tinybird

S3 Connector

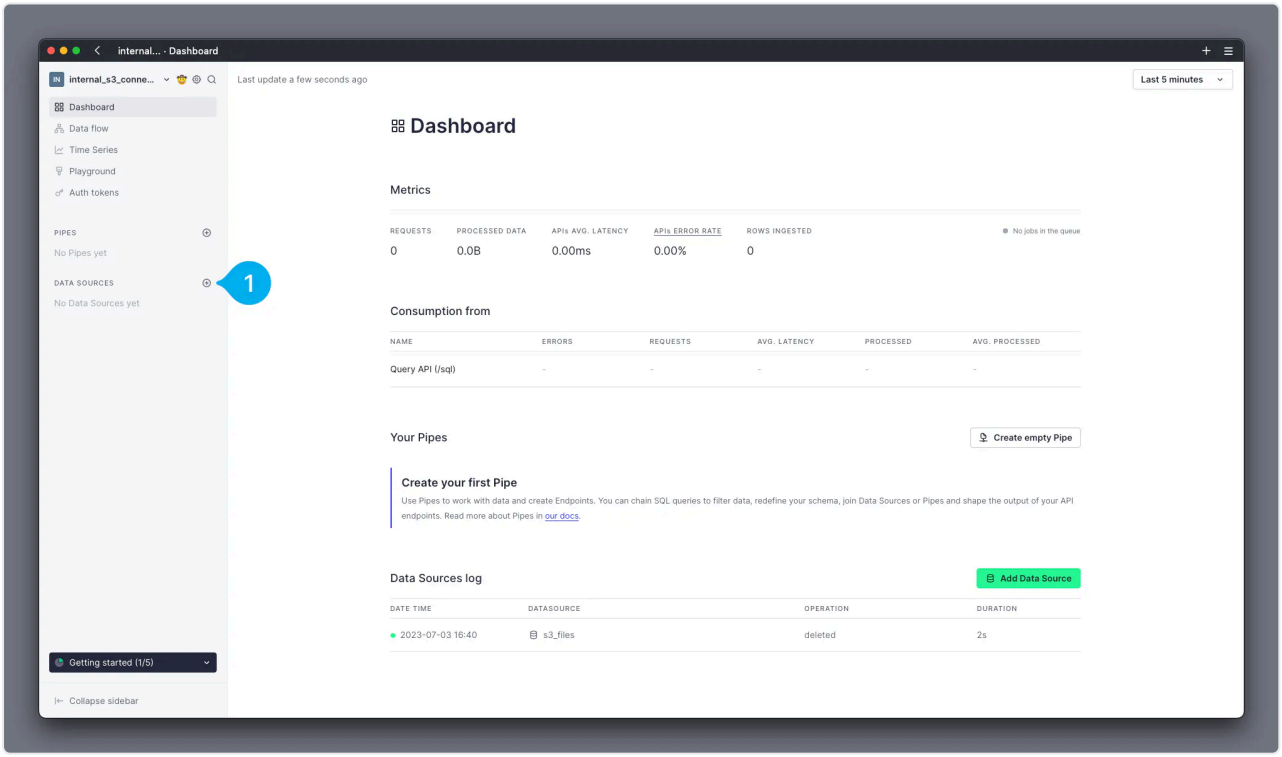
The S3 Connector allows you to ingest files from your Amazon S3 bucket into Tinybird. You can choose to load a full bucket, or to load files that match a pattern.

The S3 Connector is fully managed and requires no additional tooling. You can choose to execute the S3 Connector manually or automatically, and all scheduling is handled by Tinybird.

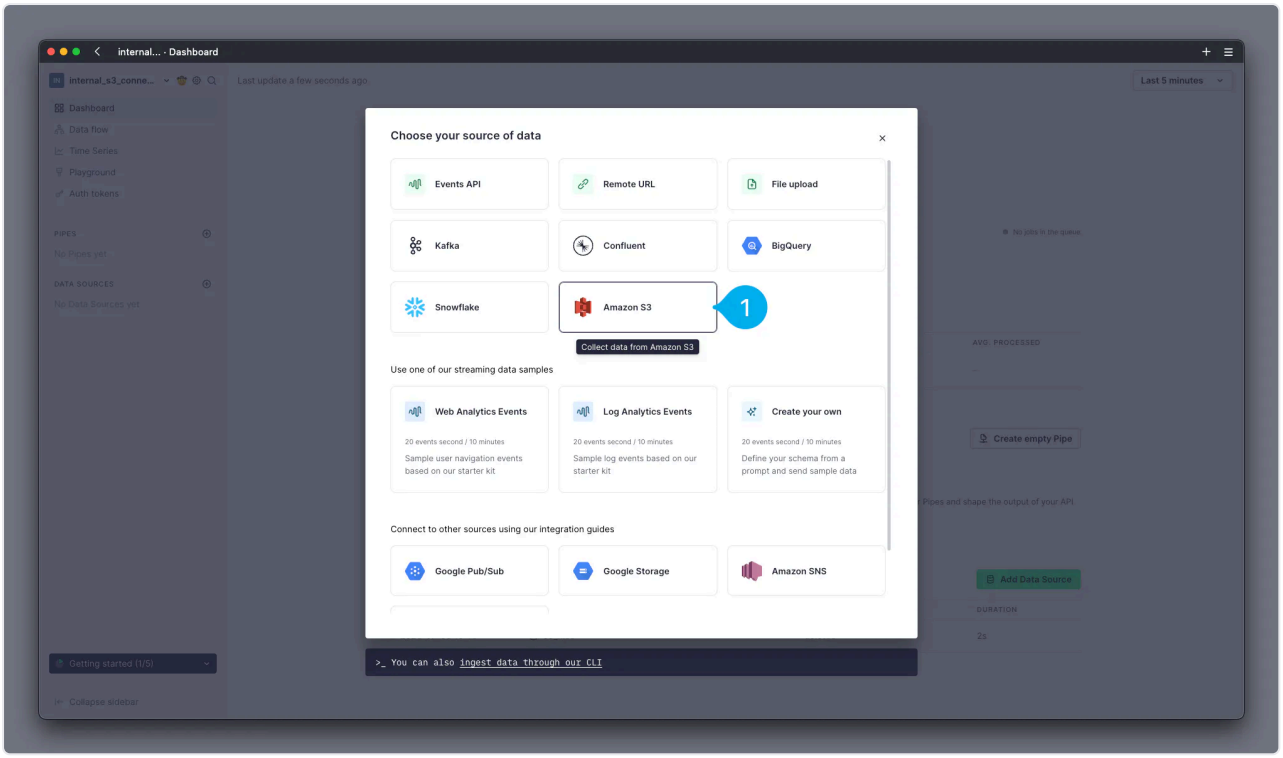
Load files from an S3 bucket

Load files from an S3 bucket in the UI

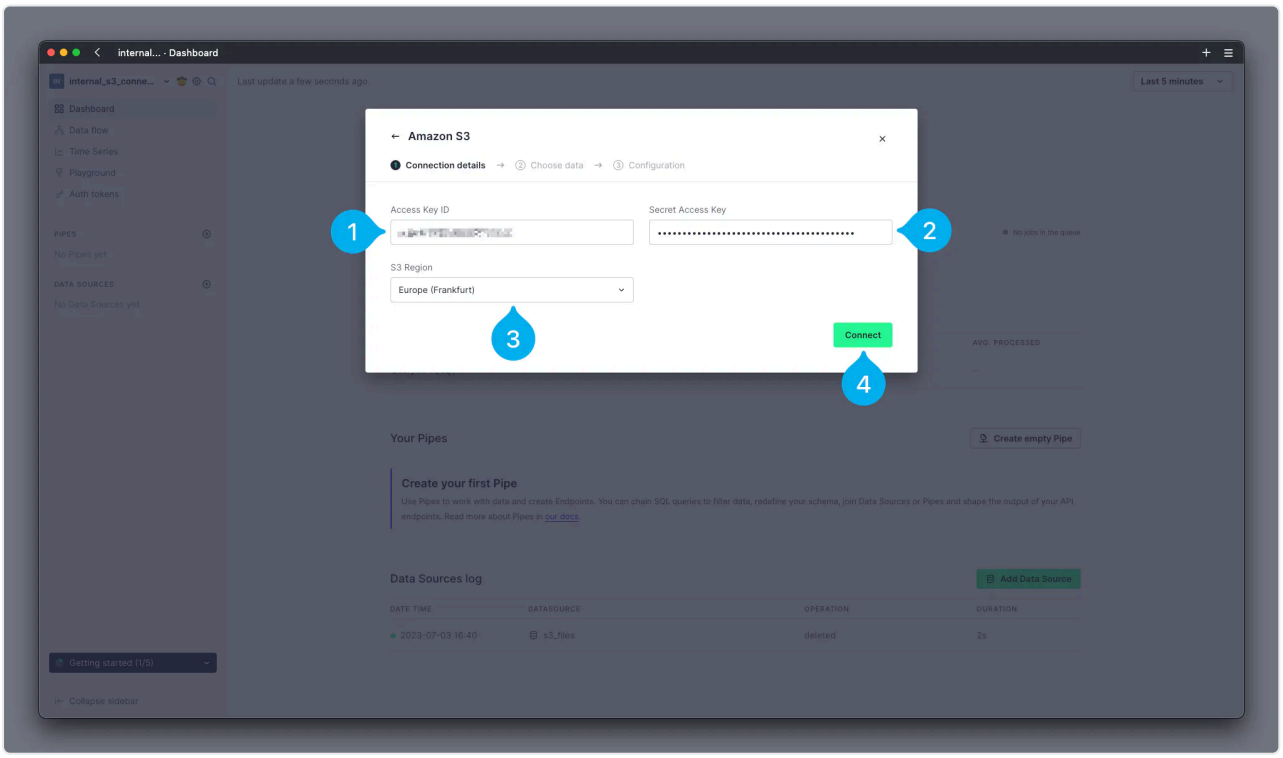
Open the Tinybird UI and add a new Data Source by selecting the + icon next to the Data Sources section on the left hand side navigation bar (see Mark 1 below).



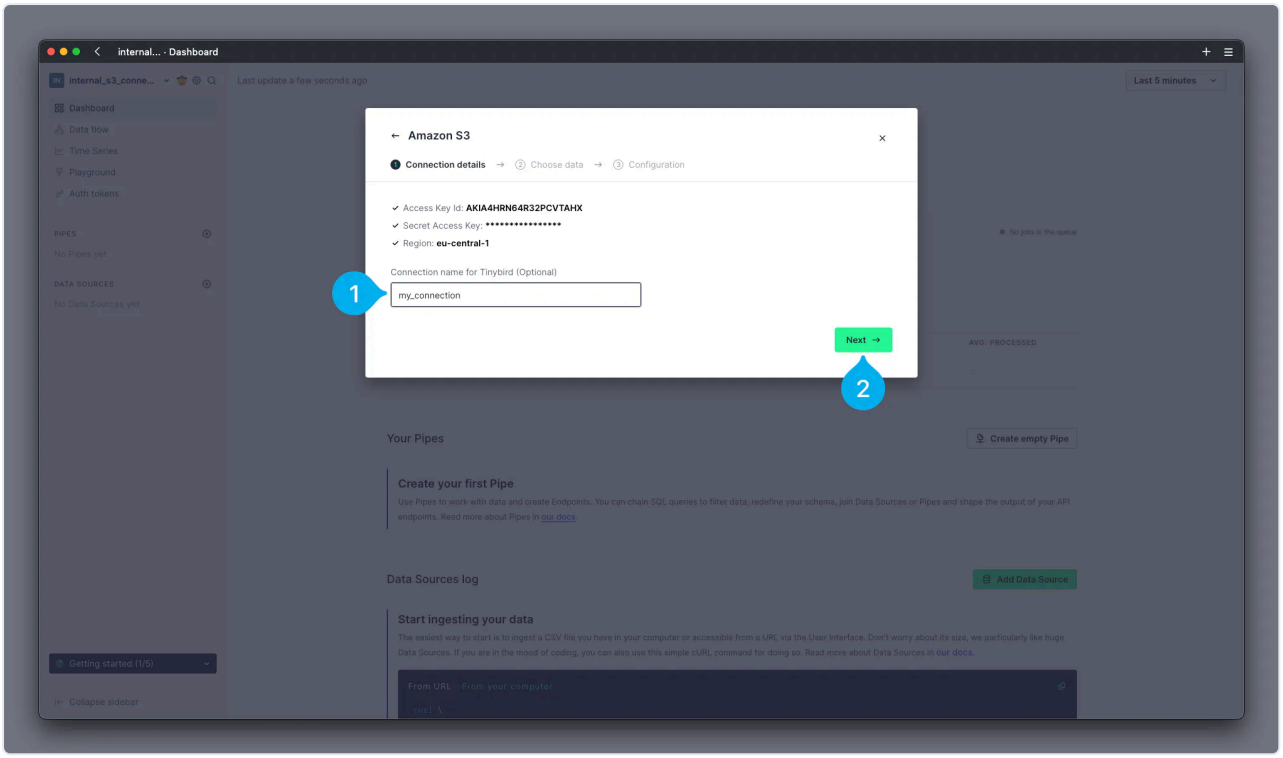
In the Data Sources modal, select the Amazon S3 box (see Mark 1 below).



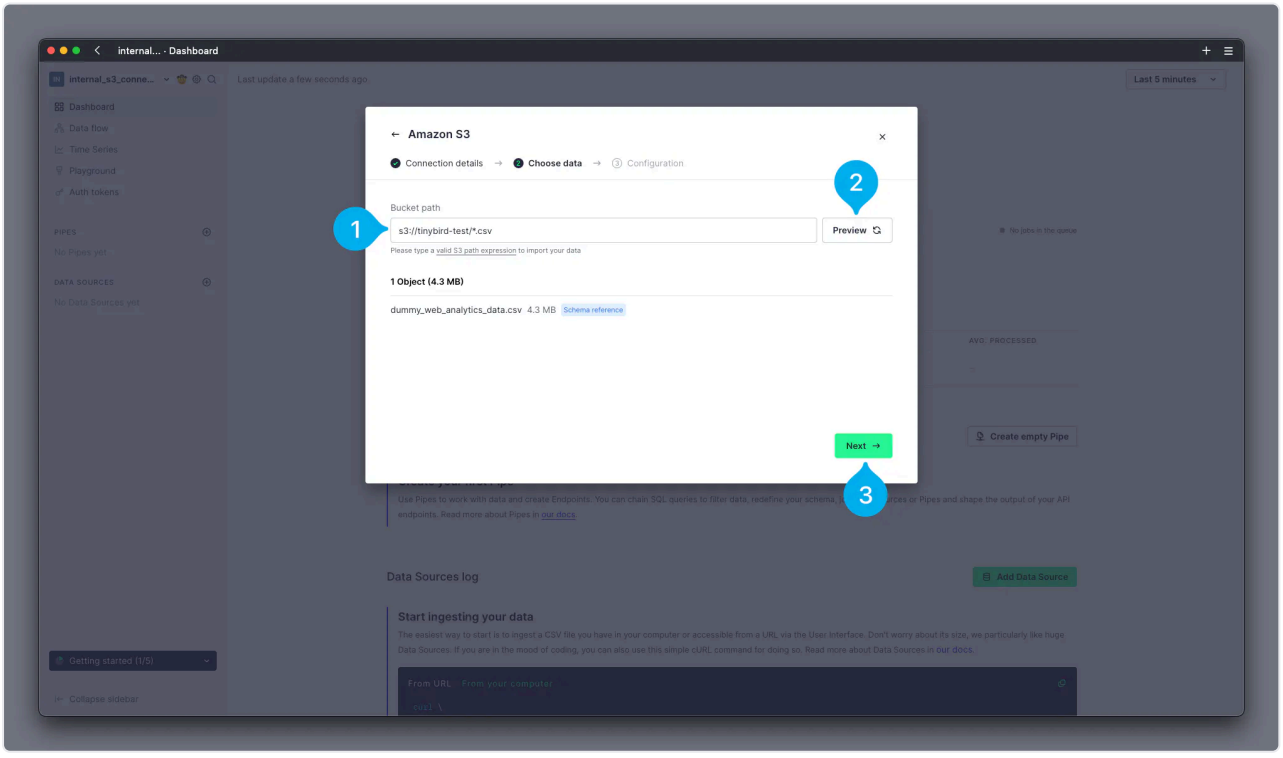
On the next screen, enter your AWS details. You will need to generate Access Keys for an AWS IAM user. Each Access Key has an Access Key ID and a Secret Access Key. Paste your Access Key ID into the Access Key ID box (see Mark 1 below). Paste your Secret Access Key into the Secret Access Key box (see Mark 2 below). Lastly, select the AWS region in which your S3 bucket is located (see Mark 3 below), and select Connect (see Mark 4 below).



The next screen shows a summary of your connection. Here, give your connection a memorable name to identify it in the Tinybird UI (see Mark 1 below). After entering a name, select Next (see Mark 2 below).



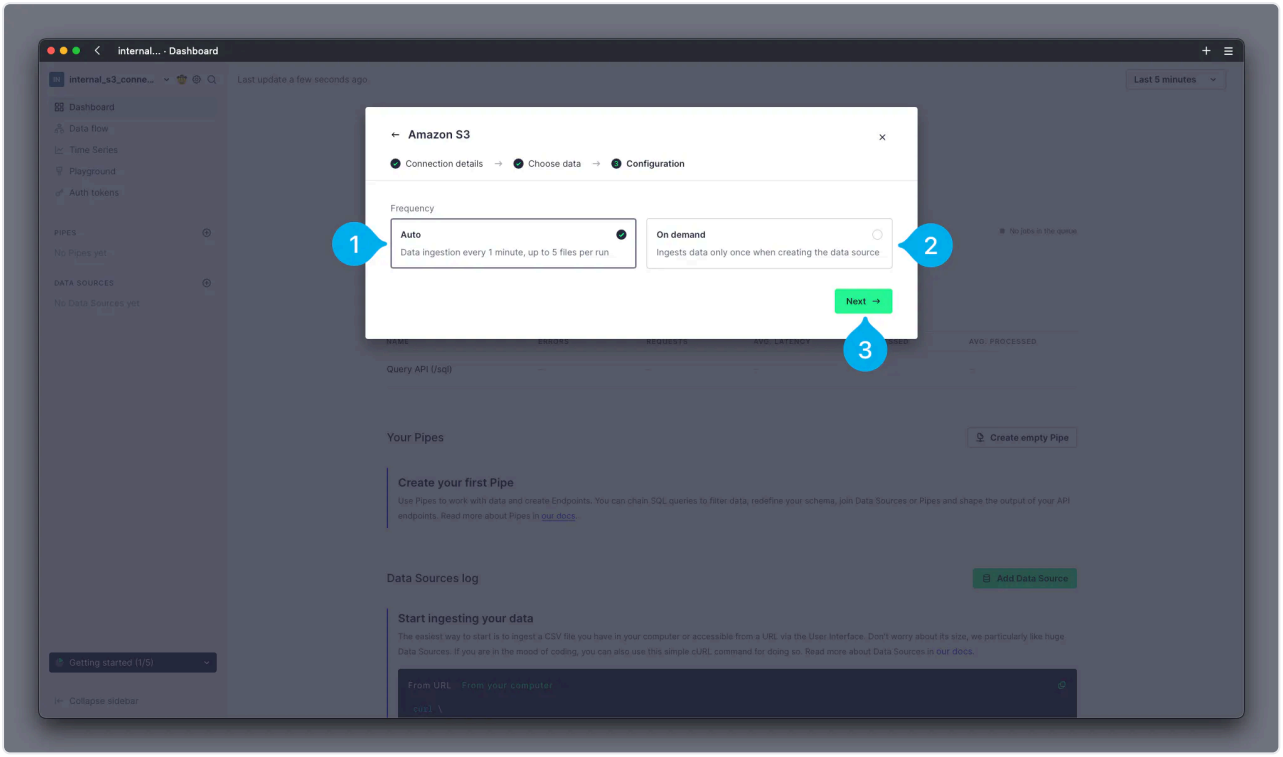
Configure the bucket path used to discover files in S3. In the Bucket path text box (see Mark 1 below) enter a full bucket path, including the `s3://` protocol, bucket name, object path and an optional pattern to match against object keys. For example, `s3://my-bucket/my-path` would discover all files in the bucket `my-bucket` under the prefix `/my-path`. You can use patterns in the path to filter objects, for example, ending the path with `*.csv` will match all objects that end with the `.csv` suffix. Click the Preview button (see Mark 2 below) to run a test discovery and review the list of files that are returned. When you're done, select Next (see Mark 3 below).



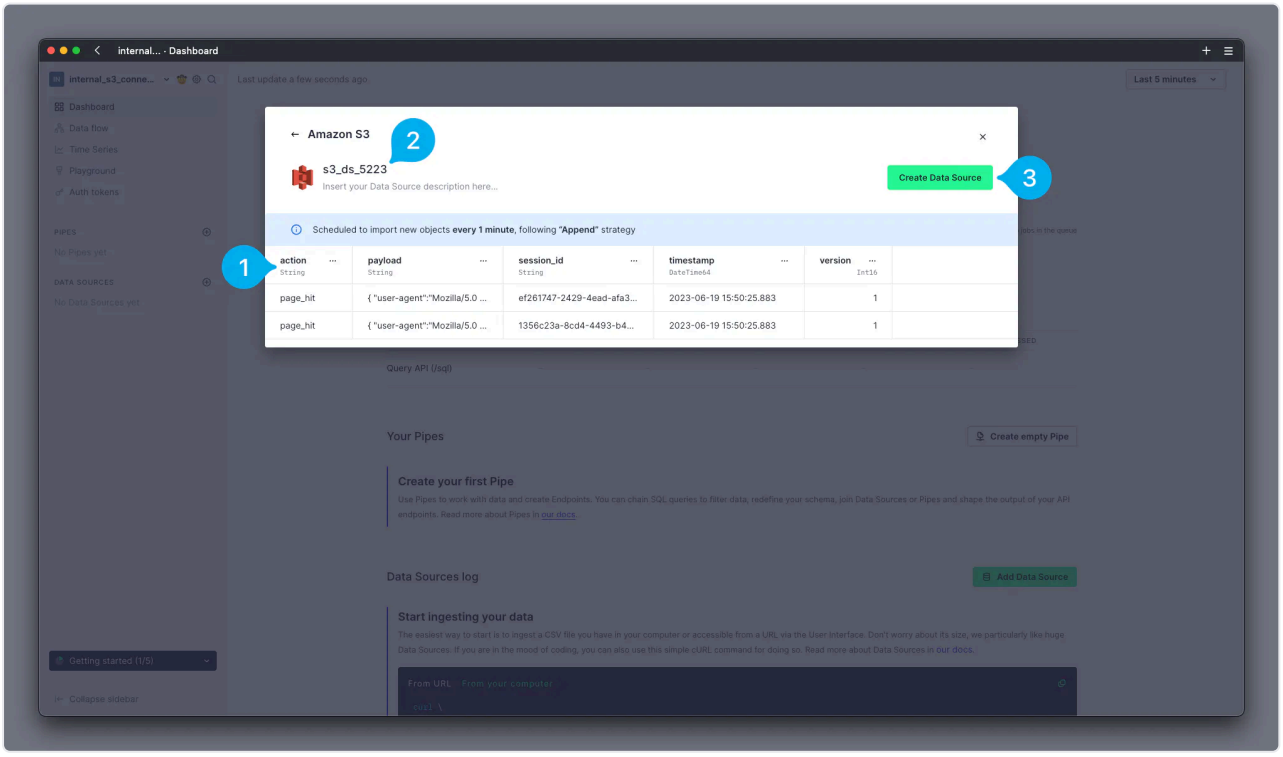
On the next screen, select the frequency to scan for new files. Select Auto (see Mark 1 below) to have Tinybird scan for new files once every minute automatically. Select On demand (see Mark 2

below) to only scan for files when manually executed. Click Next (see Mark 3 below).

> Note: When new files are discovered, data from new files is appended to any previous data in the Data Source. Replacing data is not supported.

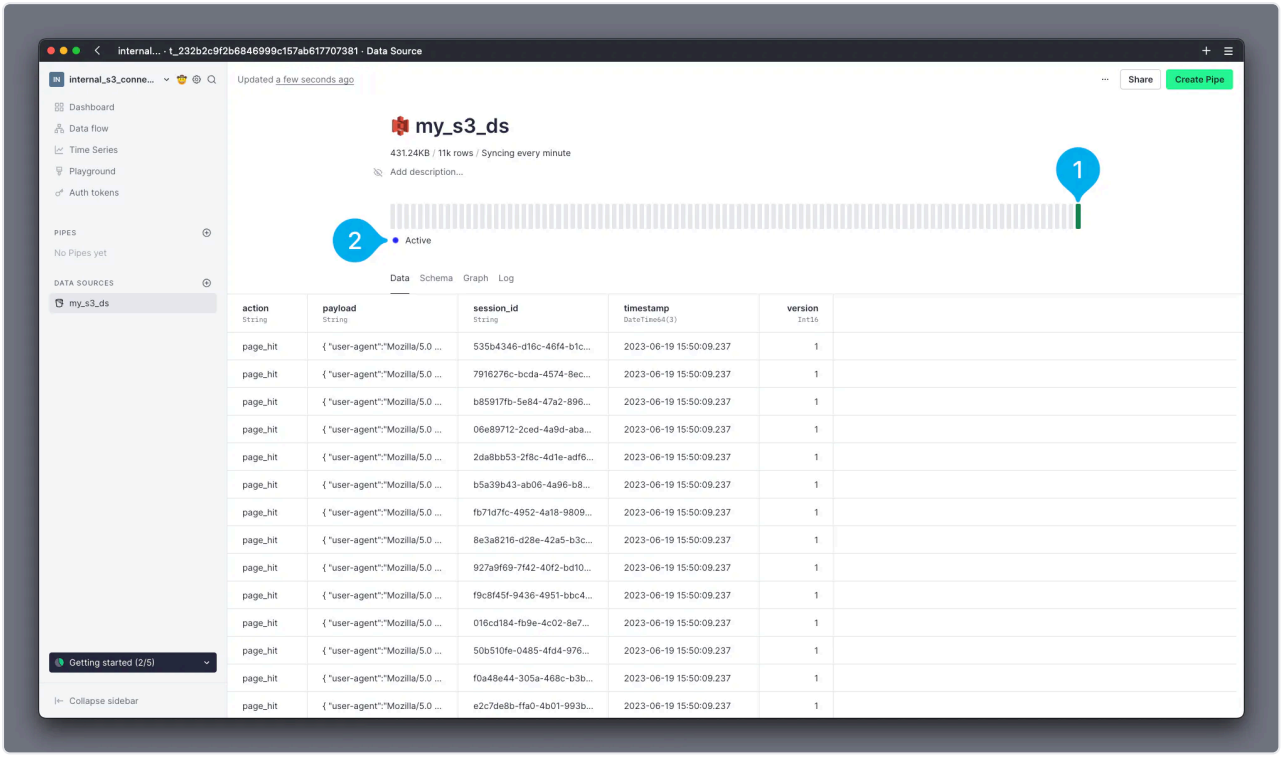


The next screen shows a preview of the incoming data. Here, you can review & modify any of the incoming columns, adjusting their names, changing their types or deleting them entirely (see Mark 1 below). You can also configure the name of the Data Source (see Mark 2 below). When done, click Create Data Source (see Mark 3 below).



You're done! On the Data Source details page, you can see the sync history in the tracker chart (see Mark 1 below) and the current status of the connection (see Mark 2 below).

> Note: For `IMPORT_STRATEGY` only `APPEND` is supported today.



Load files from an S3 bucket in the CLI

You need to create a connection before you can load files from s3 into Tinybird using the CLI. Creating a connection grants your Tinybird Workspace the appropriate permissions to view files in s3.

Authenticate your CLI and switch to the desired Workspace. Then run:

```
tb connection create s3
```

After running this command, enter your S3 details. You will need to generate Access Keys for an AWS IAM user. Each Access Key has an Access Key ID and a Secret Access Key. Paste your Access Key ID when prompted for the Key. Paste your Secret Access Key when prompted for the Secret. When prompted for the S3 region, enter the region identifier string, e.g. eu-west-3. Lastly, enter a memorable name for the connection.

A new `s3.connection` file is created in your project files.

> Note: At the moment, the `.connection` file is not used and cannot be pushed to Tinybird. It is safe to delete this file. A future release will allow you to push this file to

Tinybird to automate creation of connections, similar to Kafka connections.

Now that your connection is created, you can create a Data Source to configure the import of files from S3.

The S3 import is configured using the following options, which can be added at the end of your `.datasource` file:

- `IMPORT_SERVICE`: name of the import service to use, in this case, `s3`
- `IMPORT_SCHEDULE`: either `@auto` to sync once per minute, or `@on-demand` to only execute manually (UTC)
- `IMPORT_STRATEGY`: the strategy used to import data, only `APPEND` is supported
- `IMPORT_BUCKET_URI`: a full bucket path, including the `s3://` protocol , bucket name, object path and an optional pattern to match against object keys. For example, `s3://my-bucket/my-path` would discover all files in the bucket `my-bucket` under the prefix `/my-path`. You can use patterns in the path to filter objects, for example, ending the path with `*.csv` will match all objects that end with the `.csv` suffix.
- `IMPORT_CONNECTION_NAME`: the name of the S3 connection to use

Note: For `IMPORT_STRATEGY` only `APPEND` is supported today. When new files are discovered, data from new files will be appended to any previous data in the Data Source. Replacing data is not supported.

For example:

S3.DATASOURCE FILE

```
DESCRIPTION >
    Analytics events landing data source

SCHEMA >
    `timestamp` DateTime `json:$.timestamp`,
    `session_id` String `json:$.session_id`,
    `action` LowCardinality(String) `json:$.action`,
    `version` LowCardinality(String) `json:$.version`,
    `payload` String `json:$.payload`

ENGINE "MergeTree"
ENGINE_PARTITION_KEY "toYYYYMM(timestamp)"
ENGINE_SORTING_KEY "timestamp"
ENGINE_TTL "timestamp + toIntervalDay(60)"
```

With your connection created and Data Source defined, you can now push your project to Tinybird using:

```
tb push
```

Supported file types

The S3 Connector supports the following file types:

- CSV
- NDJSON
- Parquet

Required IAM permissions

The S3 Connector requires certain permissions to access objects in your Amazon S3 bucket. The IAM Role needs the following permissions:

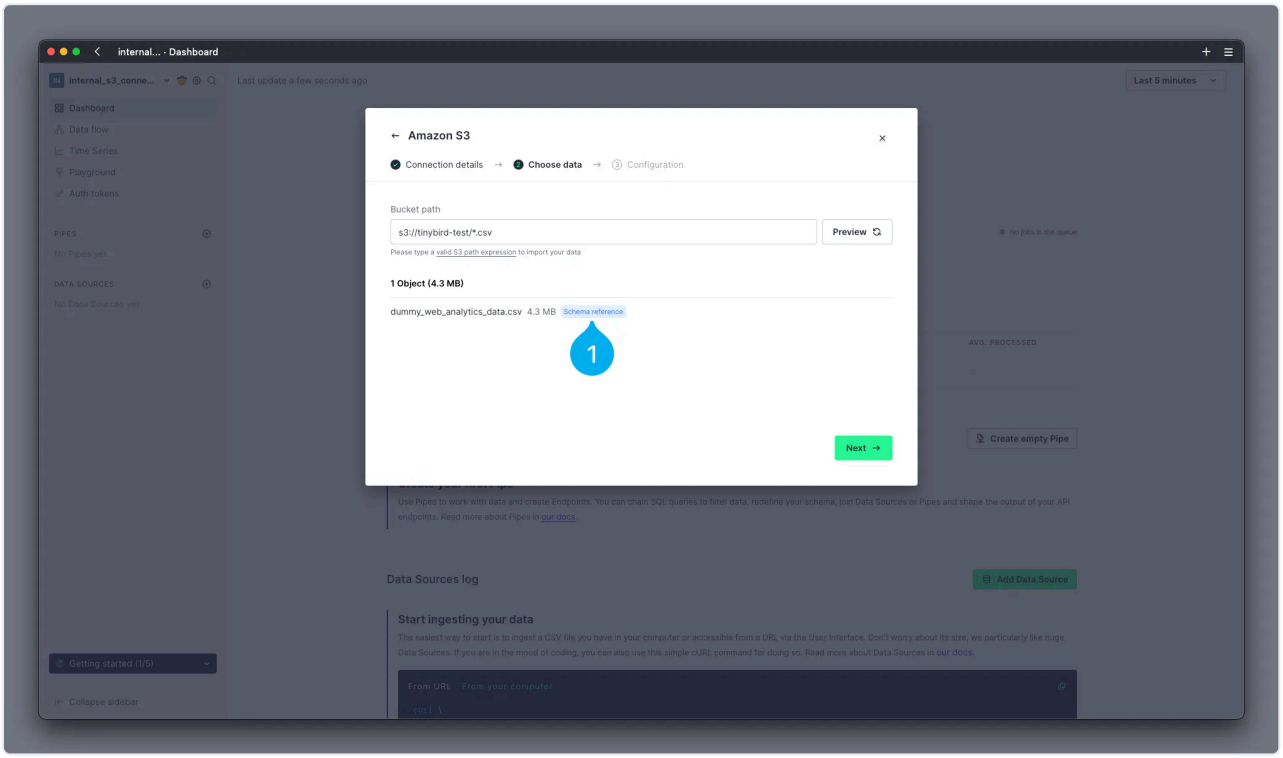
- s3:GetObject
- s3:ListBucket
- s3:ListAllMyBuckets

An example bucket policy would look like this (with bucket name and path replaced):

```
"Version": "2012-10-17",
"Statement": [
    {
        "Action": [
            "s3:GetObject",
            "s3:ListBucket"
        ],
        "Resource": [
            "arn:aws:s3:::`<bucket_name>/<path>`",
            "arn:aws:s3:::`<bucket_name>/<path>`/*"
        ],
        "Effect": "Allow"
    },
    {
        "Sid": "Statement1",
        "Effect": "Allow",
```

Schema evolution

When the S3 Connector first runs, it selects 1 file from the initial load and uses this to infer the schema of the Data Source. The file it selects is denoted by a blue "Schema reference" bubble (see Mark 1 below).



The S3 Connector supports automatic creation of new columns. This means that, if a new file contains a new column that has not been seen before, the next sync job will automatically add it to the Tinybird Data Source.

Non-backwards compatible changes, such as dropping, renaming, or changing the type of columns, are not supported and any rows from these files are sent to the Quarantine Data Source.

Limits

There are some limits applied to the S3 Connector when using the `auto` mode:

- Automatic execution of imports runs once every 1 minute.
- Each run will import at most 5 files. If there are more than 5 new files, they will be left for the next run.

> If you are regularly exceeding 5 files per minute, this limit can be adjusted. Contact us in our Slack community or email us at support@tinybird.co.

When using `on-demand`, these limits do not apply. A manual execution of the S3 connector will sync all new files available since the last run.

Spain
Calle del Dr. Fourquet, 27
28012 Madrid

USA
41 East 11th Street 11th floor
New York, NY 10003