Information Technology Policy

Introduction to Data Warehousing

ITP Number GEN-INF004A	**Effective Date** November 7, 2006
Category **Information**	**Supersedes** None
Contact RA-ITCentral@pa.gov	**Scheduled Review** May 2022

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

2. **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, and/or external agency partners.
   - Centralizing, organizing, and standardizing information in repositories such as Data Warehouses and Data Marts. This includes cleansing, 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 together 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:

   Business Intelligence Software:
   Business intelligence software is a type of application software designed to retrieve, analyze, transform and report data for business intelligence. The applications generally read data that has been previously stored, often - though not necessarily - in a data warehouse or data mart.

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.

3. Definitions:

Data Staging Area

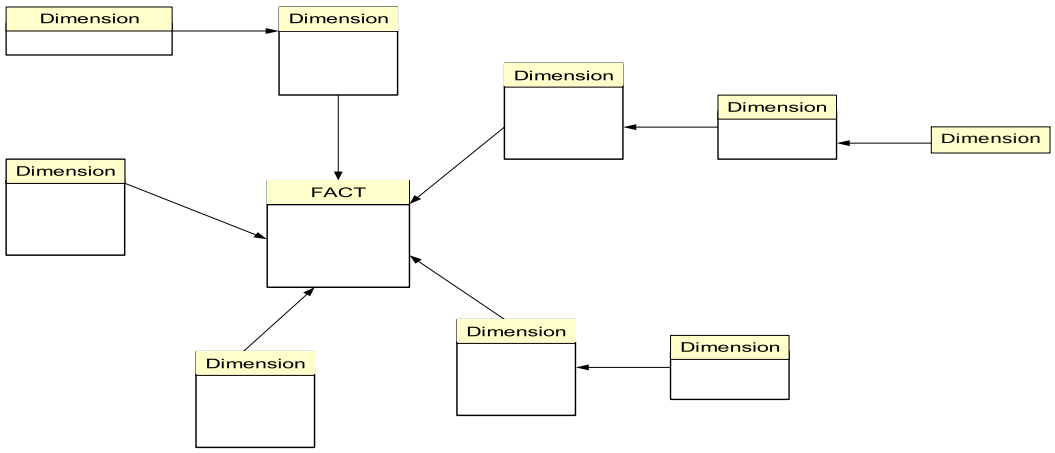
A data staging area is a system that stands between the legacy systems and the analytics system, usually a Data Warehouse and sometimes an Operational Data Store (ODS). The data staging area is considered the "back room" portion of the Data Warehouse environment. The data staging area is where the extract, 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.

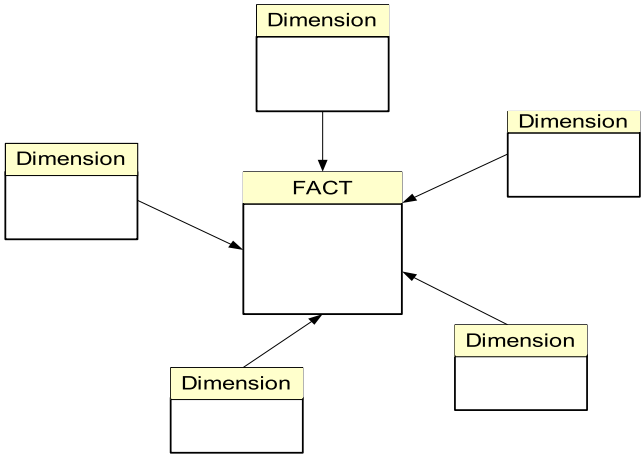
Snowflake Schema

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.



This chart contains a history of this publication's revisions:

Version	Date	Purpose of Revision
Original	11/7/2006	Base Document
Revision	11/18/2010	ITP Refresh
Revision	05/14/2021	ITP Refresh