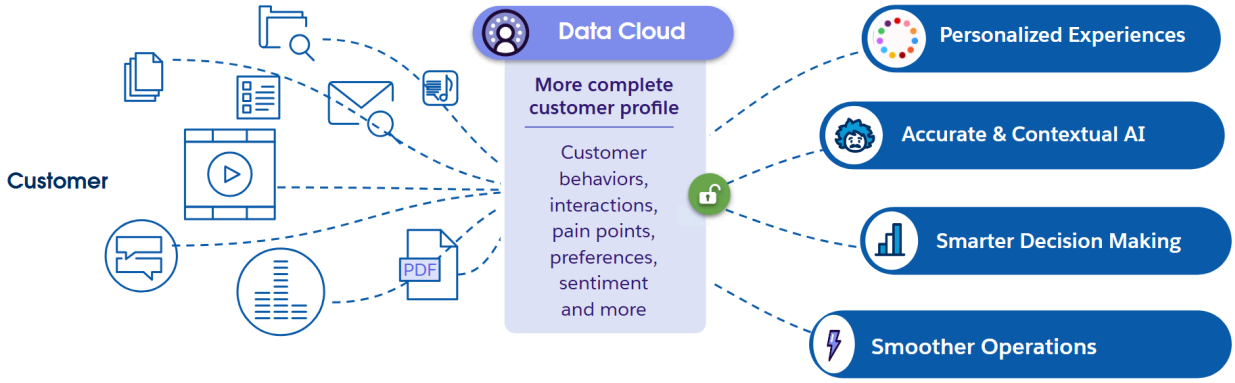
# Step-by-step Guide for Unstructured & Structured Data

## Introduction to Unstructured and Structured Data

Salesforce Data Cloud empowers you to ingest data into Data Cloud, but the way that this Data is referenced will differ based on the type of data ingested. There are two main types of data: Structured and unstructured data. You can think of structured data as organized and formatted data that is easily searchable. On the other hand, unstructured data is complex and cannot be easily analyzed as it lacks a specific or consistent form. Unstructured data can be analyzed in Data Cloud to gather critical information such as customer sentiment. Unstructured data can contain text data such as emails, social media posts, and web pages as well as multimedia content like images, audio files, and videos.

For structured data, Data Cloud utilizes a Data Lake Object (DLO) to ingest the data with a Data Stream. Data Cloud supports over 100 connectors, facilitating customers to bring structured data into Data Cloud. Structured Data is brought in by defining, mapping, and modeling the data to bring it into Data Cloud.

In contrast, unstructured data is referenced directly within Data Cloud rather than using a Data Stream. Data Cloud connects unstructured data from various sources, unpacks the data, categorizes it, stores it, and indexes it for future use in Data Cloud.




Let's explore this with an example of Sunshine Trails Hospitality, a hotel chain. They have unstructured data such as a FAQ pdf that helps guests prepare for their upcoming trip. It is important to get this data in Data Cloud, so that we can engage Agentforce to provide customers with a smooth experience when booking experiences with the hotel.

Unstructured data, due to its lack of structure, can be challenging to organize, store, and analyze. Data Cloud bridges this gap by connecting and extracting knowledge from unstructured data and enabling you to deliver deeper insights for your customers. By processing unstructured data, you leverage it for data discovery, analytics, and AI. Unstructured data contains meaningful information and with Data Cloud we can utilize this information.

## Technical Benefits of Unstructured and Structured Data

Salesforce Data Cloud takes away the complexity of understanding and difficulty in processing unstructured data allowing your technical team to focus on other areas. Data Cloud uses out-of-the-box connectors to effortlessly harmonize your data, which comes in many different forms and is spread across systems in your business. Unified Data Processing breaks down data silos and enables holistic data analysis. These prebuilt connectors link data from external sources quickly, extracting key details by converting your unstructured data into numerical formats. Data Cloud stores the unstructured data in a vector database, and indexes it for quick retrieval. Generated Chunks and Vectors are stored in new tables. Data Cloud fully manages the retrieval augmentation generation (RAG) on your behalf allowing you to access the right data from the right source at the right time with minimal effort. With Data Cloud's no-code retrievers, means that anyone in your team can access powerful analytics, and this data can be used to support automation and Agentforce use cases.

**Key Technical Features:**

- **Supports Real-Time Streaming and Batch Ingestion** of both structured and unstructured data.
- **Handles Diverse Data Sources** such as IoT devices, weblogs, call center transcripts, and customer feedback.
- **Integrates Seamlessly** with a variety of platforms and formats (e.g., JSON, XML, CSV, SQL, NoSQL, APIs).
- **Uses Connectors for External Systems** like ERP, data lakes, and social media platforms, ensuring comprehensive data coverage.
- **Automates data ingestion, transformation, and storage** for both structured and unstructured data.
- **Processes Massive Volumes of Structured and Unstructured Data** without performance degradation.
- **Transforms and Normalizes** structured and unstructured data into a common schema, enabling easy querying and analytics.
- **Reduces Infrastructure Costs** by leveraging Salesforce's cloud-native architecture and manual effort and complexity in managing diverse platforms.

**Steps to Bring Structured Data into Data Cloud**

We will follow these steps to bring Structured Data into Data Cloud from sources such as Salesforce CRM, Marketing Cloud, Amazon S3 Buckets, or Google Cloud Storage:

1. Create a Connection to the data source.
2. Create a Data Stream.
3. If using Zero Copy, define your Data Stream with Cache Acceleration.

You will then be able to Harmonize and Unify your data in Data Cloud. For more information about consolidating your structured data, refer to the step-by-step guide for Harmonization and Unification.

**Steps to Bring Unstructured Data into Data Cloud**

Now that we understand the key concepts of unstructured data, let's break them down step by step.

1. Connect to an external blob store.

2. Create a Unstructured Data Lake Object (UDLO).
3. Create a search index configuration for the unstructured data model object (UDMO) or structured Data Model Object (DMO) This creates vector objects for searching.
4. Perform Vector search queries from Prompt Builder, Einstein Copilot, or Tableau.

**Step 1: Connect to an External Blob Store**

Your unstructured data is stored in external blob stores such as Amazon S3, Microsoft Azure Blob Storage, or Google Cloud Storage, and you will reference them directly by using unstructured data lake objects (UDLO). To create a connection to an external blob store, Administrative Access to the desired external system is necessary. From there, you will log in with your credentials and create a new connection.

**Step 2: Connect Unstructured Data to Data Cloud**

Once a Connection is created, Sunshine Trails Hospitality will reference the unstructured data from external blob stores to Data Cloud. Next, you will want to create a new unstructured data lake object and select the desired external files. Then select your connector, for our example Amazon S3. You will point to the specific folder or directory in your blob storage. Then create and map the Unstructured Data Lake Object to an Unstructured Data Model Object (UDMO).

Let's take a deep dive into these steps:

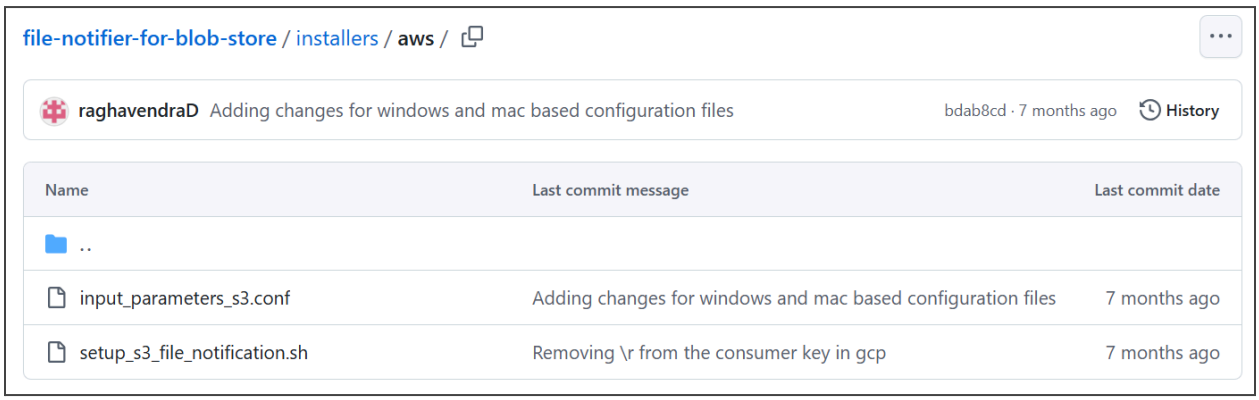
1. Navigate to Data Cloud > click **Data Lake Objects** >click **New**.
2. From the New Data Lake Object menu, select From External Files > click **Next**.
3. Choose which connector to use > click **Next**.
4. From the Select Connection dropdown list, select a connection.
5. In the Directory field, point to a specific folder or an entire directory in your blob store. All folders and subfolders in a directory are included. Optionally, use wildcard characters to specify a file name pattern for multiple files.
6. To add more directories, click **More Files**. You can include up to 5 directories > click **Next**.
7. Add an Object Name and an Object API Name for the UDLO.
8. From the Data Space dropdown list, select a data space in which to create the new UDMO or a data space from which to select an existing UDMO.
9. Map the UDLO to a UDMO.
    i. • To create a new UDMO, click **New.**
    ii. • To use an existing UDMO, click **Existing**, and select a UDMO from the list.
10. *Optionally,* leave the checkbox selected to create a search index configuration for the UDMO using system defaults that automatically selects text fields and a chunking strategy for each field. You can deselect the checkbox and create a search index configuration later if you choose not to do so now.
11. Click **Next.** (if you created a search index configuration, review the details, and save your work).

Now that you have created a connection from the external blob storage, you are ready to set up a file notification pipeline. This will notify Data Cloud whenever files are added, updated, or deleted from your blob storage.

**Step 3: Create File Notifications**

File Notifications allow you to receive notifications whenever data files are added, updated, or deleted from your external blob storage. This allows Data Cloud to know when to pick up new or updated files from your external blob storage. The steps you follow will vary depending on where your unstructured data is stored. You will want to follow the following steps in with your external blob storage:

1. Create a Private-Public Key Pair and Certificate
2. Set up a Connected App and OAuth Settings
3. Download the Installer Script and Function specific to your external blob storage
4. Install the Script Setup with macOS or Linux




**Step 4: Create a Search Index**

Follow a few automated steps to create a hybrid search index configuration for an unstructured data model object (UDMO). Data Cloud automatically applies defaults for the chunking and vectorization strategies and creates chunk and index model objects. Chunking is important for this process because we want to break it down into manageable meaningful chunks. This will allow the data to be added to the Data Cloud search index and retrieved in search results.

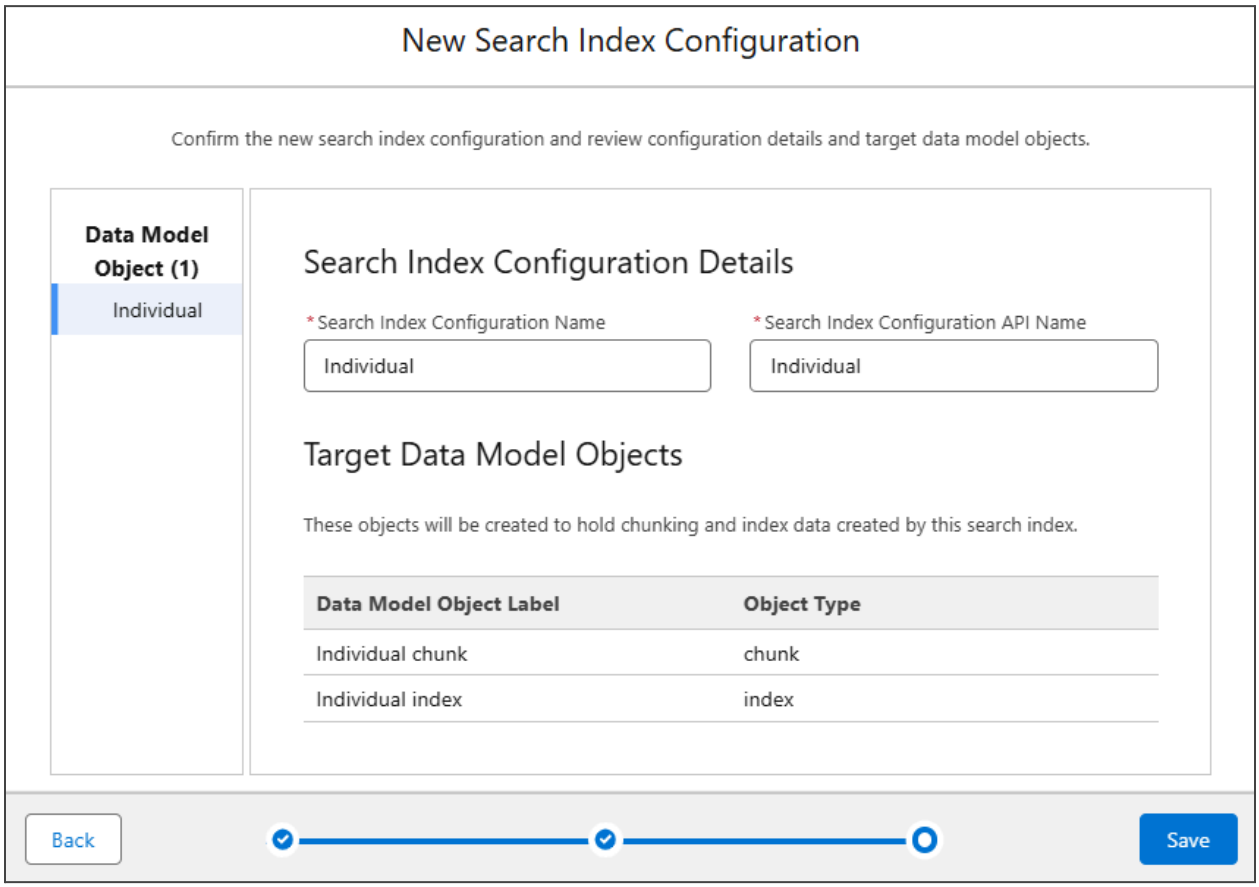
Vectorization is the process of embedding your unstructured data into numerical representations of data that machines can read. With easy setup, Salesforce will create a Search Index for us that selects these defaults:

- Chunking strategy: Passage extraction
- Embedding model: E5-Large V2
- Search Type: Hybrid Search

These steps will create a search index:

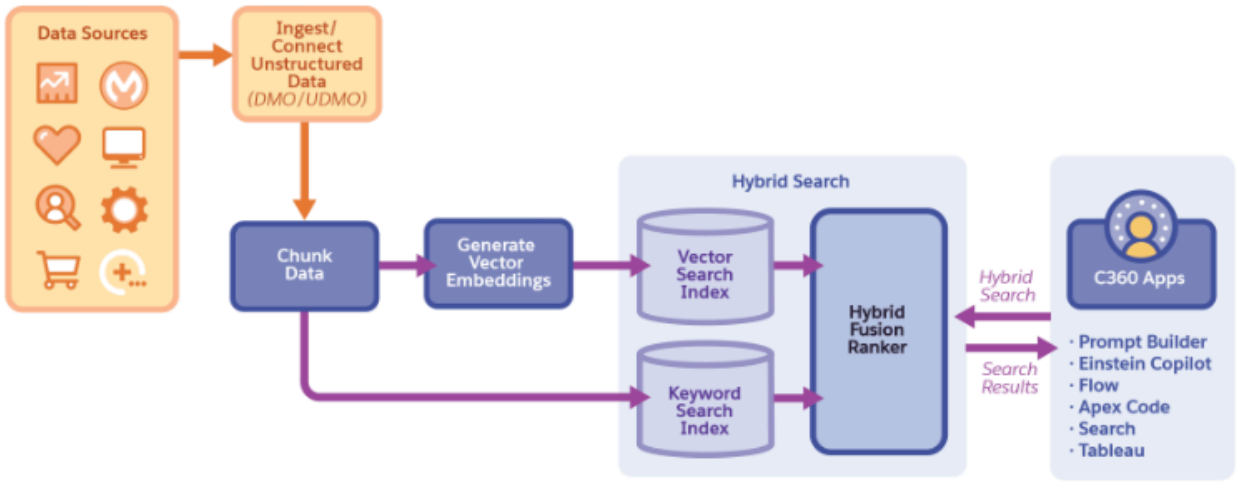
1. From App Launcher, select Data Cloud.
2. In Data Cloud, click **Search Index** > **New**.
3. Click **Easy Setup** > **Next**.

4. From the New Search Index Configuration page, select a Data Space and one or more UDMOs. A search index configuration is created for each object that you select.
5. Add a Search Index Configuration Name and Search Index Configuration API Name.
6. Click **Next**.
7. Review the configuration and the target data model objects, and click **Save**.




### Step 5: Run the Search Query

Once your unstructured data has been chunked and vectorized, Data Cloud makes it easy to perform searches on your unstructured data using prompt templates. When we created our Search Index for Sunshine Trails Hospitality, we selected the Search Type of Hybrid. Hybrid Searches merge the retrieved information including the vector search and keyword search and rank the results to show the most relevant information.




With vector search and Einstein Copilot queries, you can search across structured and unstructured data sources, and receive summarized responses that improve efficiency and accuracy.

When running a search query, Einstein Copilot will follow these steps:

1. Einstein Copilot triggers a prompt template.
2. The prompt initiates a retriever, which is sent to Data Cloud vector search.
3. Data Cloud vector search retrieves matching excerpts from relevant unstructured data.
4. The prompt template is populated and sent to a large language model (LLM) for summarization.
5. The LLM processes and returns a summarized response to the prompt template.
6. This response is then relayed back to Einstein Copilot.
7. The response is displayed to the user.

Agenforce invokes the search of your unstructured data. Once the unstructured data has been brought into Salesforce using a prompt template, you will be able to begin to retrieve the data to get relevant customer data.

**Integrating Knowledge Article Data**

Knowledge Articles contain both structured and unstructured data, this can be powerful data to ground your Service Replies and prompts to understand a customer's use case. Knowledge objects contain default mappings to the relevant DMOs in Data Cloud to save you time.

1. Create a new **Data Stream** in Data Cloud with the **Salesforce CRM** data source.
2. Select the **Knowledge** object.
3. Review the fields for the Knowledge objects included in your data stream.
    a. Standard fields from the objects in your data kit are automatically mapped to the relevant data model objects in Data Cloud. However, you must manually map any custom fields.
4. Deselect any fields not required for your data stream.
5. Review the new data streams and objects to be deployed, and click **Deploy**.

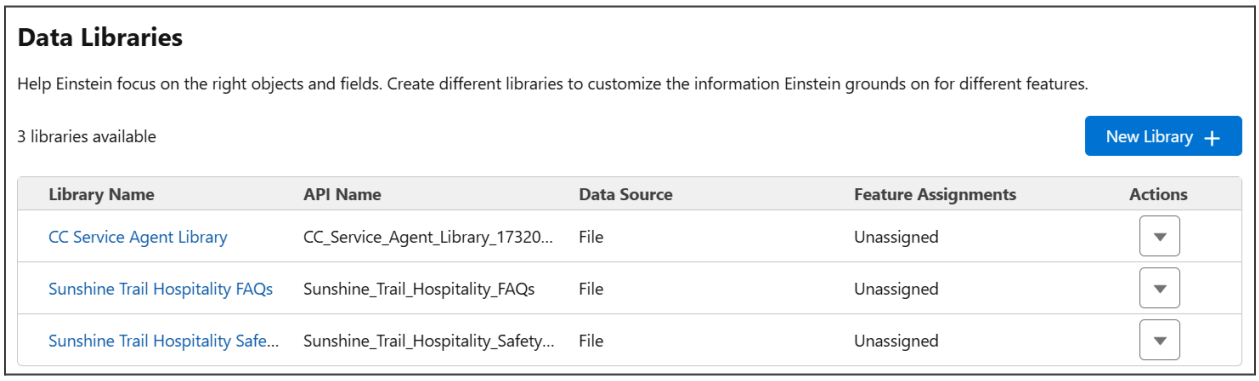
**Understanding Einstein Data Library**

Einstein Data Libraries allow you to improve accuracy, add personalization, and build trust in Einstein's responses. Einstein Data Libraries can be another way to bring unstructured data such as PDF. For our example of Sunshine Trails Hospitality, important information such as FAQs, and Safety policies are stored in PDF. By uploading these PDFs to our Data Library this allows Einstein to check the accuracy of responses against Sunshine Trails Hospitality's unique policies and information. To get started:

**Step 1: Create and Configure a Data Library**

Data Libraries allow you to create an index of Knowledge articles and fields or file uploads. This determines where Einstein knows which information to base its responses on.

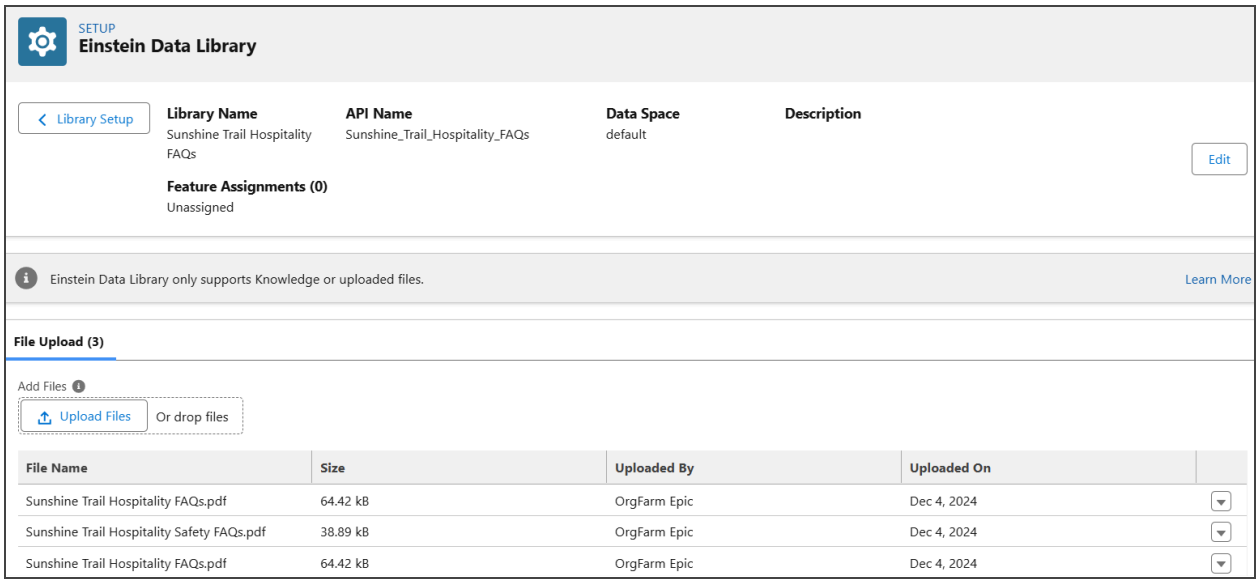
1. From the Einstein Data Library setup > select **New Library**.




2. Select a data space. After you choose a data space, you can't change it later.
3. Enter a library name, and optionally, enter a description.
4. Click **Save**.

**Step 2: Choose a Data Source for your Library**

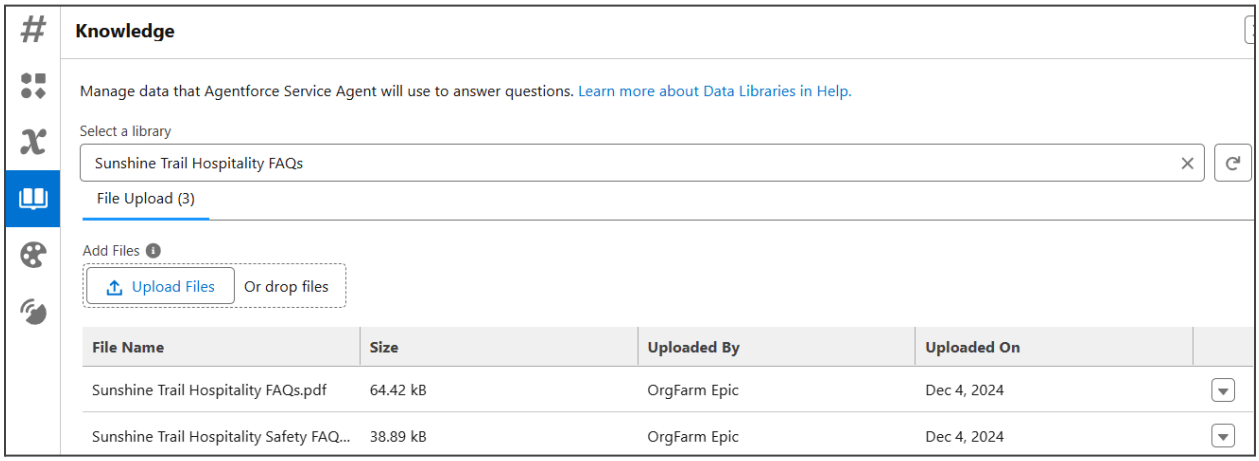
You will then select a Data Source for your Library either Knowledge or file uploads for the library's data source. A data library can't support Knowledge and file uploads simultaneously. Once you've saved your selections, data streams, a search index, and a retriever are created automatically and available to view or edit in Data Cloud.




You can create multiple data libraries and assign them to multiple features such as Agentforce, but each feature can only use one data library at a time.

### Step 3: Assign Data Libraries to Agentforce

Once created, you will use your Data Library to ground your agent responses. Navigate to Agent Builder and the Knowledge section. You will select the desired Data Source for your agent. You can always adjust this selection or manage data sources later in the Agent Builder.




### Best Practices

| Common Discovery Questions | Design Considerations |
|---|---|
| • Understanding the volume and types of data across systems. | • Filter data from the source if the data is not needed in Data Cloud |
| • Details about the systems where data is | • Ensure you choose the correct UDMO. |

| | |
|---|---|
| stored | |
| | • When unstructured data in the form of files is connected to Data Cloud from an external blob store, there are no billing implications. |
| | • In Data Cloud, unstructured data may be chunked and vectorized using an embedding model. understanding the volume and types of data across systems. |

**Conclusion**

By utilizing Data Cloud, businesses can transform complex, unstructured data into actionable insights. Data Cloud can ground Agentforce with customer insights, preferences, sentiment, and more. Data Cloud's ability to connect to external blob stores, create unstructured data lake objects, and perform search queries provides a seamless process for managing unstructured data. With the integration of Knowledge Article Data, businesses can further enhance their service responses and understand customer use cases better. Salesforce Data Cloud offers a comprehensive solution for managing and leveraging both structured and unstructured data, enabling businesses to deliver deeper insights and enhanced customer experiences.

**Resources**

Help Article: Unstructured Data in Data Cloud
Help Article: Send File Notifications from External Blob Stores to Data Cloud
Help Article: Ingest Knowledge Article Data from Salesforce CRM
Help Article: Chunking Unstructured Data
Help Article: Manage Search Indexes
Help Article: Vector Search