Open for Innovation
KNIME

KNIME最佳实践指南

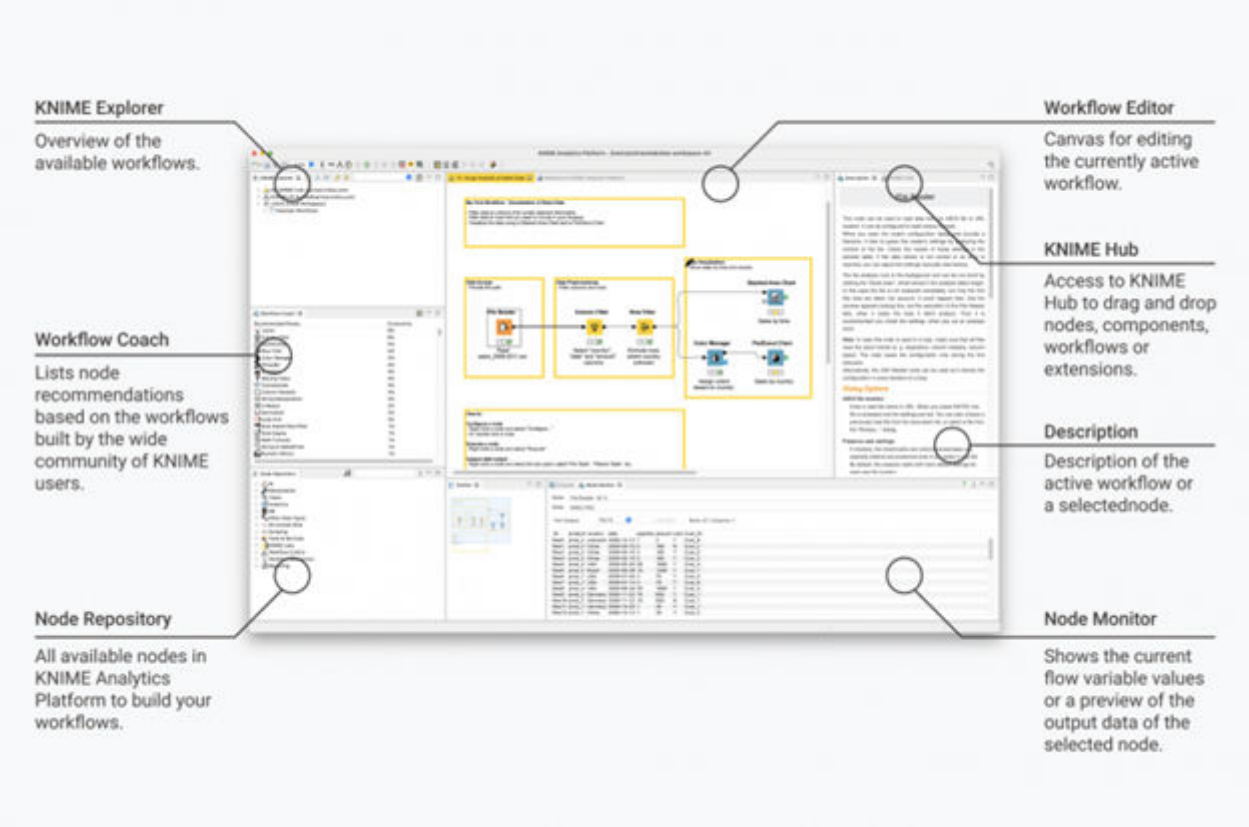
KNIME AG，苏黎世，瑞士
版本5.2（最后更新于2022年3月24日）

目录

什么是KNIME软件?

一个企业级软件平台，两个互补工具：用于创建数据科学的开源KNIME分析平台和用于生产化数据科学的商业KNIME服务器。

如果你是KNIME的新手，你可以通过内部目录或直接在KNIME网站上下载KNIME Analytics Platform。 我们建议你查看我们的入门指南，以构建你的第一个工作流程。



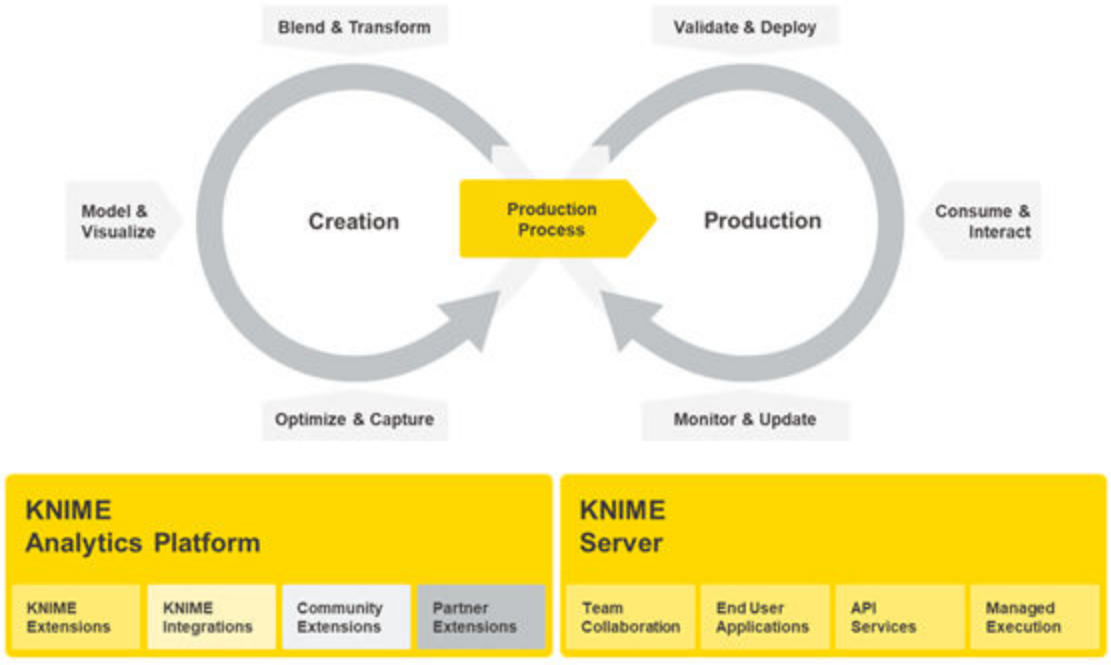
工作流程设计过程：从KNIME Analytics Platform到KNIME Server

如前所述，这些是互补工具。

KNIME Analytics Platform是我们的开源软件，用于创建数据科学。 直观、开放，并不断整合新的发展，使得理解数据和设计工作流程以及可重用组件对每个人都可访问。

KNIME Server是一款企业软件，用于团队协作、自动化、管理和部署数据科学工作流程，作为**KNIME** WebPortal应用程序、数据应用程序、**RES**T API等。

当你开始使用KNIME进行数据项目时，你将创建一个工作流程，然后可以将其上传到KNIME Server。通过工作流程，你可以设计你的数据处理方式。 一旦你的工作流程准备好，当你将其上传到KNIME Server时，它可以轻松地自动化或部署。



在构建KNIME工作流之前：项目先决条件

与任何项目一样，在构建工作流之前，了解你正在处理的范围非常重要。 这个清单可以指导你：

☐是否有明确定义的可衡量和可实现的目标?
例如：我们希望将流失率降低*15%*。

☐为达到目标需要哪些数据源? 你是否可以访问这些数据源?

☐谁应该使用你部署的数据科学应用程序，以及如何使用? 例如：你正在创建：

- ☐定期创建和发送报告的定期工作流程?
- ☐定期执行以处理新数据或应用预测模型的定期工作流程?
- ☐部署到KNIME WebPortal的数据应用程序?
- ☐通过REST API访问的预测模型?

☐你需要哪种类型的工作流程来实现你的目标?

- ☐你想构建的每个单独工作流的任务是什么?
- ☐每个单独工作流的输入和输出是什么?
- ☐你的工作流是否共享某些部分?

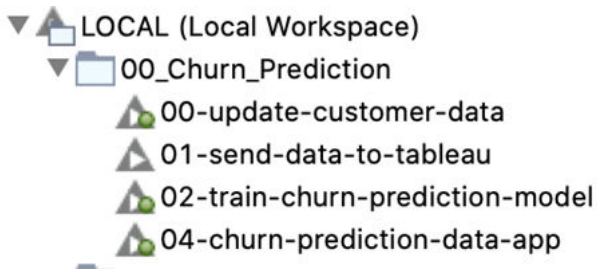
☐为创建的工作流定义要求（执行速度如何，执行频率如何等）。

☐您需要什么硬件和软件来满足所有这些要求?

在使用KNIME Analytics Platform时的最佳实践

为您的工作流或组使用适当的命名

为您的工作流使用清晰的命名约定。 从工作流名称中，应清楚地知道工作流正在做什么（例如，“项目_1”与“读取和预处理客户数据”）。如果一个工作流后面跟着另一个工作流，您可能希望使用数字作为前缀来引入顺序。



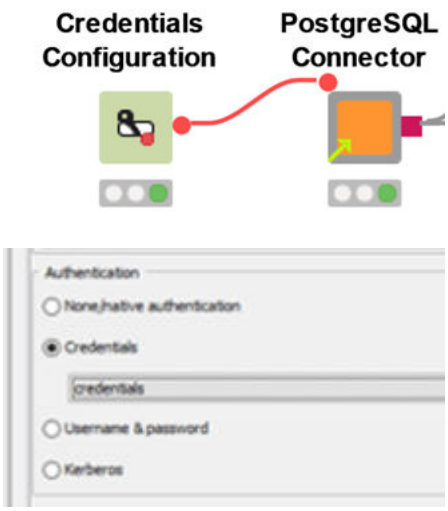
以安全、可重用和高效的方式设计您的工作流

避免使用本地文件路径

为了使您的工作流能够在您的计算机上或任何其他计算机上访问数据源，避免使用本地文件路径非常重要。 而是使用相对路径，或者如果您的数据已经在远程文件系统中可用，则直接在那里管理您的文件和文件夹。 为此，您可以使用动态文件系统连接端口和许多不同的连接器节点，例如SMB连接器或S3连接器节点。

使用凭证节点

不建议在节点和工作流中保存任何凭证或机密数据。 相反，使用凭证配置或凭证小部件节点，让用户在运行时指定凭证。



如果工作流通过KNIME服务器共享，执行者可以在启动或计划执行时提供凭证，可以通过WebPortal或执行窗口中的“配置”选项进行。 （右键单击KNIME资源管理器中的工作流并选择“执行”以打开它）。

构建和共享组件

组件可以帮助您和您的团队遵循“不要重复自己”的原则并实施最佳实践。 例如，如果您经常连接到某个特定数据库，可以创建一个包含所有设置选项的连接器组件，并共享该组件以在工作流中重用它或允许其他人使用它。

通过配置节点，您可以使您的组件更加灵活。 例如，您可以更改它，以便不必在组件内保存凭证，或允许其他用户更改某些设置。 如果您的目标是将工作流部署为KNIME WebPortal的DataApp，您可以使用小部件节点使其可配置。

为了使组件能够被他人重复使用，建议您适当地对其进行文档化，包括为预期输入和不同的设置选项添加描述（类似于KNIME节点）。 您可以通过进入组件并打开描述视图/面板来编辑组件描述。

KNIME组件指南为您提供了有关如何创建和共享组件的更多信息。

记录您的工作流程

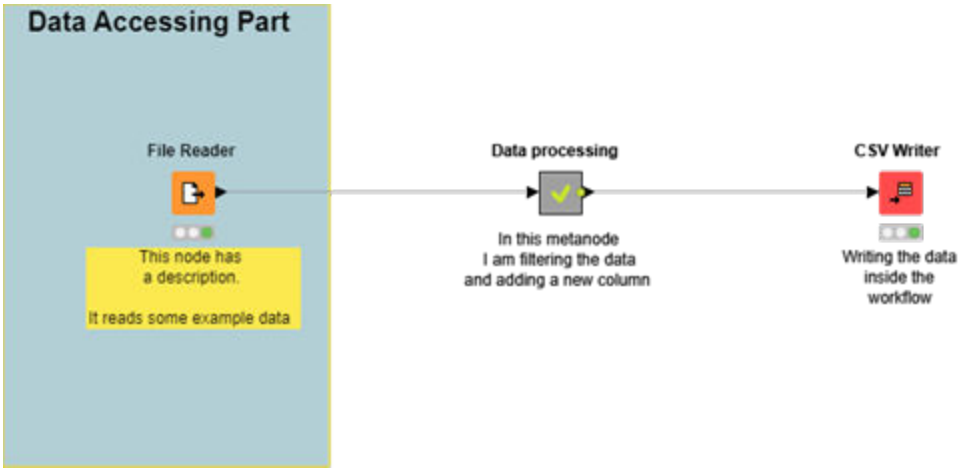
为了使您的工作流程能够被重复使用，您和您的团队需要能够快速理解它的功能。 因此，记录您的工作流程非常重要 - 如果它很大，您可以

应该将其分成不同的部分进行结构化。

为了构建一个大型工作流程，您可以使用元节点和组件来进行不同的部分，并将它们嵌套。 为了记录工作流程，您可以：

- 通过双击节点标签来更改节点标签，以定义各个节点的功能。

- 创建一个注释框。 只需在工作流程编辑器中右键单击任意位置，然后选择“新建工作流程注释”。然后输入您的解释性评论，调整注释窗口的大小以清楚地指明它所涉及的节点组，并使用上下文菜单格式化文本。



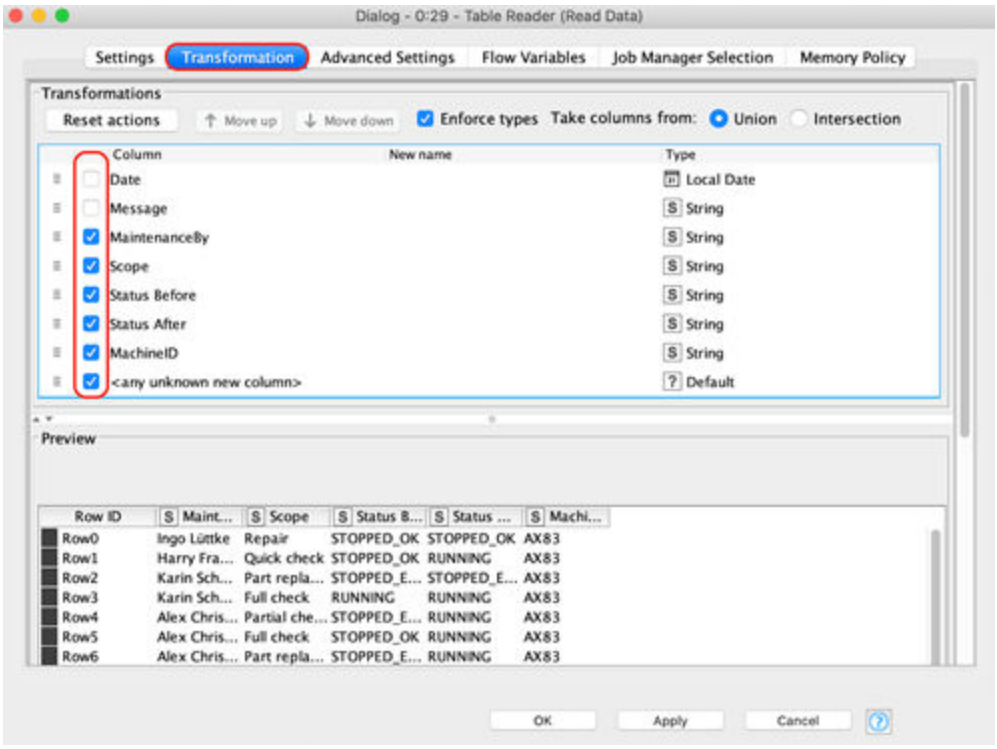
这个工作流是一个如何记录工作流程并添加描述和注释的示例。

- 通过点击工作流程中的空白区域，附加一个工作流程描述，然后进入节点描述视图，点击右上角的编辑按钮。

设计您的工作流以提高效率

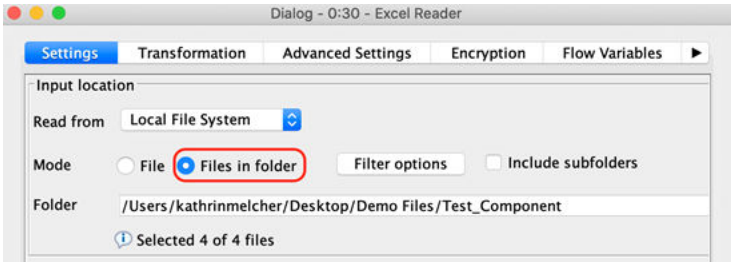
为了使工作流程的执行高效，最好遵循一些最佳实践，并考虑执行每个单独节点的计算密集程度。

- 在读取之前排除多余的列。 如果数据集的所有列都没有被使用，可以通过在读取节点的“转换”选项卡中排除这些列来避免读取它们。



- 为了使数据访问部分的工作流程高效，应避免多次读取相同的文件，而是将一个读取节点连接到多个节点。

- 使用“文件夹中的文件”选项来读取具有相同结构的多个文件。



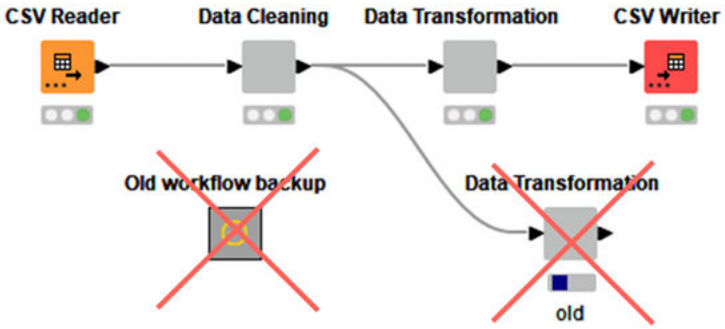
- 在执行昂贵的操作（如连接或数据聚合）之前，尽早删除冗余的行和列。

- 仅在绝对必要时使用循环。 执行循环通常是昂贵的。 尝试通过使用其他节点（例如字符串操作（多列）或数学公式（多列）节点）来避免使用循环。

- 尽可能将计算推送到数据库中。 如果你正在使用数据库，你可以通过在将数据读入KNIME Analytics Platform之前使用DB预处理节点将尽可能多的数据预处理工作推送到数据库服务器来加快工作流的执行速度。

- 尽可能将计算推送到数据库中。

- 使用DB连接关闭节点关闭数据库连接。
- 删除不需要的断开连接的节点或工作流分支。



为了进一步提高工作流程的性能，您可以使用计时器信息节点来查找执行单个节点所需的时间。

在使用KNIME服务器时的最佳实践

版本控制

在团队中共同构建和修改工作流程时，很容易通过覆盖同事的工作而丢失进展。 有时候您也会做出后来发现错误的更改，然后希望将工作回滚到以前的版本。 在编写代码时，通常会使用像GIT这样的版本控制系统来完成此任务。 这样的系统通过跟踪更改并确保不丢失重要工作来提高生产力。 关于KNIME服务器版本控制功能的视频可在YouTube上观看。

快照

KNIME服务器提供了自己的版本控制机制（请参阅我们的文档），该机制基于快照。

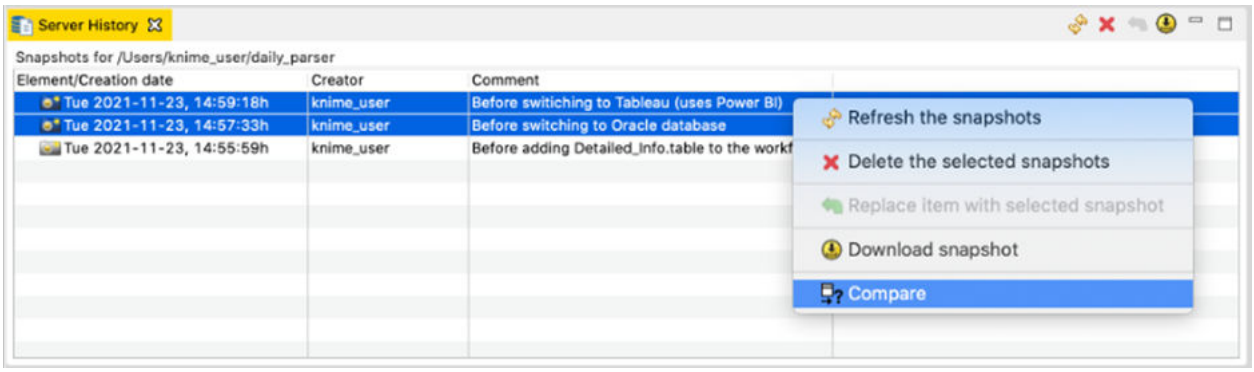
快照只是工作流程在某个特定时间点的副本。 可以通过在KNIME资源管理器中的项目上下文菜单中选择该选项来随时对服务器存储库中的项目进行快照。

此外，用户可以选择在覆盖项目时创建快照。 如果需要的话，服务器管理员还可以在每次覆盖操作时启用强制快照。

快照的一个重要部分是您为其编写的注释。在这方面，KNIME与其他版本控制系统有些不同，因此您需要小心放置什么。
在创建快照时，KNIME会要求您提供一个描述快照的小段文字。 这意味着您写的文字不应该描述所做的更改，而是快照中工作流的实际状态，在更改之前。

版本历史

除了快照外，存储库项始终具有一个当前状态。 这是您在KNIME资源管理器中看到的项目。 如果您想访问快照，您需要查看项目的历史记录。



在这里，您可以看到快照的列表，包括创建日期、时间和注释。 从这里，您可以通过选择两个工作流快照，然后右键单击并选择“比较”来进行比较。

如果您只选择一个工作流程的快照进行比较，则将其与当前版本进行比较。 除了比较，您还可以将快照下载到本地存储库中，删除快照，并使用快照覆盖当前状态。

管理对KNIME服务器项目的访问权限

当多个团队在同一个KNIME服务器上工作时，控制谁可以访问服务器存储库中的哪些项目非常重要。

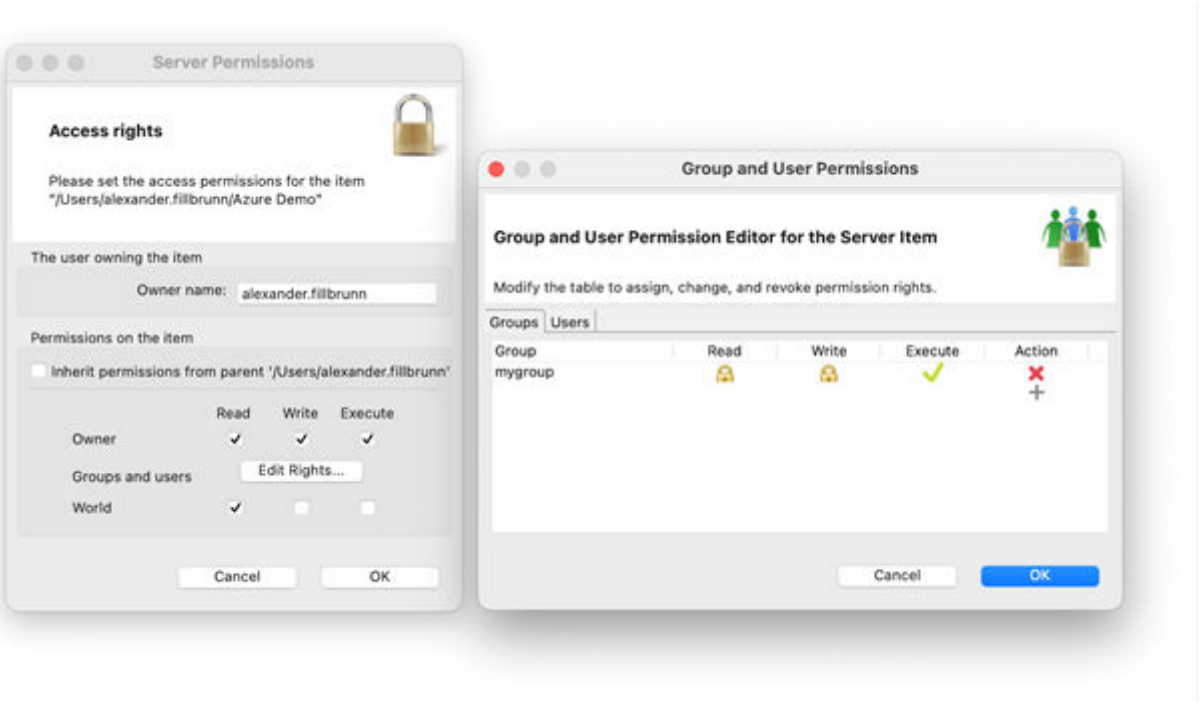
并非每个同事都应该能够看到您正在处理的内容，而其他一些同事可能会看到您的工作流程，但您不希望他们更改任何内容。 其他人只能执行您的工作流程，而无法看到底层的具体情况。 KNIME服务器具有基于组和个人用户的全面访问控制功能。

权限

存储库中的每个项目，无论是工作流程、工作流程组还是数据文件，都可以分配权限。 对于工作流程和工作流程组，这些权限涵盖了读取、写入和执行权限。 对于数据文件，只有前两个权限是相关的。

你可能会想为什么有人需要在工作流组上拥有执行权限。 这是因为，默认情况下，工作流组中的项目会继承该组上设置的权限。

因此，如果你给某个人在特定的工作流组上赋予执行权限，然后将一个工作流上传到该组中，那么该人就可以执行该工作流。 权限继承可以在每个项目上禁用，并替换为该项目特定的权限。



阅读、写入和执行权限不仅可以一次性控制每个用户，还可以控制每个个别用户和组。 为此，权限对话框提供了多个选项：“所有者”，“用户和组”和“全局”。

项目的所有者是最初的上传者，但所有者或管理员可以在以后更改。所有者是负责该项目的人，始终可以更改权限。 即使你禁用了所有者的阅读和写入权限，他们仍然可以执行这些操作。 然而，当拒绝执行权限时，所有者将无法执行工作流。

全局权限是适用于能够访问**KNIME**服务器的每个用户的权限。如果您想严格控制谁有访问权限，最好禁用此行中的所有复选框，然后为个别用户和组启用访问权限。 您可以通过点击“查看权限...”按钮来执行此操作。 然后，将打开一个新窗口，让您输入自定义用户和组（在相应的选项卡中）及其权限。 只需在新行中输入用户或组，然后单击列以授予执行读取、写入和执行操作的权限。

分配权限的最佳实践

通常，将权限设置得尽可能狭窄是有意义的，但最需要严密保护的权限是写入权限。 有了写入权限，用户可以更改和覆盖您的数据和工作流，如果他们没有相关业务，很容易造成混乱。

理想情况下，您的团队在服务器存储库中有一个只有他们拥有写入权限的文件夹。
权限继承将确保默认情况下，其他人无法更改任何内容。

但有时候你也想保护你的知识产权，或者你想避免别人复制你的工作流并在服务器上进行修改，导致项目的碎片化。 在这种情况下，最好完全关闭你的项目文件夹，甚至禁止外部人员读取权限：取消“World”下的所有复选框，然后在“用户和组”设置中明确为你的团队启用访问权限。

远程工作流编辑器

KNIME工作流通常在KNIME Analytics平台上本地开发，这样可以很容易地在每个步骤之后检查是否产生了预期的结果，并且每个节点都是绿色的。 但是一旦工作流完全完成，它就会部署到KNIME服务器上按计划执行，可以通过REST API或WebPortal进行。

如果工作流构建得完美无缺且从不失败，就没有太多理由去检查节点状态或节点的中间输出。 不幸的是，节点偶尔会失败，原因有很多——可能是意外的数据、网络故障或其他无数的原因。 为了快速解决这些问题，KNIME提供了一种查看在KNIME服务器上运行的作业的方式，就像在本地运行一样。 允许这样做的扩展名为远程工作流编辑器。

安装远程工作流编辑器

远程工作流编辑器需要通过“文件 →安装KNIME扩展...”手动安装。只需在搜索框中输入“远程工作流编辑器”，然后勾选相应的扩展进行安装。 安装编辑器后，您可以在KNIME资源管理器中双击任何作业（执行中的工作流）以在工作流编辑器中打开它。

您会注意到编辑器顶部有一个横幅，告知您这是在KNIME服务器上运行的作业。您现在可以查看节点设置和输出，以及错误和警告消息。 如果作业不是从KNIME WebPortal启动的，您甚至可以重置和重新执行节点，更改连接，并添加全新的节点。 对于WebPortal作业，这是不可能的，因为它会干扰在其浏览器中访问相同作业的用户。

远程工作流编辑器的使用案例

KNIME客户使用远程工作流编辑器进行各种操作。 有些人无法访问本地设备上的数据库和文件共享，因此他们将部分完成的工作流上传到服务器，通过右键单击 → “打开” → “作为新作业在服务器上”来执行作业，然后进行必要的配置更改以使其读取数据。一旦数据准备好，他们就可以在本地继续工作。

加载完成后，包括数据在内的作业可以通过在KNIME资源管理器中右键单击并选择“另存为工作流...”来保存在服务器上。然后可以下载此保存的工作流进行进一步编辑。

如何团队合作?

使用组件在同一项目上工作

在处理较大的项目时，您可能希望多个团队成员同时在同一项目和/或工作流上工作。 为此，您可以将整个工作流分成多个部分，每个部分都在组件内实现，您可以定义任务/范围以及预期的输入和输出列。

例如，如果您正在进行预测项目（例如客户流失预测或潜在客户评分），其中有多个数据源（例如活动记录、网站互动、电子邮件数据或合同信息），您可能希望基于每个数据表生成一些特征。 在这种情况下，您可以将预处理拆分为多个具有清晰描述的组件 - 基于输入数据集应创建哪些特征以及期望的输出（例如“ID列生成的特征”或类似）。

然后每个团队成员可以分别处理这些任务，将工作封装成一个组件，并与团队共享。 最后，不同的步骤可以合并成一个使用正确顺序的工作流程，其中使用了共享的组件。

另一个好处是个别组件可以被其他工作流程重复使用。
因此，如果你有共享组件来准备客户数据进行流失预测模型，那么在处理下一个最佳步骤模型时可以重复使用它们。

构建（共享）组件的最佳实践

构建共享组件时需要记住一些事情，以便使它们易于重复使用。 最好以KNIME节点的期望行为方式构建共享组件。这意味着它应该可以通过配置窗口进行配置，具有信息丰富的描述，并在发生错误时给出有意义的错误消息。

使用配置节点使共享组件可配置

通过使用配置节点，您可以为组件创建一个配置窗口。

这样可以通过组件的配置窗口控制单个节点的设置选项。

因此，如果组件中的某个节点有一个设置选项，您或者组件的用户将来可能想要更改，您可以使用配置节点来允许通过组件的配置窗口定义此设置选项。

组件指南的这一部分介绍了不同的配置节点及其用法。

创建组件描述

在配置KNIME节点时，描述提供有关节点的一般功能、输入和输出端口以及不同的设置选项的信息。 您可以为您的组件创建类似的描述。

要添加一般描述，您可以首先在组件的任何空白处单击，然后单击描述视图中的铅笔图标。

输入和输出端口的名称和描述可以在同一视图的下部定义。

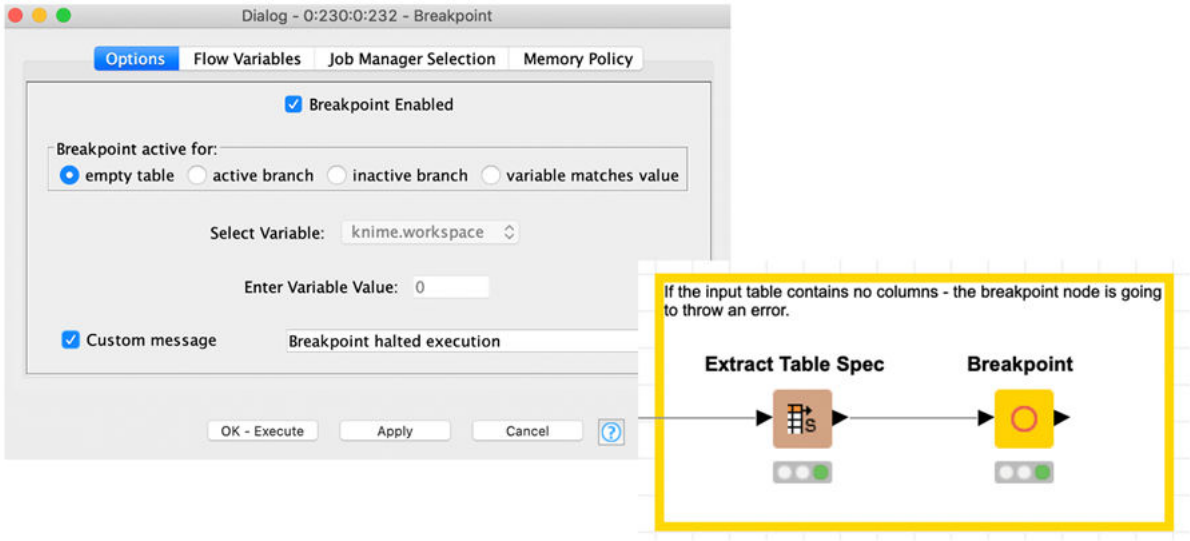
提示：在描述视图中，您还可以通过节点类型选择更改组件的背景颜色，并添加图标。

可以在相应配置节点的配置窗口中定义每个配置选项的描述。

创建自定义错误消息

断点节点允许您在满足指定条件的情况下停止执行并输出自定义错误消息。 因此，例如，如果输入表为空或某些选定的设置导致组件失败的表格，则可以提前停止执行并通过自定义消息通知用户您的组件。

在配置窗口中，您可以定义何时停止组件的执行，例如当输入表为空、节点是活动或非活动分支，或者流变量与定义的值匹配时。 可选地，在下部分您可以定义要显示的自定义消息。

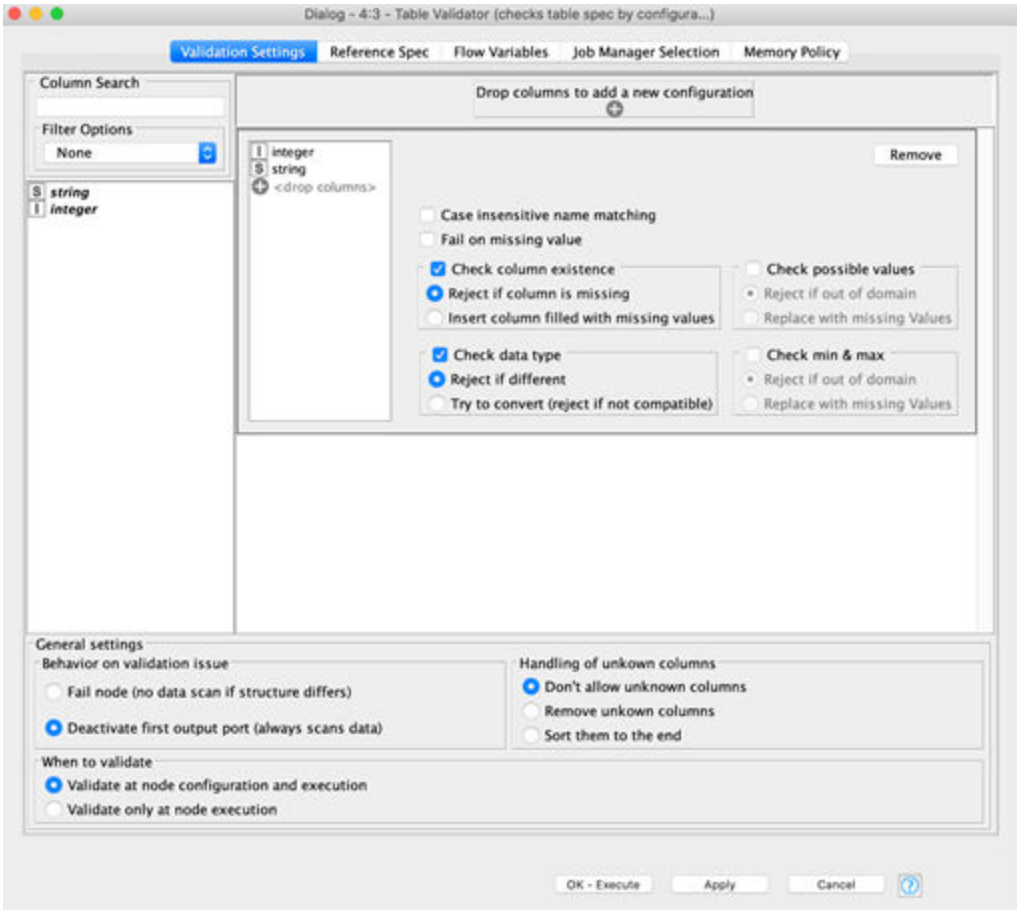


验证输入数据

表验证器节点允许您检查表中是否存在所有必要的列。
在组件中，此节点可用于检查是否存在所有必要的输入列，并通过断点节点输出自定义错误消息，以防缺少必要的列。

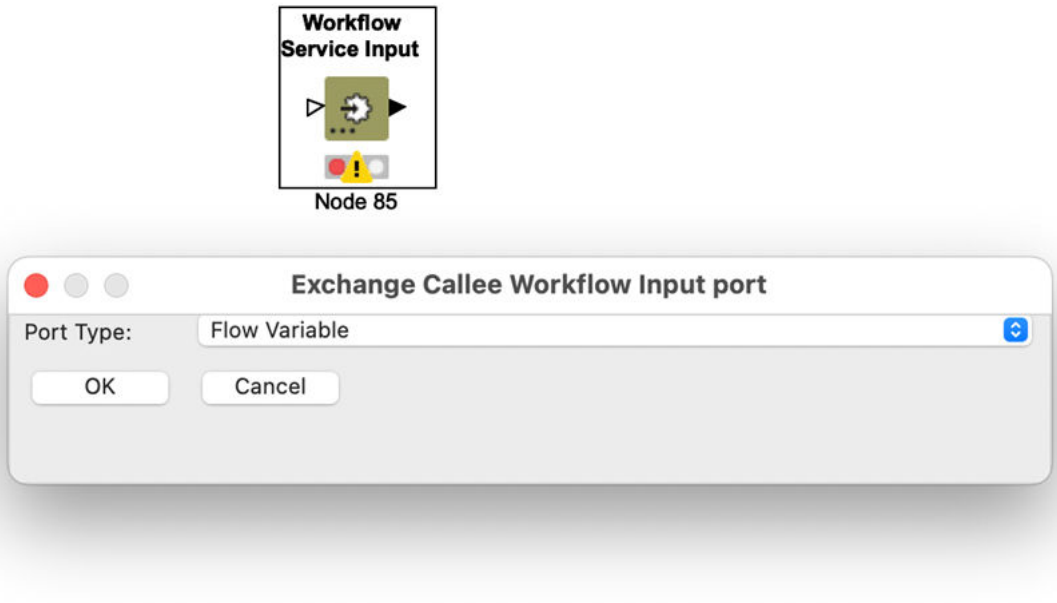
在配置窗口中，您可以指定必要的列以及

每个列需要满足的规范，例如是否存在足够的列，是否需要具有额外的数据类型，是否不允许有任何新值，或者范围是否必须与设置节点时使用的模板表相同。



工作流服务（从另一个工作流中调用工作流）

KNIME Analytics Platform中的一个新引入的功能是所谓的工作流服务。
与组件类似，工作流服务具有明确定义的输入和输出。 然而，与组件不同的是，工作流服务是独立的，并且通过将工作流服务输入或工作流服务输出节点放入工作流中来定义每个输入和输出。 节点具有动态端口，因此您可以通过单击节点内部的三个点来更改其类型。 要使用工作流服务，您可以使用调用工作流服务节点。 一旦您将其指向具有工作流服务输入和输出节点的工作流，调用工作流服务节点可以根据当前的输入和输出进行自适应，并相应地调整自己的端口。 现在您只需要连接它，当您运行它时，它所指向的工作流将被执行。
您的工作流数据被传输到被调用的工作流中，然后它返回其结果。



工作流服务与组件

那么何时应该使用工作流服务，何时应该使用组件？ 后者在添加到工作流时会被复制。组件所组成的所有节点现在都在你的工作流中，并具有自己的配置。 从存储库中删除组件不会改变工作流的行为。另一方面，工作流服务仅被引用 - 调用工作流服务节点告诉工作流要调用什么，但将要执行的确切节点在工作流服务中是隐藏的。 这样做有优点和缺点：由于包含较少的节点，因此工作流会更小，但工作流服务的发布者可以在不被察觉的情况下更改其实现。使用共享组件时，如果引用组件发生了变化，您可以决定是否要更新它们。 这在工作流服务中是不可能的。

组件也将作为所属工作流的一部分执行。另一方面，工作流服务在其所在位置执行。 在本地存储库中的工作流服务将在本地执行，但如果它位于KNIME服务器上，则服务将在服务器上运行。这意味着您可以将工作从本地工作流转移到服务器上，或者同时触发多个工作流服务，如果服务器具有用于并行运行工作流的分布式执行器，那么这些工作流服务可以非常快速地处理您的数据。

工作流服务与调用工作流和容器输入输出

除了工作流服务外，调用工作流和各种容器输入和输出
节点已经存在一段时间了。 这些和新工作流的区别

服务是指在KNIME内部使用，而前者可以通过KNIME服务器的REST API从第三方应用程序中调用。那为什么不把所有东西都用Call Workflow呢？ 这些节点通过首先将数据转换为JSON格式来传输数据，JSON是一种文本-基格式，无法处理所有KNIME数据类型，而且相当大。 另一方面，工作流服务使用KNIME内部使用的专有二进制格式来表示其数据。 这意味着不需要转换，数据大小非常小。 这导致更快的执行速度，但代价是无法理解数据格式的第三方应用程序无法使用。

工作流服务的使用案例

当你想将一个项目分解成较小的可重用工作流时，工作流服务总是一个好主意。 当在工作流中使用许多组件时，这可能会使工作流变得非常庞大，导致工作流加载缓慢。 将逻辑放入被调用的工作流中可以显著减小工作流的大小，从而减少加载时间。

工作流服务的一个特定用例是为KNIME Analytics平台的其他用户提供将计算密集型任务转移到KNIME服务器的能力。当某人在本地构建工作流，但希望训练类似深度学习模型这样的任务，却没有强大的GPU来高效完成时，他们可以将数据发送到一个工作流服务，该服务接收数据和超参数作为输入，训练模型，并输出最终结果。

但工作流服务也可以用于提供对其他数据源的访问。 只能从服务器获取的数据可以通过工作流服务向客户端公开。 由于服务可以在工作流中进行检查和转换，因此很容易以不发送任何机密数据的方式构建它。 例如，您可以在将数据提供给客户端之前对数据进行匿名化，或者可以过滤掉某些行或列。 工作流服务实际上充当实际数据源和客户端之间的抽象层。

术语表

工作流注释

一个带有彩色边框和可选文本的框，可以放置在工作流中以突出显示一个区域并提供额外的信息。 在工作流中的空白处右键单击，选择“新建工作流注释”以在该位置创建一个新的工作流注释。

节点注释

节点下方的文本。 默认情况下，只显示“节点 <id>”，其中<id>是放置在工作流中的节点的连续编号。 通过点击文本，用户可以编辑它并添加节点的功能描述。

工作流描述

可以由用户编辑的工作流描述。 当没有选择节点时，显示在节点描述的位置。 点击右上角的铅笔图标进行编辑。当用户打开工作流的页面时，工作流描述也会显示在KNIME Server WebPortal中。

服务器存储库

存储工作流、组件和数据文件的服务器位置。 可以使用权限保护存储库中的项目，防止未经授权的访问。

作业

一个用于执行的工作流的副本。 作业只能由其所有者，即启动它的用户访问。有关作业的更多信息可以在这里找到。

（共享）组件

将工作流片段包装成一个单一节点。 通过将其保存在服务器存储库中，可以与其他用户共享。 如果组件包含视图或小部件节点，则它提供自己的视图，其中包含其中所有视图。如果它包含配置节点（例如字符串配置），用户可以通过组件的配置对话框传递值。更多

信息可以在这里找到。

数据应用

包含一个或多个包含用户可以在浏览器中与之交互的视图和小部件节点的工作流。 数据应用程序部署在KNIME服务器上，并通过WebPortal启动。有关构建数据应用程序的指南可以在这里找到。

计划

执行特定工作流的计划。 一旦设置完成，KNIME服务器将确保在配置的时间执行工作流，可以是仅一次或重复执行。 更多信息请参阅此处。

REST API

使用HTTP调用（GET、POST、PUT、DELETE等）访问KNIME服务器功能的接口。REST代表表征状态转移，限制了万维网上的接口行为。 API代表应用程序编程接口，通常用于允许机器之间进行通信。 在KNIME服务器上，可以使用REST API来管理工作流和其他文件（上传和下载、移动、删除），创建计划并触发工作流的执行。 KNIME Analytics Platform也通过REST API与KNIME服务器通信，因此基本上您可以在AP中做的任何事情，也可以通过REST来做。更多信息可以在这里找到。

工作流服务

一个带有Workflow Service Input和Workflow Service Output节点的工作流，可以使用Call Workflow Service节点从其他工作流中调用。

KNIME AG
Talacker 50
8001 Zurich, Switzerland
www.knime.com
info@knime.com