

Airbyte 部署及使用

AUTHOR: 彭玲 TIME: 2022/11/10

Airbyte 部署及使用

Airbyte 本地部署

Airbyte 登录

通过 Airbyte 同步 PG 至 CH

配置 Source

配置 Destination

配置 Connection

同步数据

Airbyte 同步方式

Airbyte 本地部署

Airbyte 本地部署较为简单，提前准备好 Docker 和 Docker Compose 环境即可开始：

```
1 ! [airbyte-connection-source] (assets/airbyte-connection-source.jpg)$ git clone https://github.com/airbytehq/airbyte.git
2 $ cd airbyte
3 $ docker-compose up
```

上述命令执行成功后，打开浏览器，输入：<http://localhost:8000> 即可访问 Airbyte 服务。

Airbyte 登录

访问 <http://localhost:8000> 时，界面提示输入用户名和密码，初始 用户名/密码 为 `airbyte/password`。

通过 Airbyte 同步 PG 至 CH

这里介绍下如何通过 Airbyte 界面将 PostgreSQL 数据库（Source）同步至 ClickHouse 数据库（Destination）。3 个步骤：

1. 创建 源数据库（Source）；
2. 创建 目标数据库（Destination）；
3. 建立 源数据库至目标数据库之间的连接（Connection）。

说明：以下步骤可以通过 `onboarding` 菜单逐步进行，也可以通过 `Sources`、`Destinations` 和 `Connections` 分步进行。

配置 Source

进入 `Sources` 菜单页，点击 `+ New source` 按钮，选择 `Source type` 为 `Postgres`，然后依次填写下面的信息：

The screenshot shows the 'Source Settings' configuration page for a PostgreSQL source. The left sidebar has buttons for Connections, Sources (which is selected), Destinations, Update, Resources, and Settings. The main area shows the following fields:

- Source type:** Postgres (selected)
- Source name:** PG_166_pepca
- Host:** 10.8.30.166
- Port:** 5432
- Database Name:** pepca
- Schemas (Optional):** public
- Username:** FashionAdmin
- Password (Optional):** (redacted)
- JDBC URL Parameters (Advanced) (Optional):** (empty)
- Connect using SSL:** (checkbox checked)
- SSL Modes:** disable
- Replication Method:** Standard

填写完成后，点击 `Set up source` 完成源数据库的配置。

配置 Destination

进入 `Destinations` 菜单页，点击 `+ New destination` 按钮，选择 `Destination type` 为 `clickhouse`，然后依次填写下面的信息：

The screenshot shows the 'Destination Settings' page for a Clickhouse ALPHA destination. The left sidebar has a dark theme with icons for Home, Connections, Sources, Destinations (selected), Update, Resources, and Settings. The main area shows the following configuration:

- Destination type:** Clickhouse ALPHA (highlighted with a red box)
- Alpha connectors** are in development and support is not provided. See our [documentation](#) for more details.
- Destination name:** CH_71_airbyte_pepca
- Host:** 10.8.30.71
- Port:** 30123
- DB Name:** airbyte_pepca
- User:** default
- Password:** (Optional, empty field)
- JDBC URL Params:** (Optional, empty field)
- SSL Connection:** (Toggle switch off)
- SSH Tunnel Method:** No Tunnel

At the bottom are buttons for **Save changes and test**, **Cancel**, and **Retest destination**.

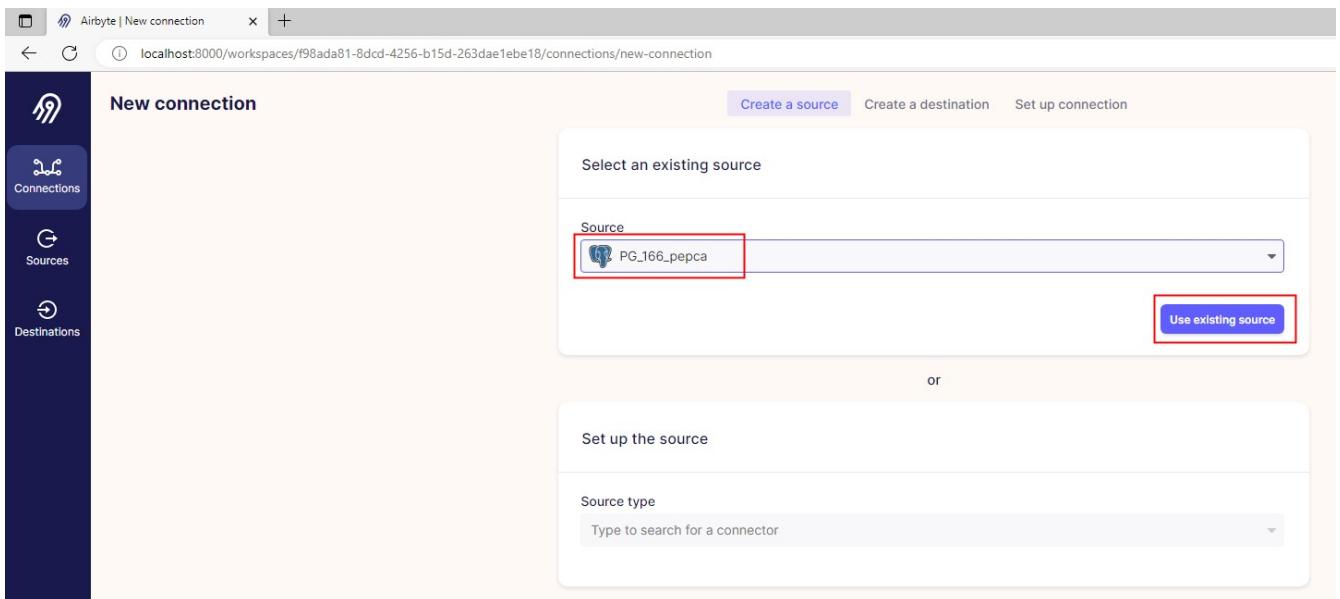
填写完成后，点击 **Set up destination** 完成目标数据库的配置。

注意：需要提前准备好 CH 数据库（如，提前创建好这里的 airbyte_pepca 数据库），否则，连接测试会失败。

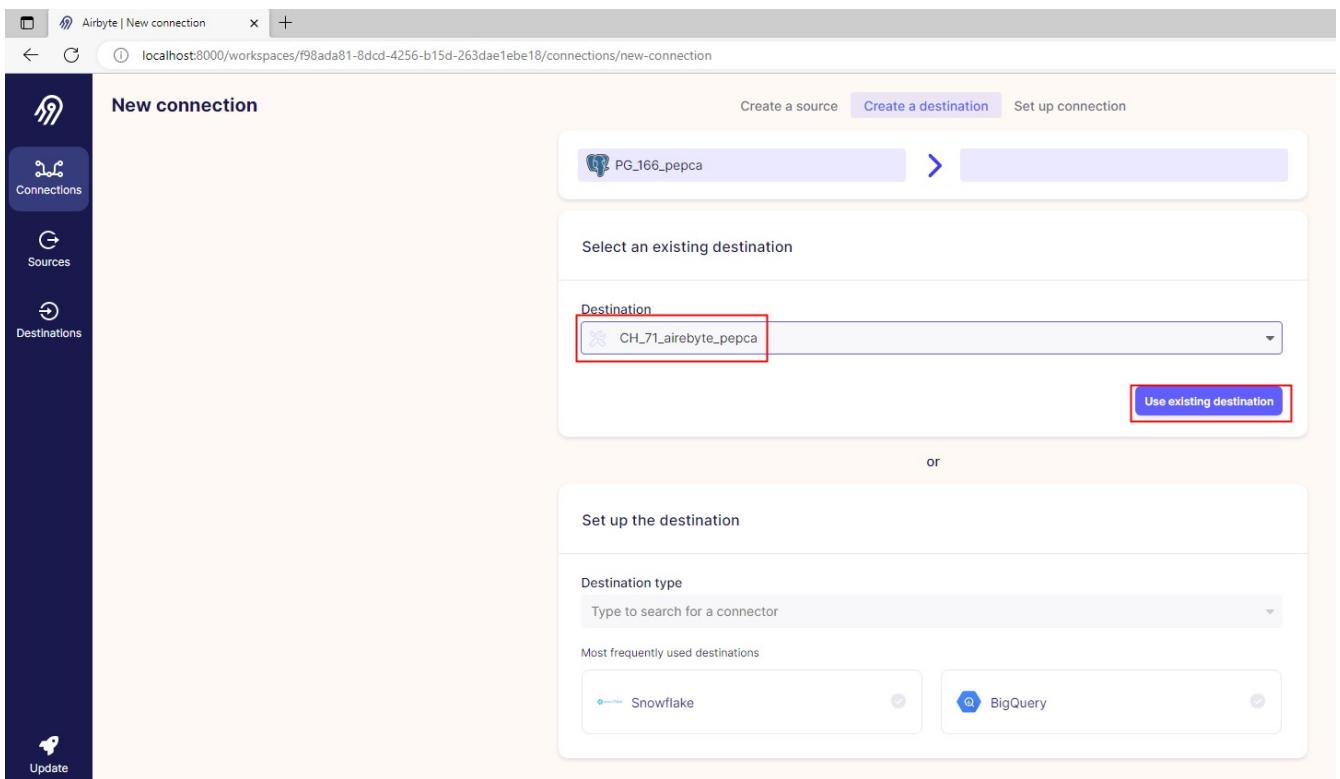
配置 Connection

配置好 Source 和 Destination 后，接下来为两者建立起连接即可。

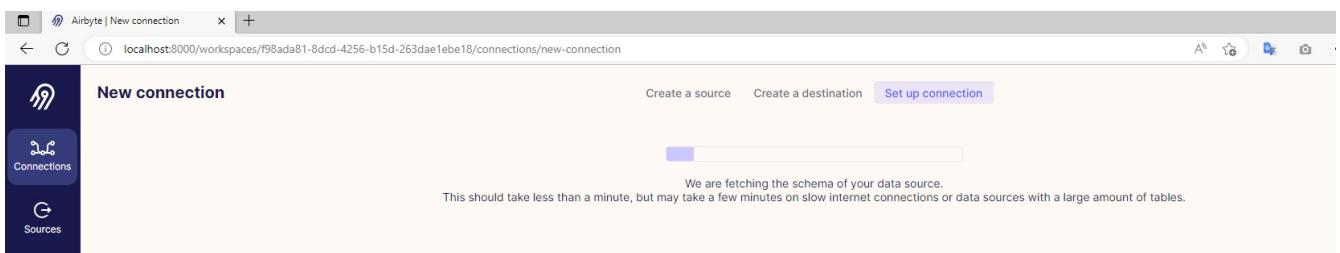
进入 **Connections** 菜单页，点击 **+ New connection** 按钮，选择 **Select an existing source**，从 **source** 下拉列表中选择前面已经配置好的源数据库：



接下来选择 `Select an existing destination`, 从 `Destination` 下拉列表中选择已经配置好的目标数据库:



然后, 系统开始建立 源数据库 和 目标数据库 之间的连接:



接下来配置 连接名称、 复制频率、 同步模式 等信息:

New connection

Transfer
Replication frequency*
Set how often data should sync to the destination
Manual

Streams
Destination Namespace*
Define the location where the data will be stored in the destination
Destination default
Destination Stream Prefix (Optional)
Add a prefix to stream names (ex. "airbyte_" causes "projects" => "airbyte_projects")
prefix

Activate the streams you want to sync

Sync	Source ⓘ	Sync mode ⓘ	Cursor field ⓘ	Primary key ⓘ	Destination ⓘ
	Namespace Stream name	Source Destination			Namespace Stream name
<input type="checkbox"/>	> public api_log	Full refresh Overwrite			<destination schema> api_log
<input type="checkbox"/>	> public application	Full refresh Overwrite			<destination schema> application
<input type="checkbox"/>	> public basicdata_category	Full refresh Overwrite			<destination schema> basicdata_category
<input type="checkbox"/>	> public basicdata_gift	Full refresh Overwrite			<destination schema> basicdata_gift
<input type="checkbox"/>	> public basicdata_goods	Full refresh Overwrite			<destination schema> basicdata_goods
<input type="checkbox"/>	> public basicdata_post	Full refresh Overwrite			<destination schema> basicdata_post

Refresh source schema

选择标准化方式 Normalized tabular data :

Normalization

Raw data (JSON)
 Normalized tabular data Map the JSON object to the types and format native to the destination. [Learn more](#)

Set up connection

同步数据

完成 Connection 配置后，即可同步数据：

The screenshot shows the Airbyte web interface for managing database connections. On the left, there's a sidebar with icons for Connections, Sources, and Destinations. The main area is titled 'Connection' and shows a connection named 'PG_166_pepca >> CH_71_airebyte_pepca'. It displays a flow from 'Postgres PG_166_pepca' to 'Clickhouse ALPHA CH_71_airebyte_pepca'. The connection status is 'ENABLED'. Below this, there are tabs for 'Status', 'Replication', 'Transformation', and 'Settings'. Under the 'Status' tab, there's a section for 'Sync History' which lists four sync events:

Event	Details	Timestamp
Sync Succeeded	484.7 MB 563,779 emitted records 5m 37s	2:26PM 11/09
Reset Succeeded (68 streams)	api_log, application, basicdata_category, basicdata_gift, basicdata_goods, basicdata_post, basicdata_task_type, ...	2:25PM 11/09
Sync Succeeded	483.95 MB 563,767 emitted records 5m 21s	10:15AM 11/09
Sync Succeeded	483.5 MB 563,767 emitted records 5m 4s	10:05AM 11/09

A 'Sync now' button is located at the bottom right of the sync history section.

数据同步完成后，可以查询 ClickHouse 数据库，验证同步情况。

Airbyte 同步方式

Airbyte 平台中给出了 4 中数据同步方式：

- Source: Full refresh | Dest: Overwrite
- Source: Full refresh | Dest: Append
- Source: Incremental | Dest: Deduped + history
- Source: Incremental | Dest: Append

经过测试，4 种方式的应用效果如下：

- `Source: Full refresh | Dest: Overwrite` 同步方式：Airbyte 从源数据库表中获取所有数据，并以覆盖的方式将数据同步至目标数据库表中。
- `Source: Full refresh | Dest: Append` 同步方式：Airbyte 从源数据库表中获取所有数据，并以追加的方式将数据同步至目标数据库表中。
- `Source: Incremental | Dest: Append` 同步方式：Airbyte 从源数据库表中获取增量数据，并以追加的方式将数据同步至目标数据库表中。
- `Source: Incremental | Dest: Deduped + history` 同步方式：这种方式与 `source: Incremental | Dest: Append` 的差别在于，针对有数据修改的记录，Airbyte 以更新而非追加的方式同步至目标数据库表中。

`Incremental` 同步方式，需要设置游标（`cursor`），Airbyte 通过游标去判断有无增量数据（新增的数据、修改的数据）。通常，会使用 `updated_at` 这样的时间字段作为游标。

另外，`Incremental` 同步方式下，当源数据表有数据删除时，不影响目标数据库表中的数据，即不会同步删除目标数据库表数据。