IBM SPSS Modeler Entity Analytics 16 用户指南

注释

在使用本资料及其支持的产品之前，请阅读第 51 页的『声明』中的信息。

产品信息

此版本适用于 IBM(r) SPSS(r) Modeler V16.0.0 以及所有后续发行版和修订版，直到在新版本中另有声明为止。

目录

前言

IBM® SPSS® Modeler 是 IBM Corp. 企业级数据挖掘工作平台。SPSS Modeler 通过深度的数据分析帮助组织改进与客户和市民的关系。组织通过借助源自 SPSS Modeler 的洞察力可以留住优质客户，识别交叉销售机遇，吸引新客户，检测欺诈，降低风险，促进政府服务交付。

SPSS Modeler 的可视化界面让用户可以应用他们自己的业务专长，这将生成更加强有力的预测模型，缩减实现解决方案所需时间。SPSS Modeler 提供了多种建模技术，例如预测、分类、分割和关联检测算法。模型创建成功后，通过 IBM SPSS Modeler Solution Publisher，在广泛的企业内交付给决策者，或通过数据库交付。

关于 IBM Business Analytics

IBM Business Analytics 软件提供完整、一致和准确的信息，决策者可以信任这些信息来提高企业业绩。企业智能、预测分析、财务业绩和战略管理的完整产品组合，和分析应用程序一起提供对当前业绩的清晰、直接和实用的洞察力，以及预测未来结果的能力。丰富的行业解决方案、久经考验的实践和专业服务，各种规模的组织均可驱动最高生产力、安心地实现决策自动化并提供更出色的业绩。

作为此产品服务组合的组成部分，IBM SPSS Predictive Analytics 软件可帮助组织预测将来的事件，并在该洞察力的基础上提前行动，驱动更好的企业成果。世界各地的商业、政府和学术客户依靠 IBM SPSS 技术作为竞争优势，吸引、挽留和增长客户，同时减少欺诈和降低风险。通过在日常活动中融入 IBM SPSS 软件，组织将变成前瞻性企业，能够指引并实现决策的自动化，以满足企业目标并实现可衡量的竞争优势。有关详细信息或要联系一位代表，请访问 http://www.ibm.com/spss。

技术支持

可以为维护客户提供技术支持。客户可就 IBM Corp. 产品使用或某一受支持硬件环境的安装帮助寻求技术支持。要寻求技术支持，请访问 IBM Corp. Web 站点(http://www.ibm.com/support)。在请求获取帮助时，请准备好标识您自己、您所在的组织以及您的支持协议。

第 1 章 实体分析

关于实体分析

IBM SPSS Modeler Entity Analytics 会将额外维度添加到 IBM SPSS Modeler 预测性分析中。预测分析尝试根据过去的数据来预测将来的行为，而实体分析侧重于通过解析记录本身中的身份冲突来提高当前数据的连贯性和一致性。身份可以指个人、组织、对象或可能不确定的任何其他实体的身份。身份解析在许多领域至关重要，包括客户关系管理、欺诈检测、反洗钱以及国家和国际安全。

假设您有来自两个不同来源的以下客户记录，并且不确定它们指的是同一个人还是不同的人。

来源 1

记录编号: 70001
姓名: Jon Smith
地址: 123 Main Street
纳税参考: 555-00-1111
驾驶执照: 0001133107
信用卡: 10229127

来源 2

记录编号: 9103
姓名: JOHNATHAN Smith
出生日期: 06/17/1934
电话: 555-1212
信用卡: 10229128
电子邮件: jls@mail.com
IP 地址: 9.50.18.77

两个记录之间的数据并不完全一致。但是，如果我们引入第三个来源，便会发现一些共同属性。

来源 3

记录编号: 6251
姓名: Jon Smith
电话: 555-1212
驾驶执照: 0001133107
信用卡: 10229132

驾驶执照号码将来源 1 和来源 3 中的记录联系在一起，而电话号码将来源 2 和来源 3 联系在一起。所以，我们可以合理地确定三个来源全部都指的是同一个人。

但是，如果不这么容易分辨怎么办？我们能作为判断依据的数据可能会非常少。请考虑下面两个记录。

来源 4

记录编号: S45286
姓名: John T Smith Jr
地址: 456 Main Street
电话: 703-555-2000
出生日期: 03/12/1984

记录编号: S45287

姓名: John T Smith
地址: 456 Main Street
电话: 703-555-2000
驾驶执照: 009900991

显然，前面两个记录中的 Smith 先生并不是同一个人，我们完全可以通过其中的差异排除这一点。但我们仍有疑问。两个不同的记录来自同一个数据源，看上去似乎都与同一个人有关。它们是重复记录吗？我们无法确定，除非我们能找到其他相关记录为我们提供更多的信息，或许会从其他来源找到。

来源 5

记录编号: 769582-2
姓名: John T Smith Sr
地址: 456 Main Street
电话: 703-555-2000
驾驶执照: 009900991
出生日期: 06/25/1959

问题到这儿解决了。来源 4 中的两个记录并非重复，而实际上是名字相同、住在同一地址并使用同一电话号码的父子俩。在手动系统上，可能需要进行数周的搜索才能找到一个可解析身份问题的记录。有了自动化实体分析系统，解析时间大大减少。

实体分析与预测分析

如果所有数据都由完整、明确并来自一个来源的记录组成，IBM SPSS Modeler 解析身份冲突要相对简单一些。如果只使用预测分析，您可以将您的数据读入 IBM SPSS Modeler 中，执行处理并获得可靠的结果。

但在现实生活中，情况通常完全不同。数据通常极不完整、常常含糊不清，经常分散在许多不同的数据源中，只是用很少的重叠字段记录许多不同的属性。实体分析的部分价值在于将来自所有不同来源的数据汇集到一个称为**存储库**的中央存储区。然后，实体分析系统会仔细检查数据以解析冲突，同时向源自同一个人或组织的记录添加唯一标识。

下表说明了两种分析类型之间的差别。

表 1. 预测分析与实体分析之间的差异.

特征	预测分析	实体分析
培训数据类型	基于相对较小的集合和数值范围	可以使用诸如名称和地址之类的大集合（无类型的字段）
培训数据大小	通常忽略大的集合（无类型的字段）	使用所有数据
广义化	广义化训练数据的算法，建立简洁模型	数据仍使用适合实体匹配和关系检测的结构
欺诈检测	如果记录具有欺诈性应用程序的典型特征，则将记录标记为可能有欺诈风险	如果记录与已知的欺诈记录有关系，或者记录源自同一个人但身份不同，则将记录标记为可能有欺诈风险

第 2 章 IBM SPSS Modeler 的实体分析

使用 IBM SPSS Modeler 的实体分析

您可能会怀疑数据的身份有问题。例如，某个人可能出现了超过一次，或者有几个人可能看上去已合并或已缺失。 IBM SPSS Modeler Entity Analytics 如何帮助您解决此问题？以下是建议的过程，但您可能需要对此进行相应的调整以满足您的特定需求。

- 将源数据读入 IBM SPSS Modeler
- 创建可存储数据的存储库
- 将 IBM SPSS Modeler 连接到存储库
- 将数据字段映射到存储库特征
- 将数据导出到存储库中，并解析身份
- 分析已解析的身份
- 解析存储库的新个案
- 生成所有所需的警告（批处理或实时）

此时，您需要了解 IBM SPSS Modeler 的工作方式。IBM SPSS Modeler 是非常用户友好的工具，基于经过许多节点的数据流图形表示。每个节点都代表工作流的特定阶段。

IBM SPSS Modeler 提供了多种节点，包括所有标准数据挖掘功能。IBM SPSS Modeler Entity Analytics 则添加了专门用于在实体分析中使用的节点。它们是“EA 导出”节点、“Entity Analytics (EA)”源节点和“流 EA”过程节点。

下图说明了此过程。

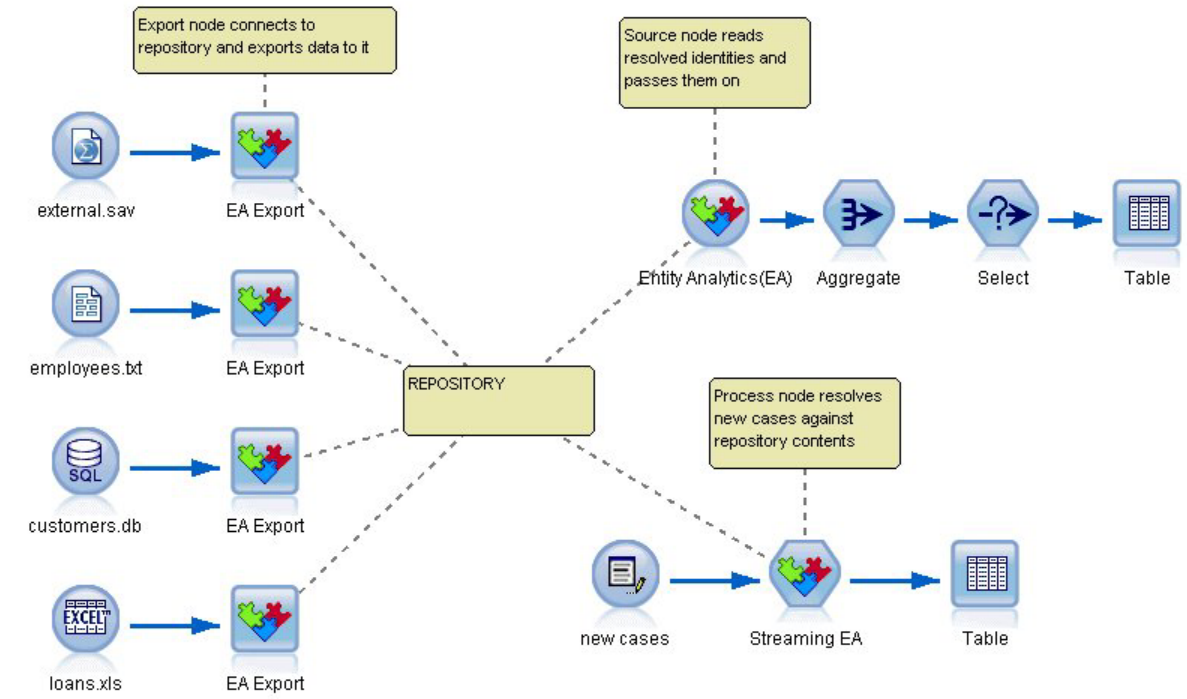

图 1. 实体分析过程

阶段 1: 将源数据读入 SPSS Modeler

首要任务是通过一个或多个源节点将数据读入 SPSS Modeler，在 SPSS Modeler 中以圆形图标表示。

数据可以使用 SPSS Modeler 支持的任何格式，例如文本文件、数据库表、电子表格、XML 文件等，但是每个不同的表格都需要相应的 SPSS Modeler 源节点。在本例中为“数据库”源节点。

每个数据源文件都必须有一个唯一标识每个记录的字段。如果数据源没有这样的字段，您可以轻松地在 SPSS Modeler 中添加一个。请参阅主题第 10 页的『添加唯一记录标识』，了解更多信息。

请参阅主题第 10 页的『连接数据源』，了解更多信息。

阶段 2: 创建存储库

所有实体分析工作的重点都是存储库 - 汇集所有数据记录的中央存储区。

要创建存储库，首先将数据源连接到由方形图标表示的“EA 导出”节点。

您可以从导出节点创建新存储库（或选择现有存储库），准备好接收导出的数据。

稍后详细介绍创建存储库的过程。请参阅主题第 9 页的『设置一个实体存储库（“EA 导出”节点）』，了解更多信息。

注: 如果要以远程服务器方式运行，那么必须在 Modeler Server 计算机中创建存储库（即，创建存储库时，必须通过 Modeler Client 连接到 Modeler Server，从而在服务器计算机上创建 EA 存储库）。

设置完存储库后，您就可以通过各种方法维护其内容。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。

阶段 3: 将 SPSS Modeler 连接到存储库

创建完存储库后，再将其连接到 SPSS Modeler 流。

请参阅主题第 11 页的『实体存储库选项』，了解更多信息。

阶段 4: 将输入字段映射到存储库特征

数据源中可以包含许多不同类型的实体信息。某些信息类型由大多数实体数据源共用，但其他类型可能由特定数据源专用。在实体存储库中，这些不同的信息类型称为**特征**。存储库会将许多特征作为标准特征提供，您也可以创建自己的特征。

存储库特征是可以与实体数据源配合使用的单独信息类型。一些特征（例如，名、姓、出生日期等）可以与许多不同的数据源配合使用，而一些特征则专用于特定数据源。一项特征通常相当于数据记录中的一个字段，或数据库表中的一列。

创建完存储库并连接到该存储库时，将输入数据的一个字段指定为**唯一键**字段，后续分析中会用到此字段。还要将输入数据字段映射到存储库中与之对应的各个特征。映射到预定义的功能会让实体存储库了解到要比较哪些节点并且重要的是如何比较这些节点。 “EA 导出”节点提供了映射表，您可以在其中创建映射。

请参阅主题第 12 页的『将输入字段映射到特征（“EA 导出”节点）』，了解更多信息。

阶段 5: 将数据导出到存储库并解析身份

每个数据源节点都需要自己的“EA 导出”节点，因此，如果您的数据分散在许多不同的源中，那么您的流可能有多个数据源，每个数据源都连接到一个单独的“EA 导出”节点。请参阅主题第 3 页的『使用 IBM SPSS Modeler 的实体分析』，了解更多信息。

如果您有多个数据源，那么可以选择从一个、部分或全部数据源读取记录。Entity Analytics 系统会分析您选择的记录，并为每个记录添加一个名为 $EA_ID 的标识字段。如果与先前模糊的身份相关的两个或多个记录现在已可以解析，那么添加到这些记录的标识在整个存储库是唯一标识。系统还会添加一个字段，以显示记录所源自的数据源。

将每个数据源节点连接到它自己的“EA 导出”节点，将输入字段映射到存储库特征，然后运行流，以将数据从 SPSS Modeler 导出到存储库中，并在一个操作中解析所有身份冲突。为对此过程进行说明，假定您有来自四个不同数据源的下列记录。

外部数据

表 2. 外部数据

名称	电话	信贷风险
Mike	555-1234	560
Joe	555-4567	780

员工

表 3. 员工

名称	地址	电话
Michael	1234 5th Street	555-1234
Fred	543 1st Avenue	555-9876

客户

表 4. 客户

名称	地址	储蓄
Susan	1234 5th Street	1234 美元
Joe	777 Oak Street	5 美元

贷款

表 5. 贷款

名称	地址	电话	贷款
Sue	1234 5th Street	555-1234	10,000 美元
Joseph	777 Oak Street	555-4567	50,000 美元

正如我们所见，您依次将每个数据源导出到存储库中。此过程中，存储库会更新每个记录的解析状态。在存储库中，每个记录前都附加了标识字段（名为 *$EA-ID*）和源指示符字段（名为 *$EA-SRC*），该字段会显示记录所源自的数据源。因此在本例中，您导出所有四个数据源后，存储库内容会如下所示。

表 6. 导出阶段后存储库内容的示例.

$EA-ID	$EA-SRC	名称	电话	地址	信贷风险	储蓄	贷款
1	员工	Michael	555-1234	1234 5th St			
1	外部	Mike	555-1234		560		
2	客户	Joe		777 Oak St		5 美元	
2	外部	Joe	555-4567		780		
2	贷款	Joseph	555-4567	777 Oak St			50,000 美元
3	员工	Fred	555-9876	543 1st Ave			
4	客户	Susan		1234 5th St		1234 美元	
4	贷款	Sue	555-1234	1234 5th St			10,000 美元

Entity Analytics 系统根据共同的电话号码已经确定了*外部*数据集中的 *Mike* 与*员工*数据集中的 *Michael* 是同一个人，并为其分配了标识 1。

*外部*数据集中 *Joe* 的情况有些更为棘手。他与*客户*中的 *Joe* 是同一个人吗？只通过这两个数据源无法进行判断，但我们有第三个源，即*贷款*，其中有一个 *Joseph*。现在，我们发现以下匹配之处：Joseph 的电话号码与*外部*数据集中 Joe 的电话号码相同。基于这一点，系统确定他们是同一个人，并为他提供标识 2。

Fred 没有多个记录，所以为他指定 ID 3。*客户*中的 *Susan* 经确定与*贷款*中的 *Sue* 是同一个人，因为她们有相同的地址，所以为她分配 ID 4。

注：这是为了说明所举出的一个乐观匹配的例子。您可以选择一个更严格的规则集，如此一来，一条简单的姓名与电话或地址本身就无法构成精确匹配，并将同一个标识分配给两条记录。

阶段 6: 分析已解析的身份

解析存储库中的身份冲突后，现在您可以对结果执行进一步分析和处理。例如，如果您怀疑同一身份存在重复的记录可能有欺诈活动，那么可能希望生成一份列出重复项的报告。

首先要创建一个 Entity Analytics (EA) 源节点，并将其链接到存储库。

节点的常见输入由以下字段组成。

- 由系统添加的标识字段（阶段 5 示例中的 *$EA-ID*）
- 由系统添加的源指示符字段（阶段 5 示例中的 *$EA-SRC*）
- 您在阶段 4 中指定的唯一键字段

另外，如果正在查看关系，那么将生成以下输出。请参阅主题第 20 页的『选择数据源』，了解更多信息。

- 实体之间的分离度 (*$EA-DEGREE*)
- 父字段 (*$EA-PARENT*)
- 子字段 (*$EA-CHILD*)
- 识别关系的规则 (*$EA-RULE*)

要查看 SPSS Modeler 中的输出，您可以附加一个 SPSS Modeler 输出节点（如“表”节点）或“报告”节点，并运行流的此部分。如果您需要汇总输出（结果可能非常巨大），那么可以包括“汇总”或“选择”节点之类的记录操作节点。

稍后详细介绍 Entity Analytics (EA) 源节点。请参阅主题第 20 页的『分析已解析身份（Entity Analytics (EA) 源节点）』，了解更多信息。

阶段 7: 解析存储库的新案例

您已经解析所有数据源中的所有记录的身份。但如果您要比较一组新记录以查看它们如何与已知信息关联，从而更好地进行评分，应该如何处理？这种情况下就需要使用“流 EA”节点。

首先您要添加新的 SPSS Modeler 数据源节点，以将您的新数据读取到流中。接下来，将此源节点连接到一个“流 EA”节点。要查看输出，请像之前一样添加“表”节点。

运行流的此部分时，“流 EA”节点会读取每条新记录，并将其与存储库内容进行比较。如果“流 EA”节点在存储库中找到匹配记录，那么该节点会输出所有匹配记录和新记录，并向新记录添加标识字段和源指示符字段。如果未找到匹配项，那么过程节点只输出添加了标识字段和源指示符字段的新记录。

为了说明此过程，假定存储库当前包含 Entity Analytics (EA) 源节点输出的内容。请参阅第 6 页的表 6。

现在我们收到以下新记录。他们与我们已知的人有关吗？

表 7. 要评分的新记录.

名称	地址	电话	贷款
Suzan	1234 5th Street	555-1234	100,000 美元
Mark	888 9th Ave	555-9999	60,000 美元

将新数据与现有存储库内容进行比较，“流 EA”节点将第一条新记录与现有记录中标识为 4 的人员进行比较。但是，未能找到第二个新记录的匹配项，所以为其分配新的唯一标识 5。

“流 EA”节点会添加标识字段和源指示符字段，并输出新记录及其所有匹配记录。因此，输出将如下所示。

表 8. “流 EA”节点的输出.

$EA-ID	$EA-SRC	名称	电话	地址	信贷风险	储蓄	贷款
4	客户	Susan		1234 5th St		1234 美元	
4	贷款	Sue	555-1234	1234 5th St			10,000 美元
4	新贷款	Suzan	555-1234	1 2 3 4 5 t h Street			100,000 美元
5	新贷款	Mark	555-9999	888 9th Ave			60,000 美元

然后，此输出可将实体分析标识作为汇总密钥进行汇总，然后传递到下游节点以作进一步处理。

稍后详细介绍“流 EA”节点。

阶段 8: 生成警报

潜在的可疑活动可能会再次显现出来。在本例中，标识为 4 的人已经有了一笔 10,000 美元的贷款，而且现在正在用另一个姓名申请另一笔十倍于此金额的贷款。当然，这可能是完全可以接受的，并且没有任何欺诈企图。不过，如果按照您的业务规则，此类行为被视为可疑行为，那么就值得查看。

例如，您可以附加并运行 SPSS Modeler“表”节点或“报告”节点，印出其输出窗口的内容，然后找人阅读并手动生成警告。或者，您可以将“流 EA”节点的输出传递给先前已经在 IBM SPSS Modeler 中创建的风险评估模型，生成更切合您的业务规则的一组得分。另一个可能的方法是将 Entity Analytics 过程节点输出导出到数据库或一些其他媒体进一步处理。使用 IBM SPSS Modeler，您将有许多操作可以选择以满足您特定的需要。

第 3 章 实体分析任务

关于任务

本节介绍以下实体分析任务。

- 设置实体存储库
- 配置实体存储库
- 分析已解析的身份
- 依据实体存储库解析新个案
- 清除实体存储库
- 删除实体存储库
- 将实体分析与其他 IBM SPSS 产品一起使用
- 管理实体分析

设置一个实体存储库（“EA 导出”节点）

设置实体存储库的过程由以下任务组成。

1. 连接到数据源。请参阅主题第 10 页的『连接数据源』，了解更多信息。
2. 创建存储库。请参阅主题第 10 页的『创建存储库』，了解更多信息。
3. 将数据源中的输入字段映射到存储库中的特征。请参阅主题第 12 页的『将输入字段映射到特征（“EA 导出”节点）』，了解更多信息。

设置完映射后，您就可以针对当前数据源或针对存储库已知的所有数据源显示这些映射。请参阅主题第 13 页的『显示字段映射（“EA 导出”节点）』，了解更多信息。

注: 自 V16 起，SPSS Entity Analytics 支持 IBM DB2 产品中的存储库。因为存储库特定于 SPSS Modeler 的某个版本，并且无法从较早版本导入；如果您已有某个存储库，并且要升级到 SPSS Entity Analytics V16，那么必须在新的 DB2 数据库中重新创建该存储库。

实体存储库

存储库提供了一个中心存储区域，用作所有实体信息的数据高速缓存。因为存储库是实时的，它具有单一状态，所以实体存储库没有版本控制的概念。存储库会保留所有输入数据的当前状态，并且可能会变得很大。

您可以通过易于使用的图形界面维护存储库内容。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。

重要事项: 从 V16 开始，IBM SPSS Modeler Entity Analytics 支持在 IBM DB2 产品上使用存储库；之前版本的 SPSS Entity Analytics 支持在 IBM solidDB 上托管的存储库。如果已存在 solidDB 存储库，那么当升级到 SPSS Entity Analytics V16 或更高版本时，将需要在在新的 DB2 数据库中创建该存储库。

注意: IBM SPSS Modeler Premium 随附的 IBM SPSS Modeler Entity Analytics 版本仅支持在与 SPSS Entity Analytics 捆绑在一起的 IBM DB2 产品上托管的单个存储库。使用此版本时，您必须先删除现有存储库，然后才能创建新存储库。此 SPSS Entity Analytics 有一个单独许可的升级版（即我们所知的 IBM SPSS Modeler

Entity Analytics Unleashed 版）允许在同一系统中同时存在多个存储库，每个存储库可以容纳超过 1000 万栏资料，且可使用四核处理器。请与当地的 IBM 支持代表联系，以获得详细信息。

连接数据源

首先要通过源节点将源数据读入 SPSS Modeler。

连接数据源

1. 从 SPSS Modeler 主窗口底部节点选项板上的“源”选项卡，双击与源数据类型对应的图标。这样可将源节点添加到屏幕画布。
2. 在屏幕画布上，双击该图标以打开其对话框。
3. 在“文件”字段中，输入源数据文件的位置和名称。
4. 根据需要完成该对话框的其余部分（单击“帮助”按钮了解详细信息），然后单击“确定”。
5. 如果源数据文件没有唯一标识每个记录的字段，可通过“派生”节点添加该字段。请参阅主题『添加唯一记录标识』，了解更多信息。

添加唯一记录标识

每个输入到实体存储库中的数据源文件都必须有一个唯一标识每个记录的字段。如果数据源文件没有这样的字段，您可以通过 SPSS Modeler“派生”节点添加一个。

要为数据源文件添加唯一记录标识

1. 在屏幕画布上，单击您在上一个任务中添加的源节点。
2. 从节点选项板上的**字段选项**选项卡中，双击**派生**图标将“派生”节点附加到源节点。
3. 在屏幕画布上，双击“派生”节点以打开其对话框。
4. 在**派生**字段，为您添加的标识字段，用有意义的名称替换默认名称（如 **ID**）。
5. 确保**派生为**字段设置为**公式**
6. 将**字段类型**设置为连续。
7. 在**公式**文本框中，输入 `@INDEX` 并单击**确定**。

创建存储库

需要创建存储库，以存储所有输入数据。

注：如果要以远程服务器方式运行，那么必须在 Modeler Server 计算机中创建存储库（即，创建存储库时，必须通过 Modeler Client 连接到 Modeler Server，从而在服务器计算机上创建 EA 存储库）。

创建存储库

1. 从 SPSS Modeler 节点选项板的“导出”选项卡，将“EA 导出”节点放在流画布上。

   注：如果您是第一次创建存储库，请使用“EA 导出”节点，并将它连接到包含您要输入到存储库中的数据的 SPSS Modeler 源节点（或者，如果您已添加“派生”节点以获取唯一标识字段，那么可以连接至该“派生”节点）。要连接节点，请执行以下操作。

   a. 右键单击 SPSS Modeler 源节点。
   b. 选择“连接”。
   c. 单击“EA 导出”节点。
2. 双击“EA 导出”节点以打开其对话框。
3. 单击**实体存储库**列表。

4. 单击**<浏览...>**以显示“实体存储库”对话框。
5. 在“实体存储库”对话框上，单击“存储库名称”字段。
6. 选择 **<创建新存储库...>** 以显示创建存储库向导。

创建存储库向导

第 1 步

在这个步骤中，选择是要使用与 IBM SPSS Modeler Entity Analytics 捆绑在一起的 IBM DB2 来创建一个本地存储库，还是为该存储库使用一个外部数据库。

创建本地存储库。为用来托管正在创建的存储库的 IBM DB2 数据库指定管理员用户名和密码。确认密码并单击**下一步**。

注: 您不能在文件名中使用连字符或下划线符号。

必须为 IBM DB2 数据库使用的凭证取决于操作系统。 Windows 用户必须使用用户名 G2user 和密码 G2password。 UNIX 用户必须使用用户名 g2user 和密码 G2password。

Entity Analytics 节点内的存储库管理任务（比如创建或破坏存储库）需要额外的许可权。在 UNIX 上，登录到 IBM SPSS Modeler Server 的用户必须是 root 用户或 g2user，并且必须是 db2iadm1 组的成员。在 Windows 上，登录到 IBM SPSS Modeler Server 的用户必须是 DB2ADMNS 组的成员，以便执行存储库管理。

如果随后您需要更改管理员凭证，可以通过数据库的命令行编辑器进行更改。请参阅主题第 25 页的『管理存储库数据库的管理员凭证』，了解更多信息。

注: 只能有一个用户名和密码组合。所有登录到存储库的用户都共享同一用户名和密码。

添加外部存储库。如果要使用外部数据库来托管存储库，请使用此选项。在**选择存储库 .ini 文件**字段键入数据库 .ini 文件的位置，然后单击**下一步**。

第 2 步

新存储库名称。为新存储库键入唯一名称。

配置导入源。 （仅适用于本地存储库）如果要基于现有存储库进行配置，请在此处选择存储库，否则请选择**默认**。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。

如果您选择现有存储库，而它们与您在上一屏幕上输入的存储库不同，请输入连接详细信息。

单击**确定**以创建新存储库，并显示“实体解析实例”对话框，您可以从此对话框连接到存储库。

实体存储库选项

“实体存储库”对话框中包含用于创建、连接、配置和维护实体存储库的许多选项。

连接到存储库。使用上述选项创建新实体存储库，或连接到现有实体存储库。

- **存储库名称。**显示当前实体存储库（如果存在）。要从多个现有存储库中另选一个存储库，请从列表中选择。

  要创建新存储库，请选择**<创建新存储库...>**。这样会启动一个指导您逐步完成创建过程的向导。
- **用户名。**输入选定存储库的有效用户名。
- **密码。** 该用户名的密码。

- **连接。**单击以连接到当前存储库。

管理存储库。此表列出了加载到当前存储库（您连接到的存储库）的数据源，显示每个数据源中的记录数。

- **刷新。** 更新表格中的数据源和大小信息，例如在添加了新的数据源或更改了现有数据源的大小时。
- **清除所有。**除去存储库中的所有源数据，但保留所有配置详细信息。如果配置信息仍然有用，但您要除去存储库中的所有数据记录，那么可以使用此选项。请参阅主题第 27 页的『清除实体存储库』，了解更多信息。
- **删除未使用的。**除去源数据中突出显示的源数据，但是保留所有配置详细信息。请参阅主题第 28 页的『删除存储库中的未使用数据』，了解更多信息。
- **重命名源。**将打开一个对话框，可以在其中更改突出显示的数据源的名称。

  注: 该选项会重命名存储库中的数据源，您将需要在现有的报告中重新选择此新数据源名称，或者从引用流节点的位置重新选择该节点。

破坏整个存储库。完全破坏当前存储库内容和配置详细信息。请参阅主题第 28 页的『删除实体存储库』，了解更多信息。

配置存储库。显示您可以在其中配置当前存储库的窗口。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。

将输入字段映射到特征（“EA 导出”节点）

存储库提供许多预定义的特征作为标准。不同的数据源可能对同一特征的信息类型使用不同的字段名称（例如，**地址 1** 或**地址行 1**）。为了避免重复，需要将输入数据源字段映射到特定的存储库特征。您不需要映射数据集中的每个字段，只需映射可能符合其他数据集中同一特征的字段。

如果数据源使用与其他类型的信息（未在存储库中预定义）对应的字段，那么您可以从“存储库配置”窗口创建新特征。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。

要将输入字段映射到特征

1. 将“EA 导出”节点附加到流画布上的数据源节点。必须将您使用的每个数据源节点附加到它们各自的“EA 导出”节点。
2. 打开“EA 导出”节点以显示“输入”选项卡，该选项卡中包含用于映射输入字段的选项。请参阅主题『映射的存储库输入选项』，了解更多信息。
3. 在“EA 导出”节点上，选择“存储库”选项卡，以查看当前数据源或所有数据源（如果您使用多个数据源）的映射分配。
4. 要保存一组映射分配（例如，使用不同数据源的导出节点），请单击**导出映射**。

您已经完成映射第一个数据源节点后，请针对您要使用的所有其他数据源节点重复此过程。

映射的存储库输入选项

“输入”选项卡包含一些选项，用于将数据源字段映射到可导出到存储库的存储库特征。设置此选项卡上的映射分配，也可单击“存储库”选项卡查看其他数据源的映射，然后单击**运行**导出数据到存储库。

如果您已经将映射集存储在 XML 文件中，您可以通过单击**导入映射**使用映射集。

Mode. 如果您想要将源文件记录添加到现有的存储库内容，请保留**添加到存储库**默认选择。如果您想在添加源记录之前清空存储库内容，但又想保存配置信息，请选择**导出前清除存储库**。

实体存储库。显示当前实体存储库（如果存在）。要从多个现有存储库中另选一个存储库，请从列表中选择。要创建新的存储库，请选择**<浏览...>**以显示可用来创建存储库的对话框。请参阅主题第 11 页的『实体存储库选项』，了解更多信息。

与实体类型映射。在存储库中定义的实体类型（即特征集）列表。从列表中选择一项，或者选择 **<添加新的实体类型...>** 以显示存储库配置窗口，您可在此窗口中定义新的实体类型。请参阅主题『配置实体存储库』，了解更多信息。

源标记。这是标记列表，指示存储库目前所知的数据源。从列表中选择一个，或选择 **<添加新的源标记...>** 以为新数据源创建一个标记。

唯一键。（必需）要用于数据记录的唯一标识的输入字段。

映射表。在此表中，您可以将每个输入字段映射到存储库中的相应特征。如果所选实体类型中不存在适合的特征，您可在此处创建新特征。

- **字段。**这是所选数据源中的输入字段集。每个字段都有一个图标，指示该字段的测量级别（即数据类型）。
- **映射到特征。**要将字段映射到特征，请在字段行双击此列（或按空白栏）并从列表中选择一个特征。如果没有适合的特征，请选择 **<添加新特征...>** 以显示存储库配置窗口，您可在此窗口中为此实体类型定义新特征。请参阅主题『配置实体存储库』，了解更多信息。
- **用法。**指示特定字段的环境，可以有多个环境，例如，家庭和工作电话号码。“地址”和“电话”特征有可用的预设用法类型，您可以为所有特征创建自己的用法类型。要设置不同于默认（**自动**）的用法，请在所需行上单击此列，然后选择一个现有用法类型（如果有）或者单击**<添加用法...>** 创建一个新的用法类型。请参阅主题第 17 页的『保留实体类型』，了解更多信息。

导入映射。从外部 XML 文件导入先前导出的一组字段到特征映射。如果不同数据源的映射需求相同，那么这可能会非常有用，因为这可避免必须针对不同的源重新定义相同的映射。

导出映射。将映射表中显示的一组字段到特征映射导出到外部 XML 文件。

显示字段映射（“EA 导出”节点）

在“存储库”选项卡上，单击**刷新**按钮，以查看输入字段映射到哪些存储库特征。您可以针对当前数据源（附加到此导出节点的源节点所控制的源）或针对所有数据源查看此内容。

显示输入。 选择一个选项以显示当前数据源的映射或存储库已知的所有数据源的映射。

刷新。 更新所选输入选项的显示。

特征。 在所显示的数据源中具有映射的所有特征的列表。不显示未映射的特征。

<数据源>。 每列会针对已为其定义映射的每个特征列出特定数据源中的映射字段。

配置实体存储库

您可以从“存储库配置”窗口维护存储库内容，该窗口为整个存储库提供了易于使用的可视界面。

如果您要使用配置相同或类似的多个存储库，那么可以设置基本配置，并将其导出到随后可以导入其他存储库中的文件。请参阅主题第 19 页的『重用存储库配置』，了解更多信息。

注: 自 V16 起，SPSS Entity Analytics 支持 IBM DB2 产品中的存储库。因为存储库特定于 SPSS Modeler 的某个版本，并且无法从较早版本导入；如果您已有某个存储库，并且要升级到 SPSS Entity Analytics V16，那么必须在新的 DB2 数据库中重新创建该存储库。

注意:
如果要对一个包含了数据的存储库的配置进行修改和保存，那么可能会提示您先清除存储库内容，然后再重新装入数据。这样做可以避免让存储库处于不一致的状态。

设置存储库配置

1. 打开任一 Entity Analytics 节点。
2. 单击**实体存储库**列表。
3. 单击 **<浏览...>** 以显示“实体解析实例”对话框。
4. 在“实体解析实例”对话框上，单击**存储库名称**列表。
5. 选择要为其设置配置的存储库。
6. 如果您尚未进行连接，请输入管理员用户名和密码，并单击**连接**。
7. 在**配置存储库**按钮启用时，单击此按钮以显示“存储库配置”窗口。
8. 按下文中的说明创建配置详细信息。

“存储库配置”窗口左侧的导航窗格中包含一个树结构，您可以从此树结构管理存储库的不同特征。

表 9. 存储库配置窗口的主要元素.

功能区	描述	
数据源	显示从所有数据源到不同存储库特征的映射。	请参阅主题『查看数据源映射』，了解更多信息。
特征	创建新特征，或复制、编辑或删除现有特征。	请参阅主题『保留存储库特征』，了解更多信息。
实体类型	创建新实体类型，或管理现有实体类型（复制、重命名、附加或删除特征、删除）。	请参阅主题第 17 页的『保留实体类型』，了解更多信息。
解析规则	设置实体匹配的阈值。	请参阅主题第 19 页的『设置实体匹配的阈值』，了解更多信息。

查看数据源映射

在“存储库配置”窗口的“数据源”部分，“所有源”条目会提供将所有数据源映射到不同的存储库特征的只读显示。

如果已经将新数据源添加到存储库，可单击**刷新**更新列表。

注: 您不能在此处将数据源添加到存储库。只能通过创建 SPSS Modeler 源节点，并将其连接到 Entity Analytics 导出节点来添加数据源。请参阅主题第 10 页的『连接数据源』，了解更多信息。

保留存储库特征

存储库特征是可以与实体数据源配合使用的单独信息类型。一些特征（例如，名、姓、出生日期等）可以与许多不同的数据源配合使用，而一些特征则专用于特定数据源。一项特征可包含一个或多个元素，而每个元素通常相当于数据记录中的一个字段，或数据库表中的一列。

在“存储库配置”窗口的“特征”部分中，“所有特征”条目提供了维护所有存储库特征的方法。您可以执行以下操作。

- 创建新特征
- 复制现有特征（例如，根据现有特征创建新特征）

- 编辑现有特征
- 删除现有特征

这些任务在本节稍后说明。

特征列表显示此存储库中已经定义的所有特征。列表中的列显示特征可能具有的各种属性。

特征。这是特征名称。特征名称旁的挂锁符号说明该特征被锁定。被锁定的功能无法被删除或复制，只能保存对隐藏属性所作的更改。

频率。 指示此功能可以具有相同值的实体数。有效值为**一个**（比如护照号）、**一些**（比如地址）或**很多**（比如生日）。

独占性。 指示一个实体通常应该只有一个此类特征。例如，出生日期或身份证号码在此处值为**是**，但地址或信用卡号码的值为**否**（因为一个实体可能有多个地址或信用卡）。

稳定性。 指示此特征的稳定性值（即，实体在其作用期限内是否*不可能*更改）。比如，出生日期特征的值为**是**，因为它是不变的，而地址特征则为**否**，因为它有可能更改，因而较不稳定。*注:* 性别在生命周期内通常是稳定的，但因为它常常由于不良数据而被错误指定，因此缺省配置将其值指定为**否**。

隐藏。指示功能是否已被隐藏。其条目或者显示**是**，或者显示**否**。请参阅第 16 页的『隐藏存储库功能』以获取更多信息。

创建新特征

1. 执行下列其中一项操作。
   - 单击“新建特征”按钮（屏幕右侧的顶部按钮）。
   - 右键单击屏幕左侧的导航窗格中的**所有特征**，并选择**新特征**。
2. 完成“添加/编辑特征”对话框。请参阅主题第 16 页的『添加或编辑特征』，了解更多信息。

复制现有特征

1. 在屏幕右侧表的**特征**列中，选择您要复制的特征。
2. 单击“复制所选特征”按钮（屏幕右侧的第二个按钮）。
3. 完成“添加/编辑特征”对话框。请参阅主题第 16 页的『添加或编辑特征』，了解更多信息。

编辑现有特征

注意如果在存储库已含有数据的情况下编辑、删除或隐去某个功能或功能元素的名称，那么随后应清除该存储库，并重新装入数据。这样做可以避免让存储库处于不一致的状态。

1. 在屏幕右侧表的**特征**列中，选择您要编辑的特征。*注:* 您只能编辑您创建的那些特征，无法编辑系统提供的特征。
2. 单击“编辑所选特征”按钮（屏幕右侧的第三个按钮）。
3. 完成“添加/编辑特征”对话框。请参阅主题第 16 页的『添加或编辑特征』，了解更多信息。

删除现有特征

注意如果在存储库已含有数据的情况下编辑、删除或隐去某个功能或功能元素的名称，那么随后应清除该存储库，并重新装入数据。这样做可以避免让存储库处于不一致的状态。

1. 在屏幕右侧表的**特征**列中，选择您要删除的特征。*注:* 您只能删除您创建的那些特征，无法删除系统提供的特征。

2. 执行下列其中一项操作。
   - 单击“删除所选特征”按钮（屏幕右侧的底部按钮）。
   - 右键单击屏幕左侧的导航窗格中的**所有特征**，并选择**删除**。
3. 单击**继续**以确认删除特征。

   注意:
   无法撤销对功能执行的删除操作。

添加或编辑特征

注意如果在存储库已含有数据的情况下编辑、删除或隐去某个功能或功能元素的名称，那么随后应清除该存储库，并重新装入数据。这样做可以避免让存储库处于不一致的状态。

在“添加/编辑特征”对话框上，您可以创建新存储库特征，或者复制或编辑现有特征。

注意: 如果现有功能被锁定了，那么将无法在此对话框中编辑其详细信息。

特征类型。这是指示与特征相关的信息类型的标签。该标签形成了特征标识的第一部分。

描述。特征类型的简要文本说明，仅供参考。

频率。 指示此功能可以具有相同值的实体数。有效值为**一个**（比如护照号）、**一些**（比如地址）或**很多**（比如生日）。

独占性。 指示一个实体通常应该只有一个此类特征。例如，出生日期或身份证号码在此处值为**是**，但地址或信用卡号码的值为**否**（因为一个实体可能有多个地址或信用卡）。

稳定性。 指示此特征的稳定性值（即，实体在其作用期限内是否*不可能*更改）。比如，出生日期特征的值为**是**，因为它是不变的，而地址特征则为**否**，因为它有可能更改，因而较不稳定。*注:* 性别在生命周期内通常是稳定的，但因为它常常由于不良数据而被错误指定，因此缺省配置将其值指定为**否**。

元素表。此特征所包含的元素列表。

- **元素。**元素名称。
- **描述。**元素所提供的内容的简要说明。
- **数据类型。**此元素可用的数据类型。可用类型有: String、Integer、Real 和 Date。

“添加新元素”按钮。将一个新行添加到元素表，以便定义新元素。

“删除元素”按钮。从元素表中删除所选行。您无法撤消此操作。

注意如果在存储库已含有数据的情况下编辑、删除或隐去某个功能或功能元素的名称，那么随后应清除该存储库，并重新装入数据。这样做可以避免让存储库处于不一致的状态。

隐藏。出于数据保护的目的，可以选择在向存储库添加数据时将其隐藏；要隐藏某功能，请选择**是**。请参阅主题『隐藏存储库功能』，了解更多信息。

隐藏存储库功能

出于数据安全的目的，可能想要在向存储库添加数据时将之隐藏，从而降低个人身份信息无意间被泄露的风险。

当隐藏数据导出到存储库中时，则需要一种隐藏方法，这种方法使得系统仍然能对隐藏数据进行实体解析。例如，某人的信用卡详细信息的两个数据记录被隐藏为“anon_s21”和“anon_s9271”，这两个记录即没有关系了；但是如果在记录之间使用一种内部的后台链接，那么系统仍然能够知道某个名称是另一个名称的缩写形式。

链接隐藏数据的后台链接和标识是在创建存储库时生成的，并且对该存储库来说是唯一的。已加密的数据会先存储在存储库内部，然后当有流连接到存储库时被读取。

在配置存储库时，可以为每个功能单独指定是否需要隐藏。如果选择隐藏某功能，那么它的所有元素都会被隐藏，并且无论其用法类型是什么，都会始终被隐藏。请参阅主题第 16 页的『添加或编辑特征』，了解更多信息。

注: 请确保未隐藏所有 SPSS Entity Analytics 的字段，或者确保无法识别返回回来的数据。建议不隐藏至少一个字段（即使只是一个行号），以便将来可以控制重新合并原始数据。

“存储库配置”窗口中功能列表的某列会显示哪些功能已设置为“隐藏”。其条目或者显示**是**，或者显示**否**。

注: 如果现有的存储库包含了任何隐藏功能前的数据，那么必须先清除所有数据，否则不会在隐藏功能和未隐藏功能之间产生任何匹配。

保留实体类型

实体类型是在逻辑上有共同归属的一组存储库特征。例如，专门供与客户数据集一起使用的实体类型可能由姓名、出生日期、性别、地址、电话号码等特征组成。

IBM SPSS Modeler Entity Analytics 存储库带有一套标准的实体类型，您也可以添加自己的类型。

“存储库配置”窗口的“实体类型”部分列出了已经创建的不同实体类型。您可以执行以下操作。

- 创建新实体类型
- 复制现有实体类型（例如，基于现有实体类型创建新实体类型）
- 将特征附加到实体类型
- 从实体类型中删除特征
- 重命名实体类型
- 删除实体类型

实体类型。所选实体类型的名称。

特征。此实体类型所包含的有效特征列表。

用法类型。（可选）指示可能使用此特征的不同上下文。双击此列可添加或编辑用法类型，用逗号和空格分隔用法类型。您在此处指定的值，可定义当用户在“输入”选项卡上单击特征的“用法”列时，显示在“EA 导出”节点或“流 EA”节点上的值。请参阅主题第 12 页的『映射的存储库输入选项』，了解更多信息。

有关用法类型的常规信息:

- • 用法类型为任意标签。
- • 您可以使用几乎任何文本条目来创建用法类型；但是，不允许输入空格和无效字符。
- • 进行输入时，输入内容会自动更改为大写。
- • 您可以有任意数量的用法类型。
- • 用法类型不必有意义，但是，如果您使用对您以及其他用户有意义的命名约定，那么在稍后进行映射时，这些用法类型将对您有帮助。

- • 进行映射时，如果您使用某个用法类型映射某些元素，并使用另一个用法类型映射其他元素，那么将会出现一条以红色字体显示的警告。

通常，尝试将两个字段映射到同一个 feature.element 时会显示错误。用法类型可用于将两个或两个以上字段映射到同一个 feature.element，并在这些字段中进行配对。

例如，如果定义了两个独立的功能：*HOMEADDRESS* 和 *WORKADDRESS*，那么在这两个功能之间不存在任何匹配。如果某个实体的 *HOMEADDRESS* 与另一个实体的 *WORKADDRESS* 相同，它们之间不存在任何匹配，因为它们属于不同的功能。但是，如果您以不同的用法类型对单个功能进行复用，那么解析将认为 *ADDRESS.WORK* 与 *ADDRESS.HOME* 相同。

您可以复用不同功能的用法类型，或者具有不同的用法类型；例如，对电话使用 *HM* 和 *WK*，对地址使用 *HOME* 和 *WORK*。这样做没有问题，因为我们不会根据地址在电话中进行配对；但是，如果用法类型能够保持一致，那么稍后将有助于对字段进行识别和分组。

将多个实体类型送入单个存储库之后，如果您要使用相同的功能，那么用法类型的内容并不重要。例如，如果将实体类型 *COMPANY* 的 *ADDRESS* 用法类型定义为 *WK* 和 *HM*，那么仍然会根据 *WORK* 和 *HOME* 将其作为 *PERSON* 的 *ADDRESS* 用法类型进行配对。

创建新实体类型

1. 右键单击屏幕左侧的导航窗格中的**实体类型**。
2. 选择**新实体类型**。
3. 输入实体类型的唯一名称，并单击“确定”。
4. 将特征附加到实体类型（参阅下一节）。

将特征附加到实体类型

1. 在屏幕左侧的导航窗格中选择实体类型。
2. 单击“附加特征”按钮（屏幕右侧的顶部按钮）。
3. 从可用特征的列表中，选择一个或多个（使用 Ctrl-单击以选择多个特征）并且单击“确定”。

从实体类型中删除特征

1. 在屏幕左侧的导航窗格中选择实体类型。
2. 从屏幕右侧的附加特征表中选择一个或多个特征。按住 Ctrl 键的同时单击，选择多个特征。
3. 单击“分离特征”按钮（屏幕右侧的底部按钮）。

复制现有实体类型

1. 在屏幕左侧的导航窗格中，右键单击您要复制的实体类型。
2. 选择**复制实体类型**。
3. 为新实体类型输入唯一名称，并单击“确定”。
4. 根据需要附加特征到实体类型，或从实体类型删除特征（参阅先前的说明）。

重命名实体类型

注意如果在存储库已含有数据的情况下编辑、删除或隐去某个功能或功能元素的名称，那么随后应清除该存储库，并重新装入数据。这样做可以避免让存储库处于不一致的状态。

1. 在屏幕左侧的导航窗格中，右键单击您要重命名的实体类型。
2. 选择**重命名**。

3. 输入实体类型的新名称，并单击“确定”。

删除实体类型

注意如果在存储库已含有数据的情况下编辑、删除或隐去某个功能或功能元素的名称，那么随后应清除该存储库，并重新装入数据。这样做可以避免让存储库处于不一致的状态。

1. 在屏幕左侧的导航窗格中，右键单击您要删除的实体类型。
2. 选择 **删除** 。
3. 单击**确定**以确认删除实体类型。

   注意:
   无法撤销对实体类型执行的删除操作。

设置实体匹配的阈值

在“存储库配置”窗口的“解析规则”部分中，您选择将出现实体匹配的阈值。

创建存储库时，匹配预设为默认阈值。

如果您在自己的记录中未找到足够的匹配项来执行实体解析，请选择**设置为主动解析**。

选择**设置为默认解析**以从其他设置之一返回默认阈值。

如果找到的匹配项太多，请选择**设置为保守解析**。

要同时为实体和关系构建一个存储库，请选择**包含关系**。请注意，只有当您拥有单独许可的升级（被称为 IBM SPSS Modeler Entity Analytics Unleashed）时，才可使用此选项。

重用存储库配置

如果您已经设置了配置，并要将其用于其他存储库，可以将现有配置导出到 XML 文件，并将该文件导入其他（目标）存储库。这只有在现有安装中才可能实现。例如，无法将存储库配置从 IBM SPSS Modeler 的一个版本迁移到另一个版本，或者从一种数据库类型迁移到另一种数据库类型。

重用现有配置

1. 显示要使用其配置的存储库的“存储库配置”窗口。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。
2. 从该窗口的菜单选择

   配置 > 导出配置。
3. 在“另存为”对话框中，选择导出 XML 文件的名称和位置。
4. 显示目标存储库的“存储库配置”窗口。
5. 从该窗口的菜单选择

   配置 > 导入配置。
6. 在“打开”对话框中，选择已导出的 XML 文件的名称和位置，然后单击**打开**。

保存您的配置更改

将更改保存到配置

从“存储库配置”窗口的菜单中，选择

文件 > 保存。

关闭配置窗口

从配置窗口中退出

从“存储库配置”窗口的菜单中，选择

文件 > 退出。

如果您尚未将更改保存到配置，请单击**确定**以保存更改并退出，或单击**取消**退出，不进行保存。

分析已解析身份（Entity Analytics (EA) 源节点）

在将数据导出到存储库中之后，可以使用 Entity Analytics (EA) 源节点将已解析的身份传递到其他 IBM SPSS Modeler 节点，进行进一步分析或处理，例如创建一份列出已解析身份的报告。

分析已解析的身份

1. 将 Entity Analytics (EA) 源节点添加到流中。
2. 打开 Entity Analytics (EA) 节点。
3. 在“数据”选项卡上，选择实体存储库，以及其中一个或多个输入数据源（单击**刷新**更新记录计数）。请参阅主题『选择数据源』，了解更多信息。
4. 将其他节点添加到流中，以执行所需的处理。请参阅主题第 21 页的『将节点添加到流』，了解更多信息。

选择数据源

在“数据”选项卡上，至少选择存储库中一个要对其执行进一步处理的数据源。要更新所列出的数据源的记录计数，请单击**刷新**。

实体存储库。显示当前实体存储库（如果存在）。要选择另一个存储库（如果存在多个存储库），请从列表中选择一个。要创建新的存储库，请选择**<浏览...>**以显示可用来创建存储库的对话框。请参阅主题第 11 页的『实体存储库选项』，了解更多信息。

包括来自数据源的记录。此表列出已输入存储库的不同数据源，以及每个源中的记录数。对于您要用于执行进一步分析和处理的数据源，请选择**包括**复选框。要选择或取消选择所有数据源，请分别单击**包括所有**或**排除所有**。

关系。选择要包含在存储库中的关系类型。请注意，只有当您拥有单独许可的更新（被称为 IBM SPSS Modeler Entity Analytics Unleashed）并且已将存储库配置为包含关系时，才可使用此选项。

- **没有关系。**未使用关系详细信息。
- **密切关系。**仅选择密切相关的实体。关系的密切程度取决于很多变量，例如，映射的功能属性、共享的功能以及是设置为保守解析还是主动解析。
- **所有关系。**选择所有相关的实体。

最大分离度。只有在选择了**密切关系**或**所有关系**时才可用。选择用于识别关系的分离度数。例如，如果 Ann 和 Bob 互相不认识，但是 John 同时认识 Ann 和 Bob，那么 Ann 和 Bob 的分离度为 2。

输出实体类型。缺省情况下，如果存储库包含详细信息，那么它显示的是存储库中列出的第一个实体类型。如果存在多个存储库，那么在此选择某个实体类型将更改“过滤器”选项卡上显示的功能，以便列出该类型的功能。可以从存储库使用的任何实体类型中进行选择。

重命名数据字段

可以使用“过滤”选项卡重命名任何传递到下游进行进一步处理的已解析身份字段。您可能希望重命名已解析身份字段，例如，与下游其他数据集合并时，保留字段名的兼容性。

字段及其原始名称如下。

表 10. 已解析身份字段

字段	描述
$EA-ID	实体标识
$EA-SRC	标识记录所源自的数据源的源标记
$EA-KEY	在数据源文件中指定为唯一键的字段

注：尽管您也可以使用“过滤”选项卡过滤字段，但不应在此进行，因为已解析身份字段是实体分析过程所需的绝对最小值。

为数据字段设置类型信息

在“类型”选项卡上，您可以查看或更改已传递到下游做进一步处理的已解析身份字段的各种属性。

您可以更改的属性与常规 SPSS Modeler “类型”节点的“类型”选项卡上的属性相同，如下所示。

表 11. 字段的“类型”属性

属性	描述
测量(M)	测量级别（即数据类型），用于描述字段中数据的特征。
值(V)	提供用于从数据集中读取数据值的选项。
缺失	用于指定字段缺失值的处理方法。
检查(K)	用于确保字段值符合特定值或范围的验证选项。
角色	指定在数据传递到建模节点或模型块时如何使用字段。

将节点添加到流

您可以将不同的 SPSS Modeler 节点添加到流，以对 Entity Analytics (EA) 源节点的输出执行分析或处理操作。例如，您可以添加一个或多个以下节点。

- 汇总或区分节点，用于汇总输出，输出可能非常大。
- 选择节点，用于选择输出的子集
- 表节点，用于从 Entity Analytics (EA) 源节点查看输出
- 报告节点，用于打印报告的输出
- SPSS Modeler 导出节点，用于将输出导出为不同的格式，例如，电子表格或数据库

有关更多信息，请参阅《*IBM SPSS Modeler 源、过程和输出节点指南*》中的有关记录操作、输出和导出节点的部分。

比较新个案与存储库（“流 EA”节点）

当您已执行存储库中的一些身份解析之后，可以使用“流 EA”节点将随后遇到的新案例与存储库内容进行比较。此节点处理新数据源中的记录，将它们与存储库中现存的已解析实体进行比较，并传递所有的匹配记录供进一步处理。匹配可以设置为精确匹配，也可设置为与现有实体松散关联。

如“EA 导出”节点一样，“流 EA”节点将单一 SPSS Modeler 源节点作为输入。但“流 EA”节点有所不同，体现在下列几个方面。尽管导出节点会输出与其输入记录相关的所有实体的记录，但“流 EA”节点仅输出与存储库中已解析的条目相关的那些实体的记录。请参阅主题第 24 页的『“流 EA”节点的输出』，了解更多信息。

比较新个案与存储库

1. 连接到包含要与现有实体进行比较的新纪录的数据源。请参阅主题第 10 页的『连接数据源』，了解更多信息。
2. 在“记录选项”选项卡上，将“流 EA”节点附加到数据源节点。
3. 双击“Entity Analytics 导出”节点以打开其对话框。
4. 单击**实体存储库**列表。
5. 单击**<浏览...>**以显示“实体存储库”对话框。
6. 在“实体存储库”对话框上，单击“存储库名称”字段。
7. 单击要使用的存储库的名称。
8. 为此存储库输入用户名和密码，然后单击**连接**。连接到存储库时请单击**确定**。
9. 在“流 EA”对话框中，选择想要映射的“实体类型”。请参阅主题第 17 页的『保留实体类型』，了解更多信息。
10. 将数据源中的输入字段映射到存储库中的特征。请参阅主题『将输入字段映射到特征（“流 EA”节点）』，了解更多信息。
11. （可选）可以在对数据进行评分时实时更新存储库中的记录。请参阅主题『将输入字段映射到特征（“流 EA”节点）』，了解更多信息。
12. 单击**输出**选项卡可查看已输入存储库并已设置了检索现有实体的选择条件的各种数据源的详细信息。请参阅主题第 23 页的『显示字段映射和数据源（“流 EA”节点）』，了解更多信息。
13. 单击**过滤器**选项卡可查看存储在存储库中的输入字段和功能的详细信息。缺省情况下，会过滤出尚未在节点中映射的所有功能，但是也可根据要求进行更改。
14. 节点设置正确时，请单击**确定**。
15. 将“表”节点附加到“流 EA”节点并运行流。

“表”节点的输出窗口列出与数据源中新纪录相匹配的所有检索到的实体。输出字段添加了前缀 **$EA-**。请参阅主题第 24 页的『“流 EA”节点的输出』，了解更多信息。

注：在运行“流 EA”节点时，可能会遇到错误“**在服务器数据模型中检测到的字段数不正确**”。如果您在创建“流 EA”节点后编辑了存储库配置，那么可能会发生此错误。在这些情况下编辑配置相当于更改了从该节点输出的字段的数目和名称。要解决此问题，请打开“流 EA”节点，然后单击**刷新**按钮。这样做会导致重新计算输出字段的数目和名称。

将输入字段映射到特征（“流 EA”节点）

“输入”选项卡包含一些选项，用于将此节点的输入中的字段映射到存储库特征。请在此选项卡上设置映射分配，或者选择**查看**选项卡以查看存储库中所有数据源的详细信息，然后单击**确定**。

如果您已经将映射集存储在 XML 文件中，您可以通过单击**导入映射**使用映射集。

实体存储库。显示当前实体存储库（如果存在）。要从多个现有存储库中另选一个存储库，请从列表中选择。要创建新的存储库，请选择**<浏览...>**以显示可用来创建存储库的对话框。请参阅主题第 11 页的『实体存储库选项』，了解更多信息。

与实体类型映射。在存储库中定义的实体类型（即特征集）列表。从列表中选择一项，或者选择 **<添加新的实体类型...>** 以显示存储库配置窗口，您可在此窗口中定义新的实体类型。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。

持久搜索。如果想要在对数据进行评分时实时更新存储库中的记录，请选择此选项。

源标记。仅在选择**持久搜索**时可用。这是标记列表，指示存储库目前所知的数据源。从列表中选择一个，或选择 **<添加新的源标记...>** 以为新数据源创建一个标记。

唯一键。仅在选择**持久搜索**时可用。它是用于数据记录唯一标识的输入字段。

映射表。在此表中，您可以将每个输入字段映射到存储库中的相应特征。如果所选实体类型中不存在适合的特征，您可在此处创建新特征。

- **字段。**这是所选数据源中的输入字段集。每个字段都有一个图标，指示该字段的测量级别（即数据类型）。
- **映射到特征。**要将字段映射到特征，请在字段行双击此列（或按空白栏）并从列表中选择一个特征。如果没有适合的特征，请选择 **<添加新特征...>** 以显示存储库配置窗口，您可在此窗口中为此实体类型定义新特征。请参阅主题第 13 页的『配置实体存储库』，了解更多信息。
- **用法。**指示特定字段的环境，可以有多个环境，例如，家庭和工作电话号码。请参阅主题第 17 页的『保留实体类型』，了解更多信息。

导入映射。从外部 XML 文件导入先前导出的一组字段到特征映射。如果不同数据源的映射需求相同，那么这可能会非常有用，因为这可避免必须针对不同的源重新定义相同的映射。

导出映射。将映射表中显示的一组字段到特征映射导出到外部 XML 文件。

显示字段映射和数据源（“流 EA”节点）

在“输出”选项卡上，可以查看已经输入到存储库中的各种数据源的详细信息。这些是针对其处理此节点的输入以搜索并检索匹配实体的数据源。单击**刷新**以更新记录计数。

包括数据源的匹配项。此表列出了存储库中各种可用数据源及其每个源的记录数。

匹配项。这些选项指定您在“输入”选项卡上指定的字段到特征映射信息与候选者记录（即，整个存储库内容）的匹配程度。匹配标准越接近，检索到的实体越少。

注: 如果找到 20 个以上的匹配项，那么将只返回找到的前 20 个。

- **只包括完全匹配项。**这是最接近的匹配标准，它导致选择的记录数最少。请在想要仅返回完全匹配的实体时使用此选项。
- **包括可能的匹配项。**当您希望返回匹配实体以及共享相同标识的实体（具有频率值配置为一个的特征，如，匹配信用卡号码、报税 ID 号等）时，可使用此选项。
- **包含所有匹配项。**当您希望查看存储库中具有共享特征的最大数量的可能实体时，可以使用此选项。这是最宽松的匹配标准，它导致选择的记录数最多。此选项可返回完全匹配和几乎共享所有特征的实体（一般是频率值为一个或几个的特征）。例如，包括具有相同报税 ID 号的实体和具有相似地址的实体。

关系。仅当已将存储库配置为包含关系时才可用。要将存储库配置为包含关系，必须拥有单独许可的升级（被称为 IBM SPSS Modeler Entity Analytics Unleashed）。选择想要包含在输出中的关系类型。

- **没有关系。**未使用关系详细信息。
- **密切关系。**仅选择密切相关的实体。关系的密切程度取决于很多变量，比如映射功能属性、共享的功能以及解析是设置为保守的还是主动的。
- **所有关系。**选择所有相关的实体。

最大分离度。只有在选择了**密切关系**或**所有关系**时才可用。选择用于识别关系的分离度数。例如，如果 Ann 和 Bob 互相不认识，但是 John 同时认识 Ann 和 Bob，那么 Ann 和 Bob 的分离度为 2。

输出实体类型。缺省情况下，如果存储库包含详细信息，那么它显示的是存储库中列出的第一个实体类型。如果存在多个存储库，那么在此选择某个实体类型将更改“过滤器”选项卡上显示的功能，以便列出该类型的功能。可以从存储库使用的任何实体类型中进行选择。

“流 EA”节点的输出

“流 EA”节点输出检索到的每条记录由下列字段组成:

字段	描述
Field1[, *Field2*[, ... *FieldN*]]	包含新纪录的数据源字段。
$EA-ID	此记录在存储库中的实体标识。
$EA-SRC	识别此记录源自的数据源的源标记。
$EA-KEY	此记录在数据源文件中的唯一键的值。
$EA-SC	匹配近似度字段，标明此记录与存储库中观察到的实体的匹配近似度，其取值范围为 1.0 （低度匹配）到 10.0 （高度匹配）。
$EA-*Feature1*[, $EA-*Feature2*[, ... $EA-*FeatureN*]]	此记录在数据源文件中所映射的特征的值。

如果在存储库中启用了关系字段，并且“输出”选项卡上的分离度显示为大于零，那么“流 EA”节点也将为每个检索的记录包含以下字段。

字段	描述
$EA-DEGREE	分离度
$EA-PARENT	计算分离的源记录标识。
$EA-CHILD	计算分离的目标记录标识。
$EA-RULE	

与其他 IBM SPSS 产品一起使用 IBM SPSS Modeler Entity Analytics

可找到安装程序让您与下列产品一起使用 IBM SPSS Modeler Entity Analytics:

- IBM SPSS Collaboration and Deployment Services
- IBM SPSS Modeler Batch for Windows
- IBM SPSS Modeler Solution Publisher

您需要先运行这些安装程序，才能与这些产品一起使用 IBM SPSS Modeler Entity Analytics 的特征。有关更多信息，请参阅《*IBM SPSS Modeler Premium 安装指南*》。

管理任务

对于那些在 Entity Analytics 中创建的存储库，是使用 IBM DB2 产品来创建新的数据库服务的。有一些与 DB2 关联的管理任务，这些任务通常由数据库管理员或系统管理员来执行，并且可:

- 配置端口分配
- 管理存储库数据库的管理员凭证

所需执行的其他管理任务可应用于所有的存储库，这些任务包括:

- 将存储库移到另一个存储目录
- 为“日期/时间”和“时间戳记”字段设置流属性
- 调整超时设置
- 在同一 Windows 系统中使用 SPSS Modeler 客户端和 SPSS 建模器服务器 服务器端运行 IBM SPSS Modeler Entity Analytics
- 清除实体存储库
- 删除实体存储库
- 在不能与存储库连接时，将存储库删除

配置端口分配

必须为每个 DB2 数据库服务分配一个端口，这个端口不能再分配给正在机器上运行的其他服务。属于运行 IBM SPSS Modeler Server 的同一机器（或当 IBM SPSS Modeler 没有与 IBM SPSS Modeler Server 连接使用时，机器运行 IBM SPSS Modeler）上的数据库服务。

缺省情况下，Entity Analytics 分配范围在 1320 到 1520 之间的端口，对于第一个创建的存储库，从端口 1320 开始。若出现冲突，您可以通过编辑文件 *<modeler server installation path>/ext/bin/pasw.entityanalytics/ea.cfg* 来配置端口分配，并且为 *min_port* 和 *max_port* 设置指定适当的值。此文件的默认内容显示如下:

```
# port range configuration for entity analytics

#

#     this port range controls which ports DB2 databases

#     (created to store Entity Analytics Repositories in)

#     may use.  Configure this if the default port range will

#     introduce a conflict on your system.

#

# default min_port = 1320

# default max_port = 1520

min_port, 1320

max_port, 1520
```

管理存储库数据库的管理员凭证

托管实体存储库的 DB2 数据库的管理员用户名和密码在创建存储库时定义。如果知道当前凭证，那么可以通过 DB2 SQL 编辑器更改这些详细信息。

要启动 DB2 SQA 编辑器

1. 在客户端机器上，打开命令提示符窗口。
2. 输入:

   cd *modeler_install_dir*\ext\bin\pasw.entityanalytics\DB2\bin

modeler_install_dir 是安装 SPSS Modeler 的目录。

3. 输入:

```
solsql -c "C:\Documents and Settings\All Users\Application Data\IBM\SPSS\Modeler\version\EA\
repositories\repos_name
```

version 是 SPSS Modeler 的安装版本号码，*repos_name* 是存储库的名称。

4. 在提示符处，输入当前数据库管理员的用户名和密码，以显示 `solsql>` 提示符。

更改数据库管理员密码

1. 在 `solsql>` 提示符处，输入:

```
alter user username identified by password;
commit work;
```

username 是数据库管理员的当前用户名，*password* 是新密码。

2. 输入 exit; 以关闭编辑器。
3. 重启 SPSS Modeler 客户端。

有关 DB2 数据库可以执行的其他管理任务的信息，请参阅 http://publib.boulder.ibm.com/ 处 IBM DB2 的恰当版本的文档。

将存储库移到另一个存储目录

存储库文件默认存放在名为 *EA* 的下列目录位置下:

- C:\Documents and Settings\All Users\ApplicationData\IBM\SPSS\Modeler*version*\EA （Windows 系统）
- *modeler_install_directory*/ext/bin/pasw.entityanalytics/EA （UNIX 系统）

由于存放存储库的文件可能变得非常庞大，您可能会需要把它们移动到其他磁盘或分区以腾出更多空间。

要将存储库移到另一个目录，请执行以下操作:

1. 退出 SPSS Modeler。
2. 将 *EA* 目录从原始位置（如前所列）移动到新位置。例如，在 Windows 中，您可能希望将它移动到类似 *F:\data\EA* 的新位置。
3. 编辑 *<modeler 服务器安装路径>/ext/bin/pasw.entityanalytics/ea.cfg* 文件来增加以下选项:

```
repository_data_directory, new_location
```

此处的 *new_location* 即为移入了 EA 目录的新位置，例如: *F:\data\EA*。

为“日期/时间”和“时间戳记”字段设置流属性

如果源数据中的字段包含日期/时间或时间戳记数据，请确保相应的流属性设置为 IBM SPSS Modeler Entity Analytics 可以识别的格式。

要设置流属性格式，请执行以下操作:

1. 在主 SPSS Modeler 菜单上，选择:

工具 > **流属性** > **选项**。

2. 选择**日期与时间**。
3. 设置**日期格式**为 **YYYY-MM-DD**。
4. 设置**时间格式**为 **HH:MM:SS**。

5. 单击**确定**。

调整超时设置

在运行速度较慢或负载很重的系统上，如果您在创建或访问存储库时遇到错误，那么可能要增大用于启动和停止实体分析引擎或实体分析数据库服务器的超时设置。

要调整实体分析引擎的超时，请执行以下操作:

1. 退出 SPSS Modeler。
2. 编辑文件 *<modeler server installation path>/ext/bin/pasw.entityanalytics/ea.cfg*，以增大以下选项的值:

```
timeout, value
```

其中 *value* 是实体分析引擎的超时值（以秒计，缺省值为 60）。

要调整实体分析数据库服务器的超时值（仅限 DB2）:

1. 退出 SPSS Modeler。
2. 编辑文件 *<modeler server installation path>/ext/bin/pasw.entityanalytics/ea.cfg*，以增大以下选项的值:

```
timeout, value
```

其中，*value* 是实体分析 DB2 数据库服务器的超时值（以秒计，缺省值为 100）。

在同一 Windows 系统中使用 SPSS Modeler 客户端和 SPSS 建模器服务器 服务器端运行 IBM SPSS Modeler Entity Analytics

若您已在同一 Windows 系统的 SPSS Modeler 客户端和 SPSS 建模器服务器 服务器端安装 IBM SPSS Modeler Entity Analytics，按默认设定，客户端和服务器端会共享同一个存储库。如您希望它们使用不同的存储库，您需要编辑客户端或服务器端的**其中之一**的配置文件 *ea.cfg*，使它们使用不同的端口范围及存储库文件夹。

注: 特别是在您使用一个 32 位的 SPSS Modeler 客户端和一个 64 位的 SPSS 建模器服务器（反之亦然）的情况下，应执行此操作。

1. 打开文件 *<modeler [server] 安装路径>/ext/bin/pasw.entityanalytics/ea.cfg* 进行编辑。
2. 更改 `min_port` 和 `max_port` 设置，以使用与其他系统不同的端口。请参阅主题第 25 页的『配置端口分配』，了解更多信息。
3. 更改 `repository_data_directory` 设置，以使用与其他系统不同的目录。
4. 保存并关闭 *ea.cfg* 文件。

清除实体存储库

如果想要清除实体存储库中的数据记录，但又想保留配置信息，那么可以清除存储库中数据。

要清除存储库的所有数据:

1. 打开 Entity Analytics 节点。
2. 单击**实体存储库**列表。
3. 单击**<浏览...>**，即显示“实体解析实例”对话框。
4. 在“实体解析实例”对话框上，单击**存储库名称**列表。
5. 选择想要清除的存储库。
6. 如果尚未连接，请输入管理员用户名和密码，然后单击**连接**。
7. 在启用了**清除全部**按钮后，请单击此按钮。

8. 在“清除全部数据源”对话框中，单击**清除**以确认清除存储库。

删除存储库中的未使用数据

如果存储库中有一个不再使用或需要在实体存储库中使用的数据源，那么可以删除该源。可以选择删除一个或多个数据源。

要从存储库中删除选定的数据源:

1. 打开 Entity Analytics 节点。
2. 单击**实体存储库**列表。
3. 单击**<浏览...>**，即显示“实体解析实例”对话框。
4. 在“实体解析实例”对话框上，单击**存储库名称**列表。
5. 选择想要删除的数据源所属的存储库。
6. 如果尚未连接，请输入管理员用户名和密码，然后单击**连接**。
7. 在**管理存储库**列表中，选择要删除的数据源。如果需要，请先按下 Ctrl，然后单击选择其他要删除的数据源。
8. 在启用了**删除未使用的**按钮后，请单击此按钮。
9. 在“删除未使用数据源”对话框中，单击**删除**以确认清除存储库。

删除实体存储库

您完全不再需要存储库时，可以完全删除它。

注意: 这会完全按照所指示的操作执行。**您无法撤销此操作。**如果您不确定，请使用**清除**按钮，以删除所有源数据。这样做不会删除存储库配置。请参阅主题第 27 页的『清除实体存储库』，了解更多信息。

注: 以下过程假定您可以从 SPSS Modeler 连接到存储库，并且知道托管存储库的数据库的管理员用户名和密码。如果情况并非如此，请在无法连接到存储库时，按照用于删除存储库的过程执行操作。请参阅主题『在不能与存储库连接时，将存储库删除』，了解更多信息。

删除存储库

1. 打开 Entity Analytics 节点。
2. 单击**实体存储库**列表。
3. 单击 **<浏览...>** 以显示“实体解析实例”对话框。
4. 在“实体解析实例”对话框上，单击**存储库名称**列表。
5. 选择要删除的存储库。
6. 如果您尚未进行连接，请输入管理员用户名和密码，并单击**连接**。
7. 在启用**删除整个存储库**按钮时，请单击此按钮。
8. 单击**删除**，以确认删除存储库。
9. 单击**确定**，以确认成功删除。

在不能与存储库连接时，将存储库删除

由于 SPSS Modeler 的连通性问题或者因为您忘记了用户名或密码，在您要删除实体存储库但不能与其连接时，可执行以下步骤。

请在托管存储库数据库的机器上,，执行此步骤。

Windows 系统

1. 打开命令提示符窗口。
2. 输入:

   ```
   cd modeler_install_dir
   cd ext\bin\pasw.entityanalytics\tools
   delete_repository.bat repos_name
   ```

   modeler_install_dir 是安装 SPSS Modeler 的目录，*repos_name* 是存储库的名称。
3. 在本节稍后将继续“完成步骤”。

UNIX 系统

1. 打开 Shell。
2. 输入:

   ```
   cd modeler_server_install_dir
   cd ext/bin/pasw.entityanalytics\tools
   ./delete_repository.sh repos_name
   ```

   modeler_server_install_dir 是安装 SPSS 建模器服务器 的目录，*repos_name* 是存储库的名称。

完成步骤（所有系统）

1. 在提示符处，输入 Y 确认删除存储库。
2. 当存储库被删除时，您可看到消息:

   信息 - 请从以下目录中除去存储库文件:
   directory_path
   （请注意，可能需要先重新引导，然后才能除去存储库文件）
3. 删除与您已删除的存储库具有相同名称的目录。如果您无法删除目录，请重启机器然后再试一次。

第 4 章 实体分析实例

关于本示例

在本例中，我们将了解添加实体分析如何进一步改善使用 IBM SPSS Modeler 获得的已令人印象深刻的结果。

本示例使用流 *loan_entity_analytics.str*，此流引用数据文件 *loan_applications.csv*。这些文件可在任何也安装了 IBM SPSS Modeler 的 IBM SPSS Modeler Entity Analytics 安装程序的 *Demos* 目录中找到。可从 Windows“开始”菜单的 IBM SPSS Modeler 程序组中访问 *Demos* 目录。*loan_entity_analytics.str* 文件在 *Entity_Analytics* 目录中。

注：必须先在系统上创建存储库，然后才能运行此示例流。请先创建存储库，然后再继续此示例。请参阅主题 第 10 页的『创建存储库』，了解更多信息。

让我们从一个熟悉的情景开始 - 一家银行的高管们担心客户是否可能拖欠一项审查中的贷款申请。这家银行是 SPSS Modeler 的长期用户，所以其员工已根据有关该银行过去所发放的 700 项贷款的现有数据创建了一个流，并构建了一个预测模型。要么这些贷款已经偿还，要么客户没有按期还款。

原始模型

下面说明了银行员工如何构建其模型以及他们从该模型中了解到了哪些信息。

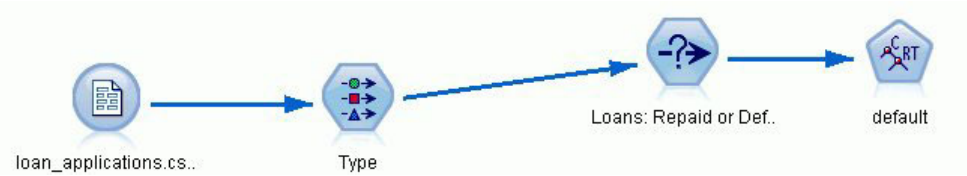

图 2. 包含建模节点的初始流

loan_applications.csv 数据集包含贷款申请仍在审查中的 150 个客户的详细信息以及过去贷款的详细信息，共计 850 条记录。

做出预测时并不会用到数据集中的所有字段，例如，可以忽略名称字段。“类型”节点可将忽略字段的角色设置为**无**将其过滤出来。将预测会使用到的字段角色设置为**输入**，然后将模型尝试预测其值的字段角色设置为**目标**。

图 3. "类型"节点中设置的字段角色

由于模型必须仅基于过去的数据进行预测，流所包含的"选择"节点仅包含那些没有标记为"待定"的贷款，因此将丢弃这 150 个审查中的贷款。

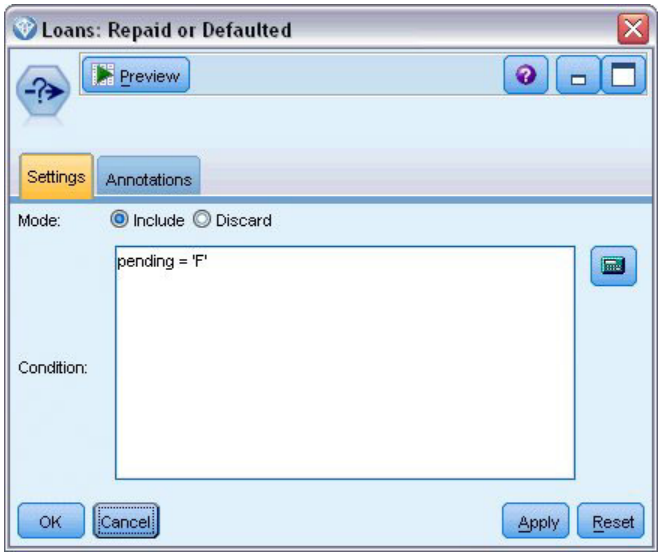

图 4. 丢弃审查中的贷款申请

丢弃了审查中的贷款后，只有剩余已经偿还或还在拖欠的 700 项贷款的详细信息传递到了建模节点。银行可以使用许多 SPSS Modeler 算法中的一种来生成适合的模型。在本例中，他们使用了"C&R 树"节点，此节点将用

于建立一个根据银行客户的过去表现预测可能违约者的模型。

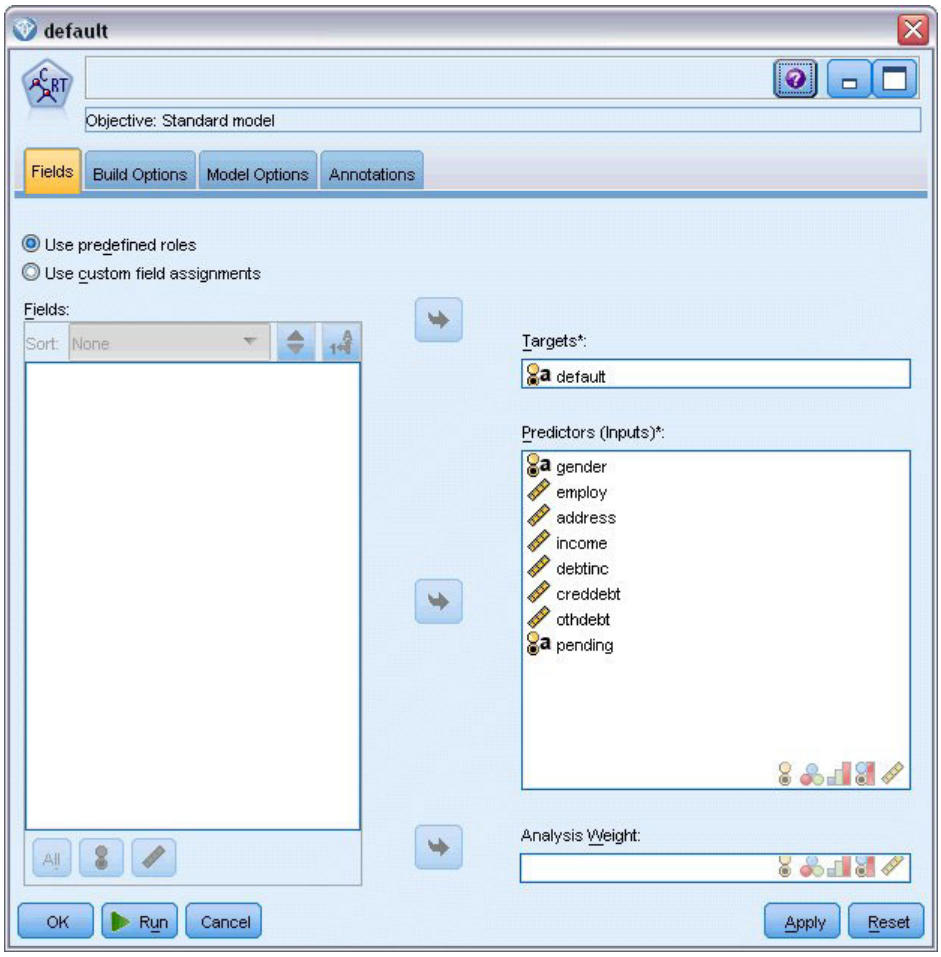

图 5. 指定预测变量和目标字段

将用于预测的字段指定为预测变量字段，并将模型尝试预测其值的字段（在本例中为 **default**）设置为目标字段，如前面“类型”节点所指定的那样。

运行此流会生成一个模型块，其中包含已从预测变量字段构建的模型。

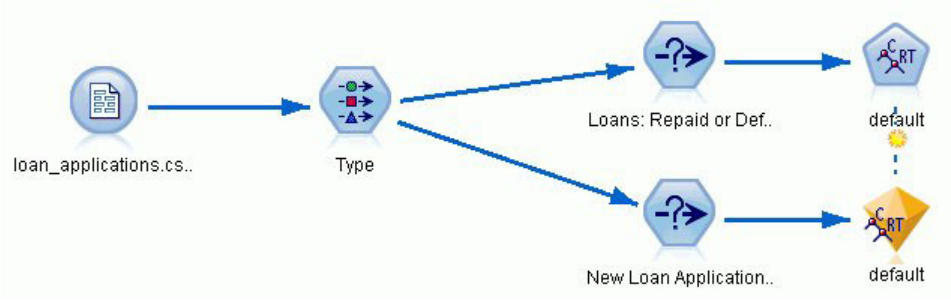

图 6. 添加了模型块的流

现在，银行分析人员可以使用此模型开始预测有应还款项的客户是否可能会拖欠还款。通过使用原始数据集，分析人员插入一个“选择”节点，这一次该节点只包含标记为“待定”的 150 条贷款记录，而不是丢弃这些记录。

分析人员直接将这些记录传递到模型，并添加一个分布节点，以直观表示模型的预测。

图 7. 用于选择新贷款申请并添加了“分布”节点的流

分布节点可显示模型中 *$R-default* 字段的值的分布。此字段由“C&R 树”节点在运行时添加到数据模型。字段包含对每个新申请人会偿还或是拖欠借款的预测，稍后也会使用此字段来比较添加实体分析的效果。

运行流的此部分时，分析人员可以从“分布”节点的输出中了解到，在 150 个新申请人中，有 137 人有望偿还贷款。预测剩余的 13 人会拖欠借款，所以分析人员很可能建议银行拒绝他们的申请。

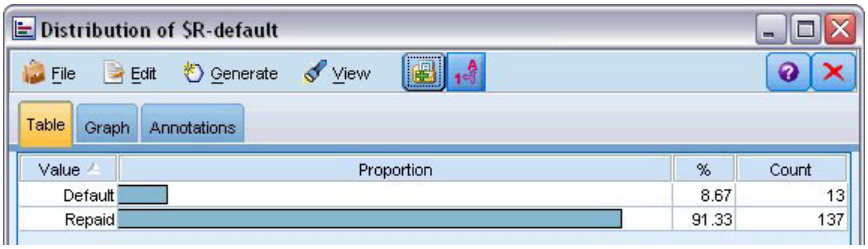

图 8. 没有实体分析的“分布”节点的输出

添加实体分析

现在让我们看一下，在方程式中添加实体分析是否会使情况有所改观。假设您是一位实体分析专家，受银行邀请来调查源数据的客户记录中可能存在的欺诈性条目。可能有因数据录入错误而导致的重复记录，但也可能是贷款申请人试图掩饰自己的身份。在任何一种情况下，银行都需要知道真实情况。

对于本示例，我们假设已经创建了实体存储库。请参阅主题第 10 页的『创建存储库』，了解更多信息。

将源数据转入存储库中

首先，您需要将“EA 导出”节点添加到数据源节点，以便可以将源数据导出到实体存储库中。

在导出数据之前，需要将数据源中的字段映射到实体存储库中的特征。这是必需的，因为不同的数据源可以对同一类型的信息使用不同的字段名称。实体存储库提供了一组标准信息类型（称为“特征”）来避免重复。

在“EA 导出”节点中，设置有关存储库的详细信息：连接详细信息、源标记（用于标识数据源，在本例中为 **TEST**）、实体类型（我们使用的特征集，名为 **PERSON**）和唯一键字段（用于唯一标识每个记录）。在本例中，使用**键**字段作为唯一键。

现在可以设置映射。在使用的特征集中，存在与下列字段相对应的特征： *fname*、*mname*、*lname*、*generation*、*dob*、*gender*、*addr1*、*city*、*country*、*postcode*、*phone*、*email*、*ssn*、*drlic* 和 *passport*。

首先设置 *fname* 的映射。在表中的 *fname* 行上双击**映射到特征**列，向下滚动到 **NAME.GIVEN_NAME** 条目，然后单击它以创建映射。

现在映射具有相应特征的剩余字段，以使整个映射集类似如下所示。

表 12. 映射到存储库特征的字段.

字段	映射到特征
fname	NAME.GIVEN_NAME
mname	NAME.MIDDLE_NAME
lname	NAME.SUR_NAME
generation	NAME.NAME_GEN
dob	DOB.DOB
gender	GENDER.GENDER
addr1	ADDRESS.ADDR1
city	ADDRESS.CITY
country	ADDRESS.COUNTRY
postcode	ADDRESS.POSTAL_CODE
phone	PHONE.PHONE_NUM
email	EMAIL_ADDR.ADDR
ssn	SSN.ID_NUM
drlic	DRLIC.ID_NUM
passport	PASSPORT.ID_NUM

单击**运行**将数据导出到存储库中。此过程需要少许时间，当“执行反馈”对话框关闭时，导出即完成。

读取已解析的身份

将数据导出到存储库时，实体分析系统便开始解析可能的身份冲突，并分配唯一的实体标识（即，您稍后将看到的 *$EA-ID* 字段）。（*注*: 此字段与“EA 导出”节点中的“唯一键”字段不同，后者仅用于唯一标识数据源记录。）

读取已解析身份的第一步是将 Entity Analytics (EA) 源节点添加到流中。在此阶段，不应将此源节点与任何内容相连。

打开 Entity Analytics(EA) 源节点，并设置“实体存储库”详细信息。随后会显示已导出到存储库的数据源列表，在本例中只有一个数据源。

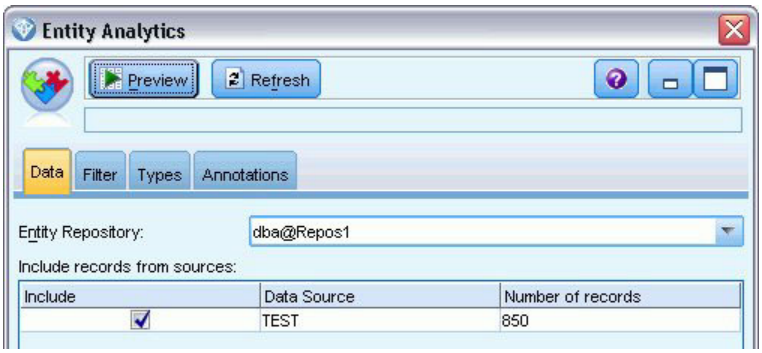

图 9. 在存储库中选择一个数据源

选中 **TEST** 数据源的复选框并单击“确定”。

让我们看看实体分析系统对数据执行了哪些操作。将“表”节点附加到 Entity Analytics (EA) 源节点，打开该“表”节点，然后单击**运行**以显示“表”节点输出窗口。

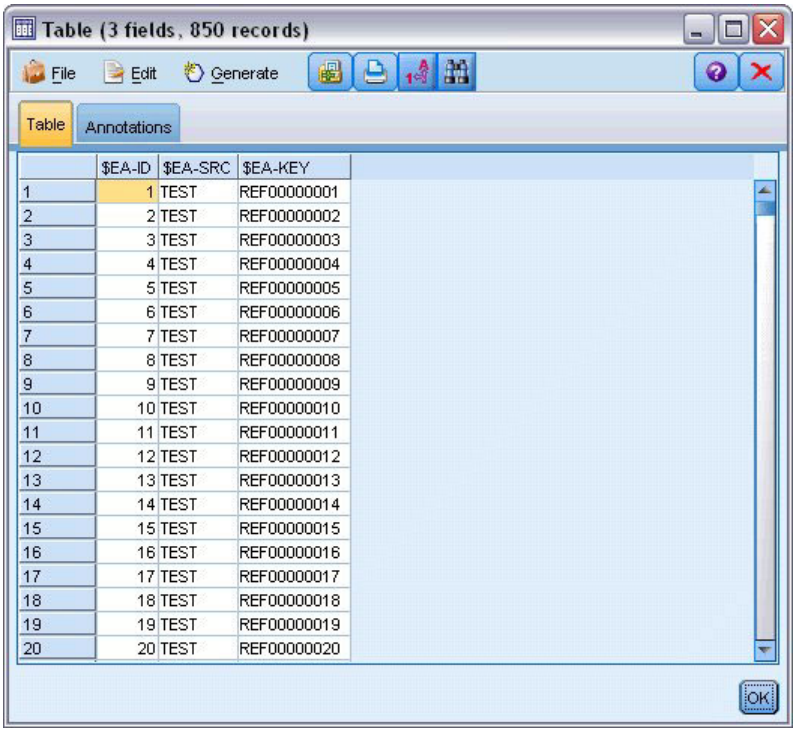

图 10. “表”节点的输出

只有一个字段看起来很熟悉，即，标有 *$EA-KEY* 的字段。它实际上是源数据中的*键*字段，之所以出现在这里是因为您在“EA 导出”导出节点中将其选为“唯一键”字段。

但系统还添加了另外两个字段。*$EA-ID* 字段是唯一的标识，不是源记录的标识，而是已解析身份的标识。稍后我们将看到具体的差异。*$EA-SRC* 字段标识数据的来源，此处它指示 **TEST**，因为这是您在“EA 导出”节点中为其分配的源标记。

源数据中的所有其他字段是如何处理的？别担心，它们还在存储库中，只是出于性能原因，Entity Analytics (EA) 源节点仅仅向下游传递了最少的一组字段以进行进一步处理。

现在，将“表”节点输出向下滚动到 385 行。

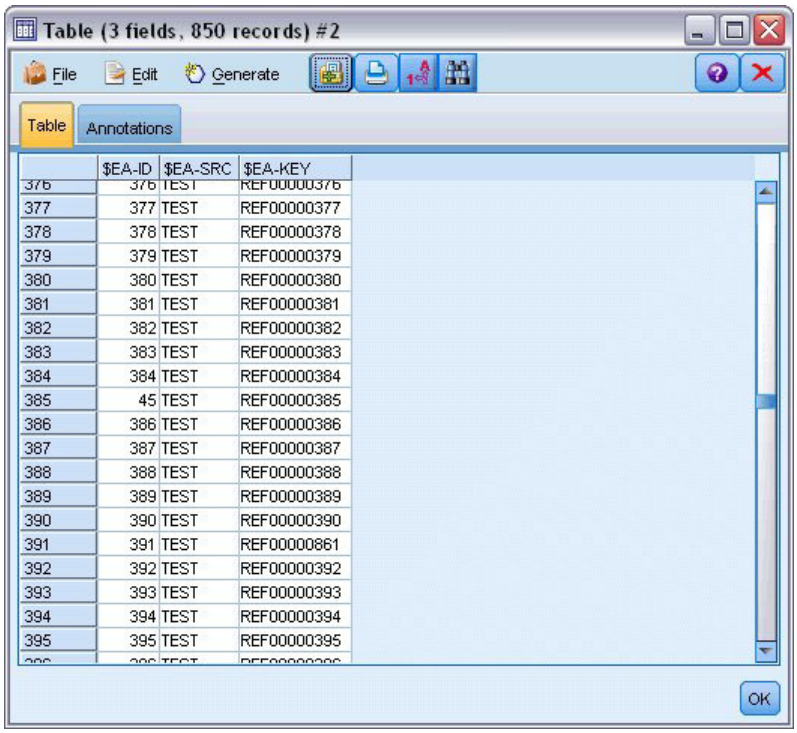

图 11. 表输出行与 $EA-ID 号码之间的差异

注意 *$EA-ID* 号码如何在此处显示为失序。实体分析系统已确定记录 REF00000385 引用标识为实体 45 的人员，该人员也拥有记录 REF00000045。继续向下滚动输出，会有更多无序号码，例如，在 485、517、520 等行。我们来仔细看看。

首先，要强调一个事实，就是通过将数据重命名为*键*，*$EA-KEY* 字段包含来自源数据中*键*字段的数据。将“过滤”节点附加到 Entity Analytics (EA) 源节点并打开该“过滤”节点。双击第二个**字段**列中的字符串 **$EA-KEY** 并输入 key。

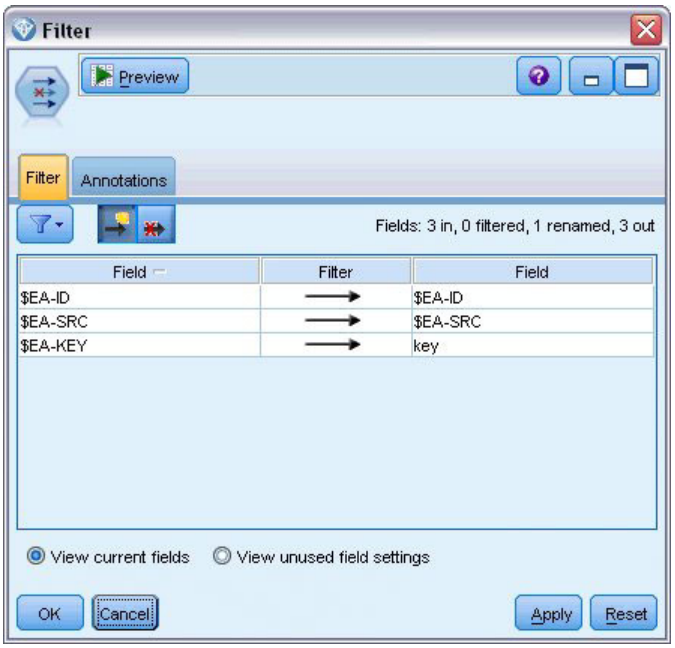

图 12. 重命名 $EA-KEY 字段

单击**确定**关闭“过滤”节点。

现在需要对 *$EA-ID* 实体 ID 进行升序排列。将“排序”节点附加到“过滤”节点。打开该“排序”节点，单击**排序方式**表旁的顶部按钮，选择 **$EA-ID** 并单击**确定**。

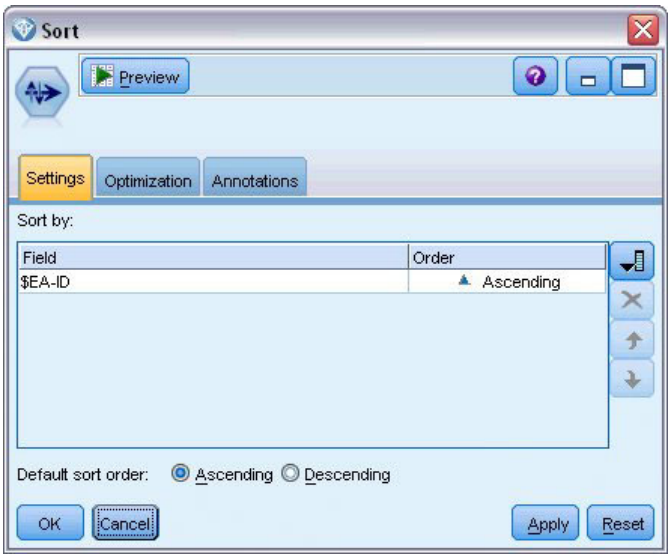

图 13. 按升序排列实体标识

将排序顺序保持为**升序**，并单击**确定**。

现在，您需要创建一个额外字段，用于指示记录是唯一还是重复的记录。将“派生”节点附加到“排序”节点。打开该“派生”节点，并将**派生字段**名称设置为 IsDuplicate。从**导出为**列表中选择**标志**，这也将**字段类型**设置为**标志**。将**真值**字段设置为**重复**并将**假值**字段设置为**唯一**。

要查找重复记录，需要使用 SPSS Modeler 随附的特殊序列函数 @OFFSET。

在 **If** 字段中键入以下内容:

```
'$EA-ID' = @OFFSET('$EA-ID',1) or '$EA-ID' = @OFFSET('$EA-ID',-1))
```

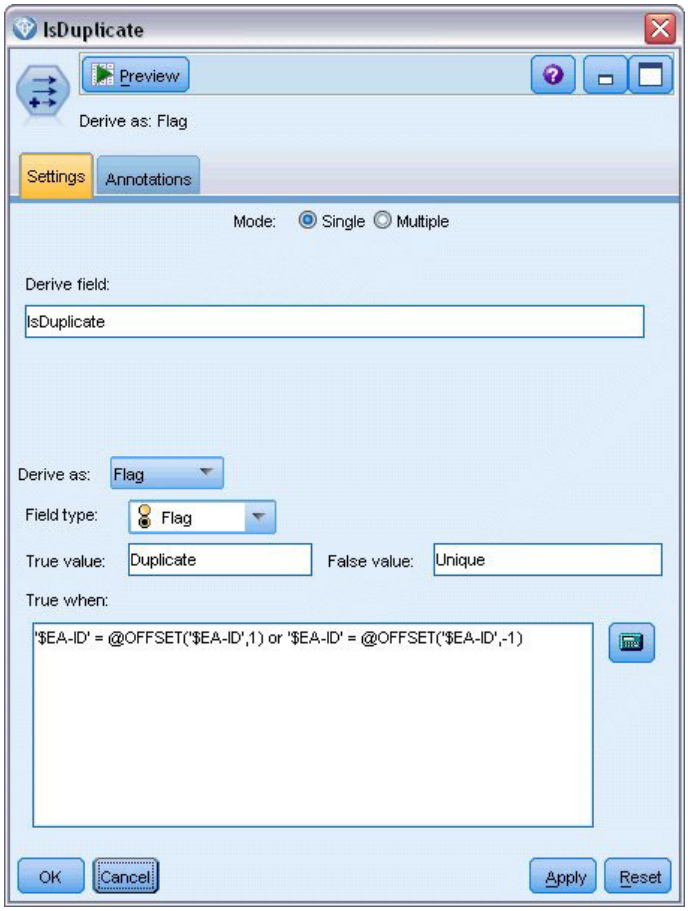

图 14. 在“派生”节点中设置条件

按升序排列实体标识后，OFFSET 函数将检查相邻的实体标识是否相同（实体标识相同时记录就重复）。如果是，其 *IsDuplicate* 值将设置为*重复*，否则将设置为*唯一*。

单击**确定**关闭此节点。

要查看“导出”节点的效果，请将“表”节点附加到“导出”节点，打开该“表”节点，然后单击**运行**。将“表”节点输出窗口向下滚动到 45 行。

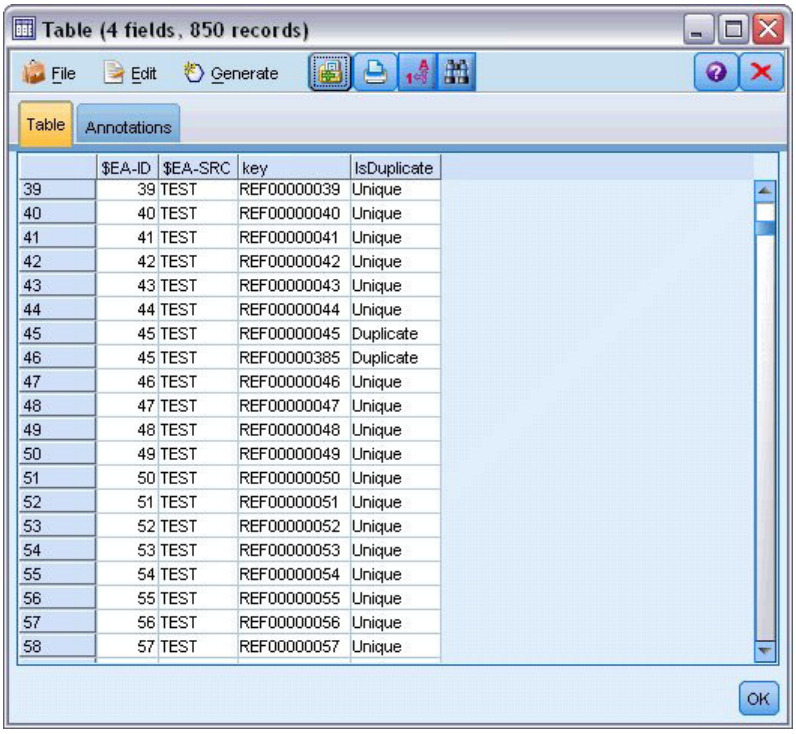

图 15. "派生"节点的输出

还记得当时我们直接从 Entity Analytics(EA) 源节点查看输出吗？系统已经确定记录 REF00000385 与实体 45 指的是同一个人。现在我们已经更进一步，发现记录 REF00000045 和 REF00000385 是重复项，因为它们都是指实体 45。

继续将输出窗口向下滚动，您会看到其他标记为重复项的记录。

要获取列出重复记录的报告，请将"报告"节点（从节点选项板的"输出"选项卡）附加到 *IsDuplicate* 导出节点。打开该"报告"节点，将以下文本复制在"模板"选项卡的输入字段中，然后单击**运行**。

```
<html>
<h1>List of duplicate customer records.

<h2>This report was generated: [@TODAY]

<h2>Duplicate records
<table>
 <tr>
   <td>Entity ID</td>
   <td>Key</td>
 </tr>

#WHERE IsDuplicate = "Duplicate"

 <tr>
   <td>['$EA-ID']</td>
   <td>[key]</td>
 <tr>
#
</table>

</html>
```

其输出如下所示。

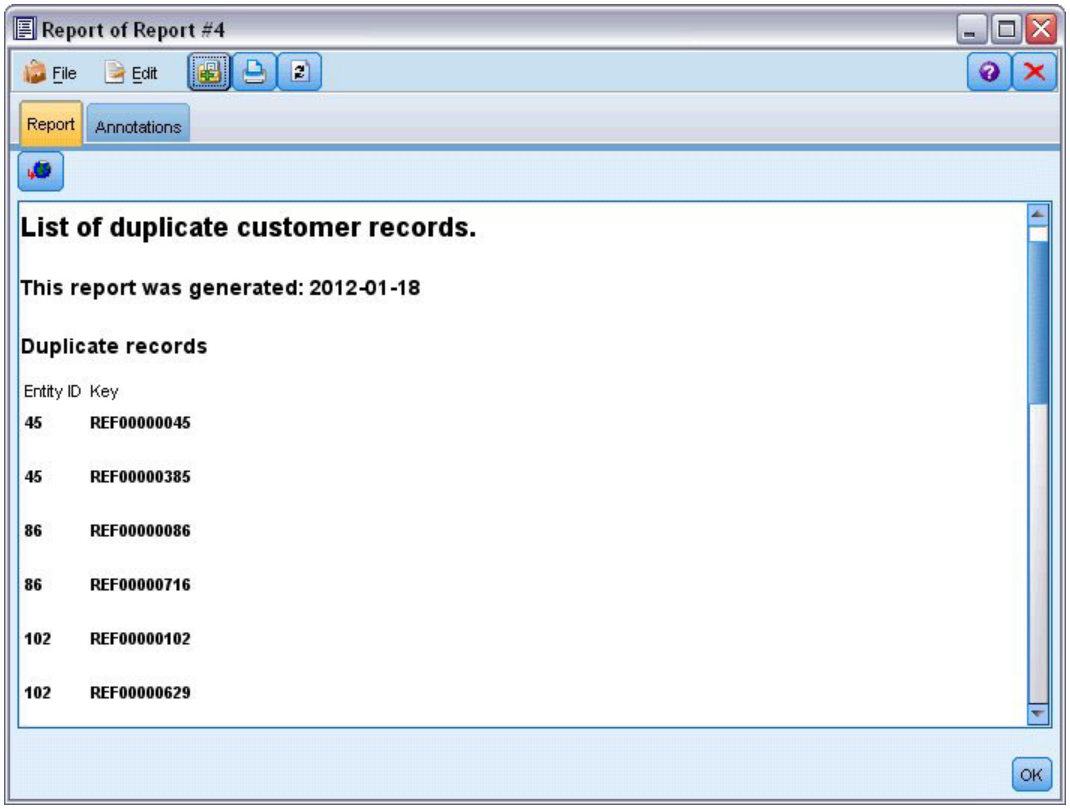

图 16. “报告”节点的输出

在本例中，报告使用 HTML 格式，但您也可以使用 XML 或 ASCII 格式。

比较实体分析输出与原始模型

本例的最后阶段是查看添加实体分析是否会给银行的原始预测带来任何改变。您可能记得，原始模型预测了 150 个待定申请中有 13 个违约者。您将使用“合并”节点将该模型的输出与来自实体分析的关于重复记录的信息进行合并，以了解这样做是否会改变预测。

首先，需要确认由实体分析添加的新字段具有正确的数据类型，或其在 SPSS Modeler 中的*测量级别*。将“类型”节点附加到“**IsDuplicate** 导出”节点，打开“类型”节点并单击**读取值**按钮。

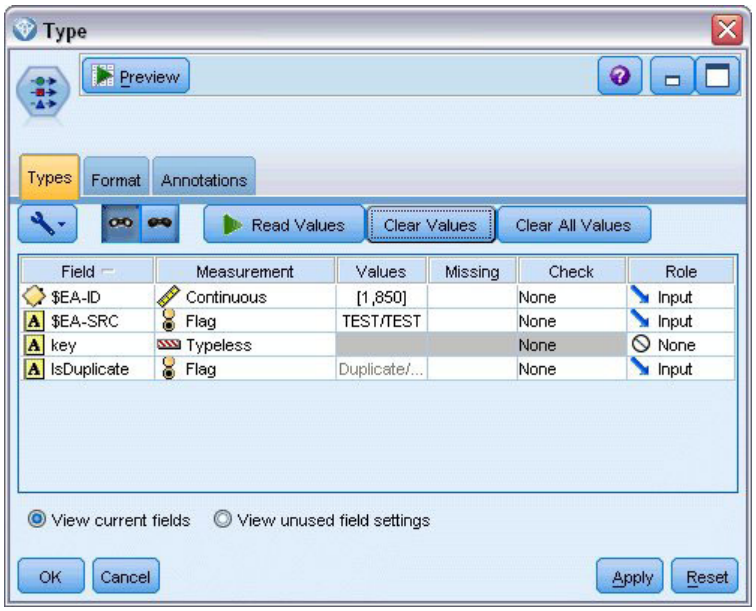

图 17. “类型”节点的设置

现在可以添加“合并”节点。将其附加到“类型”节点，并将其连接到包含原始模型的金色块。要执行此操作，请右键单击金色块，选择**连接**，然后单击“合并”节点，现在该节点应有两个输入箭头。

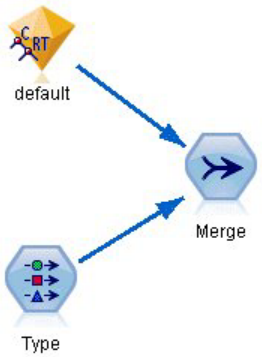

图 18. “合并”节点的输入

打开“合并”节点，将**合并方法**设置为**键**，然后单击右箭头按钮将**键**字段从**可能键**移动到**用于合并的键**，然后单击**确定**。

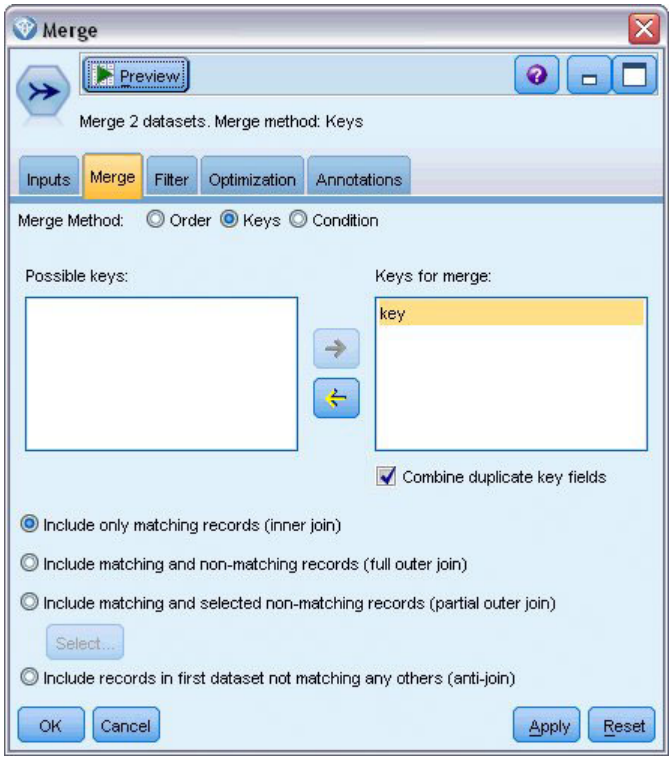

图 19. 指定合并操作的键字段

现在，差不多准备好进行比较了。但是，如果您此时要附加并运行“分布”节点，将不会看到与原始预测有任何变化。尽管现在流已将原始模型块的输出与实体分析创建的新字段合并，但数据模型中的预测字段本身（*$R-default*）尚未进行新信息更新。

为此，您将使用“填充”节点，此节点可以替换字段值。将“填充”节点附加到“合并”节点并打开该“填充”节点。

单击**填写字段**右边的顶部按钮，滚动到列表的底部，选择 **$R-default** 并单击**确定**。如果满足在此对话框的其余部分中指定的条件，那么将更改此字段的值。

要指定条件，请确保**替换**设置为**根据以下条件**，然后在**条件**字段中输入以下内容:

```
default /="default" and IsDuplicate  = "Duplicate"
```

在**替换为**字段中输入以下内容:

```
"default"
```

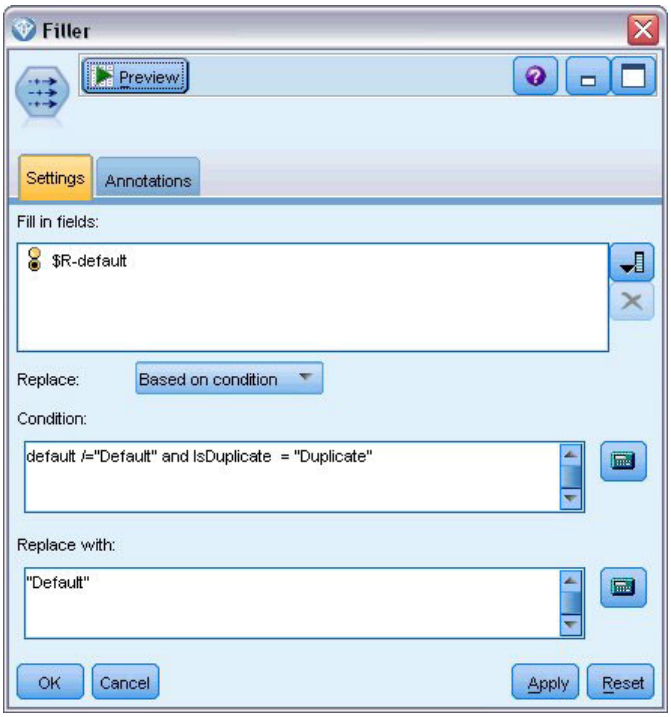

图 20. 指定替换字段值的条件

这些设置需要一些说明。条件的含义为，对于每个记录，如果原始数据集中的默认字段值不等于**默认**，并且该记录已被标记为重复，则将模型中 *$R-default* 字段的值设置为**默认**。

$R-default 字段是模型中的字段，包含关于客户是否可能会拖欠贷款的预测。这样，具有重复记录的客户将作为潜在违约者添加到模型中。

单击**确定**关闭“填充”节点。

终于可以查看实体分析所带来的改变了。从“图形”选项板中，将“分布”节点附加到“填充”节点并打开该“分布”节点。单击**字段**列表并选择 **$R-default**。

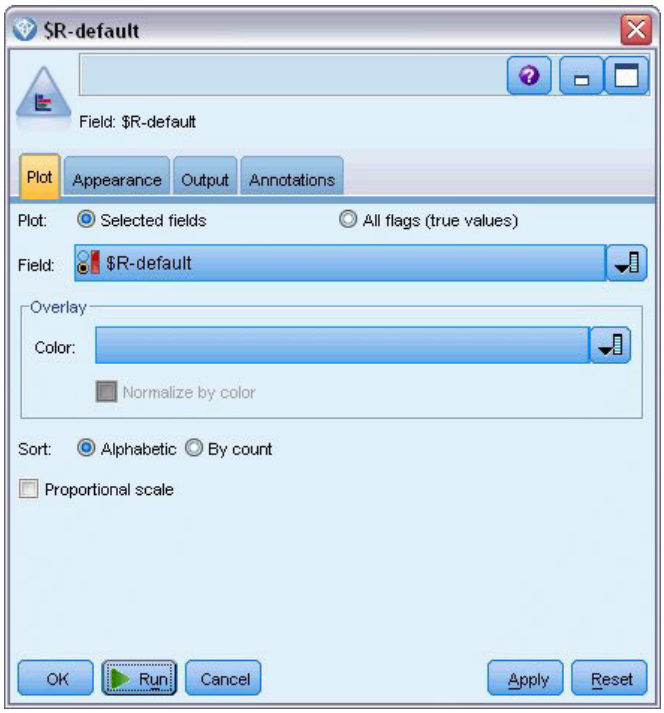

图 21. "分布"节点的设置

单击**运行**以生成新预测的图表。

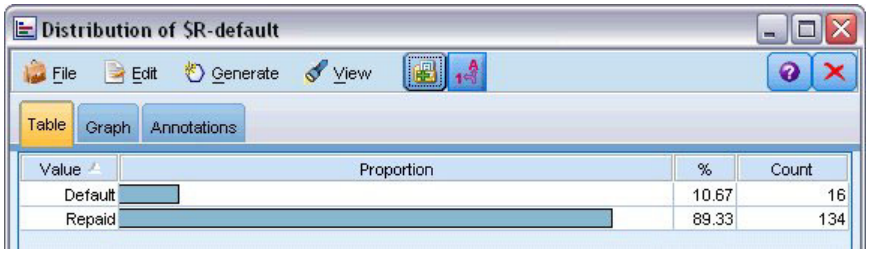

图 22. 经过实体分析后"分布"节点的输出

现在有 16 个存在风险的申请，而不是 13 个了。如果额外的申请确实拖欠了还款，损失可能会非常惨重，所以您可以图形方式向银行展示在其风险评估操作中添加实体分析的好处。

摘要

本例已经展示，使用实体分析如何可以消除人员或组织数据中的重复记录，从而提高预测质量。

注：理想情况下，在执行任何其他处理之前应该消除重复的记录。然后，使用"自动数据准备"(ADP) 节点分析数据并标识修订、筛选出有问题或可能无用的字段、适当时派生新属性并通过智能筛选技术提高性能。

合并实体分析和自动化数据准备有助于确保您正在处理的是尽可能干净的数据。

附录. IBM SPSS Modeler Entity Analytics 的脚本编制属性

使用 IBM SPSS Modeler Entity Analytics 进行脚本编制

IBM SPSS Modeler Entity Analytics 中的脚本编制是用于在用户界面上实现过程自动化的强大工具。您使用鼠标或键盘进行的操作，借助脚本同样可以完成，而且使用脚本可以自动执行那些手动执行将造成大量重复操作或高耗时的任务。有关使用脚本的说明，请参阅 IBM SPSS Modeler 提供的 *ScriptingAutomation.pdf* 指南。

公共属性

下表列出了 IBM SPSS Modeler Entity Analytics 节点的公共属性。关于特定节点的信息将在后面的章节中介绍。

表 13. 公共属性

属性名称	数据类型	属性说明
entity_repository	['*field*','*field*', ... ,'*field*']	这是存储库连接字符串。格式: ['*reposname*', '*username*', '*password*'] 示例: entity_repository = ['repos1', 'dba', 'psw1']
entity_type	*字符串*	这是要使用的实体类型（特征集）。 示例: entity_type = 'PERSON'

entityanalytics_exportnode 属性

“EA 导出”节点是一个终端节点，它从数据源中读取实体数据，然后将该数据导出到存储库以进行实体解析。

表 14. entityanalytics_exportnode 属性

entityanalytics_exportnode 属性	数据类型	属性说明
mode	Add PurgeFirst	这是导出方式。Add 将源文件记录添加到存储库的现有内容中；PurgeFirst 在导出之前除去现有内容。
source_tag	*字符串*	这是数据源标识。 示例: source_tag = 'CUST'
unique_key_field	*字符串*	这是要用作数据记录的唯一标识的输入字段。 示例: unique_key_field = 'ID'

表 14. *entityanalytics_exportnode 属性 (续)*

entityanalytics_exportnode 属性	**数据类型**	**属性说明**
field_mapping	[[*'field_name'* *'feature.element'* *'usage_type'*]...]	用于将输入字段映射到存储库中的相应特征。 示例: field_mapping = [['fname' 'NAME.GIVEN_NAME' ''] ['addr1' 'ADDRESS.ADDR1' 'PRIMARY']] 注: 要将 *usage_type* 设置为"(Auto)"的等价项，请在前面的第一个示例中使用 ''。

entityanalytics_sourcenode 属性

Entity Analytics (EA) 源节点从存储库中读取已解析的实体，然后将此数据传递到流以进行进一步处理，例如格式化为报告。

表 15. *entityanalytics_sourcenode 属性*

entityanalytics_sourcenode 属性	**数据类型**	**属性说明**
source_tags	*list*	将要从存储库中输出的数据源的标记列表。 示例: source_tags=['LOANS', 'CUSTOMERS']
relationships	None Close All	用于检索存储库中的关系详细信息的匹配条件。 None 不返回任何关系。 Close 返回根据详细信息（比如分离度）而确定的密切匹配。 All 返回所有可能的关系。
max_degree_separation	*integer*	最小值为 0，最大值为 3。
output_entity_type	*字符串*	存储库中使用的实体类型列表。

entityanalytics_processnode 属性

"流 EA"节点将新案例与存储库中的实体数据进行比较。

表 16. *entityanalytics_processnode 属性*

entityanalytics_processnode 属性	**数据类型**	**属性说明**
match	Exact ByIdentifier All	这是用于从存储库中检索实体的匹配条件。Exact 仅返回完全匹配项。 ByIdentifier 返回完全匹配项以及共享相同标识的实体。 All 返回所有可能的匹配项。
save_search_records	*boolean*	

表 16. entityanalytics_processnode 属性 (续)

entityanalytics_processnode 属性	**数据类型**	**属性说明**
relationships	None Close All	用于检索存储库中的关系详细信息的匹配条件。 None 不返回任何关系。 Close 返回根据详细信息（比如分离度）而确定的密切匹配。 All 返回所有可能的关系。
max_degree_separation	*integer*	最小值为 0，最大值为 3。
output_entity_type	*字符串*	存储库中使用的实体类型列表。

声明

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您所在区域当前可获得的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。但是，评估和验证任何非 IBM 产品、程序或服务的操作，由用户自行负责。

IBM 可能已拥有或正在申请与本文档内容有关的各项专利。提供本文档并不意味着授予用户使用这些专利的任何许可。您可以用书面形式将许可查询寄往:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

以下段落对于英国和与当地法律有不同规定的其他国家或地区均不适用: INTERNATIONAL BUSINESS MACHINES CORPORATION“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某特定用途的保证。某些国家或地区在某些交易中不允许免除明示或暗含的保证。 因此本条款可能不适用于您。

本信息可能包含技术方面不够准确的地方或印刷错误。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。IBM 可以随时对本出版物中描述的产品和/或程序进行改进和/或更改，而不另行通知。

在本信息材料中对任何非 IBM 网站的引用仅为了方便用户，并不以任何方式表明对这些网站的认可。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的: (i) 使其能够在独立创建的程序和其它程序（包括本程序）之间进行信息交换，以及 (ii) 使其能够对已经交换的信息进行相互使用，请与下列地址联系:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

本文档中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际程序许可协议或任何同等协议中的条款提供。

此处所含的性能数据均在受控环境下决定。因此，在其他操作环境中获得的结果可能差异较大。有些测量可能在开发级的系统中进行，不保证这些测量结果与常用系统上的测量结果相同。此外，有些测量结果可能通过推断来估计得出。实际结果可能有所差异。此文档的用户应针对其具体环境验证适用的数据。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的 资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问 题应当向这些产品的供应商提出。

有关 IBM 未来方向或意向的所有声明均可能未经通知即变更或撤销，并且仅代表目标和目的。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

如果您正在查阅此信息的软拷贝，照片和彩色插图可能不会显示。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。当前的 IBM 商标列表，可从 Web 站点 www.ibm.com/legal/copytrade.shtml 上『版权和商标信息』部分获取。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、 Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国和/或其他国家或地区的注册商标。

Microsoft、Windows、Windows NT 以及 Windows 徽标是 Microsoft Corporation 在美国和/或其他国家或地区的商标。

UNIX 是 The Open Group 在美国和/或其他国家或地区的注册商标。

Java 和所有基于 Java 的商标及徽标皆为 Oracle 和/或其附属公司的商标或注册商标。

其他产品和服务名称可能是 IBM 或其他公司的商标。

索引

Printed in China