Metadata-driven Data Migration for SAP Projects

Martin Oberhofer, Albert Maier, Thomas Schwarz, Manfred Vodegel

IBM Germany - Research and Development GmbH
Schoenaicherstrasse 220
71032 Boeblingen, Germany
{martino, amaier, schwarzt, vodegel}@de.ibm.com

Abstract: SAP applications are mission-critical for many enterprises today. However, projects to introduce a new SAP solution or consolidate existing SAP solutions often fail respectively overrun budget and time. A common root cause is the underestimation of data migration work. Data quality in legacy systems is often not sufficient for SAP, and specifications of the target data model often change very late in the project lifecycle, e.g. due to new business requirements or new insights about legacy systems and legacy business processes. This can cause significant re-work in the ETL jobs that extract data from source systems, cleanse that data and load it into the target SAP system(s). We apply a model-driven architecture (MDA) approach [MP10] to such data migration projects. We generate ETL infrastructure from SAP metadata. This novel approach (known as the IBM Ready-To-Launch (RTL) for SAP solution [Ibm10]) significantly reduces project risk and cost. In addition, data quality is addressed and improved. Our demo will show programmatic access to SAP metadata and its systematic exploitation throughout the data migration project, including the generation of logical and physical data models from this metadata, and the generation of ETL jobs.

1 The RTL for SAP Solution – System Architecture

The IBM InfoSphere Information Server platform is the technology foundation of the RTL for SAP solution. It delivers enterprise information integration capabilities across all integration areas such as discovery, data profiling, ETL, replication, federation [GH10] and SOA services. It also offers SAP-certified application connectors to extract data from SAP applications and load data into SAP applications. By applying an MDA approach, RTL provides several novel capabilities to SAP data migration projects. On the one hand, using MDA principles, functional data requirement specifications are linked to business process specifications. On the other hand, functional data requirement specifications are linked to programmatic SAP metadata access. RTL has a so-called Rapid Generator component which can generate ETL jobs for data exchange with SAP systems. By linking ETL jobs model-based to functional data requirement specifications, which in turn are linked to business process specifications, it is now easy to adapt to changes in the business process requirements. If, for example, a change of a business process results in an additional, new business object attribute, the ETL jobs just need to be regenerated whenever such a change happens. In a traditional approach [LN06] such changes are often not detected until system integration test and require manual, often cumbersome adjustments in the ETL code base and additional testing.

Trademarks: IBM logos are trademarks or registered trademarks of the International Business Machines Corporation in the United States, other countries, or both. SAP Netweaver, SAP ERP, SAP R/3, SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Using the conceptual system architecture shown in Figure 1, we will now explain where we apply the Rapid Generator capability. On the source system side, there might be some legacy SAP R/3 systems as well as some non-SAP systems. The target(s) are SAP Netweaver systems such as SAP ERP. While the data is moved from source to target, there are three distinct areas where the data is persisted while in transit – the areas are usually separated using database schemas within the same database. The staging (STG) area is modelled identically to the source system data models. For SAP sources, the STG area is modelled and the tables are automatically physically created exploiting programmatic access to SAP metadata. Based on the same metadata Rapid Generator generates all extract jobs extracting the data from SAP into STG. The alignment (ALG) area is programmatically created from the SAP target system data model. It is not an exact copy of the SAP target data model. For example, there might be a need to map a hierarchical IDoc model to a relational model or to not enforce foreign key relationships yet. The rationale is that we want to be able to get all records from all sources into ALG. Once the data from all sources is in ALG in a common format, we can analyze, report, and later resolve all data quality issues on it in a uniform way.

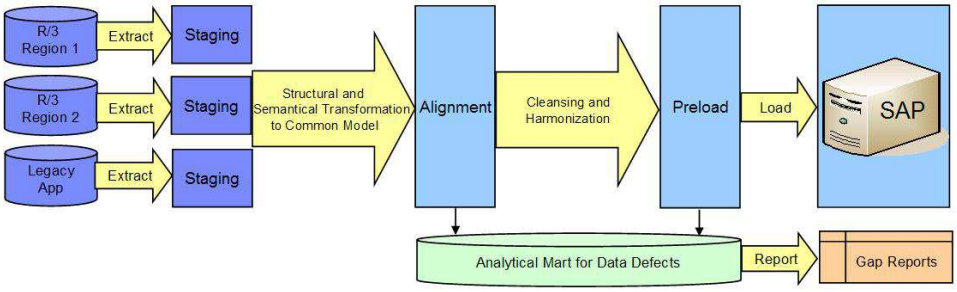

Figure 1: Conceptual System Architecture of the RTL for SAP Solution

Two steps are done when data is moved from STG to ALG: Structural alignment which means harmonizing the various source data models into a common model and semantical alignment which means transcoding the lookup values from the source system into the lookup values understood by the SAP target system. The country code value for Germany in a source system might be for example ***62*** whereas the corresponding code value in the SAP target system might be ***DE***. Once the data is in ALG, data cleansing is performed. Common cleansing operations include data standardization, matching and de-duplication. The third area known as preload (PLD) area is also generated from the data model of the SAP target system. Data types and foreign key relationships are enforced now. Moving data from ALG to PLD requires a structural transformation generated by Rapid Generator. All jobs required to load the data from PLD into SAP are also generated by Rapid Generator. In addition, Rapid Generator generates all jobs to extract the data from all lookup tables used for attributes in the model. The values of the lookup tables are needed for the Data Validity Gap Report (DVGR) which measures how many of the source records have code values that are not valid in the context of the SAP target. The DVGR is just one type of gap reports in RTL. Other gap reports are checking data completeness, data validity and field length. In general, gap reports are used to measure data quality defects while the data resides in ALG, and to measure data

load readiness in PLD. They are executed periodically during the data migration project to provide the project manager the ability to see how much progress has been made since their last execution. All gap reports are dynamically created. Technically they are based on SQL templates and take SAP metadata into account. For example, if in the SAP target system an optional attribute becomes mandatory (a normal SAP customization step), this change is visible in the SAP metadata, and the Data Completeness Gap Report would then start to report how many source records are not providing values for this attribute. Technically, SQL statements are used for these checks.

2 Demo Scenario

The key ingredient for the RTL for SAP solution is SAP metadata. Without a programmatic access to SAP metadata, the modelling of the STG, ALG and PLD areas for SAP source and target systems is cumbersome, very time consuming and, if done manually, prone to errors. An SAP Netweaver ERP solution has about 70000 logical tables – some of them are data tables where data tables can have more than 150 attributes. Dozens of these attributes could be backed by lookup tables. In addition, without an understanding of the dependencies among SAP business objects and their data models, it is difficult to define the right order of migrating data for the various SAP business objects. For example, transactional business objects like orders require that master data business objects such as product and customer have been successfully loaded. Furthermore, SAP exposes different attribute subsets for the business objects through different application-level load interfaces such as IDoc. An IDoc is a hierarchical structure representing a business object. It could easily have more than 300 attributes. Without an in-depth understanding of the technical metadata of these interfaces, and how they are related to the logical data model of the SAP business objects, it is difficult to prepare the data for load. Combining this with the fact that the functional data specifications might change while the data migration project is executed drives the need to automate the loading with a model-driven approach.

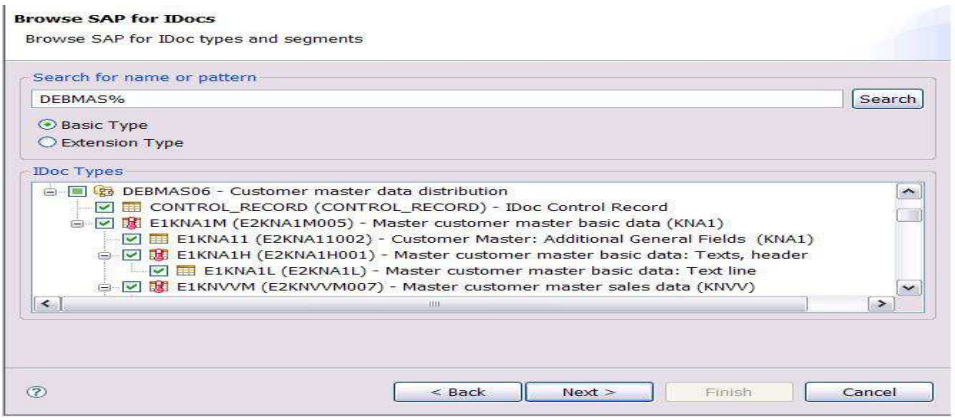

Figure 2: IBM InfoSphere Rapid Generator for SAP applications – IDoc example

The Rapid Generator approach delivers many benefits. Re-generating the physical tables for STG, ALG and PLD is seamless because the programmatic SAP metadata access reacts to data model changes in the SAP systems. Rapid Generator removes the manual efforts for extract and load jobs for SAP systems almost entirely. It automates the generation of jobs extracting a large numbers of lookup tables as well as the related lookup description tables by automatically joining them together and making them available in ALG for code value transcodings and gap reports. If the IDoc interface is used for load, the technical details in linking together all the tables at the right place in the hierarchy with the right foreign key relationships are now hidden from the ETL developer. This removes a previously tricky and error prone manual task. An example for the customer business object for SAP ERP with the corresponding DEBMAS IDoc structure is shown for a portion of the tree in Figure 2.

Among others, in the demo we will show the following steps: We demonstrate how to access SAP metadata programmatically, build corresponding logical and physical models, and persist SAP metadata in an auxiliary schema in the RTL database. Also, the functional data requirement specifications are linked to and persisted in this auxiliary schema. Then we demonstrate how the STG, ALG and PLD areas are instantiated in a model-driven way. Third, we show how Rapid Generator consumes the models and generates jobs. An example load job using the IDoc interface is shown in Figure 3.

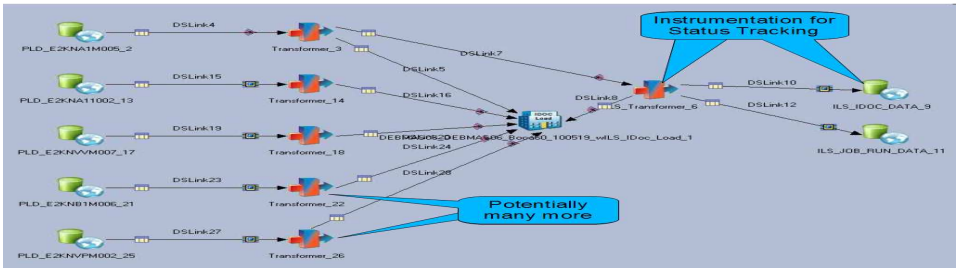

Figure 3: IBM InfoSphere Rapid Generator for SAP applications – IDoc example

References

[GH10] Godinez, M., Hechler, E., Koenig, K., Lockwood, S., Oberhofer, M., Schroeck, M.: The Art of Enterprise Information Architecture – A Systems-Based Approach for Unlocking Business Insight. Pearson, 1st Edition, 2010.

[Ibm10] IBM whitepaper: Driving successful business transformation: persistent organizational excellence with IBM Ready-to-Launch for SAP, October 2010, ftp://public.dhe.ibm.com/common/ssi/ecm/en/imw14493usen/IMW14493USEN.PDF.

[LN06] Leser, U., Naumann, F.: Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. dpunkt Verlag, 1st Edition, 2006.

[MP10] Molina, J.-C., Pastor, O.: Model-Driven Architecture in Practice: A Software Production Environment Based on Conceptual Modeling. Springer, 1st Edition, 2010.