



Informatica®

10.1.1 HotFix 1

引用数据指南

Informatica 引用数据指南

10.1.1 HotFix 1

2017 年 6 月

© 版权所有 Informatica LLC 2013, 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 Big Data Management 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/或文档的某些部分受第三方版权制约，包括但不限于：版权所有 DataDirect Technologies。保留所有权利。版权所有 (C) Sun Microsystems。保留所有权利。版权所有 (C) RSA Security Inc. 保留所有权利。版权所有 (C) Ordinal Technology Corp. 保留所有权利。版权所有 (C) Aandacht c.v. 保留所有权利。版权所有 Genivia, Inc. 保留所有权利。版权所有 Isomorphic Software。保留所有权利。版权所有 (C) Meta Integration Technology, Inc. 保留所有权利。版权所有 (C) Intalio。保留所有权利。版权所有 (C) Oracle。保留所有权利。版权所有 (C) Adobe Systems Incorporated。保留所有权利。版权所有 (C) DataArt, Inc. 保留所有权利。版权所有 (C) ComponentSource。保留所有权利。版权所有 (C) Microsoft Corporation。保留所有权利。版权所有 (C) Rogue Wave Software, Inc. 保留所有权利。版权所有 (C) Teradata Corporation。保留所有权利。版权所有 (C) Yahoo! Inc. 保留所有权利。版权所有 (C) Glyph & Cog, LLC。保留所有权利。版权所有 (C) Thinkmap, Inc. 保留所有权利。版权所有 (C) Clearpace Software Limited。保留所有权利。版权所有 (C) Information Builders, Inc. 保留所有权利。版权所有 (C) OSS Nokalva, Inc. 保留所有权利。版权所有 Edifecs, Inc. 保留所有权利。版权所有 Cleo Communications, Inc. 保留所有权利。版权所有 (C) International Organization for Standardization 1986。保留所有权利。版权所有 (C) ej-technologies GmbH。保留所有权利。版权所有 (C) Jaspersoft Corporation。保留所有权利。版权所有 (C) International Business Machines Corporation。保留所有权利。版权所有 (C) yWorks GmbH。保留所有权利。版权所有 (C) Lucent Technologies。保留所有权利。版权所有 (C) University of Toronto。保留所有权利。版权所有 (C) Daniel Veillard。保留所有权利。版权所有 (C) Unicode, Inc. 版权所有 IBM Corp. 保留所有权利。版权所有 (C) MicroQuill Software Publishing, Inc. 保留所有权利。版权所有 (C) PassMark Software Pty Ltd. 保留所有权利。版权所有 (C) LogiXML, Inc. 保留所有权利。版权所有 (C) 2003-2010 Lorenzi Davide。保留所有权利。版权所有 (C) Red Hat, Inc. 保留所有权利。版权所有 (C) The Board of Trustees of the Leland Stanford Junior University。保留所有权利。版权所有 (C) EMC Corporation。保留所有权利。版权所有 (C) Flexera Software。保留所有权利。版权所有 (C) Jinfonet Software。保留所有权利。版权所有 (C) Apple Inc. 保留所有权利。版权所有 (C) Telerik Inc. 保留所有权利。版权所有 (C) BEA Systems。保留所有权利。版权所有 (C) PDFlib GmbH。保留所有权利。版权所有 (C) Orientation in Objects GmbH。保留所有权利。版权所有 (C) Tanuki Software, Ltd. 保留所有权利。版权所有 (C) Ricebridge。保留所有权利。版权所有 (C) Sencha, Inc. 保留所有权利。版权所有 (C) Scalable Systems, Inc. 保留所有权利。版权所有 (C) jQWidgets。保留所有权利。版权所有 (C) Tableau Software, Inc. 保留所有权利。版权所有 (C) MaxMind, Inc. 保留所有权利。版权所有 (C) TMate Software s.r.o. 保留所有权利。版权所有 (C) MapR Technologies Inc. 保留所有权利。版权所有 (C) Amazon Corporate LLC。保留所有权利。版权所有 (C) Highsoft。保留所有权利。版权所有 (C) Python Software Foundation。保留所有权利。版权所有 (C) BeOpen.com。保留所有权利。版权所有 (C) CNRI。保留所有权利。

本产品包括由 Apache Software Foundation (<http://www.apache.org/>) 开发的软件和/或在不同 Apache 许可证版本（以下简称“许可证”）下许可的其他软件。您可从 <http://www.apache.org/licenses/> 获取这些许可证的副本。除非适用法律要求或者有相应书面协议，否则依据这些“许可证”分发的软件以“原样”提供，不附带任何明示或暗示的担保或条件。请参阅“许可证”中规定的具体语言管理权限和限制。

本产品包括由 Mozilla (<http://www.mozilla.org/>) 开发的软件、由 JBoss Group, LLC 开发的软件（版权所有 JBoss Group, LLC 保留所有权利）、由 Bruno Lowagie 和 Paulo Soares 开发的软件（版权所有 (C) 1999-2006 Bruno Lowagie 和 Paulo Soares）以及在 <http://www.gnu.org/licenses/lgpl.html> 网站上的不同版本 GNU Lesser General 公共许可协议下许可的软件。这些材料由 Informatica 按“原样”免费提供，不附带任何明示或暗示的担保，包括但不限于适销性和特定用途适用性的暗示担保。

本产品包括 ACE(TM) 和 TAO(TM) 软件，这些软件版权归 Douglas C. Schmidt 及其在华盛顿大学、加利福尼亚大学欧芬分校以及范德堡大学的研发团队所有（版权所有 (C) 1993-2006，保留所有权利）。

本产品包括由 OpenSSL Project 开发并在 OpenSSL Toolkit（版权所有 OpenSSL Project。保留所有权利）中使用的软件，该软件的再分发受 <http://www.openssl.org> 和 <http://www.openssl.org/source/license.html> 上规定条款之制约。

本产品包括 Curl 软件，版权所有 1996-2013，Daniel Stenberg <daniel@haxx.se>。保留所有权利。有关该软件的权限和限制受 <http://curl.haxx.se/docs/copyright.html> 上规定条款之制约。允许出于任何目的以免费或收费形式使用、复制、修改和分发该软件，但前提是所有副本均应注明上述版权声明以及本许可声明。

本产品包括由 MetaStuff, Ltd. 开发的软件，版权所有 2001-2005 ((C)) MetaStuff, Ltd. 保留所有权利。有关该软件的权限和限制受 <http://www.dom4j.org/license.html> 上规定条款之制约。

本产品包括由 Dojo Foundation 开发的软件，版权所有 (C) 2004-2007, Dojo Foundation。保留所有权利。有关该软件的权限和限制受 <http://dojotoolkit.org/license> 上规定条款之制约。

本产品包括 ICU 软件，版权所有 International Business Machines Corporation 和其他方。保留所有权利。有关该软件的权限和限制受 <http://source.icu-project.org/repos/icu/icu/trunk/license.html> 上规定条款之制约。

本产品包括由 Per Bothner 开发的软件，版权所有 (C) 1996-2006 Per Bothner。保留所有权利。<http://www.gnu.org/software/kawa/Software-License.html> 上的许可证中规定了您使用这些材料的权利。

本产品包括 OSSP UUID 软件，版权所有 (C) 2002 Ralf S. Engelschall，版权所有 (C) 2002 OSSP Project，版权所有 (C) 2002 Cable & Wireless Deutschland。有关该软件的权限和限制受 <http://www.opensource.org/licenses/mit-license.php> 上规定条款之制约。

本产品包括由 Boost (<http://www.boost.org/>) 开发的软件或在 Boost 软件许可证下许可的软件。有关该软件的权限和限制受 http://www.boost.org/LICENSE_1_0.txt 上规定条款之制约。

本产品包括由 University of Cambridge 开发的软件，版权所有 (C) 1997-2007 University of Cambridge。有关该软件的权限和限制受 <http://www.pcre.org/license.txt> 上规定条款之制约。

本产品包括由 The Eclipse Foundation 开发的软件，版权所有 (C) 2007 The Eclipse Foundation。保留所有权利。有关该软件的权限和限制受 <http://www.eclipse.org/org/documents/epl-v10.php> 和 <http://www.eclipse.org/org/documents/edl-v10.php> 上规定条款之制约。

本产品包括在 <http://www.tcl.tk/software/tcltk/license.html>、<http://www.bosrup.com/web/overlib/?License>、<http://www.stlport.org/doc/license.html>、<http://asm.ow2.org/license.html>、<http://www.cryptix.org/LICENSE.TXT>、<http://hsqldb.org/web/hsqldbLicense.html>、<http://httpunit.sourceforge.net/doc/license.html>、<http://jung.sourceforge.net/license.txt>、http://www.gzip.org/zlib/zlib_license.html、<http://www.openldap.org/software/release/license.html>、<http://www.libssh2.org>、<http://slf4j.org/license.html>、<http://www.sente.ch/software/OpenSourceLicense.html>、<http://fusesource.com/downloads/license-agreements/fuse-message-broker-v-5-3-license-agreement>、<http://antlr.org/license.html>、<http://aopalliance.sourceforge.net/>、<http://www.bouncycastle.org/licence.html>、<http://www.jgraph.com/jgraphdownload.html>、<http://www.jcraft.com/jsch/LICENSE.txt>、http://jotm.objectweb.org/bsd_license.html、<http://www.w3.org/Consortium/Legal/2002/copyright-software-20021231>、<http://www.slf4j.org/license.html>、<http://nanoxml.sourceforge.net/orig/copyright.html>、<http://www.json.org/license.html>、<http://forge.ow2.org/projects/jaservice/>、<http://www.postgresql.org/about/licence.html>、<http://www.sqlite.org/copyright.html>、<http://www.tcl.tk/software/tcltk/license.html>、<http://www.jaxen.org/faq.html>、<http://www.jdom.org/docs/faq.html>、<http://www.slf4j.org/license.html>、<http://www.iodbc.org/dataspace/iodbc/wiki/IODBC/License>、<http://www.keplerproject.org/md5/license.html>、<http://www.toedter.com/en/jcalendar/license.html>、<http://www.edankert.com/bounce/index.html>、<http://>

www.net-snmp.org/about/license.html、<http://www.openmdx.org/#FAQ>、http://www.php.net/license/3_01.txt、<http://srp.stanford.edu/license.txt>、<http://www.schneier.com/blowfish.html>、<http://www.jmock.org/license.html>、<http://xsom.java.net>、<http://benalman.com/about/license/>、<https://github.com/CreateJS/EaselJS/blob/master/src/easeljs/display/Bitmap.js>、<http://www.h2database.com/html/license.html#summary>、<http://jsoncpp.sourceforge.net/LICENSE>、<http://jdbc.postgresql.org/license.html>、<http://protobuf.googlecode.com/svn/trunk/src/google/protobuf/descriptor.proto>、<https://github.com/rantav/hector/blob/master/LICENSE>、<http://web.mit.edu/Kerberos/krb5-current/doc/mitK5license.html>、<http://jibx.sourceforge.net/jibx-license.html>、<https://github.com/lyokato/libgeohash/blob/master/LICENSE>、<https://github.com/hjiang/jsonxx/blob/master/LICENSE>、<https://code.google.com/p/lz4/>、<https://github.com/jedisct1/libsodium/blob/master/LICENSE>、<http://one-jar.sourceforge.net/index.php?page=documents&file=license>、<https://github.com/EsotericSoftware/kryo/blob/master/license.txt>、<http://www.scala-lang.org/license.html>、<https://github.com/tinkerpop/blueprints/blob/master/LICENSE.txt>、<http://gee.cs.oswego.edu/dl/classes/EDU/oswego/cs/dl/util/concurrent/intro.html>、<https://aws.amazon.com/asl/>、<https://github.com/twbs/bootstrap/blob/master/LICENSE> 和 <https://sourceforge.net/p/xmlunit/code/HEAD/tree/trunk/LICENSE.txt> 下许可的软件。

本产品包括在 Academic 免费许可证 (<http://www.opensource.org/licenses/afl-3.0.php>)、通用开发和分发许可证 (<http://www.opensource.org/licenses/cddl1.php>)、通用公共许可证 (<http://www.opensource.org/licenses/cpl1.0.php>)、Sun Binary Code 许可协议补充许可条款、BSD 许可证 (<http://www.opensource.org/licenses/bsd-license.php>)、新 BSD 许可证 (<http://opensource.org/licenses/BSD-3-Clause>)、MIT 许可证 (<http://www.opensource.org/licenses/mit-license.php>)、Artistic 许可证 (<http://www.opensource.org/licenses/artistic-license-1.0>) 以及原始开发者公共许可证版本 1.0 (<http://www.firebirdsql.org/en/initial-developer-s-public-license-version-1-0/>) 下许可的软件。

本产品包括由 Joe Walnes 和 XStream Committers 开发的软件，版权所有 (C) 2003-2006 Joe Walnes, 2006-2007 XStream Committers。保留所有权利。有关该软件的权限和限制受 <http://xstream.codehaus.org/license.html> 上规定条款之制约。本产品包括由 Indiana University Extreme! Lab 开发的软件。有关详细信息，请访问 <http://www.extreme.indiana.edu/>。

本产品包括软件版权所有 (c) 2013 Frank Balluffi 和 Markus Moeller。保留所有权利。有关此软件的权限和限制受 MIT 许可证上规定条款之制约。

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。如果您发现本文档中存在任何问题，请以书面形式将问题报告给我们，邮寄地址是 Informatica LLC 2100 Seaport Blvd. Redwood City, CA 94063。

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: 2018-05-16

目录

| | |
|-------------------------------|----|
| 前言 | 8 |
| Informatica 资源 | 8 |
| Informatica Network | 8 |
| Informatica 知识库 | 8 |
| Informatica 文档 | 8 |
| Informatica 产品可用性矩阵 | 9 |
| Informatica Velocity | 9 |
| Informatica Marketplace | 9 |
| Informatica 全球客户支持部门 | 9 |
| 第 1 章：引用数据简介 | 10 |
| 引用数据概览 | 10 |
| Informatica 引用数据 | 11 |
| 用户定义的引用数据 | 11 |
| 引用表 | 12 |
| 引用表结构 | 12 |
| 引用数据仓库的特权 | 12 |
| 参数和引用表 | 12 |
| 引用数据对象和版本控制 | 13 |
| 第 2 章：Analyst 工具中的引用表 | 14 |
| Analyst 工具引用表概览 | 14 |
| 引用表属性 | 14 |
| 引用表常规属性 | 15 |
| 引用表列属性 | 15 |
| 在引用表编辑器中创建引用表 | 16 |
| 从配置文件数据创建引用表 | 16 |
| 从配置文件列数据创建引用表 | 16 |
| 从值模式创建引用表 | 18 |
| 从平面文件创建引用表 | 19 |
| Analyst 工具平面文件属性 | 19 |
| 从平面文件创建引用表 | 19 |
| 从数据库表创建引用表 | 20 |
| 从数据库表创建引用表 | 20 |
| 在已添加版本的模型存储库中处理引用表 | 21 |
| 引用表更新 | 21 |
| 管理列 | 22 |
| 管理行 | 22 |
| 查找和替换值 | 23 |
| 导出引用表数据 | 23 |

| | |
|---|-----------|
| 启用和禁用在非受管引用表中进行编辑. | 24 |
| 刷新引用表值. | 24 |
| 审计跟踪事件. | 25 |
| 查看审计跟踪事件. | 25 |
| 引用表的规则和准则. | 25 |
| 第 3 章： Developer 工具中的引用数据. | 27 |
| Developer tool 引用数据概览. | 27 |
| 引用数据和转换. | 27 |
| 在已添加版本的模型存储库中处理引用数据对象. | 28 |
| 签出引用数据对象. | 28 |
| 签入引用数据对象. | 28 |
| 引用表. | 29 |
| 引用表数据属性. | 29 |
| 创建引用表对象. | 30 |
| 从平面文件创建引用表. | 31 |
| 从关系源创建引用表. | 32 |
| 内容集. | 33 |
| 字符集. | 33 |
| 分类器模型. | 34 |
| 模式集. | 34 |
| 概率模型. | 34 |
| 正则表达式. | 35 |
| 标志集. | 35 |
| 概率模型和分类器模型的规则和准则. | 37 |
| 创建内容集. | 37 |
| 在内容集中创建引用数据对象. | 38 |
| 第 4 章： 分类器模型. | 39 |
| 分类器模型概览. | 39 |
| 分类器模型结构. | 40 |
| 分类器得分. | 40 |
| 分类器转换示例. | 40 |
| 分类器模型选项. | 41 |
| 分类器模型引用数据. | 42 |
| 分类器模型标签数据. | 42 |
| 分类器模型标签管理. | 43 |
| 分类器模型配置. | 44 |
| 创建分类器模型. | 44 |
| 将数据源中的数据附加到分类器模型. | 45 |
| 将引用数据行添加到分类器模型. | 45 |
| 将标签添加到分类器模型. | 45 |
| 为引用数据行分配标签. | 46 |

| | |
|-----------------------------|-----------|
| 识别未使用的标签值. | 46 |
| 从分类器模型中删除行. | 46 |
| 从分类器模型中删除标签. | 47 |
| 编译分类器模型. | 47 |
| 筛选操作和查找操作. | 47 |
| 使用数据值筛选引用数据行. | 47 |
| 使用标签值筛选引用数据行. | 47 |
| 在引用数据行中查找值. | 48 |
| 复制和粘贴操作. | 48 |
| 将分类器模型复制到其他内容集. | 48 |
| 从其他内容集导入分类器模型. | 48 |
| 第 5 章： 概率模型. | 50 |
| 概率模型概览. | 50 |
| 概率模型结构. | 51 |
| 标签创建器转换示例. | 51 |
| 解析器转换示例. | 52 |
| 概率模型选项. | 52 |
| 概率模型“数据”视图. | 53 |
| 概率模型“标签”视图. | 54 |
| 概率模型引用数据. | 55 |
| 概率模型标签数据. | 55 |
| 溢出标签. | 56 |
| 概率模型属性. | 56 |
| 概率模型配置. | 57 |
| 创建空的概率模型. | 57 |
| 从数据对象创建概率模型. | 58 |
| 将数据源中的数据附加到概率模型. | 58 |
| 向概率模型添加引用数据行. | 59 |
| 将标签添加到概率模型. | 59 |
| 为引用数据值分配标签. | 60 |
| 为多个数据值分配标签. | 60 |
| 从概率模型中删除行. | 61 |
| 从概率模型中删除标签. | 61 |
| 编译概率模型. | 61 |
| 在概率模型中查找数据行. | 61 |
| 按标签分配筛选引用数据值. | 62 |
| 查找未使用的标签值. | 62 |
| 复制和粘贴操作. | 62 |
| 将概率模型复制到其他内容集. | 62 |
| 从其他内容集导入概率模型. | 63 |
| 将引用数据行复制到剪贴板. | 63 |

| | |
|--|----|
| 附录 A: 引用数据和 Informatica Big Data Management..... | 64 |
| 引用数据和 Informatica Big Data Management 概览. | 64 |
| 用于地址验证的引用数据. | 64 |
| 安装地址引用数据文件. | 65 |
| 索引..... | 66 |

前言

《Informatica 引用数据指南》包含可在 Informatica Developer 和 Informatica Analyst 中使用的引用数据对象和文件的相关信息。该指南面向的读者是数据分析人员、数据管理者以及想要使用引用数据来验证并提高组织数据的准确性和可用性的任何人。

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

引用数据简介

本章包括以下主题：

- [引用数据概览, 10](#)
- [Informatica 引用数据, 11](#)
- [用户定义的引用数据, 11](#)
- [引用表, 12](#)
- [引用数据对象和版本控制, 13](#)

引用数据概览

Informatica 转换可以使用引用数据来分析和更新数据。您可以在 Developer tool 和 Analyst 工具中创建引用数据对象。还可以将引用数据对象和文件导入模型存储库和文件系统。您可以使用 Data Quality 内容安装程序导入引用数据对象和安装引用数据文件。

可以创建和编辑以下类型的引用数据：

引用表

引用表包含一组数据值的标准版本和替代版本。您将引用表添加到 Developer tool 中的某个转换，以验证源数据值是否准确并且格式正确。

大多数引用表至少包含两列。一列包含标准或首选版本的值，其他列包含替代版本。将引用表添加到转换时，该转换将在输入端口数据中搜索同时显示在引用表中的值。您可以使用任何对于所处理的数据项目有用的数据来创建表。

内容集

内容集是在模型存储库或文件中指定引用数据值的模型存储库对象。将内容集添加到某个转换中时，该转换在输入数据中搜索与内容集中的数据模式匹配的值。

Data Quality 内容安装程序可以安装以下类型的引用数据：

Informatica 引用表

Informatica 开发的存储库对象和数据文件。将加速器对象导入到模型存储库中时，将导入 Informatica 引用表。引用信息的类型包括电话区号、邮政编码格式、名字、职业和首字母缩略词。您可以编辑 Informatica 引用表。

Informatica 内容集

Informatica 开发的存储库对象和数据文件。将加速器对象导入到模型存储库中时，将导入内容集。内容集包含不同类型的引用数据，您可以将其用于对数据质量转换执行搜索操作。

地址引用数据文件

包含某个国家/地区的可投递地址数据的引用数据文件。地址验证器转换将读取该引用数据。无法创建或编辑地址引用数据文件。

地址引用数据在限定的一段时间内是最新的，因此您必须定期刷新数据（例如每个季度刷新一次）。

标识填充文件

包含有关个人、家庭和公司标识的信息的引用数据文件。匹配转换和比较转换使用填充文件在输入数据中查找潜在标识。您不能创建或编辑标识填充文件。

Informatica 引用数据

可以从 Informatica 购买并下载地址引用数据和标识填充数据。

可以购买某个国家/地区的地址数据年度订阅，可以在订阅期间随时从 Informatica 下载最新的地址数据。

内容安装程序用户从应用程序单独下载和安装引用数据。有关您的系统上安装的引用数据，请联系用户管理员。

用户定义的引用数据

可以使用数据对象中的值创建引用数据对象。

例如，可以选择包含特定于某个项目或组织的值的数据对象或配置文件列。从列值创建自定义引用数据对象。

可以从数据列构建引用数据对象以确认以下内容：

- 列中的数据行包含同一类型的信息。
- 源值有效。引用对象可能包含一列有效值，引用对象也可能包含一列无效值。

下表列出了可以包含引用数据的项目数据列的常见示例：

| 信息 | 引用数据示例 |
|---------------|--|
| 库存单位 (SKU) 代码 | 使用 SKU 列创建组织的有效 SKU 代码的引用表。使用引用表在数据集中查找正确或错误的 SKU 代码。 |
| 员工代码 | 使用员工代码或员工 ID 列创建有效员工代码的引用表。使用引用表在员工数据中查找错误。 |
| 客户帐号 | 对客户帐户列运行配置文件，以标识帐号模式。使用配置文件创建错误数据模式的标志集。使用该标志集查找不符合正确的帐号结构的帐号。 |
| 客户名称 | 当客户名称列包含名、中间名和姓时，可以创建用于定义列中字符串所需的结构的概率模型。使用概率模型查找不属于该列的数据字符串。 |

引用表

在 Analyst 工具和 Developer tool 中创建和更新引用表。

引用表在模型存储库中存储元数据。引用表可以在引用数据仓库或其他数据库存储列数据。当由引用数据仓库存储列数据时，Informatica 服务会将表标识为受管引用表。当由其他数据库存储列数据时，Informatica 服务会将表标识为非受管引用表。

内容管理服务存储引用数据仓库的数据库连接。您可以指定 IBM DB2 数据库、Microsoft SQL Server 数据库或 Oracle 数据库作为引用数据仓库。

当您从另一个数据库导入引用数据仓库时，请使用本地连接或 ODBC 连接来导入数据。当您指定非受管数据库作为引用表的数据源时，请使用本地连接来连接到数据库。

引用表结构

大多数引用表至少包含两个列。一个列包含数据值的正确或所需的版本。其他列包含这些值的其他版本，包括可能显示在源数据中的替代版本。

包含正确或所需值的列称为有效列。当某个转换读取映射中的引用表时，该转换在无效列中查找值。当该转换找到无效值时，将返回有效列中的对应值。还可以将转换配置为返回单个通用值，而不是有效值。

有效列可以包含形式正确的数据，如邮政编码。有效列可以包含与项目相关的数据，如某个组织特有的库存单位 (SKU) 编号。还可以从错误数据（如包含您要搜索的已知数据错误的值）创建有效列。

例如，创建包含一列某零售组织中有效 SKU 编号的引用表。将引用表添加到某个标签创建器转换并使用该转换创建映射。使用产品数据库表运行映射。当该映射运行时，标签创建器创建一个用于标识不包含有效 SKU 编号的产品记录。

引用表和解析器转换

创建单列引用表，以便在基于模式的解析操作中使用表数据。您配置解析器转换来执行基于模式的解析，并将引用数据导入转换配置。

引用数据仓库的特权

内容管理服务使用特权来限制用户对引用表的操作。在 Administrator 工具中使用“安全”选项来查看或更新服务特权。

要使用引用表，您必须在内容管理服务中拥有以下特权：

- 创建引用表
- 编辑引用表数据
- 编辑引用表元数据

要在非受管引用表中编辑数据，还需要确认您已将引用表对象配置为允许编辑。

注意：如果您编辑了数据库应用程序中非受管引用表的元数据，请使用 Analyst 工具将模型存储库与该表同步。在 Developer tool 中使用非受管引用表之前，必须先同步模型存储库与该表。

参数和引用表

您可以使用参数来标识模型存储库中的引用表。您可以在 Developer tool 中创建参数以用于标识引用表，或者也可以将引用表位置添加到参数文件。

如果在 Developer tool 中创建参数，则您可以将其添加到映射中的转换。如果将引用表位置添加到参数文件，则您可以在命令提示符中运行映射时指定文件。在每种情况下，数据集成服务都会在您运行映射时读取该参数所标识的引用表。

您可以将标识引用表的参数添加到以下转换：

- 大小写转换器转换
- 标签创建器转换
- 标志解析模式的解析器转换
- 标准创建器转换

注意：使用 *infacmd ms runMapping* 命令可在命令提示符中运行映射。

引用数据对象和版本控制

如果存储引用数据对象的模型存储库与版本控制应用程序集成，则可以向对象应用版本控制。可以将版本控制应用于引用表和内容集。

可以从支持版本控制的模型存储库签入和签出引用数据对象。可以撤销签出、检索对象的早期版本以及将对象还原到早期版本。引用数据对象未应用版本控制时，模型存储库会锁定您所编辑的引用数据对象。其他用户无法编辑您正在处理的锁定对象。您关闭对象时，模型存储库会释放锁，其他用户即可编辑此对象。

注意：版本控制适用于模型存储库为非受管引用表对象存储的元数据。版本控制不适用于非受管引用表中的数据。无法查看或还原早期版本的非受管引用表中的引用数据。

第 2 章

Analyst 工具中的引用表

本章包括以下主题：

- [Analyst 工具引用表概览, 14](#)
- [引用表属性, 14](#)
- [在引用表编辑器中创建引用表, 16](#)
- [从配置文件数据创建引用表, 16](#)
- [从平面文件创建引用表, 19](#)
- [从数据库表创建引用表, 20](#)
- [在已添加版本的模型存储库中处理引用表, 21](#)
- [引用表更新, 21](#)
- [审计跟踪事件, 25](#)
- [引用表的规则和准则, 25](#)

Analyst 工具引用表概览

在 Analyst 工具的设计工作区中创建引用表。

可以从平面文件、模型存储库中的数据源和其他数据库中的表创建引用表。

可以从配置文件列或配置文件列中数据的子集创建引用表。还可以从您从配置文件中选择的列模式创建引用表。

创建或更新引用表时，将配置该表及其包含的数据列上的属性。

引用表属性

可以在 Analyst 工具中查看和更新引用表属性。引用表显示常规属性和列属性。常规属性包括引用表名称、创建日期、数据库连接名称和有效列名称。列属性包括列名称、精度值和小数位数值。

可以在只读模式下查看属性。要更新属性，请编辑或签出引用表。

引用表常规属性

常规属性包含引用表对象的有关信息。

下表介绍常规属性：

| 属性 | 说明 |
|--------|---------------------------|
| 名称 | 引用表名称。 |
| 说明 | 用户为引用表输入的说明。 |
| 位置 | 引用表对象在模型存储库中的位置。 |
| 有效列 | 引用表中有效列的名称。 |
| 创建日期 | 引用表名称的创建日期和时间。 |
| 创建者 | 创建引用表的用户的登录名。 |
| 上次修改时间 | 最后一次更新引用表的日期和时间。 |
| 上次修改者 | 最后一次执行更新操作的用户的登录名。 |
| 连接名称 | 存储引用数据值的数据库的连接名称。 |
| 类型 | 引用表的类型。引用表可以处于受管状态或非受管状态。 |

引用表列属性

列属性包含有关列元数据的信息。

下表介绍了列属性：

| 属性 | 说明 |
|------|--|
| 名称 | 列名称。 |
| 数据类型 | 每个列中数据的数据类型。可以选择以下数据类型之一： <ul style="list-style-type: none">- 长整型- 日期/时间- 小数- 双精度型- 整型- 字符串 创建空引用表或从平面文件创建引用表时，无法选择双精度数据类型。 |
| 精度 | 每个列的精度。精度是列可以容纳的最大位数或最大字符数。 您配置的精度值取决于数据类型。 |
| 小数位数 | 每个列的小数位数。小数位数是列可以在小数点右侧容纳的最大位数。适用于小数列。 您配置的小数位数取决于数据类型。 |
| 说明 | 每个列的可选说明。 |

| 属性 | 说明 |
|----|---|
| 可空 | 指示列是否可以包含空值。 |
| 键 | 标识键列。如果您从指定了键列的表中导入引用数据，Analyst 工具可以确定键列。 |

在引用表编辑器中创建引用表

在引用表编辑器中定义表结构并将数据添加到引用表中。

1. 单击**新建 > 引用表**。
此时将打开**新建引用表**向导。
2. 选择**使用引用表编辑器**选项，然后单击**下一步**。
3. 使用**添加新列**选项，将列添加到表中。
4. 配置每个列的属性。
属性包括列名称、数据类型、精度和小数位数。
如果列包含转换可以在引用数据搜索中返回的数据，请选择“有效”选项。
5. （可选）添加列以包含低级别说明，作为引用表中的元数据。
6. （可选）输入表的审计说明。
审计说明显示在审计跟踪日志中。
7. 单击**下一步**。
8. 输入引用表的名称，在模型存储库中选择引用表对象的位置。
9. 单击**完成**。

从配置文件数据创建引用表

可以使用配置文件数据创建与配置文件中的源数据相关的引用表。使用引用表查找源数据中不同类型的信息。

可以使用配置文件通过以下方法创建或更新引用表：

- 在配置文件中选择一个列并将其添加到引用表。
- 浏览某个配置文件列并将该列的一个子集添加到引用表。
- 在该配置文件中选择一个列并将该列的模式值添加到引用表。

从配置文件列数据创建引用表

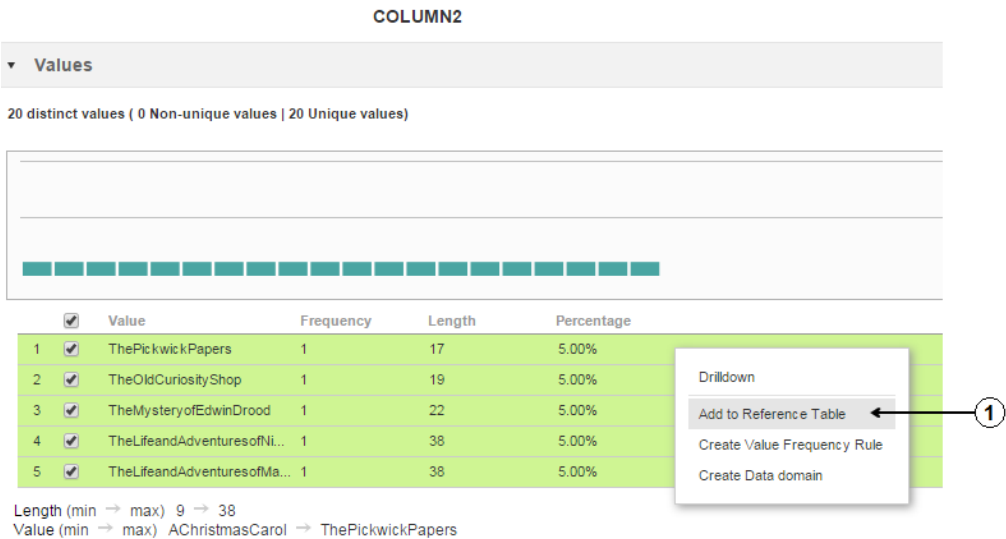
可以从配置文件数据列中的一个或多个值创建引用表。选择配置文件中的某个列，然后选择要添加到引用表中的列值。

1. 在 Analyst 工具中打开**库**工作区。
2. 选择**配置文件**资产类别。

库显示模型存储库中配置文件的列表。

- 3. 打开包含要添加到引用表中的列的配置文件。
配置文件概览会列出配置文件列名称。
- 4. 检查列数据。
要查看列数据，请单击列名称。
- 5. 在配置文件详细视图中，选择要添加到引用表的数据值。可以逐个选择值，也可以选择所有值。
- 6. 右键单击列名称，然后选择**添加到引用表**。

下图显示配置文件详细视图中的数据列：



数字 1 指出了图像中的**添加到引用表**选项。

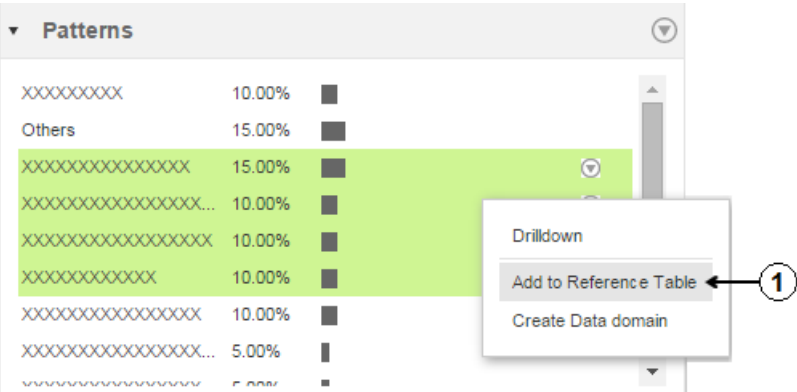
- 7. 此时将打开**添加到引用表**向导。
选择**创建引用表**选项。
注意：也可以选择将数据添加到当前引用表的选项。
- 8. 单击**下一步**。
列名称默认显示为引用表名称。（可选）更新名称。
- 9. （可选）输入说明和默认值。
Analyst 工具为任何不包含值的表记录使用该默认值。
- 10. 单击**下一步**。
- 11. 确认列属性。
（可选）选择为低级别描述性元数据创建列。
- 12. 单击**下一步**。
- 13. 检查引用表名称和说明。
（可选）输入审计说明。
- 14. 为引用表对象选择一个模型存储库位置。
- 15. 单击**完成**。

从值模式创建引用表

可以从配置文件列中的列模式创建引用表。模式表示一个或多个列字段中数据值的组成形式。选择配置文件中的某个列，然后选择要添加到所创建的引用表中的模式。

- 1. 在 Analyst 工具中打开库工作区。
- 2. 选择配置文件资产类别。
库显示模型存储库中配置文件的列表。
- 3. 打开包含要添加到引用表中的值模式的配置文件。
配置文件概览会列出配置文件列名称。
- 4. 选择用于定义要添加到引用表中的模式数据的列。
- 5. 查看列数据模式。
要查看列数据，请单击列名称。
- 6. 在配置文件详细视图中，选择要添加的列模式。
- 7. 右键单击所选择的模式，然后选择**添加到引用表**。

下图显示配置文件详细视图中某个列的数据模式：



数字 1 指出了图像中的**添加到引用表**选项。

- 8. 此时将打开**添加到引用表**向导。
选择**创建引用表**选项。
注意: 也可以选择将数据添加到当前引用表的选项。
- 9. 单击**下一步**。
列名称默认显示为引用表名称。（可选）更新名称。
- 10. （可选）输入说明和默认值。
Analyst 工具为任何不包含值的表记录使用该默认值。
- 11. 单击**下一步**。
- 12. 确认列属性。
（可选）选择为低级别描述性元数据创建列。
- 13. 单击**下一步**。
- 14. 检查引用表名称和说明。
（可选）输入审计说明。
- 15. 为引用表对象选择一个模型存储库位置。
- 16. 单击**完成**。

从平面文件创建引用表

可以从 CSV 文件导入引用数据。使用**新建引用表**向导导入文件数据。

必须为您用于创建引用表的每个平面文件配置属性。

Analyst 工具平面文件属性

将某个平面文件作为引用表导入时，必须为该文件中的每个列配置属性。您配置的选项决定 Analyst 工具从该文件读取数据的方式。

下表介绍了当您为引用表导入文件数据时可以配置的属性：

| 属性 | 说明 |
|-------|--|
| 分隔符 | 用于分隔数据列的字符。使用“其他”字段输入其他分隔符。 分隔符必须是可打印字符，并且必须不同于转义符和引号字符（如果已选择）。 不能选择非打印多字节字符作为分隔符。 |
| 文本限定符 | 定义文本字符串边界的引号字符。 选择“无引号”、“单引号”或“双引号”。 如果选择引号字符，向导将忽略引号对内的分隔符。 |
| 列名称 | 从第一行导入列名称。如果列名称显示在第一行中，请选择该选项。 向导在预览中使用第一行中的数据作为列名称。 默认情况下不启用。 |
| 值 | 选择从某一行开始进行值导入。当向导导入文件时，指示在预览中向导开始读取的行号。 |

从平面文件创建引用表

从平面文件创建引用表数据时，该表使用文件的列结构并导入文件数据。

1. 单击**新建 > 引用表**。

此时将显示**新建引用表**向导。

2. 选择**导入平面文件**选项。

3. 单击**下一步**。

4. 单击**选择文件**以选择平面文件。

5. 选择与平面文件中的数据相匹配的代码页。

6. 单击**上载**以上载文件数据。

7. 单击**下一步**。

8. 配置平面文件属性。

属性标识文件使用的分隔符以及文件的第一行是否包含列名称。

9. 要预览您配置的属性，请刷新**预览**窗格。

10. 单击**下一步**。

11. 配置每个列的属性。

属性包括列名称、数据类型、精度和小数位数。

如果列包含转换可以在引用数据搜索中返回的数据，请选择“有效”选项。

12. （可选）添加列以包含低级别说明，作为引用表中的元数据。
13. （可选）输入表的审计说明。
审计说明显示在审计跟踪日志中。
14. 单击**下一步**。
15. 输入引用表的名称，在模型存储库中选择引用表对象的位置。
16. （可选）输入表的说明。
17. 单击**完成**。

从数据库表创建引用表

当您从数据库表中创建引用表时，请在模型存储库中创建元数据对象。可以选择将表数据导入引用数据仓库。

创建受管引用表时，请将列数据导入引用数据仓库。创建非受管引用表时，请确定存储列数据的数据库表。您可以从 ODBC 连接或本地连接创建受管引用表。可以从本地连接创建非受管引用表。

创建引用表之前，请先验证 Informatica 域中是否存在与包含引用数据的数据库的连接。如果域不包含到该数据库的连接，则可以在 Analyst 工具中定义一个连接。

要定义数据库连接，请单击**管理 > 连接**。

从数据库表创建引用表

要创建引用表，请连接到数据库并选择含有引用数据的表。

1. 选择**新建 > 引用表**。
此时将显示**新建引用表**向导。
2. 选择**连接到关系表**选项。
要创建不在引用数据仓库中存储数据的引用表，请选择**非受管表**。
要使用户能够编辑非受管引用表，请选中**可编辑**选项。
单击**下一步**。
3. 从连接列表中选择数据库连接。
单击**下一步**。
4. 在**表面板**中，选择一个表。
5. 检查**属性**面板中的表属性。
（可选）单击**数据预览**查看表数据。
单击**下一步**。
6. 在**列属性**面板中，选择有效的列。
如果您创建受管引用表，则可以在**列属性**面板上执行以下操作：
 - 编辑引用表列名称。
 - 添加用于行级说明的列。
7. （可选）添加列以包含低级别说明，作为引用表中的元数据。
8. （可选）输入表的审计说明。

审计说明显示在审计跟踪日志中。

9. 单击**下一步**。
10. 输入引用表的名称，在模型存储库中选择引用表对象的位置。
11. （可选）输入引用表的说明。
12. 单击**完成**。

在已添加版本的模型存储库中处理引用表

以只读模式打开引用表。要处理引用表，必须进入编辑模式，或必须从模型存储库签出引用表。

1. 在 Informatica 工具栏中，单击**打开**。
此时将打开资产库。
2. 选择**引用表**资产类别，然后选择引用表名称。
引用表以只读模式打开。
3. 要编辑引用表的当前版本，请单击**编辑**。
要编辑已添加版本的模型存储库中的引用表，请签出此引用表。
4. 完成对引用表的处理后，单击**完成**。Analyst 工具会将更改保存到引用表中。
如果此引用表是从已添加版本的模型存储库签出的，请签入此对象。已添加版本的模型存储库在您签入此对象后才会更新引用表版本。

引用表更新

引用表包含的业务数据可能随时间而变化。在引用表中查看并更新数据和元数据，以确认该表包含准确的信息。在 Analyst 工具中更新引用表。可以在受管引用表和非受管引用表中更新数据和元数据。

可以对引用表数据和元数据执行以下操作：

管理列

可以添加列，删除列和编辑列属性。

管理行

可以将数据行添加到引用表。

编辑引用数据值

可以编辑引用数据值。

替换数据值

使用**查找和替换**选项替换不再准确或与组织相关的数据值。可以查找列中的某个值并将其替换为其他值。可以使用单个值替换列中的所有值。

导出引用表

将引用表导出到逗号分隔的值 (CSV) 文件、字典文件或 Excel 文件。

启用或禁用对非受管表的编辑

更新非受管引用表，以便启用或禁用对表数据和元数据的编辑。

刷新引用表数据

将引用表数据重新加载到 Analyst 工具中以查看对数据的最新更改。

管理列

可以将列添加到引用表中并更新列属性。也可以更新非受管引用表的可编辑状态。

1. 单击**打开**。
此时将打开资产库。
2. 选择**引用表**资产类别，然后选择引用表名称。
引用表以只读模式打开。
3. 要编辑引用表的当前版本，请单击**编辑**。
要编辑已添加版本的模型存储库中的引用表，请签出此引用表。
4. 打开**操作**菜单，然后选择**修改列属性**。
此时将打开**修改列属性**对话框。使用对话框选项执行以下操作：
 - 添加列。
 - 更改表中的有效列。
 - 更改列名称。
 - 更新列的描述性文本。
 - 更新非受管引用表的可编辑状态。
 - 更新表的审计说明。
5. 完成操作后，单击**确定**。

管理行

可以添加、编辑或删除引用表中的行。

1. 单击**打开**。
此时将打开资产库。
2. 选择**引用表**资产类别，然后选择引用表名称。
引用表以只读模式打开。
3. 要编辑引用表的当前版本，请单击**编辑**。
要编辑已添加版本的模型存储库中的引用表，请签出此引用表。
4. 编辑数据行。可以通过以下方法编辑数据行：
 - 要添加行，请选择**操作 > 添加行**。
在**添加行**对话框中，在有效列和至少一个其他列中输入值。（可选）输入审计说明。
单击**确定**以添加行。
 - 要更新单个数据值，请单击该值并更新其数据。
更新数据后，使用行级选项接受或拒绝数据。直接在数据行中输入数据时无法输入审计说明。
 - 要更新行中的数据值，请选择**操作 > 编辑行**。
在**编辑行**对话框中，在一个或多个列中输入值。（可选）输入审计说明。
单击**应用**以更新所选列中的数据。

- 要更新多个行中的值，请选择要编辑的行，然后选择**操作 > 编辑行**。
在**编辑多行**对话框中，在一个或多个列中输入值。（可选）输入审计说明。
单击**确定**以更新所选列中的数据。
- 要删除行，请选择要删除的行并单击**操作 > 删除**。
在**删除行**对话框中，输入审计说明。
单击**确定**以删除行。

注意: 使用 Developer tool 编辑大型引用表中的行数据。例如，如果引用表包含 500 个以上的行，则在 Developer tool 中编辑该表。

查找和替换值

可以在引用表中查找和替换数据值。当表包含您必须更新的数据值的一个或多个实例时，使用查找和替换选项。

1. 单击**打开**。
此时将打开资产库。
2. 选择**引用表**资产类别，然后选择引用表名称。
引用表以只读模式打开。
3. 要编辑引用表的当前版本，请单击**编辑**。
要编辑已添加版本的模型存储库中的引用表，请签出此引用表。
4. 单击**操作 > 查找和替换**。
此时将显示**查找和替换**工具栏。
5. 在该工具栏中输入搜索条件：
 - 在**查找**字段中输入一个数据值。
 - 选择要搜索的列。默认情况下，该操作会搜索所有列。
 - 在**替换为**字段中输入一个数据值。
6. 使用以下选项逐个替换值或替换所有值：
 - 使用**下一个**和**上一个**选项逐个查找值。
 - 要替换某个值，请选择**替换**。
 - 要显示该值的所有实例，请选择**全部突出显示**。
 - 要替换该值的所有实例，请选择**全部替换**。

导出引用表数据

将引用表中的数据导出到逗号分隔的文件、字典文件或 Microsoft Excel 文件。可以采用只读模式导出数据。

1. 单击**打开**。
此时将打开资产库。
2. 选择**引用表**资产类别，然后选择引用表名称。
引用表以只读模式打开。
3. 单击**操作 > 导出数据**。
此时将打开**将数据导出到文件**对话框。

下表介绍了此对话框的选项：

| 选项 | 说明 |
|--------------|---|
| 文件名 | 要包含数据的文件的名称。导出操作创建文件。 |
| 文件格式 | 要包含数据的文件的格式。选择下列其中一种格式： <ul style="list-style-type: none">• csv.逗号分隔的文件。默认格式。• xls.Microsoft Excel 文件。• dic.Informatica 字典文件。 |
| 将字段名称作为第一行导出 | 列名称选项。选择该选项可指示文件的第一行包含列名称。 |
| 代码页 | 引用数据的代码页。默认代码页为 UTF-8。 |

4. 单击**确定**导出文件。

启用和禁用在非受管引用表中进行编辑

可以启用或禁用更新非受管引用表中的数据值和列。

在更改引用表的可编辑状态之前，保存该表。

1. 单击**打开**。
此时将打开资产库。
2. 选择**引用表**资产类别，然后选择引用表名称。
引用表以只读模式打开。
3. 要编辑引用表的当前版本，请单击**编辑**。
要编辑已添加版本的模型存储库中的引用表，请签出此引用表。
4. 打开**操作**菜单，然后选择**修改列属性**。
此时将打开**修改列属性**对话框。
5. 选中或清除**可编辑**选项。

注意：以下条件适用于允许用户更新的非受管引用表：

- 引用表必须使用简单数据类型，例如字符串和数字。
- 不要对引用表元数据定义任何约束，也不要为任何列指定默认值。

刷新引用表值

您可能需要刷新 Analyst 工具显示的引用表值。

要重新加载引用表值，请单击**操作 > 刷新**。Analyst 工具从数据库检索数据值的当前版本。

审计跟踪事件

您可以查看用户对引用表所做更改的审计跟踪。在引用表上使用“审计跟踪”视图可查看审计跟踪事件。可以筛选 Analyst 工具显示的审计跟踪事件。

下表介绍了可以指定的筛选选项：

| 选项 | 说明 |
|----|---|
| 日期 | 要显示的操作的开始日期和结束日期。请使用日历选项设置日期。 |
| 类型 | 审计跟踪事件的类型。您可以查看以下事件类型： <ul style="list-style-type: none">- 数据。与引用表中的数据值相关的事件。这些事件包括添加行、删除行和更新行的操作。- 元数据。与引用表元数据相关的事件。事件包括创建引用表、添加或删除列和签入引用表的操作。 注意： 不能同时查看数据和元数据事件。 |
| 用户 | 编辑过引用表的用户。筛选器会显示用户的全名和登录名。 |
| 状态 | 审计跟踪日志事件状态。该状态对应于您在引用表编辑器中执行的操作。例如，状态可能表示用户创建了引用表或添加了行。 |

审计跟踪日志事件还包括审计跟踪注释和您插入、更新或删除的列值。

查看审计跟踪事件

查看审计跟踪事件，以了解用户对引用表所做的更新。可以在只读模式下查看审计跟踪事件。

1. 单击**打开**。
此时将打开资产库。
2. 选择**引用表**资产类别，然后选择引用表名称。
引用表以只读模式打开。
3. 单击**审计跟踪**。
4. 配置筛选选项。
可以按更新日期、更新类型、更新状态和执行更新的用户名称进行筛选。
5. 单击**显示**
将显示指定筛选选项的日志事件。

引用表的规则和准则

在 Analyst 工具中使用引用表时，使用以下规则和准则：

- 当您从 Oracle、IBM DB2 或 Microsoft SQL Server 数据库导入引用表时，如果表、视图、架构、同义词或列名称包含混合大小写字符或小写字符，Analyst 工具将无法显示预览。
要预览位于区分大小写的数据库中的表内的数据，请将“支持混合大小写标识符”属性设置为 True。
- 从使用一种格式的推理的列模式创建引用表时，Analyst 工具将通过使用另一种格式的列模式填充引用表。

例如，为列模式 X(5) 创建引用表时，Analyst 工具将针对引用表中的列模式显示以下格式：XXXXX。

- 导入 Oracle 数据库表时，确认表中任何 VARCHAR2 列的长度。Analyst 工具无法导入包含长度大于 1000 的 VARCHAR2 列的 Oracle 数据库表。
- 要读取引用表，您需要在到存储表数据值的数据库的连接上具有执行权限。例如，如果引用数据仓库存储这些数据值，您需要在到该引用数据仓库的连接上具有执行权限。您需要执行权限才能以读取或写入模式访问引用表。数据库连接权限适用于数据库中的所有引用数据。
- 运行带有指定引用表转换的映射时，该映射将使用模型存储库中当前版本的引用表。配置转换时，不能选择历史版本的引用表。

如果其他用户将引用表恢复到并行 Developer tool 会话中的早期版本，则各个会话中的引用表版本将不再相同。如果配置并运行使用引用表的映射，则该映射可能会失败，因为当前会话不能识别当前引用表版本。要确保映射使用当前引用表，请在运行映射前刷新模型存储库。

- 将非受管引用表配置为允许编辑时，请验证该引用表是否使用简单数据类型（例如字符串和数字）。另请验证该引用表是否未对引用表元数据定义任何约束或未使用列的默认值。

第 3 章

Developer 工具中的引用数据

本章包括以下主题：

- [Developer tool 引用数据概览, 27](#)
- [引用数据和转换, 27](#)
- [在已添加版本的模型存储库中处理引用数据对象, 28](#)
- [引用表, 29](#)
- [内容集, 33](#)

Developer tool 引用数据概览

您可以在 Developer tool 中创建、更新和查看引用数据对象的配置属性。

使用 Developer tool 创建和更新以下类型的对象：

引用表

引用表包含一组数据值的标准版本和替代版本。您将引用表添加到 Developer tool 中的某个转换，以验证源数据值是否准确并且格式正确。

内容集

内容集是在模型存储库或文件中指定引用数据值的模型存储库对象。内容集包含不同类型的引用数据，您可以将其用于在数据质量转换中执行搜索操作。

您还可以在 Developer tool 中使用地址引用数据文件和标识填充文件。配置地址验证器转换时，选择地址引用数据文件。为标识匹配分析配置匹配转换时，选择标识填充文件。

引用数据和转换

多个转换读取引用数据以执行数据质量任务。

以下转换可以读取引用数据：

- 地址验证器。读取地址引用数据以确认地址的准确性。
- 大小写转换器。读取引用数据表以确定必须更改大小写的字符串。
- 分类器。读取内容集数据以确定字符串中信息的类型。
- 比较。在重复分析过程中读取标识填充数据。

- 标签创建器。读取内容集数据以标识字符串并为其添加标签。
- 匹配。在重复分析过程中读取标识填充数据。
- 解析器。读取内容集数据以根据字符串包含的信息解析字符串。
- 标准创建器。读取引用数据表以按照通用格式标准化字符串。

数据质量内容安装程序文件集包含您可以导入的 Informatica 引用数据对象。

在已添加版本的模型存储库中处理引用数据对象

如果在已添加版本的模型存储库中处理引用表或内容集，此存储库可能会向对象应用版本控制。要向对象应用版本控制，用户应将此对象签入到模型存储库中。

如果引用表或内容集未应用版本控制，可以在版本控制系统外部打开并更新此对象。您打开对象时，模型存储库会锁定此对象，从而使其他用户无法处理此对象。

如果引用表或内容集已应用版本控制，您可以在只读模式下打开此对象。要处理对象，请从模型存储库签出此对象。或者，也可以签出对象，然后再将其打开。签入对象，以创建包含您的最新更改的对象版本。

签出引用数据对象

要处理用户已签入到模型存储库中的引用表或内容集，请从存储库签出此对象。

1. 在 Object Explorer 中，浏览到某个引用表或内容集。
2. 右键单击对象名称，然后单击**打开**。
对象将以只读模式打开。
3. 右键单击对象名称，然后单击**签出**。
此时可以编辑此对象。

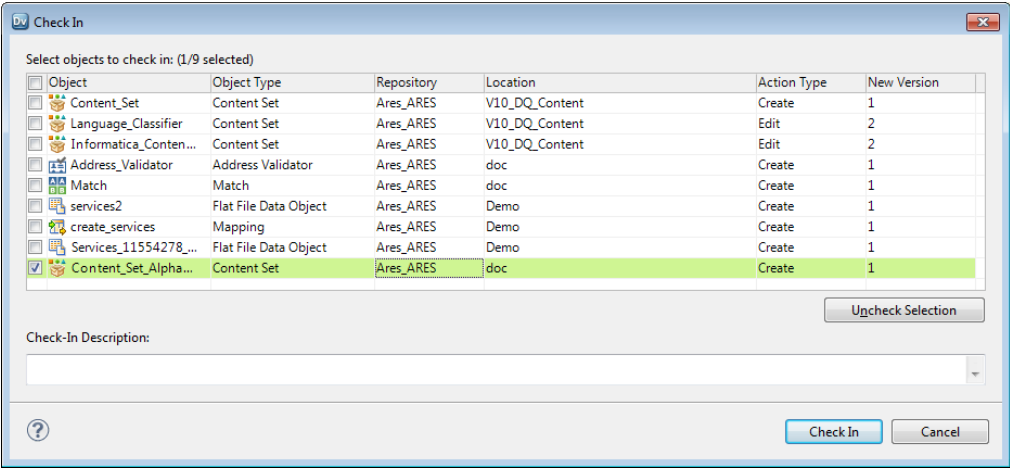
签入引用数据对象

对于从模型存储库签出的引用表或内容集的相关工作完成后，请签入此对象。

要查看当前已签出对象的列表，请打开引用表编辑器下方的**已签出对象**选项卡。

1. 保存对引用表或内容集所做的更改。
2. 在 Object Explorer 中，浏览到引用表或内容集。
3. 右键单击对象名称，然后单击**签入**。
此时将显示**签入**对话框。

下图显示此对话框：



4. 选择一个或多个要签入到存储库中的对象。
- 注意：**可以签入未在当前会话中打开的对象。可以签入任何处于已签出状态的对象。
5. （可选）输入操作的说明。
6. 单击**签入**。
- 签入操作会更新对象版本号。如果首次签入某个对象，模型存储库为此对象创建版本 1。

引用表

您可以在 Developer tool 中向转换添加引用表。您可以配置转换在输入数据中查找引用表值，然后将引用表中对应的有效值作为输出写入。

要在 Developer tool 中创建引用表，请使用以下方法之一：

- 创建空引用表并输入数据值。
- 使用平面文件中的数据创建引用表。
- 使用数据库表、同义词或视图中的数据创建引用表。

引用表数据属性

您可以在 Developer tool 中查看引用表数据和元数据的属性。从模型存储库打开引用表时，Developer tool 将显示这些属性。

引用表显示常规属性和列属性。您可以在 Developer tool 中查看引用表属性。可以在 Analyst 工具中查看和编辑引用表属性。

下表介绍了引用表的常规属性：

| 属性 | 说明 |
|----|-----------|
| 名称 | 引用表的名称。 |
| 说明 | 引用表的可选说明。 |

下表介绍了引用表的列属性：

| 属性 | 说明 |
|------------|----------------------------|
| 有效 | 确定包含有效引用数据的列。 |
| 名称 | 每个列的名称。 |
| 数据类型 | 每个列中数据的数据类型。 |
| 精度 | 每个列的精度。 |
| 小数位数 | 每个列的小数位数。 |
| 说明 | 列的内容说明。创建引用表时，可以选择添加说明。 |
| 包含用于行级说明的列 | 指示引用表包含用于列数据说明的列。 |
| 默认值 | 列中字段的默认值。创建引用表时，可以选择添加默认值。 |
| 连接名称 | 与包含引用表数据值的数据库的连接的名称。 |

创建引用表对象

需要创建空引用表并手动添加值时选择该选项。

1. 从 Developer 工具菜单中选择 **文件 > 新建 > 引用表**。
2. 在新建表向导中，选择 **空引用表**。
3. 输入表的名称。
4. 选择用于存储表元数据的项目。

在“位置”字段中，单击**浏览**。此时将打开**选择位置**对话框并显示存储库中的项目。选择所需的项目。

单击**下一步**。

5. 将两个或更多列添加到表中。单击**新建**选项创建列。

下表介绍了每个列的属性：

| 属性 | 默认值 |
|------|---------|
| 名称 | 列 |
| 数据类型 | 字符串型 |
| 精度 | 10 |
| 小数位数 | 0 |
| 说明 | 空。可选属性。 |

6. 选择包含有效值的列。可以更改所创建的列的顺序。

7. 下表介绍了可选属性：

| 属性 | 默认值 |
|------------|-----|
| 包含用于行级说明的列 | 已清除 |
| 审计说明 | 空 |
| 默认值 | 空 |

单击**完成**。

此时将在 Developer 工具工作区中打开引用表。

从平面文件创建引用表

可以从平面文件中存储的数据创建引用表。

1. 从 Developer 工具菜单中选择**文件 > 新建 > 引用表**。
2. 在新建表向导中，选择**从平面文件创建的引用表**。
3. 浏览至您要用做表的数据源的文件。
4. 输入表的名称。
5. 选择用于存储表元数据的项目。

在“位置”字段中，单击**浏览**。此时将打开**选择位置**对话框并显示存储库中的项目。选择所需的项目。

单击**下一步**。

6. 将 UTF-8 设置为代码页。
7. 指定平面文件使用的分隔符。
8. 如果平面文件包含列名称，则选择用于从文件的第一行导入列名称的选项。
9. 下表介绍了可选表属性：

| 属性 | 默认值 |
|--------------|--------------|
| 文本限定符 | 无引号 |
| 导入起始行 | 第 1 行 |
| 行分隔符 | \012 LF (\n) |
| 将连续分隔符视为一个整体 | 已清除 |
| 转义符 | 空 |
| 在数据中保留转义符 | 已清除 |
| 要预览的最多行数 | 500 |

单击**下一步**。

10. 选择包含有效值的列。

11. 下表介绍了可选属性：

| 属性 | 默认值 |
|------------|-----|
| 包含用于行级说明的列 | 已清除 |
| 审计说明 | 空 |
| 默认值 | 空 |
| 要预览的最多行数 | 500 |

单击**完成**。

此时将在 Developer 工具工作区中打开引用表。

从关系源创建引用表

您可以从关系表、同义词或视图创建引用表。

创建受管引用表时，请将列数据导入引用数据仓库。创建非受管引用表时，请确定存储列数据的数据库表。您可以从 ODBC 连接或本地连接创建受管引用表。可以从本地连接创建非受管引用表。

创建引用表之前，请先验证 Informatica 域中是否存在与包含引用数据的数据库的连接。

可以在连接浏览器中配置数据库连接。如果 Developer tool 不显示连接浏览器，请从 Developer tool 菜单中选择**窗口 > 显示视图 > 连接浏览器**。

从关系源创建引用表

要创建引用表，请连接到数据库并选择含有引用数据的表。

1. 从 Developer tool 菜单中选择**文件 > 新建 > 引用表**。
2. 在表创建向导中，选择**来自关系源的引用表**。

单击**下一步**。

3. 选择数据库连接。

在“连接”字段中，单击**浏览**。此时将打开**选择连接**对话框并显示可用的数据库连接。

选择连接时单击**确定**。

4. 选择数据库资源。

在“资源”字段中，单击**浏览**。此时将打开**选择资源**对话框并显示数据库连接上的资源。浏览数据库并选择一个数据库表、同义词或视图。

您可以选择预览资源上的实体信息。

5. 输入表的名称。

6. 为引用表对象选择一个位置。

在“位置”字段中，单击**浏览**。此时将打开**选择位置**对话框并显示存储库中的项目。

选择一个位置并单击**下一步**。

7. 要创建不在引用数据仓库中存储数据的引用表，请选择**非受管表**。

要使用户能够编辑非受管引用表，请选中**可编辑**选项。

单击**下一步**。

8. 选择包含有效值的列。

9. 下表介绍了您可以指定的可选属性：

| 属性 | 默认值 |
|------------|-----|
| 包含用于行级说明的列 | 已清除 |
| 说明 | 已清除 |
| 默认值 | 空 |
| 审计说明 | 空 |
| 要预览的最多行数 | 500 |

10. 单击**完成**。

内容集

内容集是存储其他引用数据对象的数据或元数据的模型存储库。内容集可以包含字符集、模式集、标志集、正则表达式、概率模型和分类器模型。使用内容集定义和组织与单个项目、信息类型或业务目的相关的引用数据对象。

Developer tool 包含不显示在模型存储库中的系统定义的字符集和标志集。要查看和使用系统定义的对象，请在标签创建器转换、解析器转换或标准创建器转换中配置策略。

字符集

字符集包含可标识特定字符和字符范围的表达式。可以在使用为“字符添加标签”模式的标签创建器转换中使用字符集。

字符范围指定字符代码的顺序范围。例如，字符范围 “[A-C]” 与大写字符 “A”、“B” 和 “C” 相匹配。该字符范围与小写字符 “a”、“b” 或 “c” 不匹配。

使用字符集将特定字符或字符范围标识为“添加标签”操作的一部分。例如，可以在包含电话号码的列中为所有数字添加标签。为所有数字添加标签后，可以使用解析器转换标识模式，并将有问题的模式写入单独的输出口。

字符集属性

为字符集配置用于确定“为字符添加标签”操作的属性。

下表介绍了用户定义的字符集的属性：

| 属性 | 说明 |
|------|---------------------------|
| 标签 | 定义标签创建器转换应用至与字符集匹配的数据的标签。 |
| 标准模式 | 启用包含开始范围和结束范围字段的简单编辑视图。 |
| 开始范围 | 指定字符范围中的第一个字符。 |

| 属性 | 说明 |
|-------|--|
| 结束范围 | 指定字符范围中的最后一个字符。对于具有单个字符的范围，将该字段留空。 |
| 高级模式 | 启用高级编辑视图，您可以在该视图中使用范围字符和分隔符字符手动输入字符范围。 |
| 范围字符 | 临时更改表示字符范围的符号。当您关闭字符集时，范围字符将还原为默认字符。 |
| 分隔符字符 | 临时更改分隔符字符范围的符号。当您关闭字符集时，分隔符字符将还原为默认字符。 |

分类器模型

分类器模型分析输入字符串并确定字符串最可能包含的信息类型。在分类器转换中使用分类器模型。

分类器模型包含引用数据行和标签值。行代表可能连接到分类器转换的端口上的输入数据。标签值描述数据行包含的信息类型。配置分类器模型时，应为模型中的每个引用数据行分配标签。

要将引用数据行链接到分类器模型中的标签，您可以对该模型进行编译。编译过程会在数据行与标签值之间生成一系列逻辑关联。当您运行读取模型的映射时，数据集成服务会将模型逻辑应用到分类器转换输入数据。数据集成服务将返回对每个输入数据字段中的信息描述最准确的标签。

请在 Developer tool 中创建分类器模型。模型存储库存储分类器模型对象。Developer tool 将数据行、标签和编译数据写入 Informatica 目录结构中的一个文件。

模式集

模式集包含表达式，可用于标识“为标志添加标签”操作的输出中的数据模式。可以使用模式集分析标志化的数据输出端口，并将匹配的字符串写入一个或多个输出端口。在使用模式解析模式的解析器转换中使用模式集。

例如，可以将解析器转换配置为使用可标识姓名和首字母的模式集。在“为标志添加标签”模式下，该转换使用模式集分析标签转换的输出。可以将解析器转换配置为将输出中的姓名和首字母写入单独的端口。

模式集属性

配置用于确定模式集中的模式的属性。

下表介绍了用户定义的模式集的属性：

| 属性 | 说明 |
|----|--|
| 模式 | 定义模式解析器搜索的模式。可以为一个模式集输入多个模式。可以输入从通配符、字符和字符串的组合构造的模式。 |

概率模型

概率模型分析输入数据值并确定这些值最可能包含的信息类型。请在标签创建器转换和解析器转换中使用概率模型。

概率模型包含引用数据值和标签值。引用数据值代表连接到转换的输入端口上的数据。标签值描述引用数据值包含的信息类型。请为模型中的每个引用数据值分配标签。

要将引用数据值链接到概率模型中的标签，您可以对该模型进行编译。编译过程会在数据值与标签之间生成一系列逻辑关联。当您运行读取模型的映射时，数据集成服务会将模型逻辑应用到转换输入数据。数据集成服务将返回最能准确描述输入数据值的标签。

您可以在 Developer tool 中创建概率模型。模型存储库存储概率模型对象。Developer tool 将数据值、标签和编译数据写入 Informatica 目录结构中的一个文件。

正则表达式

在内容集的上下文中，正则表达式是可以在解析和添加标签操作中使用的表达式。使用正则表达式标识输入数据中的一个或多个字符串。可以在使用标志解析模式的解析器转换中使用正则表达式。还可以在使用“为标志添加标签”模式的标签创建器转换中使用正则表达式。

解析器转换使用正则表达式匹配输入数据中的模式，并将所有匹配的字符串解析到一个或多个输出。例如，您可以使用正则表达式标识输入数据中的所有电子邮件地址，并将每个电子邮件地址组件解析到不同的输出。

标签创建器转换使用正则表达式匹配输入模式并创建单个标签。包含多个输出的正则表达式不会生成多个标签。

正则表达式属性

配置确定正则表达式如何标识和写入输出字符串的属性。

下表介绍了用户定义的正则表达式的属性：

| 属性 | 说明 |
|--------|---|
| 输出数 | 定义正则表达式写入的输出端口的数量。 |
| 正则表达式 | 定义解析器转换用于匹配字符串的模式。 |
| 测试表达式 | 包含您输入以测试正则表达式的数据。在该字段中键入数据后，该字段将突出显示与正则表达式匹配的字符串。 |
| 下一个表达式 | 移动至下一个与正则表达式匹配的字符串并将该字符串的字体更改为粗体。 |
| 上一个表达式 | 移动至上一个与正则表达式匹配的字符串并将该字符串的字体更改为粗体。 |

标志集

标志集包含能够标识特定标志的表达式。您可以在使用“为标志添加标签”模式的标签创建器转换中使用标志集。还可以在使用标志解析模式的解析器转换中使用标志集。

使用标志集可以将特定标志标识为添加标签和解析操作的一部分。例如，您可以使用标志集为所有采用“AccountName@DomainName”格式的电子邮件地址添加标签。为标志添加标签后，您可以使用解析器转换将这些电子邮件地址写入指定的输出端口。

标志集属性

为标志集配置用于确定添加标签操作的属性。

下表介绍了用户定义的字符集的属性：

| 属性 | 标志集模式 | 说明 |
|----|-------|-----------|
| 名称 | N/A | 定义标志集的名称。 |
| 说明 | N/A | 描述标志集。 |

| 属性 | 标志集模式 | 说明 |
|--------|-------|---|
| 标志集选项 | N/A | 定义标志集使用正则表达式模式还是字符模式。 |
| 标签 | 正则表达式 | 定义标签创建器转换应用至与标志集匹配的数据的标签。 |
| 正则表达式 | 正则表达式 | 定义标签创建器转换用于匹配字符串的模式。 |
| 测试表达式 | 正则表达式 | 包含您输入以测试正则表达式的数据。在该字段中键入数据后，该字段将突出显示与正则表达式匹配的字符串。 |
| 下一个表达式 | 正则表达式 | 移动至下一个与正则表达式匹配的字符串并将该字符串的字体更改为粗体。 |
| 上一个表达式 | 正则表达式 | 移动至上一个与正则表达式匹配的字符串并将该字符串的字体更改为粗体。 |
| 标签 | 字符 | 定义标签创建器转换应用至与字符集匹配的数据的标签。 |
| 标准模式 | 字符 | 启用包含开始范围和结束范围字段的简单编辑视图。 |
| 开始范围 | 字符 | 指定字符范围中的第一个字符。 |
| 结束范围 | 字符 | 指定字符范围中的最后一个字符。对于单个字符范围，请将该字段留空。 |
| 高级模式 | 字符 | 启用高级编辑视图，您可以在该视图中使用范围字符和分隔符字符手动输入字符范围。 |
| 范围字符 | 字符 | 临时更改表示字符范围的符号。当您关闭字符集时，范围字符将还原为默认字符。 |
| 分隔符字符 | 字符 | 临时更改分隔符字符范围的符号。当您关闭字符集时，分隔符字符将还原为默认字符。 |

概率模型和分类器模型的规则和准则

模型存储库中的每个概率模型和分类器模型都标识了 Informatica 目录结构中的一个文件。该文件包含数据值以及您在 Developer tool 中向模型添加的标签。文件中还包含定义数据值与标签之间关联的编译逻辑。

使用概率模型或分类器模型时，请注意以下规则和准则：

- 当您运行包含模型的映射时，数据集成服务会将编译的模型逻辑应用到转换输入数据。数据集成服务不会在映射运行时读取模型中的数据值或标签。
- 您可以选择从概率模型或分类器模型中删除数据值和标签。例如，您可能会决定从模型中删除敏感数据或专有数据。您可以在 Developer tool 中删除个别数据值和标签。您可以在从模型存储库中导出模型时删除所有数据值和标签。

注意：如果您从模型中删除所有数据值和标签，将无法编译该模型。

- 当您从模型删除一个或多个数据值或标签时，编译的模型逻辑将不再代表模型文件中的当前数据。要同步模型逻辑与数据值和标签，请重新编译模型。如果要保持当前的模型逻辑，则不要编译模型。
- 要保护分类器模型或概率模型中的数据，请备份 Informatica 目录结构中的模型文件。请在从模型中删除所有数据值和标签之前备份该文件。
- 在内容管理服务主机计算机中找到模型文件。

概率模型文件的默认位置 and 文件扩展名如下：

<Informatica 安装目录>/tomcat/bin/ner/<文件名>.ner

分类器模型文件的默认位置 and 文件扩展名如下：

<Informatica 安装目录>/tomcat/bin/classifier/<文件名>.classifier

- 如果您升级了 Informatica 安装，在将模型用于映射之前，可能需要先对概率模型和分类器模型进行编译。如果模型不包含任何数据，可使用包含数据的备份文件替换 Informatica 目录结构中的当前文件。

管理分类器模型和概率模型中的标签

要检查并更新概率模型或分类器模型中的标签，请使用**管理标头**对话框。

1. 打开包含该分类器模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 打开**管理标签**对话框。

此对话框会列出模型中的标签。

相关主题：

- [“分类器模型标签管理” 页面上 43](#)
- [“概率模型配置” 页面上 57](#)

创建内容集

创建内容集，以管理引用单个项目、信息类型或业务目的的引用数据对象。

1. 在**对象浏览器**视图中，选择用于存储内容集的项目或文件夹。
2. 单击**文件 > 新建 > 内容集**。
3. 输入内容集的名称。
4. （可选）选择**浏览**以更改内容集的模型存储库位置。
5. 单击**完成**。

在内容集中创建引用数据对象

可以在内容集中创建字符集、模式集、标志集、正则表达式、概率模型和分类器模型。

1. 在编辑器中打开内容集并选择**内容**视图。
2. 选择引用数据对象类型。
3. 单击**添加**。
4. 输入引用数据对象的名称。
(可选) 输入对象的说明。
5. 配置引用数据对象属性。
6. 单击**完成**。

提示: 可以将一个内容集中的引用数据对象复制到另一个内容集中。使用**复制到**和**粘贴自**选项可以创建内容集中某个对象的副本。使用 **Ctrl** 键可以选择多个内容集对象。

第 4 章

分类器模型

本章包括以下主题：

- [分类器模型概览, 39](#)
- [分类器模型结构, 40](#)
- [分类器得分, 40](#)
- [分类器转换示例, 40](#)
- [分类器模型选项, 41](#)
- [分类器模型引用数据, 42](#)
- [分类器模型标签数据, 42](#)
- [分类器模型配置, 44](#)
- [筛选操作和查找操作, 47](#)
- [复制和粘贴操作, 48](#)

分类器模型概览

分类器模型是内容集中的引用数据对象。使用分类器模型分析包含多个值的长文本字符串。分类器模型确定每个字符串中最常见的信息类型。

将分类器模型添加到分类器转换中。转换搜索分类器模型数据和每个输入行中的数据之间的共有值。转换使用这些共有值对每一行代表的信息类型进行分类。

当输入数据具有以下特性时使用分类器模型：

- 输入数据包含文本。分类器模型将自然语言处理应用至文本数据以确定文本中信息的类型。自然语言处理检测输入字符串中的相关单词。自然语言处理忽略不相关的单词。
- 输入数据字符串包含多个值。例如，可以创建在每个字段中包含电子邮件内容的数据列。

分类器转换读取字符串数据类型。转换对输入字符串的长度没有限制。

在 Developer 工具中编译分类器模型。编译模型时，将在模型中的相似数据值之间创建关联。分类器转换使用编译的数据搜索输入数据中的信息。

分类器模型结构

分类器模型包含引用数据值和标签值。引用数据值代表要分类的数据。标签值指定分类器转换可从数据中识别的信息类型。

分类器模型还包含编译数据。分类器转换使用编译数据来测量模型中的引用数据和转换输入数据之间的相似度。对分类器模型进行编译时，可以创建或更新编译数据。将输入数据与模型数据相比较后，分类器转换会返回描述每行输入数据的标签值。

Developer tool 会将引用数据值、标签值和编译数据写入到 Informatica 目录结构中的一个文件。模型存储库中的分类器模型对象存储了文件名。保存分类器模型后，当前的引用数据值和标签值随即写入到文件中。对模型进行编译时，将更新文件中的编译数据。您可以从 Developer tool 中的模型属性读取文件名。

分类器得分

分类器转换将输入数据中的每一行与分类器模型中引用数据的每一行进行比较。转换计算每个比较的得分。得分代表输入行与引用数据行之间的相似度。

运行某个包含分类器转换的映射时，该映射返回使用最高得分标识引用数据行的标签。得分范围为从 0 到 1。高分指示输入数据与模型数据之间的强匹配。

查看分类器得分以确认标签输出准确描述输入数据的每一行。还可以查看得分以确认分类器模型适合输入数据。如果转换输出包含很大比例的低得分，则分类器模型可能不适合。要改进比较，请再次编译模型。如果已编译的模型不能提高得分，请替换转换中的模型。

分类器转换示例

可以使用分类器模型和分类器转换根据电子邮件包含的文本对电子邮件进行分类。

例如，假设您在某家软件制造商的客户支持中心担任数据管理者一职，您负责审阅支持中心从客户处收到的电子邮件。您所在组织的客户来自许多国家/地区，因此支持中心收到用多种不同语言撰写的电子邮件。您决定按语言对这些电子邮件进行排序，以便可以将每封电子邮件发送到可向客户做出最佳回复的部门。

要对电子邮件进行排序，请执行以下步骤：

1. 将电子邮件写入到单个文件或一个数据库表。
2. 在模型存储库中创建读取文件或数据库表的数据对象。
3. 在模型存储库中为邮件所用的每种语言创建数据对象。
4. 创建包含每种语言的示例文本的分类器模型。

注意：可以使用电子邮件数据中的示例数据作为模型的源数据。

5. 将分类器模型添加到可重用的分类器转换中。
6. 配置映射，以将分类器转换应用到邮件数据。

要配置映射，请执行以下步骤：

- 将分类器转换和数据对象添加到映射中。
- 将分类器转换输入端口连接到源数据对象。
- 将分类器转换输出端口连接到目标数据对象。

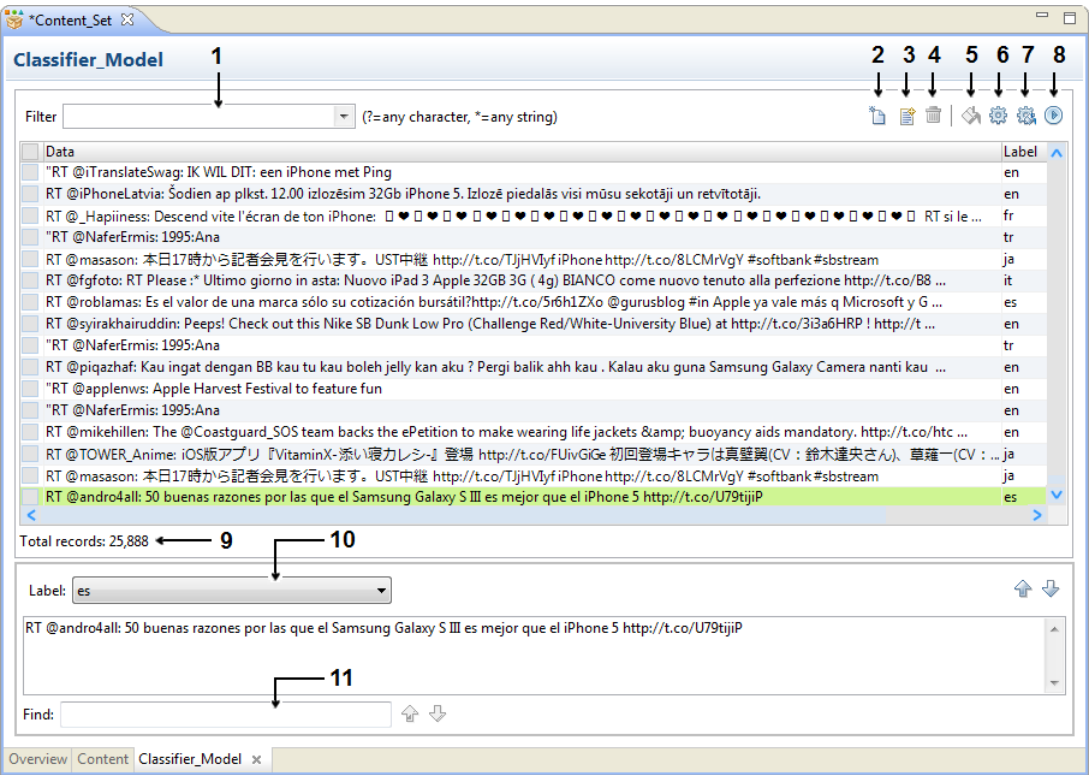
运行映射时，分类器转换分析电子邮件并将电子邮件文本写入正确的数据目标。您可与每个部门的团队成员共享数据目标。

分类器模型选项

Developer tool 会在包含上方窗格和下方窗格的编辑器中显示分类器模型数据。上方窗格显示每个引用数据行以及为数据分配的任何标签。下方窗格显示所选行的内容。

使用上方窗格可审阅引用数据行以及识别任何未使用标签的行；使用下方窗格可审阅行的内容以及为该行分配标签。上方窗格的每一行大概显示数据的 100 个字符。下方窗格显示所选行中的所有数据。

下图显示了分类器模型编辑器：



编辑器包含以下选项：

1. “筛选器”字段
基于指定的数据值或标签筛选引用数据行的列表。
2. 添加行
插入空的引用数据行。
3. 附加数据
将数据对象中的数据导入到模型存储库。
4. 删除
删除所选引用数据行。可以使用复选框来选择行。
5. 分配标签

将标签分配给选定的一个或多个引用数据行。可以使用复选框来选择行。

6. 编辑属性

显示分类器模型属性。

7. 管理标签

打开**管理标签**对话框。您可以使用此对话框向分类器模型添加标签值或从中删除标签值。

8. 编译

编译分类器模型。

9. 记录总数

指示分类器模型中的引用数据行的数量。

10. “标签”字段

显示可以应用到当前引用数据行的标签值。

11. “查找”字段

在当前引用数据行中查找指定的数据值。

分类器模型引用数据

分类器模型包含引用数据列，该列可以包含文本句子、段落或页面。引用数据代表分类器转换可以在映射中读取的不同类型的文本输入。创建模型时，确认引用数据包含您希望在运行映射时查找的文本类型。

可以使用映射源数据创建分类器模型。选择一个源数据示例并将该数据示例复制到模型中。

在使用分类器模型引用数据时，请考虑以下规则和准则：

- 引用数据字段可以具有任何长度。可以将文本页面输入每个数据字段。
- 从数据对象导入引用数据。
- 无法编辑引用数据值。但是，可以删除数据行。
- 编译分类器模型时，编译进程将忽略引用数据中的任何数字值。

分类器模型标签数据

分类器模型包含一个或多个描述性标签，这些标签概括了引用数据行中的信息类型。请为每个引用数据行分配标签。

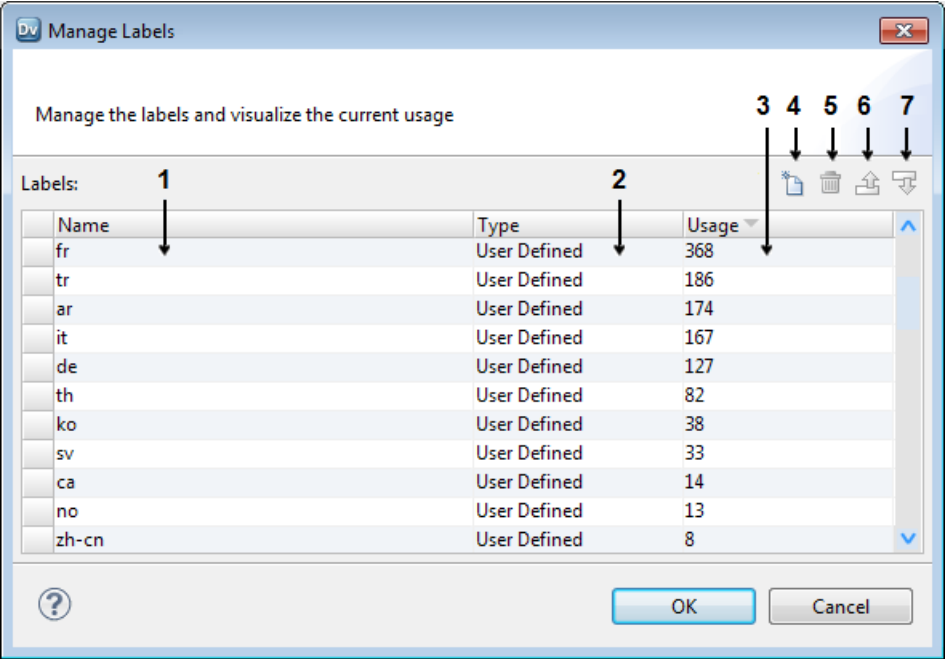
将数据源中的数据添加到分类器模型时，您可以将某个列指定为标签数据列。还可以在模型中创建标签。

标签独立于其描述的引用数据值。删除使用某个标签的引用数据行时，不会从模型中删除该标签。如果删除某个标签，不会删除与该标签关联的引用数据值。

分类器模型标签管理

您可以使用**管理标签**对话框来检查和更新分类器模型中的标签值，还可以对标签值进行排序和更新。

下图显示了**管理标签**对话框：



管理标签对话框包含以下元素：

1. “名称”列。
包含分类器转换可以应用到输入数据行的标签值。可以按名称对标签进行排序。
2. “类型”列。
标识标签值的源。分类器模型会将所有标签标识为用户定义的值。
3. “使用情况”列。
指示使用各个标签的引用数据行的数量。可以按行数对标签进行排序。
4. “添加”按钮。
请将标签添加到分类器模型，并在行上的“名称”列中输入标签值。
注意: 要更新标签值，请双击该值并输入所需的值。
5. “删除”按钮。
从分类器模型中删除标签。
6. 向上箭头。
在对话框中将标签向上移动一行。
7. 向下箭头。
在对话框中将标签向下移动一行。

分类器模型配置

配置分类器模型的第一步是选择要分类的数据。添加到模型的引用数据的内容必须反映连接到分类器转换的数据。转换会将输入数据中的数据值和模式与分类器模型中的数据值和模式相比较。

要创建可在分类器转换中使用分类器模型，请执行以下任务：

1. 确定要添加到模型中的引用数据值和标签值。
可以使用要分类的数据片段。请在模型存储库中创建读取该数据片段的数据对象。
2. 创建内容集，并将分类器模型添加到该内容集。
3. 将引用数据值添加到模型中。
4. 将标签值添加到模型中。
您可以将数据对象中的数据导入到模型存储库，也可以输入单个引用数据行或标签。
5. 为每个引用数据行分配标签。
您可以通过单个操作为多个行分配标签。
6. 对模型进行编译。

对分类器模型进行编译后，便可以在分类器转换中使用该模型。

创建分类器模型

可将数据对象用作分类器模型数据的源。

将分类器转换的输入数据用作模型引用数据的源时，分类器模型的性能最佳。

1. 在对象浏览器中，打开或创建一个内容集。
2. 选择**内容**视图。
3. 选择**分类器模型**，然后单击**添加**。
此时将打开分类器模型向导。
4. 输入分类器模型的名称。
或者，输入模型的文本说明。
5. 浏览模型存储库并选择包含要导入的数据的数据对象。
请勿选择社交媒体数据对象。
单击**下一步**。
6. 检查数据对象中的列，然后选择一个或多个列以添加到模型中。可以在同一个操作中添加引用数据列和标签列。
 - 要将数据列导入为引用数据，请选择列名称并单击**数据**。
可以选择多个数据列。Developer tool 会将所选列的内容合并为单个列。
 - 要将数据列导入为标签值，请选择列名称并单击**标签**。导入引用数据和标签值时，Developer tool 会将每一行上的标签分配给同一行上的引用数据字符串。选择列之前，您可以预览数据。创建模型后，您可以更改标签分配。
单击**下一步**。
7. 选择要从数据源导入的行数。
默认情况下，Developer tool 会导入数据源中的所有行。如果输入数字，则模型会从数据集的开头对行进行计数。
8. 单击**完成并保存模型**。

创建模型后，请验证标签分配并对模型进行编译。

将数据源中的数据附加到分类器模型

您可以通过单个操作将多行引用数据值或标签值导入到分类器模型。

1. 打开包含该分类器模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 单击**附加数据**。

此时将打开分类器模型向导。

4. 浏览模型存储库并选择包含要导入的数据的数据对象。

请勿选择社交媒体数据对象。

单击**下一步**。

5. 检查数据对象中的列，然后选择一个或多个列以添加到模型中。可以在同一个操作中添加引用数据列和标签列。

- 要将数据列导入为引用数据，请选择列名称并单击**数据**。

可以选择多个数据列。Developer tool 会将所选列的内容合并为单个列。

- 要将数据列导入为标签值，请选择列名称并单击**标签**。

导入引用数据和标签值时，Developer tool 会将每一行上的标签分配给同一行上的引用数据字符串。选择列之前，您可以预览数据。创建模型后，您可以更改标签分配。

单击**下一步**。

6. 选择要从数据源导入的行数。

默认情况下，Developer tool 会导入数据源中的所有行。如果输入数字，则模型会从数据集的开头对行进行计数。

7. 单击**完成**并保存模型。

将引用数据行添加到分类器模型

您可以向分类器模型添加单个引用数据行。

1. 打开包含该分类器模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 单击**添加行**。

Developer tool 将在引用数据中当前行的下方添加一个行。

4. 将引用数据值输入到该行中。

您可以使用 Windows 快捷键将数据粘贴到行中。

将标签添加到分类器模型

您可以向分类器模型添加单个标签。

1. 打开包含该分类器模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 打开**管理标签**对话框。

此对话框会列出模型中的标签。

4. 单击**新建**。

Developer tool 将在标签列表的底部添加一个行。

5. 双击“名称”列中的默认值，然后输入标签名称。
6. 单击**确定**。

创建标签后，您可以将该标签分配给一个或多个引用数据行。**管理标签**对话框中的“使用情况”列显示了使用该标签的行的数量。

为引用数据行分配标签

您可以通过单个操作为一个或多个引用数据行分配标签。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 选择一个或多个引用数据行。可以使用复选框选项来选择行。
注意：您可以使用筛选器选项来显示所有包含指定的数据值的行。使用“全选”复选框选项可选择所有包含相应值的行。
4. 单击**分配标签**。
Developer tool 将显示分类器模型中的标签的列表。
5. 选择标签值，然后单击**分配**。
Developer tool 将用该标签值更新所选引用数据行。
(可选) 对模型进行编译，以将标签名称添加到分类器模型逻辑。

识别未使用的标签值

您可以使用**管理标签**对话框查找分类器模型中任何仍未使用的标签。**管理标签**对话框会显示标签值在分类器模型中的使用情况数据。您可以基于使用情况数据验证使用某个标签值的引用数据行的数量，以及查找未使用的标签值。

1. 打开包含该分类器模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 打开**管理标签**对话框。此对话框会列出分类器模型中的标签。
4. 检查每个标签的“使用情况”列数据。
“使用情况”列会列出使用该标签的引用数据行的数量。如果某个标签值未使用，其“使用情况”列的值将为零。

从分类器模型中删除行

您可以通过单个操作从分类器模型中删除一个或多个引用数据行。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 选择一个或多个引用数据行。可以使用复选框选项来选择行。
4. 单击**删除**。
Developer tool 将从分类器模型中删除所选行。
要撤消操作，请按键盘上的 Ctrl+Z。

从分类器模型中删除标签

您可以使用**管理标签**对话框从分类器模型中删除标签。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 打开**管理标签**对话框。
4. 单击**删除**。
5. 单击**是**以确认操作。

Developer tool 将从模型中删除该标签。Developer tool 不会删除任何使用该标签的引用数据行。

6. 单击**确定**关闭对话框。

要撤消操作，请按键盘上的 Ctrl+Z。

编译分类器模型

每次在分类器模型中编辑标签值或引用数据值时，必须编译模型。编译模型时，将更新模型中的编译数据。

- 要更新编译数据，请在 Developer 工具中打开模型并单击**编译**。

筛选操作和查找操作

使用筛选器选项可显示或隐藏满足指定的条件的引用数据行。应用筛选器时，您可以对分类器模型显示的数据行执行其他操作。例如，您可以将标签值应用到所有数据行。

使用筛选器选项可执行以下任务：

- 查找包含输入的值的引用数据行。
- 查找使用所选标签的引用数据行。
- 查找未使用标签的引用数据行。

您还可以在引用数据行中搜索数据值。

使用数据值筛选引用数据行

使用筛选器可验证一个或多个引用数据行是否包含预期的数据值。

1. 打开包含该分类器模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 在“筛选器”字段中输入值。

输入的值可以包含通配符。

Developer tool 将显示包含筛选器文本的引用数据行。

使用标签值筛选引用数据行

使用筛选器可显示或隐藏使用所选标签的引用数据行。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。

3. 从“筛选器”菜单中选择标签值。

Developer tool 将显示使用该标签值的引用数据行。

注意: 要查找任何未使用标签的引用数据行, 请从“筛选器”菜单中选择**无标签**选项。

在引用数据行中查找值

使用“查找”字段可在所选行中搜索数据值。

1. 打开包含该模型的内容集。
2. 选择模型名称, 然后单击**编辑**。
3. 选择引用数据行。
4. 在“查找”字段中输入值。

模型会突出显示值在引用数据行中的实例。

5. 使用向上箭头或向下箭头查找值在行中的其他实例。

复制和粘贴操作

可以将模型存储库中一个内容集中的分类器模型复制到另一个内容集中。复制分类器模型, 以便与其他 Developer 工具用户共享资源。

可以将模型复制到其他内容集, 也可以将模型导入当前内容集。可以在单个操作中从存储库中的多个内容集导入多个模型。

复制模型时, 内容管理服务在服务计算机上创建模型数据文件的副本。每个模型使用不同的数据文件。

将分类器模型复制到其他内容集

可以将模型存储库中一个内容集中的分类器模型复制到另一个内容集中。复制分类器模型时, 应指定模型对象以及源内容集和目标内容集。

1. 打开包含该分类器模型的内容集。
2. 选择一个分类器模型并单击**复制到**。
3. 浏览模型存储库并选择一个内容集。

可以将分类器模型复制到当前项目或其他项目中的内容集。

4. 单击**确定**。

Developer 工具将分类器模型复制到所选的内容集中。

从其他内容集导入分类器模型

在模型存储库中, 可以将一个内容集中的分类器模型导入另一个内容集。导入分类器模型时, 应指定一个或多个模型对象以及源内容集和目标内容集。

1. 打开包含该分类器模型的内容集。
2. 选择一个分类器模型并单击**粘贴自**。
3. 浏览模型存储库并选择一个分类器模型。

可以从当前项目或其他项目中的内容集粘贴分类器模型。

4. 单击**确定**。

Developer 工具将分类器模型粘贴到当前内容集中。

第 5 章

概率模型

本章包括以下主题：

- [概率模型概览, 50](#)
- [概率模型结构, 51](#)
- [标签创建器转换示例, 51](#)
- [解析器转换示例, 52](#)
- [概率模型选项, 52](#)
- [概率模型引用数据, 55](#)
- [概率模型标签数据, 55](#)
- [概率模型属性, 56](#)
- [概率模型配置, 57](#)
- [复制和粘贴操作, 62](#)

概率模型概览

概率模型是您在内容集中创建的引用数据对象。使用概率模型可以分析包含多个数据值的数据字符串。概率模型标识字符串中每个值中的信息类型。可以将概率模型添加到标签创建器转换和解析器转换中。

使用标签创建器转换中的概率模型为输入字符串中的每个值分配一个描述性标签。标签创建器转换将标签写入单个输出端口。在解析器转换中使用概率模型可以将输入字符串中的每个值写入代表值中的信息的端口。解析器转换为每种类型的信息创建一个输出端口。

您可以在 Developer tool 中设计和编译概率模型。定义概率模型时，将一系列数据行添加到模型中并为每个行中的每个值分配一个标签。编译概率模型时，Developer tool 在数据值和您添加的标签之间创建关联。标签创建器转换和解析器转换使用自然语言处理将概率模型数据与输入端口数据进行比较。

自然语言处理使用以下技术来确定数据值中信息的类型：

- 自然语言处理可以识别相似的数据值，并为这些值应用相同的标签。
- 自然语言处理可以将某个数据值与字符串中的相邻值进行比较。自然语言处理分析值序列以了解每个字符串的用法并确认字符串代表的信息类型。

概率模型结构

概率模型包含引用数据值行和标签值。引用数据值代表可能出现在转换输入数据中的各个值。标签值标识您希望输入数据包含的信息类型。

概率模型还包含编译数据。标签创建器转换和解析器转换使用编译数据来测量模型中的引用数据和转换输入数据之间的相似度。对概率模型进行编译时，将创建或更新编译数据。

一个数据行可以包含单个值或多个值。每个数据行可能具有不同的结构。您可以将同一个标签分配给数据行中的多个值。或者，您也可将出现在一个行中不同位置的同一个值分配不同的标签。运行映射时，数据集成服务会将值在输入字符串中的相对位置考虑在内。在编译概率模型之前，将每个标签分配给至少一个数据值。

Developer tool 会将引用数据值、标签值和编译数据写入到 Informatica 目录结构中的一个文件。模型存储库中的概率模型对象存储了文件名。保存概率模型后，当前的引用数据值和标签值随即写入到文件中。对模型进行编译时，将更新文件中的编译数据。您可以从 Developer tool 中的模型属性读取文件名。

注意: 要优化概率模型的功能，请验证每个数据行是否包含多个引用数据值。每个行中值的顺序必须尽可能与值在转换输入数据中出现的顺序一致。如果数据行包含单个引用数据值，则标签创建器转换或解析器转换无法在概率分析期间应用自然语言处理。

标签创建器转换示例

某保险组织中的客户数据库包含多个数据条目错误。您是该保险组织的数据管理者。您使用标签创建器转换配置一个映射，以确定每个列包含的不同数据类型。

下表介绍了客户数据库中的示例数据：

| 行 ID | 字段 1 | 字段 2 | 字段 3 |
|------|-------------------|--------------------------------|------------------|
| 1 | 19132954 | AIM SECURITIES | PETRIE TAYBRO |
| 2 | 10110169 | JASE TRAPANI | BANK OF NEW YORK |
| 3 | 10111786 | WANGER ASSET MANAGEMENT, LLP | JAN SEEDORF |
| 4 | 10112299 | FELIX LEVINGER | HARVARD MAGAZINE |
| 5 | 10112036 | DESCHÊNES & FILS LTÉE (QUEBEC) | RICHARD TREMBLAY |
| 6 | BERGER ASSOCIATES | 10111101 | DAREEN HULSMAN |
| 7 | 19131385 | EAGLE FINANCIAL GROUP INC | PATRICK MCKINNIE |
| 8 | LAKENYA PASKETT | WHITEHALL FINANCIAL GROUP | 15954710 |

当您运行该映射时，标签创建器转换将输入数据与概率模型引用数据进行比较。标签创建器转换会为每个输入端口上的数据选择一个标签。该转换将这些标签写入一个输出端口。每个输出行包含一组用于定义对应输入行上的数据结构的标签。

下表介绍了标签创建器转换添加到输出端口的标签：

| 行 ID | 输出标签 |
|------|-----------|
| 1 | 数字 组织 联系人 |
| 2 | 数字 联系人 组织 |
| 3 | 数字 组织 联系人 |
| 4 | 数字 联系人 组织 |
| 5 | 数字 组织 联系人 |
| 6 | 组织 数字 联系人 |
| 7 | 组织 数字 联系人 |
| 8 | 联系人 组织 数字 |

解析器转换示例

某超市在数据库表的单个列中存储产品说明。这些产品说明包含多个代表不同类型的信息的数据值。您是该超市的数据管理者。您需要为这些产品说明中不同类型的信息创建列。

您需要使用解析器转换配置一个映射，以便将数据值组织到正确的字段中。

以下数据片段包含橙汁的产品说明：

Sunnydream Orange Juice Unsweetened 12 oz

下表介绍了解析器转换从输入数据创建的输出数据：

| 产品名称 | 产品类型 | 产品详细信息 | 产品规格 |
|------------|--------------|-------------|-------|
| Sunnydream | Orange Juice | Unsweetened | 12 oz |

概率模型选项

编辑概率模型时，可以在“数据”视图或“标签”视图中工作。

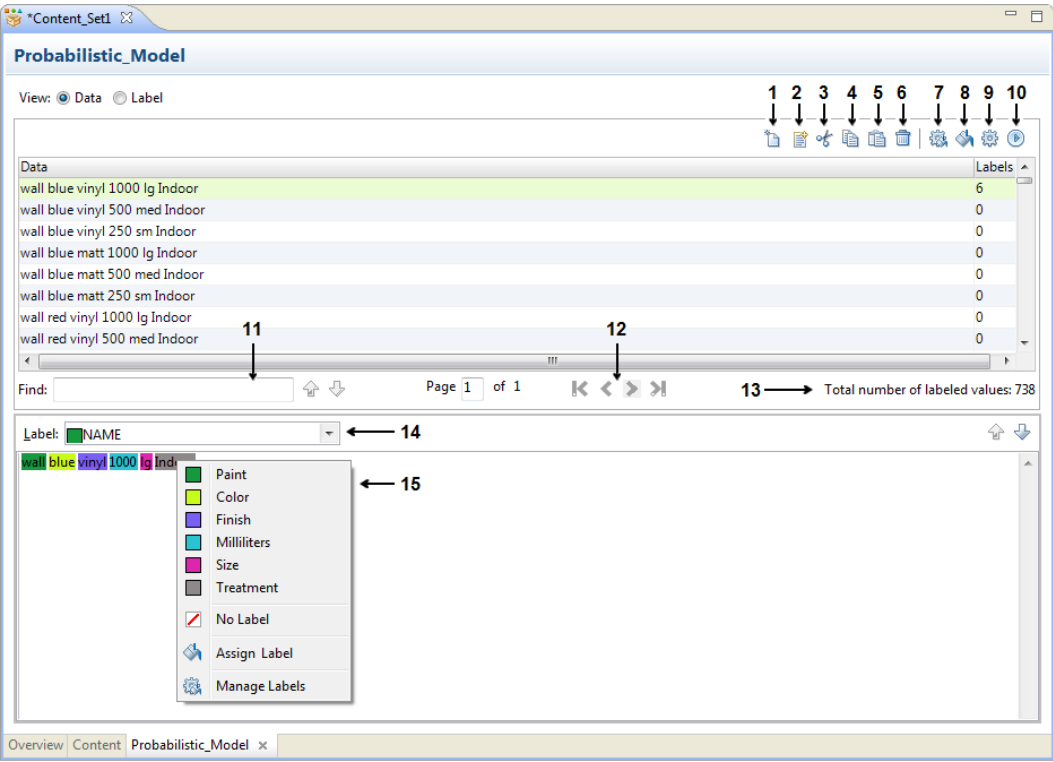
使用“数据”视图可将引用数据行添加到模型中，以及为每个行中的数据值分配标签。使用“标签”视图可查看有关标签值在模型中的使用情况的详细信息。您可以在“数据”视图和“标签”视图中向概率模型添加标签。

概率模型 “数据” 视图

“数据” 视图显示了概率模型中的引用数据行，以及为每个行分配的标签值的个数。“数据” 视图还显示您为当前模型中的值分配的标签总数。

选择引用数据行后，该行中的值会显示在“查找” 字段下方的编辑器中。要为行中的引用数据值分配标签，请在编辑器中右键单击该值并选择标签值。

下图显示了选择 “数据” 视图时可用的概率模型选项：



“数据” 视图包含以下选项：

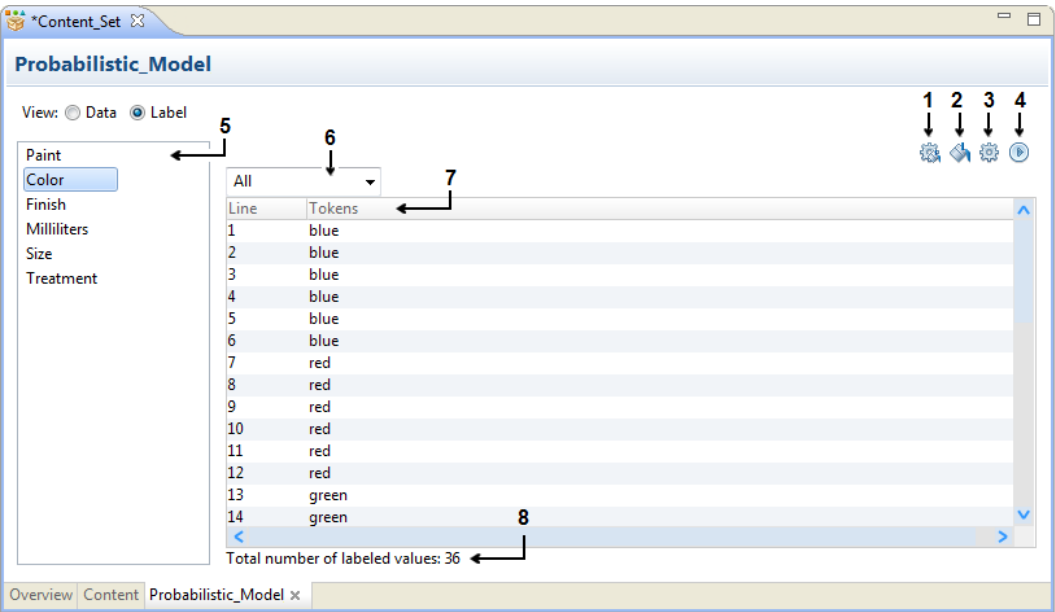
1. 添加行
插入空的数据行。
2. 附加数据。
将数据对象中的数据导入到模型存储库。
3. 剪切
将数据行从概率模型中删除并添加到剪贴板。
4. 复制
将数据行复制到剪贴板。
5. 粘贴
将剪贴板中的数据行粘贴到概率模型。
6. 删除
从概率模型中删除数据行。
7. 管理标签
打开**管理标签**对话框。您可以使用此对话框向概率模型添加标签值或从中删除标签值。

- 8. 分配标签
将标签分配给选定的一个或多个引用数据值。您可以使用此选项将标签分配给引用数据值在模型中的所有实例。
- 9. 编辑属性
显示概率模型属性。
- 10. 编译
对概率模型进行编译。
- 11. “查找” 字段
在模型中查找包含输入的引用数据值的行。可以使用向上箭头和向下箭头移到包含该值的行。
- 12. 向右箭头和向左箭头
在模型中的数据值行中向前和向后移动。
- 13. 标签值总数
指示使用某个标签的引用数据值的个数。
- 14. “标签” 字段
显示可以应用到所选引用数据值的标签值。
- 15. 标签菜单
显示可用于将标签分配给一个或多个引用数据值的选项的列表。要打开该菜单，请在引用数据编辑器中右键单击引用数据值。

概率模型 “标签” 视图

“标签” 视图列出您在概率模型中定义的标签。选择一个标签后，“标签” 视图会显示您为每个行中的标签分配的数据值。

下图显示了选择 “数据” 视图时可用的概率模型选项：



“标签”视图包含以下选项：

1. 管理标签
打开**管理标签**对话框。您可以使用此对话框向概率模型添加标签值或从中删除标签值。
2. 分配标签
将标签分配给选定的一个或多个引用数据值。
在单次操作中，可以将标签分配给单个数据值，也可以将标签分配给多个值。
3. 编辑属性
显示概率模型属性。
4. 编译
对概率模型进行编译。
5. 标签值的列表
列出可以分配给模型中的引用数据值的标签。
6. 分配筛选器
筛选使用所选标签的引用数据值的列表。筛选器选项会根据您用于为数据值分配标签的方法显示或隐藏引用数据值。
应用筛选器时，“标签”视图中带标签值的总数反映满足筛选条件的值数量。
7. 引用数据值列
列出使用当前标签的引用数据值。
8. 标签值总数
指示使用当前标签的引用数据值的个数。

概率模型引用数据

概率模型中的引用数据值代表可以连接到映射中的转换的输入数据类型。

您可以在 Developer 工具中添加、编辑和删除引用数据行。您可以从剪贴板粘贴数据，也可以从数据源导入数据。添加引用数据值后，应为每个行中的每个数据值分配标签。

概率模型标签数据

概率模型中的标签值代表引用数据值可能包含的信息类型。将引用数据行添加到模型时，应为每个行中的每个值分配标签。添加到模型的标签将显示在“标签”视图以及“数据”视图的菜单选项中。

可以将模型中的任何标签分配给任何引用数据值。如果同一个值在引用数据的不同的行中具有不同的含义，您可以在每一行中为该值分配不同的标签。

标签值的范围可与标签创建器转换或解析器转换在概率分析期间读取的输入端口的范围一致。概率模型必须包含至少一个标签值，以便转换可将标签值应用到每个输入端口上的数据值。

例如，某个仓库可能将库存数据存储在定义了八个列的逗号分隔文件中。您可以设计将库存数据解析到数据库表的映射，并创建每个数据列都包含标签值的概率模型。当您运行该映射时，解析器转换会将输入数据中的每个值写入到目标表中正确的列。

下表显示了您可能在概率模型中创建的库存数据列和标签值：

| 库存列名称 | 标签名称 |
|--------------|--------------------|
| Product_Name | Product_Name |
| 数量 | 数量 |
| 位置 | 位置 |
| 条形码 | 条形码 |
| SKU | Stock_Keeping_Unit |
| Arrival_Date | Arrival_Date |
| Cost_Price | Cost_Price |

注意：您可以使用输入列名称或其他名称。名称不必匹配。

溢出标签

当转换无法将标签应用到输入数据值时，它会将该数据值视为溢出数据。标签创建器转换会为其无法识别的任何数据值应用溢出标签。解析器转换会将其无法识别的任何数据值写入到溢出端口。

下表显示了解析器转换如何使用溢出端口解析概率模型无法识别的地址数据元素：

| 输入数据 | Street_Name 端口 | Street_Descriptor 端口 | 溢出端口 |
|------------------------|----------------|----------------------|--------|
| Park Place | Park | Place | 无溢出数据 |
| Park Avenue | Park | Avenue | 无溢出数据 |
| Madison Avenue | Madison | Avenue | 无溢出数据 |
| Central Park | Central | Park | 无溢出数据 |
| Washington Square Park | 华盛顿 | Square | Park |
| Madison Square Garden | Madison | Square | Garden |

当输入值的个数大于模型中的标签数时，解析器转换也会将值写入到溢出端口。在转换中使用概率模型之前，请检查输入数据并验证该模型包含的标签值个数是否正确。

概率模型属性

您可以检查概率模型的常规属性和高级属性。

要打开属性编辑器，请在“数据”视图或“标签”视图中选择**编辑属性**选项。

常规属性显示概率模型的名称、模型的任何说明以及模型数据文件的名称。高级属性显示 Developer 工具用于编译概率模型的计算属性。

概率模型编译中的基本元素是 *n-gram*。n-gram 是一系列位于其他字母之后或之前以完成某个单词的字母。当映射运行时，标签创建器转换或解析器转换将为概率模型引用数据列中的每个值创建多个 n-gram。转换将输入数据值与引用数据值和 n-gram 进行比较。概率模型的高级属性决定概率模型如何处理 n-gram 和其他模型特征。

注意：高级属性的默认值代表概率分析和概率模型编译的首选设置。如果您编译高级属性，则可能会对概率分析的准确性产生不利影响。除非您了解所进行的更改的作用，否则不要编辑高级属性。

相关主题：

- [“概率模型和分类器模型的规则和准则” 页面上 37](#)

概率模型配置

配置概率模型的第一步是选择要执行的分析类型。在标签创建器转换中使用概率模型可识别输入字符串中每个值包含的信息类型。在解析器转换中使用概率模型可将输入字符串中的数据值解析到不同的输出端口。

您可以使用同一个概率模型来创建数据标签和解析数据。在标签创建器转换中使用该模型时，转换会为选定的每个输入端口创建单个输出端口。在解析器转换中使用该模型时，转换会为选定的每个输入端口创建单个输出端口。

要创建概率模型，请执行以下任务：

1. 确定要添加到模型中的引用数据值和标签值。
可以使用要分析的数据片段。请在模型存储库中创建读取该数据片段的数据对象。
 2. 创建内容集，并将概率模型添加到该内容集。
 3. 将引用数据值添加到模型中。
 4. 将标签值添加到模型中。
您可以将数据对象中的数据导入到模型存储库，也可以输入单个引用数据行或标签。
要使用概率模型来解析数据，请验证该模型是否为转换必须创建的每个输出端口分配了标签值。
 5. 为每个行中的每个引用数据值分配标签。
您可以通过单个操作为多个引用数据值分配标签。
 6. 对模型进行编译。
- 对概率模型进行编译后，便可以在转换中使用该模型。

创建空的概率模型

您可以创建不包含任何引用数据或标签数据的概率模型对象。请创建空模型，然后将数据添加或导入到该模型。

1. 在对象浏览器中，打开或创建一个内容集。
2. 选择内容视图。
3. 选择**概率模型**，然后单击**添加**。
此时将打开概率模型向导。
4. 选择**概率模型**选项。
单击**下一步**。
5. 输入概率模型的名称。

或者，输入模型的文本说明。

6. 单击**完成**。

从数据对象创建概率模型

可以将数据对象用作概率模型数据的源。

将标签创建器转换或解析器转换的输入数据用作模型引用数据的源时，概率模型的性能最佳。

1. 在对象浏览器中，打开或创建一个内容集。
2. 选择**内容**视图。
3. 选择**概率模型**，然后单击**添加**。

此时将打开概率模型向导。

4. 选择**来自数据对象的概率模型**选项。
单击**下一步**。

5. 输入概率模型的名称。

或者，输入模型的文本说明。

6. 浏览模型存储库并选择包含要导入的数据的数据对象。

请勿选择社交媒体数据对象。

单击**下一步**。

7. 检查数据对象中的列，然后选择一个或多个列以添加到模型中。可以在同一个操作中添加引用数据列和标签列。

- 要将数据列导入为引用数据，请选择列名称并单击**数据**。

可以选择多个数据列。Developer tool 会将所选列的内容合并为单个列。

- 要将数据列导入为标签值，请选择列名称并单击**标签**。

导入引用数据和标签值时，Developer tool 会将每一行上的标签分配给同一行上的引用数据字符串。选择列之前，您可以预览数据。创建模型后，您可以更改标签分配。

单击**下一步**。

8. 选择要从数据源导入的行数。

默认情况下，Developer tool 会导入数据源中的所有行。如果输入数字，则模型会从数据集的开头对行进行计数。

9. 为导入的数据值指定分隔符。

可为引用数据值和标签值指定不同的分隔符。默认分隔符是一个字符空格。

10. 单击**完成**并保存模型。

创建概率模型后，请验证标签分配并对模型进行编译。

将数据源中的数据附加到概率模型

您可以通过单个操作将多行引用数据值和标签值导入到概率模型。

1. 打开包含该概率模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 单击**附加数据**。

此时将打开概率模型向导。

4. 浏览模型存储库并选择包含要导入的数据的数据对象。
请勿选择社交媒体数据对象。
单击**下一步**。
5. 检查数据对象中的列，然后选择一个或多个列以添加到模型中。可以在同一个操作中添加引用数据列和标签列。
 - 要将数据列导入为引用数据，请选择列名称并单击**数据**。
可以选择多个数据列。Developer tool 会将所选列的内容合并为单个列。
 - 要将数据列导入为标签值，请选择列名称并单击**标签**。导入引用数据和标签值时，Developer tool 会将每一行上的标签分配给同一行上的引用数据字符串。选择列之前，您可以预览数据。创建模型后，您可以更改标签分配。
单击**下一步**。
6. 选择要从数据源导入的行数。
默认情况下，Developer tool 会导入数据源中的所有行。如果输入数字，则模型会从数据集的开头对行进行计数。
7. 为导入的数据值指定分隔符。
可为引用数据值和标签值指定不同的分隔符。默认分隔符是一个字符空格。
8. 单击**完成**并保存模型。

向概率模型添加引用数据行

您可以使用“数据”视图向概率模型添加空行。

1. 打开包含该模型的内容集。
选择模型名称，然后单击**编辑**。
 2. 选择“数据”视图。
 3. 要向模型添加空行，请单击**新建**。
 4. 选择已添加的行，然后将一个或多个引用数据值输入到该行。
 5. 保存概率模型。
- 保存模型后，请为行中的每个值分配标签。或者，您也可以对模型进行编译。

将标签添加到概率模型

您可以向一个概率模型添加一个标签。为模型数据值代表的每一个信息类型添加一个标签。如果在解析器转换中使用概率模型，则为您希望该转换创建的每个输出端口添加一个标签。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 在“数据”视图或“标签”视图中，单击**管理标签**。
此时将显示**管理标签**对话框。
4. 在**管理标签**对话框中，单击**新建**。
此时将在该对话框的第一个空行中显示一个标签。
5. 编辑标签名称。（可选）更新标签的颜色。
6. 单击**确定**将标签添加到模型中。
7. 保存概率模型。

添加标签后，至少将标签分配给一个数据值。

为引用数据值分配标签

可为引用数据行中的单个数据值分配标签。

您可为出现在同一行中不同位置或不同行中的同一个数据值分配不同的标签。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 选择“数据”视图。
4. 查找没有标签或标签不正确的数据值。使用标签的数据值用颜色标记。
5. 选择包含该数据值的数据行。
此时将在编辑器中显示该行。
6. 在编辑器中右键单击某数据值并从上下文菜单中选择一个标签。
Developer 工具将为该数据值分配此标签。
7. 保存概率模型。

保存概率模型后，可以选择对模型进行编译。

为多个数据值分配标签

您可以通过单个操作为多个引用数据值分配标签。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 单击**分配标签**。
此时将显示**将标签分配给多个值**对话框。
4. 在“查找”字段中输入一个或多个字符。
可以在“查找”字段中输入通配符。
5. （可选）选择其他搜索条件。
可以选中或清除以下选项：
 - 匹配大小写。
指定搜索操作区分大小写。请勿将通配符与此选项结合使用。
 - 匹配完整字符串。指定搜索操作查找引用数据值中的字符和输入的字符之间的完整匹配项。请勿将通配符与此选项结合使用。
 - 忽略标签值。
指定搜索操作跳过任何使用标签的引用数据值。
6. 选择一个标签以分配给与搜索条件匹配的引用数据值。
您也可以选择**无标签**选项。选择此选项可从包含所输入的字符的引用数据值删除标签。
7. 单击**开始**。
Developer tool 会将标签分配给所有与定义的搜索条件匹配的引用数据值。
注意：要查看您在单次操作中标记的引用数据值，请在“标签”视图中使用**已批量分配**筛选器。

从概率模型中删除行

您可以通过单个操作从概率模型中删除一个或多个引用数据行。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 在“数据”视图中，选择一个或多个引用数据行。
4. 单击**删除**。

Developer tool 将从分类器模型中删除所选行。

要撤消操作，请按键盘上的 Ctrl+Z。

从概率模型中删除标签

当您从模型中删除标签值时，任何使用该标签的引用数据值将保留在模型中。请为每个引用数据值分配另一个标签值。

1. 打开包含该模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 在“数据”视图或“标签”视图中，单击**管理标签**。
4. 在**管理标签**对话框中，选择一个标签值。
5. 单击**删除**。
6. 单击**确定**删除标签。
7. 保存概率模型。

注意：标签是概率模型中的结构元素。如果您在将模型添加到转换之后添加或删除标签，使用该模型的操作将变得无效。要使用您更新的模型，请删除并重新创建该转换操作。

编译概率模型

更新概率模型中的数据或标签分配时，您可以对该模型进行编译。对模型进行编译可用当前引用数据值和当前标签值之间的关联来更新模型逻辑。

对概率模型进行编译之前，请验证每个标签值是否标识至少一个引用数据值。

- 要编译模型，请在 Developer 工具中打开模型并单击**编译**。

在概率模型中查找数据行

使用“数据”视图可查找包含输入的值的引用数据行。

1. 打开包含该概率模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 选择“数据”视图。
4. 在**查找**字段中输入一个或多个字符。
“数据”视图显示模型中第一个包含您输入的值的行。
5. 使用向上箭头或向下箭头移到包含该值的其他行。

按标签分配筛选引用数据值

使用“标签”视图查找使用所指定标签的引用数据值。根据用于分配标签的方法筛选结果。

1. 打开包含该概率模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 在“标签”视图中，选择标签值。

概率模型将显示使用该标签的引用数据值的列表。模型还会显示使用该标签的数据值的数量。

4. 向使用此标签的引用数据值列表应用筛选器。

选择下列一种筛选器：

- 全部。列出使用标签的引用数据值。“全部”为默认选项。
- 用户已分配。显示您分配标签时逐个选定的任何引用数据值。
- 已批量分配。显示批量分配操作过程中向其分配标签的引用数据值。

概率模型显示满足筛选条件的引用数据值。

查找未使用的标签值

使用“标签”视图可查找任何未分配给引用数据值的标签值。您必须将每个标签分配给至少一个引用数据值。

1. 打开包含该概率模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 在“标签”视图中，选择标签值。

概率模型将显示使用该标签的引用数据值的列表。模型还会显示使用该标签的数据值的总数。

如果数据值的总数为零，则表示您未将该标签分配给概率模型中的任何引用数据值。

复制和粘贴操作

在模型存储库中，可以将一个内容集中的概率模型复制到另一个内容集中。复制概率模型，以便与其他 Developer tool 用户共享资源。

可以将模型复制到其他内容集，也可以将模型导入当前内容集。可以在单个操作中从存储库中的多个内容集导入多个模型。

复制模型时，内容管理服务会在 Informatica 服务主机计算机上创建一份模型数据文件的副本。每个模型使用不同的数据文件。

将概率模型复制到其他内容集

在模型存储库中，可以将一个内容集中的概率模型复制到另一个内容集中。复制概率模型时，应指定模型对象以及源内容集和目标内容集。

1. 打开包含该概率模型的内容集。
2. 选择一个概率模型并单击**复制到**。
3. 浏览模型存储库并选择一个内容集。

可以将概率模型复制到当前项目或其他项目中的内容集。

4. 单击**确定**。

Developer 工具将概率模型复制到所选的内容集中。

从其他内容集导入概率模型

在模型存储库中，可以将一个内容集中的概率模型导入另一个内容集。导入概率模型时，应指定一个或多个模型对象以及源内容集和目标内容集。

1. 打开要包含该概率模型的内容集。
2. 选择一个概率模型并单击**粘贴自**。
3. 浏览模型存储库并选择一个概率模型。
可以从当前项目或其他项目中的内容集粘贴概率模型。
4. 单击**确定**。

Developer 工具将概率模型粘贴到当前内容集中。

将引用数据行复制到剪贴板

您可以将概率模型中的一个或多个引用数据行复制到剪贴板，还可以将行粘贴到其他概率模型。

1. 打开包含该概率模型的内容集。
2. 选择模型名称，然后单击**编辑**。
3. 在“数据”视图中，选择一个或多个引用数据行。
4. 按 Ctrl+C 将行复制到剪贴板。

此操作会复制引用数据以及为引用数据分配的标签值。

按 Ctrl+V 可将行粘贴到文本编辑器或其他概率模型的“数据”视图。

附录 A

引用数据和 Informatica Big Data Management

本附录包括以下主题：

- [引用数据和 Informatica Big Data Management 概览, 64](#)

引用数据和 Informatica Big Data Management 概览

Informatica Big Data Management(R) 是一个大数据解决方案，它将 Informatica 域和客户端应用程序与 Hadoop 群集结合使用。您可以从 Developer tool 将映射下推至群集，然后在群集中的节点上运行映射。

当推送的映射包含用于读取引用数据的转换时，下推操作可以复制转换所使用的任何引用数据。下推操作会将引用表数据、内容集数据和标识填充数据复制到群集。在映射运行后，群集会删除下推操作随该映射一起复制的引用数据。

注意：下推操作不会复制地址验证引用数据。如果推送用于执行地址验证的映射，您必须在运行该映射的每个数据节点上安装地址验证引用数据文件。在地址验证映射运行后，群集不会删除地址验证引用数据文件。

用于地址验证的引用数据

在 Hadoop 环境中运行地址验证映射时，地址引用数据文件必须驻留在运行该映射的每个数据节点上。Informatica Big Data Management 会安装一个 shell 脚本，您可以使用该脚本在数据节点上安装文件。

使用 shell 脚本可通过单个操作在数据节点上安装地址引用数据文件。该脚本会读取一个文件，其中包含节点的名称或 IP 地址。该脚本会将地址引用数据文件复制到此文件标识的每个节点。

该脚本的名称是 `copyRefDataToComputeNodes.sh`。

可以在 Informatica Big Data Management 安装中的以下目录找到该脚本：

`<Informatica 安装目录>/tools/dq/av`

下表描述了该脚本使用的选项：

| 选项 | 说明 |
|----|--|
| -n | 一个文件，其中包含 Hadoop 群集中的数据节点的名称或 IP 地址的列表。请在文件中单独的行上输入每个节点名称或 IP 地址。 默认情况下，该脚本将从 <code>\$BASEDIR/HadoopDataNodes</code> 目录（其中 <code>\$BASEDIR</code> 是 shell 脚本的位置）读取文件。 |
| -p | 一个提示，提示您确认是否要安装地址引用数据文件。 默认情况下，该脚本会显示一个提示，提示您确认是否要将源目录中的文件复制到数据节点上的目标目录。如果按计划运行 shell 脚本，则可以禁用提示。 默认选项值为 Y。要禁用提示，请将值设置为 N。 |
| -s | 该脚本复制到节点的地址引用数据文件的源目录。 默认情况下，该脚本将从本地计算机上的 <code>/reference_data</code> 目录读取文件。 注意： 地址引用数据文件使用文件扩展名 <code>.MD</code> 。源目录必须仅包含地址引用数据文件，不能包含任何其他文件。 |
| -t | 每个节点上的一个目录，该脚本会将地址引用数据文件复制到此目录。 默认情况下，该脚本会将文件复制到每个节点上的 <code>/reference_data</code> 目录。 |
| -u | 运行该脚本的用户的用户名。用户必须拥有节点的无密码安全 shell 访问权限。 |

安装地址引用数据文件

要在 Hadoop 群集中的数据节点上安装地址引用数据文件，请运行 `copyRefDataToComputeNodes.sh` shell 脚本。或者，定义一个作业，以便按指定的时间间隔在作业计划程序应用程序中运行 shell 脚本。

在运行脚本或定义作业之前，请检查为脚本指定的选项值。您可以接受默认值，也可以更新值。

通过命令提示符安装地址引用数据文件

要通过命令提示符安装这些文件，请执行以下步骤：

1. 在命令提示符下，打开以下目录：
`<Informatica 安装目录>/tools/dq/av`
2. 运行 `copyRefDataToComputeNodes.sh`。
（可选）为脚本选项输入一个或多个值。如果不为选项输入值，脚本将使用该选项的默认值运行。
默认情况下，脚本会提示您确认是否要安装文件。要安装文件，请输入 Y。

通过已计划作业安装地址引用数据文件

您可以定义一个作业，以便按指定的时间间隔运行 shell 脚本。将该作业添加到作业计划程序应用程序中。如果通过定义一个作业来安装文件，您必须禁用确认安装的提示。

要禁用提示，请在 shell 脚本上设置以下选项：

`-p n`

索引

A

Analyst 工具
查找和替换引用数据值 [23](#)

B

版本控制
Analyst 工具中的引用表 [21](#)
Developer tool 中的引用表 [28](#)
内容集 [13](#), [28](#)
引用表 [13](#)
标志集 [35](#)
Big Data Management
安装地址引用数据 [65](#)
地址引用数据安装脚本 [64](#)
引用数据的要求 [64](#)

C

查看审计表事件
引用表 [25](#)
从列模式创建引用表
引用表 [18](#)
从配置文件列数据创建引用表
引用表 [16](#)

D

导出引用表
引用表 [23](#)
导入引用表
引用表 [19](#)

F

非受管引用表
定义 [12](#)
启用和禁用编辑 [24](#)
与模型存储库同步 [12](#)
分类器模型
规则和准则 [37](#)
位于内容集中 [34](#)

G

概率模型
规则和准则 [37](#)
位于内容集中 [34](#)
管理列
引用表 [22](#)

管理行
引用表 [22](#)

H

Hadoop 环境
安装地址引用数据 [65](#)
地址引用数据安装脚本 [64](#)
引用数据的要求 [64](#)

M

模式集 [34](#)

N

内容管理服务
引用表特权 [12](#)
内容集
版本控制 [13](#), [21](#), [28](#)
标志集 [35](#)
分类器模型 [34](#)
概率模型 [34](#)
模式集 [34](#)
正则表达式 [35](#)
字符集 [33](#)

S

手动创建引用表
引用表 [16](#)
受管引用表 [12](#)

T

特权
内容管理服务 [12](#)

Y

引用表
Analyst 工具概览 [14](#)
Analyst 工具中的属性 [14](#)
版本控制 [13](#), [21](#), [28](#)
查看审计跟踪表 [25](#)
从列模式创建引用表 [18](#)
从配置文件列创建引用表 [16](#)
导出引用表 [23](#)
导入引用表 [19](#)
Developer tool 概览 [29](#)

引用表 (续)

Developer tool 中的属性 [29](#)

非受管引用表 [12](#)

管理列 [22](#)

管理行 [22](#)

基于模式的解析 [12](#)

内容管理服务 [12](#)

手动创建引用表 [16](#)

受管和非受管 [12](#)

受管引用表 [12](#)

特权 [12](#)

引用表 (续)

引用数据仓库 [12](#)

在 Analyst 工具中查找和替换值 [23](#)

在 Analyst 工具中刷新 [24](#)

Z

正则表达式 [35](#)

字符集 [33](#)