pennsylvania
OFFICE OF ADMINISTRATION

Information Technology Policy

Best Practice Approach to Data Warehousing

Number
BPD-INF004B

Effective Date
November 7, 2006

Category
Information

Supersedes
None

Contact
RA-ITCentral@pa.gov

Scheduled Review
March 2024

1. Introduction

Data warehouse systems are powerful tools used in decision making. However, they can also be costly and time consuming. With the proper approach and design, these tools can be successful and beneficial to the entire enterprise. This document describes some high-level recommendations for ways to specialize the software engineering process for Data Warehousing. The following steps are recommended when developing Data Warehousing projects.

- Creating a Vision
- Identifying Business Objectives
- Understanding and Creating the Architecture
- Determining Roles and Responsibilities
- Creating an Implementation Plan

The sections below describe each of these steps. Although these steps are presented sequentially, they will rarely be executed in a waterfall fashion. Rather, a data warehouse will often be developed iteratively. In many instances, the use of prototyping or proofs-of concept will serve as a vital technique towards determining the vision, objectives, and architecture of the data warehouse and the expected output.

2. Best Practices

2.1 Creating a Vision:

A data warehouse requires the creation of a common vision to address needs at the enterprise and agency level. This vision shall also be properly communicated throughout the agency or enterprise as it evolves.

2.2 Identifying Business Objectives:

The identification of critical business information assists in the measurement of objectives against performance and establishes the necessary metrics required by agency business managers and other stakeholders prior to any business decision. For

this purpose, the agency's senior managers and other stakeholders shall be identified and interviewed and any existing information architecture shall be reviewed as well.

The following questions are samples that can be posed to stakeholders during project startup:

Do agency managers possess the information necessary to:
- measure, manage, and monitor the agency's business on a regular basis?
- identify and find causes behind the poor performance?
- monitor the decisions made and fine tune them as they evolve?
- determine the difference to the bottom line had this information been available beforehand?

During the interview process, several things shall be considered:
- Does the agency have a "vision statement"?
- Are business goals and objectives a direct translation of the agency's vision statement?
- Has the agency identified its key performance indicators?
- How does the proposed solution help measure each performance indicator?
- How is critical business information classified (e.g., citizen, education, criminal, driver)? These classifications serve as the foundation for the information model.
- What are the decisions agency decision makers like to make with the help of the Data Warehousing environment?
- What are the decisions analysts and end users like to make with the help of the proposed solution?

2.3 Understanding and Creating the Architecture:

2.3.1 Conceptual Information Architecture:

A conceptual model is created to help define and prioritize data and the sources of data in a decision support environment. (Please see *ITP-INF003A, Data Modeling Standards*). Conceptual models for data warehouses also describe the following types of information:
- How data will be grouped within the data warehouse and Data Marts.
- How information in the Warehouse is organized for access by users. For example, labels, definitions, and data item groups shall be aligned with the way business professionals approach business analysis.

Knowledge about the design of the database or the peculiarities associated with each data source is not a requirement for the business managers and analysts. All information requirements for decision making, regardless of source, shall conform to a standard data structure in which data is organized into dimensions and measures (or facts) that correlate to the business event.

2.3.2 Data Architecture:

Data architecture is created to define the mapping of the source data and its transformation into the data warehouse. It also addresses how data will be managed continuously and presented to the end user.

The following sections describes each aspect of the data architecture:

Data sourcing:

A data sourcing architecture describes the flow of data throughout a data warehouse. It reflects data origination (application databases) as well as eventual

data residence (data warehouse/Data Marts). It is comprised of three components:

- *Data flow* describes the data movement path from the point of extraction from source systems to the point of delivery into a specific data warehouse or Data Mart.

- A *data repository* is a central place where data is stored and maintained.

- *Metadata* is data about data. It documents the rules by which systems interoperate and provides descriptive information about data. Metadata can be categorized as the following:

    - Business metadata provides functional or business description of data elements to help users to locate, understand and access information in a data warehouse environment. It contains information on calculations used in the creation of the data element, graphs, or charts, along with time and date of creation.
    - Technical metadata provides technical descriptions of data along with the schedule to extract and move data to its destination. It contains information about the source and type of the data, destination, as well as rules used to extract, cleanse, and transform the data.

Data sourcing architecture standardization enhances consistency, content, timeliness, and minimizes administration and infrastructure problems within the data warehouse or Business Intelligence environment.

Data management:
The data management architecture describes the set of technologies and processes necessary to manage and maintain the following:

- Extraction of data
- Transformation of data
- Validation and integration of data
- Summarization of data

Data Lakes are often utilized for storage and processing of information as part of a data warehousing initiative. Data Lake contents may be used or accessed directly in the Data Lake or advanced into the data warehouse.

Data delivery and presentation:
Data delivery and presentation architectures describe technologies and processes. Data delivery and presentation architectures focus on data delivery processes and presentations to the end user. End users employ data delivery tools to gather data from the appropriate source(s), sometimes resulting in the creation of a Data Mart. End users, then utilize data presentation tools to view and analyze data.

Data Quality:
A data warehouse, being a decision-support information system, shall provide data that is not only accurate, but of high quality. There is a direct correlation between data quality and the effectiveness of IT and business operations that rely on this data. A high level of data quality is critical to the success of strategic business initiatives.

Data quality refers to more than finding and fixing missing or inaccurate data. It means delivering comprehensive, consistent, relevant, and timely information to the organization regardless of its application, use, or origin. An effective data quality initiative will encompass the following elements:

- Coherency – integrity constraints shall be observed. For example, the conversion of values to the same measurement unit is required to perform coherent computations.
- Completeness - the percentage of data found in a data store, with respect to the necessary amount of data that should be there.
- Timeliness – refers to how current the data is.
- Accuracy – often dependent on the data source and proper Aggregation of the data.
- Accessibility – the ability for the user to access the data warehouse from wherever and whenever the data is needed.
- Availability – refers to how long it takes for source data to get to the warehouse, and whether the warehouse captures the data sought by the user.
- Performance – refers to how difficult it is for the user to acquire the sought-after data, and how quickly they can acquire that data.

Data quality assurance measures are often applied during a data staging process where the data is tested for consistency, completeness, and fitness to publish to the user community.

Best practices suggest using a phased, iterative, ongoing process for improving data quality incorporating the following three steps:

1. Data Quality Assessment – Determine the current state of data quality. This will allow the development of a business case for the data quality initiative and provide a baseline from whence to begin. List all issues, prioritized by maximum impact on the business.
2. Data Quality Planning – The next step is to develop an incremental project plan to resolve existing issues and challenges as well as prevent future ones. Specify ways to ensure new applications incorporate data quality principles from the start.
3. Data Quality Strategy and Implementation – Selecting the best strategy requires balancing the cost of each data quality initiative against its impact.

Technology Architecture:

The technology architecture specifies and describes the necessary infrastructure to support the requirements necessary to deliver the information. The following are important to consider:

- Network architecture
- Hardware
- Software
- Cloud services

2.4 Determine Roles and Responsibilities:

This describes the organizational structures and processes necessary to manage and administer the data warehouse. Key roles to consider include business process owners, data owners, data warehouse project managers, database architects, business analysts, and support personnel.

2.5 Creating an Implementation Plan:

The implementation strategy for a data warehouse solution shall take into consideration the unique aspects of the organization's culture. The strategy and implementation plan shall focus on delivering all critical success factors. The plan shall include steps based on priority, along with a timeframe to revisit the strategy.

3. Other Best Practices

- 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 data warehouse on an iterative basis;
- Developing a scalable architecture to serve as the data warehouse's technical and application foundation and identifying and selecting the hardware, software, and middleware components to implement it;
- Defining the correct level of consolidation to support 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;
- Establishing governance processes for ensuring logging, auditing, security, and access controls associated with the data warehouse.

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
Revision	03/27/2023	• Minor grammatical and formatting edits • General non-policy information from the ITP was moved to section 3 of this document