| | |
|---|---|
| Standard ID: | **IN-DAT-001** |
| Title: | **Data Integration** |
| Domain: | **Information** |
| Discipline: | **Integration** |
| Revision Date:<br>Revision no.:<br>Original date: | **5/15/2013**<br>**1**<br>**11/10/2011** |

## I. Authority, Applicability and Purpose

A. **Authority –** Title 29, Chapter 90C provides broad statutory authority to the Department of Technology and Information to implement statewide and interagency technology solutions, policy, standards and guidelines for the State of Delaware's technology infrastructure. "Technology" means computing and telecommunications systems, their supporting infrastructure and interconnectivity used to acquire, transport, process, analyze, store and disseminate information or data electronically. The term "technology" includes systems and equipment associated with e-government and Internet initiatives.

B. **Applicability –** Applies to all State of Delaware communications and computing resources. DTI is an Executive Branch Agency and has no authority over the customers in Legislative and Judicial Branches, as well as School Districts, and other Federal and Local Government entities that use these resources. However, all users, including these entities, must agree to abide by all policies, standards promulgated by DTI as a condition of funding, access and continued use of these resources.

C. **Purpose --** Due to the importance of the information managed by the State’s technology solutions, it is necessary to establish common guidelines for integrating / processing the data. This document provides approaches and best practices for data integration. This document will address data consolidation and data propagation (near real time only).

## II. Scope

A. **State of Delaware** – All communications and computing resources involving data owned by the State of Delaware

B. **Areas Covered –** This standard covers all aspects of integration where data is owned by the State of Delaware.

C. **Environments –** This standard applies to data integration between the State of Delaware and any entity outside of the State of Delaware. Also, it addresses data integration between and within State organizations.

## III. Process

**A. Adoption** – These standards have been adopted by the Department of Technology and Information (DTI) through the Technology and Architecture Standards Committee (TASC) and are applicable to all Information Technology use throughout the state of Delaware.

**B. Revision** – Technology is constantly evolving; therefore the standards will need to be regularly reviewed. It is the intent of the TASC to review this standard annually. The TASC is open to suggestions and comments from knowledgeable individuals within the state, although we ask that they be channeled through your Information Resource Manager (IRM).

**C. Contractors** – Contractors or other third parties are required to comply with these standards when proposing technology solutions to DTI or other state entities. Failure to do so could result in rejection by the Delaware Technology Investment Council. For further guidance, or to seek review of a component that is not rated below, contact the TASC at dti_tasc@state.de.us.

**D. Implementation responsibility –** DTI and/or the organization's technical staff will implement this standard during the course of normal business activities, including business case review, architectural review, project execution and the design, development, or support of systems.

**E. Enforcement –** DTI will enforce this standard during the course of normal business activities, including business case and architectural review of proposed projects and during the design, development, or support of systems. This standard may also be enforced by others during the course of their normal business activities, including audits and design reviews.

**F. Contact us** – Any questions or comments should be directed to dti_tasc@state.de.us.

## IV. Definitions/Declarations

### A. Definitions

1. **Change Data Capture** -- It identifies changed records from a source dataset. Only the changed records are moved to the target reducing volume. The change data capture can be done by comparing the values to the prior set from the source, reading the database logs, or using a value in the dataset to identify that a record has changed.

2. **Data Integration** – Data integration involves combining data residing in different sources and provides users with a unified view of this data. Business processing requirements that can influence specific considerations include:

   - Volume
     - 1 record
     - 1000 records
     - 1mm records
   - Latency
     - Real-time
     - Near real-time
     - Hourly

  - Daily
  - Weekly
  - Monthly
  - Quarterly
  - Annually
  - As needed
- Transformation
  - None
  - Minimal
  - Simple
  - Complex
  - Many
- Lookups
- Audit requirements
- Logging requirements
- Restart requirements
- Guaranteed delivery requirements

Data integration can also be used to move data from within a location. For example, when dealing with decayed or decaying data (data that was once correct but through the passage of time has ceased to be accurate), data integration can be used to improve the quality of the data by moving it to another location within the same database.

Data Integration can be leveraged in a number of traditional techniques: ETL (Extract, Transform, Load), EAI (Enterprise Application Integration), and EII (Enterprise Information Integration). Other techniques can also be built or used such as ELT (Extract, Load, Transform). The latency can create significant differences in deciding which technique can be or should be considered (e.g., real time vs. monthly load).

3. **Data Integration Techniques** -- There are three main techniques used for integrating data: consolidation, federation, and propagation.

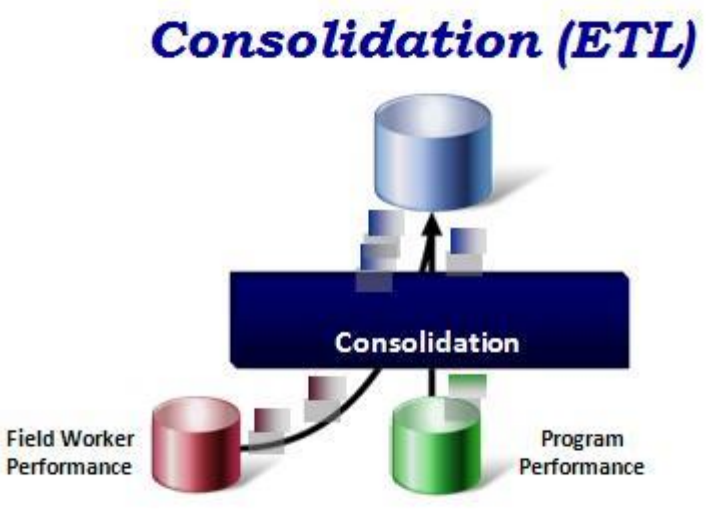
**Consolidation** is physically combining and transforming data from one or more sources to create a single view and this technique may be used:

- If there can be delay or latency between the source and target systems.
- To capture data from multiple source systems and integrate to single system
- When handling large volumes of data
- For populating enterprise data warehouses and data marts (this is a widely used approach)
- With both Pull and Push technologies.
- Technologies used for Consolidation are ETL and ECM (Enterprise Content Management)
    - ETL is the most common technology used for data warehousing
    - ECM is used for consolidating and managing unstructured data (documents, reports and web pages)

**Drawbacks for Consolidation**

- It needs more computing resources to process massive volumes of data
- Possibly requires large amounts of hard disk space for the temporary storage of staging files used in the process.

An example of consolidation can be found in Appendix A

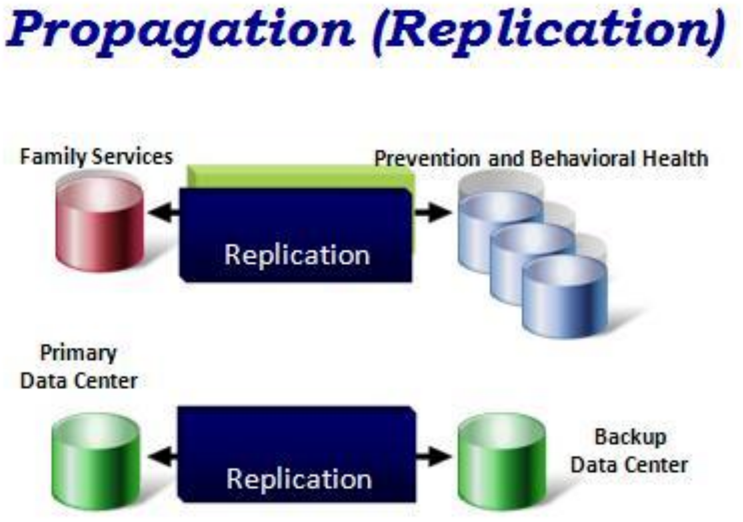


**Propagation** is also referred to as replication when data is copied from one location to another. It is typically event-driven (push technology). Propagation (properly implemented and managed) may :

- Be used to share real time information access among systems.
- Streamline business processes and help raise organizational efficiency
- Maintain information integrity across multiple systems

**Drawbacks for Propagation**

- High Initial development costs, especially for small and mid-sized firms
- Requires a large amount of time for business design and oversight
- Ease of development and maintenance.



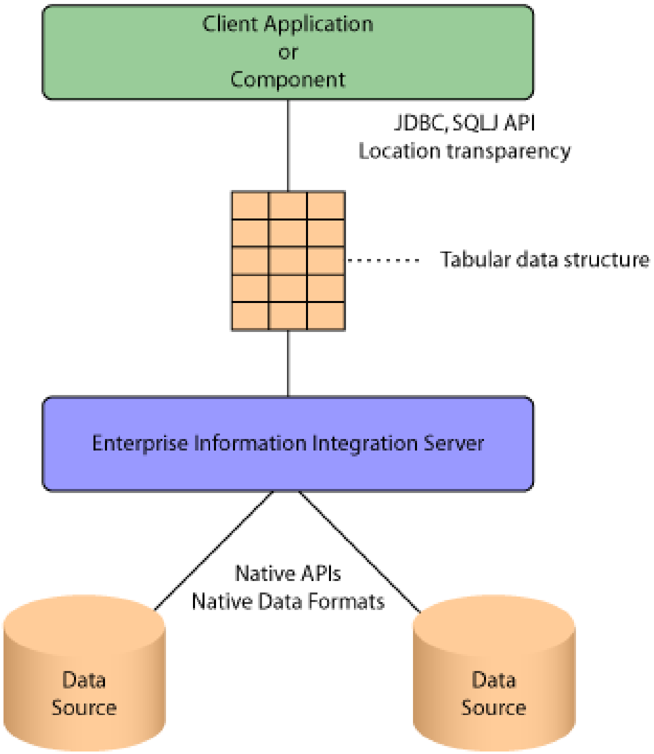
An example of propagation can be found in Appendix B

**Federation**, also known as Enterprise Information Integration (EII), is a process of integrating information into a single federated database in real time and to provide a unified view of data. It:

- Provides access to the current data without a need for consolidation.
- Always uses a pull technology (On-Demand).
- When there can be no delay or zero latency between the source and target systems.
- When cost of data consolidation outweighs business benefits.
- Security or License problems exist in replicating / copying the data.
- EII is an example of federation approach.

### Drawbacks for Federation

- Not suited for retrieving and reconciling large volume of data
- Not suited when the data needs to be cleansed and transformed.
- Performance impact on the source systems.
- Overhead of accessing multiple source systems at run time



4. **Data Integration Patterns** – There are various deployment patterns for data integration. For more details, reference appendix C.

   - Source-to-Target (Same Vendor (Oracle – Oracle) No Transformations)
   - Source-to-Target Different Vendor (Oracle – SQL Server)
   - Source-to-Target databases from same vendor – Near real time
   - Source-to-Target databases from different vendor – Near real time
   - Source-to-Target databases with medium – complex transformations & Lookups

5. **Data Integration Components** -- Data Integration consists of multiple processes and staging zones. Each of these components is the "building block" for an extensible and leverage-able data integration environment. The use of these components will depend on the business and functional requirements. They include the following:

   - Extract/Subscribe Processes – Extract/Subscribe is the set of processes that capture data, transactional or bulk, structured or unstructured, from various sources and lands it in an Initial Staging Zone.
   - Initial Staging Zone – Initial Staging is an optional "landing zone" where the copy of the data from the sources is landed as a result of the extract/subscribe processing. One of the purposes for the Initial Staging Zone is to persist source data in non-volatile storage to achieve the "pull it once from source" goal. Data from real time sources that is intended only for real time targets is not passed through Extract/Subscribe and may not land in an Initial Staging Zone.
   - Data Quality Processes – Data Quality Processes are the processes that qualifies and cleanses the data, based upon technical and business process rules. Regardless of the data quality rules the Data Quality Process should provide the following functionality:
     - Data Cleansing: using the data quality criteria
     - Data Rejection: The reason for failure must be documented
   - Clean Staging Zone – The Clean Staging Zone contains records that have passed all data quality checks. This data may be passed to processes that builds load ready files or used in transformation processes that produce new data sets.
   - Transformation Processes – There are a several types of transformations which include:
     - Calculations – Mathematical processes are applied to the data to derive new data elements.
     - Split - The architecture supports splitting data sets. Splitting is a technique used to divide a data set into subsets of fields that are then stored individually. Vertical filtering is used to pass only the

data elements the target needs and horizontal filtering passes only the records that conform to the target's rules.

- Joins - combining fields from multiple sources and storing the combined set
- Lookups - combining fields from records with values from reference tables and storing the combined set
- Aggregations - creation of new data sets derived from the combination  of multiple sources and/or records

- Load-Ready Publish Staging Zone – Load-ready Staging Zone is utilized to store target-specific load-ready files.
- Load/Publish Processes – Load/Publish is a set of standardized processes for executing loads. There are five types of loads:
  - SFTP to Target:  In this type of load, the process is only responsible for depositing the output to the target environment.
  - Piped data:  This process executes a load routine on the target that takes the data directly piped from the Target Specific Filter.
  - RDBMS Utilities:  Bulk Load on the target, but the source is Load-Ready Staging Area.
  - SQL: SQL writes directly to the target database.
  - Message Publishing: For loading real time data feeds to message queues

5. **Data Integrity Techniques** – Below are some of the defined techniques to help ensure the integrity of data.
   - Source systems – Ensure the final / target data file match the number of records retrieved from the source tables when appropriate
   - Source systems – On creation of the data file, create a manifest file which will contain the number of records created and size of the data file. This information will help the target to ensure the data movement and loads worked fine.
   - When a file is SFTP'd check the file sizes between the source and target.
   - Ensure the target file size match with the manifest file before loading the file to the target tables (Staging) or any activity.
   - Ensure the total records loaded to Staging match the records mentioned in the manifest file. (Manifest file can also contains sum of measures in the facts (Ex) Sum of Sales Amount)
   - Create Quality Control checks in the packages to ensure the records are moved consistently across. Keep track of records dropped and reasons for the drop in a table or a log file so, analysts / operations can use the logs to understand the issue and probably fix the records and re-load them (based on the requirements).

6. **Extract Transform and Load (ETL)** – It is a process used especially in data warehousing that involves: Extracting data from source systems and transforming it to fit operational needs (includes data quality) and loading it to the end target (database or data warehouse or a data file)

7. **Manifest File** -- File created by the source system for Audit and Quality check purposes. This file will be used by the downstream systems to ensure data is moved correctly.

8. **Restartability** -- This is the ability to restart a process that moves data from start to end. Start-to-end is defined as moving all necessary data from the original source to the loading of the final target. The process must be able to be restarted at any point where a failure may occur.

9. **Reusability** -- Reusability means that a dataset that needs to be shared with more than one target environment can be done so without a specifically written solution for each of the target environments. Other guides towards reusability include:

   - Use late binding techniques to provide metadata at run time – this means that file names are not hard coded
   - Repeated algorithms (transformations, cleansing, standardization, etc.) on common attributes are made available through subroutines.
   - Designs are decomposed to enhance reuse opportunities
   - Leveraging techniques to align and harmonize data can be used to improve reuse opportunities. Alignment is where data is first made consistent within its processing area and harmonization is where data is then made consistent across processing areas.

## B. Declarations

**Data Integration implementations must**

- Ensure the target datastore enforces at least the same level of data security, data privacy, etc as the source datastore
- Have documented permission to use the data
- Ensure that only the required data is sent to the target datastore
- Ensure that quality data is integrated
- Provide an audit trail at the appropriate level
- Ensure data integrity especially during the movement from the source datastore to the target datastore.
- Ensure that error and operational logging is in place
- Be designed with reusability
- Utilize a centralized repository of metadata
- Implement a consistent and repeatable methodology for mitigating risks
- Minimize the impact to operational datastores
- Leverage existing metadata and data models from source and target datastores.
- Document the data integration and follow establish naming conventions.

**Data Integration implementations may**

- Utilize a Probabilistic Matching Engine for data standardization jobs
- Implement a Parallel Processing Engine for scalability
- Implement a shared GUI Interface between major components of the platform
- Provide SOA Capabilities
- Handle Change Data Capture
- Have Connectivity/Adapter capabilities for current vendors for the ability to interact with:
    - Relational and Non relational Databases
    - Packaged applications
    - SaaS and cloud base applications
    - Message Queues
    - Industry standard message formats

- Receive and create various file formats including XML
- Have a real time data and event capture
- Provide monitoring and debugging
- Support restartability
- Divide the load activities into smaller jobs in order to achieve faster load times. Likewise, large load files may be divided into several smaller load files to achieve faster load times.
- Minimize or eliminate any changes to the data when accessing an operational datastore.
- Utilize landing and staging tables to achieve business and functional requirements.
- Address error handling in the following ways
  - All errors from a data integration component may be written to an error file.
  - Log and error ports of database components may be connected to data files.
  - Log components may be used to gather information from the log port of any component.
  - All component error thresholds may be ‘abort on first reject’. The use of ‘never abort’ as the error threshold is very dangerous and hence is strongly discouraged.
  - Use flow record counts to perform data consistency and data integrity checks. Data integration summary files may be used for getting record counts of data files or data flows.
  - All error-resolution history may be stored for analysis and reference.
  - Use a centralized error checking procedure / function that evaluates the return codes within an application.

## V. Definition of Ratings

| Individual components within a Standard will be rated in one of the following categories.<br>**COMPONENT RATING** | **USAGE NOTES** |
|---|---|
| • **STANDARD** – DTI offers internal support and has arranged for external vendor support as well (where applicable). DTI believes the component is robust and can be expected to enjoy a useful life of 5+ years from the Effective Date. | These components can be used without explicit DTI approval for both **new projects** and **enhancement** of existing systems.<br>*[1] Note the useful life concern for the "Acceptable" rating.* |
| • **ACCEPTABLE** – DTI offers internal support and has arranged for external vendor support as well (where applicable). DTI believes the component is stable, but has a useful life [1] of less than 5 years from the Effective Date. | |
| • **EMERGING** – DTI considers the component to be a likely candidate for future classification as STANDARD or ACCEPTABLE within the state pending further investigation. | These components must be explicitly approved by DTI for **all projects**. They must not be used for **minor enhancement** and **system maintenance** without explicit DTI approval. |
| • **DECLINING** – Deprecated – DTI considers the component to be a likely candidate to have support discontinued in the near future. A deprecated element is one becoming invalid or obsolete. | |
| • **LIMITED SUPPORT** – DTI has limited or no internal support capability for the component; or has no arrangement for vendor support for the product. Users must arrange for adequate overall support of the component through their own efforts. | |
| • **NOT SUPPORTED BY DTI** – DTI offers no internal support and has no arrangement for vendor support. Users must arrange for all support of the component through their own efforts. | |
| • **DISCONTINUE** – For reasons of overall risk, product support, high TCO, or other issues, the use of this technology is discouraged. All current instances of this technology should have a plan developed for its retirement. DTI expects to work aggressively with the users of such technologies to devise a collaborative plan. | No waiver requests for new solutions with this component rating will be considered. |
| • **DISALLOWED** – DTI declares the component to be unacceptable for use and will actively intervene to disallow its use when discovered. | No waiver requests for new solutions with this component rating will be considered. |

A. **Applicability of Ratings** – The ratings and usage notes are intended to encourage technology decisions to move toward components that enjoy the full support of DTI. However, acknowledging that mass replacement of lower rated components is not feasible, DTI will allow continued maintenance, enhancement, and possibly limited new development using these components. In making such determinations, DTI may require that the requestor demonstrate that they have adequate support arrangements in place.

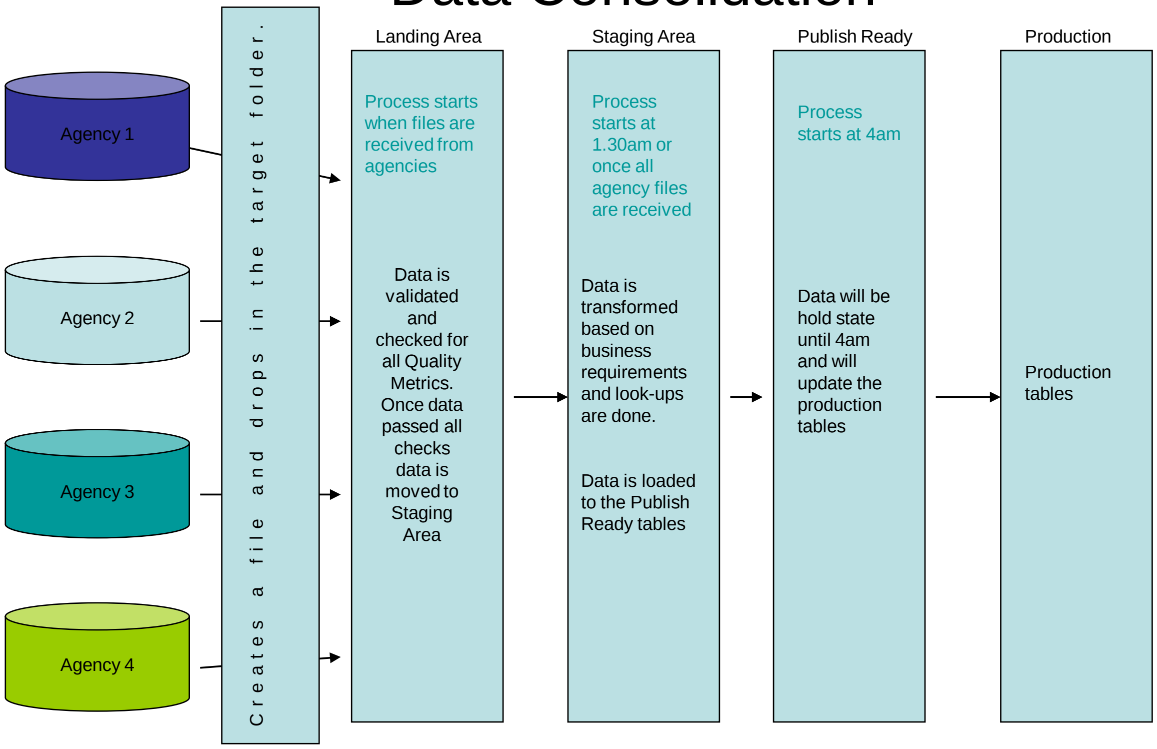
B. **Missing Components –** No conclusions should be inferred if a specific component is not listed. Instead, contact the TASC to obtain further information.

## VI. Component Assessments

| # | Component | Rating | Comments |
|---|---|---|---|
| **1** | **Data Integration Middleware** | | |
| a) | IBM DataStage | Standard | General Release Levels |
| b) | DPSync | Limited Support | |
| c) | Natquery | Limited Support | Used with DataStage |
| d) | SQL Server Integration Services (SSIS) | Limited Support | |
| e) | iWay | Limited Support | |

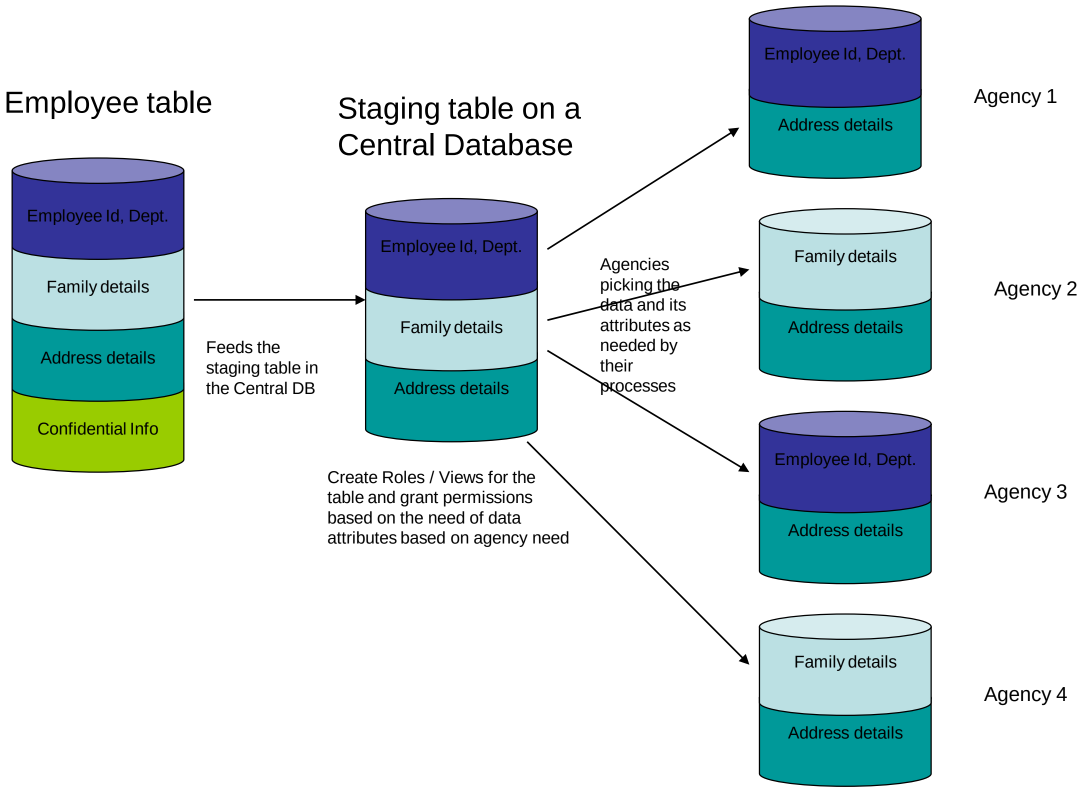
# Appendix A

## Data Consolidation


Agency 1
Agency 2
Agency 3
Agency 4
Creates a file and drops in the target folder.
Landing Area
Process starts when files are received from agencies
Data is validated and checked for all Quality Metrics. Once data passed all checks data is moved to Staging Area
Staging Area
Process starts at 1.30am or once all agency files are received
Data is transformed based on business requirements and look-ups are done.
Data is loaded to the Publish Ready tables
Publish Ready
Process starts at 4am
Data will be hold state until 4am and will update the production tables
Production
Production tables

**Requirement**: Agencies should send the files between 10pm – 1am. Process should start immediately after receiving the files and should be validated and QC. The transformation should be started after receipt of all agency files or at 1.30am. The data for publish should be ready by 3.30am and should not be published until 4am

# Appendix B - Propagation



# Appendix C – Data Integration Patterns

**Pattern 1: (**Source and Target databases from same vendor)

- Backup the Source database
- Transfer the Source backup to the target server.
- Ensure target server has enough disk space before restoring the backup.
- Backup the target database if requirements indicate.
- Restore the Source backup.
- Check to ensure backup restored correctly.
- Ensure all the user permissions are intact in the target database as needed by the downstream applications.

**Pattern 2: (**Source and Target databases from different vendor)

- Export data from Source tables to flat files (Suggested target data file format is Delimited file).
- Create a Manifest file which contains at a minimum Table name and total records.
  - For Facts create sums for the measure columns (ex. Sum of Sales / Purchase Amount)
- Ensure Target database has enough space allocated.
- Disable the constraints (Only disable if it does impose any RI problems)
- Load the data from flat files to the tables (Bulk Load)
- Ensure the totals records in the target match the manifest.
- Enable constraints (Only if disabled)
- Ensure all the target database users and their roles are in place.

**Pattern 3: (**Source-to-Target databases from same vendor – Near real time)

Following methods could be used:

- Snapshots
- Log shipping
- ETL by using Change Data Capture (CDC)
  - Extract records which changed since last extract from source table to Data file (Delimited - Preferred)

- Create Manifest file.
- Transfer the files to target server.
- Ensure the files transferred successfully
- Load Data file to Target tables
- Validate the Data loaded with Manifest file.
- Publish

**Pattern 4: (**Source-to-Target databases from different vendor – Near real time)

- ETL by using Change Data Capture (CDC)
  - Follow same steps as defined in the Scenario 3 ETL by using Change Data Capture (CDC)

**Note**: Ensure the data types are compatibility between the vendor databases/systems or else a transformation is required (ex Dates and Numeric).

**Pattern 5: (**Source-to-Target databases with medium – complex transformations & Lookups)

- Create Source data files from one or more systems or tables
- Create a Manifest file for each source data file created.
- Copy the source files to target server
- Ensure the files have been transmitted correctly
- Load the source files to their corresponding Landing tables
- Validate the records loaded and sums using the manifest file
- Perform transformations and lookups based on the business and functional requirements and insert into the staging / holding tables.
- Validate the tables to ensure all the source records have been processed. Source Records = Staging Records + Dropped records
- Publish the records to Production table / target data file etc.
- Perform QC check to ensure data is matched
- Backup the Source files (If this is a requirement for Audit purposes)
- Send email or notify the support personnel once job is completed / failed (If this is a part of a requirement)