Information Technology Policy
Commonwealth of Pennsylvania
Governor's Office of Administration/Office for Information Technology

ITP Number:	**ITP-INF004**
ITP Title:	**Data Warehousing Policy**
Issued by:	**Deputy Secretary for Information Technology**
Date Issued: November 7, 2006	**Date Revised: November 18, 2010**
Domain:	**Information**
Discipline:	**Data Warehousing**
Technology Area:	**Data Warehousing**
Revision History Date:	**Description:**
3/23/09	**Added the use of IES Data Warehouse for ERP related applications.**
11/18/2010	**ITP Refresh**

Abstract:

The purpose of this Information Technology Policy (ITP) is to establish enterprise-wide standards and guidance for Data Warehousing.

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 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 data warehouse;
- Extracting, cleansing, aggregating, transforming, and validating the data to ensure accuracy and consistency;
- Identifying and prioritizing subject category areas to be included in the data warehouse;
- Managing the scope of each subject area implemented into the warehouse on an iterative basis;
- Developing a scalable architecture to serve as the data warehouses technical and application foundation, and identifying and selecting the hardware/software/ middleware components to implement it;
- Defining the correct level of consolidation in support of business decision making;
- 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 data in the warehouse;
- Educating users about the business benefits made possible through data warehousing;
- Establishing data warehouse 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 data warehouse.

Data Warehouse Technology Components

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

- **Extract, Transform and Load (ETL) tools** address the process of extracting, transforming, and loading data from various agency application sources into the operational data store/data warehouse using either custom-developed utilities or existing marketplace products.

- **Data Warehouse repository tools** address products used for physical storage of data warehouse information. These tools range from products currently available in the marketplace; tools used in the Commonwealth; and those recommended for deployment.

- **Enterprise reporting tools** address the end-user reporting tools to be used to satisfy agency-specific reporting requirements.

- **Architecture standards** address the various architecture patterns and available modeling techniques; techniques currently used in the Commonwealth; and emerging modeling techniques.

- **Data exchange and security standards** address the information exchange mechanisms for horizontal and vertical information exchange (e.g., exchange of information within the Commonwealth, as well as with local, state and federal agencies). Security standards address authentication and access to the data warehouse for agency employees, and approval and audit processes by authorized agency members, subject to agency/enterprise security and privacy policies.

GEN-INF004A: *Introduction to Data Warehousing* provides an introduction to data warehousing technology.

BPD-INF004B: *Best Practice Approach to Data Warehousing* documents best practices in data warehousing.

STD-INF004C: *Data Warehousing Product Standards* provides guidance to agencies on the current standards and the status of other data warehousing solutions that are being used or being considered for use.

GEN-INF004D*: Data Warehousing Product Availability* provides information on the availability and licensing of current data warehousing product standards.

General:

This ITP applies to all departments, boards, commissions and councils under the Governor's jurisdiction. Agencies not under the Governor's jurisdiction are strongly encouraged to follow this policy.

Policy:

Agencies are to utilize existing data warehousing solutions or build and implement a Data Warehousing solution when a business requirement necessitates reporting which summarizes or combines data from multiple sources.

Agencies are to leverage the data warehousing solution provided by Integrated Enterprise Systems (IES) for Enterprise Resource Planning (ERP) applications and/or applications that associated with ERP data when the IES data warehousing solution meets agencies business requirements. ERP applications include financial, human resources, customer relationship management, supplier relationship management, platform life cycle management, supply chain management and material management enterprise applications.

Data warehousing standards are defined for Integrated Enterprise Systems in STD-INF004C*.*

The data warehouse is to enable access to centrally stored information to accommodate business reporting requirements.

The data warehouse is to contain a subset of information from operational systems optimized for data retrieval and reporting to support performance measurement against agency/enterprise goals and objectives.

Access to the data warehouse and the information within is to be based upon each user's job requirements and access level approved by the authorized agency officer, subject to agency/enterprise security and privacy policy.

To promote and maintain consistency across Commonwealth agencies, any data warehouse model is to follow the *Core Citizen Data Model and Data Elements* (ITP-INF003D) for citizen-centric common data elements described in the citizen model. In addition, the data warehouse model is to follow database standards referenced in ITP-INF001, *Database Management Systems*.

Data warehouse solutions are to enable data federation to support horizontal and vertical exchange (between Commonwealth agencies and potential future centralized Commonwealth-wide data warehouses).

Any data warehouse is to reside on hardware separate from operational and transaction-related systems, thereby mitigating potential performance issues with these systems.

Any data warehouse is to have an efficient extracting/harvesting process from operational and transaction-related systems in order to minimize the impact in either performance or availability of these systems.

Any custom development done for ETL is to adhere to existing Commonwealth standards.

Standard methods such as ANSI-SQL, Open Database Connectivity (ODBC), and Java Database Connectivity (JDBC) will be used to access the any data warehouse.

Due to privacy and security constraints, data warehousing solutions will physically operate on Commonwealth infrastructure.

Data Quality is critical in order to establish the integrity of the information and user confidence in the validity of the resulting output. Agencies are responsible for taking appropriate automated and manual measures to maintain a high degree of quality of the information in the warehouse.

Training requirements for each model of the data warehouse are to be met before a user will be granted access to data.

Agencies will determine the level of mission criticality of their data warehouse. The infrastructure and operational procedures necessary to support the Data Warehouse will be designed and implemented commensurate to the level of mission criticality of the data warehouse.

Definitions of Terms:

Ad Hoc Query
Any query that cannot be determined prior to the moment the query is issued. Also, any query which consists of dynamically constructed SQL, usually constructed by desktop-resident query tools.

Administrative Data
The data used by a warehouse administrator to manage data in a data warehouse. Examples of administrative data are *user profiles* and *order history data*.

Aggregations

The process of consolidating data values into a single value. For example, sales data could be collected on a daily basis and then aggregated at the week and/or month level. The data can then be referred to as *aggregate data*. Aggregation is synonymous with *summarization*, and *aggregate data* is synonymous with *summary data*.

Analyst
A user who creates views for analytic interpretation of data, performs calculations, and distributes the resulting information in the form of reports.

Data Integration
The consolidation and reconciliation of dispersed data maintained by organizations in multiple, heterogeneous systems for analytical purposes. Data can be accessed, extracted, moved, loaded, validated, and transformed.

Data Loading
Data loading is the process of populating the Data Warehouse. Data loading is provided by Database Management System (DBMS)-specific load processes, DBMS insert processes, and independent fast-load processes.

Data Mapping
Mapping is the assignment of a source data element to a target data element.

Data Mart
A data warehouse designed for a particular line of business, such as sales, marketing, or finance. In a dependent data mart, the data can be derived from an enterprise-wide data warehouse. In an independent data mart, data can be collected directly from sources.

Data Migration
Data migration is the process of transferring data from one repository to another.

Data Mining
Data mining is the process of sifting through large amounts of data to produce data content relationships. It also refers to the technique by which a user utilizes software tools to look for particular patterns or trends. This technique can uncover future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. It is also known as *data surfing*.

Data Model
A data model is a logical map that represents the inherent properties of the data independent of software, hardware or machine performance considerations. The model shows data elements grouped into records, as well as the association around those records.

Dimensions
Dimension tables describe the business entities of an enterprise, represented as hierarchical, categorical information such as *time, departments, locations, and products*. Dimension tables are sometimes called *lookup* or *reference tables.*

ETL
Extract, Transform, and Load. ETL refers to the methods involved in accessing and manipulating source data and loading it into a data warehouse.

Facts
Data, usually numeric and additive, that can be examined and analyzed. Examples include *sales, cost, and profit*. A table in a star schema contains *facts*. A fact table typically has two types of columns: Those that contain facts (e.g., numbers), and those that are foreign keys to dimension tables. The primary key of a fact table is usually a composite key constructed with all of its foreign keys.

Snowflake Schema

A snowflake schema is a set of tables comprised of a single, central fact table surrounded by normalized dimension hierarchies. Each dimension level is represented in a table. Snowflake schemas implement dimensional data structures with fully normalized dimensions. Snowflake schemas are an alternative to star schemas.

Star Schema

A star schema is a relational schema whose design represents a multidimensional data model. The star schema consists of one or more fact tables and one or more dimension tables related through foreign keys.

Refresh Schedule:

All standards identified in this ITP are subject to periodic review and possible revision, or upon request by the Enterprise Architecture Standards Committee (EASC).

Exemption from This Policy:

In the event an agency chooses to seek an exemption, for reasons such as the need to comply with requirements for a federally mandated system, a request for waiver may be submitted via the Commonwealth of PA Procurement and Architectural Review (COPPAR) process. Requests are to be entered into the COPPAR Tool located at http://coppar.oa.pa.gov/. Agency CIO approval is required. Contact your agency CoP Planner for further details or assistance.

Questions:

Questions regarding this policy are to be directed to RA-ITCentral@pa.gov.

Policy Supplements:

GEN-INF004A: Introduction to Data Warehousing
BPD-INF004B: Best Practice Approach to Data Warehousing
STD-INF004C: Data Warehousing Product Standards
GEN-INF004D: Data Warehousing Product Availability

References:

ITP-INF001: Database Management Systems
BPD-INF003D: Core Citizen Data Model and Data Elements

1. Process/Project DWH - Data Warehouse Process, <http://www.gantthead.com/process/processMain.cfm?ID=2-2357-2> (14 April 2006)