SAS Enterprise Data Integration Server - A Complete Solution Designed To Meet the Full Spectrum of Enterprise Data Integration Needs

[1]*Silvia BOLOHAN,* [2]*Sebastian CIOBANU*
[1]Marketing Manager, [2]Implementation Project Manager
SAS Analytical Solutions Romania

This paper is about why is Data Integration important for organisations around the world. Organizations struggle daily with the challenges of large distributed data volumes, inconsistently defined data across disparate systems and the high expectations of data consumers who depend on information to be correct, complete and available when they need it. SAS Enterprise Data Integration Server provides a comprehensive solution that enables organizations to solve these challenges in a timely, cost-effective manner with the ability to efficiently manage data integration projects on an enterprise scale, increasing overall productivity and reducing the total cost of ownership.
Keywords: *metadata, data cleansing, data integration, ETL, ELT*

1 Introduction

What does SAS[®] Enterprise Data Integration Server do?

SAS Enterprise Data Integration Server, featuring DataFlux[®] technology, is a powerful, configurable and comprehensive solution designed for managing big data. It can meet a wide variety of data integration requirements, from small tactical projects to strategic business initiatives.

• Access virtually all data sources.
• Extract, cleanse, transform, conform, aggregate, load and manage data.
• Support data warehousing, migration, synchronization, federation and provisioning initiatives.
• Support both batch-oriented and real-time master data management solutions.
• Create real-time, reusable data integration services in support of service-oriented architectures and data governance.

Why is SAS[®] Enterprise Data Integration Server important?

It enables organizations to efficiently manage data integration projects on an enterprise scale in a timely, cost-effective manner and meet the high data quality expectations of information consumers.

For who is SAS[®] Enterprise Data Integration Server designed?

It is designed for organizations in all industry sectors that are implementing one or more data integration projects, dealing with changing business landscapes and business-driven IT initiatives, trying to meet regulatory requirements, or implementing data governance.

2 SAS[®] Enterprise Data Integration Server

Product overview

SAS offers the only comprehensive enterprise data integration environment that is built from the ground up to meet the full spectrum of enterprise data integration needs. Instead of linking and managing technologies from different vendors, SAS Enterprise Data Integration Server provides a collaborative design environment promoting object reuse and sharing, administrative controls, wizard-driven design process workflow, and ease of use and maintenance. This flexible, reliable solution can access data from virtually any system in any form, transform and cleanse data even in real time, and handle data migration, synchronization and federation projects all through a versatile services environment that is easy to deploy and maintain.

Interactive data Integration Development Environment

A graphical user interface (GUI) simplifies and speeds projects with wizards, extensive built-in transformations and powerful productivity enhancements, all while providing a single point of control for managing complex enterprise data integration processes. SAS Data Integration Studio is easy to learn and use. It provides a collaborative environment that lets you build reusable processes to speed data integration development both now and in the future. It automatically captures and manages standardized metadata from any source, and enables you to easily display, visualize and understand enter-prise metadata and your data integration processes.

Connectivity and Data Access

Most organizations struggle with accessing the plethora of data sources (legacy, relational, flat files, XML, cloud data, text, etc.) that are necessary to support analytical systems. SAS Enterprise Data Integration Server provides connectivity to virtually all types of data sources and types, operating systems and hardware environments using both native access and open standards. It also supports the reading and writing of data from message queues and the sending and receiving of data to and from Web services.

Metadata Management

SAS provides a shared metadata environment that is both independent (for data integration) and part of SAS' comprehensive platform. Technical, business, process and administrative metadata is stored and managed in a way that leverages and facilitates reuse of existing table definitions, business rules and more. Navigational tools help users understand how the data was derived and where it is stored and used. Shared metadata provides a consistent definition across data sources to speed integration projects, simplify design and reduce maintenance costs.

Data Cleansing and Enrichment

There is an increased awareness, driven by compliance mandates and data breaches, of how data quality and data governance can directly affect the bottom line. This puts increasing pressure on IT to address potential data quality issues. SAS Enterprise Data Integration Server provides a single environment that seamlessly integrates data quality within the data integration process, taking users from profiling and rules creation through execution and monitoring of results. From data de-duplication (for example, within database marketing applications) to cleaning up data (for example, before storing in a data warehouse), SAS provides an enterprise approach that lets you develop and share a library of data rules and processes between projects and across the entire data integration solution. Organizations can transform and combine disparate data, remove inaccuracies, standardize on common values, parse values and cleanse dirty data to create consistent, reliable information.

Extraction, Transformation and Load (ETL) and Extraction, Load and Transformation (ELT)

Loading data warehouses and data marts within their allotted time windows, quickly building analytical marts for special projects, and creating extract files for reporting and analysis applications are tasks IT organizations face each day. SAS Enterprise Data Integration Server includes an intuitive point-and-click Design Editor window that allows developers to easily build logical process workflows, quickly identify the input and output data stores and create business rules in metadata, enabling the rapid generation of data warehouses, data marts and data streams. Users can also choose to have many transformations and processes take place inside a connected database, data warehouse or storage system. This is referred to as ELT, push-down or in-database processing, and can substantially speed up overall processing times by reducing unnecessary data movement. SAS

Enterprise Data Server uses visual SQL push-down to select the optimal processing approach.

Migration and Synchronization

Moving data from system to system is a constant activity in most organizations. Mergers and acquisitions result in multiple, overlapping systems containing information that often needs to be synchronized and ultimate migrated. Moving legacy data during upgrades and conversions is an on-going process, as is the movement of data into and out of ERP systems.

SAS Enterprise Data Integration Server provides the capability to migrate, synchronize and replicate data across different operational systems and data sources. The point-and-click process design editor makes it easy to document migration and synchronization processes in workflows that can be reused and modified for other projects. Powerful data transformations are available for altering, reformatting and consolidating information during these processes.

You also can build a library of reusable business rules ensuring that bad data is never spread from system to system. In this way, information delivered across all applications, systems, environments and geographies is up-to-date, consistent and accurate.

Data Federation

Organizations today have data stored and scattered in and across numerous data sources. Often what's needed is fast access to the most current operational data to support real-time analytics and reporting needs. SAS Enterprise Data Integration Server provides the ability to query and use data across multiple systems without the physical reconciliation or movement of source data. The logical semantic layer shields business users from the complexities of the underlying physical data. By avoiding unnecessary data replication and movement, it is possible to quickly and cost-effectively deliver up-to-date data that is consistent and accurate.

Master Data Management Support

SAS Enterprise Data Integration Server includes data-mastering capabilities that provide a basis for implementing master data management projects that enable you to identify, standardize and correct common master data such as customer and product data. Unsurpassed data access, profiling, enrichment, clustering and consolidation to clean, standardize and enhance data, and an intuitive development environment that is adaptable to each organization's technologies and standards, increase productivity and produce more rapid results.

SOA and Message Queue Integration

Organizations are challenged to improve operational efficiency, streamline processes and be more agile. Using a service-oriented architecture (SOA) approach helps IT ensure that various applications can communicate with each other to better meet changing business requirements. SAS Enterprise Data Integration Server delivers easily maintained data services that enable developers to build sophisticated data services once and deploy them across the enterprise for reuse.

Message queue integration is another way to reduce maintenance, integration costs and bridge new technologies. Using SAS Enterprise Data Integration Server, you can access message queues in batch or real time without the need for custom programming. Integration developers simply treat message queues as any other source and target. [1]

With SAS Enterprise Data Integration Server, you can define the propagation of information from table to table in your transformations. This shows an example of the default mapping rules being applied when mapping numeric to character columns, and character to numeric columns. This also shows the intelligence of the mapping display that uses color to indicate the presence of a transformational expression between source and target. Default mapping rules are predefined but can be extended to meet business-specific needs.
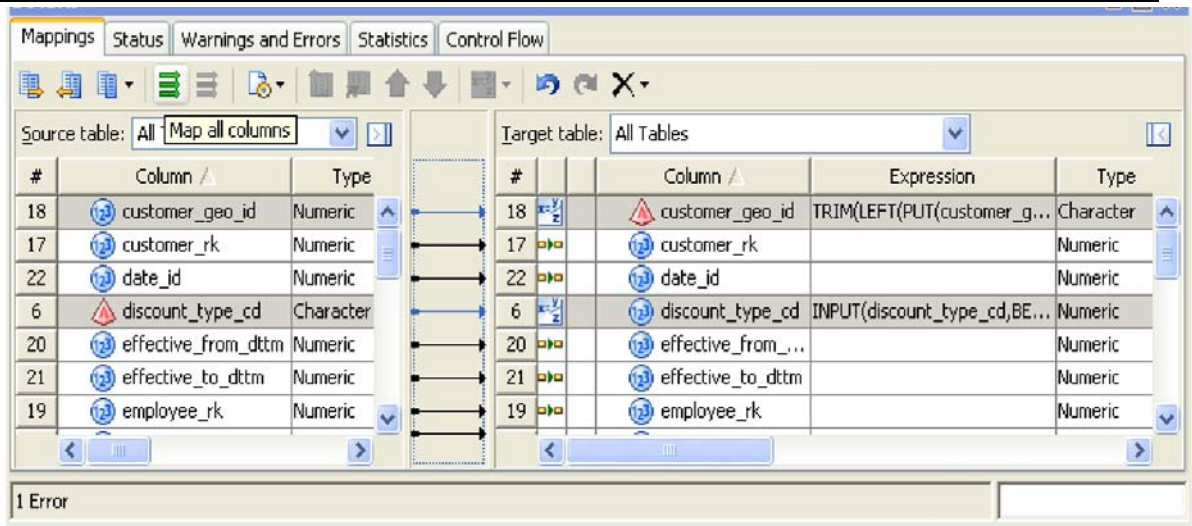
Fig. 1. Default mapping rules chart

3 Key Features

3.1. Interactive data integration development environment

• An easy-to-use, point-and-click GUI uses an intuitive set of configurable windows for managing data integration development processes.

- A visual, end-to-end process designer lets developers quickly build and edit processes.
- Drag-and-drop functionality eliminates programming.
- Wizards make it easy to access source systems, create target structures, import and export metadata, and build and execute ETL process flows.
- The multiple-user, multiple-level design environment supports collaboration on large, enterprise projects.
- Role-based permissions show users only what they are authorized to see.
- Customizable metadata tree views let users display, visualize and understand metadata.
- Dedicated GUI for profiling data to identify and repair source system issues while retaining the business rules for use in other ETL processes.
- Interactive debugging and testing of jobs during development and full access to logs.
- Check-in/check-out of jobs, related

tables and objects; and job status viewing.
- Audit history lets designers see which jobs or tables were changed, when and by whom.
- Ability to distribute data integration tasks to nearly any platform and to connect virtually any source or target data store.
- Integration with third-party vendors Subversion and CVS provides enhanced version and source control features such as archiving, differencing and rollback.
- Job status and performance reports provide the ability to track metrics such as CPU use, memory, I/O, etc.
- Automated job deployment allows the use of common scripting languages to deploy SAS Data Integration Studio batch jobs in an automated manner.
- Enhanced SAS code import capabilities give current SAS users an easy way to import their SAS jobs and SAS code into SAS Data Integration Studio. Includes logging and error checking.
- Command-line job deployment for deploying single and multiple jobs.
- Enhanced data integration job orchestration (process flow).
- The ability to surface in-database scoring models within SAS Data Integration Studio.
- Enhanced connectivity to Aster Data, EMC Greenplum, Hadoop and Sybase IQ

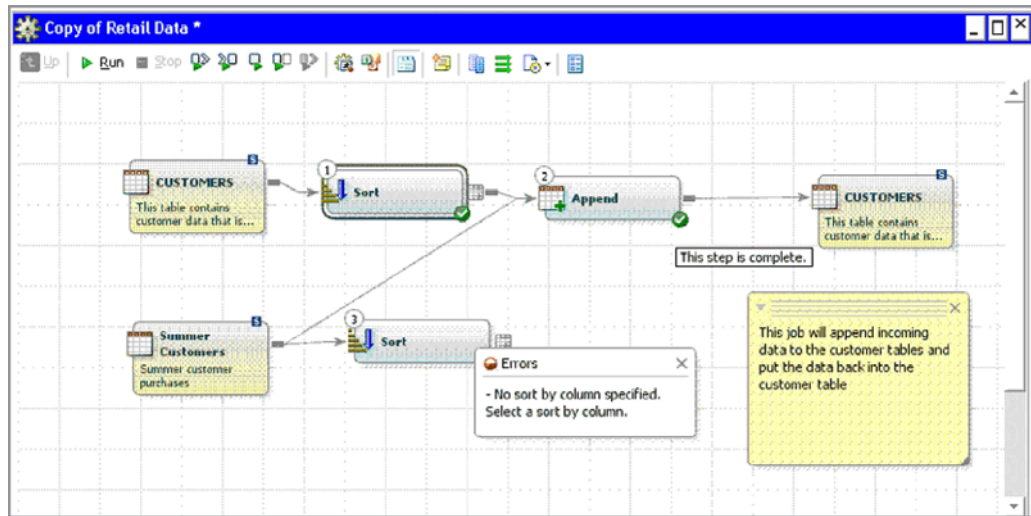databases with the ability to push down more processing to the databases. [2]

Fig.2. Retail Data ChartSAS Enterprise

Data Integration Server includes an easy-to-use and informative GUI. You build jobs by dragging and dropping data objects into the diagram area. You can add transformations such as sorts, joins and loads from a library and draw arrows to connect the objects together. Self-documentation is provided using annotated data, and yellow notes containing further information can be added by users.

3.2 Connectivity and data access
• Provides connectivity in batch or through message queues in real time to more data sources on more platforms than any other solution.
• Data access engines are available for enterprise applications, nonrelational databases, RDBMSs, data warehouse appliances, PC file formats and more.
• Specialized table loaders provide optimized bulk loading of Oracle, Teradata and DB2.
• File reader/writer available for Hadoop file system (HDFS).
• Support for Hadoop's MapReduce, Pig and Hive within flows.
• Data movement capabilities to and from Hadoop.

• A complete and shared metadata environment provides consistent data definition across all data sources.
• Native access methods deliver the best performance and reduce the need for custom coding.
• Support for message-oriented middleware, including WebSphere MQ from IBM, MSMQ from Microsoft, Java Message Service (JMS) and Tibco's Rendezvous.
• Support for unstructured and semi-structured data to parse and process files.
• Access to static and streaming data for sending and receiving via Web services.
• Expanded support for MPP databases: AsterData nCluster, EMC Greenplum and Sybase IQ, enabling more ELT pushdown and support for bulk-load utilities.
• Native support for SQL-based processing.

3.3 Metadata management
• Metadata is captured and documented throughout transformations and data integration processes, and is available for immediate reuse.

• Sophisticated metadata mapping column definitions from sources to targets, and for creating automated, intelligent table joins.
• New metadata search tool.
• Impact analysis for assessing the scope and impact of making changes to existing objects such as columns, tables and process jobs before they occur.
• Ability to determine the path, processes and transformations taken to produce the resulting information.
• Data lineage (reverse impact analysis), which is critical for both validating processes and building user confidence in data.
• Change analysis for metadata change

technologies for quickly propagating discovery, comparison, analysis and selective propagation.
• Multiple-user collaboration support includes object check-in and check-out.
• Promotion and replication of metadata across development, test and production environments.
• Wizard-driven metadata import and export.
• Wizard for metadata column standardization.
• Metadata-driven deployment flexibility so that process jobs can be deployed for batch execution, as reusable stored processes or as Web services.
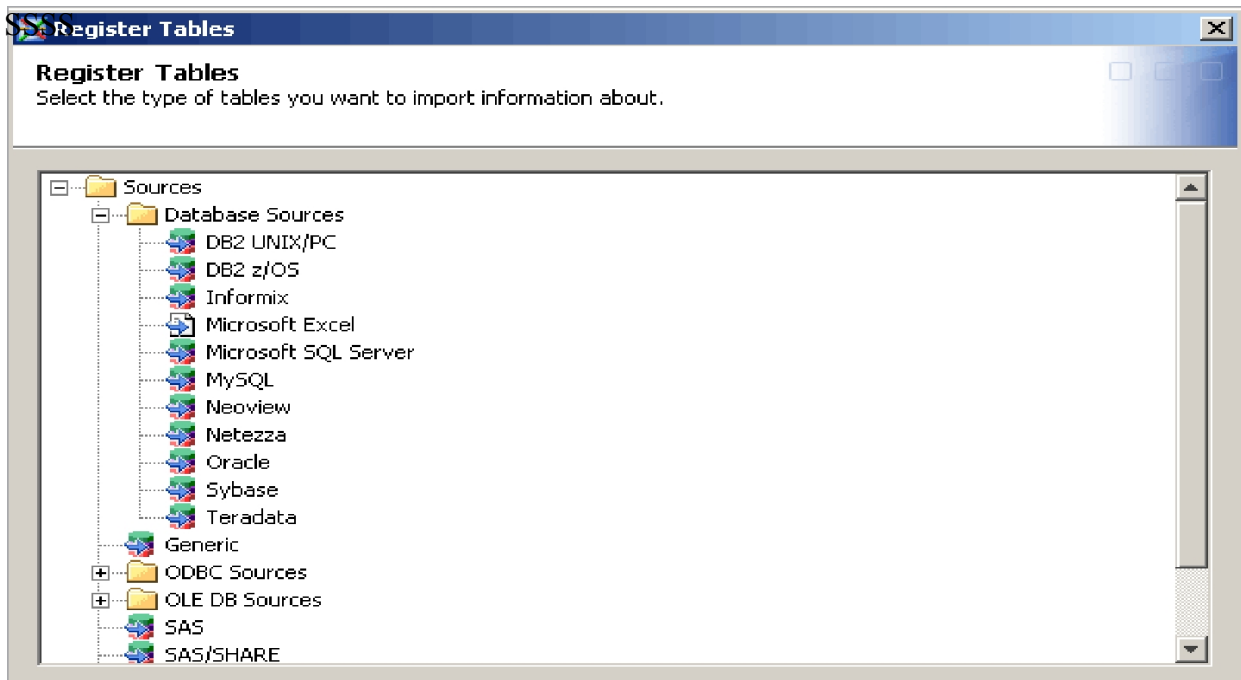

Fig.3. Register Tables Chart

The Register Tables wizard makes it easy to access data from many different systems, as well as read and manage metadata from external sources.

4. Conclusions
From legacy systems to ERP applications to data stored in Hadoop, data from virtually any hardware platform or operating system can be accessed,

cleansed and processed. New source systems can easily be added and security is managed centrally. This saves time, shortens learning curves and gives decision makers the complete information they need.

Another conclusion is that a common repository enables the centralized storage, management and reuse of work based on user authorizations, reducing both

development and maintenance time. A GUI environment that is easy to use provides a standard interface for building and documenting work. Collaboration is encouraged and manual coding is available when needed. New team members can get up to speed quickly on work done by others, which is important when documentation is inadequate or missing.

A huge benefit of this tool is that you can manage security and administration at all levels. Reusable templates make it quick and easy to provide role-based authorizations and administrative privileges at the user, department or enterprise level.

SAS processes data fast! Organizations can take advantage of a grid-enabled load-balanced, multithreaded parallel processing architecture that can quickly trans-form and move data between different platforms and systems. SAS also supports zero data movement by using SQL pass-through into popular database appliances, including Oracle, DB2, Teradata, Netezza, SQL Server, AsterData and Hadoop.

Consistently getting correct data when and where it is needed provides increased confidence in the accuracy and timeliness of operational and business information. Data quality auditing tools monitor the quality of data in processes and source systems. Users can see where data originated and how it was transformed. Optional enrichment components can add value and ensure everyone receives the best possible data. Deliver consistent, trusted and verifiable information.

SAS offers the only comprehensive enterprise data integration solution that is built from the ground up to meet the full spectrum of data integration needs. It eliminates the piecemeal approach of linking and managing technologies from different vendors and provides lower overall cost, reduced risk and faster results.

References
[1] SAS Data Integration Studio 4.4: User's Guide, http://support.sas.com/documentation/cdl/en/etlug/65016/HTML/default/viewer.htm#titlepage.htm
[2] Base SAS 9.3 Procedures Guide, http://support.sas.com/documentation/cdl/en/proc/63079/HTML/default/viewer.htm#titlepage.htm

Silvia BOLOHAN is Marketing Manager at SAS Romania for 7 seven years. She leads and functions in the creation or production of marketing content for internal and external use in area of assignment. Silvia is responsible for developing and executing marketing strategy and/or programs for SAS products and services.

Silvia has given support for data analysis based projects such as customer segmentation, attrition modelling, customer lifetime value, etc. She contributes to the efforts of building and maintaining a comprehensive reporting and tracking strategy for campaign response.

Sebastian CIOBANU is SAS consultant for over 3 years in the Business Intelligence and Data Integration domain for the Banking sector. Projects he has worked for include Analytical CRM solutions, Data Mining and Sales Data marts. He has a BA in Economic Informatics and MsC on Databases from the Academy of Economic Studies of Bucharest. His areas of interest are: Databases, Data Modeling, Business Intelligence solutions and the Banking area.