SAS® GLOBAL FORUM 2018

USERS PROGRAM

Introduction to ETL with SAS®

Presented by Vasilij Nevlev
Analytium Ltd

April 8 – 11 | Denver, CO
#SASGF

Introduction to ETL with SAS®

Presented by Vasilij Nevlev

Analytium Ltd

Why ETL is important?

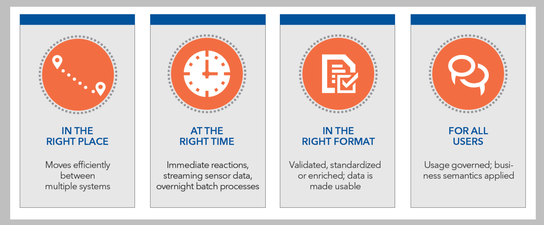
If you are here, at SAS Global Forum, you are probably involved in data management or data consumption in one or more ways.

It might sound obvious, by difficult to achieve, the data is often required to be:

- In the right place
- At the right time
- In the right format
- Available for the rights users

ETL framework existing to meet these objectives

Data Management Objectives



What is ETL?

- ETL is an acronym for “Extract, Transform, Load” which describes key stages and their order in a typical data management process.
- ETL process is often, but not always, implemented at an enterprise level as a data warehouse
- “A data warehouse is a system that extracts, cleans, conforms and delivers sources data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making”

Source: Ralph Kimball, Joe Caserta: The Data Warehouse ETL Toolkit; Whiley 2004

- The most important part to the business is “querying and analysis”
- The most complex and time consuming part is “extracts, cleans, conforms and delivers”.

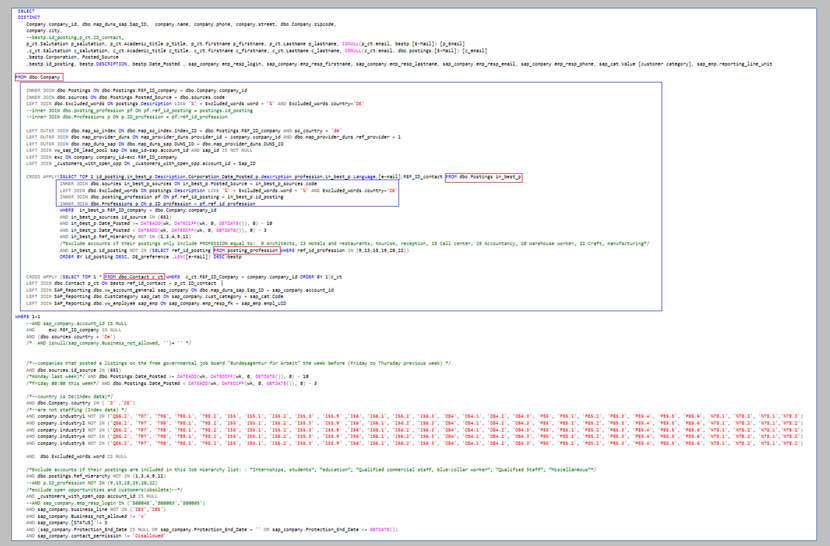
Well developed ETL process helps you to:

- Access the data you need
- Improve productivity
- Reduce development and maintenance time
- Govern and secure your data
- Work faster and meet time constrains
- Eliminate overlapping and redundant tools

When there is no managed ETL

When there is no defined approached to ETL, organisations end up with creating multiples views of their data (demonstrated below). These are monolithic chunks of code that no one understands, difficult to manage and often produce unexpected results.

Queries like these often have key ETL components within them, but are intermingled within a single view while siloed and duplicated with other queries of views often within the same database.



Side note: ETL and Scale of Data Management Initiative

- ETL process is applicable at any scale
  - Enterprise Scale: Enterprise Data Warehouse
  - Business Unit Scale: ETL to deliver specific data to another operational tool
  - Developer Level: individual development tasks
- Everyone does it, even if they don’t realise it
- Some do it better than the others

Introduction to ETL with SAS®

ETL Key Subsystems and Components

Presented by Vasilij Nevlev

(E)xtract data subsystems

Prepare To Start:

- Business Requirements
- Logical Data Map
- Business Terms
- Naming conventions

Judge Data:

- Data Profiling
- You need to know what data attributes you are dealing with in advance

Change Data Capture

- In many cases you need to track and version changes.

Result: Extracted Table including Format Conversion

SAS Tools:

- Profiling tools in SAS Data Quality Studio
- Profiling tools in SAS Enterprise Miner
- Profiling tools in SAS Enterprise Guide
- Full support of different databases, such as Oracle, Terradata and Netezza
- Full range of transformations aimed at data extraction
- Full support out of the box SCD1 and SCD2
- Implicit and explicit in-database processing

(T)ransform: Clean and Conform

Cleaning Machinery:

- Cleansing and data quality

Cleaning Controls:

- Error event schemas
- Audit dimension

Integration:

- Deduplication and conforming systems

Keys:

- Surrogate Key Generator

Result: Cleaned Tables and Conformed Dimensions

SAS Tools:

- Look up in memory and as a data merge in SAS DI
- Complex data quality processing rules in SAS DQ
- Exception handling and reporting in SAS DI
- Data exception tracking in SAS DI
- Out if the box, no code required, Surrogate Key Generators

(L)oad: Prepare for consumption

Time variance:

- SCD Management

Bridges:

- Bridge Tables
- Multi valued dimensions
- Special dimensions

Aggregations:

- Aggregate tables (datamarts)
- OLAP Cubes

Result: Datamarts, Fact and Dimension Tables ready for consumption

SAS Tools and Techniques:

- SCD Transformations in SAS DI Studio
- SAS Scalable Performance Data Server
- SAS OLAP Cubes powered by:
  - SAS OLAP Cube Server
  - SAS OLAP Cube Studio

(M)anage – ETL/M?

Control:

- Scheduling

Protect:

- Backup
- Recovery/Restart

Control:

- Version/Control
- Migration

Other:

- Workflow monitoring
- Performance/Scalability
- Lineage and Dependency
- Problem Escalation
- Pipeline/Parallelize
- Security
- Compliance
- Metadata Repository

SAS Tools and Techniques:

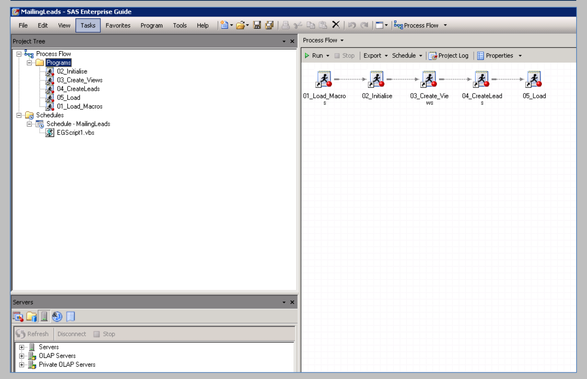
- SAS Metadata Server
- SAS Management Console
- Scheduling Plugin in SAS SMC
- Integration with IBM Suite
- Building In Process Scheduling
- Building OS Scheduling Integration
- Visual editing of flows in SAS SMC
- Visual editing of jobs in SAS DI Studio

Introduction to ETL with SAS®

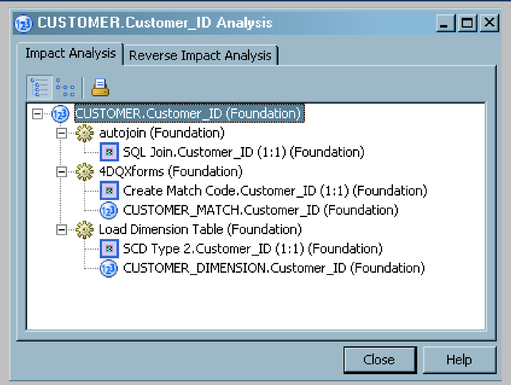
ETL Key Subsystems with SAS

Presented by Vasilij Nevlev

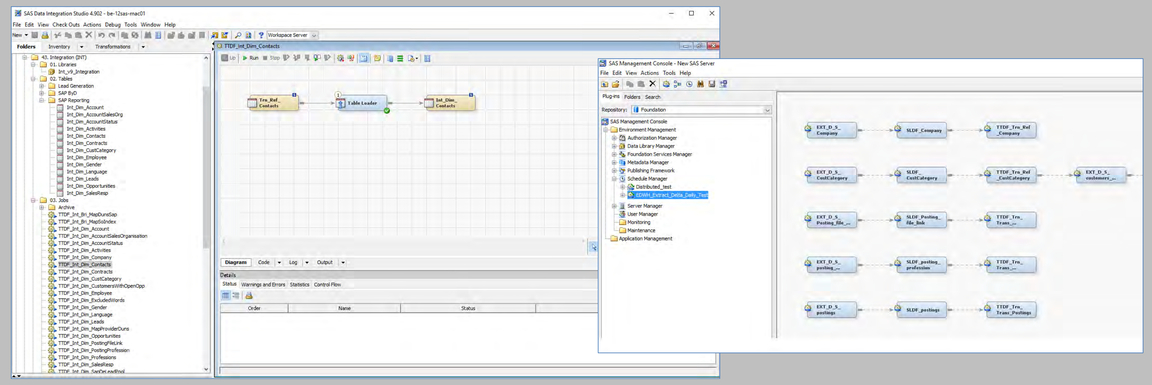
Simple ETL in SAS EG



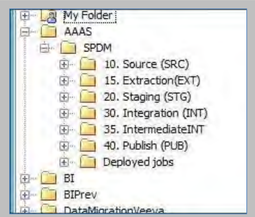
Data Lineage in SAS DI

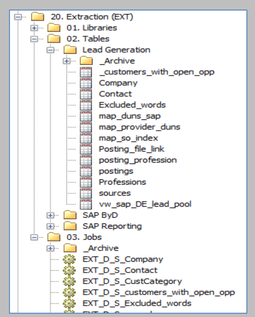


SAS DI Studio Job and SAS Management Console Flows

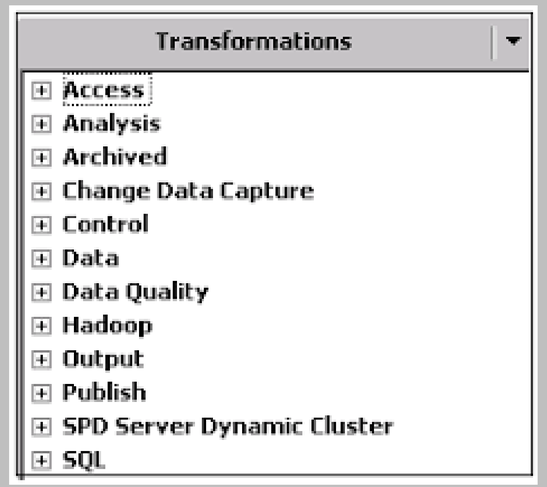


Metadata Organization in SAS

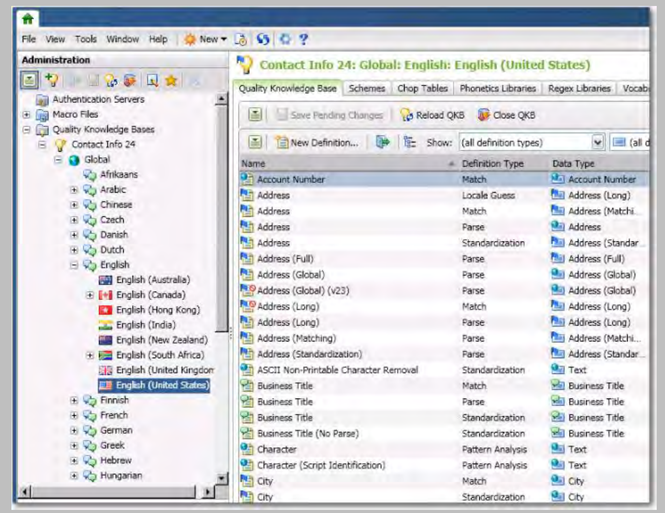




Data Transformations in SAS DI



Conclusion



SAS Offers all the tools necessary to implement all of the major ETL subsystems, from being able to profile the data sources, to organize metadata structures and orchestrate data flows according to business requirements.

Provides Knowledge Bases (shows to the left) that assesses data quality based on the region and locale. This can applied to addresses, names and lots of other areas.

SAS has the capability to almost every single ETL subsystem described earlier, so then the whole company ETL could be managed within a single ecosystem of tools and systems.

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF