User Guide

Generated on 2025-05-05

Contents

Overview

Welcome to the user guide for Analytics Platform (AP from now on). AP lets you ingest and merge, in real time, data from multiple and varied data sources together in a scalable data warehouse. The AP is designed with easy-to-use connectors (data pipelines) for platforms, systems and tools commonly used by governments and development organizations supporting them.

Publicly available and global datasets ranging from population and demographic, health, nutrition, agriculture and food security, geological, and economic data are made available through AP to enhance your programmatic data, enabling data triangulation within and across sectors to generate better insights.

Data in the warehouse are available for advanced analytics, machine learning and predictive analytics, and widespread sharing using popular third party business intelligence (BI) tools.

AP offers a user-friendly interface and seamless flow, from data ingestion to visualization, so that organizations can reduce staff time spent on curating, managing, and manipulating data, and instead focus on generating actionable insights from their data to inform programmatic decision making.

Key platform features

- **Data ingestion:** The AP offers data pipelines to systems, databases and tools commonly used in the health and international development sector.
- **Data transformation:** Data can be transformed and enriched using SQL statements upon ingestion. Furthermore, data sets can be parsed and/or joined to create unique data views for enhancing analysis.
- **Data warehousing:** Data is organized and stored in a scalable cloud-based warehouse. AP integrates with ClickHouse, PostgreSQL, SQL Server, Amazon Redshift, Azure SQL Database and Anzure Synapse.
- **Import of public data sets:** The platform offers easy import of publicly accessible data sets. A range of datasets exist within the library, including from the UN, WHO Global Health Observatory and World Bank.
- **User management:** Users and user groups can be managed, and fine-grained access control provided through a multi-dimensional security model.
- **Logging and alerts:** The platform provides logging and alerts on failures so that issues can be immediately detected and corrected.
- **Analytics and BI tool integration:** The platform supports most leading analytics and business intelligence tools, including Power BI, Tableau, and Superset, to create customized visualizations and dashboards.
- **Security:** Data is encrypted during transit and at rest in the data warehouse. AP offers firewall management for BI tool connections.

ETL vs ELT

Until recently, expensive data storage and underpowered data warehouses meant that accessing data involved building and maintaining fragile ETL (Extract, Transform, Load) pipelines that pre-aggregated and filtered data down to a consumable size. This meant you had to decide up front which data elements and fields were to be ingested. Technological advances now makes the life of data analysts easier. Practically free cloud data storage and a lot more powerful, modern, columnar cloud data warehouses make fragile ETL pipelines a relic of the past. Modern data architecture is ELT (Extract, Load, Transform): Extract and load the raw data into the destination, then transform and model it after load. ELT has many benefits, including increased versatility and usability.

Data pipelines

AP offers turn-key data pipelines to popular information systems, databases and public cloud blob stores. The data pipelines are designed to just work, meaning automatically adapting to changes in the source system, such as new data fields becoming available and changes to existing fields. The primary value is that you can define a data pipeline and forget it, allowing the platform to keep it up to date. Data will be

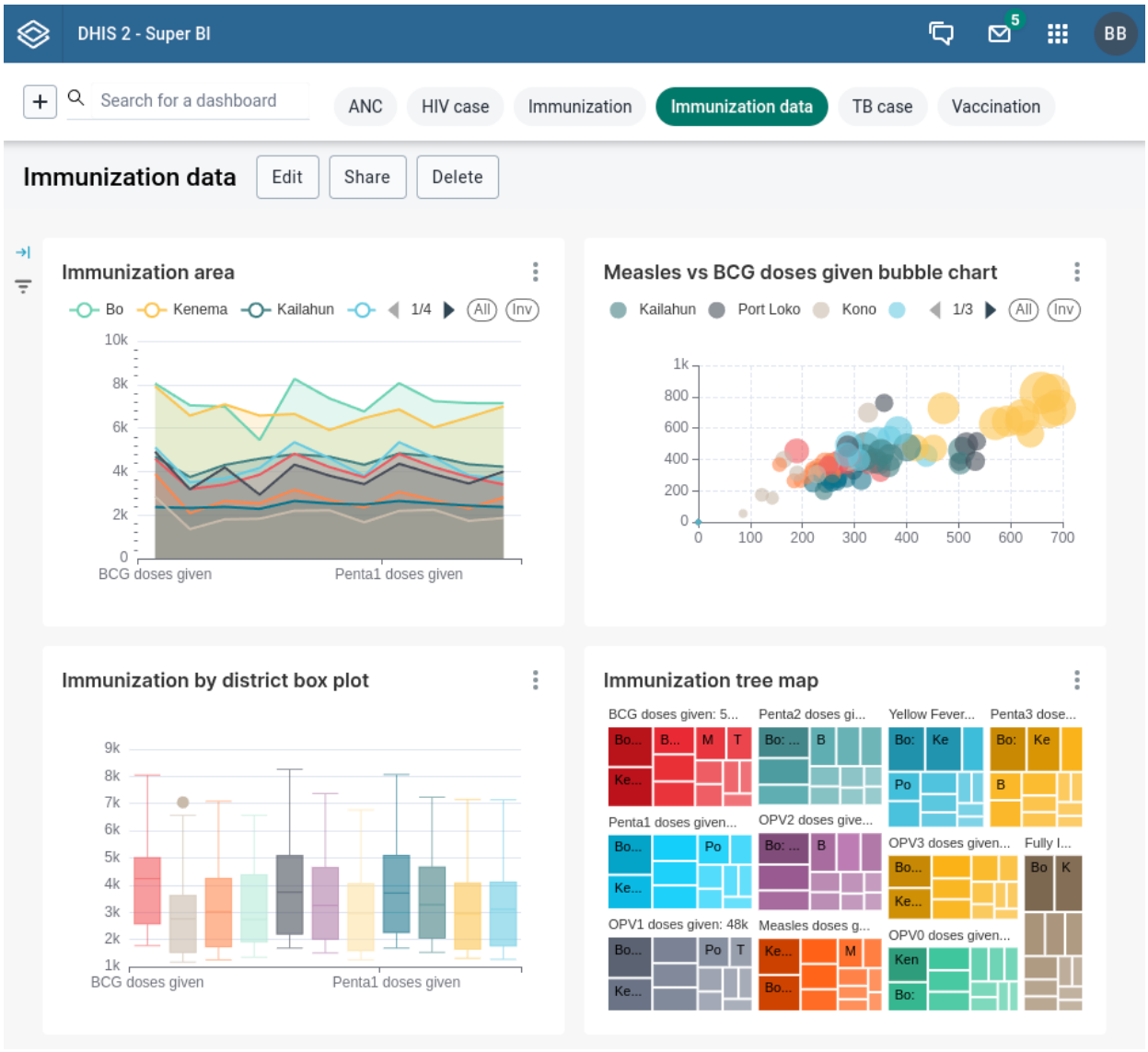

Figure 1: Superset dashboard

loaded through full refreshes or incremental updates. In the case of a data pipeline failing, e.g. because the authentication is no longer valid, AP provides alerts so that you can take timely action.

Data pipelines use a combination of API calls and database connections, depending on the nature and capability of the data source. AP offers fast data synchronization, ensuring you have data that are correct and up-to-date.

AP offers strong security. The platform encrypts secrets such as passwords and API keys before they are stored, using strong algorithms and encryption keys. Communication with data sources and data warehouses are encrypted using TLS/SSL.

Data flow

AP allows for ingesting data from a variety for data sources including systems, databases and files using data pipelines. The data is stored in the data storage area and loaded into the data warehouse of the platform. This makes the data available for analytics using a variety of tools. This includes BI tools, such as PowerBI and Tableau, data exploration tools like Apache Superset and the Super BI web app for DHIS2. Data can be aggregated and loaded back to DHIS2 using destinations. A high level diagram with a typical data flow is found below.

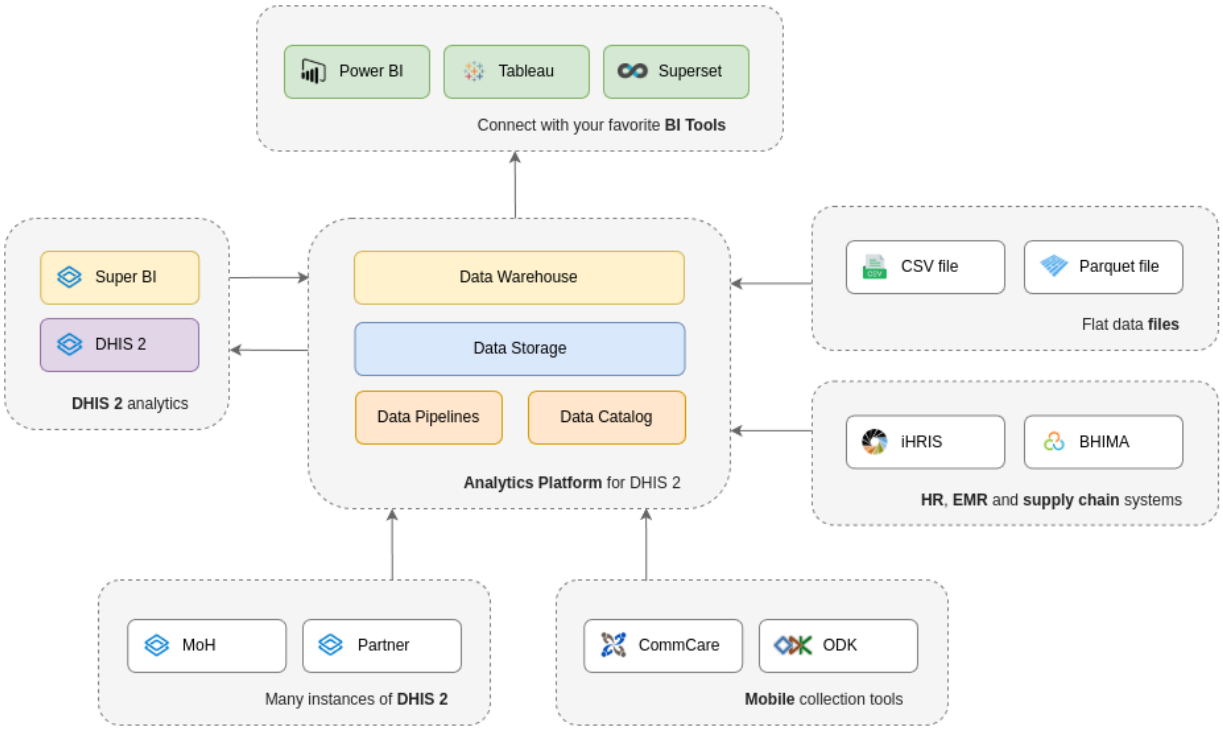

Figure 2: AP high-level data flow

Integrated data repository

AP follows the *ELT* (extract, load and transform) approach for data loading and integration. Data pipelines are responsible for retrieving data of interest from source systems and loading it into the platform. From there, data can be mapped, transformed and aggregated using views in order to produce data analytics and insights. This approach reduces the challenges of complex, fragile and slow *ETL* (extract, transform and load) jobs, where you have to decide up front which datasets and fields to ingest. The diagram below illustrates a typical integration scenario.

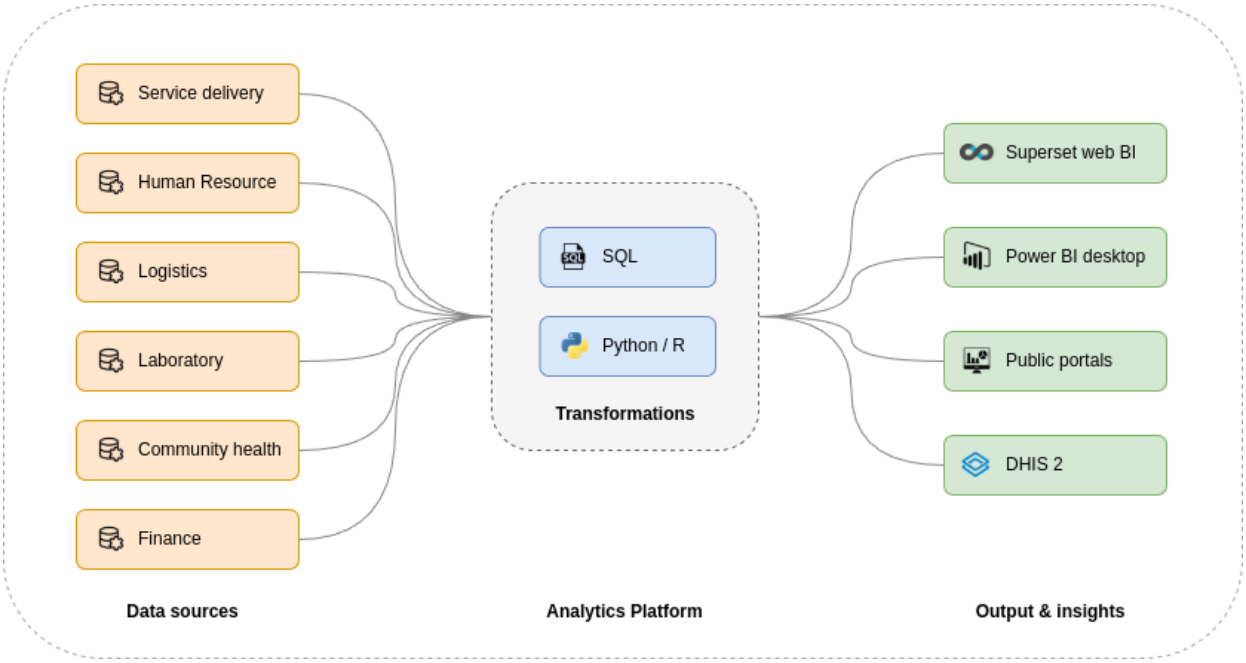

Figure 3: AP integrated data repository

Bring your own analytics tools

AP integrates and streamlines your data and makes it easy to consume from a variety of BI and data visualization tools.

- **Apache Superset:** Superset is integrated as the default data exploration tool, providing comprehensive and flexible data visualizations.
- **BI tools:** Users can easily connect both the cloud and desktop versions of popular BI tools such as Power BI and Tableau.
- **Super BI for DHIS2:** Dashboards can be embedded within DHIS2 with the Super BI web app, even without loading the data into DHIS2.

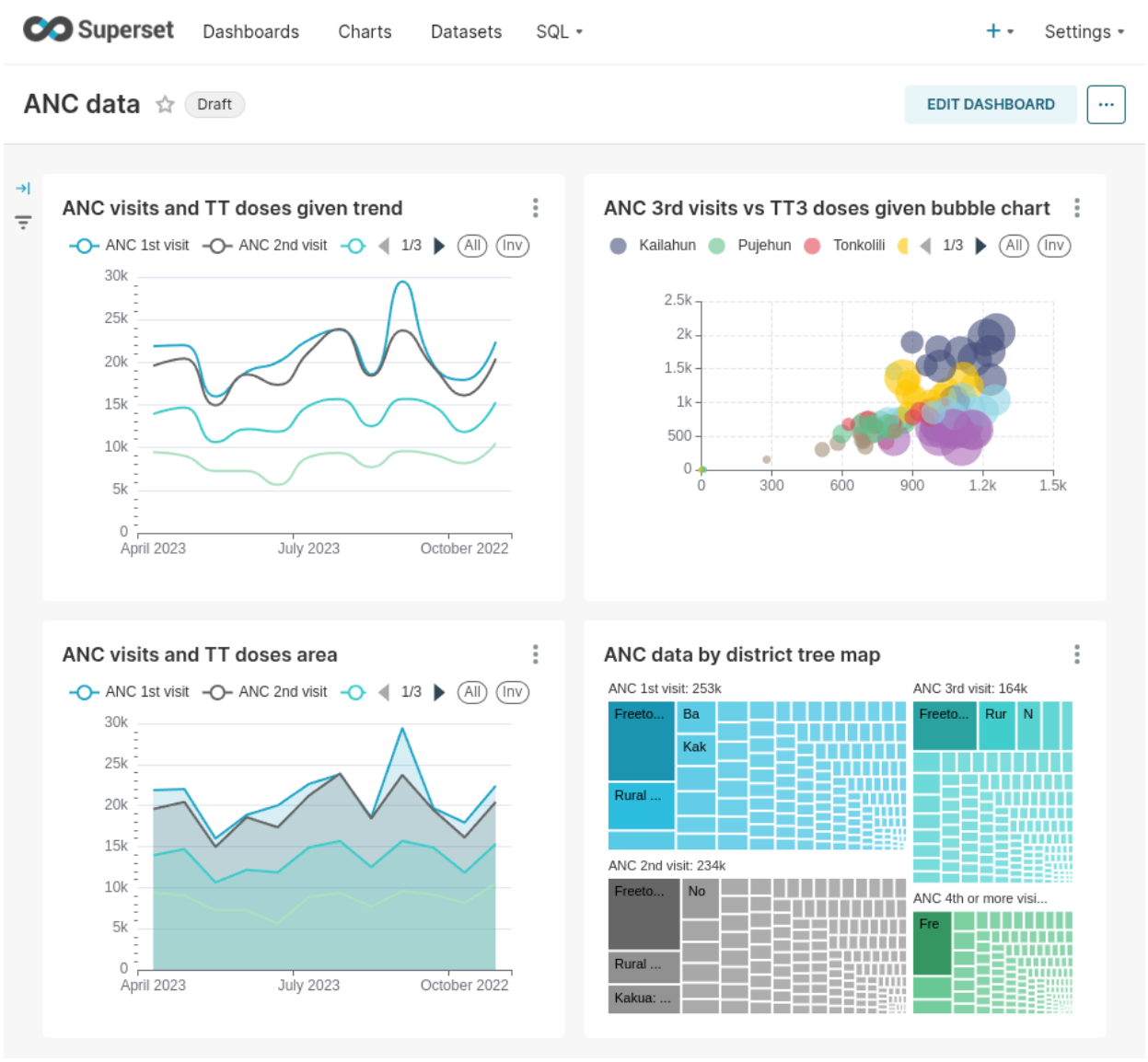

Figure 4: Superset ANC dashboard

Spaces

The AP front-end is composed of two main spaces.

- Analytics Platform
- Users

You can navigate to each space by clicking on the app menu in the header bar, followed by **Analytics Platform** or **Users**. Which spaces are visible depends on the permissions of the user.

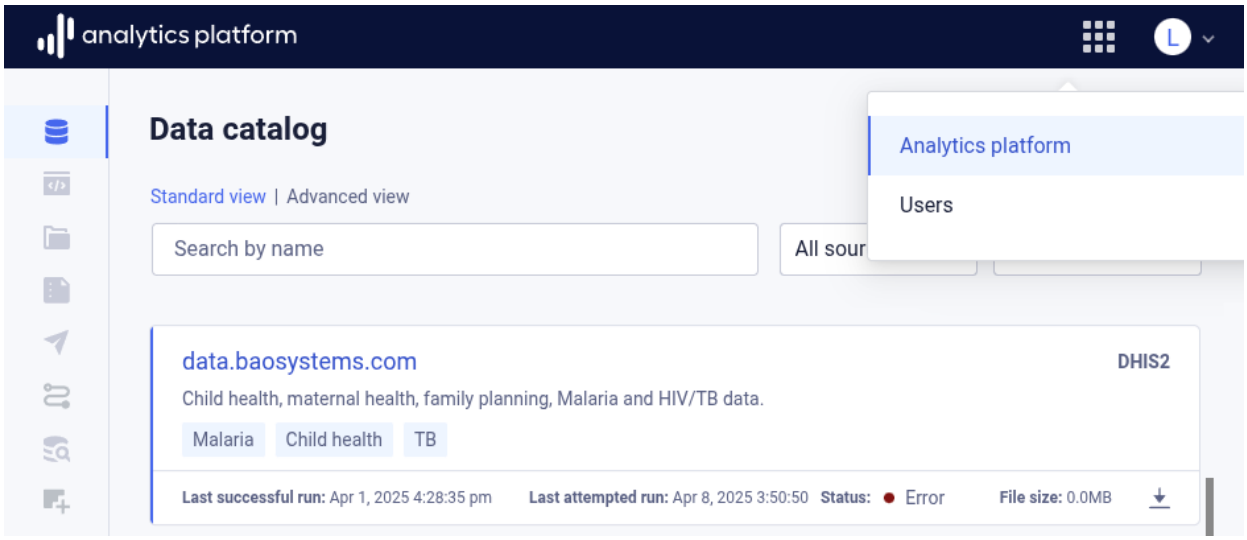

Figure 5: App menu

Data catalog

Overview

The data catalog in Analytics Platform (AP) is a comprehensive inventory system that organizes and manages information about your data assets. This system is integral for users to understand the types, sources, and characteristics of data integrated within the platform.

A data catalog serves as a central repository where all your data assets are systematically cataloged. In AP, the data catalog provides metadata, management features, and search capabilities, enabling users to quickly locate and understand data across various sources. It details the data origin, format, and the relationships between different datasets, making it easier to navigate and manage large volumes of information within the organization.

The primary utility of the data catalog lies in its ability to provide a central inventory of datasets and sources, which simplifies data governance and enhances the efficiency of data management practices. It

ensures that users have access to reliable and up-to-date data descriptions, fostering better decision-making and streamlining data utilization across projects. By centralizing data knowledge, the data catalog reduces redundancy and improves data quality.

The terms dataset, data source and data pipelines are used somewhat interchangeably in this guide.

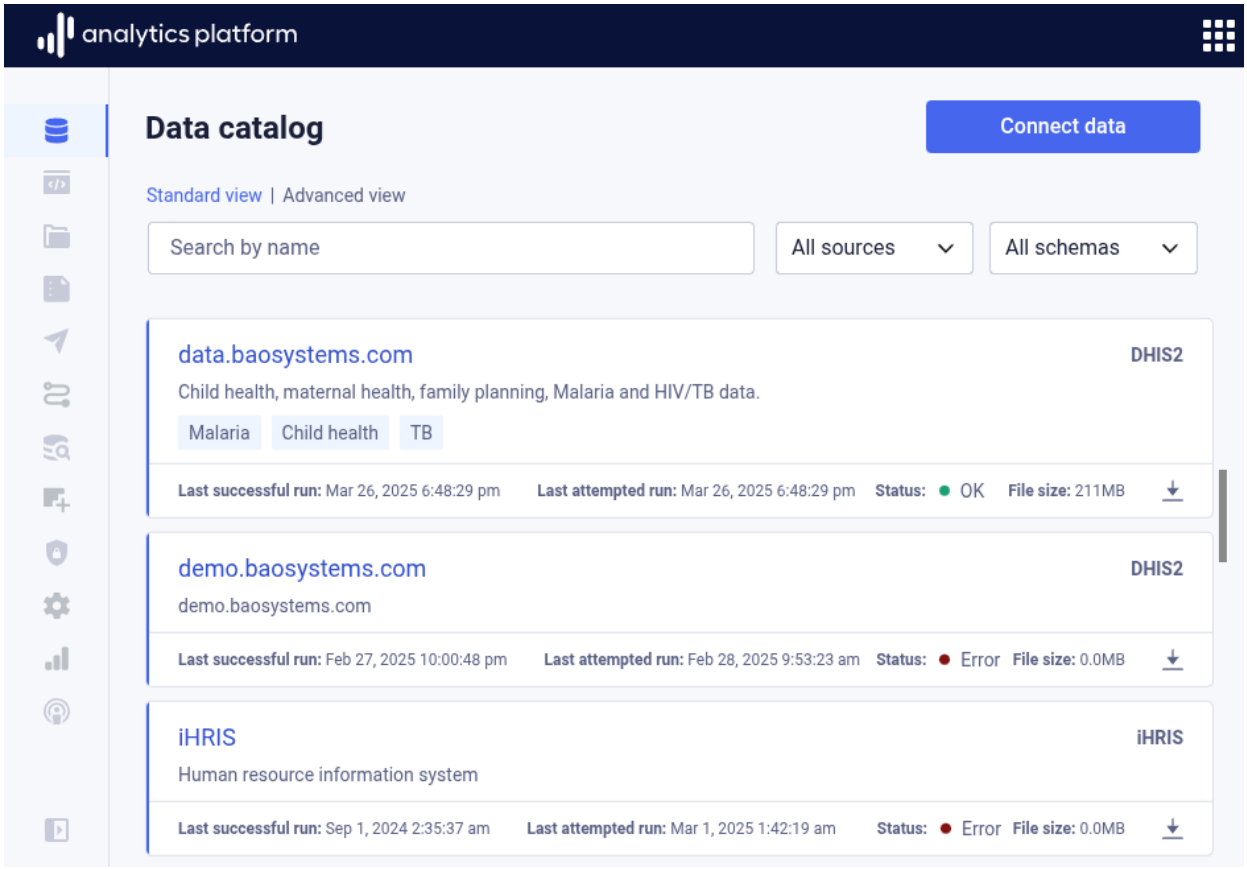

Figure 6: Data catalog

Audience

The data catalog is designed for use by various stakeholders within an organization. Data engineers and data integration specialists benefit from it by gaining insights into available data sources and how they can be best utilized and integrated. Analysts and data scientists use the catalog to find relevant datasets for their analytical work, ensuring that they are working with the most appropriate and up-to-date data. Additionally, business users and decision-makers rely on the catalog to verify that the data they base their strategic decisions on is accurate and comprehensive.

View data catalog

- Click **Data catalog** from the left-side menu to open the data catalog and view datasets.
- Use the source *All sources* drop-down at the top of the page to filter datasets by source type.
- Use the schema *All schemas* drop-down at the top of the page to filter datasets by schema.
- Click the name of the dataset to view more information.

Connect data

To connect data sources and bring datasets into the data catalog, click **Connect data** from the top-right corner. This will open the data pipeline dialog. Consult the *Data pipelines* page to learn more about connecting data sources.

View dataset

The dataset overview screen provides comprehensive information about the dataset.

Details

The *Details* tab displays metadata such as owner and URL, and the username of the user who created and last modified the dataset.

Tables

The *Tables* tab shows a list of all tables for the data source. Data pipelines can generate one or many tables. As an example, a DHIS2 data pipeline will typically generate a large number of tables, including metadata, data, enrollment and event tables. The tables list allows you to get an overview of which tables exist. Clicking on table will display the data structure, meaning the list of columns for the table.

Data structure

The *Data structure* tab shows the structure of the table as a list of columns for the selected table. For each column, the data type, number of distinct values, null (blank) values, min and max value are displayed. Min and max value only apply to numeric data fields.

Data preview

The *Data preview* tab displays the first 50 rows of the table. This is useful to get an overview over what type of data exists in the table.

Change log

The *Change log* tab displays an overview of data load tasks for the data pipeline. For each task, the start time, data load strategy, duration and status are displayed. The data load strategy can be *Full replace* or *Incremental append*. The *Incremental append* strategy is only relevant for data pipelines for which data is

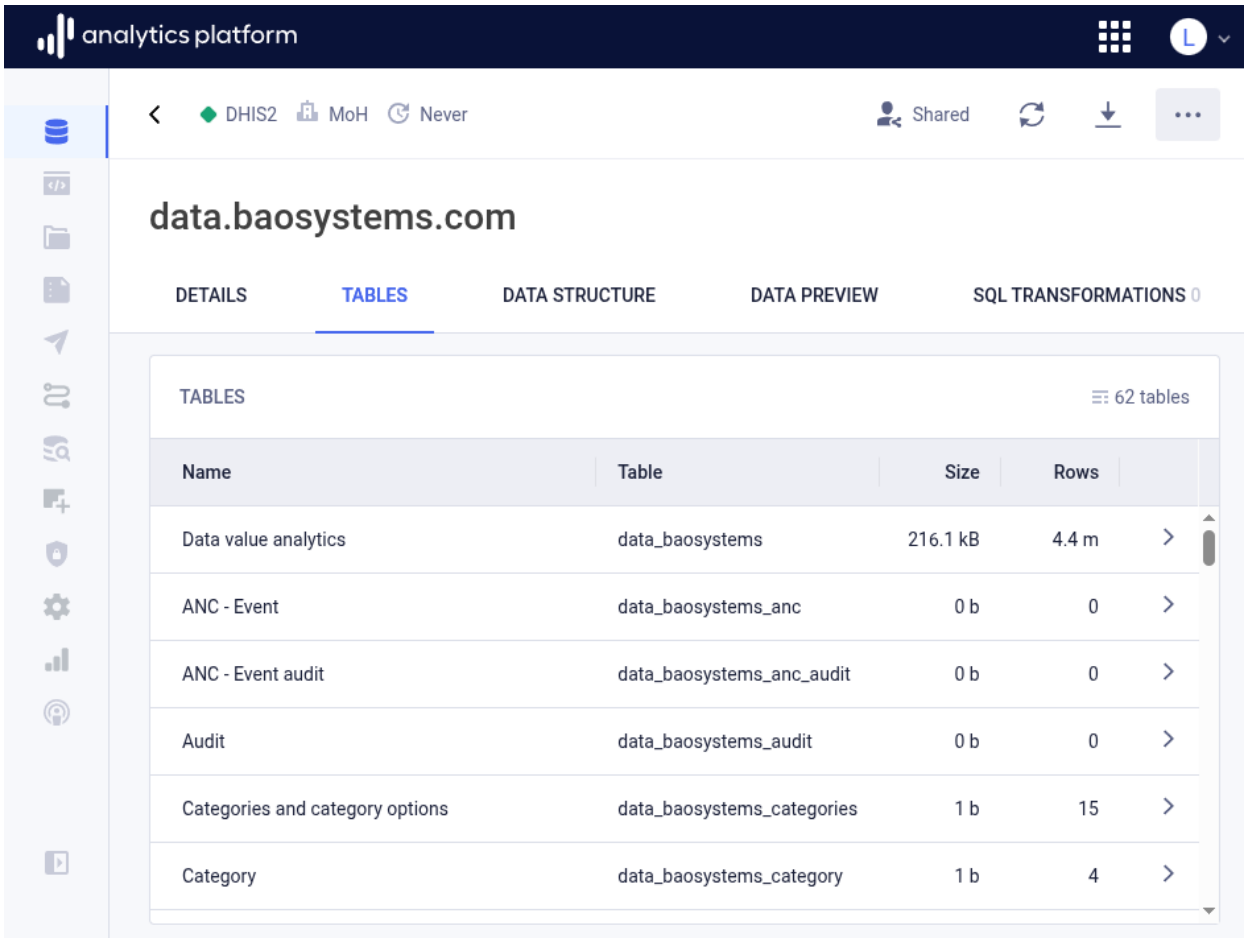

Figure 7: Dataset tables

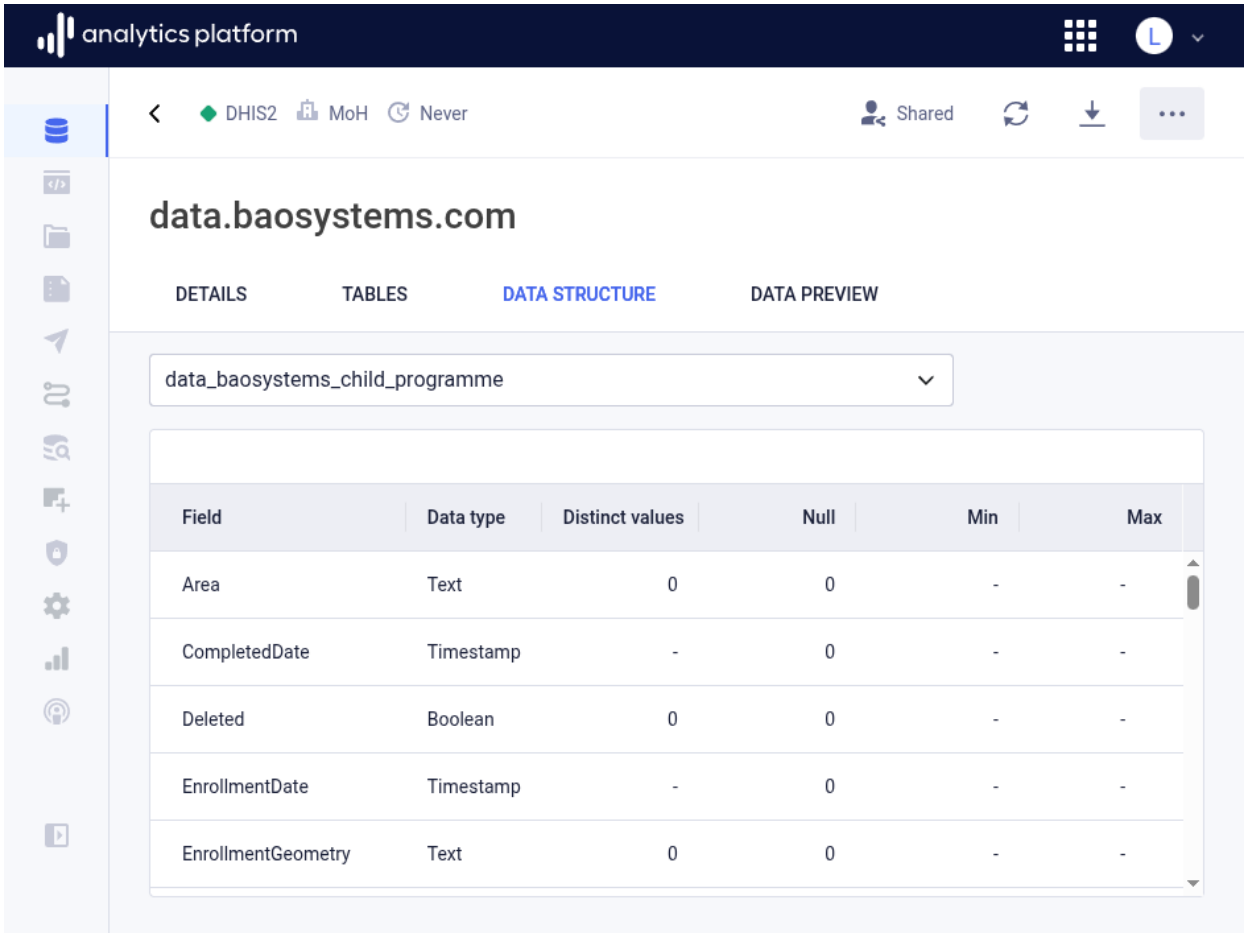

Figure 8: Dataset data structure

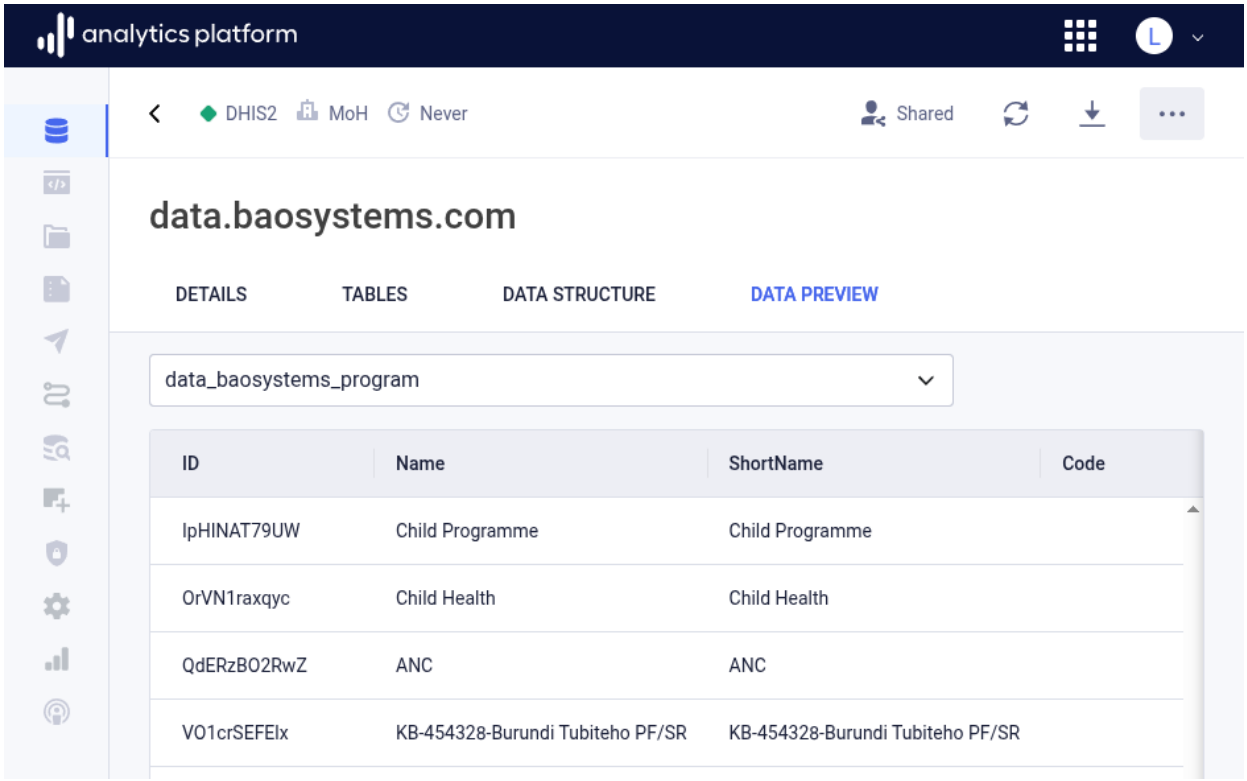

Figure 9: Dataset data preview

continuously updated. The status will be *In progress* when a task is running, then *Successful* when a task completed successfully, and *Failed* when a task failed with an error.

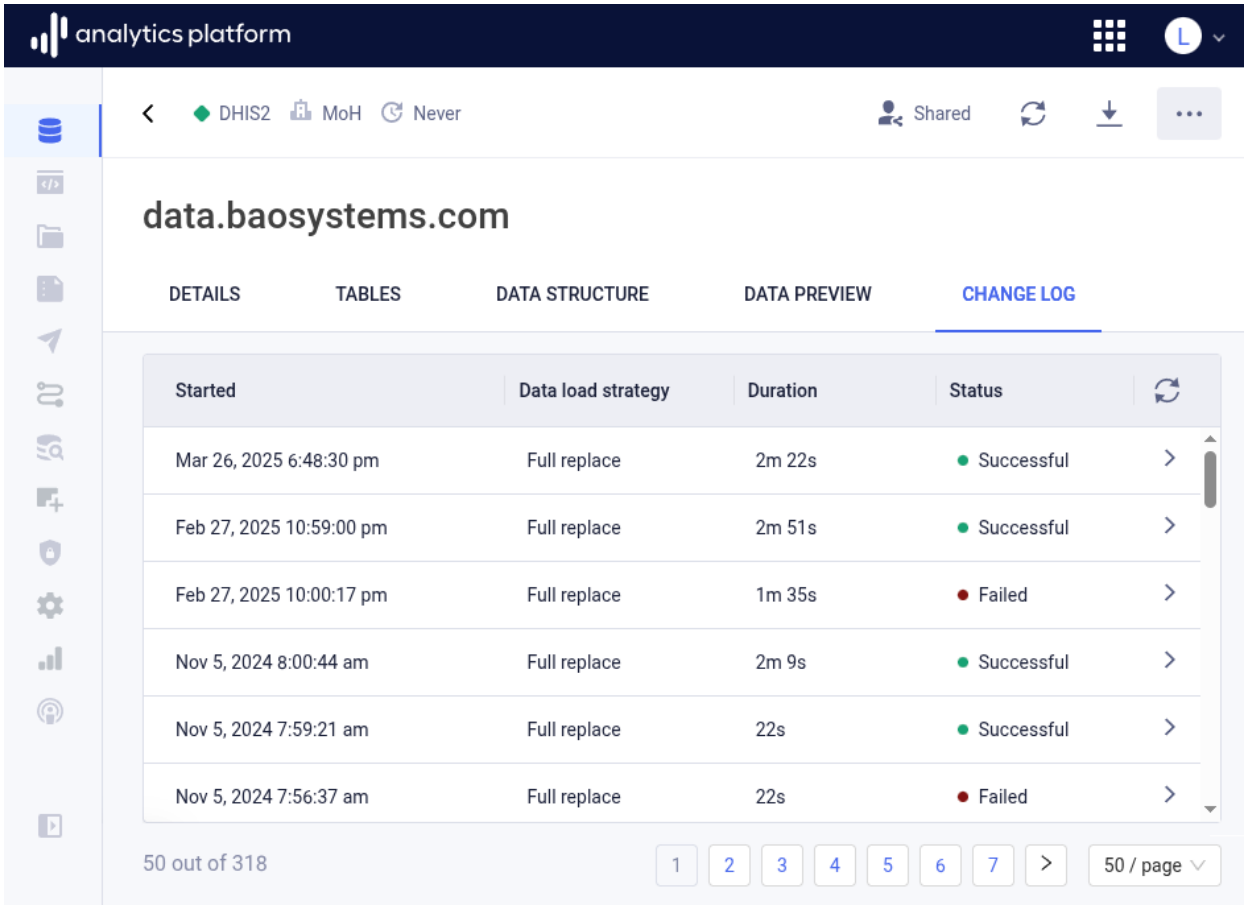

Figure 10: Dataset change log

Task log

You can click on a task row to view logs for the task. The logs provides detailed information about the data load process, and includes the tables which were created and loaded, the count of data records, the runtime for each table data load, and more.

Edit dataset

A dataset can be edited after it has been created.

1. Open the context menu by clicking the icon in the top-right corner.
2. Click **Edit dataset**.

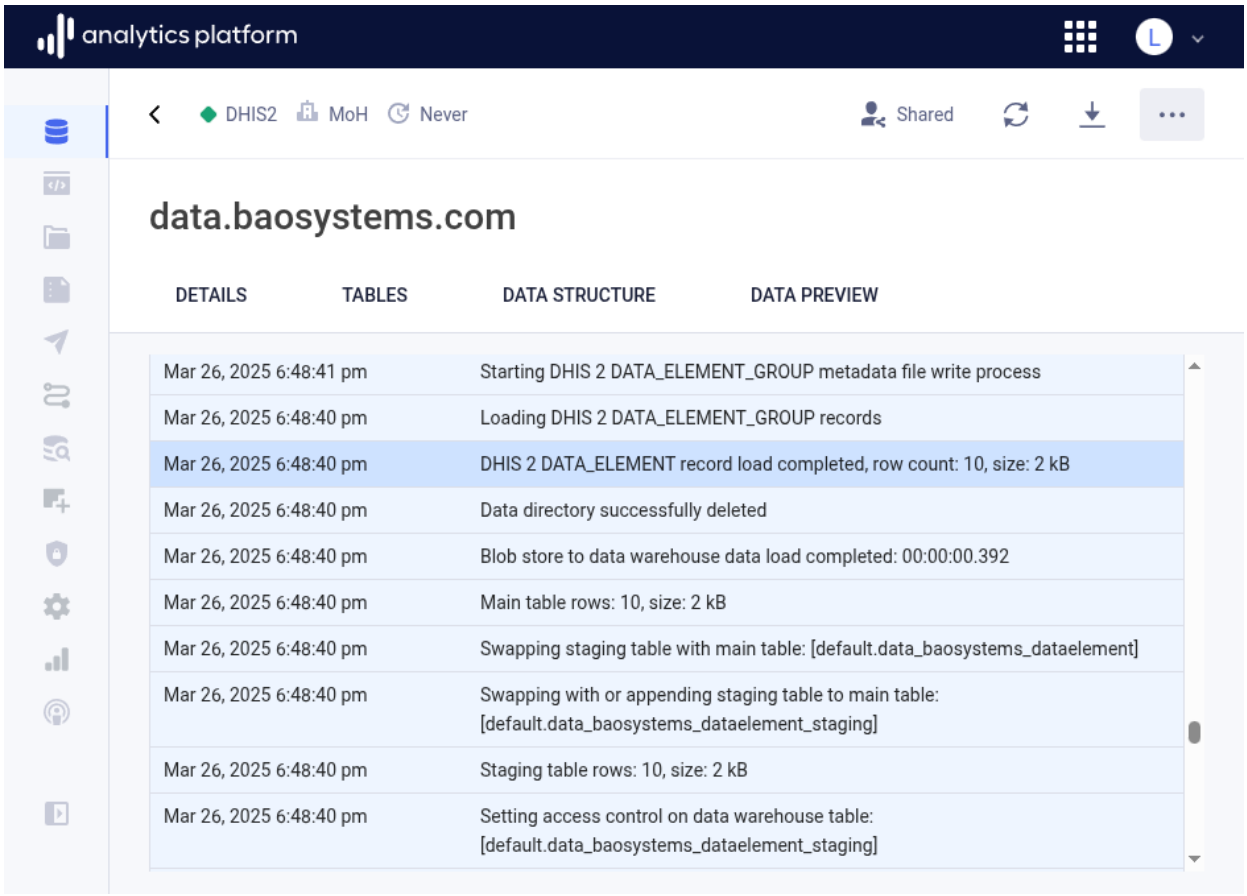

Figure 11: Dataset change log

3. Edit values in the relevant sections.
4. Click **Save** at the bottom of the section.
5. Close the dialog by clicking the close icon in the top-left corner.

Share

Access to a dataset can be controlled by setting the appropriate sharing permissions. A dataset can be shared with everyone in the organization, referred to as *public access*, with user groups and with users. Users can be given view access or edit access. Edit access implies view access. Refer to the sharing page for sharing and access control documentation.

Download dataset

The data files used to load data to the data warehouse is available for download in CSV format.

1. Open the context menu by clicking the icon in the top-right corner.
2. Click **Download data** in the context menu. This will open a dialog that displays the available data files for download.
3. Click the download icon next to a file to download it.
4. Click the link icon next to a file to copy the link / URL to the file.

Note that the downloadable data files are in compressed CSV format. The files are compressed with the gzip tool. Tools for decompression exist for all operating systems. For MS Windows, 7-Zip is a free alternative. For Mac and Linux, use a terminal with the `gunzip` command, e.g. `gunzip data.csv.gz`.

Download metadata

Metadata for the dataset is available for download in JSON format.

1. Open the context menu by clicking the icon in the top-right corner.
2. Click **Download metadata**.

Refresh data

Data for the data pipeline can be manually refreshed. This will load data from the data source into the platform and data warehouse. Note that data pipelines will typically be scheduled to refresh automatically. This can be set in the create and update data pipeline screens.

1. Open the context menu by clicking the icon in the top-right corner.
2. Click **Refresh data**.

Test connection

After setting up a data pipeline, it is useful to be able to test that the connection is valid.

1. Open the context menu by clicking the icon in the top-right corner.

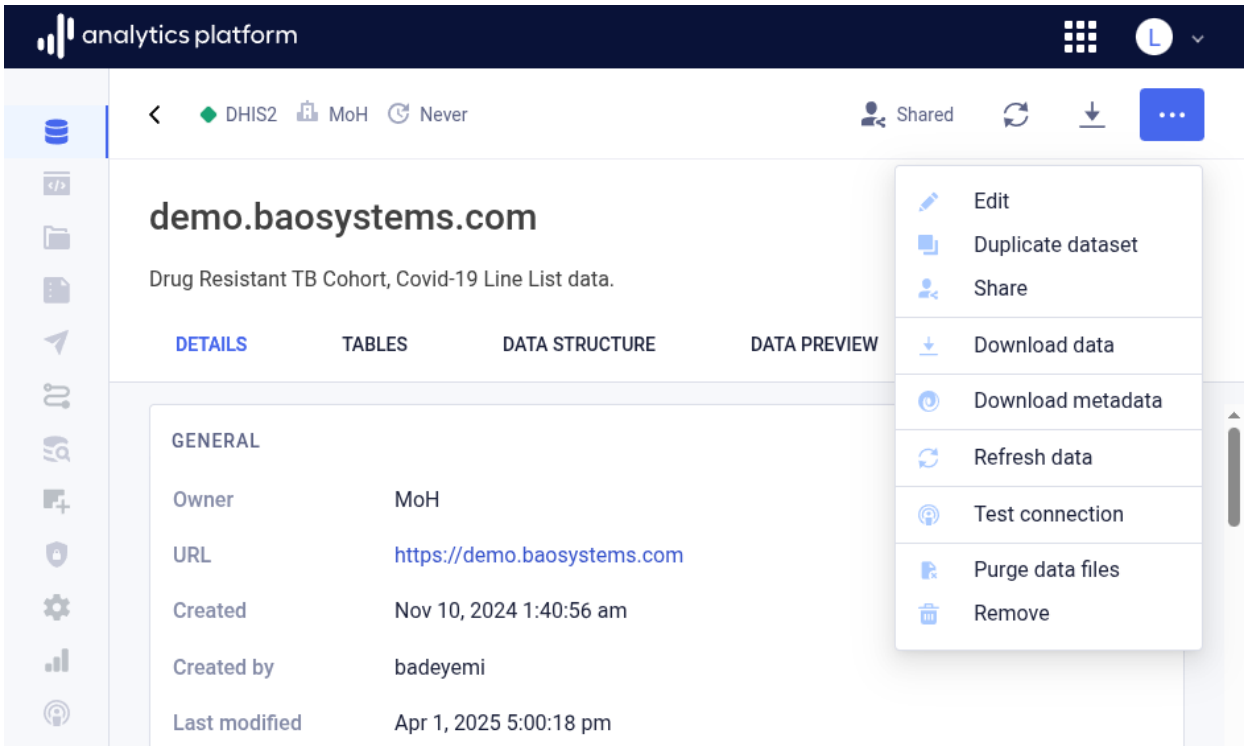

Figure 12: Dataset context menu

2. Click **Test connection**.

Purge data files

Every time data for a data pipeline is loaded into the platform, the data files used to stage and load data are retained. If underlying data has changed, for data protection and compliance reasons, it may be necessary to purge the data files for each data load process in order to have a fresh start.

1. Open the context menu by clicking the icon in the top-right corner.
2. Click **Purge data files**.

Remove data pipeline

A data pipeline, including data files and data warehouse tables, can be removed when no longer needed.

1. Open the context menu by clicking the icon in the top-right corner.
2. Click **Remove**.

Data pipelines

Overview

Analytics Platform (AP) offers *data pipelines* for ingesting data from a variety of data sources into the platform. A data pipeline is a mechanism for moving data from data sources into the AP. Within the AP, the data is ingested into a data catalog, data store and data warehouse. A data source can be an application, a blob (file) store and a data file. Data pipelines are *pull* based, meaning data will be pulled from the data source before being loaded in the platform.

The following data pipelines are supported.

Applications

- **BHIMA** An open-source hospital and logistics information management system used for electronic medical records (EMR), inventory and commodity tracking and billing in low-resource settings.
- **CommCare:** A mobile-based platform for managing front-line health programs, providing case management, data collection, and real-time decision support for community health workers.
- **DHIS2:** A flexible, an open-source, web-based platform for collecting, analyzing and visualizing health data, widely used for managing and monitoring health programs, particularly in low-resource settings.
- **FHIR:** Fast Healthcare Interoperability Resources (FHIR) is a healthcare data standard and framework with an API for representing and exchanging electronic health records (EHR).
- **Google Sheets:** A cloud-based spreadsheet application within Google Drive that enables users to create, edit, and format spreadsheets online while collaborating in real time with others.
- **iHRIS:** An open-source human resources information system designed for managing health workforce data, allowing organizations to track employee information.

- **Kobo Toolbox:** An open-source data collection and management tool that enables field researchers and humanitarian organizations to design surveys and collect data offline or online using mobile devices.
- **ODK:** A suite of open-source tools used for mobile data collection, especially in challenging environments, with features like offline data capture and GPS integration.
- **Ona:** An open-source data collection and analysis platform designed for mobile and web-based surveys, commonly used for monitoring and evaluating projects in health, agriculture, and development sectors.
- **Talamus:** A hybrid cloud-based software for patients and health-care providers to facilitate high quality health-care, providing hospitals, labs, pharmacies and imaging centers with a digital platform.

Blob stores

- **Amazon S3:** A scalable and durable object storage service in the AWS cloud.
- **Azure Blob Storage:** A scalable and durable object storage in the Microsoft Azure cloud.

Databases

- **SQL Server:** A relational database management system with robust data management capabilities developed by Microsoft.
- **MySQL:** An open-source and flexible relational database management system widely used for web applications and large datasets.
- **Oracle RDBMS:** A powerful, enterprise-grade relational database management system designed for complex transactional processing.
- **PostgreSQL:** A sophisticated, widely adopted and open-source relational database management system with a rich ecosystem of extensions.
- **Amazon Redshift:** A managed and scalable data warehouse designed for fast querying of large-scale datasets in the AWS cloud.

File formats

- **CSV file upload:** A simple, text-based file format used to store tabular data, where each line represents a data record and each field is separated by a comma.
- **Parquet file upload:** A columnar storage file format designed for high-performance data querying and storage optimization in analytics workloads.

Schedule data refresh

Data pipelines can be scheduled to automatically refresh data from the data source. The refresh schedule is set in the add and update data pipeline pages. There are two types of scheduling.

- **Regular interval:** Next to the **Refresh schedule** label, the preferred interval for when to refresh the data source can be selected from the drop-down field. Data will be regularly refreshed from the data source at the selected interval. Data in AP will be *fully replaced* for every refresh, meaning, existing data will be entirely replaced with data from the source system.

- **Continuously updated**: Next to the **Refresh schedule** label, the **Continuously updated** checkbox can be selected. Data will be continuously refreshed from the data source. Data in AP will be *incrementally appended* for every refresh with a one minute delay in between. In addition, data in AP will be *fully replaced* once per night. For the incremental appends, new, updated and removed data records in the source system are first loaded in a staging table in AP, then corresponding data records are removed, then finally the new and updated data records are loaded in AP. Continuous update scheduling is currently only supported for the DHIS2 data pipeline.

Authentication

Data pipelines need to authenticate to the data source for secure exchange of data records. Data pipelines typically use the following types of authentication.

- **API token:** This method uses a token that is sent in the HTTP header of requests to authenticate with APIs. The token acts as a secure key to access the API, ensuring that only authorized users or services can interact with the data source. It is often used for RESTful APIs and provides a straightforward way to handle authentication without exposing user credentials.
- **API username/password:** A simple authentication scheme built into the HTTP protocol, also known as *basic auth*. It involves sending a username and password with each HTTP request. These credentials are typically *base64* encoded for transmission but are easily decoded, making this method less secure unless used in conjunction with TLS encryption to protect the credentials in transit.
- **Database username/password:** This method involves connecting to a database using a username and password. The credentials are used to establish a JDBC or similar database connection, ensuring that only authenticated users can execute queries and access data. This traditional form of database access control is widely used due to its simplicity and direct support in most database systems.

General metadata

All data pipeline types have a common section titled **General settings** with various metadata fields. This section allows you to store extensive metadata for each data pipeline (dataset). Maintaining descriptive metadata for each dataset allows for a comprehensive data catalog, allowing users to get an overview of which datasets exists for your organization. The metadata fields include name, description, owner, URL, tags, reference, link to source and link to terms of use. Tags are entered as free text. Tags which already exist will be suggested as you type in a tag name.

Data source connections

AP pulls data from your data sources using a set of fixed IP addresses. To ensure that AP can connect to your data sources, you must allow list these IP addresses in your firewall. This typically only applies to databases, where a connection is made directly with the database. It typically does not apply to HTTP API-based applications which are already accessible on the Internet. It also only applies for the managed AP offering from BAO Systems, not for other deployment models.

AP environment	Region	IP address
Production	US East	3.93.131.28
Test	US East	54.173.36.156

Manage data pipelines

The following section covers how to view, create, update and remove data pipelines.

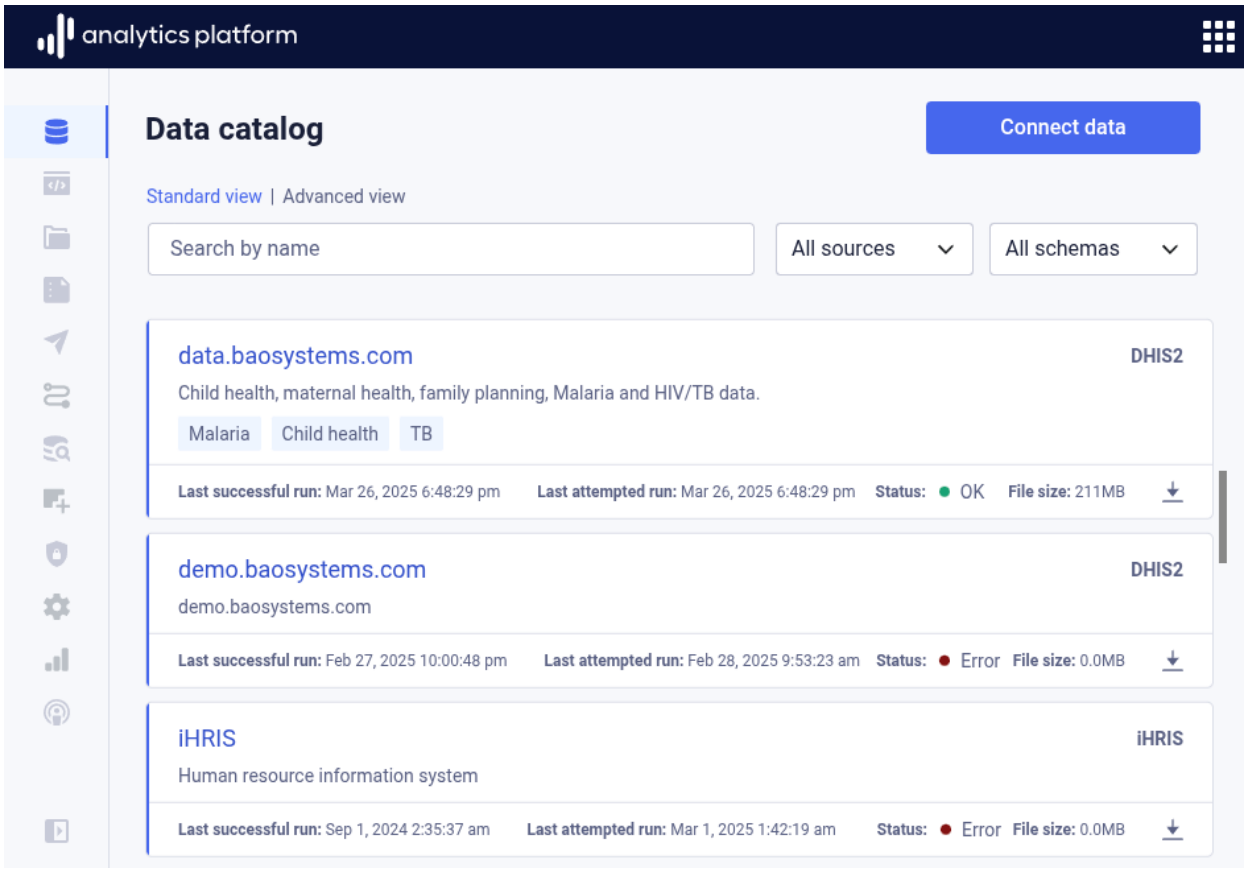

Figure 13: Data catalog

View data pipeline

1. Click **Data catalog** in the left-side menu to view all data pipelines.
2. Click the name of a data pipeline to view more information.

Create data pipeline

The starting point for creating a new data pipeline is the data catalog. The data catalog displays the existing data pipelines, also referred to as datasets.

1. Click **Connect data** from the top-right corner.
2. Choose the data source for which you want to create a data pipeline.

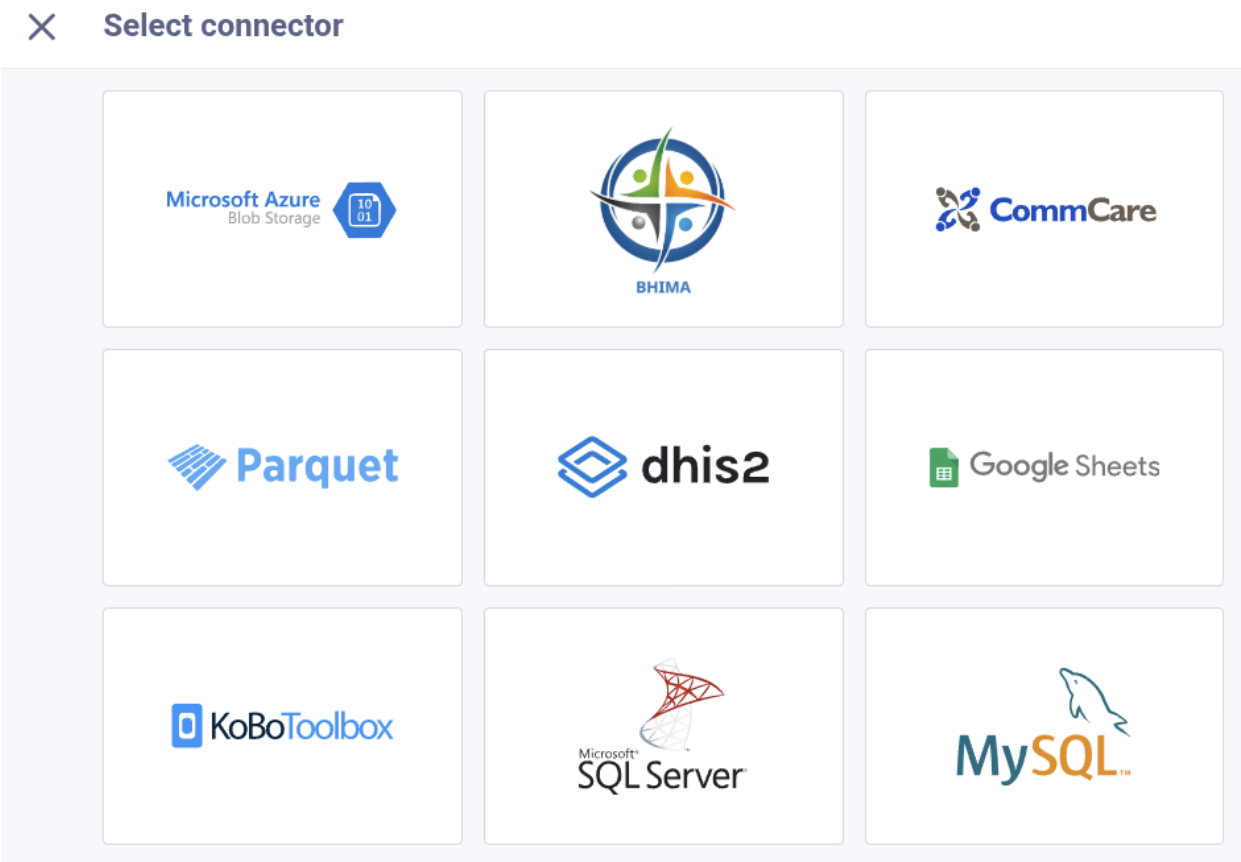

Figure 14: Data pipeline types

General settings

In the *General settings* section, enter the following information. This section is present for all data pipeline types.

Field	Description
Name	The name of the data pipeline (required)
Refresh schedule	The interval for when to refresh data from the data source (required)
Description	A description of the data pipeline
Owner	The owner of the source data or system
URL	A URL to the source data or system
Tags	Free text tags which categorizes the source data
Disable pipeline	Whether to disable loading of source data
Reference	A reference text for the data source
Link to source	A URL refering to information about the data source
Link to terms of use	A URL refering to terms of use for the data source

The following section describes steps for creating each type of data pipeline.

Data warehouse target

In the *Data warehouse* enter the following information. This section appears last of all sections, and is present for all data pipeline types.

Field	Description
Table shema	The data warehouse schema in which to create tables
Table name	The table name, for multi-table data pipelins, the base name for all tables

BHIMA

Topic	Value
Connection	Web API
Authentication	API username/password
Data model	Project

Connection

Field	Description
URL	The URL for the BHIMA instance
Username	Username for the BHIMA account
Password	Password for the BHIMA account
Project	Project for which to exchange data

Settings

Stock usage

Field	Description
Depot	The depot to load data for
Inventory	The inventory to load data for
Avg consumption algorithm	The algorithm to use for average consumption calculation
Monthly interval	The monthly interval

Stock satisfaction rate

Field	Description
Start date	Start date for satisfaction rates
End date	End date for satisfication rates
Depots	Depots

CommCare

Topic	Value
Connection	Web API
Authentication	API username/token

Topic	Value
Data model	Application and form

Settings

Field	Description
Domain	The domain (project) to load data for
Application	The application to load data for
Hash column names	Whether to hash column names to ensure uniqueness

DHIS2

Topic	Value
Connection	Web API and database connection
Authentication	API token, API username/password, database username/password
Data model	Aggregate data, program, event and enrollment

Web API

Field	Description
Base URL to web API	Base URL to web API for DHIS2 instance, do not include `/api`
Username	Username for DHIS2 user account
Password	Password for DHIS2 user account

Database

Providing a database connection URL and credentials will drastically improve performance, and is required to load enrollment and event data. If a database connection cannot be provided, the database section can be skipped, and the data pipeline will work with an API connection only. For the API connection, only metadata, data set completness and aggregate data are supported.

Field	Description
Hostname	The hostname to the PostgreSQL DHIS2 database server, do not include a protocol prefix
Port	Port number to the database server, default is 5432
SSL	Whether to enable SSL encryption for the database connection
Trust server certificate	Whether to trust the server SSL certificate for the database connection
Database name	The name of the database
Database username	The username of the database user
Database password	The password of the database user

Data types

The data types section provides selections for the data types to load.

Field	Description
Aggregate data	Include aggregate data values and complete data set registrations
Program	Include events and enrollments, use the drop-down to specify which programs to include, or leave the drop-down blank to include all current and future programs

Field	Description
Metadata	Include metadata

Data filters

The data filters section provides filters for the data to load. All filters are optional.

Field	Description
Data element groups	The data element groups to include
Data elements	The data elements to include
Organisation units	The organisation units to include
Data sets	The data sets to include
Period last time unit	The last time periods of the specified unit to load
Include soft deleted data	Whether to include soft deleted data records
Skip wide aggregate data table	Whether to skip the wide aggregate data table
Include zero data values	Whether to include zero data values
Include narrow event table for programs	Whether to include narrow event table for programs

FHIR

The FHIR data pipeline allows for retrieving information, typically related to electronic health records.

Learn more about FHIR in general at fhir.org and FHR development at build.fhir.org.

Topic	Value
Connection	Web API
Authentication	API username/password

The following FHIR resources are supported.

Resource	Documentation
Code system	build.fhir.org/codesystem.html

Resource	Documentation
Condition	build.fhir.org/condition.html
Encounter	build.fhir.org/encounter.html
Location	build.fhir.org/location.html
Medication	build.fhir.org/medication.html
Observation	build.fhir.org/observation.html
Organization	build.fhir.org/organization.html
Patient	build.fhir.org/patient.html
Person	build.fhir.org/person.html
Practitioner	build.fhir.org/practitioner.html
Questionnaire	build.fhir.org/questionnaire.html
Questionnaire response	build.fhir.org/questionnaireresponse.html
Value set	build.fhir.org/valueset.html

Settings

Field	Description
Questionnaires	The questionnaires to include

Google Sheets

Topic	Value
Connection	Web API
Authentication	API token part of connection URL
Data model	Sheet

Connection

Field	Description
URL	URL to sheet including API token

Settings

Field	Description
Schema	Sheets in JSON format

iHRIS

Topic	Value
Connection	Database (JDBC)
Authentication	Database username/password
Data model	Form

MySQL database

Field	Description
Hostname	The hostname for the iHRIS database
Port	The port for the iHRIS database, often 3306
SSL	Whether to enable SSL encryption for the database connection
Trust server certificate	Whether to trust the server SSL certificate for the database connection
Database name	The name of the database
Database username	The username of the database user
Database password	The password of the database user

Settings

Field	Description
Forms	The forms to load data for
References	The references to load data for
Include record history	Whether to load record history

Kobo Toolbox

Topic	Value
Connection	Web API
Authentication	API token

Topic	Value
Data model	Survey

Connection

Field	Description
URL	URL to Kobo instance
Auth token	Authentication token for Kobo user account

Settings

Field	Description
Survey	The survey to load data for

ODK

Topic	Value
Connection	Web API
Authentication	API username/password
Data model	Project and form

Connection

Field	Description
URL	URL to ODK instance
Username	Username for ODK user account
Password	Password for ODK user account

Settings

Field	Description
Project	The project to load data for
Form	The form to load data for

Ona

Topic	Value
Connection	Web API
Authentication	API token
Data model	Form

Connection

Field	Description
URL	URL to Ona instance
Auth token	Authentication token for Ona user account

Settings

Field	Description
Form	The form to load data for

Talamus

Topic	Value
Connection	Web API
Authentication	API token
Data model	Facility

Connection

Field	Description
URL	URL to Kobo instance
Auth token	Authentication token for Kobo user account

Settings

Field	Description
Facilities	The facilities to load data for
Start date	The start date of the time range to load data for
End date	The end date of the time range to load data for

Amazon S3

Amazon S3 refers to files as objects.

Topic	Value
Connection	Web API
Authentication	Access key/secret key
Data model	Bucket and object

Source

Field	Description
Bucket	The bucket name
Object key	The key for the object to load
Access key	The IAM access key
Secret key	The IAM secret key

Azure Blob Storage

Azure Blob Storage refers to files as blobs.

Topic	Value
Connection	Web API
Authentication	Connection string
Data model	Container and blob

Source

Field	Description
Container name	The container name
Blob path	The path to the blob to load

Field	Description
Connection string	The connection string for the container

SQL Server

Topic	Value
Connection	Database (JDBC)
Authentication	Database username/password
Data model	Table

SQL Server

Field	Description
Hostname	The hostname for the database
Port	The port for the database, often 1433
SSL	Whether to enable SSL encryption for the database connection
Trust server certificate	Whether to trust the server SSL certificate for the database connection
Database name	The name of the database
Database username	The username of the database user
Database password	The password of the database user

Data source

Field	Description
SQL query	The SQL query for retrieving data to load
Tables	The database tables to load

MySQL

Topic	Value
Connection	Database (JDBC)
Authentication	Database username/password
Data model	Table

MySQL

Field	Description
Hostname	The hostname for the database
Port	The port for the database, often 1433
SSL	Whether to enable SSL encryption for the database connection
Trust server certificate	Whether to trust the server SSL certificate for the database connection
Database name	The name of the database
Database username	The username of the database user
Database password	The password of the database user

Data source

Field	Description
SQL query	The SQL query for retrieving data to load
Tables	The database tables to load

Oracle RDBMS

Topic	Value
Connection	Database (JDBC)
Authentication	Database username/password
Data model	Table

Oracle RDBMS

Field	Description
Hostname	The hostname for the database
Port	The port for the database, often 1433
SSL	Whether to enable SSL encryption for the database connection
Trust server certificate	Whether to trust the server SSL certificate for the database connection
Database name	The name of the database
Database username	The username of the database user
Database password	The password of the database user

Data source

Field	Description
SQL query	The SQL query for retrieving data to load
Tables	The database tables to load

PostgreSQL

Topic	Value
Connection	Database (psql)
Authentication	Database username/password
Data model	Table

PostgreSQL

Field	Description
Hostname	The hostname for the database
Port	The port for the database, often 1433

Field	Description
SSL	Whether to enable SSL encryption for the database connection
Trust server certificate	Whether to trust the server SSL certificate for the database connection
Database name	The name of the database
Database username	The username of the database user
Database password	The password of the database user

Data source

Field	Description
SQL query	The SQL query for retrieving data to load
Tables	The database tables to load

Amazon Redshift

Topic	Value
Connection	JDBC
Authentication	Database username/password, IAM ARN
Data model	Table

Amazon Redshift

Field	Description
Hostname	The hostname for the database
Port	The port for the database, often 1433
SSL	Whether to enable SSL encryption for the database connection

Field	Description
Trust server certificate	Whether to trust the server SSL certificate for the database connection
Database name	The name of the database
Database username	The username of the database user
Database password	The password of the database user

Data source

Field	Description
SQL query	The SQL query for retrieving data to load
Tables	The database tables to load

CSV file upload

Topic	Value
Connection	File upload
Authentication	-
Data model	Table

Settings

Field	Description
CSV files	One or more CSV data files to load
Delimiter	The CSV file delimiter

File format requirements

- If uploading multiple files, the schema (columns) must be the same for all files
- The first row should be the header defining the column names
- Column names must be unique within the file
- Column names are recommended to contain only letters and digits and start with a letter
- The filename is recommended to contain only letters and digits and start with a letter

Parquet file upload

Topic	Value
Connection	File upload
Authentication	-
Data model	Table

Settings

Field	Description
Parquet file	Parquet data file to load

Edit data pipeline

1. Find and click the data pipeline to edit in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Edit**.
4. Edit values in the relevant sections.
5. Click **Save** at the bottom of the section.
6. Close the dialog by clicking the close icon in the top-left corner.

Remove data pipeline

1. Find and click the data pipeline to remove in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Remove**.

Views

Overview

Views are SQL queries that act like virtual tables in a database system and display results returned from a specified SQL query. In AP, views enable users to manage and analyze data more efficiently, serving as a powerful abstraction layer that simplifies complex SQL queries into reusable and manageable components.

SQL views are versatile tools in data management, offering numerous data capabilities:

- *Join:* Views can seamlessly integrate data from multiple tables through SQL JOIN operations. This capability is invaluable when you need to combine related datasets for comprehensive analyses.
- *Filter:* Views can filter data to focus on specific records, making it easier to work with subsets of data pertinent to particular analyses or reports.

- *Aggregate*: With views, you can perform aggregation queries, such as SUM, AVG, MAX, and COUNT, to summarize data. This is particularly useful for generating high-level reports from detailed data records.
- *Enrich:* Views can also enhance data by incorporating calculated columns or by formatting existing data in a way that is more suitable for user requirements or specific analyses.

Types of views

AP provides utilize two primary types of SQL views.

- *Logical views*: These are the standard types of views that do not store data physically. They are essentially SQL queries which execute every time the view is accessed. Logical views are ideal for real-time data analysis and scenarios where data changes frequently.
- *Materialized views:* Unlike logical views, materialized views store the query result as a physical table on the disk. This type of view is particularly useful for datasets that do not change frequently but require fast read access. Materialized views improve performance by storing the computed result, reducing the load on compute resources during each query execution. Note that in AP, materialized views are created as regular tables. A benefit of tables over "native" materialized views is the ability to drop tables which are part of the SQL query of the view.

Manage views

The following section covers how to view, create, update and remove views.

View views

1. Click **Views** in the left side menu to list all views.
2. Click the name of a view to see more information.

Create view

1. Click the **Create new** button from the top-right corner.
2. Enter the following information.

Field	Description
Name	The name of the view
Description	A description of the view
Tags	Free text tags which categorizes the view
Schema	The schema in which to store the view

Field	Description
View name	The name of the data warehouse view, meaning the name as it will appear in the data warehouse
SQL query	The SQL query uses to retrieve data for the view

Edit view

1. Find and click the view to edit in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Edit**.
4. Edit values in the relevant sections.
5. Click **Save** at the bottom of the section.
6. Close the dialog by clicking the close icon in the top-left corner.

Edit SQL query

1. Find and click the view to edit in the list.
2. Click the context menu in the top-right corner.
3. Click **Edit the SQL query**.
4. In the SQL editor, edit the SQL query.
5. Click **Save**.

Remove view

1. Find and click the view to remove in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Remove**.

Destinations

Overview

AP offers *destinations* for loading data from the platform back into operational systems. This is valuable in order to enrich the destination system with integrated and harmonized data from a variety of sources.

The following destinations are supported.

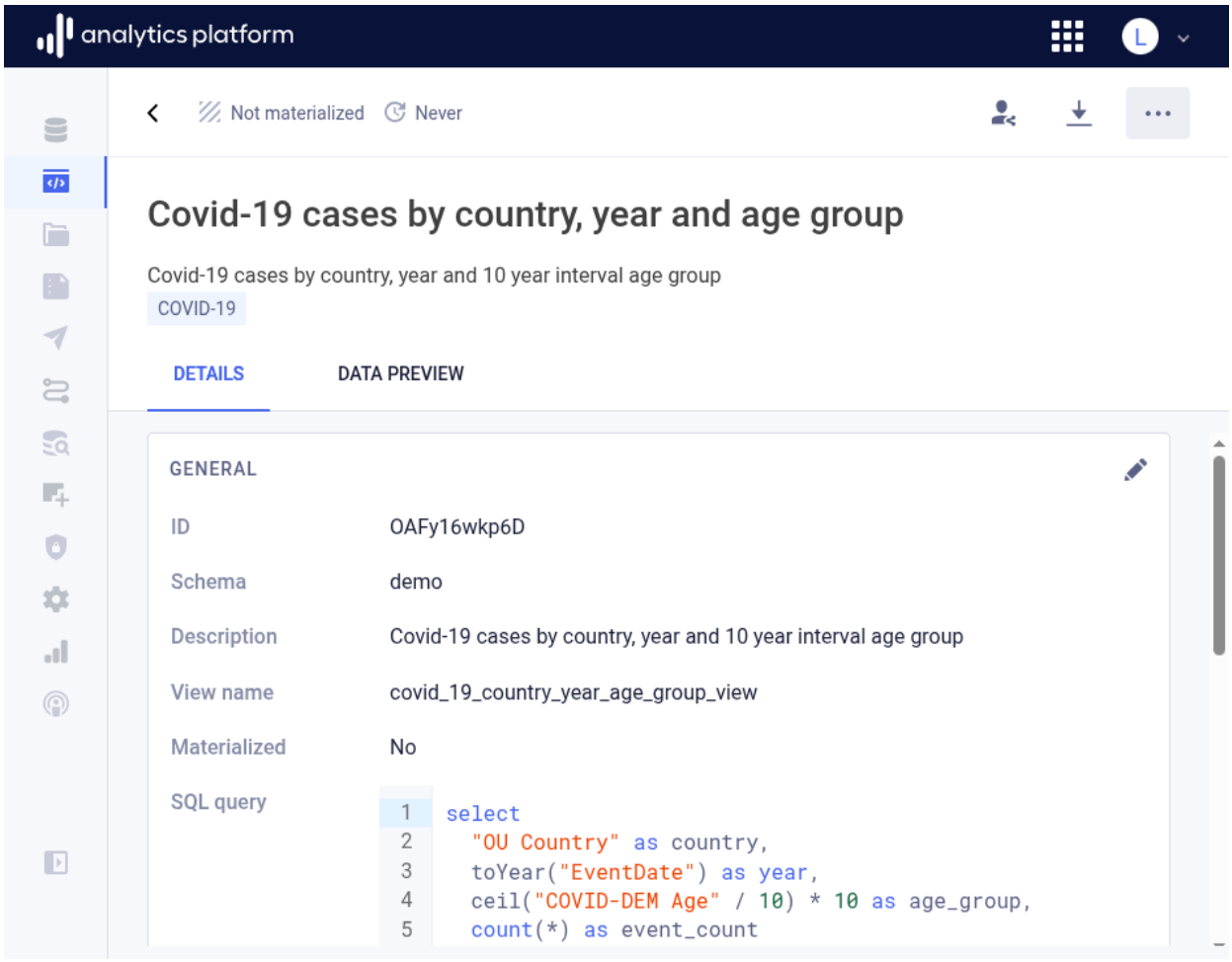

Figure 15: View overview

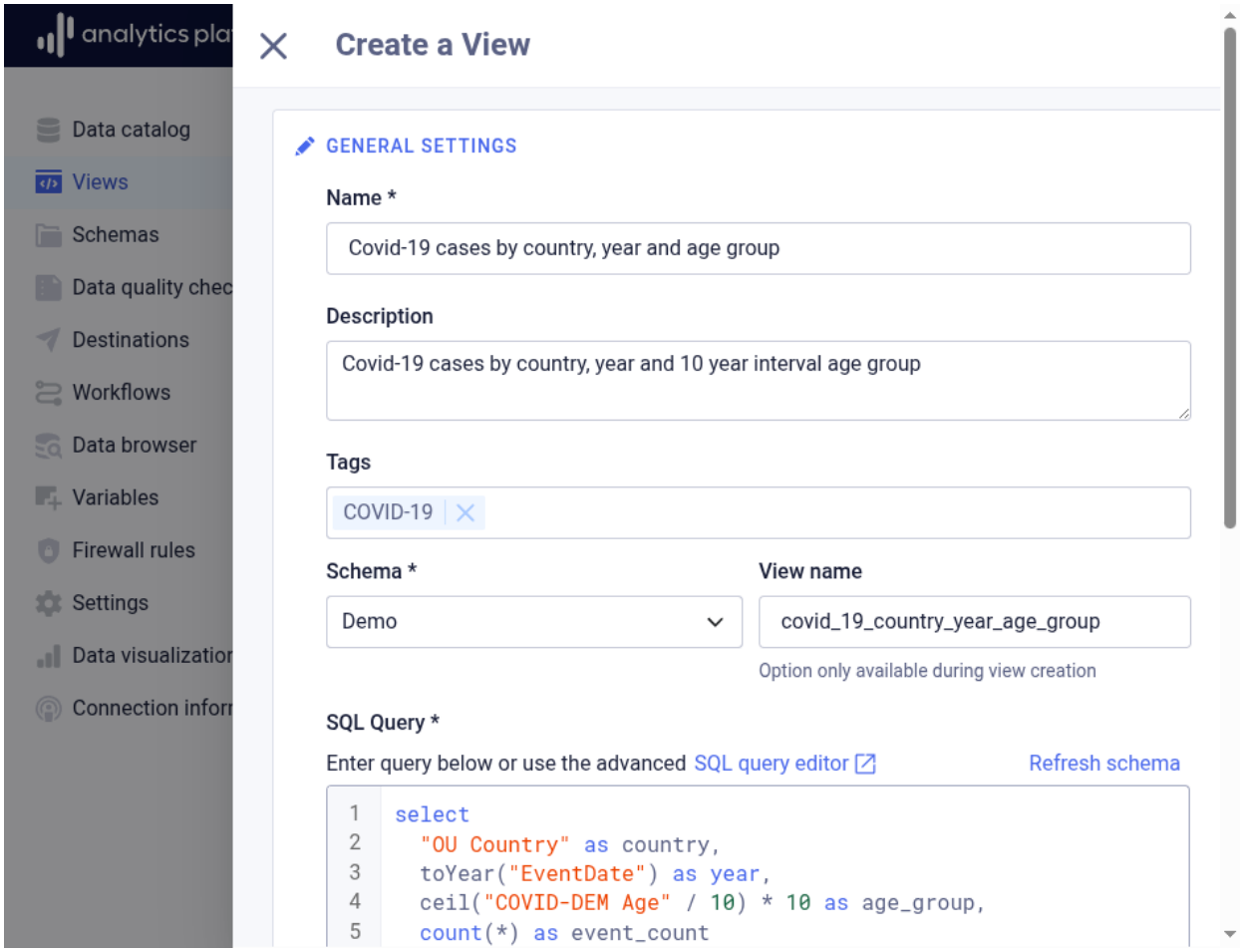

Figure 16: Create view

Applications

- **DHIS2:** A flexible, an open-source, web-based platform for collecting, analyzing and visualizing health data, widely used for managing and monitoring health programs, particularly in low-resource settings.

Manage destinations

The following section covers how to view, create, update and remove destinations.

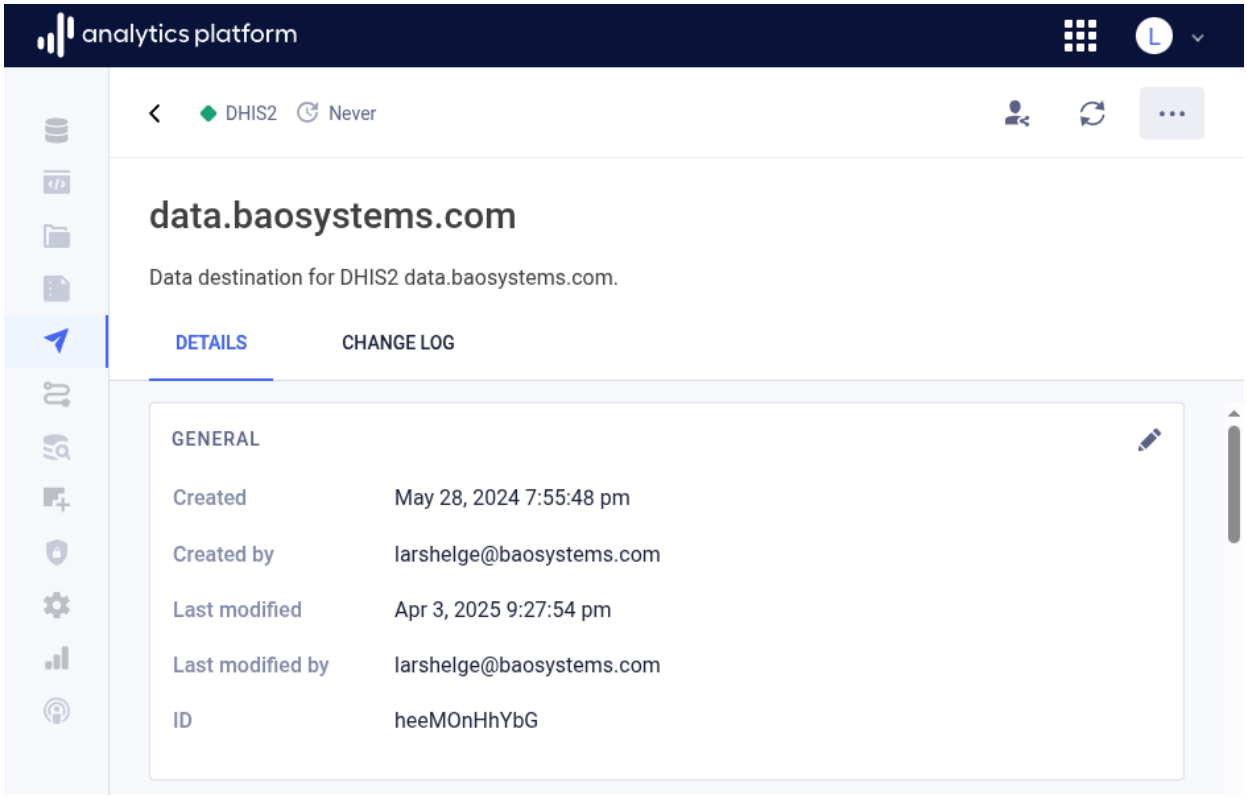

Figure 17: Destination overview

View destination

1. Click **Views** in the left side menu to view all destinations.
2. Click the name of a view to see more information.

Create destination

1. Click **Create new** from the top right corner.

2. Chose the type of destination.

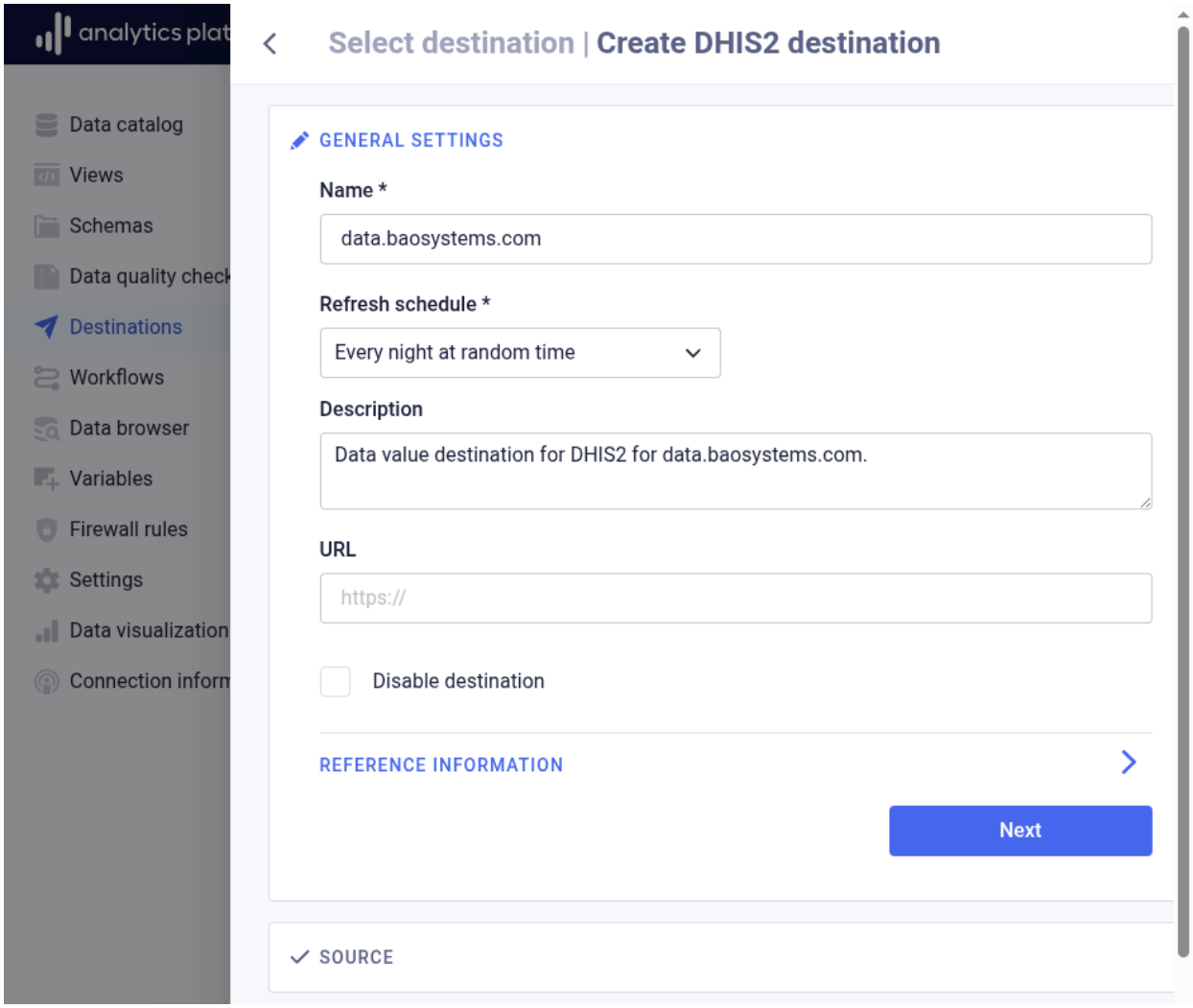

Figure 18: Create destination

General settings

In the *General settings* section, enter the following information. This section is present for all data pipeline types.

Field	Description
Name	The name of the destination (required)
Refresh schedule	The interval for when to refresh data from the data source (required)
Description	A description of the data pipeline
URL	A URL to the source data or system
Disable destination	Whether to disable loading of destination data
Reference	A reference text for the data source
Link to source	A URL refering to information about the data source
Link to terms of use	A URL refering to terms of use for the data source

Source

In the *Source* section, select the view to use to retrieve destination data.

Field	Description
View	The view to use to retrieve destination data

DHIS2 Web API

Field	Description
Base URL to web API	Base URL to web API for DHIS2 instance, do not include `/api`
Username	Username for DHIS2 user account
Password	Password for DHIS2 user account

Import options

Field	Description
Data element ID scheme	The data element ID scheme to use for data import
Org unit ID scheme	The organisation unit ID scheme to use for data import
Cat opt combo ID scheme	The category option combo ID scheme to use for data import
General ID scheme	The general ID scheme to use for data import
Dry run	Whether to make a dry run import without saving data in the destination
Skip audit	Whether to skip generating audit records during data import in the destination

Destination view

The following columns are supported for SQL views for DHIS2 destinations.

Field	Column name	Required
Data element	data_element_id	Yes
Period	period_id	Yes
Org unit	org_unit_id	Yes
Category option combo	category_option_combo_id	Yes
Attribute option combo	attribute_option_combo_id	Yes
Value	value	Yes
Stored by	stored_by	No
Comment	comment	No
Follow-up	comment	No

The column name matching is permissive and tolerates variations. Matching is case insensitive, and allows names with or without underscore and `_id` suffix. Using data element as an example, the following column names are valid:

Column name variation
data_element_id
dataelementid
data_element

Column name variation
DataElementID
DataElement

Examples

An example of a SQL query to use with a view as source for a DHIS2 destination.

```
select
  dv."DataElementID",
  dv."PeriodID",
  dv."OrgUnitID",
  dv."CatOptComboID",
  dv."AttOptComboID",
  dv."Value",
  dv."Deleted"
from
  dhis2.data_datavalue dv;
```

Edit destination

1. Find and click the destination to edit in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Edit**.
4. Edit values in the relevant sections.
5. Click **Save** at the bottom of the section.
6. Close the dialog by clicking the close icon in the top-left corner.

Remove destination

1. Find and click the destination to remove in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Remove**.

Workflows

Overview

The Analytics Platform (AP) offers a comprehensive workflow management system designed to orchestrate complex data pipelines efficiently. Workflows in AP automate the movement, transformation, and integration of data across various sources and systems into a unified data store, enabling integrated data analysis

through your chosen visualization tools. This section explains the architecture of workflows within AP and their critical role in streamlining data operations.

Benefits

Workflows and job orchestration in AP bring several key benefits.

- **Efficiency:** Automating data tasks reduces manual efforts and speeds up the data transformation processes.
- **Consistency:** Scheduled workflows ensure that data handling is performed consistently without gaps or overlaps, leading to reliable data integrity.
- **Scalability:** As organizational needs grow, workflows can be scaled to handle new data sources, increasing data volumes and complexity without compromising performance.

Workflow Model

- **Workflow**: A workflow is a structured sequence of operations designed to automate processes for data loading, transformation and integration. Each workflow consists of one or many steps.
- **Step:** A step encapsulates work to be done. Each step contains one or many jobs.
- **Job:** A job represents a specific task such as data extraction, transformation, or loading.

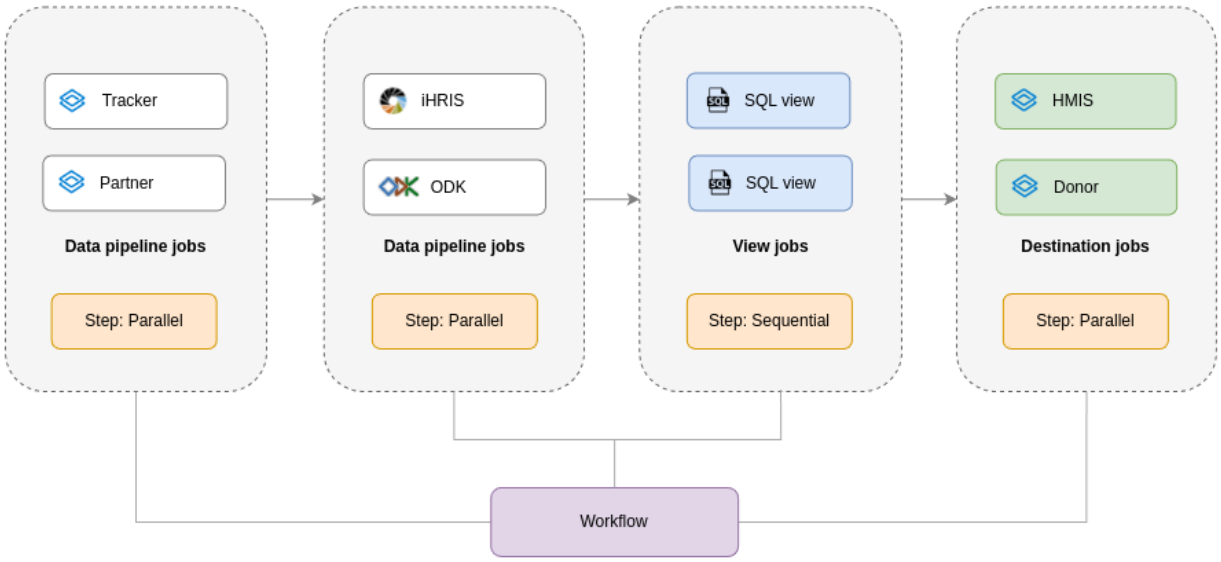

Figure 19: Workflow diagram

Scheduling

A workflow can be scheduled to run at specific intervals, which automates the recurring tasks and ensures data freshness without manual intervention. This scheduling capability is crucial for maintaining up-to-date data views and operational readiness in dynamic business environments. Workflows can be set to continuous updates. This will apply to jobs which support continuous updates, such as DHIS2 data pipelines.

Steps

Steps can be configured to run jobs either in sequence or in parallel within a step. Running jobs in parallel will reduce the total runtime of a workflow, thereby enhancing the efficiency of data processing tasks. If jobs have dependencies on each other, meaning the output of one view is required as input for another view, running jobs in serial, meaning in sequence, is recommended.

Jobs

Jobs within a workflow are categorized into three types:

- **Data pipeline:** Responsible for extracting and moving data from source systems into the data store.
- **View:** Handles joining and integration of datasets to prepare them for analysis.
- **Destination:** Manages loading of processed data back into operational systems or other destinations for further use.

Manage workflows

The following section covers how to create, update and remove workflows.

View workflow

1. Click **Workflows** in the left-side menu.
2. Click the name of a workflow to see more information.

Create workflow

1. Click the **Create new** button from the top-right corner.
2. In the *General settings* section, enter the following information.

Field	Description
Name	The name of the view
Refresh schedule	The interval for when to refresh data from the data source (required)
Description	A description of the view

Field	Description
Disable workflow	Whether to disable the workflow

Create step in workflow

1. In the *Steps* section, click **+ Add step** to add one or more steps.
2. In the *Add step* section, enter the following information.

Field	Description
Execution mode	Whether to execute jobs in the step in serial or in parallel
Disable workflow	Whether to disable the workflow

3. Click **+ Add job** to add one or more jobs.
4. In the section for new job, select the following information.

Field	Description
Type	The type of job
Source/Target	For data pipelines, the data source type; for destinations, the target type
Name	The data pipeline, view or destination

5. Click the check icon to save the job.
6. Repeat from 3. to create additional jobs.
7. Click **Save** to save the step.
8. Repeat to create additional steps.

Edit workflow

1. Find and click the workflow to edit in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Edit**.

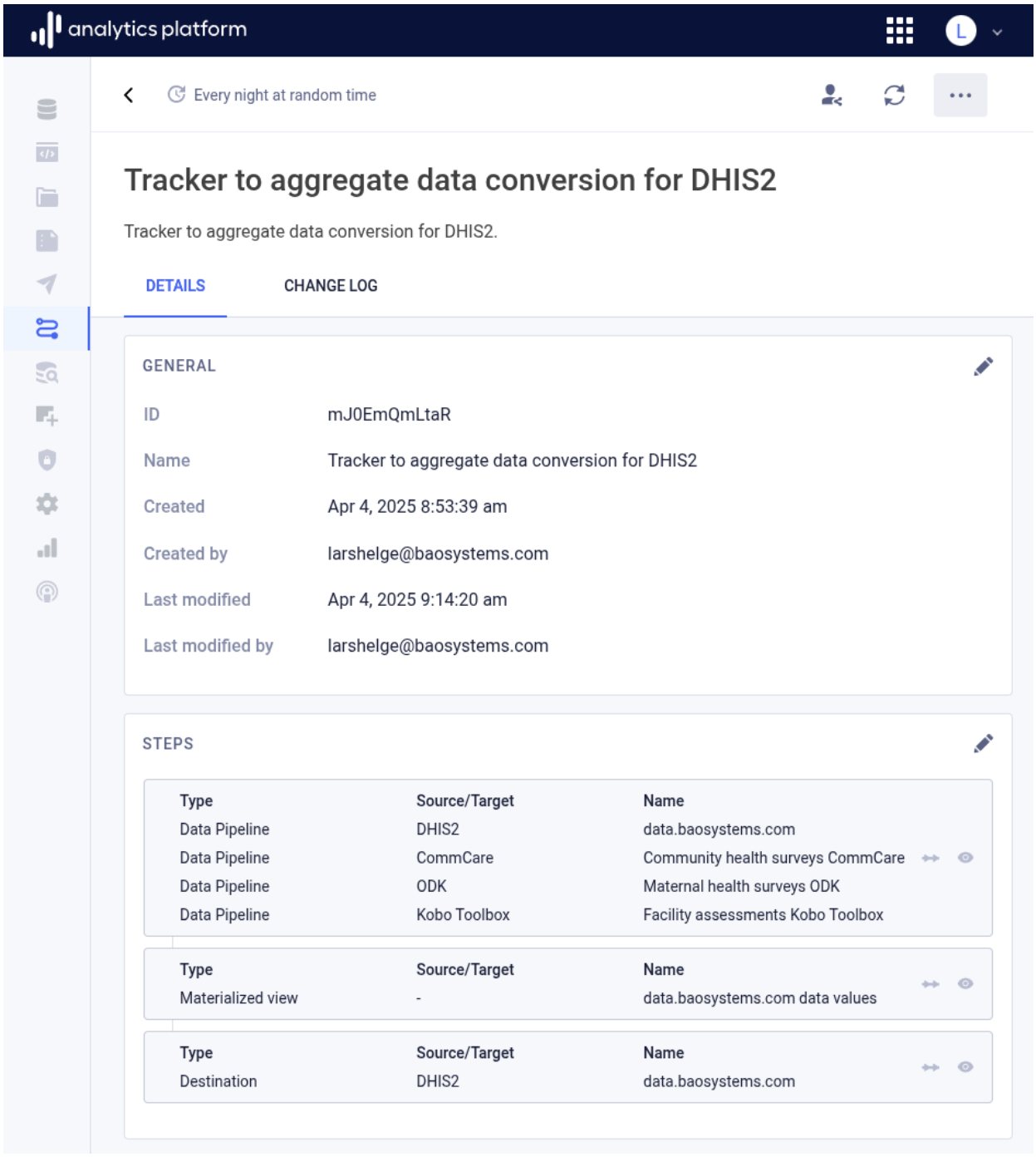

Figure 20: Workflow overview

4. Edit values in the relevant sections.
5. Click **Save** at the bottom of the section.
6. Close the dialog by clicking the close icon in the top-left corner.

Remove workflow

1. Find and click the workflow to remove in the list.
2. Open the context menu by clicking the icon in the top-right corner.
3. Click **Remove**.

Refresh data

Workflows can be manually triggered.

1. Open the context menu by clicking the icon in the top-right corner.
2. Click **Refresh data**.

Data browser

Overview

Data browser is a core component of AP, designed to empower users to engage directly with their data through interactive querying across all available datasets. This allows for exploring data in near real-time, enabling users to derive insights and make informed decisions rapidly.

In AP, data pipelines automates the ingestion of data from multiple sources into the central data store. These pipelines are configured to handle diverse data formats and sources, such as applications, databases and data files, ensuring that the data is up-to-date and readily accessible. Once the data is in the platform, users can create queries which span all data sources and datasets, allowing for integrated data exploration and analytics.

The main query language of the bata browser is SQL, or Structured Query Language. SQL is a standardized language for managing and manipulating databases and data warehouses. SQL provides a powerful means to execute queries on data in a *declarative* style. It allows users to specify exactly what data they need from a database without requiring detailed knowledge of how the database is structured or stored. Typical operations are selecting specific data, aggregating data across dimensions, filtering data on particular values and joining tables together. SQL is widely known and used among data professionals, making it a common language for data exploration and analysis.

Schema navigator

The schema navigator is placed on the left-side panel of the data browser. It outlines the entire schema of the data available to the user. The schema is displayed as a hierarchy, where the first level represents table *schemas*, the second level represents *tables* and the third level represents table *columns*.

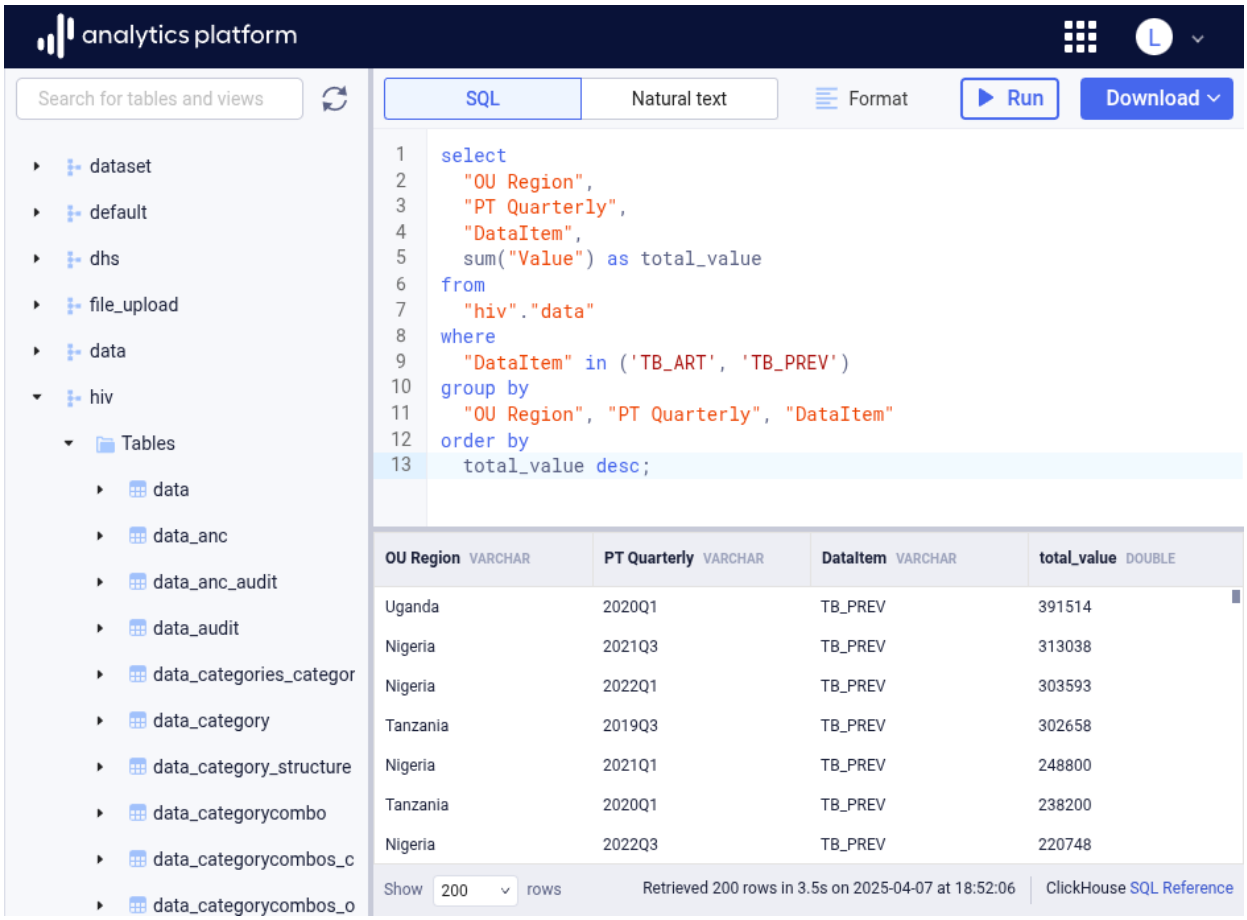

Figure 21: Data browser overview

Level	Description
1	Schema
2	Tables
3	Columns

Expanding an item in the hierarchy will reveal items at the next level. The column *data type* is displayd next to each column name.

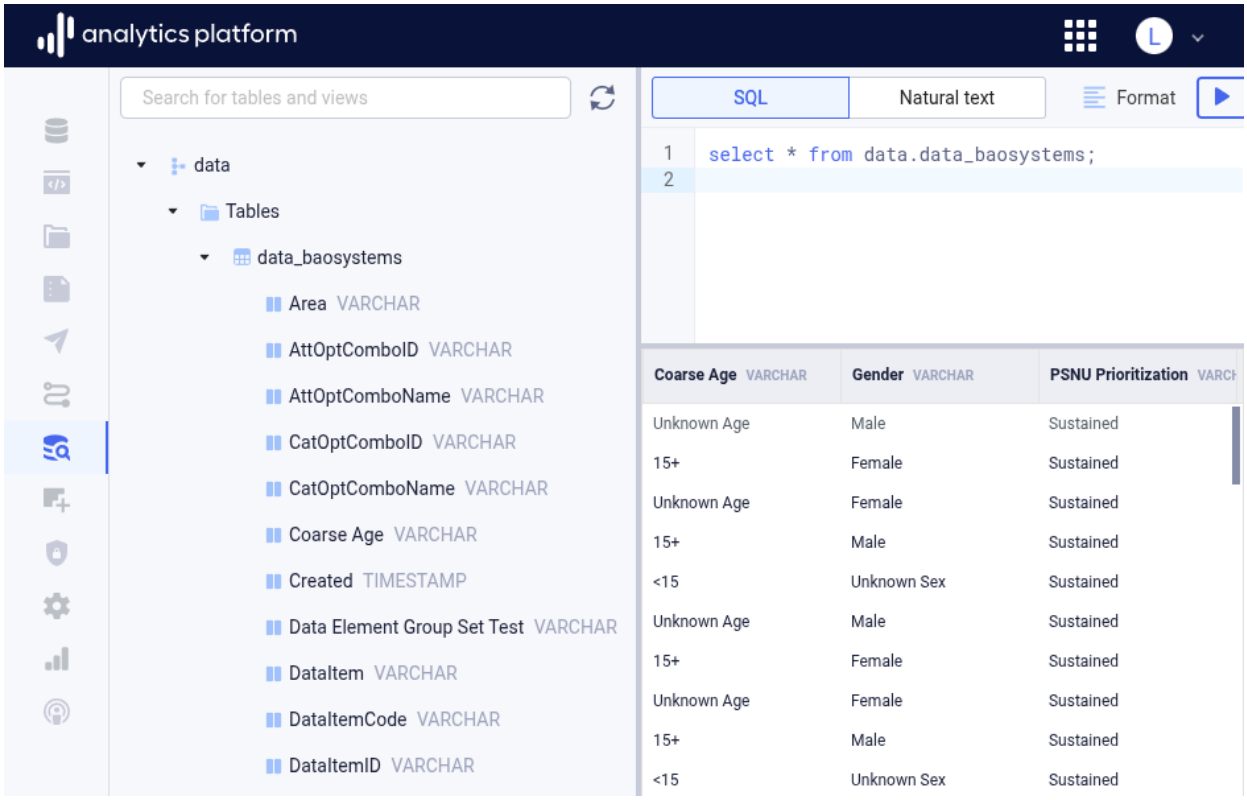

Figure 22: Data browser schema navigator

Query editor

The query editor is placed at the center of the data browser. This is the area where the query can be specified. There are two types of queries: **SQL** and **Natural text**. The type of query to work with can be selected at the top bar.

SQL queries

Users can write their SQL queries directly into the query editor area. The editor supports auto-completion of SQL statements to make writing more efficient. To activate auto-complete, press **Ctrl + Space** (Windows/Linux) or **Command + Space** (macOS), with the cursor at the relevant position of the query.

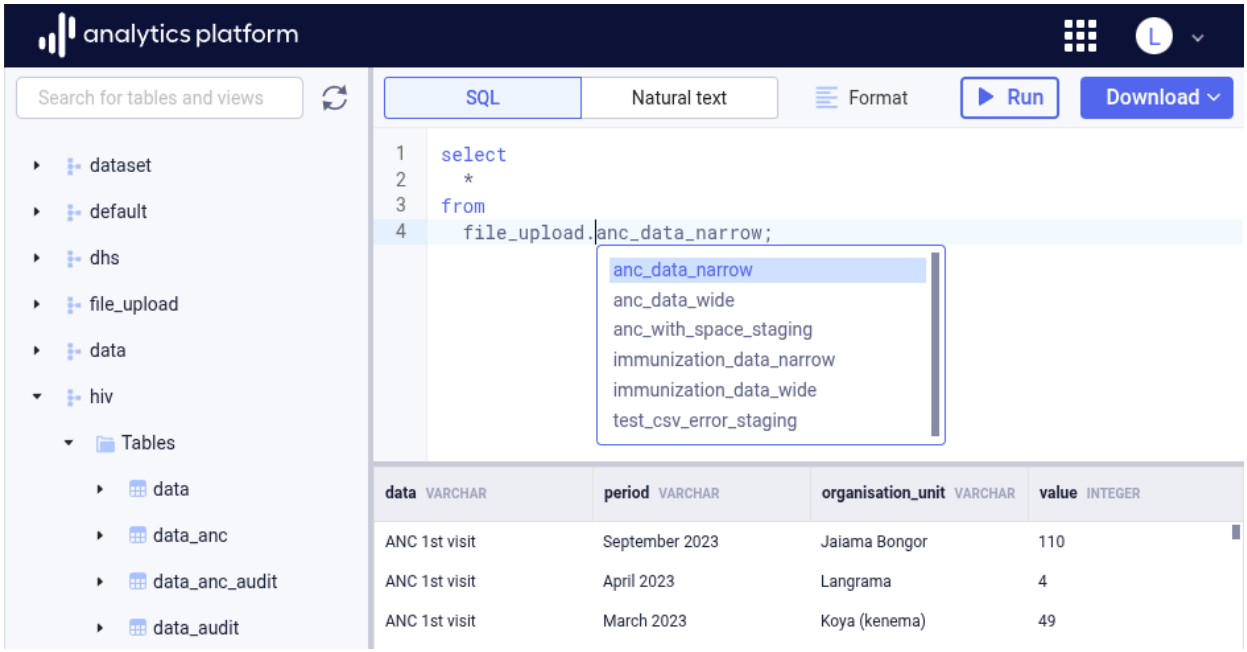

Figure 23: Data browser auto completion

After writing a SQL query, click **Run**.

SQL queries will vary depending on the schema. A simple example of a SQL query that summarizes data values and groups by data item, quarterly time periods and countries:

```
select
  d."DataItem",
  d."PT Quarterly",
  d."OU Country",
  sum(d."Value")
from
  demo.demo d
group by
  d."DataItem",
  d."PT Quarterly",
  d."OU Country";
```

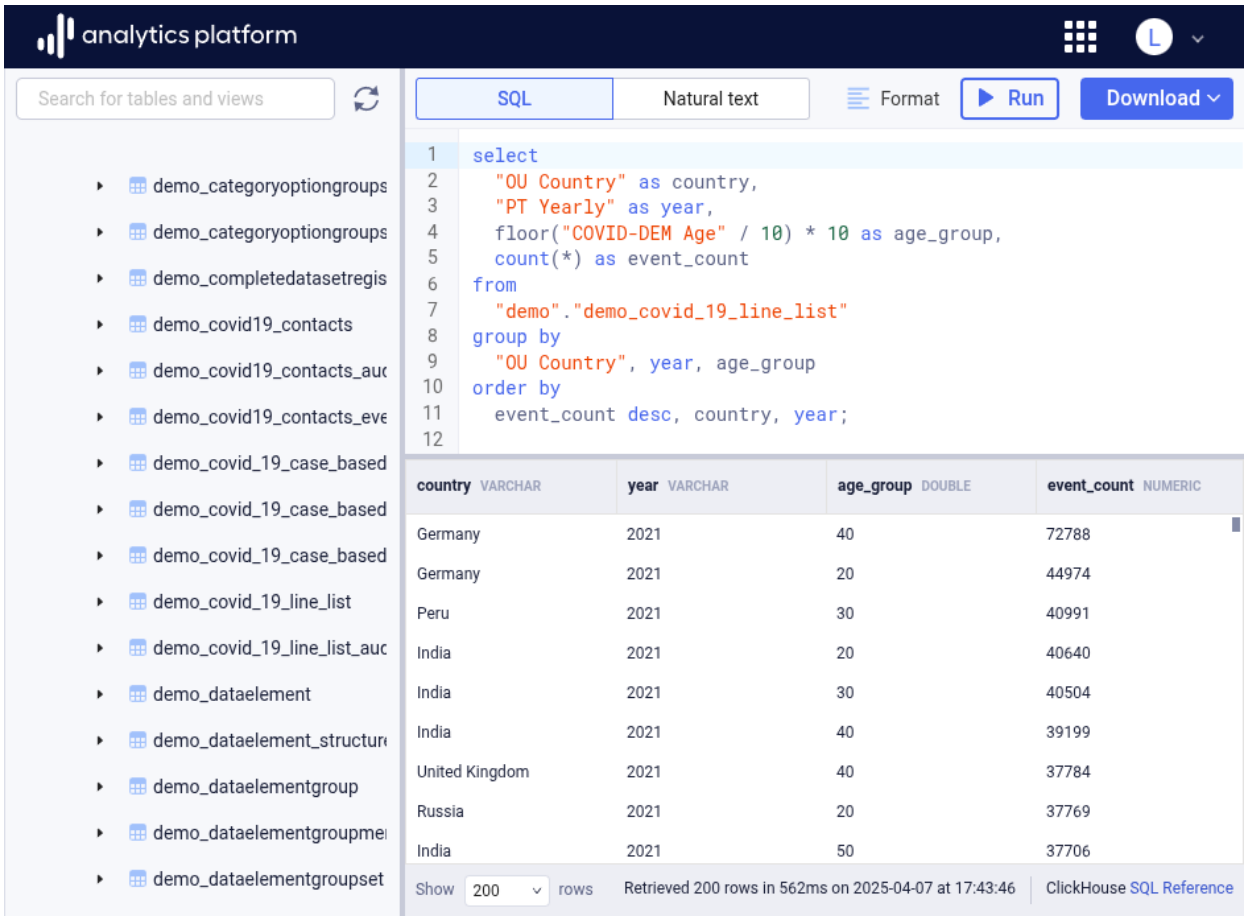

Figure 24: Data browser SQL query

Formatting queries

To format a SQL query to make it more readable, click **Format** on the top bar.

Viewing results

The query response will appear in the result area. The query result is displayed as a table, with the name and the data type of each column displayed on the header row. By default, the first 200 rows of the result are displayed. The number of rows to display can be changed from the bottom bar drop-down to 500 or 1000.

Natural text queries

Users who are not proficient in SQL can write queries in natural language text. Click **Natural text** in the top bar to switch to natural text queries. Select one or more schemas from the schema selector at the top bar to narrow down the part of the schema to retrieve data from.

With text queries, a user can ask simple questions about metadata, for example:

```
Tell me about the ID, names and code of all data elements.
```

The result of metadata queries can be used to ask more sophisticated data questions, for example:

```
Give me the sum of data values for data items related to TB_ART and TB_PREV by
quarter and OU Region level. Include the data item. Order by data value descending.
```

Download query result

After a query has run successfully, the result of the query can be downloaded to a data file in CSV format. The user is provided with two options.

Download preview

Downloads the rows which are visible in the result area. The number of rows can be changed from the bottom bar drop-down. The format is uncompressed CSV.

Download full dataset

Downloads the entire set of rows produced by the query. This download option will *stream* results to the web browser. The format is Gzip compressed CSV. Note that downloading extremely large datasets is not recommended.

Tools for decompressing Gzip files are pre-installed on MacOS and Linux. The 7-Zip tool is recommended for Microsoft Windows.

SQL reference

SQL is a standard query language defined by ANSI which ensures interoperability across data warehouses supported by AP. Numerous courses and guides exist online for learning purposes.

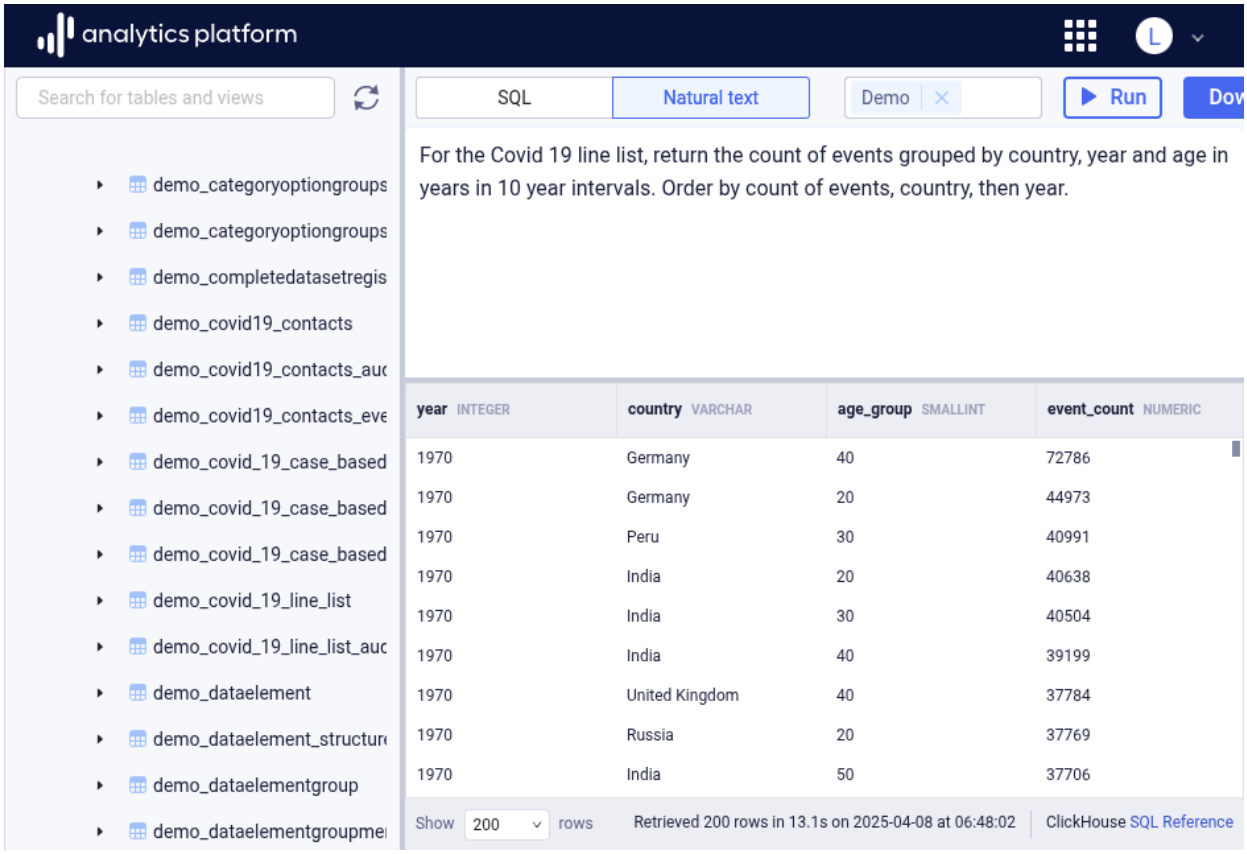

Figure 25: Data browser natural text query

However, every data warehouse provides a range of specific features and functions. Users writing SQL queries can learn about data warehouse specific functions by consulting the respective SQL reference guides listed below. You can observe the type of data warehouse from the right-side label at the bottom bar, and explore the respective SQL guide by clicking the **SQL reference** link next to it, or from the table below.

Data warehouse	SQL reference guide
PostgreSQL	www.postgresql.org
Amazon Redshift	docs.aws.amazon.com
ClickHouse	clickhouse.com
SQL Server	learn.microsoft.com
Synapse	learn.microsoft.com

Text-to-SQL

The following diagram describes the tetx-to-SQL solution at a high level.

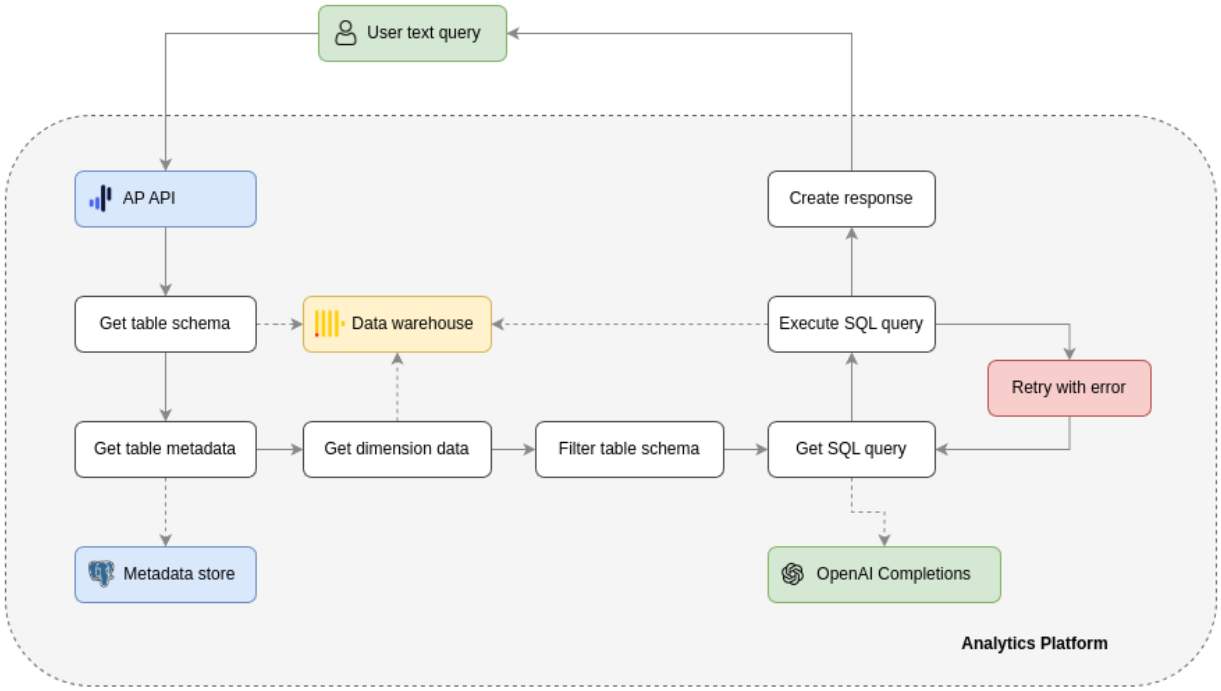

Figure 26: Text-to-SQL solution architecture

Data quality checks

Overview

Ensuring high data quality is crucial for any organization that relies on data for decision-making, analysis, and strategic planning. High-quality data can significantly enhance accuracy in reporting, consistency in analytics, and reliability in automated decisions. Conversely, poor data quality can lead to misguided decisions based on inaccurate, incomplete, or outdated information.

AP provides data quality checks to ensure the integrity and accuracy of your data. These checks allow users to define specific criteria that data must meet before it is considered valid for analysis and reporting. This functionality includes:

- **Outlier detection:** Identify data points that deviate significantly from the norm. Outliers may indicate data entry errors or unusual events that could skew analysis results.
- **Relationship:** Ensure that relationships between data items make sense. For example, the number of tests should most likely not exceed positive tests.
- **Data completeness:** Verify that all required data fields are populated and that data spans the required time frames or categories.
- **Consistency:** Compare data across time, category and sources to ensure data has a consistent format, is free from duplicates and uses the same coding system.

Duality checks in AP is based on SQL queries which define and enforce these rules. By writing a SQL query, users can precisely specify the conditions under which data is considered valid. Conversely, SQL query will reveal conditions which are in violation of the check. When a SQL query identifies data that violates a quality check, AP can trigger alerts or even prevent the integration of flawed data into your reports and analyses.

Manage data quality checks

The following section covers how to view, create, update and remove data quality checks.

Create data quality check

1. Click **Create new** from the top-right corner.
2. Enter the following information.

Field	Description
Name	The name of the schema (required)
Description	A description of the schema
Code	A code of the schema
Labels	One or many labels on the format `key:value`

Field	Description
SQL query	A SQL query which specifies the conditions under which data is considered valid

3. Click **Create**.

Edit data quality check

1. Find and click the data quality check to update in the list.
2. Click the context menu in the top-right corner.
3. Click **Edit**.
4. Update the relevant fields.
5. Click **Save**.

Edit SQL query

1. Find and click the data quality check to edit in the list.
2. Click the context menu in the top-right corner.
3. Click **Edit the SQL query**.
4. In the SQL editor, edit the SQL query.
5. Click **Save**.

Remove data quality check

1. Find and click the data quality check to remove in the list.
2. Click the context menu in the top-right corner.
3. Click **Remove**.

Schemas

Overview

A schema refers to a namespace with the platform and data warehouse. It defines the structure of the data and represents how tables and views are organized. This setup is similar to having folders within a single file system, where each folder can contain files with the same names and provide clarity and structure.

Schemas help in segregating database objects according to their use, type, access level, or any other criteria that suits the business. This allows for a cleaner and more organized data structure, making it easier for users to locate and manage their data. Since objects are contained within schemas, users can avoid naming conflicts in a shared database environment.

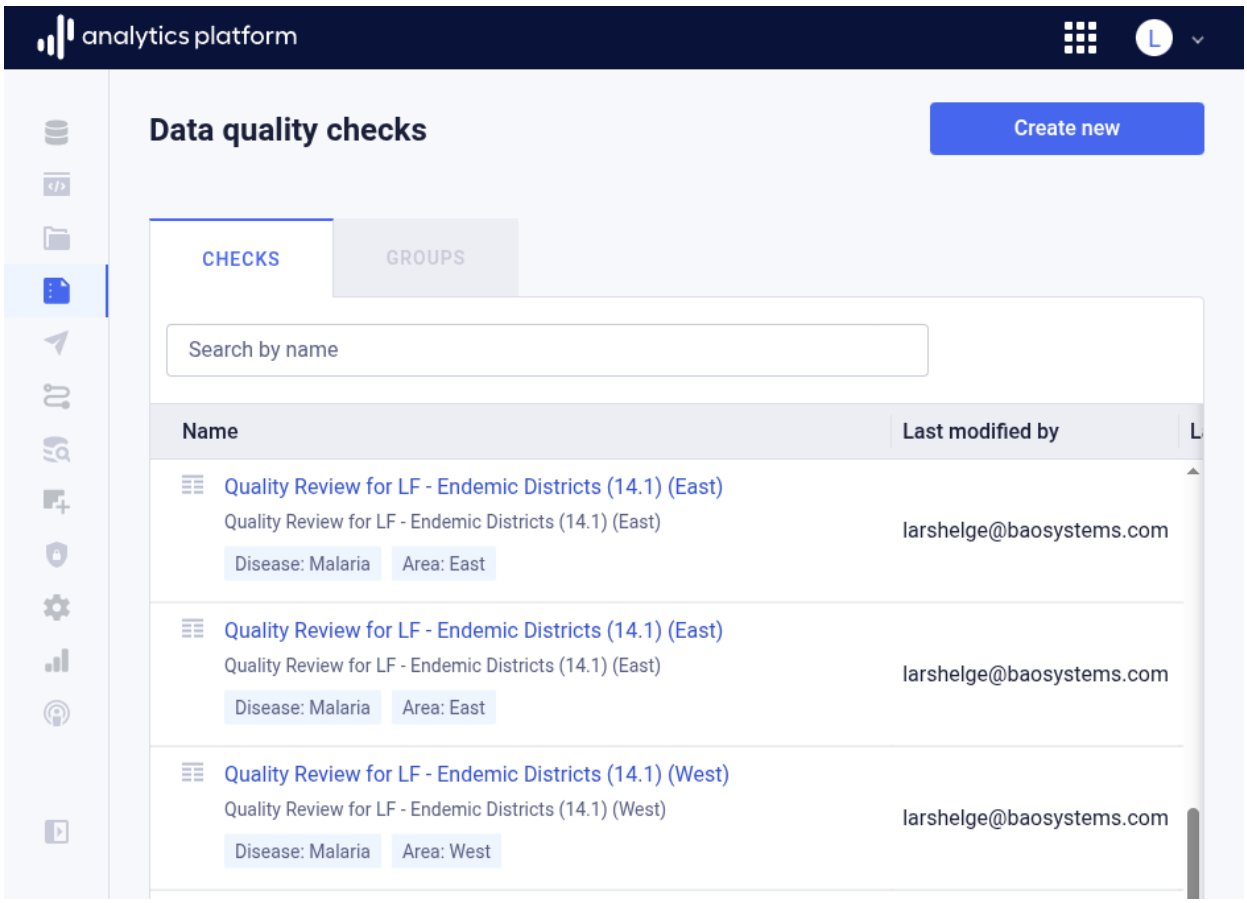

Figure 27: Data quality check overview

When creating data pipelines and views, a schema must be selected in the *Data warehouse target* section. For multi-table data pipeline types, like DHIS2, it is advisable to create a schema per DHIS2 instance for improved organization of tables.

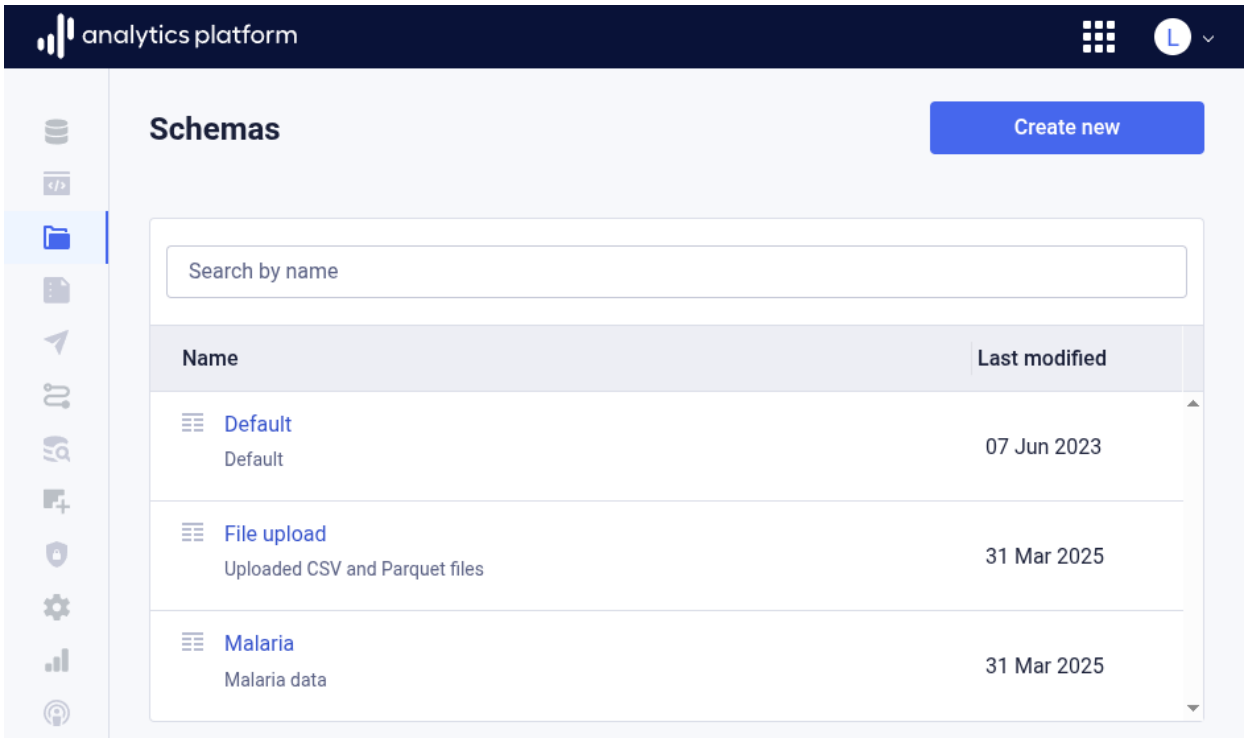

Figure 28: Schema list

Manage schemas

The following section covers how to view, create, update and remove schemas.

View schema

1. Click **Schemas** in the left side menu to list all schemas.
2. Click the name of a schema to view more information.

Create schema

1. Click **Create new** from the top-right corner.
2. Enter the following information.

Field	Description
Name	The name of the schema (required)
Description	A description of the schema
Tags	One or many tags which describe the schema

3. Click **Create**.

Edit schema

1. Find and click the schema to edit in the list.
2. Click the context menu in the top-right corner.
3. Click **Edit**.
4. Update the relevant fields.
5. Click **Save**.

Remove schema

1. Find and click the schema to remove in the list.
2. Click the context menu in the top-right corner.
3. Click **Remove**.

Permanent schemas

In AP, one or more schemas are defined as *permanent*. This schema is built into the data warehouse and cannot be renamed or removed. It is advisable to not overly use the permanent schemas and instead create more specific schemas for each type of tables and views. The table below describes the permanent schemas for each data warehouse type.

Data warehouse	Name	UID
ClickHouse	default	J5bHYonzwDY
PostgreSQL	public	SRndd67ndLP
Amazon Redshift	public	TPUfm314K8k
Microsoft SQL Server	dbo	B7zjADK2Jin
Azure Synapse	dbo	Qxkm9zeMGPl
Azure Synapse	guest	P91CTQou2sN
Azure Synapse	sys	ZMX22Oo4UK1

Variables

Overview

Variables in the AP are named key-value pairs that you can define once and reuse across your data pipelines, integrations, and configurations. They are especially useful for managing values that change between environments, like development, staging, and production, or that are used repeatedly in different workflows, such as database credentials and API tokens.

The variables page in AP provides a central location where you can create, update, and view all available variables in your workspace. Each variable has a name and a corresponding value, and once defined, it can be referenced anywhere in your configuration using the `${VARIABLE_NAME}` format. For example, referencing `${DB_USERNAME}` in a connection or script will automatically substitute the value of the `DB_USERNAME` variable at runtime.

Variables can be either plain or secure. Secure variables are always hidden from logs and user interfaces to protect sensitive information. To mark a variable as secure, check the secure checkbox when creating or editing it. This is recommended for variables that contain secrets like passwords, API keys or access tokens. Plain variables, on the other hand, are visible in the interface and logs and are suitable for non-sensitive values such as configuration flags or environment labels. Both secure and plain values are encrypted at rest.

Using variables in AP brings several benefits. It reduces duplication by allowing you to define a value once and reuse it everywhere. It also enhances security by minimizing the risk of exposing sensitive values in logs or user interfaces. Most importantly, it makes your workflows more portable and easier to maintain, as if a value changes, you only need to update it in one place.

Manage variables

The following section covers how to view, create, and remove variables.

View variables

1. Click **Variables** to list all variables.

Create variable

1. Enter the following information in the input fields at the top of the page.

Field	Description
Name	Variable name
Value	Variable value
Secured	Whether to always hide the value

2. Click **Create**.

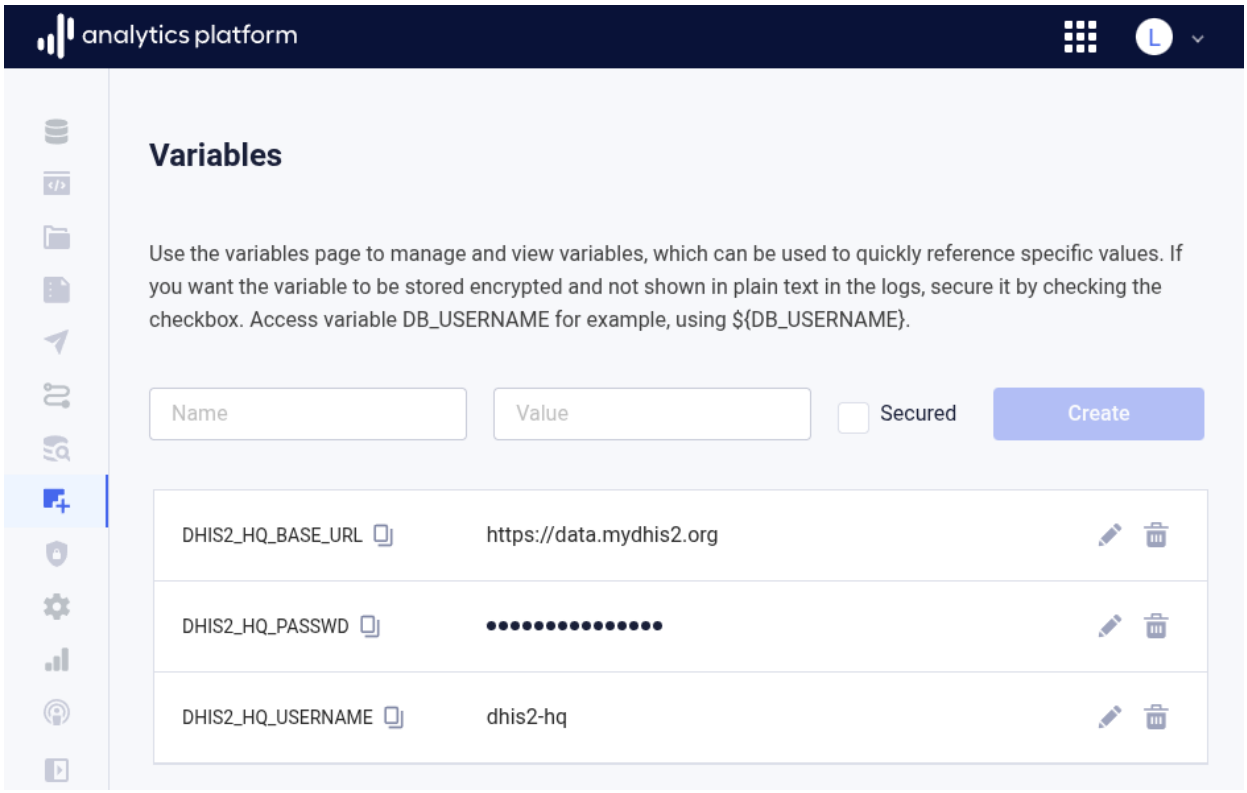

Figure 29: Variable list

Update variable

1. Find the variable to edit in the list.
2. Click the **Update** icon next to the variable.
3. Edit the variable name or value.
4. Click the **Done** icon next to the variable.

Remove variable

1. Find the variable to remove in the list.
2. Click the **Remove** icon next to the variable.

Example

1. Click **Variables**.
2. Create a variable with name `DHIS2_HQ_USERNAME` with the name of a DHIS2 user account.
3. Click **Data catalog**.
4. Create or edit a DHIS2 data pipeline.
5. Click **Web API**.
6. In the **Username** field, enter `${DHIS2_HQ_USERNAME}`.
7. Click **Save**.

Firewall rules

Overview

!!! tip “Note” Firewall rules are only supported when using MS SQL Server and Azure Synapse as data warehouse

A firewall is a network security system that controls and restricts incoming and outgoing network traffic based on predefined security rules. In the context of the Analytics Platform (AP), firewall rules play a key role in protecting your data warehouse by ensuring that only trusted sources can establish direct connections.

By default, direct access to the data warehouse is disabled for all external sources. This is a security-first approach to ensure that your data is protected against unauthorized access, potential breaches, and misuse. All data processing, ingestion, and analysis tasks are performed within the AP environment unless explicit access has been granted.

In cases where you need to connect to the data warehouse from external desktop applications, such as Power BI, Tableau, or other Business Intelligence (BI) tools, you can define firewall IP rules to allow specific IP addresses to connect. This enables secure and controlled access from your local environment, allowing you to build dashboards, run custom queries, or analyze data directly from your preferred tools. Always ensure that only trusted IP addresses are added to minimize security risks.

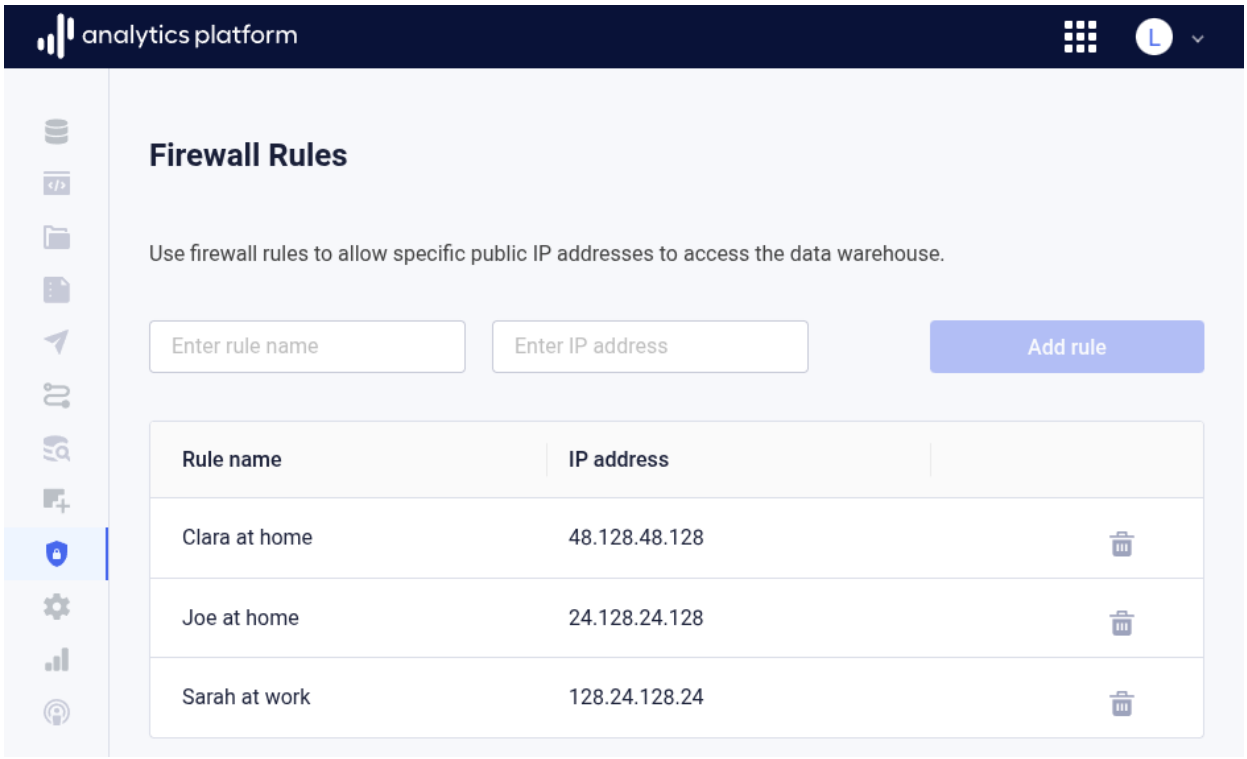

Figure 30: Firewall rule list

Manage firewall rules

The following section covers how to view, create, and remove firewall rules.

View firewall rule

1. Click **Firewall rules** in the left side menu.

Create firewall rule

1. Enter the following information in the input fields at the top of the page.

Field	Description
Name	Rule name
Address	IP address to allow connections from

2. Click **Add rule**.

Remove firewall rule

1. Find the variable to remove in the list.
2. Click the **Remove** icon next to the variable.

Sharing

Overview

AP provides an access control model referred as *sharing*. The sharing model works on the *object* level, where objects means specific instances of the various entities in AP. The following entities support sharing in AP.

- Data pipelines
- Views
- Data quality checks
- Destinations
- Workflows

The sharing model controls which users can view and edit specific objects in AP. Sharing is in other words two-dimensional.

Who has access

The first dimension defines *who* has access to an object. The following three levels exist.

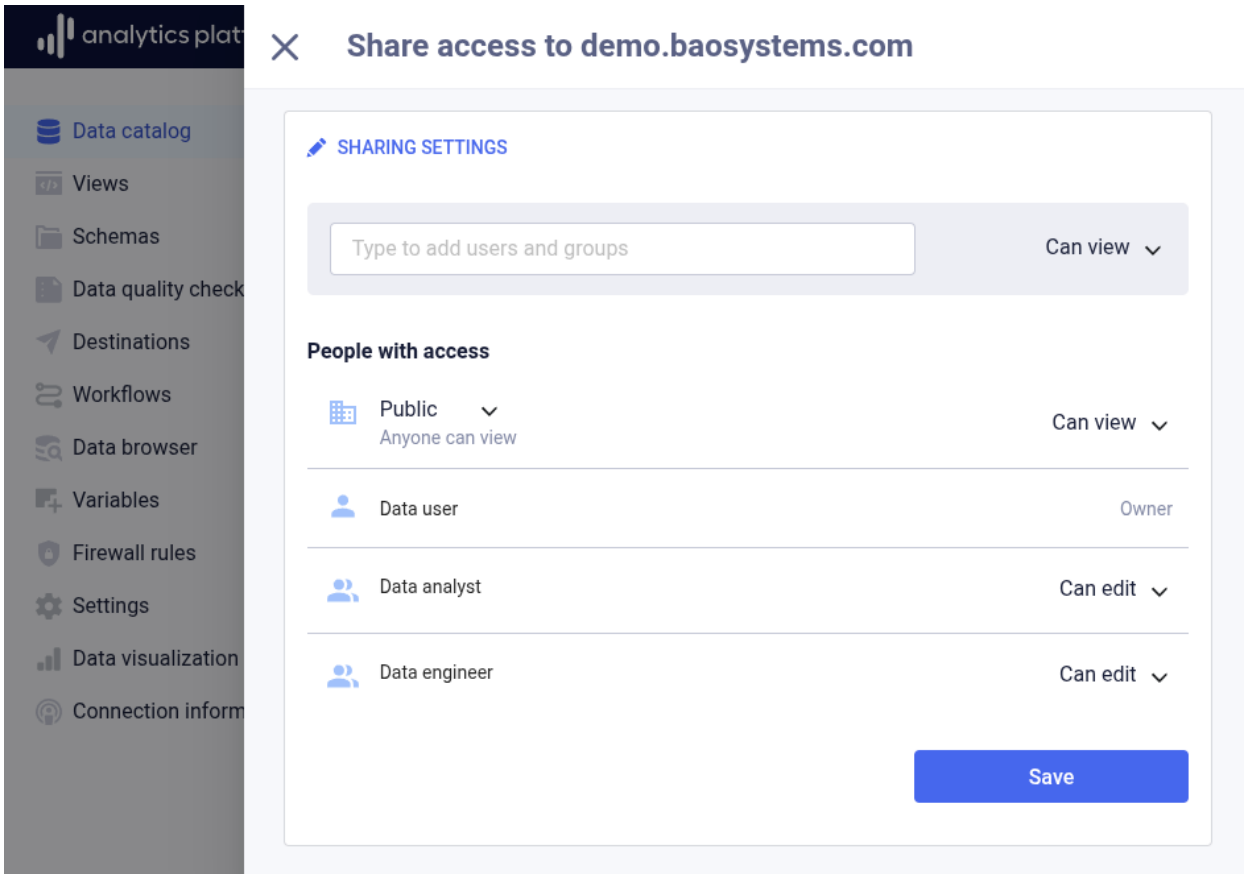

Figure 31: Sharing overview

Level	Description
Public	All authenticated users within the organization (**not** anyone on the Internet)
User group	Users which are members of a specific group
User	Specific users

What actions are allowed

The second dimension defines *what actions* a user are allowed on an object. The following three levels exist.

Level	Description
Can view	Read permission
Can edit	Read and write permission
None	No permission

The combination of who has access and what actions are allowed defines the sharing model for objects in AP.

Managing sharing

The following section covers how to set and update sharing for an object.

Open sharing dialog

1. In the list of objects (e.g. data pipelines), click the name of the object to view more information.
2. Click the context menu in the top-right corner.
3. Click **Share**.

Set who has access

1. Enter the name of the user group or user in the seach input field.
2. Check the checkbox next to the user group or user to share the object with.
3. Click anywhere outside the search dialog to close it.

Set what actions are allowed

1. Next to the **Public** label, select *Can view*, Can edit* or None* from the drop-down. To remove public access altogether, select *Restricted* from the **Public** drop-down.
2. Next to each user group and user, select *Can view*, *Can edit* or *Remove access* from the drop-down.
3. Click **Save** to store the sharing settings.

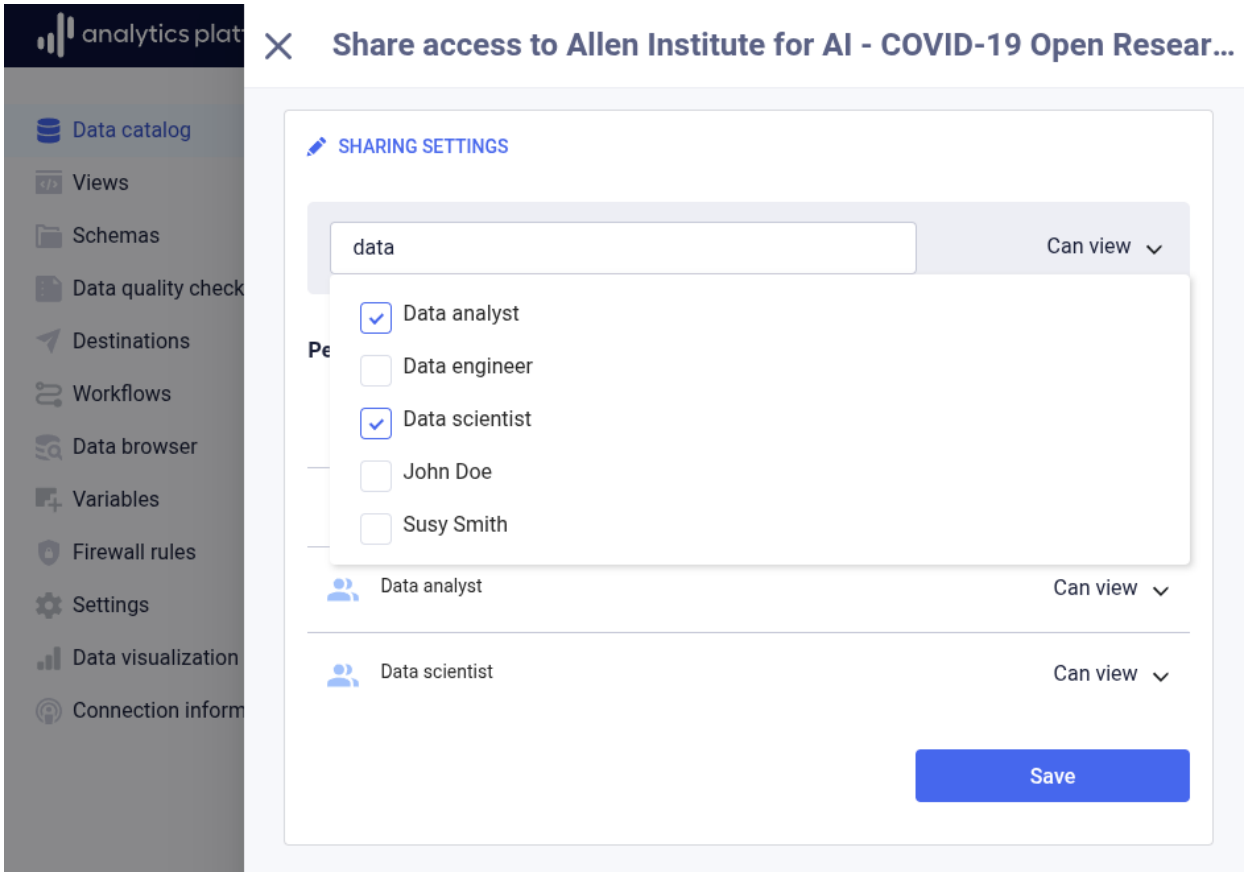

Figure 32: Sharing user

Settings

Overview

The settings page allows for specifying desirable platform behavior. Settings work at the organization level, also called client or tenant.

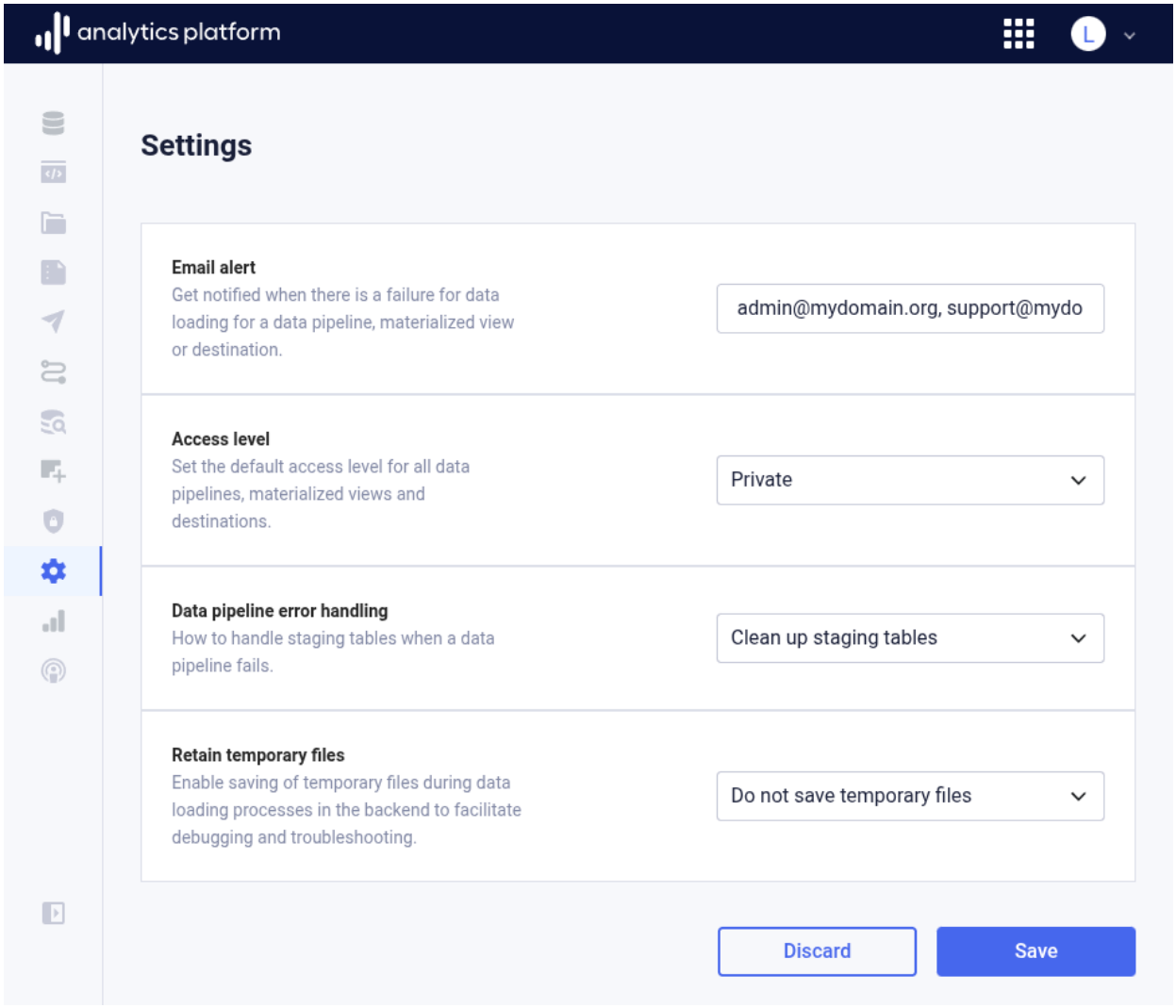

Figure 33: Setting overview

Settings

The following settings are available.

Setting	Description
Email alert	Who should receive an email notification if a pipeline or materialized view fails to refresh due to an error. Enter value as one or many email addresses, separated by comma.
Access level	The default sharing level for new objects, such as data pipelines, views and destinations, to be either public or private.
Data pipeline error handling	Whether staging tables should be cleaned up, meaning removed, or retained when a data pipeline data load operation fails. Staging tables are temporary database tables which are created as part of the data loading processs. Retaining staging tables can be helpful for troubleshooting. As standard practice, cleaning up staging tables is advisable.
Retain temporary files	Whether temporary data files generated during data loading processes in the platform backend should be retained to facilitate debugging and troubleshooting. Note that enabling this property should only be done by a system administrator for short periods of time.

After specifying one or many system settings, click **Save** to have the setting changes saved, or click **Discard** to have the setting changes discarded.

Connection information

Overview

The connection information page provides you with information about the data warehouse integrated in AP. This information is useful when connecting desktop applications, like Power BI, Tableau and other BI tools, or cloud services, like AWS and Azure.

When connecting directly to the AP data warehouse, it is typically required to open a port in the firewall. Make sure to follow best security practices when allowing direct connections.

If your AP user account has permission for accessing the data warehouse, this means that a corresponding user account exists in the data warehouse. You can authenticate to the data warehouse using the same password as you use for logging in to AP.

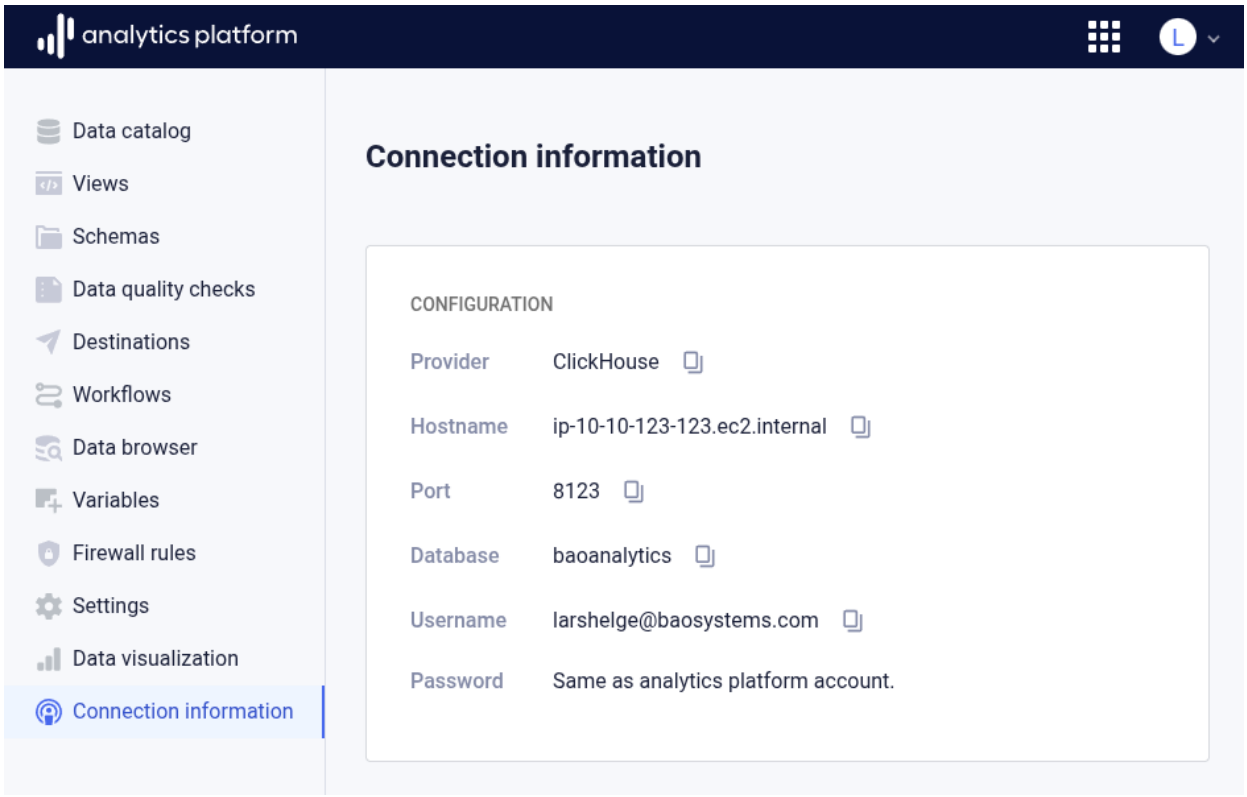

Figure 34: Connection information

Information

The connection information page offers the following information.

Field	Description
Provider	Data warehouse management system
Hostname	Hostname of data warehouse
Port	Port of data warehouse
Database	Database name
Username	Database username
Password	Use the AP user account password

Data warehouses

Overview

AP supports the following data warehouses.

- ClickHouse
- PostgreSQL
- Amazon Redshift
- Azure SQL Database
- Azure Synapse
- Microsoft SQL Server

Notes:

- ClickHouse is the default data warehouse for AP.
- Azure SQL Database is a managed database service in the Azure cloud based on Microsoft SQL Server.
- Azure Synapse is a cloud data warehouse which largely adheres to Microsoft SQL Server data types and SQL syntax.

ClickHouse

ClickHouse is an open-source columnar database management system optimized for online analytical processing. It enables fast data insertion and real-time query performance, making it well-suited for handling large volumes of data.

Topic	Value
Default port	8123
Default schema	default

Data type mapping

AP	ClickHouse
Small int	Int16
Integer	Int32
Big int	Int64
Numeric	Decimal
Real	Float32
Double	Float64
Boolean	Bool
Char	String
NChar	String
Varchar	String
NVarchar	String
Text	String
NText	String
Date	String
Timestamp	DateTime64
Timestamp with timezone	DateTime64
Time	DateTime64
Time with timezone	DateTime64
Geometry	String
JSON	String
Binary	FixedString

PostgreSQL

PostgreSQL is a powerful, open-source object-relational database system known for its robustness, scalability, and support for advanced SQL compliance. It offers a wide range of features, including complex queries, foreign keys, triggers, views, transactional integrity.

Topic	Value
Default port	5432
Default schema	public

Data type mapping

AP	PostgreSQL
Small int	smallint
Integer	integer
Big int	bigint

AP	PostgreSQL
Numeric	numeric
Real	real
Double	double precision
Boolean	boolean
Char	char
NChar	char
Varchar	varchar
NVarchar	varchar
Text	varchar
NText	varchar
Date	date
Timestamp	timestamp
Timestamp with timezone	timestamptz
Time	time
Time with timezone	timetz
Geometry	geometry
JSON	json
Binary	bytea

Microsoft SQL Server

Microsoft SQL Server is a relational database management system developed by Microsoft, designed to support a wide range of data applications, including transaction processing, business intelligence, and analytics.

Topic	Value
Default port	1433
Default schema	dbo

Data type mapping

AP	SQL Server
Small int	smallint
Integer	int
Big int	bigint
Numeric	numeric
Real	real
Double	float
Boolean	varchar

AP	SQL Server
Char	char
NChar	nchar
Varchar	varchar
NVarchar	nvarchar
Text	varchar
NText	nvarchar
Date	date
Timestamp	datetime2
Timestamp with timezone	datetimeoffset
Time	time
Time with timezone	time
Geometry	varbinary
JSON	nvarchar
Binary	varbinary

Super BI

Overview

AP and the *Super BI* web app allow for embedded data visualizations and business intelligence (BI) with Apache Superset integrated within DHIS2.

Apache Superset

Apache Superset is an open-source data exploration and visualization platform designed to be intuitive and highly accessible for business intelligence purposes. Superset is integrated in AP in the following ways.

- **Single Sign-On:** SSO provides a seamless user experience as users can sign in once and later navigate between AP and Superset without having to log in again.
- **Datasets for views:** When a user creates a view in AP, a corresponding dataset is automatically created in Superset. The dataset can be used as basis for Superset charts and dashboards.
- **Embedded dashboards:** Superset dashboards can be embedded in DHIS 2 with the Super BI web app, allowing for exploration of data stored in AP from within DHIS2.

The following is an overview of the data analytics model in Apache Superset.

- **Dataset:** A data table, can be *physical*, meaning based on a data warehouse table, or *virtual*, meaning based on a SQL query.
- **Chart:** A visualization, such as a column chart, bar chart, line chart, bubble chart, box plot, tree map, table or pivot table.
- **Dashboard:** A collection of visualizations which are organized and arranged to provide a comprehensive view of your data at a glance.

For more information about Apache Superset, consult the official documentation.

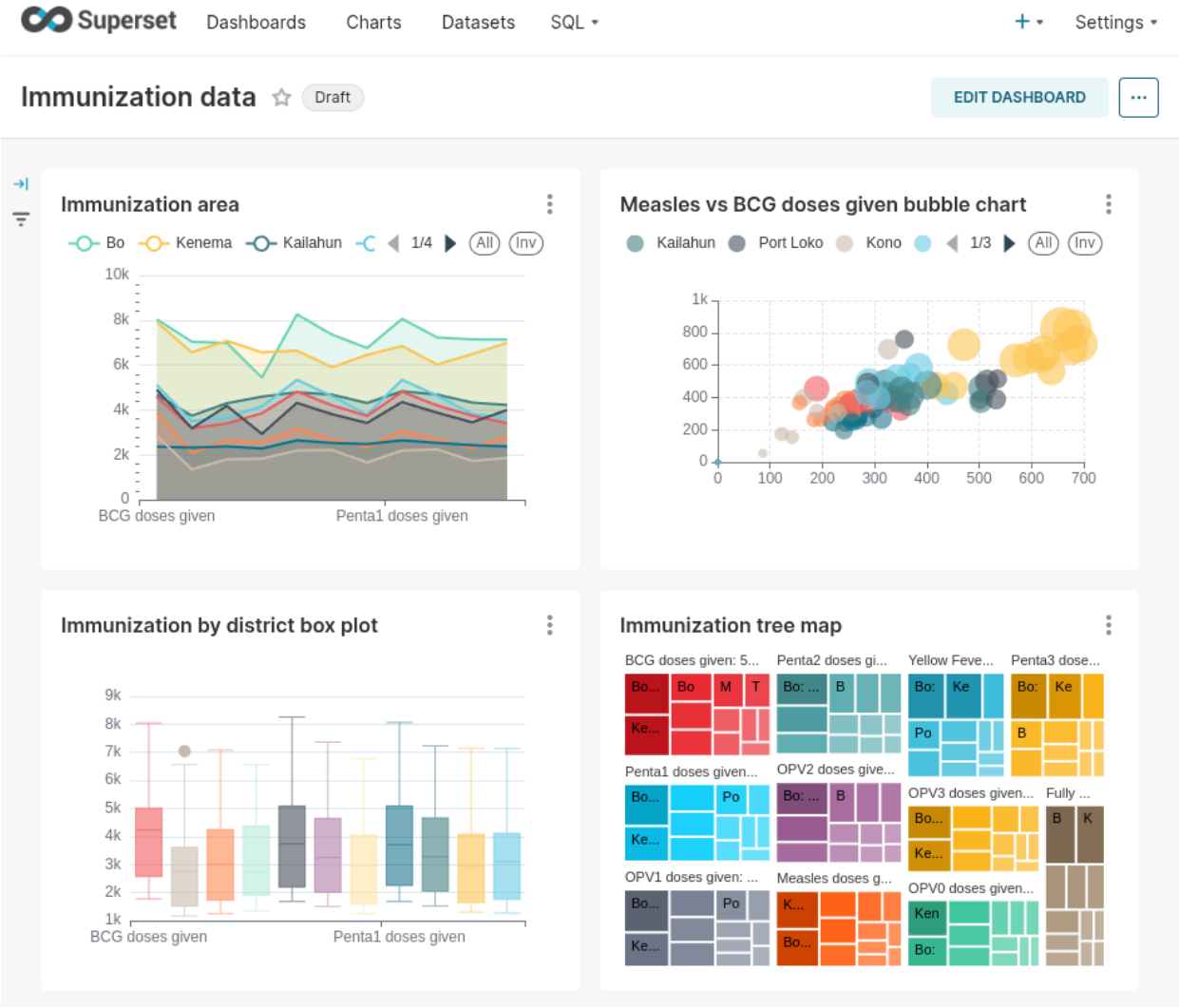

Figure 35: Superset dashboard

Manage embedded dashboards

The following section covers how to view, create, share and remove embedded dashboards.

Create embedded dashboard

The following describes the high-level flow for embedding a dashboard with Super BI.

1. Create a dashboard in Apache Superset.
2. Enable embedding for the dashboard and take note of the embed ID.
3. Create a dashboard in Super BI and use the embed ID to embed the Superset dashboard.

The following describes the steps in more detail.

Create Superset dashboard

1. In Apache Superset, click **Dashboards**.
2. Click the + **Dashboard** button from the top-right corner, which will open the dashboard screen.
3. Drag charts from the right-side bar.
4. Click **Save**.

Enable embedding for Superset dashbboard

1. Click the three-dot context menu in the top-right corner.
2. Click **Embed dashbard**, which will open the embed dialog.
3. Click **Enable embedding**, which will reveal an embed ID.
4. Copy and store the embed ID.

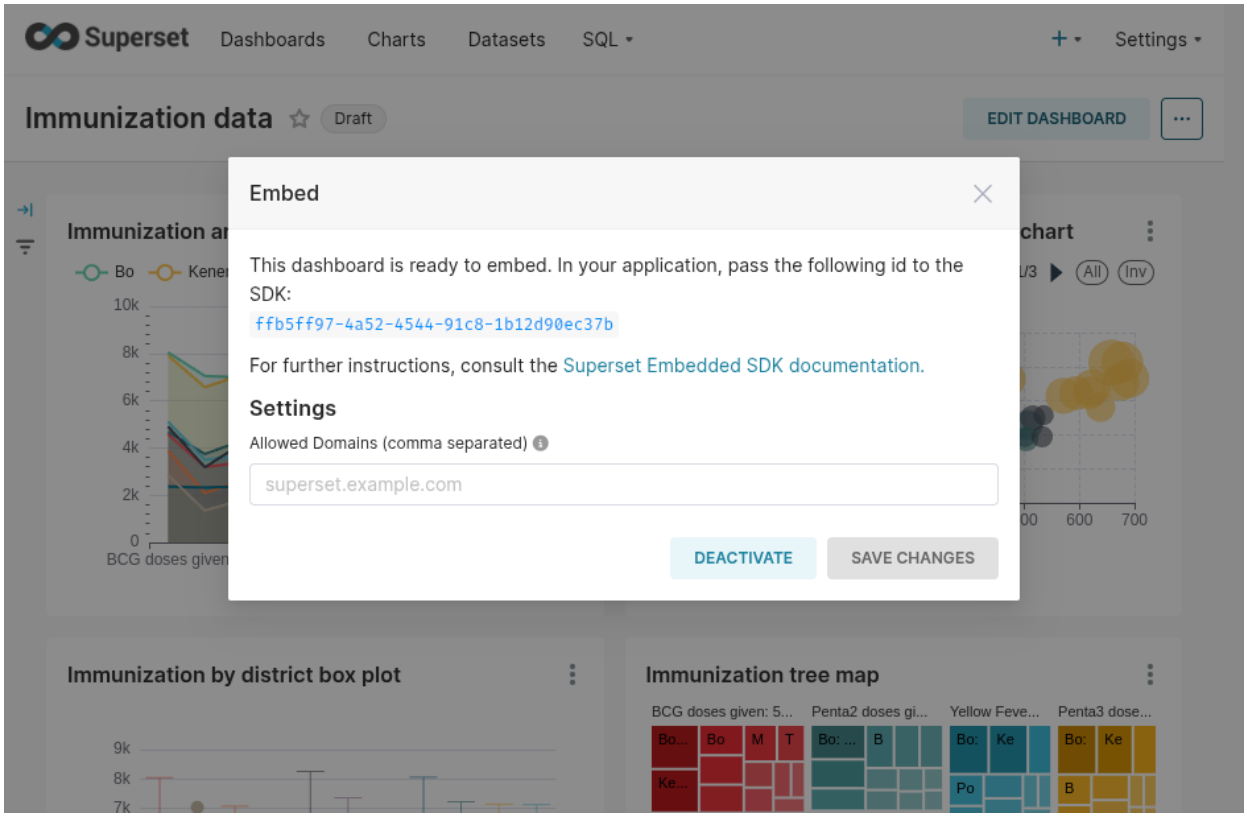

Figure 36: Superset dashboard embed ID

Create Super BI dashboard

1. In the Super BI web app for DHIS2, click the + button in the top-left corner.
2. Enter the following information.

Field	Description
Name	The name of the dashboard (required)
Superset embed ID	The Superset embed ID previsouly retrieved from Superset (required)

3. Click **Save**.

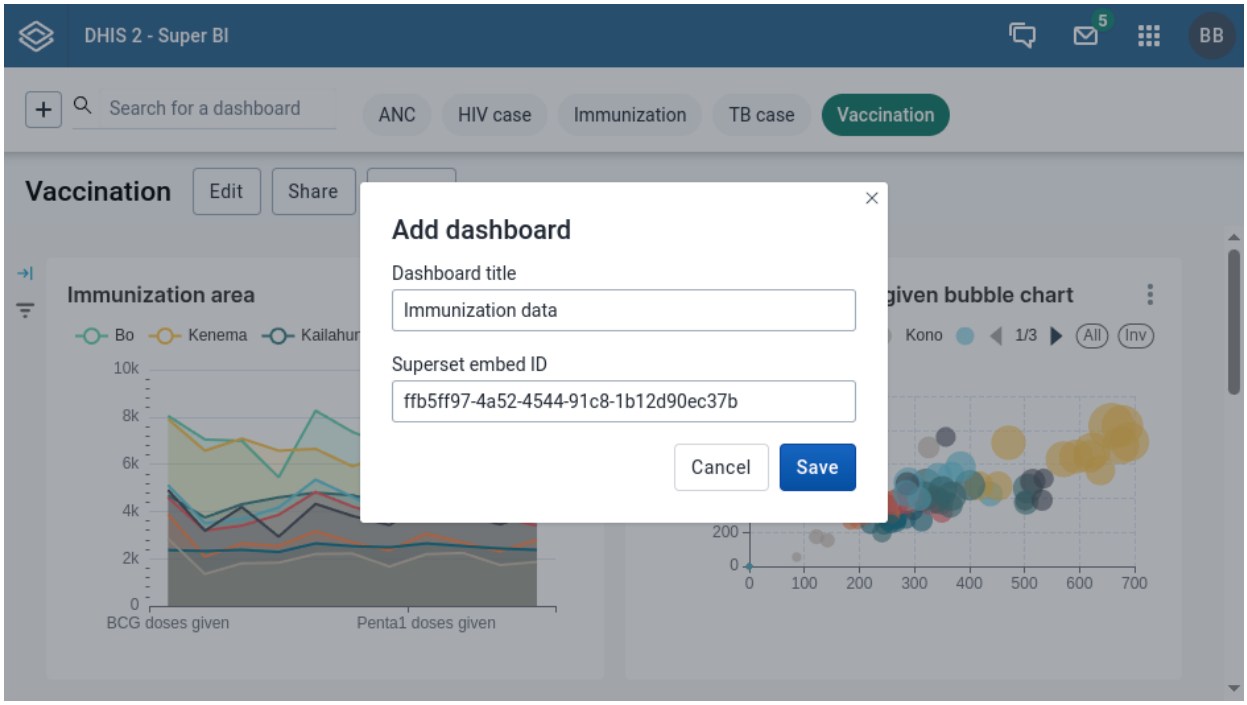

Figure 37: Super BI create dashboard

View Super BI dashboard

To view an existing Super BI dashboard, click the name of the dashboard in the top bar.

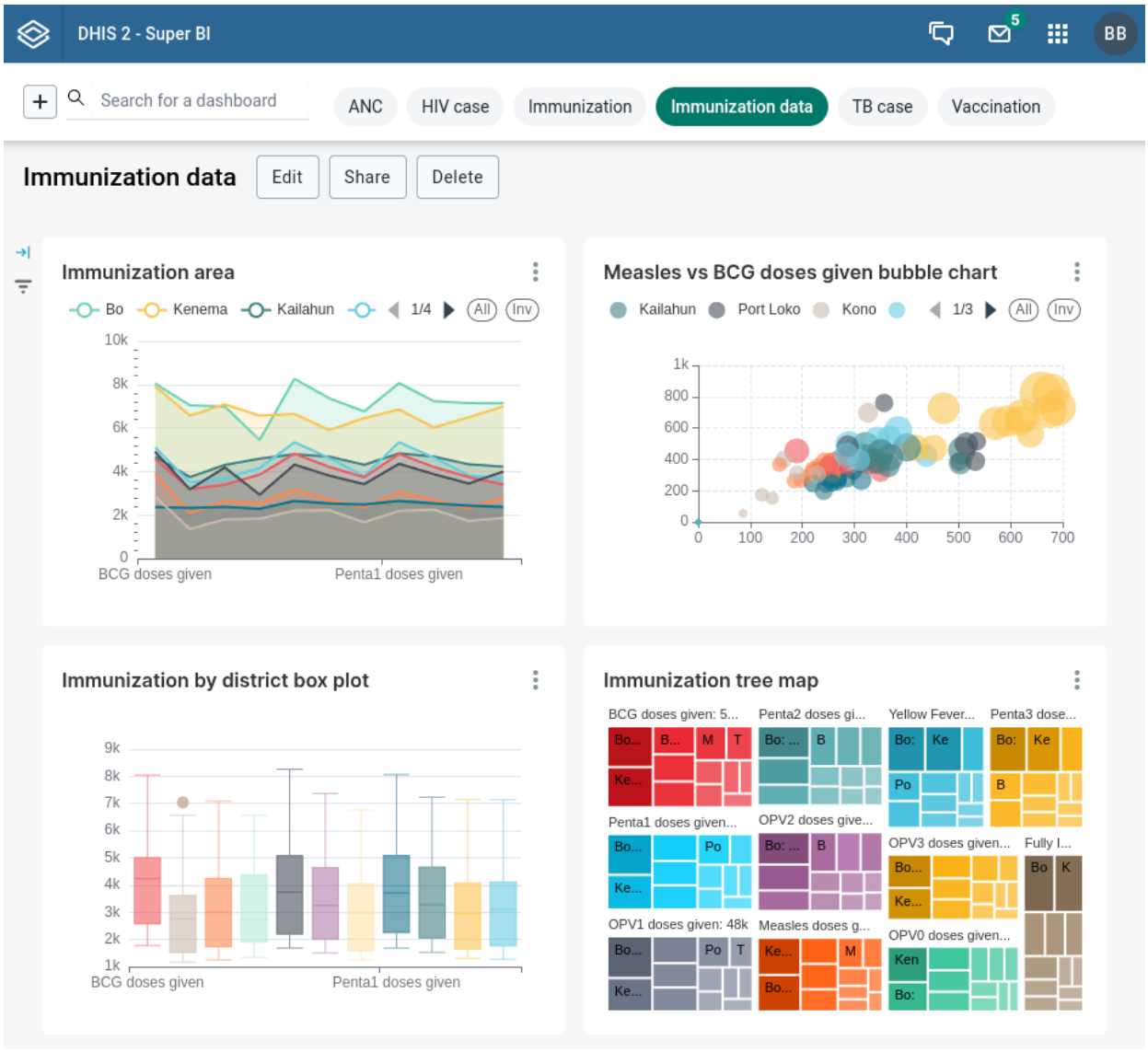

Figure 38: Super BI dashboard for immunization

Share Super BI dashboard

Super BI dashboards support the regular DHIS2 sharing access control model of DHIS2. This means that dasboards can be shared publicly, with user groups and users. View and edit permissions can be granted for each subject.

1. Select the dashboard from the top bar.
2. Click **Share** from the dashboard bar.
3. Share the dashboard with the appropriate subjects and permissions.
4. Click **Close**.

Remove Super BI dashboard

1. Select the dashboard from the top bar.
2. Click **Delete**.
3. In the confirmation dialog, click **Delete** again.

Solution

The Super BI web app and solution in AP provides several benefits with regards to data storage, access, visualization and analytics.

- The comprehensive data visualization capabilities of Apache Superset can be utilized within DHIS2.
- Users can utilize their existing DHIS2 account, removing the to introduce a new set of user accounts.
- Users will see and use a regular DHIS2 web app, minimizing the need for training.
- Access to the Super BI web app and dashboards is controlled with regular DHIS2 user roles.
- Data can be queried and processed directly in the high-performance AP data warehouse.
- Data can be analyzed without having to be loaded into the DHIS2 database.
- The DHIS2 data pipeline in AP provides near real-time access to new and update data.

Architecture

The following diagram describes the DHIS2 / Superset / AP architecture.

Examples

This section provides examples of Super BI dashboards.

Users

Overview

AP provides user and user group management.

Consult the *Sharing* section regarding object level access control.

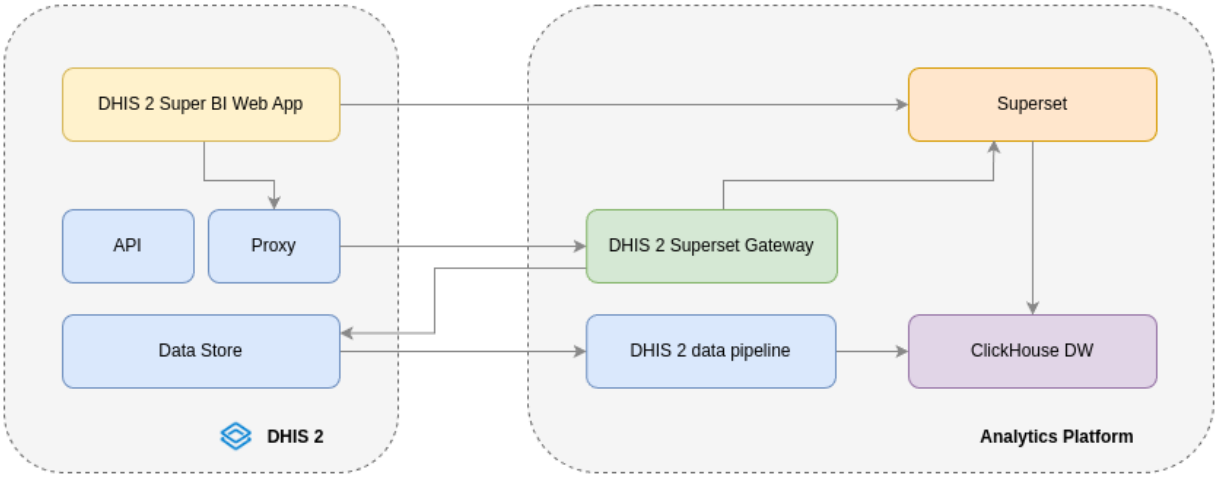

Figure 39: DHIS2 Superset architecture

Permissions

The authorization model in AP is based on granting user accounts individual permissions to perform actions in the platform.

Most permissions follow a *View* and *Manage* model.

- *View* means the ability to view information about objects of a specific entity.
- *Manage* means the ability to create new objects, edit existing objects and remove objects of a specific entity. The *Manage* permission includes the *View* permission, in other words, if a user is granted *Manage*, the user is implicitly granted *View* permission.

The following permissions are supported.

Admin

Permission	Description
Super Admin	Perform all actions in the system (super-user)

Analytics Platform

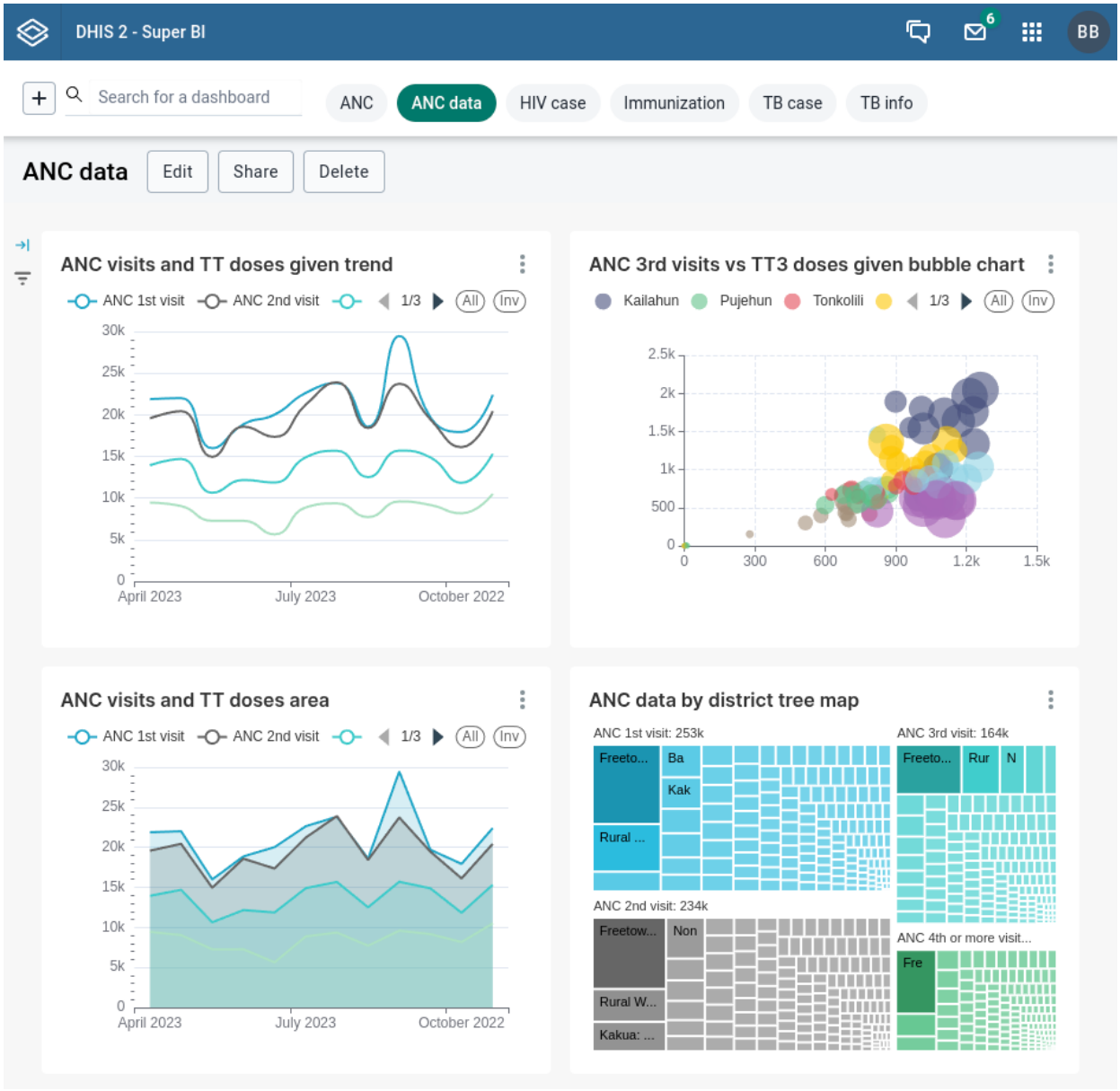

Figure 40: Super BI dashboard for ANC

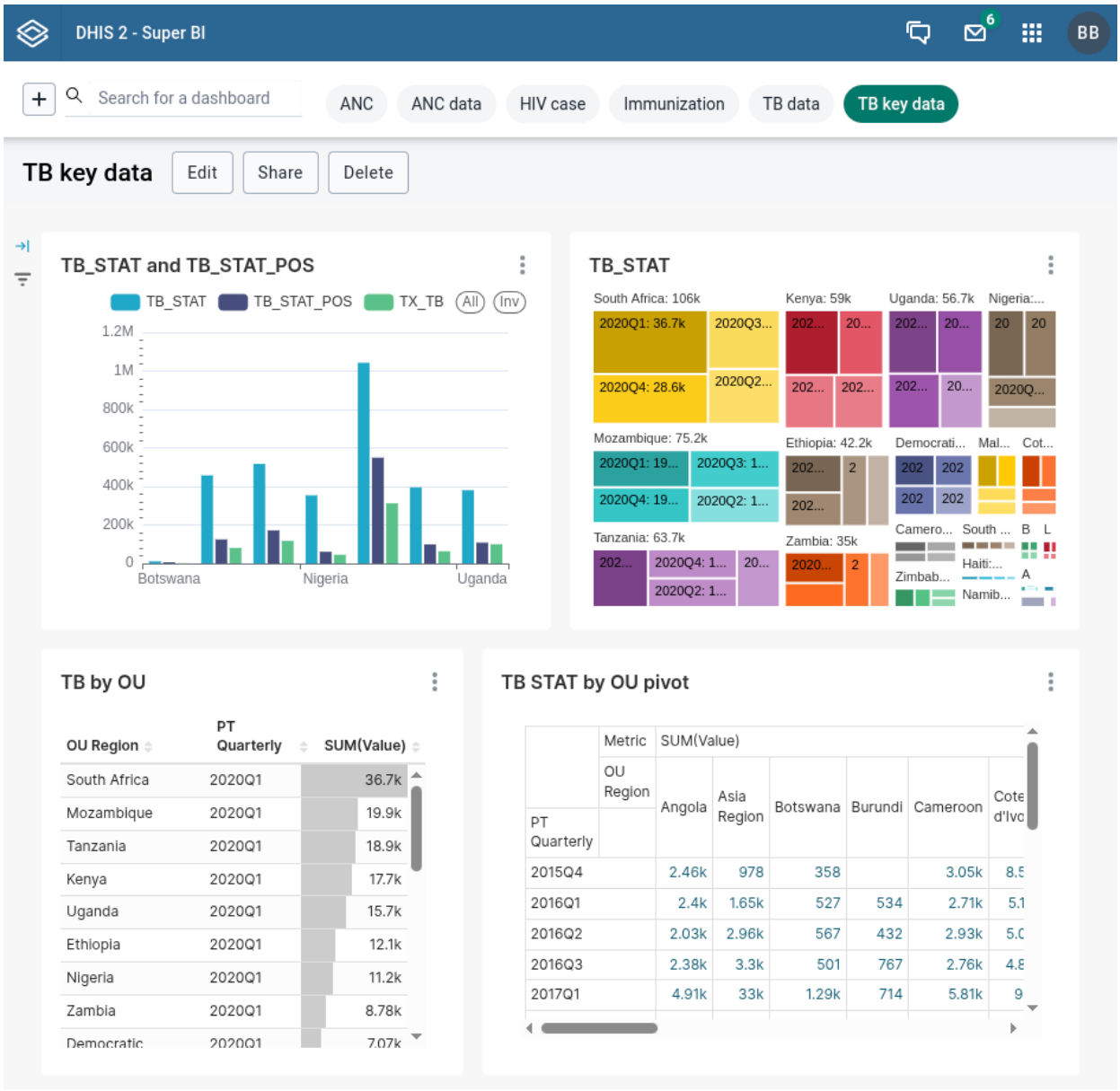

Figure 41: Super BI dashboard for TB

Permission	Description
Access to Analytics Platform	Access to AP and a corresponding user account is created in the AP data warehouse
View data for all data pipelines	Whether the data warehouse user account can view all data tables
Data pipelines	View or manage data pipelines
Schemas	View or manage schemas
Variables	View or manage variables
Settings	Manage settings
Views	View or manage views
Data quality checks	View or manage data quality checks
Data quality check groups	View or manage data quality check groups
Firewall rules	Manage firewall rules
Workflows	View or manage workflows
Destinations	View or manage destinations

Users

Permission	Description
Users	Manage users
User groups	Manage user groups

Managing users

The following section covers how to view, create, update and remove users.

View user

1. Click **Users** in the left-side menu to list all users.
2. Click the name of a user to view more information.

Create user invitation

Users in AP are primarily created by sending an invitation to create a user account over email to the relevant person. This allows the person to type in their own password, avoiding the need to send the password with

out-of-band communication.

1. Click **Add new user**.
2. Enter the following information.

Field	Description
Name	The full name of the user (required)
Username	The username of thser (required and unique)
Email	The email address of the user (required)
Start page	The space to use as start page when the user logs in
Enable SSO	Whether to enable Single Sign-On for the user account
Permissions	Select the permissions to grant to the user

3. Click **Send invitation**.

Edit user

1. Find and click the user to edit in the list.
2. Find the section which contains the information to edit.
3. Click the edit icon in the top-right corner of the section.
4. Update the relevant fields.
5. Click **Save**.

Reset password

1. Find and click the user for which to reset the password in the list.
2. Click the context menu in the top-right corner.
3. Click **Reset password**.

Disable user

1. Find and click the user to disable in the list.
2. Click the context menu in the top-right corner.
3. Click **Disable**.

Remove user

1. Find and click the user to remove in the list.
2. Click the context menu in the top-right corner.
3. Click **Remove**.

Managing groups

The following section covers how to view, create, update and remove user groups.

View group

1. Click **Groups** in the left-side menu to list all user groups.
2. Click the name of a user to view more information.

Create group

1. Click **Add new group**.
2. Enter the following information.

Field	Description
Name	The name of the group (required and unique)
Code	The code of the group (required)
Description	A description of the group

3. Click **Add new group**.

Edit group information

1. Find and click the group to edit in the list.
2. Click the edit icon in the top-right corner of the group information section.
3. Update the relevant fields.
4. Click **Save**.

Add and remove user group members

1. Find and click the group to edit in the list.
2. Click the edit icon in the top-right corner of the users section.
3. Enter the search criteria for the user to add or remove as a member in the search input field.
4. Select or unselect the checkboxes next to the names of the users to add or remove.
5. Click **Done**.

Remove group

1. Find and click the group to remove in the list.
2. Click the context menu in the top-right corner.
3. Click **Remove**.

Terminology

Term	Description
AP	Analytics Platform (AP) is a software platform for data integration and advanced analytics.
Apache Superset	Apache Superset is an open-source data exploration and visualization platform designed to be intuitive and highly accessible for business intelligence purposes.
BI	Business Intelligence (BI) refers to the technologies, applications and practices used to collect, analyze, integrate, and present business information.
Data pipeline	A series of processing steps to move data from a source system into the AP. The AP implements "ELT" (extract, load, and transform) pipelines.
Data catalog	An inventory of the data warehouse's datasets (tables) as well as a dataset's metadata such as table name, description, data types.
Destination	Allows for users to make data available to downstream/destination systems.

Term	Description
Logical view	A SQL query which provides the instructions for creating a virtual table, where the table is not stored in the database.
Materialized view	A SQL view that is stored in the database as a table.
Schema	A schema provides a mechanism to organize objects such as tables and views.
SQL	Structured Query Language, a standardized programming language that is used to manage relational databases and perform various operations on the data in them.
Variable	Text-based placeholders that are proxies for secrets, such as as passwords and API tokens.