

Client Logo

<Client Name>
Data Movement Best Practice
Topics to Consider

Version 0.1

<Date>

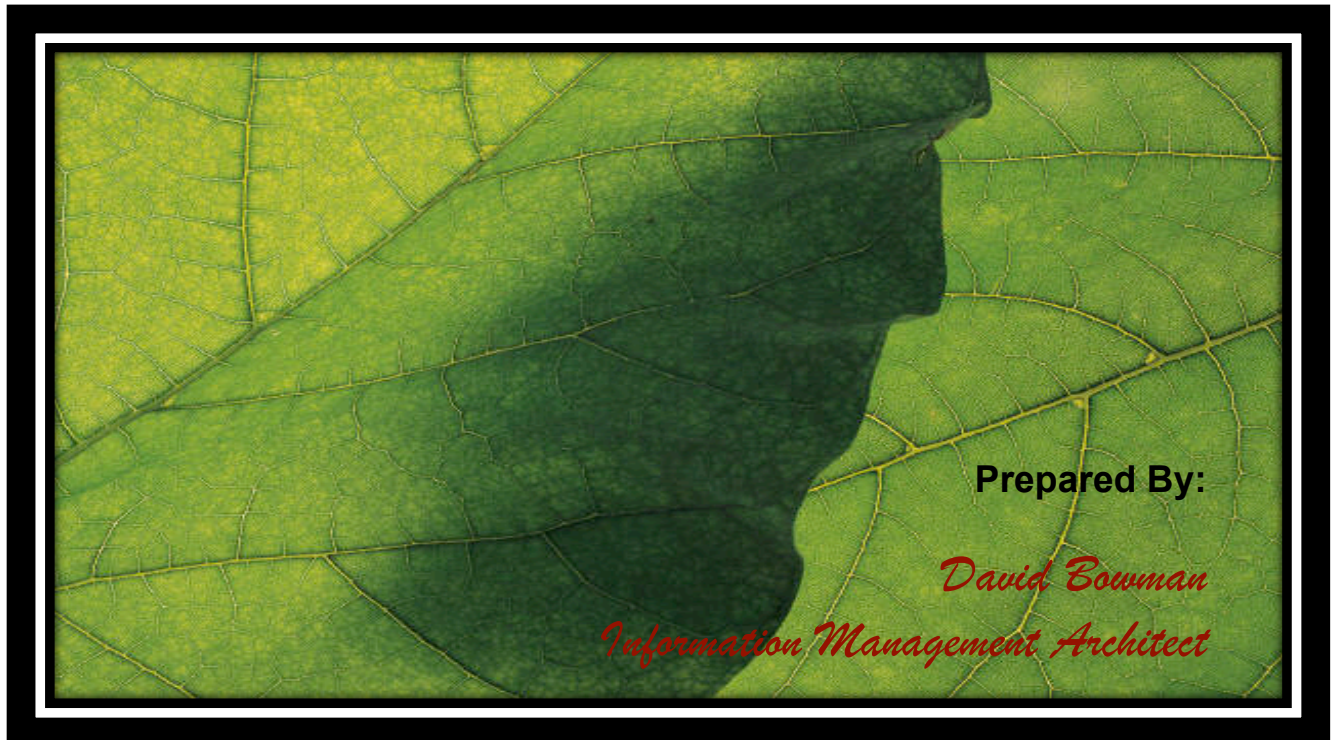


Table of Contents

- Revision History3
- Objective 4
- Architecture And Design Best Practices..... 4
 - Technical Design Qualities..... 4
 - Functional Qualities..... 4
 - Non-functional Qualities..... 4
- Constraints..... 4
 - Run-time Qualities, including..... 4
 - Development-time Qualities, including:..... 4
 - Separation of Concerns;.....5
 - Process Flow;5
 - Designing Transactions at an Atomic Level;.....5
 - Restart-ability;5
 - Error Handling Approach;.....5
 - Parallelization; and.....5
 - Job Scheduling5
- Build Phase Best practices5

Revision History

Version	Date	Description	Author

Objective

Data Movement best practices provide a guide for the analysis, design, and development of data movement processes that are consistent, usable and of high quality.

This document provides an outline of some technical best practice topics that should be covered in a best practices guide.

Architecture And Design Best Practices

Data Movement architecture and design best practices should be defined for each of the following:

Technical Design Qualities

Functional Qualities

Non-functional Qualities

Constraints

Run-time Qualities, including

Configurability and Supportability:

Correctness, Reliability, Availability:

Quality of Service:

Safety Properties:

Operational Scalability: and

Accuracy and Completeness.

Development-time Qualities, including:

Localizability:

Modifiability or Extensibility:

Scalability:

Reusability: and

Efficiency.

Separation of Concerns;

Process Flow;

Designing Transactions at an Atomic Level;

Restart-ability;

Error Handling Approach;

Parallelization; and

Job Scheduling

Build Phase Best Practices

Design Goals

- Best practices should be defined to meet the following architecture and design goals:
 - Move data from source once i.e. store data from source to the ETL environment quickly so that the source is only accessed once. The target architecture should enforce reuse of a single copy of data drawn from production sources. This will minimize resource utilization on the source system.
 - Standardize data quality: Provide a facility for standard centralized data quality checks. The facility should allow for required checks and optional checks (each target determines if they will subscribe to the optional checks). Systems of record are responsible for data quality.
 - Increase ETL Processing efficiency: Provide a capability (through metadata and process control flow) for tracking data dependencies

and driving data processing as soon as possible based on the availability of required input files.

- Maximize reuse: Provide a facility for storing and cataloging clean data (both source and derived from calculations and/or joins). By storing the clean data, it can be utilized by new processes. This will minimize the need for new processes to re-source the data recheck quality or re-compute derived values.
- Focus on Metadata: Use of a centralized metadata repository to drive data quality and integrity, and ETL workflow.

Components

- Best practices should be developed for the following components:
 - **Staging Area Components** – provide facility for storage and retrieval of interim/staged ETL data sets as flat files.
 - **Filtering Components** – provide a mechanism for creating target specific outputs by eliminating unwanted records and fields.
 - **Computation Components** – provide a mechanism for reformatting and calculation, creating new derived values.
 - **Data Movement Components** – provide mechanisms for moving data from sources to the ETL environment and from the ETL environment to targets.
- Best practices should be developed for the following components:
 - **Data Set Level Validation** – verifies the integrity of copied data sets (e.g., checksums).
 - **Data Quality** – a standard and centralized method for testing and validating data at the row level.
 - **Workflow Management** – a mechanism for moving data through ETL components as quickly as is feasible, allowing for jobs to start as soon as all inputs are available.
 - **Metadata** – technical metadata to support these processes.
 - **Other Services** – Data Security, Disaster Recovery, Notifications.
- Server Configurations
- Guidelines for Data Security in Production Access Lockdown Environment
- Environment Variables Naming Conventions
- Technical Documentation
- Construction Error Handling
- File Transfer Protocol (FTP) Guidelines Overview

- Continuous Data Movement
- Coding Standards
- Graph Standards
- Component Standards
- File Naming Standards
- Code Reuse Standards
- Data Movement Testing Processes