IBM SPSS Analytic Server V3.0

用户指南

注释

在使用本信息及其支持的产品之前，请先阅读第 29 页的『声明』中的信息。

产品信息

此版本是用于 IBM SPSS Analytic Server 的 V3.0.0.0 以及后续发行版和修订版，直至在新版本中另有说明为止。

目录

第 1 章 分析服务器 控制台

分析服务器 提供了用于管理数据源和项目的瘦客户机界面。

登录

1. 在浏览器的地址栏中输入 分析服务器 的 URL。 该 URL 可从服务器管理员处获取。
2. 输入登录该服务器所使用的用户名。
3. 输入与指定用户名关联的密码。

登录后，会显示控制台主页。

浏览控制台

- 标题显示产品名、当前登录的用户名和到帮助系统的链接。 当前登录的用户的名称是包含注销链接的下拉列表的标题。
- 内容区域显示您可从控制台主页中执行的操作。

数据源

数据源为记录的集合以及数据模型，它定义用于分析的数据集。 记录源可以是 HDFS 上的文件（定界文本、固定宽度文本 和 Excel）、关系数据库、HCatalog 或地理空间。 数据模型定义了分析数据必需的所有元数据（字段名称、存储、测量级别等）。 数据源所有者可以授予或限制对数据源的访问权。

数据源列表

主“数据源”页面，提供数据源列表，当前用户属于其中数据源的成员。

- 单击数据源名称以显示其详细信息和编辑其属性。
- 在搜索区域输入以对列表进行过滤，从而仅显示其名称中包含搜索字符串的数据源。
- 单击**新建**以使用在**添加新数据源**对话框中指定的名称和内容类型来创建新的数据源。
  - 请参阅第 19 页的『命名规则』，以了解有关可以给数据源提供的名称的限制。
  - 可用的内容类型是“文件”、“数据库”、HCatalog 和“地理空间”。

    注: 仅当 Analytic Server 已配置为使用这些数据源时，HCatalog 选项才可用。

    注: 选择后，将无法编辑内容类型。
- 单击**删除**将除去该数据源。 此操作将保持与数据源关联的所有文件完整无缺。
- 单击**刷新**以更新列表。
- “操作”下拉列表可执行选定的操作。
  1. 选择**导出**可创建数据源的归档并将其保存到本地文件系统。 此归档包含已添加至处于**项目**方式或**数据源**方式的数据源的任何文件。
  2. 选择**导入**可导入使用“导出”操作创建的归档。
  3. 选择**复制**可创建数据源的副本。

个别数据源详细信息

内容区域分为几个部分，这些部分取决于数据源的内容类型。

详细信息

这些设置对所有内容类型通用。

名称 这是一个可编辑的文本字段，用于显示数据源名称。

显示名称

这是一个可编辑的文本字段，显示其他应用程序中显示的数据源的名称。 如果该字段为空，那么该“名称”用作为显示名称。

描述 这是一个可编辑的文本字段，用于提供数据源解释性文本。

公开 这是一个复选框，用于指示是任何人都可以查看数据源（选中情况下）还是必须将用户和组作为成员明确添加到数据源中才可以查看数据源（未选中情况下）。

定制属性

应用程序可以通过使用定制属性来将属性附加到数据源中，例如，该数据源是否是临时数据源。 这些属性显示在 分析服务器 控制台中，以提供对应用程序如何使用数据源的进一步的洞察。

单击**保存**以保留设置的当前状态。

共享

这些设置对所有内容类型通用。

通过添加用户和组作为“作者”或“读者”，您可以共享数据源的所有权。

- 在文本框中输入时，会对其名称中包含搜索字符串的用户和组进行过滤。 从下拉列表中选择**作者**或**读者**可以分配用户在该数据源中的角色。 单击**添加成员**可以将他们添加至成员列表。
- 要除去参与者，请选择成员列表中的用户或组，然后单击**除去参与者**。

注: 具有**管理员**角色的用户对每个数据源都具有读写访问权，而不考虑他们是否明确地作为成员列出。

文件输入

专门用于定义具有文本内容类型的数据源的设置。

文件查看器

显示了可包含在数据源中的文件。 选择**项目**方式可查看 分析服务器 项目结构中的文件，选择**数据源**可查看数据源中存储的文件，或者选择**文件系统**可查看文件系统（通常为 HDFS）。 您可以浏览任一文件夹结构，但是无法编辑 HDFS，在**项目**方式下，您无法添加文件、创建文件夹或删除处于根级别的项，但是仅限于已定义的项目中。 要创建、编辑或删除项目，请使用项目。

- 单击**上载**以将文件上载到当前数据源或项目/子文件夹。 您可以在单个目录中浏览并选择多个文件。

  注: 将文件上载到分布式文件系统。 您可以在相应租户、数据源或项目（取决于所选方式）以及子文件夹下的 `/analytic-root` 目录结构中找到上载的文件。 例如，在以下情况下:

  1. 登录租户 `ibm`
  2. 创建名为 `fraudDetection` 的数据源
  3. 选择**数据源**方式
  4. 创建名为 `historicalData` 的子文件夹

5. 上载文件 charges2015.csv

然后，可以在分布式文件系统上的 /analytic-root/ibm/.datasource/fraudDetection/historicalData/charges2015.csv 中找到该文件。或者，在以下情况下:

1. 登录租户 ibm
2. 创建名为 fraudDetection 的数据源
3. 选择**项目**方式
4. 选择名为 creditProcessing 的现有项目
5. 创建名为 historicalData 的子文件夹
6. 上载文件 charges2015.csv

然后，可在分布式文件系统上的 /analytic-root/ibm/creditProcessing/historicalData/charges2015.csv 中找到该文件。

- 单击**新建文件夹**以在当前文件夹下创建新文件夹，以“新建文件夹名称”对话框中指定的名称命名。
- 单击**下载**以将选定的文件下载到本地文件系统。
- 单击**删除**以除去选定的文件/文件夹。

数据源定义中所包含的文件

使用移动按钮可将选中的文件和文件夹添加到数据源或从数据源移除。 对于数据源中每个选中的文件或文件夹，单击设置可定义读取文件的规范。

当在某个数据源中包含多个文件时，这些文件必须共享公用元数据；即，每个文件包含的字段数量必须相同，在每个文件中必须按相同顺序解析这些字段，并且在所有文件之间，每个字段必须具有相同的存储。 文件之间的不匹配可能导致在 分析服务器 读取文件时控制台无法创建预览和元数据，或者将有效的值解析为无效的值（空值）。

数据库选择

为包含记录内容的数据库指定连接参数。

数据库 选择要连接到的数据库的类型。 从以下类型中进行选择: DB2、Greenplum、Amazon Redshift、MySQL、Netezza、Oracle、SQL Server、Sybase IQ 或 TeraData。 如果未列出您正在寻找的类型，请要求您的服务器管理员使用相应的 JDBC 驱动程序配置 分析服务器。

服务器地址

输入托管数据库的服务器的 URL。

服务器端口

数据库将侦听的端口号。

数据库名称

要连接到的数据库的名称。

用户名 如果数据库受密码保护，请输入您的用户名。

密码 如果数据库受密码保护，请输入您的密码。

表名称 从要使用的数据库输入表的名称。

最大并行读取数

输入对要从数据源中指定的表读取的、可从 分析服务器 发送到数据库的并行查询数量的限制。

HCatalog 选择

指定用于访问受管于 Apache HCatalog 的数据的参数。

数据库 HCatalog 数据库名称。

表名称 从要使用的数据库输入表的名称。

过滤器 这是表的分区过滤器（将表创建为分区表的情况下）。 HCatalog 过滤仅在类型为字符串的 Hive 分区键上受支持。

注: !=、<> 和 LIKE 运算符在某些 Hadoop 版本中无效。 这是 HCatalog 与这些版本之间的兼容性问题。

HCatalog 字段映射

向数据源中的字段显示 HCatalog 中的元素的映射。 单击编辑可修改字段映射。

注: 在创建基于 HCatalog 的数据源以显示来自 Hive 表的数据之后，您可能会发现当从大量数据文件形成 Hive 表时，每次 分析服务器 开始从数据源读取数据时会发生显著延迟。 如果您注意到此类延迟，请重新构建 Hive 表，减少使用的大型数据文件的数量，并将文件数减少至 400 或更少。

地理空间选择

指定用于访问地理数据的参数。

地理空间类型

地理数据可以来自联机映射服务或形状文件。

如果要使用映射服务，请指定服务的 URL，然后选择要使用的映射层。

如果要使用形状文件，请选择或上载该文件。 请注意，形状文件实际上是一组具有公共文件名并存储在同一目录中的文件。 选择具有 SHP 后缀的文件。 Analytic Server 将查找并使用其他文件。 具有 SHX 和 DBF 后缀的两个附加文件必须始终存在；根据形状文件，还可能存在许多其他文件。

预览和元数据

为数据源指定设置之后，请单击预览和元数据以检查并确认数据源规范。

输出 通过从 分析服务器 上运行的流中输出，可附加具有文件或数据库内容类型的数据源。 选择**可写**以启用附加功能，并且:

- 针对具有数据库内容类型的数据源，选择将要写入输出数据的输出数据库表。
- 针对具有文件内容类型的数据源:
  1. 选择写入新文件的输出文件夹。

     提示: 针对每个数据源使用单独的文件夹，以简化跟踪文件与数据源之间的关联的过程。
  2. 选择文件格式；选择 **CSV**（逗号分隔的值）或者**可分割的二进制格式**。
  3. （可选）选择**生成序列文件**。 如果要创建可在下游 MapReduce 作业中复用的可分割的压缩文件，那么该功能很有用。
  4. 如果您的输出为 CSV 并且具有包含嵌入式换行符或回车符的字符串字段，请选择**可以对换行符进行转移**。 这将导致每个换行符、回车符和反斜杠分别以后跟字母“n”的反斜杠、后跟字母“r”的反斜杠和两个连续的反斜杠的形式进行写入。 此类数据必须以相同设置进行读取。 我们强烈建议在处理包含换行符或回车符的字符串数据时使用可拆分的二进制格式。
  5. 选择压缩格式。 该列表包含已配置为配合安装 分析服务器 使用的所有格式。

     注: 压缩格式和文件格式的某些组合会导致无法分割输出，从而导致不适合 MapReduce 进行进一步处理。 进行此类选择时，分析服务器 会在“输出”部分生成一条警告。

设置（文件数据源）

“设置”对话框允许您定义读取基于文件的数据的规范。 这些设置适用于所有选定的文件，以及匹配**文件夹**选项卡上的条件的选定文件夹中的所有文件。

为文件指定错误的解析器设置可能导致在 分析服务器 读取文件时控制台无法创建预览和元数据，或者将有效的值解析为无效的值（空值）。

“设置”选项卡

“设置”选项卡允许您指定文件类型和特定于该文件类型的解析器设置。

您可以针对任何受支持的文件格式使用压缩文件来定义数据源。 受支持的压缩格式包括 Gzip、Deflate、Bz2、Snappy 和 IBM CMX。

定界的文件类型

定界的文件是自由字段文本文件，其记录包含的字段数保持不变，但是每个字段包含的字符数可变。 定界的文件通常带有 *.csv 或 *.tab 文件扩展名。 有关更多信息，请参阅 第 6 页的『定界的文件类型设置』。

固定文件类型

固定字段文本文件是其中字段未定界但是从相同位置开始并且长度固定的文件。 固定字段文本文件通常带有 *.dat 文件扩展名。 有关更多信息，请参阅 第 7 页的『固定文件类型设置』。

半结构化文件类型

半结构化文件（如 *.log）是文本文件，带有可预测的结构，其结构可通过正则表达式映射至字段，但是结构化程度不及定界的文件的结构化程度高。 有关更多信息，请参阅 第 7 页的『半结构化文件类型设置』。

文本分析文件类型

文本分析文件是可使用 SPSS Text Analytics 进行分析的文档（例如，*.doc、*.pdf 或 *.txt）。

跳过空行
指定是否忽略抽取的文本内容中的空行。 缺省值为**否**。

行分隔符
指定定义换行的字符串。 缺省为换行字符“\n”。

SPSS Statistics 文件类型

SPSS Statistics 文件（*.sav 和 *.zsav）是包含数据模型的二进制文件。 针对此类文件类型，无需在“设置”选项卡上进行进一步设置。

可分割的二进制格式文件类型

指定文件类型为可分割的二进制格式文件 (*.asbf)。 此文件类型可以表示所有 分析服务器 字段类型（不同于 CSV，后者完全无法表示列表字段并且需要特殊设置才能处理嵌入式换行符和回车符）。 针对此类文件类型，无需在“设置”选项卡上进行进一步设置。

序列文件类型

序列文件 (*.seq) 是按键/值对构造的文本文件。 这些文件通常用作为 MapReduce 作业中的中间格式。

Excel 文件类型

指定文件类型为 Microsoft Excel 文件（*.xls 和 *.xlsx）。 有关更多信息，请参阅 第 8 页的『Excel 文件类型设置』。

定界的文件类型设置:

可以为定界的文件类型指定以下设置。

字符集编码

文件的字符编码。 选择或指定 Java 字符集名称，如“UTF-8”、“ISO-8859-2”和“GB18030”。 缺省值为 **UTF-8**。

字段分隔符

一个或多个标记字段边界的字符。 每个字符都被视为独立的分隔符。 例如，如果选择**逗号**和**制表符**（或者选择**其他**并输入 **,\t**），那么它表示表示逗号或制表符标记字段边界。 如果控制字符对字段进行了定界，那么除了控制字符之外，也将此处指定的字符视为定界符。 如果控制字符未对字段进行定界，那么缺省为“,”；否则缺省值为空字符串。

控制字符对字段进行定界

设置是否将 ASCII 控制字符（LF 和 CR 除外）视为字段分隔符。 缺省为**否**。

首行包含字段名称

设置是否使用首行来确定字段名称。 缺省为**否**。

要跳过的起始字符数

文件开头将被跳过的字符数。 非负整数。 缺省值为 0。

合并空格

设置是否将相邻出现的多个空格和/或制表符作为一个字段分隔符。 如果空格或制表符都不是字段分隔符，那么没有作用。 缺省为**是**。

行尾注释字符

一个或多个用于标记行尾注释的字符。 将忽略记录上该注释后的字符和所有内容。 每个字符都被视为独立的注释标记。 例如，“/*”表示注释开头的斜杠或星号。 无法定义诸如“//”之类的多字符注释标记。 空字符串表示未定义注释字符。 如果已定义，那么在处理引号或跳过起始字符之前，先检查注释字符。 缺省值为空字符串。

无效字符

确定无效字符（与编码中的字符不对应的字节序列）的处理方式。

废弃 废弃无效的字节序列。

替换为 将每个无效的字节序列替换为指定的单个字符。

单引号 指定单引号（撇号）的处理。 缺省为**保留**。

保留 单引号没有特殊含义，可将其视为任何其他字符。

删除 除非是引用内容，否则删除单引号

成对 将单引号视为引用字符，成对单引号之间的字符将失去任何特殊含义（将其视为引用内容）。 单引号本身是否可以出现在带有单引号的字符串中，这取决于设置**可以用双重引号括起引号**。

双引号 指定双引号的处理。 缺省为**成对**。

保留 双引号没有特殊含义，可将其视为任何其他字符。

删除 除非是引用内容，否则删除双引号

成对 将双引号视为引用字符，成对双引号之间的字符将失去任何特殊含义（将其视为引用内容）。 双引号本身是否可以出现在带有双引号的字符串中，这取决于设置**可以用双重引号括起引号**。

可以用双重引号括起引号

当设置为**成对**时，指示双引号是否可以显示在带有双引号的字符串中，以及单引号是否可以显示在带有单引号的字符串中。 如果**是**，那么双引号可以在使用双重双引号的字符串中进行转义，以及单引号可以在使用双重单引号的字符串中进行转义。 如果**否**，那么不能在带有双引号的字符串中使用双引号，并且不能在带有单引号的字符串中使用单引号。 缺省为**是**。

可以转义换行

指示解析器是否将后跟字母“n”、字母“r”或另一个反斜杠的反斜杠分别解释为换行符、回车符或反斜杠字符。 如果未对换行符进行转义，那么这些字符将按字面读取为后跟字母“n”的反斜杠等。 缺省值为**否**。

固定文件类型设置:

可以为固定文件类型指定以下设置。

字符集编码

文件的字符编码。 选择或指定 Java 字符集名称，如“UTF-8”、“ISO-8859-2”和“GB18030”。 缺省值为**UTF-8**。

无效字符

确定无效字符（与编码中的字符不对应的字节序列）的处理方式。

废弃 废弃无效的字节序列。

替换为 将每个无效的字节序列替换为指定的单个字符。

记录长度

指示如何定义记录。 如果选择**通过换行来定界**，那么记录由换行、文件开始或文件结束来定义（定界）。 如果选择**特定长度**，那么记录由记录长度（以字节计）来定义。 指定正数值。

要跳过的起始记录数

文件开头将被跳过的记录数。 指定非负整数。 缺省值为 0。

字段 此部分定义了文件中的字段。 单击**添加字段**并指定字段名称、字段值开始的列，以及字段值的长度。 文件中的列从 0 开始编号。

半结构化文件类型设置:

半结构化文件的设置包含将文件内容映射至字段的规则。

规则表 个别规则会从记录中提取信息以创建字段；通过与规则表相结合，可以定义可从数据源中的每条记录提取的所有字段。

表中的规则按顺序应用于每条记录；如果表中的所有规则均匹配记录，那么无需任何其他规则即可处理该记录，并处理下一条记录。 如果表中的任何规则不匹配，那么将废弃表中先前规则提取的所有字段值；如果存在另一个规则表，那么该表中的规则将应用于此记录。 如果没有匹配记录的表，那么将应用“不匹配”规则。

不匹配 您可以选择**跳过**不匹配任何规则表的记录，或者将该记录中的所有字段值设置为**缺失**（空值）。

导出规则

您可以保存当前可见的规则表以供复用。 导出的表会保存在服务器上。

导入规则

您可以将保存的规则表导入当前可见的规则表。 这将覆盖您针对该表定义的任何规则，因此最好创建一个新表，然后导入规则表。

规则编辑器

规则编辑器允许您为单个字段创建抽取规则。

匿名捕获组

这是一条字段捕获规则，通常开始从先前规则停止的位置上的记录抽取数据。 当半结构化数据源中的两个字段之间存在无关联的信息时，可将其用于定义匿名捕获组，以将解析器置于下一个字段开始的位置。 选择**匿名捕获组**时，禁用对捕获组进行命名和添加标签 的控件，但是其余对话框可正常运作。

字段名称

输入字段的名称。 这用于定义数据源元数据。 字段名称在规则表中必须唯一。

规则名称

（可选）输入规则的描述性标签。

描述 （可选）输入规则的详细描述。

定义规则

存在两种规则定义方法。

使用抽取规则控件

这样可简化抽取规则的创建。

1. 指定抽取字段数据的起点；**当前位置**将从上一条规则停止的位置开始，**跳至**将从记录开始处开始忽略所有字符，直至达到文本框中指定的字符为止。 如果希望字段数据包含位于开始位置处的字符，请选择**包含**。
2. 从**捕获**下拉列表中选择字段捕获组。
3. （可选）选择抽取字段数据的停止点；**空格**将在遇到任何空格字符（例如，空格或跳格）时停止，**位于字符**将在指定的字符串处停止。 如果希望字段数据包含位于停止位置处的字符，请选择**包含**。

手动定义正则表达式规则

如果对编写正则表达式语法无异议，请选择此项。 在**正则表达式**文本框中输入一个正则表达式。

添加字段捕获组

这允许您保存正则表达式以供将来使用。 保存的捕获组会显示在**捕获**下拉列表上。

在已应用规则表中先前所有规则后，“规则编辑器”会显示由该规则从第一条记录抽取的数据的预览。

***Excel** 文件类型设置:*

可以为 Excel 文件指定以下设置。

工作表选择

选择要用作为数据源的 Excel 工作表。 指定数字索引（第一个工作表的索引为 0）或工作表名称。 缺省为使用第一个工作表。

用于导入的数据范围选择。

可以从第一个非空白行或单元格的显式范围开始导入数据。

- **从第一个非空白行上开始的范围**。 找到第一个非空白单元格，并将其用作为数据范围的左上角。

- 或者按行和列指定单元格的显式范围。 例如，要指定 Excel 范围 A1:D5，可以在第一个字段中输入 A1，在第二个字段中输入 D5（或者输入 R1C1 和 R5C4）。 这样会返回指定范围内的所有行，包括空白行。

首行包含字段名称
指定所选单元格范围的第一行是否包含字段名称。 缺省值为**否**。

遇到空白行后停止读取
指定在遇到多个空白行后是停止读取记录还是继续读取所有数据直至工作表末尾（包括空白行）。 缺省值为**否**。

格式

“格式”选项卡允许您为解析的字段定义格式化信息。

字段转换设置

删除空格
将字符串字段开头和/或结尾的空格字符移除。 缺省为**无**。 支持以下值:

无 不必移除空格字符。

左侧 将字符串字段开头的空格字符移除。

右侧 将字符串字段结尾的空格字符移除。

两侧 将字符串字段开头和结尾的空格字符都移除。

语言环境
定义语言环境。 缺省为服务器语言环境。 必须将语言环境字符串指定为: <language>[_country[_variant]]，其中:

language
是 ISO-639 所定义的有效的两个小写字母代码。

country
是 ISO-3166 所定义的有效的两个大写字母代码。

variant
是专门用于供应商或浏览器的代码。

十进制分隔符
设置用作为小数符号的字符。 缺省为特定于语言环境的设置。

分组符号
设置是否必须使用用于千位分隔符且特定于语言环境的字符。

缺省日期格式
定义缺省日期格式。 支持由 Unicode 语言环境数据标记语言 (LDML) 规范定义的所有格式模式。

缺省时间格式
定义缺省时间格式。

缺省时间戳记
定义缺省时间戳记格式。

缺省时区
设置时区。 缺省为 UTC。 此设置适用于不具有显式指定的时区的时间和时间戳记字段。

字段覆盖

本部分允许您为个别字段分配格式化指示信息。 从数据模型中选择字段，或者输入字段名称，然后单击**添加**以将其添加到具有个别指示信息的字段列表中。 单击**除去**以将其从列表中除去。 对于列表中选中的字段，可以设置该字段的以下属性。

存储 设置字段的存储。

十进制分隔符
对于具有实存储的字段，设置用作为小数符号的字符。 缺省为特定于语言环境的设置。

分组符号
对于具有整数或实存储的字段，设置是否应使用用作为千位分隔符的特定于语言环境的字符。

格式 对于具有“日期”、“时间”或“时间戳记”存储的字段，设置格式。 从下拉列表中选择格式。

“字段顺序”选项卡

对于定界的文件类型和 Excel 文件类型，“字段顺序”选项卡允许您定义文件的字段解析顺序。 当在数据源中有多个文件时，这是很重要的，因为在各文件中字段的实际顺序可能不同，但是字段的解析顺序必须相同才能创建一致的数据模型。

对于固定的文件类型和半结构化文件类型，在“设置”选项卡上定义顺序。

当数据源中有单个文件时，或者所有文件的字段顺序都相同，那么可以使用缺省**字段顺序匹配数据模型**。 如果在数据源中有多个字段，并且文件中的字段顺序不匹配，那么请定义**特定字段顺序**用于解析文件。

1. 要将字段添加到顺序列表，请输入字段名称，或者从数据模型提供的列表中选择字段。 您可通过单击**添加全部**来同时添加数据模型中的所有字段。 只能在顺序列表中添加一次字段名称。
2. 使用箭头按钮按期望对字段进行排序。

当使用**特定字段顺序**时，未添加到列表中的任何字段都不属于该文件的结果集的一部分。 如果在此对话框中未列出的数据模型中包含字段，那么在结果集中这些值为空。

“文件夹”选项卡

为文件夹指定解析器设置时，“文件夹”选项卡允许您选择文件夹中哪些文件要包含在数据源中。

匹配所选文件夹中的所有文件
数据源将包含顶层文件夹中的所有文件；不包含子文件夹中的文件。

匹配使用正则表达式的文件
数据源将包含顶层文件夹中匹配指定正则表达式的所有文件；不包含子文件夹中的文件。

匹配使用 Unix 文件名替换表达式的文件（潜在递归）
数据源包含匹配指定的 Unix 文件名替换表达式的所有文件；该表达式可包含位于所选文件夹的子文件夹中的文件。

HCatalog 字段映射

HCatalog 模式
显示指定表的结构。 HCatalog 可以支持高度结构化的数据。 要在此类数据上定义 分析服务器 数据源，必须对结构进行序列化，从而转换成简单行和简单列。 选择模式中的元素，然后单击移动按钮，可以将元素映射到字段以进行分析。

不是所有节点都可以进行映射的。 例如，将复杂类型的阵列或映射视为“父级”，并且无法进行直接映射；HCatalog 的阵列或映射中的每个简单元素必须单独添加。 可以通过树中以 ...:array:struct 或 ...:map:struct 结尾的标签来确定这些节点。

例如:

- 对于整数数组，为数组中的值分配字段：bigintarray[45]，而不能为数组本身分配字段：bigintarray
- 对于映射，可以为映射中的值分配字段：datamap[″key″]，而不能为映射本身分配字段：datamap
- 对于整数数组，可以为值分配字段 bigintarrayarray[45][2]，但不能为数组本身分配字段 bigintarrayarray[45]。

因此，为数组或映射元素分配字段时，元素定义必须包含 index 或 key：bigintarray[index] 或 bigintmap[″key″]。

字段映射

HCatalog 元素
双击单元以进行编辑。 HCatalog 元素为阵列或映射时，必须对单元进行编辑。 对于阵列，根据您想要将其映射到某个字段的阵列的成员来指定整数。 对于映射，根据您想要将其映射到某个字段的关键字来指定加引号的字符串。

映射字段
分析服务器 数据源中出现的字段。 双击单元以进行编辑。 不允许复制“映射字段”列中的值，否则将引起错误。

存储 字段的存储。 存储来自 HCatalog，并且无法进行编辑。

注: 单击编辑和元数据来最终完成 HCatalog 数据源时，没有编辑选项。

原始数据
显示 HCatalog 中存储的记录；这可以帮助您确定如何将 HCatalog 模式映射到字段。

注: HCatalog 选项中指定的任何过滤可应用于原始数据的视图。

使用 HCatalog 数据源

分析服务器 为 HCatalog 数据源提供支持。 本节描述了如何设置各种底层 NoSQL 数据库。

在大多数情况下，您应该参考供应商文档以了解 Hive 集成。

Apache Accumulo
https://cwiki.apache.org/confluence/display/Hive/AccumuloIntegration

Apache Cassandra
第 12 页的『Apache Cassandra』

Apache HBase
https://cwiki.apache.org/confluence/display/Hive/HBaseIntegration

MongoDB
https://github.com/mongodb/mongo-hadoop/wiki/Hive-Usage

Oracle NoSQL
https://docs.oracle.com/cd/E57371_01/doc.41/e57351/bigsql.htm#BIGUG21115

XML 数据源
第 13 页的『XML 数据源』

Apache Cassandra

分析服务器 为在 Apache Cassandra 中具有底层内容的 HCatalog 数据源提供支持。

Cassandra 可提供结构化键值存储。 键映射到多个值，这些值分组为多个列族。 列族是在创建数据库时固定的，但是可以随时将列添加到族中。 此外，列仅添加到指定的键中，因此不同的键在任何给定族中可包含不同数量的列。 来自每个键的列族的值存储在一起。

由两种方法可用于定义 Cassandra 表: 使用旧 Cassandra 命令行界面 (cassandra-cli) 和新的 CQL shell (csqlsh)。

如果表是使用旧 CLI 创建的，那么请使用以下语法在 Hive 中创建外部 Apache Cassandra 表。

```
CREATE EXTERNAL TABLE <hive_table_name>  (<column specifications>)
STORED BY 'org.apache.hadoop.hive.cassandra.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<cassandra_column_family>",
"cassandra.host"="<cassandra_host>","cassandra.port" = "<cassandra_port>")
TBLPROPERTIES ("cassandra.ks.name" = "<cassandra_keyspace>");
```

例如，针对以下 CLI 表定义:

```
create keyspace test
with placement_strategy = 'org.apache.cassandra.locator.SimpleStrategy'
and strategy_options = [{replication_factor:1}];

create column family users with comparator = UTF8Type;

update column family users with
        column_metadata =
        [
        {column_name: first, validation_class: UTF8Type},
        {column_name: last, validation_class: UTF8Type},
        {column_name: age, validation_class: UTF8Type, index_type: KEYS}
        ];

assume users keys as utf8;

set users['jsmith']['first'] = 'John';
set users['jsmith']['last'] = 'Smith';
set users['jsmith']['age'] = '38';
set users['jdoe']['first'] = 'John';
set users['jdoe']['last'] = 'Dow';
set users['jdoe']['age'] = '42';

get users['jdoe'];
```

... 将显示的 Hive 表 DDL 如下:

```
CREATE EXTERNAL TABLE cassandra_users  (key string, first string, last string, age string)
STORED BY 'org.apache.hadoop.hive.cassandra.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "users",
"cassandra.host"="<cassandra_host>","cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

如果表是使用 CQL 创建的，那么请使用以下语法在 Hive 中创建外部 Apache Cassandra 表。

```
CREATE EXTERNAL TABLE <hive_table_name>  (<column specifications>)
STORED BY 'org.apache.hadoop.hive.cassandra.cql.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<cassandra_column_family>",
"cassandra.host"="<cassandra_host>","cassandra.port" = "<cassandra_port>")
TBLPROPERTIES ("cassandra.ks.name" = "<cassandra_keyspace>");
```

例如，针对以下 CQL3 表定义:

```
CREATE KEYSPACE TEST WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 2 };
USE TEST;
```

```
CREATE TABLE bankloan_10(
  row int,
  age int,
  ed int,
  employ int,
  address int,
  income int,
  debtinc double,
  creddebt double,
  othdebt double,
  default int,
  PRIMARY KEY(row)
);

INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (1,41,3,17,12,176,9.3,11.359392,5.008608,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (2,27,1,10,6,31,17.3,1.362202,4.000798,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (3,40,1,15,14,55,5.5,0.856075,2.168925,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (4,41,1,15,14,120,2.9,2.65872,0.82128,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (5,24,2,2,0,28,17.3,1.787436,3.056564,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (6,41,2,5,5,25,10.2,0.3927,2.1573,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (7,39,1,20,9,67,30.6,3.833874,16.668126,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (8,43,1,12,11,38,3.6,0.128592,1.239408,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (9,24,1,3,4,19,24.4,1.358348,3.277652,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (10,36,1,0,13,25,19.7,2.7777,2.1473,0);
```

... Hive 表 DDL 如下:

```
CREATE EXTERNAL TABLE cassandra_bankloan_10 (row int, age int,ed int,employ int,address int,
   income int,debtinc double,creddebt double,othdebt double,default int)
STORED BY 'org.apache.hadoop.hive.cassandra.cql.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "bankloan_10","cassandra.host"="<cassandra_host>",
   "cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

XML 数据源

分析服务器 通过 HCatalog 为 XML 数据提供支持。

示例

1. 根据以下规则，通过 Hive 数据定义语言 (DDL) 将 XML 模式映射到 Hive 数据类型。

```
CREATE [EXTERNAL] TABLE <table_name> (<column_specifications>)
ROW FORMAT SERDE "com.ibm.spss.hive.serde2.xml.XmlSerDe"
WITH SERDEPROPERTIES (
    ["xml.processor.class"="<xml_processor_class_name>",]
    "column.xpath.<column_name>"="<xpath_query>",
     ...
    ["xml.map.specification.<element_name>"="<map_specification>"
     ...
    ]
)
STORED AS
    INPUTFORMAT "com.ibm.spss.hive.serde2.xml.XmlInputFormat"
    OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat"
[LOCATION "<data_location>"]
TBLPROPERTIES (
    "xmlinput.start"="<start_tag ",
    "xmlinput.end"="<end_tag>"
);
```

注: 如果您的 XML 文件是使用 Bz2 压缩进行压缩的，那么 INPUTFORMAT 映射至为 com.ibm.spss.hive.serde2.xml.SplittableXmlInputFormat。 如果是使用 CMX 压缩进行压缩的，那么映射至为 com.ibm.spss.hive.serde2.xml.CmxXmlInputFormat。

例如，以下 XML...

```
<records>
  <record customer_id="0000-JTALA">
    <demographics>
      <gender>F</gender>
      <agecat>1</agecat>
      <edcat>1</edcat>
      <jobcat>2</jobcat>
      <empcat>2</empcat>
      <retire>0</retire>
      <jobsat>1</jobsat>
      <marital>1</marital>
      <spousedcat>1</spousedcat>
      <residecat>4</residecat>
      <homeown>0</homeown>
      <hometype>2</hometype>
      <addresscat>2</addresscat>
    </demographics>
    <financial>
      <income>18</income>
      <creddebt>1.003392</creddebt>
      <othdebt>2.740608</othdebt>
      <default>0</default>
    </financial>
  </record>
</records>
```

...将使用以下 Hive DDL 来表示。

```
CREATE TABLE xml_bank(customer_id STRING, demographics map<string,string>, financial map<string,string>)
ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'
WITH SERDEPROPERTIES (
    "column.xpath.customer_id"="/record/@customer_id",
    "column.xpath.demographics"="/record/demographics/*",
    "column.xpath.financial"="/record/financial/*"
)
STORED AS
    INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'
    OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'
TBLPROPERTIES (
    "xmlinput.start"="<record customer",
    "xmlinput.end"="</record>"
);
```

有关更多信息，请参阅 『XML 到 Hive 数据类型映射』。

2. 在 分析服务器 控制台中使用 HCatalog 内容类型创建 分析服务器 数据源。

限制

- 当前仅支持 XPath 1.0 规范。
- 处理 Hive 字段名称时使用元素和属性的限定名的本地部分。 忽略名称空间前缀。

XML 到 Hive 数据类型映射: 以 XML 建模的数据可使用以下记录的约定转换为 Hive 数据类型。

结构

XML 元素可直接映射到 Hive 结构类型，以使所有属性都成为数据成员。 元素的内容成为基本或复杂类型的额外成员。

XML 数据

```
<result name="ID_DATUM">03.06.2009</result>
```

Hive DDL 和原始数据

```
struct<name:string,result:string>
```

```
{"name":"ID_DATUM", "result":"0.3.06.2009"}
```

阵列

元素的 XML 序列可表示为基本或复杂类型的 Hive 阵列。 以下示例显示了用户如何使用 XML <result> 元素的内容来定义字符串的阵列。

XML 数据

```
<result>03.06.2009</result>
<result>03.06.2010</result>
<result>03.06.2011</result>
```

Hive DDL 和原始数据

```
result array<string>
{"result":["03.06.2009","03.06.2010",...]}
```

映射

XML 模式不为映射提供本机支持。 在 XML 中对映射建模有三种常用方法。 为适应不同的方法，我们使用以下语法:

```
"xml.map.specification.<element_name>"="<key>-><value>"
```

其中

元素名称
要视为映射条目的 XML 元素的名称

键 映射条目键 XML 节点

值 映射条目值 XML 节点

应在 Hive 表创建 DDL 中的 SERDEPROPERTIES 部分下定义给定 XML 元素的映射规范。 可使用以下语法来定义键和值:

@attribute
@attribute 规范允许用户使用属性值作为映射的键或值。

元素 元素名称可用作为键或值。

#content
元素内容可用作为键或值。 由于映射键只能采用基本类型，复杂内容将转换为字符串。

以 XML 表示映射的方法及其对应的 Hive DDL 和原始数据如下所示。

元素名称到内容
元素的名称用作为键，内容用作为值。 这是常用的技术之一，缺省情况下，将 XML 映射到 Hive 映射类型时使用。 此方法的明显限制在于映射键只能采用字符串类型。

XML 数据

```
<entry1>value1</entry1>
<entry2>value2</entry2>
<entry3>value3</entry3>
```

映射、Hive DDL 和原始数据

在此情况下，无需指定映射，因为缺省情况下元素名称用作为键，内容用作为值。

```
result map<string,string>
{"result":{"entry1": "value1", "entry2": "value2", "entry3": "value3"}}
```

属性到元素内容
使用属性值作为键，使用元素内容作为值。

XML 数据

```
<entry name="key1">value1</entry>
<entry name="key2">value2</entry>
<entry name="key3">value3</entry>
```

映射、Hive DDL 和原始数据

```
"xml.map.specification.entry"="@name->#content"

result map<string,string>

{"result":{"key1": "value1", "key2": "value2", "key3": "value3"}}
```

属性到属性

XML 数据

```
<entry name="key1" value="value1"/>
<entry name="key2" value="value2"/>
<entry name="key3" value="value3"/>
```

映射、Hive DDL 和原始数据

```
"xml.map.specification.entry"="@name->@value"

result map<string,string>

{"result":{"key1": "value1", "key2": "value2", "key3": "value3"}}
```

复杂内容

用作为基本类型的复杂内容将转换为有效的 XML 字符串，方法是添加称为 <string> 的根元素。 请考虑以下 XML:

```
<dataset>
 <value>10</value>
 <value>20</value>
 <value>30</value>
</dataset>
```

XPath 表达式 /dataset/* 将导致返回多个 <value> XML 节点。 如果目标字段为基本类型，那么实施将把查询结果转换为有效的 XML，方法是添加 <string> 根节点。

```
<string>
 <value>10</value>
 <value>20</value>
 <value>30</value>
</string>
```

注: 如果查询结果是单个 XML 元素，那么实施将不会添加根元素 <string>。

文本内容

如果 XML 元素的文本内容仅包含空格，将被忽略。

预览和元数据（数据源）

单击**预览和元数据**会显示记录样本和数据源的数据模型。 此时您可以查看基本元数据信息。

预览 “预览”选项卡显示了记录的小样本及其字段值。

编辑

“编辑”选项卡会显示基本字段元数据。 对于带有“文件”内容类型的数据源，数据模型是从小型记录样本生成的，可以在该选项卡上手动编辑字段元数据。 对于带有 HCatalog 内容类型的数据源，数据模型是基于 HCatalog 字段映射生成的，不能在该选项卡上编辑字段存储。

字段 双击字段名称可对其进行编辑。

度量 这是用于描述给定字段中数据特征的测量级别。

角色 用于告知建模节点，机器学习过程的字段将是“输入”（预测变量字段）还是“目标”（预测字段）。 “两者皆是”和“无”也是可与“分区”一起使用的角色，指示了用于将记录划分到单独样本以进行培训、测试和验证的字段。 Split 值指定为字段的每个可能的值构建单独的模型。 “频率”可指定字段值应用作为每条记录的频率权重。 “记录标识”用于识别输出中的记录。

存储 存储描述了将数据存储在字段中的方式。 例如，值为 1 和 0 的字段用于存储整数数据。 这与测量级别不同，测量级别描述了数据的使用情况，而不会影响存储。 例如，要设置整数字段的测量级别，可将值标记为 1 和 0。通常 1 = True，0 = False。

值 显示带有分类度量的字段的个别值或者带有连续度量的字段的值范围。

结构 指示字段中的记录包含单一值（原语）还是值列表。

深度 指示列表深度；0 为原语的列表，1 为列表的列表，以此类推。

扫描所有数据值

该选项允许您启动并取消数据源数据值的扫描，以确定类别值和范围限制。 如果扫描正在进行中，那么请单击该按钮以**取消数据扫描**。 扫描所有数据值可确保元数据正确，但是如果数据源具有众多字段和记录，那么可能需要一些时间。

项目

项目是用于存储输入和访问作业输出的工作空间。 项目提供了用于包含文件和文件夹的顶级组织结构。 项目可以与单个用户和组分享。

项目列表

主“项目”页面，提供项目列表，当前用户属于其中项目的成员。

- 单击项目名称以显示其详细信息和编辑其属性。
- 在搜索区域输入以对列表进行过滤，从而仅显示其名称中包含搜索字符串的项目。
- 单击**新建**以使用在**添加新项目**对话框中指定的名称创建新项目。 请参阅第 19 页的『命名规则』，以了解有关可以给项目提供的名称的限制。
- 单击**删除**以除去所选项目。 此操作会从 HDFS 除去该项目并删除与该项目关联的所有数据。
- 单击**刷新**以更新列表。

个别项目详细信息

内容其余划分为**详细信息**、**共享**、**文件**和**版本**可折叠部分。

详细信息

名称 这是一个可编辑的文本字段，显示项目名称。

显示名称

这是一个可编辑的文本字段，显示其他应用程序中显示的项目的名称。 如果该字段为空，那么该“名称”用作为显示名称。

描述 这是一个可编辑的文本字段，用于提供项目解释性文本。

要保留的版本数量

版本数量超过指定数量时，自动删除最早提交的项目版本。 缺省值为 25。

注: 清除过程并非即时完成的过程，而是每 20 分钟在后台运行一次。

公开 这是一个复选框，用于指示是任何人都可以查看项目（选中情况下）还是必须将用户和组作为成员明确添加到项目中才可以查看项目（未选中情况下）。

单击**保存**以保留设置的当前状态。

共享 通过将用户和组作为作者或查看者添加到项目中，可以共享该项目。

- 在文本框中输入时，会对其名称中包含搜索字符串的用户和组进行过滤。 选择共享级别，然后单击**添加成员**以添加到成员列表中。
  - 作者是项目的正式成员，可修改该项目以及该项目中的文件夹和文件。 通过 IBM® SPSS® Modeler 连接到 分析服务器 时，这些组的这些用户和成员对此项目具有写（分析服务器"导出"节点）访问权。
  - 查看者可以查看项目中的文件夹和文件，并通过项目中的对象来定义数据源，但是不能修改该项目。
- 要移除作者，请在"作者"列表中选择用户或组，然后单击**移除成员**。

注: 无论是否将管理员明确列为成员，他们对每个项目都具有读写访问权。

注: 会自动立即应用对"共享"所作的更改。

文件

项目结构窗格

右侧窗格显示当前所选项目的项目/文件夹结构。 您可以浏览文件夹结构，但是不能对其进行编辑，除非通过这些按钮。

- 单击**将文件下载到本地文件系统**按钮时，会将选中的文件下载到本地文件系统。
- 单击**删除选中的文件**按钮时，会移除选中的文件/文件夹。

文件查看器

显示当前项目的文件夹结构。 文件夹结构只能在定义的项目中编辑。 即，您无法在**项目**方式的 root级别上添加文件、创建文件夹或删除项。 要创建或删除某个项目，请返回至"项目"列表。

- 单击**将文件上载到 HDFS** 以将文件上载到当前项目/子文件夹。
- 单击**创建新文件夹**以在当前文件夹下创建新文件夹，新文件夹名称为您在**新文件夹名称**对话框中指定的名称。
- 单击**将文件下载到本地文件系统**按钮时，会将选中的文件下载到本地文件系统。
- 单击**删除选中的文件**按钮时，会移除选中的文件/文件夹。

版本

项目的版本号基于对文件和文件夹内容的更改。 对项目属性的更改（如描述、是否公共的以及与谁共享）不需要新版本。 添加、修改或删除文件或文件夹不需要新版本。

项目版本控制表

该表显示现有项目版本、其创建和提交日期、负责每个版本的用户以及父版本。 父版本是所选版本基于的版本。

- 单击**锁定**可以对所选项目版本内容做出更改。
- 单击**提交**将保存对项目做出的所有更改并使此版本成为项目的当前可视状态。
- 单击**放弃**将放弃对锁定项目做出的所有更改并将项目的可视状态返回到最近提交的版本。

- 单击**删除**按钮将除去所选版本。

用户管理

管理员可以通过“用户”页面管理用户和组的角色。

内容区域划分为**详细信息**和**主体**可折叠部分。

详细信息

名称 这是一个不可编辑的文本字段，显示租户的名称。

描述 可编辑的文本字段，您可以提供关于该租户的说明性文本。

URL 这是将提供给用户通过 分析服务器 控制台登录租户的 URL。

状态 **处于活动状态的**租户目前正在使用中。 使租户**处于不活动状态**将使用户无法登录该租户，但是不会删除任何底层的信息。

主体

主体是从配置过程中设置的安全提供程序获取的用户和组。 您可以将主体的角色更改为“管理员”、“用户”或“读者”。

度量 使您能够配置租户的资源限制。 报告租户目前使用的磁盘空间。

- 您可以为租户设置最大磁盘空间配额；当达到该限制时，不能再向该租户上的磁盘写入任何数据，直到清理出足够的磁盘空间，以使租户磁盘空间使用量低于配额。
- 您可以为租户设置磁盘空间警告级别；当超过配额时，该租户上的主体不能提交分析作业，直到清理出足够的磁盘空间，以使租户磁盘空间使用量低于配额。
- 您可以设置在该租户上一次能够运行的最大并行作业数；当超过配额时，该租户上的主体不能提交分析作业，直到目前正在运行的作业完成。
- 您可以设置数据源能够具有的最大字段数。 每当创建或更新数据源时，都会检查该限制。
- 您可以设置数据源能够具有的最大记录数。 每当创建或更新数据源时（例如，添加新文件或更改文件的设置时），都会检查该限制。
- 您可以设置最大文件大小（以兆字节为单位）。 上载文件时，会检查该限制。

安装提供程序配置

使您能够指定用户认证提供程序。 **缺省**使用在安装和配置期间设置的缺省租户提供程序。 **LDAP** 允许您使用诸如 Active Directory 或 OpenLDAP 之类的外部 LDAP 服务器来认证用户。 为提供程序指定设置并且可选择指定过滤器设置，以控制“主体”部分中提供的用户和组。

命名规则

对于在 分析服务器 中可给予唯一名称的一切对象（例如，数据源和项目），以下规则适用于这些名称。

- 在单个租户中，名称必须在同类型的对象中唯一。 例如，两个数据源不能同时命名为 insuranceClaims，但是一个数据源和一个项目可分别命名为 insuranceClaims。
- 名称区分大小写。 例如，insuranceClaims 和 InsuranceClaims 均被视为唯一名称。
- 名称忽略前置和后置空格。
- 以下字符在名称中无效。

  ~, #, %, &, *, {, }, \\, :, <, >, ?, /, |, ", \t, \r, \n

第 2 章 SPSS Modeler 集成

SPSS Modeler 是一种具有可视化分析方法的数据挖掘工作台。 用画布上的节点表示作业中的每个不同操作（从访问数据源到合并记录、以及写入新文件或构建模型）。 将这些操作链接起来形成一个分析流。 要构建与 分析服务器 一起运行的分析流:

1. 流必须以 分析服务器 源节点为起点。
2. 选择 分析服务器 支持的进程节点（字段或记录操作），像平时一样在 Modeler 界面中构建流的中间部分。 Modeler 选用板中存在显示受支持节点的 分析服务器 面板。
3. 有几个用于完成流的选项。
   - 选择 分析服务器 支持的终端节点（输出、图形、导出或建模）。在这种情况下，Modeler 将整个流推送到 分析服务器。 分析服务器 在 Hadoop 集群上编排必需的作业，并使结果可用于 Modeler。 Modeler 接受这些结果并向您提供这些结果，就像在本地处理流一样。
   - 如果选择 分析服务器 不支持的终端节点，Modeler 将向 分析服务器 推送尽可能多的流，然后开始从 Hadoop 获取记录。 请注意，目前无法使用 分析服务器 构建的模型可以由 分析服务器 评分。 这表示您可以使用 分析服务器 构建流以获得大数据的有效统计子样本，然后在 Modeler“本地”构建模型。 产生的模型块可包含在完全在 分析服务器 中运行的评分流中。

注: 您可以在 分析服务器 流属性中设置 SPSS Modeler 将从 Hadoop 下载的最大记录数。

受支持的节点

HDFS 上支持执行许多 SPSS Modeler 节点，但是某些节点的执行可能有所不同，另有一些节点目前尚不受支持。本主题详细描述了目前的支持级别。

注: 请参阅 SPSS Modeler 文档以获取有关这些节点的常规操作的信息。

概述

- 分析服务器 将不接受某些在带引号的建模器字段名称中通常可接受的字符。
- 对于要在 分析服务器 中运行的“建模器”流，它必须以一个或多个 分析服务器“源”节点开始，并以单个“建模”节点或 分析服务器“导出”节点结束。
- 建议将连续目标的存储设置为实数而非整数。评分模型始终将实数值写入连续目标的输出数据文件，而评分的输出数据模型会跟在目标存储后面。 因此，如果连续目标是整数存储，那么写入的值和评分的数据模型将不匹配，这种情况会在尝试读取评分数据时造成错误。

源

- 将在本地运行以 分析服务器 源节点之外的任何节点开始的流。

记录操作

支持所有记录操作，“流式方法 TS”节点和“空间时间框”节点除外。 以下是有关受支持的节点功能的进一步注释。

选择

- 支持衍生节点所支持的同一组函数。
- 将“选择”节点与废弃选项配合使用时，将废弃结果集中具有空值的字段。 例如: 如果条件是废弃 OCCUPATION = ″Retired″ 的行，那么将废弃 OCCUPATION = ″Retired″ 并且 OCCU-

PATION = null 的所有行。 您应该将选择条件修改为添加“not(field = undef)”。 例如: 将选择条件更新为 ((OCCUPATION = ″Retired) 和 not(OCCUPATION = undef))。 此结果集将包含 OCCUPATION 字段为空的行。

样本

- 不支持块级采样。
- 不支持“复合抽样”方法。

聚合

- 不支持连续密钥。 如果要复用设置用于对数据进行排序的现有流，然后在“汇总”节点中使用该设置，请更改流以除去“排序”节点。
- 顺序统计（中间值、第一个四分位值和第三个四分位值）采用近似计算，通过“优化”选项卡提供支持。

排序

- 不支持“优化”选项卡。

在分布式环境中，保留由“排序”节点建立的记录顺序的操作数量有限。

- “排序后接导出”节点会生成经过排序的数据源。
- 带有**第一条**记录采样的“排序后接样本”节点会返回前 *N* 条记录。

总之，应将“排序”节点放置在尽可能靠近需要经过排序的记录的操作的位置。

合并

- 不支持按顺序合并。
- 不支持“优化”选项卡。
- 分析服务器 不会连接空字符串键；即，如果正在合并的某个键包含空字符串，那么包含该空字符串的任何记录将会从已合并的输出中删除。
- 合并操作相对较慢。 如果 HDFS 中存在可用空间，那么一次合并您的数据源并在以下流中使用合并的源可以比在每个流中合并数据源更快。

R 变换

节点中的 R 语法应包含一次一记录操作。

字段操作

支持所有字段操作，“变换”节点、“时间间隔”节点和“历史记录”节点除外。 以下是有关受支持的节点功能的进一步注释。

自动数据准备

- 不支持培训节点。 支持将受过培训的自动数据准备节点中的转换应用于新数据。

类型

- 不支持检查列。
- 不支持“格式”选项卡。

衍生

- 支持除序列功能之外的所有派生功能。
- 在拆分时使用拆分字段的同一个流中无法派生这些字段；您将需要创建两个流；一个流派生拆分字段，另一个流在拆分时使用该字段。
- 在比较中标记字段不能被它自己所使用；即，if (flagField) then ... endif 将导致错误；使用 if (flagField=trueValue) then ... endif 即可解决此问题

- 建议您使用 ** 运算符将指数指定为实数（如 x**2.0），而不是 x**2，以便与 Modeler 中的结果匹配。

填料

- 支持衍生节点所支持的同一组函数。

分箱 不支持以下功能。

- 最优分箱
- 等级
- 平铺 -> 平铺：值之和
- 平铺 -> 间距：保持当前值并随机分配
- 平铺 -> 定制 N：值大于 100，其中 100 % N 的任何 N 值均不等于零。

RFM 分析

- 不支持用于处理间距的“Keep in current”选项。 RFM 近度 (recency)、频度 (frequency)和值度 (monetary) 评分与“建模器”用同一数据计算出的评分不一定始终匹配。 评分范围相同，但是评分赋值（bin 号）可能不同。

图形 支持所有的“图形”节点。

建模 支持以下建模节点：时间序列、TCM、树 AS、C&R 树、请求、CHAID、线性、线性 AS、神经网络、GLE、LSVM、二阶 AS、随机树、STP 及关联规则。 以下是有关这些节点功能的进一步注释。

Linear 对大数据构建模型时，通常希望将对象更改为非常大的数据集或者指定拆分。

- 不支持对现有 PSM 模型进行连续性培训。
- 仅当定义了 split 字段时才建议使用“标准”模型构建目标，从而每个拆分中的记录数量不会太大。其中“太大”的定义取决于 Hadoop 集群中个别节点的能力。 相比之下，您也需要注意确保勿将拆分定义得过于精确而使用于构建模型的记录过少。
- 不支持 Boosting 目标。
- 不支持 Bagging 目标。
- 如果记录很少，那么不建议使用“非常”大的数据集目标；它通常不会构建模型或不会构建降级的模型。
- 不支持 Automatic Data Preparation。 尝试用含有许多缺失值的数据构建模型时，可能会出现问题；通常，缺失值会插补到 Automatic Data Preparation 中。 变通方法是使用树模型或具有“高级”设置的神经网络来插补所选的缺失值。
- 未计算拆分模型的准确性统计。

Neural Net

对大数据构建模型时，通常希望将对象更改为非常大的数据集或者指定拆分。

- 不支持现有标准或 PSM 模型的继续培训。
- 仅当定义了 split 字段时才建议使用“标准”模型构建目标，从而每个拆分中的记录数量不会太大。其中“太大”的定义取决于 Hadoop 集群中个别节点的能力。 相比之下，您也需要注意确保勿将拆分定义得过于精确而使用于构建模型的记录过少。
- 不支持 Boosting 目标。
- 不支持 Bagging 目标。
- 如果记录很少，那么不建议使用“非常”大的数据集目标；它通常不会构建模型或不会构建降级的模型。
- 如果数据中存在许多缺失值，请使用“高级”设置来插补这些缺失值。

- 未计算拆分模型的准确性统计。

C&R Tree、CHAID 和 Quest

对大数据构建模型时，通常希望将对象更改为非常大的数据集或者指定拆分。

- 不支持对现有 PSM 模型进行连续性培训。
- 仅当定义了 split 字段时才建议使用“标准”模型构建目标，从而每个拆分中的记录数量不会太大。其中“太大”的定义取决于 Hadoop 集群中个别节点的能力。 相比之下，您也需要注意确保勿将拆分定义得过于精确而使用于构建模型的记录过少。
- 不支持 Boosting 目标。
- 不支持 Bagging 目标。
- 如果记录很少，那么不建议使用“非常”大的数据集目标；它通常不会构建模型或不会构建降级的模型。
- 不支持交互式会话。
- 未计算拆分模型的准确性统计。
- 存在拆分字段时，在 Modeler 本地构建的树模型与 分析服务器 所构建的树模型略有不同，因此会生成不同的评分。 这两种情况下的算法都有效；分析服务器 所使用的算法更新。 假定树算法倾向于具有许多启发式规则，两个组件之间存在差异很正常。

模型评分

针对所有支持建模的模型，同样支持评分。 此外，支持针对以下节点进行本地构建的模型块进行评分: C&RT、Quest、CHAID、Linear 和 Neural Net（无论模型是标准模型、带有 boost bag 的模型还是非常大的数据集的模型）、Regression、C5.0、Logistic、Genlin、GLMM、Cox、SVM、Bayes Net、TwoStep、KNN、Decision List、Discriminant、Self Learning、Anomaly Detection、Apriori、Carma、K-Means、Kohonen、R 和 Text Mining。

- 不会对原始倾向或已调整的倾向进行评分。 因为通过使用派生节点手动计算原始倾向，您可以获得相同的效果，表达式如下所示: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif
- 对模型进行评分时，分析服务器 不会检查模型中所用的所有字段是否都位于数据集中，因此模型在分析服务器 中运行之前，请先确保模型中所用的所有字段都位于数据集中。

R 块中的 R 语法应包含一次一记录操作。

输出 支持“矩阵”节点、“分析”节点、“数据审计”节点、“转换”节点、“统计”节点、“平均值”节点和“表”节点。以下是有关受支持的节点功能的进一步注释。

数据审计

“数据审计”节点无法为连续字段生成方式。

平均值 “平均值”节点无法生成标准误差或 95% 的置信区间。

表 通过编写包含上游操作的临时 分析服务器 数据源支持“表”节点。 然后“表”节点会翻阅该数据源的内容。

导出 流可以 分析服务器 源节点开始，并以导出节点而非 分析服务器 导出节点结束，但是数据会从 HDFS 移动到 SPSS Modeler Server，并最终移动到导出位置。

最佳实践

回送至 HCatalog/Hive

在处理分区 Hive 表中的数据时，您可以构造 Modeler 流以将期望的分区选择回送至 Hive。

1. 通过引用 HCatalog/Hive 数据源的 分析服务器 源节点开始流。
2. 连接到仅针对用作 Hive 表中的分区字段的字段选择记录的“选择”节点。如果在此“选择”节点的表达式中引用不用作分区字段的字段，那么不会将流回送至 HCatalog/Hive。
3. 像平时一样连接到其他节点。

第 3 章 故障诊断

本节描述了一些常见用法问题以及它们的解决方法。

数据源

不支持在 HCatalog 数据源中的分区列中定义的过滤器

这是在某些版本的 Hive 中遇到的问题，并且可能会在下列情况下发生。

- 您定义了 HCatalog 数据源并在数据源定义中指定了过滤器。
- 创建的 Modeler 流包含引用分区表列的“过滤”节点。

变通方法是将“派生”节点添加至 Modeler 流，此流将创建具有等于分区列的值的新字段。 “过滤”节点应该引用此新字段。

声明

本信息是为在美国国内供应的产品和服务而编写的。

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

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

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

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

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

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

本信息可能包含技术方面不够准确的地方或印刷错误。 本信息将定期更改；这些更改将编入本信息的新版本中。IBM 可以随时对本出版物中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。 那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

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

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

IBM Software Group
ATTN: Licensing

200 W. Madison St.
Chicago, IL; 60606
U.S.A.

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费，都可获得这方面的信息。

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

此处包含的任何性能数据都是在受控环境中测得的。 因此，在其他操作环境中获得的数据可能会有明显的不同。有些测量可能是在开发级的系统上进行的，因此不保证与一般可用系统上进行的测量结果相同。 此外，有些测量是通过推算而估计的， 实际结果可能会有差异。 本文档的用户应当验证其特定环境的适用数据。

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

所有关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

所有 IBM 的价格均是 IBM 当前的建议零售价，可随时更改而不另行通知。 经销商的价格可与此不同。

本信息仅用于规划的目的。在所描述的产品上市之前，此处的信息会有更改。

本信息包含在日常业务操作中使用的数据和报告的示例。 为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。 所有这些名字都是虚构的，若现实生活中实际业务企业使用的名字和地址与此相似，纯属巧合。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明:

本信息包含在日常业务操作中使用的数据和报告的示例。 为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。 所有这些名字都是虚构的，若现实生活中实际业务企业使用的名字和地址与此相似，纯属巧合。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明:

©（贵公司的名称）（年）。 此部分代码是根据 IBM Corp. 公司的样本程序衍生出来的。

如果您正在查看本信息的软拷贝，图片和彩色图例可能无法显示。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。 在 Web 站点 www.ibm.com/legal/copytrade.shtml 上的“Copyright and trademark information”部分中提供了 IBM 商标的最新列表。

Adobe、Adobe 徽标、PostScript 以及 PostScript 徽标是 Adobe Systems Incorporated 在美国和 / 或其他国家或地区的注册商标或商标。

IT Infrastructure Library 是 Central Computer and Telecommunications Agency 的注册商标，该企业现已成为 Office of Government Commerce 的一部分。

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

Linux 是 Linus Torvalds 在美国和@3B72其他国家或地区的注册商标。

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

ITIL 是一个注册商标，是 Minister for the Cabinet Office 的共同体注册商标，并且已在 U.S. Patent and Trademark Office 进行注册。

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

Cell Broadband Engine 是 of Sony Computer Entertainment, Inc. 在美国和/或其他国家或地区的商标并且在当地许可证下使用。

Linear Tape-Open、LTO、LTO 徽标、Ultrium 和 Ultrium 徽标是 HP、IBM Corp 和 Quantum 在美国和其他国家或地区的商标。

Printed in China