

ETL Case Study



SCOTT HEFFRON
SLC SQL SERVER USER GROUP
11 JANUARY 2010

Who Am I?



- Scott Heffron
- Director of Engineering, Empowered Solutions Group
- URL: www.EmpoweredSolutionsGroup.com
- Personal Web Site: www.CTR-SQL.com

Agenda



- Explain Extract, Transform and Load
- Case Study
- Questions and Suggestions

Extract, Transform, and Load



A process is used to enable companies to move data from multiple sources, reformat and cleanse it, and then load the data into another area for analysis or operational system for support of the organizations business process.

Extract, Transform, Load (ETL)



- **Extract** – The process of reading data from an external source
- **Transform** – The process of converting the extracted data from its previous form into the form that is needed to complete the journey.
- **Load** – The ability to write the data into the target area.

ETL Life Cycle



- Initiation
- Build Reference Data
- Extract
- Validate
- Transform
- Stage
- Load
- Audit
- Archive
- Clean Up

Initiation



- Create a conversion working area to use for the transformation process
- Configure environment variables that are needed

Build Reference Data



- **Default Values** - These are values to use if the field is empty and the column has been designated as allowing a default value.
- **Translation Data** - These are mapping data elements to allow the system to know what the source data is and what target data should be substituted.

Extract



- **Pull the desired data from external sources and place them into preliminary data structures.**

Validate



- This is used to check that data falls within the appropriate parameters defined by the systems. A judgement as to whether data is valid is made possible by the validation program, but it cannot ensure complete accuracy.

Transform



- Clean
- Apply business rules
- Check for data integrity
- Create aggregates
- Format data as desired

Stage



- These table are used to create the new primary keys.
- Make sure referential integrity is in place be between the data that is being uploaded.

Load



- **Manage Triggers to speed up the load of data**
- **Manage Constraints to speed up the load of data**
- **Manage Primary Key Configuration**

Audit Report



- **Compliance with business rules**
- **Case of failure, helps to diagnose/repair**
- **Perform Random Spot Testing**

Clean Up



- **Archive Scripts**
- **Remove Un-needed Database Objects**
- **Get Sign-off**

Case Study



Business Process



A social service organization decides to purchase a case-management software suite. The organization will have client data that they are using to help their workers assist the clients with their day to day lives. The client data may be stored in several different storage mechanisms. This data needs to be uploaded into the new case-management package.

Environment Restrictions



- **SQL Server 2005 or 2008**
- **No SSIS**
- **External party needs to run the process**
- **External party will not do post validation**

Old Architecture



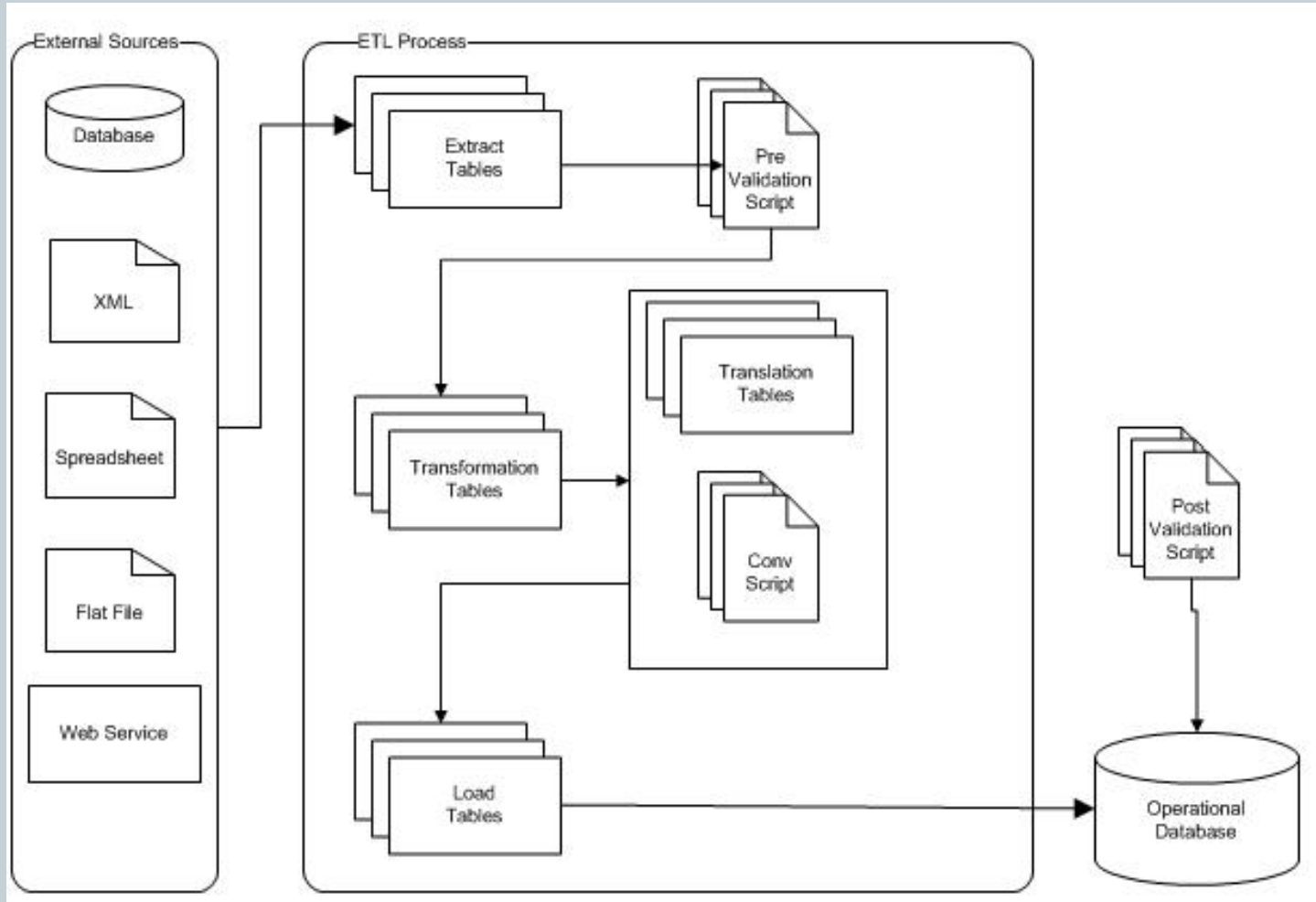
- **Each Conversion was brand new**
- **Hard coded translations with CASE and CAST**
- **No Testing**

ETL Architecture



- **T-SQL**
- **Run as a Batch Process (Windows)**
- **Use tables to allow system to translate source data to target data elements**

ETL Process Overview



Process Batch Structure - #1



```
SET DBServer=CTRSQL-01\SQLEXPRESS  
SET FileDirectory=M:\Projects\ESG\Goodwill\Conversion\
```

```
echo Start ETL Process!  
echo Y | time | find "urre"
```

```
-- Initialize Environment  
-- Extracting Area  
-- Transformation Area  
-- Load Area
```

```
echo ETL Process Complete!  
echo Y | time | find "urre"
```

```
-- Conversion Completed - Please test  
Pause
```

Extract Process Template Steps



- **Go to the correct Database**
- **Setup Environment**
- **Create Storage Area**
- **Define Variables needed for the script**
- **Upload source data to upload area**
- **Place Data into conversion area**
- **Check to see if records were uploaded**
- **Remove un-needed records**
- **Reset Environment**

Conversion Process Template Steps



- **Setup the environment**
- **Go to the correct Database**
- **Create Conversion Area**
- **Translate Data after Pre-Validation**
- **Collect Default Items and data**
- **Use Translation tables to Substitute Source data for Target data**
- **Process Business Logic**

Load Process Template Steps



- **Find Primary Keys for target table**
- **Create Load Table**
- **Assign Target Primary Key to Load Table**
- **Insert Records to Load Table from Transformation Tables**
- **Update Transformation Target PK with Actual PK**
- **Manage Triggers & Constraints from Target DB**
- **Insert Records from Load to Target Table**

Table Commands Used



- **Manage Identity constrained field**
 - SET IDENTITY_INSERT <TargetTable> ON
 - SET IDENTITY_INSERT <TargetTable> OFF
- **Manage Constraints**
 - ALTER TABLE <TargetTable> NOCHECK CONSTRAINT <Constraint Name>
 - ALTER TABLE <TargetTable> WITH NOCHECK CHECK CONSTRAINT <Constraint Name>
- **Reseeding the Identity Field**
 - DBCC CHECKIDENT('<TargetTable>', RESEED, @IDBase)

Table Commands Used



- **Manage Triggers**

- ALTER TABLE <TargetTable> DISABLE TRIGGER <Trigger Name>
- ALTER TABLE <TargetTable> ENABLE TRIGGER <Trigger Name>

Pro's



- Easy to Test
- Easy to Run
- See Results Quickly
- Easy to Add Another Area to the Process
- No Human Intervention During the Process
- A Deeper Understanding of how the system setup

Con's



- No Visual Tools
- Need Better Error Handling
- Need Better Rollback Scenarios

Brain Storming



HOW CAN WE MAKE THIS FASTER

End



- Questions?
- Thank You!!
 - Additional questions?
 - ✦ Scott@EmpoweredSG.net