
CS –32

**Data Warehousing with
SQL Server 2012**

**04. Enforcing Data Quality,
Extending SQL Server Integration
Services**

Enforcing Data Quality, Extending SQL Server Integration Services

- Data quality refers to the overall utility of a dataset(s) as a function of its ability to be easily processed and analyzed for other uses, usually by a database, data warehouse, or data analytics system.

What do I need to know about data quality?

- Quality data is useful data. To be of high quality, data must be consistent and unambiguous. Data quality issues are often the result of database merges or systems or cloud integration processes in which data fields that should be compatible are not due to schema or format inconsistencies.

What do I need to know about data quality?

- Data that is not high quality can undergo data cleansing to raise its quality.

Inconsistencies. Data that is not high quality can undergo data cleansing to raise its quality.

What activities are involved in data quality?

- Data quality activities involve data rationalization and validation.
- Data quality efforts are often needed while integrating disparate applications that occur during merger and acquisition activities, but also when soled data systems within a single organization are brought together for the first time in a data warehouse or big data lake.

What activities are involved in data quality?

- Data quality is also critical to the efficiency of horizontal business applications such as enterprise resource planning (ERP) or customer relationship management (CRM).

What are the benefits of data quality?

- When data is of excellent quality, it can be easily processed and analyzed, leading to insights that help the organization make better decisions.
- High-quality data is essential to business intelligence efforts and other types of data analytics, as well as better operational efficiency.

- Before you start any data quality or MDM project, you have to understand the sources and destinations of problematic data.
- Therefore, data quality activities must include overviews. You must make an extensive overview of all of the schemas of the databases that pertain to the problem data.
- You should interview domain experts and users of data. This is especially important for gaining a comprehension of the quality of the schema dimensions.

- In addition, after this step, you should also have a clear understanding of the technology that the enterprise is using.
- If necessary, you might need to plan to include appropriate technology experts in the project.
- During the overview of the data, you should also focus on the data life cycle so that you can understand retention retention periods and similar.

Using Data quality service to cleanse data:

- Data cleansing is the process of analyzing the quality of data in a data source, manually approving/rejecting the suggestions by the system, and thereby making changes to the data.
- Data cleansing in data Quality Services (DQS) includes a computer-assisted ...Cont

Using Data quality service to cleanse data:

process that analyzes how data conforms to the knowledge in a knowledge base, and an interactive process that enables the data steward to review and modify computer-assisted process results to ensure that the data cleansing is exactly as they want to be done.

Using Data quality service to cleanse data:

- The data steward can also perform data cleansing in the Integration Services packaging process.
- In this case, the data steward would use the DQS Cleansing component in Integration Services that automatically performs data cleansing using an existing knowledge base.

The data cleansing feature in DQS has the following benefits:

- Identifies incomplete or incorrect data in your data source (Excel file or SQL Server database), and then corrects or alerts you about the invalid data.
- Provides two - step process to cleanse the data:
 - Computer-assisted
 - Interactive

The data cleansing feature in DQS has the following benefits:

- **Computer-assisted :**
 - The computer-assisted process uses the knowledge in a DQS knowledge base to automatically process the data, and suggest replacements/corrections.

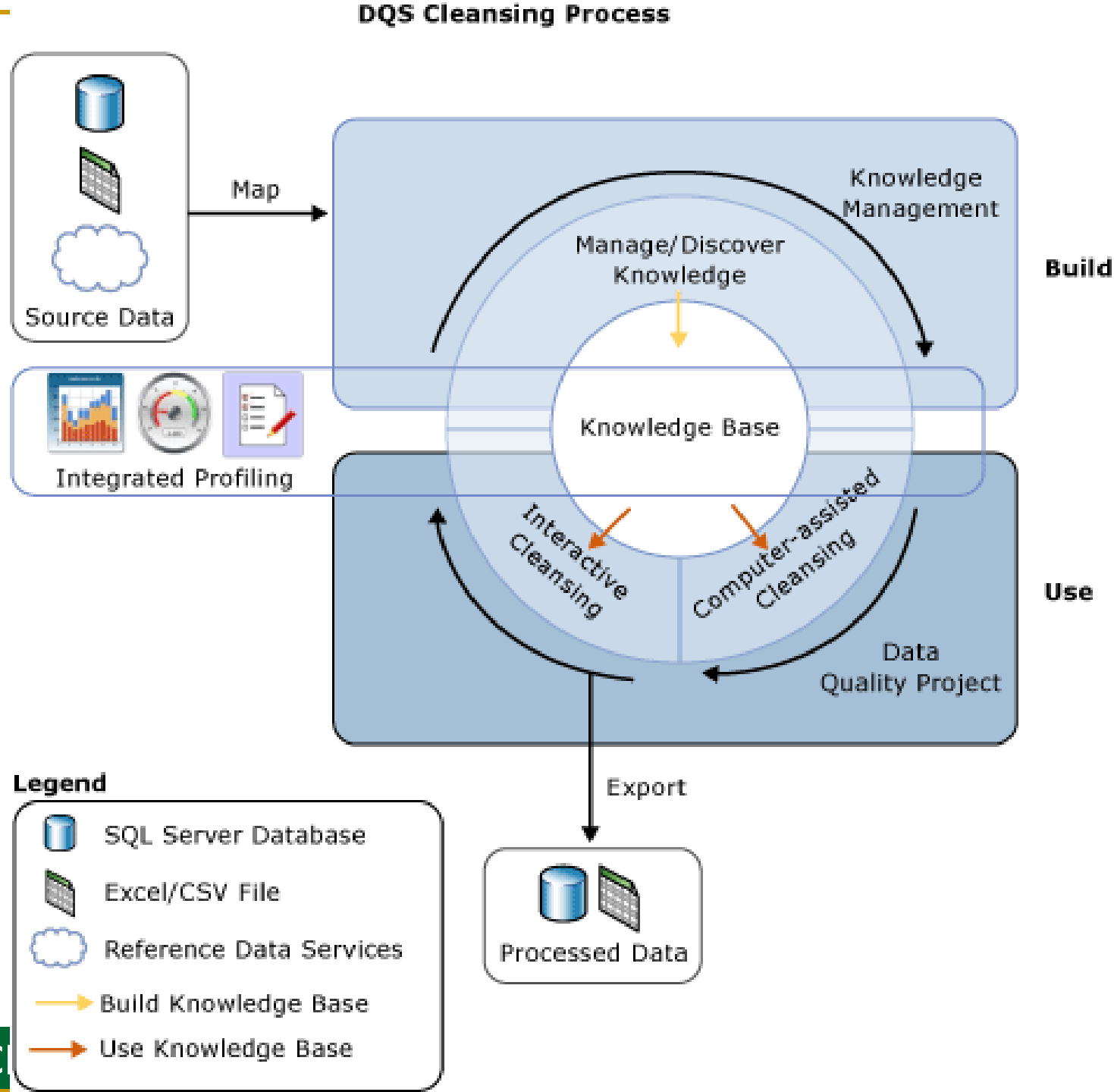
The data cleansing feature in DQS has the following benefits:

- **Interactive :**
 - The next step, interactive, allows the data steward to approve, reject, or modify the changes proposed by the DQS during the computer-assisted cleansing.

The data cleansing feature in DQS has the following benefits:

- Standardizes and enriches customer data by using domain values, domain rules, and reference data.
- For example,
 - ▣ Standardize term usage by changing “St.” to “Street”, enrich data by filling in missing elements by changing.
- Provides a simple, intuitive, and consistent wizard-like interface to the user to navigate data and inspect errors amongst a very large set of data.

The following illustration displays how data cleansing is done in DQS:



Computer-assisted Cleansing :

- The DQS data cleansing process applies the knowledge base to the data to be cleansed, and proposes changes to the data.
- The data steward has access to each proposed change, enabling him or her to assess and correct the changes.

Computer-assisted Cleansing :

- To perform data cleansing , the data steward proceeds as follows:
 1. Create a data quality project, select a knowledge base against which you want to analyze and cleanse your source data, and select the **Cleansing** activity. Multiple data quality projects can use the same knowledge base.

Computer-assisted Cleansing :

2. Specify the database table/view or an Excel file that contains the source data to be cleansed. The database or the Excel file can be the same one that was used for knowledge discovery, or it can be a different database or Excel file.

Computer-assisted Cleansing :

3. Map the data fields to be cleansed to appropriate domains/composite domains in the knowledge base. If you map a field to a composite domain, the mapping happens between the field and the composite domain, and not with the individual domains in the ...Cont

Computer-assisted Cleansing :

composite domain, and not with the individual domains in the composite domain. Also, the data cleansing for the mapped field is done based on the rules specified for the composite domain, and not for the individual domain in the composite domain.

Computer-assisted Cleansing :

4. Run the computer-assisted cleansing process by clicking **Start** on the **Cleanse** page.

- The data cleansing process finds the best match of an instance of data to known data domain values. The process applies data quality knowledge to all source data, unlike the knowledge discovery process, which runs on a percentage of the sample data.

Computer-assisted Cleansing :

- The computer-assisted process display data quality information in Data Quality Client that will be used for the interactive cleansing process.
- Apart from the adherence to the syntax error rules, DQL also uses reference data and advanced algorithms to categorize data using confidence level.

Computer-assisted Cleansing :

- The confidence level indicates the extent of certainty of DQS for the correction or suggestion. The confidence level indicates the extent of certainty of DQS for the correction or suggestion.

Computer-assisted Cleansing :

- The confidence level is based on the following threshold values:
 - **An auto-correction threshold** value above which DQS will suggest a change and make it unless the data steward rejects it.
 - We can specify the auto correction threshold value in the **General Settings** tab in the **Configuration** screen.

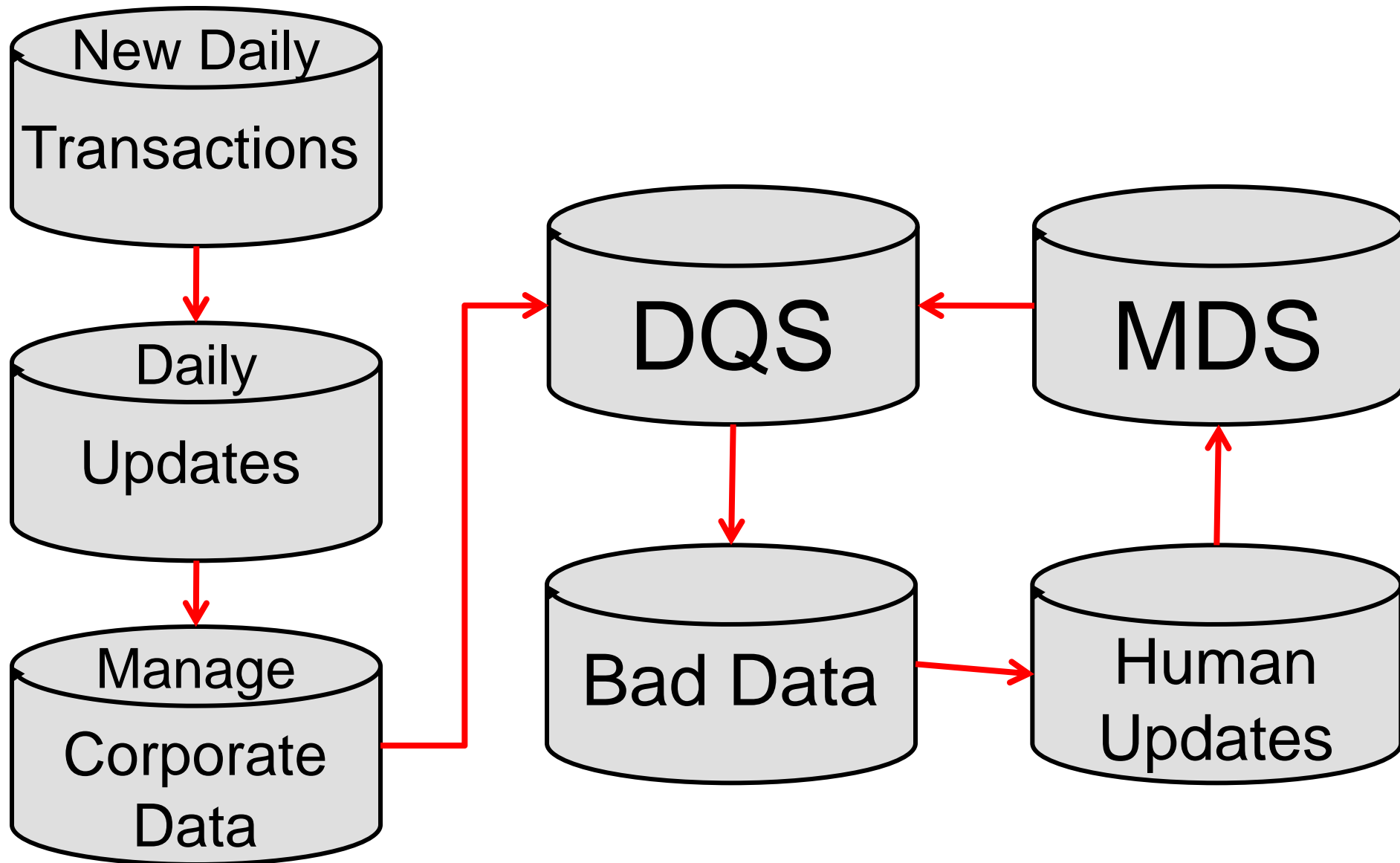
Computer-assisted Cleansing :

- ❑ **An auto-suggestion threshold** value, below the auto-correction threshold, above which DQS will suggest a change, and make it if the data steward approves it. You can specify the auto suggestion threshold value in the **General Settings** tab in the **Configuration** screen.

High Level Strategy Approach

- Any value having a confidence level below the auto-suggestion threshold value is left as is by DQS unless the data steward specifies a change.
- This is where “Data Quality Services” comes into play. Now that we have the necessary background of my client’s issue, let us look at the high level view of the strategy that we need to follow. it is **High Level Strategy Approach.**

High Level Strategy Approach



High Level Strategy Approach

- The diagram above shows the “High Level Strategy Approach” that we recommended to the client.
- For our discussion, we begin by downloading the master data services (MDS) data to Data Quality Services (DQS).

High Level Strategy Approach

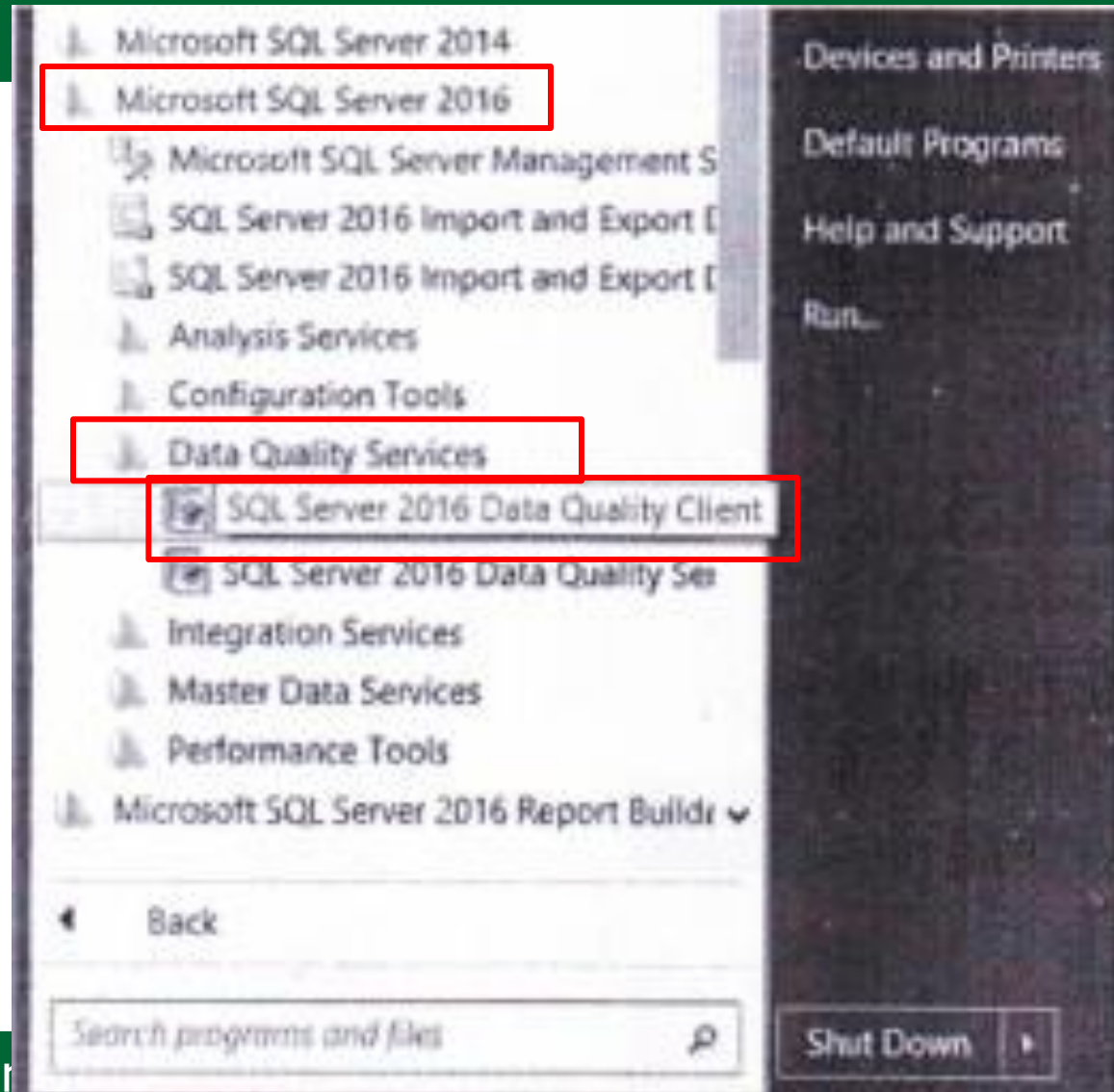
- For those of us who are not all the familiar with the product, Data quality services is an intelligent piece of software that learns to recognise data patterns and in doing so is able to flag anomalous data.
- In fact DQS is able to send errors to an “Error” tables so that Business Analysts and Data Stewards are able to inspect the data and to rectify any errors.

Setting up the necessary infrastructure within Data Quality Services.

Start Menu

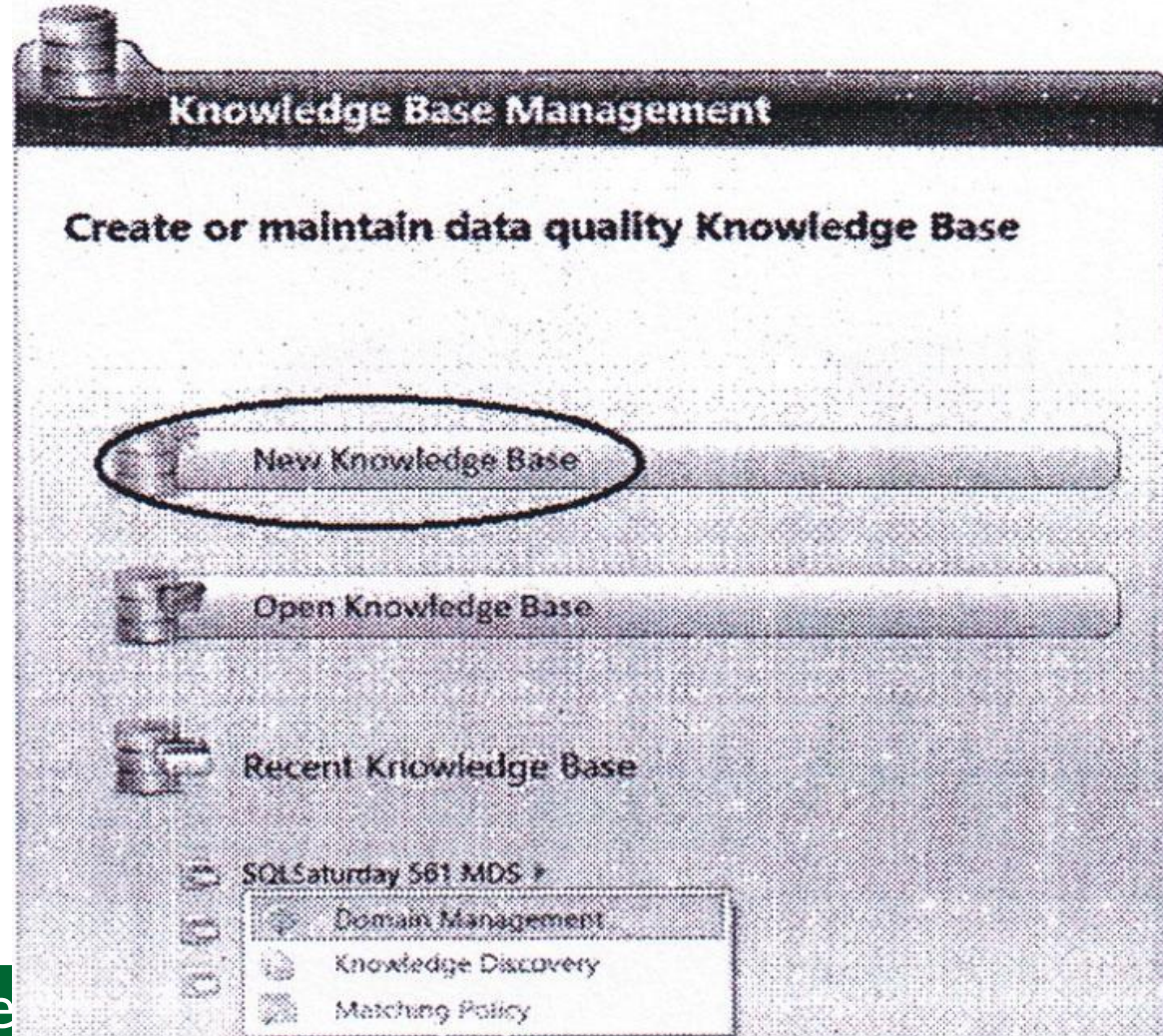
SQL Server

Data Quality
Client



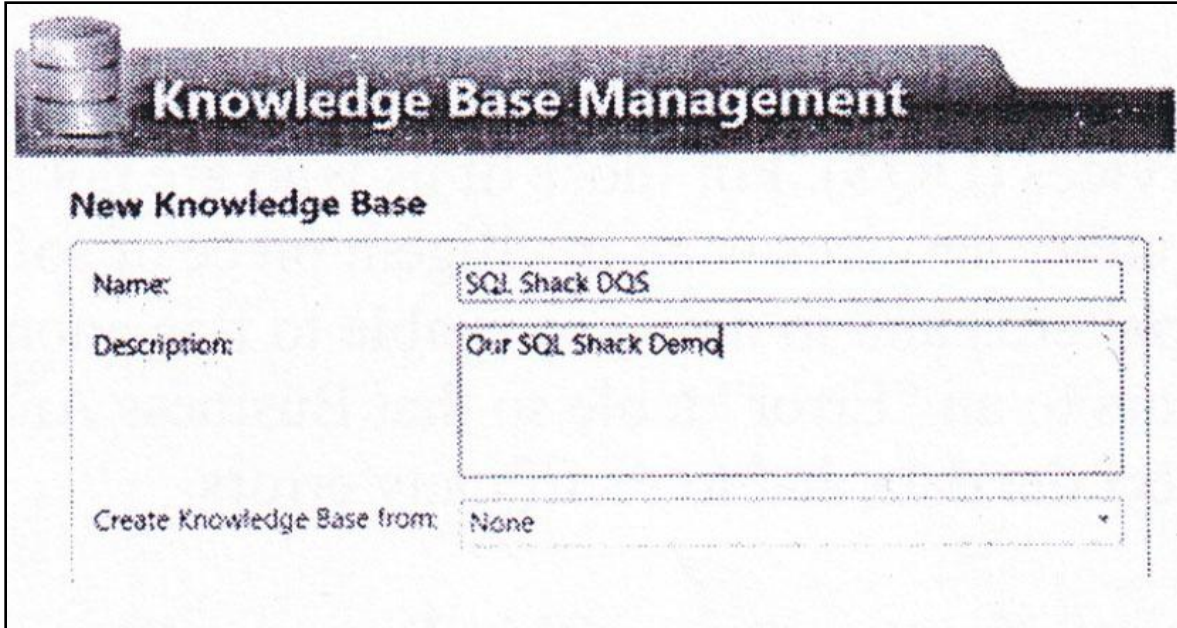
Setting up the necessary infrastructure within Data Quality Services.

- Opening the Data Quality Services Client we arrive at the home page. We click “New Knowledge Base”.



Setting up the necessary infrastructure within Data Quality Services.

- We give our new “Knowledge Base” a name (see above)

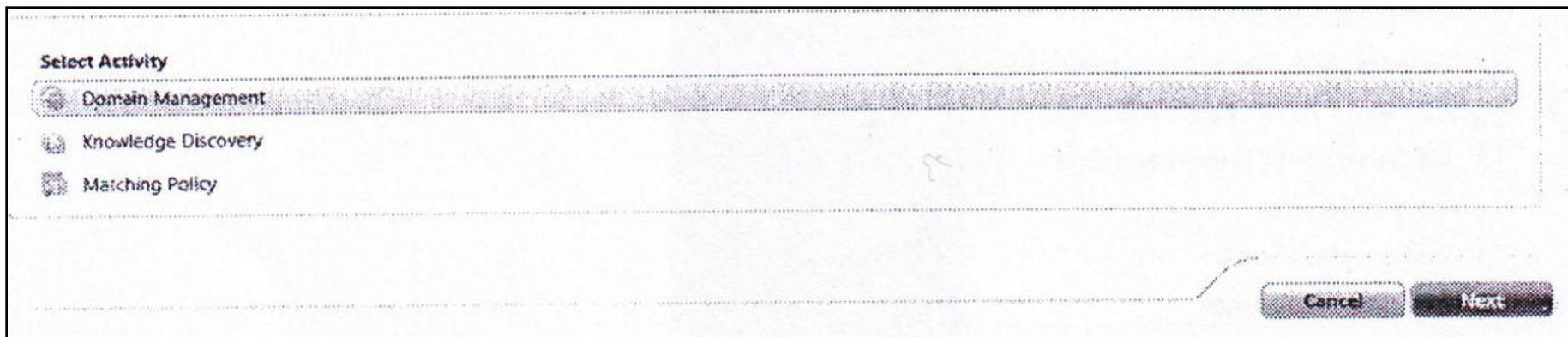


The screenshot displays the 'Knowledge Base Management' interface. At the top, there is a header with a database icon and the text 'Knowledge Base Management'. Below this, the 'New Knowledge Base' form is visible. The form contains three fields: 'Name:' with the value 'SQL Shack DQS', 'Description:' with the value 'Our SQL Shack Demo', and 'Create Knowledge Base from:' with a dropdown menu set to 'None'.

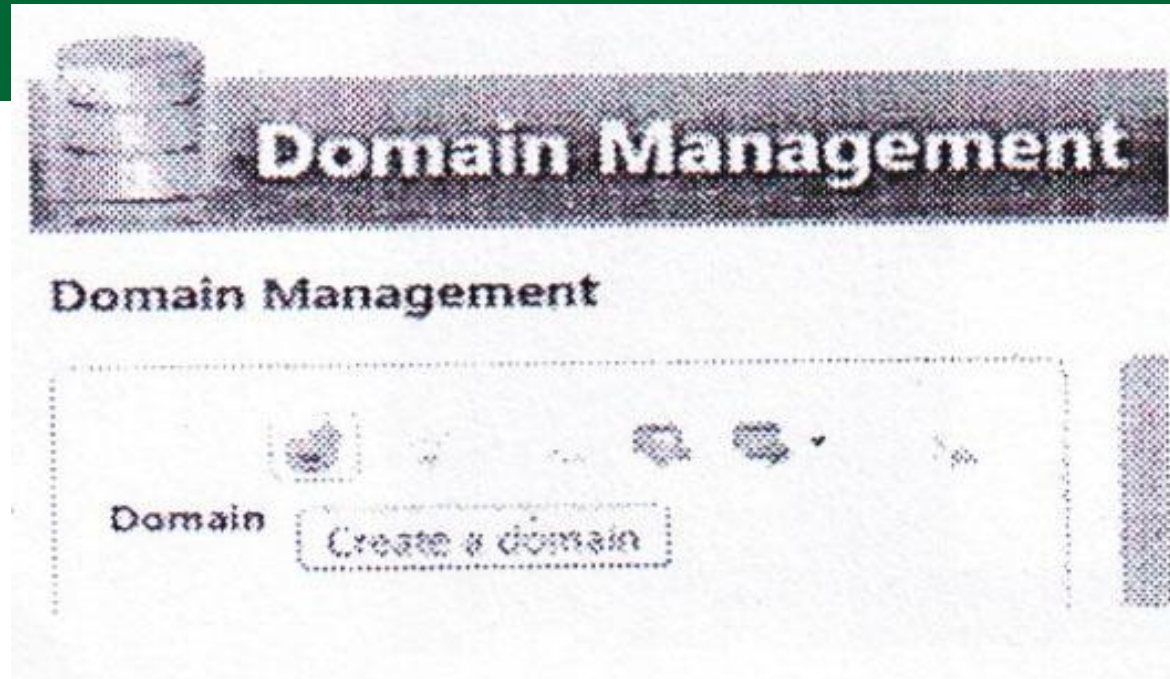
New Knowledge Base	
Name:	SQL Shack DQS
Description:	Our SQL Shack Demo
Create Knowledge Base from:	None

Setting up the necessary infrastructure within Data Quality Services.

- We click "NEXT" to continue.



Setting up the necessary infrastructure within Data Quality Services.



- We are taken to the "Domain Management" page. A domain is a nice word for a "table" (see above).

Setting up the necessary infrastructure within Data Quality Services.

- The important point being that ***these domains will contain the ONLY permissible attribute values.***
- If the attribute value that is entered by the user is NOT one in this table, then the whole record will be flagged by DQS as an error and thus invalid. We click "Create a domain".

Data Quality Services

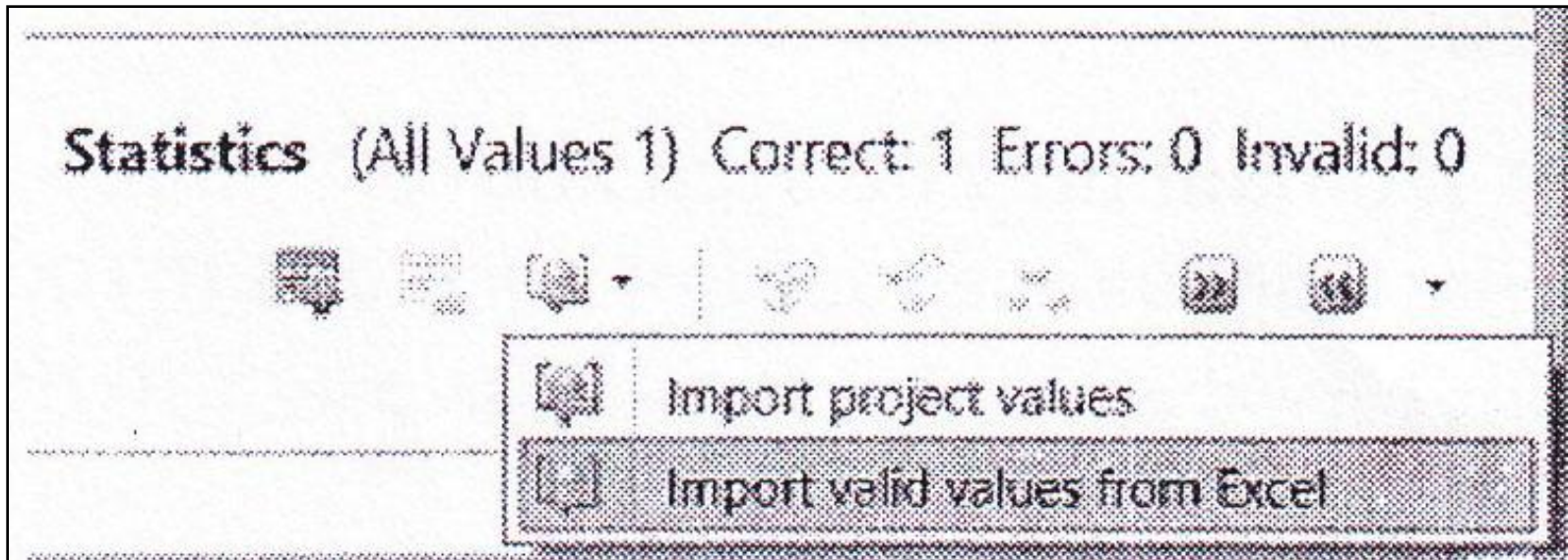
■ We give our "Domain" a name and define its type. Color will be an Entity on its own HOWEVER it is also an attribute of "Product" as our products have "Color". We click "OK" to continue.

The screenshot shows a "Create Domain" dialog box with the following fields and options:

- Domain Name: Color (Required)
- Description: (Empty text area)
- Data Type: String
- Use Leading Values
- Normalize String
- Format Output to: None
- Language: English
- Enable Speller
- Disable Syntax Error Algorithms

Buttons at the bottom: OK, Cancel, Help

Data Quality Services



■ We find ourselves on the "Color" design surface. We click the "Domain Values" tab. Now here is the weird part. We need to import our "Master List" of colours. Thus far, the only way to do so is from a spreadsheet. I have brought this issue up with Microsoft a few years back. The "Color" spreadsheet may be seen below:

Data Quality Services

A1	A	B
1	Color	
2	Black	
3	Blue	
4	Brown	
5	Green	
6	Grey	
7	Light Green	
8	N/A	
9	Navy	
10	Pink	
11	Purple	
12	Red	
13	Tan	
14	Teal	
15	White	
16	Yellow	

Color

Domain Properties Reference Data Domain Rules Domain

Find: Filter: All Values Show Only New

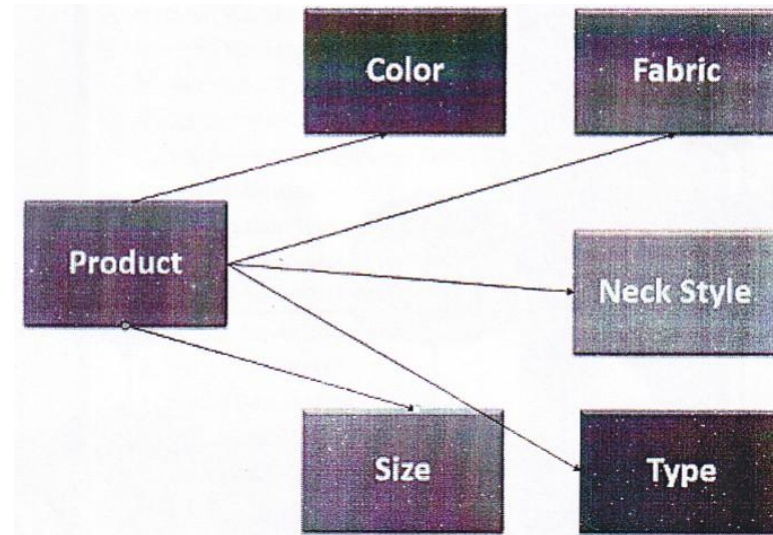
Value	Type	Correct to
Black	✓	-
Blue	✓	-
Brown	✓	-
Green	✓	-
Grey	✓	-
Light Green	✓	-
N/A	✓	-
Navy	✓	-
Pink	✓	-
Purple	✓	-
Red	✓	-
Tan	✓	-
Teal	✓	-

Data Quality Services

- The populated Color Domain may be seen above.
- Let us now set a "Domain Rule" that a colour cannot be "Pers" (Purple).



Data Quality Services



- Product is our main table (see above) and each product has several attributes (Colour, Fabric, Neck Style, Type and Size). Each of these attributes forms its own Entity (table).

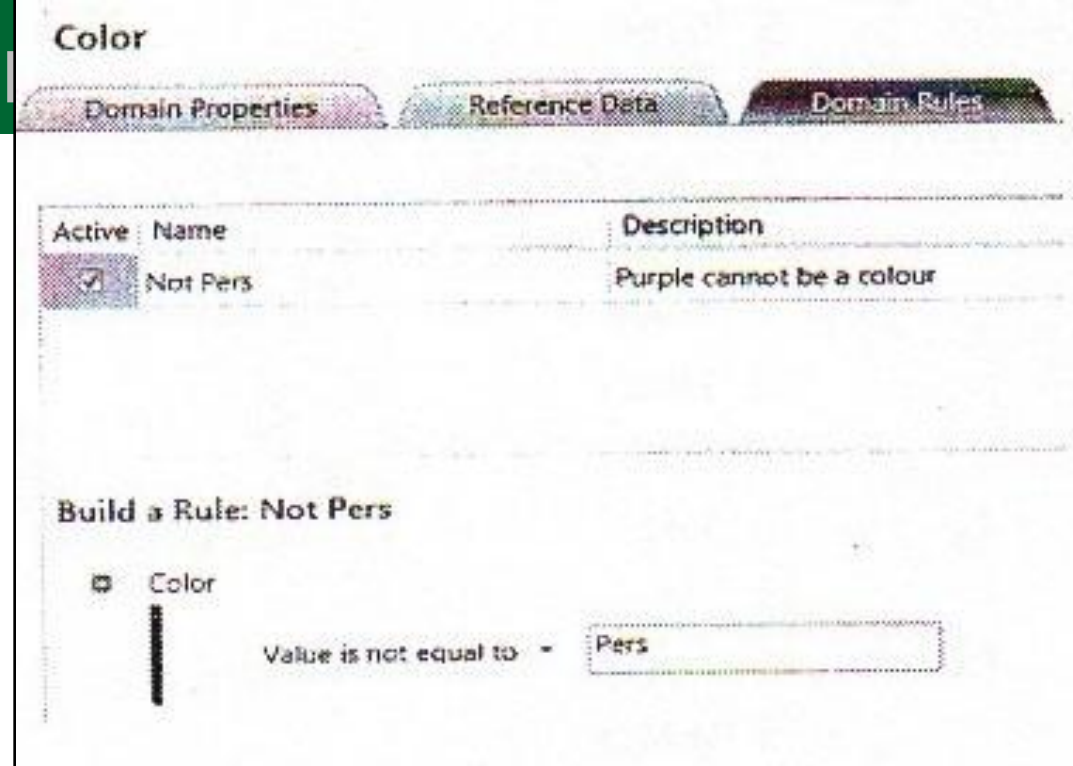
Data Quality Services

- We choose the “Product” Entity (or for us “older folks” a table).
- We see the data from the product entity (see above).
- In our simple example, users enter raw data (in the form of transactions) and this data is eventually moved to master data services. Now should a user have entered “Pers” or purple for the colour, this value must be deemed “Invalid”.

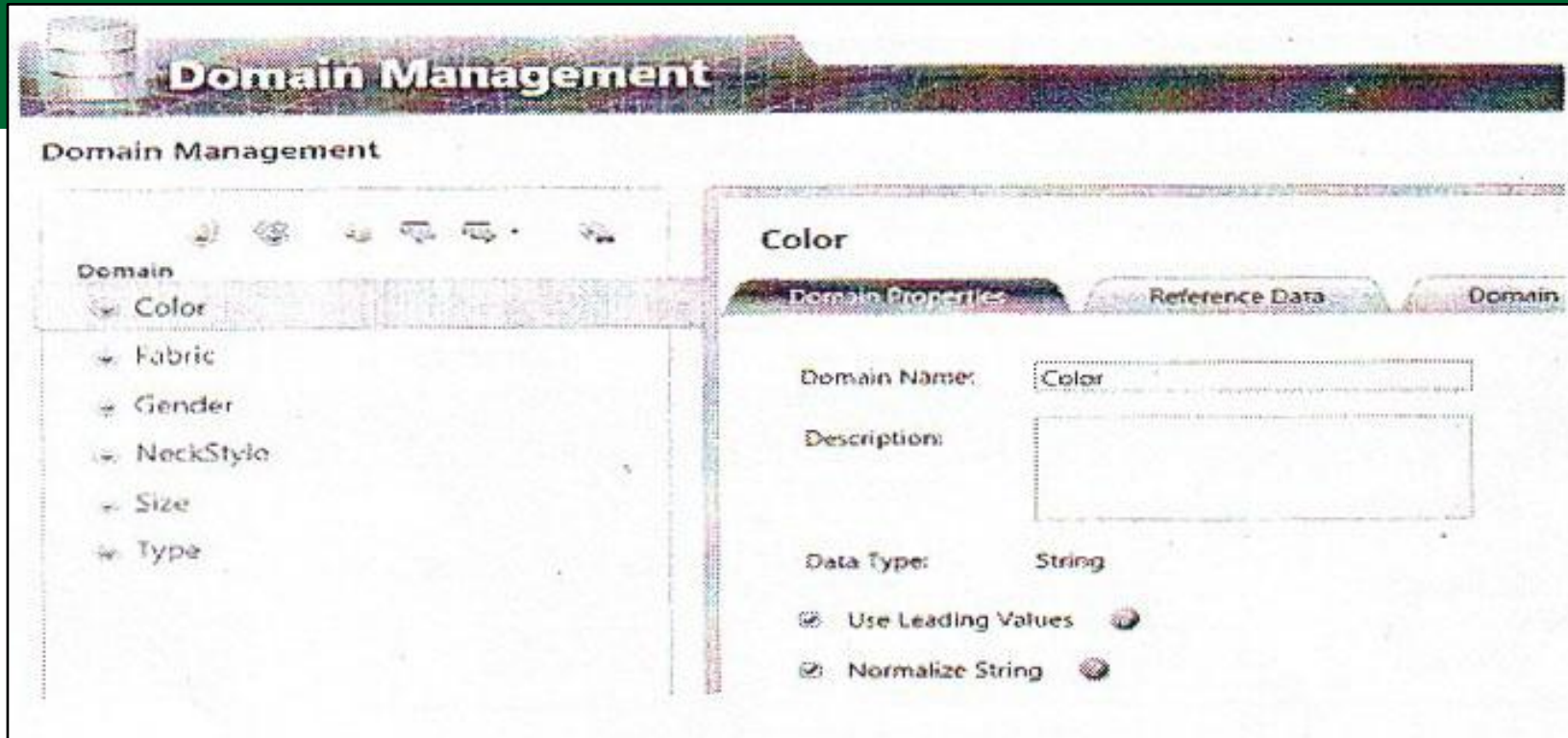
Data Quality Services

- So how do we recognize that we have an issue and how do we flag this type of data issue so that the necessary folks are able to rectify the data?
- This is where “Data quality services” comes into play.
- Now that we have the necessary background of my client’s issue, let us look at a high level view of the strategy that we need to follow.

Data Quality Se

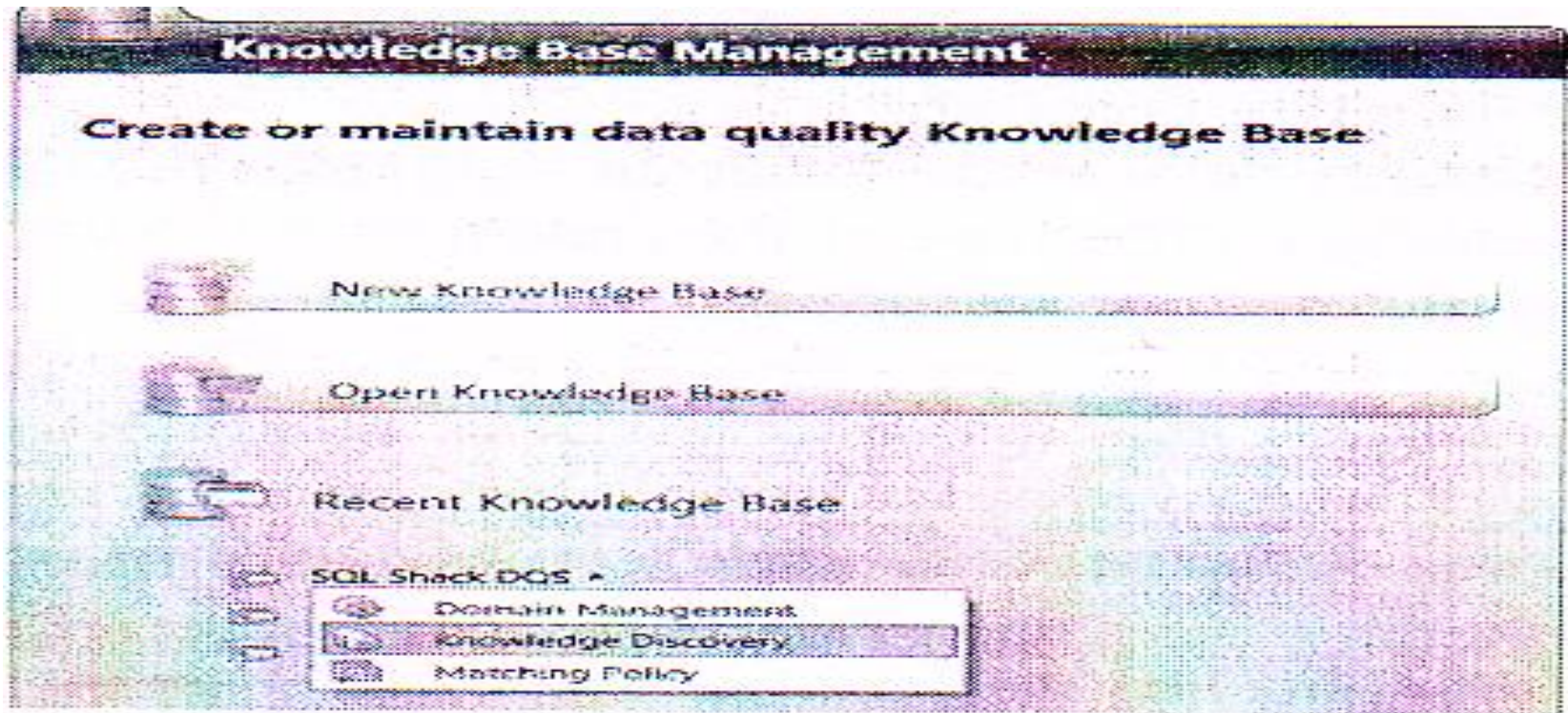


- The new "Domain Rule" dialog box opens. We give our rule the name "Not pers" and give the rule a description. Finally we set the rule: "Value is not equal to "Pers" (see above). We then "Apply the rule" to leave the screen. We repeat the same process for the remaining Entities / Domains as may be seen below:



- Having created the "Fabric", "Gender", "NeckStyle", "Size" and "Type" Domains we click "Finish" to leave the Domain Management Screen. We click "Publish".

Knowledge discovery and matching



- Having completed the "Domain Management" portion of the process, we are now in a position to cover the "Knowledge Discovery" and "Matching Policy" portions of the process.

Knowledge discovery and matching

- As I have covered “Knowledge Discovery” and “Matching Policy” exhaustively within another “fire side chat” that we had a year or so back, the reader is referred to the article “How clean is your data”.
- At this point, we shall assume that our “Knowledge Base” is complete (having finished the “Knowledge Discovery” and “Matching Policy” processes).

Using Data quality service to match data :

- The Data Quality Services (DQS) data matching process enables you to reduce data duplication and improve data accuracy in a data source. Matching analyzes the degree of duplication in all records of a single data source, returning weighted probabilities of a match between each set of record compared. You can then decide which records are matches and take the appropriate action on the source data.

Using Data quality service to match data :

- The DQS matching process has the following benefits:
 - Matching enables you to eliminate differences between data values that should be equal, determining the correct value and reducing the errors that data differences can cause.
 - For example, names and addresses are often the identifying data for a data source, particularly customer data, but the data can become dirty and deteriorate over time.

Using Data quality service to match data :

- The DQS matching process has the following benefits:
 - Performing matching to identify and correct these errors can make data use and maintenance much easier.
 - Matching enables you to ensure that values that are equivalent, but were entered in a different format or style, are rendered uniform.

Using Data quality service to match data :

- The DQS matching process has the following benefits:
 - Matching identifies exact and approximate matches, enabling you to remove duplicate data as you define it. You define which fields are assessed for matching, and which are not.

Using Data quality service to match data :

- The DQS matching process has the following benefits:
 - DQS enables you to create a matching policy using a computer-assisted process, modify it interactively based upon matching results, and add it to a knowledge base that is reusable.

Using Data quality service to match data :

- The DQS matching process has the following benefits:
 - You can re-index data copied from the source to the staging table, or not re-index, depending on the state of the matching policy and the source data
Not re-indexing can improve performance.

Using Data quality service to match data :

- The DQS matching process has the following benefits:
 - You can perform the matching process in conjunction with other data cleansing processes to improve overall data quality. You can also perform data de-duplication using DQS functionality built into Master Data Services. For more information, see Master Data Services Overview (MDS).