

ETL in SAP BW using DataStage



Applies to:

This applies to SAP BI (3.5, 7.0) For more information, visit the [Business Intelligence homepage](#).

Summary

This document helps the BW teams who uses complex calculations & transformations on the frequent basis & uses SAP BW & DataStage to extract, transform and load the data from various data sources including the legacy systems.

Author: Mayank Tyagi

Company: Infosys Technologies Limited

Created on: 05- March-2010

Author Bio

Mayank Tyagi has been working in data warehousing domain from last 3 years with 1.5 years in the SAP BW technology. He is working as a Senior System Engineer with Infosys Technologies Limited. You can reach him at Mayank_tyagi@infosys.com/ Mayank_tyagi@zapak.com .

Table of Contents

SAP BW vs DataStage	3
Transformations.....	3
General Transformations	3
Connecting DataStage to SAP BW.....	7
Advantages and Disadvantages of using Datastage	9
<i>Business advantages of using DataStage as an ETL tool:</i>	9
<i>Technical advantages:</i>	9
<i>Disadvantages:</i>	9
Related Content.....	11
Disclaimer and Liability Notice.....	12

SAP BW vs DataStage

SAP is the most popular Enterprise Business Warehouse solution and one of the biggest advantages of SAP BW is that it can be integrated in any enterprise. In that way SAP BW is an open data warehouse environment. This allows users to connect it to different datasource systems through Flat Files Interfaces, DB-Link Interfaces & Application Programming Interfaces.

The sole purpose of any Business Intelligence tool is extracting, transforming & loading of the data from the data source system to the target system & generate report based on that. So there are 2 parts ETL (extraction, transformation & loading) & reporting. Although BW in itself is capable of doing all of these things, but some times doing ETL on the data is not so easy in SAP BW. IBM DataStage is one the ETL tool available in the market and it can be integrated with the SAP to do many of the ETL tasks performed in SAP BW in much simpler & efficient way, & makes the developers life easy.

Transformations

Datastage or any other ETL can not replace SAP BW completely since there is always a need of BW update and transfer rules but using third party ETL tools makes most of the transformations easier.

Since BW is the main tool which is directly integrated with the ultimate data so all the enterprise wide, business & functional level transformations & data cleansing activities should be done in SAP BW only. All the transactions & data cleansing and formatting which is associated with datasource should be done in DataStage.

To clear out the confusion where to use Datastage and where to use SAP BW, IBM and SAP has defined some quality levels, which are as follows

All the datasource related transformations are done using Datastage & this data is further provided to SAP BW in a staging area. For example the gender value for male in the legacy system should have a value of either M or F. In the cases where it is X, dataStage converts the value to an M or F.

DataStage is responsible for ensuring that data is complete & accurate from the local source system perspective. For example, if any values are missing in any fields, DataStage would handle them. If the gender status field is left empty & must be derived, a value of G may be derived in order to ensure completeness of the data.

Once all the data has been cleansed by Datastage and brought to SAP staging area then all enterprise wide cleansing & transformations are done using BW transfer rules. For example date format dd-mm-yyyy does not confirm to the BW domain values. An SAP transfer rule is needed to make this transformation.

All the lookup checks & consistency of data are performed in transfer rules of SAP BW.

All the functional & business requirement related changes are done using SAP BW update rules.

General Transformations

IBM DataStage & data warehousing solutions provide graphical user interface & libraries of predefined transformations so in that way it has reduced the effort needed to actually implement & test the transformations.

Common Transformation Operations

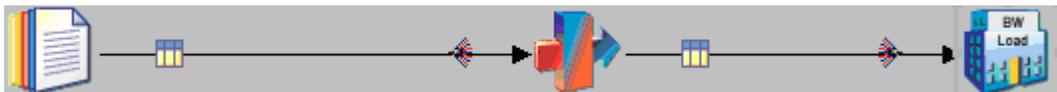
Data integration & application logic are the two main determinants for implementing data warehouse. All the technical & semantic differences between data from source systems are eliminated keeping data integration in mind where as application logic transformations transform integrated data into an application specific format optimized for a specific purpose.

Data integration Transformations

The most important transformations involved in technical data integration are:

Source to Target Direct Mapping:

The most common transformation where mapping between two semantically and technically identical sources and target fields, can be done this way.



Mapping

Constraint:

Destination	Column Name
In_LgPOGRInvoRcpt_Hist.DU54L_B_PD	RPO_DATE
In_LgPOGRInvoRcpt_Hist.DU54L_N_PD	DI_EBELN
In_LgPOGRInvoRcpt_Hist.DU54L_D_PO_PO	PGR_PSTOT
In_LgPOGRInvoRcpt_Hist.ORDER001	PMNRCPSTQ
In_LgPOGRInvoRcpt_Hist.DU54L_N_VNDT	PMNRGVSND
-	FISCVAPNT
In_LgPOGRInvoRcpt_Hist.CALDAY	CALDAY
In_LgPOGRInvoRcpt_Hist.DU54L_N_DC	PLANT
In_LgPOGRInvoRcpt_Hist.DU54L_N_ITEM	MATERIAL

Source

Col	SQL type	Extended	Length	Scale	Nullable	width
1	VarChar	Unicode	8	No	Prefix	
2	VarChar	Unicode	4	No	invoer	
3	VarChar	Unicode	10	No	pushi	
4	VarChar	Unicode	7	No	Mokm	
5	VarChar	Unicode	30	No	Buwig	
6	VarChar	Unicode	5	No	calenr	
7	VarChar	Unicode	10	No	verdo	
8	VarChar	Unicode	4	No	order	
9	VarChar	Unicode	5	No	ASV v	
10	VarChar	Unicode	25	No	name	
11	VarChar	Unicode	1	No	item ty	
12	VarChar	Unicode	11	No	verdo	
13	VarChar	Unicode	1	No	type s	
14	VarChar	Unicode	5	No	invent	
15	VarChar	Unicode	2	No	stock	
16	VarChar	Unicode	2	No	order	
17	VarChar	Unicode	4	No	numbr	
18	Decimal		17	2	The S	
19	Decimal		17	2	The O	
20	Decimal		17	3	order	

Target

Column name	extended	Length	Scale	Nullable	Description
1	<input type="checkbox"/>				RPO Date
2	<input type="checkbox"/>				Receiving
3	<input type="checkbox"/>				GR Posting
4	<input type="checkbox"/>				Receipt Se
5	<input type="checkbox"/>				Original PO
6	<input type="checkbox"/>				Fiscal Year
7	<input type="checkbox"/>				Calendar d
8	<input type="checkbox"/>				Plant
9	<input type="checkbox"/>				Material
10	<input type="checkbox"/>				Vendor
11	<input type="checkbox"/>				Material gr
12	<input type="checkbox"/>				Order Unit
13	<input type="checkbox"/>				NDC Numbr
14	<input type="checkbox"/>				Purchasing
15	<input type="checkbox"/>				Created on
16	<input type="checkbox"/>				Base Unit
17	<input type="checkbox"/>				Order unit
18	<input type="checkbox"/>				The numb
19	<input type="checkbox"/>				Currency
20	<input type="checkbox"/>				Net price

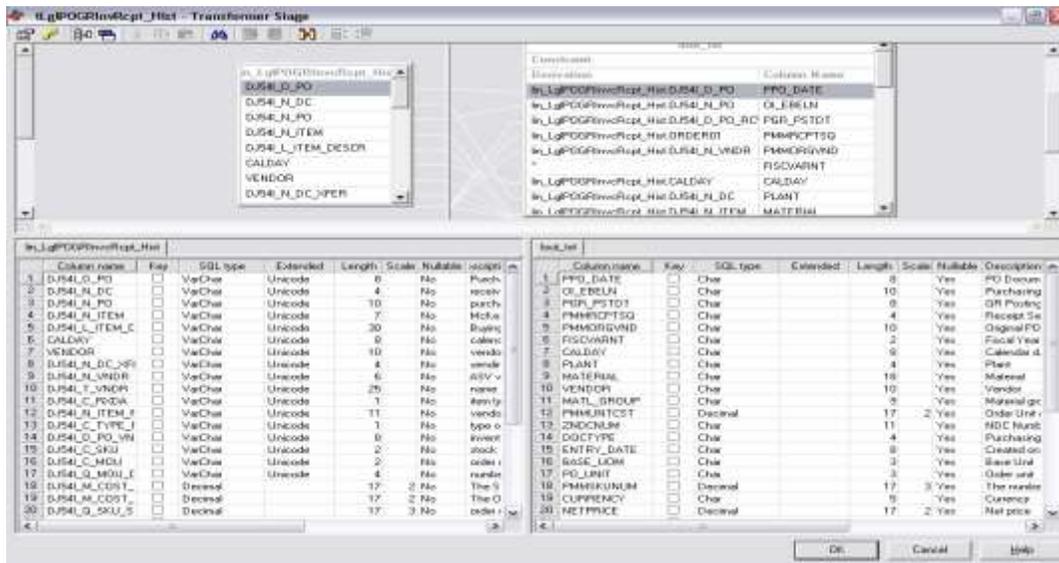
OK Cancel Help

Data Type Conversion:

Since all the data types should be according to the ultimate business requirement so data types used in the operational environment are converted to those used in the data warehouse environment. SAP BW automatically performs compatible type conversion usually require the more complex transformation to be applied.

Unicode Conversion:

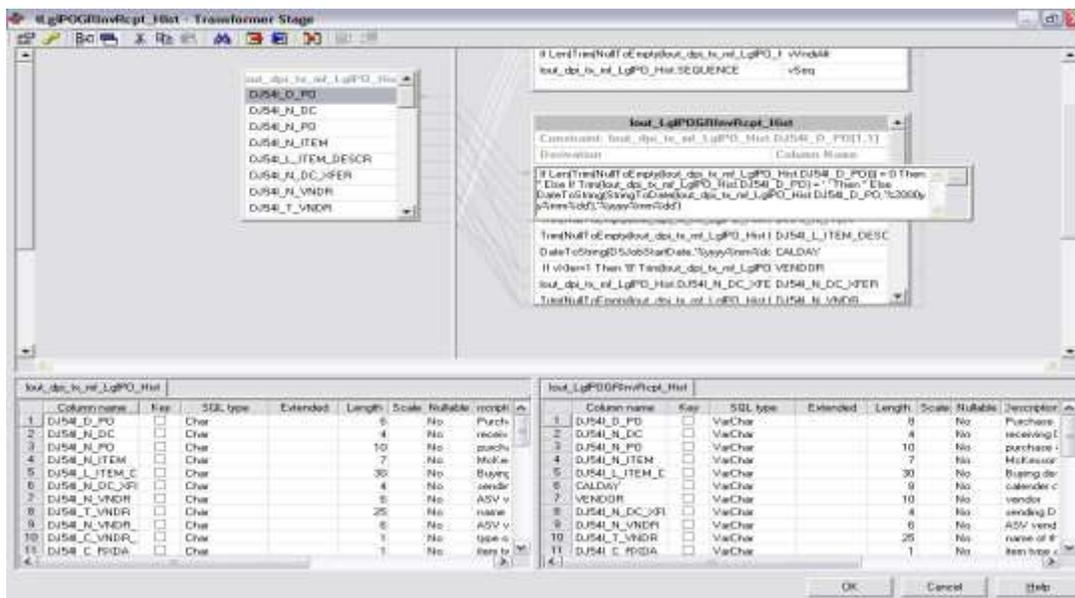
Data coming from different source systems can be in different format & languages so to overcome this issue DataStage provides UNICODE system to convert different ACII Character to one single representation in data warehouse. Unicode conversion also has the facility of supporting various national languages.



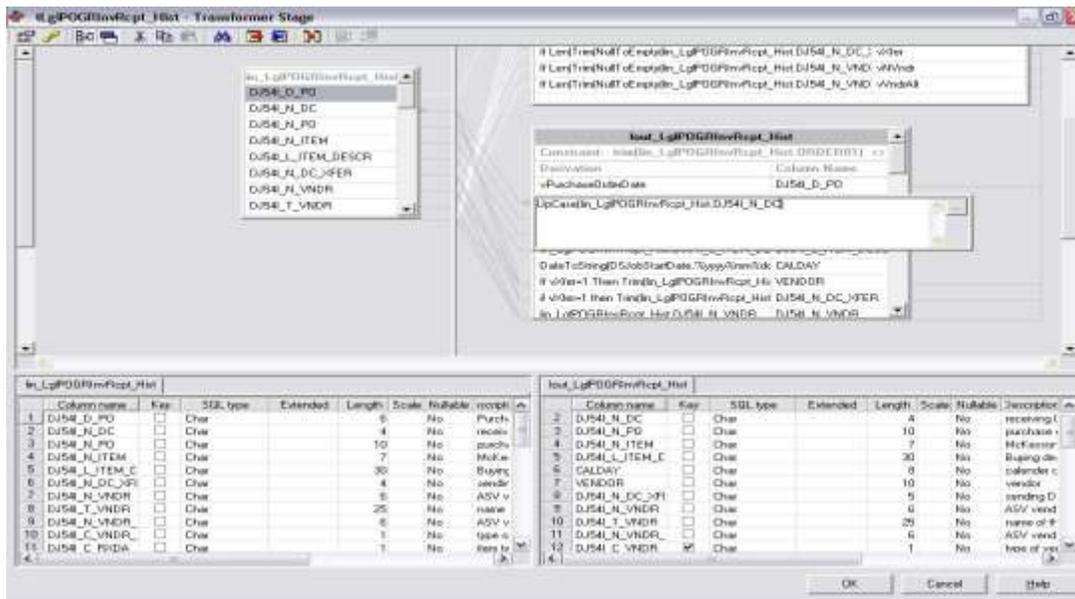
Format Changing:

Changing the technical representation of values from the source system to the target system is called format changing.

For example date format mm-dd-yyyy is required to be changed to dd-mm-yyyy.

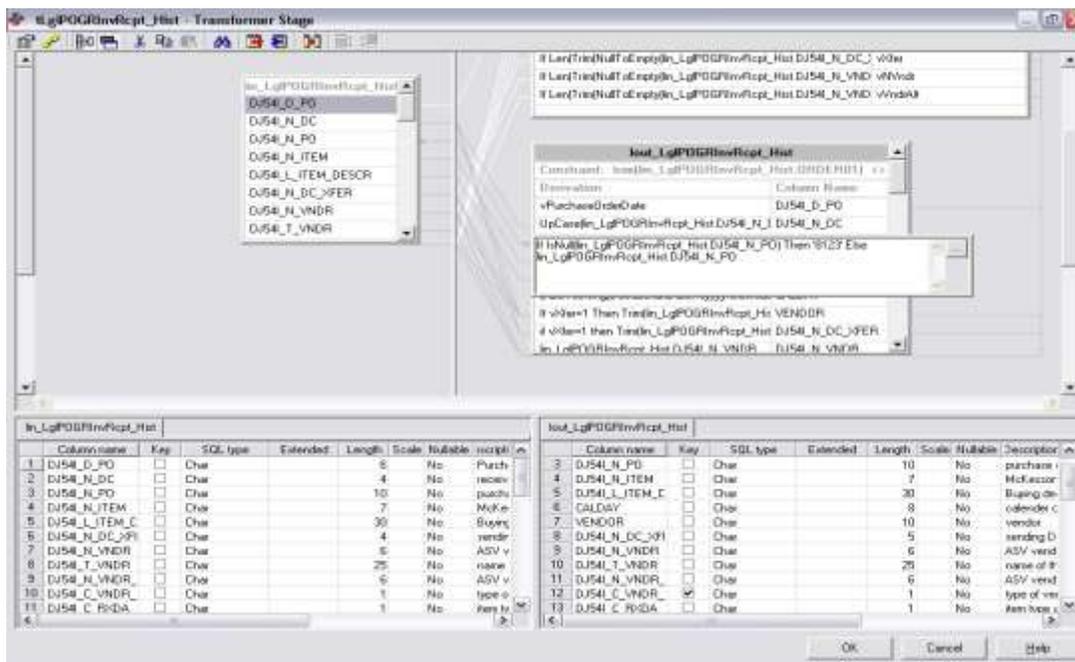


Some of the key values may need to be converted from mixed case “ABcDEf” to uppercase “ABCDEF”.



Replacing Missing Values:

Sometimes we don't get meaningful data for all the required fields from the source systems. In these cases some default values which may be retrieved from lookup tables have to be determined and these determined values can replace the missing values. For example, for non-SAP source systems to send date information without information about fiscal year variants, since fiscal year variants tend to be an SAP-specific data element.



access & mining are done using object oriented business application programming interface (BAPI). SAP BW also supports a flat file interface to import data from common flat file data formats, such as a comma delimited file or Microsoft Excel file.

For the compatibility check of all the third party tools, SAP & vendors do a check to ensure that the third party tool is in compliance with available API's for each version of SAP BW.

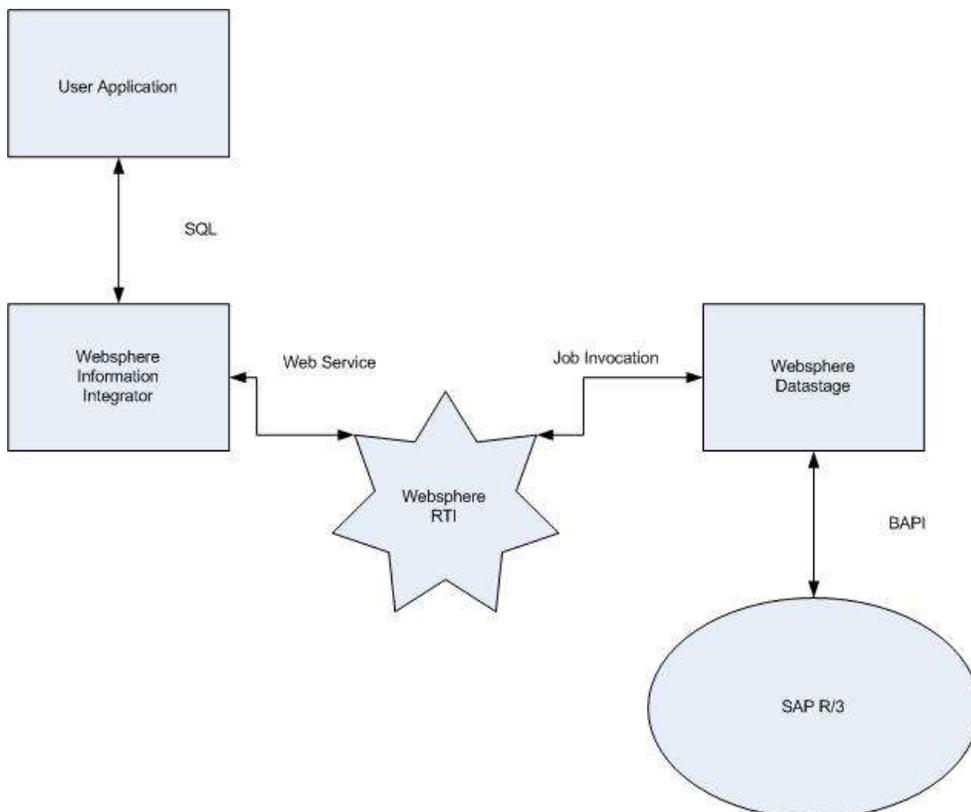
The data Extraction, Transformation, and Load (ETL) process is common to all data warehouses. SAP BW provides built-in methods to perform ETL as part of the business content for SAP R/3 OLTP. However, when data sources are not generated from SAP R/3 OLTP, third-party ETL tools can be used to perform ETL functions along with SAP BW data transformation capabilities.

The most common data load API for SAP BW is called the Staging BAPI (SBAPI). The reason behind the name "staging" is that such BAPI's do not load data directly in a database data stores; they interact with the database engine through the SAP BW Staging engine, which handles data manipulations and loads data in the reporting data stores. The basic principle behind Staging BAPI's is that the third-party tools communicate with the SAP BW staging engine through a Remote Function Call (RFC) server. The RFC server is registered in SAP BW as part of the third-party data source definition.

IBM Websphere DataStage is one of the third party ETL tools, which can be used to perform ETL functions along with SAP BW data transformation capabilities.

Datastage supports Staging BAPI and allow you exchange meta data with and transfer data to SAP BW system. The benefit for SAP not only lies in extending its offering by ETL technology but also in extending the reach of the business content by jointly providing integrated solutions for ERP & CRM solutions provided by other vendors like PeopleSoft, Siebel & Oracle etc.

From the technology point of view, the main advantage of DataStage is that it not only provides access to disparate systems & technologies including mainframe databases & merges their data, they also provide powerful & flexible transformation & sometimes even data quality maintenance functionality, all easily defined using graphical user interface designer software.



The need for a robust transformation engine is greater when disparate source system data must be merged. When you are loading from SAP system only, this is less of a necessity, since the system is integrated and shares common meta data. For example, for all the applications in SAP R/3, the same checklist is used for plants. In another system there may be an entirely different list of codes for plants. A tool must transform those records in a foreign system into a conformed set of plant codes.

Advantages and Disadvantages of using Datastage

There are several advantages of using datastage for the ETL task e.g.

Data Quality: The data that builds a data warehouse often comes from various data sources. The structure of the legacy data is often not documented and the data quality is poor. The WebSphere Information Analyzer product analyzes your data and determines the data structure and quality. It helps you understand your data. The WebSphere Quality Stage product standardizes and matches any type of information to create high quality data.

Data Volume: There is often a huge amount of data that needs to be processed regularly for a data warehouse environment. Sometime the data volume grows beyond expectations. The issue needs to be addressed with a scalable ETL architecture. IBM Information Server leverages the pipeline and partition technologies to support high data throughput. IBM Information Server can be deployed on symmetric multiprocessing (SMP) and massively parallel processing (MPP) computer systems to achieve the maximum scalability.

Business advantages of using DataStage as an ETL tool:

Hand coding is minimal so the return of investment is good and it saves a lot of time.

Development takes very less time and the maintenance is also easier with the help of GUI tool.

Datastage can easily be integrated with so many data warehouse such as SAP, Cognos, Oracle, Teradata etc. Development Partnerships - easy integration with top market products interfaced with the data warehouse, such as SAP, Cognos, Oracle, Teradata, and SAS.

Complex transformations can be done very easily and it is capable of transferring bulk data easily.

Technical advantages:

So many heterogeneous applications can be integrated with the Single interface.

Datastage provides flexible development environment which reduces training needs and enhances reuse. Developers can follow data integrations quickly through a graphical work.

Datastage has the ability to join data both at the source and at the integration server and to apply any business rule from within a single interface without having to write any procedural code.

Same data infrastructure can be used for data movement and data quality.

Datastage Enterprise Edition has the facility of parallel processing engine which provides unlimited performance and scalability.

The datastage server performs very well on both Windows and UNIX servers.

Disadvantages:

Due to the big architectural differences between Server and Enterprise edition the migration takes lot of time.

The error handling and locking resolution is manual here and every time you faces these kind of issues you either need to go to Datastage director or manger and resolve manually from there

No Unix Datastage client - the Client software available only under Windows and there are different clients for different datastage versions. The good thing is that they still can be installed on the same windows pc and switched with the Multi-Client Manager program.

Small & medium size company may not be able to invest that much of money.

Related Content

www.Dsxchange.com

www.sdn.sap.com

www.ibm.com

www.ittoolbox.com

For more information, visit the [Business Intelligence homepage](#).

Disclaimer and Liability Notice

This document may discuss sample coding or other information that does not include SAP official interfaces and therefore is not supported by SAP. Changes made based on this information are not supported and can be overwritten during an upgrade.

SAP will not be held liable for any damages caused by using or misusing the information, code or methods suggested in this document, and anyone using these methods does so at his/her own risk.

SAP offers no guarantees and assumes no responsibility or liability of any type with respect to the content of this technical article or code sample, including any liability resulting from incompatibility between the content within this document and the materials and services offered by SAP. You agree that you will not hold, or seek to hold, SAP responsible or liable with respect to the content of this document.