Data freshness

Data Lakes and Warehouses offer different sync frequencies:

- Warehouses can sync up to once an hour, with the ability to set a custom sync schedule and selectively sync collections and properties within a source to Warehouses.
- Data Lakes offers 12 syncs in a 24 hour period, and doesn't offer custom sync schedules or selective sync.

Duplicates

Segment's 99% guarantee of no duplicates for data within a 24 hour look-back window applies to data in Segment Data Lakes and Warehouses.

Warehouses and Data Lakes also have a secondary deduplication system to further reduce the volume of duplicates to ensure clean data in your Warehouses and Data Lakes.

Object vs event data

Warehouses support both event and object data, while Data Lakes supports only event data.

See the table below for information about the source types supported by Warehouses and Data Lakes.

	Warehouses	Data Lakes
Website Libraries	≣	≣
Mobile	≣	≣
Server	≣	≣
Object Cloud Sources	≣	≣≣
Event Cloud Sources	≣	≣
HTTP	≣	≣
Pixel	≣	≣

Schema

Data types

Warehouses and Data Lakes both infer data types for the events each receives. Since events are received by Warehouses one by one, Warehouses look at the first event received every hour to infer the data type for subsequent events. Data Lakes uses a similar approach, however because it receives data every hour, Data Lakes is able to look at a group of events to infer the data type.

This approach leads to a few scenarios where the data type for an event may be different between Warehouses and Data Lakes. Those scenarios are:

- **Schema evolution** - Events reach Warehouses and Data Lakes at different times, due to their differing sync schedules. As a result, there is no way to guarantee that data types do not change since the field may have varying data types.
- **Different data type inferred based on sample size** - Warehouses and Data lakes use a different number of events to infer the schema. Warehouses receive one event at a time, and use the first received event to infer the data type. Data Lakes receive events in batches, and use a larger number of events to more accurately infer the data type.

Variance in data types between Warehouses and Data Lakes don't happen often for booleans, strings, and timestamps, however it can occur for decimals and integers.

If a bad data type is seen, such as text in place of a number or an incorrectly formatted date, Warehouses and Data Lakes attempt a best effort conversion to cast the fields to the target data type. Fields that cannot be casted may be dropped. Contact Segment Support if you want to correct data types in the schema and perform a replay to ensure no data is lost.

Tables

Tables between Warehouses and Data Lakes will be the same, except for in these two cases:

- `tracks` - Warehouses provide one table per specific event (`track_button_clicked`) in addition to a summary table listing all `track` method calls. Data Lakes also creates one table per specific event, but does not provide a summary table. Learn more about the `tracks` table in the Warehouses schema docs.
- `users` - Both Warehouses and Data Lakes create an `identifies` table (as seen in the Warehouses schema docs), however Warehouses also create a `users` table just for user data. Data Lakes does not create this, since it does not support object data. The `users` table is a materialized view of users in a source, constructed by data inferred about users from the identify calls.

- `accounts` - Group calls generate the `accounts` table in Warehouses. However because Data Lakes does not support object data (Groups are objects not events), there is no `accounts` table in Data Lakes.
- *(Redshift only)* **Table names which begin with numbers** - Table names are not allowed to begin with numbers in the Redshift Warehouse, so they are automatically given an underscore (_) prefix. Glue Data Catalog does not have this restriction, so Data Lakes don't assign this prefix. For example, in Redshift a table name may be named `_101_account_update`, however in Data Lakes it would be named `101_account_update`. While this nuance is specific to Redshift, other warehouses may show similar behavior for other reserved words.

Columns

Similar to tables, columns between Warehouses and Data Lakes will be the same, except for in a few specific scenarios:

- `event` and `event_text` - Each property within an event has its own column, however the naming convention for these columns differs between Warehouses and Data Lakes. Warehouses snake case the original payload value and preserves the original text within the `event_text` column. Data Lakes use the original payload value as-is for the column name, and does not need an `event_text` column.
- `channel`, `metadata_*`, `project_id`, `type`, `version` - These columns are Segment internal data which are not found in Warehouses, but are found in Data Lakes. Warehouses is intentionally very detailed about it's transformation logic and does not include these. Data Lakes does include them due to its more straightforward approach to flatten the whole event.
- *(Redshift only)* `uuid`, `uuid_ts` - Redshift customers will see columns for `uuid` and `uuid_ts`, which are used for de-duplication in Redshift; Other warehouses may have similar columns. These aren't relevant for Data Lakes so the columns won't appear there.
- `sent_at` - Warehouses computes the `sent_at` value based on timestamps found in the original event in order to account for clock skews and timestamps in the future. This was done when the Segment pipeline didn't do this on it's own, however it now calculates for this so Data Lakes does not need to do any additional computation, and will send the value as-is when computed at ingestion.
- `integrations` - Warehouses does not include the integrations object. Data Lakes flattens and includes the integrations object. You can read more about the `integrations` object in the filtering data documentation.

This page was last modified: 03 Aug 2023

Need support?

Questions? Problems? Need more info? Contact Segment Support for assistance!

Visit our Support page

Help improve these docs!

Edit this page

Request docs change

Was this page helpful?

Yes

No

Get started with Segment

Segment is the easiest way to integrate your websites & mobile apps data to over 300 analytics and growth tools.

Your work e-mail

Request Demo

or

Create free account