CAMBRIDGE SEMANTICS

5.2 Getting Started Guide

Last Updated: 4/10/2023

© 2023 Cambridge Semantics, Inc.

Table of Contents

2

Anzo 5.2 Getting Started Guide Home **5**

Anzo Concepts and Vocabulary **6**

Anzo Glossary 9

Introduction to the Anzo Application **12**

General Interface Elements 14

Onboard 17

Model 21

Blend 22

Access 24

Provenance 27

Connecting to a File Store **28**

Creating an Anzo Data Store **33**

Making a Basic Connection to AnzoGraph **36**

Tutorial: Building a Sample Solution from Scratch **38**

Prepare the Environment 39

Download the Sample Data 40

Create a CSV Data Source and Import the Sample Data 41

Ingest the Imported Data Set 44

Run the ETL Pipeline 46

Create and Activate the Movie Data Graphmart 47

© 2023 Cambridge Semantics, Inc.

Explore the Data in a Hi-Res Analytics Dashboard 50

Review the New Solution's Artifacts 61

© 2023 Cambridge Semantics, Inc.

Anzo 5.2 Getting Started Guide Home

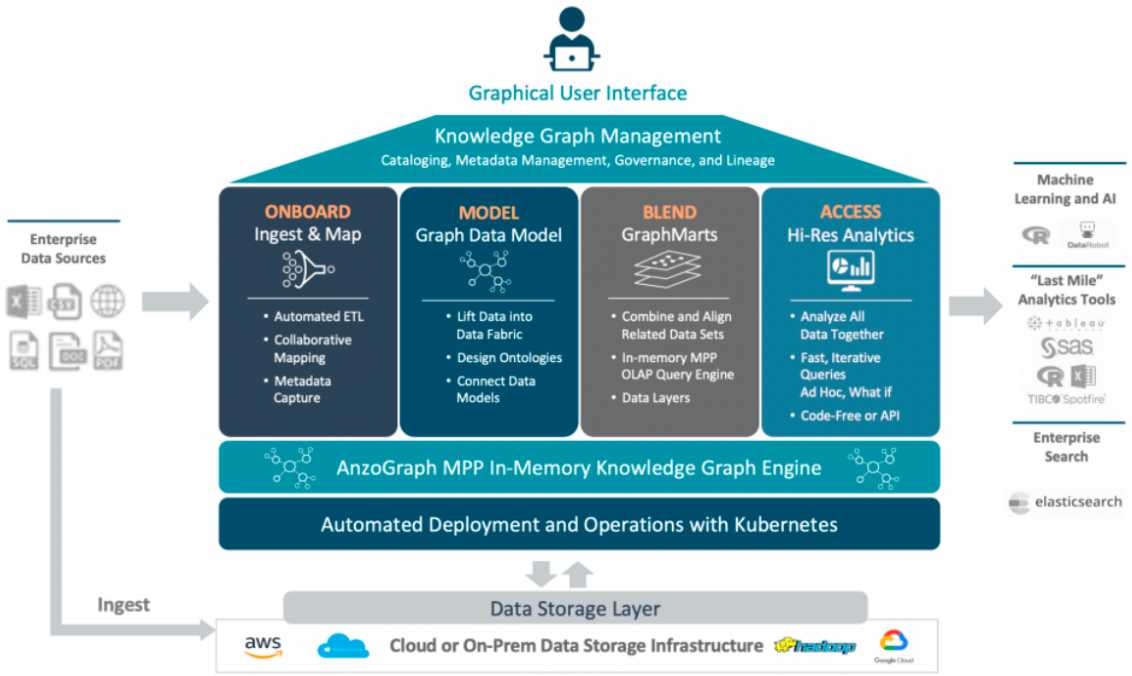
Welcome to the Anzo 5.2 Getting Started Guide! This guide helps you get started with Anzo. It includes an introduction to Anzo concepts and the user interface, describes basic setup information, and provides a tutorial that guides you through the process of building a sample solution from start to finish.

- Anzo Concepts and Vocabulary
- Introduction to the Anzo Application
- Connecting to a File Store
- Creating an Anzo Data Store
- Making a Basic Connection to AnzoGraph
- Tutorial: Building a Sample Solution from Scratch

© 2023 Cambridge Semantics, Inc.

Anzo Concepts and Vocabulary

This topic introduces you to key features, concepts, and vocabulary to know when working with Anzo. The diagram below shows a high-level overview of Anzo components and concepts. Details about the components in the image are described below, followed by a glossary that defines common Anzo terms and phrases.



Component	Description
Enterprise Data Sources	Anzo onboards data from many structured and unstructured data sources. Structured data sources such as relational databases or flat files are onboarded using Anzo's built-in pipelines. These pipelines natively support CSV, JSON, XML, and SAS files, along with all common database connections, including SQL, Oracle, MySQL, HIVE and others. Unstructured data sources, such as documents, PDFs, text snippets, web pages, and content from knowledgebases, are onboarded using configurable unstructured data pipelines. These pipelines onboard unstructured source files, integrate metadata about those files into the Anzo catalog, and make the full text in those files and key facts available as part of graph data models.
Data Storage Layer	The Anzo platform components, AnzoGraph, Spark, Elasticsearch, etc., share a file system for maintaining onboarded graph data and supporting files. Anzo supports storage systems such as HDFS, AWS S3, and NFS.

© 2023 Cambridge Semantics, Inc.

Component	Description
Metadata Catalog	Anzo's catalog is a special-purpose graph data model within Anzo. It combines traditional technical, operational, and business metadata with a semantic layer to describe all aspects of enterprise data elements. The catalog enables Anzo's unique use of semantics and graph models and is the system of record for data in Anzo. Anzo collects and generates metadata at every stage in the data discovery and integration process. Metadata in the catalog documents how data is converted during the onboarding process from its original format into a graph model. Subsequent data blending, transformation, and preparation steps are captured as additional metadata. Anzo also captures new metadata to describe all actions taken against data within Anzo. Anzo uses the metadata to enable users to visualize their data, understand business contexts, identify connections, and blend and prepare data.
Onboard	When data is onboarded from its source platform to Anzo, it is converted from its original format to a new format that describes the data as a graph data model. This format, Resource Description Framework (RDF), captures each data value and relationship. Anzo stores the converted RDF data in files that Anzo's catalog manages. RDF data, in Turtle (TTL) format, is efficiently laid out on disk for optimal loading into Anzo's in-memory graph engine, AnzoGraph. Mappings describe how data from source systems is transformed into Anzo's RDF format. These mappings can be automatically generated from the source system's schema or custom-defined to perform additional transformation steps as part of the onboarding process. At run time, Anzo converts these mappings into the code that is executed on the ETL engine.
Metadata Dictionary	Metadata Dictionaries enable users to automate, accelerate, and simplify the process of mapping data from enterprise data sources into semantic graph business models. A dictionary is a centralized repository of the concepts that link the logical business models to the physical schemas of the data sources that feed it. The dictionary structure becomes the basis for creating and reusing models and mappings across data sources.

© 2023 Cambridge Semantics, Inc.

Component	Description
Model	Anzo establishes the semantic layer by enabling users to convert diverse enterprise data models into graph data models and then enhance the data by adding new business definitions, names, and tags. Further insight is added when data from separate graph data maps are linked, connecting shared business definitions across previously siloed sources. Anzo employs open World Wide Web Consortium (W3C) standards, including Web Ontology Language (OWL), RDF, and SPARQL to model, connect, and query interconnected graphs.
Blend	When users select one or more data sets from the Anzo Dataset catalog to blend and access, Anzo loads the corresponding files from the file store into memory for rapid analysis and manipulation. Anzo delivers blending and access through “graphmarts” that give users the flexibility to combine and analyze any subset of data in Anzo.
Access	Once data has been onboarded, modeled, and blended into the dataset catalog and graphmarts, users have several options for accessing and analyzing the data. Anzo’s Hi-Res Analytics application enables users to create dashboards for exploring and visualizing the data without needing to have specialized query knowledge. The Query Builder provides access for finding specific statements or writing and running SPARQL queries. And the Anzo Data on Demand service provides access to data from business intelligence tools (see Last Mile Analytics Tools).
Graphical Application Interface	The graphical application interface is organized and compartmentalized by the concepts and processes described above. The design accommodates use cases where users with different permissions and responsibilities build various parts of a solution. For a tour of the graphical user interface, see Introduction to the Anzo Application.
Last Mile Analytics Tools	In line with its open standard architecture, Anzo graphmarts can be accessed using modern application program interfaces (APIs). In addition to using SPARQL-compliant query endpoints, Anzo offers standards-compliant Open Data Protocol (OData)-based REST data feed endpoints as part of its data on demand service.

© 2023 Cambridge Semantics, Inc.

Anzo Glossary

The table below defines commonly used Anzo terms and phrases.

Phrase	Description
Anzo for Office	Anzo for Office (AFO or A4O) is the Microsoft Excel plugin that enables you to create and edit source to target ETL mappings.
Anzo Data Store	An Anzo data store (previously known as a graph data source) defines an endpoint for writing data. It specifies the file store and directory on the file store where Anzo can generate file-based linked data sets (see File-Based Linked Data Set). It also defines write properties such as the maximum file size and whether files should be compressed.
AnzoGraph	AnzoGraph (AZG) is Anzo's in-memory massively parallel processing (MPP) graph OLAP engine.
Data Layers	Data layers enable users to enhance graphmarts dynamically. Users can create layers to load additional data sets, clean, conform, or transform data, infer new information, or export data to a file-based linked data set (FLDS).
ELT	In addition to traditional ETL, Anzo's data layers capability enables users to transform, blend, and prepare any data that has been added to the catalog into analytics-ready data sets using an extract, load, transform (ELT) flow. Data layers are Anzo's mechanism for flexibly transforming data in memory.
ETL	The extract, transform, and load (ETL) process takes source data and converts it to the graph data model using a source to target mapping. Anzo's mapping tool enables users to define field-level transformations, including type casting, date conversions, unit conversions, etc., as data is onboarded to Anzo.
File-Based Linked Data Set	When the onboarding process is complete, Anzo creates a data set in the Dataset catalog. The data set in the catalog is registered in the Anzo system data source (see Journal or Volume) and includes metadata about the data, including a pointer to the data store location for the RDF files generated by the ETL pipeline. The catalog data set and the files on disk are known as a file-based linked data set (FLDS).

© 2023 Cambridge Semantics, Inc.

Phrase	Description
File Store	A file store is the file storage system, such as NFS, HDFS, or cloud storage, that is shared between servers in an Anzo solution. Anzo, AnzoGraph, Elasticsearch, and other systems share data in a file store.
Graph Data Interface	AnzoGraph's Graph Data Interface is a SPARQL service that enables users to query data from external endpoints that are accessible over JDBC or HTTP. Information from the external sources can augment data in the Anzo Dataset catalog without having to onboard the data to Anzo.
Graphmart	Graphmarts are collections of Datasets that users can blend and enhance. Graphmarts can combine any subset of data in Anzo for analysis.
Hi-Res Analytics	Hi-Res Analytics enable users to explore and ask questions across all of their data. Using model-guided dashboards, users can perform computations across multi-dimensional data. Hi-Res Analytics dashboards generate complex graph queries dynamically based on user input.
IRI	An Internationalized Resource Identifier (IRI) is similar to URI but allows a greater range of characters. URI and IRI are often used interchangeably.
Journal or Volume	A journal, also known as a volume, refers to data that is stored in Anzo's embedded graph store. The graph store is transactional and is used to persist metadata, which is written to disk in a .jnl file. The system volume (or system data source) is the default, required volume where Anzo stores ontologies as well as system configuration, data set, catalog, registry, and access control metadata. Users can create secondary local volumes that are used for more compartmentalized data and can be created and deleted without affecting the core system.
Linked Data Set	A linked data set (LDS) is a fundamental concept. Anzo organizes all data, including system data, into linked data sets. An LDS is associated with a data model and can be searched, discovered, shared, and protected with access control. For example, graphmarts are organized in a linked data set or registry of graphmarts, pipelines are organized in a linked data set, the Activity Log is a linked data set, data source configurations exist in a linked data set, and so on.

© 2023 Cambridge Semantics, Inc.

Phrase	Description
NLP	Anzo onboards unstructured data using natural language processing (NLP) to find and extract data.
OData	Open Data Protocol (OData) facilitates the creation of interoperable RESTful APIs. The Anzo Data on Demand service provides OData-based feeds that can be used to query graphmart data from third-party business intelligence tools.
OSGi	The Open Service Gateway Initiative (OSGi) is the open-standard architecture upon which Anzo is built. It is a Java framework for developing and deploying software programs and libraries. OSGi enables Cambridge Semantics to compartmentalize Anzo into "bundles" that can be deployed, activated, and removed independently without affecting other bundles in the system.
Provenance	Anzo retains and displays the provenance of all onboarded structured data. The provenance explorer provides an overview of the relationships across various sources and models. Users can search for data entities and view associated pipelines, data sources, models, and schemas.
Registry	Anzo manages configurations in system-level registries. Each registry is a collection of application and system component configurations of the same type. Like data, registries are stored and managed with RDF named graphs according to ontologies. Technically, a registry is a Linked Data Set.
URI	A Uniform Resource Identifier (URI) is a globally unique identifier for a piece of information. A URL (Uniform Resource Locator) is a URI that specifies a location, such as a web address.

Related Topics

Introduction to the Anzo Application

© 2023 Cambridge Semantics, Inc.

Introduction to the Anzo Application

This topic provides instructions for accessing the Anzo user interface and gives a general overview of the design and layout. For information about the Anzo concepts and procedures that influence the user interface's design and use, see Anzo Concepts and Vocabulary.

To access the user interface:

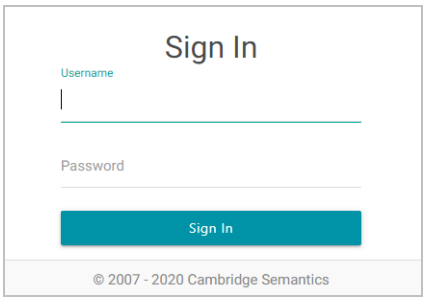
1. Go to the following URL in your browser:

   ```
   https://hostname
   ```

   Where **hostname** is the Anzo server DNS name or IP address.

   > **Note**
   > Your browser may warn you that the server's SSL certificate has not been signed by a trusted authority. This is normal behavior. To continue, click the **Advanced** link at the bottom of the page and then click the **Proceed** link. If you require a trusted site certificate, you can obtain one from a Certificate Authority and add it to the Anzo server.

2. On the Sign In screen, type your username and password and then click **Sign In**.



The user interface is organized and compartmentalized by the processes involved with building solutions. The design accommodates use cases where users with different permissions and responsibilities build various parts of a solution. The sections below introduce you to each of the user interface elements.

> **Note**
> The following sections show an administrator view of the user interface. Depending on their assigned system role, some users might not see all components. For information about default Anzo roles and permissions, see Predefined Anzo Roles and Permissions in the Anzo Deployment and User Guide.

- General Interface Elements
- Onboard
- Model

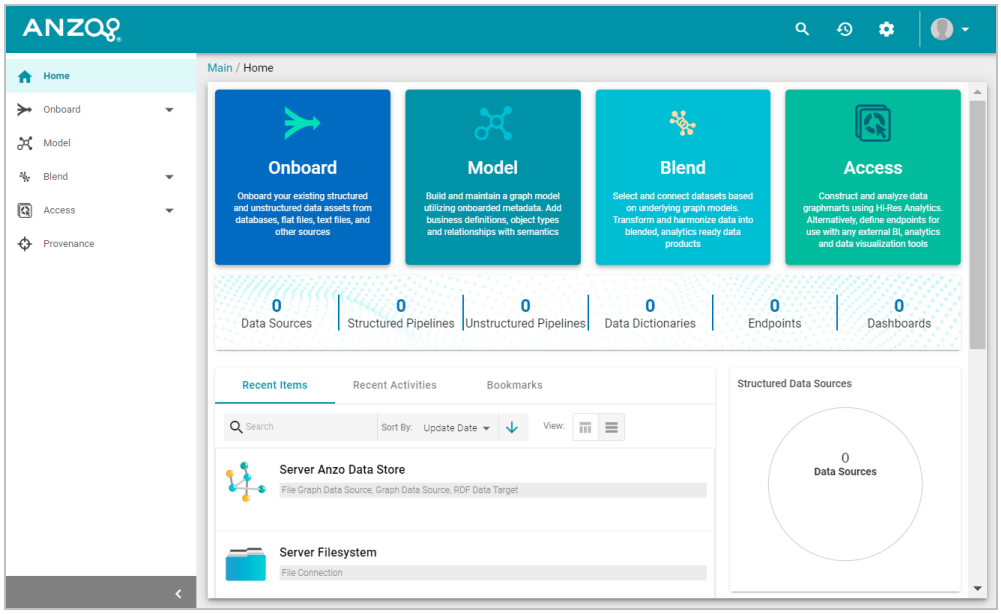
© 2023 Cambridge Semantics, Inc.

- Blend
- Access
- Provenance

© 2023 Cambridge Semantics, Inc.

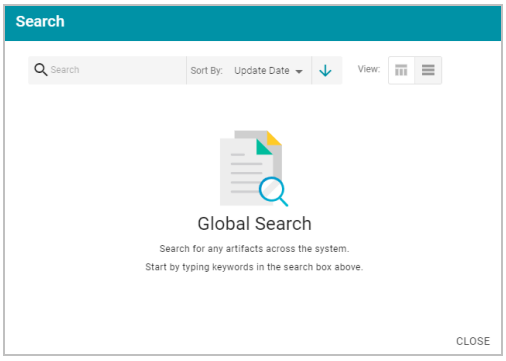
General Interface Elements

Each time you log into the application, the Home page is displayed and provides links to the features the logged in user has access to. Below the icons is a dashboard that displays an overview of the system artifacts and recently updated or bookmarked items. The dashboard becomes populated once you start onboarding Data Sources.



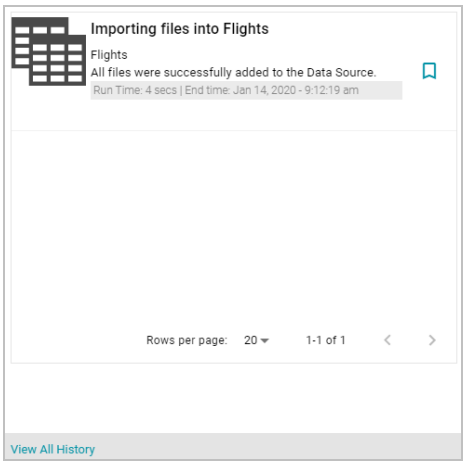
On the bottom left, you can click the collapse icon (<) to expand and collapse the left navigation menu. The right side of the top menu bar includes the following options:

- The Search icon ( ) opens the Search dialog box where you can do a global search for a resource, such as a Data Source, Dataset, Graphmart, or Schema, by title. The Search field accepts wildcard characters.

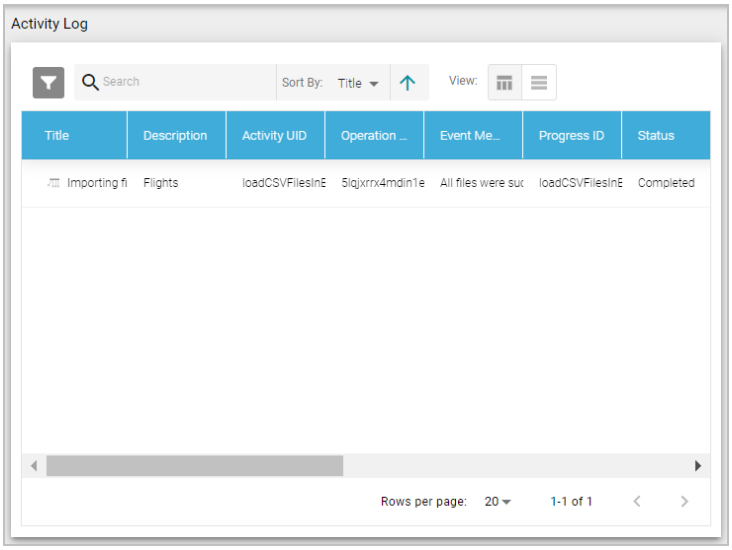


- The Activity Log icon ( ) opens the Activity Log, which offers a high-level view of individual activities, such as ETL engine and AnzoGraph usage and pipeline activity. The ability to filter activities based on their status and type enables users to drill down to activities of interest as well as profile user behavior. Additionally, the ability to include system data expands the scope of monitoring and aids in troubleshooting.

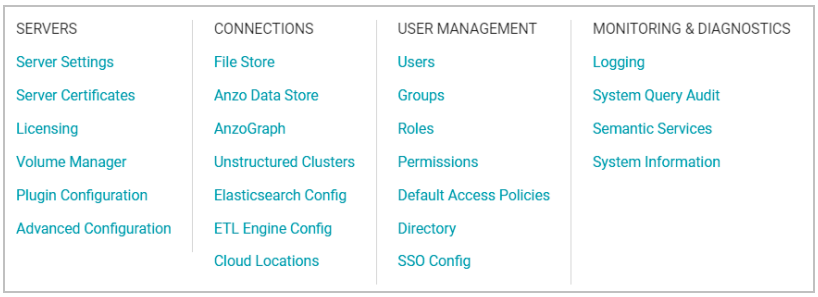
© 2023 Cambridge Semantics, Inc.



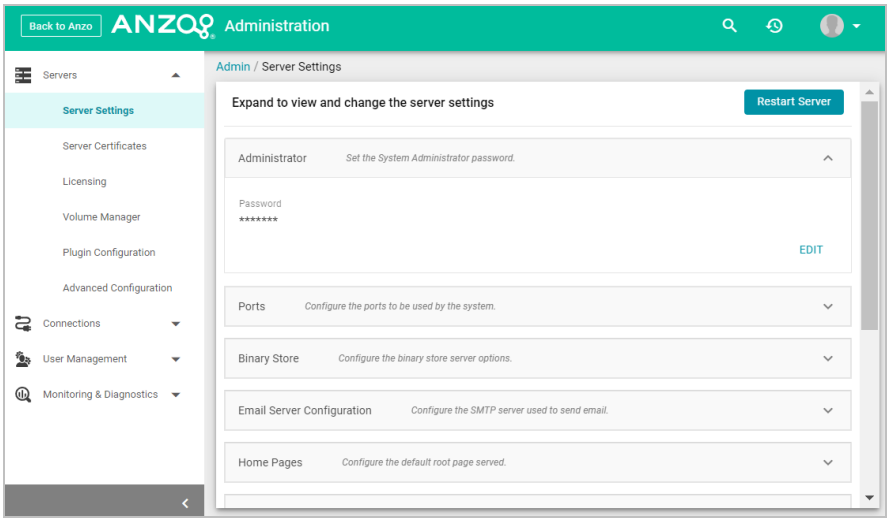
Once you start using Anzo and there are events that show up in the log, you can click **View All History** to open the full log for searching and filtering.



- The Administration icon ( ) opens the Administration menu. Clicking a menu option opens the Administration user interface for managing Anzo server settings, connections, users, and other components.

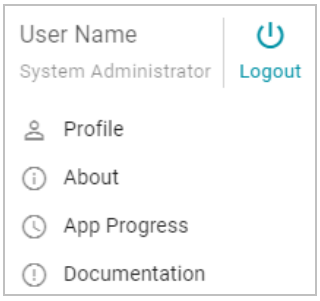


© 2023 Cambridge Semantics, Inc.

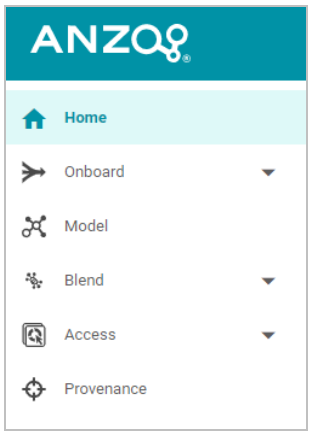


For information about the Administration application, see the Administration Guide in the Anzo Deployment and User Guide

- The User menu ( ) provides access to your user profile, the About screen, and the Application Progress window, which lists recent application activity. It also includes the ability to log out of the application and a Documentation link that opens this guide.



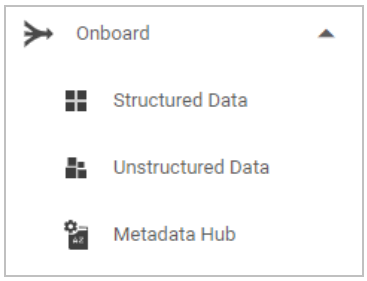
The navigation menu on the left provides access to all of the Anzo features.



© 2023 Cambridge Semantics, Inc.

Onboard

The Onboard menu provides access to the components that users configure to ingest data from various sources.

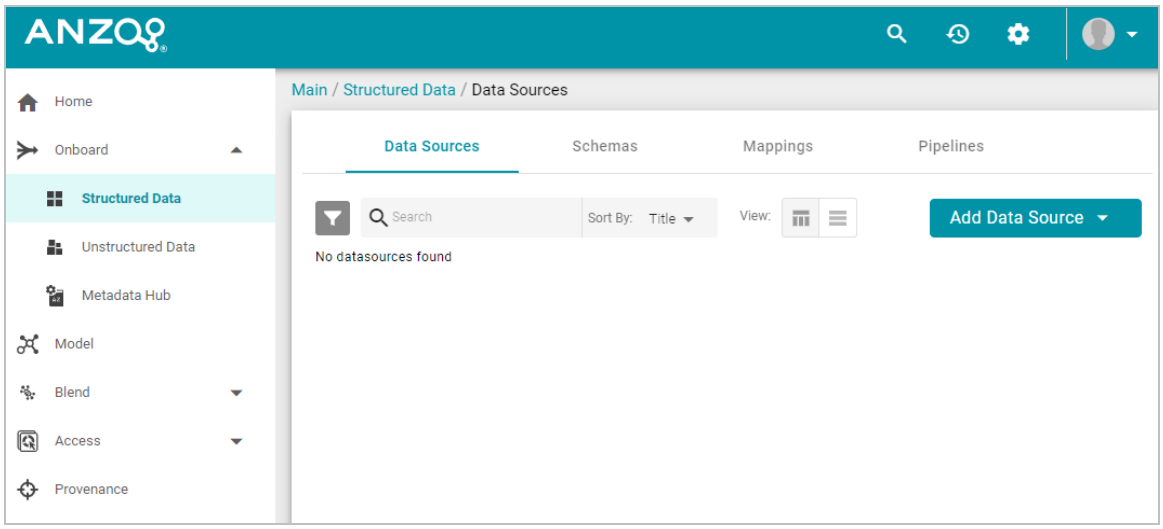


Structured Data

From **Structured Data**, users access Data Sources, Schemas, Mappings, and Pipelines for structured data sources: database connections, CSV, JSON, XML, Parquet, and SAS files.

Data Sources

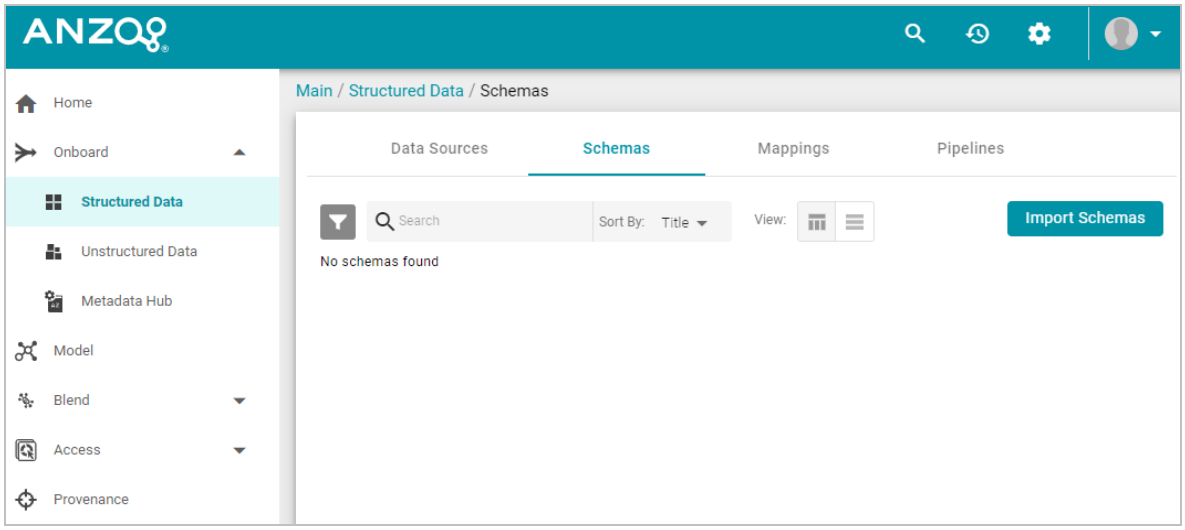
From the **Data Sources** tab, users connect to the files and databases that contain the data to onboard:



Schemas

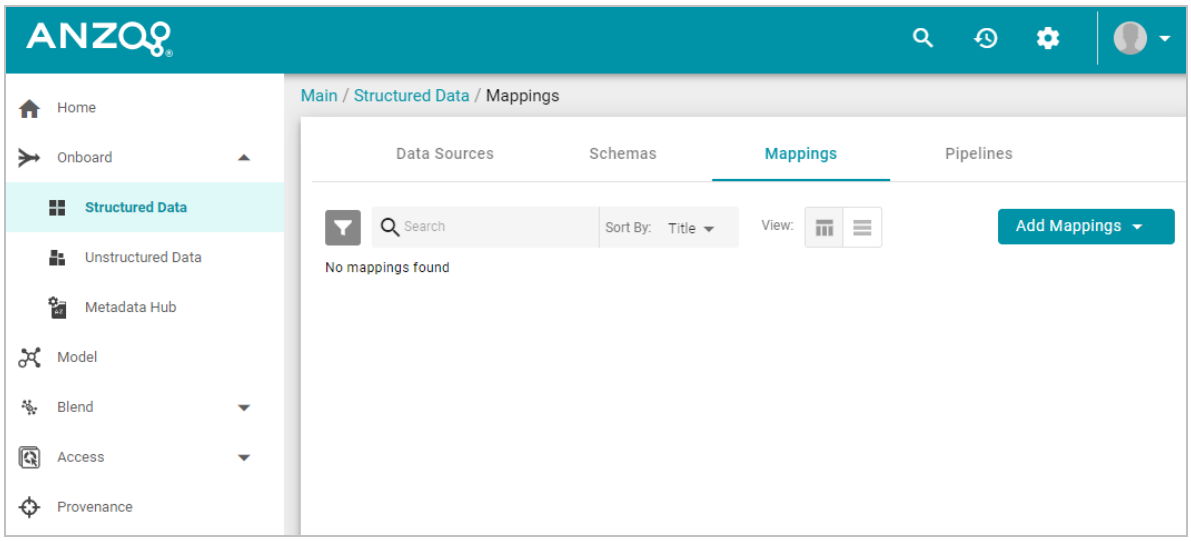
From the **Schemas** tab, users view, create, and manage the Schemas that define the data to onboard:

© 2023 Cambridge Semantics, Inc.



Mappings

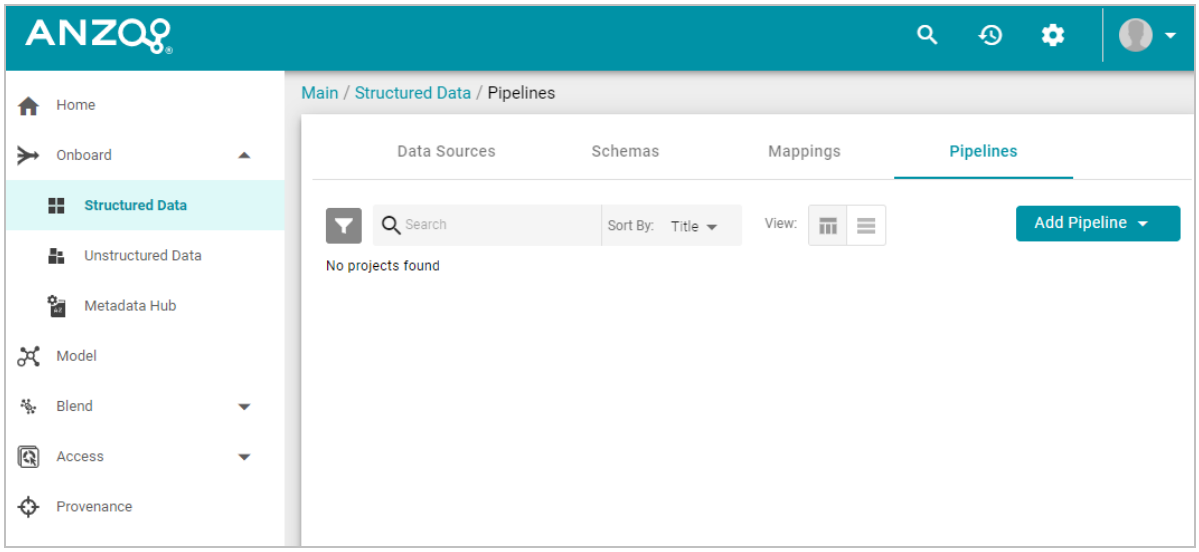
From the **Mappings** tab, users view, create, and manage the Mappings that describe the relationships between Schemas and Data Models as well as perform optional transformations on the source data:



Pipelines

From the **Pipelines** tab, users view, create, and manage the Pipelines that run ETL jobs to onboard data into Anzo:

© 2023 Cambridge Semantics, Inc.

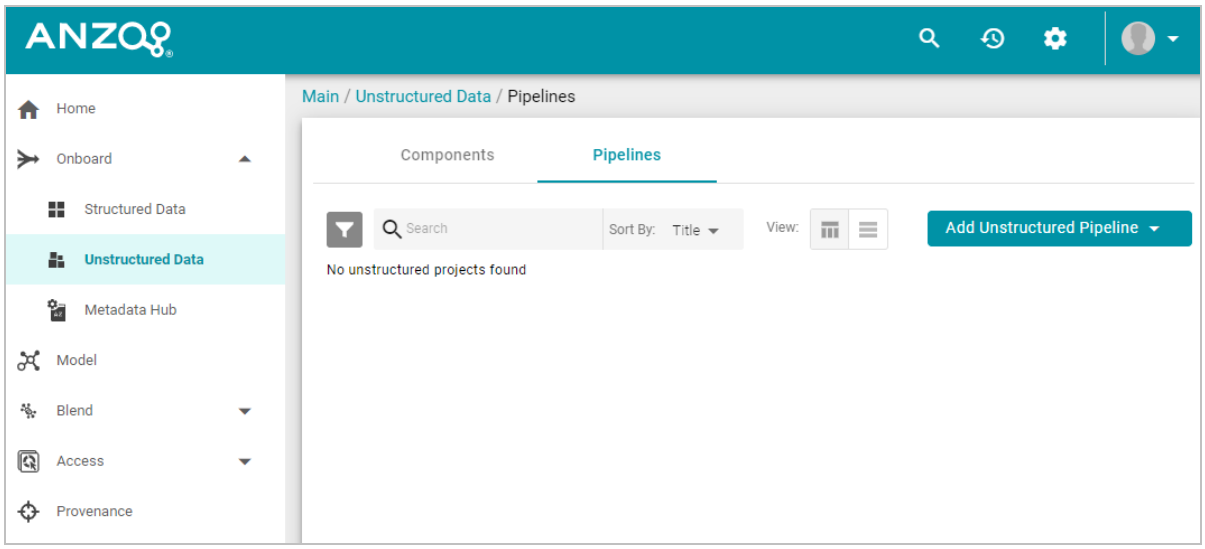


Unstructured Data

From **Unstructured Data**, users access and create Pipelines for onboarding data from Unstructured sources such as Office documents, PDFs, web pages, email messages, and knowledgebases.

Pipelines

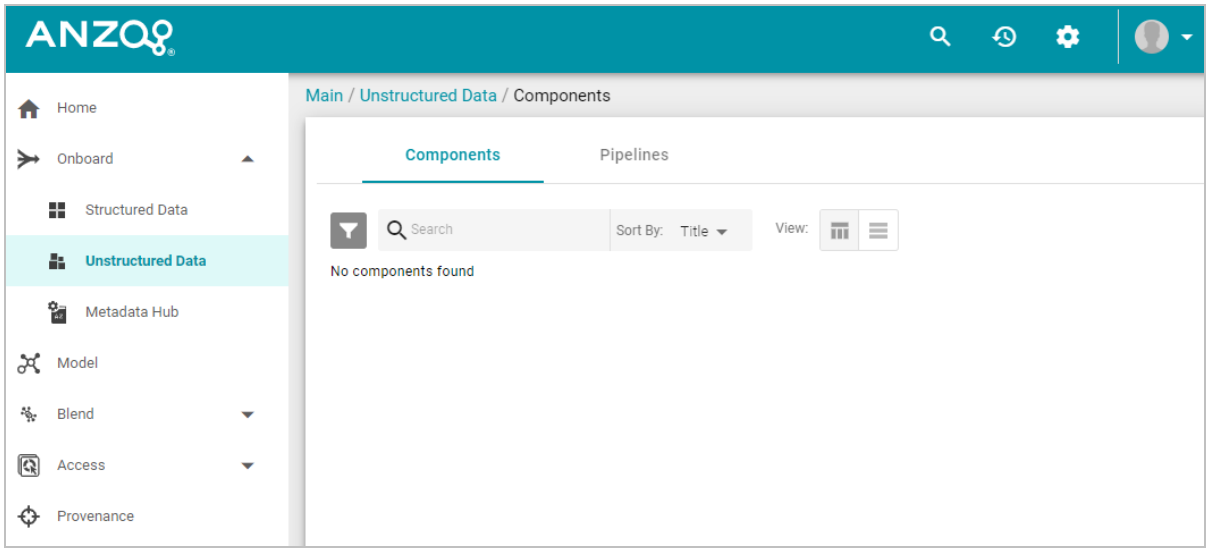
From the **Pipelines** tab, users can create and manage unstructured pipelines:



Components

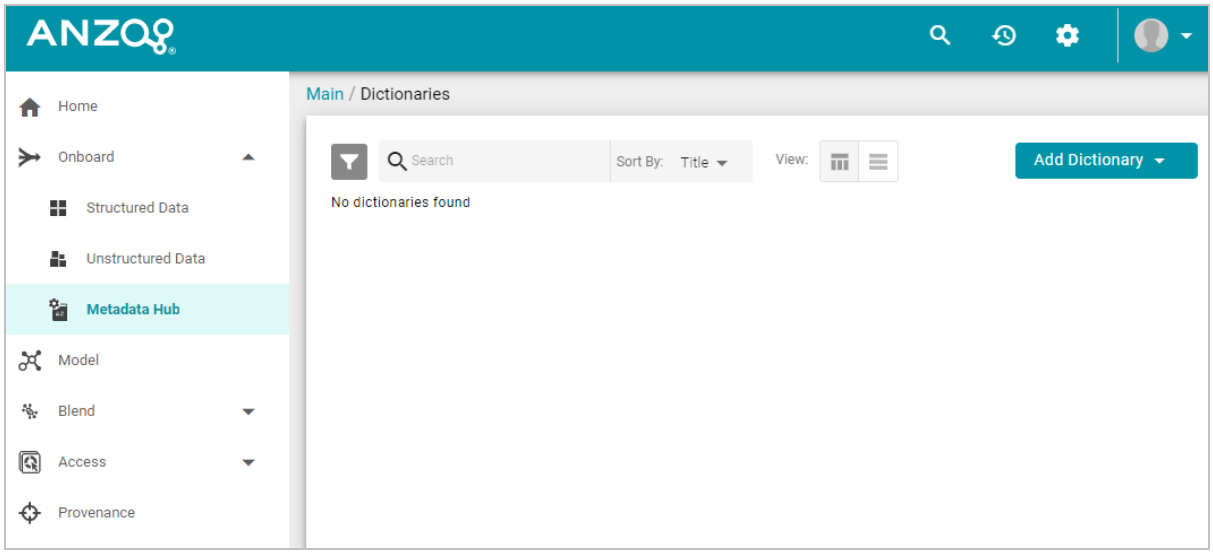
From the **Components** tab, users can view and configure Unstructured Pipeline components:

© 2023 Cambridge Semantics, Inc.



Metadata Hub

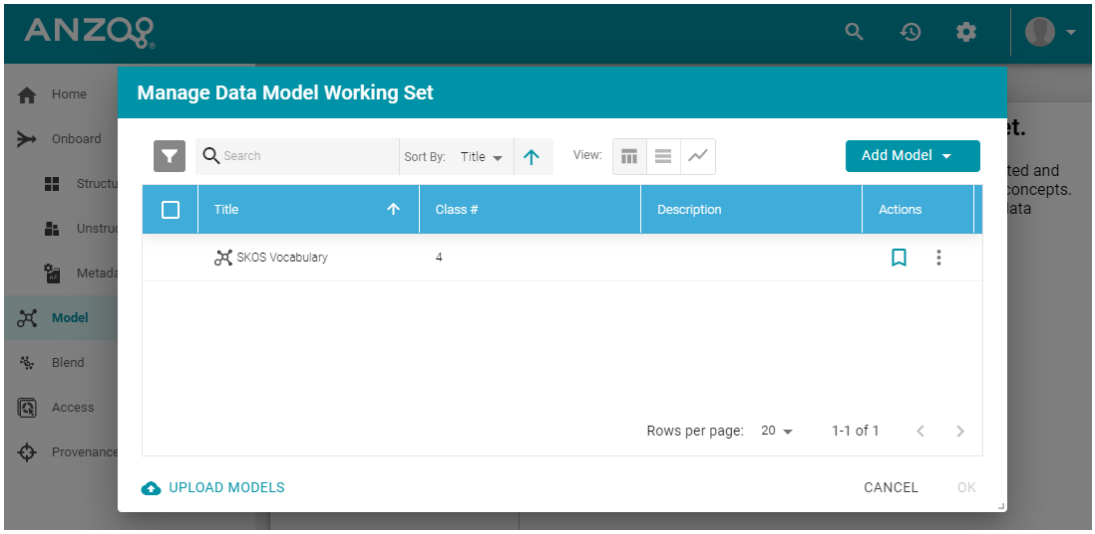
From **Metadata Hub**, users view, create, and manage Metadata Dictionaries, which define and normalize concepts across Data Sources.



© 2023 Cambridge Semantics, Inc.

Model

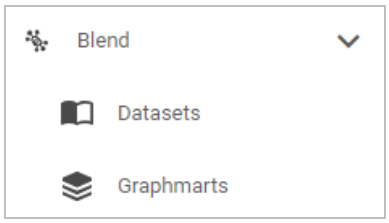
The **Model** manager enables users to view, create, and manage Data Models, which describe the concepts, attributes, and relationships in or across Datasets.



© 2023 Cambridge Semantics, Inc.

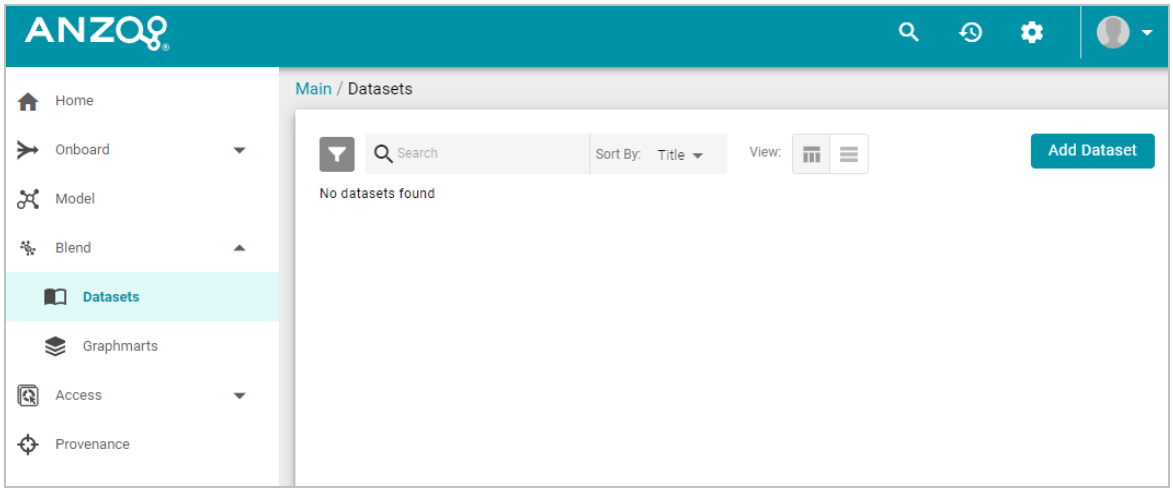
Blend

The **Blend** menu provides access to features that combine various datasets from different sources.



Datasets

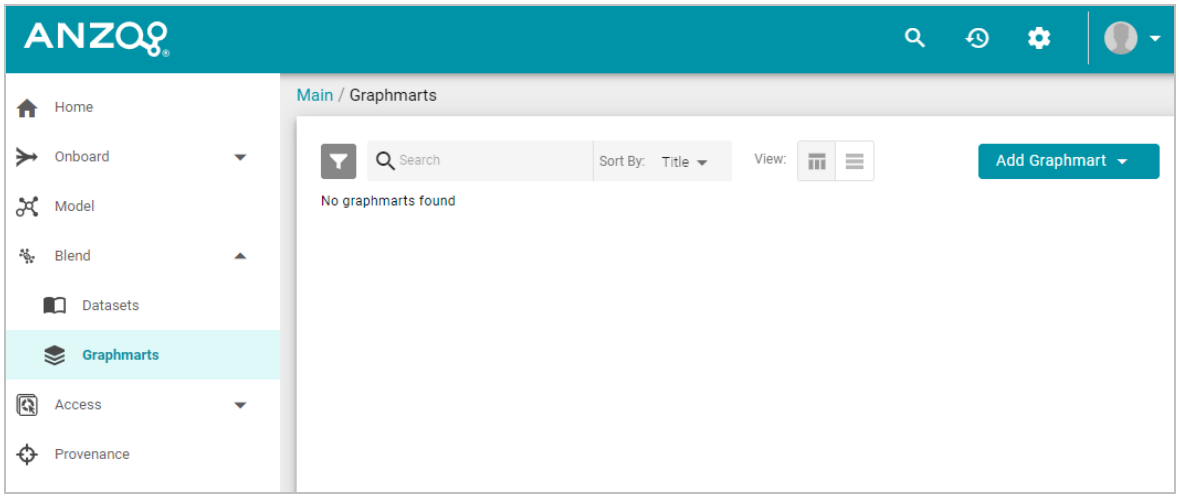
The **Datasets** catalog is an inventory of all of the data in Anzo. Users can sort the list by tags, classes, and Dataset creators. Users can also add Datasets to Graphmarts for loading into AnzoGraph and then designing Hi-Res Analytics dashboards.



Graphmarts

The **Graphmarts** page lists all of the existing Graphmarts. Users can designate favorites and sort the list by tags and Graphmart creators. Users can click a Graphmart in the list to view details such as the Datasets included in the Graphmart, Data Layer details, and associated dashboards.

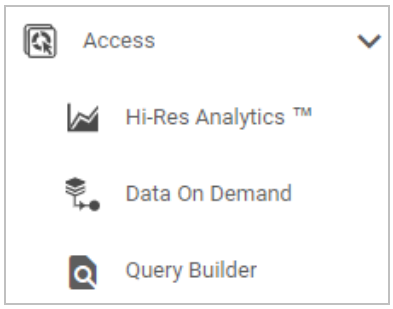
© 2023 Cambridge Semantics, Inc.



© 2023 Cambridge Semantics, Inc.

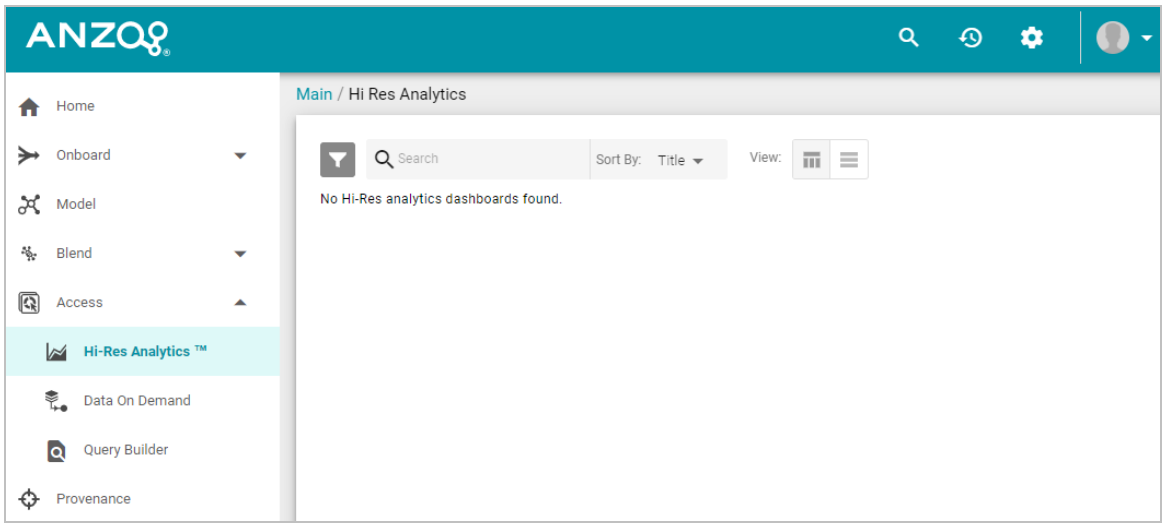
Access

The **Access** menu provides quick access to commonly used features.



Hi-Res Analytics

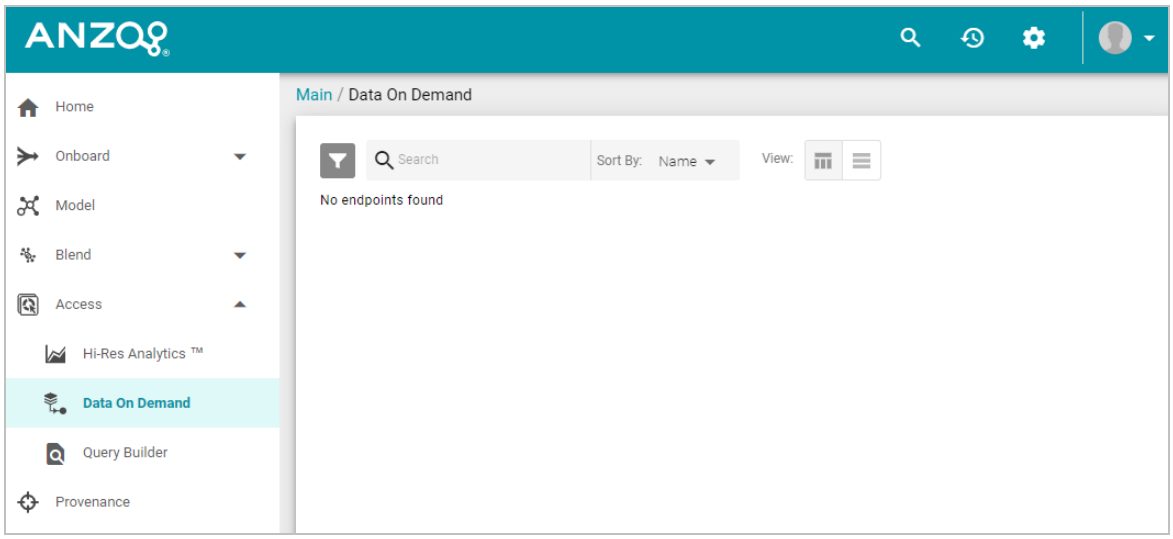
The **Hi-Res Analytics** page lists all of the existing dashboards. Users can click a dashboard in the list to open it in the Hi-Res Analytics application. Users can designate favorites and sort the list by tags, Graphmarts, and dashboard creators.



Data On Demand

The **Data On Demand** page provides a list of the Data on Demand endpoints that have been created.

© 2023 Cambridge Semantics, Inc.

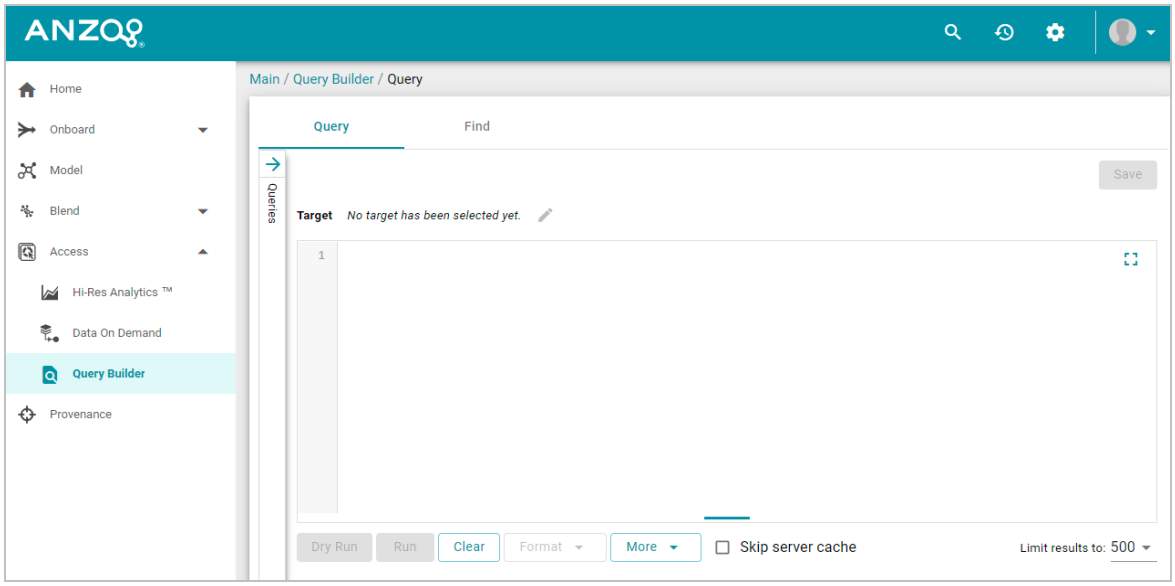


Query Builder

The **Query Builder** provides options to write and run SPARQL queries or find quads.

Query

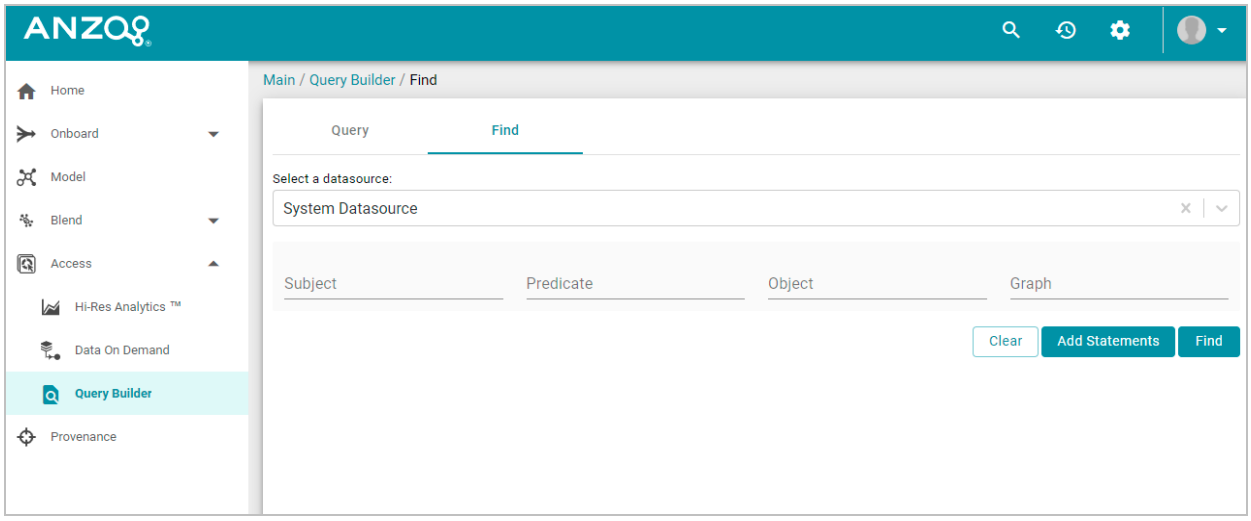
The **Query** editor provides syntax assistance, type-ahead suggestions for Model entity names, and automated prefix creation and query formatting for readability. It also includes the option to save queries for later use.



Find

The **Find** page enables users to search for quads by specifying a single subject, object, predicate, or graph name.

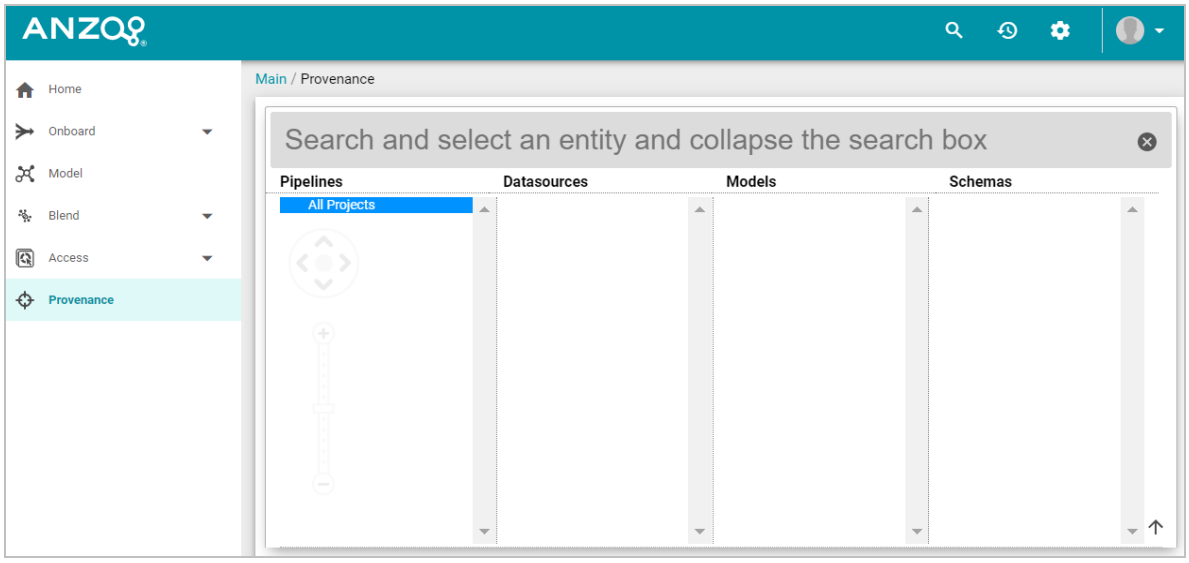
© 2023 Cambridge Semantics, Inc.



© 2023 Cambridge Semantics, Inc.

Provenance

The **Provenance** viewer enables users to view the lineage of all ingested structured data. The Provenance explorer provides an overview of the relationships across various sources and models. Users can search for data entities and view associated Pipelines, Data Sources, Data Models, and Schemas.



> **Note**
>
> By default, the Provenance viewer is available only to the sysadmin user. Whether you can access Provenance depends on whether the sysadmin user has enabled the feature for additional users or groups.

Related Topics

Connecting to a File Store

Creating an Anzo Data Store

Making a Basic Connection to AnzoGraph

Tutorial: Building a Sample Solution from Scratch

© 2023 Cambridge Semantics, Inc.

Connecting to a File Store

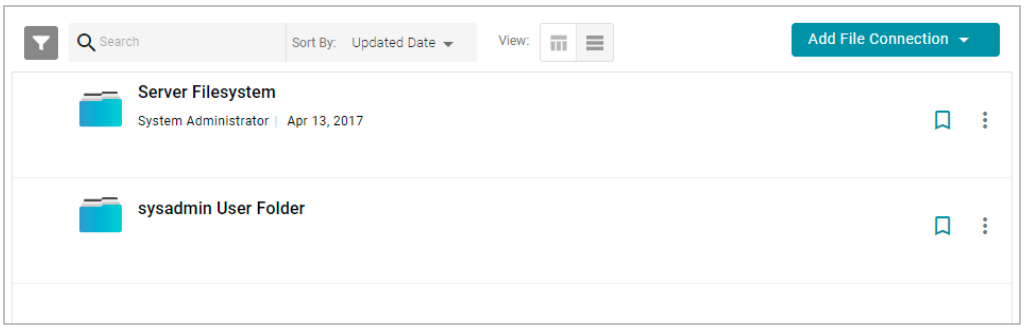
This topic provides instructions for connecting to the file system that all Anzo components will read from and write to during the onboarding processes. At least one file store needs to be shared between Anzo, AnzoGraph, and any Anzo Unstructured, Elasticsearch, or Spark servers. In almost all cases, organizations create an NFS to mount to all of the servers in the Anzo environment. Mounted file systems typically offer the best performance for reading and writing files. For more information, see Deploying the Shared File System in the Anzo Deployment and User Guide.

> **Note**
> The Anzo Server File System location is configured and accessible by default. If you store files on a storage system that is mounted directly onto the Anzo, AnzoGraph, Elasticsearch, Anzo Unstructured, and Spark servers, you are not required to configure that location.

Anzo supports reading from and writing to local or mounted file systems (such as NFS), Hadoop Distributed File Systems (HDFS), File Transfer Protocol (FTP or FTPS) systems, Google Cloud Platform (GCP) storage, and Amazon Simple Cloud Storage Service (S3).

> **Note**
> Administrator privileges are required to complete this task. Specifically, the **Manage File Stores** and **Administer System Setup** permissions are required.

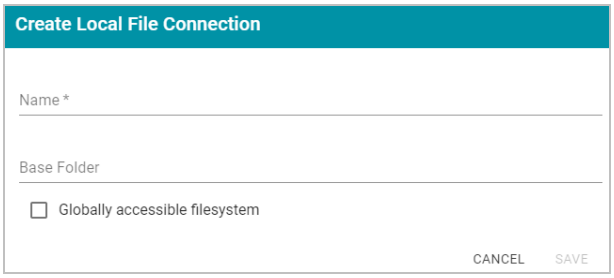
1. In the Administration application, expand the **Connections** menu and click **File Store**. Anzo displays the File Store screen, which lists existing file store connections. For example:



2. Click the **Add File Connection** button and select the type of file connection that you want to create. For the local disk or mounted NFS, choose **Local File Connection**. Anzo displays the create connection screen for the type of connection you chose.
3. On the connection screen, provide the file system details. The settings that display depend on the type of file connection that you chose. The list below describes the settings for each file connection type.

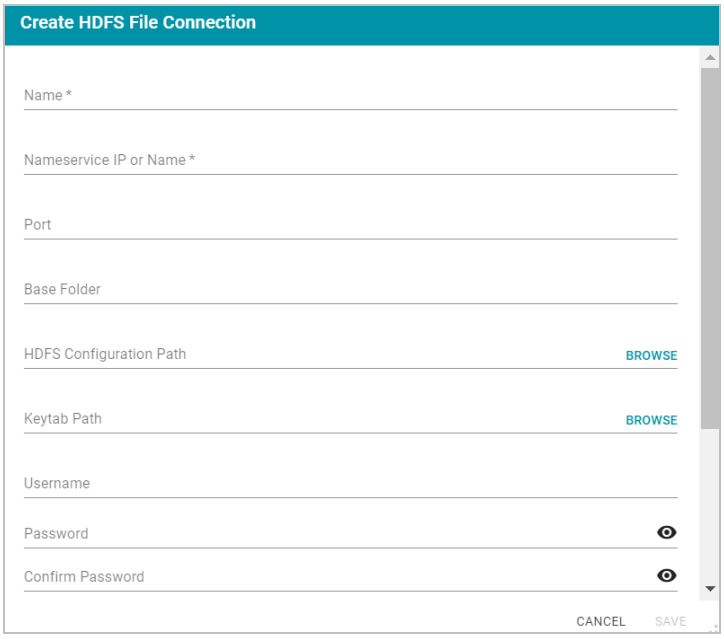
   Local File Connection

© 2023 Cambridge Semantics, Inc.



- **Name**: The name to use to describe this file connection within Anzo.
- **Base Folder**: The base or root folder on the file system where you want Anzo to either read or write files. Each time Anzo generates new files it creates a new subdirectory under this base location.
- **Globally accessible filesystem**: Select this option if this file store is accessible by all of the servers in an AnzoGraph cluster. If only the AnzoGraph leader server can access this system, leave this option blank.

HDFS File Connection



- **Name**: The name to use to describe this file connection within Anzo.
- **Nameservice IP or Name**: The IP address or host name for the storage system.
- **Port**: The RPC port to access the server on. The default RPC port is 8020.
- **Base Folder**: The base or root folder on the file system where you want Anzo to either read or write files. Each time Anzo generates new files it creates a new subdirectory under this base location.
- **HDFS Configuration Path**: Enter the full path to the configuration files.
- **Keytab Path**: The full path to the keytab file.
- **Username**: The user name for the account used to access the server.
- **Password** and **Confirm Password**: The password for the account used to access the server.

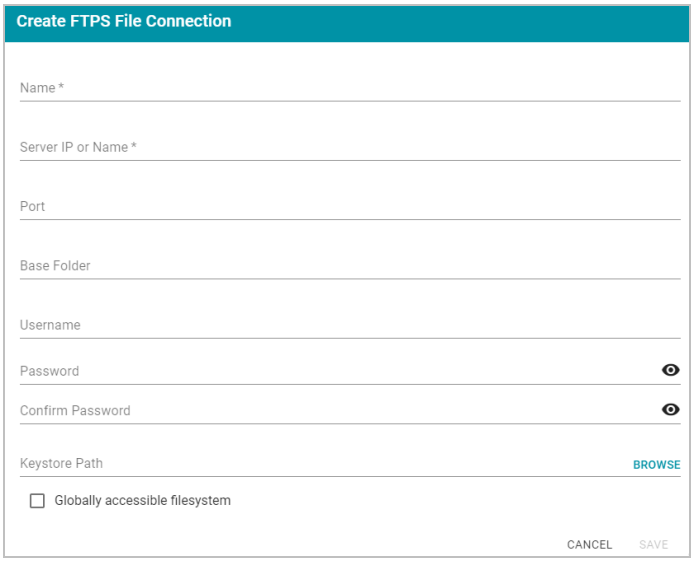
© 2023 Cambridge Semantics, Inc.

- **Nameservice Rest IP or Name**: The HTTP REST IP address or host name. Typically this value is the same as the Nameservice IP or Name.
- **Nameservice Rest Port**: The HTTP port. AnzoGraph uses this port to access HDFS and load the FLDS. The default HTTP port for the namenode is 9870.
- **Nameservice Rest Protocol**: The protocol to use for requests. Specify one of the following values:
  - **hdfs**: Specify **hdfs** for non-secure HTTP protocol.
  - **shdfs**: Specify **shdfs** for secure HTTPS protocol.
  - **khdfs**: Specify **khdfs** for non-secure HTTP protocol with Kerberos authentication.
  - **kshdfs**: Specify **kshdfs** for secure HTTPS protocol with Kerberos authentication.

> **Important**
> If you use Kerberos Authentication with HDFS, you must also configure your AnzoGraph cluster to authenticate with Kerberos. For instructions, see Configuring AnzoGraph for Kerberos Authentication in the Anzo Deployment and User Guide.

- **Globally accessible filesystem**: Select this option if this file store is accessible by all of the servers in an AnzoGraph cluster. If only the AnzoGraph leader server can access this system, leave this option blank.

FTP or FTPS File Connection

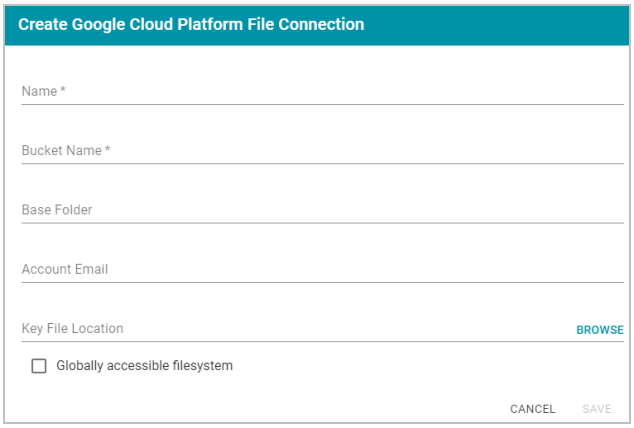


- **Name**: The name to use to describe this file connection within Anzo.
- **Server IP or Name**: The IP address or host name for the storage system.
- **Port**: The port to access the server on.
- **Base Folder**: The base or root folder on the file system where you want Anzo to either read or write files. Each time Anzo generates new files it creates a new subdirectory under this base location.

© 2023 Cambridge Semantics, Inc.

- **Username**: The user name for the account used to access the server.
- **Password** and **Confirm Password**: The password for the account used to access the server.
- **Keystore Path**: For FTPS connections, the full path to the keystore file.
- **Globally accessible filesystem**: Select this option if this file store is accessible by all of the servers in an AnzoGraph cluster. If only the AnzoGraph leader server can access this system, leave this option blank.

Google Cloud Platform File Connection

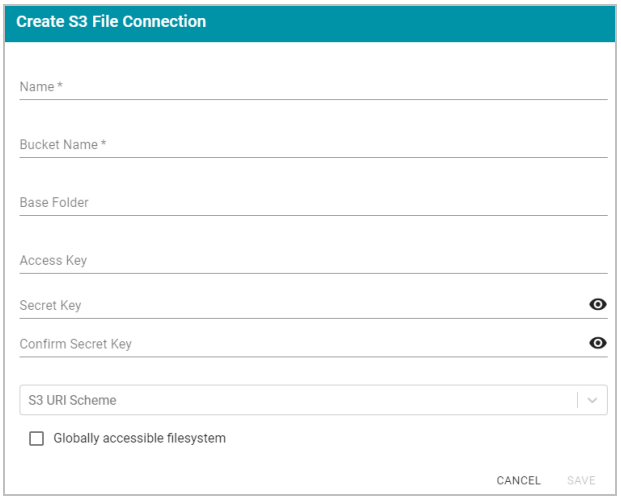


- **Name**: The name to use to describe this file connection within Anzo.
- **Bucket Name**: The name of the bucket to store files in.
- **Base Folder**: The base or root folder on the file system where you want Anzo to either read or write files. Each time Anzo generates new files it creates a new subdirectory under this base location.
- **Account Email**: The email address for the account used to access the storage.
- **Key File Location**: The full path to the keystore password file.
- **Globally accessible filesystem**: Select this option if this file store is accessible by all of the servers in an AnzoGraph cluster. If only the AnzoGraph leader server can access this system, leave this option blank.

S3 File Connection

> **Important**
> When using Amazon S3 for file storage, do not use client-side encryption, where data is encrypted before it is sent to Amazon S3. Anzo cannot read files on S3 if the object store uses client-side encryption.

© 2023 Cambridge Semantics, Inc.



- **Name**: The name to use to describe this file connection within Anzo.
- **Bucket Name**: The name of the bucket to store files in.
- **Base Folder**: The base or root folder on the file system where you want Anzo to either read or write files. Each time Anzo generates new files it creates a new subdirectory under this base location.
- **Access Key**: The Access Key ID to use for accessing the S3 location.
- **Secret Key** and **Confirm Secret Key**: The Secret Key ID for the Access Key.
- **S3 URI Scheme**: Specifies whether the URI scheme is S3, S3 Native, or S3A.
- **Globally accessible filesystem**: **Required**. Enable this option for S3 file stores.

4. Click **Save** to save the configuration. The file store connection that you specified becomes available as a choice when you create graph data stores or select source files to onboard.

See Creating an Anzo Data Store for instructions on designating a directory on the file store where file-based linked data sets and other files can be created during the ETL process (see File-Based Linked Data Set for more information).

Related Topics

Introduction to the Anzo Application

Creating an Anzo Data Store

Making a Basic Connection to AnzoGraph

Tutorial: Building a Sample Solution from Scratch

© 2023 Cambridge Semantics, Inc.

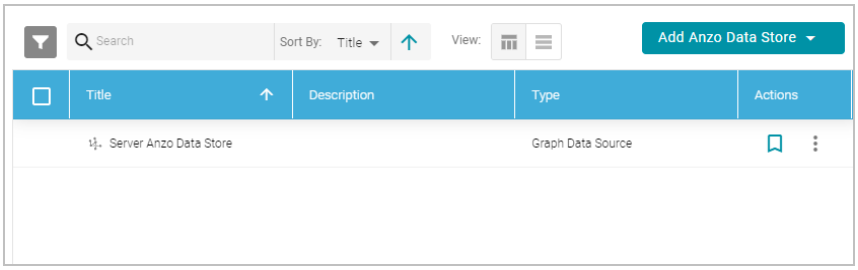
Creating an Anzo Data Store

This topic provides instructions for creating an Anzo Data Store (sometimes called a graph data source). Creating a Data Store means that you designate a directory on the File Store where file-based linked data sets and other files can be created and shared during the ETL process. All installations require at least one Data Store. You can create one Data Store and configure all pipelines to write to that store (each ETL run automatically creates a new sub-directory under the Data Store directory) or you can create multiple Data Stores to use for different Datasets.

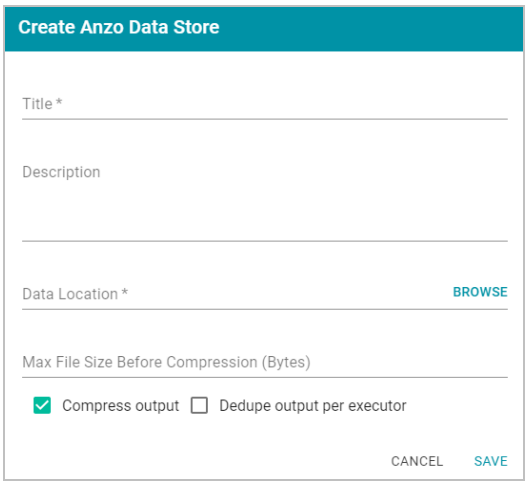
For information about setting up a connection to the shared file system that will host the Data Store, see Connecting to a File Store.

> **Note**
> Administrator privileges are required to complete this task. Specifically, the **Create Anzo Data Stores** and **Administer System Setup** permissions are required.

1. In the Administration application, expand the **Connections** menu and click **Anzo Data Store**. Anzo displays the Anzo Data Store screen, which lists any existing data stores. For example:



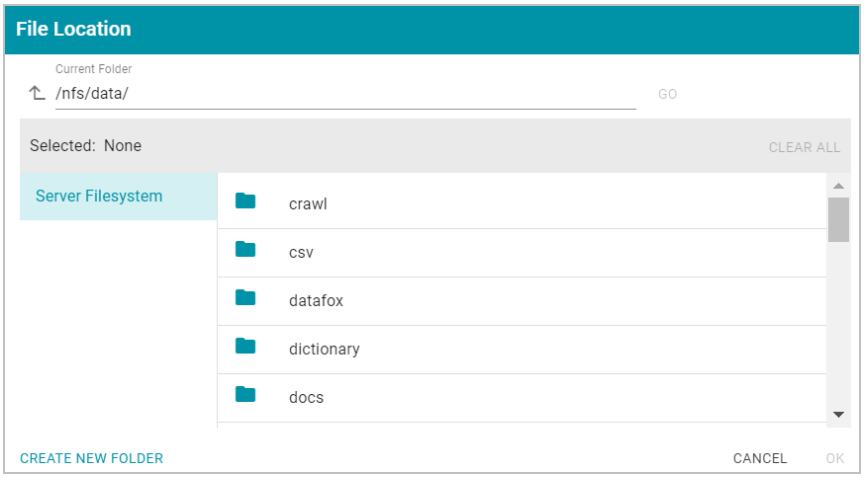
2. On the Anzo Data Store screen, click the **Add Anzo Data Store** button and select **Add Anzo Data Store**. Anzo opens the Create Anzo Data Store screen.



3. Type a **Title** and optional **Description** for the Data Store.

© 2023 Cambridge Semantics, Inc.

4. Click in the **Data Location** field. Anzo opens the File Location dialog box.



5. On the left side of the screen, select the File Store on which to create this Data Store. On the right side of the screen, navigate to the directory that you want to designate as the data location. Select a directory, and then click **OK**. Or click **Create New Folder** to create a new directory. Each time a pipeline is run for this Data Store, a new subdirectory is created under the specified data location.

> **Note**
> The Data Location needs to be a directory on the file store that is shared between Anzo, AnzoGraph, and any Anzo Unstructured, Elasticsearch, or Spark servers. If you want Anzo to generate files for this data store in one location and then load the files into AnzoGraph from another location, specify the file generation location in this field, and then specify the AnzoGraph load location in the **Alternate Data Location** field that is displayed on the Details screen after you save the data store.

6. Specify whether to compress the generated load files. By default, the **Compress output** checkbox is selected, indicating that Anzo generates .ttl.gz files when writing to this graph data source. If you clear the checkbox, Anzo generates uncompressed .ttl files. To preserve disk space and reduce read times when loading data into memory, Cambridge Semantics recommends that you accept the default configuration and compress load files.
7. The ETL engine does not remove duplicates by default when running Pipelines. If the source contains a significant number of duplicate entities, you have two options for deduplicating the data:
   - **Deduplicate the data during the ETL process**: To deduplicate the data while running the jobs that will generate this graph source, select the **Dedupe output per executor** option. Enabling the dedupe option limits the number of duplicates to one duplicate per executor node. For example, if the Spark configuration has 10 executor nodes, the resulting data set can contain a maximum of 10 duplicate entities.

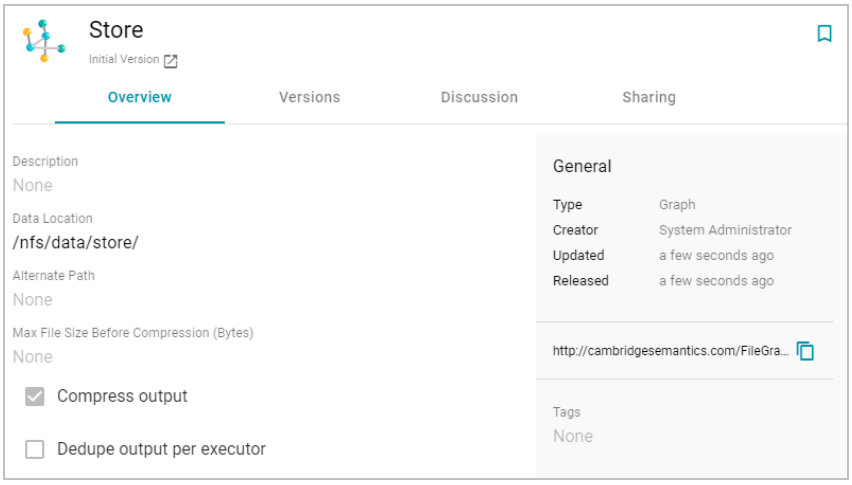
© 2023 Cambridge Semantics, Inc.

> **Important**
> Deduplication is based on primary keys and URI templates. If the source does not employ templating, do not enable the dedupe option. In addition, enabling this option substantially increases the time it takes to run the jobs for this data store.

- **Deduplicate the data after loading it to AnzoGraph**: AnzoGraph deduplicates data during a "vacuum" process that runs automatically after data is loaded into memory. If you leave the **Dedupe output per executor** option disabled, duplicates will be removed by AnzoGraph.

> **Note**
> Deduplicating data with AnzoGraph streamlines the ETL process but can increase load time and temporary memory usage in AnzoGraph during the load.

8. Click **Save** to create the data store. Anzo saves the store and displays the data store overview. For example:



   You can click a field to edit a value. Click the check mark icon (✓) to save changes to an option, or click the X icon (✕) to clear the value for an option.

9. If you plan to load files into AnzoGraph from a location that is different than the **Data Location** that you specified, edit the **Alternate Data Location** field and select the location for AnzoGraph load files.

See Making a Basic Connection to AnzoGraph for instructions on connecting Anzo to AnzoGraph for loading the RDF files that are generated when Pipelines are published.

Related Topics

Introduction to the Anzo Application

Connecting to a File Store

Making a Basic Connection to AnzoGraph

Tutorial: Building a Sample Solution from Scratch

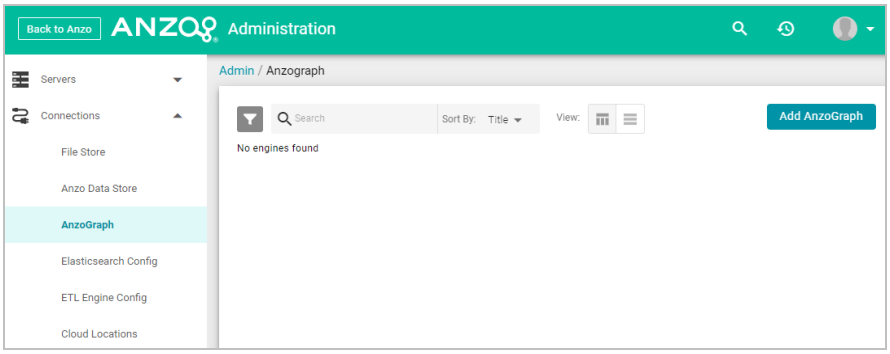
© 2023 Cambridge Semantics, Inc.

Making a Basic Connection to AnzoGraph

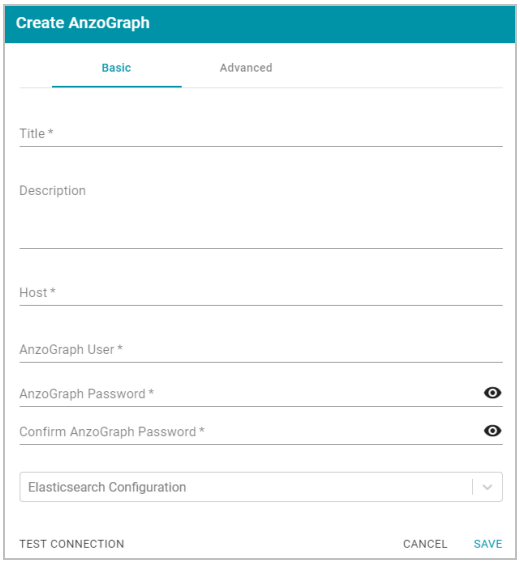
This topic provides instructions for quickly configuring a basic connection to AnzoGraph. For information about all of the AnzoGraph connection options, including the advanced options, see Connecting to AnzoGraph in the Anzo Deployment and User Guide.

> **Note**
> Administrator privileges are required to complete this task. Specifically, the **Manage AnzoGraph** and **Administer System Setup** permissions are required.

1. In the Administration application, expand the **Connections** menu and click **AnzoGraph**. Anzo opens the AnzoGraph connection overview screen, which lists any existing connections. For example:



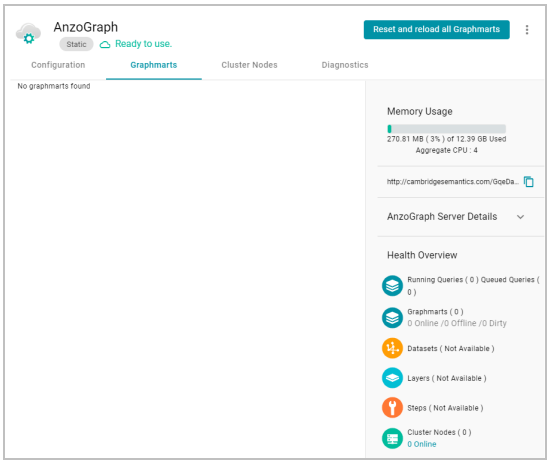
2. On the AnzoGraph screen, click **Add AnzoGraph** to add a connection. Anzo displays the Create AnzoGraph dialog box.



3. On the Basic tab, type a name for the engine in the **Title** field.

© 2023 Cambridge Semantics, Inc.

4. In the optional **Description** field, type a description for the graph query engine. If you leave this field blank, Anzo creates a description when you save the configuration.
5. In the **Host** field, type the AnzoGraph server host name or IP address. If you have a cluster, type the name or IP address of the leader server.
6. In the **AnzoGraph User** field, type the username that was created when AnzoGraph was installed.
7. Type the password for the AnzoGraph user in the **AnzoGraph Password** and **Confirm Password** fields.
8. Leave the **Elasticsearch Configuration** field unset. The Elasticsearch configuration is used with Anzo Unstructured.
9. Click **Test Connection** to check if Anzo can connect to AnzoGraph. If the connection fails, make sure that AnzoGraph is running and that you typed the correct username and password.
10. Click **Save** to save the configuration. Anzo connects to AnzoGraph and opens the Graphmarts tab. For example:



See Tutorial: Building a Sample Solution from Scratch for next steps and instructions for creating a sample solution.

Related Topics

Introduction to the Anzo Application

Connecting to a File Store

Creating an Anzo Data Store

Tutorial: Building a Sample Solution from Scratch

© 2023 Cambridge Semantics, Inc.

Tutorial: Building a Sample Solution from Scratch

This topic helps get you started with Anzo by teaching you the basic steps to follow when creating a complete solution. The instructions below provide a sample CSV file and guide you through the steps required from importing the file to visualizing the data in a Hi-Res Analytics dashboard.

1. Prepare the Environment
2. Download the Sample Data
3. Create a CSV Data Source and Import the Sample Data
4. Ingest the Imported Data Set
5. Run the ETL Pipeline
6. Create and Activate the Movie Data Graphmart
7. Explore the Data in a Hi-Res Analytics Dashboard
8. Review the New Solution's Artifacts

© 2023 Cambridge Semantics, Inc.

Prepare the Environment

To give the tutorial continuity and make it straightforward to follow, it helps to complete a few general setup steps before you start building the solution. This section provides background information and describes the environment preparation.

This tutorial guides you through importing data from a file, running an ETL pipeline to create RDF files for AnzoGraph, and loading the data into AnzoGraph. Completing those steps requires a File Store that Anzo and AnzoGraph can access, a designated location (Anzo Data Store) on that storage system to place the RDF files that the ETL process generates, and a connection to AnzoGraph for loading the data.

Confirm that the following setup steps are complete before proceeding:

1. AnzoGraph and AnzoGraph have access to the same File Store and its location is configured in Anzo. For instructions, see Connecting to a File Store.

   > **Note**
   > The Anzo Server File System is configured and accessible by default. If you store files on a file system that is mounted directly to the Anzo and AnzoGraph servers, you do not need to configure a new file system location.

2. An Anzo Data Store has been created for the RDF files that are generated during the ETL process. For instructions on creating a Data Store, see Creating an Anzo Data Store.
3. Anzo has a connection to AnzoGraph. For instructions on configuring the connection, see Making a Basic Connection to AnzoGraph.

© 2023 Cambridge Semantics, Inc.

Download the Sample Data

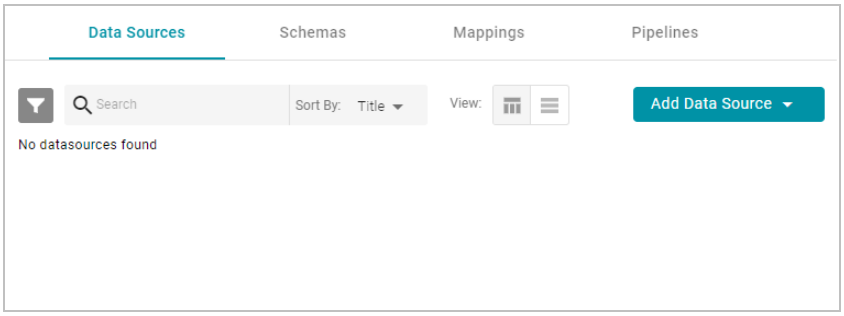
This tutorial uses a small sample data set from Data World. The data is in CSV format and includes IMDB movie data from 2006 – 2016. Click the link below to download IMDB-Movie-Data.zip to your computer. Then extract the ZIP file to make **IMDB-Movie-Data.csv** available for importing.

IMDB-Movie-Data.zip

© 2023 Cambridge Semantics, Inc.

Create a CSV Data Source and Import the Sample Data

1. In the Anzo application, expand the **Onboard** menu and click **Structured Data**. Anzo displays the Data Sources screen, which lists any existing data sources:

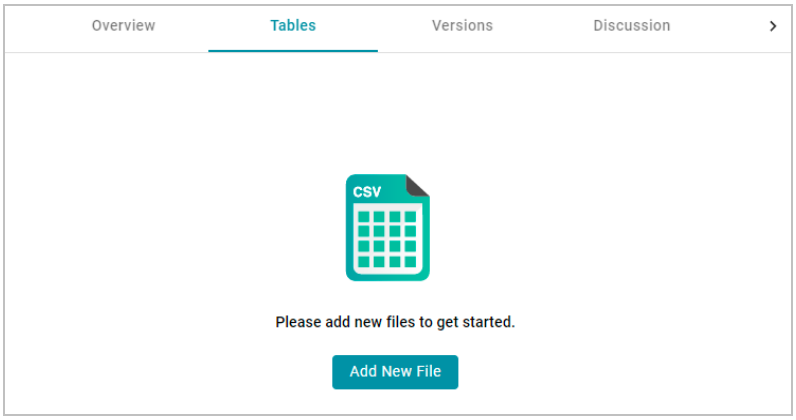


2. Click the **Add Data Source** button and select **File Data Source** > **CSV Data Source**. Anzo opens the Create CSV Data Source screen.

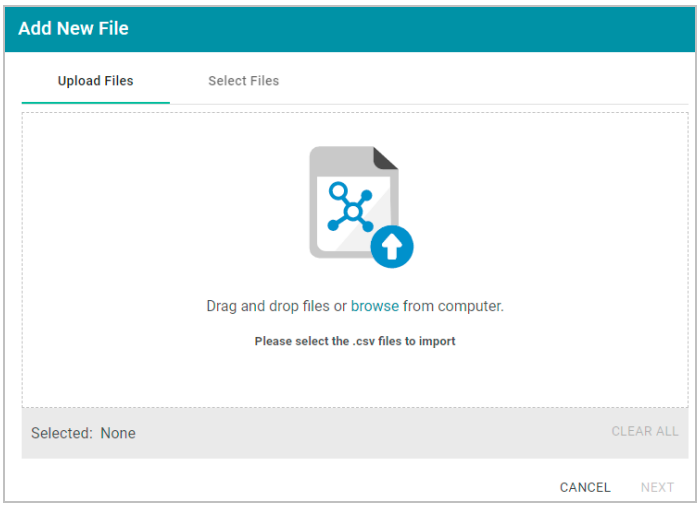
3. Type a name for the source in the **Title** field. For example, "Sample Movie Data." Then type an optional description in the **Description** field. For example:

© 2023 Cambridge Semantics, Inc.

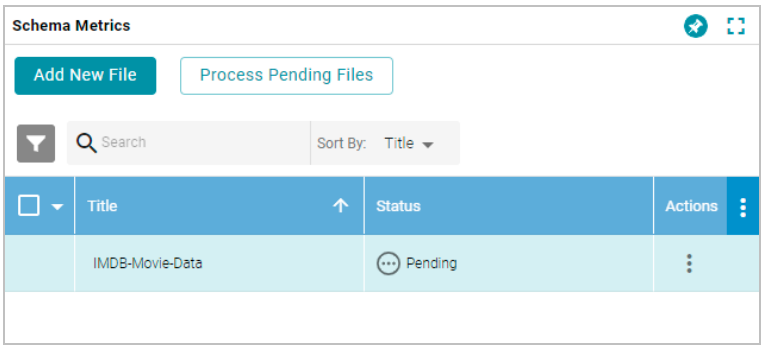
4. Click **Save**. Anzo saves the source and displays the Tables tab for the new Data Source.



5. On the left side of the screen, click the **Add New File** button. Anzo displays the Add New File dialog box, and the **Upload Files** tab is selected.



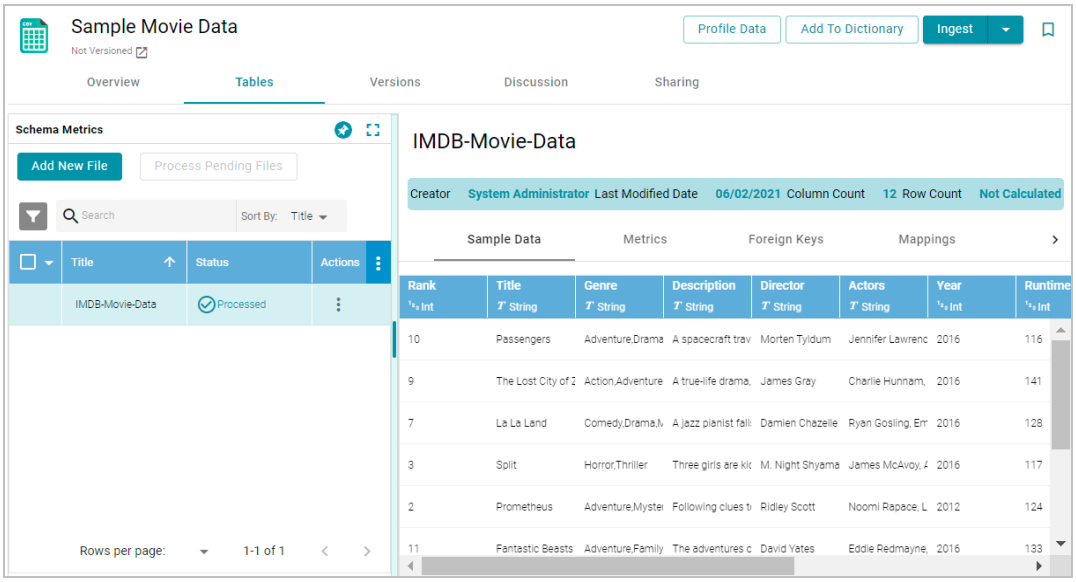
6. On the **Upload Files** tab, drag and drop the IMDB-Movie-Data.csv file onto the screen or click **browse** and navigate to the file and double-click it to select it. Anzo attaches the file and the Next button becomes active. Click **Next**. Anzo lists the file on the left side of the screen with a status of Pending. For example:



7. Click the **Process Pending Files** button to import the IMDB data to Anzo. Anzo imports the data and the status changes to **Processed**. Additional buttons are also displayed on the right side of the screen. You can click the

© 2023 Cambridge Semantics, Inc.

table row for IMDB-Movie-Data to display the schema details.

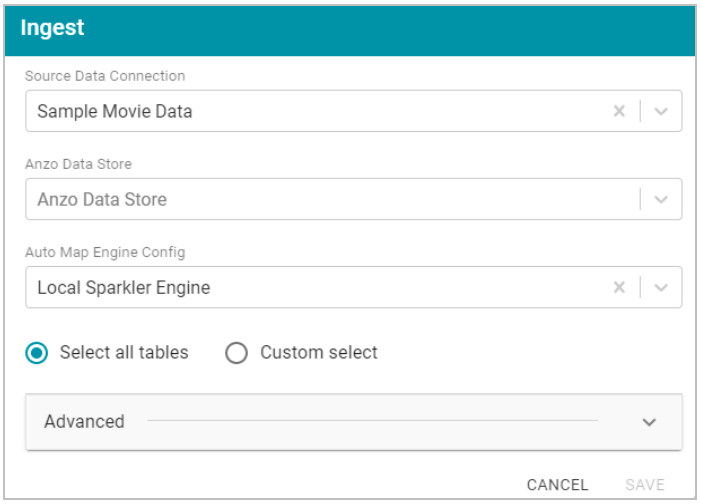


The data is now imported as a CSV Data Source. Proceed to the next step to ingest the data and automatically derive a Data Model, generate a Mapping, and create an ETL Pipeline for transforming the data to the RDF graph model.

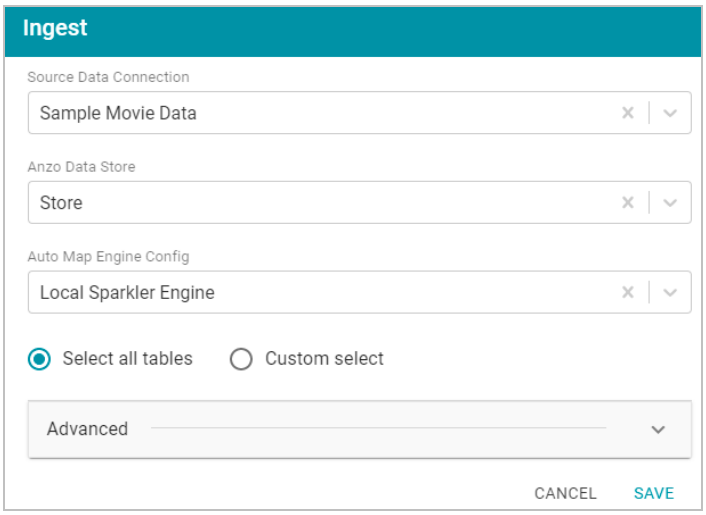
© 2023 Cambridge Semantics, Inc.

Ingest the Imported Data Set

1. On the Tables screen (from the last step of the procedure above), click the **Ingest** button. Anzo opens the Ingest dialog box and automatically populates the Data Source Connection value. If there is only one configured Data Store, the Anzo Data Store value is also auto-populated. In addition, if the default ETL Engine is configured for the system, the Auto Map Engine Config field is also populated. For example, in the image below the Anzo Data Store field is not populated because there are multiple available choices. The ETL Engine field is populated because the Local Sparkler Engine is configured as the default ETL Engine:



2. If necessary, click the **Anzo Data Store** field and select the Data Store for this pipeline.
3. If necessary, click the **Auto Map Engine Config** field and select the ETL Engine to use. For this tutorial, select **Local Sparkler Engine**. That means the pipeline will use the SPARQL-driven Sparkler ETL compiler to process the source data for the Spark engine.
4. Leave the **Select all tables** radio button selected. For example, the image below shows the completed screen:



© 2023 Cambridge Semantics, Inc.

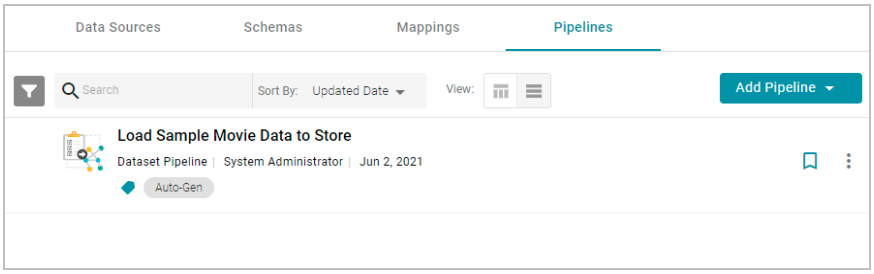
5. Click **Save**. Anzo creates a Pipeline for the Data Source and displays a "Pipeline was successfully created" confirmation message.

As part of creating the Pipeline, Anzo also generates a Data Model and Mapping for the sample data set. The last step in this tutorial describes each of the components. Proceed to the next step to run the Pipeline and generate the RDF File-Based Linked Data Set (FLDS).

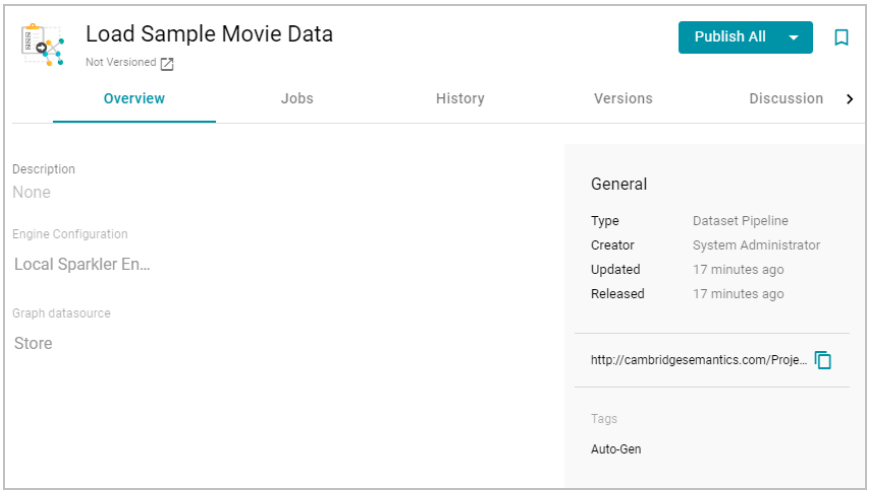
© 2023 Cambridge Semantics, Inc.

Run the ETL Pipeline

1. In the Onboard menu, click **Structured Data**. Then click the **Pipelines** tab. Anzo displays the Pipelines screen, which lists the Pipeline for the sample movie data set. For example:



2. Click the Pipeline title to open the pipeline Overview tab. For example:



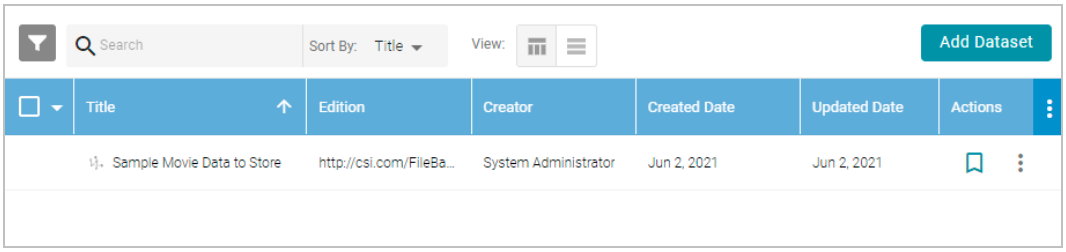
3. Click the **Publish All** button to run the Pipeline, which transforms the data to RDF graph format and generates the FLDS in the specified Data Store.

Anzo publishes the ETL files and displays confirmation messages as each step in the process completes. When the job execution is finished, the sample movie data set becomes available in the **Dataset** catalog. Proceed to the next step to create a Graphmart for the new Dataset and load it into AnzoGraph.

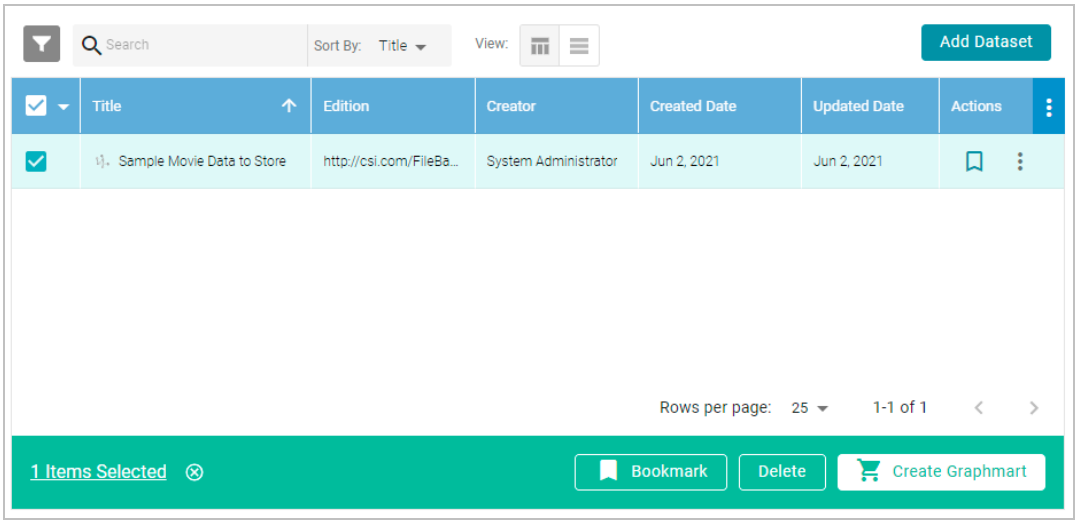
© 2023 Cambridge Semantics, Inc.

Create and Activate the Movie Data Graphmart

1. Expand the **Blend** menu and click **Datasets**. Anzo displays the Datasets catalog, which shows an inventory of all of the onboarded data in Anzo. It lists the sample movie data set. For example:

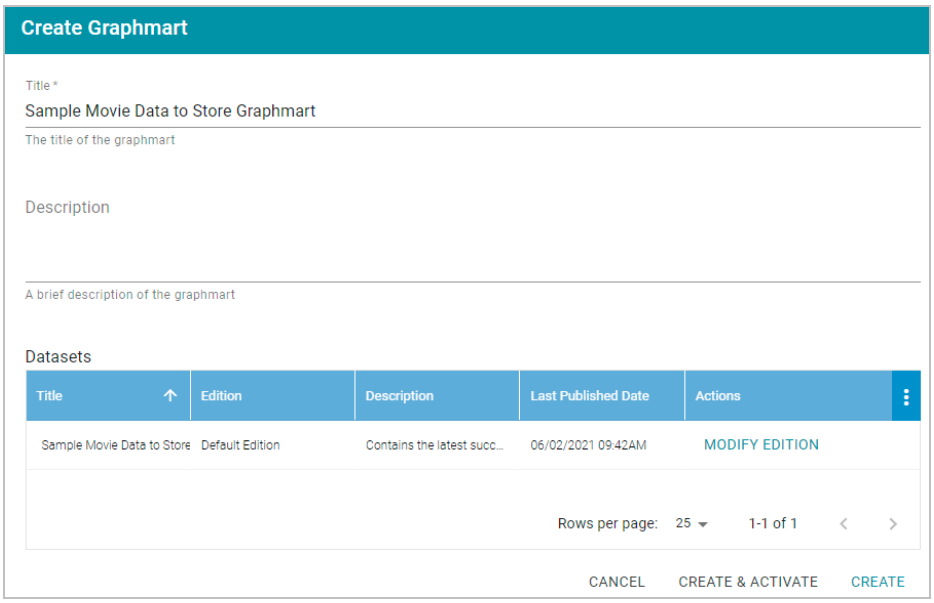


2. Hover the pointer over the Sample Movie Data item to display a checkbox in the left column, and then select the checkbox. Anzo adds the Dataset to the shopping cart and additional buttons become available at the bottom of the screen. For example:



3. Click the **Create Graphmart** button at the bottom of the screen. Anzo displays the Create Graphmart screen. Anzo populates the **Title** field by appending "Graphmart" to the Dataset name. For example:

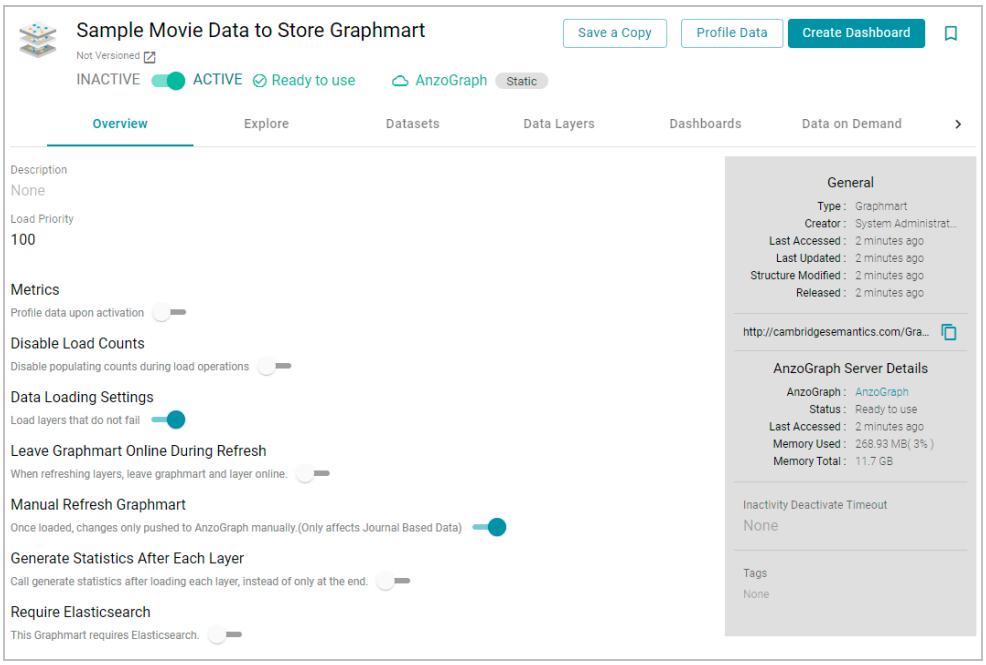
© 2023 Cambridge Semantics, Inc.



> **Tip**
>
> By default, the current working edition of the Dataset, called the Default Edition, is selected for the Graphmart. The Default Edition is managed by Anzo (the Managed Edition) and always contains the Data Components from the last successful run of the Pipeline. Since this is a new Data Source and the first run of the Pipeline, no other editions exist. To learn more about Dataset Editions, see Managing Dataset Editions in the Anzo Deployment and User Guide.

4. On the Create Graphmart screen, you have the option edit the Title and type an optional Description. When you are ready to create the Graphmart and load the data to AnzoGraph, click **Create & Activate** at the bottom of the screen. Anzo creates and activates the Graphmart and displays the Overview screen for the new Graphmart. For example:

© 2023 Cambridge Semantics, Inc.



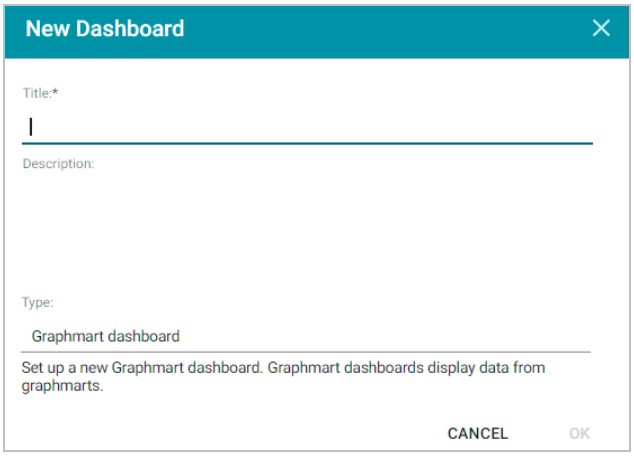
> **Tip**
>
> If you want to learn about the Graphmart settings that are shown in the image above, see Creating a Graphmart in the Anzo Deployment and User Guide.

Now that the data is loaded into AnzoGraph, it is available to explore and analyze in dashboards. Proceed to the next step to explore the sample data in the Hi-Res Analytics application.

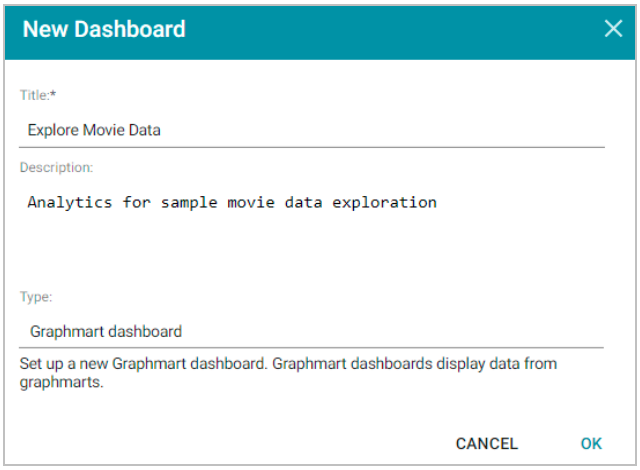
© 2023 Cambridge Semantics, Inc.

Explore the Data in a Hi-Res Analytics Dashboard

1. On the Graphmart screen (as shown in the last step above), click the **Create Dashboard** button. Anzo opens the Hi-Res Analytics application and displays the New Dashboard dialog box:

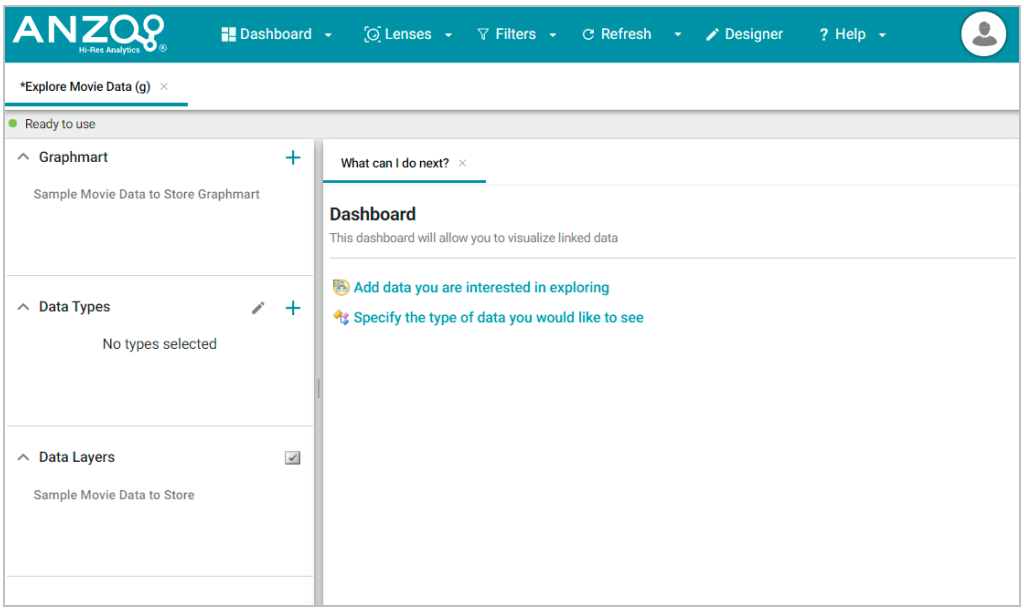


2. On the New Dashboard dialog box, specify a **Title** for the dashboard and add an optional **Description**. Leave the default **Graphmart dashboard** value in the **Type** field. For example:

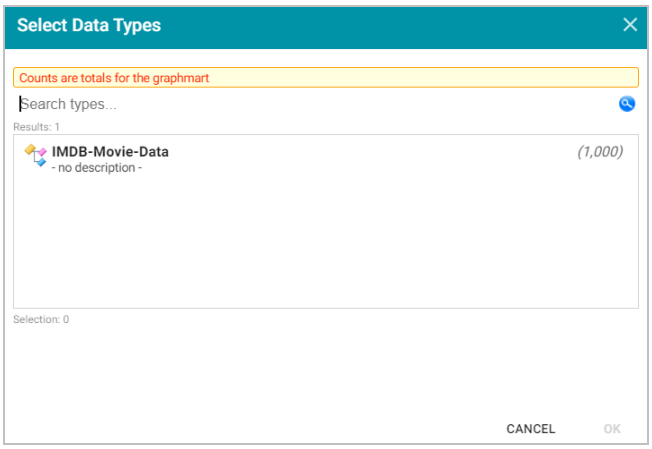


3. Click **OK** to create the dashboard. The new dashboard appears as a new tab on the screen and contains a sub-tab titled **What can I do next?**. This tab acts as a wizard to guide you through the initial dashboard creation. Anzo populates the Graphmart and Data Layers panels with the sample movie data Graphmart name and default Data Layer for the Graphmart.

© 2023 Cambridge Semantics, Inc.



4. To define the type of data to display on the dashboard, click the **Specify the type of data you would like to see** link on the What can I do next tab. The Select Data Types dialog box appears and displays the available data types. Since the sample data set has one type or class of data, only **IMDB-Movie-Data** is listed. The value in parentheses shows the total number of instances of that type.

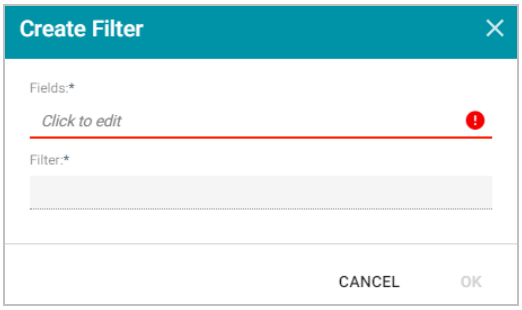


5. Click **IMDB-Movie-Data** to select it, and then click **OK** to add the data type to the dashboard. The selected type is added to the Data Types panel and additional options appear on the What can I do next tab.
6. In the main Hi-Res Analytics toolbar, click the **Dashboard** button and select **Save** to save the dashboard.

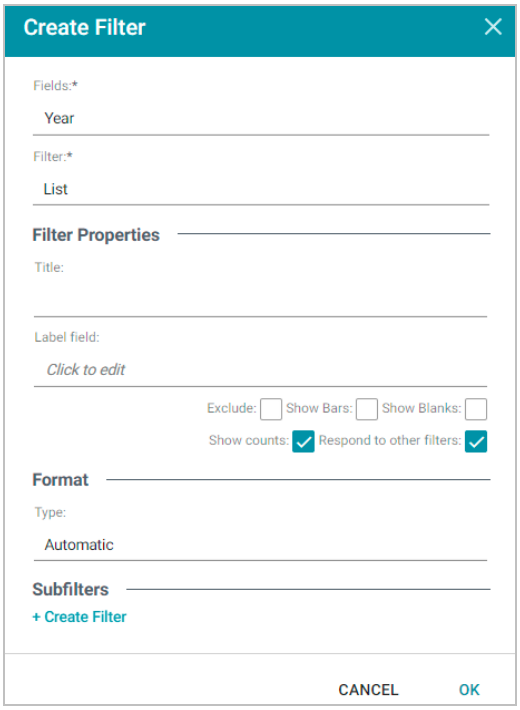
> **Tip**
>
> To start to explore the data and see what values exist for the properties in the IMDB-Movie-Data class, it can help to add a filter to the dashboard. Filters reveal the values associated with properties. Learning more about the values enables you to start making decisions about what properties to group on, for example, what properties have relationships, and what results you want to visualize.

© 2023 Cambridge Semantics, Inc.

7. To create a filter, click the **Add facets to filter the data** link on the What can I do next tab. Anzo displays the Create Filter dialog box:

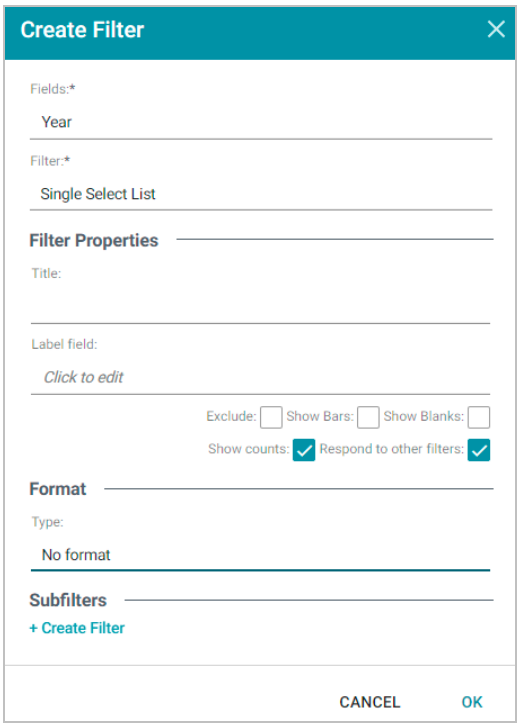


8. In the Create Filter dialog box, click the **Fields** field and browse the available properties. For this tutorial, select **Year**, and then click **Close** to close the Fields drop-down list. Additional options are displayed:



9. Click the **Filter** drop-down list and browse the available filter types. For this tutorial, select **Single Select List**.
10. At the bottom of the screen under **Format**, click the **Type** drop-down list and select **No format**. Removing formatting from the filter ensures that commas are not displayed for the Year values. The image below shows the completed dialog box.

© 2023 Cambridge Semantics, Inc.

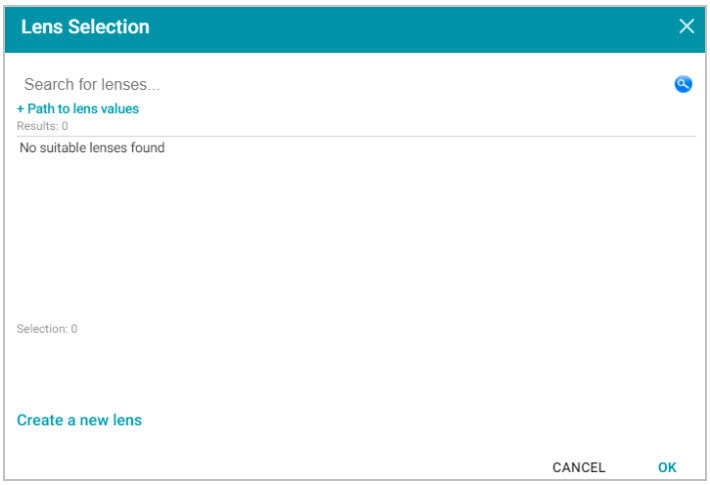


11. Click **OK** to close the dialog box and add the new filter to the dashboard. Anzo adds the Year filter to the left panel. The new filter will enable you to narrow the scope of the data when it is visualized.

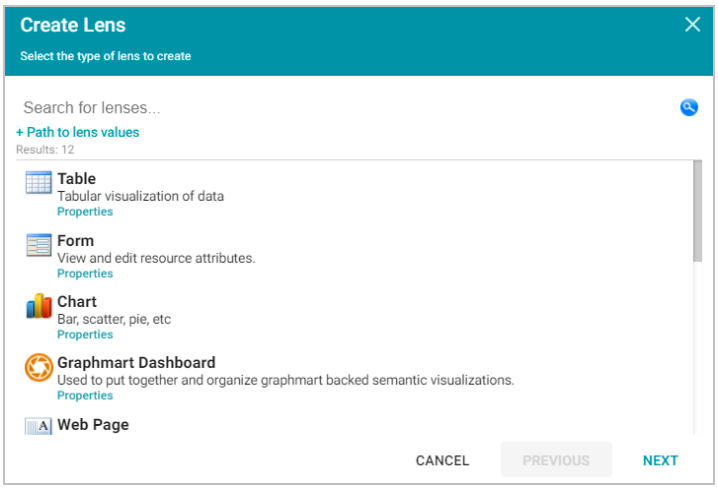
> **Tip**
>
> Once you have a good understanding of the values and relationships that exist in the data set, you can experiment with lenses and decide on the most appropriate way to display the data. Creating a Table lens is a quick way to get started with learning about the values and relationships in the data and to see how the new filter works.

12. To create a Table lens, click the **Select or create visualizations of your data** link on the What can I do next tab. Anzo displays the Lens Selection dialog box. Since lenses have not been created previously, the list of lenses is empty:

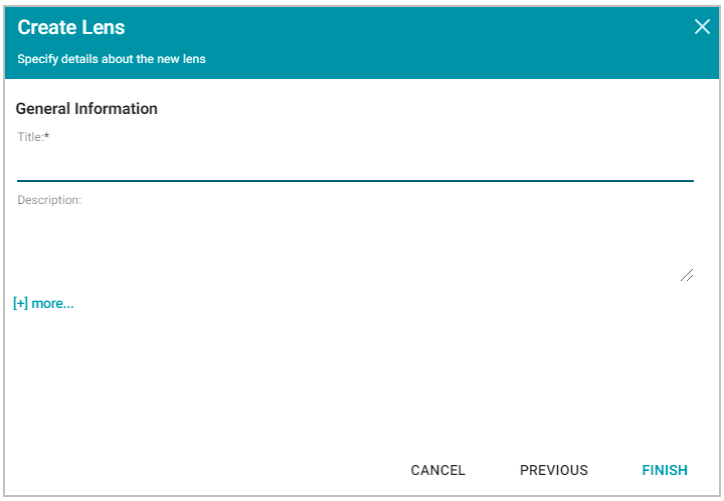
© 2023 Cambridge Semantics, Inc.



13. Click the **Create a new lens** link at the bottom of the screen. Anzo displays the Create Lens dialog box:

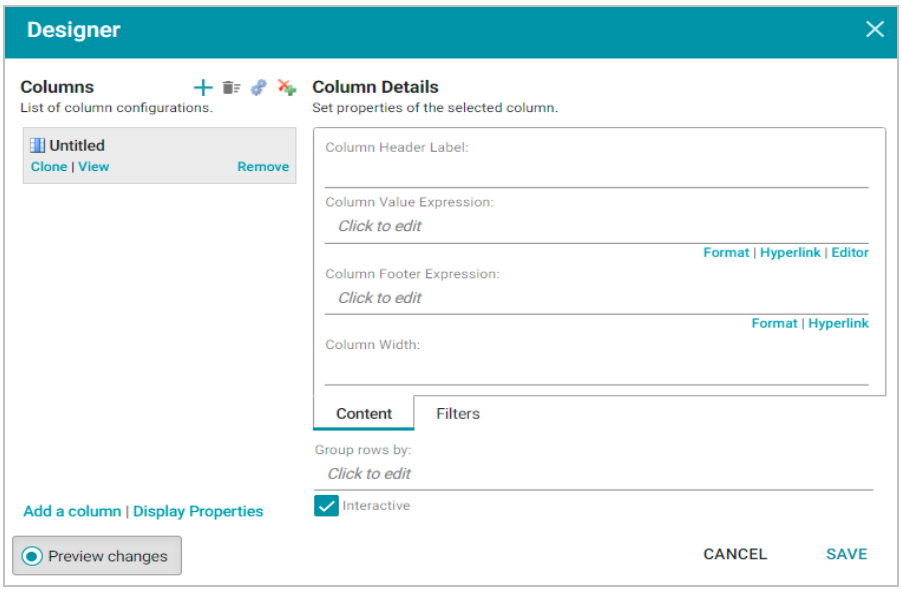


14. Select the **Table** lens and then click **Next**. Anzo displays the General Information dialog box:



15. Type a **Title** for the lens, for example "Movie Data Table," and then click **Finish**. Anzo adds the lens as a new sub-tab on the dashboard and opens the Table Designer:

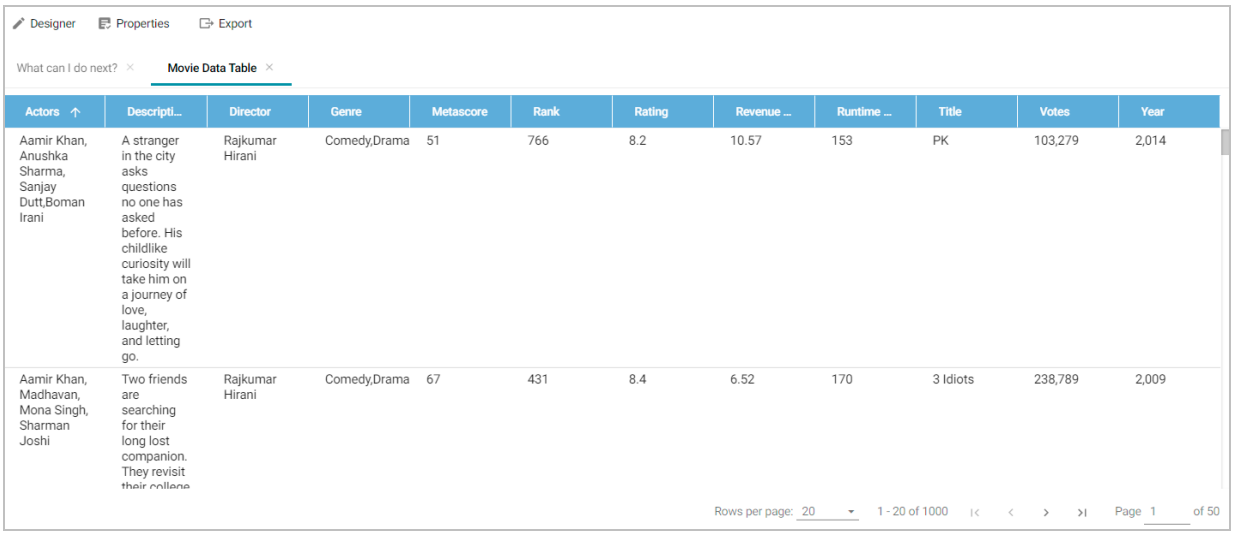
© 2023 Cambridge Semantics, Inc.



16. In the Designer, click the **Auto-generate columns** icon ( ) to add all available columns to the table.

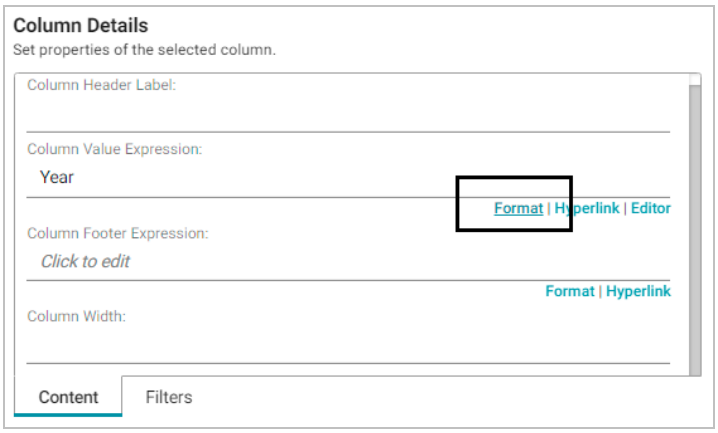
> **Tip**
>
> Since the sample data set includes only 12 columns, it is not overwhelming to view all columns at once. For larger data sets, you might want to be more selective when adding columns to table lenses.

17. Click **Save**. The table lens is added to the dashboard and is populated with data. For example:



18. If you want to make adjustments to the table, such as to reformat the **Year** column to remove the commas from the values, click the **Designer** button for the lens. In the Designer, select the **Year** column, and then click **Format** under Column Value Expression in the Column Details:

© 2023 Cambridge Semantics, Inc.

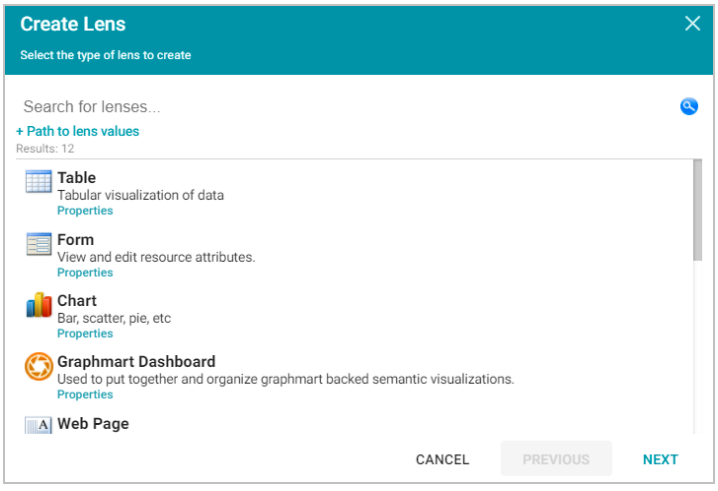


In the Format Column dialog box, click the **Type** drop-down list and select **Year** or **No format**, and then click **Close**. Click **Save** to change the lens change and close the Designer.

19. Now, you can filter the data and reduce the number of results that are displayed by clicking a Year value in the **Year** filter. In the filter, click any one of the years in the list to display only the data from that year. You can click another year to filter the table again, and so on.
20. In the Hi-Res Analytics main toolbar, click the **Dashboard** button and select **Save** to save the changes. Remember to save dashboards periodically.

Now that you can view a summary of the data in a table, it can help you determine how to further narrow or expand the results by adding, changing, or removing filters. In addition, you can experiment by adding other lenses to the dashboard to find the ideal way to display the data to answer the questions that you have. For example, the table includes a **Revenue** column. It might be interesting to see which movies made the most revenue. The next steps guide you through creating a chart lens to display movie revenue.

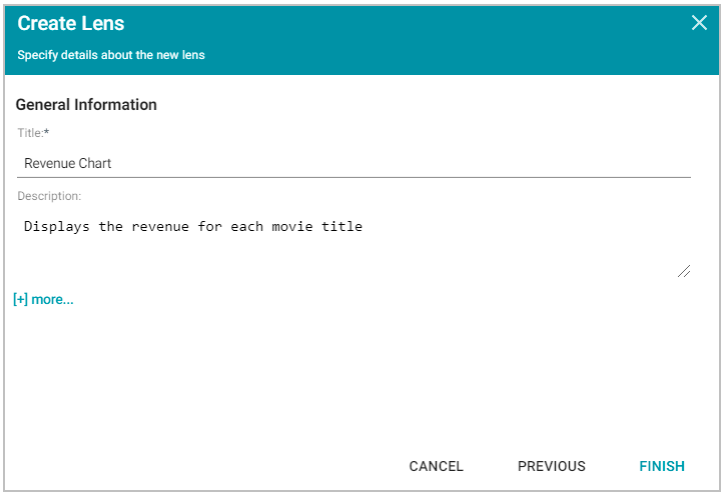
1. In the Hi-Res Analytics main toolbar, click **Lenses** and select **New**. The Create Lens dialog box opens.



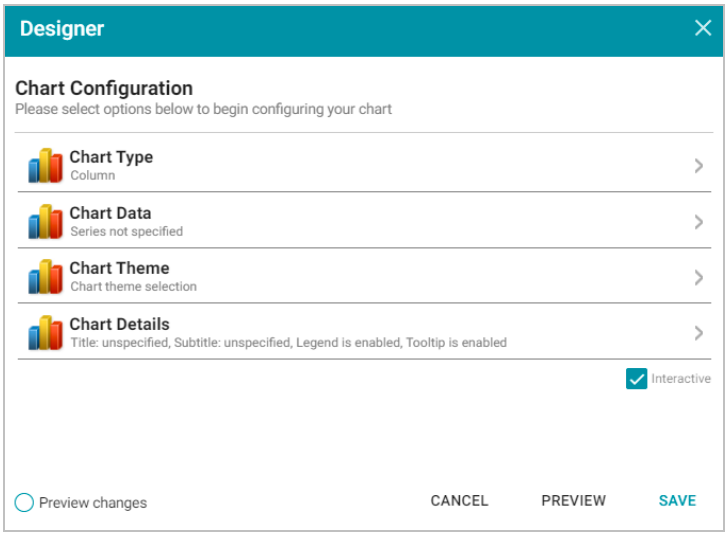
2. Select **Chart** and click **Next**. Anzo displays the General Information dialog box.

© 2023 Cambridge Semantics, Inc.

3. Specify a **Title** and add an optional **Description** for the new lens. For example:

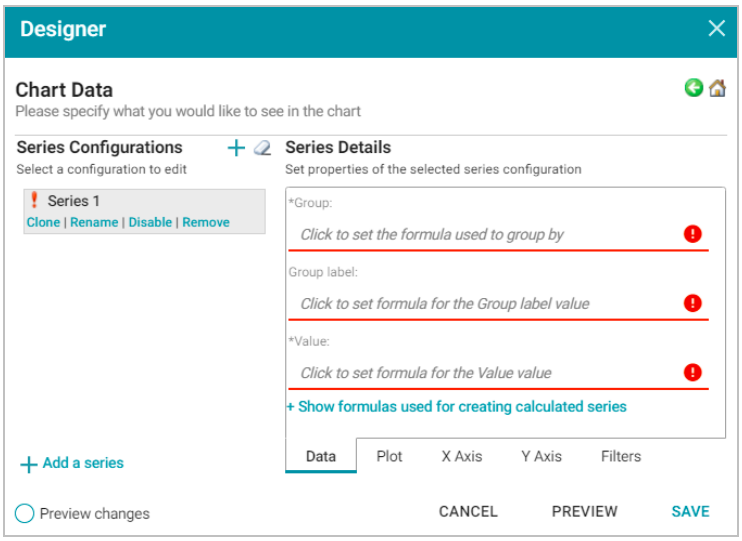


4. Click **Finish**. Anzo displays the Chart Designer, which enables you to configure the type of chart to create, such as column, pie, or line, specify the data that will populate the chart, and customize details such as the chart's theme and fonts, legend, title, and tooltip formats.

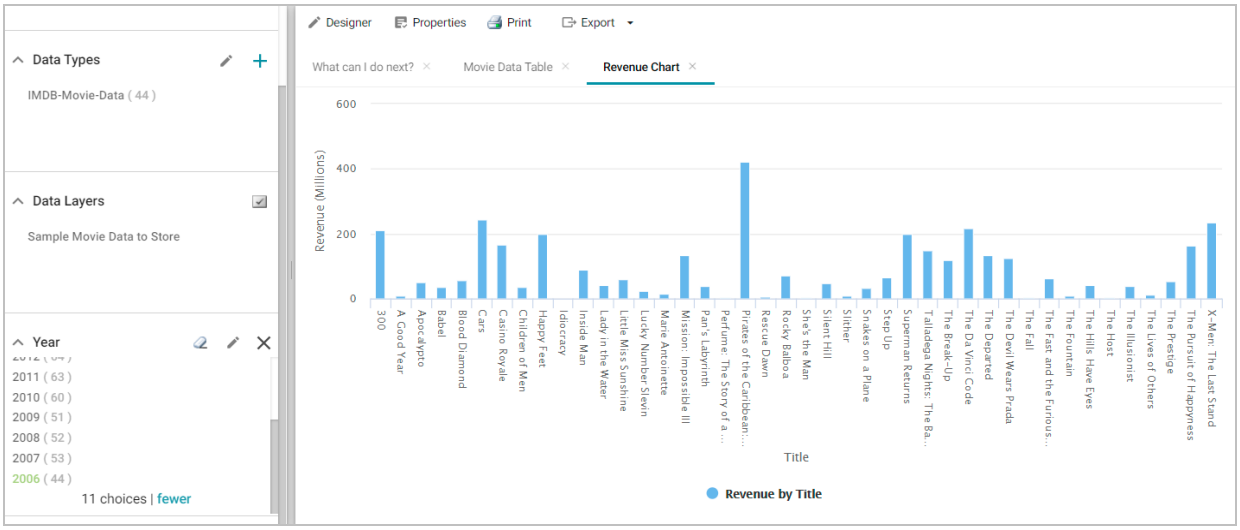


5. For this tutorial, accept the default Chart Type of **Column**, and click **Chart Data** to view the Chart Data screen and configure the data to display.

© 2023 Cambridge Semantics, Inc.



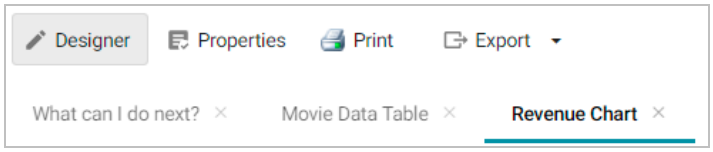
6. On the left side of the screen, click **Rename** under **Series 1**. Specify a new name in the Rename Series field, for example, "Revenue by Title," and then click **OK**.
7. On the right side of the screen, click the **Group** field and select the **Title** property from the drop-down list. The Group field sets the values for the X Axis on the column chart. Then click **Close** to close the drop-down list. Anzo also populates the Group label with the Title property.
8. Click the **Value** field and select the **Revenue (Millions)** property from the drop-down list. The Value field sets the Y Axis values. Click **Close** to close the drop-down list.
9. Click **Save**. Anzo adds the new Chart lens to the dashboard. To narrow the results that are displayed and filter on one year at a time, click a year in the Year filter. For example, by clicking **2006** in the filter, the chart lists the revenue for titles that were released in 2006 and shows that "Pirates of the Caribbean: Dead Man's Chest" had the most revenue for that year:



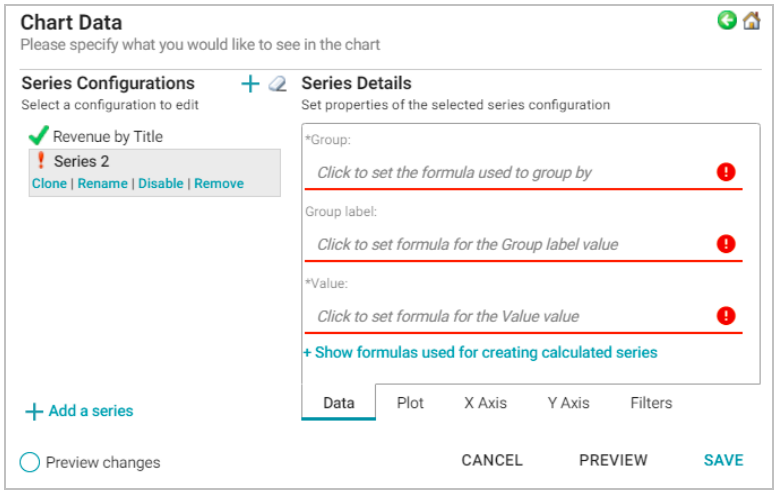
Tip You can hover the pointer over a bar in the chart to view details.

© 2023 Cambridge Semantics, Inc.

10. If you want to add more data to the Chart, such as to compare Revenue versus Rating for each Title, follow these steps:
    a. Open the Chart Designer by clicking the **Designer** button above the lens tabs.

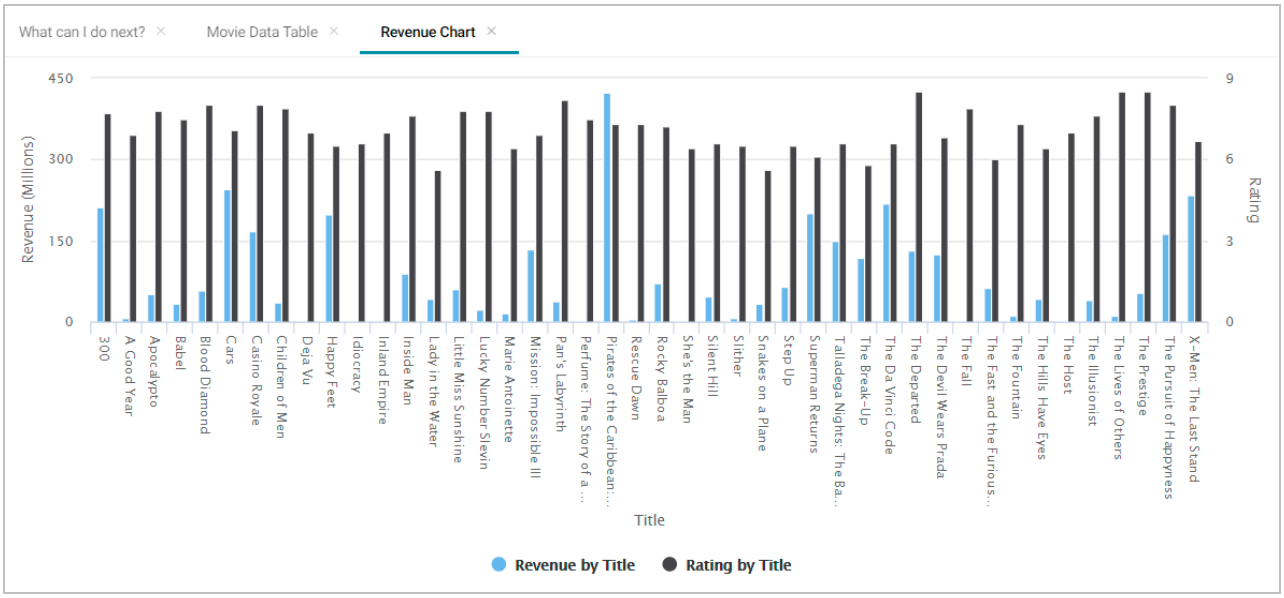


    b. In the Designer, click **Chart Data**.
    c. On the Chart Data screen, click the **Add a series** link at the bottom of the screen. A new series, named Series 2, is added to the list.



    d. Under **Series 2**, click **Rename** and specify a name for the series, such as "Rating by Title." Then click the **Group** field on the right side of the screen and select the **Title** property from the drop-down list. Click **Close** to close the list.
    e. Click the **Value** field and select the **Rating** property from the drop-down list. Close the list.
    f. Click **Save** to save the changes. The Chart is updated to display the Rating values in addition to the Revenue. For example:

© 2023 Cambridge Semantics, Inc.



Note

If you filtered the data in a previous step, the lens remains filtered on the chosen Year. You can de-select the Year in the filter or choose an alternate Year to explore the Ratings and Revenue for additional Titles.

Tip

For more information about working with dashboards, see Analyzing Data with Hi-Res Analytics in the Anzo Deployment and User Guide.

© 2023 Cambridge Semantics, Inc.

Review the New Solution's Artifacts

This section briefly reviews the artifacts that were created as part of the new sample solution and provides references to more details about each of the artifacts.

During the process of automatically ingesting the sample movie data from the supplied CSV file, Anzo generated the following components:

- A Model, which describes the structure and type of data that the movie data set contains. For more information and instructions for viewing the derived model, see Modeling Data in the Anzo Deployment and User Guide.
- A Mapping, which is input to the ETL job and maps the source data to the target elements defined in the data model. For more information and instructions for viewing the generated mapping, see Working with Mappings in the Anzo Deployment and User Guide.
- A Dataset Pipeline, which contains the ETL job that generates the target data set. For more information and instructions for viewing the generated pipeline, see Configuring Pipelines in the Anzo Deployment and User Guide.

Related Topics

Introduction to the Anzo Application

© 2023 Cambridge Semantics, Inc.