Develop mapping models with IBM InfoSphere Data Architect

Jorge Machado
Hermann Voellinger

January 27, 2011

Designing the mappings for an extract, transform, and load (ETL) process is a critical step in a data warehouse project. Mappings must be easy to modify, capable of version control, easily reported, and easily exported to other formats. This tutorial illustrates how to develop a complete source-to-target mapping model using InfoSphere™ Data Architect. You will also learn about the reporting functions that InfoSphere Data Architect provides.

Before you start

Introduction

The goal of this tutorial is to introduce a method for developing mapping models for data warehouse (DWH) projects using the functionality available in InfoSphere Data Architect (IDA), Version 7.5.2. The design of the mappings for the extract, transform, and load (ETL) process is an absolutely critical part of each data warehouse project. This design of mappings should be stored in a format that fulfills several requirements, including the following:

- The mapping can easily be changed.
- Changes can be versioned.
- It is easy to report the mappings.
- It can be exported to other formats.

You will learn in this tutorial how these goals can be reached with IDA. The tutorial builds describes all the necessary steps to build your first mapping model and to create a report for this model. The high-level steps are:

1. Create a project in IDA
2. Create the source data model (using the DB2 SAMPLE database)
3. Create the target data model
4. Create mappings between source and target tables
5. Analyze your model using the reporting functions
This tutorial describes each of these five steps in detail. At the end, you will have your first meaningful mapping report in PDF style by using the inherited functionality of the IDA tool.

**Product overview and prerequisites**

The IBM InfoSphere Data Architect (previously called Rational Data Architect) is an enterprise data modeling and integration design tool that provides a collaborative data design solution that you can use to discover, model, relate, and standardize diverse and distributed data assets.

IDA is based on an open-standard Eclipse integration.

In order to work through this tutorial you need to install the following:

- InfoSphere Data Architect, Version 7.5.2
- IBM DB2 9.7 Enterprise Server Edition with its sample database
- An instance of IBM DB2 Enterprise Edition and the standard sample database installed and accessible on the same computer

See [Related topics](#) for links to trial versions of both DB2 and IDA.

**Creating a project in IDA**

All the design work you do in IDA is grouped into projects. In this section, you will learn how to use the IDA wizard to create a new project for your mapping model. Complete the following steps to create a new project.

1. Select **File > New > Project** from the InfoSphere Data Architect menu bar. This launches the New Project wizard to help you create a new project in the workbench, as shown in Figure 1.
2. Browse through the list of projects, and open the folders to find out about all types of projects you can create. To assist in locating a particular wizard, the text field can be used to show only the wizards that match the entered text.
3. Enter **data** as a key word in the text field to limit the list of projects.
4. Select **Data Design Project**. The Data Design wizard launches, as shown in Figure 2. The wizard guides you through the basic setup steps.

![Figure 1. Creating a new project](image-url)
Figure 2. Data Design Project wizard

5. Click **Next**.
6. Enter a new name for the project, such as **MyMappingTutorial**, as shown in Figure 3, and click **Finish**.
7. Click **Yes** when prompted to open the data perspective.

**Figure 4. Open associated perspective**

Your new project is now ready.

**Creating the source data model**

In this section, learn to build a physical source data model by reverse-engineering part of the DB2 SAMPLE database. You use only a few of the existing tables. You build a subset by using the DELETE function of IDA. Finally, add annotations to a few fields in the model to extend the description of these fields.

**Select data sources**

In this step, connect to the data source that contains the source tables: the DB2 SAMPLE database.
1. Go to the Data Project Explorer, which is located on the left side of the screen. (If it is not visible, go to the menu and select Window > Show View > Data Project Explorer.)
2. Right-click Data Models, and then click New > Physical Data Model in the popup menu, as shown in Figure 5.

**Figure 5. Creation of the physical data model**

![Data Project Explorer](image)

3. Enter the name of the source database model, such as **Source Database Model 1**.
4. Select V9.7 for the DB2 version, and select Create from reverse engineering, as shown in Figure 6.
5. Click Next.

**Figure 6. Specifying the source database model**

![New Physical Data Model](image)

6. Leave the next screen unchanged, indicating that you want to reverse engineer from a database.
7. Click the New button to add a new database connection to the project, as shown in Figure 7.
8. Enter the user name and password, and select the **Save password** check box.
9. Click the **Test Connection** button to verify that a connection can successfully be established, as shown in Figure 8.

**Figure 8. Specify connection parameters**

10. Click **OK** when you see the information box confirming that the connection succeeded.
11. Leave the Filter screen unchanged, and click **Finish**.
12. Click on the newly created database connection **SAMPLE** on the Select Connection screen, as shown in Figure 9.
13. Click **Next**.

14. Apply a filter on the schemas you want to have in the source database model by clicking the check box next to the schema that has the same name as the Windows® user you were when you created the DB2 SAMPLE database. In the example for this tutorial, this is the schema **MACHADO**, as shown in Figure 10.

**Figure 10. Filter objects for source database model**

15. Click **Next**.

16. Select the element types within the schema that you want to have included in your source database model. For the tutorial, be sure that only **Tables** is selected, as shown in Figure 11.
Figure 11. Select database elements

17. Click **Next**.
18. Select the following three check boxes, as shown in Figure 12:

- **Overview**
  You can generate diagrams from the wizard, or you can generate diagrams after you complete the wizard by using the context menu options that are available from the Diagrams folder.

- **Infer implicit primary keys**
  Select this option to find implicit primary keys in the form of unique indexes. Both unique indexes and un-enforced primary key holders are created in the model when a unique index is present.

- **Infer implicit relationships**
  Select this option to find relationships among tables that are not explicitly modeled by using foreign keys. If you select this option, these relationships are represented by an un-enforced foreign key holder in the model.
19. Click **Finish**. The physical model of the source database is created.

After the creation process, you see the main window where you can review the results of the previous steps. Your newly created model is now available in the Data Project Explorer in the upper left area of the screen. You can also find the new database connection to the sample database in the Data Source Explorer in the lower left area. The main window shows general properties of the new project, as shown in Figure 13.

**Figure 12. Select more options for model**

![Select more options for model](image)

**Figure 13. Screen after creation of source data model**

![Screen after creation of source data model](image)

(View a [larger version](#) of Figure 13.)
Select the table subset from source data model

All elements of the schema MACHADO have been added to the model. You can view them in the data diagram that belongs to the source database model by clicking Data Project Explorer > Data Diagrams > Source Database Model 1.dbm > MACHADO. The data diagrams offer a graphical representation of the models you create, as shown in Figure 14.

Figure 14. Data diagram of Source Database Model 1

(View a larger version of Figure 14.)

You can now browse through the model and see the results of the reverse engineering process.

You can also use the outline widget to help you see the big picture of the model and quickly navigate throughout the pane.

Remove unnecessary objects

Because you need only a subset of those tables, you can remove all unnecessary objects. To remove unnecessary objects, go to the pane in the center of the screen, and remove every table except Employee and Project by right-clicking on each table and selecting Delete from Model from the context menu, as shown in Figure 15.
Figure 15. Remove tables from data diagram

Keep in mind that Delete from Diagram leaves the object in the data model, while Delete from Model removes the object from both the data diagram and the data model (but leaves it in the source database). The result of selecting Delete from Model should look similar to Figure 16.

Figure 16. Cleansed source table model

Add annotations to every element of source data model

The source model you created in the previous section contains most of the specification used by application developers on a technical level. However a more descriptive annotation to every element of the model down to column level is required. For example, not every developer would
interpret EMPNO char(6) exactly the same way. Complete the following steps to add an exemplary annotation to the EMPNO column in the table field, as shown in Figure 17.

1. If the columns are not visible, right-click the project table and select Filters > Show/Hide compartment.
2. Select the key and column.
3. Repeat this for the employee table.
4. Click EMPNO in the table employee column.
5. Click Properties in the lower information area.
6. Click Annotation.
7. Click on the golden rhombus in the upper left corner of the lower information area.
8. Rename the new annotation property to Description.
9. Type a meaningful description in the Value field, such as The unique serial number of an employee of company xyz. This serial remains personally assigned even if the employee leaves the company and later decides to return.

Figure 17. Add annotations to every element

(View a larger version of Figure 17.)

Creating the target data model

To build the physical target data model that you will use in the mapping tutorial, you first create a logical data model from scratch. Later you will transform this logical to a physical data model.

Create a logical target data model

Complete the following steps to create the target data model from a logical level. This will be the input for the physical model described later.
1. Go to the Data Project Explorer, and right-click **Data Diagrams**.
2. Click **New > Logical Data Model**.
3. Click **Finish**.
4. Change the name of the newly created diagram by right-clicking on the new ldm-file and naming it **Logical Target Data Model**. Figure 18 shows the resulting window in your Data Project Explorer:

**Figure 18. Data Project Explorer with new logical target data model**

![Data Project Explorer with new logical target data model](image)

Note that a corresponding data model has automatically been created.

5. Right-click the working pane of the new diagram, which is named `diagram1`. This diagram is part of Logical Target Data model.ldm.
6. Add a new entity to the diagram by clicking **Add Data Object > Entity**, as shown in Figure 19.

**Figure 19. Adding a new logical entity**

![Adding a new logical entity](image)

4. Name this new entity **Ent_proj_cost**, as shown in Figure 20.
5. Add the attributes by right-clicking the new entity and choosing either Add Primary Key, as shown in Figure 21, or Add Attribute for the non-key attributes.

This entity contains the results of the transformations from the source tables that provide information about the costs caused by the salaries of the employees assigned to the respective project.

Create physical target data model

This section describes how to derive the physical target data model from the logical model you created in the previous section. Complete the following steps.

1. Right-click Logical Target Data Model, and click Transform to Physical Data Model, as shown in Figure 22. You see a series of screens in which you will specify physical characteristics of your model.
2. Change the database to **DB2 for Linux, UNIX, and Windows**, change the Version to **V9.7**, and update the file name to **Physical Target Data Model**, as shown in Figure 23.

**Figure 22. Initialization of physical model creation**

**Figure 23. Specify filename and database version of physical data model**
3. Click **Next**. The resulting window prompts you to specify key attributes, as shown in Figure 24 and described in the following steps.
4. Specify the Name Case for physical names to be **Use lower case**.
5. Enter the Schema name as **reporting**.
6. Select **Transform diagrams**.
7. Select **Column ordering (key columns first)**.
8. Click **Next** after you have finished specifying attributes.

**Figure 24. Transform to physical data model**

![](image)

9. Click **Next**.

You have now successfully created the physical data model. Note that an entry for this new model Physical Target Data Model.dbm has been created in the Data Project Explorer, as shown in Figure 25.
Figure 25. New physical model created

Customize your view

To customize how your model is displayed, click the **Properties** tab, and select all of the following options on the Properties screen, as shown in Figure 26:

- Show key
- Show non-key
- Show name
- Show label
- Show data types
- Show foreign key decoration

Figure 26. Display options for models

Your physical data model is now ready to be used as a target in the transformations and mappings you will do in the next section.
Creating mappings between source and target tables

This section describes how to do the actual mapping between the source and target tables. IDA provides mapping models for this purpose. Complete the following steps to create such a model using the functionality of IDA.

1. Go to Data Project Explorer, and right-click the Mappings folder.
2. Click New > Mapping Model. The resulting window is shown in Figure 27.

Figure 27. Naming the mapping model

3. Change the file name to source_to_target_mapping.msl. At this point, if you wanted to create a link to an existing mapping model, you would click Advanced, but that is not needed for this tutorial.
4. Click Next.
5. Click Add, and specify Source Database Model 1 for the source data model for the mapping you want to create, as shown in Figure 28.
6. Select **MACHADO** as the database schema for the mapping model, as shown in Figure 29.
7. Click Finish.
8. Repeat steps 2-7 for the target data model Physical Target Data Model.dbm, as shown in Figure 30.
You have now created the mapping model with the tables required to proceed with the actual mapping between the source and the target tables.

**Create a mapping to fill the target table**

Complete the following steps to create a mapping to fill the target table with data on the costs caused by the salaries of each project member aggregated by project.

1. Click the column, hold down **Shift**, and select the following fields:
   - employee.empno
   - project.respemp
   - ent_proj_cost.cost_per_project
2. Right-click the **Mappings** pane, and select **Create Mapping** from the context menu. IDA creates graphical links between the selected source and the target columns.
3. Click the **Properties** tab in the middle lower area of the screen, and scroll down to the transformation box.
4. Manually enter the transformation rule in Listing 1, substituting the user name that you used when creating the SAMPLE database for **machado**, if needed.
Listing 1. Transformation rule

SELECT projname, sum(Salary) as summe
FROM machado.employee, machado.proj
where employee.empno = proj.respemp
group by  projname

Listing 1 also contains the selection and aggregation criteria.

Figure 31 shows your model.

Figure 31. Mapping model with first transformation rule defined

(View a larger version of Figure 31.)

5. Repeat steps 1-4 to create a one-to-one mapping for the following fields:
   • Project.Projectno > ent_proj_cost.projno
   • Project.projname > ent.proj_cost.proj_name

6. Add the text **1:1** in the transformation boxes of the respective mappings.

You have now successfully mapped all columns for this tutorial.

Analyzing your model using the reporting functions

IDA provides a series of pre-configured reports that can easily be used to produce documentation. You can now produce a report that shows details of the mapping model you built in the previous section.

1. Go the menu bar, and select Run > Report > Report Configurations. The Report Configuration screen appears, as shown in Figure 32. From this screen, you can configure, manage, or run a report:
2. Double-click the **Data Model Classic XSLT Report** to create a new blank report configuration page.

3. Rename the report to be **MyTutorialMappingReport**.

4. Go to the Report Input box, and click **Browse**.

5. Select **/MyMappingTutorial/source_to_target_mapping.msl** as input for the report.

6. Click **Browse** for the Built-In field, and click the mapping model **Classic Mapping Model Report**. The resulting window is shown in Figure 33.
7. In the Report Output box, type the Location where you want the report to be stored, naming the report `xslt_classic_mapping_model.pdf`, as shown in Figure 34. Currently PDF is the only supported output format for the report.
Create an actual report

You have two ways to create an actual report:

- Click **Report** at the lower end of the configuration dialog (mandatory for the initial run).
- Click **Close** at the lower end of the configuration dialog.

To continue the tutorial, click **Run > Report > MyTutorialMappingReport** from the menu bar. A PDF file opens containing the mapping report data. Because your mapping model is very simple, the report also displays a few mappings.

The mapping model consists of two main parts:

**Header information**
- The name of the mapping model on which the report is based
- The source and target tables
- A table of contents based on the target objects
The header information for MyTutorialMappingReport is shown in Figure 35.

**Figure 35. Header area of mapping report**

![Mapping Report](image)

**Detailed mapping information**
- Which source column is mapped to which target column and based on which transformation rule
- The join criteria

The detailed mapping information for MyTutorialMappingReport is shown in Figure 36.

**Figure 36. Detailed mapping report**

![Detailed Mapping Report](image)

**Conclusion**

This tutorial described how to get a first meaningful mapping model for the design of ETL jobs with just a few steps.
You started by creating a new project folder for mapping models in InfoSphere Data Architect (IDA). Next, you created a physical source data model (using the standard SAMPLE database in DB2). You extended the description of the fields in this model. After this you defined a new target physical data models with IDA from scratch. Then you started defining mapping models. You designed a few simple mappings from source to target, which showed the possibilities for designing ETL mappings with IDA. In the last section, you learned how to report the mappings in a first classical xslt report.

You can now use this approach for more complex scenarios with hundreds of mappings. Then you can take advantage of the automated features for easy changing, easy versioning, and better reporting functionality by using the data modeling tool InfoSphere Data Architect (IDA) from IBM.
Related topics

- See the tutorial "Scoping the IBM Industry Model for banking using Enterprise Model Extender and InfoSphere Data Architect" (developerWorks, Mar 2010) for how the process of scoping can yield business objects that will guide you to developing a meaningful logical data model.
- View the InfoSphere Data Architect page on developerWorks for more resources for growing your IDA skills.
- Go to the book Mastering Data Warehouse Design by Claudia Imhoff, Nicholas Galemno, and Jonathan Geiger for details on a balanced approach to data warehousing.
- Download an evaluation copy of InfoSphere Data Architect.
- Download an evaluation copy of DB2 for Linux, UNIX, and Windows.

© Copyright IBM Corporation 2011
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)