CENTER *for the* INTEGRATION *of* IDEA DATA

DATA INTEGRATION TOOLKIT

# The CIID Data Integration Toolkit

## STEP 5 Implement and Perform Extract, Transform, and Load (ETL) Procedures

## Task 5.3: Create an ETL requirements document for the programmers.

### Overall Description of the Task

As programmers prepare to write the code needed to extract, transform and load (ETL) the integrated data, they will need clear and complete requirements on where the data reside, the rules of transforming and loading, and rules for validating the process.

### Why Is This Task Important?

- Requirements documents will provide the programmers the information needed to successfully code ETL.
- For a successful ETL, an accurate and timely process needs to be executed, with the appropriate validation steps at strategic points in the process. A requirements document can provide information needed for those validations.
- Data integration is only as good as the accuracy of the ETL.

### Activities

**5.3.1 Identify the specific data elements that will be integrated.**

Based on the original element documentation produced in Steps 3 and 4 and the program requirements in Task 5.2, identify the individual data elements that will support the purpose and functionality required for data integration. For each data element, perform the following actions:

- Verify the source of the data element, including database, table names, source attribute or column;
- Confirm the data element type and length;
- Describe the data element's destination table and attribute or column;
- Determine the years/semester/terms, etc. that the data are available;
- Allow for appropriate access to the data;
- Define any edits that will take place on the extracted data elements;

- Define a process for responding if records are rejected because of accuracy or validity issues;
- Define any edit reports that will be used for extract validations; and
- Define the format of the extracted data elements that will be used for transformation.

### 5.3.2 Define and document transformation rules.

Transformation rules are used to foster the integration of data from multiple sources by resolving differences in data definitions and code sets. Create specific rules for transforming or cleaning the data in order to provide the programmer with necessary information for putting the code in place to make these transformations. Use the following steps for defining transformation rules:

- Determine the input format for the transformation;
- Determine the business rules for transforming data from the original source format to the transformed format;
- Define a process to capture the pre- and post-transformation data for validation purposes;
- Define any edit reports to be used for transformation validation;
- Determine how records with errors will be handled;
- Define the format of the transformed data elements that will be used for loading the integrated data; and
- Document other data elements that may be integrated during transformation.

### 5.3.3 Describe the load process.

Define processes that describe the database and tables where the data will be loaded. Make sure these processes include accompanying documentation or metadata that support any cross validation that may be needed. These processes should help confirm that all the expected data were loaded. Use the following steps for defining the load process:

- Document where the data are to be loaded;
- Identify any audit trail data that will need to be captured as the data are loaded; and
- Ensure that proper database backup and recovery procedures are documented and referenced for execution before the ETL takes place.

### 5.3.4 Compile the ETL requirements document.

Compile the information from activities 5.3.1-5.3.3 to create an updated ETL requirements document. Ensure the information is sufficient and includes the appropriate technical details so the programmers will be able to write and test the code to efficiently and accurately integrate and load the needed data. Document the sign off process confirming the accuracy of the data as loaded when compared to its original state in the source database.

## What Is the Timeline for This Task?

| Activities in the Task | Participants | Duration |
|---|---|---|
| 5.3.1 Identify the specific data elements that will be integrated. | Project lead, business analysts, programmers | 1-2 weeks* |
| 5.3.2 Define and document transformation rules. | Project lead, business analysts, programmers | 3-5 days* |
| 5.3.3 Describe the load process. | Project lead, business analysts, programmers | 1-2 days* |
| 5.3.4 Compile the ETL requirements document. | Project lead, business analysts, programmers, data stewards | 3-5 days* |

**Note: The duration of these activities will vary based on the project scope.*

## What Are the Lessons Learned From Previous Efforts?

- Be careful to apply business rules accurately.
- Have the necessary documentation and ETL statistics in order to know the transformations and loading are done correctly. If the data are not accurately transformed correctly, integration and loading errors can occur.
- Provide sufficient detail so that the ETL programmer knows the specific data to extract, its location, precise specifications on the process of transforming the data, directions on where that data are to be loaded, and instructions for handling errors (both data errors and process errors).
- Execute proper database backup and recovery procedures before running the ETL.

**Contact Us:** By email at CIIDTA@aemcorp.com.
Visit the CIID website for more information at CIIDTA.org.