
BigPrivacy®

SaveYourData®

User Guide

Rev 20183008.4

Information contained herein is subject to change and is not guaranteed to be error-free.

Anonos has been actively engaged in research and development to advance the state of the art in global data protection, privacy and security technology since 2012. Anonos BigPrivacy systems and technology are protected by an intellectual property portfolio that includes, but is not limited to: (2018) SYSTEMS AND METHODS FOR ENHANCING DATA PROTECTION BY ANONOSIZING STRUCTURED AND UNSTRUCTURED DATA AND INCORPORATING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE IN CLASSICAL AND QUANTUM COMPUTING ENVIRONMENTS – NO. 10,043,035; (2017) SYSTEMS AND METHODS FOR ANONOSIZING DATA – NO. 9,619,669; (2016) SYSTEMS AND METHODS FOR CONTEXTUALIZED DATA PROTECTION – NO. 9,361,481; (2015) DYNAMIC DE-IDENTIFICATION AND ANONYMITY – NO. 9,129,133; (2015) DYNAMIC DE-IDENTIFICATION AND ANONYMITY – NO. 9,087,216; (2015) DYNAMIC DE-IDENTIFICATION AND ANONYMITY – NO. 9,087,215; plus 60+ additional U.S. and international patent applications.

See anonos.com/patents for more information.

Contents

About the User Guide

This document provides guidance for user-facing features provided in the BigPrivacy® SaveYourData® computer software application ("SYD"). It provides for connecting to source (ingest) databases, extracting data, and pseudonymising the data as the initial step in transitioning "to GDPR compliance by basing data processing on a different

lawful basis while ensuring that continued processing is fair and accounted for" by supporting "legitimate interests" as a lawful basis for ongoing processing.

Document Conventions

The following conventions are used throughout this document to highlight important cautions and notes.

- The BigPrivacy SaveYourData software application is routinely referred to herein as “SaveYourData” or “SYD.”
- Labels of clickable elements in the user interface, like button names, are always **Bold.**
- Screen names and field labels are capitalized, like Audit screen.
- Text entered at the command line or in configuration files uses the `Courier New font`.
- Important notes are tagged with the notes icon 

Additional Resources

For stakeholders in the installation process, or admin personnel responsible for maintaining the system, please also see,

- SaveYourData Installation Guide

Topics Covered

The User Guide provides instructions for extracting and processing ("pseudonymising") data using the SaveYourData application.

The major topics are,

- About the User Guide
- Introduction
- Getting Started
- Managing Connection Profiles
- Auditing User Activity
- Troubleshooting

Who Should Read This Guide?

The primary audience for this guide consists of the users responsible for starting and monitoring ingest jobs in the data center. Ideally, they will have some familiarity with database administration and relevant concepts, but anyone with valid credentials for an ingest database can run the application.

Furthermore, any stakeholder with an interest in the structure, functionality, and design of SaveYourData may find portions of this document informative.

Version and Release Notes

This guide supports SaveYourData version 1.0.0, the first GA release.

Release 1.0.0 Highlights

- **Regulatory Pseudonymisation**
  Pseudonymisation of ingest datasets up to 50 terabytes.

- **Support for Multiple Simultaneous Ingest Jobs**
  Up to four simultaneous jobs. Automatically queues and starts additional jobs started by the user.

- **Extensive Database Support**
  Full JDBC read-only support for ingest (source) databases.

- **Saved Connection Profiles**
  Ability to save database connection profiles for repeated use.

- **Audit Record**
  Logging of events and user activity.

- **Table Selection**
  Option to select all tables, or a subset of tables for ingestion.

- **Active Directory Integration**
  Authentication against Active Directory for user login.

Getting Technical Support

Support is available from Hitachi by phone or email.

Phone	+44-203-608-9365
Email	hitachisupport@bigprivacy.com

To accelerate resolution of your issue, please confirm basic connectivity and availability of customer-owned components prior to contacting support.

Introduction

TOPICS

- What is SaveYourData?
- Important Concepts
- Architectural Notes
- Support for Specific Data Types
- Getting Support

1

What is SaveYourData?

SYD keeps an organization's options open for potential ongoing use of personal data collected using (now under the GDPR) legally non-compliant broad-based consent ("Legacy Personal Data") without requiring (i) deletion or (ii) anonymisation of the data.

Regulatory guidelines issued in connection with the GDPR require organizations that previously relied upon consent to review their pre-GDPR consents to ensure they are compliant with new heightened GDPR requirements for consent. If the review reveals that the prior consent is not valid under the GDPR, data controllers are provided the following "one off" opportunity to get GDPR compliant consent or to change the lawful basis on which the relevant processing takes place:

> *If a controller finds that the consent previously obtained under the old legislation will not meet the standard of GDPR consent, then controllers must undertake action to comply with these standards, for example by refreshing consent in a GDPR-compliant way. Under the GDPR, it is not possible to swap between one lawful basis and another. If a controller is unable to renew consent in a compliant way and is also unable – **as a one off situation** – **to make the transition to GDPR compliance by basing data processing on a different lawful basis while ensuring that continued processing is fair and accounted for,** the processing activities must be stopped. In any event the controller needs to observe the principles of lawful, fair and transparent processing.*

Many organizations are being advised to delete Legacy Personal Data because they are unaware that technical solutions such as SaveYourData exist to enable data controllers and processors to "pseudonymise" data in accordance with the requirements of Article 4(5) of the EU General Data Protection Regulation (GDPR) to help support lawful secondary uses of data, like iterative analytics and artificial intelligence (AI), when consent does not provide a valid legal basis under the GDPR because processing cannot be described with specificity and unambiguity at the time of data collection.

If an organisation is subject to regulatory retention or reporting obligations, this may mean locking up Legacy Personal Data so that it is accessible only in response to regulator inquiry. Alternatively, organizations are being advised to "delete" or "anonymise" their Legacy Personal Data so that relinking (directly or indirectly) to identifying data is no longer possible. In either situation, this means that access to Legacy Personal Data for analytics, artificial intelligence, machine learning, or digital transformation may no longer be possible.

SaveYourData provides a third option instead of (i) deleting or (ii) anonymising Legacy Personal Data. SYD represents a "one off" opportunity to transform Legacy Personal Data for potential future secondary use such as analytics, artificial intelligence, machine learning or digital transformation. SaveYourData accomplishes the initial step in transitioning "to GDPR compliance by basing data processing on a different lawful basis while ensuring that continued processing is fair and accounted for" by pseudonymising personal data to help support "legitimate interests" as a lawful basis for processing.

Important Concepts

Listed below are the key concepts related to taking the initial step in transitioning “to GDPR compliance by basing data processing on a different lawful basis while ensuring that continued processing is fair and accounted for" by supporting "legitimate interests" as a lawful basis for ongoing processing, the function provided by SaveYourData. Please Reference Exhibit 1 for Details on SYD Regulatory Background.

Connection Profile

The data record SYD uses to connect to ingest databases in the end-user IT environment. Connection profiles are displayed in the user interface of the application, on the Main screen. Each profile is represented visually as a user-named 'card', each card representing a single database and table selection.

Customer

The end-user organization that purchased SYD. The customer provides the host environment for SYD. They are typically managed and supported by a systems integrator or similar VAR.

Destination database

The customer-provided PostgreSQL(10) database that SYD writes pseudonymised data to. It contains the output data from an ingest job. That data is referred to variously as "pseudonymised" or "GDPR-compliant" data.

ETL

Extract, Transform, and Load. The general category of processing performed by SYD.

GDPR

The EU General Data Protection Regulation that defines "pseudonymisation" and provides guidance that defines what is and what is not compliant data.

Ingest database

The database used as the original source of data process by an ingest job.

Ingest Job

The process of connecting to a database, ingesting its dataset, processing that dataset for GDPR compliance, and writing the pseudonymised version of the dataset to the destination database.

Installation Team

The person or persons from the systems integrator who perform the on-site installation for the customer.

JDBC

Java database connector, an API that abstracts logical database operations from the platform.

Mosaic Effect

The “Mosaic Effect” occurs when a person is indirectly identifiable due to a phenomenon referred to by the Article 29 Working Party as “unique combinations” where, notwithstanding the lack of identifiers that directly single out of a particular person, the person is still “identifiable” because that information may be combined with other pieces of information known to relate to the same individual (whether the latter is retained by the data controller or not) to create a “mosaic” of the person, enabling the individual to be distinguished from others. To help address this issue, SaveYourData replaces each occurrence of the same data element with a different dynamically generated token to separate the information value of data from the risk of re-identification via the Mosaic Effect.

MFA

Multi-factor authentication.

Pseudonymisation

GDPR Article 4(5) defines “Pseudonymisation” as requiring separation of the information value of data from the risk of re-identification. To benefit from GDPR statutory/regulatory incentives and rewards for pseudonymisation, this separation is necessary. Replacing multiple occurrences of the same personal data elements with “static” (or persistent) tokens fails to separate the information value of data from the risk of re-identification because re-identifying correlations and linkage attacks (aka the “Mosaic Effect”) are possible because “static” (or persistent) identifiers are used instead of dynamic de-identifiers. The “Mosaic Effect” occurs when a person is indirectly identifiable due to a phenomenon referred to by the Article 29 Working Party as “unique combinations” where, notwithstanding the lack of identifiers that directly single out of a particular person, the person is still “identifiable” because that information may be combined with other pieces of information known to relate to the same individual (whether the latter is retained by the data controller or not) to create a “mosaic” of the person, enabling the individual to be distinguished from others. To help address this issue, SaveYourData replaces each occurrence of the same data element with a different dynamically generated token to separate the information value of data from the risk of re-identification via the Mosaic Effect.

SYD

An acronym for SaveYourData. It's used throughout this document.

SYD Access Group

The group defined in the corporate directory (AD or other LDAP 3+) whose members consist of users who are permitted to access SYD.

SYD Installation Image

VMware ESXi-compatible SaveYourData (SYD) virtual machine (VM) image provided by Anonos to the installation team.

Systems Integrator

The umbrella term for the type of company that manages and performs the SYD installation, manages the customer, and provides front-line support for the customer and their users. (also, system integrator).

Architectural Notes

This section provides some initial context in a brief overview of SYD's major components, dataflow, and supported data types.

The SYD application is designed to support ingest and pseudonymisation of large data sets from RDBMS sources that contain private or otherwise sensitive data. It is compatible with any data platform that supports the JDBC API for platform-agnostic, read-only access. SYD runs in your VM environment in containerized modules that are lightweight and straightforward to maintain and update.

Three containers are deployed during installation. They are,

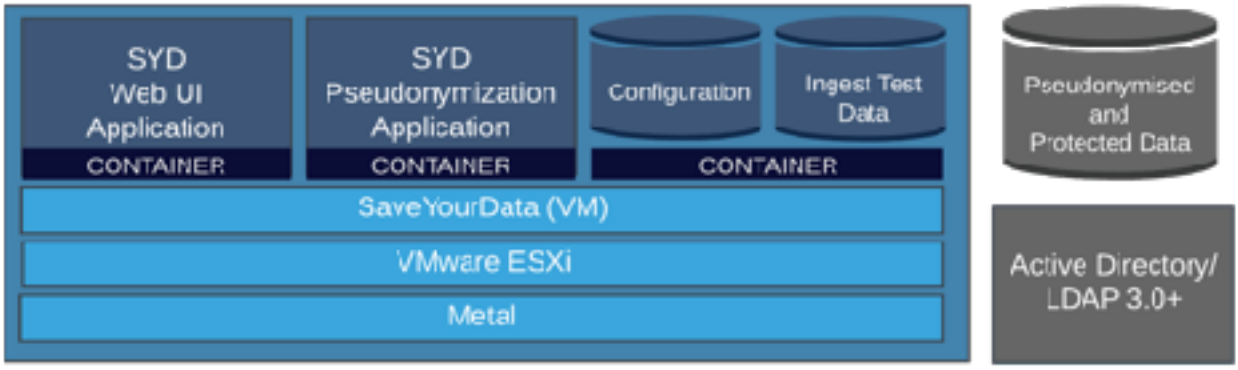
1. **Web UI Application**
   This is the presentation layer of the application, and the single interface for end-users.

2. **Pseudonymisation Application**
   This is the core engine responsible for pseudonymising ingest data and writing it to the destination database.

3. **Configuration and Ingest Test Database**
   A container with two database schemas – a lightweight configuration database with preconfigured values, and a data source to use for initial post-install validation tests.

   The configuration tables also store the connection profiles displayed on the Ingest Data page, and the audit events that are shown on the Audit page.

   Note that some configuration parameters are also stored separately, in flat files that are edited at the time of installation.

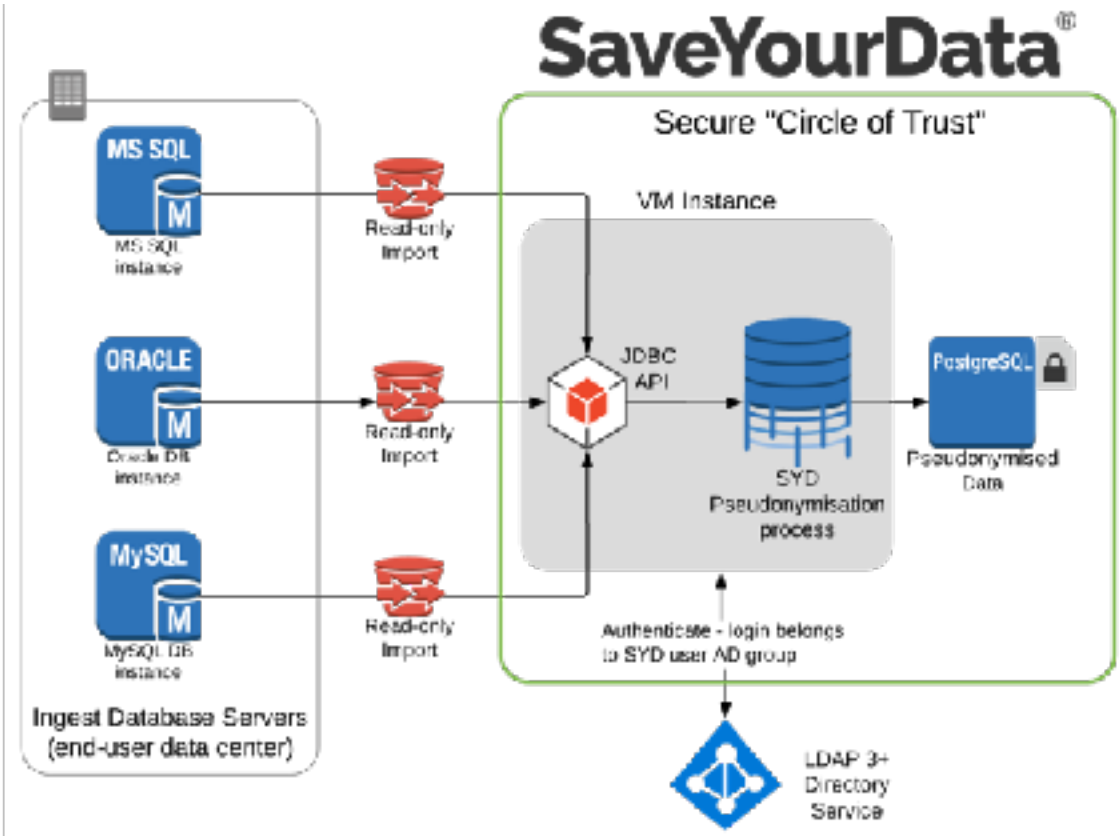


Major SYD Architectural Components

The diagram below summarizes the data flow in a deployed system, starting from the customer's data center, through pseudonymisation in SYD, and ultimately back into the data center in pseudonymised form.

SYD End-to-End Data Path

Level1 DFD



NOTES

1. The system connects with read-only access to any JDBC-compliant RDBMS in the customer's data-center. Up to four connections can be active at a given time, and the system can store connection profiles for N total ingest data platforms without restriction.

2. Directory services conforming to the LDAP 3.0+ standard are used to authenticate users, and confirm their membership in the group authorized for SYD access.

3. SYD Pseudonymisation process (or "ingest job") is initiated by a user, and pseudonymised data is written to the designated output RDBMS in the end-user data center.

Compatibility with Specific Data Types

SYD's ingest system supports the data types specified in the SQL-92 standard:

Supported Numeric Types

```
SMALLINT, INTEGER, INT
DECIMAL, NUMERIC, FLOAT
REAL, DOUBLE PRECISION
BIT, BIT VARYING
```

Supported String Types

```
CHARACTER, CHAR
NATIONAL CHARACTER, NATIONAL CHAR, NCHAR
CHARACTER VARYING, CHAR VARYING, VARCHAR
NATIONAL CHARACTER VARYING, NATIONAL CHAR VARYING, NCHAR VARYING
```

Supported Date and Time Types

```
DATE, TIME, INTERVAL
TIME WITH TIME ZONE, TIMESTAMP, TIMESTAMP WITH TIMEZONE
```

If the system encounters unsupported data types, like XML or LOB variations, the performance of the ingest process may become unpredictable.

Getting Started

TOPICS

- Overview
- Logging in For The First Time
- Main Screen at a Glance
- Preparing to Run an Ingest Job
- Connecting to a Data Source
- Running Your First Ingest Job
- Deleting Ingested Data

2

Overview

This section covers basic usage of the *SaveYourData* application – specifically how to use its primary function: transforming your existing datasets into pseudonymised data, a process referred to as running an "ingest job".

Before looking at the detailed mechanics behind pseudonymising data, a high-level orientation will help introduce the process and define some of the language used throughout this document. After reviewing the steps involved in running an ingest job below, we'll log in to the application.

Steps to Running an Ingest Job

1. **Define a Connection Profile**

WHAT IS IT?	A connection profile is a user-defined object that contains the access information for a single ingest database. They're displayed as individual profile 'cards' on the Ingest Data page, and one is required in order to run an ingest job. They're saved for repeated use.
HOW TO DO IT?	Click the **CREATE NEW DATABASE** button on the Ingest Data screen. Enter database connection info into the form.
TERM TO KNOW?	*Profile cards* – The selectable tiles on the main screen that each represents a connection profile.

2. **Connect to an Ingest database**

WHAT IS IT?	The ingest database is an RDBMS platform in your data-center that provides source data for a given ingest job. You must connect to it before running an ingest job.
HOW TO DO IT?	Click the **CONNECT** link on its profile card on the Ingest Data screen and enter its password at the prompt. IMPORTANT – Please pay particular attention to the instructions related to selecting and saving ingest tables.
TERM TO KNOW?	*Ingest database* – The database providing source data for a given ingest job. Access to it is defined in a profile.

3. **Run the Ingest Job**

WHAT IS IT?	The process of ingesting data, pseudonymising it, and writing it to a separate database.

HOW TO DO IT?	Click the **START DATA INGEST** link profile card for the desired ingest database.
TERM TO KNOW?	*Ingest job*

Logging in for the First Time

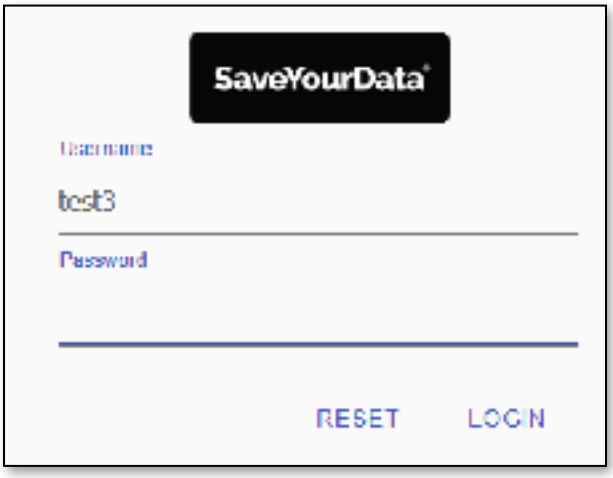
After you log in and begin to explore the interface, the high-level process on the previous page will start to take on more meaning.

Your supervisor, vendor, or support channel can provide you with the URL assigned to access the application on your network.

Because SYD authenticates against your directory service, your existing username and password should let you login, provided you're also part of the SYD authorized user group defined in the directory. If the proper directory permissions aren't yet set up, your vendor or support channel can provide a temporary username a password to use for login.

To log in, browse to the SaveYourData URL provided.

The login screen appears.



Enter your username and password and then click **LOGIN**.

The login authenticates against your LDAP3-compliant directory service, and each user must be part of the designated group of SYD users in ActiveDirectory before they can access the application. (The group name for authorized users is set at installation.)

The **RESET** button clears text from the form fields.

Login Errors

If a login error occurs, it's displayed in red below the login form. The two possible errors are essentially 1) invalid credentials, and 2) the user doesn't belong to the LDAP group of authorized SYD users.

`error: Authentication credentials are incorrect, please try again`
This indicates that the system failed to find the username and password combination in the directory.

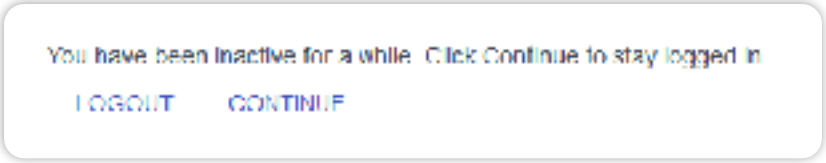
`error: User <USERNAME> is not part of <GROUPNAME>`
This indicates that the system found the user in the directory but did not recognize that user as a valid member of the SYD authorized users group. Note that the name your company uses is configurable and is set at installation. It can be any valid LDAP group name.

Logout

To log out of a SYD session, click the **Logout** link in the left-hand side navigation pane. Logging out does not affect ingest activity in any way, so jobs that are running when you log out continue to run to completion.

Login Timeout

If the system detects no activity for 5 minutes, a popup dialog warns the user they're about to be logged out and offers the choice to remain logged in or log out of the application.



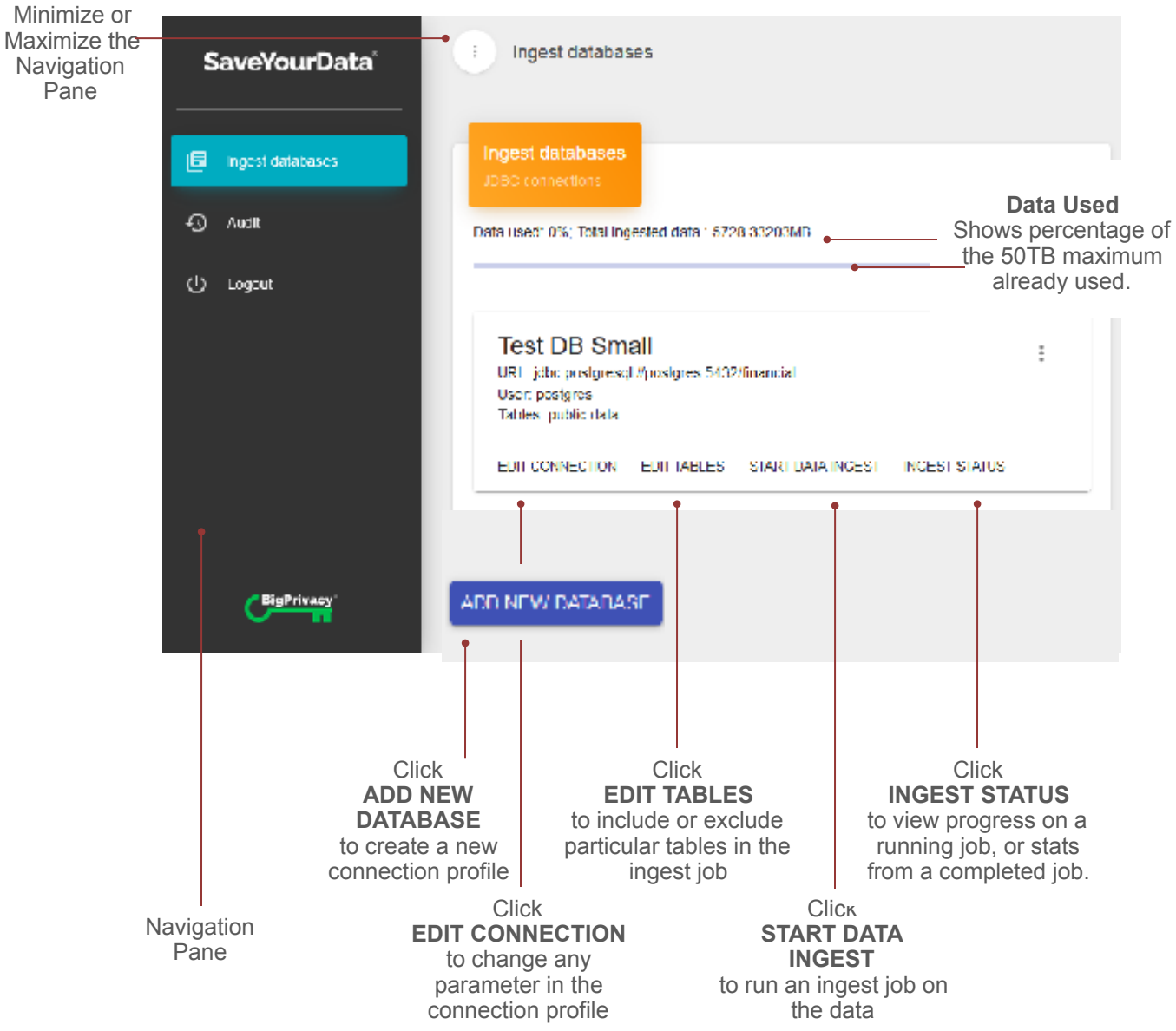
Click **CONTINUE** to reset the timer and remain logged in.

Ingest Database Screen at a Glance

By default, the Ingest Database screen appears immediately after login. All ingest jobs are run from this screen. It's also the primary interface for managing your database connections.

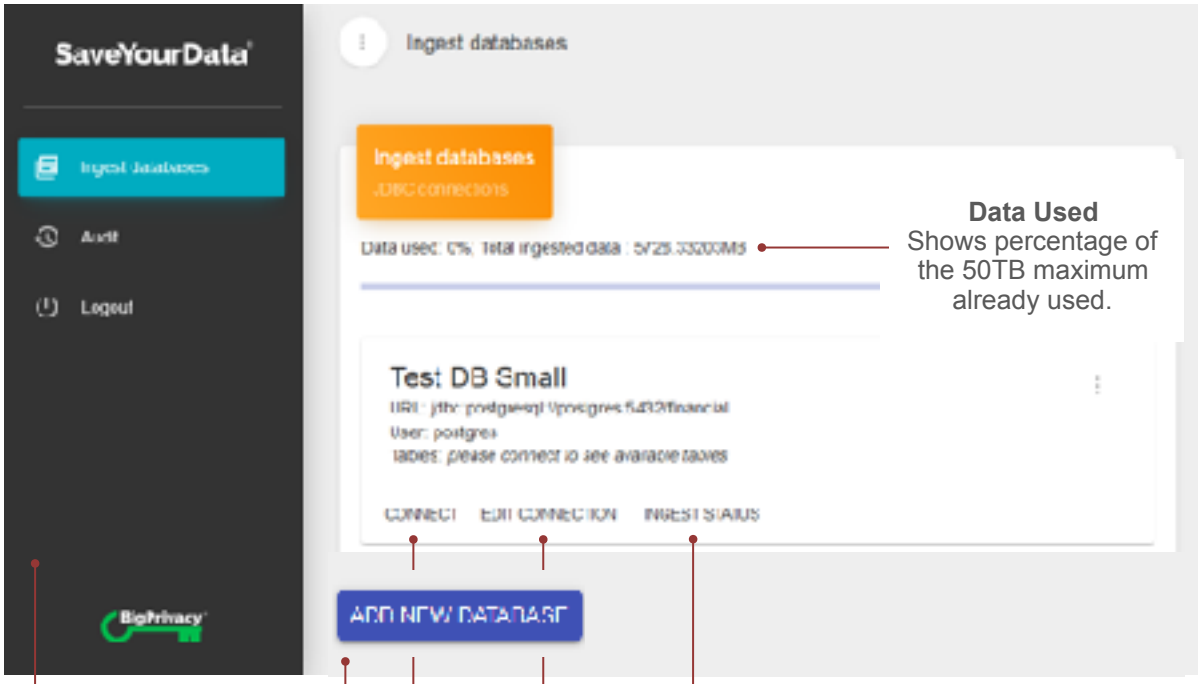
The screen shot below shows the Ingest Database screen with just one profile displayed, although there is no limit to the number of profiles you can create. All the 'action' controls for that database are along the bottom of the profile 'card'. Note that the system is currently connected to the database, as indicated by the fact that the **START DATA INGEST** link is active.

Ingest Databases Screen with "Test DB Small" profile – connected



Below is the same data profile card shown on the previous page, without an active connection to the database, as reflected by the different set of options along the bottom of the card, including the **CONNECT** option.

Ingest Databases Screen with "Test DB Small" profile – NOT connected



Data Used
Shows percentage of the 50TB maximum already used.

Navigation Pane

Click **ADD NEW DATABASE** to create a new connection profile

Click **EDIT CONNECTION** to change any parameter in the connection profile

Click **CONNECT** to connect to the database prior to ingestion

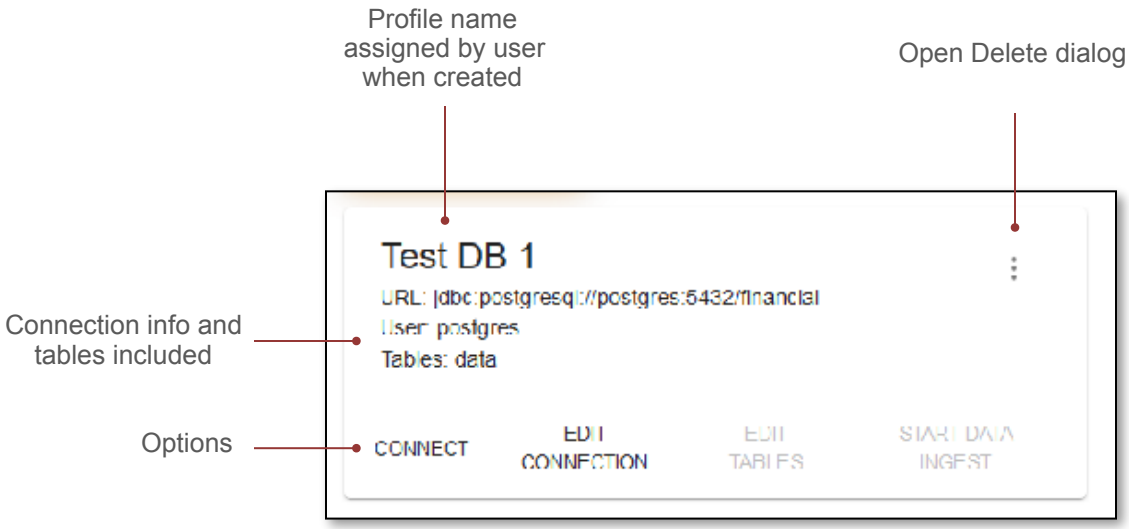
Click **INGEST STATUS** To see stats and a running or completed ingest job

Minimize/Maximize the Navigation Pane

Note the *more options* icon at the top of the page, just to the right of the navigation pane. Click this to toggle the navigation pane between minimized and maximized states.

Database Connection Profiles

All saved connections are displayed in a grid of connection profile 'cards' on the Ingest Data screen. Each card contains links to perform all the available functions on an ingest database – connect to the data source, test the connection, edit the profile, and run and monitor ingest jobs.



Profile Card

Navigation Pane

The navigation pane on the left allows you to navigate between the main Data Ingest screen, and the Audit screen, which shows a timeline of user activity.

TO COLLAPSE OR EXPAND THE NAVIGATION PANE

1. Click the *more options* icon ⋮ on the top left of the screen and the navigation pane collapses.
2. Click again to toggle back.

Preparing to Run an Ingest Job

From a user's perspective, running an 'ingest job' refers to simply clicking **START DATA INGEST** on the relevant profile card. That triggers the entire process of ingesting and converting source data to a pseudonymised dataset, then writing it to the destination database.

There are three factors to consider carefully prior to running a job.

1. **Consideration #1: Job Duration**
   Estimating the job duration may help set expectations for management and can inform scheduling decisions.

2. **Consideration #2: Compatibility and Availability**
   Make sure your database is compatible with the JDBC standard and SQL-92 data types. Also consider how the ingest platform is currently deployed and used, and how running an ingest job may affect other users of that platform.

3. **Consideration #3: Tables Selected**
   Understand the implications of selecting or deselecting specific tables for inclusion in the ingest job.

Each of these considerations is discussed in more detail on the following pages.

Consideration #1: Job Duration

The maximum data transfer rate across your network dictates the minimum possible duration of a given ingest job.

That rate ultimately depends on multiple factors, some of which are related to the infrastructure between SYD and a given ingest database. Although processing the data takes CPU time, the job duration is governed by data throughput. That means that the transfer rate across your network, and the I/O rate to the destination database are the limiting factors.

The table below shows some sample job estimates based on transfer rate and data volume.

Ingest Job Duration Estimates

Ingest Data Volume - Rate	1 Gbps	10 Gbps	40 Gbps
1 GB	0:00:00:09	0:00:00:01	>0:00:00:01
2 GB	0:00:00:17	0:00:00:02	>0:00:00:01
50 GB	0:00:07:09	0:00:00:43	0:00:00:11
100 GB	0:00:14:19	0:00:01:26	0:00:00:21
1 TB	0:02:26:36	0:00:14:40	0:00:03:40
2 TB	0:04:53:12	0:00:29:19	0:00:07:20
50 TB	5:02:10:05	0:12:13:00	0:03:03:15

Format: days:hh:mm:ss

Of course, multiple factors that can affect network performance are transient in nature. Even if you have an established bench-mark for the data transfer rate in your data center, there may be dynamic variations in load across your network that degrade the actual performance of any given job.

A convenient resource for calculating transfer times in various units can be found at, https://www.expedient.com/file-transfertime-calculator.

Jobs Continue Running After Users Log Out
Logging out of the application doesn't stop jobs that are already running.
Jobs run until the entire ingest dataset is processed, regardless of user login status.

Consideration #2: Compatibility and Availability

To help ensure the success of an ingest job, there are three points of compatibility to consider before running it.

JDBC Compatibility

Because SYD uses an open, standards-based API (JDBC) for database connectivity, it's important to ensure that the correct JDBC drivers are deployed in the SYD VM.

If SYD is using deprecated drivers, it will not be compatible with your RDBMS platforms, even if it's listed as compatible in the dropdown on the Add New Database dialog. (See the *SaveYourData Installation Guide* for more information.)

Compatible Data Types

SYD supports SQL92-compliant data types. Some data types that came into use after that standard was established are not yet supported by the SYD ingest engine. When SYD encounters an unsupported data type, it will attempt to convert it to a string or automatically exclude it from the job, however it may also cause the job to fail, so reviewing data-type compatibility up front is recommended.

Incompatible data types include,

- XML
- LOB types
- User-defined types

See also the section entitled *Architectural Notes* in the *Introduction*.

Database Load

Before running an ingest job, consider the usage and load factors on the ingest database, and how that might impact other clients that use it. You may have the hostname and access credentials for the source database, but how familiar are you with the way it's being used in other critical processes in your organization?

SYD's ingest process can put considerable load on RDBMS resources on the ingest platform. Especially if the source database is in production or gets regular use by other

clients, the potential for service disruption is nontrivial. If there's a potential conflict, consider scheduling ingest at off-peak times, or using an intermediate platform.

It's one thing to ingest an archival database that's under relatively light load, it's another to pull data in competition with other clients, especially those in mission critical roles.

Consideration #3: Specific Tables Selected

By default, all tables in the ingest database are selected for inclusion upon creation of a new connection profile.

But SYD also lets users specify which tables are included in a job by clicking the **EDIT TABLES** link on the profile card and selecting or deselecting tables individually.

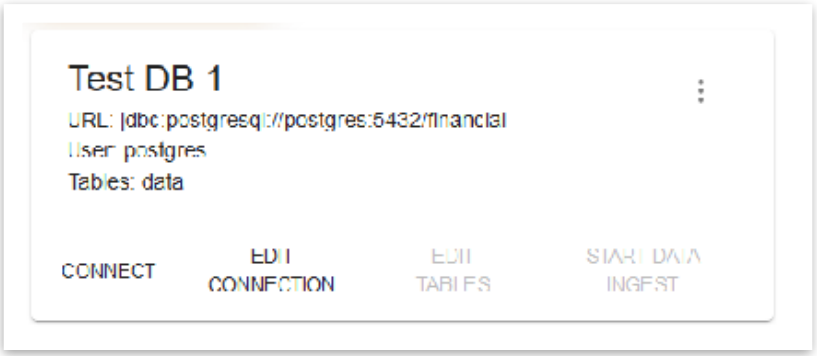
Verify that the desired tables are included in a job prior to initiating it. The included tables are displayed on the profile card, and best practice is to check the card prior to running a job.

Running a long job doesn't benefit from loading tables that won't be used. Conversely, running a long job then realizing that it's missing mission-critical tables means starting again from the beginning.

Speed up ingest jobs by excluding unneeded data
Transferring data that won't be used unnecessarily extends the job duration. Consider deselecting those tables that will go unused after pseudonymisation.

Connecting to an Ingest Data Source

SYD keeps connection information for each database in a user-created 'connection profile'. A valid connection profile is required for the ingest (source) database before you can run an ingest job on it.



Each saved connection profile is represented visually by a profile 'card' on the Ingest Data screen. Profile cards are laid-out on the screen in a grid, and each card describes a single connection profile for a single ingest database.

The user names the profile when it's created, and that name, along with the database URL is displayed on each card to identify it uniquely.

Each connection profile contains all the information required to connect to a database and run an ingest job from start to finish. That information includes the unique, user-defined profile name, the database server host name or IP address, the username, the RDBMS type, and the particular tables to be included in the ingest job.

Creating a new profile is as simple as entering the information in the Add New Database form. Be sure to have the username the host name or IP address for the database server at hand before you start.

You can create and save as many connection profiles as you need. There is no practical limit.

The procedure on the following page describes the steps required to create a profile.

TO CREATE A CONNECTION PROFILE

1. On the Ingest Data screen, click the **ADD NEW DATABASE** button.

   The Add New Database dialog opens.

2. Enter a descriptive name for your profile. The profile name is used only as a reference for users. It's not used as part of the connection string, but it should be something easily associated with the data source and particulars of the job.
3. Select the type of database from the database **Type** dropdown.
4. Enter the host name or IP address of the database server.
5. Enter the IP port number to use for the connection.
6. Enter the name of the database schema to connect to.
7. Enter the username that will be used to login to the database.
8. Click **Create.**

The new profile should appear in a card on the main screen, labeled with the name you gave it.

Now that you've created a connection profile, there's one additional and crucial step required before you can run an ingest job.

Selecting Specific Tables for an Ingest Job

When you create a new connection profile, all tables are selected for ingestion by default. However, when you define a new profile, you must click **EDIT TABLES**, which takes you to the Ingest Tables screen, and from there you must click **SAVE** before returning to the Ingest Data screen. The START DATA INGEST link simply won't become available until you've performed this task, whether you select particular tables or not

You only have to go to the Ingest Tables screen and click **SAVE** once for a given profile card – after you create it, but before you run an Ingest Job. You can run subsequent ingest jobs from the card without the requirements to go to the Ingest Table screen at all.

Deselected Tables are NOT processed
Any tables not selected for ingest will not be ingested or processed by SYD in any way and are not represented in the output dataset.

Let's take a look at table selection in more detail...

TO MANAGE TABLE INCLUSION

1. Click the **EDIT TABLES** link on the profile card. The Ingest Tables screen opens.

2. If you just created the profile, you should see all tables in the database selected. That's the default setting, HOWEVER, on a newly created profile, EVEN IF YOU DON'T EDIT TABLE SELECTIONS, YOU MUST GO TO THE EDIT TABLES SCREEN AND CLICK **SAVE** BEFORE YOU RUN AN INJEST JOB AGAINST THE PROFILE.

3. If you'd like to select particular tables to be included or excluded from the job, click the checkbox to the left of the table name to select or deselect the table for inclusion.

4. You can click the top checkbox to the left of 'Table Name' to toggle all tables between a checked and an unchecked state. If you want to include only a few tables, use the top checkbox to toggle all to unchecked, then select the ones you need.

5. Be sure to click **SAVE** before leaving this screen whether you changed any checkboxes or not.

When Defining a new Connection Profile, you MUST go to the Ingest Tables Screen
Even if you're not selecting or deselecting particular tables, you must click **EDIT TABLES** to go to the Ingest Tables screen, and then click SAVE before you can use your new profile.

Running Your First Ingest Job

Recall that ingest jobs are initiated from the Ingest Data screen by clicking the **START DATA INGEST** link directly on the profile card. That link is enabled only after connecting to the database AND clicking SAVE on the Ingest Tables screen, so you'll have to first click **CONNECT** if you haven't yet done so.

If the database is already configured in a connection profile, and you're connected, and have saved the Ingest Tables at least once, it takes just a single mouse-click to run the job.

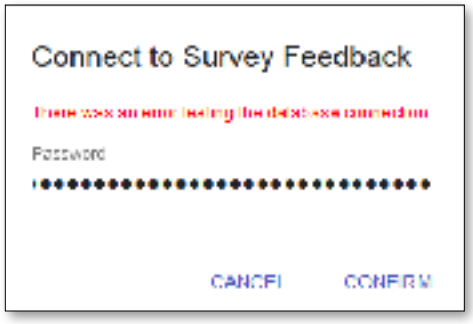
If this is your first run, we suggest using a small dataset, less than 1 GB is ideal, and although the system allows you to select a subset of tables to include in the job, it's recommended that the first run include all the tables in the ingest database.

TO RUN AN INGEST JOB

1. On the Ingest Data screen, connect to the ingest database by clicking the **CONNECT** link on its profile card. The database login dialog appears (database passwords are not stored in connection profiles.).

2. Enter the password for this database and click **OK**

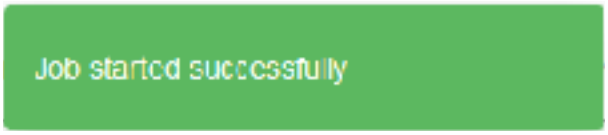
3. If the connection test fails, an error dialog opens saying "There was an error testing the database connection".



   Double-check your credentials, as well as the host name, and IP port, then retry. You can also troubleshoot the connection externally to SYD to help isolate the issue. See also *Errors Connecting to an Ingest Database*, following this section.

4. Click the **EDIT TABLES** link on the profile card and review the selected tables. If this profile is newly created, you must click **SAVE** on this screen to enable the ingest functionality for the database. Click on the **Ingest Data** button in the navigation pane to return to the database card view.

5. To run the job, click the **START DATA INGEST** link on the card. The link should gray out and remain inactive while the job runs, and you should see the green popup below.



Datasets not to Exceed 50TB
SYD can accommodate up to 50TB from the ingest database. If you attempt to ingest data that exceeds that size, the ingest job will fail without warning.

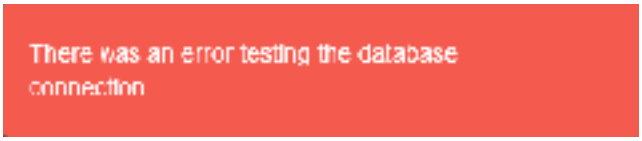
While the ingest job runs, its profile card displays a spinning "Ingestion in progress" animation.

Note also that while the job is running the **INGEST STATUS** link remains active. Click that at any time to see real time, table-by-table progress of the active job.

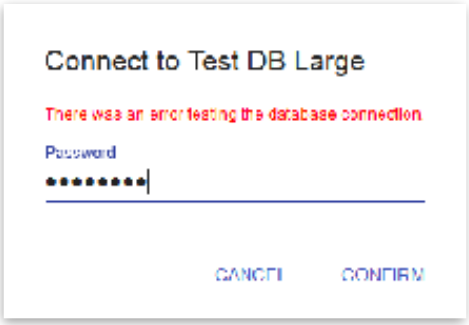
Errors Connecting to an Ingest Database

When you enter the database password and click **CONFIRM**, the system tries to connect to the database defined in the profile.

If the connection fails, whether it's a network fault, the wrong password, an invalid connection string, or just that the target RDBMS platform is offline, the error message shown below will pop up.



The password dialog will also indicate the connection failure.



Follow the same troubleshooting pattern you would for any connection failure between platforms in the data center – double-check the password and connection parameters, ping the target platform to confirm LAN connectivity, and try logging in to the database, and running a simple query on it through your DB management tool. Use the same credentials that are in SYD.

Ingest Job FAQs

Can I start more than one job at a time?

You can run up to four ingest jobs simultaneously. If you start additional jobs while four are already running, each will be queued and automatically run as other jobs complete.

Is it possible to accidently delete data from our ingest databases using the SYD interface?

No. SYD connects to ingest databases as a read-only client. To manage pseudonymised data in the destination database, SYD provides some tools to delete an entire dataset, or to delete individual records that hold data matching user-specified values.

If tables from multiple databases are written to the same destination database, how does the back-end process determine the table's origin?

Tables with pseudonymised data are written to the destination database under a table name that deterministically identifies the original database and table name.

Managing Data & Connections

TOPICS

- Overview
- Securing Pseudonymised Data
- Deleting Pseudonymised Data
- Managing Connection Profiles

3

Overview

The need to maintain appropriate access restrictions and security in the VM is emphasized at installation, but users should also be aware of important best practices in regard to protecting the contents of the destination database. Each SaveYourData user is responsible for implementing and maintaining necessary security procedures and practices in a manner compliant with industry standards, laws, rules and regulations to protect the output of SaveYourData processing from unauthorized access, destruction, use, modification, or disclosure.

SYD offers some simple tools to help manage both the data profiles, and the pseudonymised data in the backend database, and the next pages discuss the available options.

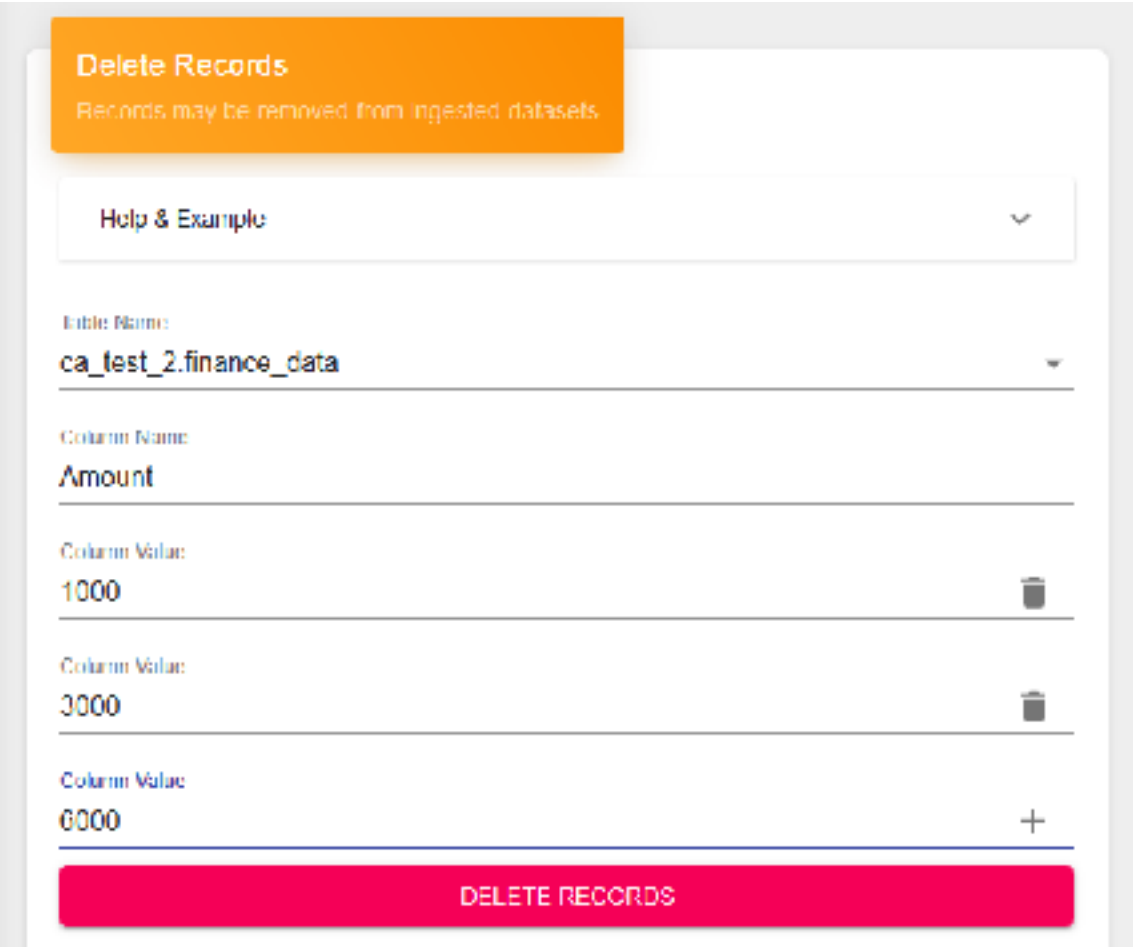
Deleting Pseudonymised Data

After running a job, the ingested data remains stored in the destination database until it's deleted. Although that data is fully pseudonymised, SYD gives you the option to delete rows that have fields matching particular values, such as your internal policies dictate.

The driving motivation behind the delete option is that SYD is designed to respect the spirit of the "right to be forgotten", a common tenet underlying modern thinking about data privacy regulation. It does so by allowing users to delete the entirety of a pseudonymised dataset, or particular rows based on columns that contain data matching specified values.

Deleting Records Based on Field Values

To delete only selected rows (records) based on specific values in a given field (column), select **Delete Ingested Records.** That opens the Delete Records dialog, a form that lets you specify a table, a column, and one or more values for that column that correspond to records you want to delete.



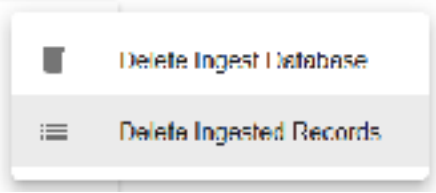
From here you select a table, enter a column name, then enter the values to match for records you want deleted. There's no limit to the number of distinct values you can enter, except for the eventual performance degradation of the delete process.

All the records that have a listed value in the specified column will be deleted from the destination database. Note that the values are a straight match. No wildcard or alternate syntaxes are supported.

If the column type is a string, simply enter the string without quotes or other modification. For boolean values, enter *true* or *false*. Notes while boolean values are not case-sensitive, string values are.

TO SELECTIVELY DELETE INGESTED RECORDS

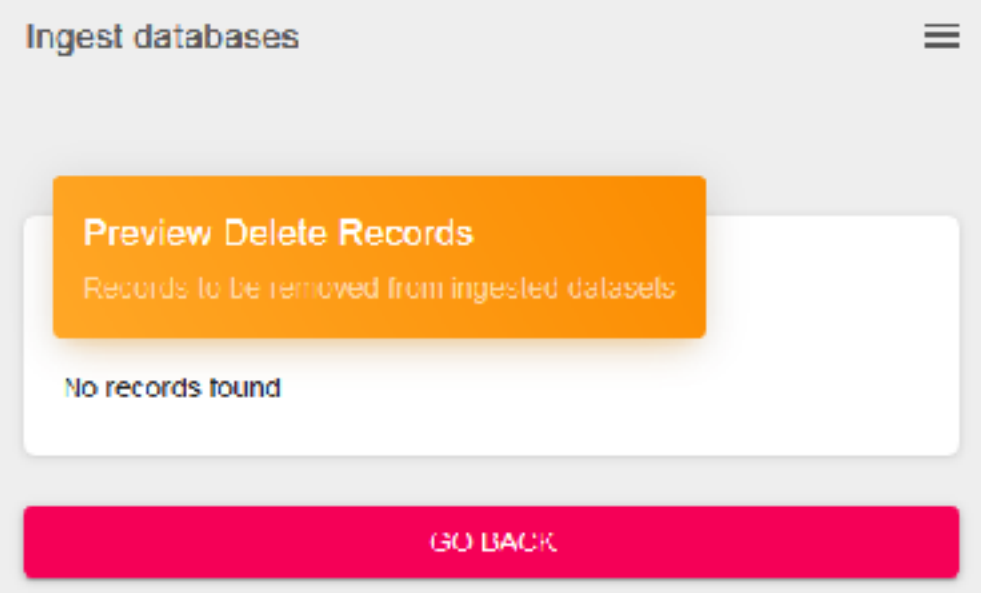
1. Click the *more options* icon ⋮ on the upper right of the profile card, then click **Delete Ingested Records.**



2. The Delete Records form opens.
3. Select the desired table from the Table Name dropdown.
4. Type a valid column name (you must know the column names in the table)
5. Type a value to search for in Column Value.
6. Click **+** to add additional values.
7. Click 🗑 to delete a value.
8. When you're satisfied with the values entered, click **DELETE RECORDS** at the bottom of the form.

Limit on the Number of Column Values
While there is no upper limit enforced on the number of distinct column values you can enter, there may be a performance penalty for loading too many column values. The more values there are, the longer the delete process will take.

If SYD doesn't find any of the values you entered in the specified column in the database, the following screen appears.



Managing Connection Profiles

Previous sections touched on connection profiles and how to create them. This section summarizes all the options available for managing profiles.

Recall that connection profiles are saved for reuse, and that all previously saved ingest database profiles are displayed in 'cards' on the main screen.

Users can add, delete, and edit profiles.

By default, when you create a new connection, all tables in the database are selected, and will be included when you run a ingest Job.

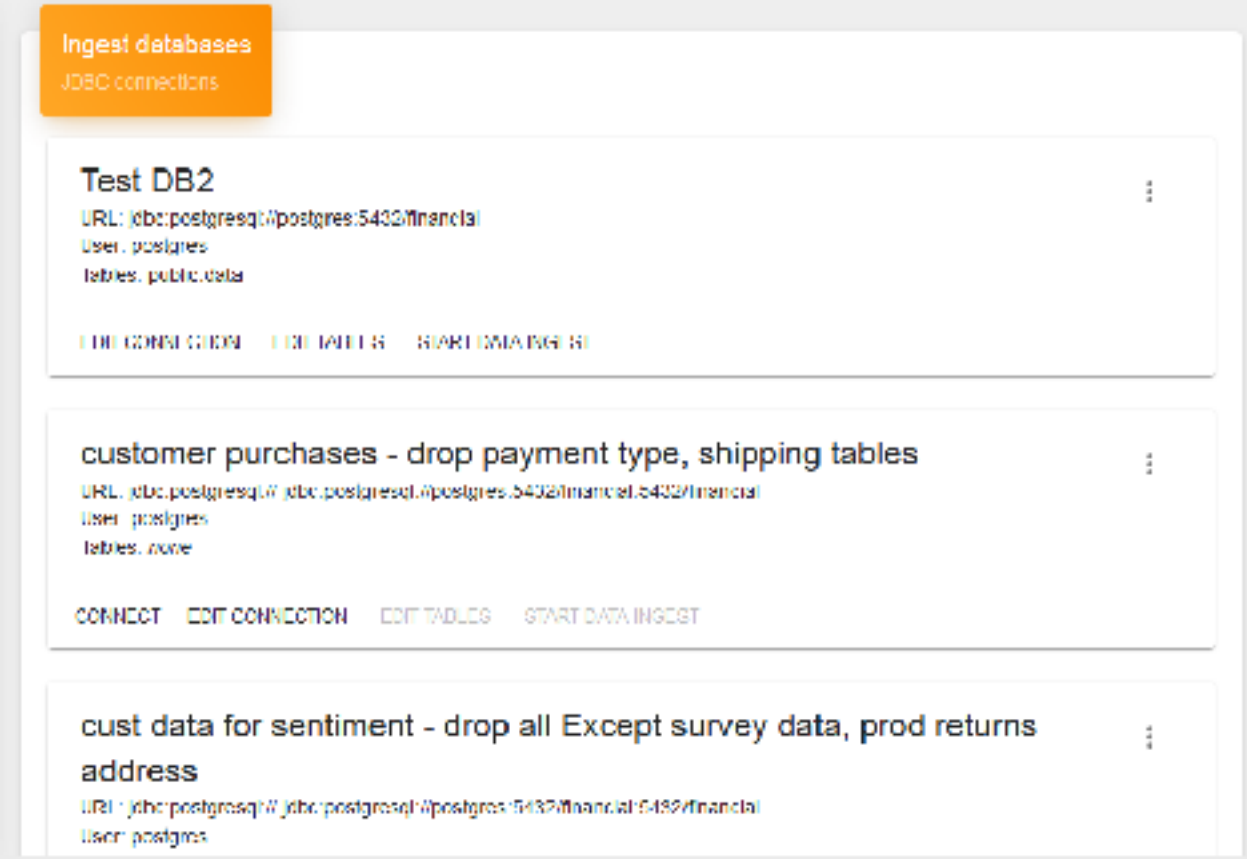
But as we just saw on the previous pages, you can select a subset of tables to participate in the ingest job by clicking **EDIT TABLES** and selecting or deselecting the desired tables.

Your selections will be saved in the profile along with the connection information.

Multiple Profiles for the Same Database

Although each profile card can reference only one database, you can have an unlimited number of cards that each reference the same database, provided each has a unique profile name.

The most useful aspect of that fact is the ability to save different sets of tables on each card, allowing you maintain distinct combinations of tables on the same database to tailor the exact tables ingested to specific requirements.



There is no hard limit imposed on the number of database connection profiles you can create and save.

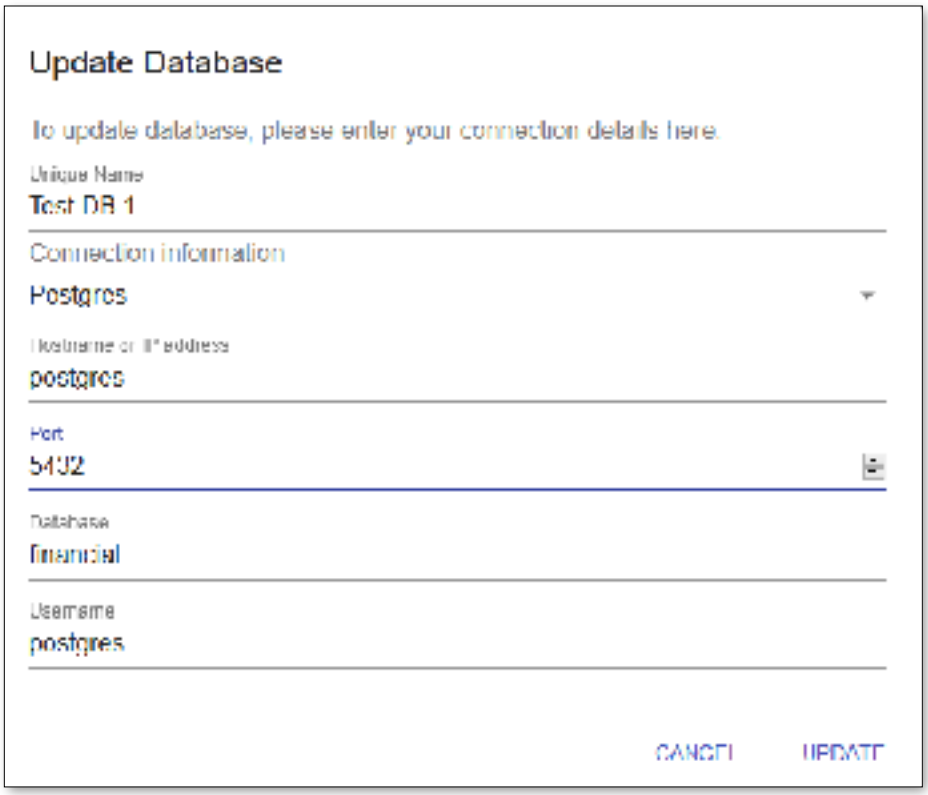
If you regularly run multiple ingest jobs using different sets of tables within the same database, you can save each variation as a unique profile. The table selection is included in each saved connection. Note that by selecting particular tables, the non-selected tables will be automatically excluded from the ingest job.

Updating Connection Profiles

When the credentials or other connection specifics change on a particular ingest database, simply update the existing profile(s) that refer to it.

To update or modify and existing connection profile, click **EDIT CONNECTION** on the profile card.

The Update Database dialog appears.



Edit the form directly by clicking on the desired fields, then be sure to click **UPDATE** to save changes when you're done.

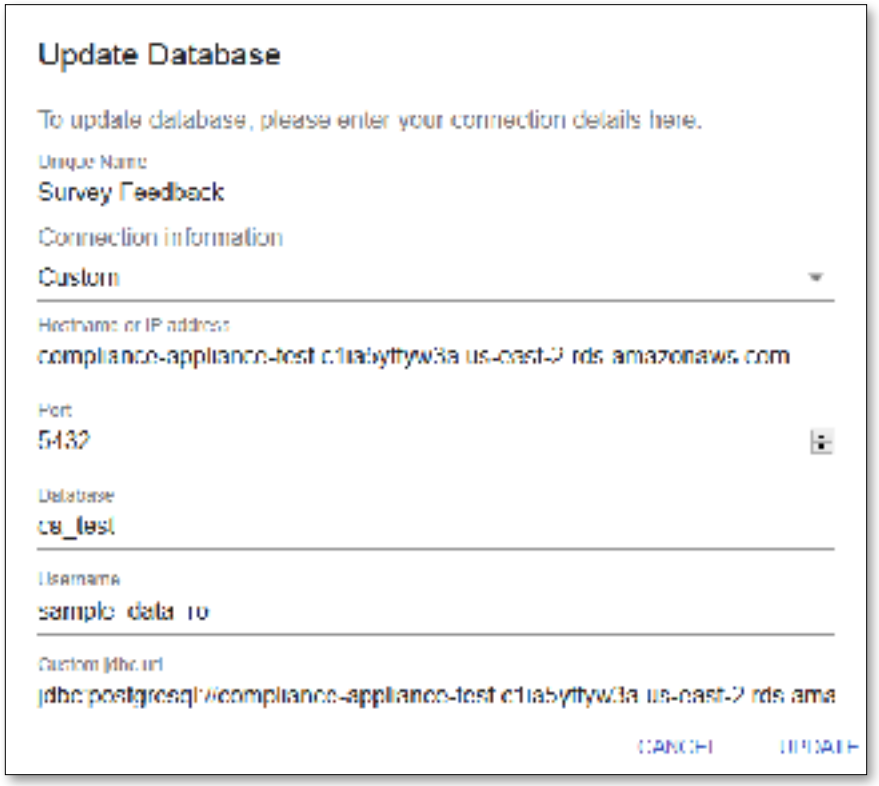
Each profile is a unique and persistent data entity, so changing the profile name and clicking **UPDATE**, for example, does not create a new profile instance.

Stranding Data by Modifying Profiles
Note that if you are storing data ingested from a particular profile, making major changes to that profile means it may no longer match the data in the destination database that it was originally used to ingest. That isn't a problem in and of itself, but it might make it difficult to later delete that data from the destination unless you can readily recreate the profile.

Custom Database Connection

Not all RDBMS platforms support the JDBC standard connection URL. MS SQL, for example, does not.

For platforms not specifically listed in the *Connection information dropdown*, select *Custom type*. Doing so adds the *Custom JDBC URL* field to the bottom of the form. From there you can edit the URL directly to conform to the format required by the relevant ingest database.

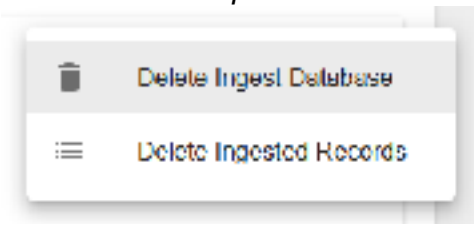


Deleting Connection Profiles and Matching Data

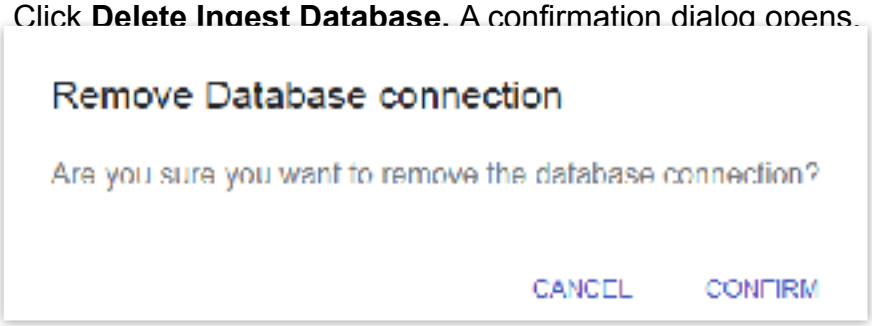
You can delete the ingest database profile from the system directly from the profile card if it's no longer in use. This has no effect on data in the customer-owned ingest database, but it does delete all of pseudonymised data in the destination database that matches this profile.

TO DELETE A CONNECTION PROFILE

1. Click the *more options* icon ⋮ on the upper right of the profile card.



2. Click **Delete Ingest Database**. A confirmation dialog opens.



3. Click **CONFIRM** to delete the profile, or **CANCEL** to return to the Ingest Data page without deleting.

Note that when the data matching the profile is deleted, the total data ingested against the 50TB limit is also reduced appropriately.

Monitoring & Auditing

TOPICS

- Overview
- Monitoring Ingest Status
- Audit Screen
- System Log Files

4

Overview

SYD provides three distinct mechanisms for monitoring the system in real time and reviewing historical events – the job status view, the Audit screen, and plain-text log files written directly to the server.

Within the SYD user interface, users can see real time status of ingest jobs in progress by clicking the INGEST STATUS link on the relevant profile card. For jobs that are completed, that same screen will show summary metrics describing the duration and total record count for the job on a table by table basis.

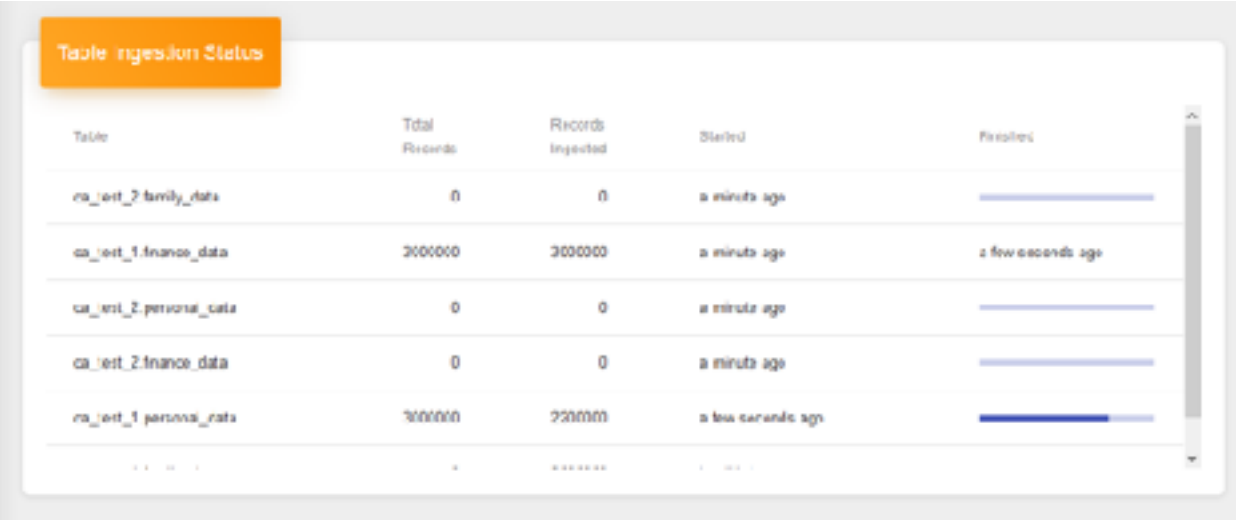
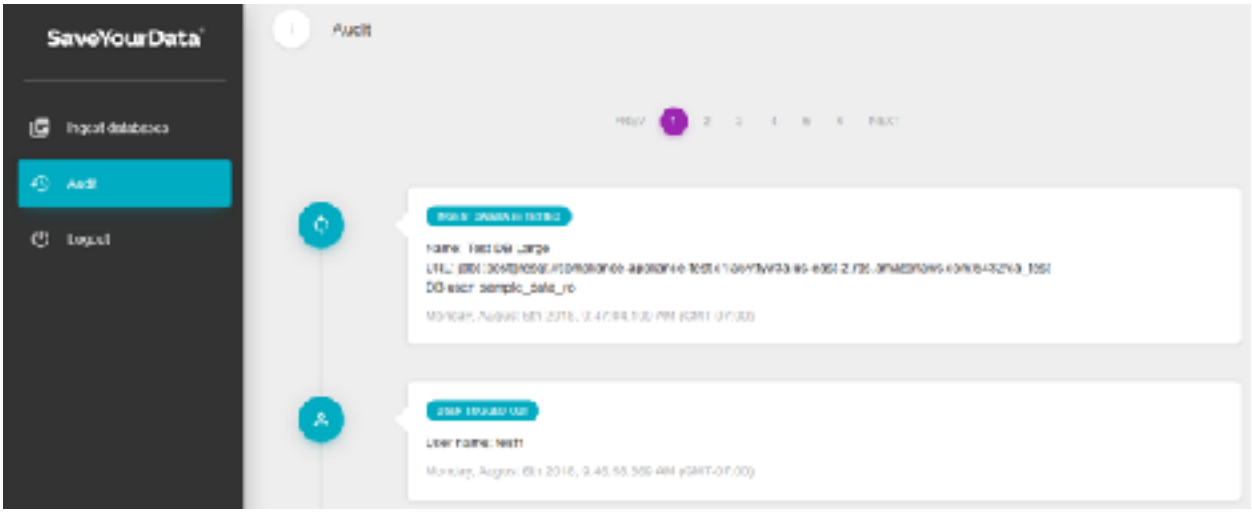


Table Ingestion Status

The Audit screen provides a more global perspective of time-stamped events related to all users and every ingest job.



Audit screen

Monitoring Ingest Status

You can check the status of any ingest job while it's running, or after a job completes, to review metrics of interest, such as the job duration and the number of records ingested.

To see the current status and key metrics for a job, click the **INGEST STATUS** link on the relevant profile card.

The Table Ingestion Status panel displays the ingest progress or disposition individually for each table in the database.

Under that, the Ingestion Summary shows total tables and records ingested, as well as the duration of the job.

Below is a sample status display for a completed job. Note that after completion the *Total Records* equals *Records Ingested*.

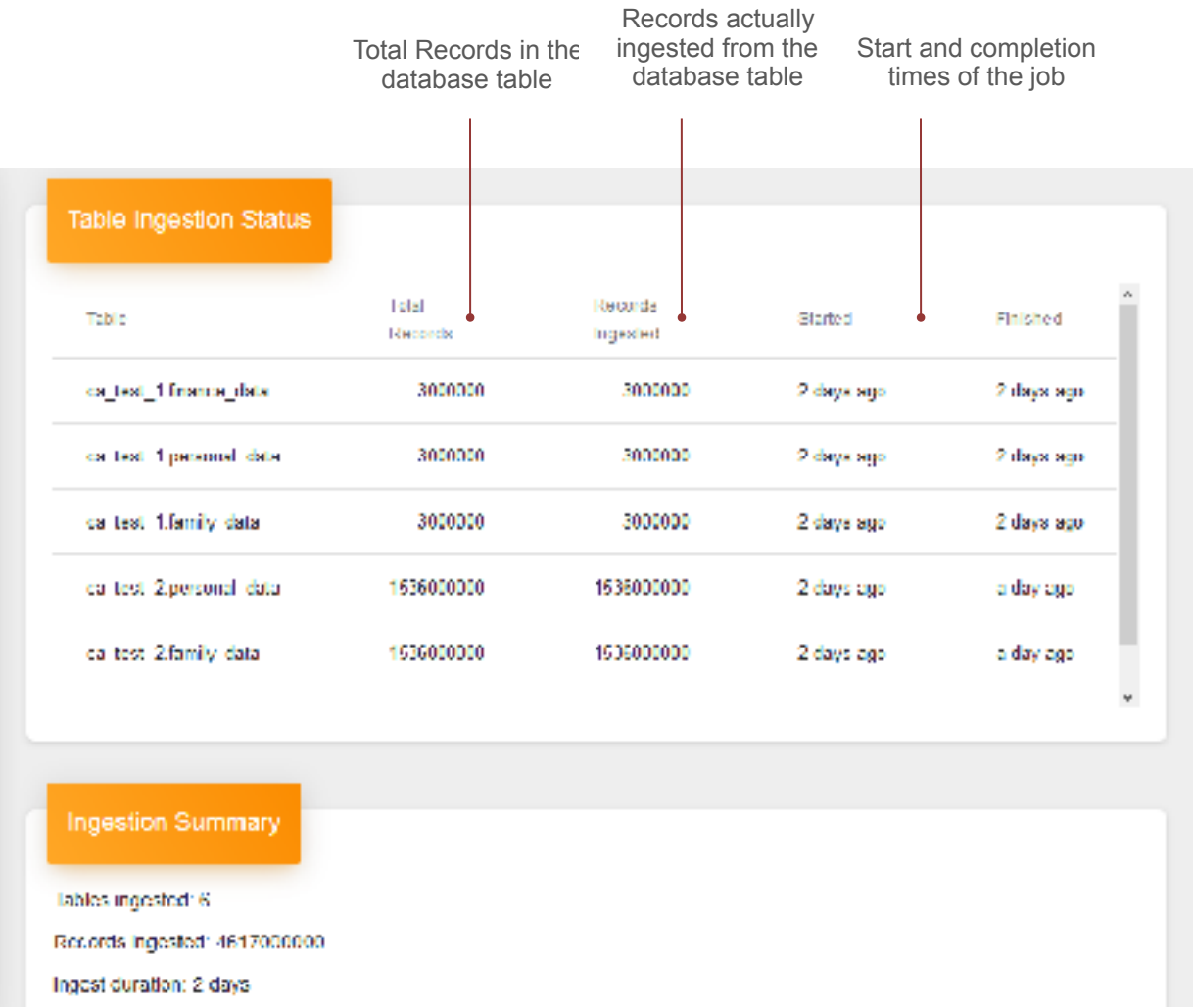


Table Ingestion Status for a Completed Job

Ingestion Summary for a Completed Job

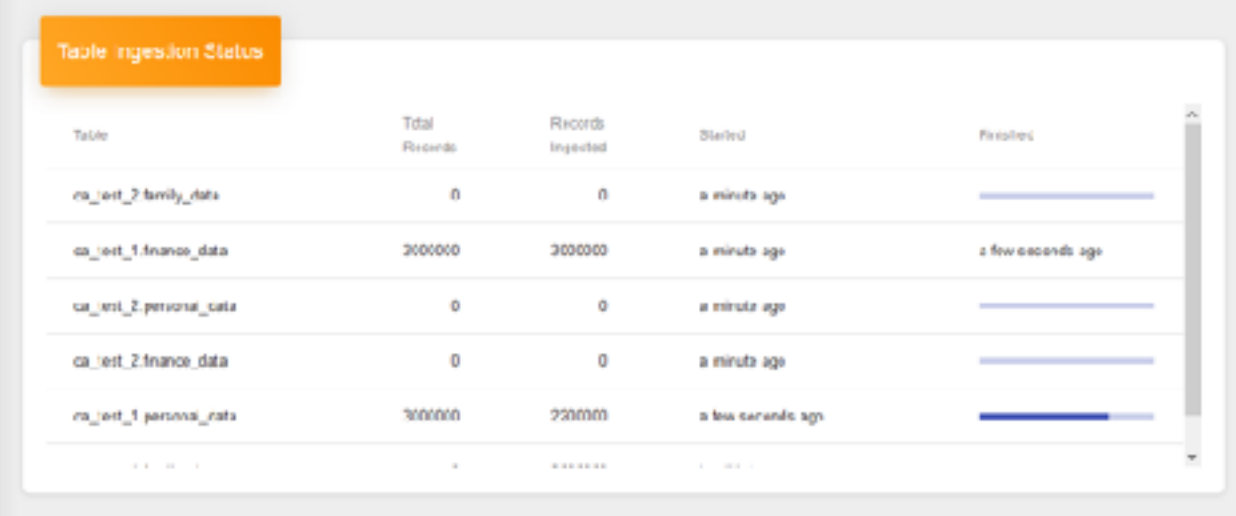
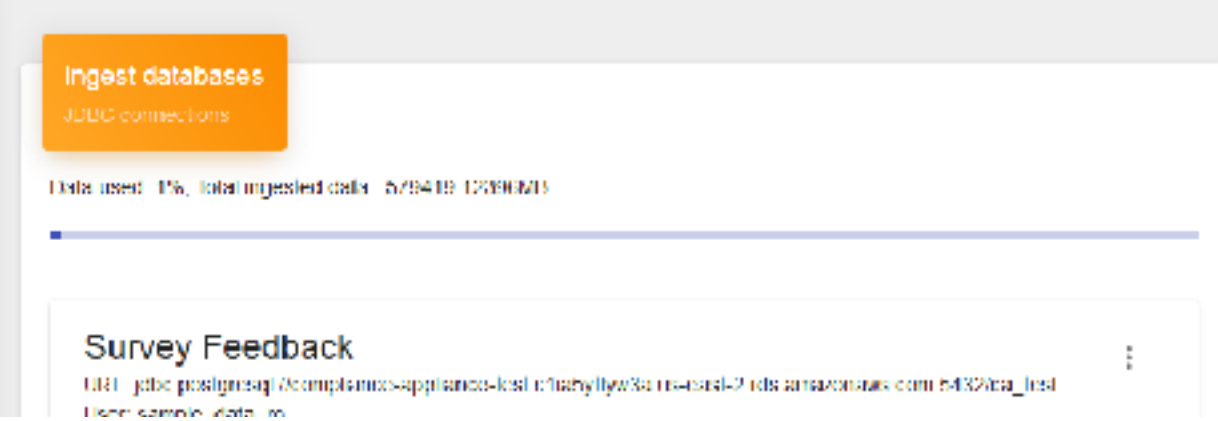


Table Ingestion Status for a Job in Progress

As mentioned in the Getting Started section, the cumulative total percentage of data ingested under the current 50TB license is shown at the top of the Ingest Databases screen.

The quantity of ingested data is also shown in megabytes (MB).

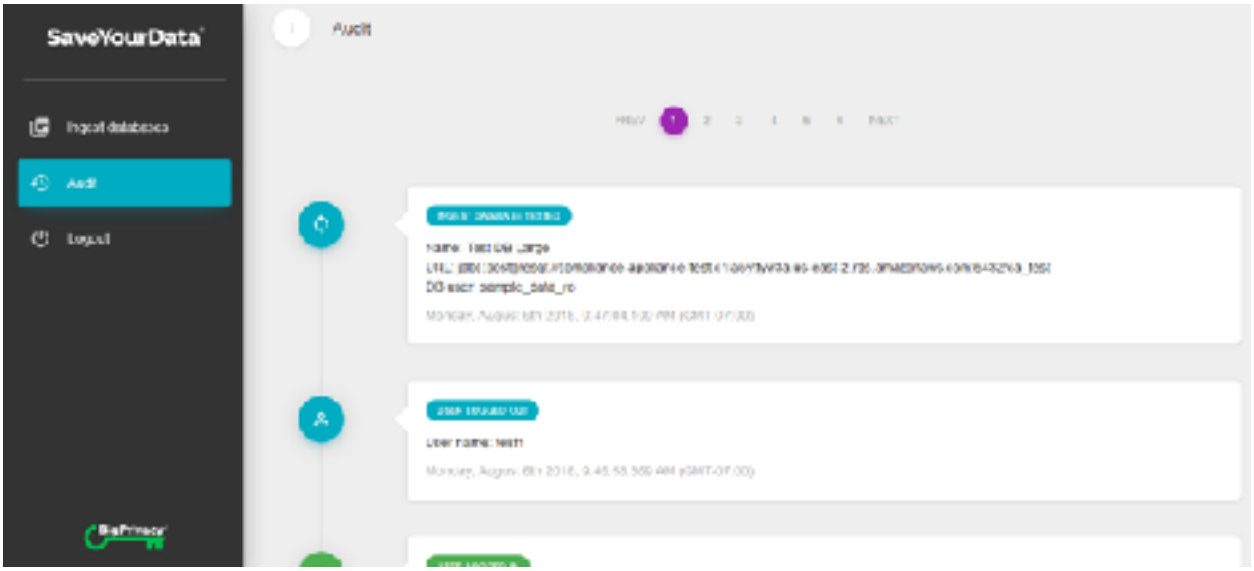


Total Ingested Data Indicator at the top of the Ingest Databases screen

Audit Screen

Another way to see system and user activity at a glance is on the Audit Screen.

Click **AUDIT** in the navigation panel to see the Audit Screen. You'll see a paginated vertical timeline of user-related events.



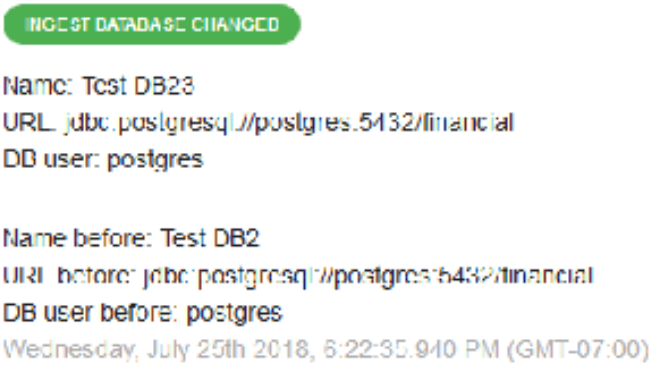
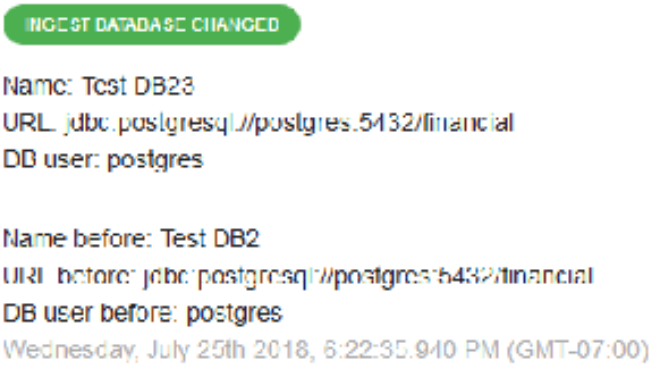
This view includes events associated with all users and connection profiles, including user logins, job start and completion events, and other similar high-level transactions.

Events Tracked on the Audit Screen

The vertical timeline on the Audit screen provides an event-based summary of user interaction and ingest job events.

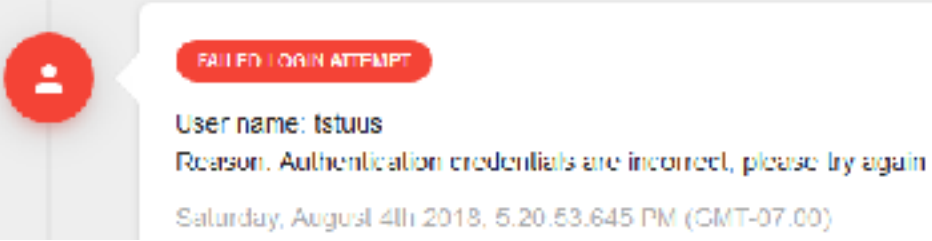
Each entry in the timeline is a discrete event or action taken by the user, along with the time and date it was recorded.

Icon	Event	Notes
	Ingest database deleted	A profile for an ingest database was deleted by a user.
	New ingest database added	The user clicked the **ADD NEW DATABASE** button on the main screen, then configured and saved a new profile.
	Ingest database tested	The user clicked **CONNECT** on the profile card, and the connection test was successful.
	Ingest job started	The user clicked **START DATA INGEST** on the profile card.
	Databases tables Edited	The user clicked **EDIT TABLES** on the profile card, and clicked **SAVE** on the Ingest Tables screen.

Ingest Database changed	Note that the settings prior to the change are also shown. 
User logged in	A valid user logged in successfully 
User Login Failed (not a member of auth group)	The user that tried to login is not part of the authorized SYD user group in the directory service.
User Login Failed (bad credentials)	The user clicked on **CONNECT** on the profile card, and the connection test was successful.

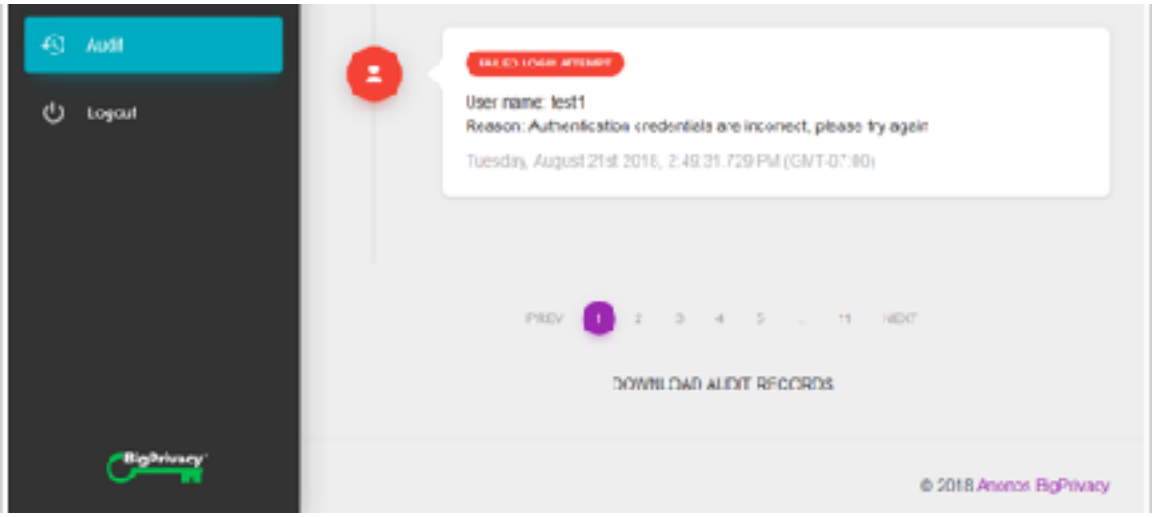
Note that the system also writes log files to the server in JSON format. Your IT team has the option to monitor or tail those logs as needed. See the *SaveYourData Installation Guide* for more information.

Both types of login errors are recorded as events in the timeline.



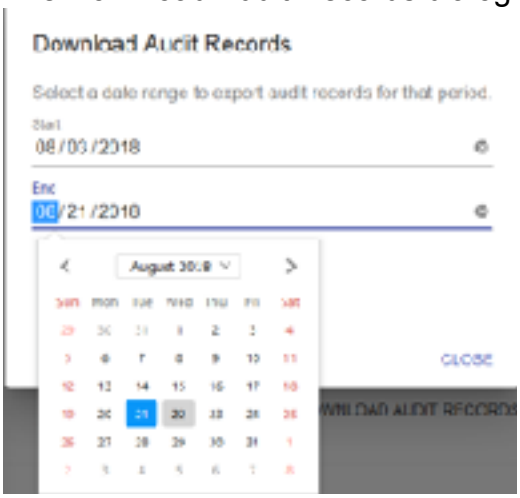
Exporting Events to a File

You can export audit events for a specified date range into a file. The file is in JSON format and it includes all events in the audit record, for all users, and all connection profiles.



TO DOWNLOAD AUDIT RECORDS

1. Click the **DOWNLOAD AUDIT RECORDS** at the bottom of the Audit screen. The calendar widget appears.
2. The Download Audit Records dialog opens.



3. Click the Start and End fields to open the date widget and set the timeframe to pull records from. Then click **Download**.
4. The file downloaded to your default download location is SYD-audit-archive-[YYYY]-[MM]-[DD].json.

System Log Files

In addition to the Audit feature, SYD writes plain-text, JSON-formatted log files to a designated folder.

While these aren't viewable from inside the application, your company has access to the relevant folder directly on the server.

You can tail or monitor the logs as your IT practices dictate.

Application log files for each Docker container are stored on the VM at,

`/var/lib/docker/containers/[CONTAINERID]/[CONTAINER_ID]-json.log.`

You can identify a Docker container by running the command ‘docker ps’ at the command line, which will list all running containers and their id.

Troubleshooting

TOPICS

- Troubleshooting SYD Application Issues
- Errors and Warnings

5

Troubleshooting SYD Application Issues

SYD relies heavily on network infrastructure for performance and proper functioning, so any fault in the chain of connected systems has the potential not only to interrupt proper functioning, but to manifest itself in a component other than that it originated in.

Because of the diverse components involved in data flow, the most effective general strategy for troubleshooting issues is "divide and conquer" – that is, to first isolate the component that is the root cause, prior to focusing on any one element. When an issue arises, the customer is encouraged to confirm basic network connectivity and functionality before engaging their support channel. Even if it becomes necessary open a support ticket, having verified some basic system connectivity up front will ultimately accelerate resolution.

The following troubleshooting issues are addressed.

- Troubleshooting Login Issue 1: User Does not Belong to Group
- Troubleshooting Login Issue 2: User Credentials not Valid
- Troubleshooting Database Connection Issues

Troubleshooting Login Issue 1: User Does not Belong to Group

SYD authenticates user login attempts against the customer's LDAP 3-compliant directory service. For access to SYD, users must have both a valid username and password, and belong to the group designated for authorized SYD users.

Symptoms

When the user attempts to log in, an error appears that user does not belong to the group required for SYD access.

Typical Root Cause(s)

The error indicates that,

- The system is connecting to the directory properly and reading user records.
- The system found the user credentials in the directory, but that user is not a member of the SYD access group.

Resolution

If ALL users get this message,

- Verify they are members of the SYD access group in the customer's directory.
    - If not, add them to that group and retry.
    - If they are... next step.

- Verify that the name of the SYD access group in the customer's directory matches the name assigned to `DIRECTORY_GROUP` in the .env file.
    - If not, add them to that group and retry.
    - Otherwise...next step.

If only some users get this message,

- Verify that the affected users are members of the SYD access group in the customer's directory.
    - If not, add them to that group and retry.
    - If they are... next step.

- Consider other restrictions in LDAP that may prevent the user from logging in as an artifact of other settings.

Troubleshooting Login Issue 2: User Credentials not Valid

SYD authenticates user login attempts against the customer's LDAP 3-compliant directory service. For access to SYD, users must have both a valid username and password, and belong to the group designated for authorized SYD users.

Symptoms

When the user attempts to log in, an error appears saying user credentials are not valid.

Typical Root Cause(s)

- The user is not configured in the directory used to authenticate SYD logins.
- The user appears to be configured in the directory used to authenticate SYD logins, but is entering the wrong username and/or password.

This error indicates that the user record is not found at all in the directory. It is unrelated to group membership.

Resolution

If ALL users get this message,

- Verify that both the username and password SYD uses to connect to the directory server is correct in the .env file. The strings assigned to `DIRECTORY_USER` and `DIRECTORY_PASSWORD` should match the correct username and password for directory access.
    - If not, correct the entry or advise the user of their proper credentials and retry.
    - If they are... next step.

If only some users get this message,

- Verify that both the username and password they enter at login matches what is in the directory.
    - If not, correct the entry or advise the user of their proper credentials and retry.
    - If they are... next step.

- Consider other restrictions in LDAP that may prevent the user from logging in as an artifact of other settings.

Troubleshooting Database Connection Issues

Symptoms

An error appears when you click **CONNECT** on a connection profile, and the profile **CONNECT** link doesn't switch to **EDIT THIS CONNECTION**.

Typical Root Cause(s)

- LAN connectivity is not available
- The connection is available, but the connection string is not valid.
- The RDBMS platform is down.

Resolving connectivity issues

1. Check the database name and credentials and try test connection again.
2. Test access to the RDBMS target outside of the SYD interface.
    a) ping the endpoint to ensure that it is reachable.
    b) try a manual connection to the database with your DBMS application.

A failure to connect can be anywhere in the integrated system.

- Retry once to verify that it wasn't just a transient error that caused the issue.
- Verify that the database URL and username are correct.
- Verify that outside of SYD, you can see the URL.
- Verify that outside of SYD, you can connect to the database on login using your database management application

- SaveYourData uses JDBC to maintain standards-based compatibility with any data platform that supports JDBC. Nevertheless, platform-specific drivers must be installed prior to use.

The application fails to connect to an ingest data source after the user clicks the connect link.

The majority of failure scenarios involve driver incompatibility or LAN/WAN infrastructure external to SaveYourData.

Errors and Warnings

The following errors may occur in the context of either connecting to the ingest database or running an ingest job.

See the notes provided for recommended actions below.

Message	Description/Resolution
Cannot connect to database	This error appears if you click CONNECT on a connection card and SYD cannot access the database. Troubleshoot **#1 Root cause**: Any mistyped parameter from the connection string to username and password could cause this issue. OR a network or database **#2 Root cause**: SYD has a job already running. **#3 Root cause**: Network or database platform fault. **1** Click EDIT CONNECTION and review the connect profile carefully for errors. **2** If no errors are found, log out of SYD, and test the data connection form your workstation to the database server for TCP/IP connectivity and database login. **3** If the tests fail without SYD in the loop, it's likely an infrastructure issue, and your IT policy will guide you to the next step in troubleshooting. **4** If the tests succeed without SYD in the loop, log back in, and try again to connect it's likely an infrastructure issue, and your IT or helpdesk policy will determine the next step in troubleshooting.

Message	Description/Resolution
There are no ingested records to delete.	This error appears if you click the traffic light menu in the top corner of the connection card, then select **Delete Records** and the system doesn't find records to delete. Troubleshoot **#1 Root cause**: There are no records in the destination database because all tables were deselected in the connection profile OR because all tables appear to be selected but the save button inside the Edit Tables screen was never clicked when the profile was created. It must be clicked when you define a new profile, WHETHER OR NOT YOU MAKE CHANGES TO THE SELECTED TABLES. This isn't necessarily an error. It's simply saying that there are no records in the destination database to delete. If you just completed a job, and there SHOULD be records in the destination database, login with your database management tool. If there are records there, go to step 2, if not, try the SYD process again to make sure it's writing to the destination database. If the problem persists, contact your support channel. Be sure to tell them when you ran the job, and whether the problem is consistent across all database servers or just particular ones.

Message	Description/Resolution
There is no ingested database to delete.	This error appears if you click the traffic light menu in the top corner of the connection card, then select **Delete Database** and the system doesn't find data to delete. Troubleshoot **#1 Root cause**: There are no records in the destination database because all tables were deselected in the connection profile OR because all tables appear to be selected but the save button inside the Edit Tables screen was never clicked when the profile was created. It must be clicked when you define a new profile, WHETHER OR NOT YOU MAKE CHANGES TO THE SELECTED TABLES. This isn't necessarily an error. It's simply saying that there are no records in the destination database to delete. If you just completed a job, and there SHOULD be records in the destination database, login with your database management tool. If there are records there, go to step 2, if not, try the SYD process again to make sure it's writing to the destination database. If the problem persists, contact your support channel. Be sure to tell them when you ran the job, and whether the problem is consistent across all database servers or just particular ones.

Conclusion

Many data uses that were lawful for decades become illegal under the GDPR without newly required data-centric controls enforced as close as possible to the data.

Contact LearnMore@BigPrivacy.com if you are interested in learning more about how BigPrivacy dynamic data-centric controls can convert global data from a liability to an asset.