Information Technology Policy
Data Warehousing Policy

Number
ITP-INF004

Effective Date
November 7, 2006

Category
Information

Supersedes
None

Contact
RA-ITCentral@pa.gov

Scheduled Review
March 2024

1. Purpose

This Information Technology Policy (ITP) establishes enterprise-wide standards and guidance for Data Warehousing.

2. Scope

This ITP applies to all offices, departments, boards, commissions, and councils under the Governor's jurisdiction (hereinafter referred to as "agencies"). Agencies not under the Governor's jurisdiction are strongly encouraged to follow this ITP.

Third-party vendors, licensors, contractors, or suppliers shall meet the policy requirements of the Commonwealth's ITPs that are applicable to the products and services provided to the Commonwealth.

3. Definitions

Aggregations: The process of consolidating data values into a single value. For example, sales data could be collected daily and then aggregated at the week 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.

Data Integration Technology: Technology that facilitates the consolidation and reconciliation of dispersed data maintained by agencies in multiple, heterogeneous systems for analytical purposes. Data can be accessed, extracted, moved, loaded, validated, and transformed.

Data Lake: A system or repository of data stored in its natural or raw format, usually object blobs or files. A Data Lake is usually a single store of data including raw copies of source system data, sensor data, social data, etc., and transformed data used for tasks such as reporting, visualization, advanced analytics, and machine learning. A Data

Lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs), and binary data (images, audio, video).

Data Mart: A subset of the enterprise data warehouse that is designed for a particular line of business, such as sales, marketing, or finance.

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. Often performed by leveraging Artificial Intelligence or Machine Learning.

Data Model: An abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of entities. **Data Warehouse:** A storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprisewide data analysis and reporting for predefined business needs.

Data Warehousing: 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.

Dimension Tables: Dimension Tables describe the entities, represented as hierarchical, categorical information such as time, departments, locations, and products. Dimension Tables are sometimes called *lookup* or *reference tables.*

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

Fact: A value or measurement, which represents a Fact about the managed entity or system. Facts, as reported by the reporting entity, are said to be at raw level; Examples include sales, cost, and profit.

Fact Table: A table in a Star Schema that 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 its foreign keys.

Online Analytical Processing (OLAP): A type of software used to perform rapid multidimensional analysis on large volumes of data from a data warehouse or some other centralized data store. This is accomplished by extracting data from multiple relational data sets and reorganizing it into a multidimensional format that enables fast processing.

Snowflake Schema: A snowflake schema is a multi-dimensional Data Model that is an extension of a star schema, where Dimension Tables are broken down into subdimensions. Snowflake schemas are commonly used for business intelligence and reporting in OLAP data warehouses, Data Marts, and relational databases.

Star Schema: A star schema is a multi-dimensional Data Model used to organize data in a database so that it is easy to understand and analyze. Star schemas can be applied to data warehouses, databases, Data Marts, and other tools.

4. Policy

When a business requirement necessitates reporting that summarizes or combines data from multiple sources, Agencies shall consider utilizing data warehouse technology.

Prior to building or implementing a new data warehouse, agencies shall determine if a data warehouse exists that meets their needs. If a data warehouse exists that contains all the required data, agencies shall utilize it rather than implementing a new data warehouse.

Agencies shall leverage the Data Warehousing solution provided by Integrated Enterprise Systems (IES) for Enterprise Resource Planning (ERP) applications or ERP data when the IES Data Warehousing solution meets agencies business requirements. Additional information regarding the Data Warehousing solution offered by IES for ERP applications is available in *ITP-SFT008, Enterprise Resource Planning (ERP) Management.*

Agencies shall use one of the current standards for Data Warehousing as defined in *STD-INF004C, Data Warehousing Product Standards*. Information regarding the availability and licensing of current Data Warehousing product standards is available in *GEN-INF004D, Data Warehousing Product Availability*.

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

Agencies shall take appropriate automated and manual measures to continually improve the quality and accuracy of the information in the data warehouse working towards a goal of 95% or greater. Agencies shall maintain metrics regarding the current quality and accuracy of the information in their data warehouses. Data quality and accuracy are critical to establishing the integrity of the information and user confidence in the validity of the resulting output.

Additionally, agencies shall ensure:

- All data warehouse models follow the database standards referenced in *ITP-INF001, Database Management Systems*.
- All data warehouse models follow *BPD-INF003D, Core Citizen Data Model and Data Elements* for citizen-centric common data elements described in the citizen model.
- Any custom development done for ETL adheres to existing Commonwealth standards.
- Data Warehousing solutions physically operate within Commonwealth-approved

environments.

- Any data warehouse contains a subset of information from operational systems optimized for data retrieval and reporting to support performance measurement against agency or enterprise goals and objectives.
- Data warehouses utilize Commonwealth approved IT resources separate from operational and transaction-related systems, thereby mitigating potential performance issues with these systems.
- Any data warehouse has an efficient extracting or harvesting process separate from operational and transaction-related systems in order to minimize the impacts to the performance or availability of these systems.
- Standard methods such as ANSI-SQL, Open Database Connectivity (ODBC), Java Database Connectivity (JDBC), or OData (Open Data Protocol) RESTful APIs are used to access any data warehouse.
- Training requirements are established and met for each model of the data warehouse (Snowflake and Star Schema) before a user will be granted access to data.
- Access to the data warehouse and the information within the data warehouse shall be based upon each user's job requirements and access level approved by the authorized agency officer and in accordance with *Management Directive 205.34 Amended Commonwealth of Pennsylvania Information Technology Acceptable Use Policy*.
- Data federation is enabled to support horizontal and vertical exchange between agencies and centralized Commonwealth-wide data warehouses utilizing data sharing agreements and role-based access control.

Agencies are encouraged to review the best practices contained in *BPD-INF004B, Best Practice Approach to Data Warehousing.*

5. Responsibilities

5.1 Agencies shall:

Comply with the requirements as outlined in this ITP.

5.2 Office of Administration, Office for Information Technology shall:

Comply with the requirements as outlined in this ITP.

5.3 Third-party vendors, licensors, contractors, or suppliers shall:

Comply with the requirements as outlined in this ITP that are applicable to the products or services provided to the Commonwealth.

6. Related ITPs/Other References

- Definitions of associated terms of this policy are published on the Office of Administration's public portal: *http://www.oa.pa.gov/Policies/Pages/Glossary.aspx*
- Commonwealth policies, including Executive Orders, Management Directives, and IT Policies are published on the Office of Administration's public portal: *http://www.oa.pa.gov/Policies/Pages/default.aspx*
- *Management Directive 205.34 Amended, Commonwealth of Pennsylvania Information Technology Acceptable Use Policy*

- *ITP-ACC001, Information Technology Digital Accessibility Policy*
- *ITP-INF001, Database Management Systems*
- *BPD-INF003D, Core Citizen Data Model and Data Elements*
- *BPD-INF004B, Best Practice Approach to Data Warehousing*
- *STD-INF004C, Data Warehousing Product Standards*
- *GEN-INF004D, Data Warehousing Product Availability*

7. Authority

Executive Order 2016-06, Enterprise Information Technology Governance

8. Publication Version Control

It is the Authorized User's responsibility to ensure they have the latest version of this publication, which appears on https://itcentral.pa.gov for Commonwealth personnel and on the Office of Administration public portal: http://www.oa.pa.gov/Policies/Pages/default.aspx. Questions regarding this publication shall be directed to RA-ITCentral@pa.gov.

9. Exemption from this Policy

In the event an agency chooses to seek an exemption from the guidance within this ITP, a request for a policy waiver shall be submitted via the enterprise IT policy waiver process. Refer to *ITP-BUS004, IT Policy Waiver Review Process* for guidance.

This chart contains a history of this publication's revisions. Redline documents detail the revisions and are available to CWOPA users only.

Version	Date	Purpose of Revision	Redline Link
Original	11/7/2006	Base Document	N/A
Revision	03/23/2009	Added the use of IES Data Warehouse for ERP related applications.	N/A
Revision	11/18/2010	ITP Refresh	N/A
Revision	05/14/2021	• ITP Refresh • Added to ITP template • Added Third-Party vendor to Scope and Responsibilities Updated Exemption Section	N/A
Revision	03/27/2023	Updated definitions Removed objective section Informational content regarding data warehousing technology components and activities removed Reorganized policy section Added requirement for Agencies to maintain metrics regarding the current quality and accuracy of the information in their data warehouses Updated references Rescinding GEN-INF004A	Revised IT Policy Redline <03/27/2023>