Data Integration and Interoperability

1 Introduction

DII describes processes relating to the movement and consolidation of data within and between data stores, applications and organisations. **Data Integration** consolidates data into consistent forms. **Data Interoperability** is the ability for multiple systems to communicate.

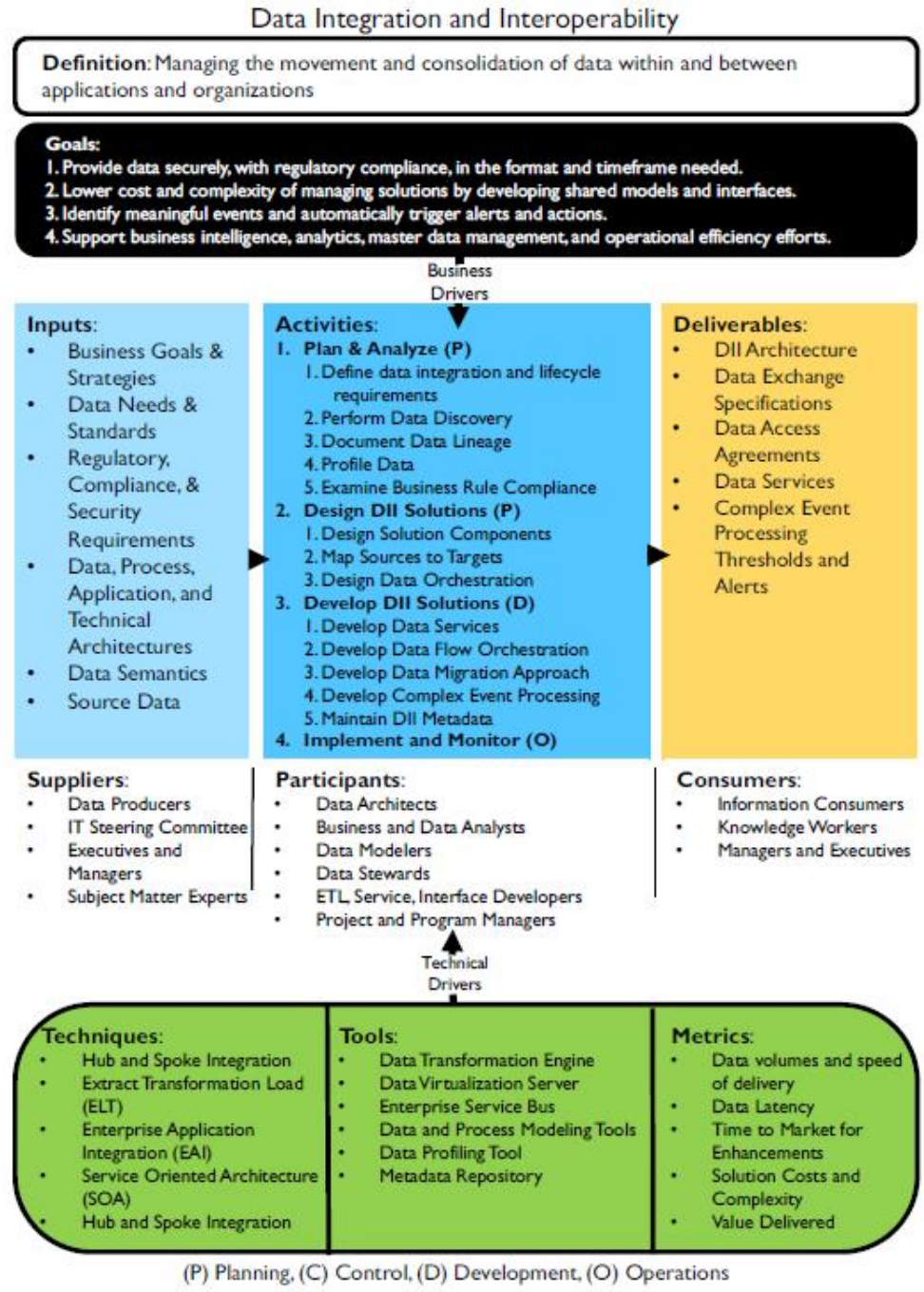
DII provides:

- Data migration and conversion
- Data consolidation into hubs or marts
- Integration of vendor packages into organisation
- Data sharing across applications or organisations
- Distributing data across data stores and data centres
- Archiving data
- Managing data interfaces
- Obtaining and ingesting external data
- Integrating structured and unstructured data
- Providing operational intelligence and management decision support

DII is dependent on these other areas of data management:

- **Data Governance:** The transformation rules and message structures
- **Data Architecture**: Designing solutions
- **Data Security**: ensuring solutions protect the security of data
- **Metadata:**
    - Tracking technical inventory of data (persistent, virtual and in motion)
    - Business meaning of data
    - Business rules for transforming data
    - Operational history and lineage of the data
- **Data Storage and Operations:** Managing the physical instantiation of the solutions
- **Data Modelling and Design:** Designing the physical and virtual data structures, and messages passing information between applications and organisations

DII is critically important **to Data Warehousing and Business Intelligence** as well as **Reference Data and Master Data Management**.



1.1 Business Drivers

- The need to manage data movement efficiently
- Purchased applications come with its own data stores that must integrate with the other data store in the organisation.
- An enterprise view of data integration is more cost effective than point to point solutions
- Data hubs such as data warehouses and Master Data solutions
- Managing the cost of support by using standard tools and reducing the complexity of interface management
- DII supports the organisation's ability to comply with data handling standards and regulations.

1.2 Goals and Principles

Goals:
1. Provide data securely, with regulatory compliance, in the format and timeframe needed.
2. Lower cost and complexity of managing solutions by developing shared models and interfaces.
3. Identify meaningful events and automatically trigger alerts and actions.
4. Support business intelligence, analytics, master data management, and operational efficiency efforts.

When implementing DII follow these principles:

- Design should take an enterprise perspective (for future extensibility), but implement iteratively and incrementally
- Balance local data needs with enterprise data needs, including support and maintenance
- Ensure business accountability for DII design and activity.

1.3 Essential concepts

1.3.1 Extract, Transform and Load (ETL)

The essential steps in moving data around

- **Extract:** Select required data and extract it from its source, and stage it physically or in memory
- **Transform:** Make the data compatible with the structure of the target store. May be done in batch or real-time:
  - **Format changes:** Technical format e.g. EBCDIC to ASCII
  - **Structure changes:** e.g. denormalised to normalised
  - **Semantic conversion:** Conversion of values to maintain consistent semantic representation
  - **De-duping:** If rules require unique values, scan target and remove duplicate rows
  - **Re-ordering:** Change to order of the file data elements to fit a pattern
- **Load:** Physically store the result of the transformation in the target system

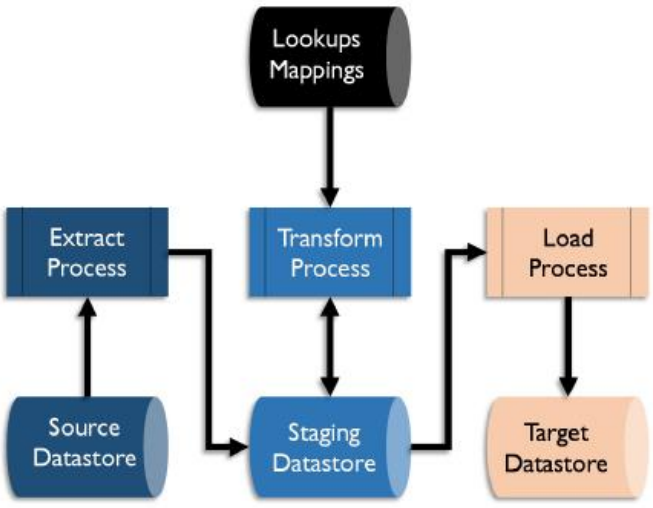


Figure 67 ETL Process Flow

- **ELT (Extract, Load and Transform):** Used if the target system has more transformation capability. also allows data to be instantiated on the target as raw data.

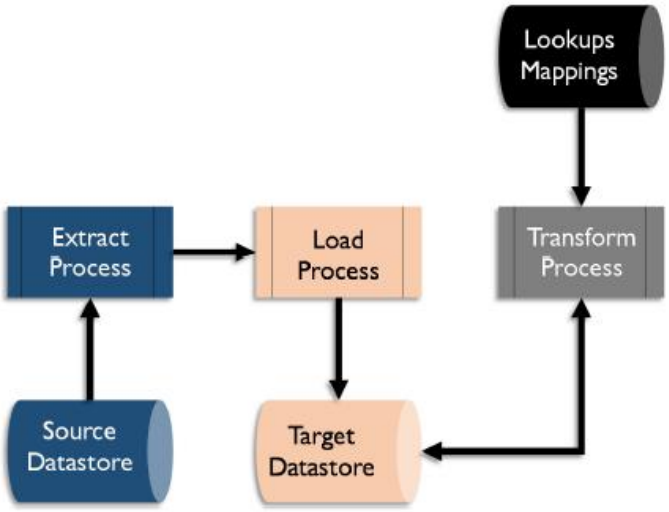


Figure 68 ELT Process Flow

- **Mapping:** A synonym for transformation, the process of developing a lookup matrix from source to target structures, and the result of that process.

1.3.2 Latency

The time difference between when data is generated in the source system and when it is available in the target system. Can be high (batch), low (event driven) to very low (real-time synchronous).

- **Batch:** Data moving in clumps or files periodically. Called a **batch** or **ETL**.
  - **Snapshot:** Full set of data at a point in time
  - **Delta:** Data that has changed values since the last time data was sent.
  - High latency
  - Used for data conversions, migrations, archiving and extracting from and loading data warehouses and marts
- **Change data Capture:** Filtering data to include only the data that has changed in a given timeframe (the delta) and passes it on to data consumers.
- **Near-real-time and Event Driven:** Data is processed in smaller sets spread across the day, or when and event such as an update occurs. Lower latency than batch.
- **Asynchronous:** The source does not wait for the target to acknowledge before continuing processing. The target need not be available.
- **Real-time, Synchronous:** No time delay or other differences between source and target are acceptable. The executing process waits for confirmation before executing its next transaction.
- **Low Latency or Streaming:** Extra hardware costs. Need extremely fast transfer of large amounts of data over large distances.

1.3.3 Replication

Applications maintain exact copies of data sets on multiple physical locations. Better response times for international users. Use DBMS replication utilities. Not recommended if changes to the data occur at more than one site.

1.3.4 Archiving

ETL functions can transport and possibly transform infrequently used data to a cheaper storage solution.

1.3.5 Enterprise Message Format/ Canonical Model

A canonical model is the common model used by an organisation to standardise the format in which data will be shared. Transformations need only be done to and from the canonical model.

- Hub-and-spoke
- All systems interact with central information hub
- Data is transformed based on the enterprise message format of the organisation
- Reduces transformations as each system only needs to transform data to and from the central canonical model
- Reduces complexity of DII in the enterprise
- Lowers cost of support
- Complex to develop
- Justified in managing more than 3 systems and critical for more than 100

1.3.6 Interaction Models

Describe ways to make connections between systems:

- **Point-to-point:**
  - Systems pass data directly to each other.
  - Suitable in small systems.
  - Can impact processing
  - Many interfaces to manage
  - Multiple interfaces can lead to inconsistent data
- **Hub-and-spoke:**
  - Consolidates shared data in a central hub.
  - Examples are Data Warehouses, Data Marts. Operational Data Stores and Master Data Management Hubs.
  - Easy to add more systems.
  - Enterprise Service Buses (ESB) – near real-time sharing of data where the hub is a virtual canonical model.
- **Publish-subscribe:** Systems push out (publish) data and other systems pull data in (subscribe).

1.3.7 DII Architecture Concepts

- **Application coupling:**
  - **Tight coupling**: Synchronous interface, one waits for the other to respond
  - **Loose coupling**: Preferred interface design as data is passed without waiting for a response and without causing both systems to be unavailable if one is unavailable. e.g. Service Oriented Architecture using an Enterprise Service Bus

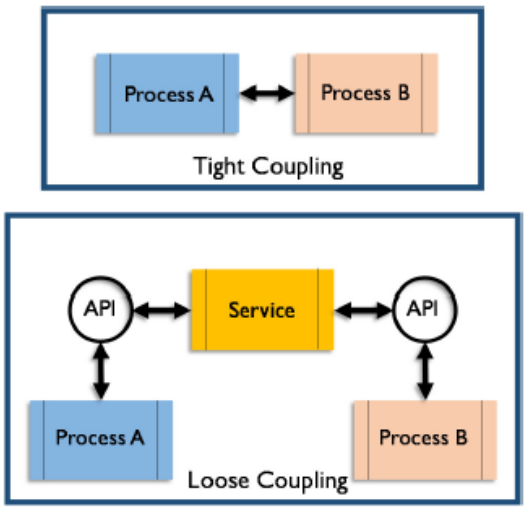


Figure 69 Application Coupling

- **Orchestration and Process Controls:**
  - **Orchestration:** How multiple processes are organised and executed in a system. All systems handling messages must be able to manage the order of those processes.
  - **Process Controls:** The components that ensure shipment, delivery, extraction and loading of data is complete. Include:
    - Database activity logs
    - Batch job logs
    - Alerts
    - Exception logs
    - Job dependence charts with remediation options, standard responses
    - Job clock information – length of jobs and computing window time.
- **Enterprise Application Integration (EAI):** Software modules only interact with each other through APIs.
- **Enterprise Service Bus (ESB):** An intermediary system, passing messages between other systems.

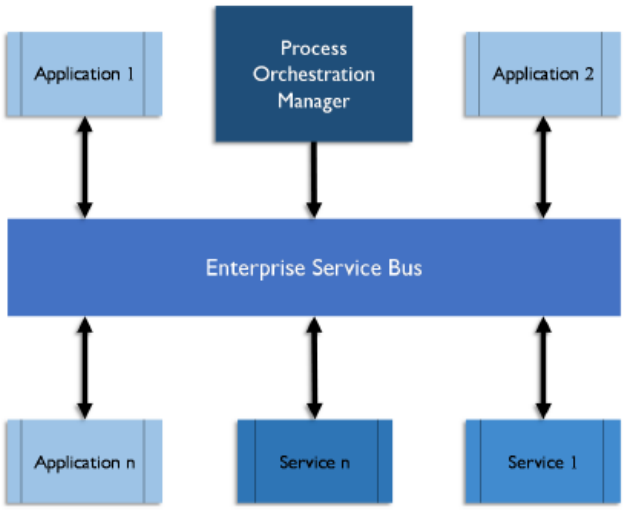


Figure 70 Enterprise Service Bus

- **Service-Oriented Architecture (SOA):** The functionality of providing data or updating data is provided through well-defined service calls between applications enabling application independence. Usually implemented through APIs.
- **Complex Event Processing (CEP):** Tracks and analyses data about events from multiple sources to predict meaningful events (threats, opportunities) to predict behaviour or automatically trigger a response.
- **Data Federation and Virtualisation:** Provides access to a combination of individual data stores. Virtualisation enables distributed databases to be accessed as a single database.
- **Data-as-a-Service (DaaS):** Data licensed from a vendor and provided on demand.
- **Cloud-based Integration:** Also called Integration Platform-as-a-service (IPaaS) and are usually run as DaaS applications at the data centres of vendors

1.3.8 Data Exchange Standards

Data Exchange Standards are formal rules for the structure of data elements. ISO or any common model used by an organisation or data exchange group. It simplifies data interoperability, lowers support costs and increases understanding of the data.

2 Data Integration Activities

Data Integration activities follow a development lifecycle (plan, design, development, testing and implementation)

2.1 Plan and Analyse

- Define data integration and lifecycle requirements:
  - Data and technology required to meet business objectives
  - Laws and restrictions on data contents
  - Defined by Business Analysts, data stewards and various architects
  - Creates and uncovers valuable Metadata – manage throughout lifecycle
- Perform Data Discovery:
  - Identify potential sources of data for the integration effort
  - High level assessment of data quality
  - Maintain the inventory of organisational data in a Metadata repository
  - Plan for acquiring and integrating external data
- Document Data Lineage:
  - How data under analysis is acquired or created by the organisation, where it moves, how it is changed, how it is used by the organisation for analytics, decision making or triggering.
- Profile Data:
  - Data profiles reveals actual data structure and contents
  - Assess the quality of the data
  - Can reveal differences from what is assumed leading to early intervention
  - Basic profiling involves analysis of:
    - Data format (data structures and inferred from actual data)
    - Data population (NULLS, blanks and defaults)
    - How data values correspond to a set of valid values
    - Patterns and relationships internal to the data set
    - Relationships to other data sets
  - The requirement to profile data must be balanced with an organisation's privacy and security rules (see Chapter 13)

- Collect Business Rules: rules harvesting or mining
    - A Business Rule is a statement that defines or constrains and aspect of business processing.
    - Four categories:
        - Definition of business terms
        - facts relating terms to each other
        - Constraints or action assertions
        - Derivations
    - Use business rules to support DII to:
        - Assess data in potential source and target data sets
        - Direct the flow of data in the organisation
        - Direct when to automatically trigger events and alerts

2.2 Design Data Integration Solutions

2.2.1 Design Data Integration Architecture

DI solutions should be specified at both enterprise and individual solution level. Enterprise standards save time as planning has been done and group licences and resource sharing and reusing DII components saves money.

- **Solution architecture indicates**:
    - Techniques and technologies to be used
    - Inventory of involved data structures
    - Orchestration and frequency of data flow
    - Regulatory and security concerns and remediation
    - Operating concerns around backup and recovery, archiving and retention
- **Select Interaction model:** hub-and-spoke, point-to-point or publish-subscribe
- **Design data services or exchange patterns:** Start with industry standards or existing patterns.

2.2.2 Model Data Hubs, Interfaces, Messages and Data Services

Model persistent and transient datatypes.

2.2.3 Map Data Sources to Targets

Specify the rules for transforming data from one location and format to another. For each attribute mapped specify:

- Technical format of source and target
- Transformations required for intermediate staging points
- How each attribute in the final or intermediate data store will be populated
- Whether data values need to be transformed
- calculations required

2.2.4 Design Data Orchestration

Pattern of data flows from start to finish, including intermediate steps.

- **Batch**: Frequency of data movement and transformation is usually coded into a scheduler
- **Real-time**: Usually triggered by an event such as an update.

2.3 Develop Data Integration Solutions

- **Develop Data Services:** Can be tools or vendor suites

- **Develop data flows:**
  - ETL flows developed in a tool such as a scheduler for batch which manages the order, frequency and dependency of executing the data pieces
  - Real-time: Monitor for events that trigger services to acquire, transform or publish data
- **Develop data migration approach:** Moving data needs proper analysis and testing
- **Develop a publication approach:** Push changed data using common message definitions (canonical model) and notify the recipients
- **Develop complex event processing flows:**
  - Preparing historical data needed for the predictive model
  - Pre-populate predictive model and identify meaningful events
  - Executing triggered action in response to a prediction
- **Maintain DII Metadata:**
  - Manage Metadata uncovered during the development of DII solutions
  - Document all data structures involved
  - ETL vendors have Metadata repositories
  - Service-Oriented Architecture registry

2.4 Implement and Monitor

- Activate the data services that have been developed and tested.
- Real-time data processing needs real-time monitoring for issues, human or automated.
- Monitor and service at the level of the most demanding target application or consumer.

3 Tools

- **Data Transformation Engine/ETL Tool:** Primary tool that supports operation and design.
- **Data Virtualisation Server:** Perform data extract, transform and integrate virtually. A data warehouse is often input to a data virtualisation server
- **Enterprise Service Bus:** Middleware to support near real-time messaging between heterogeneous sources in the same enterprise
- **Business Rules Engine:** Allows non-technical users to manage business rules implemented by software
- **Data and Process Modelling Tools:** Used to design target and intermediate data structures
- **Data Profiling tool:** Use a tool to profile large amounts of data
- **Metadata Repository:** Tools have Metadata repositories which store the Rules for transformation, lineage and processing, as well as the instructions for scheduled processes and triggers.

4 Techniques

Described in Essential Concepts

5 Implementation Guidelines

5.1 Readiness Assessment / Risk Assessment

As all organisations already have DII in place assess for readiness/risk around tool implementation or interoperability. An enterprise data integration solution supports the movement of data between many applications and organisations.

Working DII solutions shouldn't be replaced. Focus on where none exists. Additional use of data integration adds to the investment in a data warehouse to Master Data Management hub.

It is necessary to sponsor the implementation of an enterprise data integration program by a high level of authority over solution design and technology purchase, to prevent local data integration solutions from developing. It may be perceived the cost of these is less than the enterprise wide solution.

Don't' become too focussed on the tool and lose focus on the business needs.

5.2 Organisation and Cultural change

- Local teams understand data in their applications.
- Central teams know tools and techniques.
- Enterprise solutions should be overseen by a Centre of Excellence
- Although technical, data integration solutions must be based on deep business knowledge to successfully deliver value.
- Modelling and data analysis should be done by business resources
- Canonical message model (consistent standard for how data is shared in the organisation) development requires technical and business resources
- Business SMEs review transformation mapping design and changes

6 DII Governance

Business is responsible for:

- Decisions about the design of data messages, data models and data transformation rules
- Defining the rules for loading and transforming data
- approve changes to these rules (which are captured as Metadata for cross enterprise analysis)
- Identifying and verifying predictive models and defining the actions they trigger

Governance controls to support trust that the DII will perform as promised:

- Determine what events trigger governance reviews (exceptions or critical events)
- Map each trigger to reviews that engage with governance bodies
- Event triggers may be part of SDLC at Stage Gates or part of User Stories

Controls come from governance-driven management routines such as mandated review of models, Metadata audits, gating of deliverables or required approval of changes to the transformation rules.

Include real-time operational data integration solutions in SLAs and Business Continuity/Disaster Recovery plans on the same tier as the most critical system to which they provide data.

Policies need to be established to ensure the organisation benefits from an enterprise approach to DII.

6.1 Data Sharing agreements

A data sharing agreement or memorandum of understanding (MOU) must be put in place before developing interfaces, and be approved by business data stewards, which stipulates:

- Responsibilities
- Acceptable use of data to be exchanged

- Anticipated use and access to the data
- Restriction on use
- Expected service levels
- Required system up times and response times

6.2 DII and Data Lineage

Governance to ensure that knowledge of data origins and movement is documented. Data lineage must be managed as it is critical Metadata.

Compliance standards require an organisation be able to describe where its data originated, and how it has changed as it moves through systems.

Impact analysis when making changes to data structures, data flows or data processing requires forward and backward data lineage.

6.3 Data Integration metrics

To measure scale and benefits of DII solutions:

- Data Availability (of data requested)
- Data volumes and speed
    - Volumes of data transported and transformed
    - Volumes of data analysed
    - Speed of transmission
    - Latency between update and availability
    - Latency between event and triggered action
    - Time to availability of new data sources
- Solution costs and complexity
    - Cost of developing and managing solutions
    - Ease of acquiring new data
    - Complexity of solutions and operations
    - Number of systems using data integration solutions