SAP BusinessObjects

Text Data Processing Data Quality Management Blueprints User's Guide

- SAP Data Services 4.2 (14.2.0)

2013-05-09

SAP

2013-05-09

Contents

Introduction

1.1 Documentation set for SAP Data Services content objects

You should become familiar with all of the pieces of documentation that relate to the SAP Data Services blueprints and other content objects.

Document	What this document provides
Content Objects Summary	Lists all of the available blueprints and other content objects and the jobs and other objects that they contain.
Content Objects What's New	Highlights the new and enhanced blueprints and other content objects available for this release.
Data Quality Management Custom Functions User's Guide	Contains instructions for downloading and importing custom functions.
Data Quality Management Match Blueprints User's Guide	Contains a list of available Data Quality Management Match blueprints and instructions for downloading, configuring, and running them.
Data Quality Management Product Blueprints User's Guide	Contains a list of available Data Quality Management product blueprints and instructions for downloading, configuring, and running them.
Data Quality Management Regional Blueprints User's Guide	Contains a list of available Data Quality Management regional blueprints and instructions for downloading, configuring, and running them.
Text Data Processing Data Quality Management Blueprints User's Guide	Contains a list of available Text Data Processing Data Quality Management blueprints and instructions for downloading, configuring, and running them.
Text Data Processing Entity Extraction Dictionary File Generator User's Guide	Contains instructions for installing and using the Excel spreadsheet to generate and compile dictionary XML files used by the Entity Extraction transform.
Text Data Processing Language Blueprints User's Guide	Contains a list of available Text Data Processing Language blueprints and instructions for downloading, configuring, and running them.

Document	What this document provides
Text Data Processing Miscellaneous Blueprints User's Guide	Contains a list of available Text Data Processing Miscellaneous blueprints and instructions for downloading, configuring, and running them.

1.2 SAP information resources

A global network of SAP technology experts provides customer support, education, and consulting to ensure maximum information management benefit to your business.

Useful addresses at a glance:

Address	Content
Customer Support, Consulting, and Education services http://service.sap.com/	Information about SAP support programs, as well as links to technical articles, downloads, and online forums. Consulting services can provide you with information about how SAP can help maximize your information management investment. Education services can provide information about training options and modules. From traditional classroom learning to targeted e-learning seminars, SAP can offer a training package to suit your learning needs and preferred learning style.
Product documentation http://help.sap.com/bods/	SAP product documentation.
Supported Platforms (Product Availability Matrix) https://service.sap.com/PAM	Get information about supported platforms for SAP Data Services. Use the search function to search for Data Services. Click the link for the version of Data Services you are searching for.
SAP Data Services Community Network http://scn.sap.com/community/data-services	Get online and timely information about SAP Data Services, including forums, tips and tricks, additional downloads, samples, and much more. All content is to and from the community, so feel free to join in and contact us if you have a submission.
Blueprints http://scn.sap.com/docs/DOC-8820	Blueprints for you to download and modify to fit your needs. Each blueprint contains the necessary SAP Data Services project, jobs, data flows, file formats, sample data, template tables, and custom functions to run the data flows in your environment with only a few modifications.

1.3 Introduction to SAP Data Services 4.2 Content Objects

Welcome to SAP Data Services 4.2 version 14.2.0 Content Objects.

Data Services overview

SAP Data Services delivers a single enterprise-class solution for data integration, data quality, data profiling, and text data processing that allows you to integrate, transform, improve, and deliver trusted data to critical business processes. It provides one development UI, metadata repository, data connectivity layer, run-time environment, and management console—enabling IT organizations to lower total cost

of ownership and accelerate time to value. With SAP Data Services, IT organizations can maximize operational efficiency with a single solution to improve data quality and gain access to heterogeneous sources and applications.

Data Services Content Objects overview

We've identified a number of common scenarios that you are likely to perform with SAP Data Services. For each scenario, we've included a blueprint that is already set up to solve the business problem in that scenario. Each blueprint contains the necessary project, jobs, data flows, file formats, sample data, template tables, and custom functions to run the data flows in your environment with only a few modifications.

You can download the blueprint packages from the SAP Community Network. On the website, we periodically post new and updated blueprints, custom functions, best practices, whitepapers, and other content. You can refer to this site frequently for updated content and use the forums to provide us with any questions or requests you may have. We've also provided the ability for you to upload and share any content that you've developed with the rest of the SAP Data Services development community (for instructions on uploading content, see *How to Contribute* at https://www.sdn.sap.com/irj/scn/submitcontent).

Instructions for downloading and installing the content objects are also located on the SAP Community Network website.

Blueprints Overview

The Text Data Processing Data Quality Management blueprints illustrate a common text data processing with data quality scenario that you are likely to perform with SAP Data Services when you want to process unstructured text. For the scenario, we've included a blueprint that is already set up to demonstrate the text data processing Entity Extraction transform usage in that scenario. The blueprint contains the necessary project, jobs, data flows, file formats, sample data, template tables, and custom functions to run the jobs in your environment with only a few modifications.

Downloading Blueprint Packages

3.1 Blueprint versions

The following table shows the version of the Data Quality Management and Text Data Processing blueprints that can be used for SAP Data Services. The blueprint version is displayed on the SAP Data Services Blueprints page of the SAP Community Network website.

SAP Data Services version	Blueprint version	Blueprints available
4.2	4.2	Data Quality Management Text Data Processing
4.1.1	4.1.1	Data Quality Management Text Data Processing
4.1	4.1	Data Quality Management
XI 4.0	XI 4.0	Data Quality Text Data Processing
XI 3.2	XI 3.2	Data Quality
XI 3.1	XI 3.0	Data Quality
XI 3.0	XI 3.0	Data Quality

3.2 Available Text Data Processing Data Quality Management blueprints

Each blueprints package contains sample jobs configured to illustrate best practice settings for common use cases of Text Data Processing in conjunction with Data Quality Management. It also helps you visualize the extracted concepts and sentiments using an SAP BusinessObjects BI 4.0 Universe and SAP BusinessObjects Web Intelligence reports.

To see the contents of each blueprint, including jobs and custom functions, see the *Content Objects Summary*. To help you compare the available blueprints and decide which to download, see the following table.

Blueprint	Description
Text Data Processing Blueprints – Data Quality Management	Contains sample jobs configured to illustrate the use of Text Data Processing in conjunction with Data Quality Management. It also helps you visualize the extracted concepts and sentiments using an SAP BusinessObjects BI 4.0 Universe and SAP BusinessObjects Web Intelligence reports.

3.3 Downloading and setting up blueprints

These are the general steps for downloading and setting up Text Data Processing Data Quality Management blueprint packages for SAP Data Services.

1. To access the SAP Community Network website, go to https://www.sdn.sap.com/irj/boc/blueprints in your web browser.
2. Log into your account using your username and password, or create a new account.
3. Review the list of available blueprint packages and other content objects and their descriptions to decide which to download.
4. Select the blueprint package that you want to download. A new page opens.
5. Click the **View this Code Sample** button.
6. In the File Download window, save the `.zip` file to the `Tutorial Files` folder in your installed SAP Data Services path. By default, this folder is installed to `\Program Files\SAP BusinessObjects\Data Services\Tutorial Files` for 32-bit Windows and `\Program Files (x86)\SAP BusinessObjects\Data Services\Tutorial Files` for 64-bit Windows.

   If you are running on UNIX, the `Tutorial Files` folder exists only on the Windows client workstation, and you should download the `.zip` file there.
7. In the `Tutorial Files` folder in Windows Explorer, right-click the `.zip` file and select to extract the compressed (zipped) folders to the Tutorial Files folder. For example, if you use WinZip for file compression, right-click the .zip file and select **WinZip > Extract to here**.

   Extracting creates subfolders and places the files in the appropriate location. The `.atl` file is saved to the `Text Data Processing Samples` folder, and the sample data files are saved to the `Text Data Processing Samples\<subfolder>` folder.

8. In the Designer, import the `.atl` file. In the Passphrase window, enter the name of the `.atl` file without the extension (for example, when importing `tdp_blueprints_data_quality.atl`, enter the passphrase tdp_blueprints_data_quality) and click **Import**. Click **OK** to close the warning window.

   Importing the file adds a project called TextDataProcessingBlueprintsDataQuality to your object library. The project contains jobs whose names begin with TdpBlueprintDqm*Xxx* and contain in their name the Text Data Processing use case that they illustrate. Each job contains a data flow. The import also adds two datastores called TextDataProcessingBlueprintsDqm and TextDataProcessingBlueprintsDqmIQ to your object library, and file formats called TdpDqmIn*Xxx* and TdpDqmOut*Xxx*, where *Xxx* is the name of the Text Data Processing use case for the sample input and output data.
9. If you are running on UNIX, copy the input files to the job server machine and create the same folder structure that is on the Windows client workstation.
10. If you imported the blueprint `.atl` files using a Data Services Designer on 32-bit Windows and use a job server on 64-bit Windows, then you must copy the blueprint files to the Data Services installation of the job server machine.

The Text Data Processing Blueprints - Data Quality Management blueprint is packaged with an SAP BusinessObjects BI 4.0 Universe and SAP BusinessObjects Web Intelligence reports for visualizing the output of the TdpBlueprintDqm_VocMatch or TdpBlueprintDqm_VocMatchIQ job. To see the reports, you must have SAP BusinessObjects BI 4.0 installed.

Related Topics

- Editing the datastore

Configuring and Running Jobs

4.1 Editing the datastore

After you download the blueprint `.zip` file to the appropriate folder, unzip it, and import the `.atl` file in the Designer, you must edit the TextDataProcessingBlueprintsDqm or TextDataProcessingBlueprintsDqmIQ datastore.

Typically, you would decide whether to use the Microsoft SQL Server or the Sybase IQ version, and then edit one of the datastores. However, the blueprint package has been created in such a way that you can configure both datastores and run both jobs without overwriting anything.

The database that you use for running the blueprints does not need to be the same database that is used for the SAP Data Services repository. It can be a locally installed database system or any shared database system that you have access to create tables in and read from those tables.

Related Topics

- Microsoft SQL Server
- Sybase IQ
- Other database types

4.1.1 Microsoft SQL Server

If you have access to write and read data to tables in Microsoft SQL Server 2000, 2005, or 2008, complete the following steps.

1. Select the **Datastores** tab of the Local Object Library, right-click the TextDataProcessingBlueprintsDqm datastore, and select **Edit**.
2. In the **Edit TextDataProcessingBlueprintsDqm** window, enter your repository connection information in place of the four **CHANGE_THIS** values.
3. Click **OK**. If the window closes without an error message, then the database is successfully connected.

4.1.2 Sybase IQ

If you have access to write and read data to tables in Sybase IQ 15.0, 15.1, 15.2, 15.3, or 15.4, complete the following steps.

Note:

- **dbspace Sizing**
  - Current version of Sybase IQ: By default, an "iqdemo" database on Sybase IQ 15.3 or 15.4 has 25MB in the IQ_SYSTEM_TEMP dbspace, and 100MB in both the iq_main and IQ_SYSTEM_MAIN dbspaces. All three dbspaces have 200MB of reserve space. This configuration should be sufficient to run the Text Data Processing Blueprints - Data Quality Management blueprint against a standard Sybase IQ installation with the default "iqdemo" database running.
  - Previous versions of Sybase IQ 15.x (15.0, 15.1, and 15.2): it is recommended that you increase the size of the IQ_SYSTEM_TEMP dbspace. The SQL syntax used to increase the size of the IQ_SYSTEM_TEMP dbspace is:

    ```
    alter dbspace IQ_SYSTEM_TEMP add file IQ_SYSTEM_TEMP_2 'iqdemo_2.iqtmp'
    size 25 mb reserve 200 mb
    ```

- **Table Owner**

  By default, the table owner is set to dba for the Sybase IQ database tables used in the TdpBlueprintDqm_VocMatchIQ job. If you require a different table owner, follow the steps in the Other database types section.

1. Select the **Datastores** tab of the Local Object Library, right-click the TextDataProcessingBlueprintsDqmIQ datastore, and select **Edit**.
2. In the **Edit TextDataProcessingBlueprintsDqmIQ** window, enter your repository connection information in place of the three **CHANGE_THIS** values (Data Source, User Name, and Password).
3. Click **OK**. If the window closes without an error message, then the database is successfully connected.

4.1.3 Other database types

If you have access to write and read data to tables in another database system (other than Microsoft SQL Server or Sybase IQ), complete the following steps.

1. Select the **Datastores** tab of the Local Object Library, expand the TextDataProcessingBlueprintsDqm or TextDataProcessingBlueprintsDqmIQ datastore, and expand the **Template Tables** subfolder.
2. Make note of the names of the datastore, template tables, and dataflows in which the template tables are used. In this blueprint, all of the template tables are used only in the TdpBlueprintDqm_VocMatch or TdpBlueprintDqm_VocMatchIQ data flow.

3. Delete all of the template tables. Right-click a template table (for example, TDP_BLUEPRINTS_DQM_VOCMATCH_FEEDBACK), select **Delete**, and select **Yes** to confirm your selection.
4. Delete the appropriate datastore. (You can update either job, since both can be used with other database types. It depends on which datastore that you want to update, and based on that, which datastore should be deleted, re-created, and used to create new template tables.) Right-click TextDataProcessingBlueprintsDqm or TextDataProcessingBlueprintsDqmIQ, select **Delete**, and select **Yes** to confirm your selection.
5. Create a new datastore with the same name as the one you just deleted. In the **Datastores** tab of the Local Object Library, right-click in the white space and select **New**. In the **Datastore** name field, enter the name TextDataProcessingBlueprintsDqm or TextDataProcessingBlueprintsDqmIQ, depending on which job you are updating. In the Database type field, select your database system. Complete the remaining fields with the connection information to the database that you have access to.
6. Click **OK** to close the Create New Datastore window.
7. Open the TdpBlueprintDqm_VocMatch or TdpBlueprintDqm_VocMatchIQ dataflow and delete the target. Then add a new template table with the same name by selecting the Template Table icon from the buttons on the right menu and clicking the dataflow canvas. In the Create Template window, enter the name of the template table that you deleted and select the TextDataProcessingBlueprintsDqm or TextDataProcessingBlueprintsDqmIQ datastore in the In datastore field. Click **OK** to close the Create Template window. Connect the last transform to the template table.
8. Repeat step 7 for each of the target tables.

4.2 Verifying the substitution parameters

Before you run the sample jobs, verify that the **[$$SamplesInstall]** substitution parameter is set to the DataServices installation directory.

4.3 Running the jobs

Before you run the TdpBlueprintDqm_AddressDataCleanse job, you should have already completed the following tasks:

1. Copy the U.S. address cleanse reference files.
2. Install the PERSON_FIRM cleansing package.
3. Set the accurate value in the substitution parameter configuration Configuration1:
   - $$RefFilesAddressCleanse—Enter the path location where you copied the address cleanse reference files.

4.4 Viewing job output data using Interactive Analysis Desktop reports

After you run the TdpBlueprintDqm_VocMatch or TdpBlueprintDqm_VocMatchIQ job and the tables are created in your datastore, follow the steps below to view the output data using Interactive Analysis Desktop reports. You can use the same Universe and set of reports with Microsoft SQL Server or Sybase IQ; the only difference is the connection setup in the Universe, which points to either a Microsoft SQL Server database or a Sybase IQ database.

1. Launch the SAP BusinessObjects BI 4.0 Universe design tool, log in to your CMS repository, and open the `VocMatch_Universe.unv` universe, by default located in the `\Program Files (x86)\SAP BusinessObjects\Data Services\Tutorial Files\Text Data Processing Samples\Data Quality\VocMatch\Universe` folder.
2. To connect the Universe to your datastore, select **File > Parameters**. In the Universe Parameters window, click **New** and create a new secured connection to your datastore where all the target tables of the TdpBlueprintDqm_VocMatch or TdpBlueprintDqm_VocMatchIQ job are created.
3. After you create the secured connection, click **Test** in the Universe Parameters window to verify the database connection. Click **OK**.
4. To export the Universe to your CMS, select **File > Export**. In the Export Universe window, select the root of the CMS repository as the domain and click **OK**.

   The Universe should be exported successfully.
5. Launch the SAP BusinessObjects Enterprise XI 4.0 Interactive Analysis Desktop, log in to your CMS repository, and open the `Most-Mentioned-Concepts.wid` and `Concept-Details.wid` reports, by default located in the `\Program Files (x86)\SAP BusinessObjects\Data Services\Tutorial Files\Text Data Processing Samples\Data Quality\VocMatch\Reports`.
6. To refresh the data in the reports, click **Refresh**.

   The reports are pre-configured to use the VocMatch_Universe universe.
7. In the Concept-Details report, select the concepts from the Concept list or search for concepts and add them to the Concept(s) box. Click **OK**.

   More information about what each of the reports shows is explained through information boxes within the reports.

Index