

# COMMONWEALTH OF PENNSYLVANIA

## DEPARTMENT OF HUMAN SERVICES

### **INFORMATION TECHNOLOGY STANDARD**

Name Of Standard: <b>Data Warehouse Standards</b>	Number: <b>STD-EKMS001</b>
Domain: <b>Enterprise Knowledge Management</b>	Category: <b>Data Warehouse</b>
Date Issued: <b>8/30/2002</b>	<b>DHS Bureau of Information Systems</b>
Date Revised: <b>2/19/2016</b>	

**Abstract:**

The purpose of this Standard is to establish enterprise-wide standards and guidance for Data Warehousing and its usage.

Data Warehousing is a process for building decision support systems and a knowledge-based application environment in support of both everyday tactical decision making and long-term business strategy. Data Warehouses and Data Warehouse applications are designed primarily to support the decision-making process by providing the decision makers with access to accurate and consolidated information from a variety of sources.

The primary objective of Data Warehousing is to collate information from disparate database sources and place the information in a format conducive to the decision-making process. This objective necessitates a set of activities more complex than merely collecting data and reporting against it. Data Warehousing requires both business and technical expertise and typically involves the following activities:

- Accurately identifying information to be placed in the Enterprise Data Warehouse (EDW);
- Identifying and prioritizing subject category areas to be included in the EDW;
- Managing the scope of each subject area implemented into the EDW on an iterative basis;
- Defining the correct level of consolidation in support of business decision making;
- Extracting, cleansing, aggregating, transforming, and validating the data to ensure accuracy and consistency;
- Defining and capturing metadata on all elements within the EDW;
- Defining record retention criteria and policies for all records on the EDW;
- Developing a scalable architecture to serve as the EDW's technical and application foundation, and identifying and selecting the hardware/software/ middleware components to implement it;

- Establishing a refresh program that is consistent with business needs, timing and cycles;
- Providing powerful and user-friendly tools at the desktop to access the EDW;
- Educating users about the realm of possibilities available to them through the EDW;
- Establishing EDW support mechanisms and training users to effectively utilize the desktop tools;
- Establishing processes for maintaining, enhancing, and ensuring the ongoing success and applicability of the EDW.

## General:

These standards apply to all developers, both commonwealth employees and business partners, who develop against the DHS Enterprise Data Warehouse. This policy ensures that all developed and implemented applications will facilitate enterprise-wide interoperability and standardization.

## Standard:

### Data Warehouse Technology Components:

Since Data Warehousing encompasses many technologies, it is not limited to one specialized area. Typically, the technical aspects of EDW are divided into the following areas:

**Extract, Transform and Load (ETL) tools** address the process of extracting, transforming, and loading data from various sources into the EDW using either custom-developed utilities or existing marketplace products. The DHS tool standard for this technology is Informatica Power Center.

**Data Warehouse repository tools** address products used for physical storage of EDW information. These tools range from products currently available in the marketplace; tools used in the commonwealth; and those recommended for deployment. The DHS Enterprise Data Warehouse utilizes an Oracle database.

**Enterprise reporting tools** address the end-user reporting tools to be used to satisfy reporting requirements. The standard reporting tools for the DHS Enterprise Data Warehouse are from Cognos.

**Architecture standards** address the various architecture patterns and available modeling techniques. ErWin is the DHS standard tool for modeling.

**Security standards** Security standards address authentication and access to the EDW for agency employees, and approval and audit processes by authorized agency members, subject to agency/enterprise security and privacy policies. The DHS Enterprise Data Warehouse uses Netegrity Site Minder for authentication and the security features in the Cognos, Oracle and Informatica tools for authorization.

### Data Warehouse Development:

The development of data warehouse subject areas or the addition of elements to an already defined data warehouse subject follows a defined development lifecycle – Data Warehouse Development Life Cycle (DWSDLC). Although the phases in a DWSDLC are similar to the phases in a typical Systems Development Life Cycle (SDLC), the tasks and methodology for completing the tasks in each phase are different. That is because the goals and the data structures of an operational system, that supports the running of a business, and a data warehouse, that supports analysis of the business, are different. The following describes the phases and tasks and deliverables to be completed in the DWSDLC.

## **PHASE 1 PLANNING**

This phase encompasses developing the overall scope and effort it will take to implement the subject area into the data warehouse.

- Defining the project scope
- Creating the project plan
- Review and establish the technical infrastructure
- Defining the necessary resources
- Defining the tasks and deliverables
- Initial capacity plan
- Architecture review

The deliverable from this phase is a Project Work Plan developed in Microsoft Project.

## **PHASE 2 GATHERING DATA REQUIREMENTS AND MODELING**

The tasks for this phase are as follows:

- Define all user business requirements including
  - What business questions are to be answered
  - Record retention strategy
  - Performance metrics
  - Criticality of data

The deliverable for this phase is General System Design (GSD) Document.

## **PHASE 3 DETAILED SYSTEM DESIGN**

This phase translates the user-oriented General System Design into a technical, computer-oriented detailed systems design. Data structures and processing are designed to the level of detail necessary to begin coding and includes the following tasks.

- Refine Data Requirements
- Develop Proof of Concept/Prototype
- Define Appropriate Reporting Applications

The Deliverables to be produced in this phase are the Detailed Systems Design (DSD) document, the Requirements Traceability Matrix, and a Logical Data Model.

## **PHASE 4 DATABASE DESIGN AND DEVELOPMENT**

This phase covers the database design and denormalization, and includes the following tasks.

- Designing the database, including fact tables, relationship tables and lookup tables.
- Denormalizing the data.
- Identifying keys.
- Creating index strategies.
- Creating database objects

The Deliverables to be produced in this phase are the Data Base Request Form and Erwin Physical Data Model. The Database Request Form and Erwin Data Model should adhere to all DHS Database standards as defined within both the Data Domain and the Knowledge Management Domain on the BIS website (<http://mydhs/oa/bis/index.htm>). These deliverables (final or draft versions) must be submitted to the Enterprise Knowledge Management Section (EKMS) two (2) business days before a formal review with the BIS Database Section is held. This time will be used to review and identify required changes before the formal meeting.

## **PHASE 5          DATABASE MAPPING AND TRANSFORMATION**

This phase is started in conjunction with the database design phase. This phase encompasses identifying the sources of the data, determining the transformation logic and mapping the source data to the target data warehouse table design.

Tasks in this phase include:

- Defining the sources systems
- Determining file layouts
- Developing transformation specifications
- Mapping source to target data
- Defining the test plan and queries that will be used to validate the data

The deliverables to be produced in this phase are:

- Table Metadata Document and Package Metadata Document
- Source data acquisition processes which will be developed in accordance with the System Development Life Cycle standards in place of all application program development
- Test plan

## **PHASE 6          DATA EXTRACTION AND LOAD**

This phase covers the actual extraction of the source data, running of the transformation logic, and loading the development database.

Tasks in this phase include:

- Developing procedures to extract and move the data
- Developing procedures to load the data
- Developing extract programs
- Developing the transformation logic
- Develop monthly load programs if necessary
- Incorporate metadata into the EKMS metadata application

Testing the extract, transformation and loading procedures.

The Deliverables to be produced in this phase is the Data Extract and Load Procedures Document.

## **PHASE 7          TEST PLAN**

This phase includes the configuring the interface tools that will be used to access that data warehouse and determining the reports and queries that will be used by the testers for data validation and approval of BI reports and cubes.

Tasks in this phase include:

- Configuring the interface tools.
- Defining the test scenarios for user data validation
- Defining and develop the reports and queries to be created for data validation
- Defining test queries for validation of the table data.

The deliverable out of this phase is a comprehensive user test plan that is developed by both the developers and the testers

## **PHASE 8 DATA VALIDATION AND TESTING**

This phase has a few components. It addresses both the user acceptance testing and validation of the data that was loaded into the data warehouse. Standard data validation processes are included throughout the data extraction, transformation, load development. The test conditions identified in Phase 7 should be used during this phase. Once the data has been validated the confirmation of the readiness of the BI reports and cubes is received, the applications then go into User Acceptance testing.

Tasks in this phase include:

- UAT of loaded data
  - Run data validations
  - Reviewing all errors
  - Determine how to and correct data that is in error.
  - Re-Run reports and validate results.

The deliverables from this phase are:

- Test Report
- User Sign Off

## **PHASE 9 PRE-IMPLEMENTATION/ AUTOMATING PROCESS**

This phase addresses the automation of the extract, transformation and load process. It can be started after Phase 6.

Tasks in this phase include:

- Creating operational procedures if applicable.
- Creating scripts to automate the various processes.
- Providing information to the help desk.

The Deliverables to be produced in this phase are:

- Operational Procedures
- Backup and Recovery Procedures
- Database Scripts

## PHASE 11 TRAINING

This phase focuses on developing training for the user community. Users will have existing business knowledge; training will focus on tool capabilities and usage.

Tasks in this phase include:

- Creating procedures for User Support
- Designing training programs or contracting out for those services

The deliverables from this phase are:

- User Support Procedures
- Training Plan
- All training material

## PHASE 12 ROLLOUT THE DATA WAREHOUSE ADDITIONS

This phase includes the necessary tasks to deploy the data warehouse to the user community.

Tasks in this phase include:

- Installing components necessary to connect users to the data warehouse.
- Installing middleware and end user software.
- Implementing the extract programs and scripts.

The deliverables for this phase are:

- Element Transfer Request Form
- Completed User Request Forms
- Deployment Playbook Creation and Review
- Informatica Deployment Request Form

### Exemptions from this Standard:

There will be no exemptions to this standard.

### Refresh Schedule:

All standards and referenced documentation identified in this standard will be subject to review and possible revision annually or upon request by the DHS Information Technology Standards Team.

### Standard Revision Log:

Change Date	Version	Change Description	Author and Organization
08/30/02	1.0	Initial Creation	Unknown

04/02/07	2.0	Revised content and format	Arlene DiMarco DTE/EKMS
02/19/16	3.0	Reviewed for content. Changed 'Public Welfare' to 'Human Services'; changed 'DPW' to 'DHS'; removed cube, report, and map information (to be addressed elsewhere).	Don Pidich EKMS