Extraction, Transformation, and Load (ETL) Specification

***Bureau of Information Systems***

<Project Name>

Table of Contents

[Revision History 3](#_Toc306868833)

[1. Introduction 4](#_Toc306868834)

[1.1 Summary 4](#_Toc306868835)

[1.2 Scope of Initiative 4](#_Toc306868836)

[1.3 Resources 4](#_Toc306868837)

[1.4 Definitions, Acronyms and Abbreviations 4](#_Toc306868838)

[2. Transformation Component Documentation 5](#_Toc306868839)

[2.1 Data Import/Extract 5](#_Toc306868840)

[2.2 Data Import Method 5](#_Toc306868841)

[2.3 Data Acquisition 5](#_Toc306868842)

[2.4 Data Integrity 6](#_Toc306868843)

[2.5 Data Refresh Frequency 6](#_Toc306868844)

[2.6 Data Security 7](#_Toc306868845)

[3. Data Quality 8](#_Toc306868846)

[3.1 Performance Measures and Standards 8](#_Toc306868847)

[3.2 Resources 8](#_Toc306868848)

# Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Description** | **Author** |
| <dd/mm/yy> | <x.x> | <details> | <name> |
| 2/23/2016 | 1.2 | Reviewed for content; changed DPW to DHS. | Don Pidich - EKMS |

# Introduction

The purpose of the Extraction, Transformation, and Load (ETL) Specification Document is to capture details that pertain specifically to ETL development to be used by the developer as an aid in ETL development.

## Summary

Address in the section:

* Who is the business owner (The Client)?
* What is the business that the Client does?
* What business need, problem or objective will be addressed?
* Where do they want the data?
* When do they need it by?
* Were previous attempts made to integrate the data?

## Scope of Initiative

Address in this section:

* The business process status
* Expected business objective
* Describe need for integration with any external system

## Resources

Address in this section:

* Requesting Business Client
* Project Manager or Business Analyst
* Detailed specification of the operational source systems

## Definitions, Acronyms and Abbreviations

Add items and terms that need to be defined in this section

# Transformation Component Documentation

[This section outlines a more detailed description of the processes that are currently utilized and the proposed processes developed to achieve the objectives of this initiative.]

## Data Import/Extract

Address in this section:

* Database tables names
* File Exports names
* Transactions processes
* Special: parameter or control files

## Data Import Method

Address in this section:

* Direct access by table or file on source system
* Extract from source
* File transferred
* Required permissions

## Data Acquisition

Address in this section:

* What data is needed?
* Will technical data specifications be provided?
	+ Fixed length
	+ Comma Separated
* How will the data sets be identified and selected?
* How often are source systems updated?
* How much data can be expected?
* How often will the data be received?
* What state is the source system?
	+ Production?
	+ Testing – when will the design be finalized?
* When can a production extract be ready for use by the developer?
* Are there any specific bottlenecks to getting the data?

Data Transform Specification

Address in this section:

* Has any technical analysis been previously performed?
* Do design specification or data models exist?
* What kind of history or trending information is needed?
* Are there any known fields that require a derived calculation?

## Data Integrity

Address in this section:

* How reliable is the source system data (e.g., empty fields, dates stored as text, invalid code values, text fields with odd or control characters, etc.)?
* Are the source data sets used by more than one data owner?
* How often does the source system change?
* How will changes be communicated to the users and development team?

## Data Refresh Frequency

Address in this section:

* How often should the Data Warehouse process refresh?
	+ Daily
	+ Weekly
	+ Monthly
	+ Quarterly
	+ Semi-Annually
	+ Other

## Data Security

Address in this section:

* Do monitoring requirements need to be satisfied?
* Are there any federally mandated HIPAA considerations?
* Is there need to build in additional privacy?
	+ Encryption
	+ Data masking
	+ Auditing
	+ Backups

Data Target/Result

Address in this section:

* How long will be data retention?
* How often can the data be purged?
* What interface will the Client/Users access the data?
	+ Cognos
	+ Reports
	+ Training
	+ Adhoc
	+ other…
* Does the project data have an expiration date?

# Data Quality

## Performance Measures and Standards

Address in this section:

* What measure will determine the success of this project?
* Do specific (Federal or otherwise) compliance measures exist?
* Do other specific outcomes exist?
	+ Expected Reports
	+ Totals and expected counts
	+ Technical comparison (e.g., using SQL, etc.)

## Resources

Address in this section:

* Who will be performing the user acceptance testing?
* Do they have a timeline or window for testing?
* When will they be doing it?