



Informatica®
10.4.0

Data Discovery 指南

Informatica Data Discovery 指南

10.4.0

2019 年 12 月

© 版权所有 Informatica LLC 2011, 2020

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。

Informatica 和 Informatica 标志是 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。

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。

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。

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: 2020-02-04

目录

前言	13
Informatica 资源	13
Informatica Network	13
Informatica 知识库	13
Informatica 文档	13
Informatica 产品可用性矩阵	14
Informatica Velocity	14
Informatica Marketplace	14
Informatica 全球客户支持部门	14
第 I 部分：Data Discovery 简介	15
第 1 章：剖析简介	16
剖析概览	16
剖析体系结构	18
数据发现过程	19
第 2 章：数据发现	20
数据发现概览	20
配置文件和分析类型	20
剖析组件	21
配置文件结果	22
第 3 章：列配置文件概念	23
列配置文件概念概览	23
列配置文件选项	24
存储库配置文件锁定和受版本控制的配置文件管理	24
结果卡	24
第 4 章：数据域发现概念	25
数据域发现概念概览	25
数据域	26
数据域组	26
数据域词汇表	26
数据域发现过程	27
Spark 引擎上的数据域发现	27
第 5 章：内容管理概念	28
内容管理概念概览	28
分析师和开发人员的内容管理	28

内容管理任务.	29
第 II 部分：使用 Informatica Analyst 的 Data Discovery.	30
第 6 章：Informatica Analyst 中的列配置文件.	31
Informatica Analyst 中的列配置文件概览.	31
列剖析过程.	32
配置文件选项.	32
采样选项.	32
向下钻取选项.	33
运行时环境.	33
本地环境.	34
Hadoop 环境.	34
Informatica Analyst 中的操作系统配置文件概览.	34
选择操作系统配置文件.	35
存储库资产锁定和基于团队的开发概览.	35
在 Informatica Analyst 中创建列配置文件.	35
编辑列配置文件.	36
运行配置文件.	37
在 Spark 引擎上运行配置文件.	37
同步选项.	37
在 Informatica Analyst 中同步平面文件数据对象.	38
在 Informatica Analyst 中同步关系数据对象.	39
第 7 章：Informatica Analyst 中的规则.	41
Informatica Analyst 中的规则概览.	41
预定义规则.	41
预定义规则进程.	42
应用预定义规则.	42
表达式规则.	42
创建表达式规则.	43
使用规则规范创建表达式规则.	44
第 8 章：Informatica Analyst 中的筛选器.	46
Informatica Analyst 中的筛选器概览.	46
创建筛选器.	46
创建简单筛选器.	47
创建高级筛选器.	48
创建 SQL 筛选器.	49
管理筛选器.	49
第 9 章：Informatica Analyst 中的列配置文件结果.	51
Informatica Analyst 中的列配置文件结果概览.	51

摘要视图.	52
摘要视图属性.	52
摘要视图中的默认筛选器.	53
详细视图.	54
详细视图窗格.	54
统计信息.	55
数据预览.	55
数据类型.	56
离群值.	56
模式.	57
值.	58
配置文件运行的类型.	61
最新配置文件运行.	61
历史配置文件运行.	61
已合并配置文件运行.	61
选择配置文件运行.	61
比较多个配置文件结果概览.	62
比较多个配置文件结果.	62
配置文件结果比较的摘要视图.	63
配置文件结果比较的详细视图.	65
列配置文件向下钻取.	66
向下钻取行数据.	67
应用筛选器以向下钻取数据.	67
Analyst 工具中的内容管理.	67
批准数据类型和数据域.	67
拒绝数据类型和数据域.	68
Informatica Analyst 中的列配置文件导出文件.	68
CSV 文件格式的配置文件导出结果.	68
Microsoft Excel 格式的配置文件导出结果.	68
从 Informatica Analyst 中导出配置文件结果.	69
 第 10 章：Informatica Analyst 中的业务术语、注释和标记.	 70
Informatica Analyst 中的业务术语、注释和标记概览.	70
业务术语.	70
为列分配业务术语.	70
注释.	71
向配置文件或列添加注释.	71
标记.	71
向配置文件或列添加标记.	72
 第 11 章：Informatica Analyst 中的结果卡.	 73
Informatica Analyst 结果卡概览.	73
Informatica Analyst 结果卡进程.	74

在 Informatica Analyst 中创建结果卡.	74
向现有结果卡添加列.	76
向现有结果卡添加列.	76
运行结果卡.	77
查看结果卡.	77
编辑结果卡.	77
度量.	78
度量权重.	78
数据质量的值.	78
定义阈值.	79
度量组.	79
创建度量组.	79
将得分移至度量组.	80
编辑度量组.	80
删除度量组.	80
对列进行向下钻取.	81
趋势图表.	81
得分趋势图表.	82
成本趋势图表.	82
查看趋势图表.	83
导出趋势图表.	83
Informatica Analyst 中的结果卡仪表板.	84
结果卡(按项目).	85
结果卡运行趋势.	86
具有结果卡的数据对象.	87
累积度量趋势.	88
Informatica Analyst 结果卡导出文件.	89
从 Informatica Analyst 导出结果卡结果.	89
Microsoft Excel 格式的结果卡导出结果.	90
结果卡通知.	90
通知电子邮件模板.	91
设置结果卡通知.	92
配置结果卡通知的全局设置.	92
结果卡沿袭.	93
在 Informatica Analyst 中查看结果卡沿袭.	93
第 12 章：Informatica Analyst 中的数据域发现.	94
Informatica Analyst 中的数据域发现概览.	94
Informatica Analyst 中的数据域词汇表.	94
在 Informatica Analyst 中创建数据域组.	95
在 Informatica Analyst 中创建数据域.	95
在 Informatica Analyst 中基于配置文件结果创建数据域.	96
在 Informatica Analyst 中查找数据域和数据域组.	96

Informatica Analyst 中的数据域发现选项.	96
Informatica Analyst 中的数据域列选择.	97
Informatica Analyst 中的数据域选择.	97
Informatica Analyst 中的数据域推理选项.	97
在 Informatica Analyst 中创建自定义配置文件以执行数据域发现.	99
在 Informatica Analyst 中编辑列配置文件和数据域发现.	100
运行配置文件以执行数据域发现.	100
Informatica Analyst 中的数据域发现结果.	101
批准数据域.	101
拒绝数据域.	101
Informatica Analyst 中的数据域发现导出文件.	102
Microsoft Excel 中的数据域发现结果.	102
从 Informatica Analyst 中导出数据域发现结果.	102
 第 13 章： Informatica Analyst 中的企业发现.	 103
Informatica Analyst 中的企业发现概览.	103
Informatica Analyst 中的企业发现过程.	103
企业发现的配置选项.	104
数据域发现设置.	104
列配置文件设置.	105
在 Informatica Analyst 中创建企业发现配置文件.	105
编辑企业发现选项.	106
 第 14 章： Informatica Analyst 中的企业发现结果.	 108
Informatica Analyst 中的企业发现结果概览.	108
摘要视图.	109
摘要视图配置文件结果.	109
查看数据域发现结果.	110
查看列配置文件结果.	110
数据类型冲突.	111
查看数据类型冲突.	111
配置文件视图.	111
查看配置文件属性.	111
 第 15 章： Informatica Analyst 中的发现搜索.	 112
在 Informatica Analyst 中执行发现搜索概览.	112
发现搜索必备条件.	113
Informatica Analyst 中的发现搜索过程.	113
发现搜索选项.	113
发现搜索条件.	114
搜索资产.	114
Informatica Analyst 中的发现搜索结果.	114
发现搜索结果面板.	115

筛选发现搜索结果.	116
匹配类型.	116
直接匹配.	116
间接匹配.	116
查看匹配项信息.	116
打开发现搜索结果中的资产.	117
相关资产.	117
每种资产类型的相关资产.	117
查看相关资产.	118
常见问题.	118
 第 16 章： Informatica Analyst 中的 Business Glossary 桌面版.	119
业务术语.	119
管理 Metadata Manager Business Glossary 中的业务术语.	120
在 Business Glossary 桌面版中查找业务术语.	120
 第 III 部分： 使用 Informatica Developer 执行数据发现.	121
 第 17 章： Informatica Developer 配置文件.	122
Informatica Developer 配置文件概览.	122
Informatica Developer 配置文件视图.	123
存储库对象锁定和使用受版本控制的对象进行基于团队的开发.	124
 第 18 章： 数据对象配置文件.	125
数据对象配置文件概览.	125
Informatica Developer 中的列配置文件.	126
筛选选项.	126
采样选项.	127
运行时环境.	128
本地环境.	128
Hadoop 环境.	128
主键发现.	129
主键推理属性.	129
推理的主键属性.	130
键冲突属性.	130
功能相关性发现.	130
功能相关性推理属性.	131
推理的功能相关性属性.	131
功能相关性冲突属性.	132
Informatica Developer 中的操作系统配置文件.	132
选择操作系统配置文件.	132
在 Informatica Developer 中创建单个数据对象配置文件.	132
在 Informatica Developer 中创建多个数据对象配置文件.	133

编辑配置文件.	134
同步选项.	134
在 Informatica Developer 中同步平面文件数据对象.	134
在 Informatica Developer 中同步关系数据对象.	136
注释.	136
在 Informatica Developer 添加注释.	137
 第 19 章：基于半结构化数据源的列配置文件.	 138
基于半结构化数据源的列配置文件概览.	138
JSON 和 XML 数据对象.	138
从 JSON 或 XML 数据源创建数据对象.	139
HDFS 中半结构化数据源的复杂文件数据对象.	139
HDFS 中来自 JSON 或 XML 数据源的复杂文件数据对象.	139
HDFS 中来自 Avro 或 Parquet 数据源的复杂文件数据对象.	140
创建 HDFS 连接.	140
从 HDFS 中的 JSON 或 XML 文件创建复杂文件数据对象.	140
从 Avro 或 Parquet 数据源创建复杂文件数据对象.	141
创建基于半结构化数据源的列配置文件.	142
 第 20 章：Informatica Developer 中的规则.	 144
Informatica Developer 中的规则概览.	144
在 Informatica Developer 中创建规则.	145
在 Informatica Developer 中应用规则.	145
 第 21 章：Mapplet 和映射剖析.	 146
Mapplet 和映射剖析概览.	146
对 Mapplet 或映射对象运行配置文件.	146
比较映射或 Mapplet 对象的配置文件.	147
从配置文件生成映射.	147
 第 22 章：Informatica Developer 中的列配置文件结果.	 148
Informatica Developer 中的列配置文件结果.	148
列值属性.	149
列模式属性.	149
列统计信息属性.	149
列数据类型属性.	150
Informatica Developer 中的内容管理.	150
批准数据类型.	151
拒绝数据类型.	151
从 Informatica Developer 中导出配置文件结果.	151
 第 23 章：Informatica Developer 中的结果卡.	 153
Informatica Developer 中的结果卡概览.	153

创建结果卡.	153
为结果卡沿袭导出资源文件.	154
查看 Informatica Developer 中的结果卡沿袭.	154
 第 24 章： Informatica Developer 中的数据域发现.	155
Informatica Developer 中的数据域发现概览.	155
Informatica Developer 中的数据域词汇表.	155
在 Informatica Developer 中创建数据域组.	156
在 Informatica Developer 中创建数据域.	156
在 Informatica Developer 中基于配置文件结果创建数据域.	157
在 Informatica Developer 中查找数据域.	157
导入数据域.	158
导出数据域.	158
Informatica Developer 中的数据域发现选项.	159
Informatica Developer 中的数据域选择.	159
Informatica Developer 中的数据域列选择.	159
Informatica Developer 中的数据域推理选项.	160
在 Informatica Developer 中创建配置文件以执行数据域发现.	161
在 Informatica Developer 中编辑配置文件.	161
在 Informatica Developer 中运行配置文件以执行数据域发现.	162
Informatica Developer 中的数据域发现结果.	162
按数据域组查看.	163
按列查看.	163
确认结果.	163
批准数据域.	164
拒绝数据域.	164
从 Informatica Developer 中导出数据域发现结果.	164
 第 25 章： Informatica Developer 中的企业发现.	166
Informatica Developer 中的企业发现概览.	166
企业发现进程.	167
企业发现的配置文件选项.	167
企业发现的数据域选择.	167
企业发现的列配置文件采样选项.	168
运行时环境选项.	169
企业发现的主键推理选项.	169
企业发现的外键推理选项.	169
外键推理的自动内容管理参数.	170
在 Informatica Developer 中创建企业发现配置文件.	171
编辑配置文件.	172
运行企业发现配置文件.	173
外键发现.	173
定义父对象和子对象关系.	173

发现数据对象之间的外键关系.	174
外键分析结果.	174
联接分析.	175
创建联接配置文件.	175
联接分析结果.	175
将联接配置文件结果导出到文件.	176
重叠发现.	176
重叠发现结果.	176
发现重叠数据.	177
DDL 脚本文件.	178
从企业发现配置文件创建 DDL 脚本.	178
同步企业发现配置文件.	178
同步企业发现配置文件.	179
第 26 章：企业发现结果.	180
企业发现结果概览.	180
关系视图.	181
搜索数据对象.	181
导航到外键剖析视图.	182
外键剖析视图.	182
查看数据对象关系.	182
放大和缩小视图.	182
查找数据对象.	183
查看列关系.	183
将实体关系图另存为图像.	183
从“外键剖析”视图查看数据对象配置文件结果.	184
表格视图.	184
表详细信息窗格.	184
验证企业发现结果.	184
管理列关系.	185
将结果提交至模型存储库.	185
数据域视图.	185
查看数据域发现结果.	185
验证数据域发现结果.	186
对行进行向下钻取.	186
在数据域视图中查看数据对象配置文件结果.	186
列配置文件视图.	187
查看数据对象配置文件结果.	187
在企业发现运行期间查看列配置文件结果.	187
在企业发现运行期间查看数据域发现结果.	187
查看企业发现的运行时状态.	188
企业发现导出文件.	188
导出企业发现结果.	188

第 27 章： Informatica Developer 中的 Business Glossary 桌面版.....	189
Business Glossary 搜索.....	189
查找业务术语.....	189
自定义热键以查找业务术语.....	190
附录 A： 基于剖析仓库连接的功能支持.....	191
剖析功能支持.....	191
索引.....	192

前言

如果要了解如何创建和运行配置文件以分析数据源的内容、质量及结构，请阅读《*Informatica Data Discovery 指南*》。了解如何执行数据发现来发现一个或多个数据源的各列之间存在关系的源系统的元数据。可以使用 Developer tool 和 Analyst 工具来创建、管理及运行列配置文件、数据域发现配置文件或企业发现配置文件。

本指南的目标读者为数据分析师和开发人员。

Informatica 资源

Informatica 通过 Informatica Network 和其他在线门户为您提供一系列产品资源。使用这些资源，可以充分利用 Informatica 产品和解决方案，并向其他 Informatica 用户和主题专家学习。

Informatica Network

在 Informatica Network 中可以获得许多资源，包括 Informatica 知识库和 Informatica 全球客户支持。要进入 Informatica Network，请访问 <https://network.informatica.com>。

作为 Informatica Network 成员，您可以选择以下服务：

- 在知识库中搜索产品资源。
- 查看产品可用性信息。
- 创建并检查您的支持案例。
- 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 知识库可查找产品资源，例如操作方法文章、最佳实践、视频教程以及常见问题的答案。

要搜索知识库，请访问 <https://search.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

使用 Informatica 文档门户可浏览大量当前与最近产品版本的文档库。要浏览文档门户，请访问 <https://docs.informatica.com>。

如果您对产品文档有任何疑问、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、数据库以及数据源和目标的类型。您可以在以下网址中浏览 Informatica PAM:

<https://network.informatica.com/community/informatica-network/product-availability-matrices>。

Informatica Velocity

Informatica Velocity 是由 Informatica 专业服务根据数百个数据管理项目的实际经验所开发出来的，其中汇集了大量使用技巧和最佳实践。Informatica Velocity 代表了 Informatica 顾问的集体知识，这些顾问与世界各地的组织合作，共同计划、开发、部署和维护成功的数据管理解决方案。

您可以在以下网址中找到 Informatica Velocity 资源：<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扩展和增强您的 Informatica 实施。利用 Informatica 开发人员和合作伙伴在 Marketplace 中提供的数以百计的解决方案，可提高您的工作效率并加快项目实施时间。您可以在以下网址中找到 Informatica Marketplace：<https://marketplace.informatica.com>。

Informatica 全球客户支持部门

您可以通过电话或 Informatica Network 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：<https://www.informatica.com/services-and-training/customer-success-services/contact-us.html>。

要在 Informatica Network 上查找在线支持资源，请访问 <https://network.informatica.com>，然后选择 eSupport 选项。

第 I 部分： Data Discovery 简介

本部分包含以下章节：

- [剖析简介, 16](#)
- [数据发现, 20](#)
- [列配置文件概念, 23](#)
- [数据域发现概念, 25](#)
- [内容管理概念, 28](#)

第 1 章

剖析简介

本章包括以下主题：

- [剖析概览, 16](#)
- [剖析体系结构, 18](#)
- [数据发现过程, 19](#)

剖析概览

使用剖析查找应用程序、架构或企业的数据源的内容、质量和结构。数据源内容包括值频率和数据类型。数据源结构包括键和功能相关性。

在发现过程中，可以创建并运行配置文件。配置文件是一个存储库对象，可查找和分析企业内数据源中的所有数据不规范问题和使数据项目处于危险中的隐藏数据问题。通过在企业内的任何数据源中运行配置文件，可以很好地了解企业数据和元数据的优势和劣势。

可以使用 Informatica Analyst 和 Informatica Developer 分析源数据和元数据。分析人员和开发人员可以使用这些工具进行协作、识别数据质量问题以及分析数据关系。您可以根据自己的职位选择使用 Analyst 工具或 Developer tool 的功能。您可以执行的剖析度因使用工具的不用而有所不同。

您可以在 Developer tool 和 Analyst 工具中执行以下任务：

- 执行列剖析。该过程包括发现列中唯一值、空值和数据模式的数量。
- 执行数据域发现。您可以发现企业内的关键数据特性。
- 管理配置文件结果，包括数据类型、数据域、主键和外键。
- 创建结果卡以监视数据质量。
- 选择一个操作系统配置文件，然后根据在操作系统配置文件中定义的操作系统用户的权限，创建并运行列配置文件、企业发现配置文件以及结果卡。
- 使用存储库资产锁定可防止其他用户覆盖所做的工作。
- 使用版本控制系统可保存配置文件的多个版本。
- 创建标记并将其分配给数据对象。
- 在 Business Glossary Desktop 中查找对象名称作为业务术语的含义。例如，可以查找列名或配置文件名称的含义，以了解其业务要求和当前的实现。

可以在 Developer tool 中执行以下任务：

- 发现数据源中两个数据列之间的潜在联接度。
- 确定一个或多个数据源内列中成对重叠数据的百分比。

- 比较列剖析的结果。
- 从配置文件中生成映射对象。
- 发现数据源中的主键。
- 发现一个或多个数据源中的外键。
- 发现数据源中各列之间的功能相关性。
- 对多个连接中的大量数据源运行数据发现任务。数据发现任务包括列配置文件、主键和外键关系推理、数据域发现以及生成数据关系的合并图形摘要。

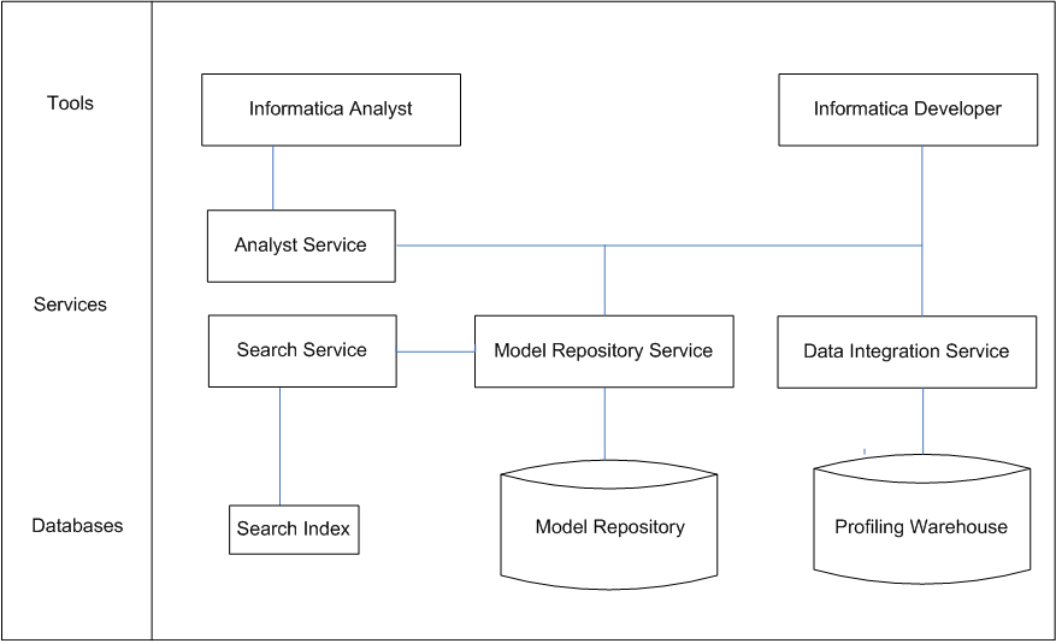
可以在 Analyst 工具中执行以下任务：

- 对多个连接中的大量数据源执行企业发现。可以查看列元数据和数据域的合并的发现结果摘要。
- 执行发现搜索以查找数据和元数据在企业中存在的位置。您可以搜索特定资产，例如数据对象、规则和配置文件。发现搜索查找资产并识别与数据库和企业架构中的其他资产之间的关系。
- 查看最新配置文件运行的配置文件结果。
- 比较列配置文件中两次配置文件运行的配置文件结果。
- 查看每个结果卡度量和度量组的结果卡沿袭。
- 查看结果卡仪表板。
- 向配置文件或配置文件中的列添加注释。
- 为配置文件或配置文件中的列分配标记。
- 为配置文件中的列分配业务术语。

剖析体系结构

剖析体系结构由工具、服务和数据库组成。工具部分包含客户端应用程序。服务组件包含管理工具、执行数据集成任务和管理配置文件对象的元数据所需的应用程序服务。数据库组件由模型存储库和剖析仓库组成。

下图显示了用于剖析的体系结构组件：



运行配置文件时，分析服务或 Developer tool 会从模型存储库服务中接收配置文件定义。然后，分析服务或 Developer tool 会在数据集成服务中调用剖析插件。接下来，剖析插件会处理配置文件作业并将该作业提交至数据集成服务。此时，数据集成服务便会生成剖析结果。然后，数据集成服务会将剖析结果写入到剖析仓库中。

发现搜索使用搜索服务。搜索服务会在搜索索引（而不是模型存储库或剖析仓库）中执行各项搜索。搜索服务会根据模型存储库和剖析仓库中的内容生成搜索索引。搜索服务包含提取器可提取每个存储库中的内容。

下表描述了体系结构组件：

组件	说明
Informatica Analyst	基于 Web 的客户端应用程序，可用来发现、分析和报告数据源的数据和元数据。
Informatica Developer	用于执行高级数据发现（如主键发现、外键发现和企业发现）的客户端应用程序。
分析服务	运行 Analyst 工具并管理服务组件和 Analyst 工具用户之间的连接的应用程序服务。
搜索服务	在 Analyst 工具中管理搜索的应用程序服务。默认情况下，搜索服务会从模型存储库中返回搜索结果，如数据对象、配置文件、映射规范、引用表、规则和结果卡。
搜索索引	位于自定义目录中的文件系统，用于存储搜索服务从模型存储库和剖析仓库中提取的索引内容。
模型存储库服务	管理模型存储库的应用程序服务。
数据集成服务	针对 Analyst 工具、Developer tool 和外部客户端执行数据集成任务的应用程序服务。

组件	说明
模型存储库	用于存储在 Analyst 工具或 Developer tool 中创建的项目的元数据的关系数据库。
剖析仓库	用于存储剖析信息的数据库，如配置文件结果和结果卡结果。

数据发现过程

开始数据集成项目时，剖析通常是第一步。可以创建配置文件以分析数据源的内容、质量和结构。作为剖析过程的一部分，可以发现数据源的元数据。

针对不同类型的数据分析使用不同的配置文件，如列配置文件、主键发现、外键发现和数据域发现。发现并记录数据质量问题。完成下列任务以执行数据发现：

1. 查找并分析数据源中的数据内容。包括数据类型、值频率、模式频率和数据统计信息，如最小值和最大值。
2. 发现数据的结构。包括键、功能相关性和外键。
3. 检查并验证配置文件结果。
4. 向下钻取配置文件结果。
5. 精细搜索配置文件结果。
6. 创建引用数据。
7. 记录数据问题。
8. 创建并运行规则。
9. 创建结果卡以监视数据质量。

可以使用以下工具管理发现过程：

Informatica Administrator

管理用户、组、权限和角色。可以管理分析服务并在 Informatica Analyst 中管理项目和对象的权限。可以在 Informatica Developer 中使用此工具控制访问权限。

Informatica Developer

在此工具中创建并运行配置文件以查找和分析包括发现各列间关系的一个或多个数据源的元数据。使用向导创建配置文件。

Informatica Analyst

在 Analyst 工具中可以在数据对象上运行列配置文件、执行数据域发现以及执行企业发现。运行配置文件后，可以对数据源中的数据行进行向下钻取。

第 2 章

数据发现

本章包括以下主题：

- [数据发现概览, 20](#)
- [配置文件和分析类型, 20](#)
- [剖析组件, 21](#)
- [配置文件结果, 22](#)

数据发现概览

数据发现是指发现源系统中包括内容和结构的元数据的过程。内容指的是数据值、频率和数据类型。结构包括候选键、主键、外键和功能相关性。您可以创建并运行配置文件以发现数据源的内容和结构。

您可以定义配置文件以分析单个数据对象或多个数据对象中的数据。向配置文件中添加注释，以便可以有效跟踪剖析的处理过程。

运行配置文件可以评估数据结构并验证数据列中是否包含您所需要的信息类型。您可以在已剖析的数据中对数据进行向下钻取。如果配置文件结果显示数据存在问题，则可以应用规则以修复结果集。可以创建结果卡以便在应用规则前后跟踪和度量数据质量。如果配置文件的外部源元数据或结果卡发生更改，则可以将这些更改与其数据对象同步。

配置文件和分析类型

根据您需要执行的分析类型创建配置文件。创建的配置文件类型应您执行的分析类型相对应。例如，要执行主键分析，则应创建主键配置文件。

可以创建下列配置文件以执行数据分析和发现：

列配置文件

用于分析表或文件的所选列中的数据质量。可以在 Analyst 工具和 Developer 工具中为列分析定义配置文件。

数据域发现

用于发现企业范围内的关键数据特性。数据域发现会根据列值或名称标识与该列关联的所有数据域。在发现过程中，您可以手动创建数据规则和列名称规则以验证值或列名称是否属于数据域。然后，可以在创建数据域后关联这些规则。还可以根据列配置文件结果中的值和模式来创建数据域。

主键配置文件

用于发现表或文件中各列之间的主键关系。可以在 Developer 工具中为主键分析定义配置文件。

功能相关性配置文件

用于发现表或文件中各列之间的功能相关性。可以在 Developer 工具中为功能相关性分析定义配置文件。

外键配置文件

用于发现多个表或多个文件中各列之间的外键关系。可以在 Developer 工具中为外键分析定义配置文件。

联接配置文件

用于确定在一个数据源或多个数据源中各列之间的潜在联接程度。可以在 Developer 工具中为联接分析定义配置文件。结果会显示在韦恩图中。

重叠发现

用于确定在一个数据源或多个数据源内各列对中的重叠数据百分比。可以通过 Developer 工具中的编辑器运行重叠发现任务。可以在韦恩图中验证结果并进行查看。

企业发现

用于在大量数据源中运行多个数据发现任务并生成配置文件结果的合并摘要。包括运行列配置文件、数据域发现以及发现主键和外键关系。企业发现会自动处理大量数据源的配置文件过程。

注意: 在 Analyst 工具中对配置文件所做的更改不会显示在 Developer tool 中，除非刷新与模型存储库的 Developer tool 连接。

剖析组件

配置文件具有多个组件，您可以使用这些组件有效分析数据源的内容和结构。

配置文件具有以下组件：

筛选器

用于创建满足特定条件的原始数据源的子集。随后，可以对该示例数据运行配置文件。

规则

定义您在运行配置文件时应用于数据的条件的业务逻辑。可以向配置文件中添加规则以验证数据。

标记

根据业务目的在模型存储库中定义对象的元数据。创建标记可以根据对象的业务用途来分组对象。您可以在 Analyst 工具中为配置文件或配置文件中的列分配标记。

注释

有关配置文件的说明。使用注释可与其他 Analyst 工具用户和 Developer tool 用户共享配置文件的相关信息。您可以在 Analyst 工具中向配置文件或配置文件中的列添加注释。

结果卡

在配置文件结果中以图形方式显示列的有效值或规则输出。使用结果卡测量数据质量进度。

配置文件结果

可以在运行配置文件之后查看配置文件结果。可以查看配置文件中的列和规则的值、模式和统计信息摘要。同时可以查看配置文件中列和规则的属性。可以预览配置文件数据。

下表介绍了各个配置文件类型的配置文件结果：

配置文件类型	结果
列配置文件	<ul style="list-style-type: none">- 列中的空值、相异值和非相异值的个数与百分比以及列值的推理的数据类型。- 所选列中数据值的频率和字符模式以及该列的统计摘要。- 通过分析列数据推理出的数据类型。- 已记录的数据的数据类型。- 最大值和最小值。- 运行配置文件的日期和时间。
主键配置文件	<ul style="list-style-type: none">- 推理出的主键候选键的唯一值、重复值和空值的数量和百分比。- 推理出的主键候选键中的键冲突数。
功能相关性配置文件	<ul style="list-style-type: none">- 推理出的功能相关性。- 功能相关性冲突数。
外键配置文件	<ul style="list-style-type: none">- 满足您定义的主外键推理条件的主键列和外键列。- 主键和外键之间匹配的数据值数量（以百分比表示）。- 配置文件运行之前为主键列和外键列定义的关系类型。
联接配置文件	<ul style="list-style-type: none">- 显示各列之间关系的韦恩图。- 列中孤立值、空值和联接值的数量和百分比。
重叠发现	<ul style="list-style-type: none">- 两列之间重叠的百分比。- 显示各列之间重叠的韦恩图。
数据域发现	<ul style="list-style-type: none">- 与预定义数据域匹配的列名称和数据。- 列所属的数据域组及其数据类型。
企业发现	<ul style="list-style-type: none">- 列配置文件结果。- 数据域发现结果。- 主键发现结果。- 以图形和表格方式显示的外键配置文件结果。

可以使用第三方报告工具读取配置文件仓库中的配置文件结果。Informatica 提供了一组配置文件视图，可用于对您要读取的配置文件统计信息进行自定义。这些视图基于配置文件统计信息和配置文件结果分析的常用类型。

第 3 章

列配置文件概念

本章包括以下主题：

- [列配置文件概念概览, 23](#)
- [列配置文件选项, 24](#)
- [存储库配置文件锁定和受版本控制的配置文件管理, 24](#)
- [结果卡, 24](#)

列配置文件概念概览

列配置文件可确定数据源中列的特性，如值频率、百分比和模式。

列剖析发现数据的以下相关情况：

- 每个列中的空值、相异值和非相异值的个数，用数字和百分比表示。
- 每个列中数据的模式以及这些值出现的频率。
- 有关列值的统计信息，例如，每个列中值的最大和最小长度，以及第一个值和最后一个值。
- 记录的数据类型、推理的数据类型以及两者之间的潜在冲突。
- 模式离群值和值频率离群值。

创建或编辑配置文件时可以配置以下选项：

- 列配置文件选项。您可以选择要在上面运行配置文件的列，以及选择采样选项和向下钻取选项。
- 添加、编辑或删除筛选器和规则。

在配置文件结果中，可以向配置文件和配置文件中的列添加注释和标记。可以为列分配业务术语。

模型存储库会使用存储库配置文件锁定来锁定配置文件，以防止用户覆盖所做的工作。版本控制系统会保存配置文件的多个版本，并为每个版本分配一个版本号。可以先签出配置文件，然后在进行更改后再签入配置文件。重新签入配置文件之前，可以撤消配置文件的签出操作。

创建结果卡以定期查看数据质量。在为配置文件应用规则之前和之后均需创建结果卡，以便查看列中有效值的图形表示形式。

使用计划程序服务调度配置文件运行和结果卡运行，以便在特定时间或间隔运行。计划程序服务会管理配置文件、结果卡、已部署映射和已部署工作流的计划。您可以在 Informatica Administrator 中创建、管理和运行计划。

列配置文件选项

创建配置文件时，您可以使用配置文件向导定义筛选器、规则、向下钻取选项、采样选项和连接。这些选项将确定配置文件从源数据读取行的方式。

您可以在列配置文件、数据域发现配置文件或企业发现配置文件中定义以下选项：

- 筛选器。您可以为配置文件创建并应用筛选器。
- 规则。创建配置文件时可以添加规则。您可以重复使用在 Analyst 工具或 Developer tool 中创建的规则。
- 向下钻取选项。您可以选择读取数据源中的当前数据，或者读取暂存在剖析仓库中的配置文件数据。
- 采样选项。您可以选择一个采样选项，确定要对其运行配置文件的行数。
- 连接。您可以在本地或 Hadoop 运行时环境中运行配置文件。在 Hadoop 运行时环境中可以选择 Blaze 或 Spark 引擎。

存储库配置文件锁定和受版本控制的配置文件管理

模型存储库会锁定配置文件以防止用户覆盖所做的工作。当您开始编辑配置文件时，该配置文件会被锁定，使其他用户无法保存对其所做的更改。保存配置文件后，便会释放锁定。受版本控制的配置文件管理会创建配置文本版本，您可以查看版本历史记录。

在 Developer tool 或 Analyst 工具中编辑配置文本时，模型存储库会锁定该配置文件。如果该工具意外停止，锁定仍会保留，这样，当您再次连接到模型存储库时，便可查看已锁定的配置文件。您可以继续编辑配置文件，也可以解除配置文件锁定。

模型存储库与版本控制系统集成时，您可以管理配置文件的版本。例如，您可以签出和签入配置文件，撤消签出，查看配置文件的特定历史版本以及已签出的配置文件。有关 Analyst 工具中的存储库资产锁定和受版本控制的资产管理的信息，请参阅《Analyst 工具指南》。有关 Developer tool 中的存储库对象锁定和受版本控制的对象管理的信息，请参阅《Developer Tool 指南》。

结果卡

结果卡是配置文件结果中列的有效值或规则的输出的图形表示形式。使用结果卡测量数据质量进度。可以通过配置文件创建结果卡，然后监视数据质量随时间推移的进展情况。

结果卡包含多个组件，如度量、度量组和阈值。在运行配置文件后，可将源列作为度量添加到结果卡，然后为度量配置有效值。结果卡可以帮助组织通过在度量和结果卡级别跟踪错误数据的成本，从而度量数据质量的值。要度量每个度量的错误数据的成本，请为该度量分配一个成本单位，然后设置固定或可变成本。在运行结果卡后，结果卡结果将包括每个度量的错误数据成本，以及所有度量的总成本值。

使用度量组将结果卡中的相关度量分类到一个集合中。阈值以百分比的形式确定记录中的列可接受的错误数据的范围。可以为正常、可接受或不可接受的数据范围设置阈值。

在运行结果卡后，配置是否要对实时数据或暂存数据向下钻取得分度量。在运行结果卡并查看得分后，向下钻取每个度量，以确定有效数据记录和无效记录。还可以查看结果卡中每个度量或度量组的结果卡沿袭。要有效跟踪数据质量，可以使用得分趋势图表和成本趋势图表。这些图表可以监视得分和错误数据的成本在一段时间内如何变化。

剖析仓库存储结果卡统计信息和配置信息。可以配置第三方应用程序来获取结果卡结果并运行报告。还可以在 Web 应用程序、门户或报告（如业务情报报告）中显示结果卡结果。

第 4 章

数据域发现概念

本章包括以下主题：

- [数据域发现概念概览, 25](#)
- [数据域, 26](#)
- [数据域组, 26](#)
- [数据域词汇表, 26](#)
- [数据域发现过程, 27](#)
- [Spark 引擎上的数据域发现, 27](#)

数据域发现概念概览

您需要确定和了解关键源数据的含义，以便能够采取措施对其进行有效处理。数据域发现是根据数据的语义发现数据源中数据的功能含义的过程。

创建配置文件以执行数据域发现，您可以识别企业内的关键数据特性。然后可以为数据应用更多数据管理策略（如数据质量或数据屏蔽）。例如，发现产品代码或说明，以分析需要应用哪些数据质量标准以及解析规则，以使数据有用且可靠。另一个示例是查找敏感客户数据，如信用卡号、电子邮件 ID 和电话号码。可能要屏蔽这些信息以对其加以保护。

您可以创建并运行配置文件，以便在 Analyst 和 Developer tool 中执行数据域发现。可以定义配置文件，以根据以下规则执行数据域发现：

- 数据规则。查找包含与规则中定义的特定逻辑相匹配的数据的列。
- 列名称规则。查找与规则中定义的列名称逻辑相匹配的列。

可以从列配置文件结果中的值和模式创建数据域。然后可以使用这些数据域跨多个数据系统或在整个企业内发现关键数据。

您可以创建一个利用采样选项和筛选器来执行数据域发现的配置文件。运行该配置文件时，即会对数据源应用采用选项和筛选器，同时生成一个数据集。数据域发现过程会使用该数据集来发现数据域。

数据域

数据域是基于列数据或列名称的语义的预定义或用户定义模型存储库对象。例如，社会保障号、信用卡号、电子邮件 ID 和电话号码可以是个人数据域。

数据域可帮助您在数据源中找到仍未发现的重要数据。例如，您可能在“注释”字段中包含社会保障号的旧数据系统。您需要先找到此信息并加以保护，然后才能将其移至新的数据系统。

您可以选择源行的最低百分比或源行的最小数量作为数据域匹配的遵从性条件。在列配置文件中执行数据域发现时，您也可以排除空值。

您可以将逻辑数据域分组为多个数据域组。数据域词汇表列出所有数据域和数据域组。使用 Developer tool 中的“首选项”菜单在数据源词汇表中导入和导出数据域。

使用规则定义与源数据和元数据匹配的数据和列名称模式。创建数据域时，Analyst 工具或 Developer tool 将关联的规则和其他相关对象复制到数据域词汇表中。使用 Developer tool 管理数据域，包括从数据源词汇表中导入和导出数据域。还可以使用 Developer tool 管理数据域的规则逻辑。

注意：您可能希望保存一个项目或文件夹中的所有数据域规则。在导出数据域之后并且需要编辑规则和其他关联的数据对象时，该步骤会有帮助。

数据域组

数据域组可帮助您将数据域分类为特定组。例如，可以将个人健康信息 (PHI) 数据域组下的数据域 first_name、last_name 和 account_number 分组。

可以创建包含社会保障号、名字和姓氏的个人身份识别信息 (PII) 数据域组。数据域可以是多个数据域组的组成部分。例如，社会保障号可以同时属于支付卡行业 (PCI) 和 PII 数据域组。数据域组可以包含数据域，不包含其他数据域组。

注意：如果在安装后导入数据域文件 Informatica_IDE_DataDomain.xml，数据域词汇表将显示预定义的数据域组和数据域。然后可以根据需要创建更多数据域组。要查看和更改与数据域关联的规则，可导入 Informatica_IDE_DataDomainRule.xml 文件。

数据域词汇表

数据域词汇表是所有域组和数据域的容器。可以使用数据域词汇表创建、管理和删除数据域及数据域组。

可以在数据域词汇表内搜索特定域和域组。还可以将数据域导出到 XML 文件，以及将数据域从 XML 文件导入到数据域词汇表。数据域词汇表包含复制的规则和与数据域关联的所有引用数据。您无法编辑数据域词汇表中的规则。

可以从 Developer 工具中的“首选项”菜单以及从 Analyst 工具中的“管理”菜单查看数据域词汇表。使用模型存储库服务特权**管理数据域**确定创建、编辑和删除数据域和数据域组的人员。

数据域发现过程

可以定义并运行配置文件，以便根据您的作业角色在 Analyst 工具或 Developer 工具中执行数据域发现。配置了数据域发现选项并运行配置文件后，可以对结果进行验证和向下钻取。如果从编辑器内运行数据域发现，可以将结果添加到数据模型中。

完成以下步骤执行数据域发现：

1. 创建或导入数据域和域组。
2. 或者，合并相应域组下的数据域。
3. 创建配置文件以执行数据域发现。首先选择运行列配置文件以及数据域发现，还是仅运行数据域发现。
4. 选择列、域和相应的采样选项。
5. 运行配置文件。
6. 验证、向下钻取配置文件结果，并根据需要将结果添加到数据模型中。

Spark 引擎上的数据域发现

在 Spark 引擎上运行配置文件来执行数据发现时，引用表会暂存在 Hadoop 群集上。要确保所有数据域的引用表都暂存在群集上，可以执行以下步骤：

先决条件：

执行数据域发现时，必须具有模拟 HDFS 用户的权限。

下载 JDBC .JAR 文件

1. 获取所用的引用数据库的 JDBC .jar 文件。可以从数据库供应商网站下载这些文件。
2. 将下载的文件复制到以下位置：<INFA_HOME>/externaljdbcjars

在数据集成服务上配置自定义属性

1. 启动 Informatica Administrator，然后在域导航器中选择数据集成服务。
2. 单击属性选项卡中的自定义属性选项。
3. 设置以下自定义属性来暂存数据域的引用表：

属性名称	属性值
AdvancedProfilingServiceOptions.ProfilingSparkReferenceDataHDFSDir	hdfs://<Namenode>:<Port>/tmp/cms
ExecutionContextOptions.SparkRefTableHadoopConnectorArgs	--connect <JDBC thin 驱动程序连接 URL>

4. 确保群集中存在 hdfs://<Namenode>:<Port>/tmp/cms 目录。如果不存在此目录，请创建 hdfs://<Namenode>:<Port>/tmp/cms 目录或在要暂存数据的位置创建一个自定义目录。默认情况下，引用数据会暂存在 hdfs://<Namenode>:<Port>/tmp/cms 目录中。
5. 再次应用数据集成服务。
6. 打开 Analyst 工具或 Developer tool，并确保在第一次运行配置文件时选择所有数据域以暂存引用数据。

注意：如果在第一次运行配置文件时没有选择所有数据域，之后又在下一次运行配置文件时选择了其他数据域，则配置文件运行将失败。

第 5 章

内容管理概念

本章包括以下主题：

- [内容管理概念概览, 28](#)
- [分析师和开发人员的内容管理, 28](#)
- [内容管理任务, 29](#)

内容管理概念概览

内容管理是验证和管理数据源的已发现元数据的过程，以便元数据适合使用和报告。

可以对以下推理配置文件结果进行内容管理：

- 数据类型
- 数据域
- 主键
- 外键

可以对推理配置文件结果进行内容管理，以使关于数据库和架构中的列、数据域和数据对象关系的元数据准确无误。随后可以在使用发现搜索跨多个存储库搜索信息时，找到最相关的元数据。还可以在企业发现结果中查看外键关系图表时，找到最相关的元数据。

可以对配置文件作为配置文件运行的一部分生成的特定元数据推理进行内容管理。例如，可以批准或拒绝列配置文件结果和数据域发现结果中的推理的数据类型。还可以批准或拒绝企业发现结果中的推理主键和外键。

分析师和开发人员的内容管理

作为数据分析师或数据管理者，您可以在 Analyst 工具中管理列配置文件结果和数据域发现结果。可以管理配置文件结果以准备好准确的配置文件信息，以便进行发现搜索和进一步验证数据资产。

作为开发人员或数据架构师，您可以在 Developer 工具中管理列配置文件结果、数据域发现结果、主键发现结果和外键发现结果。

内容管理示例

以开发人员身份执行企业发现时，Developer 工具将为整个数据集处理选定的数据域。此操作可能会导致产生多个数据域推理，如将电话号码数据推理为社会保障号数据域。当列中的部分数据与其他数据域相匹配时，会发生多个数据域推理。例如，如果一个 10 位数的电话号码缺少一位数，则其模式可能会与社会保障号相同。发生数

据域推理指示列中可能存在数据质量问题，或者多个数据域之间存在匹配的模式。在本例中，Developer 工具可能会同时推理电话号码数据域和社会保障号数据域。您可以管理配置文件结果以便选择最适合的数据域并批准该数据域。在本示例中，电话号码是相关数据域，因为发生社会保障号推理是由于存在数据质量问题。

运行企业发现时，Developer tool 可能会针对日期列推理多个数据类型，如日期、字符串和变长字符型。作为数据架构师，建议您选择并批准“日期”数据类型，因为此数据类型与日期列最相关。

Developer 工具中的企业发现可能会根据列数据来推理所有的数据对象关系。其中部分数据对象关系包括所发现候选键中无效的数据对象关系。例如，Developer 工具可能会推理将序列表示为可能的键的列，并发现与包含类似列的表的关系。这些数据对象关系在数据库中可能不是有效的关系。在此类情况下，可以在内容管理过程中评估、验证和批准最适合的推理配置文件结果。

内容管理任务

可以在配置文件运行后对配置文件结果进行内容管理。还可以扭转您在先前运行配置文件时采取的内容管理决策。

可以在 Analyst 工具中执行以下内容管理任务：

- 批准或拒绝多个列和数据域的推理的数据类型。
- 将已批准或已拒绝的数据类型还原为已推理状态。
- 将已批准或已拒绝的数据域还原为推理状态。
- 查看或隐藏已拒绝的结果行。
- 根据特定的元数据首选项（如已批准的数据类型和数据域）将列从配置文件运行中排除。

可以在 Developer 工具中执行以下内容管理任务：

- 批准或拒绝多个列的推理的数据类型。
- 将已批准或已拒绝的数据类型还原为已推理状态。
- 将已批准或已拒绝的数据域还原为推理状态。
- 查看或隐藏已拒绝的结果行。
- 批准或拒绝主键发现结果中的数据对象。
- 批准或拒绝企业发现结果，包括外键发现结果。
- 根据特定的元数据首选项（如已批准的数据类型和数据域）将列从配置文件运行中排除。

第 II 部分： 使用 Informatica Analyst 的 Data Discovery

本部分包含以下章节：

- [Informatica Analyst 中的列配置文件, 31](#)
- [Informatica Analyst 中的规则, 41](#)
- [Informatica Analyst 中的筛选器, 46](#)
- [Informatica Analyst 中的列配置文件结果, 51](#)
- [Informatica Analyst 中的业务术语、注释和标记, 70](#)
- [Informatica Analyst 中的结果卡, 73](#)
- [Informatica Analyst 中的数据域发现, 94](#)
- [Informatica Analyst 中的企业发现, 103](#)
- [Informatica Analyst 中的企业发现结果, 108](#)
- [Informatica Analyst 中的发现搜索, 112](#)
- [Informatica Analyst 中的 Business Glossary 桌面版, 119](#)

第 6 章

Informatica Analyst 中的列配置文件

本章包括以下主题：

- [Informatica Analyst 中的列配置文件概览, 31](#)
- [列剖析过程, 32](#)
- [配置文件选项, 32](#)
- [运行时环境, 33](#)
- [Informatica Analyst 中的操作系统配置文件概览, 34](#)
- [存储库资产锁定和基于团队的开发概览, 35](#)
- [在 Informatica Analyst 中创建列配置文件, 35](#)
- [编辑列配置文件, 36](#)
- [运行配置文件, 37](#)
- [在 Spark 引擎上运行配置文件, 37](#)
- [同步选项, 37](#)

Informatica Analyst 中的列配置文件概览

创建配置文件时，您应选择要运行的配置文件所基于的数据对象中的列。您可以配置采样选项和向下钻取选项，以提高剖析速度。您可以选择运行时环境。创建配置文件时，您可以向该配置文件添加规则和筛选器。在运行配置文件后，可以检查剖析统计信息，以了解数据。

可以剖析最多包含 1000 列的宽表和平面文件。创建或运行配置文件时，您可为配置文件选择所有列，也可选择各个列。您可以选择对所有列进行向下钻取，并可查看这些列的值频率。如果列名称超过 245 个字符，则无法选择带分隔符的文件中的列来运行配置文件。

在 Spark 引擎上，无法在半结构化的数据源上运行配置文件。

可以在 Informatica Analyst 中使用以下方法创建列配置文件：

- 右键单击库工作区中的数据对象来创建配置文件。
- 使用默认选项创建默认列配置文件。
- 自定义配置文件设置来创建自定义配置文件。

注意：您可以查看并运行基于 Avro、JSON、Parquet 和 XML 数据源的配置文件。您可以在 Informatica Developer 中创建并编辑基于 Avro、JSON、Parquet 和 XML 数据源的配置文件。

列剖析过程

在列剖析过程中，可以选择包括所有源列进行剖析，或者选择特定列进行剖析。您也可以接受默认配置文件选项，或配置采样选项、向下钻取选项和运行时环境。

以下步骤描述了列剖析过程：

1. 为列配置文件选择名称、说明和位置。
2. 选择要运行配置文件的导入数据对象或外部源。
3. （可选）预览源数据。
4. 选择要运行配置文件的列。
5. 确定要使用默认选项创建配置文件还是更改默认选项。可以配置的选项包括采样选项、向下钻取选项以及运行时环境。
6. （可选）在创建配置文件时添加规则和筛选器。
7. 运行配置文件。

注意：对于列名称以及剖析多语言和 Unicode 数据，请考虑以下规则和准则：

- 可以从不同的源剖析多语言数据，并基于浏览器中的区域设置查看配置文件结果。Analyst 工具将根据浏览器的区域设置更改“日期时间”、“数值”和“十进制”数据类型。
- 对多语言数据进行排序。您可以对多语言数据进行排序。Analyst 工具将根据浏览器的区域设置显示排序顺序。
- 要剖析 DB2 数据库中的 Unicode 数据，请设置数据库中的 DB2CODEPAGE 数据库环境变量，并重新启动数据集成服务。

配置文件选项

配置文件选项包括数据采样选项和数据向下钻取选项。可以在为数据对象创建或编辑列配置文件时配置这些选项。

您可以在**发现**工作区中配置配置文件选项。可以选择使用默认的列选项、采样选项和向下钻取选项创建配置文件。使用向下钻取选项可在实时数据和暂存数据之间进行选择。

采样选项

采样选项可以确定 Analyst 工具选择对其运行配置文件的行数。可以在定义配置文件或运行配置文件时配置采样选项。

下表介绍了配置文件的各个采样选项：

选项	说明
所有行	对数据对象的所有行运行配置文件。 本地、Blaze 和 Spark 运行时环境中支持此选项。
对前 <number> 行进行采样	对从数据对象第一行开始的采样行运行配置文件。最多可以选择 2,147,483,647 行。 本地和 Blaze 运行时环境中支持此选项。

选项	说明
对 <number> 行随机采样	对数据对象中随机选取的若干行运行配置文件。最多可以选择 2,147,483,647 行。 本地和 Blaze 运行时环境中支持此选项。
随机采样(自动)	对基于数据对象中的行数计算出的采样行运行配置文件。 本地和 Blaze 运行时环境中支持此选项。
限制 n <数字> 行	根据数据对象中的行数运行配置文件。选择在 Hadoop 验证环境中运行配置文件时，Spark 引擎会从数据对象的多个分区收集样本并将这些样本推送到单个节点来计算采样大小。“限制 n”采样选项支持 Oracle、SQL Server 和 DB2 数据库。不能对“限制 n”采样选项使用高级筛选器。 Spark 运行时环境中支持此选项。
随机百分比	对数据对象中某一百分比的行运行配置文件。 Spark 运行时环境中支持此选项。
在后续运行配置文件时， 从数据类型和数据域推理 中排除已批准的数据类型 和数据域	在下次运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型或数据域。

选择对随机的行样本运行配置文件后，随机采样算法将以随机方式在数据对象中选择行来运行配置文件。为列配置文件选择随机采样选项时，Analyst 工具将对暂存的数据执行向下钻取。这会影响向下钻取性能。为数据域发现配置文件选择随机采样选项时，Analyst 工具将对实时数据执行向下钻取。

向下钻取选项

可以在定义或编辑配置文件时配置向下钻取选项。

下表介绍了配置文件的各个向下钻取选项：

选项	说明
实时	对实时数据进行向下钻取，以读取数据源中的当前数据。
暂存	对暂存数据进行向下钻取，以读取暂存在剖析仓库中的配置文件数据。
选择列	标识要向下钻取但并未选择进行剖析的列。

运行时环境

您可以选择本地或 Hadoop 作为列配置文件的运行时环境。在 Hadoop 运行时环境中，可以选择 Blaze 或 Spark 引擎。选择运行时环境后，Informatica Analyst 会在配置文件定义中设置该运行时环境。

本地环境

在本地运行时环境中运行配置文件时，Analyst 工具会向剖析服务模块提交配置文件作业。剖析服务模块随后将配置文件作业拆分成一组映射。数据集成服务会在运行数据集成服务的同一台计算机上运行这些映射，并将配置文件结果写入剖析仓库。默认情况下，所有配置文件都在本地运行时环境中运行。

您可以使用本地源在本地环境中创建并运行配置文件。本地数据源是非 Hadoop 源，例如平面文件、关系源或大型机源。您还可以使用本地环境中的 Hive 或 HDFS 数据源，在映射规范或逻辑数据源中运行配置文件。

Hadoop 环境

在 Hadoop 运行时环境中，可以选择 Blaze 引擎或 Spark 引擎来运行配置文件。

选择 Blaze 或 Spark 后，可以选择 Hadoop 连接。数据集成服务将配置文件逻辑推送到 Hadoop 群集上的 Blaze 或 Spark 引擎来运行配置文件。

在 Hadoop 环境中运行配置文件时，Developer tool 会向剖析服务模块提交配置文件作业。剖析服务模块随后将配置文件作业拆分成一组映射。数据集成服务会通过 Hadoop 连接将映射推送至 Hadoop 环境。Blaze 引擎或 Spark 引擎会处理映射，并且数据集成服务会将配置文件结果写入剖析仓库。

Sqoop 数据源的列配置文件

您可以在使用 Sqoop 的数据对象上运行列配置文件。选择 Hadoop 作为验证环境后，可以选择 Hadoop 连接中的 Blaze 或 Spark 引擎来运行列配置文件。

在逻辑数据对象或自定义数据对象上运行列配置文件时，您可以配置 num-mappers 参数以实现并行处理并优化性能。您还必须配置 split-by 参数，以指定 Sqoop 必须据其拆分工作单元的列。

请使用以下语法：

```
--split-by <column_name>
```

如果主键的值并没有在最小值和最大值范围内均匀分布，则可以将 split-by 参数配置为指定另一个具有平衡的数据分布的列来拆分工作单元。

如果没有定义 split-by 列，Sqoop 会根据以下条件拆分工作单元：

- 如果数据对象包含单个主键，Sqoop 会将主键用作 split-by 列。
- 如果数据对象包含复合主键，则 Sqoop 会默认处理不包含 split-by 参数的复合主键。有关详细信息，请参阅 Sqoop 文档。
- 如果数据对象包含具有相同列的两个表，您必须使用表限定名称定义 split-by 列。例如，如果表名称是 CUSTOMER，列名称是 FULL_NAME，请将 split-by 列定义如下：

```
--split-by CUSTOMER.FULL_NAME
```
- 如果数据对象不包含主键，m 参数和 num-mappers 参数的值默认为 1。

如果您使用 Teradata 提供技术支持的 Cloudera 连接器或 Hortonworks Connector for Teradata，并且 Teradata 表不包含主键，则需要配置 split-by 参数。

Informatica Analyst 中的操作系统配置文件概览

您可以在 Analyst 工具中选择操作系统配置文件。选择一个操作系统配置文件后，数据集成服务即会根据操作系统配置文件用户的权限创建并运行列配置文件、企业发现配置文件和结果卡。

该 Analyst 工具会使用默认配置文件来运行配置文件和结果卡。如果只有一个操作系统配置文件，则默认情况下即会选择该操作系统配置文件。如果具有多个操作系统配置文件，您可以选择操作系统配置文件之一。

选择操作系统配置文件

您可以在 Informatica Analyst 中选择操作系统配置文件。数据集成服务使用操作系统配置文件用户的权限来运行剖析作业。

1. 在 Informatica Analyst 标头区域中，单击 **<用户名> > 设置**。
此时将显示**设置**对话框。
2. 选择操作系统配置文件。单击**保存**。

存储库资产锁定和基于团队的开发概览

模型存储库会锁定配置文件，以防止用户覆盖其他用户所做的工作。如果模型存储库与版本控制系统集成在一起，它会保存多个资产版本并为版本分配版本号。可以签出和签入配置文件，还可以撤消签出。可以查看所签出配置文件的特定版本。

当您开始在 Analyst 工具中编辑配置文件时，模型存储库会锁定此配置文件，使其他用户无法对其进行编辑。保存此配置文件时，锁定仍会保留。关闭此配置文件时，模型存储库会解除配置文件锁定。

模型存储库会通过受版本控制的资产管理来保护配置文件不被开发团队的其他成员覆盖。当您试图编辑另一用户已经签出的配置文件时，您会收到一条通知，告知您哪位用户已经签出此配置文件。可以在只读模式下打开已签出的配置文件，也可以使用其他名称来保存此配置文件。

可以在“配置文件属性”对话框中选择配置文件的一个版本来查看此版本的配置文件定义。可以在“操作”菜单中访问“配置文件属性”选项。有关存储库资产锁定和受版本控制的资产管理的详细信息，请参阅《*Analyst 工具指南*》。

在 Informatica Analyst 中创建列配置文件

既可创建自定义配置文件，也可创建默认配置文件。创建自定义配置文件时，您可以配置列、采样行和向下钻取选项。创建默认配置文件时，将对整个数据集的所有数据域运行列配置文件和数据域发现。

1. 在**发现**工作区中，单击**配置文件**，或从表头区域中选择**新建 > 配置文件**。
注意：可以右键单击**库**工作区中的数据对象并创建一个配置文件。在此配置文件中，配置文件名称、位置名称和数据对象将从数据对象属性中提取。可以创建默认配置文件，也可以通过自定义相关设置来创建自定义配置文件。
此时将显示**新建配置文件**向导。
2. 默认情况下将选择**单源**选项。单击**下一步**。
3. 在**指定常规属性**屏幕中，输入配置文件的名称和可选说明。在“位置”字段中，选择要在其中创建配置文件的项目或文件夹。单击**下一步**。
4. 在**选择源**屏幕中，单击**选择**以选择数据对象，或单击**新建**以导入数据对象。单击**下一步**。
 - 在**选择数据对象**对话框中，选择数据对象。单击**确定**。
“属性”窗格将显示所选数据对象的属性。“数据预览”窗格将显示数据对象中的列。
 - 在**新建数据对象**对话框中，可以选择要为其创建配置文件的链接、架构、表或视图，选择位置，然后创建用于导入数据对象的文件夹。单击**确定**。
5. 在**选择源**屏幕中，选择要对其运行配置文件的列。或者，选择**名称**以选择所有列。单击**下一步**。

默认情况下将选择所有列。Analyst 工具将显示每个列的列属性，例如，名称、数据类型、精度、小数位数、可空性和主键的参与方。

6. 在**指定设置**屏幕中，选择要运行列配置文件还是数据域发现或同时运行列配置文件和数据域发现。默认情况下将选择列配置文件选项。
 - 选择**运行列配置文件**可运行列配置文件。
 - 选择**运行数据域发现**可执行数据域发现。在**数据域**窗格的**编辑数据域发现的列选择**对话框中，选择要发现的数据域，接着选择遵从性条件，然后选择要执行数据域发现的列。
 - 选择**运行列配置文件和运行数据域发现**可同时运行列配置文件和数据域发现。在**数据域**窗格中选择数据域选项。

注意: 默认情况下，所选列将用于列配置文件和数据域发现。单击**编辑**可选择或取消选择用于数据域发现的列。
 - 选择“数据”、“列”或“数据和列”以对其运行数据域发现。
 - 选择采样选项。在**针对以下对象运行配置文件**窗格中，可以选择**所有行(完整分析)**、**采样前**、**随机采样**、**随机采样(自动)**、**限制 n** 或**随机百分比**作为采样选项。采样选项应用于列配置文件和数据域发现。
 - 选择向下钻取选项。可以在**向下钻取**窗格中选择**实时**或**暂存**向下钻取选项，也可以选择**关闭禁用**向下钻取。或者，单击**选择列**，以选择要进行向下钻取的列。可以选择跳过对数据类型或数据域已经过批准的列进行数据类型和数据域推理。
 - 选择**本地**、**Blaze** 或 **Spark** 作为运行时环境。如果选择 **Blaze** 或 **Spark**，请单击**选择**以在**选择 Hadoop 连接**对话框中选择 Hadoop 连接。
7. 单击**下一步**。

此时将打开**指定规则和筛选器**选项卡。
8. 在**指定规则和筛选器**屏幕中，可以执行以下任务：
 - 创建、编辑或删除规则。您可以将现有规则应用到配置文件。
 - 创建、编辑或删除筛选器。

注意: 基于此配置文件创建结果卡时，可以重用为配置文件创建的筛选器。
9. 单击**保存并完成**以创建配置文件，或单击**保存并运行**以创建并运行配置文件。

编辑列配置文件

可以在运行列配置文件后对其进行更改。

1. 在**库**工作区中，选择包含配置文件的项目，或者在**资产**窗格中选择配置文件。
2. 单击配置文件名称。

摘要视图将显示在**发现**工作区中。
3. 如果启用了版本控制系统，请单击**操作 > 签出**以签出配置文件。
4. 单击**操作 > 编辑配置文件**。

此时将显示**配置文件**向导。
5. 根据要进行的更改，选择以下页面选项之一：
 - **指定常规属性**。更改基本属性，例如名称、说明和位置。
 - **选择源**。选择其他匹配的数据源和列以对其运行配置文件。

- **指定设置。**选择要运行列配置文件还是同时运行列配置文件和数据域发现。选择要发现的数据域，并修改数据域发现、采样和向下钻取选项。
 - **指定规则和筛选器。**创建、编辑或删除规则和筛选器。
6. 单击**保存并完成**以完成配置文件编辑，或单击**保存并运行**以编辑并运行配置文件。
 7. 如果启用了版本控制系统，必须执行以下任务：
 - 单击**保存并完成**以完成配置文件编辑。
 - 在摘要视图中，单击**签入**以签入配置文件。
 - 单击**操作 > 运行配置文件**，以运行该配置文件。

运行配置文件

运行配置文件，以针对内容和结构分析数据源，然后选择要用于向下钻取的列和规则。可以针对列和规则对实时数据或暂存数据进行向下钻取。执行初始配置文件运行后，您可以仅对某个列或规则运行配置文件，而不对所有源列运行配置文件。

1. 在**库**工作区中，在“项目”窗格中选择包含配置文件的项目或文件夹，或者在“资产”窗格中选择配置文件。
2. 单击**操作 > 打开**。
摘要视图将显示在**发现**工作区中。
3. 单击**操作 > 运行配置文件**。
Analyst 工具将执行配置文件运行，并在摘要视图中显示配置文件结果。
4. 在摘要视图中，单击一个列以查看列结果。
此时将显示详细视图。

在 Spark 引擎上运行配置文件

在 Spark 引擎上通过 JDBC 连接运行配置文件时，配置文件运行将失败。

在 Spark 引擎上运行配置文件之前，请执行以下步骤：

1. 创建 JDBC 仓库连接。
2. 获取用于提取数据的数据库的 Data Direct JAR 文件。
3. 将文件复制到以下位置：<INFA_HOME>/externaljdbcjars

同步选项

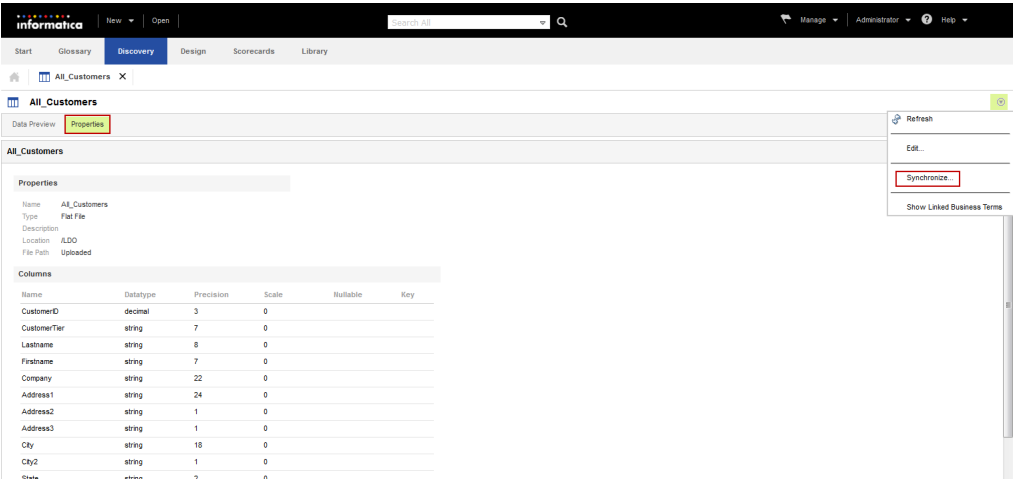
更改外部数据源的元数据时，默认情况下不会更新模型存储库中的数据对象元数据。可以使用“同步”选项将数据对象元数据与数据源元数据同步。可以对列配置文件、企业发现配置文件和结果卡使用“同步”选项。外部数据源可以是关系数据源或平面文件数据源。

在 Informatica Analyst 中同步平面文件数据对象

可以将外部平面文件数据源的更改与其在 Analyst 工具中的数据对象同步。使用**同步平面文件**向导同步数据对象。

- 1. 打开**库**工作区。
- 2. 在**项目**部分中，选择项目中的平面文件数据对象。
- 3. 从“操作”菜单中，单击**同步**。

下图显示了“属性”选项卡和“操作”菜单中的“同步”选项：



将显示**同步平面文件**向导。

- 4. 选择浏览某个位置或输入网络路径以导入平面文件。
 - 要浏览位置，请单击**选择文件**以从计算机可访问的目录中选择平面文件。
 - 要输入网络路径，请选择**输入网络路径**并配置文件路径和文件名。

下图显示了“同步平面文件”向导：

Synchronize Flat File: Step 1 of 5

Specify a location to import the flat file from and specify how to import the flat file.

☒ Browse and Upload:

Browse...

No file selected.

☐ Enter a Network Path:

☐ Hadoop File System

Description

Upload files from a local machine. Recommended for smaller files up to 10 MB. The Analyst tool uploads a copy of the file to the node on which the Analyst Service runs. Upload the file again if you modify the file.

?

BackNextFinishCancel

5. 单击**下一步**。
6. 选择导入带分隔符或固定宽度的平面文件。
 - 要导入带分隔符的平面文件，请接受**带分隔符**选项。
 - 要导入固定宽度的平面文件，请选择**固定宽度**选项。
7. 单击**下一步**。
8. 为带分隔符或固定宽度的平面文件配置平面文件选项。
9. 单击**下一步**。
10. （可选）更改列属性。
11. 单击**下一步**。
12. 接受默认名称或为平面文件输入其他名称。
13. （可选）输入说明。
14. 单击**完成**。
- 此时将显示一条同步消息，提示您确认操作。
15. 单击**是**可同步平面文件。
- 将显示一条消息指明同步已完成。要查看元数据更改的详细信息，请单击**显示详细信息**。
16. 单击**确定**。

在 Informatica Analyst 中同步关系数据对象

您可以将外部关系数据源的更改与其表数据对象进行同步。外部数据源更改包括添加、更改和删除源列和规则列。

1. 打开**库**工作区。

2. 在**项目**部分中，选择项目中的表数据对象。
Analyst 工具将在**属性**选项卡中显示表属性。
3. 从“操作”菜单中，单击**同步**。
此时将显示一条消息，提示您确认操作。
4. 要完成同步过程，请单击**是**。
将显示同步状态消息。
5. 将显示一条消息指明同步已完成。
要查看元数据更改的详细信息，请单击**显示详细信息**。
6. 单击**确定**。

第 7 章

Informatica Analyst 中的规则

本章包括以下主题：

- [Informatica Analyst 中的规则概览, 41](#)
- [预定义规则, 41](#)
- [表达式规则, 42](#)

Informatica Analyst 中的规则概览

规则是业务逻辑，用于定义在运行列配置文件时对源数据应用的条件。可以在配置文件中添加规则以验证数据。

在不同情况下，您可能需要使用规则。可以添加规则，以清理一个或多个数据列。可以添加查找规则，以提供源数据无法提供的信息。可以添加规则，为数据质量或数据集成项目验证清理规则。

创建或编辑列配置文件时，可以创建规则并将其添加到配置文件中，或者也可以将现有规则应用到配置文件。可以在列配置文件中使表达式规则或预定义规则。

运行配置文件后，Analyst 工具会在摘要视图中显示规则列的配置文件结果。您可以在详细视图中查看规则的列结果。规则的输出可以是一个或多个虚拟列。虚拟列存在于配置文件结果中。Analyst 工具将对虚拟列运行配置文件。例如，您使用某一自定义规则来拆分某一列，该列在 FIRST_NAME 和 LAST_NAME 虚拟列中包含名字和姓氏。Analyst 工具将对 FIRST_NAME 和 LAST_NAME 列运行配置文件。

注意：如果您删除了其他对象类型引用的某一规则对象，则 Analyst 工具将显示一条消息，列出这些对象类型。在删除规则之前，请确定删除它会造成什么影响。

预定义规则

预定义规则是在 Developer tool 中创建的规则，或 Developer tool 和 Analyst 工具随附提供的规则。将预定义规则应用到列配置文件，以修改或验证源数据。

预定义规则使用转换来定义规则逻辑。可以将预定义规则与多个配置文件配合使用。在模型存储库中，一条预定义规则就是一个 Maplet，其中包含用于定义规则逻辑的输入组、输出组和转换。

预定义规则进程

可以使用**新建规则向导**将预定义规则应用于配置文件。

要应用预定义规则，可以执行以下步骤：

1. 打开某一配置文件。
2. 选择一个预定义规则。
3. 审阅该规则的参数。
4. 选择输入列。如果想将该规则应用于多个列，可以选择多个列。
5. 配置剖析选项。

应用预定义规则

在应用预定义规则时，首先选择该规则，然后为该规则配置输入列和输出列。应用一个预定义规则，以使用一条被升级为可重用规则的规则，或者使用由开发人员创建的规则。

1. 在**库**工作区中，选择包含配置文件的项目，或者在**资产**窗格中选择配置文件。
2. 单击**操作 > 打开**，以打开配置文件。
摘要视图将显示在**发现**工作区中。
3. 单击**操作 > 编辑配置文件**。
此时将显示**配置文件**向导。
4. 单击**指定规则和筛选器**。
5. 在**指定规则和筛选器**屏幕的**规则**面板中，单击**操作 > 应用现有规则**。
此时将显示**应用规则向导**对话框。
6. 选择一个规则，然后单击**下一步**。
7. 单击**添加**。
此时将显示**选择输入端口的列**对话框。
8. 选择一个字段和一个输入列。单击**确定**。
输入列和输出列将显示在**应用规则向导**对话框中。
9. 在**应用规则向导**对话框中，单击**确定**。
该规则将显示在**指定规则和筛选器**屏幕中。

表达式规则

表达式规则使用表达式函数和列来定义规则逻辑。可以在 **Analyst** 工具中创建表达式规则，然后将它们添加到列配置文件中。

可以使用表达式规则更改或验证列配置文件中的列的值。可以创建一个或多个表达式规则在配置文件中使用。表达式函数与用于转换源数据的 SQL 函数相似。可以使用以下类型的函数创建表达式规则逻辑：

- 字符
- 转换
- 数据清理

- 日期
- 编码
- 财务
- 数值
- 科学计数
- 特殊
- 测试

可以使用以下方法创建表达式规则：

- 配置文件向导。创建或编辑列配置文件时，可以在配置文件向导中创建和应用表达式规则。通过将规则升级为可重用规则，您可以在多个配置文件中使用该规则。
- 规则规范。可以在 Analyst 工具中配置规则规范，并在列配置文件中使用该规则规范。配置规则规范时，请将业务规则的要求转换为一个或多个规则语句。规则语句表示用于确定数据集是否符合业务规则的逻辑。从规则规范生成 Mapplet，然后在 Developer tool 中创建的列配置文件中使用该 Mapplet。

可以使用表达式编辑器来添加表达式函数、将列配置为函数的输入、验证表达式，以及配置返回类型、精度和小数位数。创建并验证表达式规则后，可以编辑输出规则列的精度值。默认情况下，输出规则列的精度值设置为 10。输出规则列超出所设置的精度值时，精度值会被截断。

表达式规则的输出是虚拟列，该列使用规则名称作为列名称。Analyst 工具将对该虚拟列运行列配置文件。例如，您使用某一表达式规则来验证邮政编码。如果邮政编码有效，则该规则返回 1；如果邮政编码无效，则返回 0。Informatica Analyst 将对该规则的 1 和 0 输出值运行列配置文件。

创建表达式规则

可以使用**配置文件**向导来创建表达式规则，然后将其添加到配置文件。创建表达式规则，以验证配置文件中各列的值。

1. 打开某一配置文件。
2. 在摘要视图中，单击**操作 > 编辑配置文件**，以打开**配置文件**向导。
3. 单击**指定规则和筛选器**。
4. 在“规则”窗格中，单击**操作 > 添加规则**。

此时将显示**新建规则**对话框。

5. 在**新建规则**对话框中，输入规则的名称和可选说明。您可以在“函数”面板或“列”面板中创建规则。
 - 在“函数”面板中，选择函数类别，然后单击向右箭头(>>)按钮。在对话框中，指定参数并单击**确定**。函数将与列和值一起显示在“表达式”面板中。
 - 在“列”面板中，选择列，然后单击向右箭头(>>)按钮。该列将显示在“表达式”面板中。您可以添加函数、表达式和值，以创建规则。
6. 要验证规则，请单击**验证**。
7. 或者，选择将规则升级为可重用规则，然后配置项目和文件夹位置。如果将规则升级为可重用规则，则您或其他用户可以将该规则在其他配置文件中用作预定义规则。
8. 单击**确定**。

此时将显示**指定规则和筛选器**屏幕，并在“规则”窗格中列出规则。

使用规则规范创建表达式规则

可以在 Informatica Analyst 中使用规则规范创建表达式规则。可以将规则添加到列配置文件中来验证数据。

1. 在标题区域中，单击**新建 > 规则规范**。

此时将显示**新建规则规范**向导。
2. 在**新建规则规范**向导中，输入规则的名称和可选说明。
3. 在**位置**字段中，单击**浏览**以选择要将规则保存到的项目或文件夹。
4. 单击**继续**。

规则规范将显示在**设计**工作区中。


5. 要输入规则的属性，请选择该规则中的顶级八角形，然后单击**属性**。
6. 要配置主规则集，请单击该规则中的下一级矩形。
7. 要输入规则集的输入，请单击**属性 > 输入**。

此时将显示**输入管理**对话框。

8. 在**输入管理**对话框中，单击**添加输入**，然后键入输入的名称、数据类型、最大长度和说明。（可选）可以输入多个输入。
9. 单击**确定**。


输入将显示在**属性**部分中。

10. 要定义规则逻辑，请单击**规则逻辑**，输入运算符和条件，然后在**操作**列表中选择操作。
11. （可选）根据需要输入多个规则集。

12. 要验证规则，请单击**验证** () 图标。

13. 要在列配置文件中保存并使用规则规范，请单击**保存并完成**。

14. 要保存并继续处理该规则，请单击**保存并继续**。

15. 要在 Developer tool 中使用该规则规范，请单击**生成规则** () 图标以生成 Mapplet。

Analyst 工具将在模型存储库中创建 Mapplet。使该 Mapplet 成为有效的规则，然后在 Developer tool 中创建的列配置文件中使用该 Mapplet。

第 8 章

Informatica Analyst 中的筛选器

本章包括以下主题：

- [Informatica Analyst 中的筛选器概览, 46](#)
- [创建筛选器, 46](#)
- [管理筛选器, 49](#)

Informatica Analyst 中的筛选器概览

可以创建一个筛选器，以便能够创建符合筛选条件的原始数据源子集。随后，可以对筛选后的数据运行配置文件。

您可以创建筛选器，以查看满足筛选条件的配置文件结果。您可以基于摘要视图中可用的默认筛选器查看配置文件结果。

创建筛选器

可以创建一个筛选器，以便能够创建符合筛选条件的原始数据源子集。

1. 打开某一配置文件。
2. 在摘要视图中，单击**操作 > 编辑配置文件**。
此时将显示**配置文件**向导。
3. 单击**指定规则和筛选器**。
4. 在**筛选器**窗格中，单击**操作 > 添加筛选器**。
此时将显示**新建筛选器**对话框。
5. 创建简单筛选器、高级筛选器或 SQL 筛选器。
注意：对于日期列上的简单或高级筛选器，请以 YYYY/MM/DD HH:MM:SS 格式提供条件。
数据预览窗格将显示满足筛选条件的原始数据源的子集。
6. 单击**确定**。
此时将显示**指定规则和筛选器**屏幕，并在**筛选器**窗格中列出筛选器。

创建简单筛选器

可以使用 =、!=、>、< 等条件运算符创建简单的筛选器。使用筛选器可创建原始数据源的子集。

- 1. 在**新建筛选器**对话框中，单击**简单**。

下图显示了**新建筛选器**对话框中可用来创建简单筛选器的选项：

New Filter

Create a filter. The filter is used to create a subset of the data rows before profiling.

Name*

Description:

Choose the filter type*

Simple

Advanced

SQL

Columns	Operator	Values(s)
<div>-Select-</div>	<div>-Select-</div>	<div>+</div>

Filter Preview

?

Ok

Cancel

- 2. 输入名称和可选说明。
- 3. 选择一列。
- 4. 选择条件运算符。
- 5. 输入一个值。
- 6. （可选）单击加号 (+) 图标添加更多筛选器。
- 7. 单击**确定**。

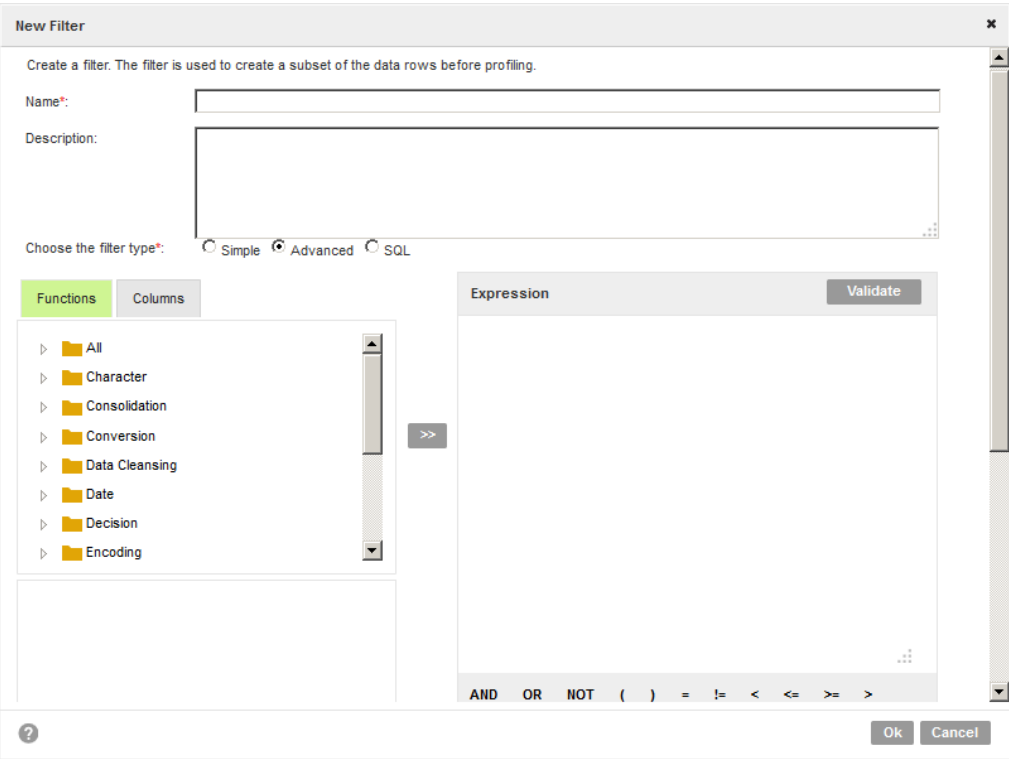
此时将显示**指定规则**和**筛选器**页面，并在“筛选器”窗格中列出筛选器。

创建高级筛选器

您可以使用诸如 AND、OR 和 NOT 等表达式创建高级筛选器，以创建原始数据源的子集。

1. 在**新建筛选器**对话框中，单击**高级**。

下图显示了**新建筛选器**对话框中的高级筛选器选项。



2. 输入高级筛选器的名称和可选说明。
3. 您可以在“函数”面板或“列”面板中创建高级筛选器。
 - 在“函数”面板中，选择函数类别，然后单击向右箭头 (>>) 按钮。
在对话框中，指定参数并单击**确定**。函数将与列和值一起显示在“表达式”面板中。
 - 在“列”面板中，选择列，然后单击向右箭头 (>>) 按钮。该列将显示在“表达式”面板中。
您可以添加函数、表达式和值，以创建高级筛选器。
4. 要验证高级筛选器，请单击**验证**。
5. 单击**确定**。

此时将显示**指定规则和筛选器**屏幕，并在“筛选器”窗格中列出筛选器。

创建 SQL 筛选器

您可以使用 SQL 查询创建 SQL 筛选器。可以为关系数据源创建 SQL 筛选器。

1. 在**新建筛选器**对话框中，单击 **SQL**。

下图显示了**新建筛选器**对话框中的 SQL 筛选器选项。

The screenshot shows the 'New Filter' dialog box with the 'SQL' tab selected. The dialog has a title bar 'New Filter' with a close button. Below the title bar is a subtitle: 'Create a filter. The filter is used to create a subset of the data rows before profiling.' There are two input fields: 'Name*' and 'Description:'. Below these fields are three radio buttons for 'Choose the filter type*': 'Simple', 'Advanced', and 'SQL' (which is selected). Below the radio buttons is a section titled 'Type or paste a SQL statement' with a 'Validate' button. At the bottom of the dialog is a 'Filter Preview' section with a refresh icon. The bottom of the dialog has a help icon, 'Ok', and 'Cancel' buttons.

2. 输入 SQL 筛选器的名称和可选说明。
3. 在文本框中，键入或粘贴 SQL 查询。
4. 单击**验证**以验证 SQL 查询。
5. 单击**确定**。

此时将显示**指定规则和筛选器**页面，并在“筛选器”窗格中列出 SQL 筛选器。

管理筛选器

可以编辑和删除筛选器。

1. 在**库**工作区中，选择包含配置文件的项目，或者在**资产**窗格中选择要筛选的配置文件。
2. 打开某一配置文件。
3. 在摘要视图中，单击**操作 > 编辑配置文件**，以打开**配置文件**向导。
4. 单击**指定规则和筛选器**。
5. 在“筛选器”窗格中，选择筛选器，然后单击**操作 > 编辑筛选器**。
此时将显示**编辑筛选器**对话框。
6. 编辑筛选器设置，然后单击**确定**。

7. 要删除筛选器，请选择该筛选器，然后单击**操作 > 删除筛选器**。

第 9 章

Informatica Analyst 中的列配置文件结果

本章包括以下主题：

- [Informatica Analyst 中的列配置文件结果概览, 51](#)
- [摘要视图, 52](#)
- [详细视图, 54](#)
- [统计信息, 55](#)
- [配置文件运行的类型, 61](#)
- [比较多个配置文件结果概览, 62](#)
- [列配置文件向下钻取, 66](#)
- [Analyst 工具中的内容管理, 67](#)
- [Informatica Analyst 中的列配置文件导出文件, 68](#)

Informatica Analyst 中的列配置文件结果概览

查看配置文件结果可了解并分析数据的内容、结构和质量。您可以在摘要视图中查看配置文件中的所有列和规则。可以在详细视图中查看列或规则的详细属性。

您可以在**发现**工作区下查看配置文件结果。视图标题显示了配置文件类型、配置文件中的列数和规则数以及采样数据与创建日期和时间。

在摘要视图中，每个列的属性可以显示为值、水平条形图或百分比。您可以查看列属性，例如，空值、相异值、非相异值、模式、数据类型和数据域。您可以基于默认筛选器在摘要视图中查看配置文件结果。

在详细视图中，您可以在各个窗格中查看空值、相异值、非相异值、推理的数据类型、推理的数据域、推理的模式、值、业务术语，以及预览数据。

您可以查看最新运行、历史运行与合并的运行的配置文件结果。还可以比较两次配置文件运行的配置文件结果，并在摘要视图和详细视图中查看结果。您可以查看配置文件统计信息并管理数据。配置文件统计信息包括列和规则的值、模式、数据类型、离群值和统计信息。您可以对数据执行数据发现和向下钻取。

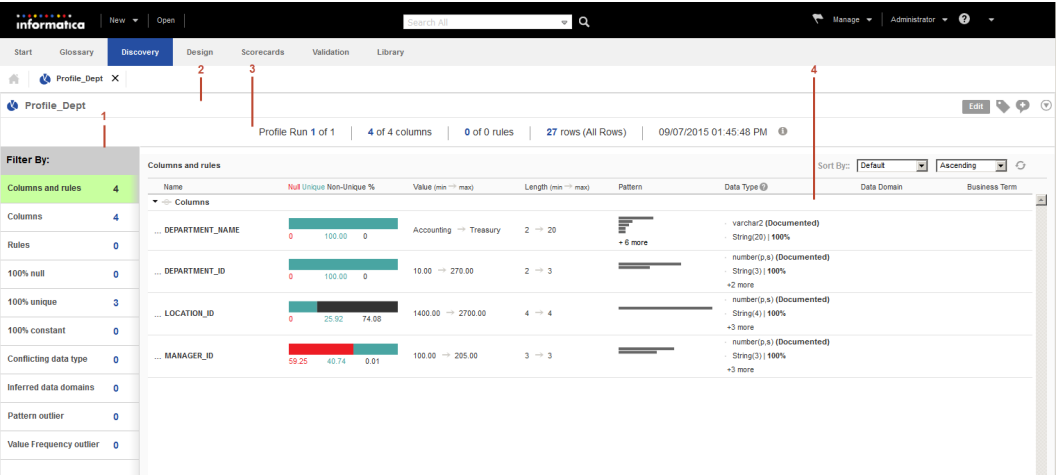
注意：您可以查看并运行基于 Avro、JSON、Parquet 和 XML 数据源的配置文件。还可以查看最新运行、历史运行与合并的运行的配置文件结果，以及比较两次配置文件运行的配置文件结果。

您可以将值频率、模式频率、向下钻取数据、注释、标记和业务术语导出到 CSV 文件。可以将配置文件摘要信息导出到 Microsoft Excel 文件，以便在一个文件中查看所有数据并进行进一步分析。您可以在配置文件结果中查看规则信息。显示的配置文件结果取决于配置文件配置和采样选项。

摘要视图

配置文件结果摘要以网格格式显示在摘要视图中。您可以在摘要视图中使用默认筛选器查看特定统计信息。例如，选择“规则”时，摘要视图将显示配置文件中的所有规则。

下图显示了摘要视图的图形视图示例：



1. 默认筛选器。在摘要视图中，可以基于默认筛选器查看配置文件结果。
2. 配置文件表头。可以在表头中查看配置文件名。可以使用“编辑”按钮编辑配置文件，使用标记和注释图标来添加或编辑标记和注释，以及从“操作”菜单中选择相应选项。
3. 摘要视图表头。可以在摘要视图表头中查看配置文本特定信息。可以查看配置文件运行编号、配置文件总运行次数、列数、规则数以及配置文件中的行数。
4. 摘要视图。可以查看配置文件中所有列和规则的属性。

在摘要视图中，可以运行或编辑配置文件，检测模式离群值或值频率离群值，向结果卡添加列，选择配置文件运行，比较两次配置文件运行，将配置文件结果或数据域发现结果导出到 Microsoft Excel 电子表格，验证多个列的推理结果，添加或删除注释和标记，或者查看配置文件属性。

摘要视图属性

摘要视图显示了配置文件中所有列和规则的属性。摘要视图包含属性的可视化表示形式。您可以单击每个摘要属性，对属性值进行排序。

下表介绍了配置文件结果摘要属性：

属性	说明
名称	显示配置文件中的列或规则的名称。
空值 相异 非相异百分比	以百分比形式显示列或规则输出的空值、相异值和非相异值。值可以显示为水平条形图。
模式	将列中的多个模式显示为水平条形图。将鼠标指针悬停于条形图上方时，列中的模式字符和相似模式的个数可以显示为百分比。
值	显示列或规则输出的最小值和最大值。
长度	显示列或规则输出中的值的最小长度和最大长度。

属性	说明
数据类型	<p>显示列或规则的已记录的数据类型。在您将鼠标指针悬停于字段上方时显示推理的数据类型。Analyst 工具可以推理以下数据类型：</p> <ul style="list-style-type: none"> - 字符串 - 变长字符型 - 小数 - 整型 - 日期 <p>您也可以基于推理的数据类型查看遵从性百分比。</p> <p>注意: Analyst 工具无法从精度高于 38 的数值列的值派生数据类型。Analyst 工具无法从精度高于 255 的字符串列的值派生数据类型。如果您基于日期列创建了年份值早于 1800 年的列配置文件，则推理的数据类型可能会以固定长度字符串的形式显示。根据需要更改 InferDateTimeConfig.xml 中最小年份参数的默认值。</p>
数据域	显示与列关联的数据域的名称，以及遵从性百分比和遵从行数。
业务术语	显示分配给列的业务术语。

摘要视图中的默认筛选器

您可以基于默认筛选器在摘要视图中查看配置文件结果。

摘要视图默认显示所有源列、虚拟列和规则列的配置文件结果。“筛选依据”窗格会显示可将默认筛选器应用到的列的数量。

在摘要视图中，您可以使用以下默认筛选器选项来查看配置文件结果：

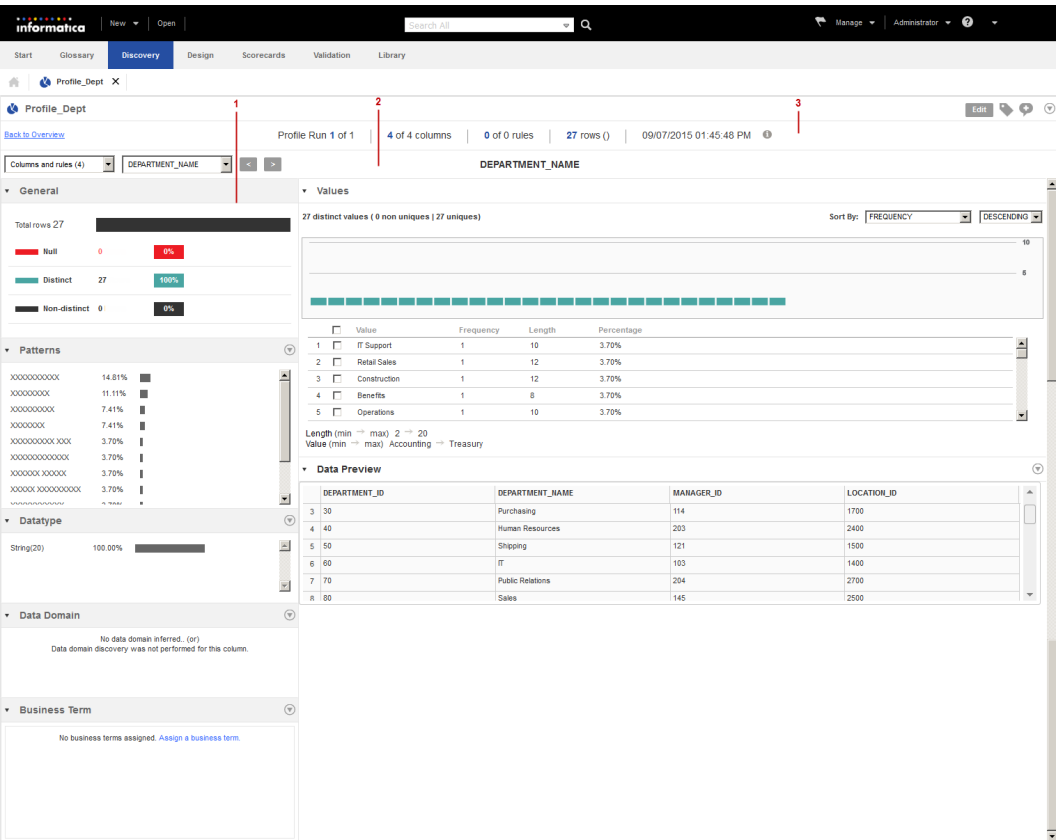
默认筛选器选项	说明
列和规则	显示源列和规则列的配置文件结果。您可以展开和折叠源列与规则列来查看结果。
列	显示源列的配置文件结果。
规则	显示规则列的配置文件结果。
100% 空值	显示值全部为空值的列的配置文件结果。
全部相异	显示值全部为相异值的列的配置文件结果。
100% 常量	显示所有记录都具有相同值的列的配置文件结果。例如，如果“国家/地区”列仅包含值“USA”，则“100% 常量”筛选器将包括该列的配置文件结果。
冲突数据类型	显示已记录的数据类型与推理的数据类型不匹配的列的配置文件结果。例如，筛选器将显示列 CustomerTier，因为该列的已记录的数据类型为“整数 (2)”，而推理的数据类型为“字符串”。
推理的数据域	显示推理的数据域与配置的数据域相同的列的配置文件结果。
模式离群值	显示具有模式离群值的列的配置文件结果。
值频率离群值	显示具有值离群值或频率离群值的列的配置文件结果。

详细视图

列结果将显示在详细视图中。可以详细查看列属性。

单击摘要视图中的某个列后，即会显示该列的详细视图。

下图是详细视图中列属性的一个图形视图示例：



1. 窗格。可以在窗格中查看常规属性、列值、数据预览、推理的模式、推理的数据类型、推理的数据域以及业务术语。
2. 列详细信息表头。可以通过在下拉列表中选择列或者使用导航按钮来查看列结果。
3. 摘要视图表头。可以在摘要视图表头中查看配置文本特定信息。可以查看配置文件运行、列数、规则、配置文件运行中的行数以及配置文件运行的时间和日期。

在详细视图中，可以运行或编辑配置文件，向结果卡添加列，选择配置文件运行，比较两次配置文件运行，将配置文件结果导出到 Microsoft Excel 电子表格，将值频率、模式频率、数据类型、所选值的向下钻取数据或所选模式的向下钻取数据导出到 csv 文件，将注释和标记添加到列或从列中将其删除，以及查看配置文件属性。

使用每个窗格中的“操作”菜单可对列属性执行进一步的操作。可以折叠或展开窗格。

详细视图窗格

详细视图会在各个窗格中显示如下列属性：相异值、非相异值和空值的个数与百分比以及模式、推理的数据类型、推理的数据域、值、数据预览和链接的业务术语。

单击列或规则时，将打开该列或规则的详细视图。

下表介绍了详细视图中的窗格：

窗格	说明
常规	以各种颜色显示包含空值、相异值和非相异值的行的数量。您可以查看以百分比表示的值。您可以通过迷你图查看每次连续的配置文件运行的常规值增加和减少情况。走势图会在折线图中显示最近连续五次运行配置文件时空值、相异值或非相异值的数量变化。将指针移动到每个配置文件运行对应的走势图时，您可以查看值数量和值百分比。可以向列添加标记和注释。
模式	显示列值的模式。模式在列中出现的频率显示为水平条形图，并以百分比表示。您可以对模式进行向下钻取，还可以将模式添加到引用表或使用所选模式创建数据域。
数据类型	显示列的推理的数据类型。数据类型在列中出现的频率显示为水平条形图，并以百分比表示。您可以对数据类型进行向下钻取，还可以批准、拒绝或充值所选的推理的数据类型。 显示拒绝项 选项显示了拒绝的推理的数据类型。
数据域	显示列的推理的数据域。您可以对数据域进行向下钻取，了解遵从行数、不遵从行数或具有空值的行数。可以批准、拒绝或重置数据域值。 显示拒绝项 选项显示了拒绝的数据域。您可以验证数据域值。
业务术语	显示列的已分配业务术语。您可为列分配或取消分配业务术语。
值	以图形表示形式显示列中的所有值以及频率、长度和百分比。您可以对每个值进行向下钻取。您可以将值添加到引用表，还可以创建值频率规则以及创建数据域。
数据预览	显示所选模式、数据类型、数据域或值的向下钻取数据。

统计信息

您可以查看配置文件中的列和规则的如下统计信息：值、模式、数据类型、数据域和离群值。

您可以在摘要视图中查看配置文件统计信息，而在摘要视图和详细视图中查看列统计信息。可以查看最新配置文件运行、历史配置文件运行与合并的配置文件运行的统计信息。还可以比较两次配置文件运行的配置文件结果，并在摘要视图和详细视图中查看配置文件和列的统计信息。

数据预览

您可以在“数据预览”窗格中查看所选模式、数据类型、数据域或值的向下钻取数据。

您可以在详细视图中查看“数据预览”窗格。单击摘要视图中的某列时即会显示详细视图，而且默认情况下会收起“数据预览”窗格。要查看列数据，请单击**操作 > 显示预览**。

下表说明了“数据预览”窗格中操作菜单中的选项：

选项	说明
添加到筛选器	创建向下钻取筛选器来筛选向下钻取数据，从而对分析结果子集中数据的不规范问题进行分析。
保存筛选器	保存向下钻取筛选器。
显示预览	显示源行。
导出数据	将向下钻取结果导出到 CSV 文件或 Microsoft Excel 文件。

数据类型

数据类型包括配置文件结果中每个列的所有推理的数据类型。

可以在摘要视图和详细视图查看数据类型。在摘要视图中，您可以查看已记录的数据类型和推理的数据类型。**冲突数据类型**筛选器可显示已记录的数据类型和推理的数据类型之间存在冲突的列。在详细视图中，您可以查看列的推理的数据类型。数据类型在列中出现的频率显示为水平条形图，并以百分比表示。您可以向下钻取、批准、拒绝或重置选定的推理的数据类型。“显示拒绝项”选项显示了拒绝的推理的数据类型。

下表介绍了数据类型的属性：

属性	说明
数据类型	显示配置文件中列的已记录的和推理的数据类型的列表。
频率	显示数据类型在列中出现的次数，以数字表示。
百分比	显示数据类型在列中出现的百分比。
向下钻取	根据列数据类型向下钻取到特定的源行。 注意: 如果选择了多个推理的数据类型，则无法执行向下钻取操作。
状态	指示数据类型的状态。状态有“已推理”、“已批准”或“已拒绝”。 已推理 指示 Analyst 工具推理的列数据类型。 已批准 指示列的已批准数据类型。批准数据类型后，该数据类型随即提交到模型存储库。 已拒绝 指示列的已拒绝数据类型。

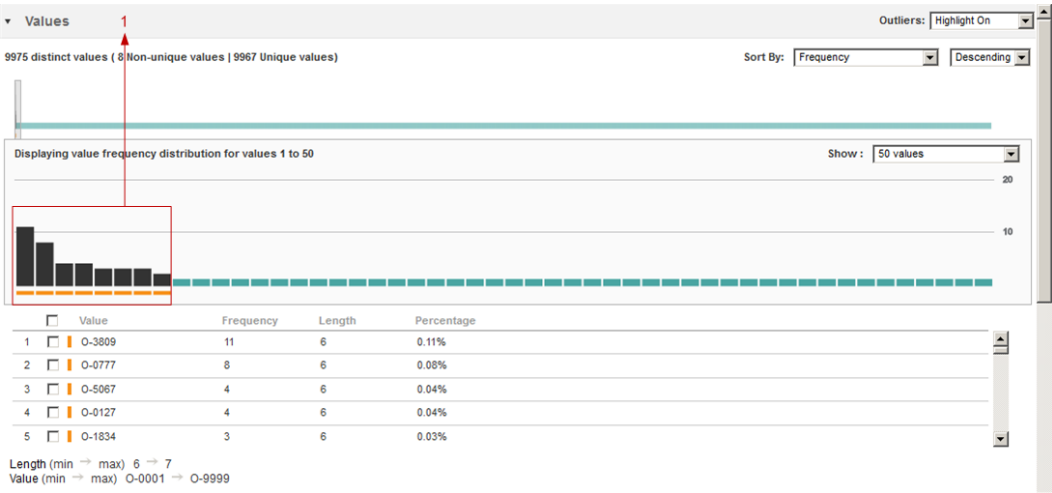
离群值

离群值是指配置文件结果中的列的模式、值或频率不在预期值范围内。

数据集成服务的剖析插件将运行某种算法，以确定不在列中大多数值的范围内的值。任何不在列中大多数值的预期范围内的模式、值或频率将被视为离群值。

默认情况下，Analyst 工具不会确定配置文件结果中的离群值。在摘要视图中，可以运行离群值以查看离群值结果。“模式”离群值筛选器会基于列中的模式显示离群值。“值频率”离群值筛选器会基于列中的值或频率显示离群值。离群值检测会在后台进行，以便您能在摘要视图中执行其他操作。

在详细视图中，如果从列表中选择了**突出显示**选项，则您可以在“值”窗格中查看离群值。离群值以带橙色下划线的竖条显示。要仅查看离群值，必须从列表中选择**筛选器**选项。



1. 离群值。离群值以带橙色下划线的竖条显示。

运行离群值

运行离群值可确定列中不在预期值范围内的模式、值或频率。

- 在摘要视图中，单击**操作 > 检测离群**。
筛选依据窗格中的“模式离群值”和“值频率离群值”会从 N/A 变为检测到的离群值数量。
- 在筛选依据窗格中，单击**模式离群值**。
具有模式离群值的列将显示在摘要视图中。
- 在筛选依据窗格中，单击**值频率离群值**。
具有值离群值或频率离群值的列将显示在摘要视图中。
- 在详细视图中，从离群值下拉列表中选择**突出显示**。
在“值”窗格中，离群值将显示为带橙色下划线的竖条。
- 单击“离群值”下拉列表中的**筛选器**，以只查看离群值。

模式

可以在摘要视图和详细视图中查看列值的模式以及这些模式的出现频率。

在摘要视图中，列中的多个模式可以显示为水平条形图。将鼠标指针悬停于条形图上方时，列中的模式字符和相似模式的个数可以显示为百分比。在详细视图中，模式在列中出现的频率可以显示为水平条形图，并以百分比表示。您可以执行向下钻取，还可以将模式添加到引用表或使用所选模式创建数据域。

默认情况下，剖析仓库最多可以存储 16,000 个唯一的最高频率值，包括配置文件结果的空值。如果配置文件结果中至少有一个空值，则 Analyst 工具可能会将空值显示为模式。

注意: Analyst 工具无法为精度高于 38 的数字列派生模式。Analyst 工具无法为精度高于 255 的字符串列派生模式。

下表说明了列模式的属性：

属性	说明
模式	显示配置文件中列的模式。
频率	显示模式在列中出现的次数，以数字表示。
百分比	显示模式在列中出现的百分比。

下表介绍了模式字符及其表示的含义：

字符	说明
“B” 或 “b” 或 “ ”	表示空格。
“C” 或 “c”	表示任何字符。
“L” 或 “l”	表示任何小写字母字符。
“T” 或 “t”	表示 Tab 键。
“U” 或 “u”	表示任何大写字母字符。
9	表示任何数字字符。Informatica Analyst 以 “9” 的格式单独显示最多 3 个字符。Analyst 工具将 3 个以上字符显示为一个以括号括起的值。例如，格式 “9(8)” 表示包含 8 位数的数值。
“X” 或 “x”	表示任何字母字符。Informatica Analyst 以 “X” 的格式单独显示最多 3 个字符。Analyst 工具将 3 个以上字符显示为一个以括号括起的值。例如，格式 “X(6)” 可以表示值 “Boston”。 注意： 模式字符 X 不区分大小写，并且可以表示源数据中的大写字符或小写字符。
“P” 或 “p”	表示 “(”，即左括号。
“Q” 或 “q”	表示 “)”，即右括号。

注意：列模式可以包含特殊字符。例如，~、[、]、=、-、?、=、{、*、-、>、< 和 \$。

值

您可以查看列的值，以及值在列中出现的频率。

可以在摘要视图中查看列中的最小值和最大值。在详细视图中，您可以查看列的值属性。

摘要视图中的值

您可以在摘要视图中查看最新配置文件运行、历史配置文件运行与合并的配置文件运行的所有列和规则的最小值和最大值。

示例

假设某家零售店的数据库的“员工”表中名为“员工 ID”的列填充了范围在 100 到 250 之间的员工 ID，并且该数据库还包含姓名（例如 Bob 和 Robert），则当您为“员工”表运行列配置文件时，摘要视图中的“员工 ID”的“值”列将显示 “100 --> Robert”。

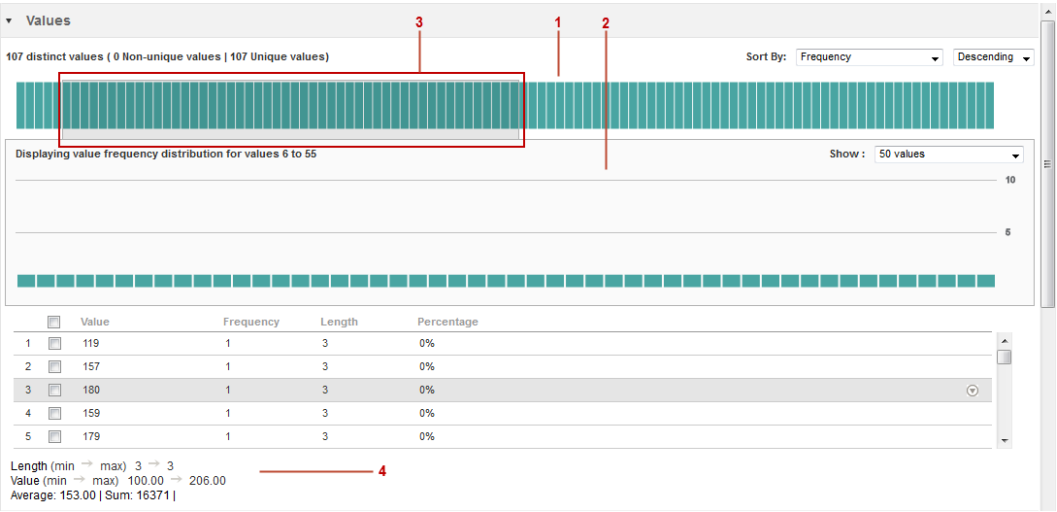
详细视图中的值

详细视图中的列值包括列的值，以及值在列中出现的频率。

值窗格以图形表示形式显示列值。您可以查看每个值的频率、长度和百分比。可以基于值或频率对值进行排序。可以对数据执行向下钻取以及将值添加到引用表，还可以创建值频率规则或创建数据域。 您可以查看用红色竖条表示的空值，用黑色竖条表示的值频率，以及用橙色竖条突出显示的离群值。您可以突出显示离群值、禁用离群值或将结果筛选为只显示列中的离群值。
“值”窗格包含图形布局和价值部分。

图形布局分为两个面板。

下图显示了详细视图中的“值”窗格：



1. 上方面板。您可以查看用垂直条形图表示的各种值。您可以按频率和值对各种值进行排序。您可以按照升序或降序顺序对值进行排序。您可以查看用橙色竖条突出显示的离群值。
2. 下方面板。您可以在下方面板中查看滑块中的值，其中的每个值都用竖条表示。可以对值执行向下钻取以及将值添加到引用表，还可以创建值频率规则，以及为值创建数据域。您可以同时查看 50、75 或 100 个值。
3. 滑块。在上方面板中，您可以将滑块滑动至各个值上方。下方面板会显示滑块中的值。
4. 值属性。值属性部分会显示值和属性。

下表介绍了图形布局中的面板：

面板	说明
上方面板	以垂直条形图形式显示所有值。上方面板最多可以显示 16,000 个值。 可以使用滑块查看一系列值。
下方面板	显示您在上方面板中选择的批次的值。默认情况下，Analyst 工具显示 50 个值。您可以选择同时查看 75 或 100 个值。

下表介绍了值部分中的列值的属性：

属性	说明
值	显示您在上方面板中选择的批次的值列表。 注意: Analyst 工具会从列值中排除 CLOB、BLOB、“原始”和“二进制”等数据类型。
频率	显示值在列中出现的次数，以数字表示。
长度	显示列值的长度。
百分比	显示值在列中出现的百分比。

下表介绍了所选列的统计信息：

统计信息	说明
长度(最小值 - 最大值)	显示列中最短值和最长值的长度。
值(最小值 - 最大值)	显示列中的最小值和最大值。
平均值	显示列的值的平均值。
总和	显示列中所有值的总和。

配置文件结果比较的详细视图中的值

配置文件结果比较的详细视图中的“值”窗格显示了如下值属性：非重复值的个数、最小值、最大值、最大长度、最小长度、平均值、标准偏差以及值总和。

配置文件结果比较的列的详细视图在水平条形图中显示值属性、值和值频率。

下表介绍了比较两次配置文件运行的结果时详细视图中的列值的属性：

属性	说明
非重复值的个数	显示列中非重复值的个数。
最小值	显示列中的最小值。
最大值	显示列中的最大值。
长度(最小 - 最大)	显示列中最短值和最长值的长度。
平均值	显示列的值的平均值。
标准偏差	显示所有列值之间的标准偏差或变异性。
总和	显示列中所有值的总和。

配置文件运行的类型

可以查看最新配置文件运行、历史配置文件运行与合并的配置文件运行的配置文件结果。您可以在摘要视图中查看配置文件运行结果。

最新配置文件运行

可以在摘要视图中查看最新配置文件运行的配置文件结果。

在以下情况下，可以在摘要视图中查看最新配置文件运行的配置文件结果：

- 创建、保存和运行配置文件。
- 在库工作区中打开之前运行的配置文件。
- 对于合并的配置文件运行，在摘要视图或详细视图中单击[返回最新的配置文件运行](#)链接。
- 对于历史配置文件运行，在摘要视图或详细视图中单击[返回最新的配置文件运行](#)链接。
- 在[选择配置文件运行](#)对话框中选择最新配置文件运行，然后单击**确定**。

历史配置文件运行

可以在摘要视图中查看先前的配置文件运行的配置文件结果。

剖析仓库会保存配置文件的所有运行的配置文件结果。您可以通过在“选择配置文件运行”对话框中选择先前版本的配置文件运行来查看该配置文件运行的结果。

已合并配置文件运行

可以在摘要视图中查看配置文件中每个列的最新配置文件结果。

在合并的配置文件运行中，您可以查看配置文件中每个列的最新结果。如果在[选择配置文件运行](#)对话框中选择已合并配置文件运行，则剖析仓库会从所有配置文件运行中检索最新列结果。您可以在摘要视图中查看结果，摘要视图标题会显示增量配置文件运行。

示例

作为数据分析人员，您可以查看配置文件中每个列的最新结果。例如，您可以选择列 1、2 和 3 以执行配置文件运行 A，并选择列 3、4 和 5 以执行配置文件运行 B。要查看所有列的最新结果，您可以在“选择配置文件运行”对话框中选择合并的配置文件运行。摘要视图会显示运行 A 返回的列 1 和 2 的结果，并显示运行 B 返回的列 3、4 和 5 的结果。

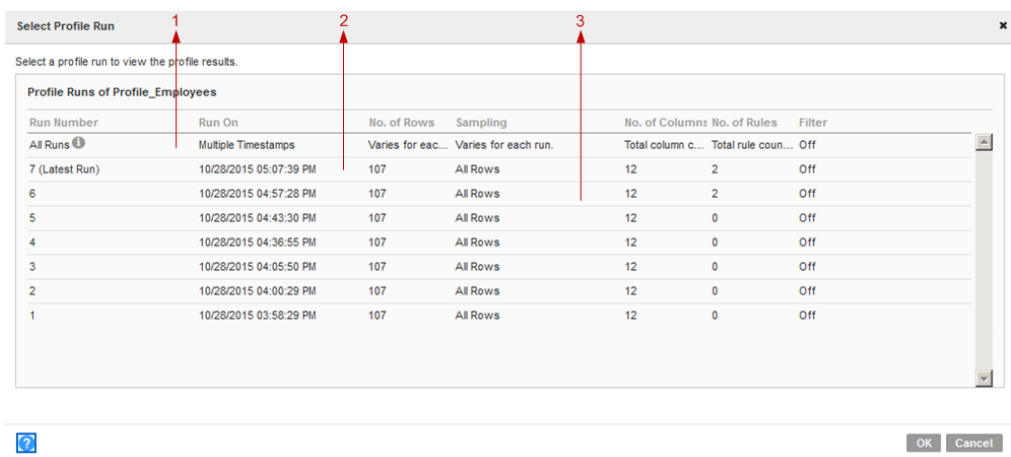
选择配置文件运行

您可以选择历史配置文件运行、最新配置文件运行或合并的配置文件运行，以查看配置文件结果。您可以在摘要视图中查看配置文件结果，而在详细视图中查看列结果。

1. 在库工作区中，选择包含配置文件的项目或文件夹，或者在资产窗格中选择配置文件。
2. 单击**操作 > 打开**，以打开配置文件。
摘要视图将显示在发现工作区中。
3. 在摘要视图中，单击**操作 > 选择配置文件运行**。

此时将显示[选择配置文件运行](#)对话框。

下图显示了[选择配置文件运行](#)对话框。



1. 合并的配置文件运行。选择此配置文件运行时，可以在摘要视图中查看每列的最新配置文件结果。
2. 最新配置文件运行。选择此配置文件运行时，您可以在摘要视图中查看配置文件的最新配置文件结果。
3. 历史配置文件运行。选择此配置文件运行时，您可以在摘要视图中查看先前的配置文件运行的历史配置文件结果。
4. 在**选择配置文件运行**对话框中，选择其中一次配置文件运行以查看其配置文件结果：
 - 要查看最新配置文件运行的配置文件结果，请选择最新配置文件运行，然后单击**确定**。
 - 要查看历史配置文件运行的配置文件结果，请选择最新配置文件运行以外的配置文件运行，然后单击**确定**。
 - 要查看合并的配置文件运行的配置文件结果，请单击**全部运行**，然后单击**确定**。每个列的最新配置文件结果将显示在摘要视图中。

Analyst 工具将执行配置文件运行，并在摘要视图中显示配置文件结果。

5. 在摘要视图中，单击一个列以查看列结果。
此时将显示详细视图。

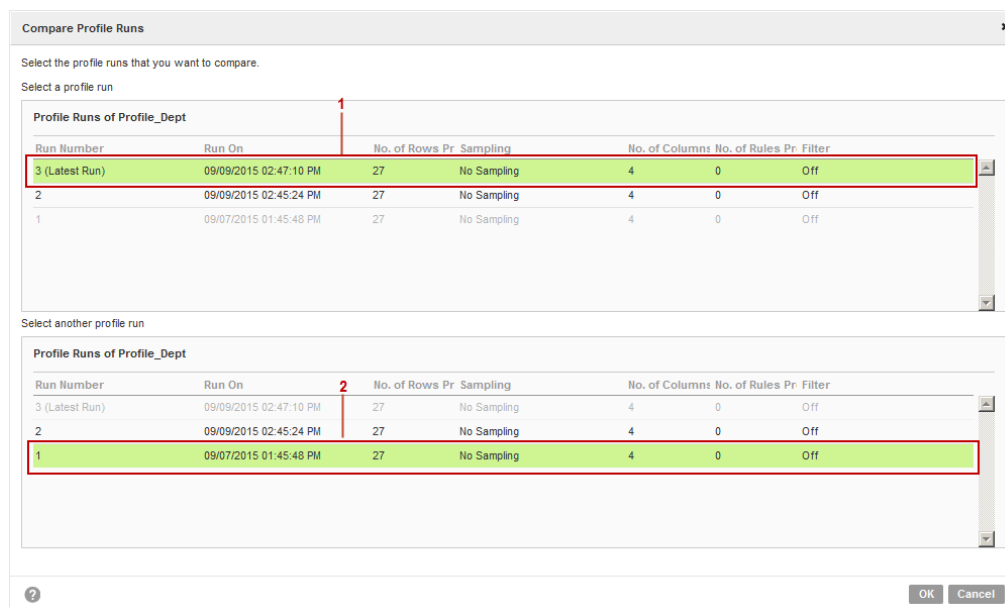
比较多个配置文件结果概览

可以比较两次配置文件运行的配置文件结果。您可以在摘要视图中查看比较结果，在详细视图中查看列结果。在摘要视图中，您可以查看两次配置文件运行的所有列的比较结果。

比较多个配置文件结果

比较两次配置文件运行后，您可以在摘要视图中查看配置文件结果比较。

1. 在摘要视图中，单击**操作 > 比较配置文件运行**。
下图显示了**比较配置文件运行**对话框。



1. 运行 A。选择一次配置文件运行作为“运行 A”。
2. 运行 B。选择一次配置文件运行作为“运行 B”。

此时将显示**比较配置文件运行**对话框。

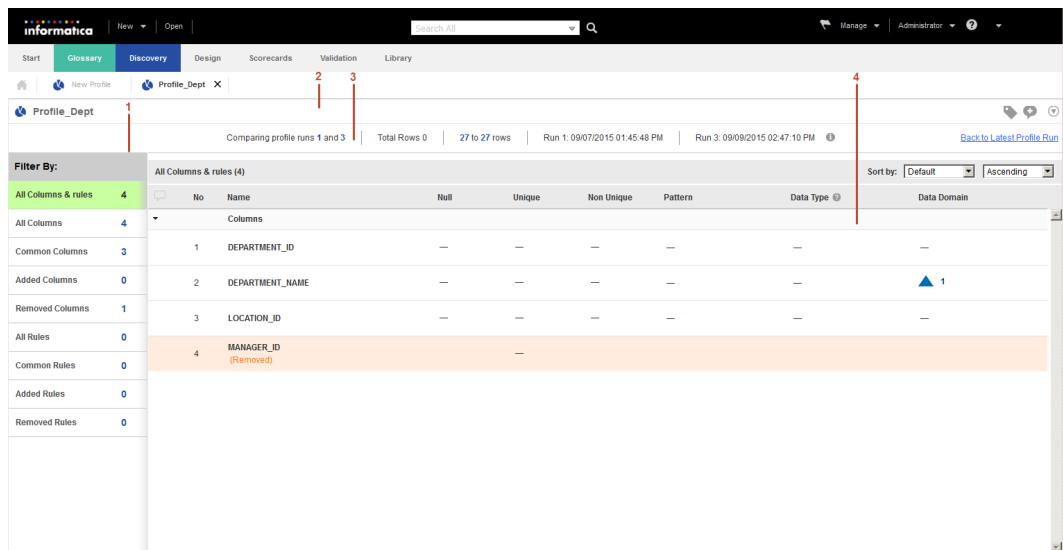
2. 从**运行 A** 窗格中选择一个配置文件，并从**运行 B** 窗格中选择另一个配置文件。
3. 单击**确定**。

摘要视图将显示配置文件结果的合并视图。

配置文件结果比较的摘要视图

比较两次配置文件运行后，您可以在摘要视图中查看网格格式的结果。您可以在摘要视图使用默认筛选器查看特定统计信息。

下图显示了摘要视图中两次配置文件运行的配置文件比较结果：



1. 默认筛选器。在摘要视图中，您可以基于默认筛选器查看配置文件比较结果。
2. 配置文件表头。可以在表头中查看配置文件名。
3. 摘要视图表头。可以在摘要视图表头中查看配置文本特定信息。可以查看所比较的配置文件运行、配置文件各次运行相比行数的增加或减少、配置文件中的行数，以及配置文件运行的时间和日期。
4. 摘要视图。可以查看两次配置文件运行中列之间的比较结果。

配置文件结果比较的摘要视图属性

配置文件结果比较的摘要视图属性包括相异值、非相异值和空值的个数与百分比以及模式、推理的数据类型、推理的数据域和链接的业务术语。摘要视图包含属性的可视化表示形式。您可以单击每个摘要属性，对属性值进行排序。

在摘要视图中，数据集成服务按升序为所有列和规则分配编号。

注意: 带有计数的向上箭头指示其中一次配置运行的属性值比另一次增加了。带有计数的向下箭头指示其中一次配置运行的属性值比另一次减少了。

下表介绍了配置文件结果比较的摘要属性：

属性	说明
编号	显示列或规则的编号。
名称	显示配置文件中的列或规则的名称。
空值	显示空值的增加或减少。
相异	显示相异值的增加或减少。
非相异	显示非相异值的增加或减少。
模式	显示两次配置文件运行之间的模式差异。
数据类型	显示两次配置文件运行之间列或规则的推理的数据类型的差异。
数据域	显示两次配置文件运行之间与列或规则关联的推理的数据域的差异。

摘要视图中配置文件结果比较的默认筛选器

您可以基于默认筛选器在摘要视图中查看配置文件结果。

在摘要视图中，您可以查看源列和虚拟列。规则的输出在摘要视图中显示为虚拟列。更改规则的输出端口并将配置文件运行与历史运行进行对比时，历史规则输出列会显示在**移除的规则**筛选器中，并且新规则输出列会显示在**添加的规则**筛选器中。如果更改单个输出规则的规则逻辑，或者如果更改配置文件运行中的多个规则输出的输入，并将其与历史运行进行比较，**添加的规则**和**移除的规则**筛选器输出不会发生更改。筛选器输出不会发生更改是因为筛选器仅将对列的名称更改视为对筛选器的有效输入。

您可以使用以下默认筛选器选项来查看满足特定条件的配置文件结果：

默认筛选器选项	说明
所有列和规则	显示源列、虚拟列和规则列的配置文件结果。您可以展开和折叠源列与规则列来查看结果。
所有列	显示源列和虚拟列的配置文件结果。
共同列	显示在两个配置文件运行结果中都可用的列。
添加的列	显示在最新配置文件运行中可用的列。例如，如果将运行 5 与运行 3 相比较，则“添加的列”会显示在运行 5 中可用而在运行 3 中不可用的列。
移除的列	显示在历史配置文件运行中可用的列。例如，如果将运行 5 与运行 3 相比较，则“移除的列”会显示在运行 3 中可用而在运行 5 中不可用的列。
所有规则	显示所有规则列的配置文件结果。
添加的规则	显示在最新配置文件运行中可用的规则。例如，如果将运行 5 与运行 3 相比较，则“添加的规则”会显示在运行 5 中可用而在运行 3 中不可用的规则。
移除的规则	显示在历史配置文件运行中可用的规则。例如，如果将运行 5 与运行 3 相比较，则“移除的规则”会显示在运行 3 中可用而在运行 5 中不可用的规则。

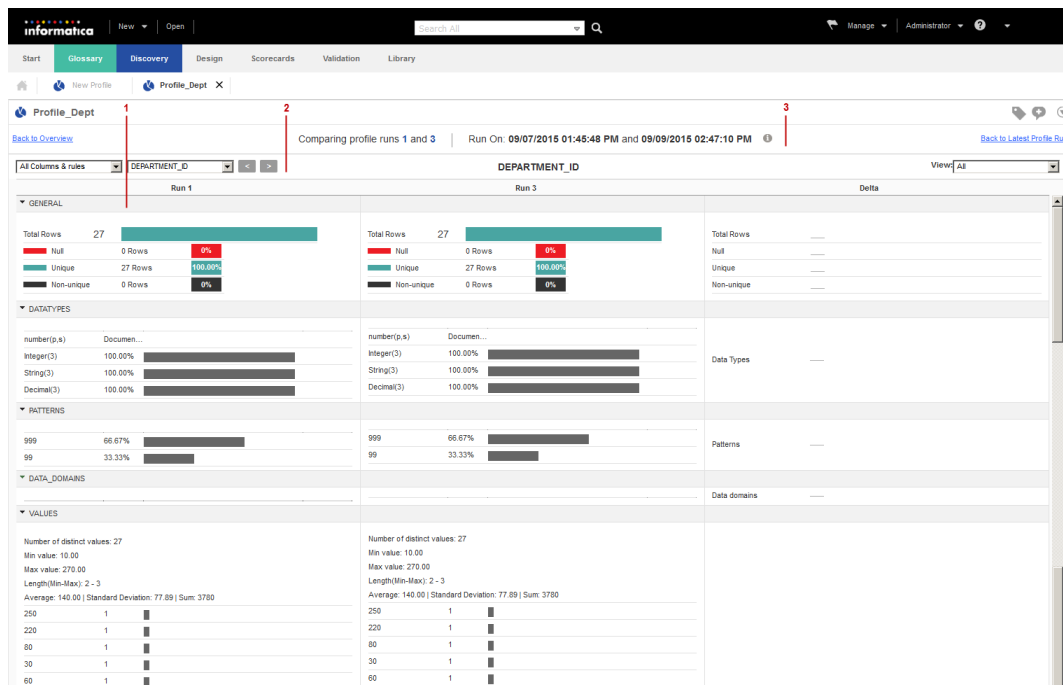
摘要视图默认显示所有源列和虚拟列的配置文件结果。

配置文件结果比较的详细视图

列结果以网格格式显示在详细视图中。列详细信息包含常规信息，例如，相异值、非相异值、空值、模式、数据类型、数据域、业务术语、值和数据预览。

单击列名称时，将显示该列的详细视图。您可以在单独的列中查看运行 A 和运行 B 的列结果，两列数据的比较结果将显示在增量列中。

下图显示了详细视图中某个列的配置文件结果比较：



1. 窗格。在窗格中，可以查看配置文件结果以及两次配置文件运行中列的统计信息，还可查看两次配置文件运行中列的增量信息。
2. 配置文件表头。可以通过在下拉列表中选择列或者使用导航按钮来查看列结果。可以查看列名称，还可使用“视图”下拉列表中的选项来查看特定结果。
3. 摘要视图表头。可以在摘要视图表头中查看配置文本特定信息。可以查看所比较的配置文件运行以及配置文件运行的时间和日期。

配置文件结果比较的详细视图窗格

详细视图详细地显示了两次配置文件运行的配置文件结果和某个列的比较结果。

详细视图显示了运行 A 和运行 B 的列结果，而数据比较会显示在增量列中。要查看其他列结果，您可以从筛选器下拉列表中选择筛选器，或从列下拉列表中选择列。

列配置文件向下钻取

在列配置文件中，使用向下钻取选项可根据列值向下钻取到数据源中的特定行。可以选择读取数据源中的当前数据进行向下钻取，或者读取暂存在剖析仓库中的配置文件数据。当向下钻取到暂存配置文件数据中的特定行时，Analyst 工具将创建一个向下钻取筛选器以用于匹配列值。在向下钻取后，可以编辑、重新调用、重置和保存向下钻取筛选器。

可以选择要进行向下钻取的列，即使您并未选择这些列进行剖析也可以这样做。可以选择读取数据源中的当前数据进行向下钻取，或者读取暂存在剖析仓库中的配置文件数据。在对某一列值执行向下钻取后，可将所选值或模式的向下钻取数据导出到位于所选位置的 CSV 文件。虽然 Informatica Analyst 仅显示向下钻取数据的前 200 个值，但该工具会将所有值导出到 CSV 文件。

向下钻取行数据

运行配置文件后，可以向下钻取到与列值、数据类型或模式匹配的特定行。

1. 运行配置文件。
配置文件结果将显示在摘要视图中。
2. 在摘要视图中，单击一个列名称。
列结果将显示在详细视图中。
3. 在详细视图中，右键单击**值**窗格中的值，然后选择**向下钻取**。
数据预览窗格将显示向下钻取数据。

应用筛选器以向下钻取数据

可以按照迭代方式筛选向下钻取数据，以便能够针对配置文件结果的子集分析数据不规范性。

1. 在**值**选项卡上选择一个列值。
2. 右键单击并选择**向下钻取**。
向下钻取结果将显示在**数据预览**窗格中。
3. 要添加筛选条件，请右键单击**数据预览**窗格中的某一列值，然后选择**添加到筛选器**。
此时将显示**向下钻取筛选器**对话框，其中包含筛选条件。
4. 添加所需的筛选条件，然后单击**确定**。
无法将向下钻取筛选器应用于推理数据类型。
5. 要保存筛选器，请单击**操作 > 保存筛选器**。
6. 要清除向下钻取筛选器，请单击**操作 > 刷新**。
7. 要将向下钻取数据导出到 Microsoft Excel 电子表格，请单击**操作 > 导出数据**。

Analyst 工具中的内容管理

内容管理是验证和管理数据源的已发现元数据的过程，以便元数据适合使用和报告。在 Analyst 工具中管理元数据时，可以批准、拒绝和重置配置文件结果中的推理的数据类型或数据域。

您可为一个列批准一个数据类型和一个数据域。可以对列隐藏拒绝的数据类型或数据域。批准或拒绝推理的数据类型或数据域后，您可以重置数据类型或数据域以还原已推理状态。

批准数据类型和数据域

配置文件结果包含数据源中每个列的推理的数据类型和数据域。您可以在 Analyst 工具中选择并批准每个列的单个数据类型和单个数据域。

1. 运行配置文件。
配置文件结果将显示在摘要视图中。
2. 在摘要视图中，单击一个列名称。
列结果将显示在详细视图中。
3. 在详细视图中，在**数据类型**窗格中选择数据类型，或者在**数据域**窗格中选择数据域。

- 4. 单击**操作 > 批准**。
- 5. 要还原数据类型或数据域的已推理状态，请选择该数据类型或数据域，然后单击**操作 > 重置**。

拒绝数据类型和数据域

在详细视图中，您可以拒绝数据类型和数据域，还可以显示或隐藏拒绝的数据类型和数据域。

- 1. 运行配置文件。
配置文件结果将显示在摘要视图中。
- 2. 在摘要视图中，单击一个列名称。
列结果将显示在详细视图中。
- 3. 在详细视图中，在**数据类型**窗格中选择数据类型，或者在**数据域**窗格中选择数据域。
- 4. 单击**操作 > 拒绝**。
Analyst 工具会从数据类型列表中删除拒绝的数据类型。
- 5. 要查看拒绝的数据类型，请单击**操作 > 显示拒绝项**。

Informatica Analyst 中的列配置文件导出文件

可以将列配置文件结果导出到 CSV 文件或 Microsoft Excel 文件，这取决于您选择的是部分配置文件结果还是完整的结果摘要。

可将所选值和模式的值频率、模式频率、数据类型或向下钻取数据导出到 CSV 文件。可以将所有列的剖析结果摘要导出到 Microsoft Excel 文件。使用数据集成服务权限**向下钻取和导出结果**来确定是由用户还是由组来导出配置文件结果。

CSV 文件格式的配置文件导出结果

可以导出值频率、模式频率、数据类型或向下钻取数据，以在文件中查看数据。Analyst 工具会将信息保存在 CSV 文件中。

在导出推理列模式时，Analyst 工具将导出不同格式的列模式。例如，在导出推理列模式 X(5) 时，Analyst 工具将在 CSV 文件中显示以下格式的列模式：XXXXX。

Microsoft Excel 格式的配置文件导出结果

在导出完整配置文件结果摘要时，Analyst 工具会将信息保存至一个 Microsoft Excel 文件内的多个工作表中。Analyst 工具将以“xlsx”格式保存该文件。

下表描述了导出文件中的每个工作表上显示的信息：

选项卡	说明
列配置文件	摘要信息将在配置文件运行后从摘要视图中导出。示例包括列名称、规则名称、相异值数量、空值数量、推理的数据类型，以及上次运行配置文件的日期和时间。
值	列和规则的值以及这些值在每一列显示的频率。

选项卡	说明
模式	运行配置文件的列和规则的值模式以及这些模式显示的频率。
数据类型	列的所有数据类型、每种数据类型的频率、百分比值以及数据类型的状态（例如，“已推理”、“已批准”或“已拒绝”）。
统计信息	每一列和每条规则的统计信息。例如平均值、长度、最高值、最低值和标准偏差。
属性	“属性”视图信息，包括配置文件名称、类型、采样策略和行计数。

从 Informatica Analyst 中导出配置文件结果

可将配置文件的结果导出到“.csv”或“.xlsx”文件中，以便在一个文件中查看数据。

1. 在库工作区中，选择包含配置文件的项目或文件夹。
2. 单击配置文件将其打开。
配置文件结果将显示在摘要视图中。
3. 在摘要视图中，单击**操作 > 导出数据**。
此时会显示**将数据导出到文件**对话框。
4. 在**将数据导出到文件**对话框中，输入文件名。或者，使用默认文件名。
5. 选择**所有(摘要、值、模式、统计信息、属性)**或**数据域发现结果**，然后选择**代码页**。单击**确定**。
数据将导出到 Microsoft Excel 电子表格。
6. 单击摘要视图中的列。
列结果将显示在详细视图中。
7. 在详细视图中，单击**操作 > 导出数据**。
此时会显示**将数据导出到文件**对话框。
8. 在**将数据导出到文件**对话框中，输入文件名。或者，使用默认文件名。
9. 选择以下选项之一：
 - 所有（摘要、值、模式、统计信息、属性）
 - 选定列的值频率。
 - 选定列的模式频率。
 - 选定列的数据类型。
 - 选定值的向下钻取数据。
 - 选定模式的向下钻取数据。
 - 选定数据类型的向下钻取数据。
10. 输入文件格式。对于**全部**选项，格式为**Excel**；对于其余选项，格式为**CSV**。可以选择将字段名称导出为文件中的第一行。
11. 选择文件的代码页。
12. 单击**确定**。
数据将导出到文件。

第 10 章

Informatica Analyst 中的业务术语、注释和标记

本章包括以下主题：

- [Informatica Analyst 中的业务术语、注释和标记概览, 70](#)
- [业务术语, 70](#)
- [注释, 71](#)
- [标记, 71](#)

Informatica Analyst 中的业务术语、注释和标记概览

您可以向配置文件或配置文件中的列添加业务术语、注释和标记。可以在摘要视图和详细视图中分配、查看和编辑业务术语、注释和标记。

业务术语

可以在 Analyst 工具中为配置文件中的列分配业务术语。您可以编辑资产链接或删除列的业务术语。业务词汇表是一系列使用业务语言为企业用户定义概念的术语。业务术语提供了概念的业务定义和用法。

您可以在摘要视图和详细视图的列中分配、查看或删除业务术语。要在**词汇表**工作区中查看业务术语，请在详细视图中单击业务术语。

您可以编辑业务术语的资产链接的属性。您可以将业务术语作为虚拟列与配置文件结果一起导出到 CSV 文件。

为列分配业务术语

可以在摘要视图和详细视图中为配置文件中的列分配业务术语。您可以在**业务术语**面板中删除列的业务术语。可以在**编辑资产链接**对话框中编辑业务术语的资产链接的属性。

1. 在摘要视图中，右键单击列名称并选择**管理业务术语**。在详细视图中，从**业务术语**选项卡的**操作菜单**中选择**管理业务术语**。
此时将显示**业务术语**面板。
2. 单击加号 (+) 图标或**分配业务术语**链接，以添加业务术语。

此时将显示**分配业务术语**面板。

3. 从**分配业务术语**面板的业务术语列表中选择业务术语。输入资产名称。或者，也可为该资产添加上下文和说明。单击**确定**。

该业务术语将显示在**业务术语**面板中。

注释

您可以向配置文件和配置文件中的列添加注释，以便可以提供更多信息用于进行进一步协作和分析。

在配置文件级别，您可以添加有关配置文件、配置文件定义或配置文件元数据的注释。您可以在摘要视图中查看配置文件注释。

可以在摘要视图和详细视图中添加和查看列注释。

您可以对注释执行以下任务：

- 将注释作为虚拟列与配置文件结果一起导出到 CSV 文件。该 CSV 文件包含配置文件及其中的列的所有注释。
- 使用关键字在注释列中搜索配置文件结果。
- 向配置文件中的源列和虚拟列添加注释。

注意：如果您未选择任何列，或者未添加任何列注释，则摘要视图中的注释面板将显示配置文件注释。

向配置文件或列添加注释

可以在**注释**面板中添加或查看注释。

1. 可以在摘要视图或详细视图中添加注释。
 - 在摘要视图中，要添加配置文件注释，请单击**操作 > 显示注释**。
 - 在摘要视图中，要添加列注释，请右键单击列，然后选择**显示注释**。
 - 在详细视图中，单击**常规**窗格中的**添加注释**。

此时将显示**注释**面板。

2. 单击**添加注释**。

此时将在**注释**面板中显示一个文本框。

3. 添加描述性注释文本，然后单击**保存**。

该注释将与当前用户名以及创建日期和时间一起显示在**注释**面板中。

标记

您可为配置文件或配置文件中的列分配标记，以根据业务用途将对象分组。

可以在摘要视图中查看或分配配置文件标记。可以在摘要视图和详细视图中查看或分配列标记。

您可以对列标记执行以下任务：

- 将标记作为虚拟列与配置文件结果一起导出到 CSV 文件。该 CSV 文件包含配置文件及其中的列的所有标记。
- 为配置文件中的源列和虚拟列分配标记。

注意: 如果您未选择任何列, 或者未添加任何列标记, 则摘要视图中的标记面板将显示配置文件标记。

向配置文件或列添加标记

您可以在摘要视图中向配置文件添加标记。可以在摘要视图和详细视图中向列添加标记。

1. 可以在摘要视图或详细视图中添加标记。
 - 在摘要视图中, 要为配置文件分配标记, 请单击**操作 > 显示标记**。
 - 在摘要视图中, 要向列添加标记, 请右键单击列, 然后单击**显示标记**。
 - 在详细视图中, 单击**常规**窗格中的**添加标记**。

此时将显示**标记**面板。

2. 单击加号 (+) 图标或**分配标记**链接, 以分配标记。

此时将显示**分配标记**对话框。

3. 选择一个或多个标记以分配给配置文件或列。单击**确定**打开**标记**面板。

注意: 要创建标记, 请单击**分配标记**面板中的**添加新标记**。

第 11 章

Informatica Analyst 中的结果卡

本章包括以下主题：

- [Informatica Analyst 结果卡概览, 73](#)
- [Informatica Analyst 结果卡进程, 74](#)
- [在 Informatica Analyst 中创建结果卡, 74](#)
- [向现有结果卡添加列, 76](#)
- [向现有结果卡添加列, 76](#)
- [运行结果卡, 77](#)
- [查看结果卡, 77](#)
- [编辑结果卡, 77](#)
- [度量, 78](#)
- [度量组, 79](#)
- [对列进行向下钻取, 81](#)
- [趋势图表, 81](#)
- [Informatica Analyst 中的结果卡仪表板, 84](#)
- [Informatica Analyst 结果卡导出文件, 89](#)
- [结果卡通知, 90](#)
- [结果卡沿袭, 93](#)

Informatica Analyst 结果卡概览

结果卡是配置文件中某个列的有效值的图形表示形式。可以创建结果卡，然后向下钻取实时数据或暂存数据。

使用结果卡测量数据质量进度。例如，可以创建一个结果卡，用于在应用数据质量规则之前度量数据质量。应用数据质量规则后，可以创建另一个结果卡，以比较这些规则对数据质量的影响。

结果卡将各列的值频率显示为得分。得分可以反映各列中有效值的百分比。运行配置文件后，可将配置文件中的列作为度量添加到结果卡中。可以创建度量组，以便能将相关度量分组到一个实体中。可以定义阈值，指定记录中的列可接受的错误数据的范围；还可以为每个度量分配度量权重。在运行结果卡时，Analyst 工具将为每个度量组生成加权平均值。要进一步评估数据质量，还可以为每个度量分配一个固定或可变成本。在运行结果卡时，Analyst 工具将为每个度量计算错误数据成本总和，然后显示总成本。

创建或编辑结果卡时，可以基于源数据创建结果卡筛选器。使用结果卡筛选器，可以基于筛选条件重新计算度量得分。要标识有效数据记录和无效记录，可以向下钻取每个度量。可以使用趋势图表来跟踪度量得分，以及度量中无效数据的成本在一段时间内如何变化。可以在结果卡中重复使用配置文件筛选器。

在 Analyst 工具中启用版本控制系统时，可以为结果卡创建多个版本并查看结果卡的版本历史记录。默认情况下，结果卡在创建后会签出。您必须签入结果卡，其他用户才能编辑结果卡。

可以在**结果卡**工作区中查结果卡仪表板。在结果卡仪表板中，您可以查看具有结果卡的数据对象、项目中的结果卡、过去六个月的結果卡运行趋势以及一个月内所有结果卡运行的良好、可接受和不可接受的度量汇总。

您可以在 Informatica Analyst 中配置和管理结果卡的电子邮件通知。请使用电子邮件服务管理电子邮件通知。电子邮件服务是一项系统服务，可在 Informatica Administrator 中配置。

Informatica Analyst 结果卡进程

可以在 Developer tool 和 Analyst 工具中创建和编辑结果卡。可以在 Analyst 工具中运行结果卡。可以针对数据对象中的当前数据或暂存在剖析仓库中的数据运行结果卡。

可以在**结果卡**工作区中查看结果卡。在运行结果卡后，可以在**结果卡**面板上查看得分。可以选择数据对象，然后从结果卡中的某一得分导航到该数据对象。Analyst 工具将在另一个选项卡中打开该数据对象。

在使用结果卡时，可以执行以下任务：

1. 在 Developer tool 或 Analyst 工具中创建结果卡，然后从配置文件中添加列。
2. 在 Analyst 工具中打开结果卡。
3. 在运行配置文件后，将配置文件的列作为度量添加到结果卡中。
4. 或者，基于源数据创建结果卡筛选器。
5. 或者，为每个度量配置无效数据的成本。
6. 运行结果卡以生成各列的得分。
7. 查看结果卡，以查看记录中每一列的得分。
8. 对列进行向下钻取以获得得分。
9. 编辑结果卡。
10. 为结果卡中的每个度量设置阈值。
11. 创建组，以添加或移动结果卡中的相关度量。
12. 根据需要编辑或删除组。
13. 查看每个得分的得分趋势图表，以监视得分如何随着时间推移而变化。
14. 或者，查看每个度量的成本趋势图表，以监视数据质量的值。
15. 查看每个度量或度量组的结果卡沿革。
16. 查看您拥有读取权限的结果卡的合并信息。

在 Informatica Analyst 中创建结果卡

创建结果卡，然后将配置文件中的列添加到该结果卡中。必须首先运行配置文件，然后才能将列添加到结果卡中。

1. 在**库**工作区中，选择包含配置文件的项目或文件夹。
2. 单击配置文件将其打开。

配置文件结果将显示在**发现**工作区的摘要视图中。

3. 单击**操作 > 添加到结果卡**。

此时将显示**添加到结果卡**向导。

4. 在**添加到结果卡**屏幕中，可以选择创建新结果卡，也可以编辑现有结果卡以向预定义结果卡添加列。默认情况下将选择**新建结果卡**选项。单击**下一步**。

5. 在**第 2 步(共 8 步)**屏幕中，输入结果卡的名称。还可以选择输入结果卡的说明。选择用于保存结果卡的项目和文件夹。单击**下一步**。

默认情况下，结果卡向导将选择配置文件中定义的列和规则。无法添加未包括在配置文件中的列。

6. 在**第 3 步(共 8 步)**屏幕中，选择要作为度量添加到结果卡的列和规则。或者，单击左侧列表头中的复选框，以选择所有列。或者，选择**列名称**，以对列名称进行排序。单击**下一步**。

7. 在**第 4 步(共 8 步)**屏幕中，可以将筛选器添加到度量。

可以将为配置文件创建的筛选器应用于度量，或创建新筛选器。在**度量筛选器**窗格中选择一个度量，然后单击**管理筛选器**图标以打开**编辑筛选器: 列名称**对话框。在**编辑筛选器: 列名称**对话框中，可以选择执行以下任务之一：

- 选择为配置文件创建的筛选器。单击**下一步**。
- 选择现有筛选器。单击编辑图标，以在**编辑筛选器**对话框中编辑筛选器。单击**下一步**。
- 单击加号 (+) 图标，以在**新建筛选器**对话框中创建筛选器。单击**下一步**。

或者，您可以选择将选定的筛选器应用于结果卡中的所有度量。

筛选器将显示在**度量筛选器**窗格中。

8. 在**第 4 步(共 8 步)**屏幕中，单击**下一步**。

9. 在**第 5 步(共 8 步)**屏幕中，选择**度量**窗格中的每个度量以执行下列任务：

- 配置有效值。在**正在使用的得分: 值**窗格中，选择**可用值**窗格中的一个或多个值，然后单击向右箭头按钮，将这些值移至**有效值**窗格中。此时将在**可用值**窗格顶部显示某一度量的有效值的总数。
- 配置度量阈值。在**度量阈值**窗格中，为**正常**、**可接受**和**不可接受**得分设置阈值。
- 配置无效数据的成本。要为度量的成本分配常量值，请选择**固定成本**。要将数字列作为可变成本附加到度量，请选择**可变成本**，然后单击**选择列**，以选择某一数字列。或者，单击**更改成本单位**，以更改成本的单位。如果不希望为度量配置无效数据的成本，请选择**无**。

10. 单击**下一步**。

11. 在**第 6 步(共 8 步)**屏幕中，可以选择要将度量添加到的度量组，或创建新度量组。要创建新度量组，请单击组图标。单击**下一步**。

12. 在**第 7 步(共 8 步)**屏幕中，指定组中度量的权重以及组的阈值。

13. 在**第 8 步(共 8 步)**屏幕中，选择**本地**或**Hadoop**作为运行结果卡的运行时环境。在 Hadoop 运行时环境中，可以选择**Blaze**或**Spark**引擎。如果选择**Blaze**或**Spark**引擎，请单击**浏览**以选择 Hadoop 连接来运行配置文件。

14. 单击**保存**以保存结果卡，或者单击**保存并运行**以保存并运行结果卡。

结果卡将显示在**结果卡**工作区中。

向现有结果卡添加列

运行配置文件后，可将配置文件结果中的列添加到现有结果卡中。您可以添加度量或度量组，为列配置有效值，并为每个度量添加无效数据的成本。如果使用**所有行**以外的采样选项将配置文件中的某一列添加到结果卡中，则结果卡可能不会反映配置文件结果。

当您可以向现有结果卡添加列时，在**添加到结果卡**向导中无法编辑结果卡的现有度量或度量组。要修改结果卡中的现有度量，请导航到结果卡工作区，编辑结果卡，然后根据需要更新度量或度量组。

向现有结果卡添加列

运行配置文件后，可向现有结果卡添加列。

1. 单击某一配置文件以打开它。
配置文件结果将显示在摘要视图中。
2. 选择一列。单击**操作 > 添加到结果卡**。
此时将显示**添加到结果卡**向导。
注意: 在将列添加到结果卡中之前，使用以下规则和准则：
 - 如果列名称与结果卡名称均匹配，则无法将列添加到结果卡中。
 - 即使更改列名称，也无法将同一列两次添加到一个结果卡中。
3. 选择**现有结果卡**，以将列添加到预定义结果卡中。单击**下一步**。
4. 在**第 2 步(共 7 步)** 屏幕中，选择要添加列的结果卡。单击**下一步**。
您可以查看与结果卡关联的现有度量和度量组。
5. 在**第 3 步(共 7 步)** 屏幕中，选择要作为度量添加到结果卡的列和规则。或者，单击左侧列表头中的复选框，以选择所有列。单击**列名称**，以对列名称进行排序。单击**下一步**。
6. 在**第 4 步(共 7 步)** 屏幕中，可以为度量创建筛选器。也可以将为配置文件创建的筛选器应用到度量。
7. 在**第 5 步(共 7 步)** 屏幕中，可以执行以下任务：
 - 在**度量窗格**中，选择各个度量并配置其他窗格中的度量值。
 - 在**正在使用的得分:值**窗格中，选择**可用值**窗格中的多个值，然后单击向右箭头按钮，将这些值移至**有效值**窗格中。
此时将在**可用值**窗格顶部显示某一度量的有效值的总数。
 - 在**度量阈值: 窗格**中，可以为**正常、可接受和不可接受**得分设置阈值。
 - 在**无效数据的成本**中，可以：
 - 选择每个度量，然后为度量配置无效数据的成本。
 - 选择**固定成本**选项，以为度量的成本分配常量值。可以单击**更改成本单位**，以更改成本的单位。
 - 选择**可变成本**选项，以将数字列作为可变成本附加到度量。可以单击**选择列**以选择数字列。
8. 单击**下一步**。
9. 在**第 6 步(共 7 步)** 屏幕中，可以执行以下任务：
 - 选择要添加度量的度量组。
 - 在**默认 - 度量**窗格中，可以双击默认度量权重 0，以更改该值。
 - 在**度量阈值:窗格**中，可以为 **正常、可接受和不可接受**得分设置阈值。

10. 单击**下一步**。
11. 在**第 7 步(共 7 步)** 屏幕中，选择运行时环境。
12. 单击**保存**以保存结果卡，或者单击**保存并运行**以保存并运行结果卡。

运行结果卡

可以运行结果卡，以生成各列的得分。

1. 在**资产**面板中，选择要运行的结果卡。
2. 单击该结果卡以打开它。
结果卡显示在**结果卡**工作区中。
3. 单击**操作 > 运行结果卡**。
4. 从**度量**窗格中选择得分，然后从**列**窗格中选择要对其进行向下钻取的列。
5. 在**向下钻取**选项中，选择对实时数据或暂存数据进行向下钻取。
要获得最佳性能，请对实时数据进行向下钻取。
6. 单击**运行**。

查看结果卡

运行结果卡可以查看每个度量的得分。结果卡将以百分比和条形图的形式显示得分。可以查看有效数据或无效数据。还可以查看结果卡信息，如度量权重、度量组得分、得分趋势和数据对象的名称。

1. 运行结果卡以查看得分。
2. 选择包含要查看的得分的度量。
3. 单击**操作 > 向下钻取**，以查看相应列的有效数据行或无效数据行。
默认情况下，Analyst 工具将在**向下钻取**部分中显示无效数据行。

编辑结果卡

可以在结果卡中编辑度量的有效值。必须运行结果卡，然后才能对其进行编辑。

1. 在**库**工作区中，在**资产**窗格中单击要编辑的结果卡。
结果卡将显示在**结果卡**工作区中。
2. 如果启用了版本控制系统，则单击**操作 > 签出**。
3. 单击**操作 > 编辑 > 常规**。
此时将显示**编辑结果卡**对话框。
4. 在**常规**选项卡中，您可以根据需要，编辑结果卡的名称和说明。

5. 单击**度量**选项卡。
6. 选择**度量**窗格中的一个得分，然后在**正在使用的得分: 值**窗格中，配置所有值的列表中的有效值。
7. 在**度量阈值**窗格中，根据需要更改得分阈值。
8. 检查每个度量的无效数据的成本，然后根据需要进行更改。
9. 单击**结果卡筛选器**选项卡。
10. 您可以添加、编辑或删除筛选器。
11. 单击**度量组**选项卡。
12. 您可以创建、编辑或删除度量组。
还可以在**度量组**选项卡上编辑度量权重和度量阈值。
13. 单击**通知**选项卡。
14. 根据需要更改结果卡通知设置。
可以为度量和度量组设置全局和自定义设置。
15. 选择**本地**、**Blaze** 或 **Spark** 作为运行时环境。如果选择 **Blaze** 或 **Spark** 作为运行时环境，请单击**浏览**以选择 Hadoop 连接。
16. 单击**保存**将更改保存到结果卡中；或者单击**保存并运行**保存更改并运行结果卡。
17. 单击**签入**。

度量

度量是数据源的某一列，或者是构成结果卡的某一规则的输出。在创建结果卡时，可为每个度量分配一个权重。可以创建度量组，以将结果卡中的相关度量分类到一个集合中。

度量权重

在创建结果卡时，可为每个度量分配一个权重。权重的默认值为 0。

当运行结果卡时，Analyst 工具会根据度量得分和您分配给每个度量的权重计算每个度量组的加权平均值。

例如，为度量 M1 分配权重 W1，然后为度量 M2 分配权重 W2。Analyst 工具将使用下面的公式来计算加权平均值：

$$(M1 \times W1 + M2 \times W2) / (W1 + W2)$$

数据质量的值

源数据中数据质量的度量是管理组织中的数据资产的关键信息。无效数据的成本以度量的方式在结果卡中表示，可以帮助组织在监视源数据的数据质量过程中派生相应的值。作为数据分析人员，您可能想将某个值（如货币单位或任何自定义单位）与度量和度量组关联起来。随后，可以运行结果卡，以查看源数据中无效数据的总成本。

可以根据业务需要定义某一度量的成本单位。还可以在创建或编辑结果卡时，为每个度量配置可变或固定成本。

固定成本

固定成本是可以分配给结果卡中某一度量的常量值。既可选择预定义成本单位，也可以创建符合业务需要的自定义成本单位。

可变成本

可变成本是根据数据源中某一数字列中的值分配给某一度量的值。数据集成服务将根据该数字列或分配给成本的虚拟列来计算度量的可变成本。

示例

作为抵押贷款管理人员，您需要为客户提供缴款书，以便客户能够提交抵押贷款付款。您可以使用结果卡来度量客户地址的准确性，以确保交付缴款书。您可能想为“地址准确性”度量的“每月付款金额”列设置可变成本。运行结果卡，以计算如果客户没有按时支付每月应付金额，抵押贷款组织亏损的总成本。

定义阈值

可以为结果卡中的每个得分设置阈值。阈值以百分比的形式指定记录中列可接受的错误数据的范围。可以为正常、可接受或不可接受的数据范围设置阈值。可以在将列添加到结果卡或编辑结果卡时，为每个列定义阈值。

为结果卡中的列定义阈值之前，需要完成以下先决条件任务之一：

- 打开配置文件，然后在**添加到结果卡**对话框中将该配置文件中的列添加到结果卡中。
 - 或者，在**库**工作区中单击结果卡，选择**操作 > 编辑**，在**编辑结果卡**对话框中编辑结果卡。
1. 在**添加到结果卡**对话框或**编辑结果卡**对话框中，选择**度量**窗格中的每个度量。
 2. 在**度量阈值**窗格中，输入代表不可接受范围上界以及正常范围下界的阈值。
可以设置最多两个小数位的阈值。
 3. 单击**下一步**或**保存**。

度量组

可以创建度量组，将结果卡中的相关得分分类到一个集合中。默认情况下，Analyst 工具会对默认度量组中的所有得分进行分类。

在创建度量组后，可以将得分移出默认度量组，移至其他度量组中。可以编辑度量组以更改其名称和说明，包括默认度量组。可以删除不再使用的度量组。无法删除默认度量组。

创建度量组

可以创建度量组，以将结果卡中的相关得分添加到该组中。

1. 在**库**工作区中，在**资产**窗格中单击要编辑的结果卡。
结果卡将显示在**结果卡**工作区中。
2. 单击**操作 > 编辑**。
此时将显示**编辑结果卡**窗口。
3. 单击**度量组**选项卡。
度量组面板中将显示默认组，而**度量**面板中则显示该默认组中的得分。

4. 单击**新建组**图标以创建度量组。
此时将显示**度量组**对话框。
5. 输入名称和可选说明。
6. 单击**确定**。
7. 单击**保存**以将更改保存到结果卡中。

将得分移至度量组

在创建度量组后，可以将相关得分移至该度量组。

1. 在**库**工作区中，在**资产**窗格中单击要编辑的结果卡。
结果卡将显示在**结果卡**工作区中。
2. 单击**操作 > 编辑**。
此时将显示**编辑结果卡**窗口。
3. 单击**度量组**选项卡。
度量组面板中将显示默认组，而**度量**面板中则显示该默认组中的得分。
4. 从**度量**面板中选择度量，然后单击**移动度量**图标。
此时将显示**移动度量**对话框。
注意: 要选择多个得分，请按住 Shift 键。
5. 选择要向其中移动得分的度量组。
6. 单击**确定**。

编辑度量组

可以编辑度量组，以更改名称和说明。可以更改默认度量组的名称。

1. 在**库**工作区中，在**资产**窗格中单击要编辑的结果卡。
结果卡将显示在**结果卡**工作区中。
2. 单击**操作 > 编辑**。
此时将显示**编辑结果卡**窗口。
3. 单击**度量组**选项卡。
度量组面板中将显示默认度量组，而**度量**面板中将显示该默认度量组中的度量。
4. 在**度量组**面板上，单击**编辑组**图标。
此时将显示**编辑**对话框。
5. 输入名称和可选说明。
6. 单击**确定**。

删除度量组

可以删除不再有效的度量组。在删除度量组时，可以选择将该度量组中的得分移至默认度量组。无法删除默认度量组。

1. 在**库**工作区中，在**资产**窗格中单击要编辑的结果卡。
结果卡将显示在**结果卡**工作区中。

2. 单击**操作 > 编辑**。
此时将显示**编辑结果卡**窗口。
3. 单击**度量组**选项卡。
度量组面板中将显示默认度量组，而**度量**面板中将显示该默认度量组中的度量。
4. 在**度量组**面板中选择度量组，然后单击**删除组**图标。
此时将显示**删除组**对话框。
5. 在删除度量组之前，请选择相应选项以删除度量组中的度量，或者选择相应选项以将度量移至默认度量组。
6. 单击**确定**。

对列进行向下钻取

在某一得分的列上进行向下钻取，以选择在查看有效数据行或无效数据行时显示的列。选择对其进行向下钻取的列将显示在**向下钻取**面板中。

1. 运行结果卡以查看得分。
2. 选择包含您要查看的得分的列。
3. 单击**操作 > 向下钻取**，以查看相应列的有效数据行或无效数据行。
4. 单击**操作 > 向下钻取列**。

相应的列将显示在选定得分对应的**向下钻取**面板中。默认情况下，Analyst 工具将显示相应列的有效数据行。或者，单击**无效**以查看无效数据行。

趋势图表

可以使用趋势图表来监视度量得分和度量中无效数据的成本在一段时间内如何变化。

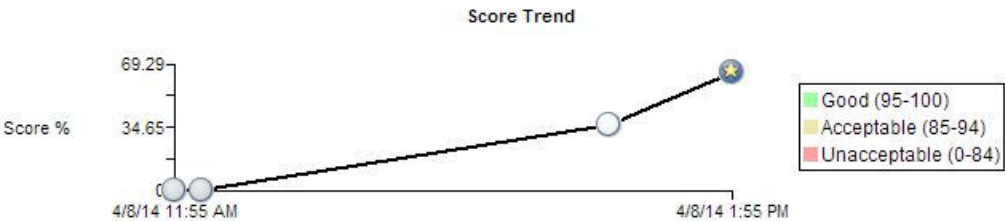
趋势图表包含得分图形和成本图形，在纵轴上绘制了得分值或成本值，而在横轴上则绘制了所有结果卡运行。默认情况下，趋势图表将显示来自最近 10 次结果卡运行的数据。可以在趋势图表中查看相应度量的总行数 and 无效行数。趋势图表还会显示得分和成本趋势是保持不变还是根据上次结果卡运行上移或下移。

Analyst 工具会将历史结果卡运行数据用于每个日期和最新有效得分值，以计算得分。Analyst 工具将使用该图表中的最新阈值设置来描绘得分点的颜色。可以查看得分的“正常”、“可接受”和“不可接受”阈值。编辑结果卡中的得分值后，每次运行结果卡都会更改阈值。在导出结果卡时，Analyst 工具会将趋势图表信息（其中包括得分和成本信息）包括在导出的文件中。

得分趋势图表

得分趋势图表以图形方式表示了度量得分如何随着多次配置文件运行而变化。得分趋势图表在纵轴上绘制度量得分值，而在横轴上绘制所有结果卡运行。

下图显示了得分趋势图表的一个示例：



示例

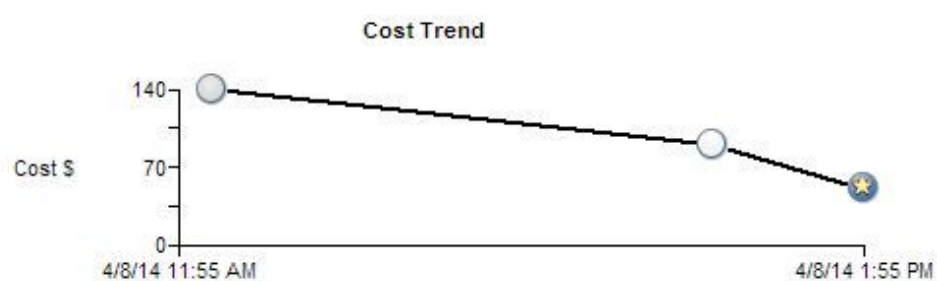
作为数据分析人员，您可以监视数据质量，以分析映射和其他进程更改是否使得数据质量得分提高。在度量数据质量中的更改之后，即可报告组织的数据质量变化，以供分析和使用。例如，在多次结果卡运行结束时，“社会保障号码”列中有效值的百分比可能已从 84 移至 90。您可以使用可视化图表的形式报告数据质量中的这一变化，以供快速分析。

成本趋势图表

成本趋势图表以图形方式表示了度量中无效数据的成本如何随着多次配置文件运行而变化。成本趋势图表可以度量数据质量在组织中的影响。成本趋势图表在纵轴上绘制成本值，而在横轴上绘制所有结果卡运行。还可以查看成本趋势图表下某一网格中的度量的无效数据以及无效值的总成本。

成本趋势图表可以帮助您跟踪无效数据对高价值记录的影响。有时，在使用固定成本计算无效数据时，可能会错过了解无效数据对高价值记录的影响的机会。发生这种问题的原因在于趋势图表可能会显示随着多次结果卡的运行，得分的提高以及总体成本的下降。不过，结果卡中显示的数据质量问题较少的情况可能存在于高价值记录中。

下图显示了成本趋势图表的一个示例：



示例

在某一金融机构中，您有几位高余额客户，在银行内存有大量存款和投资（如 1,000 万美元）。您还拥有大量低余额客户。得分趋势图表可以显示在一段时间内得分的提高。不过，一些高余额客户帐户的地址或性别不正确，可能会影响与该组织最宝贵的客户之间的关系。您可以将“帐户余额”列设置为可变成本列，用于计算无效数据。如果由于该列而使无效数据的成本较高，则可认为总值出于风险之中，并应立即采取更正操作。

查看趋势图表

可以查看每个度量的趋势图表，以监视得分或无效数据的成本如何随着时间推移而变化。

1. 在库工作区中，选择包含结果卡的项目或文件夹。
2. 单击该结果卡以打开它。
结果卡显示在结果卡工作区中。
3. 在结果卡视图中，选择某一度量。
4. 单击操作 > 显示趋势图表。

此时将显示趋势图表详细信息对话框。
下图显示了趋势图表详细信息对话框：



可以查看随时间改变的得分和成本值。在对话框顶部，可以查看总行数和无效行数。Analyst 工具使用每个日期的历史结果卡运行数据和最新有效得分值计算得分。在得分趋势图表和成本趋势图表下，可以查看度量的有效值以及无效数据的成本。

导出趋势图表

可将得分趋势图表和成本趋势图表导出到“.xlsx”文件，以便在一个文件中查看数据。

1. 打开结果卡。
2. 选择一个度量，然后单击操作 > 显示趋势图表。

此时将显示趋势图表详细信息对话框。

3. 单击**导出数据**图标。
此时会显示**将数据导出到文件**对话框。
4. 在**文件名**字段中，输入文件名。（可选）可以使用默认文件名。
默认文件格式为 Microsoft Excel。
5. 在**无效行数字**字段中，输入要导出的无效行数。可以在该字段中输入最大为 100,000 的值。
6. 在**代码页**字段中，可以选择文件的代码页。
7. 单击**确定**。

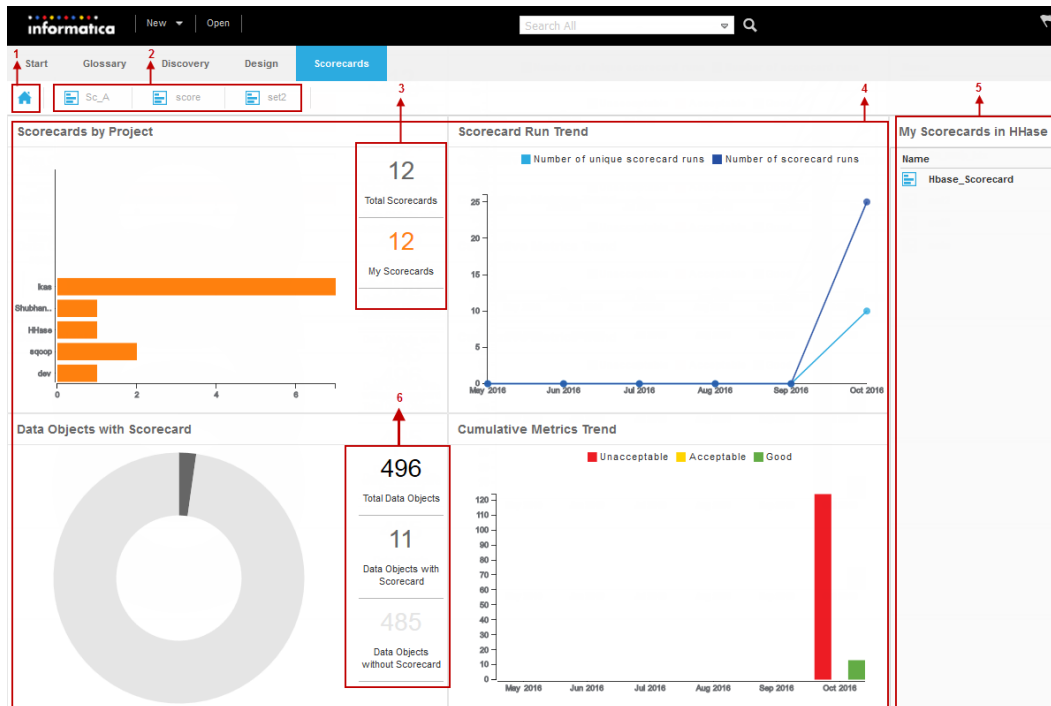
Informatica Analyst 中的结果卡仪表盘

Informatica Analyst 中的**结果卡**工作区会显示结果卡仪表盘。在结果卡仪表盘中，您可以查看具有结果卡的数据对象、过去六个月的结果卡运行趋势、项目中的结果卡、一个月内所有结果卡运行的良好、可接受和不可接受的度量汇总以及资产列表窗格。

如果其他用户修改了结果卡，您的计算机上的结果卡仪表盘不会自动刷新。请使用 F5 功能键或在工作区或结果卡结果选项卡之间切换来刷新结果卡仪表盘。

您可以在窗格中以数据系列或数据点形式查看数据。在图表中，数据点显示为不透明的小圆圈，数据系列显示为水平条、垂直条或切片。

下图显示了**结果卡**工作区中的结果卡仪表盘和资产窗格：



1. 结果卡仪表板图标。显示结果卡仪表板。
2. 结果卡结果选项卡。显示开放结果卡的结果卡结果。
3. “结果卡(按项目)”窗格中的图例。显示所有项目中的结果卡总数以及所有项目中您拥有读取访问权限的结果卡总数。
4. 结果卡仪表板。按照项目、结果卡运行趋势、具有结果卡的数据对象以及仪表板中的累积度量趋势窗格来显示结果卡。
5. 资产列表窗格。显示与图表中的图例、数据系列或数据点关联的结果卡或数据对象列表。
6. “具有结果卡的数据对象”窗格中的图例。显示数据对象总数、具有结果卡的数据对象的数量以及不具有结果卡的数据对象的数量。

在结果卡仪表板中单击数据点或数据系列后，映射到数据点或数据系列的结果卡会显示在资产列表窗格中。在资产列表窗格中单击结果卡后，结果卡结果会显示在**结果卡工作区**的选项卡中。资产列表窗格显示了您拥有读取访问权限的结果卡。

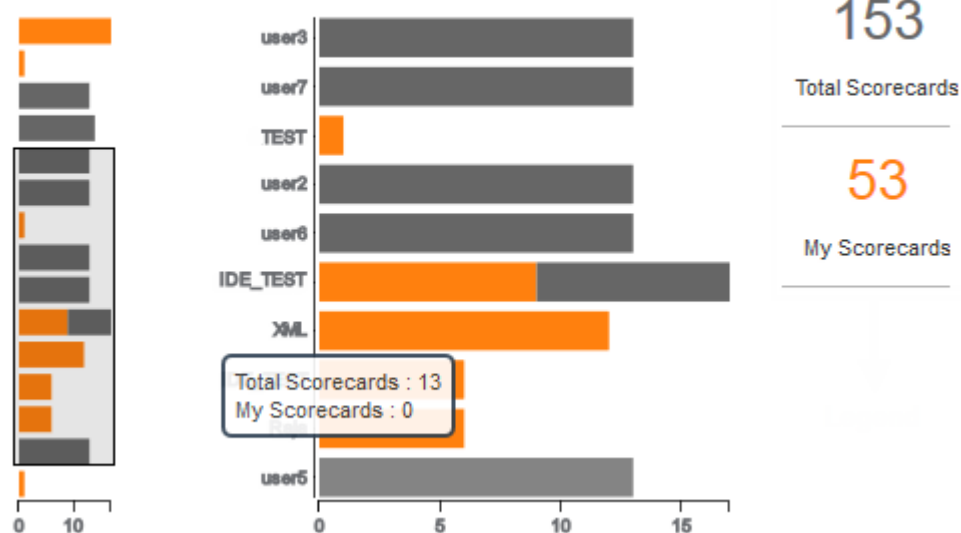
结果卡(按项目)

结果卡(按项目) 窗格会显示模型存储库中的项目，以及以条形图显示每个项目的结果卡数。条形图中的线条表示项目。图表中的 x 轴显示结果卡的数量，而 y 轴显示具有结果卡的项目。

项目中的结果卡在条形图中显示为灰色，而您拥有读取访问权限的结果卡在条形图中显示为橙色。图例中的**结果卡总数**部分会显示模型存储库中的结果卡总数。图例中的**我的结果卡**部分会显示您在模型存储库中拥有读取访问权限的结果卡数量。

下图显示了结果卡仪表板中的**结果卡(按项目)**窗格：

Scorecards by Project



您可以在窗格中查看以下图表：

- 详细图表。显示模型存储库中具有结果卡的所有项目，以及每个项目中的结果卡数量。如果项目的数量大于 10，**结果卡(按项目)** 窗格将显示一个滑块。
- 缩微图。在详细图表中显示滑块中的所有项目以及每个项目的结果卡数。

将指针移到缩微图上时，数据标签中会显示结果卡的总数以及您在项目中拥有权限的结果卡数量。

要查看您在其中拥有读取访问权限的项目的结果卡，请单击水平条的橙色部分。要查看您在模型存储库中拥有读取访问权限的所有结果卡，请在条形图中单击**我的结果卡**。结果卡会显示在资产列表窗格中。在资产窗格中单击结果卡以查看结果卡结果。

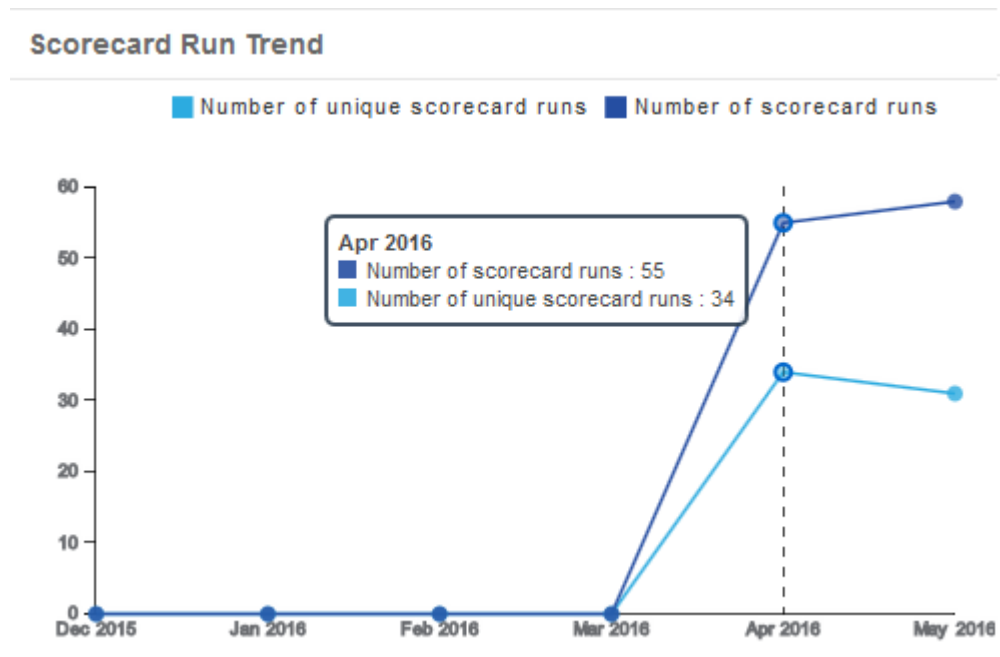
结果卡运行趋势

结果卡运行趋势 窗格会以带标记的折线图显示当前月份和过去五个月的结果卡运行趋势。图表中的 x 轴显示了当前月和过去五个月，y 轴显示了结果卡的数量。标记是折线图中的数据点。在将指针移到图表的标记上时，当月的结果卡运行摘要会显示在数据标签中。

您可以在窗格中查看以下标记：

- **结果卡运行数**。该标记会显示对应月份的结果卡运行总次数。
- **唯一结果卡运行数**。该标记会显示对应月份的唯一结果卡运行总次数。

下图显示了结果卡仪表板中的**结果卡运行趋势**窗格：



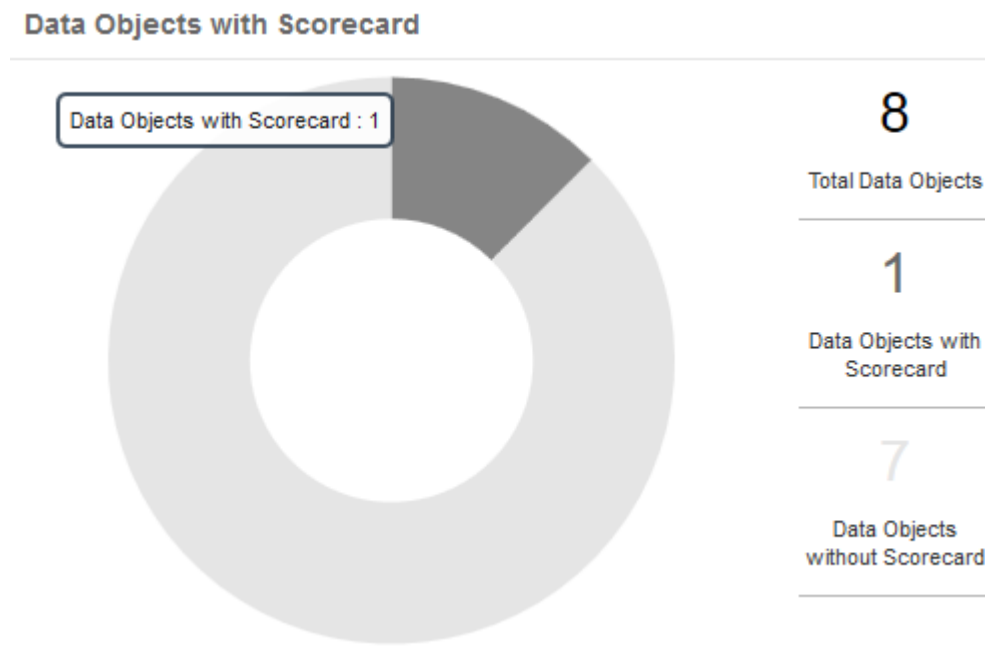
在窗格中单击标记时，映射到标记的结果卡会显示在资产列表窗格中。可以查看您拥有读取权限的结果卡。在资产列表窗格中单击结果卡，然后在**结果卡**工作区中查看结果卡结果。

具有结果卡的数据对象

具有结果卡的数据对象窗格会显示圆环图。在该图表中，可以查看以切片显示的具有和不具有结果卡的数据对象的数量。

在将指针移动到圆环图上时，映射到切片的数据会显示在数据标签中。

下图显示了结果卡仪表板中**具有结果卡的数据对象**窗格：



具有结果卡的数据对象窗格中的图例会显示以下数据统计信息：

- 数据对象总数。显示库工作区的**资产**窗格中的**数据对象**文件夹中的数据对象总数。数据对象包括逻辑数据对象和自定义数据对象。
- 具有结果卡的数据对象。显示具有结果卡的数据对象数量。
- 不具有结果卡的数据对象。显示不具有结果卡的数据对象数量。

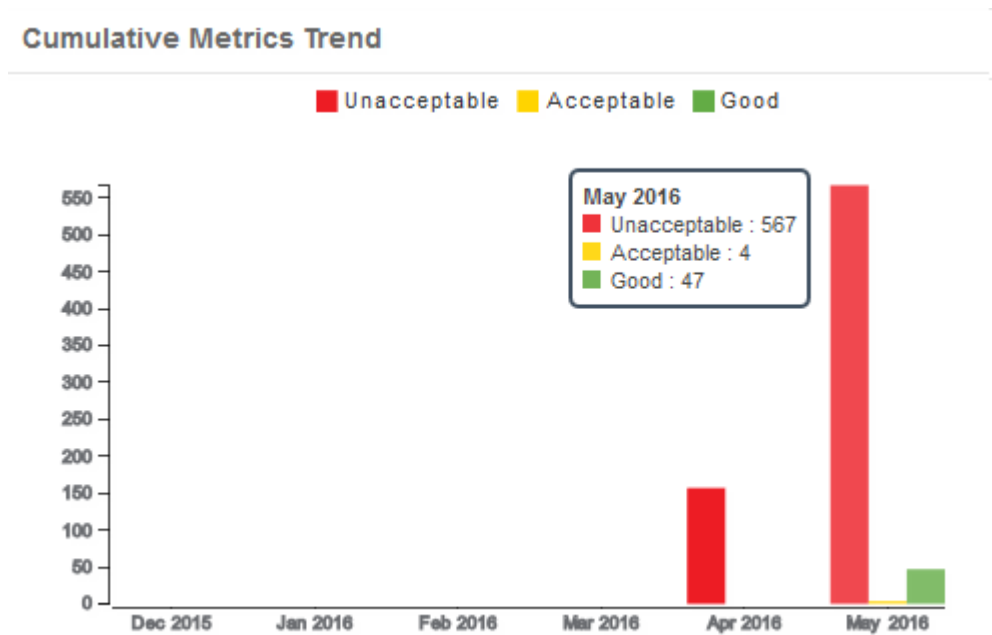
单击圆环图中的切片或**具有结果卡的数据对象**以及**不具有结果卡的数据对象**图例后，映射到圆环图中切片或图例的结果卡会显示在资产列表窗格中。

累积度量趋势

累积度量趋势窗格会显示列图表。您可以在图表中查看以垂直条显示的一个月内所有结果卡运行的良好、可接受和不可接受度量汇总。您可以使用**累积度量趋势**窗格以查看并分析当前月和过去五个月的度量趋势。

在垂直条上移动指针时，数据标签中会显示当月的度量摘要。在窗格中单击垂直条时，资产列表窗格中会显示相关的结果卡。可以查看您拥有读取权限的结果卡。在资产列表窗格中单击结果卡以查看结果卡结果。

下图显示了结果卡仪表板中的**累积度量趋势**窗格：



当结果卡趋势在一个月内在随着时间发生更改时，资产列表窗格可能会在不可接受的指标列表和良好的指标列表中显示几个结果卡。要分析指标，请打开结果卡来查看结果卡结果。

示例

您是加利福尼亚州一系列零售商店的区域经理。您在销售表中创建了 Sales_SC 结果卡，

并在 Sales_SC 结果卡中为 Sales_amt 度量设置了以下阈值：

- 不可接受 = 0% 至 40%
- 可接受 = 41% 至 89%
- 良好 = 90% 至 100%

为了每天捕获销售数据，您使用计划程序服务每晚针对销售表运行结果卡。您编制了五月份的月度管理报告，并且使用结果卡仪表板来验证您的报告。对于五月份，当您单击图表中不可接受的度量和良好度量的垂直条时，Sales_SC 结果卡会显示在资产列表窗格中。

在分析该月份的销售情况时，您发现以下趋势：

1. 从 5 月 1 日至 5 月 25 日，Sales_amt 度量计算值都低于 40%，并且被标记为不可接受的度量。
2. 而在 5 月的最后一周，由于销售量增加，Sales_amt 度量计算值高于 98%，并且被标记为良好度量。

Informatica Analyst 结果卡导出文件

可将结果卡结果导出到 Microsoft Excel 文件。Analyst 工具将以 XLSX 格式保存文件。导出结果卡后，Microsoft Excel 文件将在多个工作表中显示结果卡摘要、趋势图表、无效行数和结果卡属性。

导出结果卡时，可以配置以下选项：

数据

选择以下选项之一，以导出数据：

- 全部。将结果卡摘要、趋势图表、无效行数和结果卡属性导出到 Microsoft Excel 文件。
- 摘要视图。将结果卡摘要、趋势图表和结果卡属性导出到 Microsoft Excel 文件。

无效行数

输入要导出的无效行数。可以在该字段中输入最大为 100,000 的值。默认情况下，该字段显示 100。Analyst 工具最多可以将每个指标 100 个无效行导出到工作表。

如果选择为指标导出超过 100 个无效行，数据集成服务将执行以下步骤以导出剩余的数据：

1. 在 <INFA_HOME>/tomcat/bin/reject 位置为结果卡创建文件夹。数据集成服务将基于结果卡名称和文件创建时间采用 <scorecard_name>_HH_MM_SS 格式命名此文件夹。
例如，如果结果卡名称为 SD1，文件创建时间为 2:23:15，数据集成服务会将该文件夹命名为 SD1_02_23_15。
2. 为结果卡中的每个指标创建一个子文件夹。
例如，如果结果卡 SD1 具有名为 M1、M2 和 M3 的指标，数据集成服务将创建名为 M1、M2 和 M3 的三个子文件夹。
3. 在指标子文件夹中创建 Microsoft Excel 文件。这些文件将基于指标名称和一个增量编号采用 <metric_name>_<IncrementalNumberStartingWith0> 格式命名。创建的最后一个文件采用 <metric_name>_Remaining 格式。每个 Excel 文件可以包含最多 1 万个无效行。
例如，如果指标 M1 具有 3 万个无效行，数据集成服务将创建名为 M1_0、M1_1 和 M1_Remaining 的三个 Microsoft Excel 文件，并在每个文件中保存 1 万个无效行。
4. 数据集成服务将为结果卡中的所有其他指标重复执行步骤 3。

代码页

选择代码页。

从 Informatica Analyst 导出结果卡结果

可将结果卡结果导出到“.xlsx”文件，以便在一个文件中查看数据。

1. 打开结果卡。

2. 单击**操作 > 导出数据**。
此时将显示**将数据导出到文件**对话框。
3. 输入文件名称。或者，使用默认文件名。
默认文件格式为 Microsoft Excel。
4. 选择文件的代码页。
5. 单击**确定**。

Microsoft Excel 格式的结果卡导出结果

在导出结果卡结果时，Analyst 工具会将信息保存至 Microsoft Excel 文件内的多个工作表中。结果卡摘要、趋势图表、无效行和结果卡属性在文件中显示为工作表。Analyst 工具将以“xlsx”格式保存该文件。

下表描述了导出文件中的每个工作表上显示的信息：

选项卡	说明
结果卡摘要	导出的结果卡结果的摘要信息。这些信息包括结果卡名称、每列的总行数、无效的行数、得分和度量权重。
趋势图表	得分的趋势图表。
无效行数	每列的无效行的详细信息。Analyst 工具最多可以将 100 行导出到工作表。 在 将数据导出到文件 对话框中选择 数据 > 全部 选项时，将显示“无效行”工作表。
属性	结果卡属性，如名称、类型、说明和位置。

结果卡通知

配置结果卡通知设置，以使 Analyst 工具能在特定度量得分、度量组得分或度量成本跨越阈值时发送电子邮件。度量得分或度量组得分可能会跨越阈值，也可能会保持在特定的得分范围内，如“不可接受”、“可接受”和“正常”。度量成本值可以超过您设置的成本阈值上限和下限。

可以为单个度量得分、度量组和度量成本配置电子邮件通知。如果将全局设置用于得分，则 Analyst 工具将在特定度量得分从得分范围“正常”跨越阈值到达“可接受”以及从“可接受”跨越阈值到达“错误”时，发送通知电子邮件。如果得分在结果卡连续运行期间保持在“不可接受”得分范围内，也会收到通知电子邮件。如果将全局设置用于度量成本，则 Analyst 工具将在选定度量中无效数据的成本跨越阈值上限和下限时，发送通知电子邮件。

可以自定义通知设置，以使结果卡用户在得分从“不可接受”移至“可接受”以及从“可接受”移至“正常”得分范围时，能够收到电子邮件通知。可以选择如果度量得分或度量成本在每次结果卡运行时都保持在特定范围内，是否发送电子邮件通知。可以在通知设置中根据可以设置成本阈值的度量来查看每个度量的无效数据的当前成本。

配置结果卡以发送电子邮件通知之前，管理员必须在 Administrator 工具中配置电子邮件服务。

通知电子邮件模板

可以设置 Analyst 工具作为结果卡通知的组成部分发送给收件人的电子邮件的邮件文本和结构。电子邮件模板包含可选的介绍性文本部分、只读邮件正文部分，以及可选的结束文本部分。

下表介绍了电子邮件模板中的标记：

标记	说明
ScorecardName	结果卡的名称。
ObjectURL	指向结果卡的超级链接。需要提供用户名和密码。
MetricGroupName	度量所属的度量组的名称。
CurrentWeightedAverage	当前结果卡运行中度量组的加权平均值。
CurrentRange	当前结果卡运行中度量组的得分范围，如“不可接受”、“可接受”、和“正常”。
PreviousWeightedAverage	上一次结果卡运行中度量组的加权平均值。
PreviousRange	上一次结果卡运行中度量组的得分范围，如“不可接受”、“可接受”、和“正常”。
MetricName	度量的名称。
MetricGroupName	度量组的名称。
CurrentScore	最新结果卡运行的得分。
CurrentRange	在最新结果卡运行中，当前得分所处的得分范围。
PreviousScore	上一次结果卡运行的得分。
PreviousRange	上一次结果卡运行的得分范围。
CurrentCost	在最新结果卡运行中，度量中无效数据的成本。
PreviousCost	在上一次结果卡运行中，度量中无效数据的成本。
ColumnName	为度量分配的源列的名称。
ColumnType	源列的类型。
RuleName	规则名称。
RuleType	规则的类型。
DataObjectName	源数据对象的名称。

设置结果卡通知

既可以在度量级别也可以在度量组级别设置结果卡通知。全局通知设置适用于不包含单个通知设置的度量和度量组。

1. 在 Analyst 工具中运行结果卡。
2. 单击**操作 > 编辑**。
3. 单击**通知**选项卡。
4. 选择**启用通知**，以开始配置结果卡通知。
5. 选择某一度量或度量组。
6. 单击**通知**复选框，以便为该度量或度量组启用全局设置。
7. 选择**使用自定义设置**以更改该度量或度量组的设置。

可以选择在得分处于**不可接受**、**可接受**和**正常**范围内以及跨越阈值时发送通知电子邮件。还可以在度量成本跨越阈值上限或下限时发送通知电子邮件。

8. 要编辑结果卡通知的全局设置，请单击**编辑全局设置**图标。
在可以编辑设置（包括电子邮件模板）的位置，将显示**编辑全局设置**对话框。

配置结果卡通知的全局设置

如果选择全局结果卡通知设置，则 Analyst 工具会在得分处于**不可接受**范围内时向目标用户发送电子邮件。还可以将通知设置配置为在度量得分或度量成本超过阈值时发送电子邮件。可以为结果卡配置电子邮件模板，包括电子邮件地址和邮件文本。

1. 在 Analyst 工具中运行结果卡。
2. 单击**操作 > 编辑 > 通知**，以打开**编辑结果卡**对话框。
3. 选择**启用通知**，以开始配置结果卡通知。
4. 单击**编辑全局设置**图标。
在可以编辑设置（包括电子邮件模板）的位置，将显示**编辑全局设置**对话框。
5. 使用**得分**和**得分移动**复选框选择想在何时为度量得分发送电子邮件通知。
6. 使用**成本趋势**复选框选择想在何时为度量成本发送电子邮件通知。
7. 在**电子邮件收件人**字段中，输入收件人的电子邮件 ID。使用分号分隔多个电子邮件 ID。
默认发件人电子邮件 ID 是在域 SMTP 属性中配置的**发件人电子邮件地址**。
8. 输入电子邮件主题的文本。
9. 在**正文**字段中，添加电子邮件的介绍性文本和结束文本。
10. 要应用全局设置，请选择**将设置应用于所有度量和度量组**。
11. 单击**确定**。

结果卡沿袭

结果卡沿袭显示数据的来源、介绍路径，并显示数据如何针对度量或度量组流动。可以使用结果卡沿袭来分析度量或度量组中出现不可接受的得分差异的根本原因。可以在 Analyst 工具中查看结果卡沿袭。

要查看结果卡沿袭，请完成以下任务：

1. 在 Informatica Administrator 中，将 Metadata Manager 服务与分析服务关联起来。
2. 选择某一项目，然后使用 Developer 工具中的“导出资源文件供 Metadata Manager 使用”选项或 infacmd 工具的 exportResources 命令，将该项目中的结果卡对象导出到 XML 文件。
3. 在 Metadata Manager 中，使用导出的 XML 文件创建资源，然后加载该资源。

注意：在 Metadata Manager 中创建并加载的资源文件的名称必须使用以下命名约定：<MRS 名称>_<项目名称>。有关如何创建并加载资源文件的更多信息，请参阅《Informatica PowerCenter Metadata Manager 用户指南》。

4. 在 Analyst 工具中，打开结果卡，然后选择度量或度量组。
5. 查看结果卡沿袭。

在 Informatica Analyst 中查看结果卡沿袭

可以查看度量或度量组的结果卡沿袭图表。首先必须在 Metadata Manager 中加载结果卡沿袭和元数据，然后才能在 Analyst 工具中查看结果卡沿袭图表。

1. 在库工作区中，在资产窗格中单击要查看的结果卡。
结果卡将显示在结果卡工作区中。
2. 在结果卡视图中，选择某一度量或度量组。
3. 右键单击然后选择显示沿袭。

此时将在新窗口中显示结果卡沿袭图表。

重要说明：如果没有使用导出的结果卡对象 XML 文件在 Metadata Manager 中创建并加载资源，则会看到一条错误消息，指出目录中的资源不可用。有关为结果卡沿袭导出 XML 文件的更多信息，请参阅[“为结果卡沿袭导出资源文件”页面上 154](#)。

第 12 章

Informatica Analyst 中的数据域发现

本章包括以下主题：

- [Informatica Analyst 中的数据域发现概览, 94](#)
- [Informatica Analyst 中的数据域词汇表, 94](#)
- [Informatica Analyst 中的数据域发现选项, 96](#)
- [在 Informatica Analyst 中创建自定义配置文件以执行数据域发现, 99](#)
- [在 Informatica Analyst 中编辑列配置文件和数据域发现, 100](#)
- [运行配置文件以执行数据域发现, 100](#)
- [Informatica Analyst 中的数据域发现结果, 101](#)
- [Informatica Analyst 中的数据域发现导出文件, 102](#)

Informatica Analyst 中的数据域发现概览

创建配置文件以执行数据域发现时，可以选择源列、希望匹配列数据和列名称的数据域以及采样选项。您可以为数据域发现选择遵从性条件，并可以在数据域发现过程中排除空值。

您可以创建一个利用采样选项和筛选器来执行数据域发现的配置文件。运行该配置文件时，即会对数据源应用采用选项和筛选器，同时生成一个数据集。数据域发现过程会使用该数据集来发现数据域。

Informatica Analyst 中的数据域词汇表

数据域词汇表列出了数据域和数据域组。可以按数据域或数据域组对列表进行排序。使用数据域词汇表可搜索、添加、编辑和删除数据域以及数据域组。可以在 Developer 工具中查看和更改与数据域关联的规则逻辑。

在 Informatica Analyst 中创建数据域组

数据域组将数据域组织到特定组中，例如个人健康信息 (PHI)、个人身份信息 (PII) 或与项目相关的任何其他概念组。

1. 单击**管理 > 数据域词汇表**。
数据域词汇表在选项卡中打开，列出当前的数据域和数据域组。
2. 在导航器中，单击**操作 > 新建 > 数据域组**。
此时将显示**创建数据域组**对话框。
3. 输入数据域组的名称和说明。
4. 单击**下一步**。
5. 在**可用数据域**窗格中，选择要添加到数据域组的数据域，然后单击**添加**。
Analyst 工具将选定的数据域移至**所选数据域**窗格。
6. 单击**完成**。
Analyst 工具将数据域组添加到数据域词汇表。

在 Informatica Analyst 中创建数据域

可以创建数据域、将数据域添加到数据域词汇表以及将数据域编组到一个或多个数据域组中。要创建数据域，可以使用预定义的数据规则和列名称规则，或者基于列配置文件结果中的值和模式生成数据域。

创建数据域时，Analyst 工具将规则以及与数据域关联的其他相关对象复制到数据域词汇表。要编辑与数据域关联的规则，必须先转至原始规则，然后对该规则进行更改。然后，可以将已修改的规则重新与该数据域关联。

1. 单击**管理 > 数据域词汇表**。
数据域词汇表在选项卡中打开，列出当前的数据域和数据域组。
2. 在导航器中，单击**操作 > 新建 > 数据域**。
此时将显示**创建数据域**对话框。
3. 输入数据域的名称和说明。
4. 单击**数据规则**复选框以根据列数据发现数据域。还可以选中**列名称规则**复选框以根据数据源中的列标题发现数据域。
此时将启用**选择**按钮。
5. 单击**选择**打开**选择规则**对话框。
6. 选择相应的规则，然后单击**确定**。
所选规则将在**数据规则**和**列名称规则**字段中显示。
7. 单击**下一步**。
8. 在**可用数据域组**窗格中，选择要包含数据域的数据域组，然后单击**添加**。
Analyst 工具将选定的数据域组移至**所选数据域组**窗格。
9. 单击**完成**。
Analyst 工具将数据域添加到数据域词汇表。

在 Informatica Analyst 中基于配置文件结果创建数据域

运行列配置文件可查看源数据的值和模式。之后您可以验证配置文件结果并基于这些结果创建一个数据域。

1. 运行列配置文件以查看其结果。
配置文件结果将显示在摘要视图中。
2. 在摘要视图中，单击一个列以在详细视图中查看列结果。
3. 在**值**窗格或**模式**窗格中，选择创建数据域时所依据的值或模式。
4. 右键单击该值或模式，然后选择**创建数据域**。
此时将显示**创建数据域**对话框。
5. 输入数据域名称和可选说明。
6. 单击**创建**。
数据域便会添加到数据域词汇表中。

在 Informatica Analyst 中查找数据域和数据域组

可以在数据域词汇表中搜索特定的数据域和数据域组。可以在**数据域**视图与**数据域组**视图之间进行选择，以查看数据域词汇表中的数据域列表。

例如，您的数据域 **Zipcode** 可能已添加到 **PII** 数据域组。可以通过以下方式查找与 Zipcode 及其数据域组 PII 有关的更多信息：

搜索数据域。

在导航器顶部的文本字段中键入数据域名称的一部分，例如 **zip** 或 **code**。如果您在**数据域组**视图中，Analyst 工具将列出 **PII**，这是包含 **Zipcode** 的数据域组。如果您在**数据域**视图中，Analyst 工具将列出包含搜索字符串 **Zip** 或 **code** 的所有数据域。

注意：搜索不区分大小写。

查看所有数据域组以及其中包含的数据域。

在导航器中，单击**显示数据域组视图**。

查看所有数据域。

在导航器中，单击**显示数据域视图**。

查看数据域的属性。

确认您在**显示数据域视图**中。在导航器中，单击 **Zipcode** 以在右侧窗格中查看其属性。可以查看名称、类型、说明、关联规则及其所属的域组，在这种情况下为 **PII**。

查看数据域组的属性。

确认您在**显示数据域组视图**中。在导航器中，单击 **PII** 以在右侧窗格中查看其属性。可以在 **PII** 中查看名称、说明和数据域的列表，包括 **Zipcode**。

刷新数据域词汇表。

在导航器中，单击**操作** > **刷新**。数据域词汇表将根据您所在的视图显示数据域或数据域组的当前列表。

Informatica Analyst 中的数据域发现选项

使用数据域发现选项可为数据域发现选择列、数据域和推理选项。推理选项包括选择是否要根据列数据、列名称或两者应用的规则来运行数据域发现。

Informatica Analyst 中的数据域列选择

可以通过单击**指定设置**屏幕中的**编辑**来选择要作为数据域发现的一部分运行的列。可以在配置文件向导的**选择源**屏幕中查看数据源中的所有列。可以为列配置文件和数据域发现选择不同的列。

下表介绍了数据域发现的**编辑**对话框属性：

选项	说明
名称	显示列名称。
类型	显示记录的列数据类型。
精度	显示列的最大精度。
小数位数	显示列的小数位数。
可空	指示列可以包含空值。
键	指示列是记录为主键还是外键。

Informatica Analyst 中的数据域选择

指定设置屏幕中的**数据域**窗格会列出数据域词汇表中的所有数据域。可以选择要作为数据域发现的一部分运行的数据域。

下表介绍了用于数据域发现的**数据域**属性：

选项	说明
名称	显示数据域名称。可以选择一个或多个数据域或数据域组。
说明	显示数据域的说明。
DomainGroups	显示数据域所属的数据域组的名称。

Informatica Analyst 中的数据域推理选项

推理选项决定了数据域发现是必须在列数据、列名称还是两者上运行。您可以指定配置文件可以分析的最高源行数。您可以为数据域发现选择遵从性条件。您可以从数据域发现中排除空值。可以在配置文件向导的**指定设置**屏幕中设置数据域推理选项。

下表介绍了数据域发现的推理选项：

选项	说明
数据	对列数据运行配置文件。
列	对列标题运行配置文件。
数据和列	同时对列数据和列标题运行配置文件。
最低行数百分比	数据域匹配所需的数据集中最低遵从行数百分比。

选项	说明
最小行数	数据域匹配所需的数据集中最低遵从行数。
排除数据域发现期间的空值	从数据域发现的数据集中排除空值。
编辑	选择用于数据域发现的列。
所有行	对源中的所有行运行配置文件。
先采样	选择可运行配置文件的最大行数。Analyst 工具将从源中的第一行开始选择行。最多可以选择 2,147,483,647 行。
随机采样	从数据源中选择随机采样的行。最多可以选择 2,147,483,647 行。
随机采样(自动)	Analyst 工具将基于源数据的大小选择随机采样的行。
在后续运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型和数据域	在下次运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型或数据域。

最低遵从性百分比

您可以选择数据集中的行数最低百分比作为数据域发现的遵从性条件。

遵从性百分比是指匹配的行数除以总行数所得出的比率。

注意: Analyst 工具将空值视为非匹配行。包含大量空值的列可能不会导致执行数据域推理，除非您为最低遵从性百分比指定了值。

示例

您的一个数据源中含有 10,000 行，其中注释列在 2,500 行中含有社保号码。您创建了一个列配置文件和数据域发现，并将最低行数百分比设置为 30% 以作为遵从性条件。运行该配置文件时，配置文件结果不会将社保号码显示为推理的数据域，因为最低遵从性条件为 30% 行数，即数据源中要有 3,000 行。

最低遵从行数

您可以选择数据集中的最低行数作为数据域发现的遵从性条件。

示例

您的一个数据源中含有 10,000 行，其中注释列在 3 行中含有电子邮件地址。您创建了一个列配置文件和数据域发现，并将最低行数设置为 1 以作为遵从性条件。运行该配置文件时，配置文件结果会将电子邮件地址显示为推理的数据域，并显示三个遵从行以及其他推理的数据域。

在数据域发现过程中排除空值

对数据源执行数据域发现时，您可以排除空值。选择最低行数百分比和排除空值选项时，遵从性百分比的计算方式为：将匹配行数除以总行数与列中空值数之差。

选择**从数据域发现中排除空值**选项和多个采样选项或筛选器时，数据域发现过程有所不同。

以下场景介绍了选择排除空值选项与采样选项和筛选器时的数据域发现结果：

- 选择**所有行**作为采样选项，且未使用筛选器。数据域发现过程会忽略列中的所有空值。

- 选择一个采样选项，且未使用筛选器。数据域发现过程会忽略采样数据中的所有空值，并对其余采样数据运行。
- 选择**所有行**作为采样选项，且使用筛选器。数据域发现过程会忽略筛选后数据中的所有空值，并对其余筛选后数据运行。
- 选择一个采样选项，且使用筛选器。数据域发现过程会忽略采样的筛选后数据中的所有空值，并对其余筛选后数据运行。

示例

您的一个数据源中含有 10,000 行，其中 3,000 行的“Comments”列中含有社保号码。您创建了一个列配置文件和数据域发现，并选择了以下选项：

- 选择**从数据域发现中排除空值**选项。
- 选择**所有行**作为采样选项。
- 选择**最低行数百分比**选项并将该选项配置为 12%。

运行该配置文件时，该配置文件会在数据集上运行，并在数据域发现过程中忽略空值。

在 Informatica Analyst 中创建自定义配置文件以执行数据域发现

至少必须先创建一个数据域，然后才能在 Analyst 工具中创建列配置文件以执行数据域发现。配置文件可以同时发现匹配预定义数据域的列名称和列数据。

1. 在**发现工作区**中，单击**配置文件**，或从 Analyst 工具中的任意位置选择**新建 > 配置文件**。
此时将显示**新建配置文件**向导。
2. 默认情况下将选择**单源**选项。单击**下一步**。
3. 在**指定常规属性**屏幕中，输入配置文件的名称和可选说明。在“位置”字段中，选择要在其中创建配置文件的项目或文件夹。单击**下一步**。
4. 在**选择源**屏幕中，单击**选择**以选择数据对象，或单击**新建**以导入数据对象。单击**下一步**。
5. 在**指定设置**屏幕中，选择要运行列配置文件还是数据域发现或同时运行列配置文件和数据域发现。默认情况下将选择列配置文件选项。
 - 选择**运行数据域发现**可执行数据域发现。在**数据域**窗格中选择数据域选项。
 - 选择**运行列配置文件和运行数据域发现**可同时运行列配置文件和数据域发现。在**数据域**窗格中选择数据域选项。

注意：默认情况下，选择用于列配置文件的列也适用于数据域发现。无论选择了哪些列用于列配置文件，您都可以单击**编辑**以选择或取消选择用于数据域发现的列。

 - 选择“数据”、“列”或“数据和列”以对其运行数据域发现。
 - 在**运行配置文件**窗格中选择采样选项。
 - 在**向下钻取**窗格中选择向下钻取选项。或者，单击**选择列**，以选择要进行向下钻取的列。您可以选择跳过对数据类型或数据域已经过批准的列进行数据类型和数据域推理。
 - 选择遵从性条件，然后可以选择**从数据域发现中排除空值**选项。
 - 选择**本地**或**Hadoop**作为运行时环境。在 Hadoop 运行时环境中可以选择 Blaze 或 Spark 选项。如果选择 Blaze 选项，请单击**选择**以在**选择 Hadoop 连接**对话框中选择 Hadoop 连接。如果选择 Spark 选项，请单击**选择**以在**选择 Hadoop 连接**对话框中选择 Hadoop 连接。

6. 在**指定规则和筛选器**屏幕中，可以为配置文件添加、编辑或删除规则和筛选器。
7. 单击**保存并完成**以创建配置文件，或单击**保存并运行**以创建并运行配置文件。

在 Informatica Analyst 中编辑列配置文件和数据域发现

运行后，可以更改配置文件的属性。如果随数据域发现运行列配置文件，可以更改列配置文件设置。

1. 在**库**工作区中，选择包含配置文件的项目，或者在**资产**窗格中选择配置文件。
2. 单击配置文件名称。
摘要视图将显示在**发现**工作区中。
3. 如果启用了版本控制系统，请单击**操作 > 签出**以签出配置文件。
4. 单击**操作 > 编辑配置文件**。
此时将显示**配置文件**向导。
5. 根据要进行的更改，选择以下页面选项之一：
 - **指定常规属性**。更改基本属性，例如名称、说明和位置。
 - **选择源**。选择其他匹配的数据源和列以对其运行配置文件。
 - **指定设置**。选择要运行列配置文件还是同时运行列配置文件和数据域发现。编辑数据域选项、采样选项和向下钻取选项。
 - **指定规则和筛选器**。创建、编辑或删除规则和筛选器。
6. 单击**保存并完成**以编辑配置文件，或单击**保存并运行**以编辑并运行配置文件。
7. 如果启用了版本控制系统，必须执行以下任务：
 - 单击**保存并完成**以完成配置文件编辑。
 - 在摘要视图中，单击**签入**以签入配置文件。
 - 单击**操作 > 运行配置文件**，以运行该配置文件。

运行配置文件以执行数据域发现

作为数据域发现的一部分运行配置文件可查看匹配数据域规则模式的列。

1. 在**库导航器**中，在“项目”窗格中选择包含配置文件的项目或文件夹，或者在“资产”窗格中选择配置文件。
2. 单击**操作 > 打开**。
摘要视图将显示在**发现**工作区中。
3. 单击**操作 > 运行配置文件**。
Analyst 工具将执行配置文件运行，并在摘要视图中显示配置文件结果。
4. 在摘要视图中，单击一个列以查看列结果。
此时将显示详细视图。

Informatica Analyst 中的数据域发现结果

可以在摘要视图和详细视图中查看数据域发现结果。

数据域字段显示有关与数据域匹配的列的统计信息。在摘要视图中，您可以查看推理的数据域，以及遵从行百分比和遵从行数。

在详细视图中，可以执行以下任务：

- 在水平条形图中，查看推理的数据域，以及遵从行百分比和遵从行数。
- 向下钻取遵从行数、不遵从行数和空值的结果。
- 批准、拒绝或重置数据域。
- 显示或隐藏拒绝的数据域。
- 对数据源中的所有行运行数据域发现，以发现推理的数据域。

批准数据域

可以在 Analyst 工具中批准多个数据域。

1. 在库工作区中，选择包含配置文件的项目或文件夹。
2. 单击配置文件将其打开。
配置文件结果将显示在摘要视图中。
3. 单击要批准数据域的列。
列结果将显示在详细视图中。
4. 在详细视图的**数据域**窗格中，选择数据域。单击**操作 > 批准**。
列或数据域的状态更改为“已批准”。
5. 要还原列或数据域的已推理状态，请选择该数据域并单击**操作 > 重置**。

拒绝数据域

打开配置文件结果时，Analyst 工具默认会显示已批准的数据域。可以显示或隐藏已拒绝的数据域。

1. 在库导航器中，选择包含配置文件的项目或文件夹。
2. 单击配置文件将其打开。
配置文件结果将显示在摘要视图中。
3. 单击要拒绝数据域的列。
列结果将显示在详细视图中。
4. 要拒绝已推理的数据域，请单击**操作 > 拒绝**。
Analyst 工具将从数据域发现结果中删除已拒绝的数据域。
5. 要查看已拒绝的数据域，请单击**操作 > 显示拒绝项**。
6. 要隐藏已拒绝的数据域，请单击**操作 > 隐藏拒绝项**。

Informatica Analyst 中的数据域发现导出文件

从 Analyst 工具中导出数据域发现结果时，可以指定文件名和代码页值。可以将数据域发现结果导出到 Microsoft Excel 文件。

Microsoft Excel 文件包含多个根据列、数据域和数据域组分隔发现结果的工作表。“属性”工作表显示配置文件属性，例如名称、说明、类型、位置、上次更改配置文件的日期和时间以及指向配置文件的链接。

Microsoft Excel 中的数据域发现结果

将数据域发现结果导出到 Microsoft Excel 时，Analyst 工具将保存列名称、匹配数据域的名称、遵从性条件和空值。Excel 文件还包含每个数据域的数据域组名称以及记录的列数据类型。

下表介绍了导出文件中的每个工作表：

选项卡	说明
按列查看	按数据源列排序的数据域发现结果。
按数据域查看	按数据域排序的数据域发现结果。
按数据域组查看	按数据域组排序的数据域发现结果。
属性	基本配置文件属性，例如名称、说明、类型、位置、上次更改配置文件的日期和时间以及指向配置文件的链接。

从 Informatica Analyst 中导出数据域发现结果

可以将数据域发现结果导出到 .xlsx 文件，以便能够查看文件中的数据并在企业内部分发以供将来使用。

1. 运行配置文件以执行数据域发现。
2. 在摘要视图或详细视图中，单击**操作 > 导出数据**。
此时会显示**将数据导出到文件**对话框。
3. 输入文件名称。或者，使用默认文件名。
4. 选择文件的代码页。
5. 单击**确定**。

第 13 章

Informatica Analyst 中的企业发现

本章包括以下主题：

- [Informatica Analyst 中的企业发现概览, 103](#)
- [Informatica Analyst 中的企业发现过程, 103](#)
- [企业发现的配置选项, 104](#)
- [在 Informatica Analyst 中创建企业发现配置文件, 105](#)
- [编辑企业发现选项, 106](#)

Informatica Analyst 中的企业发现概览

企业发现是从大量架构和外部关系连接中的多个数据源发现列元数据和数据域的过程。对导入到模型存储库中的数据源以及来自外部关系连接的数据源均可以执行企业发现。

作为数据分析人员，您可以在 Analyst 工具中执行企业发现，以便从大量数据源中推理特定元数据特性。您可能还想查看与预定义的数据域匹配的源数据。这样就可以管理推理的企业发现结果，并为发现搜索和数据质量计划准备好数据。Analyst 工具中的企业发现生成配置文件结果的合并结果摘要。

企业发现结果包括列配置文件统计信息，如模式、唯一值和数据类型冲突的列。数据域发现识别与预定义的数据域相匹配的源列。

您可以在 Informatica Analyst 中选择操作系统配置文件。选择一个操作系统配置文件后，数据集成服务即会根据在操作系统配置文件中定义的操作系统用户的权限创建并运行企业发现配置文件。

Informatica Analyst 中的企业发现过程

可以创建、编辑和删除企业发现配置文件。可以在发现工作区中运行企业发现配置文件。在运行企业发现配置文件之前，需要配置列配置文件和数据域发现的推理选项。

完成以下步骤，以便在 Analyst 工具中执行企业发现：

1. 配置企业发现配置文件的常规属性。
2. 从要包含在企业发现配置文件中的模型存储库选择数据对象。
3. 从外部数据连接导入关系数据源。
4. 为企业发现配置文件配置数据推理选项和发现选项。
5. 保存更改，然后运行企业发现配置文件。

- 6. 监控配置文件运行，根据需要查看 Analyst 工具运行的配置文件任务的状态。
- 7. 查看企业发现结果摘要。结果显示在摘要和配置文件面板中。

企业发现的配置选项

企业发现的配置选项包括数据域发现选项、列配置文件采样选项和常规配置文件属性，如名称和说明。

可以选择运行列配置文件或配置文件，以执行数据域发现。还可以选择同时运行列配置文件和配置文件，以作为配置的一部分执行数据域发现。

数据域发现设置

数据域发现设置包括选择是否必须对列数据、列名称或者列数据和列名称这两者运行数据域发现。可以选择数据域并指定数据域发现是否需要处理数据源中的所有行。您可以为数据域发现选择遵从性条件。您可以排除数据域发现中的空值。

下表描述了可以在 Analyst 工具中为企业发现配置的数据域发现设置：

选项	说明
启用数据域发现	将数据域发现作为企业发现的一部分执行。
对数据运行数据域发现	对列数据执行数据域发现。
对列名称运行数据域发现	对每个列的名称执行数据域发现。
最低遵从性百分比	数据域匹配所需的数据集中最低遵从行数百分比。遵从性百分比是指匹配的行数除以总行数所得出的比率。 注意: Analyst 工具将空值视为非匹配行。
最低遵从行数	数据域匹配所需的数据集中最低遵从行数。
从数据域发现中排除空值	从数据域发现的数据集中排除空值。
排除包含已批准数据域的列	从配置文件运行的数据域推理中排除包含已批准数据域的列。
所有行	对所有源行执行数据域发现。
第一行	配置文件可以在其中运行的最大行数。Analyst 工具从源中的第一行开始选择行。最多可以选择 2,147,483,647 行。

列配置文件设置

采样选项决定 Analyst 工具对数据源的所有行还是对有限数量的行运行列配置文件。

下表描述了可以为企业发现配置文件配置的列配置文件设置：

选项	说明
启用列剖析	将列配置文件作为企业发现的一部分运行。
在后续运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型和数据域	在下次运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型或数据域。

下表描述了可以为企业发现配置文件配置的运行时环境选项：

选项	说明
Native	分析工具会将配置文件作业提交给剖析服务模块。剖析服务模块随后将配置文件作业拆分成一组映射。数据集成服务会运行这些映射，并将配置文件结果写入剖析仓库。
Blaze	数据集成服务将配置文件逻辑推送到 Hadoop 群集上的 Blaze 引擎来运行配置文件。
Spark	数据集成服务将配置文件逻辑推送到 Hadoop 群集上的 Spark 引擎来运行配置文件。

下表描述了可以为企业发现配置文件配置的采样选项：

选项	说明
所有行	对数据源中的所有行运行列配置文件。 本地、Blaze 和 Spark 运行时环境中支持此选项。
前 <number> 行	对从数据对象第一行开始的采样行运行配置文件。最多可以选择 2,147,483,647 行。 本地和 Blaze 运行时环境中支持此选项。
限制 n <数字> 行	根据数据对象中的行数运行配置文件。选择在 Hadoop 验证环境中运行配置文件时，Spark 引擎会从数据对象的多个分区收集样本并将这些样本推送到单个节点来计算采样大小。“限制 n”采样选项支持 Oracle、SQL Server 和 DB2 数据库。不能对“限制 n”采样选项使用高级筛选器。最多可以选择 2,147,483,647 行。 Spark 运行时环境中支持此选项。
随机百分比	对数据对象中某一百分比的行运行配置文件。 Spark 运行时环境中支持此选项。

在 Informatica Analyst 中创建企业发现配置文件

可以在 Informatica Analyst 中将列配置文件和数据域发现作为企业发现的一部分运行。

1. 在发现工作区中，选择新建 > 配置文件。

此时将显示新建配置文件向导。

2. 选择**企业发现**。单击**下一步**。
此时将显示**指定常规属性**选项卡。
3. 在**指定常规属性**选项卡中，输入企业发现配置文件的名称和可选说明。在“位置”字段中，选择要在其中创建配置文件的项目或文件夹。单击**下一步**。
此时将显示**选择数据对象**选项卡。
4. 在**选择数据对象**选项卡中，单击**选择**。
此时将显示**选择数据对象**对话框。
5. 在**选择数据对象**对话框中，选择一个或多个要添加到配置文件的数据对象。单击**保存**。
数据对象将显示在**数据对象**窗格中。
6. 单击**下一步**。
此时将显示**选择资源**选项卡。
7. 在**选择资源**选项卡中，单击**选择**打开**选择资源**选项卡。
可以从多个关系数据源导入数据。
8. 在**选择资源**选项卡中，选择要包含在配置文件中的连接、架构、表和视图。单击**保存**。
对话框左侧的窗格中 Informatica 域的下面列出所有内部和外部连接、架构、表和视图。
这些资源将显示在**资源**窗格中。
9. 单击**下一步**。
此时将显示**指定设置**选项卡。
10. 在**指定设置**选项卡中，可以配置列配置文件选项和数据域发现选项。单击**保存并完成**以保存企业发现配置文件，或单击**保存并运行**以运行配置文件。
可以在**指定设置**选项卡中执行以下任务。
 - 启用数据域发现。单击**选择**以选择要在**选择数据域**对话框中发现的数据域。选定的数据域将显示在 **用于数据域发现的数据域**窗格中。
 - 在数据或列名或同时在数据和列名上运行数据域。
 - 选择数据源中的所有行，也可以选择要运行域发现的最大行数。
 - 为数据域发现选择最低遵从性百分比或指定最小遵从行数。
 - 启用列配置文件设置，然后在列配置文件的数据源中选择所有行或前几行。可以在列配置文件中排除针对包含已批准数据类型的列的数据类型推理。
 - 选择**本地**、**Blaze** 或 **Spark** 作为运行时环境。选择 **Blaze** 或 **Spark** 后，选择 Hadoop 连接来运行配置文件。可以在**摘要**和**配置文件**选项卡下查看企业发现结果。

编辑企业发现选项

执行企业发现后，可以更改企业发现选项。可以重新命名配置文件并更改数据对象选择、数据域选择和推理选项。

1. 打开运行过的配置文件以执行企业发现。
配置文件结果显示在**发现工作区**。
2. 如果启用了版本控制系统，请单击**操作 > 签出**以签出配置文件。

3. 单击**编辑配置文件**。
4. 在**指定常规属性**选项卡中，根据需要更新配置文件属性。
5. 要更改数据对象选择，请单击**选择数据对象**选项卡。
6. 要更改企业发现的外部数据源，请单击**选择资源**选项卡。
7. 要更改数据域推理选项和列配置文件设置，请单击**指定设置**选项卡。
8. 要对企业发现配置文件中的所有数据域配置文件任务和列配置文件任务应用配置更改，请选择**对所有配置文件使用全局设置**。如果不选择该选项，所进行的配置文件设置更改仅应用于新添加到配置文件中的数据对象或资源。

默认情况下，所做的更改将应用于企业发现配置文件中新添加的数据对象。
9. 要撤消更改，请单击**取消**。
10. 单击**保存并运行**保存更改并再次运行配置文件。
11. 如果启用了版本控制系统，必须执行以下任务：
 - 单击**保存并完成**以完成配置文件编辑。
 - 在摘要视图中，单击**签入**以签入配置文件。
 - 单击**操作 > 运行配置文件**，以运行该配置文件。

第 14 章

Informatica Analyst 中的企业发现结果

本章包括以下主题：

- [Informatica Analyst 中的企业发现结果概览, 108](#)
- [摘要视图, 109](#)
- [数据类型冲突, 111](#)
- [配置文件视图, 111](#)

Informatica Analyst 中的企业发现结果概览

可以在**摘要**和**配置文件**视图中查看企业发现的结果。

摘要视图显示列配置文件结果和数据域发现结果。**数据域发现**部分列出包括在配置文件运行中的数据域和具有数据域匹配项的列数。**列剖析**部分显示资源列的统计信息。可以单击每个配置文件结果行，以在**摘要**视图右侧窗格中查看详细信息。

摘要视图

摘要视图显示列配置文件结果和数据域发现结果的摘要。 可以查看列中有匹配项的数据域名称以及具有数据域匹配的列数。 列统计信息包括前 10 个模式匹配的列数、全部唯一值和全部空值。 列统计信息还包括推理的数据类型与记录的数据类型之间有数据类型冲突的列数。

摘要视图配置文件结果

“摘要”视图在“数据域发现”和“列剖析”部分显示企业发现结果。

数据域发现

下表描述了数据域发现结果中的列：

列名称	说明
名称	数据域的名称。
在列中发现	具有数据域匹配项的总列数。
配置文件	包含匹配列的配置文件的名称。
列名称	匹配列的名称。
数据遵从性百分比	数据域匹配所需的最低遵从行数百分比。
连接名称	关系数据库连接的名称。
源名称	数据源的名称。
推理状态	数据域推理状态。状态包括 已接受 、 已拒绝 和 已推理 。
空值百分比	列的空值的百分比。
总行数	总行数。
确认行数	数据域匹配所需的最低行数。
列名称匹配	用于指示列名称是否与数据域名称相匹配。
已记录的数据类型	声明用于配置文件对象中的列的数据类型。
已验证	用于指示数据域匹配在数据源的所有行上是否已通过验证。
上次运行时间	上次运行配置文件的日期和时间。

列剖析

下表描述了列配置文件结果中的列：

列名称	说明
名称	配置文件结果类型的名称，如模式、全部空值和全部唯一。
在列中发现	具有匹配配置文件结果类型的列的总数。
配置文件	包含匹配列的配置文件的名称。
连接	关系数据库连接的名称。
数据源	配置文件的数据源。
列数	配置文件中具有匹配配置文件结果类型的列数。

查看数据域发现结果

可以单击数据域名称以查看其数据域发现结果。可以从数据域发现结果打开特定配置文件。

1. 运行配置文件以执行企业发现。
2. 验证是否在**摘要**视图中。
3. 单击**数据域发现**部分下面的数据域，以查看其发现结果。
将在右侧窗格中显示包含该数据域的配置文件的列表。
4. 如果需要，在右侧窗格中选择行。
指向配置文件的超链接以蓝色显示。
5. 单击配置文件名称链接或者列名称链接，以打开配置文件。
配置文件将打开，并显示数据域发现结果。Analyst 工具突出显示在结果中包含数据域的行。如果需要，可以管理配置文件结果以便更有效地使用，如发现搜索。
6. 要返回**摘要**视图，请单击**返回企业发现**。

查看列配置文件结果

可以在**摘要**视图中查看企业发现的列配置文件结果。可以从数据域发现结果打开特定配置文件。

1. 运行配置文件以执行企业发现。
2. 验证是否在**摘要**视图中。
3. 要查看推理的模式的信息，可单击**列剖析**部分下面的前 10 个模式之一。
右侧窗格中显示包含推理的模式结果的配置文件列表。
4. 要查看诸如全部空值、全部唯一值或数据类型冲突等信息，请单击**全部空值**、**全部唯一**或**已推理与已记录的数据类型冲突**。
配置文件的匹配列表显示在右侧窗格。
5. 单击配置文件名称链接或者列名称链接，以打开配置文件。
配置文件将打开，并显示列配置文件结果。
6. 要返回**摘要**视图，请单击**返回企业发现**。

数据类型冲突

企业发现可识别列中的数据类型冲突。数据类型冲突是指在运行企业发现之后，推理的列数据类型与记录的列数据类型之间存在不匹配情况。推理的数据类型是 Analyst 工具基于列数据为数据源列派生的数据类型。记录的数据类型是为源数据库中的列声明的数据类型。

企业发现可能会根据列数据与列的记录数据类型的比较结果为列推理一个不同的数据类型。例如，企业发现可以将数据类型为记录的字符串的列推理为日期数据类型。可以查看数据类型冲突，为列选择最合适的日期数据类型，然后批准。

查看数据类型冲突

从**摘要**视图打开数据类型有冲突的配置文件时，Analyst 工具会以红色突出显示数据类型冲突。

1. 运行配置文件以执行企业发现。
2. 验证是否在**摘要**视图中。
3. 在**列剖析**部分下面，单击**已推理与已记录的数据类型冲突**以查看列配置文件结果中的数据类型冲突。
将在右侧窗格中显示数据类型有冲突的列属于配置文件的列表。
4. 如果需要，在右侧窗格中选择行。
指向配置文件的超链接以蓝色显示。
5. 单击配置文件名称链接或者列名称链接，以打开配置文件。
此时配置文件将打开，并以红色显示数据类型冲突。可以选择管理推理的数据类型，以解决数据类型冲突。
6. 要管理数据类型，请选择数据类型冲突的行，并单击**数据类型**视图。
7. 单击**操作**，然后选择**批准**或**拒绝**。
8. 要返回**摘要**视图，请单击**返回企业发现**。

配置文件视图

配置文件视图显示 Analyst 工具作为企业发现的一部分运行的全部单个数据对象配置文件的列表。该配置文件列表还显示每个配置文件的运行状态。可以打开每个配置文件，以查看列配置文件结果和数据域发现结果。

查看配置文件属性

可以在**配置文件**视图中查看属于企业发现的组成部分的配置文件的列表。如果需要，可以打开每个配置文件并管理配置文件结果。

1. 运行配置文件以执行企业发现。
2. 验证是否在**配置文件**视图中。
3. 要查看配置文件的配置文件属性，可单击配置文件名称。
配置文件属性显示在右侧窗格中。配置文件属性包括源数据对象的名称、连接名称和行计数。
4. 要查看配置文件结果，可单击**打开配置文件**。
该配置文件显示列配置文件结果。
5. 要返回到**配置文件**视图，可单击**发现**工作区左上角的文件夹或项目名称链接。

第 15 章

Informatica Analyst 中的发现搜索

本章包括以下主题：

- [在 Informatica Analyst 中执行发现搜索概览, 112](#)
- [发现搜索必备条件, 113](#)
- [Informatica Analyst 中的发现搜索过程, 113](#)
- [发现搜索选项, 113](#)
- [Informatica Analyst 中的发现搜索结果, 114](#)
- [匹配类型, 116](#)
- [相关资产, 117](#)
- [常见问题, 118](#)

在 Informatica Analyst 中执行发现搜索概览

发现搜索查找资产并识别与数据库和企业架构中的其他资产之间的关系。企业用户可以使用发现搜索查找数据和元数据在企业中存在的位置。您可以搜索特定资产，例如数据对象、规则和配置文件。

如果执行全局搜索，Analyst 工具将对数据对象、数据源和文件夹执行基于文本的搜索。如果执行发现搜索，则除文本匹配项外，搜索结果中还包含与匹配搜索条件的对象有关系的对象。发现搜索还包括基于配置文件元数据的匹配项，例如数据类型和数据模式。例如，您可以查找包含特定数据模式且名称中包含特定关键字的对象。

发现搜索的搜索结果中包含以下类型的信息：

模型存储库中的对象

查找与匹配发现搜索条件的对象相关的主要对象。例如，当您搜索配置文件时，配置文件结果中将包含配置文件的数据对象。

配置文件仓库结果

包含基于配置文件的推理结果，例如数据域或数据模式。

Business Glossary 术语

包含搜索中的元数据，例如与规则关联的业务术语，具体取决于许可证。

发现搜索示例

您是企业内的一名数据管理者，负责确保妥善屏蔽敏感企业数据。您可能希望识别您或数据架构师已对其执行企业发现的不同架构和数据库中的个人身份信息 (PII)。您可能已创建数据域以识别数据源中仍未被发现的重要数据。您对“SSN”字符串执行搜索。Analyst 工具将显示社会保障数据域以及数据源中的所有匹配列。此外，发现搜索字符串可能还查找说明或名称中包含“SSN”的其他列或表。要缩小搜索范围，可以对映射规范执行筛选，以显示引用匹配数据对象的映射规范。可以根据项目或用户应用其他筛选器以筛选其他映射规范。您可能还希望打开结果中的映射规范，以确认映射规范是否符合企业的隐私政策。

发现搜索必备条件

能够在企业中的不同数据库中执行有效发现搜索之前，请对企业中的数据库和架构执行企业发现。

执行企业发现后，Analyst 工具将在剖析仓库中存储所有配置文件结果。请确认所需的所有数据源都在模型存储库中。或者，请确认模型存储库中的相应资产具有与资产关联的业务术语。执行发现搜索时，搜索服务将根据模型存储库资产和剖析仓库结果检索搜索索引信息。搜索服务随后使用编制了索引的信息根据恰当的对象元数据和关系显示搜索结果。

Informatica Analyst 中的发现搜索过程

可以根据文本、模式和数据类型等条件在配置文件结果中搜索资产。搜索将返回与搜索字符串相关的资产列表。

要在 Analyst 工具中执行发现搜索，请完成以下步骤：

1. 对企业中的数据源执行企业发现并运行所需的单个数据对象配置文件。执行发现搜索时，Analyst 工具将搜索配置文件结果和模型存储库对象中的信息。
2. 选择要查找的信息类型。例如，您可能希望查找与敏感数据或特定数据模式的数据域定义关联的所有资产。
3. 执行搜索。
4. 分析搜索结果以识别资产以及这些资产与其他资产的关系。
5. 如果需要，请验证发现的数据是否满足业务要求。

发现搜索选项

可以执行全局搜索或发现搜索以查找资产并识别这些资产与其他资产的关系。全局搜索从模型存储库和 Business Glossary（可选）中检索结果。除基于剖析仓库中的配置文件结果的配置文件外，发现搜索还从模型存储库和 Business Glossary 中检索结果。

可以搜索资产，例如数据对象、配置文件和映射规范。请输入搜索字符串以搜索匹配该搜索字符串且与该搜索字符串关联的资产。搜索资产时，可以使用通配符。

搜索资产时，可以使用以下通配符：

*（星号）

添加到搜索字符串的结尾以查找开头为该字符串的所有资产名称。例如，要搜索开头为字符串“emp”的所有资产名称，可以在搜索字段中键入“emp*”。

?（问号）

包含在搜索字符串中以表示字母数字字符。

注意：搜索资产时，不能将通配符用作搜索字符串的开头。搜索不区分大小写。

要作为一个短语一起搜索两个或多个单词，请将这些单词包含在双引号中。使用字符 + 表示 AND 运算符并搜索必须在搜索结果中显示的术语。例如，如果搜索字符串为 +sensitive +data，搜索服务将查找包含这两个术语的元数据。使用空格表示 OR 运算符。例如，如果搜索字符串为 sensitive data，搜索服务将查找包含其中一个术语的元数据。

如果搜索字符串中包含连字符 (-)、下划线 (_) 或混合大小写字符，搜索服务将查找整个单词以及该分隔符分隔的部分单词。例如，如果搜索 Profile_Customer，搜索引擎将在存储库中查找 Profile、Customer 和

Profile_Customer。要在搜索字符串中包含特殊字符（例如 * 和 ?），请用双引号引起包含特殊字符的搜索字符串。

可以执行包含关键字搜索和发现筛选的发现搜索。例如，您可能希望查找使用格式 <FirstNameInitial><LastNameInitial>-<SSN> 的员工 ID 列，以便能够识别数据安全风险。要搜索员工 ID 列，请在“库”工作区的搜索面板中输入 Employee ID，并将模式筛选器设置为 XX-999999999 <= 100%。

发现搜索条件

使用发现搜索条件可根据条件搜索信息，例如模式、数据类型、唯一值和空值。可以在搜索中使用条件运算符 =、>= 或 <=。

下表介绍了可用于发现搜索的发现搜索条件：

选项	说明
搜索	要搜索的文本表达式。
清除	清除搜索字符串以及之前选择的所有其他搜索条件。
模式	要包含在搜索中的列模式和百分比。 注意: 此选项不接受模式中的控制字符。
数据类型	要包含在搜索中的列数据类型和百分比。
唯一值	要包含在搜索中的列的唯一值的百分比。
空值	要包含在搜索中的列的空值的百分比。

搜索资产

可以搜索库工作区中的资产。搜索结果中包含同时在 Developer 工具和 Analyst 工具中创建的资产。

1. 打开库工作区。
2. 验证您是否位于发现搜索部分中。
3. 在搜索字段中，键入要搜索的搜索字符串。
4. 配置搜索筛选器以缩小搜索范围。
筛选器包括模式、数据类型、唯一值和空值。
5. 单击搜索图标。

Informatica Analyst 中的发现搜索结果

发现搜索查找发现搜索的所有许可存储库中的资产，例如模型存储库和剖析仓库。

发现搜索结果中包含匹配项的总数和匹配项列表。可以展开每个匹配项以查看匹配项属性、直接匹配项信息、间接匹配项信息和相关资产总数（如果有）。直接匹配项是指包含匹配搜索查询的资产的部分或所有元数据的匹配项。间接匹配项是指链接到直接匹配搜索查询的资产的资产匹配项。

显示的搜索结果的顺序取决于以下因素：

- 匹配搜索条件的对象属性。对象名称的优先级高于对象说明。对象说明的优先级高于其他对象属性。
- 对象类型。数据域和数据域组的优先级低于其他对象。
- 内容管理。管理的配置文件结果的优先级高于不管理的配置文件结果。
- 搜索条件与对象的匹配次数（包括直接匹配项和间接匹配项）。
- 关键字的相对频率。

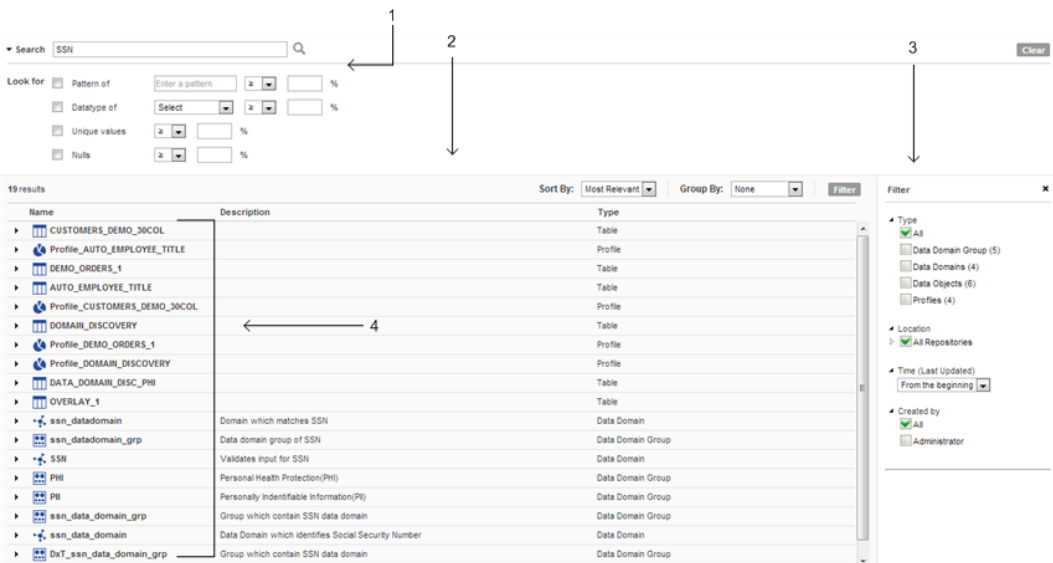
要查看搜索结果，您需要对包含直接匹配项和间接匹配项的对象具有相应的权限。

发现搜索结果面板

发现搜索结果中包含资产的名称、资产类型和资产说明。使用筛选器可缩小搜索结果的范围。

搜索结果在库工作区的结果网格中显示。可以根据相关性对结果进行排序。还可以根据资产类型、存储库位置、时间和创建资产的用户对结果进行分组。

下图显示了发现搜索结果界面：



1. 搜索条件
2. 结果网格
3. 筛选器
4. 搜索结果

发现搜索结果面板默认显示以下部分：

搜索条件

根据您能够设置的配置文件元数据来显示搜索字段（包括搜索筛选器），以缩小搜索范围。搜索字段显示在 Analyst 工具的顶部。

结果网格

根据您在搜索条件字段中选择的条件来显示匹配项总数和匹配项列表。结果网格中还包含对象说明、对象类型和下拉列表以对搜索结果进行排序和分组。

筛选器

显示您可以设置的筛选器以筛选搜索结果。**筛选器**部分在 Analyst 工具的右侧区域显示。

搜索结果

根据包含满足搜索条件的对象的搜索字符串来显示匹配搜索结果。搜索结果部分中包含在您展开匹配项时显示的匹配项属性、**直接匹配项**部分和**间接匹配项**部分。还可以在结果网格中查看匹配项的相关资产总数。

筛选发现搜索结果

可以根据资产类型、存储库位置、时间和创建资产的用户对搜索结果进行筛选。如果您安装了 Business Glossary，还可以对 Business Glossary 中的业务术语、类别和策略使用资产特定的筛选器。

1. 在**库**工作区的**发现搜索**部分中执行全局搜索或发现搜索。
2. 单击结果网格中的**筛选器**打开**筛选器**部分。
3. 在**筛选器**部分中，选择所需筛选器和相关设置。
4. 修订的搜索结果根据您选择的筛选器设置在结果网格中显示。
5. 要清除所有筛选器设置，请单击结果网格顶部的**全部清除**。

匹配类型

发现搜索结果包括直接匹配项和间接匹配项。直接匹配项是指包含匹配搜索查询的资产的部分或所有元数据的匹配项。间接匹配项是指链接到直接匹配搜索查询的其他资产的资产匹配项。

如果搜索查询包含多个搜索条件，搜索结果可能会直接、间接或通过这两种方式来满足搜索条件。可以在只读或编辑模式下打开搜索结果中的部分直接匹配项和间接匹配项。

直接匹配

直接匹配项是指包含匹配搜索查询的资产的部分或所有元数据的匹配项。例如，如果您搜索名为“Customer”的所有资产，Analyst 工具可能会列出名为“Customer”的数据对象和配置文件作为直接匹配项。执行发现搜索后，显示的匹配项列表中将包含指向部分对象的链接。

可以展开搜索结果中的资产以查看与直接匹配项有关的详细信息，例如资产属性。

间接匹配

间接匹配项是指链接到直接匹配项的匹配项。例如，结果卡使用的规则包含搜索关键字。发现搜索将返回该规则作为直接匹配项，返回结果卡作为间接匹配项。结果卡属于间接匹配项，因为它引用该规则。

使用间接匹配项信息可识别对象之间的隐藏关系以及更好地了解对象关系。还可以根据间接匹配项结果来了解发现搜索返回某个对象的原因。

查看匹配项信息

执行发现搜索后，可以查看包括直接匹配项和间接匹配项的匹配项信息。还可以查看资产属性，例如资产类型、说明和相关资产。可以打开搜索结果中的部分资产并根据需要对其进行更改。

1. 在**库**工作区的**搜索**部分中执行全局搜索或发现搜索。
2. 在结果网格中，单击资产名称开头的“展开”图标。
资产属性和匹配项信息在资产名称下方的部分中显示。
3. 检查直接匹配项和间接匹配项信息。

您可以看到资产关系及其他信息，例如相关资产总数。资产关系包括直接匹配项和间接匹配项。

- 4. 如果资产信息包含超链接，请单击超链接以打开其他工作区中的资产。
- 5. 再次单击“展开”图标以关闭匹配项信息部分。

打开发现搜索结果中的资产

您需要具有所需的项目、资产和许可证权限，才能查看发现搜索结果中的资产。

- 1. 在**库**工作区的**搜索**部分中执行全局搜索或发现搜索。
- 2. 在结果网格中，右键单击资产名称。
此时将显示一个快捷菜单。
- 3. 要在只读模式下查看工作区中的资产，请选择**打开**。
- 4. 要更改工作区中的资产，请选择**编辑**。
- 5. 要从搜索结果中删除资产，请选择**删除**。
从搜索结果中删除资产时，Analyst 工具将从模型存储库中删除该资产。
- 6. 要导航回**库**工作区，请单击**库**。

相关资产

可以查看搜索结果中的资产的相关资产。相关资产是指与搜索结果中的选定资产关联的模型存储库或 Business Glossary 中的资产。相关资产与搜索结果中的资产共享某些元数据。数据源可以将配置文件、推理的数据域和映射作为相关资产。

例如，配置文件可以作为搜索结果的一部分。可以查看配置文件的相关资产，例如配置文件的规则和数据源。可以查看**相关资产**工作区中的相关资产。显示的相关资产因资产类型而异。例如，查看规则的相关资产时，可以查看的资产包括关联业务术语、映射规范和配置文件。

每种资产类型的相关资产

Analyst 工具中显示的资产的相关资产取决于搜索的资产类型。

下表介绍了每种资产的相关资产：

资产类型	相关资产
业务术语	数据域、数据对象、Mapplet 和规则。
数据域	业务术语、数据域组、数据对象和配置文件。
数据域组	数据域、数据对象和配置文件。
数据对象	业务术语、数据域、数据域组、映射规范、配置文件、结果卡、映射和 Mapplet。
企业发现配置文件	数据对象和配置文件。

资产类型	相关资产
映射 注意: 可以在 Developer 工具中打开此对象。	数据对象、映射规范、Mapplet 和规则。
映射规范	数据对象、配置文件、结果卡、映射、Mapplet 和规则。
Mapplet 注意: 可以在 Developer 工具中打开此对象。	业务术语、数据对象、映射规范、映射、Mapplet 和规则。
配置文件	数据域、数据域组、数据对象、规则和映射规范。 注意: 结果卡不包含在配置文件的相关资产中。
规则	数据对象、规则、业务术语、映射规范、配置文件、结果卡和映射。

查看相关资产

查看搜索结果中的资产匹配项信息时，可以查看相关资产的总数。

1. 在**库**工作区中执行发现搜索。
2. 在结果网格中，单击展开图标并单击相关资产计数的链接，或者右键单击资产名称并选择**显示相关资产**。
所有相关资产的列表将在**相关资产**工作区中显示。
3. 要查看资产的详细信息，请单击资产名称，或者右键单击并选择**打开**。
4. 要查看相关资产的相关资产，请右键单击资产名称，然后选择**显示相关资产**。
相关资产信息将在工作区中显示。
5. 要在多个相关资产工作区之间导航，请从**相关资产**工作区中选择最近打开的其中一项资产。

常见问题

我为什么无法查看期望看到的部分搜索结果？

搜索结果可能因多种原因而不显示。请验证搜索条件是否满足以下准则：

- 搜索结果中显示的资产取决于项目权限。
- 发现搜索结果中不包含列配置文件结果中的值相关性。
- 搜索结果不包括您在管理配置文件结果时拒绝的配置文件结果。
- 查看的搜索结果取决于搜索索引的提取时间间隔和搜索索引中的资产的可用性。

我可以保存发现搜索结果以供将来使用或者与其他用户共享结果吗？

不可以。不能保存或共享发现搜索结果。

我为什么在搜索结果的顶部或底部看到部分发现搜索结果？

Analyst 工具显示搜索结果的顺序取决于多种因素。部分因素包括对象类型、管理的配置文件结果、主要与搜索条件匹配的对象属性以及每个对象的内部搜索排名。

我可以导出发现搜索结果吗？

不可以。不能导出搜索结果。

第 16 章

Informatica Analyst 中的 Business Glossary 桌面版

本章包括以下主题：

- [业务术语, 119](#)
- [管理 Metadata Manager Business Glossary 中的业务术语, 120](#)
- [在 Business Glossary 桌面版中查找业务术语, 120](#)

业务术语

可以在 Business Glossary 桌面版中查找业务术语。可以基于 Metadata Manager 的许可证查看业务术语并执行业务术语任务。

业务词汇表是一系列使用业务语言为企业用户定义概念的术语。业务术语提供了业务定义和概念的用法。

Business Glossary 桌面版是连接到托管业务词汇表的 Metadata Manager 服务的客户端。必须先将 Business Glossary 桌面版打开，然后才能查找 Analyst 工具对象名称。在 Business Glossary 桌面版中查找某个 Analyst 工具对象名称作为业务术语的含义，以了解其业务要求和当前实施。

Metadata Manager 托管业务词汇表。必须将 Metadata Manager 服务与分析服务相关联，才能从 Analyst 工具浏览 Metadata Manager 业务词汇表。可以在业务词汇表中查看业务术语，或者查看按类别分组的业务术语。可以编辑 Metadata Manager 业务术语。

可以通过 Metadata Manager 业务术语在 Metadata Manager 存储库中搜索 Metadata Manager 对象。可以从搜索结果中选择 Metadata Manager 对象，并将这些对象作为数据对象导入到 Analyst 工具中。无法将 Metadata Manager 业务术语添加到 Metadata Manager 业务词汇表中。

管理 Metadata Manager Business Glossary 中的业务术语

可以从 Analyst 工具访问 Metadata Manager Business Glossary，以便管理 Metadata Manager 业务术语。

1. 在 Analyst 工具标题上，单击**管理 > 管理术语**。
Metadata Manager 和 Metadata Manager Business Glossary 即在其他选项卡中打开。Metadata Manager 业务术语显示在 Metadata Manager 的**词汇表**视图上。
2. 要选择业务词汇表，请从“显示”列表中选择一个词汇表。
3. 要查看按类别分组的业务术语，请单击**操作 > 查看 > 类别**。
4. 要按字母顺序查看业务词汇表中的所有业务术语，请单击**操作 > 查看 > 字母表**。
5. 要查看以特定字母开头的所有业务术语，请单击该字母。
6. 要编辑业务术语，请选择业务术语然后单击**操作 > 编辑属性**。

在 Business Glossary 桌面版中查找业务术语

在 Business Glossary 桌面版中查找作为业务术语的某个 Analyst 工具对象名称，以了解其业务要求和当前实施。

您必须在计算机上安装 Business Glossary 桌面版。

1. 突出显示对象的名称。
2. 使用热键组合在 Business Glossary 桌面版中查找作为业务术语的某个对象的名称。
默认热键组合为 SHIFT+ALT+Q。

第 III 部分：使用 Informatica Developer 执行数据发现

本部分包含以下章节：

- [Informatica Developer 配置文件, 122](#)
- [数据对象配置文件, 125](#)
- [基于半结构化数据源的列配置文件, 138](#)
- [Informatica Developer 中的规则, 144](#)
- [Mapplet 和映射剖析, 146](#)
- [Informatica Developer 中的列配置文件结果, 148](#)
- [Informatica Developer 中的结果卡, 153](#)
- [Informatica Developer 中的数据域发现, 155](#)
- [Informatica Developer 中的企业发现, 166](#)
- [企业发现结果, 180](#)
- [Informatica Developer 中的 Business Glossary 桌面版, 189](#)

第 17 章

Informatica Developer 配置文件

本章包括以下主题：

- [Informatica Developer 配置文件概览, 122](#)
- [Informatica Developer 配置文件视图, 123](#)
- [存储库对象锁定和使用受版本控制的对象进行基于团队的开发, 124](#)

Informatica Developer 配置文件概览

在 Informatica Developer 中创建并运行配置文件可发现数据集中的数据质量问题并了解数据集中的列关系。

可以为以下类型的数据分析创建配置文件：

- 列剖析
- 对半结构化数据源执行列剖析
- 主键发现
- 功能相关性发现
- 外键发现
- 联接分析
- 重叠发现
- 数据域发现
- 企业发现

在 Developer tool 中通过向导创建配置文件。配置创建向导可向您提供**配置文件**、**多个配置文件**以及**企业发现配置文件**选项来创建配置文件。

配置文件

为单个数据对象创建配置文件。对于单个配置文件，可以为列剖析定义筛选器、规则和向下钻取选项。还可以选择高级选项以针对数据域发现创建列配置文件、主键配置文件和功能相关性配置文件。结果会显示列配置文件、主键推理、功能相关性和数据域推理。您可以为平面文件数据对象、关系数据对象和半结构化数据对象创建一个列配置文件。

多个配置文件

为多个对象创建配置文件集。Developer tool 可为每个对象创建配置文件并同时运行这些配置文件。一次可创建多个配置文件，但不能分析对象间的数据。

企业发现配置文件

从多个数据对象中构建数据模型，然后创建分析各对象间的数据的配置文件。创建企业发现配置文件，然后向其中添加您要共同剖析的物理数据对象。可以创建数据对象配置文件、外键配置文件和联接配置文件。对于企业发现配置文件中的每个数据对象，可以配置常规属性、要剖析的列、键和关系。可以在一个数据源或多个数据源中发现重叠的数据。

还可以运行企业发现，该企业发现可用于创建和运行数据发现任务，如列配置文件、数据域发现、主键配置文件和外键配置文件。Enterprise 发现会在多个连接间的大量数据源中运行。

下表列出了您可以使用各个配置文件类型执行的操作：

配置文件选项	配置文件操作
配置文件	<ul style="list-style-type: none">- 运行列配置文件。- 查找主键。- 查找功能相关性。- 标识数据域
多个配置文件	同时在多个对象中创建并运行列配置文件。
企业发现配置文件	<ul style="list-style-type: none">- 在单个数据集中运行列配置文件。- 查找主键。- 查找外键。- 查找功能相关性。- 执行联接分析。- 发现两列之间的重叠。- 运行企业发现。

Informatica Developer 配置文件视图

可以使用**概览**、**定义**、**注释**和**结果**视图查看和添加有关 Informatica Developer 中的配置文件的信息。

通过**对象浏览器**视图打开配置文件时，右侧窗格中的编辑器将在以下视图下显示配置文件信息：

概览

查看和提供有关配置文件的常规信息，例如名称、说明和位置。

定义

查看和设置配置文件定义。

该信息包括分配给配置文件的筛选器和规则列表、向下钻取选项以及在配置文件运行期间启用的配置文件函数。

结果

显示配置文件运行的结果。可以在运行配置文件后导出结果。

注释

查看注释并将其添加到配置文件中。

存储库对象锁定和使用受版本控制的对象进行基于团队的开发

模型存储库会锁定配置文件，以防止用户覆盖其他用户所做的工作。如果模型存储库与版本控制系统集成在一起，它会保存多个资产版本并为版本分配版本号。可以签出和签入配置文件，撤消签出，以及查看已签出的配置文件。

如果 Developer tool 意外停止，模型存储库会保持对象锁定。当您再次连接到模型存储库时，可以查看已锁定的对象。您可以继续编辑对象，也可以解除对象锁定。可通过**已锁定对象**对话框查看和解锁已锁定的对象。要查看**已锁定对象**对话框，请单击**视图 > 已锁定对象**。

如果模型存储库与版本控制系统集成在一起，则可使用受版本控制的对象管理来管理 Developer tool 中的对象版本。可以执行诸如签出和签入对象、查看和检索对象的历史版本以及撤消签出等操作。

模型存储库保护对象不被开发团队的其他成员覆盖。如果您打开已被其他用户签出的对象，将收到一条通知，告诉您将该对象签出的用户。您可以在只读模式下打开已签出对象，也可以使用不同名称保存它。

修改对象后，模型存储库会创建新的对象版本。

管理数据类型、主键、外键或数据域后，模型存储库会递增版本号。

还原版本时，“结果”视图中将显示最新的配置文件结果，而非还原后版本的配置文件结果。这是因为版本控制系统会在模型存储库中保留所有版本的配置文件定义，而配置文件结果将从剖析仓库中提取。有关存储库对象锁定和受版本控制的对象管理的详细信息，请参阅《*Developer Tool 指南*》。

第 18 章

数据对象配置文件

本章包括以下主题：

- [数据对象配置文件概览, 125](#)
- [Informatica Developer 中的列配置文件, 126](#)
- [运行时环境, 128](#)
- [主键发现, 129](#)
- [功能相关性发现, 130](#)
- [Informatica Developer 中的操作系统配置文件, 132](#)
- [在 Informatica Developer 中创建单个数据对象配置文件, 132](#)
- [在 Informatica Developer 中创建多个数据对象配置文件, 133](#)
- [编辑配置文件, 134](#)
- [同步选项, 134](#)
- [注释, 136](#)

数据对象配置文件概览

数据对象配置文件在数据源中发现有关列数据和元数据的信息。可以在 Informatica Developer 中对单个数据对象和多个数据对象运行配置文件。单个数据对象配置文件分析一个数据源。多个数据对象配置文件分析多个数据源。创建多个数据对象配置文件时，可以对这些配置文件运行列配置文件。

下表描述了可以针对单个数据对象配置文件执行的数据发现任务：

任务	说明
列剖析	发现数据的特性，如频率、百分比和模式。可以添加筛选器以确定配置文件在运行时读取的行。配置文件不处理不满足筛选条件的行。
主键发现	发现值可以唯一标识数据源中的行的列。
功能相关性发现	发现数据源中成对的列之间的相关性。
数据域发现	基于列值或列名称识别列的所有数据域。

下表描述了在使用**企业发现配置文件**选择创建数据模型时可以对多个数据对象执行的数据发现任务：

任务	说明
外键发现	发现所含值与另外一个数据源中的主键值相匹配的列。
联接分析	发现一个数据源中或两个数据源之间的两个列中的数据间的可能联接程度。
重叠发现	发现一个数据源或多个数据源中成对列之间重叠数据的百分比。
企业发现	发现分布于多个连接或架构之间的大量数据源中的列配置文件统计信息、数据域、主键和外键。

Informatica Developer 中的列配置文件

可以使用列配置文件来分析数据源中各列的特性，如值百分比和值模式。可以添加筛选器以确定配置文件在运行时读取的行。配置文件不处理不满足筛选条件的行。

可以发现关于对其运行配置文件的列的以下类型的信息：

- 某个值在列中出现的次数。
- 某一列中每个值出现的频率，用百分比或行数表示。
- 某一列中各个值的字符模式。
- 统计信息，如某一列中各个值的最大长度和最小长度，以及第一个值和最后一个值。
- 推理的数据类型、频率、数据域发现的遵从性条件以及数据类型推理状态。

可以为模型存储库中的映射、Mapplet 或对象的数据对象定义列配置文件。存储库中的对象可以位于一个或多个数据对象配置文件中，也可以位于企业发现配置文件中。

您可以为列配置文件选择采样选项、向下钻取选项和运行时环境。可以向列配置文件添加规则和筛选器。

筛选选项

您可以添加高级筛选器或 SQL 筛选器来确定在您运行配置文件时列配置文件使用的行。配置文件不会处理不满足筛选条件的行。

创建高级筛选器

您可以使用诸如 AND、OR 和 NOT 等表达式创建高级筛选器，以创建原始数据源的子集。

1. 创建或打开单个数据对象配置文件。
2. 选择**筛选器**视图。
3. 单击**添加**。
此时将显示**选择向导**对话框。
4. 在**选择向导**对话框中，单击**高级筛选器**。
此时将显示**筛选器**对话框。
5. 输入高级筛选器的名称和可选说明。
6. 选择**设置为活动状态**以将筛选器应用到配置文件中。单击**下一步**。
7. 选择**筛选器定义**来定义筛选器。

8. 您可以使用**函数**面板或**列**面板创建高级筛选器。
 - 在**函数**面板中，选择函数类别，然后单击向右箭头 (>>) 按钮。
在对话框中，指定参数并单击**确定**。函数将与列和值一起显示在**表达式**面板中。
 - 在**列**面板中，选择列，然后单击向右箭头 (>>) 按钮。该列将显示在**表达式**面板中。
您可以添加函数、表达式和值，以创建高级筛选器。
9. 要验证高级筛选器，请单击**验证**。
10. 创建或编辑筛选器后，请选择**数据预览**以查看筛选后的数据。您可以设置**要预览的最大行数**选项。
11. 单击**完成**。
此时会出现**新建配置文件**向导，并且筛选器位于**筛选器**视图中。

创建 SQL 筛选器

您可以使用 SQL 查询创建 SQL 筛选器。可以为关系数据源创建 SQL 筛选器。

1. 创建或打开单个数据对象配置文件。
2. 选择**筛选器**视图。
3. 单击**添加**。
此时将显示**选择向导**对话框。
4. 在**选择向导**对话框中，单击 **SQL 筛选器**。
此时将显示**筛选器**对话框。
5. 输入高级筛选器的名称和可选说明。
6. 选择**设置为活动状态**以将筛选器应用到配置文件中。单击**下一步**。
7. 选择**筛选器定义**来定义筛选器。
8. 使用**列**面板中的列来创建 SQL 筛选器。
9. 要验证筛选器，请单击**验证**。
10. 创建或编辑筛选器后，请选择**数据预览**以查看筛选后的数据。您可以设置**要预览的最大行数**选项。
11. 单击**完成**。
此时会出现**新建配置文件**向导，并且筛选器位于**筛选器**视图中。

采样选项

采样选项决定了 Developer tool 运行配置文件的行数。可以在定义配置文件或运行配置文件时配置采样选项。

下表介绍了配置文件的各个采样选项：

属性	说明
所有行	对数据对象的所有行运行配置文件。 本地、Blaze 和 Spark 运行时环境中支持此选项。
对前 <数字> 行进行采样	对从数据对象第一行开始的采样行运行配置文件。最多可以选择 2,147,483,647 行。 本地和 Blaze 运行时环境中支持此选项。
对 <数字> 行随机采样	对数据对象中随机选取的若干行运行配置文件。最多可以选择 2,147,483,647 行。 本地和 Blaze 运行时环境中支持此选项。

属性	说明
随机采样(自动)	对基于数据对象中的行数计算出的采样行运行配置文件。 本地和 Blaze 运行时环境中支持此选项。
限制 n <数字> 行	根据数据对象中的行数运行配置文件。选择在 Hadoop 验证环境中运行配置文件时，Spark 引擎会从数据对象的多个分区收集样本并将这些样本推送到单个节点来计算采样大小。“限制 n”采样选项支持 Oracle、SQL Server 和 DB2 数据库。不能对“限制 n”采样选项使用高级筛选器。 Spark 运行时环境中支持此选项。
随机百分比	对数据对象中某一百分比的行运行配置文件。 Spark 运行时环境中支持此选项。
在后续运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型和数据域	在下次运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型或数据域。

选择对随机的行样本运行配置文件后，随机采样算法将以随机方式在数据对象中选择行来运行配置文件。为列配置文件选择随机采样选项时，Developer tool 将对暂存的数据执行向下钻取。这会影响向下钻取性能。为数据域发现配置文件选择随机采样选项时，Developer tool 将对实时数据执行向下钻取。

运行时环境

选择本地或 Hadoop 作为列配置文件的运行时环境。在 Hadoop 运行时环境中，可以选择 Blaze 或 Spark 选项。选择运行时环境后，Informatica Developer 会在配置文件定义中设置该运行时环境。

本地环境

在本地运行时环境中运行配置文件时，Developer tool 会向剖析服务模块提交配置文件作业。剖析服务模块随后将配置文件作业拆分成一组映射。数据集成服务会在运行数据集成服务的同一台计算机上运行这些映射，并将配置文件结果写入剖析仓库。默认情况下，所有配置文件都在本地运行时环境中运行。

您可以使用本地源在本地环境中创建并运行配置文件。本地数据源是非 Hadoop 源，例如平面文件、关系源或大型机源。您还可以使用本地环境中的 Hive 或 HDFS 数据源，在映射规范或逻辑数据源中运行配置文件。

Hadoop 环境

在 Hadoop 运行时环境中可以选择 Blaze 或 Spark 选项来运行配置文件。

选择 Blaze 选项后，可以选择 Hadoop 连接。数据集成服务将配置文件逻辑推送到 Hadoop 群集上的 Blaze 引擎来运行配置文件。

在 Hadoop 环境中运行配置文件时，Developer tool 会向剖析服务模块提交配置文件作业。剖析服务模块随后将配置文件作业拆分成一组映射。数据集成服务会通过 Hadoop 连接将映射推送至 Blaze 引擎。Blaze 引擎会处理映射，并且数据集成服务会将配置文件结果写入剖析仓库。

选择 Spark 选项后，可以选择 Hadoop 连接。数据集成服务将配置文件逻辑推送到 Hadoop 群集上的 Spark 引擎来运行配置文件。在 Hadoop 环境中运行配置文件时，Developer tool 会向剖析服务模块提交配置文件作业。剖析服务模块随后将配置文件作业拆分成一组映射。数据集成服务会通过 Hadoop 连接将映射推送至 Spark 引擎。Spark 引擎会处理映射，并且数据集成服务会将配置文件结果写入剖析仓库。

Sqoop 数据源的列配置文件

您可以在使用 Sqoop 的数据对象上运行列配置文件。选择 Hadoop 作为验证环境后，可以选择 Hadoop 连接中的 Blaze 或 Spark 引擎来运行列配置文件。

在逻辑数据对象或自定义数据对象上运行列配置文件时，您可以配置 num-mappers 参数以实现并行处理并优化性能。您还必须配置 split-by 参数，以指定 Sqoop 必须据其拆分工作单元的列。

请使用以下语法：

```
--split-by <column_name>
```

如果主键的值并没有在最小值和最大值范围内均匀分布，则可以将 split-by 参数配置为指定另一个具有平衡的数据分布的列来拆分工作单元。

如果没有定义 split-by 列，Sqoop 会根据以下条件拆分工作单元：

- 如果数据对象包含单个主键，Sqoop 会将主键用作 split-by 列。
- 如果数据对象包含复合主键，则 Sqoop 会默认处理不包含 split-by 参数的复合主键。有关详细信息，请参阅 Sqoop 文档。
- 如果数据对象包含具有相同列的两个表，您必须使用表限定名称定义 split-by 列。例如，如果表名称是 CUSTOMER，列名称是 FULL_NAME，请将 split-by 列定义如下：
--split-by CUSTOMER.FULL_NAME
- 如果数据对象不包含主键，m 参数和 num-mappers 参数的值默认为 1。

如果您使用 Teradata 提供技术支持的 Cloudera 连接器或 Hortonworks Connector for Teradata，并且 Teradata 表不包含主键，则需要配置 split-by 参数。

主键发现

主键发现从您指定的列生成主键候选。

主键是可以唯一识别数据源中的行的一个列或列组合。主键发现可识别满足特定可信度的列和列的组合。可以为 主键标识编辑可信度以及要组合的最大列数。

主键发现可以通过识别主键候选中不唯一的行来突出显示潜在数据质量问题。这在主键发现组合了很多列时尤其有用，因为非相容记录可能包含重复信息。

主键推理属性

创建单个数据对象配置文件时，可以使用**主键剖析**视图配置主键推理属性。

下表描述了**主键剖析**视图中的主键推理属性：

属性	说明
替代默认推理选项	允许您为主键推理配置自定义设置。
最大键列数	可构成主键的最大列数。
最大行数	要剖析的行数。
遵从性条件	在您确定主键时，配置文件允许键冲突的最大百分比或者最大行数。

属性	说明
使用已记录的用户定义键排除数据对象	使用已记录的主键或用户定义的主键排除数据对象。
使用已批准键排除数据对象	使用已批准主键排除数据对象。

推理的主键属性

运行单个数据对象配置文件后，可以使用**主键剖析**视图查看数据源中推理主键的详细信息。

下表描述了**主键剖析**视图中的推理主键属性：

属性	说明
列	配置文件中的列的名称。
遵从百分比	列中唯一值的百分比。
重复百分比	列的重复值百分比。
空值百分比	列的空值的百分比。
已验证	确定列是否是主键列。
推理状态	列的推理状态。
上次运行时间	主键配置文件上次运行的日期和时间。

键冲突属性

运行单个数据对象配置文件后，可以使用**主键剖析**视图查看数据源中主键冲突的详细信息。

下表描述了**主键剖析**视图中的键冲突属性：

属性	说明
列	配置文件从中推理候选主键的列的名称。
键冲突数	主键候选中的键冲突数量。

功能相关性发现

功能相关性发现提供有关数据源中成对列之间的相关性的信息。

如果通过一对列中一个列的值可以可靠地预测另一个列中的值，则这对列具有功能相关性。例如，如果数据集包含一个“员工 ID”列和一个“生日”列，则在包含指定“员工 ID”的所有行中生日都应相同。

功能相关性可以通过识别不遵从列功能相关性的记录突出显示潜在的数据质量问题。例如，如果数据源中 99.8% 的行具有功能相关性，则其余行极有可能包含错误信息。

功能相关性推理属性

功能相关性剖析视图提供有关列之间的功能相关性的信息。

下表描述了**功能相关性剖析**视图中的功能相关性推理属性：

属性	说明
替代默认推理选项	允许配置功能相关性推理的自定义设置。
最大决定列数	配置文件可以组合的列数，用以查找决定。
最大行数	要剖析的行数。
返回的相关性数	配置文件显示的相关性数。默认为 最小覆盖范围 ，其显示最小的相关性集合，其中的每个列在相关性或决定中至少出现一次。
返回的最大相关性数	配置文件显示的最大相关性数。
遵从性条件	在确定功能相关性时，配置文件所允许的相关性冲突的最小百分比或最大行数。

推理的功能相关性属性

运行单个数据对象配置文件后，可以使用**功能相关性推理**视图查看数据源中列之间的推理的功能相关性的详细信息。

下表描述了**功能相关性推理**视图中的推理的功能相关性属性：

属性	说明
决定列	已针对功能相关性进行了分析的列的名称。
相关列	与决定列相关的列的名称。
空值百分比	列的空值的百分比。
遵从百分比	功能相关性匹配的百分比。
已验证	确定列是否具有功能相关性。
上次运行时间	功能相关性配置文件上次运行的日期和时间。

功能相关性冲突属性

该视图提供有关列之间的功能相关性的信息。运行单个数据对象配置文件后，可以使用**功能相关性推理**视图查看数据源中功能相关性冲突的详细信息。

下表描述了**功能相关性推理**视图中的功能相关性冲突属性：

属性	说明
决定列	已针对功能相关性进行了分析的列的名称。
相异相关项	唯一功能相关项的数量。

Informatica Developer 中的操作系统配置文件

您可以在 Developer tool 中选择一个操作系统配置文件。选择一个操作系统配置文件后，数据集成服务即会根据操作系统配置文件用户的权限创建并运行列配置文件和企业发现配置文件以及创建结果卡。

选择操作系统配置文件

您可以在 Informatica Developer 中选择操作系统配置文件。数据集成服务使用操作系统配置文件用户的权限来运行剖析作业。

- 在 Informatica Developer 中，单击**窗口 > 首选项**。
此时将显示**首选项**对话框。
- 单击 **Informatica > 运行配置 > 映射**。
此时将显示**映射**对话框。
- 在**映射**对话框中，清除**使用默认数据集成服务**选项。
- 单击**浏览**在列表中选择一个操作系统配置文件。
- 单击**确定**。

在 Informatica Developer 中创建单个数据对象配置文件

您可以为数据对象中的一个或多个列创建单个数据对象配置文件，然后将该配置文件对象存储在模型存储库中。

- 在**对象浏览器**视图中，选择要剖析的数据对象。
- 单击**文件 > 新建 > 配置文件**以打开配置文件向导。
- 选择**配置文件**，然后单击**下一步**。
- 输入配置文件的名称并验证项目位置。如果需要，浏览到新位置。
- 或者，输入配置文件的文本说明。
- 验证所选数据对象的名称是否显示在**数据对象**部分。

7. 单击**下一步**。
8. 配置要执行的配置文件操作。可以配置以下操作：
 - 列剖析
 - 主键发现
 - 功能相关性发现
 - 数据域发现

注意: 要启用配置文件操作, 请为该操作选择**已在“运行配置文件”操作中启用**。默认情况下将启用列剖析。

9. 查看配置文件的选项。

可以编辑所有配置文件类型的列选择。查看列配置文件的筛选器和采样选项。可以查看主键、功能相关性和数据域发现的推理选项。还可以查看数据域发现的数据域选择。
10. 查看向下钻取选项, 并根据需要进行编辑。默认情况下, **启用行向下钻取**选项为选中状态。您可以编辑列配置文件的向下钻取选项。这些选项还可确定向下钻取操作是从数据源读取还是从暂存数据读取, 以及配置文件是否存储上一次配置文件运行的结果数据。
11. 在**运行设置**部分, 选择验证环境。选择**本地**或 **Hadoop** 作为验证环境。可以选择**本地**、**Blaze** 或 **Spark** 作为运行时环境。选择 **Blaze** 或 **Spark** 后, 可以选择 Hadoop 连接。
12. 单击**完成**。

在 Informatica Developer 中创建多个数据对象配置文件

对多个数据对象运行多个数据对象配置文件时, Developer tool 将使用默认的列剖析选项为一个或多个数据对象生成列配置文件。您也可以选择创建一个企业发现配置文件, 以对多个数据对象运行一个配置文件。

1. 在**对象浏览器**视图中, 选择要剖析的数据对象。
2. 单击**文件 > 新建 > 配置文件**, 以打开**新建配置文件**向导。
3. 在**新建**向导中, 选择**多个配置文件**选项, 然后单击**下一步**。
4. 在**多个配置文件**窗口中, 选择要在其中创建配置文件的位置。可以在与其剖析对象相同的位置创建各个配置文件, 也可以为配置文件指定一个通用位置。
5. 验证所选数据对象的名称是否显示在**数据对象**部分。

或者, 单击**添加**以添加其他数据对象。
6. 或者, 指定要剖析的行数, 然后选择是否在向导完成时运行配置文件。
7. 单击**下一步**。
8. 在**验证环境**部分中, 选择**本地**。

注意: 仅选择本地选项以运行多个数据对象配置文件。在 Hadoop 运行时环境中, 要在 Blaze 或 Spark 引擎上运行多个数据对象, 可以选择企业发现配置文件。
9. 单击**完成**。
10. 或者, 输入要添加到配置文件名称的前缀和后缀字符串。
11. 单击**确定**。

编辑配置文件

可以编辑单个数据对象配置文件或多个数据对象配置文件。如果启用了版本控制系统，默认情况下将签出该配置文件。

1. 在**对象浏览器**视图中，右键单击配置文件，然后单击**打开**。
此时将显示**结果**视图。
2. 在**定义**视图中，根据需要更新属性。
3. 单击**团队 > 签入**以签入配置文件。
4. 右键单击配置文件，然后单击**运行配置文件**以运行配置文件。
配置文件结果将显示在**结果**视图中。

同步选项

更改外部数据源的元数据时，默认情况下不会更新模型存储库中的数据对象元数据。可以使用“同步”选项将数据对象元数据与数据源元数据同步。

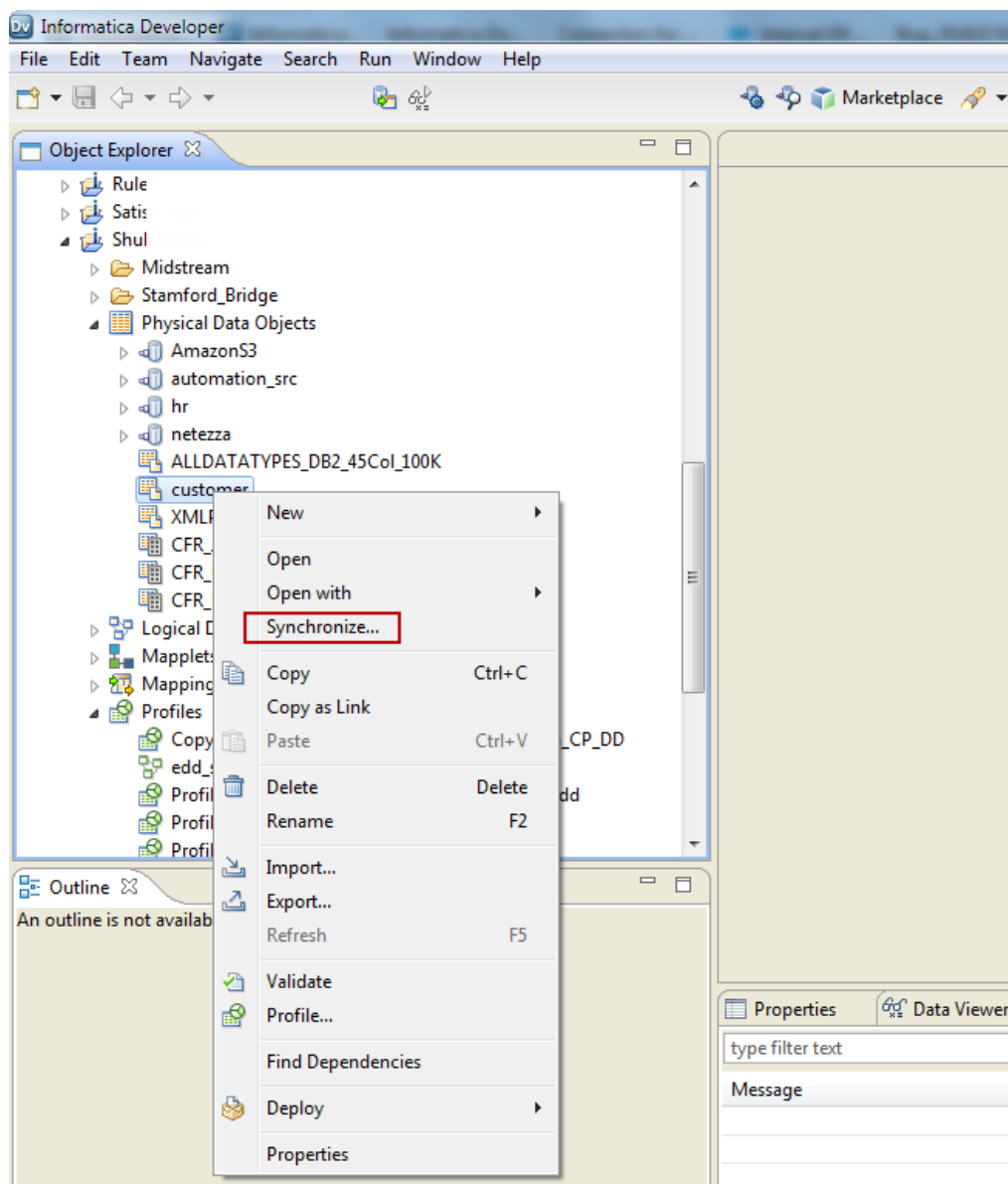
在 Developer tool 中使用“同步”选项后，当您打开使用该数据对象的配置文件或结果卡时，编辑器中的配置文件名称或结果卡名称旁边会显示星号。星号指示配置文件或结果卡的数据对象元数据已更改。请打开并保存该配置文件或结果卡，以更新模型存储库中的配置文件定义。请注意，当您在同步配置文件或结果卡的数据对象后打开该配置文件或结果卡时，Analyst 工具中不会显示任何可见更改。可以对列配置文件、企业发现配置文件和结果卡使用“同步”选项。外部数据源可以是关系数据源或平面文件数据源。

在 Informatica Developer 中同步平面文件数据对象

可以将外部平面文件数据源的更改与其在 Informatica Developer 中的数据对象同步。使用**同步平面文件**向导同步数据对象。

1. 在**对象浏览器**视图中，选择平面文件数据对象。
2. 右键单击，然后选择**同步**。

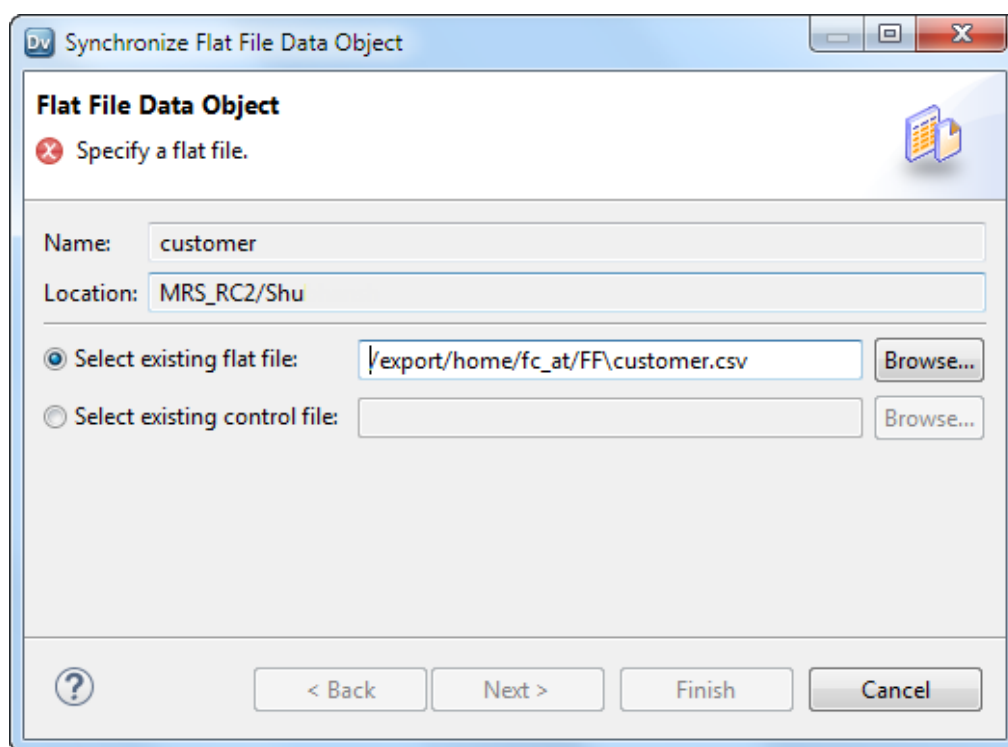
下图显示了数据对象的“同步”选项：



此时将显示同步平面文件数据对象向导。

3. 在同步平面文件数据对象向导中，验证选择现有平面文件字段中的平面文件路径。

下图显示了“同步平面文件数据对象”向导：



4. 单击**下一步**。
5. 或者，选择代码页、格式、带分隔符的格式属性以及列属性。
6. 单击**完成**，然后单击**确定**。

在 Informatica Developer 中同步关系数据对象

可以将关系数据源的外部数据源更改与其在 Informatica Developer 中的数据对象同步。外部数据源更改包括添加、更改和删除列以及规则更改。

1. 在**对象浏览器**视图中，选择关系数据对象。
2. 右键单击，然后选择**同步**。
此时将显示一条消息，提示您确认操作。
3. 要完成同步进程，请单击**确定**。
此时将显示同步过程状态消息。
4. 显示**同步完成**消息时，请单击**确定**。
该消息将显示数据对象的元数据更改的摘要。

注释

您可以添加说明以用作配置文件的注释，还可以向列配置文件结果中的列添加注释。

可以向一个列添加多个注释。您可以在 Developer tool 的**注释**视图中添加和查看注释。

在 Informatica Developer 添加注释

您可以在列配置文件结果中向列添加注释。导出配置文件结果时，Developer tool 会将注释包括在内。

1. 在**对象浏览器**视图中，打开一个配置文件。
2. 或者，运行配置文件以更新配置文件结果。
3. 选择**注释**视图。
4. 单击**添加**打开**添加注释**对话框。
5. 从列表中选择配置文件名称或其中一列。
如果此前添加了注释，则您可以在此对话框中查看注释。
6. 在**注释**字段中，输入说明。
7. 单击**确定**。
Developer tool 将在**注释**视图中显示注释。
8. 要删除注释，请在**注释**视图中选择注释，然后单击**删除**。

第 19 章

基于半结构化数据源的列配置文件

本章包括以下主题：

- [基于半结构化数据源的列配置文件概览, 138](#)
- [JSON 和 XML 数据对象, 138](#)
- [HDFS 中半结构化数据源的复杂文件数据对象, 139](#)
- [创建 HDFS 连接, 140](#)
- [从 HDFS 中的 JSON 或 XML 文件创建复杂文件数据对象, 140](#)
- [从 Avro 或 Parquet 数据源创建复杂文件数据对象, 141](#)
- [创建基于半结构化数据源的列配置文件, 142](#)

基于半结构化数据源的列配置文件概览

您可以基于 Avro、JSON、Parquet 和 XML 数据源创建数据对象，然后在数据对象上创建列配置文件。

Avro、JSON、Parquet 和 XML 格式均为半结构化数据源。要使用半结构化数据源创建列配置文件，您可以执行以下任务：

1. 基于半结构化数据源创建物理数据对象。
2. 在该物理数据对象上创建并运行列配置文件。

您可以为 JSON 或 XML 数据源创建平面文件数据对象。您可以为采用 Hadoop 分布式文件系统 (HDFS) 的 Avro、JSON、Parquet 和 XML 数据源创建复杂文件数据对象。

JSON 和 XML 数据对象

您可以从 JSON 或 XML 数据源创建平面文件数据对象或复杂文件数据对象。您可以在数据对象上创建并运行列配置文件。

创建含有 JSON 或 XML 数据源路径的文本文件，然后将该文本文件用作数据源，以此来创建平面文件数据对象。您也可以将多个 JSON 或多个 XML 数据源的文件路径添加到文本文件中。

您可以使用复杂文件读取器，从 JSON 或 XML 数据源创建复杂文件数据对象。复杂文件读取器会为数据处理器转换提供输入，然后由数据处理器转换解析文件并将源数据转换为逗号分隔值平面记录。

注意：Developer tool 不支持使用 UTF-8 编码的 JSON 数据源。

从 JSON 或 XML 数据源创建数据对象

您可以从 JSON 或 XML 数据源创建平面文件数据对象或复杂文件数据对象。

1. 在 Developer tool 的**对象浏览器**视图中，选择要在其中创建数据对象和列配置文件的项目。
2. 单击**文件 > 新建 > 数据对象**。
此时将显示**新建**对话框。
3. 您可以选择创建平面文件数据对象或复杂文件数据对象。
 - 要创建平面文件数据对象，请执行以下任务：
 1. 选择**物理数据对象 > 平面文件数据对象**，然后单击**下一步**。
此时将显示**新建平面文件数据对象**对话框。
 2. 选择**从现有平面文件创建**，然后单击**浏览**以选择文本文件。单击**下一步**。
 3. 验证代码页是否为 **MS Windows Latin 1 (ANSI)**，**Latin 1 的超集**，并验证格式是否设置为带分隔符。单击**下一步**。
 4. 验证分隔符是否设置为**逗号**。单击**完成**。
 - 要创建复杂文件数据对象，请执行以下任务：
 1. 选择**物理数据对象 > 复杂文件数据对象**，然后单击**下一步**。
此时将显示**新建复杂文件数据对象**对话框。
 2. 输入数据对象的名称。选择**文件**作为访问类型。
 3. 单击**浏览**以选择 JSON 或 XML 文件。单击**完成**。
Developer 服务器位于 Linux 中时，必须将数据源的文件路径更新为服务器中的位置。要更新文件路径，请选择复杂文件数据对象，单击**数据对象操作**选项卡中的**读取**，然后在**数据对象操作详细信息**窗格中的**高级**选项卡中添加文件路径。

数据对象将显示在项目文件夹中。

HDFS 中半结构化数据源的复杂文件数据对象

您可以对使用 HDFS 的 Avro、JSON、Parquet 或 XML 文件创建并运行列配置文件。要读取 HDFS 格式的 JSON 或 XML 文件，您需要使用复杂文件读取器将 JSON 或 XML 输入传递给数据处理器进行转换。

HDFS 中来自 JSON 或 XML 数据源的复杂文件数据对象

您可以从 JSON 或 XML 文件创建复杂文件数据对象您可以在数据对象上创建并运行列配置文件。

请先与 HDFS 建立连接，然后再为 HDFS 格式的 JSON 或 XML 文件创建数据对象。

从 HDFS 格式的 JSON 和 XML 文件创建数据对象时，您可以使用以下方法之一：

- 基于 JSON 或 XML 文件创建复杂文件数据对象。
- 基于含有多个 JSON 或多个 XML 文件的文件夹创建复杂文件数据对象。

创建数据对象后，您可以在数据对象上创建并运行列配置文件。

HDFS 中来自 Avro 或 Parquet 数据源的复杂文件数据对象

您可以从 HDFS 格式的 Avro 或 Parquet 数据源创建复杂文件数据对象。您可以使用该数据对象创建并运行列配置文件。

可以从 Avro 或 Parquet 文件创建复杂文件数据对象，也可以基于包含多个 Avro 或多个 Parquet 文件的文件夹创建。可以从 Avro 和 Parquet 数据源创建访问类型为文件或连接且资源格式为二进制、Avro 或 Parquet 的复杂文件数据对象。您需要先创建 HDFS 连接，然后再从 Avro 和 Parquet 数据源创建复杂文件数据对象。

注意：您只能为平面结构化 Avro 和 Parquet 数据源选择 **Avro** 或 **Parquet** 资源格式。

从 HDFS 格式的 Avro 和 Parquet 文件创建数据对象时，您可以选择以下选项之一：

- 将访问类型选择为文件，将资源格式选择为二进制。
- 将访问类型选择为文件，将资源格式选择为 Avro 或 Parquet。
- 将访问类型选择为连接，将资源格式选择为 Avro 或 Parquet。

创建 HDFS 连接

在 Informatica Developer 中配置 HDFS 连接，以便基于 HDFS 格式的 Avro、JSON、Parquet 和 XML 数据源创建列配置文件。创建 HDFS 连接后，您可以创建复杂文件数据对象。

1. 单击**窗口 > 首选项**。
2. 选择 **Informatica > 连接**。
3. 展开域。
4. 选择连接类型**文件系统 > Hadoop 文件系统**，然后单击**添加**。
5. 输入连接名称。
6. 或者，也可以输入连接说明。
7. 单击**下一步**。
8. 配置连接属性。
9. 单击**测试连接**以验证与 HDFS 的连接。
10. 单击**完成**。

从 HDFS 中的 JSON 或 XML 文件创建复杂文件数据对象

可以从使用 HDFS 的 JSON 或 XML 源文件创建复杂文件数据对象，然后基于该数据对象创建列配置文件。

1. 在 Developer tool 的**对象浏览器**视图中，选择要在其中创建物理数据对象和列配置文件的项目。
2. 单击**文件 > 新建 > 数据对象**。
此时将显示**新建**对话框。
3. 选择**物理数据对象 > 复杂文件数据对象**，然后单击**下一步**。
此时将显示**新建复杂文件数据对象**对话框。
4. 输入数据对象的名称。选择**连接**作为访问类型。

5. 可以从 JSON 或 XML 文件创建数据对象，也可以基于包含多个 JSON 文件或多个 XML 文件的文件夹创建。
 - 要从 JSON 或 XML 文件创建复杂文件数据对象，请执行以下步骤：
 1. 单击**浏览**以选择连接。
 2. 在**添加资源**对话框中，单击**添加**以选择 JSON 或 XML 文件。
 3. 单击**完成**。
数据对象将显示在项目文件夹中。
 - 要基于包含多个 JSON 文件或多个 XML 文件的文件夹创建复杂文件数据对象，请执行以下步骤：
 1. 单击**浏览**以选择连接。
 2. 在**添加资源**对话框中，单击**添加**以选择文件夹中的 JSON 或 XML 文件。
 3. 单击**完成**。
数据对象将显示在项目文件夹中。
 4. 选择项目文件夹中的数据对象，然后单击**高级 > 运行时: 读取 > 源文件目录**。
 5. 删除文件路径中的文件名，而保留其中的文件夹名称。

从 Avro 或 Parquet 数据源创建复杂文件数据对象

可以从 Avro 和 Parquet 数据源创建访问类型为**文件**或**连接**的复杂文件数据对象。可以基于该数据对象创建列配置文件。

1. 在**对象浏览器**视图中，选择一个项目。
2. 单击**文件 > 新建 > 数据对象**。
此时将显示**新建**对话框。
3. 选择**物理数据对象 > 复杂文件数据对象**，然后单击**下一步**。
此时将显示**新建复杂文件数据对象**对话框。
4. 输入数据对象的名称。
5. 您可以选择访问类型为**连接**或**文件**。
 - 如果选择**连接**作为访问类型，请执行以下步骤：
 1. 单击**浏览**选择 HDFS 连接。
 2. 在**选择连接**对话框中，选择数据源，然后单击**确定**。
 3. 在**新建复杂文件数据对象**对话框中，单击**完成**。
数据对象将显示在项目文件夹中。
 - 如果选择**文件**作为访问类型并选择**二进制**作为资源格式，请执行以下步骤：
 1. 单击**浏览**选择本地计算机上的 Avro 或 Parquet 文件。
 2. 在**新建复杂文件数据对象**对话框中，单击**完成**。
数据对象将显示在项目文件夹中。
 3. 选择项目文件夹中的数据对象，然后单击**数据对象操作**视图。
 4. 在**数据对象操作**视图中，单击**读取 > 高级**选项卡。
 5. 在**高级**选项卡中，在**文件路径**字段中输入 Linux 或 Windows 计算机上数据源的文件路径。
 6. 将文件格式输入为**自定义输入**。

7. 在 Avro 数据源的**输入格式**字段中输入 **com.informatica.avro.AvroToXML**，以及在 Parquet 数据源的**输入格式**字段中输入 **com.informatica.parquet.ParquetToXML**。添加输入格式时，数据处理器的转换会在运行时将 Avro 或 Parquet 格式的数据源处理并转换成 XML 格式数据源。
- 如果选择**文件**作为访问类型并选择 **Avro** 或 **Parquet** 作为资源格式，请执行以下步骤：
 1. 单击**浏览**选择本地计算机上的 Avro 或 Parquet 文件。
 2. 在**新建复杂文件数据对象**对话框中，单击**完成**。
数据对象将显示在项目文件夹中。
 3. 创建数据对象后，导航到**数据对象操作 > 读取 > 高级选项卡**，然后确认**文件路径**字段中的文件路径是否与 Linux 或 Windows 计算机中的数据源对应。

注意: 您只能为平面结构化 Avro 和 Parquet 数据源选择 **Avro** 或 **Parquet** 资源格式。

您可选择具有多个 Avro 或多个 Parquet 文件的文件夹来创建数据对象。创建数据对象后，导航到**数据对象操作 > 读取 > 高级选项卡**，然后验证**文件路径**字段中的文件路径是否指向 Linux 或 Windows 计算机中的数据源文件夹。

创建基于半结构化数据源的列配置文件

基于 Avro、JSON、Parquet 或 XML 数据源创建平面文件数据对象或复杂文件数据对象后，您可以基于该数据对象创建和运行列配置文件。

1. 在**对象浏览器**视图中，为 Avro、JSON、Parquet 或 XML 文件选择数据对象。
2. 单击**文件 > 新建 > 配置文件**。
此时将显示**新建**对话框。
3. 选择**配置文件**。单击**下一步**。
此时将显示**新建配置文件**对话框。
4. 在**新建配置文件**对话框中，为配置文件添加名称和可选说明。
5. 选择**处理扩展文件格式**选项。单击**下一步**。

下图显示的是选择了**处理扩展文件格式**选项的**新建配置文件**向导：

1. 处理扩展文件格式。选择此选项可处理半结构化数据源。

注意: 为“资源格式”选择 **Avro** 或 **Parquet** 时，Avro 和 Parquet 数据源不会显示**处理扩展文件格式**选项。

6. 在**单个数据对象配置文件**页面中，根据需要选择**列选择**和**数据域发现**下的列和选项。单击**完成**。

注意: 如果 Developer tool 安装在 Linux 计算机上，而且 JSON 或 XML 物理数据对象为含有文本文件的平面文件数据对象，请执行以下任务：

1. 在**概览**选项卡中，更新**精度**值，以便将字符数加入服务器中数据源的文件路径。
 2. 在平面文件数据对象上创建配置文件后，将数据源的文件路径更新为服务器中的相应位置。要更新文件路径，请单击**高级**选项卡中的**运行时: 读取 > 源文件目录**，然后添加文件路径。
7. 右键单击配置文件，然后选择**运行配置文件**。
此时将显示配置文件结果。

第 20 章

Informatica Developer 中的规则

本章包括以下主题：

- [Informatica Developer 中的规则概览, 144](#)
- [在 Informatica Developer 中创建规则, 145](#)
- [在 Informatica Developer 中应用规则, 145](#)

Informatica Developer 中的规则概览

规则是业务逻辑，用于定义在运行列配置文件时对源数据应用的条件。可以在配置文件中添加规则以验证数据。可以在列配置文件中使用的经验证为有效的规则、预定义规则或可重用规则的 Maplet。

可以通过以下方法在列配置文件中使用的规则：

- 在 Developer tool 中，创建 Maplet 并使其成为有效的规则。此规则在 Analyst 工具中显示为可重用规则。可以在 Analyst 工具和 Developer tool 中将规则应用到列配置文件。
- 可以在列配置文件中使用的预定义规则。Informatica 随 Developer tool 和 Analyst 工具一起提供了预定义规则。
- 在 Analyst 工具中，创建规则规范并生成 Maplet。可以在 Analyst 工具中将规则规范应用到列配置文件。在 Developer tool 中，使该 Maplet 成为有效的规则。该规则将显示为可重用规则，您可以在列配置文件中使用该规则。

注意：在 Developer tool 中，无法在列配置文件中添加、编辑或删除规则规范。

规则必须符合以下要求：

- 必须包含一个输入转换和一个输出转换。不能在规则中使用数据源。
- 可以包含表达式转换、查找转换和被动数据质量转换。不能包含其他任何类型的转换。例如，由于规则是主动转换，因而无法包含匹配转换。
- 它不指定输入组之间的基数。

在 Informatica Developer 中创建规则

需要将 Maplet 作为规则进行验证才能在 Developer 工具中创建规则。

在 Developer 工具中创建 Maplet。

1. 右键单击 Maplet 编辑器。
2. 选择**验证为 > 规则**。

在 Informatica Developer 中应用规则

可以将规则添加到已保存的列配置文件中。无法将规则添加到配置为用于联接分析的配置文件中。

1. 浏览**对象浏览器**视图并查找所需的配置文件。
2. 右键单击配置文件并选择**打开**。
配置文件将在编辑器中打开。
3. 单击**定义**选项卡，然后选择“规则”。
4. 单击**添加**。
此时将打开**应用规则**对话框。
5. 单击**浏览**找到要应用的规则。
从存储库项目中选择一个规则，然后单击**确定**。
6. 单击**输入值**下的**值**列以选择规则的输入端口。
7. 或者，单击**输出值**下的**值**列以编辑规则输出端口的名称。
规则将显示在**定义**选项卡中。

第 21 章

Mapplet 和映射剖析

本章包括以下主题：

- [Mapplet 和映射剖析概览, 146](#)
- [对 Mapplet 或映射对象运行配置文件, 146](#)
- [比较映射或 Mapplet 对象的配置文件, 147](#)
- [从配置文件生成映射, 147](#)

Mapplet 和映射剖析概览

可以为 Mapplet 或映射中的对象定义列配置文件。如果要验证映射或 Mapplet 的设计而不保存配置文件结果，请对 Mapplet 或映射对象运行配置文件。还可以从配置文件生成映射。

对 Mapplet 或映射对象运行配置文件

在 Mapplet 或映射对象上运行配置文件时，配置文件在所有数据列上运行，并为针对数据对象暂存的数据启用向下钻取操作。可以在具有多个输出端口的 Mapplet 或映射对象上运行配置文件。

配置文件对通过映射并流向选定对象的输出端口的源数据进行跟踪。如果运行了映射，则配置文件会分析将显示在这些端口上的数据。

1. 打开 Mapplet 或映射。
2. 验证 Mapplet 或映射是否有效。
3. 右键单击数据对象或转换，然后选择**立即剖析**。

如果转换有多个输出组，将显示**选择输出组**对话框。如果转换有一个输出组，则配置文件结果将显示在配置文件的**结果**选项卡上。

4. 如果转换有多个输出组，则根据需要选择输出组。
5. 单击**确定**。

配置文件结果将显示在配置文件的**结果**选项卡上。

比较映射或 Mapplet 对象的配置文件

可以创建一个配置文件来分析 Mapplet 或映射中的两个对象，并且可以比较这些对象的列配置文件的结果。

与单个映射或 Mapplet 对象的列配置文件相似，配置文件比较在所有数据列上运行，并为针对数据对象暂存的数据启用向下钻取操作。将数据从源表移至目标表后，可以比较配置文件来验证数据迁移。还可以比较随时间变化的数据源上的配置文件。

与单个映射或 Mapplet 对象的配置文件相似，配置文件比较在所有数据列上运行。

1. 打开 Mapplet 或映射。
2. 验证 Mapplet 或映射是否有效。
3. 按 **CTRL** 键并单击编辑器中的两个对象。
4. 右键单击其中的一个对象，然后选择**比较配置文件**。
5. 或者，对配置文件比较进行配置，以匹配一个对象的列与其他对象的列。
6. 或者，单击一个对象中的列并将其拖到其他对象的列上以匹配列。
7. 或者，选择配置文件是分析所有列还是仅分析匹配的列。
8. 单击**确定**。

从配置文件生成映射

可以从配置文件创建映射对象。使用所创建的映射对象以开发有效的映射。所创建的映射具有一个基于已剖析对象的数据源，并可基于配置文件规则逻辑包含转换。创建映射后，添加对象以完成映射。

1. 在**对象浏览器**视图中，找到要在其中创建映射的配置文件。
 2. 右键单击配置文件名称，然后选择**生成映射**。
此时将显示**生成映射**对话框。
 3. 输入映射名称。或者，输入映射的说明。
 4. 确认映射的文件夹位置。
默认情况下，Developer 工具在配置文件所在的项目中的**映射**文件夹内创建映射。单击**浏览**为映射选择不同的位置。
 5. 确认 Developer 工具用于创建映射的配置文件定义。要使用其他配置文件，请单击**选择配置文件**。
 6. 单击**完成**。
映射将显示在**对象浏览器**中。
- 将对象添加到映射中以完成映射。

第 22 章

Informatica Developer 中的列配置文件结果

本章包括以下主题：

- [Informatica Developer 中的列配置文件结果, 148](#)
- [列值属性, 149](#)
- [列模式属性, 149](#)
- [列统计信息属性, 149](#)
- [列数据类型属性, 150](#)
- [Informatica Developer 中的内容管理, 150](#)
- [从 Informatica Developer 中导出配置文件结果, 151](#)

Informatica Developer 中的列配置文件结果

列配置文件分析通过突出显示数据的值频率、模式和统计信息来提供有关数据质量的信息。

下表介绍了每一类分析的配置文件结果：

配置文件类型	配置文件结果
列配置文件	<ul style="list-style-type: none">- 唯一值和空值的百分比和计数统计信息- 推理的数据类型- 数据源声明用于数据的数据类型- 最大值和最小值- 配置文件运行的最新日期和时间- 列中每个唯一数据元素的百分比和计数统计信息- 列中每个唯一字符模式的百分比和计数统计信息
主键配置文件	<ul style="list-style-type: none">- 推理的主键- 键冲突
功能相关性配置文件	<ul style="list-style-type: none">- 推理的功能相关性- 功能相关性冲突

列值属性

列值属性显示已剖析列中的值以及每个值在各列中出现的频率。频率以数字、百分比和条形图的形式显示。

要查看列值属性，请从**显示**列表中选择“值”。双击某一列值以向下钻取到包含该值的行。

下表介绍了列值的属性：

属性	说明
值	配置文件中列的所有值的列表。
频率	值在列中出现的次数。
百分比	值在列中出现的次数，表示为占列中所有值的百分比。
图	百分比的条形图。

列模式属性

列模式属性显示已剖析列中数据的模式以及模式在各列中出现的频率。模式显示为数字、百分比和条形图。

要查看模式信息，请从**显示**列表中选择“模式”。双击某一模式以向下钻取到包含该模式的行。

下表介绍了列值模式的属性：

属性	说明
模式	选定列的模式。
频率	模式在列中出现的次数。
百分比	模式在列中出现的次数，表示为占列中所有值的百分比。
图	百分比的条形图。

列统计信息属性

列统计信息包括各种属性，例如最大长度和最小长度，以及第一个值和最后一个值。

要查看统计信息，请从**显示**列表中选择“统计信息”。

下表介绍了列统计信息属性：

属性	说明
最大长度	列中最长值的长度。
最小长度	列中最短值的长度。
底部	列中的最后五个值。
顶部	列中的前五个值。
总和	数据类型为数值的列中所有值的总和。

注意：配置文件还将显示 Integer 类型的列的平均差和标准差统计信息。

列数据类型属性

列数据类型包括配置文件结果中每个列的所有推理的数据类型。

要查看数据类型信息，请从**显示**列表中选择**数据类型**。双击数据类型以向下钻取到包含该数据类型的行。

下表介绍了列数据类型的属性：

属性	说明
数据类型	配置文件中列的所有推理的数据类型的列表。
频率	数据类型在列中出现的次数，以数字表示。
遵从性百分比	数据类型在列中出现的百分比。
状态	<p>指示数据类型的状态。状态有“已推理”、“已批准”或“已拒绝”。</p> <p>已推理</p> <p>指示 Developer tool 推理的列数据类型。</p> <p>已批准</p> <p>指示列的已批准数据类型。批准数据类型后，该数据类型随即提交到模型存储库。</p> <p>已拒绝</p> <p>指示列的已拒绝数据类型。</p>

Informatica Developer 中的内容管理

内容管理是验证和管理数据源的已发现元数据的过程，以便元数据适合使用和报告。在 Informatica Developer 中管理元数据时，可以批准、拒绝和重置配置文件结果中的推理的数据类型或数据域。

您可为一个列批准一个数据类型或数据域。可以对列隐藏拒绝的数据类型或数据域。批准或拒绝推理的数据类型或数据域后，您可以重置数据类型或数据域以还原已推理状态。

批准数据类型

配置文件结果包括数据源中每个列的推理数据类型、频率、遵从性百分比以及推理状态。您可以为每一列选择和批准单一数据类型。

1. 在**对象浏览器**视图中，选择配置文件并将其打开。
2. 验证您是否位于**结果**选项卡中。
3. 在**列剖析**视图中，选择一列以在右侧面板中查看值频率、模式、数据类型和统计信息。
4. 在**详细信息**面板下，从**显示**列表中选择**数据类型**。
显示列的推理的数据类型。
5. 右键单击要批准的列并单击**批准**。
数据类型的状态将更改为**已批准**。
6. 要还原数据类型的推理状态，请右键单击该数据类型，然后单击**重置**。

拒绝数据类型

默认情况下，Informatica Developer 会在配置文件结果中显示推理的数据类型。您可以拒绝已推理或已批准的数据类型。可以选择显示或隐藏拒绝的数据类型。

1. 在**对象浏览器**视图中，选择一个配置文件。
2. 双击该配置文件将其打开。
配置文件将在选项卡中打开。
3. 在**列剖析**视图中，选择一行。
4. 要拒绝推理的列数据类型，请在右侧面板中选择**数据类型**视图。选择要拒绝的推理的数据类型，接着右键单击相应的行，然后选择**拒绝**。
Informatica Developer 会在数据类型列表中灰显拒绝的数据类型。
5. 要隐藏拒绝的数据类型，请右键单击相应的行，然后选择**隐藏拒绝项**。
6. 要查看拒绝的数据类型，请右键单击其中一行，然后选择**显示拒绝项**。

从 Informatica Developer 中导出配置文件结果

您可以将列配置文件结果导出为 .csv 文件或 Microsoft Excel 文件。将配置文件结果导出为 Microsoft Excel 文件时，Developer tool 会将信息保存至 .xlsx 文件。

1. 在**对象浏览器**视图中，打开一个配置文件。
2. 或者，运行配置文件以更新配置文件结果。
3. 选择**结果**视图。
4. 选择一列。
5. 在**详细信息**下，选择**值**、**模式**或**数据类型**，然后单击**导出**图标。
此时将打开**将数据导出到文件**对话框。
6. 接受或更改默认文件名。
7. 选择要导出的数据的类型。您可以选择**选定列的值**、**选定列的模式**、**选定列的数据类型**或**所有(摘要、值、模式、数据类型、统计信息、属性)**。

8. 单击**浏览**选择一个位置，并将文件本地保存在计算机上。
9. 如果不想将字段名称作为第一行导出，请清除**将字段名称作为第一行导出**复选框。
10. 单击**确定**。

第 23 章

Informatica Developer 中的结果卡

本章包括以下主题：

- [Informatica Developer 中的结果卡概览, 153](#)
- [创建结果卡, 153](#)
- [为结果卡沿袭导出资源文件, 154](#)
- [查看 Informatica Developer 中的结果卡沿袭, 154](#)

Informatica Developer 中的结果卡概览

结果卡是配置文件中对质量度量的一种图形表示形式。可以在 Developer 工具中查看结果卡。在 Developer 工具中创建结果卡后，可以连接到 Analyst 工具，以打开结果卡进行编辑。对数据对象中的当前数据或暂存在剖析仓库中的数据运行结果卡。

可以在 Analyst 工具中编辑、运行结果卡，还可以查看度量或度量组的结果卡沿袭。

创建结果卡

创建结果卡，然后将配置文件中的列添加到该结果卡中。必须首先运行配置文件，然后才能将列添加到结果卡中。

1. 在**对象浏览器**视图中，选择要在其中创建结果卡的项目或文件夹。
2. 单击**文件 > 新建 > 结果卡**。
此时将显示**新建结果卡**对话框。
3. 单击**添加**。
此时将显示**选择配置文件**对话框。选择包含要添加的列的配置文件。
4. 单击**确定**，然后单击**下一步**。
5. 选择要添加到结果卡中的列。
默认情况下，结果卡向导将选择配置文件中定义的列和规则。无法添加未包括在配置文件中的列。
6. 单击**完成**。

Developer 工具将创建结果卡。

7. 或者，单击**使用 Informatica Analyst 打开**以连接到 Analyst 工具，然后在 Analyst 工具中打开结果卡。

为结果卡沿袭导出资源文件

可以导出包含结果卡和相关对象的项目，作为资源文件供 Metadata Manager 使用。在 Metadata Manager 中以 XML 格式使用已导出的资源文件，以为结果卡沿袭创建并加载资源。

1. 要打开**导出向导**，请单击**文件 > 导出**。
2. 选择 **Informatica > 资源文件(供 Metadata Manager 使用)**。
3. 单击**下一步**。
4. 单击**浏览**以选择包含需要导出的结果卡对象和沿袭的项目。
5. 单击**下一步**。
6. 选择要导出的结果卡对象。
7. 输入导出文件的名称和文件位置。
8. 要查看**导出向导**与选定对象一起导出的相关项目，请单击**下一步**。
导出向导将显示相关对象。
9. 单击**完成**。

Developer 工具将对象导出到 XML 文件中。

查看 Informatica Developer 中的结果卡沿袭

要查看 Developer 工具中的度量或度量组的结果卡沿袭，请启动 Analyst 工具。

1. 在**对象浏览器**视图中，选择包含结果卡的项目或文件夹。
2. 双击该结果卡将其打开。
此时结果卡将显示在选项卡中。
3. 单击**使用 Informatica Analyst 打开**。
Analyst 工具将在浏览器窗口中打开。
4. 在 Analyst 工具的**结果卡**视图中，选择某一度量或度量组。
5. 右键单击然后选择**显示沿袭**。
此时将在对话框中显示结果卡沿袭图表。

第 24 章

Informatica Developer 中的数据域发现

本章包括以下主题：

- [Informatica Developer 中的数据域发现概览, 155](#)
- [Informatica Developer 中的数据域词汇表, 155](#)
- [Informatica Developer 中的数据域发现选项, 159](#)
- [在 Informatica Developer 中创建配置文件以执行数据域发现, 161](#)
- [在 Informatica Developer 中编辑配置文件, 161](#)
- [在 Informatica Developer 中运行配置文件以执行数据域发现, 162](#)
- [Informatica Developer 中的数据域发现结果, 162](#)

Informatica Developer 中的数据域发现概览

使用数据域词汇表管理数据域。要创建数据域，可以使用预定义的数据规则和列名称规则。还可以根据列配置文件结果中的特定值或模式生成数据域。

您可以选择源列、要与列数据和列名称匹配的数据域、采样选项、向下钻取选项以及运行时环境。您可以选择要对其执行数据域发现的最大行数。您可以为数据域发现选择遵从性条件。您可以从数据域发现中排除空值。运行配置文件后，可以验证、管理以及向下钻取结果。还可以通过 Developer tool 中的编辑器向数据模型中添加结果。

您可以创建一个利用采样选项和筛选器来执行数据域发现的配置文件。运行该配置文件时，即会对数据源应用采用选项和筛选器，同时生成一个数据集。数据域发现过程会使用该数据集来发现数据域。

Informatica Developer 中的数据域词汇表

可以在数据域词汇表中管理数据域和数据域组。可以添加、编辑和移除数据域和数据域组。还可以搜索特定数据域和数据域组。

可以将数据域从数据域词汇表导出到 XML 文件中。还可以将数据域从 XML 文件导入到数据域词汇表中。可以创建数据域组以将数据域组织到特定组（如个人健康信息 (PHI)、个人身份信息 (PII)）或任何其他与该项目相关的概念组。可以在多个数据域组中包含一个数据域。例如，电话号码可以同时属于 PII 和 PHI 数据域组。

在 Informatica Developer 中创建数据域组

可以向数据域组中添加数据域以进行有效的列数据分析。

1. 单击 **Window > 首选项**。
此时将显示**首选项**对话框。
2. 在**首选项**对话框中，选择 **Informatica > 数据域词汇表**。
此时，Developer 工具会在**数据域词汇表**面板中显示所有数据域列表。
3. 在**显示**字段中，选择**数据域组**。
数据域词汇表面板会根据数据域组重新排列数据域列表。
4. 在**数据域词汇表**面板中，选择**数据域组**。
5. 单击**添加**。
此时将显示**数据域组**对话框。
6. 输入名称和说明。
7. 单击**下一步**。
8. 单击**选择**打开**选择数据域**对话框。
9. 选择您要添加到数据域组中的数据域，然后单击**确定**。
Developer 工具会在**选择数据域**面板中列出所选数据域。
10. 单击**完成**。
Developer 工具会将数据域组添加到数据域词汇表中。

在 Informatica Developer 中创建数据域

可以创建数据域，然后将其添加到数据域词汇表中。还可以将数据域添加到一个或多个数据域组中。

1. 单击 **Window > 首选项**。
此时将显示**首选项**对话框。
2. 在**首选项**对话框中，选择 **Informatica > 数据域词汇表**。
此时，Developer 工具会在**数据域词汇表**面板中显示所有数据域列表。
3. 在**数据域词汇表**面板中，选择**数据域**。
4. 单击**添加**。
此时将显示**数据域**对话框。
5. 输入名称和说明。
6. 单击**使用数据规则**以发现基于列数据的数据域。还可以选择**使用列名称规则**以根据数据源中的列名称来发现数据域。
浏览按钮处于启用状态。
7. 单击**浏览**以打开**选择位置**对话框。
8. 选择相应的规则，然后单击**确定**。
创建数据域后，Developer 工具会将与数据域关联的规则和其他相关对象复制到数据域词汇表中。要编辑与数据域关联的规则，必须先转至原始规则，然后对该规则进行更改。然后，可以将已修改的规则重新与该数据域关联。
所选规则将在**数据规则**和**列名称规则**字段中显示。
9. 单击**下一步**。

10. 单击**选择**打开**选择数据域组**对话框。
11. 选择您要包含数据域在内的数据域组，然后单击**确定**。
Developer 工具会将所选数据域组显示在**分配给数据域组**窗格中。
12. 单击**完成**。
Developer 工具会将数据域添加到数据域词汇表中。

在 Informatica Developer 中基于配置文件结果创建数据域

运行列配置文件后，您可以查看源数据的值和模式。然后，便可以根据这些值和模式创建数据域。

1. 运行列配置文件以查看其结果。
2. 根据您要创建的数据域选择值或模式。
值、模式以及统计信息会显示在**结果**视图中。
3. 右键单击值或模式，然后选择**发送到 > 新数据域**。
此时将显示**数据域**对话框。
4. 输入数据域名称和可选说明。
默认情况下，将位置设置为数据域词汇表。
5. 单击**完成**。
数据域便会添加到数据域词汇表中。

在 Informatica Developer 中查找数据域

默认情况下，数据域词汇表会显示所有数据域。可以搜索特定数据域和数据域组。

首选项对话框中的**数据域词汇表**窗格会显示所有数据域和数据域组。可以通过以下方式搜索并查看有关数据域和数据域组的详细信息：
搜索数据域和数据域组。

在**数据域词汇表**面板的顶部字段中键入部分数据域名称或数据域组名称。如果在**数据域组视图**中，Developer 工具会列出其名称中包含搜索字符串的数据域以及与数据域关联的数据域组。如果在**数据域视图**中，Developer 工具会列出其名称中包含搜索字符串的所有数据域。

查看数据域组和数据域组中的数据域。

在**显示**字段中，选择**数据域组**。

查看所有数据域。

在**显示**字段中，选择**数据域**。

查看数据域的属性。

在**数据域词汇表**面板下方单击数据域名称以查看其属性。您可以查看说明和关联规则。要查看数据域所属的域组，请单击**在数据域组中显示**。

查看数据域组的属性。

在**数据域词汇表**面板下方单击数据域组名称以查看其说明。

导入数据域

在 Developer 工具中可以将数据域从源 XML 文件导入到数据域词汇表中。必须验证该文件是否包含您需要导入的数据域的相关信息。

1. 打开数据域词汇表。
2. 验证**数据域**或**数据域组**是否已选中。
3. 单击**导入**。
此时将显示**导入**对话框。
4. 在**文件名称**字段中, 输入您要从中导入数据域的 XML 文件名。
单击**浏览**选择文件。
5. 单击**下一步**。
此时, **选择要导入的对象**窗格会显示您可以指定源和目标的位置。
6. 在**源**面板中, 选择您要导入的数据域。
注意: 要选择多个数据域, 请按住 Shift 键。
7. 单击**自动与目标匹配**可将数据域移动至**目标**面板。
Developer 工具会尝试根据目标选择中的名称、类型和父层次结构单独匹配当前源选择的子代, 然后添加匹配的对象。
8. 单击**解决方案**以指定处理重复对象的方式。
可以重命名导入的对象、用导入的对象替换现有对象或者重用现有对象。默认情况下, Developer 工具将重命名所有重复的对象。
9. 单击**下一步**。
Developer 工具会对导入设置进行汇总以供您查看。可以在**其他导入设置**窗格中指定其他导入设置。
10. 单击**完成**。

导出数据域

可以在 Developer 工具中将数据域和数据域规则从数据域词汇表导出到 XML 文件。

1. 打开数据域词汇表。
2. 验证**数据域**或**数据域组**是否已选中。
3. 单击**导出**。
此时将显示**导出**对话框。
4. 要导出数据域, 请选择**导出数据域**。选择**导出数据域规则**以导出数据域规则。
5. 单击**下一步**。
6. 在**导出到文件**面板中, 选择您要导出的数据域或数据域规则。
7. 要导出数据域, 请单击**浏览**以选择导出文件及其位置。要将数据域规则导出到模型存储库服务中的另一个项目中, 请选择**复制到项目**, 然后选择您要将数据域规则复制到其中的项目。
8. 单击**下一步**。
依赖关系窗格会显示相关对象的列表。
9. 单击**下一步**。
此时将显示**内容导出设置**窗格。您可以选择关联的引用表以供导出使用。

10. 单击**完成**。

如果将使用引用表的规则与数据域进行关联，则可能无法在用于创建数据域的同一个 Developer 工具会话中导出引用表。在数据域词汇表中单击**导出**之后，会先断开与模型存储库服务的连接，然后再重新连接才能导入使用引用表的规则。

Informatica Developer 中的数据域发现选项

创建配置文件以执行数据域发现时，您可以选择源列、数据域和推理选项。还可以根据列的数据类型和数据长度选择将列从数据域发现中排除。

Informatica Developer 中的数据域选择

数据域选择选项会列出数据域词汇表中的所有域。您可以先搜索特定数据域并选中它们，然后再作为数据域发现的一部分运行这些数据域。

下表描述了数据域发现的**数据域选择**选项：

选项	描述
已作为“运行配置文件”操作的一部分启用	包括运行配置文件时的数据域发现选项。
名称	数据域名称。
说明	数据域の説明。
数据域组	数据域所属的数据域组的名称。
在层次结构中显示数据域组	列出每个数据域组下方对数据域进行分组的所有数据域组。

Informatica Developer 中的数据域列选择

使用**列选择**选项以选择您要作为数据域发现的一部分而运行的列。

下表描述了数据域发现的**列选择**选项：

选项	说明
列	列名称。
数据类型	列的数据类型。
精度	列的最大精度。
小数位数	列的等级。
可空	指示列可以包含空值。
说明	列の説明。

Informatica Developer 中的数据域推理选项

推理选项可用于确定域发现是否必须在列数据或列名称中运行，还是在两者中运行。您可以指定配置文件是否需要处理数据源中的所有行。您可以为数据域匹配选择遵从性条件，并可以在数据域发现过程中排除空值。

下表介绍了数据域发现的**推理**选项：

选项	说明
替代默认推理选项	允许您更改预定义的推理选项。
数据	配置文件在列数据中运行。
列名称	配置文件在列标题中运行。
数据和列名称	配置文件同时在列数据和列标题中运行。
要剖析的最大行数	配置文件可以在其中运行的最大行数。Developer tool 从源中的第一行开始选择行。
最低行数百分比	数据域匹配所需的数据集中最低遵从行数百分比。
最小行数	数据域匹配所需的数据集中最低遵从行数。
从数据域发现中排除空值	从数据域发现的数据集中排除空值。

最低遵从性百分比

您可以选择数据集中的最低行数百分比作为数据域发现的遵从性条件。

遵从性百分比是指匹配的行数除以总行数所得出的比率。

注意: Developer tool 将空值视为不匹配的行。包含大量空值的列可能不会导致执行数据域推理，除非您为最低遵从性百分比指定了值。

示例

您的一个数据源中含有 10,000 行，其中注释列在 2,500 行中含有社保号码。您创建了一个列配置文件和数据域发现，并将最低行数百分比设置为 30% 以作为遵从性条件。运行该配置文件时，配置文件结果不会将社保号码显示为推理的数据域，因为最低遵从性条件为 30% 行数，即数据源中要有 3,000 行。

最低遵从行数

您可以选择数据集中的最低行数作为数据域发现的遵从性条件。

示例

您的一个数据源中含有 10,000 行，其中注释列在 3 行中含有电子邮件地址。您创建了一个列配置文件和数据域发现，并将最低行数设置为 1 以作为遵从性条件。运行该配置文件时，配置文件结果会将电子邮件地址显示为推理的数据域，并显示三个遵从行以及其他推理的数据域。

排除空值

对数据源执行数据域发现时，您可以排除空值。选择最低行数百分比和排除空值选项时，遵从性百分比的计算方式为：将匹配行数除以总行数与列中空值数之差。

选择**从数据域发现中排除空值**选项和多个采样选项或筛选器时，数据域发现过程有所不同。

以下场景介绍了选择排除空值选项与采样选项和筛选器时的数据域发现结果：

- 选择**所有行**作为采样选项，且未使用筛选器。数据域发现过程会忽略列中的所有空值。
- 选择一个采样选项，且未使用筛选器。数据域发现过程会忽略采样数据中的所有空值，并对其余采样数据运行。
- 选择**所有行**作为采样选项，且使用筛选器。数据域发现过程会忽略筛选后数据中的所有空值，并对其余筛选后数据运行。
- 选择一个采样选项，且使用筛选器。数据域发现过程会忽略采样的筛选后数据中的所有空值，并对其余筛选后数据运行。

示例

您的一个数据源中含有 10,000 行，其中 3,000 行的“Comments”列中含有社保号码。您创建了一个列配置文件和数据域发现，并选择了以下选项：

- 选择**从数据域发现中排除空值**选项。
- 选择**所有行**作为采样选项。
- 选择**最低行数百分比**选项并将该选项配置为 12%。

运行该配置文件时，该配置文件会在数据集上运行，并在数据域发现过程中忽略空值。

在 Informatica Developer 中创建配置文件以执行数据域发现

可以在数据源中发现数据域作为单个数据对象配置文件或企业发现配置文件的一部分。执行数据域发现之后，您可以通过 Developer 工具中的编辑器对结果进行验证、向下钻取并将其添加到数据模型中。

1. 在 **Object Explorer** 视图中，选择包含配置文件的数据对象的项目。
2. 右键单击数据对象并选择**配置文件**。
此时将显示**新建向导**。
3. 选择**配置文件**。
4. 单击**下一步**。
Developer 工具会显示您可以配置配置文件的常规属性的另一个窗格。
5. 如果需要，请更改配置文件名称和说明。还可以添加或移除数据对象。
6. 单击**下一步**。
7. 选择您要在其中运行数据域发现的列以及您要与列进行匹配的数据域。
8. 根据需要更改默认推理选项。
9. 单击**完成**创建配置文件。

在 Informatica Developer 中编辑配置文件

可以在运行配置文件以执行数据域发现后对该配置文件进行更改。您可以排除具有特定数据类型的列，还可以更改列选择、数据域选择和推理选项。

1. 在 **Object Explorer** 视图中，选择包含您要编辑的配置文件的项目或文件夹。

- 2. 双击该配置文件将其打开。
配置文件定义会显示在选项卡中。
- 3. 根据需要，对列选择、数据域选择和推理选项作出更改。
- 4. 在**列选择**部分，可以单击**排除列**以根据数据类型设置排除选项。
此时将显示**排除列**对话框。
- 5. 保存更改。

在 Informatica Developer 中运行配置文件以执行数据域发现

可以选择在创建配置文件后立即运行该配置文件。还可以在创建配置文件后手动运行该配置文件。

- 1. 在 **Object Explorer** 视图中，选择包含您要运行的配置文件的项目或文件夹。
要自动运行配置文件，请在创建配置文件后在**新建配置文件**向导中选择**完成时运行配置文件**。
- 2. 双击该配置文件将其打开。
配置文件定义会显示在选项卡中。
- 3. 右键单击配置文件，然后选择**运行配置文件**。
此时将显示**运行配置文件**对话框，该对话框会显示配置文件运行状态。

Informatica Developer 中的数据域发现结果

数据域发现结果显示与数据域匹配的列的统计信息，包括数据域匹配的遵从性条件以及列名是否与数据域匹配。

可以对结果执行向下钻取以供进一步分析。还可以验证数据源的所有行中的结果，并通过 Developer tool 中的编辑器将结果添加到数据模型中。可以根据数据域、数据域组和列对结果进行排序。可以将数据域发现结果导出到 Microsoft Excel 文件。

下表描述了数据域发现结果：

列名称	说明
名称	数据域、数据域组或列的名称基于您是选择 数据域 、 数据域组 还是 列视图 。
连接	连接的名称。
状态	列的推理状态。
数据遵从性百分比	数据域匹配所需的最低遵从行数百分比。
遵从行计数	数据域匹配所需的最低行数。
空值百分比	列的空值的百分比。
总行数	总行数。

列名称	说明
列名称匹配	用于指示列名称是否与数据域名称相匹配。
数据域组	数据域所属的数据域组。
已记录的数据类型	声明用于配置文件对象中的列的数据类型。
向下钻取	选择后，向下钻取行。
已验证	用于指示数据域匹配在数据源的所有行上是否已通过验证。
上次运行时间	上次运行配置文件的日期和时间。

按数据域组查看

可以查看按数据域组排序的数据域发现结果。

1. 运行配置文件以查看其结果。
2. 单击**结果**。
3. 单击**数据域发现**。
可以在右侧面板中查看数据域发现结果。
4. 验证**数据域**选项在**显示**字段中是否处于选中状态。
5. 选择**显示数据域组层次结构**以查看按数据域组排序的结果。

按列查看

可以查看按与数据域匹配的源列进行排序的数据域发现结果。

1. 运行配置文件以查看其结果。
2. 单击**结果**。
3. 单击**数据域发现**。
可以在右侧面板中查看数据域发现结果。
4. 选择**列**以查看按与数据域匹配的源列进行排序的结果。

确认结果

运行配置文件时，会分析数据源的示例以推理配置文件结果。可以在源数据的所有行上运行配置文件以验证推理结果。

1. 运行配置文件以查看其结果。
2. 单击**结果**。
3. 单击**数据域发现**。
可以在右侧面板中查看数据域发现结果。
4. 在右侧面板中选择您要验证的列。
5. 右键单击该列，然后选择**验证**以在数据源的所有行上运行配置文件。
验证结果后，您可能会看到**数据遵从性百分比值**或**遵从行数值**发生更改。

6. 要验证多个列的推理结果，请选择多个列。然后可以右键单击并选择**全部验证**。

批准数据域

如果在单个数据对象配置文件中运行数据域发现，则可以同时批准多个列的推理数据域。如果作为企业发现的一部分运行数据域发现，则可以一次批准一个源列的数据域。要在企业发现后批准多个列的数据域，可以打开各个数据对象配置文件任务并批准数据域。

1. 在**对象浏览器**视图中，选择一个配置文件。
2. 双击该配置文件将其打开。
配置文件将在选项卡中打开。
3. 如果已运行单个数据对象配置文件，则选择**数据域发现**视图，然后选择行。行包含每列的数据域发现结果。
4. 右键单击该行，然后选择**接受**。
此时，数据域的推理状态会更改为**已接受**。
5. 如果已运行企业发现，则选择**数据域**视图，然后选择数据域。
此时与该数据域匹配的列会显示在右侧面板中。
6. 右键单击您要批准的列，然后选择**接受**。还可以根据需要选择多个已拒绝的列并批准。
此时，数据域的推理状态会更改为**已接受**。
7. 要还原数据域的推理状态，请右键单击该行，然后单击**重置**。

拒绝数据域

默认情况下，Informatica Developer 会在配置文件结果中显示推理的数据域。您可以拒绝已推理或已批准的数据域。可以选择显示或隐藏拒绝的数据域。

1. 在**对象浏览器**视图中，选择一个配置文件。
2. 双击该配置文件将其打开。
配置文件将在选项卡中打开。
3. 在**数据域发现**视图或**数据域**视图中，选择行。
4. 要拒绝推理的数据域，请右键单击该行，然后选择**拒绝**。
此时，Informatica Developer 工具会将数据域发现结果中已拒绝的数据域变灰。
5. 要隐藏拒绝的数据域，请右键单击该行，然后选择**隐藏拒绝项**。
6. 要查看拒绝的数据域，请右键单击某一行，然后选择**显示拒绝项**。

从 Informatica Developer 中导出数据域发现结果

将数据域发现结果从 Informatica Developer 导出到 .xlsx 文件后，可以将文件保存到服务器或客户端计算机中的特定位置。

1. 运行配置文件以执行数据域发现。
2. 单击**结果**视图。
3. 单击**将结果导出到文件**图标。
此时会显示**将数据导出到文件**对话框。
4. 输入文件名称。或者，使用默认文件名。

5. 在**保存**下方，选择**在客户端上保存**，然后单击**浏览**以选择位置并将文件本地保存到您的计算机中。默认情况下，Informatica Developer 会将文件写入到 Informatica Administrator 的“数据集成服务”属性的服务器位置集中。
6. 单击**确定**。

第 25 章

Informatica Developer 中的企业发现

本章包括以下主题：

- [Informatica Developer 中的企业发现概览, 166](#)
- [企业发现进程, 167](#)
- [企业发现的配置文件选项, 167](#)
- [在 Informatica Developer 中创建企业发现配置文件, 171](#)
- [编辑配置文件, 172](#)
- [运行企业发现配置文件, 173](#)
- [外键发现, 173](#)
- [联接分析, 175](#)
- [重叠发现, 176](#)
- [DDL 脚本文件, 178](#)
- [同步企业发现配置文件, 178](#)

Informatica Developer 中的企业发现概览

企业发现是指在大量数据源中发现列配置文件统计信息、数据域、主键和外键的过程。您可以跨多个连接或架构执行企业发现。

作为企业数据分析师，您可能希望发现大量数据源的重要数据特征。可能需要识别关系数据资产，在发现的数据资产上运行列配置文件，发现企业、主键和候选键中的重要数据特征，等等。您可能还希望查看数据源之间存在的外键关系，以便根据发现的关系生成数据模型。

企业发现可发现您企业中信息资产存在的问题、模式、趋势和重要数据特征。您既可以选择导入模型存储库的数据源，也可以选择来自外部关系连接的数据源。数据发现进程包括发现列配置文件统计信息、数据域分析、包含候选键的数据对象结构和包含外键的数据对象关系。在 Developer tool 中运行企业发现，它会对每个数据源执行以下任务：

- 运行列配置文件。
- 发现数据域。
- 推理主键。

运行列配置文件、数据域发现和主键配置文件后，Developer tool 将在所有数据源上运行外键配置文件。Developer tool 完成剖析和发现任务后，将生成图形格式和表格格式的合并结果摘要。

您可以在 Informatica Developer 中选择操作系统配置文件。选择一个操作系统配置文件后，数据集成服务即会根据在操作系统配置文件中定义的操作系统用户的权限创建并运行企业发现配置文件。

企业发现进程

您可以在 Developer 工具中运行企业发现配置文件以执行企业发现。运行该配置文件之前，需要为不同的配置文件类型配置数据发现选项。

Developer 工具会为选定的数据源创建数据对象，并为每个数据对象创建配置文件任务。该工具然后会运行这些配置文件任务以生成配置文件结果。

要执行企业发现，请完成以下步骤：

1. 选择导入模型存储库的多个数据对象和跨多个外部关系连接的数据源，以创建企业发现配置文件。
2. 定义数据域发现、列配置文件、主键配置文件和外键配置文件的配置设置。
3. 运行企业发现配置文件。
4. 刷新模型存储库服务。

注意：将外部连接的元数据导入模型存储库时，需要执行此操作。您需要刷新模型存储库服务，以便 Developer 工具反映对模型存储库所做的更改。

5. 监视配置文件运行，并在需要时查看 Developer 工具运行的配置文件任务的状态。
6. 查看企业发现结果摘要。该摘要包含交互式图形用户界面视图和表格视图。

企业发现的配置文件选项

在运行配置文件以执行企业发现之前，请先设置配置文件选项。配置文件选项包括数据域发现选项、列配置文件采样选项以及主键和外键的推理选项。

您可以选择先运行企业发现配置文件，然后再设置配置文件选项。也可以选择在不运行配置文件的情况下先进行设置，然后再创建配置文件任务。

企业发现的数据域选择

推理选项决定了数据域发现是必须在列数据、列名称还是两者上运行。您可以指定配置文件是否需要处理数据源中的所有行，以及为数据域匹配选择一个遵从性条件。

下表介绍了为企业发现配置的数据域推理选项：

选项	说明
替代默认推理选项	更改预定义的推理选项。
数据	配置文件在列数据上运行。
列名称	配置文件在列标题上运行。

选项	说明
数据和列名称	配置文件同时在列数据和列标题上运行。
最低行数百分比	数据域匹配所需的数据集中最低遵从行数百分比。遵从性百分比是指匹配的行数除以总行数所得出的比率。 注意: Developer tool 将空值视为不匹配的行。
最小行数	数据域匹配所需的数据集中最低遵从行数。
从数据域发现中排除空值	从数据域发现的数据集中排除空值。
所有行	配置文件在数据源的所有行上运行。
先采样	配置文件可以在其中运行的最大行数。Developer tool 从源中的第一行开始选择行。
在后续运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型和数据域	在下次运行配置文件时，从数据类型和数据域推理中排除已批准的数据类型或数据域。

企业发现的列配置文件采样选项

采样选项决定了 Developer tool 是在数据源的所有行上运行列配置文件，还是在有限数量的行上运行列配置文件。

下表介绍了为企业发现配置的列配置文件采样选项：

选项	说明
所有行	对数据对象的所有行运行配置文件。 本地、Blaze 和 Spark 运行时环境中支持此选项。
对前 <数字> 行进行采样	对从数据对象第一行开始的采样行运行配置文件。最多可以选择 2,147,483,647 行。 本地和 Blaze 运行时环境中支持此选项。
限制 N <数字> 行	根据数据对象中的行数运行配置文件。选择在 Hadoop 验证环境中运行配置文件时，Spark 引擎会从数据对象的多个分区收集样本并将这些样本推送到单个节点来计算采样大小。“限制 n”采样选项支持 Oracle、SQL Server 和 DB2 数据库。不能对“限制 n”采样选项使用高级筛选器。 Spark 运行时环境中支持此选项。
随机百分比	对数据对象中某一百分比的行运行配置文件。 Spark 运行时环境中支持此选项。
排除含有已批准数据类型的列的数据类型推理	从列配置文件运行的数据类型推理中排除具有已批准数据类型的数据类型推理。

运行时环境选项

选择本地或 Hadoop 运行时环境选项。在 Hadoop 运行时环境中，可以选择 Blaze 或 Spark 选项。选择运行时环境后，Informatica Developer 会在配置文件定义中设置该运行时环境。运行时环境不会影响配置文件结果。

下表介绍了企业发现配置文件的运行时环境选项：

选项	说明
Native	Developer tool 会将配置文件作业提交给剖析服务模块。剖析服务模块随后将配置文件作业拆分成一组映射。数据集成服务会运行这些映射，并将配置文件结果写入配置文件仓库。
Blaze	数据集成服务将配置文件逻辑推送到 Hadoop 群集上的 Blaze 引擎来运行配置文件。
Spark	数据集成服务将配置文件逻辑推送到 Hadoop 群集上的 Spark 引擎来运行配置文件。

企业发现的主键推理选项

您可以替代企业发现的默认主键推理选项。这些选项包括可运行配置文件的最大行数和最低遵从性百分比等。

下表介绍了为企业发现配置的主键推理选项：

选项	说明
替代默认推理选项	允许您为主键推理配置自定义设置。
最大键列数	可构成主键的最大列数。
最大行数	可运行配置文件的最大行数。
最小百分比	主键匹配所要求的列数据最低遵从性百分比。
最大冲突行数	确定主键时配置文件允许的具有键冲突的最大行数。

企业发现的外键推理选项

设置外键推理选项可定义用于发现数据对象之间外键关系的列设置。外键推理结果取决于为企业发现、记录的主键和用户定义的主键设置的主键推理选项。

在 Informatica Developer 中，您可以通过以下方法之一推理外键：

- 使用默认值。
- 配置外键推理选项。
- 使用外键配置文件来配置自动内容管理参数。

下表介绍了为企业发现配置的外键推理选项：

选项	说明
替代默认推理选项	更改预定义的推理选项。
在比较中使用的数据类型	主键和外键比较中使用的数据类型。 注意: 如果在外键推理之前和数据源上运行列配置文件，则适用此选项。

选项	说明
比较区分大小写	在比较列数据时区分大小写。
比较前裁剪值	确定 Developer tool 在处理时是否包含列数据中的前导空格和尾随空格。
比较过程中使用的推理的主键 使用排列前 _ 的键	Developer tool 在所有数据源上运行外键配置文件时外键推理中使用的主键数量。Developer tool 使用顶级排列方法以及记录的主键和用户定义的主键来推理外键关系。 推理键的顶级排列以舍入为一位小数精度的按降序排序的遵从性百分比为基础。例如，Developer tool 将 99.75 的遵从性百分比视为 99.8，而将 99.74 视为 99.7。 默认值为 1。如果希望 Developer tool 在外键推理中使用所有的推理键，请将该值设置为 -1。 注意: 如果主键数据源已批准主键，则 Developer tool 将不会在外键推理中使用推理的主键。
数据对象间的最大外键数	配置文件运行后 Developer tool 返回的符合外键发现要求的最大推理列数。
最低遵从性百分比	要在外键结果中包含列的最小合格性值，以百分比表示。
重新生成签名	在源数据发生更改时重新加载列签名。

外键推理的自动内容管理参数

您可以配置自动内容管理参数，在不进行手动干预的情况下，推理主键和外键关系。自动内容管理参数属于用户定义的自定义属性，经配置后，可根据特定条件识别数据关系。

发现结果包括大量主键和外键关系时，您可能会发现，要想在数百种数据关系中找到重要的数据关系十分困难。您也可能会发现，根据特定条件（例如数据匹配或数据类型）管理这些关系也十分困难。要解决此问题，您可以配置自动内容管理参数并运行企业发现配置文件。

如果数据源具有多个候选外键，并且您想提供规则以选择候选外键，则可以执行以下操作：

- 在企业发现配置文件向导中配置**数据对象间的最大外键数**和**最低遵从性百分比**选项。
- 为 ForeignKeyConfig.xml 文件的自动内容管理参数配置权重和得分。

管理员可以编辑并保存外键配置文件。在外键配置文件中配置自动内容管理参数。算法根据自动内容管理参数，推理多个数据对象之间的主键和外键关系。

外键配置文件 ForeignKeyConfig.xml 位于以下目录中：

<Informatica 安装目录>\services\DataIntegrationService\modules\ProfilingService

自动内容管理参数是数据重叠匹配、列名匹配、关系类型匹配和数据类型匹配。

数据重叠匹配

数据重叠匹配是主键和外键之间值的预估重叠。您可以在企业发现配置文件向导中使用**最低遵从性百分比**选项设置重叠匹配。默认情况下，**最低遵从性百分比**选项设置为 90。

如果数据重叠匹配没有达到最低遵从性百分比，则不会为自动内容管理考虑外键。满足数据重叠匹配的最低遵从性后，剩余参数则会用于计算调整后的得分。

名称匹配

名称匹配参数是可选参数。它使用了“编辑距离”算法来确定主键列和外键列名称的匹配程度，并将得分设置在 0 和 1 之间。如果您不想使用此参数确定主键和外键关系，请将名称匹配权重设置为 0。

关系类型匹配

关系类型匹配会确定主键列和外键列之间的关系类型，并分配一个 0 和 1 之间的固定得分。关系类型匹配是根据外键列的列类型计算的。

以下关系类型匹配可以在 ForeignKeyConfig.xml 文件中设置：

- 外键列是非键列的主键-外键关系。此关系类型匹配的默认值是 1。您可以在许多数据源中找到此关系。
- 外键列是主键列的主键-主键关系。此关系类型的默认值是 0.25。您几乎找不到此关系类型，因为它代表已执行垂直分区的表。
- 外键列是主键列并且列的数据类型是序列数据类型的主键-主键序列关系。例如，Order 表中的 OrderID 列为序列数据类型。此关系类型的默认值是 0，因为序列键可能会导致主键-主键算法试图避免的多个误报外键。如果已知数据源包含若干序列数据类型，您可以将关系类型匹配设置为更高的得分。

数据类型匹配

数据类型匹配将主键列的数据类型与外键列的数据类型进行比较，并根据列的数据类型匹配程度，分配固定的遵从性得分。

下表列出了主键和外键不同组合的固定数据类型匹配得分：

	数字外键	日期外键	字符串外键
数字主键	1.0	0.5	0.0
日期主键	0.5	1.0	0.5
字符串主键	0.0	0.0	1.0

如果需要，您可以更改默认数据类型匹配得分。

在 Informatica Developer 中创建企业发现配置文件

您可以创建有关多个连接下多个数据源的配置文件。Developer 工具可为每个源创建单独的配置文件任务。

- 在 **Object Explorer** 视图中，选择要运行配置文件的多个数据对象。
- 单击 **文件 > 新建 > 配置文件** 以打开配置文件向导。
- 选择 **企业发现配置文件**，然后单击 **下一步**。
- 输入配置文件的名称并验证项目位置。如果需要，浏览到新位置。
- 验证所选数据对象的名称是否显示在 **数据对象** 部分。如果需要，单击 **选择** 以选择更多数据对象。
- 单击 **下一步**。
此时将显示 **将资源添加到配置文件定义** 窗格。可以从该窗格中选择多个外部关系连接和数据源。
- 单击 **选择** 打开 **选择资源** 对话框。

资源窗格将列出 Informatica 域下的所有内部和外部连接以及数据对象。

8. 单击**确定**关闭对话框。
9. 单击**下一步**。
10. 配置要运行的配置文件类型。可以配置以下配置文件类型：
 - 数据域发现
 - 列配置文件
 - 主键配置文件
 - 外键配置文件

注意: 为要作为企业发现配置文件的一部分运行的配置文件类型选择**已在“运行企业发现配置文件”操作中启用**。默认情况下将启用列剖析。
11. 查看配置文件的选项。

您可以编辑列配置文件的采样选项。还可以编辑数据域、主键和外键配置文件的推理选项。
12. 选择**创建配置文件**。

Developer 工具将为每个单独的数据源创建配置文件。
13. 完成配置文件配置后，选择**完成时运行企业发现配置文件**以运行配置文件。如果已启用所有剖析操作，则 Developer 工具将在所有选定的数据源上运行列、数据域和主键配置文件。然后，Developer 工具在所有数据源上运行外键配置文件。
14. 单击**完成**。

运行企业发现配置文件后，需要刷新模型存储库服务才能查看结果。将外部连接的元数据导入模型存储库时，需要执行此步骤。您需要刷新模型存储库服务，以便 Developer 工具反映对模型存储库所做的更改。

编辑配置文件

设置企业发现配置文件后，可以对其进行更改。您可以排除具有特定数据类型的列，还可以更改列选择、数据域选择和推理选项。

1. 在 **Object Explorer** 视图中，选择包含您要编辑的配置文件的项目或文件夹。
2. 单击**团队 > 签出**以签出配置文件。
3. 双击该配置文件将其打开。
4. 单击**属性**视图。

“属性”视图位于“默认”视图下方。
5. 单击**配置文件**以查看配置文件任务。
6. 在右侧窗格中选择要编辑的配置文件任务，然后单击**打开**。

配置文件定义会显示在选项卡中。
7. 要对企业发现配置文件的全局设置进行更改，请选择**剖析任务列表**顶部的配置文件，然后单击**配置**。
8. 对配置文件定义选项进行必要的更改。
9. 保存更改。
10. 单击**团队 > 签入**以签入配置文件。

运行企业发现配置文件

您可以通过多种方式运行企业发现配置文件。可以从 **Object Explorer** 视图运行配置文件，也可以从**属性**窗口中的**配置文件**选项卡运行配置文件。可以选择运行属于企业发现配置文件的单个和多个配置文件任务。

1. 在 **Object Explorer** 视图中，选择包含您要运行的配置文件的项目或文件夹。
要自动运行配置文件，请在创建该配置文件时在**新建企业发现**向导中选择**完成时运行企业发现配置文件**。
2. 双击该配置文件将其打开。
配置文件将在选项卡中打开。
3. 在 **Object Explorer** 视图中，右键单击该配置文件并选择**运行企业发现配置文件**。
或者，也可以在**属性配置文件**，在**剖析任务列表**下选择该配置文件名称，然后单击**运行**。
注意: 运行企业发现配置文件后，需要刷新模型存储库服务才能查看结果。将外部连接的元数据导入模型存储库时，需要执行此步骤。您需要刷新模型存储库服务，以便 Developer 工具反映对模型存储库所做的更改。
4. 此时将显示**运行**对话框。可以在该对话框中对配置文件的全局设置做出更改。
默认情况下，所做的更改将应用于企业发现配置文件中新添加的数据对象。
5. 要对企业发现配置文件中的所有数据对象配置文件任务和系统生成的外键配置文件任务应用更改，请选择**对当前配置文件使用全局设置**。
Developer 工具将根据更改后的设置来更新所有数据对象配置文件任务和外键配置文件任务。
6. 要运行单个配置文件任务，请选择一个任务并单击**运行**。
7. 要运行多个配置文件任务，请单击**运行多个**。
此时将显示**运行多个**对话框。
提示: 如果加载企业发现结果需要花费的时间较长，您可能希望更新剖析仓库数据库的统计信息。运行多个企业发现配置文件可能会导致数据量和列值发生显著更改。更新统计信息后，数据库将根据最新的统计信息运行 SQL 查询的执行计划，并优化数据库操作。
8. 默认情况下将选择所有任务。清除不希望运行的任务，然后单击**确定**。

外键发现

如果一个列的数据值与另一数据对象中的主键列值相匹配，则该列就是一个外键。

您可以在 Developer tool 中对多个数据对象执行外键发现。创建企业发现配置文件可选择数据对象并定义该配置文件。

执行外键发现之前，必须在企业发现配置文件中标识父数据对象和子数据对象。该配置文件使用父对象中的一个或多个键（包括其主键）来发现子对象中的外键。定义完父对象和子对象并标识了父对象中的键后，创建并运行配置文件。

定义父对象和子对象关系

要查找两个数据对象之间的外键关系，必须选择一个父数据对象，然后指定该对象中的主键。

1. 打开包含要分析的数据对象的企业发现配置文件。
2. 选择父对象。

3. 选择父对象中的主键：
- 单击**属性**选项卡，然后单击**键**。
 - 单击**添加**，然后在“新建键”对话框中选择主键列。
 - 在**新建键**对话框中单击**确定**。验证该主键是否在**选定字段**窗格中显示以及是否选中**主键**选项。

创建外键配置文件以分析外键的子对象。

发现数据对象之间的外键关系

在 Developer 工具使用企业发现配置文件来查找两个数据对象之间的键关系

包含主键的数据对象为父对象，而包含外键的数据对象为子对象。

1. 打开包含要分析的数据对象的企业发现配置文件。
2. 右键单击数据对象的名称并选择**外键配置文件**。
3. 输入配置文件的名称并验证项目位置。如果需要，浏览到新位置。或者，输入配置文件的文本说明。
4. 选择配置文件将用于查找子对象中外键的父对象中的键。
5. 保存并运行该配置文件。

外键分析结果

运行外键配置文件后，单击建模编辑器下方的配置文件名称可查看分析结果。

结果视图列出了符合定义的主键到外键推理条件的列。单击**选项**按钮可编辑推理设置。单击列名称并选择**验证**可验证推理键是否为数据对象的有效键。

下表介绍了外键分析属性：

属性	说明
父主键	配置文件用于查找子对象中外键的父数据对象中的主键列。
子外键	配置文件推理为当前行上父主键的对应外键的列。
包含百分比	主键与外键之间匹配的数据值的数量，以百分比表示。 注意: 对外键结果中的推理列进行验证后，您可能会看到该推理列的“包含百分比”值有所变化。对于推理列，“包含百分比”是与父对象的唯一主键列值相匹配的子对象的唯一外键列值的数量。对推理列进行验证后，它将是与父对象的主键列值相匹配的子对象的外键列值的数量。
关系类型	配置文件运行之前为主键列和外键列定义的关系类型。 如果在配置文件运行之前定义关系，则即使包含百分比数字不满足为该配置文件设置的置信度阈值，该配置文件也会返回关系数据。
已验证	指示用户已验证主键到外键关系。
上次运行时间	配置文件上次运行的日期和时间。
关系类型(模型中)	指示配置文件已验证列之间的关系。

联接分析

联接分析介绍了两个数据列之间的潜在联接度。使用联接配置文件可分析单个数据源或多个数据源中的列联接。

联接配置文件将结果显示为韦恩图以及用数量和百分比表示的值。可从企业发现配置文件创建和运行联接配置文件。

创建联接配置文件

您可以分析企业发现配置文件中数据对象之间的潜在联接。联接配置文件将分析存储在模型存储库中。

1. 创建或打开企业发现配置文件。
2. 验证该企业发现配置文件是否包含需要的数据对象。
要向联接配置文件添加数据对象，请将其从 **Object Explorer** 视图拖动到建模编辑器中。
3. 选择要剖析的数据对象。
4. 右键单击这些对象并选择**联接配置文件**。
此时将打开配置文件向导。
5. 输入配置文件的名称。或者，输入配置文件的文本说明。
6. 验证数据对象的名称是否显示在向导中的**数据对象**下方。
7. 选择或清除相应选项以便**完成时运行配置文件**。
8. 单击**下一步**。
9. 选择要在配置文件中包含的数据列，然后单击**下一步**。
如果需要，向下滚动数据对象以查看所有可用列。默认情况下，配置文件在所有列上运行。
10. 单击**添加**。
此时将显示**联接条件**对话框。
11. 单击**新建**以激活列选择字段。
12. 选择要验证的数据对象和列。
定义两个列之间的联接条件。可以在一个或多个数据对象上定义多个联接条件。
13. 单击**确定**创建联接条件。
或者，单击**添加**以定义其他条件。
14. 验证左侧和右侧联接列是否将正确的数据对象名称作为其前缀。
15. 单击**完成**。

联接分析结果

联接分析**结果**选项卡提供了有关父孤行、子孤行和联接行的数量和百分比信息。联接分析结果还包含显示列之间关系的韦恩图。

下表介绍了**结果**选项卡上显示的属性：

属性	说明
左表	左表名称和联接分析中使用的列。
右表	右表名称和联接分析中使用的列。

属性	说明
仅左侧行	左表中无法联接的行数。
仅右侧行	右表中无法联接的行数。
联接行	联接中包含的行数。

选择联接条件可查看显示列之间关系的韦恩图。韦恩图下方的区域还显示列中孤立的值、空值和联接的值的数量和百分比。

双击韦恩图中的区域可查看该区域表示的记录。这些记录在“数据查看器”视图中打开。

注意: 可以将“数据查看器”视图中的记录列表导出到平面文件中。

将联接配置文件结果导出到文件

您可以将针对联接条件返回的数据行导出到带分隔符的文件中。导出左侧与右侧源之间重叠的行或单个源中的孤行。

1. 在 **Object Explorer** 视图中，打开包含联接分析的企业发现配置文件。
2. 运行该联接配置文件。
3. 选择**联接结果**视图。
4. 在**数据查看器**选项卡上，单击**将向下钻取结果导出到文件**图标。
此时将显示**导出数据**对话框。
5. 输入文件名称，然后单击**保存**。

重叠发现

重叠发现提供有关单个数据源或多个数据源内列对中的重叠数据的信息。您可以通过企业发现配置文件查找重叠数据。可以验证配置文件结果并在韦恩图中查看这些结果。

重叠发现根据默认设置或者指定的设置来标识重叠数据。可以替代默认设置并指定推理选项，包括重叠发现根据重叠百分比返回的前几个列对的最大数量。还可以指定用于确定重叠发现合格性的置信度级别。

重叠发现结果

重叠发现选项卡显示有关参与列和重叠百分比值的信息。重叠发现结果包含韦恩图，该图显示列对中的重叠数据以及上次执行重叠发现的日期和时间。

单击列并选择**验证**可以在韦恩图中查看结果。

下表介绍了重叠发现属性：

属性	说明
左侧列	与其余列进行比较以进行重叠分析的主列。
右侧列	与主列进行比较的列。

属性	说明
重叠百分比	两个列之间重叠的百分比。
已验证	指示已验证重叠结果行。
上次运行时间	上次运行重叠发现的日期和时间。

Informatica Developer 在重叠发现结果中每个重叠对显示两次。以数据源 Items 和 Orders 为例。Items 包含列“m”和“n”。Orders 包含列“p”和“q”。

下表显示了 Items 和 Orders 的重叠发现结果：

左侧列	右侧列
Items	-
m	Orders.p
m	Orders.q
n	Orders.p
n	Orders.q
Orders	-
p	Items.m
p	Items.n
q	Items.m
q	Items.m

发现重叠数据

您可以在企业发现配置文件中确定列对之间重叠的数据。重叠分析基于列中的唯一值，不考虑空值。

1. 创建或打开包含数据对象的企业发现配置文件。
2. 选择要查找重叠数据的数据对象。
您可以选择单个数据对象以查找列对中的重叠数据，也可以选择多个数据对象。
3. 右键单击这些对象并选择**重叠发现**。
此时将显示**新建重叠发现**对话框。
4. 输入名称。
5. 或者，输入重叠分析的文本说明。
6. 验证数据对象的名称是否显示在向导中的**数据对象**下方。
7. 或者，在完成配置设置时选择**完成时运行配置文件**以运行配置文件。
8. 单击**下一步**。
9. 选择要进行重叠发现的列。

10. 单击**下一步**。
对话框中将显示默认的推理选项。
11. 或者，指定重叠发现的推理选项以替代默认设置。
12. 单击**完成**。

DDL 脚本文件

数据定义语言 (DDL) 脚本文件包含 Create、Alter 和 Drop SQL 语句。

生成脚本文件时，可以指定文件名、位置和目标数据库类型。Developer 工具可将 “_create” 和 “_drop” 标签附加到脚本文件名。虚拟列不属于 DDL 脚本文件。

从企业发现配置文件创建 DDL 脚本

从企业发现配置文件生成 DDL 脚本文件时，可以选择要保存这些脚本文件的位置。还可以选择要在其中运行脚本的数据库类型。确保在企业发现配置文件中验证并提交所有必要的更改，然后再生成 DDL 脚本。

1. 在 **Object Explorer** 视图中，选择一个企业发现配置文件。
2. 右键单击该配置文件并选择**生成 DDL**。
此时将显示**生成 DDL** 对话框。
3. 单击**浏览**以打开**另存为**对话框。
默认的文件扩展名为 .sql。
4. 选择文件位置，然后输入文件名。
5. 选择目标数据库类型。
6. 单击**确定**。

Developer 工具将在指定位置生成 DDL 脚本文件。

同步企业发现配置文件

可以在 Developer tool 中同步企业发现配置文件。

从版本 9.5 或更低版本升级到版本 9.6 或更高版本后，您可以将以前版本中的配置文件迁移到升级后的版本。对于企业发现配置文件，如果您添加了以前版本中的任何用户定义的键、已记录的键或关系，这些键和关系信息只会保留在模型存储库中，而不会保留在剖析仓库中。在升级后的版本中，当您在 Developer tool 中打开企业发现配置文件时，已记录的键或用户定义的键和关系不会显示在配置文件的特选结果中。

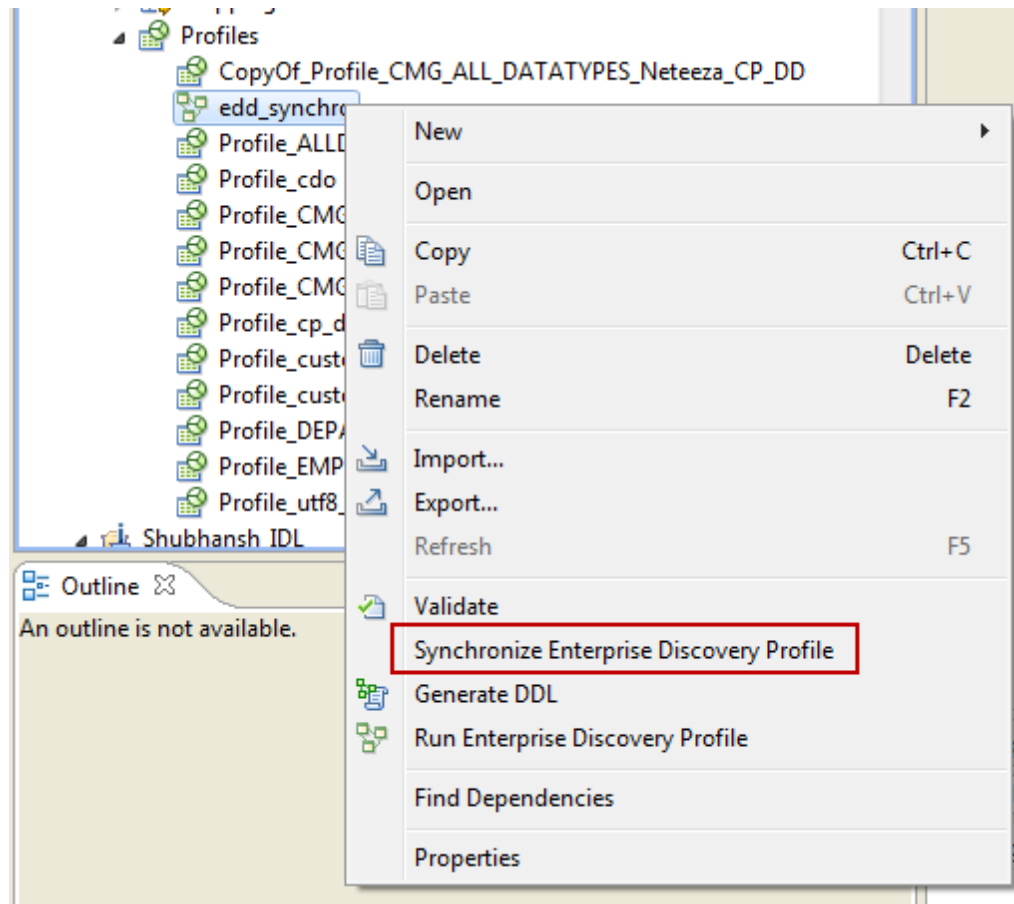
要将模型存储库中的用户定义的键、已记录的键和关系与剖析仓库同步，请使用 Developer tool 中的“同步企业发现配置文件”选项。同步企业发现配置文件后，用户定义的键和已记录的键以及关系会设置为“批准”，您可以在 Developer tool 中查看特选结果。

同步企业发现配置文件

在 Informatica Developer 中，从版本 9.5 或更低版本升级到版本 9.6 或更高版本后，您可以同步企业发现配置文件的特选结果。

1. 在 **Object Explorer** 视图中，选择一个企业发现配置文件。
2. 右键单击该配置文件，然后选择**同步企业发现配置文件**。

下图显示了 Developer tool 中的“同步企业发现配置文件”选项：



将同步配置文件的特选结果。

第 26 章

企业发现结果

本章包括以下主题：

- [企业发现结果概览, 180](#)
- [关系视图, 181](#)
- [外键剖析视图, 182](#)
- [表格视图, 184](#)
- [数据域视图, 185](#)
- [列配置文件视图, 187](#)
- [在企业发现运行期间查看列配置文件结果, 187](#)
- [在企业发现运行期间查看数据域发现结果, 187](#)
- [查看企业发现的运行时状态, 188](#)
- [企业发现导出文件, 188](#)

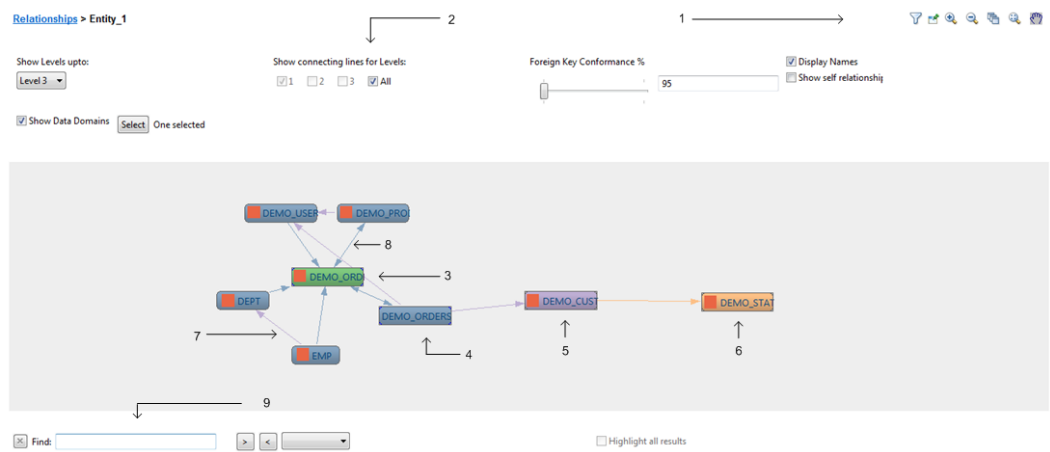
企业发现结果概览

您可以在多个视图中查看企业发现的结果。这些视图包括**关系**、**数据域**、**列配置文件**、**联接配置文件**和**重叠发现**。

关系视图将数据对象组显示为圆圈。可以从该视图启动外键配置文件结果。可以在图形视图和表格视图中查看外键配置文件结果。**数据域**视图显示数据域发现结果。**列配置文件**视图显示每个数据对象的列配置文件结果。**联接配置文件**视图显示父孤行数、子孤行数和联接中包含的行数。**重叠发现**视图提供有关参与列和重叠百分比值的信息。

数据对象之间可以存在多个关系。外键结果的图形视图显示具有最高遵从性百分比的数据对象关系。

下图显示了企业发现结果中部分示例数据对象的图形视图：



1. 工具栏图标，包括筛选、锁定数据对象、放大、缩小、全部排列、适合窗口和平移图标。
2. 筛选选项，如显示不同级别的数据对象关系、显示连接线和显示自相关数据对象。
3. 选定的数据对象，可视化编辑器根据其显示数据对象之间的其余关系。
4. 第一级数据对象关系。
5. 第二级数据对象关系。
6. 第三级数据对象关系。
7. 数据对象之间的连接器。单向箭头指示两个数据对象之间存在主键到外键关系。箭头指向具有主键的数据对象。
8. 数据对象之间的连接器。双向箭头连接器指示两个数据对象之间存在主键到主键关系。将鼠标悬停在连接器上方可查看推理关系具有最大遵从性的联接列。
9. 按 CTRL+F 显示“查找”字段并使用星号 (*) 作为通配符来查找图形视图中的数据对象。

关系视图

您可以在**关系**视图中查看企业发现结果摘要，包括实体。实体是用圆圈表示的数据对象组。实体包含源数据库的多个连接和架构中的相关数据对象、自相关数据对象和不相关的数据对象。

自相关数据对象在数据对象中具有存在关系的列。不相关的数据对象既不与源数据库中的其他数据对象存在关系，在数据对象中的列之间也不存在关系。企业发现结果中数据对象的实体关系图基于推理关系，而不是数据源中记录的关系。

搜索数据对象

您可以在**关系**视图或**外键剖析**视图中搜索数据对象。可以使用星号 (*) 作为通配符来查找数据对象。

1. 验证您是在**关系**还是**外键剖析**视图中。
2. 键入要搜索的数据对象名称的一部分，并根据搜索要求将 * 通配符添加到搜索字符串的开头或结尾。例如，要搜索以字符串“CA”开头的所有数据对象，键入“CA*”并按 **Enter** 键。要搜索名称中包含字符串“ZIP”的所有数据对象，键入“*ZIP*”。

搜索区分大小写。

导航到外键剖析视图

外键剖析视图显示运行配置文件的多个数据对象之间外键关系的统一视图。该视图中的圆圈代表实体、自相关对象和未引用的对象。

1. 验证您是否位于**关系**选项卡中。
可以在右侧窗格中查看外键配置文件链接。
2. 单击**外键配置文件**以打开视图。
此时将在新的选项卡中显示视图。视图根据数据对象的关系类型在不同的圆圈集中显示数据对象。还可以显示统一外键视图中数据对象的总数。
3. 或者，可以单击**关系**链接以返回**关系**视图。

外键剖析视图

您可以在**外键剖析**视图中以图形格式查看企业发现结果摘要。可以从该视图以表格格式打开数据对象的配置文件结果和列级别关系。

数据对象的配置文件结果包括列配置文件、主键推理、功能相关性推理以及数据域发现结果。打开数据对象的列级别关系后，可以验证和管理这些数据关系。验证数据关系时，Developer 工具将在源数据的所有行上运行配置文件以验证推理结果。可以在**外键剖析**视图中批准、拒绝和重置数据关系。

注意：当您使用 Hive 数据源来创建企业发现配置文件时，**外键剖析**视图不会显示任何数据对象。

查看数据对象关系

您可以以图形格式查看数据对象之间的关系。双击实体圆圈可在其中查看表及其关系。

1. 验证您是否在**外键剖析**视图中。
2. 要在统一图形视图中包含数据域，请选择**显示数据域**。
此时将启用**选择**按钮。
3. 单击**选择**以选择要在图形视图中包含的数据域。
此时将显示**选择数据域**对话框。
4. 选择所需的数据域，然后单击**确定**。
Developer 工具会突出显示包含选定数据域的实体圆圈。
5. 双击实体圆圈可查看该实体中表关系的可视化表示。Developer 工具以图形格式显示表，这些表代表每个数据对象与该实体中其他数据对象的关系。
与其他数据对象的关系数量最多的数据对象或从其开始导航的数据对象将突出显示为绿色。如果包含了数据域，则 Developer 工具将在每个数据对象的可视化表示的左侧突出显示数据域选择。
6. 在**外键剖析**视图的右侧窗格中验证直接关系信息和数据域信息。
7. 或者，可以单击**关系**链接以返回**关系**视图。

放大和缩小视图

您可以在**外键剖析**视图中放大数据对象关系的图形表示，以便更清楚地查看。放大时，Developer 工具会增加图像的放大级别。缩小可减小放大级别。

1. 验证您是否在**外键剖析**视图中。

2. 右键单击该视图并选择**放大**可增加图像的放大级别。
3. 要减小图形布局的放大级别，请右键单击该视图并选择**缩小**。

查找数据对象

您可以在外键结果的图形视图中搜索和查找数据对象。可以使用星号 (*) 作为通配符来查找数据对象。

1. 验证您是否在**外键剖析**视图中。
确保在该视图中打开外键结果的图形视图。
2. 按 Ctrl+F 以显示**查找**字段。
3. 在**查找**字段中，输入要搜索的数据对象名称的一部分，并根据搜索要求将 * 通配符添加到搜索字符串的开头或结尾。例如，要搜索以字符串“EMP”开头的所有数据对象，键入“EMP*”并按 **Enter** 键。要搜索名称中包含字符串“ZIP”的所有数据对象，键入“*ZIP*”。
4. 单击**下一个匹配项**按钮以移至下一个数据对象匹配项。
使用**上一个匹配项**按钮可移至上一个数据对象匹配项。
5. 选择**突出显示所有结果**可突出显示所有数据对象匹配项。
6. 要清除**查找**字段中的搜索字符串，可单击该字段旁边的**清除**按钮。

查看列关系

您可以查看数据对象中的每个列与相关数据对象中列的关系。还可以验证这些数据对象关系并将其提交至数据模型。

1. 验证您是否在**外键剖析**视图中。
2. 右键单击数据对象并选择**查看列关系**。
列关系将显示在表格视图中。该视图显示关系信息，如源数据对象、相关数据对象以及相关数据对象中的列。
3. 验证推理状态、验证状态和内容管理状态。
4. 选择**组中的所有数据对象**以查看父实体中的所有数据对象及其列关系信息。
默认情况下，该视图显示视图中选定数据对象的关系信息。
5. 或者，也可以单击视图顶部的**实体**链接以返回数据对象的图形表示。

将实体关系图另存为图像

您可以将企业发现结果中数据对象的实体关系图另存为“.png”文件。

1. 运行配置文件以执行企业发现。
2. 验证您是否在**外键剖析**视图中。
3. 从该视图切换到图形格式的数据对象关系。
4. 右键单击并选择**另存为图像**。
此时将显示**另存为**对话框。默认情况下，图像将另存为“.png”文件。
5. 选择文件位置，然后输入文件名。
6. 单击**保存**。

从“外键剖析”视图查看数据对象配置文件结果

您可以从**外键剖析**视图查看选定数据对象的列配置文件、主键和数据域发现结果。确保将数据对象锁定为画布中的选定表，以便选中该数据对象。

1. 验证您是否在**外键剖析**视图中。
2. 右键单击数据对象并选择**将数据对象锁定为焦点**以选中该表。
或者，也可以使用**锁定**图标来选择数据对象。
3. 右键单击画布中的任意位置并选择**查看数据对象配置文件**。
此时将在选项卡中显示数据对象配置文件结果。

表格视图

打开**外键剖析**视图时，Developer 工具会默认显示结果的图形视图。切换到表格视图可以以表格格式显示表及其关系详细信息。

您可以查看实体中的数据对象数量、相关表的名称、其连接信息以及两个数据对象之间的关系数量。还可以验证列关系并将其添加到数据模型。

表详细信息窗格

您可以在企业发现结果的图形视图和表格视图中查看数据对象详细信息。在图形视图中，表详细信息窗格显示与选定数据对象存在直接关系的数据对象数量和对象名称。

下表介绍了表格视图中表详细信息窗格的列：

列名称	描述
表名称	与左侧窗格中选定的数据对象存在直接关系的数据对象的名称。
连接	相关数据对象的连接的名称。
关系	左侧窗格中选定的数据对象与表详细信息窗格中的相关数据对象之间的关系数量。

验证企业发现结果

验证企业发现的结果时，Developer 工具将在数据源的所有行上运行配置文件。验证后，遵从性百分比值可能会有所不同，具体取决于数据源中所有行上的列值。

1. 运行配置文件，然后将其打开。
2. 验证您是否在**外键剖析**视图中。
3. 单击该视图顶部的**表格视图**图标。
表格视图在左侧窗格中显示实体。
4. 右键单击左侧窗格中的数据对象并选择**查看列关系**。
您可以查看选定数据对象中的列与其他数据对象中的列之间的关系。向右滚动可查看关系类型、遵从性百分比、验证状态和提交状态等详细信息。
5. 右键单击一个行并选择**验证**。
此时将显示**运行配置文件**对话框。验证完成后，选择该行可在韦恩图中查看主键和外键关系的重叠。

管理列关系

可以在**外键剖析**视图中批准、拒绝和重置数据关系。

1. 运行配置文件，然后将其打开。
2. 验证您是否在**外键剖析**视图中。
3. 在用于拒绝推理的列关系的图形视图中，选择数据对象，右键单击数据对象，然后选择以下选项之一：
 - **拒绝所有关系 > 已推理的主键。** 选择此选项可拒绝数据对象中具有推理主键的列与其他相连数据对象中具有推理外键的列之间的所有关系。
 - **拒绝所有关系 > 已推理的外键。** 选择此选项可拒绝数据对象中具有推理外键的列与其他相连数据对象中具有推理主键的列之间的所有关系。
 - **拒绝所有关系 > 已推理的主键和外键。** 选择此选项可拒绝数据对象中具有推理主键的列与其他相连数据对象中具有推理外键的列之间的所有关系，以及数据对象中具有推理外键的列与其他相连数据对象中具有推理主键的列之间的所有关系。
4. 在图形视图中，右键单击数据对象并选择**查看列关系**。
5. 选择要管理的数据对象关系。
6. 要批准该列关系，请右键单击并单击**批准**。
相应行的状态将更改为**已批准**。
7. 要还原列关系的已推理状态，请右键单击并单击**重置**。
8. 要查看已拒绝的列关系，请右键单击其中一行，然后选择**显示拒绝项**。
9. 要隐藏已拒绝的数据类型，请右键单击其中一行，然后选择**隐藏拒绝项**。

将结果提交至模型存储库

运行配置文件后，可以将数据对象之间的列关系保存到模型存储库。可以从**外键剖析**视图的表格视图将这些关系提交至模型存储库。

1. 运行配置文件，然后将其打开。
2. 验证您是否在**外键剖析**视图中。
3. 右键单击左侧窗格中的数据对象并选择**查看列关系**。
您可以查看选定数据对象中的列与其他数据对象中的列之间的关系。
4. 右键单击一个行并选择**批准**。

数据域视图

数据域视图列出了 Developer 工具在企业发现过程中发现的数据域和匹配列统计信息。您可以通过**数据域**视图验证列、对行进行向下钻取和查看数据对象配置文件结果。

查看数据域发现结果

您可以在**数据域**选项卡上查看数据域发现结果。可以搜索数据域并查看按数据域组排序的数据域。

1. 运行配置文件以执行企业发现。
2. 打开配置文件。

3. 单击**数据域**选项卡以查看数据域发现结果。
此时将在右侧窗格中显示数据对象配置文件结果。
4. 在搜索字段中键入数据域名称的一部分以查找特定的数据域。
选择在**层次结构中显示数据域组**可按照数据域组对数据域列表进行排序。

验证数据域发现结果

运行配置文件时，会分析数据源的示例以推理配置文件结果。可以在源数据的所有行上运行配置文件以验证推理结果。

1. 运行配置文件，然后将其打开。
2. 单击**数据域**选项卡以查看结果。
可以在右侧面板中查看数据域发现结果。
3. 在右侧面板中选择您要验证的列。
4. 右键单击该列，然后选择**验证**以在数据源的所有行上运行配置文件。
验证结果后，您可能会看到**数据遵从性百分比值**或**遵从行数值**发生更改。

对行进行向下钻取

您可以对数据域发现结果进行向下钻取以供进一步的数据分析。

1. 运行配置文件，然后将其打开。
2. 单击**数据域**选项卡以查看结果。
可以在右侧面板中查看数据域发现结果。
3. 在右侧面板中选择要向下钻取的行。
4. 右键单击相应的列并选择**向下钻取**以向下钻取到源行。

在数据域视图中查看数据对象配置文件结果

您可以从**数据域**视图查看选定数据对象的数据对象配置文件结果。

1. 验证您是否在**数据域**视图中。
2. 在**已剖析域**窗格中选择数据域。
3. 在右侧的**列**窗格中，选择一个列。
4. 右键单击该列并选择**打开数据对象配置文件**。
此时将在选项卡中显示数据对象配置文件结果。

列配置文件视图

列配置文件视图显示 Developer 工具在企业发现过程中运行的单个数据对象配置文件的列配置文件结果摘要。您可以查看列统计信息，例如，数据对象中每个列的唯一值、空值、数据类型以及最大值和最小值。

查看数据对象配置文件结果

企业发现包括运行数据对象配置文件以发现列数据统计信息、主键、候选键和数据域。您可以从**列配置文件**视图查看选定数据对象的数据对象配置文件结果。

1. 验证您是否在**列配置文件**视图中。
2. 在**已剖析的数据对象**窗格中选择数据对象。
3. 在右侧的**列**窗格中，选择一个列。
4. 右键单击该列并选择**查看数据对象配置文件**。
此时将在选项卡中显示数据对象配置文件结果。默认情况下将显示列配置文件结果。
5. 单击**主键推理**可查看主键配置文件结果。
6. 单击**功能相关性推理**可查看功能相关性发现结果。
7. 单击**数据域发现**可查看数据域发现结果。

在企业发现运行期间查看列配置文件结果

完成企业发现所需的时间取决于配置文件任务的数量、数据源大小和配置文件类型。当 Developer 工具继续运行数据发现任务时，您可以查看它在数据发现的初始阶段完成的列配置文件结果。

1. 运行配置文件后，在**属性**窗口中单击**配置文件**。
2. 选择要查看结果的列配置文件。确保**属性**窗口中配置文件运行的状态为**成功**。
3. 单击**打开**可在其他选项卡中查看结果。
4. 在**结果**部分，选择**列剖析**可在右侧窗格中查看结果。

在企业发现运行期间查看数据域发现结果

当 Developer 工具继续运行企业发现中包含的数据发现任务时，您可以查看 Developer 工具在企业发现的初始阶段完成的数据域发现结果。

1. 开始运行配置文件后，在**属性**窗口中单击**配置文件**。
2. 选择要查看数据域结果的配置文件。确保**属性**窗格中配置文件运行的状态为**成功**。
3. 单击**打开**可在其他选项卡中查看结果。
4. 在**结果**部分，选择**数据域发现剖析**可在右侧窗格中查看结果。

查看企业发现的运行时状态

Developer 工具的**进度**视图显示操作（如配置文件运行）的进度。您可以从**进度**视图查看企业发现任务的运行时状态。

1. 运行配置文件以对数据源执行企业发现后，单击 Developer 工具右下角的**进度视图**按钮。
如果**进度**窗格尚未打开，则此时将显示。
2. 单击**企业发现正在运行:查看任务状态**链接可打开子任务对话框。
该对话框列出了企业发现过程中的多个配置文件任务。您可以查看配置文件名称、类型及其状态。
3. 单击列标题可对配置文件任务进行排序。例如，要按照其状态对配置文件任务进行排序，可单击**状态**列标题。
4. 如果需要取消特定的配置文件任务，可选择该任务并单击**取消**。
所取消任务的状态将更改为**已终止**。

企业发现导出文件

运行企业发现配置文件后，可以导出所有的数据对象关系、数据域和单个外键任务结果等信息。您可以将数据对象关系的图形图像另存为 .jpg 文件。

导出配置文件结果时，Developer 工具会将所有的企业发现结果保存在多个 Microsoft Excel 文件中。可以在单独的文件中查看数据对象关系、列配置文件结果、数据域发现结果、实体和单个外键任务结果。

导出企业发现结果

您可以导出实体列表和每个实体的组成、单个实体的所有数据对象和列级别数据对象关系、数据域和列剖析结果。

1. 运行配置文件以执行企业发现。
2. 在**关系、数据域或列配置文件**视图中，单击窗口右上角区域的**导出**图标。
此时会显示**将数据导出到文件**对话框。
3. 输入文件名称。或者，使用默认文件名。
4. 在**保存**下方，选择**在客户端上保存**，然后单击**浏览**以选择位置并将文件本地保存到您的计算机中。默认情况下，Informatica Developer 会将文件写入在 Informatica Administrator 的数据集成服务属性中设置的位置。
5. 单击**确定**。

第 27 章

Informatica Developer 中的 Business Glossary 桌面版

本章包括以下主题：

- [Business Glossary 搜索, 189](#)
- [查找业务术语, 189](#)
- [自定义热键以查找业务术语, 190](#)

Business Glossary 搜索

在 Business Glossary 桌面版中查找某个 Developer tool 对象名称作为业务术语的含义，以了解其业务要求和当前实施。

业务词汇表是一系列使用业务语言为企业用户定义概念的术语。业务术语提供了概念的业务定义和用法。Business Glossary 桌面版是连接到托管业务词汇表的 Metadata Manager 服务的客户端。使用 Business Glossary 桌面版在业务词汇表中查找业务术语。

如果 Business Glossary 桌面版安装在计算机上，您可以在开发程序工具中选择某个对象，然后使用热键或搜索菜单在业务词汇表中查找对象的名称。您可以在 Developer tool 的视图（如**对象浏览器**视图）中查找对象名称，或者在编辑器中查找列、配置文件和转换端口的名称。

例如，开发人员想要在业务词汇表中查找与开发程序工具中的 Sales_Audit 数据对象对应的业务术语。开发人员想要查看业务术语详细信息以了解开发程序工具中的 Sales_Audit 对象的业务要求和当前实施。这可以帮助开发人员理解该数据对象的含义以及可能需要对该对象实施哪些更改。

查找业务术语

在 Business Glossary 桌面版中查找作为业务术语的某个 Developer 工具对象名称，以了解其业务要求和当前实施。

您必须在计算机上安装 Business Glossary 桌面版。

1. 选择对象。
2. 选择以使用热键或搜索菜单打开 Business Glossary 桌面版。

- 要使用热键，请使用以下热键组合：

CTRL+Shift+F

- 要使用搜索菜单，请单击**搜索 > Business Glossary**。

此时将显示 **Business Glossary 桌面版**，其中会显示匹配对象名称的业务术语。

自定义热键以查找业务术语

自定义热键可更改打开 Business Glossary 桌面版的键组合。

1. 从开发程序工具菜单中，单击**窗口 > 首选项 > 常规 > 键**。
2. 要在命令列表中查找或搜索 **Search Business Glossary**，请选择以下选项之一：
 - 要搜索键，请在搜索框中输入 Search Business Glossary。
 - 要滚动以找到键，请滚动到**命令**列下的 **Search Business Glossary** 命令。
3. 单击 **Search Business Glossary 命令**。
4. 单击**解除命令绑定**。
5. 在**绑定**字段中，输入键组合。
6. 单击**应用**，然后单击**确定**。

附录 A

基于剖析仓库连接的功能支持

本附录包括以下主题：

- [剖析功能支持, 191](#)

剖析功能支持

您可以通过 JDBC 或本地连接来连接到剖析仓库。您可以根据剖析仓库连接在剖析中执行特定的功能。

下表列出了根据在 Data Engineering Quality 和 Data Engineering Integration 中选择的剖析仓库连接类型，可执行的功能：

功能	JDBC 连接	本地连接
单个数据对象列配置文件	支持	支持
单个数据对象数据域发现	支持	支持
带有主键发现的单个数据对象配置文件	不受支持	不受支持
带有功能相关性发现的单个数据对象配置文件	不受支持	不受支持
采用先采样 <number> 行采样的配置文件	支持	支持
采用随机采样 <number> 行采样的配置文件	支持	支持
结果卡度量和值成本	不受支持	支持
带有主键和外键发现的企业发现配置文件	不受支持	不受支持
带有联接分析和重叠发现的企业发现配置文件	不受支持	不受支持
向下钻取已推理的值和值频率	支持	支持
导出配置文件结果	支持	支持
行计数映射*	不受支持	支持
*统计信息帮助器未能报告行计数时，剖析运行行计数映射。		

索引

A

Analyst 工具中的发现搜索结果
概览 [114](#)
Analyst 工具中的列配置文件结果
界面 [53](#), [63](#), [65](#)
列详细信息 [54](#), [65](#)
摘要 [52](#)
Analyst 工具中的企业发现
概览 [103](#)
进程 [103](#)
配置文件视图 [111](#)
数据类型冲突 [111](#)
摘要视图 [109](#)

B

表数据对象
同步 [39](#)

C

创建表达式规则
规则 [43](#)
创建列配置文件
配置文件 [35](#)

D

导出
结果卡沿袭为 XML [154](#)
discovery search：发现搜索
先决条件 [113](#)

F

发现搜索结果
界面 [115](#)

G

功能相关性发现
概览 [130](#)
规则
应用预定义规则 [42](#)
在 Informatica Developer 中创建 [145](#)
表达式 [42](#)
创建表达式规则 [43](#)
使用规则规范创建表达式规则 [44](#)
预定义 [41](#)
在 Informatica Developer 中应用 [145](#)

规则 (续)
在 PowerCenter Express 中应用 [145](#)

I

Informatica Analyst
列配置文件概览 [31](#), [62](#)
列配置文件结果 [51](#), [61](#)
锁定和版本管理 [35](#)
规则 [41](#)
Informatica Analyst 中的企业发现结果
概览 [108](#)
Informatica Developer
规则 [144](#)
配置文件概览 [122](#)
配置文件视图 [123](#)

J

结果卡
编辑度量组 [80](#)
查看 [77](#)
创建度量组 [79](#)
度量 [78](#)
度量权重 [78](#)
度量组 [79](#)
概览 [24](#)
固定成本 [79](#)
Informatica Analyst [73](#)
Informatica Analyst 进程 [74](#)
Informatica Developer [153](#)
可变成本 [79](#)
配置全局通知设置 [92](#)
配置通知 [92](#)
趋势图表 [81](#)
删除度量组 [80](#)
通知 [90](#)
无效数据的成本 [78](#)
向下钻取 [81](#)
移动得分 [80](#)
运行 [77](#)
编辑 [77](#)
定义阈值 [79](#)
向结果卡中添加列 [76](#)
结果卡沿袭
从 Informatica Developer 查看 [154](#)
在 Informatica Analyst 中查看 [93](#)
结果卡(按项目)窗格
Informatica Analyst [85](#)
结果卡结果
从 Informatica Analyst 中导出 [89](#)
导出 [89](#)
导出到 Excel [90](#)

结果卡仪表盘
Informatica Analyst [84](#)
结果卡运行趋势窗格
Informatica Analyst [86](#)
具有结果卡的数据对象
Informatica Analyst [87](#)

L

联接分析
概览 [175](#)
离群值
检测 [57](#)
累积度量窗格
Informatica Analyst [88](#)
列配置文件
操作系统配置文件 [35](#), [132](#)
Informatica Developer [126](#)
进程 [32](#)
向下钻取 [66](#)
概览 [23](#)
选项 [24](#)
列配置文件结果
Informatica Developer [148](#)

M

Mapplet 和映射剖析
概览 [146](#)
Metadata Manager 业务术语
管理业务术语 [120](#)
项目 [119](#)

N

内容管理
概念 [28](#)
Informatica Analyst [67](#)
Informatica Developer [150](#)
进程 [28](#)
任务 [29](#)

P

配置文件
Avro 或 Parquets 格式 [140](#)
编辑筛选器 [49](#)
创建筛选器 [46](#)
XML 和 JSON 格式 [138](#), [139](#)
运行 [37](#), [61](#), [62](#), [100](#)
组件 [21](#)
编辑列配置文件 [36](#)
创建列配置文件 [35](#)
配置文件结果
标记 [72](#)
从 Informatica Analyst 中导出 [69](#)
导出 [68](#)
Excel [68](#)
拒绝数据类型 [68](#)
拒绝数据域 [101](#)
拒绝数据域在 Informatica Developer 中 [164](#)
列数据类型 [56](#), [150](#)
列值 [58](#)
批准数据类型 [67](#)

配置文件结果 (续)
批准数据域 [101](#)
向下钻取 [67](#)
详细视图 [54](#)
业务术语 [70](#)
在 Informatica Developer 添加注释 [137](#)
在 Informatica Developer 中导出 [151](#)
在 Informatica Developer 中管理列关系 [185](#)
在 Informatica Developer 中批准数据域 [164](#)
摘要 [64](#), [66](#)
摘要视图 [52](#)
注释 [71](#)
列模式 [57](#)
批准数据类型在 Informatica Developer 中 [151](#)
在 Developer tool 中拒绝数据类型 [151](#)
配置文件选项
企业发现 [167](#)
配置选项
Analyst 工具中的企业发现 [104](#)
剖析
概览 [16](#)
锁定和版本管理 [24](#)
体系结构 [18](#)
平面文件数据对象
同步 [38](#)

Q

企业发现结果
导出 [188](#)
概览 [180](#)
另存为图像 [183](#)
企业发现配置文件
创建 DDL 脚本 [178](#)
DDL 脚本 [178](#)
运行 [173](#)
趋势图表
查看 [83](#)
成本 [82](#)
得分 [82](#)
从 Informatica Analyst 中导出 [83](#)
企业发现
编辑 [172](#)
表格视图 [184](#)
查看数据对象关系 [182](#)
概览 [166](#)
关系视图 [181](#)
进程 [167](#)
列配置文件视图 [187](#)
数据域视图 [185](#)
运行时状态 [188](#)
在 Analyst 工具中编辑 [106](#)
外键剖析视图 [182](#)
在 Informatica Analyst 中运行 [105](#)

S

筛选器
概览 [46](#)
数据对象配置文件
概览 [125](#)
注释 [136](#)
创建单个配置文件 [132](#)
创建多个配置文件 [133](#)
企业发现 [171](#)

- 数据发现
 - 概览 [20](#)
 - 进程 [19](#)
- 数据域
 - 导出 [158](#)
 - 导入 [158](#)
 - 概览 [26](#)
 - 在 Informatica Analyst 中查找 [96](#)
 - 在 Informatica Analyst 中创建 [95](#)
 - 在 Informatica Analyst 中基于配置文件结果创建 [96](#)
 - 在 Informatica Developer 中查找 [157](#)
 - 在 Informatica Developer 中创建 [156](#)
 - 在 Informatica Developer 中基于配置文件结果创建 [157](#)
- 数据域词汇表
 - 概览 [26](#)
 - Informatica Analyst [94](#)
 - Informatica Developer [155](#)
- 数据域发现
 - 概览 [25](#)
 - Informatica Analyst 概览 [94](#)
 - Informatica Developer 概览 [155](#)
 - 进程 [27](#)
- 数据域发现结果
 - 从 Informatica Analyst 中导出 [102](#)
 - 从 Informatica Developer 中导出 [164](#)
 - Informatica Analyst [101](#)
 - Informatica Developer [162](#)
 - 在 Informatica Analyst 中导出 [102](#)
- 数据域发现配置文件结果
 - Microsoft Excel [102](#)
- 数据域发现选项
 - Informatica Developer [159](#)
- 数据域组
 - 概览 [26](#)
 - 在 Informatica Analyst 中创建 [95](#)
 - 在 Informatica Developer 中创建 [156](#)
- 搜索
 - 业务词汇表 [189](#)
- Sqoop 配置
 - 剖析 [34](#), [129](#)

W

- 外键发现
 - 概览 [173](#)
- 外键配置文件
 - 发现 [174](#)

X

- 项目
 - Metadata Manager 业务术语 [119](#)

Y

- 业务术语
 - 查找 [189](#)
 - 查找业务术语 [120](#)
 - 自定义热键 [190](#)
- 预定义规则
 - 进程 [42](#)
- 映射对象
 - 运行配置文件 [146](#)
- 运行时环境
 - Analyst 工具 [34](#)
 - Hadoop [34](#), [128](#)

Z

- 在 Analyst 中执行发现搜索
 - 进程 [113](#)
- 重叠发现
 - 概览 [176](#)
 - 结果 [176](#)
 - 执行 [177](#)
- 主键发现
 - 概览 [129](#)