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

STD Number:	**GEN-INF004A**	
STD Title:	**Introduction to Data Warehousing**	
Issued by:	**Deputy Secretary for Information Technology**	
Date Issued:	**November 7, 2006**	**Date Revised: November 18, 2010**
Domain:	**Information**	
Discipline:	**Data Administration**	
Technology Area:	**Data Warehousing**	
Referenced by:	**ITP-INF004**	

Revision History Date:	**Description:**
11/18/2010	**ITP Refresh**

Introduction:

Data Warehousing:

Data Warehousing systems have reached a new level of maturity as both an IT discipline and a technology.

Main Document Content:

Data Warehouse systems assist government organizations with improved business performance by leveraging information about citizens, business partners, and internal government operations. This is done by:

- Extracting data from many sources, e.g., application databases, various local and federal government repositories, external agency partners.
- Centralizing, organizing, and standardizing information in repositories, such as Data Warehouses and Data Marts. This includes cleaning, appending, and integrating additional data.
- Providing analytical tools that allow a broad range of business and technical specialists to run queries against the data to uncover patterns and diagnose problems.

Extract, Transform and Load (ETL)

Data integration technology is generally used to extract transactional data from internal and external source applications to build the Data Warehouse. This process is referred to as ETL (Extract, Transform, Load). Data is extracted from its source application or repository, transformed to a format needed by a Data Warehouse, and loaded into a Data Warehouse. Data integration technology works hand-in-hand with technologies like Enterprise Information Integration (EII), database replication, Web Services, and Enterprise Application Integration (EAI) to bridge proprietary and incompatible data formats and application protocols.

Data Warehouses and Data Marts

A Data Warehouse, or Data Mart, stores tactical or historical information in a relational database allowing users to extract and assemble specific data elements from a complete dataset to perform analytical functions. The Data Warehouse can be constructed according to schema (e.g., star, snowflake), data composition (values and attributes), dimension levels, and descriptors. Data Marts allow additional segmentation within a broader Data Warehouse environment.

Query, Reporting and Analysis
Technical and business analysts use a variety of tools to access data, analyze information, and view the results. These include:

Query and Reporting Tools:
Most data warehouse systems allow users to perform historical, "slice-and-dice" analysis against information stored in a relational database. This type of analysis answers the "what?" and "when?" inquiries. A typical query might be, "What was the total revenue for the eastern region in the third quarter?" Often, users take advantage of pre-built queries and reports.

On-Line Analytical Processing (OLAP) and Data Mining:
OLAP analytical engines and data mining tools allow users to perform predictive, multidimensional analysis, also known as "drill-down" analysis. These tools can be used for forecasting, customer profiling, trend analysis and even fraud detection. They answer "what if" and "why?" questions, such as, "What would be the effect on the eastern region of a 15 percent increase in the price of the product?"

Information Delivery:
Query results and reports can be delivered through dedicated desktop applications, dashboards, intranets, and extranet portals.

Definitions of Terms[2]:

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

Administrative Data
In a Data Warehouse, the data that helps a warehouse administrator manage the 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 be aggregated to the week level, the week data could be aggregated to the month level, and so on. The data can then be referred to as aggregate data. Aggregation is synonymous with summarization, and aggregate data is synonymous with summary data.

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

Data Integration
Pulling together and reconciling dispersed data for analytic purposes that organizations have maintained in multiple, heterogeneous systems. Data needs to be accessed and extracted, moved and loaded, validated and cleaned, and standardized and transformed.

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

Data Mapping
Data mapping is the process of assigning a source data element to a target data element and defining the transformation rules and processes associated with the mapping.

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 repository to another.

Data Mining
A technique using software tools for identifying patterns or trends in data. Data mining is the process of sifting through large amounts of data to produce data content relationships. It can predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. This is also known as data surfing.

Data Model
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.

Data Staging Area
A data staging area is a system that stands between the legacy systems and the analytics system, usually a Data Warehouse and sometimes an Operational Data Store (ODS). The data staging area is considered the "back room" portion of the Data Warehouse environment. The data staging area is where the extract, transform and load (ETL) takes place and is out of bounds for end users.

Data Steward
The data steward acts as the conduit between information technology and the business portion of a company with both decision support and operational help. The data steward has the challenge of guaranteeing that the corporation's data is used to its fullest capacity.

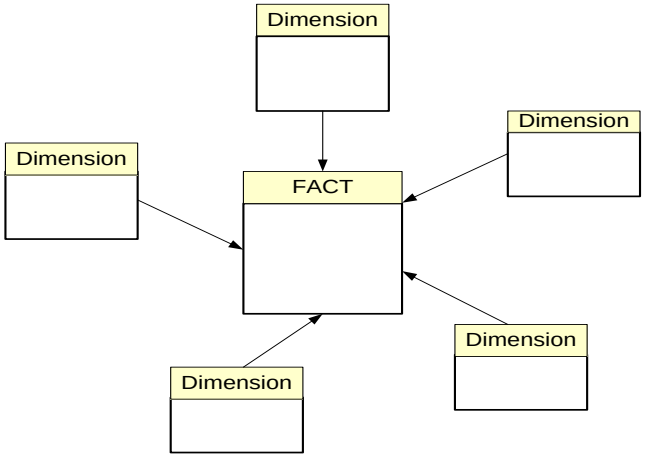
Dimensions
Dimension tables describe the business entities of an enterprise, represented as hierarchical, categorical information such as time, departments, locations, and products. Dimension tables are sometimes called lookup or reference tables.

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 and those that are foreign keys to dimension tables. The primary key of a fact table is usually a composite key that is made up of all of its foreign keys.

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.



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.