的并行化算法，同样也可以在模型工作流的算子中使用，对数据进行并行化的处理。

明略 DataInsight 支持的并行化算法列表如下：

算法名称	算法类型
决策树	分类算法
梯度提升决策树	分类算法
随机森林	分类算法
Logistic 回归	分类算法
支持向量机	分类算法
多项式朴素贝叶斯	分类算法
回归树	回归算法
梯度提升回归树	回归算法
回归森林	回归算法
线性回归	回归算法
岭回归	回归算法
Lasso 回归	回归算法
K 均值	聚类算法
主成分分析降维	降维算法
极差归一化	归一化算法
去均值归一化	归一化算法

标准归一化	归一化算法
标准向量生成	向量化算法
哈希向量生成	向量化算法
随机采样	采样算法

表 1 明略 DataInsight 并行算法

## 模型应用管理

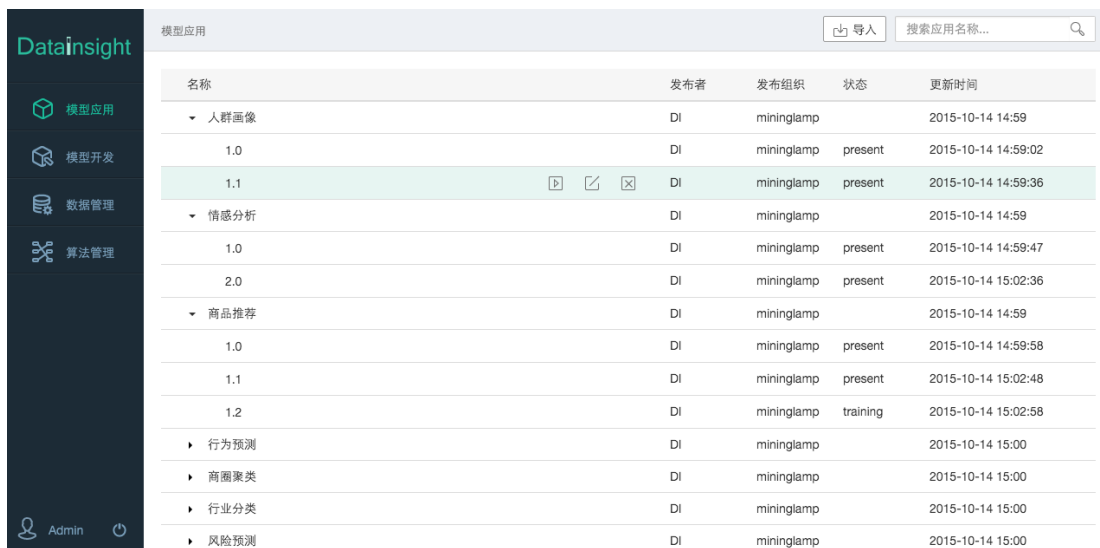
明略 DataInsight 是一款关注企业大数据挖掘落地全过程的产品。除了传统数据挖掘软件提供的建模工具之外，明略 DataInsight 还对建立完成模型的后续使用进行管理，帮助用户更好的在生产环境中应用模型，以实现期望的业务目标。

在企业生产实践中，模型效果是会随着时间而不断衰减的，而且存在模型失效的问题。为了保持模型的效果，就必须对模型进行有效的维护。在传统数据挖掘实践中，有专门的建模团队来负责模型维护，每一次模型维护都要经历模型重部署，重上线的过程，效率十分低下。

明略 DataInsight 中，最终应用于生产的模型称为应用。明略 DataInsight 通过应用更新功能使用新数据重新训练应用，使得衰减效应变缓。应用每次更新完，都会产生一个新的应用版本，这样同一应用会保留多个版本，这些版本的输入输出都保持一致，用户可以任选其中一个版本来产生最终的结果。

应用更新是一种简单而有效的方法，但是当数据发生了较大的变化时，简单的更新应用已不能维持模型效果，此时，需要深入到应用细节对应用进行调整。明略 DataInsight 可以将应用还原为工作流，用户可以在开发环境中打开工作流，在已有的工作流之上重新建模，直到模型效果满足业务需求。





The screenshot shows the 'Model Application' (模型应用) interface of DataInsight. It features a sidebar with navigation options: 模型应用 (Model Application), 模型开发 (Model Development), 数据管理 (Data Management), and 算法管理 (Algorithm Management). The main area displays a table of model applications with columns for Name, Publisher, Publishing Organization, Status, and Update Time. The table lists various models such as '人群画像' (User Portrait), '情感分析' (Sentiment Analysis), and '商品推荐' (Product Recommendation), each with multiple versions and their respective statuses (e.g., 'present', 'training').

名称	发布者	发布组织	状态	更新时间
人群画像	DI	mininglamp		2015-10-14 14:59
1.0	DI	mininglamp	present	2015-10-14 14:59:02
1.1	DI	mininglamp	present	2015-10-14 14:59:36
情感分析	DI	mininglamp		2015-10-14 14:59
1.0	DI	mininglamp	present	2015-10-14 14:59:47
2.0	DI	mininglamp	present	2015-10-14 15:02:36
商品推荐	DI	mininglamp		2015-10-14 14:59
1.0	DI	mininglamp	present	2015-10-14 14:59:58
1.1	DI	mininglamp	present	2015-10-14 15:02:48
1.2	DI	mininglamp	training	2015-10-14 15:02:58
行为预测	DI	mininglamp		2015-10-14 15:00
商圈聚类	DI	mininglamp		2015-10-14 15:00
行业分类	DI	mininglamp		2015-10-14 15:00
风险预测	DI	mininglamp		2015-10-14 15:00

图 4 DataInsight 模型应用

## 模型即服务

明略 DataInsight 作为一款企业级的大数据挖掘平台，通过 Restful API 向企业其他生产系统提供服务，外部系统可以通过调用 Restful API 实现模型的运行和更新等操作。这样，明略 DataInsight 可以作为企业统一的模型服务平台，为企业各个生产系统进行复杂的数据挖掘计算，提供各个生产系统所需的业务数据。例如，明略 DataInsight 可以通过人群画像模型对客户进行画像，将画像结果提供给企业 BI 系统，在企业 BI 系统中结合其他数据一起绘制最终的业务报表。

根据用户在 API 中指定的输入和输出，明略 DataInsight 中相应的应用就可以从输入中读取原始数据，经过数据挖掘的复杂计算后，将结果数据存放到用户指定的输出中，这样就完成了一次数据挖掘任务。例如，在商品推荐系统中，用户可以指定用户行为数据、商品数据等数据源作为输入，将在线推荐系统使用的 Redis 集群作为输出，这样，每次挖掘过后，在线推荐系统可以使用最新的推荐结果进行商品推荐。

明略 DataInsight 支持多种输入输出，包括基于大数据的 HIVE、HBASE、HDFS 等，还包括传

统关系型数据库，如 Oracle、Mysql 等，也包括一些 NoSQL 数据库，如 MongoDB、Redis 等。

明略 DataInsight 支持的 API 如下：

功能	API	说明
运行应用	/application/run	运行应用的一个版本。需要指明应用版本的 ID，运行的输入、输出，以及用户 token。系统从输入中获取数据，经过计算后将输出结果存储到用户指定的输出对象中。该 API 返回一个任务 ID，用户可以使用该 ID 进行后续操作
更新应用	/application/update	更新应用，生成一个新的应用版本。用户指明需要更新的应用 ID，新版本名称、输入的数据和用户 Token。系统通过输入数据新生成一个应用版本。该 API 返回一个任务 ID 和一个版本 ID。
停止任务	/task/kill	用户传入任务 ID 来停止一个正在运行中的任务。
查询状态	/task/status	用户传入任务 ID 来查询任务的状态。状态包括任务正在运行、运行成功、运行失败和已停止。通过查询任务状态来决定下一步操作。

表 2 明略 DataInsight Restful API

通过明略 DataInsight 提供的 API，用户可以在自己的系统中进行调用。例如，用户将一个应用加入工作流中，通过 run 命令运行一个应用，然后不断查询状态，直到状态变为已完成，再继续下一步工作。由于输入输出都是由用户指定，因此，用户可以很方便的在工作流中加入数据挖掘的步骤。

明略 DataInsight 是企业大数据挖掘量身定制的一款产品。明略 DataInsight 将企业的建模环境与最终的生产环境紧密融合，消除了模型后续的部署和二次开发的过程，极大的缩短了模型落地的周期。同时，明略 DataInsight 是一款真正意义上的大数据挖掘产品，通过并行化的架构和提供并行化的算法，真正使得用户可以在海量数据上进行数据挖掘。此外，明略 DataInsight 还提供了大量方便使用的特性，提升了用户建模效率，以及后续模型维护更新的效率。明略 DataInsight 已经帮助多个企业实现了大数据挖掘的落地，我们相信未来还有更多的中国企业将受益于该产品，大数据的价值在企业中真正得到体现。