**Discover Data, Prepare Data, and Manage Metadata with SAP Data Hub**

DAT261

Exercises / Solutions

Table of ContentS

[Before You Start 3](#_Toc525055391)

[Browser 3](#_Toc525055392)

[System Information 3](#_Toc525055393)

[Scenario Description 4](#_Toc525055394)

[1. Connection Management in SAP Data Hub 5](#_Toc525055395)

[2. Discover Data and Manage metdata in SAP Data Hub 9](#_Toc525055396)

[Metadata Terminology 9](#_Toc525055397)

[3. PREPARE DATA AND MANAGE METADATA IN SAP DATA HUB 26](#_Toc525055398)

[High Level Product Overview 26](#_Toc525055399)

[Personas 26](#_Toc525055400)

[Product Marketecture 27](#_Toc525055401)

[SAP Agile Data Preparation and SAP Data Hub Interoperability capabilities 27](#_Toc525055402)

[4. ENRICH DATA IN SAP DATA HUB 43](#_Toc525055403)

# Before You Start

## Browser

Please use supported browser(s) for your exercises. Chrome is preferred for this hands-on session.

## System Information

**SAP Data Hub Launchpad**

|  |  |
| --- | --- |
| URL (Group 01-05) | <https://ip-18-195-164-62.sapdatahub.com:9086> |
| URL (Group 06-10) |  |
| URL (Group 11-15) |  |
| URL (Group 16-20) |  |
| URL (Group 21-25) |  |
| URL (Group 26-30) |  |
| Additional |  |
| Additional |  |
|  |  |
| Tenant name | default |
| Username | TAxx  Where xx is your group number |
| Password | Welcome01 |

**Record your**

**Group ID**

## Scenario Description

You either as a business analyst or a data engineer first discover datasets available to you using SAP Data Hub’s Metadata Explorer. Once you find your dataset, it may not be ready to use out of the box. If you are a business analyst, you use SAP Agile Data Preparation to enrich your dataset found from Metadata catalogue which you in turn can republish to the catalogue for other users. A data engineer can use SAP Data Hub Modelling to create pipeline(s) to enhance your dataset and then republish them to the catalogue for other users to reuse your enriched dataset. The following exercises will walk you all the steps needed here.

**Exercise 1: Connection Management in SAP Data Hub**

**Exercise 2:** Discover Data and Manage Metadata in SAP Data Hub

**Exercise 3: Prepare Data** and Manage Metadata **in SAP Data Hub**

**Exercise 4: Enrich Data in SAP Data Hub**

# 1. Connection Management in SAP Data Hub

**Estimated time: 10 minutes**

Create a connection in SAP Data Hub, which represents an access point to a remote system or a remote data source. You create connections in the SAP Data Hub Connection Management application. Each connection is identified by a unique ID and has a type that specifies the nature of the access point. In addition to the mandatory ID attribute, every connection can have an optional description. After creating a connection, you can use the connection in the SAP Data Hub Modeler to reference an external resource.

**Objective**

In the following exercise, you will learn how to work with the SAP Data Hub Connection Management and how to manage connection of external systems in SAP Data Hub. Before you start, please make sure you have access to an instance of SAP Data Hub with necessary user and password which is expected to share prior to your session by your instructor(s).

Even though SAP Data Hub can connect to a set of large number of external systems as documented in SAP Data Hub help portal, a subset of external systems are connected with your hands-on instance.

| Explanation | Screenshot |
| --- | --- |
| 1. Click on the URL of the **SAP Data Hub Launchpad** assigned to your group.   Provide **Tenant** name, **Username** (TAxx) and **Password**.  Then click on **Login**. |  |
| 1. It will open the **SAP Data Hub Launchpad**, where you can see few applications running.   Now click on the **Connection Management** tile |  |
| 1. Find all the available connections in your hands-on instance including BW, ECC, HANA, HDFS, S3 and Vora. However, for this specific hands-on session, mostly **HDFS** connection will be used |  |
| 1. Click on the **Connection Type** on the top menu bar to explore connections supported in a particular Data Hub version. Make informed decisions about the content of your datasets after indexing, publishing and/or profiling your datasets. Note down the capabilities of each connection type. |  |
| 1. Click on **Connections** on the top menu to start using these available connections in your instance. |  |
| 1. Verify connection status of each connection. But we will check connection status of **HDFS** first. (1) click on **More Action Icon**  next to the HDFS connection and (2) click on **Check Status**.   **Info**: This check can sometimes take some time. |  |
| 1. Find the status of **HDFS** connection and then click on **OK** to close the status window. |  |
| 1. View the configuration of HDFS connection, (1) click on **More Action Icon**  next to the HDFS connection and (2) click on **Edit**. |  |
| 1. You can view the different connection parameters for this HDFS connection and when you are done, close the configuration window clicking on **Cancel.** |  |
| 1. To create a new connection in SAP Data Hub, click on **Create** on the right-hand side top. It will pop up connection parameter windows.   Based on the connection type, you will specify different parameters. |  |
| 1. §If you choose to create a new connection, enter all relevant parameters and hit Save. But here you close the windows clicking on **Cancel** as you are not attending a new connection in your test instance. |  |

Congratulations!

This conclude this exercise on Connection Management in SAP Data Hub. You are ready to continue with Metadata explorer in SAP Data Hub in the next exercises.

# 2. Discover Data and Manage metdata in SAP Data Hub

**Estimated time: 25 minutes**

across different systems by using the MetadataExplorer.

The Metadata Explorer gathers information about the location, attributes, quality, and sensitivity of data. With this information, you can make informed decisions about which datasets to publish, and determine who has access to use or view information about the datasets.

Use the Metadata Explorer to :

* preview data in the datasets
* create indexes about the dataset contents to aid in searching for datasets
* profile data to view information about the contents of different datasets
* publish datasets to allow others to view and search the data
* label the datasets with keywords, which also helps in searching for datasets

In this section, we will use the Metadata Explorer in SAP Data Hub to monitor, profile, index, label, publish metadata and search dataset.

## Metadata Terminology

**Indexing** extracts the metadata so that users can decide whether to allow others to use the dataset. The extracted information is available in the Browse window.

**Publishing** extracts the same metadata but provides additional information such as the path and folder where the data is stored and is placed in the Catalog where other users with appropriate permissions can access the published datasets.

**Metadata Extraction** can be performed either by indexing or publishing.

**Profiling** produces additional metadata about the values in the dataset. For example, you can view the unique or distinct values, the minimum and maximum values, average length, and whether there are null, blank, or zero values. This information can help you determine which datasets may need cleansing, masking, or any number of options available in the SAP Data Hub Modeler.

**Fact Sheet and Data Preview** contains both metadata extracted by indexing or publishing and profiling results, if profiling is performed and data preview allows quick viewing of data.

In this section, we will create a catalog, create a label, browse connection, index and publish a set of datasets, profile a dataset, monitor activities and search a dataset with SAP Data Hub metadata explorer. You can use the same instance throughout your hands-on session.

| Explanation | Screenshot |
| --- | --- |
| 1. In the **SAP Data Hub Launchpad**, click on **Metadata Explorer** tile |  |
| 1. It will open a new window for Metadata Explorer |  |
| 1. Click on **Overview** on the top and find all the links for Metadata Management |  |
| 1. Click on **Catalog** to explore the catalog, index, and publish datasets |  |
| 1. Click on **New Folder** to create a folder in the catalog. You will use your group number to create a new folder. |  |

| Explanation | Screenshot |
| --- | --- |
| 1. Enter Name as **Groupxx\_IoTCatalog** and Descrioption as **Groupxx IoT Catalog** and click on **Save.** |  |
| 1. A new folder is created as shown here. |  |
| 1. Navigate back to Metadata Explorer clicking on **< icon** on the left-hand top |  |
| 1. Under quick links Tab, click on **Browse Connection**. |  |
| 1. It lists the available connections in the hands-on instance. You can toggle between List View and Grid View on the right hand top. This feature is available across the product. |  |
| 1. Grid view is shown here and toggled back to list view to browse few connections starting with **HDFS** |  |
| 1. In the list view of the connections, expand clicking **HDFS** connection**.** |  |
| 1. Based on your group, select **exercises\_xx** |  |
| 1. Find device.csv file |  |
| 1. (1) Click on more actions icon next to device.csv and select **View Fact Sheet** |  |
| 1. It shows the file is not indexed, not profiled and shows metadata including columns information |  |
| 1. Navigate back by clicking on icon **<** on the top left |  |
| 1. Start profiling devices.csv by (1) clicking on **more actions icon** next to the file and selecting (2) **Start Profiling.** |  |
| 1. Confirm profiling at the prop-up window by clicking on **Yes** |  |
| 1. Click on **Notification icon** on the right hand top. |  |
| 1. Find a notification corresponding to the profiling job for devices.csv and close toggling on **notifications icon** again. |  |
| 1. Click on **Browse Connections** and click on **Monitoring** to find the status of the job. |  |
| 1. In the Metadata Monitoring, find that the profiling task has status *Active.* Once the job is completed, it will send another notification. |  |
| 1. Clicking on **the notification icon** on the right hand top, it will show that the job is completed.   **Info**: You will maybe see some other notifications, for example from other users. |  |
| 1. Go back to metadata monitoring by clicking on **notification icon.**   **Info**: To get the new status, you must refresh the screen or select **Open Monitoring View** and afterwards navigate back. |  |
| 1. It shows the profiling task is completed and click on left hand top to **navigate back** to list all the files |  |
| 1. For the row devices.csv, click on (1) **more action icon** and then (2) select **View Fact Sheet** |  |
| 1. It will show that the dataset is profiled at the top and show the profiled results including % of Null, Blank and Zero for the country field as 18.4, 0, 0. |  |
| 1. A preview of data can be obtained clicking on **Data Preview** |  |
| 1. Some of the country field shows **NULL** value. |  |
| 1. Navigate back clicking on **<** on the left hand top. Then click on (1) **more actions** and (2) select New Publication Action |  |
| 1. Find New Publication Action Window and click on **Target Folder** **Edit** pencil icon |  |
| 1. Select your group folder and click on OK. |  |
| 1. After selecting **target** and click on **Run.** |  |
| 1. Clicking on **Notifications icon** on the right-hand top which shows a new indexing and publishing task has been started and can be viewed at monitoring.   **Info**: If you get a “504 Gateway-timeout” error, try again. This is regards the current system setup for this training environment. |  |
| 1. At the metadata monitoring, it shows the publish task is completed |  |
| 1. From the metadata monitoring navigate to overview clicking on (1) **Monitoring** and then on (2) **Overview** |  |
| 1. Search Device in the overview |  |
| 1. Find Devices.csv in your Groupxx\_IoT\_Catalog |  |
| 1. Click on (1) **more action icon** and click on (2) **Edit Label**. |  |
| 1. (1) Enter your Groupxx\_MissingData and (2) Click on Save to add a label to this dataset which has missing country. The same label can be added to other datasets and searched by label. 2. Click on the back icon. |  |
| 1. From Metadata Explorer overview, search is done with a label which is *Group1\_MissingData* |  |
| 1. The search result shows the dataset with match attribute as label. |  |
| 1. (Optional) Index S3. Navigate to browser connection from Search at the top menu. |  |
| 1. (1) Click on More action next to S3, select (2) New Publication Action, select Target as Groupxx\_IoT\_Catalog and (3) click on Run. It will index and publish all the files under S3 connection. |  |

**Optional:**

You can add an additional data set to your catalog. Browse the same HDFS connection and go to your folder exercises\_xx. Profile the file customer.csv and publish it to your catalog. Have a look at the fact sheet and you will find out that the country data is always filled. In the further exercises you will use this file to

Missing Data dataset will be enhanced first by a self-service data preparation tool by a business analyst and can also be enhanced with a pipeline by a data engineer.

# 3. PREPARE DATA AND MANAGE METADATA IN SAP DATA HUB (DH native capabilities)

| Explanation | Screenshot |
| --- | --- |
| * 1. Find Devices.csv in your Groupxx\_IoT\_Catalog and choose **Prepare Data** |  |
| * 1. Choose **Change Case** for TYPE |  |
| * 1. Choose **Upper Case** and press **Apply** |  |
| * 1. Now devices types are in Upper Case |  |
| * 1. Choose **Preparations** and **Run Preparation** |  |
| * 1. You need to choose target connection and container where prepared data should be stored.   Please, choose and fill:   * Connection: * **HDFS** * Folder: * **/exercises\_<user\_number>** similar to /exercises\_01 for user ta01 * Dataset Name: * **Devices\_p<user\_number>** similar to Devices\_p01 for user ta01 * Type of the target object:   **CSV**   * Write mode:   **Overwrite**  Then press **Apply**. |  |
| 3.7 Go to Monitoring and wait until your preparation will be finished |  |
| * 1. If your preparation was completed successfully you can check your data in the target. |  |
| * 1. Go to Browse Connections and choose your target storage HDFS, your folder and see dataset prepared by you. |  |
| * 1. Please, add this dataset to metadata catalog into your folder.   Also you could profile the file. |  |
| * 1. Please, pay attention, that Device Type is in Upper Case now.   Thus we can proceed to the next exercise. |  |

Data Preparation exercise is completed. Congratulations!

# 4. ENRICH DATA IN SAP DATA HUB

**Estimated time: 25 minutes**

**Objective**

In the following exercise, you will learn how to work with the SAP Data Hub Modeler. You will create a data-driven pipeline to enrich the missing country values to the device file.

**Overview SAP Data Hub Modeler**

The SAP Data Hub Modeler tool is based on the SAP Pipeline Engine that uses a flow- based programming paradigm to create data processing pipelines (graphs).

Big Data applications require advanced data ingestion and transformation capabilities.

Some common use cases are to:

* Ingest data from source systems. For example, database systems like SAP HANA, message queues like Apache Kafka, or data storage systems like HDFS or S3.
* Cleanse the data.
* Transform the data to a desired target schema.
* Store the data in target systems for consumption, archiving, or analysis.

Users can model data processing pipelines as a computation graph, which can help to achieve the required data ingestion and transformation capabilities. In this graph, nodes represent operations on the data, while edges represent the data flow.

The SAP Data Hub Modeler tool helps users to graphically model and execute a graph. The tool also provides a runtime component to execute graphs in a containerized environment that runs on Kubernetes.

The SAP Data Hub Modeler tool provides certain predefined operators for productive use cases. These operators can help users define graphs, including non terminating, non connected, or cyclic graphs. The following example shows a simple interaction with Apache Kafka. The graph consists of two subgraphs. The first subgraph generates some data and writes the data into a Kafka message queue, while the second subgraph reads the data from Kafka, converts it to string and prints the data to a terminal.

In the following exercise you will be a technical person who is using the SAP Data Hub Modeler to create a new device file which is enriched by the missing country data which comes from the customer file.

| Explanation | Screenshot |
| --- | --- |
| 1. Click on the URL of the **SAP Data Hub Launchpad** assigned to your group.   Provide **Tenant** name, same **Username** (TAxx) and **Password**.  Then click on **Login**. |  |
| 1. It will open the **SAP Data Hub Launchpad**, where you can see few applications running.   Now click on **Modeler** tile | Home | SAP Data Hub - Google Chrome |
| 1. Click the **+ icon** to create a new Pipeline.   Ensure that you selected the **Graphs** tab. | Modeler | SAP Data Hub - Google Chrome |
| 1. Click the save button on the top of the working panel. 2. Define the name   **dx.<user>.data\_transformation and description**  **DX Device country enrichment**  and press **OK**. |  |
| 1. Select the tab **Operators**. 2. Search for the operator**(1) Data Transform** | Modeler | SAP Data Hub - Google Chrome |
| 1. Select the operator and move it via Drag and Drop to the working area. 2. Open the operator via double-click on the operator. 3. New tab will be opened in side bar containing operators of type **Data Transform**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Select the operator **Data Source** and move it via Drag and Drop to the working area. 2. Change the name to **Device** by clicking into the Operator. 3. Open the operator via double-click. | Modeler | SAP Data Hub - Google Chrome |
| 1. Browse the **Connection ID**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Select **HDFS** | Modeler | SAP Data Hub - Google Chrome |
| 1. Browse the **Source**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Double-click the the folder **exercises\_xx** to navigate into it. | Modeler | SAP Data Hub - Google Chrome |
| 1. Select the file **which you prepared in the previous exercise** and press **OK**. |  |
| 1. Check the **Data Preview** |  |
| 1. Not all country fields are filled. 2. Press **close** 3. **Save** afterwards by clicking on the save icon. |  |
| 1. Navigate back |  |
| 1. Add a second **Data Source** and name it **Customer**. 2. Open the operator via double-click. | Modeler | SAP Data Hub - Google Chrome |
| 1. Browse the **Connection ID**. 2. Select **HDFS** 3. Browse the **Source**. 4. Double-click the folder **exercises\_xx** to navigate into it. 5. Select the file **customers.csv** and press **OK**. 6. Press **Save** and navigate back. |  |
| 1. Add a **Join** operator to your pipeline. 2. Link the **Devices** output port to the first input port from the **Join** operator. | Modeler | SAP Data Hub - Google Chrome |
| 1. Link the **Customer** output port to the second input port from the **Join** operator. | Modeler | SAP Data Hub - Google Chrome |
| 1. Double-click on the **Join** operator. | Modeler | SAP Data Hub - Google Chrome |
| 1. Link **Join1\_Input1** to **Join1\_Input2**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Define as a condition:   "Join1\_Input1"."CUSTOMER" = "Join1\_Input2"."CUSTOMER"  as your join statement.  Select as Join Type: **INNER** | Modeler | SAP Data Hub - Google Chrome |
| 1. Select the following fields via pressing the round icon in front of the name:   **Join1\_Input1 (Devices)**   * Devices * Type * Customer * Time   **Join1\_Input2 (Customer)**   * Country | Modeler | SAP Data Hub - Google Chrome |
| 1. If you switch to the **Columns** tab for column mapping you will see a graphical join. 2. Save the **Join**. 3. Navigate back to the pipeline. | Modeler | SAP Data Hub - Google Chrome |
| 1. Add a **Projection** operator the pipeline. 2. Link the Join1 operator to the **Projection** operator. 3. Double-click on the **Projection** operator. | Modeler | SAP Data Hub - Google Chrome |
| 1. Map the source fields to the target fields. You can press the mapping icon to get it automatically mapped. | Modeler | SAP Data Hub - Google Chrome |
| 1. Save the mapping. 2. Navigate back to the pipeline. | Modeler | SAP Data Hub - Google Chrome |
| 1. Add a **Data Target** operator to the pipeline. 2. Link the **Projection** operator to the **Data Target**. 3. Double-click the **Data Target** operator. | Modeler | SAP Data Hub - Google Chrome |
| 1. Browse the **Connection ID**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Select **HDFS** as our target system for the new file we want to create. | Modeler | SAP Data Hub - Google Chrome |
| 1. In this scenario we would like to create a new file on the HDFS system. To easy the process, we select first the existing device file and change the target folder afterwards. 2. Browse the **Target**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Select the folder **exercises\_xx**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Select the **devices.csv** file. 2. Press Ok. 3. Confirm the Warning with OK. | Modeler | SAP Data Hub - Google Chrome |
| 1. Change the Target from **/exercises\_xx/devices.csv** to **/exercises\_xx/Pipeline** 2. All other parameters are derived from the existing file. | Modeler | SAP Data Hub - Google Chrome |
| 1. Press the auto-mapping button. 2. Save your Data Target. 3. Navigate back. | Modeler | SAP Data Hub - Google Chrome |
| 1. Save your **Data Transform** pipeline 2. Navigate back. | Modeler | SAP Data Hub - Google Chrome |
| 1. Add a Workflow Trigger operator. 2. Add a **Workflow Terminator** operator. 3. Link the **Workflow Trigger** to the **Data Transform**. 4. Link the **Data Transform** to the **Workflow Terminator**. | Modeler | SAP Data Hub - Google Chrome |
| 1. Save it. | Modeler | SAP Data Hub - Google Chrome |
| 1. Run it. | Modeler | SAP Data Hub - Google Chrome |
| 1. You will see in the **Status** that the pipeline and the subgraph are in pending. | Modeler | SAP Data Hub - Google Chrome |
| 1. Wait until your pipeline and the subgraph is completed. | Modeler | SAP Data Hub - Google Chrome |
| 1. Navigate back to the SAP Data Hub Meta Data Explorer. Browse the connection and navigate to the newly created file. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. (1) Navigate to the HDFS connection and then to your folder **exercises\_xx**. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. You will find the newly created folder **Pipeline** in your main folder exercises\_xx. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. The newly created enriched file is now available. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. Index your new file and publish it in your catalog folder.   To structure your catalog folder better, you could create a new subfolder with the name **Pipeline** and publish this file into it.  You learned how to create a folder in the exercise before. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. After the file is published you can profile it. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. Have a look at the **Fact Sheet** afterwards. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. You see that the country column has no null values anymore. | SAP Data Hub Metadata Explorer - Google Chrome |
| 1. You can also check now the Data Preview with the filled country values. | SAP Data Hub Metadata Explorer - Google Chrome |

This conclude the exercise of a data engineer to enrich a data set by joining two different source files into one new target file.