萃蓝数据管道

用户手册 v1

目录

1 一般性约定

此文档用于向您介绍如何部署并使用萃蓝数据管道来构建您的数据流。

随着产品不断迭代升级，此文档可能会发生变化，您理解并认可我们难以保证及时通知到您相关变化。您可以随时访问我们的官方网站获取最新版本的在线文档。

2 部署

2.1 使用云版本

云版本预计将于 2024 年 Q1 开放服务，我们诚挚邀请您届时通过官网获取并体验。

云版本将拥有不亚于私有化版本的安全体系，同时在资源利用率、弹性调度等方面有着非常出色的优势，能进一步提升您的数据流转效率。

2.2 Docker 部署

用于在单机上快速部署一套服务并进行功能验证。无高可用能力，不建议用于正式生产环境。

2.2.1 系统要求

- 64 位内核，Linux 操作系统
- 不小于 8*CPU、16GB 可用内存和 200GB 可用磁盘空间
- 建议最小 1000Mbps 带宽以太网卡
- Docker Engine >= 24.0.6

2.2.2 启动服务

解压您获取到的发型包，通过下列命令启动服务：

```
1 $ docker-compose up -d
```

命令行响应如图 1所示。您可进一步通过浏览器访问 https://localhost 即可验证萃蓝数据管道是否部署成功。

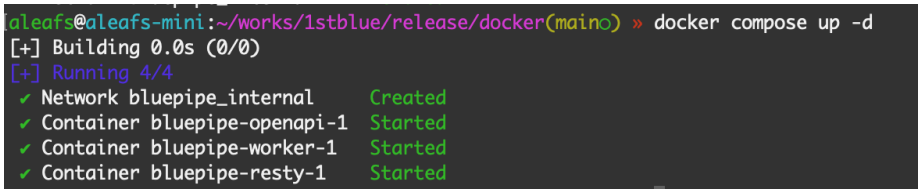

图 1: 通过 Docker 启动服务

2.2.3 停止服务

需要停止服务时，建议您通过下列命令进行：

```
$ docker compose stop
```

请注意，我们不建议使用 docker compose down 命令来停止服务，它将会导致您之前使用过程中产生的数据全部丢失。

2.3 分布式部署

当您需要在正式生产环境使用萃蓝数据管道时，我们建议您采用分布式方式进行高可用配置。

由于此项工作涉及对您业务的容量和 SLA（Service Level Aggrement）评估，建议您联系我们的实施工程师评估所需要的资源已经系统配置方案，以最大程度在高可用和资源利用率之间达到平衡。

3 快速开始

无论您采用哪种方式部署萃蓝数据管道，您都会通过一个既定的网址来访问我们的产品。我们以 Docker 部署模式为例，此时您应该能够在浏览器中通过 https://localhost 访问到产品首页。

注意：如果您的浏览器在地址栏或者其他任何地方提示”此网站不安全”（如图 2），忽略并继续访问。

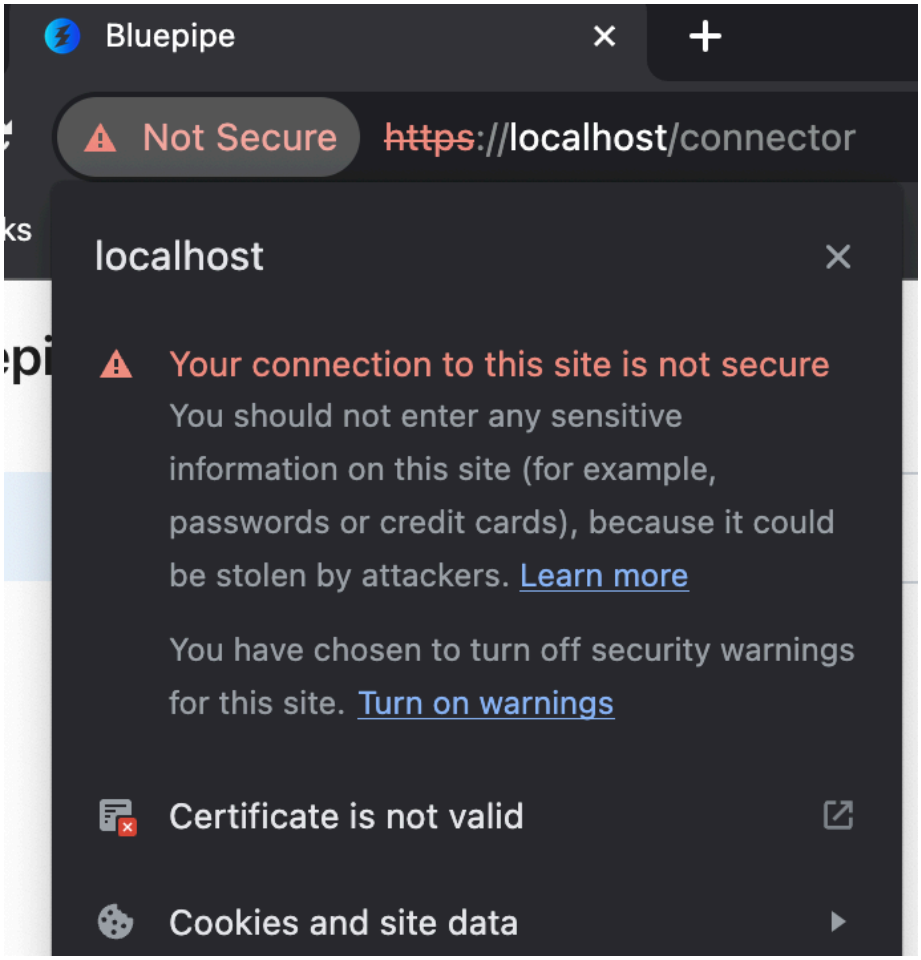

图 2: 浏览器证书警告

出现此提示是因为我们默认使用了一个自签名的数字证书来提供 HTTPS 服务，此行为对您的业务安全没有任何影响。您可通过购买 bing 部署权威 CA 签发的数字证书来消除此警告。

3.1 用户登录

首次使用萃蓝数据管道时您需要登录以验证您的合法身份。

3.1.1 使用账号密码登录

我们没有设计用户注册的流程。系统每次部署时，系统会自动生成一个管理员账号供您使用，您可以联系我们的实施工程师请教如何查看初始的用户名和密码，从而登入系统（图 3）。

图 3: 使用账号密码

3.1.2 SSO 与 LDAP

当您计划将萃蓝数据管道用于正式生产环境时，我们建议您接入您企业的单点登录系统（Single Sign On）以确保企业信息安全。

我们遵守标准 OAuth 2.0 协议来实现 SSO。因此，如果您企业的 SSO Server 也是标准 OAuth 2.0 实现，那么此项工作仅需简单配置即可；否则可能涉及额外的定制开发工作。

关于如何配置标准 OAuth 2.0 协议的 SSO 系统，请咨询我们的实施工程师以获取支持。

3.2 连接管理

您需要首先创建“连接”才能继续创建“作业”。

在萃蓝数据管道中，“连接”通常代表一个数据库实例，同一个数据库实例只能被一个“连接”使用。此约束是为了尽可能避免出现环状的数据流。

3.2.1 创建连接

依次点击左侧导航“连接”、右上角“新建连接”按钮，进入连接器选择页面（图 4）。

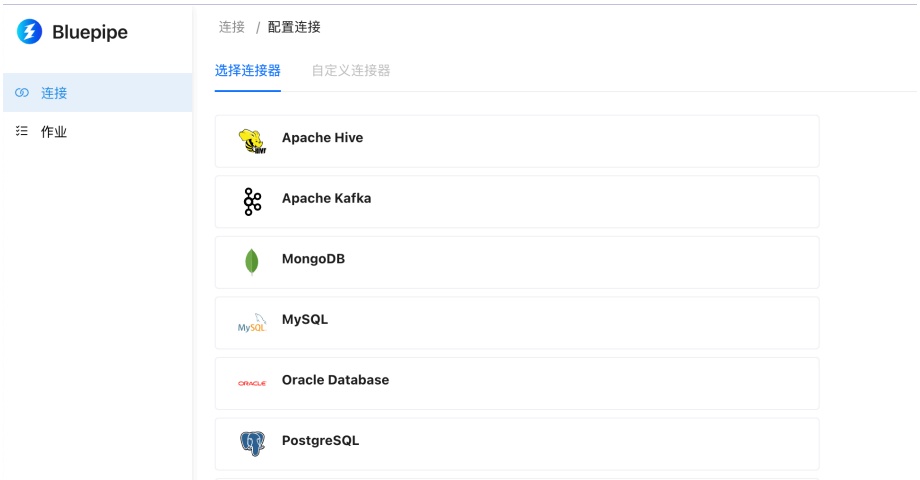

图 4: 选择连接器

点击一个连接器，进入连接配置页面（图 5）。

我们以 MySQL 为例，页面右侧文档栏详细描述了左侧表单各个字段的含义和用途，请根据您的实际情况填写。

填写完成后，点击页面下方“测试连接”按钮，右侧文档栏会切换成连接状态页。此时系统会在您的局域网内根据您输入的连接串和账号密码登信息寻找并尝试连接 MySQL，成功后需要您“确认”来完成连接的创建 (图 6)。

3.2.2 删除连接

点击左侧导航栏“连接”，此页面展现所有被成功创建的“连接”。尚未被任何“作业”关联到的“连接”可以被删除。点击相应卡片右上角删除按钮，二次确认后，此链接即被删除（图

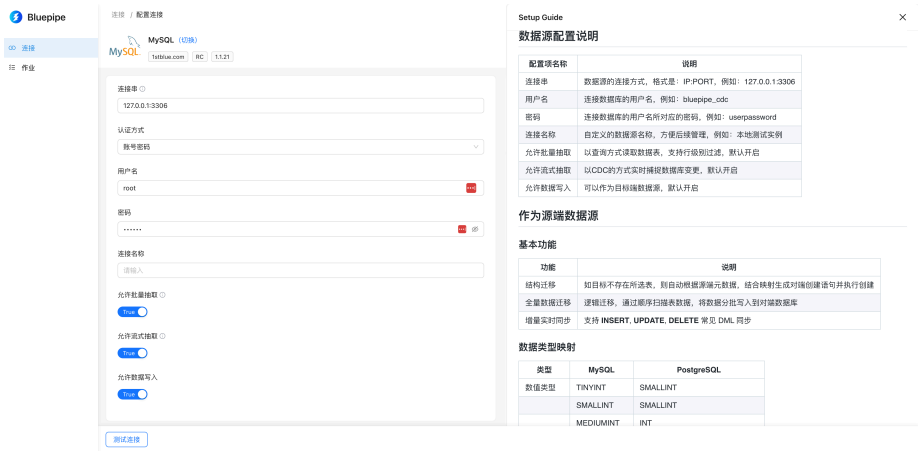

图 5: 新建连接

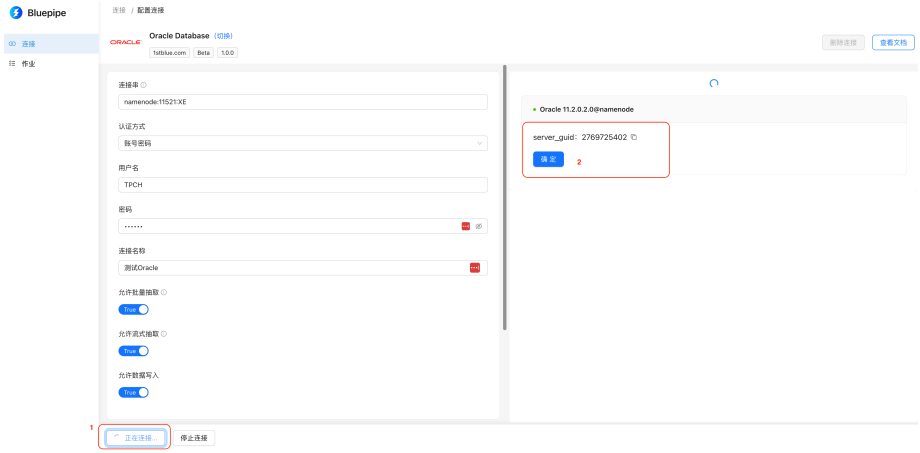

图 6: 测试并确认连接

7)。

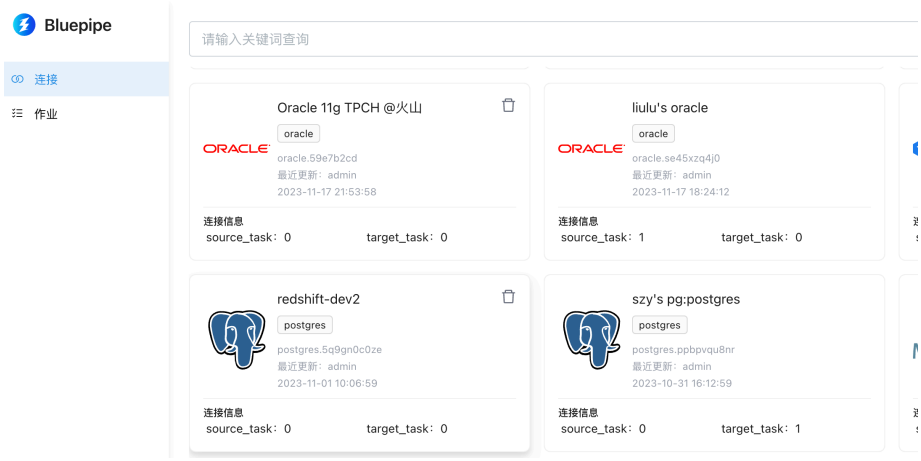

图 7: 删除连接

3.3 作业管理

3.3.1 新建作业

依次点击左侧导航栏“作业”-> 右上角“新建作业”，按照向导完成作业创建（图 8)。

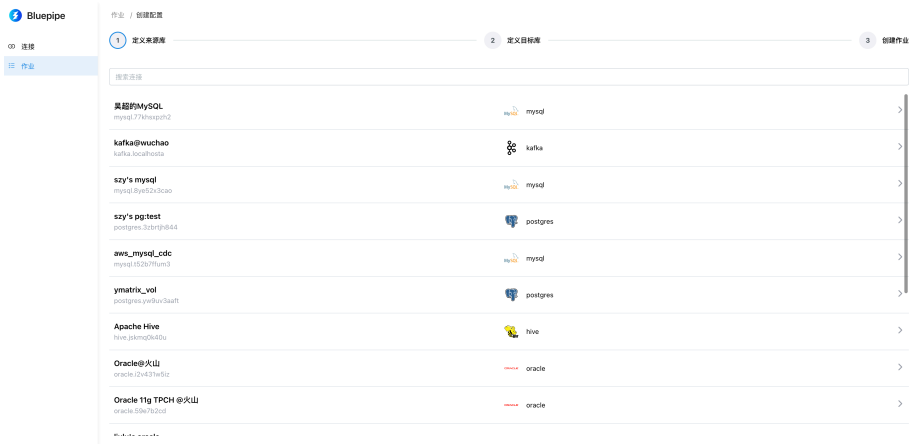

图 8: 新建作业-Step 1

选择“来源库”和“目标库”之后，系统会列出“来源库”中的所有表，默认全部勾选，意

味着对当前所有表进行同步。按需进行配置，点击右下角“确认”按钮完成作业创建（图 9）。

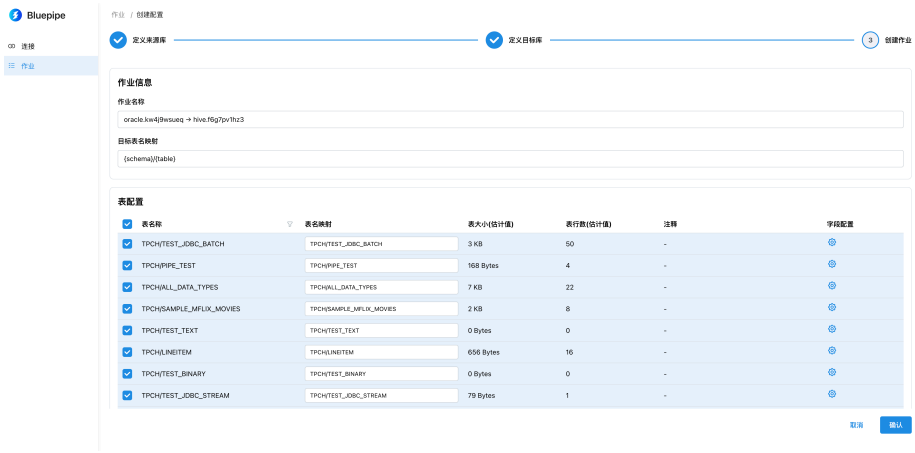

图 9: 新建作业-Step 2

3.3.2 运行作业

点击左侧导航栏“作业”，选择某个作业（占据一行）点击作业名称，进入任务状态页面（图 10）。此页面每行代表来源库中的一张表，点击右侧“操作”栏绿色三角图标“运行”此任务。

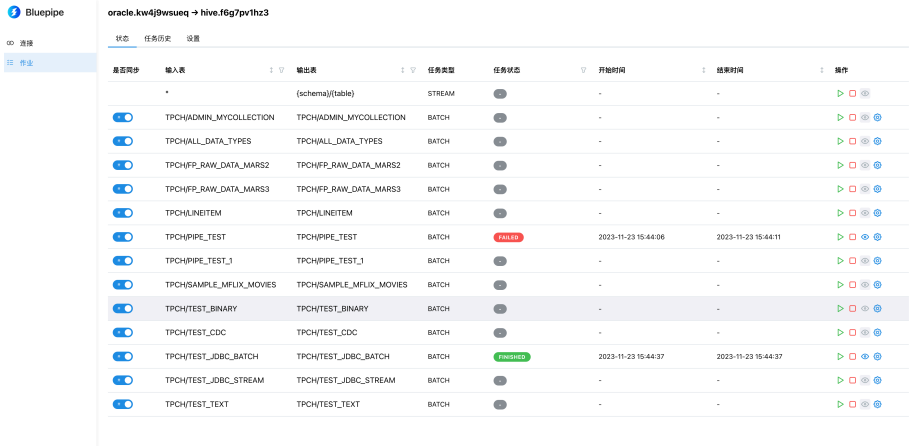

图 10: 任务状态

点击右侧“操作”栏蓝色眼睛图标可查看此行任务最后一次运行的执行计划与状态（图

11）。此执行计划比较详细地展现了任务运行过程中的关键信息，比如自动建表/Alter Table、数据切分以及运行过程中的统计信息等。点击某个 Stage，右侧抽屉展现其更详细的信息。

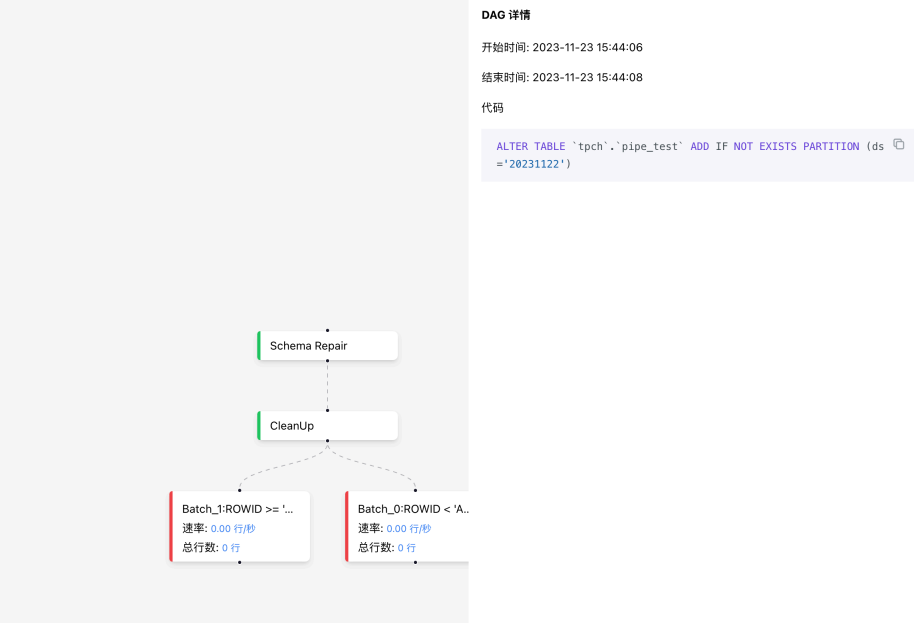

图 11: 执行计划

3.3.3 个性化配置

在任务列表页面，点击右侧“操作”栏齿轮图标，可对单张表进行个性化配置。如图 12所示，当前允许对某些列不选择从而不进行同步，也支持为写入目标库的列名重新命名。完成后“确认”，配置会在下一次运行时生效。

3.3.4 删除作业

点击左侧导航“作业”，此页面展现作业列表。点击右侧“删除”按钮，二次确认后将会删除此作业配置。请注意，此操作不可恢复，务必确认对您的业务没有影响时才可以进行。

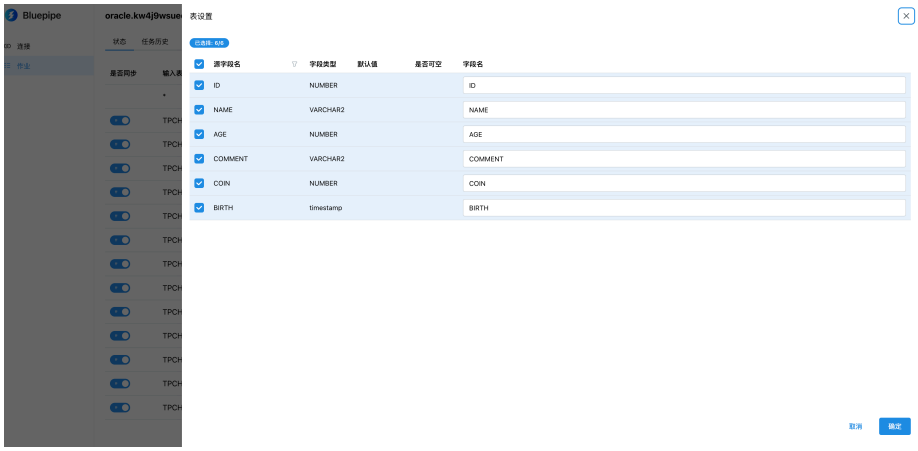

图 12: 任务配置

4 与其他系统集成

4.1 与 ETL 调度工具集成

我们支持与 Airflow、Azkaban、Dophescheduler 等开源 ETL 调度工具集成，以实现自动化的数据流编排。

如果您有需要这方面的支持，请与我们的实施工程师联系寻求帮助。