IBM SPSS Modeler Entity Analytics 15 用户指南

注意：使用本信息及其支持的产品之前，请阅读注意事项第 60 页码下的一般信息。

此版本适用于　及所有后续发布和修订，除非在新版本中另有说明。

前言

IBM® SPSS® Modeler 是 IBM Corp. 企业级数据挖掘工作平台。SPSS Modeler通过深入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，这将生成更强有力的预测模型，缩减实现解决方案所需的时间。 SPSS Modeler 提供了多种建模技术，例如预测、分类、细分和关联检测算法。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 软件为决策者提供可信赖的完整、一致和准确信息，以帮助其提升业务绩效。这一涵盖商务智能、预测分析、财务绩效与战略管理以及分析应用程序的全面组合可提供有关当前业务表现的清晰、立即和切实可行的深入见解，并能够有效预测未来结果。其中整合了丰富的行业解决方案、经过验证的做法与专业服务，以帮助各种规模的组织提升生产效率、自动化决策并取得卓越成果。

作为该软件组合的一部分，IBM SPSS Predictive Analytics 软件能够帮助各类组织有效地预测未来事件，并针对所得到的深入见解提前采取行动，以取得更优秀的业务成果。全球企业、政府和学院客户依赖 IBM SPSS 技术作为吸引、留住和增加客户数量的竞争优势，并降低欺诈和转移风险。通过将 IBM SPSS 软件融入其日常运营中，这些组织将成为“预测型”企业，即能够指引并自动化决策，以实现业务目标和取得可衡量的竞争优势。有关详细信息，或联系我们的代表，请访问 http://www.ibm.com/spss。

技术支持

我们提供有技术支持服务以维护客户。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。要获得技术支持，请访问 IBM Corp. 网站 http://www.ibm.com/support。在请求帮助时，请做好准备，以便识别您自己、您的组织以及您的支持协议。

内容

章 1

实体分析

关于实体分析

IBM® SPSS® Modeler Entity Analytics 在 IBM® SPSS® Modeler 预测分析的基础上添加了全新的维度。预测分析会尝试根据过去数据预测未来行为，而实体分析侧重于通过解析记录自身的身份冲突，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中都非常重要，包括客户关系管理、检测、反洗钱以及国家与国际安全。

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

记录编码： 70001
姓名： Jon Smith
地址： 123 Main Street
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。使用自动实体分析系统，可极大地缩短解析时间。。

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 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符。

下表说明了两种分析类型之间的差别。

表 1-1
预测分析与实体分析之间的差异

特征	预测分析	实体分析
训练	基于相对较小的集合和数值范围	基于较大集合，例如，姓名、地址
用于训练的数据	子集（训练分区）	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

章 2

IBM SPSS Modeler 的实体分析

使用 IBM SPSS Modeler 的实体分析

您意识到自己的数据可能有身份问题。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要。

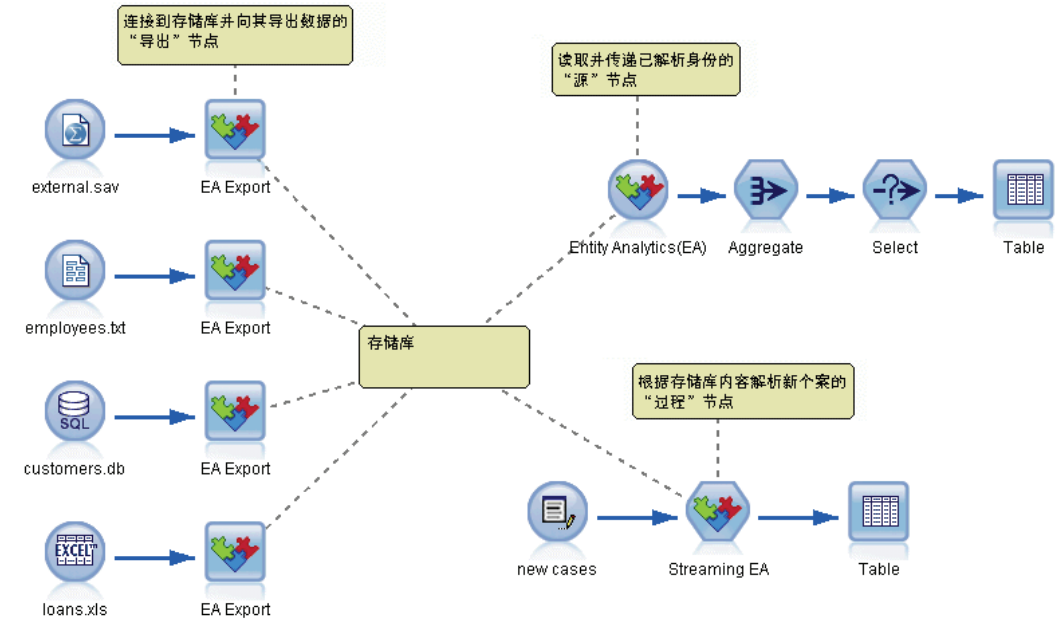
- 将源数据读入 IBM® SPSS® Modeler
- 创建可存储数据的存储库
- 将 SPSS Modeler 连接到存储库
- 将数据字段映射到存储库特征
- 将数据导出到存储库中，并解析身份
- 分析已解析的身份
- 解析存储库的新个案
- 生成所有所需的警告（批处理或实时）

此时，您需要了解 SPSS Modeler 的工作方式。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。每个节点代表工作流的特定阶段。

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。这些是 EA 导出节点，包括 Entity Analytics(EA) 节点以及流 EA 处理节点。

下图说明了此过程。

图片 2-1
实体分析过程



阶段 1：将源数据读入 SPSS Modeler

您的首要任务是通过源节点（在 SPSS Modeler 中以圆形图标表示）将您的数据读入 SPSS Modeler。

图片 2-2
IBM SPSS Modeler 源节点

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。

每个数据源文件必须具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。 有关详细信息，请参阅第 15 页码第 3 章中的添加唯一记录标识符。

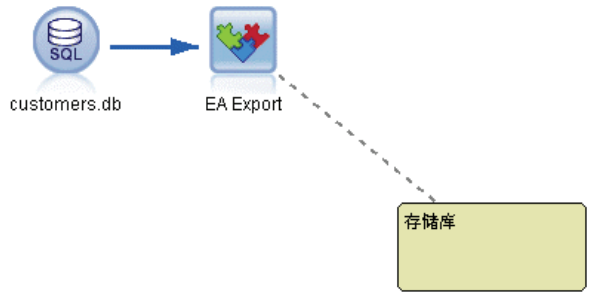
有关详细信息，请参阅第 15 页码第 3 章中的连接数据源。

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。

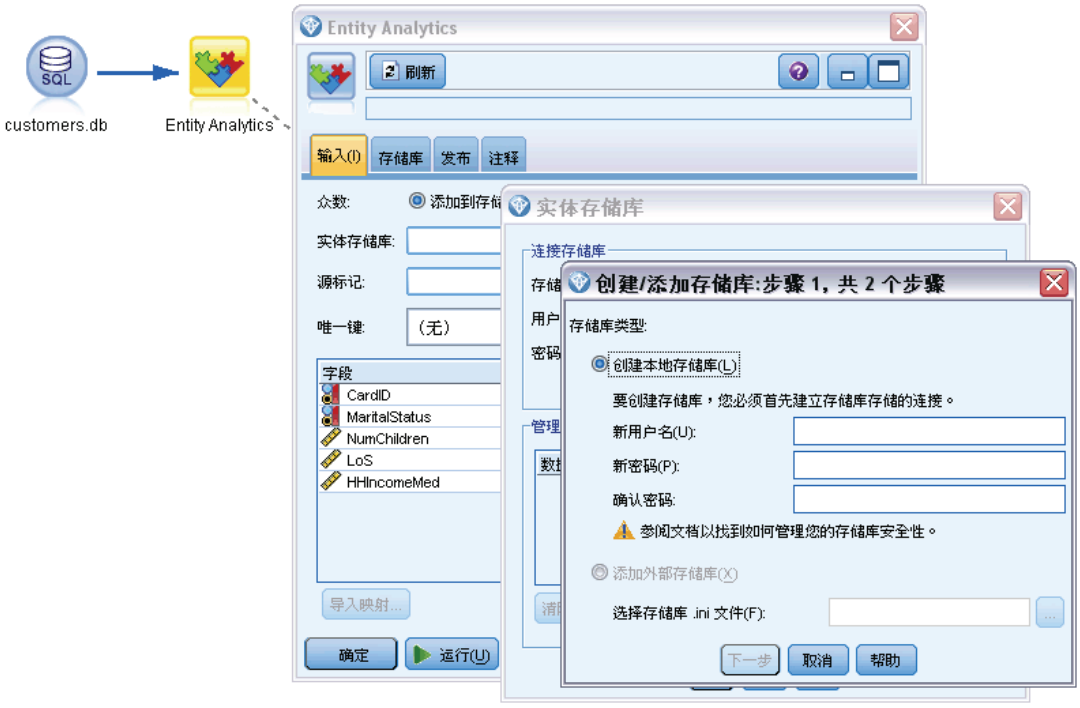
要创建存储库，首先将数据源连接到由方形图标表示的“EA 导出”节点。

图片 2-3
连接到存储库



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。

图片 2-4
创建存储库



稍后详细介绍创建存储库的过程。 有关详细信息，请参阅第 14 页码第 3 章中的设置一个实体存储库（“EA 导出”节点）。

设置完存储库后，您即可通过各种方法维护其内容。 有关详细信息，请参阅第 21 页码第 3 章中的配置实体存储库。

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。

图片 2-5
连接到存储库

有关详细信息，请参阅第 17 页码第 3 章中的实体存储库选项。

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，其他特征则专用于特定数据源。特征通常相当于数据记录中的某个字段，或数据库表中的某列。

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。这是为了避免在不同数据源对包含同种信息的字段使用不同名称时存储库中出现重复。“EA 导出”节点提供了映射表，您可以在其中创建映射。

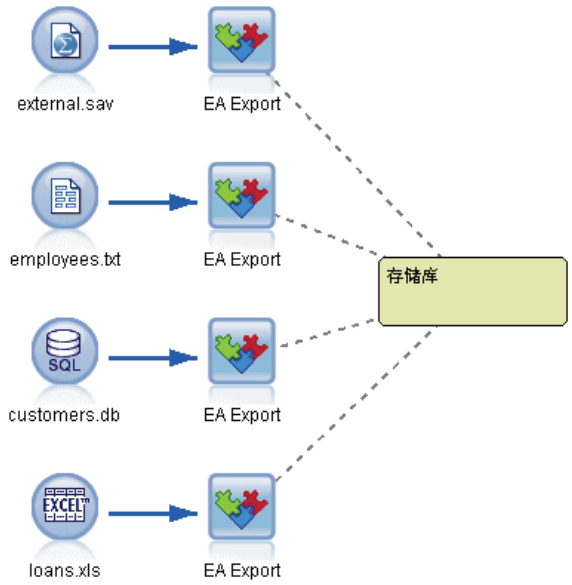
图片 2-6
将输入字段映射到特征

有关详细信息，请参阅第 18 页码第 3 章中的将输入字段映射到特征（“EA 导出”节点）。

阶段 5：将数据导出到存储库并解析冲突

每个数据源节点需要有自己的“EA 导出”节点，因此如果您的数据分散在许多不同的源中，您的流可能如下所示。

图片 2-7
将数据从多个数据源导入存储库



如果您有多个数据源，您可以选择从一个、部分或全部数据源读取记录。Entity Analytics 系统会分析您所选择的记录，并为每个记录添加一个名为 $EA_ID 的标识符字段。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，则添加到这些记录的标识符在整个存储库是唯一的标识符。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 导入存储库，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

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

员工

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

客户

姓名	地址	储蓄
Susan	1234 5th Street	$1234
Joe	777 Oak Street	$5

贷款

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

正如我们所见，您依次将每个数据源导入存储库。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符字段（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 2-1
导出阶段后存储库内容的示例

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

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 是同一人，并为其分配了 ID 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。现在我们找到了匹配项：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，并为他提供标识符 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。

注意：这是为了说明所举出的一个乐观匹配的例子。您可以选择一个更嚴格的规则集，如此一来，一条简单的姓名与电话或地址本身就無法构成精确匹配，并將同一个标识符分配給两条记录。

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，如果您怀疑相同身份存在重复的记录可能有欺诈活动，您可能希望生成一份列出重复项的报告。

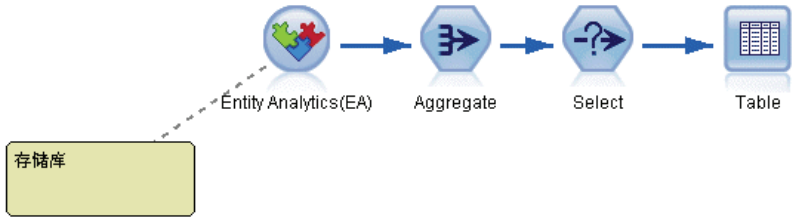
首先要创建一个 Entity Analytics(EA) 源节点，并将其链接到存储库。

源节点的输出包含以下字段。

- 由系统添加的标识符字段（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

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），则可以包括诸如“汇总”或“选择”节点的记录操作节点，如下图所示。

图片 2-8
查看输出



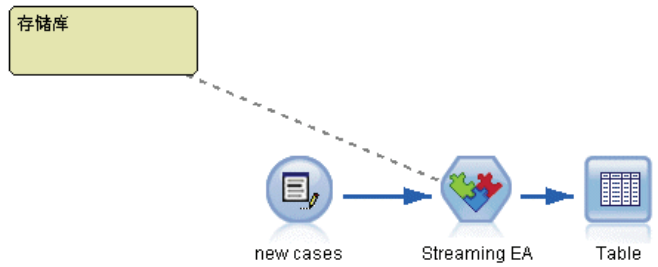
稍后详细介绍 Entity Analytics(EA) 源节点。 有关详细信息，请参阅第 29 页码第 3 章中的分析已解析身份（Entity Analytics(EA) 源节点）。

阶段 7：解析存储库的新个案

您已经解析所有数据源中的所有记录的身份。但您要使用批处理或实时模式，以查看它们是如何与已知的信息关联來取得更佳得分時，該如何处理？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，跟之前一样添加“表”节点，流的此部分现在如下所示。

图片 2-9
解析新个案



您运行此部分流时，“流 EA”节点会读取每条新记录，并将其与存储库内容比较。如果“流 EA”节点在存储库中找到匹配记录，则该节点会输出所有匹配记录和新记录，并向新记录添加 ID 字段和源指示符字段。如果未找到匹配项，则过程节点只输出添加了 ID 和源指示符字段的新记录。

为说明此过程，假定存储库当前包含 Entity Analytics(EA) 源节点输出的内容。 请参阅第 10 页码中的表 2-1。

现在我们收到以下新记录。他们与我们已知的人有关吗？

表 2-2
要评分的新记录

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

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符为 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符 5。

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

表 2-3
流 EA 节点的输出

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

此输出然后可由实体分析标示符汇总，并传递到其他下游节点进行进一步处理。

稍后详细介绍流 EA 节点。

阶段 8：生成警告

潜在的可疑活动可能会再次显现出来。在本例中，标识符为 4 的人已经有了一笔 $10,000 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，若是按照您的业务规则，此类行为被视为可疑行为的话，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 SPSS Modeler，您将有许多操作可以选择以满足您特定的需要。

章 3

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
- 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. 连接数据源。 有关详细信息，请参阅第 15 页码中的连接数据源。
2. 创建存储库。 有关详细信息，请参阅第 15 页码中的创建存储库。
3. 将数据源中的输入字段映射到存储库中的特征。 有关详细信息，请参阅第 18 页码中的将输入字段映射到特征（“EA 导出”节点）。

   设置完映射时，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。有关详细信息，请参阅第 20 页码中的显示字段映射（“EA 导出”节点）。

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。存储库会保持所有输入数据的当前状态，所以它可能变得很大。

您可以通过易于使用的图形界面维护存储库内容。 有关详细信息，请参阅第 21 页码中的配置实体存储库。

注意：IBM® SPSS® Modeler Premium 所提供的 IBM® SPSS® Modeler Entity Analytics 版本支持单一的存储库，这个存储库托管在 IBM SPSS Modeler Entity Analytics 所捆绑的 IBM solidDB 产品中。用此版本，您必须先删除一个已有的存储库才能创建一个新的。此 IBM SPSS Modeler 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工作区。
2. 在屏幕工作区，双击该图标以打开它的对话框。
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”节点添加字段。 有关详细信息，请参阅第 15 页码中的添加唯一记录标识符。

添加唯一记录标识符

每个输入实体存储库的数据源文件，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符

1. 在屏幕工作区，单击您在上一步所添加的源节点。
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. 在屏幕工作区，双击“派生”节点以打开它的对话框。
4. 在**派生**字段，为您添加的标识符字段，用有意义的名称替换默认名称（如 **ID**）。
5. 请确认将**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

您需要创建存储库，以存储所有输入数据。

创建存储库

1. 从 SPSS Modeler 节点选项板的“导出”选项卡，将“EA 导出”节点放在流工作区上。

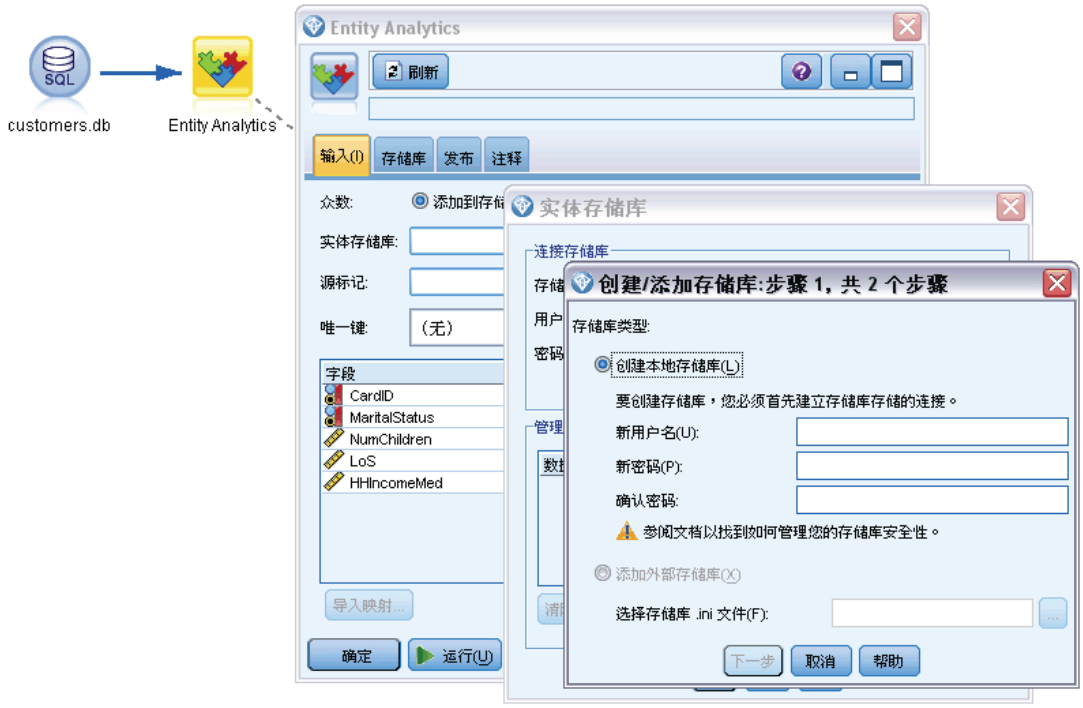
注意：如果您是第一次创建存储库，请使用“EA 导出”节点，并将其连接到含有您需要输入存储库的数据所在的 SPSS Modeler 源节点（或者，如果您已添加派生节点获取唯一标识符字段，可连接至“派生”节点）。要连接节点，请执行以下操作。

- 右键单击 SPSS Modeler 源节点。
- 选择“连接”。
- 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

图片 3-1
创建存储库向导



第 1 步

在这里，您选择是要使用绑定在 IBM SPSS Modeler Entity Analytics 中的 IBM solidDB 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。 为将要管理所建存储库的 IBM solidDB 数据库指定管理员用户名与密码详细信息。确认密码并单击**下一步**。

注意：如果随后您需要更改管理员凭证，可通过数据库的命令行编辑器进行。 有关详细信息，请参阅第 37 页码中的管理存储库数据库的管理员凭证。

添加外部存储库。 如果您想使用外部数据库来管理存储库，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。

导入配置的位置。 （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**。 有关详细信息，请参阅第 21 页码中的配置实体存储库。

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。

图片 3-2
实体存储库选项

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。**输入所选存储库的用户名。
- **密码。**该用户名的密码。
- **连接。**单击以连接到当前存储库。

管理实体存储库。此表列出了载入到当前存储库（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。**更新表中的数据源和大小信息，例如，您已经添加了新数据源或更改了现有数据源的大小。
- **清除存储库。**删除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，但您要删除存储库的所有数据记录，您可以使用此选项。 有关详细信息，请参阅第 39 页码中的清除实体存储库。

删除整个存储库。删除当前存储库的所有内容和配置详细信息。 有关详细信息，请参阅第 39 页码中的删除实体存储库。

配置存储库。显示您可以在其中配置当前存储库的窗口。 有关详细信息，请参阅第 21 页码中的配置实体存储库。

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（在存储库中未预定义）对应的字段，您可以从“存储库配置”窗口创建新特征。 有关详细信息，请参阅第 21 页码中的配置实体存储库。

要将输入字段映射到特征

1. 将“EA 导出”节点附加到“流”工作区上的数据源节点。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。 有关详细信息，请参阅第 19 页码中的映射的存储库输入选项。
3. 在“EA 导出”节点上，请选择“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. 要保存映射分配集（例如，为了使用其他数据源节点）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”选项卡中包含选项，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。

图片 3-3
将输入字段映射到特征

模式。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。 有关详细信息，请参阅第 17 页码中的实体存储库选项。

源标记。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。

实体类型。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。有关详细信息，请参阅第 21 页码中的配置实体存储库。

唯一键。（必需）要用于数据记录的唯一标识符的输入字段。

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。**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。 有关详细信息，请参阅第 21 页码中的配置实体存储库。
- **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。 有关详细信息，请参阅第 26 页码中的保留实体类型。

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。有映射要求相同的不同数据源时，这可能会非常有用，因为这会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。

图片 3-4
显示字段映射

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，您可以设置基本配置，并将其导出到可以随后导入其他存储库的文件。 有关详细信息，请参阅第 29 页码中的重用存储库配置。

警告：如果要对一个已包含数据的存储库修改并保存其配置，我们强烈建议您清除存储库的内容，然后重新载入数据。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. 选择您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 3-1
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。	有关详细信息，请参阅第 22 页码中的查看数据源映射。
特征	创建新特征，或复制、编辑或删除现有特征。	有关详细信息，请参阅第 23 页码中的保留存储库特征。
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）。	有关详细信息，请参阅第 26 页码中的保留实体类型。
解析规则	设置实体匹配的阈值。	有关详细信息，请参阅第 28 页码中的设置实体匹配的阈值。

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。

图片 3-5
配置窗口中的数据源映射

如果已经将新数据源添加到存储库，可单击**刷新**更新列表。

注意：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。 有关详细信息，请参阅第 15 页码中的连接数据源。

保留存储库特征

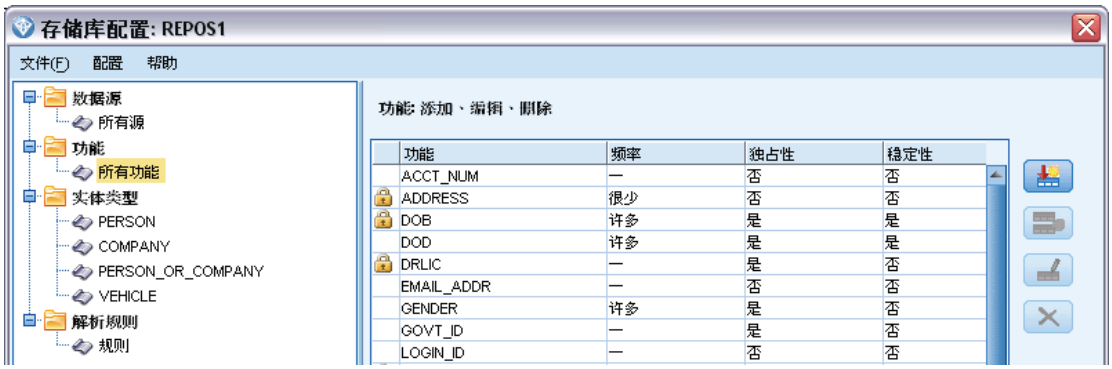
存储库特征是可以与实体数据源配合使用的单独信息类型。一些特征（例如，名、姓、出生日期等）可以与许多不同的数据源配合使用，其他特征则专用于特定数据源。一项特征可包含一个或多个元素，而每个元素通常相当于数据记录中的一个字段，或数据库表中的一列。

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。

图片 3-6
在配置窗口中列出的特征



特征列表会显示在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。 特征的名称。特征名称旁的挂锁符号说明该特征被锁定。您无法删除或复制锁定的特征，也无法保存对它们所做的更改。

频率。 指示可以对此特征具有相同值的身份的数量。有效值为**无**、**姓名**、**一个**（例如，护照号码）**几个**（例如，地址）或**许多**（例如，出生日期）。

独占性。 表示一个实体一般应仅有一个此类特征。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。注意：一般而言性别在作用期限内是稳定的，但因为它常常由于不良数据被错误指定，默认配置授予其值为**否**。

创建新特征

1. 进行以下其中一个操作。
   - 单击“创建新特征”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。 有关详细信息，请参阅第 25 页码中的添加或编辑特征。

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。 有关详细信息，请参阅第 25 页码中的添加或编辑特征。

编辑现有特征

警告：如果您在存储库已经包含数据时编辑或删除特征或特征元素，随后应清除存储库并重新载入数据。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。注意：您只能编辑您创建的那些特征，不能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。 有关详细信息，请参阅第 25 页码中的添加或编辑特征。

删除现有特征

警告：如果您在存储库已经包含数据时编辑或删除特征或特征元素，随后应清除存储库并重新载入数据。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。注意：您只能删除您创建的那些特征，不能删除系统提供的特征。
2. 进行以下其中一个操作。
   - 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：您不能撤消删除某个特征。

添加或编辑特征

警告：如果您在存储库已经包含数据时编辑或删除特征或特征元素，随后应清除存储库并重新载入数据。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。

图片 3-7
编辑特征

特征类型。表示特征所关联的信息类型的标签。该标签形成了特征标识符的第一部分。

描述。特征类型的简要文本说明，仅供参考。

频率。 指示可以对此特征具有相同值的身份的数量。有效值为**无**、**姓名**、**一个**（例如，护照号码）**几个**（例如，地址）或**许多**（例如，出生日期）。

独占性。 表示一个实体一般应仅有一个此类特征。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。注意：一般而言性别在作用期限内是稳定的，但因为它常常由于不良数据被错误指定，默认配置授予其值为**否**。

保留历史。 若设置为“是”（默认），则会保留对字段值进行的更改历史。例如，这对客户的地址字段很有用。在此情况下，如某位客户多次变更地址，则保留更改的历史则对匹配有帮助。不过，对于帐户余额这类字段，您应设置为“否”以免保留过量的数据。

元素表。此特征所包含的元素列表。

- **元素。**元素名称。

- **描述。**元素所提供内容的简要说明。
- **数据类型。** 可用于此元素的数据类型。
- **搜索。** 如选中，则表示特征的此元素本身可进行搜索，或与标有搜寻的其他特征元素一起搜索。例如，假设一个 PASSPORT 特征具有两个元素：ID_NUM 与 COUNTRY。如果您想只搜索 ID_NUM，则可选择其**搜索**选框，并清除 COUNTRY 的**搜索**选框。此特征的任何修饰符，像是国别或省份，或颁发日期与有效日期等，均不能成为搜索的一部分。
- **比较。** 指示特征中的哪些元素将送到比较例程进行实体解析。继续讨论具有 ID_NUM 与 COUNTRY 两个元素的 PASSPORT 特征的例子，由于 COUNTRY 资料不见得每次都提供，因此仅将 ID_NUM 发送到比较例程。
- **说明性的。** 指示特征中的哪些元素会出现在其内部描述中。在之前具有 ID_NUM 与 COUNTRY 两个元素的 PASSPORT 特征的例子中，如您希望两个元素都出现在特征描述中，请为两者都选择**描述性**选框。但假设该特征还有另外一个元素，ISSUE_DT。您需要清空其**描述性**选框以确保它不会出现在特征描述中。

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。

警告：如果您在存储库已经包含数据时编辑或删除特征或特征元素，随后应清除存储库并重新载入数据。这样做可以避免让存储库处于不一致的状态。

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诸如姓名、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

图片 3-8
列出属于实体类型的特征

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。

使用类型。 （可选）指示可能使用此特征的不同环境。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。 有关详细信息，请参阅第 19 页码中的映射的存储库输入选项。

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

警告：如果您在存储库已经包含数据时编辑或删除特征或特征元素，随后应清除存储库并重新载入数据。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

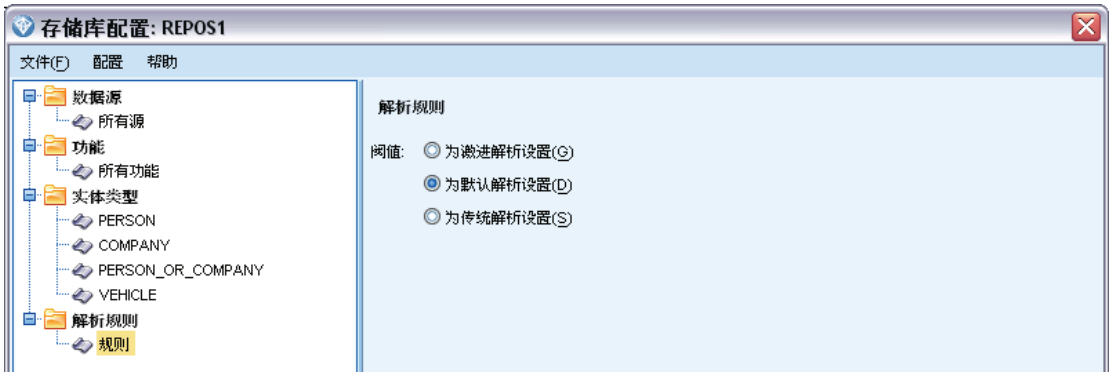
警告：如果您在存储库已经包含数据时编辑或删除特征或特征元素，随后应清除存储库并重新载入数据。这样做可以避免让存储库处于不一致的状态。

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。警告：您不能撤消删除实体类型。

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。

图片 3-9
设置实体解析规则



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**。

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。

重用现有配置

1. 显示您要使用其配置的存储库“存储库配置”窗口。 有关详细信息，请参阅第 21 页码中的配置实体存储库。
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，选择先前导出的 XML 文件的名称和位置，并单击**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）

在将至少一个数据源输入存储库时，您可以使用 Entity Analytics(EA) 源节点，将已解析的身份传递到其他 IBM® SPSS® Modeler 节点，以供进一步分析或处理，例如创建列出已解析身份的报告。

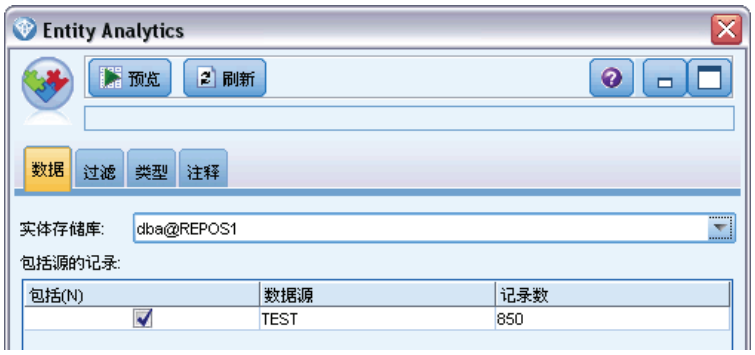
分析已解析的身份

1. 将 Entity Analytics(EA) 源节点添加到流。
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）。 有关详细信息，请参阅第 30 页码中的选择数据源。
4. 将其他节点添加到流，以执行您所需的处理。 有关详细信息，请参阅第 31 页码中的将节点添加到流。

选择数据源

在“数据”选项卡上，在存储库中至少选择一个要对其执行进一步处理的数据源。单击**刷新**更新所列出数据源的记录计数。

图片 3-10
在“源”节点上选择数据源



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。 有关详细信息，请参阅第 17 页码中的实体存储库选项。

包括源的记录。 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。

重命名数据字段

您可以使用“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 3-2
已解析身份字段

字段	描述
$EA-ID	实体标识符
$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

注意：尽管您也可以使用“过滤”选项卡过滤字段，但您不应在此进行，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 3-3
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	为从数据集读取数据值，提供选项。
缺失	用于指定字段缺失值的处理方法。
检查	确保字段值符合特定值或范围的验证选项。
角色	指定如果数据已被传递到建模节点或模型块，将如何使用字段。

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。

- 汇总节点，用于汇总输出，这可能是非常巨大
- 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）

在已经执行了存储库中的一些身份解析之后，您可以使用“流 EA”节点，将您随后遇到的新个案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输入记录相关的所有实体记录，“流 EA”节点仅输出与已在存储库中解析的条目有关的条目数据。 有关详细信息，请参阅第 35 页码中的流 EA 节点的输出。

比较新个案与存储库

1. 连接到包含您想要与现有实体比较的新纪录的数据源。 有关详细信息，请参阅第 15 页码中的连接数据源。
2. 在“记录选项”选项卡中，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，选择你要使用的“实体类型”。 有关详细信息，请参阅第 26 页码中的保留实体类型。
10. 将数据源中的输入字段映射到存储库中的特征。 有关详细信息，请参阅第 33 页码中的将输入字段映射到特征（“流 EA”节点）。
11. 或者单击**存储库**选项卡查看表格，表格显示此数据源（或所有被映射的数据源）的映射。 有关详细信息，请参阅第 34 页码中的显示字段映射（“流 EA”节点）。
12. 单击**查看**选项卡，查看已输入存储库的各种数据源的详细信息，以及设置检索现有实体的选择标准。 有关详细信息，请参阅第 34 页码中的查看数据源（“流 EA”节点）。
13. 节点设置正确时，请单击**确定**。
14. 将“表”节点附加到“流 EA”节点并运行流。

“表”节点的输出窗口会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-**。 有关详细信息，请参阅第 35 页码中的流 EA 节点的输出。

将输入字段映射到特征（“流 EA”节点）

“输入”选项卡包含：将输入到此节点的字段映射到存储库特征的选项。在此选项卡上设置映射分配，或者在**查看**选项卡上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。

图片 3-11
在存储库中将新记录中的字段映射到特征

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。 有关详细信息，请参阅第 17 页码中的实体存储库选项。

实体类型。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。有关详细信息，请参阅第 21 页码中的配置实体存储库。

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。**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。 有关详细信息，请参阅第 21 页码中的配置实体存储库。
- **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。 有关详细信息，请参阅第 26 页码中的保留实体类型。

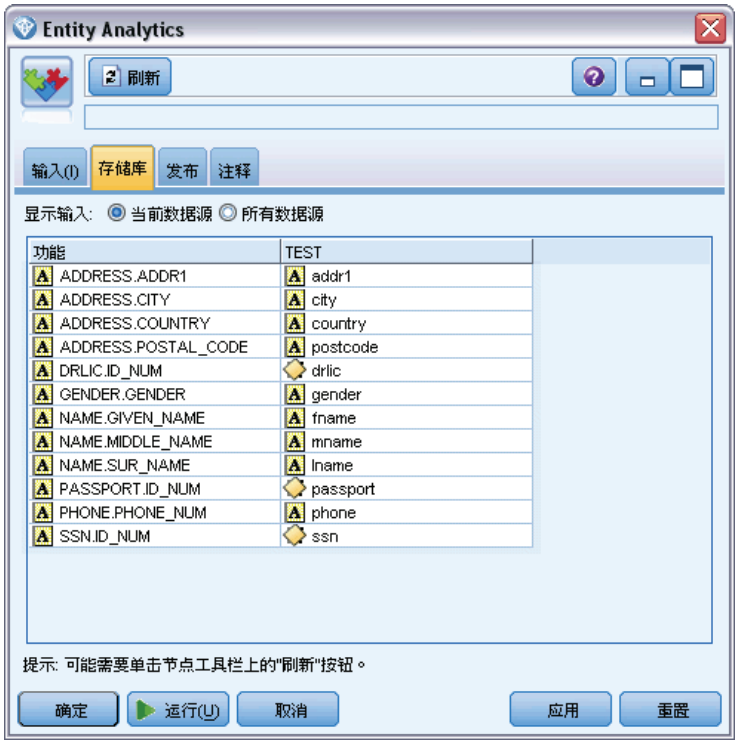
导入映射。从外部 XML 文件导入先前导出的一组字段到特征映射。有映射要求相同的不同数据源时，这可能会非常有用，因为这会避免必须针对不同的源重新定义相同的映射。

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（“流 EA”节点）

在“存储库”选项卡上，单击**刷新**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。

图片 3-12
显示字段映射



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。

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。

此节点。 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。

查看数据源（“流 EA”节点）

在“查看”选项卡上，您可查看已输入存储库的各种数据源的详细信息。此节点的输入根据这些数据源进行处理，以搜索和检索匹配实体。单击**刷新**以更新记录计数。

图片 3-13
查看数据源详细信息

匹配项。 这些选项定义：您在“输入”选项卡定义的字段到特征映射信息，与候选人记录（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。

- **只包括完全匹配项。** 这是最接近的匹配标准，选择的记录数最少。当您希望仅返回那些视为完全匹配的实体时，可使用此选项。
- **包括具有匹配标识符的备选项。** 当您希望返回匹配实体以及共享相同标识符的实体（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。
- **包括所有备选项。** 当您希望查看存储库中具有共享特征的最多可能实体数时，可使用此选项。这是最宽松的匹配标准，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。

数据源。 此表列出已载入存储库的不同数据源的源标记，以及每个源中的记录数。

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符。

字段	描述
$EA-SC	匹配近似度字段，标明此记录与存储库中观察实体的匹配近似度，其取值范围为 1.0 （低度匹配）到 10.0 （高度匹配）。
$EA-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-Feature1[, $EA-Feature2[, ...$EA-FeatureN]]	此记录在数据源文件中所映射的特征的值。

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

对于在 Entity Analytics 中创建的存储库，使用 IBM solidDB 产品创建了新的数据库服务。有一些管理任务与 solidDB 相关。这些任务一般由数据库管理员或系统管理员执行，包括：

- 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：

- 移动存储库存储目录
- 为“日期/时间”和“时间戳”字段设置流选项
- 在同一 Windows 系统中使用 SPSS Modeler 客户端和 SPSS Modeler Server 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 solidDB 数据库服务分配端口，该端口不能分配给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 SPSS Modeler Server 连接使用时，机器运行 IBM® SPSS® Modeler）上的数据库服务。

默认情况下，Entity Analytics 分配在范围 1320 至 1520 之间的端口，第一个创建的存储库从端口 1320 开始。若出现冲突，您可通过编辑以下文件来配置端口分配：<modeler 服务器安装路径>/ext/bin/pasw.entityanalytics/ea.cfg 然后为 min_port 和 max_port 设置适当的值。此文件的默认内容显示如下：

```
# port range configuration for entity analytics

#

# this port range controls which ports SolidDB 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

在创建存储库时定义：托管实体存储库的 solidDB 数据库的管理员用户名和密码。如果您知道正确的凭证，可以通过 solidDB SQL 编辑器更改这些详细信息。

开始 solidDB SQL 编辑器

1. 在客户端机器上，打开命令提示符窗口。
2. 输入：

   ```
   cdmodeler_install_dir\ext\bin\pasw.entityanalytics\solidDB\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 userusername identified bypassword;
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; 可关闭编辑器。
3. 重启 SPSS Modeler 客户端。

执行其他管理任务

有关 solidDB 数据库的其他管理任务，请访问 http://publib.boulder.ibm.com/ 参阅适当版本 IBM solidDB 的文档。

移动存储库存储目录

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。请按下面的步骤来进行移动操作。

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”字段设置流选项

如果您的源数据包含“日期/时间”或“时间戳”数据，请确保将相应的流选项设置为 IBM SPSS Modeler Entity Analytics 能识别的格式。

要设置流选项的格式

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**。

在同一 Windows 系统中使用 SPSS Modeler 客户端和 SPSS Modeler Server 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 Modeler Server 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。

注意：特别是在您使用一个 32 位的 SPSS Modeler 客户端和一个 64 位的 SPSS Modeler Server 服务器端（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。 有关详细信息，请参阅第 36 页码中的配置端口分配。
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

如果要从实体存储库清除数据记录，但保留配置信息，请使用存储库清除选项。

清除存储库

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. 选择您要清除的存储库。
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**清除存储库**按钮时，请单击此按钮。
8. 单击“清除数据源”对话框上的**清除存储库**，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。

警告：这会完全按照所指示的操作执行。**您无法撤消此操作。**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。 有关详细信息，请参阅第 39 页码中的清除实体存储库。

注意：以下步骤假设您可以从 SPSS Modeler 与存储库连接，并且知道托管存储库的数据库的管理员用户名和密码。如果并非这样，请在不能与存储库连接时，按照步骤将存储库删除。 有关详细信息，请参阅第 40 页码中的在不能与存储库连接时，将存储库删除。

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. 选择您要删除的存储库。
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：

```
cdmodeler_install_dir
cd ext\bin\pasw.entityanalytics\tools
delete_repository.batrepos_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. 打开命令解释程序。
2. 输入：

```
cdmodeler_server_install_dir
cd ext/bin/pasw.entityanalytics/tools
./delete_repository.shrepos_name
```

modeler_server_install_dir 是安装 SPSS Modeler Server 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 - 请从下列目录中将存储库文件删除：
directory_path
（请注意 - 在删除存储库文件之前可能需要先重启）
```

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。

章 4

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 所获得已令人印象深刻的结果。

本示例使用流 loan_entity_analytics.str，该流引用了数据文件 loan_applications.csv。这些文件可在任何也安装了 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 SPSS Modeler 程序组中访问 Demos 目录。loan_entity_analytics.str 文件在 Entity_Analytics 目录中。

注意：您需在系统中创建一个存储库，才能运行本示例流。请在继续本示例前执行这一操作。 有关详细信息，请参阅第 15 页码第 3 章中的创建存储库。

让我们从一个熟悉的情景开始 - 一家银行的管理人在审查客户的贷款申请时，担心他们是否可能拖欠还款。这家银行的 IT 部门是 SPSS Modeler 的长期用户，所以他们的员工已经根据自己过去所发放 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。

图片 4-1
带建模节点的初始流

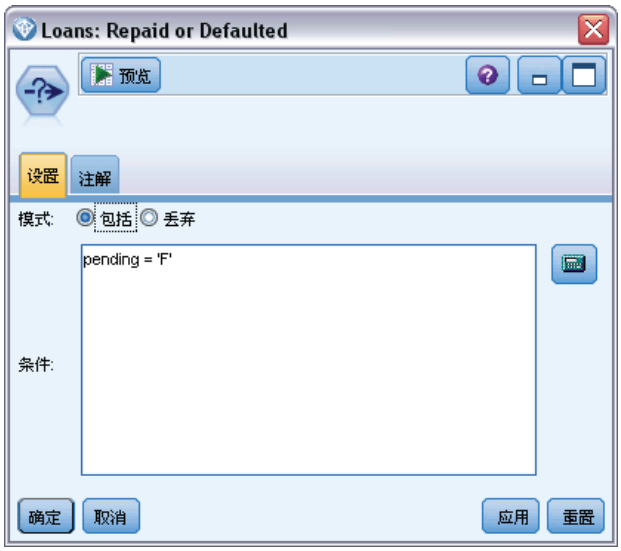
loan_applications.csv 数据集包含贷款申请仍在审查中的 150 个客户的详细信息以及过去贷款的详细信息，共计 850 条记录。

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**。

图片 4-2
“类型”节点中设置的字段角色

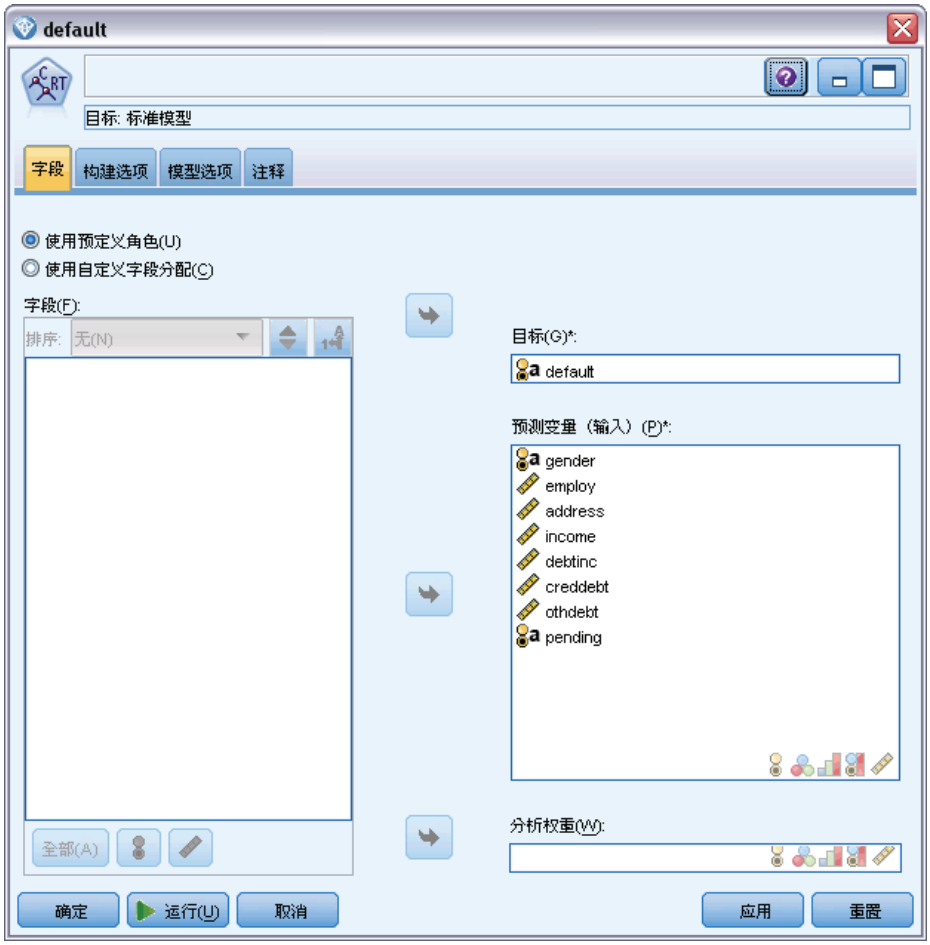
由于模型必须仅基于过去的数据进行预测，流所包含的“选择”节点仅包含那些没有标记为“待定”的贷款，因此将丢弃这 150 个审查中的贷款。

图片 4-3
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。

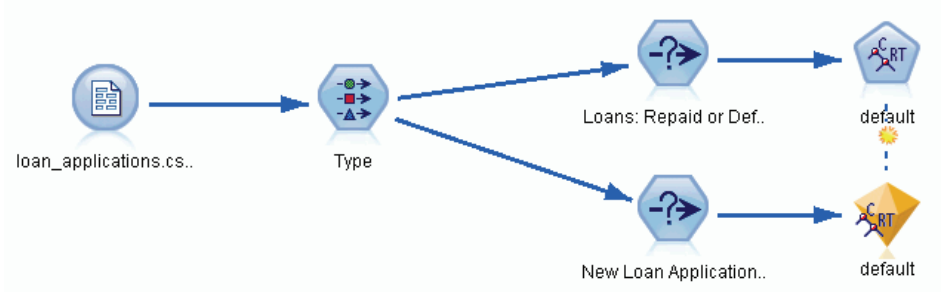
图片 4-4
指定预测变量和目标字段



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，正如先前由“类型”节点所指定的。

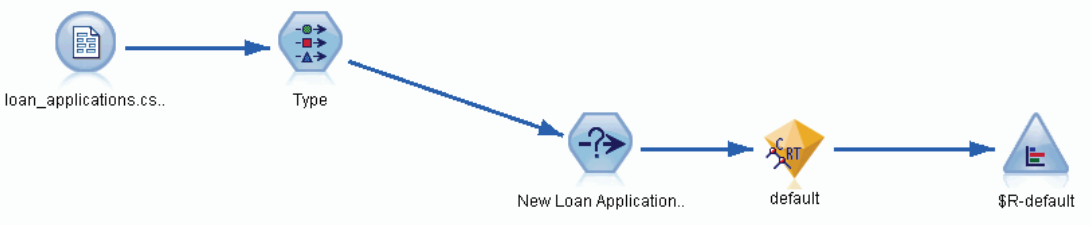
运行此流会生成一个模型块，其中包含已从预测变量字段构建的模型。

图片 4-5
添加模型块的流



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。使用原始数据集，分析人员插入一个“选择”节点，这一次该节点只包括标记为“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。

图片 4-6
用于选择新贷款申请并添加“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。此字段在“C&R 树”节点运行时，由其添加至数据模型。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。

图片 4-7
没有实体分析的分布节点输出

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可能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。

对于本例，我们将假设已经创建了实体存储库。 有关详细信息，请参阅第 15 页码第 3 章中的创建存储库。

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中，以便将源数据导入实体存储库中。

图片 4-8
将 EA 导出节点附加到数据源节点

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。实体存储库会提供一组标准信息类型（称为“特征”），以避免重复。

图片 4-9
将字段映射到特征

在“EA 导出”节点中，设置关于存储库的详细信息：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。在您使用的特征集中，有与字段 fname、mname、lname、生成、dob、性别、addr1、城市、国家、邮递区号、电话、电邮、社会安全号码、drlic 和护照相对应的特征。

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 4-1
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
生成	NAME.NAME_GEN
dob	DOB.DOB
性别	GENDER.GENDER
addr1	ADDRESS.ADDR1
城市	ADDRESS.CITY
国家	ADDRESS.COUNTRY
邮递区号	ADDRESS.POSTAL_CODE
电话	PHONE.PHONE_NUM
电邮	EMAIL_ADDR.ADDR
社会安全号码	SSN.ID_NUM
drlic	DRLIC.ID_NUM
护照	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符，即为您稍后将看到的 $EA-ID 字段。(注意：它与“EA 导出”节点中的“唯一键”字段不同，后者仅用于标识唯一的数据源记录。)

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。

图片 4-10
Entity Analytics(EA) 源节点

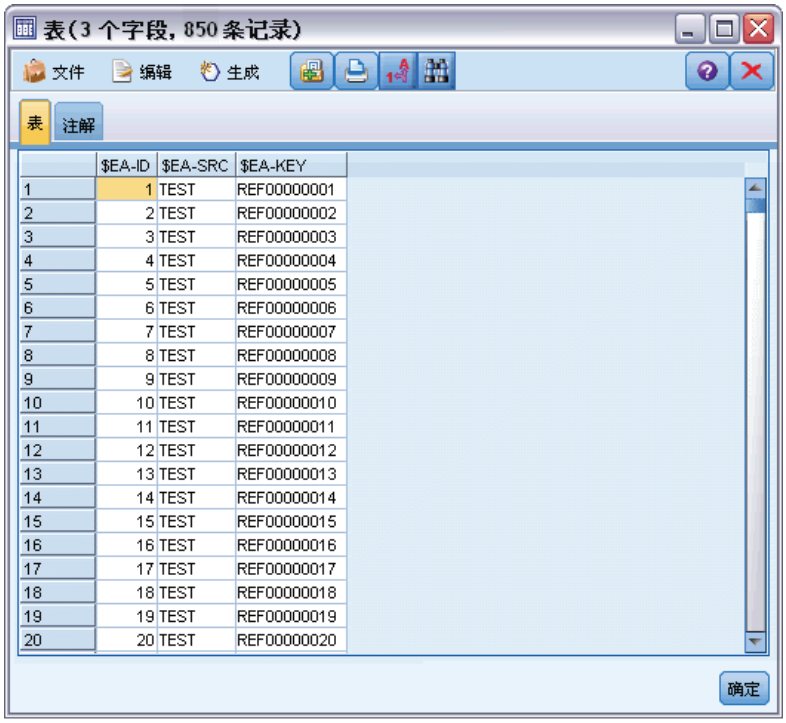
打开 Entity Analytics(EA) 源节点并设置“实体存储库”详细信息。随后会显示已导出到存储库的数据源列表，在本例中只有一个数据源。

图片 4-11
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，然后单击运行显示“表”节点输出窗口。

图片 4-12
表节点的输出



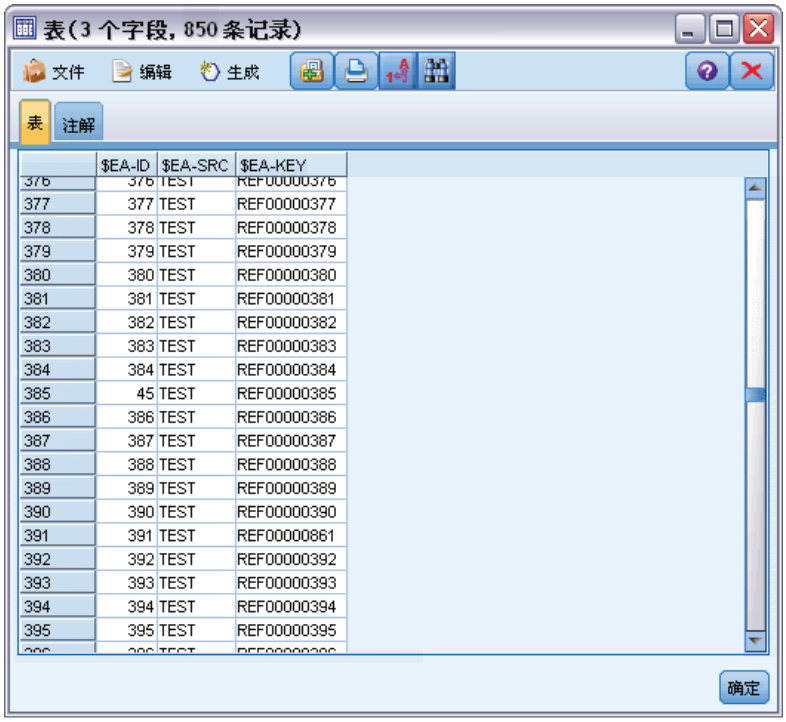
只有一个字段看起来很熟悉，即标为 $EA-KEY 的字段。它实际上是源数据中的键字段，之所以出现在这里是因为您在“EA 导出”节点中将其选为“唯一键”字段。

但系统还添加了另外两个字段。$EA-ID 字段是唯一的标识符，不是源记录的标识符，而是已解析身份的标识符。稍后我们将看到具体的差异。$EA-SRC 字段标识数据的来源，此处它指示 TEST，因为这是您已在“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用于进一步处理。

现在，将“表”节点输出向下滚动到 385 行。

图片 4-13
表输出行与 $EA-ID 号码之间的差异



	$EA-ID	$EA-SRC	$EA-KEY
377	377	TEST	REF00000377
378	378	TEST	REF00000378
379	379	TEST	REF00000379
380	380	TEST	REF00000380
381	381	TEST	REF00000381
382	382	TEST	REF00000382
383	383	TEST	REF00000383
384	384	TEST	REF00000384
385	45	TEST	REF00000385
386	386	TEST	REF00000386
387	387	TEST	REF00000387
388	388	TEST	REF00000388
389	389	TEST	REF00000389
390	390	TEST	REF00000390
391	391	TEST	REF00000861
392	392	TEST	REF00000392
393	393	TEST	REF00000393
394	394	TEST	REF00000394
395	395	TEST	REF00000395

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，他也拥有记录 REF00000045。在输出中，继续向下滚动，有更多号码失序，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，$EA-KEY 字段包含来自于源数据中键字段的数据。将“过滤”节点附加到 Entity Analytics(EA) 源节点并打开该“过滤”节点。在第二个**字段**列中双击字符串 **$EA-KEY** 并键入 **key**。

图片 4-14
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**。

图片 4-15
按升序排列实体 ID

将排序顺序保持为**升序**，并单击**确定**。

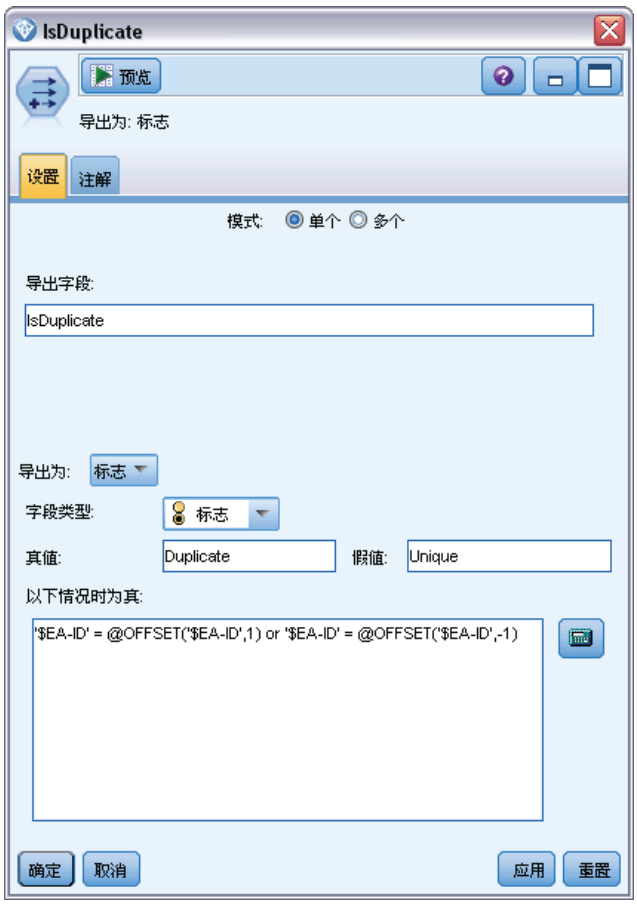
现在，您需要创建一个额外字段，用于指示记录是唯一还是重复的记录。将“派生”节点附加到“排序”节点。打开该“派生”节点并将**派生字段**名称设置为 IsDuplicate。从**派生为**列表中选择**标志**，这也将**字段类型**设置为**标志**。将**真值**字段设置为**重复**并将**假值**字段设置为**唯一**。

要查找重复的记录，需使用一个特殊的序列函数（名为 @OFFSET），它随 SPSS Modeler 一起提供。

在 If 字段中键入以下内容：

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

图片 4-16
在“派生”节点中设置条件

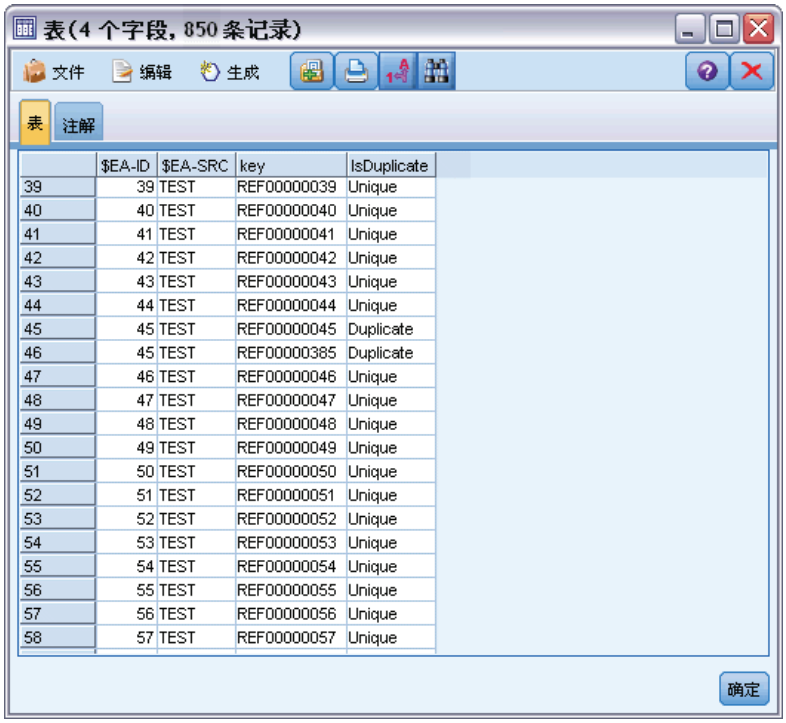


按升序排列实体 ID 后，@OFFSET 函数可用于检查相邻的实体 ID 是否相同，即记录是否重复。如果是，其 IsDuplicate 值将设置为重复，否则将设置为唯一。

单击**确定**关闭此节点。

要查看“派生”节点的效果，请将“表”节点附加到“派生”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。

图片 4-17
派生节点的输出



	$EA-ID	$EA-SRC	key	IsDuplicate
39	39	TEST	REF00000039	Unique
40	40	TEST	REF00000040	Unique
41	41	TEST	REF00000041	Unique
42	42	TEST	REF00000042	Unique
43	43	TEST	REF00000043	Unique
44	44	TEST	REF00000044	Unique
45	45	TEST	REF00000045	Duplicate
46	45	TEST	REF00000385	Duplicate
47	46	TEST	REF00000046	Unique
48	47	TEST	REF00000047	Unique
49	48	TEST	REF00000048	Unique
50	49	TEST	REF00000049	Unique
51	50	TEST	REF00000050	Unique
52	51	TEST	REF00000051	Unique
53	52	TEST	REF00000052	Unique
54	53	TEST	REF00000053	Unique
55	54	TEST	REF00000054	Unique
56	55	TEST	REF00000055	Unique
57	56	TEST	REF00000056	Unique
58	57	TEST	REF00000057	Unique

记得当时我们直接从 Entity Analytics(EA) 源节点查看输出。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，因为他们都是指实体 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。

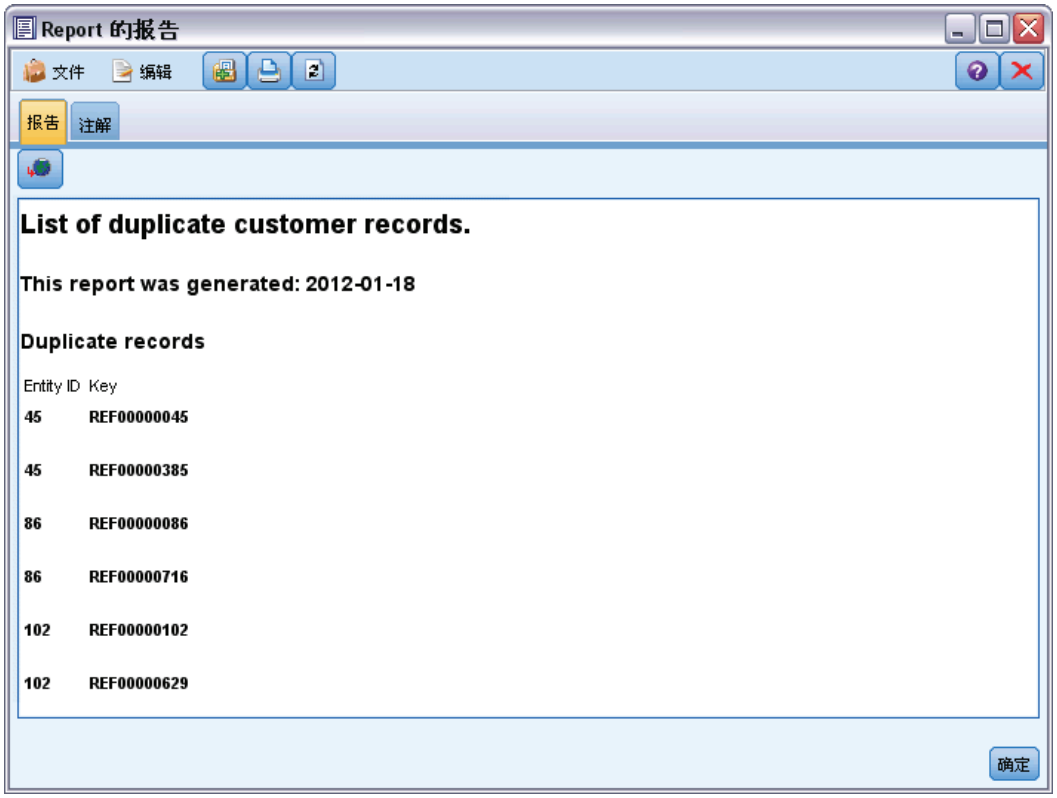
要获取列出重复记录的报告，请将“报告”节点（从节点选项板的“输出”选项卡）附加到 IsDuplicate 派生节点。打开该“报告”节点，将以下文本复制在“模板”选项卡的输入字段中，然后单击运行。

```
<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。

图片 4-18
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

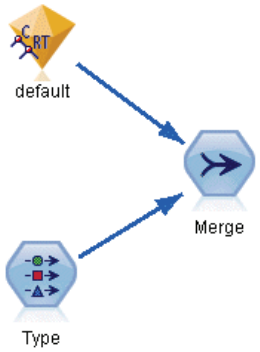
本例的最后阶段是查看添加实体分析是否会给银行的原始预测带来任何改变。您可能记得，原始模型预测了 150 个待定申请中有 13 个违约者。您将使用“合并”节点将该模型的输出与来自实体分析的关于重复记录的信息进行合并，以了解这样做是否会改变预测。

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 派生”节点，打开“类型”节点并单击**读取值**按钮。

图片 4-19
类型节点设置

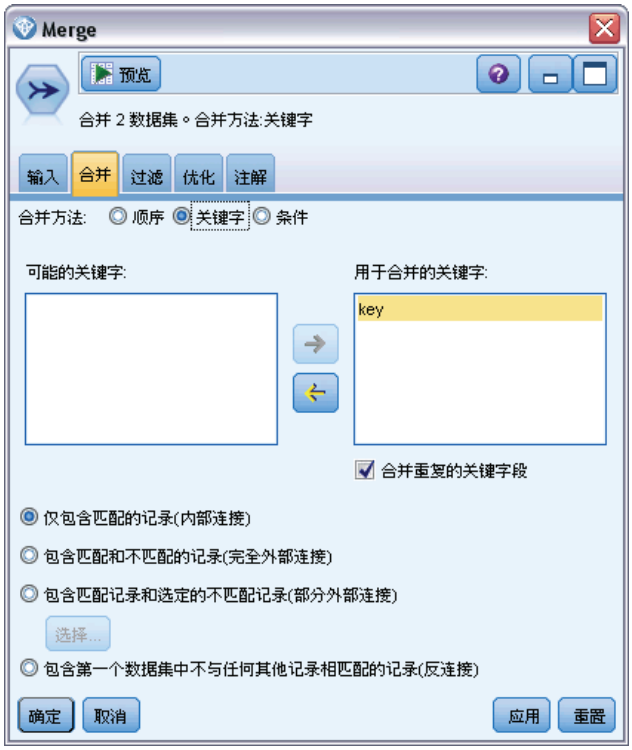
现在可以添加“合并”节点。将其附加到“类型”节点，并将其连接到包含原始模型的金色块。要执行此操作，请右键单击金色块，选择**连接**，然后单击“合并”节点，现在该节点应有两个输入箭头。

图片 4-20
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**。

图片 4-21
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。

要进行更新，您将使用“填充”节点，以便替换字段值。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，则将更改此字段的值。

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"
```

图片 4-22
指定替换字段值的条件

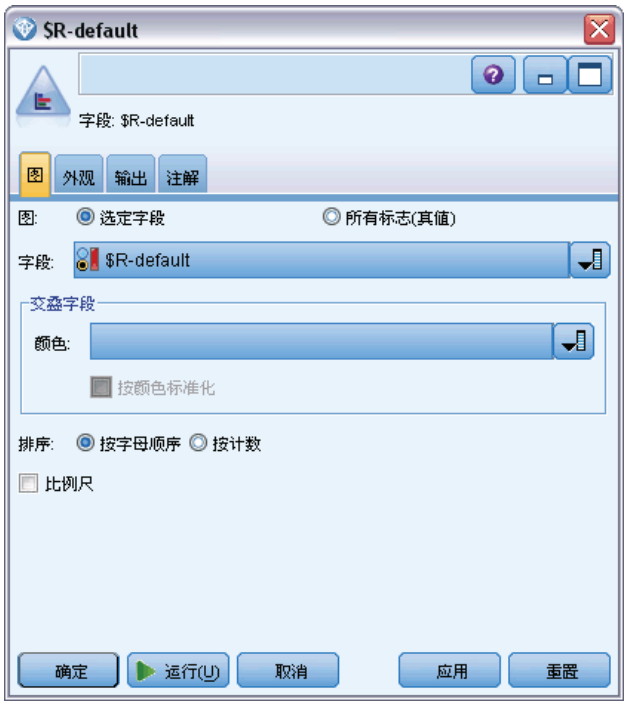
这些设置需要一些说明。条件的含义为，对于每个记录，如果原始数据集中的默认字段值不等于**默认**，并且该记录已被标记为重复，则将模型中 $R-default 字段的值设置为**默认**。

$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可能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，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**。

图片 4-23
分布节点设置



单击运行以生成新预测的图表。

图片 4-24
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。

注意：理想情况下，应在进行其他任何处理之前先消除重复的记录。要后续处理此操作，您可以使用自动数据准备 (ADP) 节点分析您的数据并找出修正，筛选出存在问题或可能无用的字段，并在适当的情况下派生新的属性，通过智能筛选技术改进性能。

合并实体分析与自动数据准备可确保您所处理的数据能尽量排除无用的数据。

附录
A

注意事项

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/地区中不提供在本文档中讨论的产品、服务或功能。请咨询您当地的 IBM 代表以了解有关您所在地区当前可用产品和服务的信息。任何对 IBM 产品、程序或服务的引用，并不意味着仅可使用这些 IBM 产品、程序或服务。作为替代，可以使用任何功能相当的产品、程序或服务，前提是不侵犯任何 IBM 知识产权。不过，用户应自行负责评估和验证任何非 IBM 产品、程序或服务的工作情况。

在本文档中介绍的主题可能涉及 IBM 的专利或申请中的专利。您获得该文档并不意味着授予您任何这些专利许可。您可以将书面许可查询函件发送至：

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

要查询双字节字符集 (DBCS) 相关许可证信息，请联系所在国家/地区中的 IBM 知识产权部门，或者以书面形式将查询函件发送至：

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

以下段落不适用于英国或此类条款与当地法律不符的其他国家/地区： INTERNATIONAL BUSINESS MACHINES 公司“按原样”提供本出版物，不保证任何明示或暗示，包括但不限于对非侵权性、适销性或对特定用途适用性的暗示担保。有些州/省不允许针对某些交易的明示或暗示免责条款，因此本声明可能不适用于您。

本信息可能含有技术误差或排版错误。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 可能随时对本出版物中所述的产品和/或程序进行改进和/或更改，恕不另行通知。

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。这些网站上的材料并非本 IBM 产品材料的一部分，您对这些网站的使用需自担风险。

IBM 可以自认为适当并且不会对您构成任何约束的任何方式使用或分发您提供的任何信息。

如果本程序的受许可方试图了解有关程序的信息以启用：(i) 在独立创建的程序和其他程序（包括本程序）之间交换信息；(ii) 相互使用交换的信息，则应联系：

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

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。

在本文档中介绍的受许可保护程序，及其所有受许可保护材料由 IBM 在双方签署的“IBM 客户协议”、“IBM 国际程序许可证协议”或任何其他等同协议下提供。

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。

非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。IBM 并未测试过这些产品，无法确认有关非 IBM 产品的性能准确性、兼容性或任何其他声明。有关非 IBM 产品功能的问题应由这些产品的供应商负责。

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、ibm.com 和 SPSS 是 IBM Corporation 的商标，在全球许多司法辖区注册。有关最新的 IBM 商标列表，请访问网页 http://www.ibm.com/legal/copytrade.shtml。

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、其他国家或地区或两者的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国、其他国家或地区或两者的商标。

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

Java 和所有基于 Java 的商标和徽标是 Sun Microsystems, Inc. 在美国、其他国家或地区或两者的商标。

其他产品和服务名称可能是 IBM 或其他公司的商标。

索引