



Informatica® PowerCenter Data Quality 集成
10.1.1 HotFix 2

用户指南

Informatica PowerCenter Data Quality 集成 用户指南
10.1.1 HotFix 2
2017 年 11 月

© 版权所有 Informatica LLC 2009, 2018

本软件和文档仅根据包含使用与披露限制的单独许可协议提供。未事先征得 Informatica LLC 同意，不得以任何形式、通过任何手段（电子、影印、录制或其他手段）复制或传播本文档的任何部分。

美国政府权利 交付给美国政府客户的程序、软件、数据库及相关文档和技术数据是指适用的联邦采购条例和政府机构特定补充条例中定义的“商业计算机软件”或“商业技术数据”。因此，使用、复制、披露、修改和改编应遵循适用的政府合同中规定的限制和许可条款、政府合同条款的适用范围以及 FAR 52.227-19 商用计算机软件许可中规定的额外权利。

Informatica、Informatica 标志和 PowerCenter 是 Informatica LLC 在美国和世界其他许多司法管辖区的商标或注册商标。欲获得 Informatica 商标的最新列表，请访问 <https://www.informatica.com/trademarks.html>。其他公司和产品名称可能是其各自所有者的商业名称或商标。

本软件和/或文档中的若干部分受第三方版权约束。所需的第三方声明随产品一起提供。

请参阅位于以下位置的专利：<https://www.informatica.com/legal/patents.html>。

免责声明：Informatica LLC 以“原样”提供本文档，不附带任何明示或暗示的担保，包括但不限于非侵权、适销性或特定用途适用性的暗示担保。Informatica LLC 不保证本软件和文档中没有错误。本软件或文档中提供的信息可能包括技术上的不准确性或排字错误。本软件和文档中包含的信息随时可能更改，恕不另行通知。

声明

本 Informatica 产品（以下称“软件”）包括由 Progress Software Corporation 的运营公司 DataDirect Technologies（以下称“DataDirect”）提供的某些驱动程序（以下称“DataDirect 驱动程序”），受以下条款和条件制约：

1. DataDirect 驱动程序以“原样”提供，不附带任何明示或暗示的担保，包括但不限于适销性、特定用途适用性以及非侵权的暗示担保。
2. 在任何情况下，DataDirect 或其第三方供应商均不对最终用户客户承担因使用 ODBC 驱动程序而引起的任何直接、间接、偶发、特殊、继发或其他损害赔偿的责任，无论是否已提前告知该种损害的可能性。这些限制适用于所有诉因，包括但不限于违反合同、违反担保、过失、严格责任、虚假陈述以及其他侵权行为。

本文档中的信息如有更改，恕不另行通知。如发现本文档中有什么问题，请通过以下电子邮件地址向我们报告：infa_documentation@informatica.com。

Informatica 产品根据对应协议的条款和条件进行担保。INFORMATICA 按“原样”提供本文档中的信息，无任何明示或暗示的担保，包括但不限于任何适销性和特定用途适用性担保，也没有任何非侵权担保或条件。

发布日期: 2018-07-03

目录

前言	4
Informatica 资源	4
Informatica Network	4
Informatica 知识库	4
Informatica 文档	4
Informatica 产品可用性矩阵	5
Informatica Velocity	5
Informatica Marketplace	5
Informatica 全球客户支持部门	5
 第 1 章： Informatica Data Quality 集成.....	6
Data Quality 集成概览	6
Data Quality 的特性和功能	6
集成安装程序	7
对象导出到 PowerCenter	7
安装引用数据	8
集成矩阵	8
产品集成的规则和准则	9
PowerCenter 中的数据质量转换行为	9
 第 2 章： 安装.....	12
安装概述	12
系统要求	12
安装前	12
安装服务器插件	13
在 Windows 平台上安装	13
在 UNIX 平台上安装	14
安装客户端插件	14
向 PowerCenter 注册插件	15
 第 3 章： 引用数据的要求.....	16
引用数据概览	16
引用数据文件的要求	17
地址引用数据的要求	17
标识填充数据的要求	19
概率模型和分类器模型的要求	20
概率模型和分类器模型文件位置	20
PowerCenter 集成服务内存分配	21
 索引.....	22

前言

PowerCenter® 用户可以使用从 Data Quality 环境中导入的映射和引用数据对象。《适用于 PowerCenter 的 Informatica Data Quality 集成用户指南》专为运行数据质量映射和管理相关引用数据对象的 PowerCenter 用户撰写。

Informatica 资源

Informatica Network

Informatica Network 囊括了 Informatica 全球客户支持部门、Informatica 知识库和其他产品资源。要访问 Informatica Network，请访问 <https://network.informatica.com>。

成员可以执行以下操作：

- 在一个位置访问您的所有 Informatica 资源。
- 在知识库中搜索文档、常见问题和最佳实践等产品资源。
- 查看产品可用性信息。
- 查看支持案例。
- 查找当地的 Informatica 用户组网络并与您的伙伴进行协作。

Informatica 知识库

使用 Informatica 知识库可在 Informatica Network 中搜索文档、入门知识文章、最佳实践和 PAM 等产品资源。

要访问知识库，请访问 <https://kb.informatica.com>。如果您对知识库有任何疑问、意见或建议，请与 Informatica 知识库团队联系，电子邮件地址为 KB_Feedback@informatica.com。

Informatica 文档

要获取有关产品的最新文档，请浏览 Informatica 知识库，网址为 https://kb.informatica.com/_layouts/ProductDocumentation/Page/ProductDocumentSearch.aspx。

如果您对此文档有任何疑问、意见或建议，请与 Informatica 文档团队联系，电子邮件地址为 infa_documentation@informatica.com。

Informatica 产品可用性矩阵

产品可用性矩阵 (PAM) 指明了产品版本支持的操作系统版本、数据库以及其他类型的数据源和目标。如果您是 Informatica Network 成员，您可以访问 PAM，网址为 <https://network.informatica.com/community/informatica-network/product-availability-matrices>。

Informatica Velocity

Informatica Velocity 收集了 Informatica 专业服务开发的一系列提示和最佳实践。Informatica Velocity 基于数以百计的数据管理项目的实际经验而开发，汇集了我们曾在世界各地组织就职的顾问在成功规划、开发、部署和维护数据管理解决方案方面的知识。

如果您是 Informatica Network 成员，您可以访问 Informatica Velocity 资源，网址为 <http://velocity.informatica.com>。

如果您对 Informatica Velocity 有任何疑问、意见或建议，请通过 ips@informatica.com 与 Informatica 专业服务联系。

Informatica Marketplace

Informatica Marketplace 是一个论坛，该论坛中提供的解决方案可补充、扩展或增强您的 Informatica 实现。您可以利用 Informatica 开发人员和合作伙伴提供的数以百计解决方案中的任何方案，提高生产率，加快项目的实现时间。您可以访问 Informatica Marketplace，网址为 <https://marketplace.informatica.com>。

Informatica 全球客户支持部门

您可以通过电话或 Informatica Network 上的联机支持与全球支持中心联系。

要查找您当地的 Informatica 全球客户支持部门电话号码，请访问 Informatica 网站，链接为：
<http://www.informatica.com/us/services-and-training/support-services/global-support-centers>。

如果您是 Informatica Network 成员，您可以使用联机支持，网址为 <http://network.informatica.com>。

第 1 章

Informatica Data Quality 集成

本章包括以下主题：

- [Data Quality 集成概览, 6](#)
- [Data Quality 的特性和功能, 6](#)
- [集成安装程序, 7](#)
- [对象导出到 PowerCenter, 7](#)
- [安装引用数据, 8](#)
- [集成矩阵, 8](#)
- [产品集成的规则和准则, 9](#)
- [PowerCenter 中的数据质量转换行为, 9](#)

Data Quality 集成概览

Informatica 用户可以在 Informatica Data Quality 中创建和运行映射。用户将映射和其他对象保存到模型存储库。用户可以导出对象以便在 PowerCenter 中使用。PowerCenter 用户可以将这些对象导入到 PowerCenter 存储库，还可以将这些对象添加到会话。

使用对象分析并增强数据的内容和结构。

您可以使用用户在 Data Quality 中创建的对象 10.1.1 HotFix 2 与以下版本的 PowerCenter：

- PowerCenter 10.1.1 安装有最新版本的 HotFix。
- PowerCenter 10.1.
- PowerCenter 9.6.1 安装有最新版本的 HotFix。

如果使用的是最新版本的 Data Quality 和 PowerCenter，则您无需执行任何额外安装即可导入来自模型存储库的对象。如果使用的是较旧版本的 PowerCenter，则您必须安装并注册最新版本的 Informatica Data Quality 集成插件。

Data Quality 的特性和功能

Informatica Data Quality 使用 Informatica 应用程序创建对象，将对象保存到模型存储库，并运行映射。

如果您使用的不是最新的产品版本，可能需要安装适用于 PowerCenter 的 Informatica Data Quality 插件。

使用 Informatica Data Quality 可设计并运行用于实现以下目标的映射：

- 剖析数据。剖析可揭示数据的内容和结构。剖析是任何数据项目中的关键一步，因为它可以识别数据中的优势和劣势，并帮助您制定项目计划。
- 创建结果卡以查看数据质量。结果卡是配置文件中对质量度量的一种图形表示形式。
- 标准化数据值。对数据进行标准化，以消除您在运行配置文件时发现的错误和不一致。您可以对标点、格式和拼写方面的变体进行标准化。例如，您可以确保城市、省/自治区/直辖市和邮政编码值一致。
- 解析记录。对数据记录进行解析，以改进记录结构并从数据中推导其他信息。您可以将一个自由格式数据字段拆分为多个包含不同信息类型的字段。您也可以将信息添加到记录中。例如，您可以将客户记录标记为个人客户或企业客户。
- 验证通信地址。地址验证可评估并提高通信地址数据的准确性及可送达性。地址验证会通过比较地址记录与国家邮政运营商提供的引用数据，更正地址中的错误并补全不完整的地址。地址验证还可以添加邮政信息，从而加快邮件送达并降低邮件成本。
- 查找重复记录。重复记录分析可对一组记录进行相互比较，找出选定的数据列中相似或一致的值。您可以设置一个表示字段值之间存在较好匹配的相似性级别。您还可以设置匹配计算中给予每个列的相对权重。例如，您可以设置姓氏信息的优先级高于名字信息。
- 创建和运行数据质量规则。Informatica 提供预设的规则，您可以运行或编辑这些规则来符合项目目标要求。您可以在 Developer 工具中创建这些规则。
- 与 Informatica 用户协作。您添加到模型存储库的规则和引用数据表可供用户在 Developer 工具和 Analyst 工具中使用。用户可以协作完成项目，并且不同用户可以在项目的不同阶段获得对象的所有权。
- 将映射导出到 PowerCenter。您可以将映射导出到 PowerCenter，以重用元数据来进行物理数据集成或创建 Web 服务。

集成安装程序

如果您使用不同版本的 Data Quality 和 PowerCenter 创建和存储对象，请安装适用于 PowerCenter 的 Informatica Data Quality 插件。从 Informatica 可以下载该插件的安装程序。

根据您要使用的 PowerCenter 版本，请安装相应的插件版本。该插件提供客户端和服务器安装程序。

您可以在 PowerCenter 客户端计算机上运行客户端安装程序，以查看存储库中的 Mapplet 和映射并将它们添加到会话。也可以在 PowerCenter 集成服务计算机上运行服务器安装程序，以便在工作流中运行会话。

在集成服务计算机上安装插件后，请向 PowerCenter 存储库注册该插件。

对象导出到 PowerCenter

使用 Developer tool 可从 Informatica 模型存储库导出 Mapplet 和映射等对象。可以将对象从模型存储库导出到 PowerCenter 存储库，并且可以将对象从模型存储库导出到文件系统。使用 PowerCenter Repository Manager 可向 PowerCenter 存储库导入 XML 对象。

- 关于从模型存储库中导出对象的信息，请参见《Informatica Developer 映射指南》和《Informatica Developer Tool 指南》。
- 有关向 PowerCenter 存储库导入对象的信息，请参阅《PowerCenter 存储库指南》。

安装引用数据

如果 Mapplet 或映射要读取引用数据文件，请在 PowerCenter 集成服务计算机上安装引用数据文件。

您可以在导出 Mapplet 或映射时使用 Developer 工具从模型存储库导出引用数据。Developer 会将引用数据导出为一个压缩文件。您提取该文件后将引用数据复制到集成服务计算机。

您也可以使用 Data Quality 内容安装程序向集成服务计算机添加引用数据。运行内容安装程序可安装地址引用数据文件和标识填充数据文件。

集成矩阵

Data Quality 与 PowerCenter 之间的集成程度取决于各自的产品版本。您可以将对象和引用数据从多个版本的 Data Quality 导出到多个版本的 PowerCenter。您不能将对象或引用数据从 PowerCenter 导出到 Data Quality。

下表显示了您可以集成的产品版本：

产品名称	PowerCenter 10.1.1 HotFix 2	PowerCenter 10.1.1 HotFix 1	PowerCenter 10.1.1	PowerCenter 10.1	PowerCenter 10.0	PowerCenter 9.6.1
数据质量 10.1.1 HotFix 2	是	否	否	安装插件	否	安装插件
数据质量 10.1.1 HotFix 1	否	是	否	安装插件	否	安装插件
数据质量 10.1.1	否	否	是	安装插件	否	安装插件
数据质量 10.1	否	否	否	是	否	安装插件
数据质量 10.0	否	否	否	否	是	安装插件
数据质量 9.6.1	否	否	否	否	否	是

表中的值代表以下类型的集成：

是

您可以将对象和文件从 Data Quality 环境中导出到 PowerCenter 环境。不需要安装或注册任何插件。您可以在升级至指定版本的 PowerCenter 存储库中运行数据质量对象。

安装插件

您可以将对象和文件从 Data Quality 环境中导出到 PowerCenter 环境。需要安装并注册适用于您所用的 PowerCenter 版本的客户端和服务端插件。您可以在升级至指定版本的 PowerCenter 存储库中运行数据质量对象。

否

无法将在 Data Quality 环境中定义的对象和文件用到 PowerCenter 环境。

产品集成的规则和准则

您可以将数据质量对象从不同版本的 Informatica Data Quality 导出到不同版本的 PowerCenter。请安装适用于您从中导出对象的 Informatica Data Quality 版本的集成插件。

将 Informatica Data Quality 与 PowerCenter 集成时，请考虑以下规则和准则：

- 如果 PowerCenter 存储库产品版本与模型存储库产品版本匹配，则不要安装集成插件。
- 无法在一台 PowerCenter 计算机上同时运行两个 Data Quality 集成插件。PowerCenter 使用最近安装的插件。例如，如果在 PowerCenter 9.5.1 计算机上安装 Data Quality 9.6.1 集成插件，则无法再从 9.5.1 版本模型存储库导入并运行数据质量对象。在安装 9.6.1 插件之前，可以继续使用已导入到 PowerCenter 存储库的任何数据质量对象。
- 无法在 PowerCenter 存储库中编辑数据质量 Mapplet 或映射。

PowerCenter 中的数据质量转换行为

当您数据质量 Mapplet 或映射从模型存储库导入 PowerCenter 存储库时，数据质量转换行为会发生变化。

合并转换的注意事项

Informatica Data Quality 在版本 9.5.1 中更新了合并转换的默认策略。在版本 9.5.1 和更高版本中，合并转换会选择具有最高行 ID 的记录作为剩余记录。在版本 9.5 和更低版本中，合并转换在您指定的列中使用最常出现的非空值以标识剩余记录。

如果您将包含合并转换的 PowerCenter 存储库升级到 9.5.1 或更高版本，则升级操作会保留最常出现的非空值作为默认设置。如果从 9.5.1 模型存储库或更高版本存储库中导入包含合并转换（具有默认设置）的映射，则转换将使用具有最高行 ID 标识剩余记录。

转换为 Mapplet

以下转换在导入到 PowerCenter 存储库时转换成 Mapplet：

- 合并
- 异常
- 键生成器
- 匹配

当您导入包含一个或多个上述转换的映射时，PowerCenter 会处理每个转换，将其转换为一个包含扩展配置的 Mapplet。

当您导入包含一个或多个上述转换的 Mapplet 时，PowerCenter 会将每个转换的配置扩展为多个转换。

为了完成转换逻辑，PowerCenter 可能会向数据质量转换或 Mapplet 扩展中添加转换。例如，PowerCenter 会使用另外的转换来扩展匹配转换。

判定转换的注意事项

如果映射或 Mapplet 包含到 PowerCenter 的判定转换且转换脚本包含系统参数，则不能导出该映射或 Mapplet。导出操作无法将系统参数转换为 PowerCenter 可以使用的值。在导出包含使用系统参数的判定转换的映射或 Mapplet 之前，请使用适当值替换参数。

异常转换的注意事项

PowerCenter 无法使用来自低于 9.5 版本 PowerCenter 存储库的异常转换。如果您将一个 PowerCenter 存储库升级到版本 9.5 或更高版本，并且该存储库包含一个具有异常转换的映射，请再次导出该映射。如果需要，请在 Developer tool 中再次创建该映射。

运行包含异常转换的映射时，您可以在 Analyst 工具中更新离散记录或重复记录输出表。如果您计划在 Analyst 工具中更新离散记录表，必须为运行映射的会话启用高精度设置。

匹配转换的注意事项

如果一个标识匹配映射会将标识索引数据读取或写入到数据库表，您不能在 PowerCenter 中运行该映射。要在 PowerCenter 中执行标识匹配分析，请在 Developer tool 中配置匹配转换，以将标识索引数据写入到临时文件。

多策略转换

在 Developer tool 中，您可以在以下每个转换内定义多个数据转换策略：

- 大小写
- 分类器
- 判定
- 键生成器
- 标签创建器
- 匹配
- 合并
- 解析器
- 标准创建器

当您导入包含多策略转换的 Mapplet 或映射时，每个策略会转换为一个单独的转换。

分区设置

在网格上运行包含数据质量映射的会话任务前，您需要先确定创建网格节点时使用的 PowerCenter 版本。如果您是在 PowerCenter 9.5.1 HotFix 2 或较早版本的 PowerCenter 中创建的节点，映射可能会失败。

要验证数据质量映射能否在网格的所有节点上运行，请执行以下任务之一：

- 查看任何向网格节点分发任务的 PowerCenter 集成服务的高级属性。将每个 PowerCenter 集成服务配置为在分发任务时忽略资源要求。默认情况下，高级属性将指定 PowerCenter 集成服务忽略资源要求。
- 重新创建您之前在 PowerCenter 9.5.1 HotFix 2 或较早版本的 PowerCenter 中创建的任何节点。重新创建节点时，PowerCenter 会将当前节点配置文件应用到该节点。

当您在网格上运行会话任务时，Data Transformation Manager 进程 (DTM) 会跨网格节点分发会话线程。某些转换不可在网格范围内分区。当一个转换不可在网格范围内分区时，DTM 会为该转换的线程创建单个分区组，并在单个节点上运行这些线程。

下表说明了数据质量转换的分区设置：

转换	分区设置
地址验证	在整个网格范围内
关联	不可分区
AV 报表	可本地分区
大小写转换器	在整个网格范围内
分类器	可本地分区
群集	不可分区
比较	在整个网格范围内
合并	不可分区
判定	在整个网格范围内
异常	不可分区
执行点	在整个网格范围内
键生成器	不可分区
密钥库	不可分区
标签创建器	在整个网格范围内
合并	在整个网格范围内
对生成器	不可分区
解析器（概率分析）	可本地分区
解析器	在整个网格范围内
标志标准创建器	在整个网格范围内
基于权值的分析器	在整个网格范围内

第 2 章

安装

本章包括以下主题：

- [安装概述, 12](#)
- [系统要求, 12](#)
- [安装前, 12](#)
- [安装服务器插件, 13](#)
- [安装客户端插件, 14](#)
- [向 PowerCenter 注册插件, 15](#)

安装概述

Data Quality 集成插件具有服务器组件和客户端组件。

请完成以下步骤安装并注册该插件：

1. 在 PowerCenter 集成服务计算机上运行服务器安装程序。
2. 在 PowerCenter 客户端计算机上运行客户端安装程序。
3. 向 PowerCenter 存储库注册 IDQZIntegration.xml 文件。

系统要求

Informatica Data Quality 集成插件的系统要求与该插件要安装到的 PowerCenter 应用程序和服务的系统要求相同。

有关 PowerCenter 的先决条件和系统要求的信息，请参阅 PowerCenter 安装文档。

安装前

如果您在 PowerCenter 中使用引用数据，请注意安装程序的以下行为。

集成插件安装程序会将属性文件写入 PowerCenter 集成服务计算机。这些属性文件指定了数据质量映射可以读取的引用数据文件的位置。

安装程序会写入以下文件：

- AD50.cfg：用于存储地址引用数据的配置属性。
- CLASSIFIER.properties：用于存储分类器模型文件的配置属性。
- IDQTx.cfg：用于存储标识填充数据的配置属性。
- NER.properties：用于存储概率模型数据的配置属性。

安装程序会将文件添加到以下目录：

[PowerCenter_installation]/server/bin

如果安装程序在 bin 目录下找到属性文件，会为该文件创建一个备份副本，然后安装新文件。要在安装后保留当前的引用数据配置，请合并新旧文件的内容。

如果您安装 PowerCenter 9.5.1 或更高版本，服务器安装过程默认会将配置文件添加到 bin 目录。

安装地址引用数据的先决条件

安装地址引用数据之前，请停止 PowerCenter 集成服务。安装数据之后再重新启动该服务。

安装服务器插件

Informatica Data Quality 集成服务器插件安装程序会将数据质量转换应用程序文件添加到 PowerCenter 集成服务目录结构。

在 Windows 平台上安装

运行 Windows Server 安装程序将插件添加到 Windows 平台的 PowerCenter 集成服务计算机。

在安装前，关闭所有应用程序。

1. 使用系统用户帐户登录计算机。
2. 提取适用于 Windows 的集成服务器文件。
3. 浏览到提取文件所在的目录并运行 install.exe。
4. 选择语言并单击**确定**。
5. 在**欢迎**屏幕上单击**下一步**。
6. 验证 PowerCenter 服务当前未在运行，然后单击**确定**。
7. 指定 PowerCenter 安装目录的路径。安装程序会为此目录提供一个默认路径。要选择其他路径，请单击**选择**。
8. 单击**下一步**。
9. 检查安装前的摘要信息。单击**上一步**修复任何错误。
10. 单击**安装**。
11. 检查安装后的摘要信息。
有关安装任务的详细信息以及要查看已安装组件的配置属性，请查看安装日志文件。
12. 单击**完成**。

在 UNIX 平台上安装

您可以运行 UNIX Server 安装程序将该插件添加到 UNIX 平台的 PowerCenter 集成服务计算机。

1. 使用系统用户帐户登录计算机。
2. 关闭所有其他应用程序。
3. 提取适用于 UNIX 的集成服务器文件。
4. 打开 UNIX Shell 并导航到提取安装文件的目录。
5. 输入 `sh install.bin`。
6. 要接受英语作为安装语言，请按 **Enter**。
7. 阅读欢迎文字并按 **Enter**。
8. 键入 PowerCenter 安装目录的绝对路径，或按 **Enter** 接受默认路径。
9. 检查安装前的摘要信息。要修复任何错误，请键入 `back` 并按 **Enter**。
10. 按 **Enter** 开始安装。
11. 检查安装后的摘要信息。

有关安装任务的详细信息以及要查看已安装组件的配置属性，请查看安装日志文件。

12. 按 **Enter** 完成安装过程。

安装客户端插件

在 PowerCenter 客户端计算机上运行 Data Quality 集成客户端安装程序。

1. 关闭所有其他应用程序。
2. 提取集成客户端文件。
3. 打开提取文件所在的目录。
4. 运行 `install.exe`。
5. 在欢迎屏幕上单击**下一步**。
安装程序将提示关闭 PowerCenter 客户端应用程序后再继续安装。
6. 验证没有 PowerCenter 客户端应用程序正在运行，然后单击**确定**。
7. 指定 PowerCenter 客户端目录的路径。安装程序会为此目录提供一个默认路径。要选择其他路径，请单击**选择**。
8. 单击**下一步**。
9. 检查安装前的摘要信息。单击**上一步**修复任何错误。
10. 单击**安装**。
11. 检查安装后的摘要信息。
有关安装任务的详细信息以及要查看已安装组件的配置属性，请查看安装日志文件。
12. 单击**完成**。

向 PowerCenter 注册插件

您需要向任何要包含数据质量对象的 PowerCenter 存储库注册插件。

集成安装程序会将 IDQZIntegration.xml 文件写入到 PowerCenter 集成服务计算机上的 \$INFA_HOME\server\bin\native 目录。请验证您是否可以访问此文件。如果需要，请将此文件复制到或通过 FTP 传输到 PowerCenter 集成服务计算机。

► 要注册插件，请运行 *pmrep RegisterPlugin* 命令。运行该命令时，请包含 -N 命令。

有关 *pmrep RegisterPlugin* 命令的详细信息，请参阅《*Informatica 命令参考*》。

注意: 如果您没有注册插件的适当权限，请联系 PowerCenter 存储库服务管理员。

第 3 章

引用数据的要求

本章包括以下主题：

- [引用数据概览, 16](#)
- [引用数据文件的要求, 17](#)
- [地址引用数据的要求, 17](#)
- [标识填充数据的要求, 19](#)
- [概率模型和分类器模型的要求, 20](#)

引用数据概览

数据质量转换可以读取引用数据。如果 PowerCenter 存储库中有任何转换要读取引用数据，请确认 PowerCenter 环境中存在引用数据文件。

PowerCenter 中的数据质量转换可以使用以下类型的引用数据：

引用数据文件

包含标准版本业务术语和替代版本术语的文本文件。您可以编辑引用数据文件。

地址引用数据文件

包含一个国家/地区所有通信地址的完整数据的数据文件。您在 PowerCenter 集成服务计算机上安装地址引用数据文件。地址引用数据文件不可编辑。

标识填充文件

用于定义不同标识类型并包含匹配转换所用算法的数据文件。您在 PowerCenter 服务计算机上安装这些填充文件。这些填充文件不可编辑。

内容集

含有一个或更多转换所指定的引用数据的数据文件。当您导出要读取内容集的映射时，导出进程会随转换元数据一起导出内容集数据。对于分类器和概率模型文件，导出过程会从 Data Quality 目录结构中导出文件数据。

引用数据文件的要求

数据质量映射可以读取您在模型存储库中定义的引用数据对象。

当 Informatica Data Quality 用户导出一个读取引用数据对象的映射时，导出进程会以一个或多个引用数据文件的形式将引用数据写入文件系统。当您在 PowerCenter 中使用该映射时，必须验证这些文件已安装在集成服务可以读取的位置。

您在导出映射时指定引用数据文件的文件位置。可以将这些文件导出到 PowerCenter 目录结构中，或将它们导出到文件系统，然后复制到 PowerCenter 集成服务计算机。

PowerCenter 目录结构中的引用数据文件位置必须与模型存储库结构中的引用表位置相对应。该目录结构必须遵循以下格式：

```
<PowerCenter_Root_Directory>\services\<Model_Repository_Project_Name>\<Model_Repository_Project_Folder_Name>
```

注意: PowerCenter 读取 UTF-8 格式的引用数据文件。Developer 工具以 UTF-8 格式将引用数据导出到 PowerCenter。

通过 Data Quality 内容安装程序安装引用数据

Data Quality 内容安装程序的文件集包括要安装到 PowerCenter 目录结构的引用数据文件。

您要在运行内容安装程序时确定 PowerCenter 安装目录。默认情况下，内容安装程序会向根目录附加以下路径：

```
\services\DQContent\INFA_Content\
```

您可以编辑该路径以符合模型存储库的结构，或者在运行内容安装程序后将引用数据文件复制到需要的路径。

使用环境变量定义引用数据文件位置

您可以在 PowerCenter 集成服务计算机上定义一个 *INFA_CONTENT* 环境变量，用于设置集成服务读取引用数据文件所使用的路径。

当您无法在 PowerCenter 集成服务计算机上需要的位置安装引用数据文件时，可使用此环境变量。请将文件安装在 PowerCenter 集成服务可以读取的位置，然后使用该安装路径创建 *INFA_CONTENT* 环境变量。

地址引用数据的要求

执行地址验证的映射会读取地址引用数据。您需要从 Informatica 购买并下载地址引用数据。请将这些数据文件复制到 PowerCenter 集成服务计算机。

下表说明了您可以安装的地址引用数据的类型：

引用数据的类型	何时使用
批处理	用于常规的地址验证操作。
CAMEO	用于向地址记录中添加消费者人口数据。

引用数据的类型	何时使用
已认证	用于验证地址记录是否符合邮件运营商规定的认证标准。 以下国家/地区规定的认证标准： <ul style="list-style-type: none"> - 澳大利亚。根据地址匹配审批制度 (Address Matching Approval System, AMAS) 标准认证邮件。 - 加拿大。根据软件评估和识别程序 (Software Evaluation And Recognition Program, SERP) 标准认证邮件。 - 法国。根据国家地址管理服务 (National Address Management Service, SNA) 标准认证邮件。 - 新西兰。根据 SendRight 标准认证邮件。 - 美国。根据编码准确性支持系统 (Coding Accuracy Support System ,CASS) 认证邮件。
建议列表	用于查找不完整的地址记录的有效备用版本。在配置地址验证映射来实时处理单个地址记录时，使用建议列表数据。
地理编码	用于向地址记录中添加纬度坐标和经度坐标。
补充	用于标识包含某个地址的地理区域或人口区域。 地址验证器转换可以向地址记录中添加以下国家/地区的补充数据： <ul style="list-style-type: none"> - 日本。标识包含该地址的市区。 - 塞尔维亚共和国。标识该地址的街道级别邮政编码。 - 英国。在地址中添加收件人地点标识符。 - 美国。标识包含该地址的人口中心。

您也可以使用 Data Quality 内容安装程序安装地址引用数据。请在 PowerCenter 集成服务计算机上运行服务器内容安装程序来安装这些文件。

Data Quality 内容安装程序将在 PowerCenter 目录结构中创建一个 /av/ 目录，然后将地址引用数据写入该目录。PowerCenter 集成服务会从 AD50.cfg 配置文件中读取地址引用数据的路径。

PowerCenter 安装程序会将 AD50.cfg 文件写入以下路径：

[PowerCenter_Installation]/server/bin

安装地址引用数据时，请验证是否已在 AD50.cfg 中设置了以下属性：

引用数据位置

在 ReferenceDataLocation 属性中输入地址引用数据的路径。

许可证密钥数据

在 LicenseKey 属性中输入许可证数据。如果有多个许可证，请以逗号分隔的字符串形式输入每个许可证。

预加载值

输入至少一个国家/地区的缩写作为映射所读取的每个类型的引用数据的预加载值。输入“全部”将为所有国家/地区应用一个预加载设置。

集成服务可以使用不同方式加载每个国家/地区的数据。例如，您可以指定对美国批处理/交互数据进行全部预加载，对英国批处理/交互数据进行部分预加载。集成服务还可以为每个类型的数据使用不同的预加载方式。例如，您可以指定对美国批处理/交互数据进行全部预加载，对美国地址编码数据进行部分预加载。

全部预加载设置会取代部分预加载设置，部分预加载设置会取代那些指示无数据预加载的设置。例如，如果您为无数据预加载设置输入“全部”，为全部预加载设置输入“USA”，集成服务会将全部美国数据加载到内存，而不加载其他任何国家/地区的数据。如果没有预加载要求，则对于计划使用的任何类型的引用数据，请为无数据预加载设置输入“全部”。

不能为补充数据指定预加载值。

您可以有选择地编辑 AD50.cfg 中的其他属性。

注意: Informatica Data Quality 9.x 不读取 AD50.cfg 文件。Informatica Data Quality 9.x 的用户通过 Administrator 工具输入引用数据位置、许可证密钥信息和预加载值。

标识填充数据的要求

执行标识匹配分析的映射会读取被称为标识填充的引用数据文件。您需要从 Informatica 购买并下载标识填充文件。将填充文件复制到 PowerCenter 集成服务计算机中或 PowerCenter 集成服务可以访问的计算机上的共享目录中。

您也可以使用 Data Quality 内容安装程序安装填充文件。请在 PowerCenter 集成服务计算机上运行服务器内容安装程序来安装这些文件。

您必须将填充文件复制或安装到名为 default 的目录。

PowerCenter 集成服务可以从 IDQTx.cfg 配置文件或 SSAPR 环境变量中读取填充文件目录的位置。PowerCenter 集成服务会在 IDQTx.cfg 文件中查找位置。如果 IDQTx.cfg 文件不存在或未指定位置，PowerCenter 集成服务将在 SSAPR 环境变量中查找位置。

注意: IDQTx.cfg 文件和 SSAPR 环境变量指定了到 default 目录的父目录的路径。路径不包括目录名称。

PowerCenter 安装程序将 IDQTx.cfg 文件写入以下路径中：

```
[PowerCenter_Installation]/server/bin
```

IDQTx.cfg 文件使用 *IdentityReferenceDataLocation* 属性指定填充文件目录的位置。请使用目录的当前位置更新该属性。

例如，您可以在 *IdentityReferenceDataLocation* 属性中设置以下路径：

```
[PowerCenter_Installation]/services/DQContent/INFA_Content/identity/
```

您可以在 *IdentityReferenceDataLocation* 属性中设置默认目录的绝对路径或相对路径。如果设置绝对路径，应指向 PowerCenter 集成服务计算机中的目录或该服务可访问的计算机中的共享目录。如果设置相对路径，应指向 PowerCenter 集成服务计算机中的目录。

PowerCenter 集成服务从以下目录读取相对路径：

```
[PowerCenter_Installation]/server/bin
```

例如，要指定一个到 bin 目录中某个填充文件目录的相对路径，请设置以下相对路径：

```
./
```

IDQTx.cfg 的属性

在 IDQTx.cfg 文件中设置 *IdentityReferenceDataLocation* 属性时，您还可以设置以下属性：

IdentityCacheDir

用于存储运行映射时所创建的临时数据文件的目录的路径。该路径标识的是一个父目录。PowerCenter 集成服务会将临时文件写入指定位置下的目录。

该属性具有以下默认值：

```
./identityCache
```

IdentityIndexDir

包含运行映射时所创建的临时索引文件的目录的路径。标识匹配分析在进行匹配分析之前，使用索引将记录排序成若干个组。该路径标识的是一个父目录。PowerCenter 集成服务会将索引文件写入您指定的位置下的目录。

该属性具有以下默认值：

./identityIndex

注意: Developer tool 用户可以在匹配转换中设置缓存目录路径和索引目录路径。如果匹配转换未指定路径，PowerCenter 集成服务将使用 IDQTx.cfg 文件中的路径。

概率模型和分类器模型的要求

概率模型和分类器模型是引用数据文件，它们标识了数据字符串中信息的类型。在运行读取概率模型或分类器模型文件的映射之前，您需要配置 PowerCenter 来读取这些文件。

请执行以下配置任务：

- 验证 PowerCenter 目录结构中的模型文件位置。
- 更新 PowerCenter 集成服务的 Java 内存分配。

概率模型和分类器模型文件位置

在导出读取概率模型或分类器模型数据的映射时，您需要为导出的模型数据文件指定一个目录位置。您可以将该文件导出到 PowerCenter 目录架构，或将文件导出到文件系统，然后复制到 PowerCenter 服务计算机。

默认情况下，PowerCenter 集成服务从以下目录读取模型文件：

[PowerCenter_Installation]/server/bin

如果文件不在默认位置，PowerCenter 集成服务将从属性文件读取文件位置。您需要使用模型文件的路径更新属性文件。

下表说明了这些属性文件：

文件名	说明
CLASSIFIER.properties	包含 <i>ClassifierFileLocation</i> 属性。更新该属性以标识分类器模型文件的路径。
NER.properties	包含 <i>NERFileLocation</i> 属性。更新该属性以标识概率模型文件的路径。

PowerCenter 集成服务从以下目录读取这些属性文件：

[PowerCenter_Installation]/server/bin

您可以在每个属性文件中设置绝对路径或相对路径来标识模型文件的位置。如果设置相对路径，PowerCenter 集成服务会将路径添加到以下目录：

[PowerCenter_Installation]/server/bin

注意: 请使用正斜杠定义模型文件在所有平台上的目录路径。

PowerCenter 集成服务内存分配

PowerCenter 集成服务使用大量内存读取分类器模型文件或概率模型文件。如果 PowerCenter 集成服务不能获取需要的内存，将无法读取模型文件，也无法处理数据。

因此，请检查 PowerCenter 集成服务的内存分配情况。如果您运行的会话需要引用分类器模型文件或概率模型文件，请将内存分配至少设置为 512 MB。

您可以使用 *Java SDK 内存上限* 进程属性来分配内存。要检查或更新该属性，请登录到 Administrator 工具，然后选择运行会话任务的 PowerCenter 集成服务。您将在**进程**视图找到 *Java SDK 内存上限* 属性。

索引

D

Data Quality 集成 [6](#)

I

Informatica Data Quality [6](#)