目录

数据源简介

工作流程 数据源这一模块为用户提供了基础的数据连接和数据处理功能，是数据分析与可视化的前置工作，用户可以在此完成数据源的连接，数据的清洗与数据模型的构建。

建立连接：获取数据源连接的相关信息，保证产品可以读取到数据表 数据处理：根据分析需求做表的合并关联，数据清洗、数据聚合、数据转置等字段级处理 数据落库：将编辑好的数据加载到数据库的过程，用户可以直接使用数据表进行后续分析 数据模型：直接基于现有落库表制作数据模型，为后续分析作准备 **模块简介** 数据连接 数据连接中用户可以连接数据源。详见连接数据 连接数据 数据准备 数据准备中用户可以进行数据处理与数据落库。详见数据准备 数据模型 数据模型中用户可以建立模型。详见建立模型

连接数据

数据源

产品目前支持的数据源有： 在使用 Elasticsearch 数据源时，需要注意问题如下： ① 要求 Elasticsearch 版本为6.3以上，并且开启了 xpack 的 SQL 支持 ② 针对 Elasticsearch 数据源的认证方式有数支持两种：无/用户名密码 ③ 只有 mapping 中每个 field 与 type 一一对应，与 RDB 中建模相同形式的 index 才能够在有数中使用

例如： { "mock_table_1": { "mappings": { "properties": { "A": { "type": "keyword" }, "SA": { "type": "long" }, "T": { "type": "date" } } } } } ④ 若mapping 中包含嵌套的结构，有数无法明确获取其 schema， 这样的 index 便不能在有数中使用

例如： { ".kibana_1": { "mappings": { "doc": { "dynamic": "strict", "properties": { "config": { "dynamic": "true", "properties": { "buildNum": { "type": "keyword" } } }, "updated_at": { "type": "date" } } } } } }

新建数据连接

在分析开始之前，我们需要做的第一步就是连接数据，如下图所示（以连接MySQL数据库为例）,对具体数据源的支持情况可以在数据源连接中查看 下面我们分步骤介绍：

1、在“数据源”模块，添加数据连接 2、选择要添加的数据连接类型 3、填写数据库信息后保存即可完成数据连接的添加 其中，用户可以对缓存有效期和查询队列并发数进行设置：

缓存有效期 为了提升访问性能，访问报告时，有数会对查询到的结果数据进行缓存，以便下次访问相同数据时加载速度更快。“缓存有效期”则指定了缓存在有数系统内存留的时间。比如缓存有效期设置为1小时，则首次访问报告会进行缓存，1小时内再次访问相同报告时会直接读取缓存数据，1小时之后再访问报告，缓存已经失效，会重新访问数据库获取最新的数据并重新进行缓存。 因此，为了提高访问性能，同时又要保证数据的时效性，我们建议缓存有效期的设置跟所连接的数据库的更新周期保持一致。（比如所连接的MySQL数据库会在每日的凌晨6点更新数据，则可以将该数据连接的“缓存有效期”设置为“1天”，“缓存失效点”设置为“06时00分”） 查询队列并发数 开启查询并发数可以设置总并发数和高优先级查询并发数，关系为：总并发数 = 高优先级查询并发数 + 普通查询并发数（其中重点用户和重点报告查询为高优先级查询，其他为普通查询）。高优先级查询并发为重点用户和重点报告查询预留，普通查询不占用高优先级查询并发，高优先级查询可以使用所有查询并发（优先使用高优先级查询并发）。

查看连接表信息

已建立的数据连接会显示在数据连接列表中，除了基本信息，我们还可以在有数系统中查看选中的数据连接的“表信息”、“字段分类”、“相关内容”、“操作记录”。

1、表信息：会显示选中数据连接的所有数据表。数据表分为“原始表”跟“自定义表”两种类型，“原始表”指的是数据库中已存在的表，“自定义表”指的是在网易有数内通过输入SQL建立的自定义视图（注：自定义视图只存储SQL逻辑，查询的数据不会落库存储）。 点击数据表的名称，可以在弹出的窗口中预览数据表的数据。 另外用户可以进行字段设置，抽取设置：

字段设置 字段设置指的是支持对数据连接下表粒度的字段配置，通过一次配置将同步至所有相关模型，提高配置效率。

抽取设置 抽取设置指的是将数据库中的表抽取至有数提供的速度更快的MPP内存数据库，具体可以查看 数据连接方式 2、字段分类：字段分类中支持对表添加描述 3、相关内容：在“相关内容”可以查看基于该数据连接建立的数据模型跟报告，点击名称可以快速跳转至对应的数据模型或报告。 4、操作记录：会显示用户对该数据连接的操作记录。我们会记录的行为包含以下几种：添加/修改数据连接、添加 / 修改 / 删除 / 暂停抽取任务、添加 / 修改 / 删除自定义表。

重要说明：通常，您需要使用AxisBI服务访问到您的数据库，您需要将AxisBI的服务IP: **120.92.12.165**添加到您的白名单中，以确保双方网络通畅。

数据准备

入门指南

产品整体示意 承接数据连接，在数据准备中，用户导入数据源后通过在画布上拖拽节点和 名词解释 画布：节点、数据流所处的操作空间，用户可以通过连接在这个空间进行编辑操作。 连接：两个节点之间的线段，代表输入关系，连线左侧节点是右侧节点的输入。 节点：节点是用来标注数据的某一个处理过程 ，用户通过节点编排数据流实现数据处理。 视图：计算机数据库中的视图，是一个虚拟表，其内容由查询定义。同真实的表一样，视图包含一系列带有名称的列和行数据。 发布：意味着将流程从开发模式提交到线上模式，对数据可以进行后续建模处理。

工作区 数据准备中工作区分为三段。左侧为数据连接，用户可以在此添加数据连接；右上为画布，用户可以通过连接和节点在这个空间进行编辑操作；右下为节点视图，点击具体的节点即可展示，用户可以在此进行节点内操作或通过视图查看数据。

1. 画布操作 从左侧拖入表生成输入节点 从节点新建节点 拖拽生成关联节点

拖拽建立连接

2. 视图 在数据准备中，一共提供了三种节点内视图，分别为字段视图、统计视图和数据视图。

字段视图仅展示字段，不展示数据。除输入、输出节点外，支持新增计算字段，支持右击字段唤起清洗：重命名、转换数据类型、值替换、数据筛选、复制字段与隐藏。 统计视图通过柱状图详细地展示了每个字段的统计信息：数值、行数与 占比。用户可以通过排序更直观地观察数据的形状。 在输入和输出节点，用户只能查看各个字段的统计信息 除输入和输出节点，用户可以新建计算字段，单击单个值进行值替换，在更多进行值筛选和值替换 数据视图以二维方式展示详细的数据信息。

输入

将原表或自定义SQL视图拖入画布。 **抽取模式下，输出节点依赖于输入节点的抽取任务，请保证输出节点执行时，输入数据已抽取完成**

在输入配置处，可以设置数据源的连接方式。包括抽取和直连。选择抽取后，点击抽取设置，允许设置抽取方式、抽取引擎、高级设置、添加定时任务。 在输入表处，显示数据源、数据库和表名。 在字段视图处，可以看到表的字段名称和注释。 在数据视图处，可以看到表的列和数据。 在统计视图处，可以看到每个字段的统计信息

数据处理节点

清洗 去掉数据表中不需要的列和行，并新增需要的列和行。 重命名：点击字段的下拉按钮，选择“重命名”，可以对字段进行重新命名。 转换数据类型：点击字段的下拉按钮，选择“转换数据类型”，可以将字段类型转换为整数、小数、字符串、日期、日期时间。 数据筛选：点击字段的下拉按钮，选择“数据筛选”，手动输入要添加的项，可以选择包含所选项以及排除所选项，点击“确定”后，满足条件的结果将会展示在数据视图中。 复制字段：点击字段的下拉按钮，选择“复制字段”，新复制的字段与数据将会展示在数据视图中。 隐藏：点击字段的下拉按钮，选择“隐藏”，字段与数据将不会显示在数据视图中。 **关联** 将两张表关联为一张宽表，并进行需要的数据处理，关联的数据在列上扩展。 建立两表间的关联关系有两种方式： 将要关联的表直接拖入已有的表中，选择关联的图标，建立两张表的关联关系。 点击已有的表，弹出“+”按钮，点击按钮，选择“关联”，将要关联的表拖入关联节点，建立两张表的关联关系。 建立关联关系后，选择两表要关联的字段，关联关系包括“等于”、“不等于”、“小于”、“小于等于”、“大于”、“大于等于”，可添加多个关联字段。 有数提供4种关联类型：内关联、左关联、右关联、外关联。 内关联：使用内关联时，生成的表将包含与两个表均匹配的值。 左关联：使用左关联时，生成的表将包含左侧表中的所有值以及右侧表中的对应匹配项。当左侧表中的值在右侧表中没有对应匹配项时，将在数据视图中看到null值。 右关联：使用右关联时，生成的表将包含右侧表中的所有值以及左侧表中的对应匹配项。当右侧表中的值在左侧表中没有对应匹配项时，将在数据视图中看到null值 外关联：使用完全外部关联时，生成的表将包含两个表中的所有值。当任一表中的值在另一个表中没有匹配项时，将在数据视图中看到null值。 两个关联的表中如果有相同的字段，将自动对字段进行重命名。 设置关联关系与关联类型之后，可以在右侧查看字段试图和数据视图。 **聚合** 根据选定的维度，在指定的度量上做数据汇总或平均。 分组：拖入字段，数据视图中将根据字段进行分组展示。 聚合：拖入字段，可以选择聚合方式。 拖入维度字段时，可供选择的聚合方式包括计数和去重计数。拖入度量字段时，可供选择的聚合方式包括求和、平均值、中位数、计数、去重计数、最小值、最大值、百分位。 用户也可以选择自定义聚合。 **行转列** 将表中具有相同值的多行数据转换成一个值的多列数据。 转置字段：拖入需要转置的字段。 聚合：拖入字段，可以选择聚合方式。

- 拖入维度字段时，可供选择的聚合方式包括计数和去重计数。拖入度量字段时，可供选择的聚合方式包括求和、平均值、中位数、计数、去重计数、最小值、最大值、百分位。
- 用户也可以选择自定义聚合。 **合并** 将两张表合并为一张表，合并的数据在行上扩展。
- 建立两表间的合并关系有两种方式：将要合并的表直接拖入已有的表中，选择合并的图标，建立两张表的合并关系。
- 点击已有的表，弹出“+”按钮，点击按钮，选择“合并”，将要合并的表拖入合并节点，建立两张表的合并关系。
- 建立合并关系后，可以选择合并主表，主表的结构将作为合并的依据，与主表一致的字段将自动合并。
- 字段视图中，展示主表的字段。同样点击字段的下拉按钮，可以对字段进行设置，包括：重命名、转换数据类型、数据筛选、复制字段、隐藏以及新建计算字段。

- 数据视图中，可以展示两张表合并后的列和数据。主表中的字段和数据将全部展示，次表中相同的字段对应的数据将在行上进行扩展，不同的字段与数据不会展示。 **输出** 将已经处理好的数据执行落库操作，用户可以选择内部输出节点或外部输出节点至数据库。内部输出将数据抽取至内部数仓，外部输出将数据输出至外部数据库表。

在内部输出节点配置，可以设置输出表名和数据更新方式。 在外部输出节点配置，可以设置输出连接、数据库、输出表和数据更新方式。输出表和数据库字段必须完全匹配。目前支持Doris、GreenPlum、ClickHouse、mysql四种数据源输出，用户需要先行建立数据连接连接数据库。 - **数据更新：**

1. 全量覆盖：每次抽取，对数据库的全部数据进行抽取，并覆盖数据库已有的数据。
2. 全量追加：每次抽取，对数据库的全部数据进行抽取，并追加在数据库中。
3. 增量更新：每次抽取，根据增量字段判断数据库中的数据是否为新增数据，对数据库的新增数据进行抽取，并追加在数据库中。
4. 增量滚动更新：每次抽取，根据日期及滚动周期将数据库中的新增以及部分历史数据抽取到数据库中，其中历史数据将会覆盖原数据。
    - 视图配置：在字段视图处，支持对字段添加注释。在数据视图处，可以查看输出表的列和数据。在统计视图，可以查看字段的统计信息，支持查看各个成员的行数与百分比。

- 用户可以在项目中心的数据任务管理中进行输出管理，具体查看 数据任务管理 **发布** 发布 用户在完成数据准备工作后，需要将流程进行发布，才能进入到执行计划的编辑、数据模型的新建。

用户也可以选择暂时不进行发布，对当前的流程进行保存。 预览态 数据准备为用户提供了两种预览态，分别为开发模式、线上模式，用户可以在开发模式查看已保存的流程，再线上模式查看已发布的流程。

- 开发模式支持编辑和发布功能。对流程的任意编辑都可以保存至开发模式，编辑中若存在节点异常，用户只能保存流程而不能进行发布。
- 线上模式呈现发布后的流程，支持新建数据模型功能。用户也可以在线上模式的输出节点中设置执行计划 **执行计划** 完成流程编辑发布后，用户可以在线上模式的输出节点对执行计划进行配置，支持立即执行、编辑执行计划与查看执行记录。 在编辑执行计划中，可以添加定时任务，支持设置依赖执行、任务频率、任务日期、任务时间、开始日期和终止日期。

建立模型

新建模型

连接完数据、完成数据准备后，用户便可以将需要的多张数据表关联成一张表，并进行需要的数据处理，建立数据模型以进行后续的数据可视化分析工作 下面我们分步骤介绍：

1、在“数据源”模块，添加数据模型 2、选择需要的数据连接，基于该连接建立数据模型 3、选择需要的一张或多张数据表，若选择多张数据表，则需要关联成一张宽表 拖入两张表时，若它们在原数据库中存在外键关联，则会自动进行关联；若无外键，系统会自动将两张表中相同名称的列设置为外键进行关联。用户也可以手动添加或修改“关联字段”。

完成关联后，下方会显示宽表中的所有字段，并将字段划分为维度、度量两种类型进行展示。 当基于数据库（比如MySQL、Oracle）类型的数据连接建立数据模型时，可以在有数内通过SQL语句建立自定义视图。 4、如果需要，可对字段进行处理，比如创建计算字段 5、保存后完成数据模型的建立

模型设置 用户可以对模型的连接方式，应用范围，是否同步复杂报表，缓存有效期进行设置。 对于连接方式为直连的模型，可以在更多内开启缓存有效期的设置。 开启后缓存有效期设置后，可对缓存时间进行设置，时间单位包括时、分、秒。 缓存规则： 首次访问后会对数据进行缓存，缓存有效期内，再次访问会直接访问缓存数据。当到达缓存失效时间点后，缓存被丢弃，之后再访问时，则需重新访问数据库获取数据并重新缓存。 例如，若首次访问时间为“08时36分”，缓存有效期设置为“3小时”，则缓存的数据会在下一个三小时整数倍时间点失效，即“09时00分”失效。若缓存有效期设置为“6小时”，则缓存的数据会在下一个六小时整数倍时间点失效，即“12时00分”失效。若缓存有效期设置为”24小时“，则当日首次访问后缓存的数据，会在当日凌晨00时00分失效，这期间查看报表都会访问缓存数据。

表关联

在数据模型中进行表关联时，支持用户对某个数据连接的多个表进行关联，也支持用户对多个数据连接中的表进行关联。

跨数据连接关联表 产品支持将不同数据连接中的表进行关联，比如一张数据表来自MySQL数据库，一张数据表来自Excel文件，要将两张数据表关联成一张宽表后分析。此时需要将不同数据连接的表抽取至产品提供的MPP数据库中 我们还可以通过“抽取设置”对抽取任务进行更灵活的设置，比如设置抽取方式，建表方式，抽取引擎，执行计划，详见数据连接方式 表关联方式

关联类型	说明	示意图

内关联	使用内关联来合并表时，生成的表将包含与两个表均匹配的值。
左关联	使用左关联来合并表时，生成的表将包含左侧表中的所有值以及右侧表中的对应匹配项；当左侧表中的值在右侧表中没有对应匹配项时，您将在数据网格中看到 null 值。
右关联	使用右关联来合并表时，生成的表将包含右侧表中的所有值以及左侧表中的对应匹配项；当右侧表中的值在左侧表中没有对应匹配项时，您将在数据网格中看到 null 值。
外关联	使用完全外部关联来合并表时，生成的表将包含两个表中的所有值；当任一表中的值在另一个表中没有匹配项时，您将在数据网格中看到 null 值。

表处理

在进行对表的处理前，我们首先需要了解有数BI划分字段的方式。

名词解释 **维度**：观察数据时，使用的粒度

度量：汇总的统计值

聚合方式：汇总的方式，比如求和、求平均、最大值、最小值

怎么理解呢？假设我们有一份明细的订单交易数据，部分数据如下： 将这份数据导入有数后，我们可以用不同的粒度观察数据，有数会自动替我们进行汇总。

比如，观察“各地区的销售额”，“地区”是维度，“销售额”是度量。每个地区都对应成百上千行数据，有数对这些数据进行了求和汇总。如下图所示： 我们也可以观察“各省的销售额”，“省/自治区”是维度，“销售额”是度量。如下图所示： 数据导入有数后，默认会把字符型的字段归类为维度，数值型的字段归类为度量，用户也可以手动更改字段的类型。

字段配置 在数据模型中，用户可以对字段进行可见性操作，重命名，复制字段，转换数据类型，数据格式设置，数据字典设置，创建层级，创建组，度量/维度转换，设置指标，新建计算度量。用户可以批量进行操作。

- 批量编辑字段 点击批量编辑字段，用户可以对字段属性等进行批量操作。
- 批量设置数据格式 点击批量设置数据格式，用户可以对数据格式进行批量操作。
- 设置指标 支持用户可以通过指标系统与自定义两种方式配置指标名称、业务与技术口径。
- 数据字典 数据字典用于修改离散字段成员的名称，只针对维度类型的字段，具体介绍可以查看数据字典。
- 层级创建 用户可以为不同的字段之间创建层级关系，具体可以查看层级。
- 创建组 用户可以为一个维度字段中所有的成员自定义组别的划分，新的划分会作为一个新的维度字段而存在。具体可以查看创建组。

数据连接方式

产品为用户提供了两种数据连接方式，分为为直连和抽取。直连指的是是直接连接用户数据库进行数据的读取，抽取指的是将表数据抽取到内置的MPP数据库中，提升查询效率。 接下来将着重对抽取进行介绍。 **数据抽取** 支持全量抽取、增量抽取两种抽取方式。 1.全量抽取：每次抽取的时候将表数据全部抽取至内置的MPP数据库，提供了全量覆盖收取和全量追加抽取两张抽取方式：

全量覆盖抽取：每次抽取，对数据库的全部数据进行抽取，并覆盖MPP数据库已有的数据，如下图所示： 全量追加抽取：每次抽取，对数据库的全部数据进行抽取，并追加在MPP数据库中，如下图所示： 2.增量抽取：每次抽取只抽取相比于上次更新增加的数据，提供了增量抽取和增量抽取（滚动覆盖）两种抽取方式

增量抽取：每次抽取，根据增量字段判断数据库中的数据是否为新增数据，对数据库的新增数据进行抽取，并追加在MPP数据库中，如下图所示： 增量抽取（滚动覆盖）：每次抽取，根据日期及滚动周期将数据库中的新增以及部分历史数据抽取到MPP数据库中，其中历史数据将会覆盖原数据。 关于增量抽取，需要注意的是： 1、只支持针对日期型（Date）和数值型（Int）的字段作为增量抽取的依据字段。 2、只支持对源表新增的数据做增量，如果源表中对数据有更新或删除操作，增量抽取的时候不会检测到这些变化。 **设置方式** 用户在数据连接、数据准备、数据模型中都可以对数据连接

方式进行设置。

在数据连接中，用户在选择需要设置的数据连接后，选择表信息模块，列表中会显示该连接内所有的数据表，可以对需要抽取的表进行抽取设置。 在数据准备中，用户可以在输入节点切换数据连接的方式。 在数据模型中，用户可以在模型信息中切换数据连接的方式。 用户选择抽取，进入抽取设置界面后，可以设置抽取方式和执行计划。 另外，在项目中心，可以对所有的抽取任务进行统筹管理。详见数据任务管理