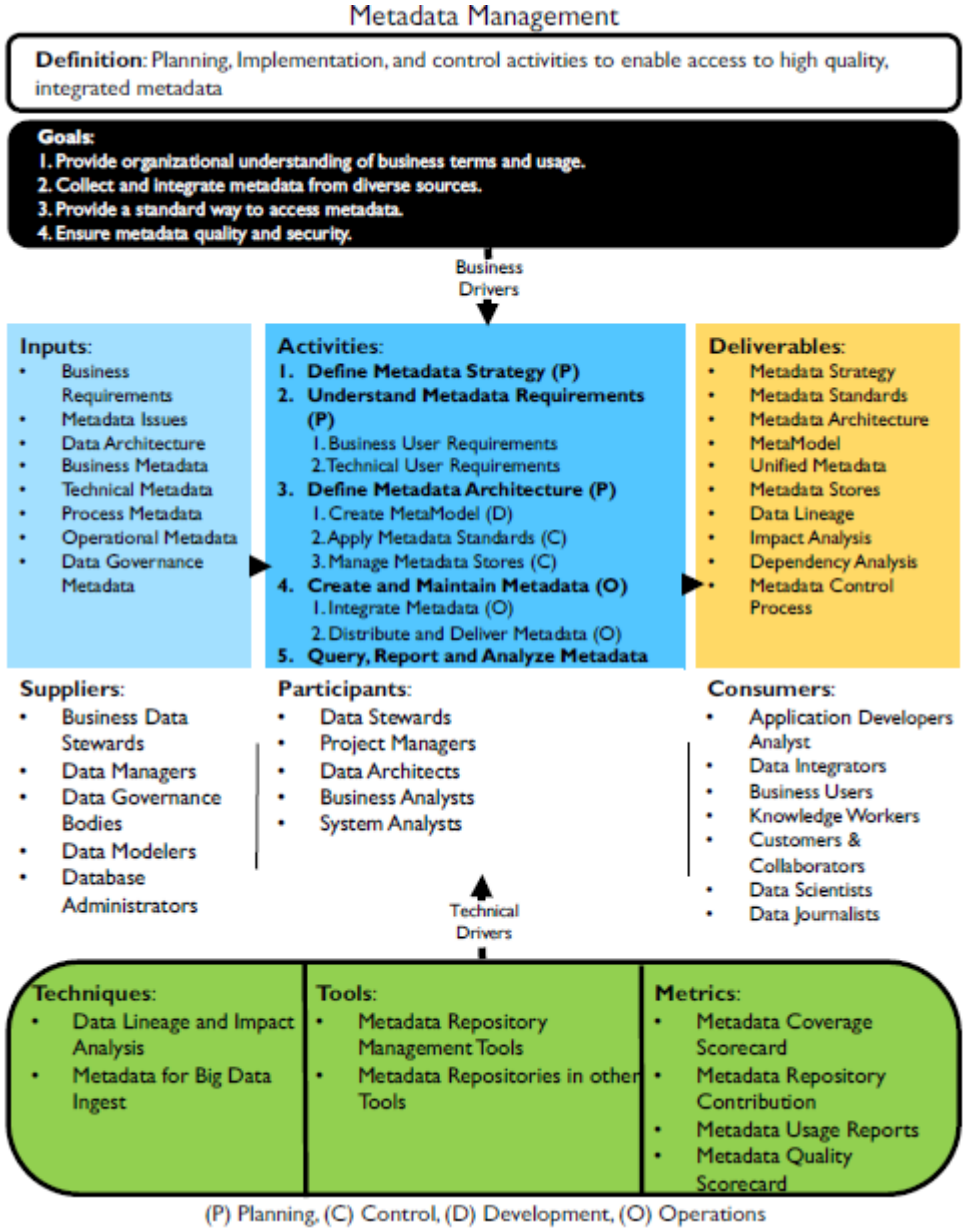
Metadata Management

1 Introduction

Metadata helps an organisation understand its data, systems and workflows. Metadata includes information about:

- Technical and business processes
- Data rules and constraints
- Logical and physical data structures
- Describes the data (databases, data elements, data models)
- Concepts the data represents
- Relationships between data and concepts



1.1 Business Drivers

Data cannot be managed without Metadata, which must also be managed. Metadata helps:

- Increase confidence in data by providing context, and measurement of data quality
- Increase value of strategic information (Master data) by enabling multiple uses
- Operational efficiency by identifying redundant data and processes
- Prevent the use of out of date or incorrect data
- Reduce data-oriented research time
- Improve communication between business and IT
- Create accurate impact analysis
- Reduce system development lifecycle time
- Support regulatory compliance

1.2 Goals and Principles

Goals:
1. Provide organizational understanding of business terms and usage.
2. Collect and integrate metadata from diverse sources.
3. Provide a standard way to access metadata.
4. Ensure metadata quality and security.

Implementation of a Metadata solution depends on the following principles:

- **Organisational commitment:** senior manager support. Metadata management is an enterprise program
- **Strategy:** How Metadata will be created, maintained, integrated and accessed. Metadata strategy must align with business
- **Enterprise perspective:** To ensure future extensibility but implement iteratively
- **Socialisation:** Communicate the necessity and purpose of each type of Metadata
- **Access:** Ensure staff members know how to access and use Metadata
- **Quality:** Process owners accountable
- **Audit:** Set, enforce and audit standards for Metadata
- **Improvement:** Feedback mechanism

1.3 Essential Concepts

1.3.1 Metadata vs. Data

Organisations should define Metadata requirements and what they need it for. What is data and what is Metadata depends on the organisation

Metadata is the "Who, What, Where, Why, When & How" of Data

Who	What	Where	Why	When	How
Who created this data?	What is the business definition of this data element?	Where is this data stored?	Why are we storing this data?	When was this data created?	How is this data formatted? (character, numeric, etc.)
Who is the Steward of this data?	What are the business rules for this data?	Where did this data come from?	What is its usage & purpose?	When was this data last updated?	How many databases or data sources store this data?
Who is using this data?	What is the security level or privacy level of this data?	Where is this data used & shared?	What are the business drivers for using this data?	How long should it be stored?	
Who "owns" this data?	What is the abbreviation or acronym for this data element?	Where is the backup for this data?		When does it need to be purged/deleted?	
Who is regulating or auditing this data?	What are the technical naming standards for database implementation?	Are there regional privacy or security policies that regulate this data?			

Global Data Strategy, Ltd. 2017 7

1.3.2 Types of Metadata

(NB If an exam question has any other type, refer to the DMBOK V1 notes, and please notify me)

Types:

- **Business Metadata:** Focus on the content and condition of data, and includes details related to Data Governance. Examples:
  - Definitions and descriptions of data sets
  - Business rules, transformation rules, calculations
  - Data models
  - Data quality rules
  - Update schedules
  - Provenance and data lineage
  - Data standards
  - Stakeholder contact details
  - Security/privacy level
  - Data usage notes
- **Technical Metadata:** Provides information about technical details of data, systems that store it and the processes that move it. Examples:
  - Physical database table and column names
  - Column properties
  - Database object properties
  - Access permissions
  - Data CRUD rules
  - Physical data models
  - Relationships between data models and physical assets
  - ETL job details
- **Operational Metadata:** Describes details of the processing and accessing of data. Examples:
  - Logs of job execution of batch jobs
  - History of extracts and results
  - Error logs

- Schedule anomalies
- Results of audit, balance, control measurements
- Reports and query access patterns, frequency, and execution time
- Patches and Version maintenance plan and execution, current patching level
- Backup, retention, date created, disaster recovery provisions
- SLA requirements and provisions
- Volumetric and usage patterns
- Data archiving and retention rules, related archives
- Data sharing rules and agreements
- Purge criteria
- Technical roles and responsibilities

1.3.3 ISO / IEC Metadata Registry Standard

ISO /IEC 11179 is structured in 6 parts:

- Part 1: Framework for the generation and standardisation of Data Elements
- Part 3: Basic Attributes of Data Elements
- Part 4: Rules and Guidelines for the Formulation of Data Elements
- Part 5: Naming and Identification Principles for Data Elements
- Part 6: Registration of Data Elements

1.3.4 Metadata for Unstructured data

Types of Metadata for unstructured data (Stored with the document):

- **Descriptive:** Catalogue information and thesauri keywords
- **Structural:** tags, field structures, format, terms on the Business Glossary
- **Administrative:** Sources, update schedules, access rights, navigation information
- **Bibliographic:** Library catalogue entries
- **Record keeping Metadata:** Retention policies
- **Preservation Metadata:** Storage, archival condition, rules for conservation, e-discovery

1.3.5 Sources of Metadata

- **Application Metadata Repositories:** The physical tables where Metadata is stored
- **Business Glossary:**
  - A document of the organisation's business concepts and terminology, definitions and the relationships between those terms. Accounts for hardware, software, database, processes and different user roles and responsibilities:
    - **Business users:** Data analysts, research analysts, management to understand terminology
    - **Data Stewards:** Manage the lifecycle of terms and definitions
    - **Technical users:** Use glossary terms to make design decisions
  - Business glossary should capture business terms attributes such as:
    - Term name, definition
    - Ownership and stewardship
- **Business Intelligence (BI) Tools:** Overview information, classes, objects, derived and calculated items
- **Configuration Management Tools:** CMDB manage Metadata related to IT assets.
- **Data Dictionaries:** Defines the contents and structure of data sets. This Metadata is embedded in database/modelling tools, and must be extracted to use

- **Data Integration Tools:** Lineage Metadata, movement of data
- **Database Management and System Catalogues:** Describe the content, sizing information, software version, deployment status, network and infrastructure uptime, availability etc.
- **Data Mapping Management Tools:** Mapping tools and data integration tools can exchange data with Metadata repositories
- **Data Quality Tools:** Can share quality scores and patterns with Metadata repositories
- **Directories and Catalogues:** Contains information about systems, sources and locations of data in the organisation. It is useful for developers and super users.
- **Event Messaging Tools:** Require a lot of Metadata to move messages between diverse systems.
- **Modelling Tools and Repositories:** Produce Metadata relevant to the design of the application or system model.
- **Service Registries:** Technical information about services and service end points e.g. APIs
- **Other Metadata Stores:** Specialised lists, repositories of repositories and business rules.

1.3.6 Types of Metadata Architecture

All architectural layers should point to the Metadata lifecycle:

- Metadata creation and sourcing
- Metadata storage in one or more repositories
- Metadata integration: Bring together Metadata from various repositories
- Metadata delivery: Formats for delivery to another system
- Metadata usage: Searching
- Metadata control and management

All require standards to be in place and enforced.

1.3.6.1 Centralised Metadata Architecture

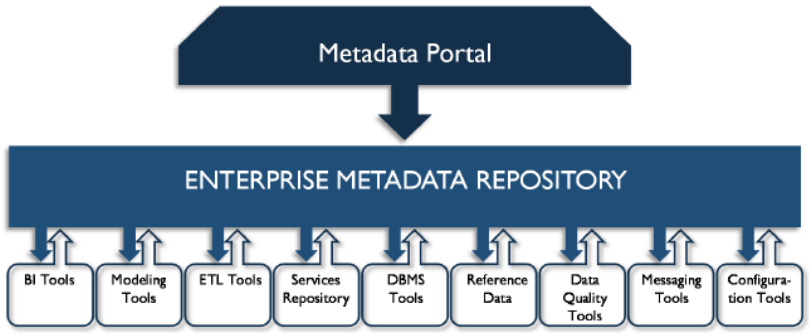


Figure 85 Centralized Metadata Architecture

A single Metadata repository containing copies of Metadata from the various sources.

Advantages of a centralised repository:

- High availability – independent of source systems
- Quick retrieval as repository and query reside together
- Resolved database structure not affected by proprietary nature of third party systems
- Quality is improved as extracted Metadata may be enhanced with Metadata from elsewhere

Limitations of a centralised repository:

- Complex processes ensure changes to source Metadata are quickly replicated into the repository
- Maintenance can be costly
- Extraction may require custom modules or middleware
- Increased demands on IT and vendor staff to maintain customised code

1.3.6.2 Distributed Metadata Architecture

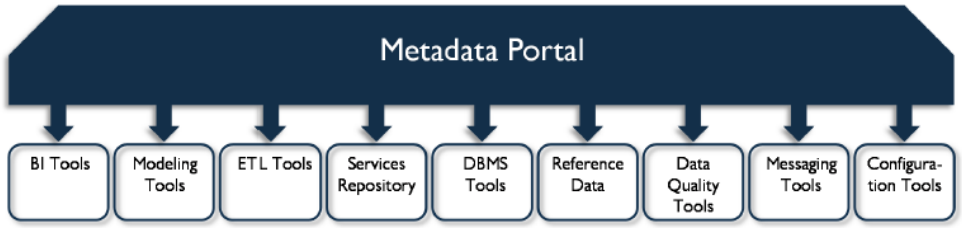


Figure 86 Distributed Metadata Architecture

No persistent repository, the portal passes user requests to the appropriate tool to execute. The Metadata retrieval engine retrieves data from source systems in real time. The Metadata management environment maintains source system catalogs and lookup information.

Advantages of the distributed Metadata architecture:

- Metadata is current and valid as it is retrieved from the source
- Queries are distributed – improved process and response time
- Implementation and maintenance effort minimised as Metadata requests from proprietary systems are queries, so no understanding of proprietary data structures is required
- Automated Metadata query processing is simpler
- Reduced batch processing

Limitations of distributed architecture:

- Not able to support user-added Metadata entries as there is no repository to put them
- Standardisation of presenting Metadata from various systems
- Query capabilities depend on the availability of the source systems
- Quality of Metadata depends on the source systems

1.3.6.3 Hybrid Metadata Architecture

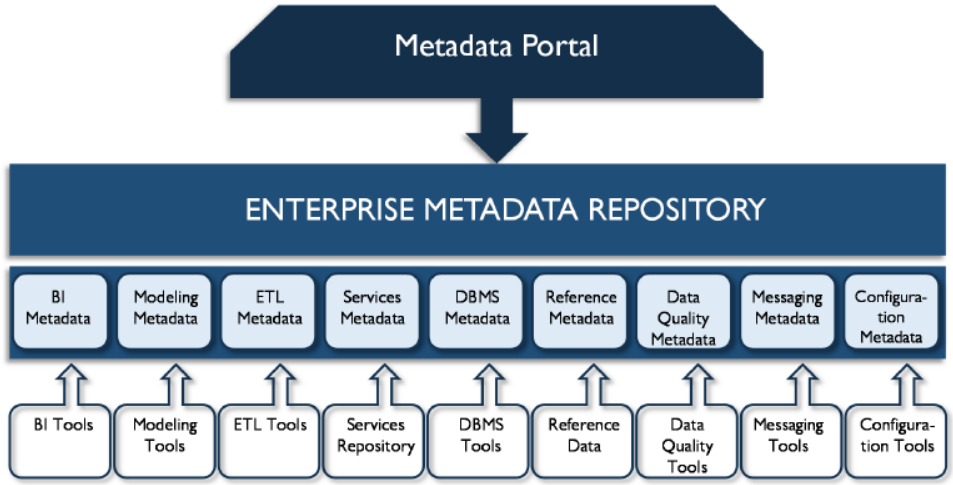


Figure 87 Hybrid Metadata Architecture

The repository design only accounts for the user-added Metadata, critical standardised items and the additions from manual sources. Near real-time retrieval of Metadata from source and enhanced Metadata when needed.

1.3.6.4 Bi-Directional Metadata Architecture

Allows Metadata to change in any part of the architecture and then feedback is coordinated from the repository to the original source. The repository is forced to contain the latest version of the Metadata source, and must manage changes to the source as well.

2 Activities

2.1 Define Metadata Strategy

Define the future state enterprise Metadata architecture and the implementation phases:

- **Initiate Metadata strategy planning:** Key stakeholders involved in planning
  - short- and long-term goals
  - Charter, scope, objectives and communications plan
- **Conduct key stakeholder interviews:** Foundation knowledge
- **Assess existing Metadata sources and information architecture:** Assess difficulty of project
  - Interview IT staff
  - Review system architecture and model documentation
- **Develop future Metadata architecture:** Develop long term architecture for the managed Metadata environment
- **Develop a phased implementation plan:** Prioritise findings from the interviews and data analyses to define a phased implementation plan.

2.2 Understand Metadata Requirements

What Metadata is needed and at what level. Functionality-focussed requirements:

- **Volatility:** Update frequency
- **Synchronisation:** Timing of updates in relation to source changes
- **History:** Do historical versions need to be retained?
- **Access rights:** Who, how and interface functionality

- **Structure:** Metadata model for storage
- **Integration:** Degree of and rules for integration
- **Maintenance:** Processes and rules for updating
- **Management:** Roles and responsibilities
- **Quality:** Metadata quality requirement
- **Security:** Some Metadata cannot be exposed as it will reveal sensitive data

2.3 Define Metadata Architecture

Metadata Management System must be able extract Metadata from many sources by scanning the sources and updating the repository, while supporting manual updates, searches and lookups by various user groups. Single access point for the Metadata repository which is transparent to users.

2.3.1 Create MetaModel

Data model for the Metadata repository:

- High-level conceptual model – relationships between systems
- Lower level metamodel that describes the elements and processes
- The metamodel is a planning tool, and Metadata itself

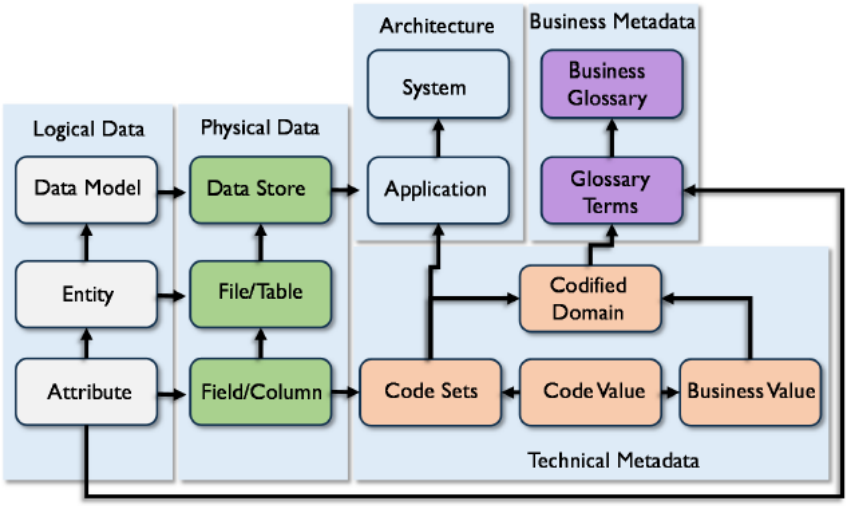
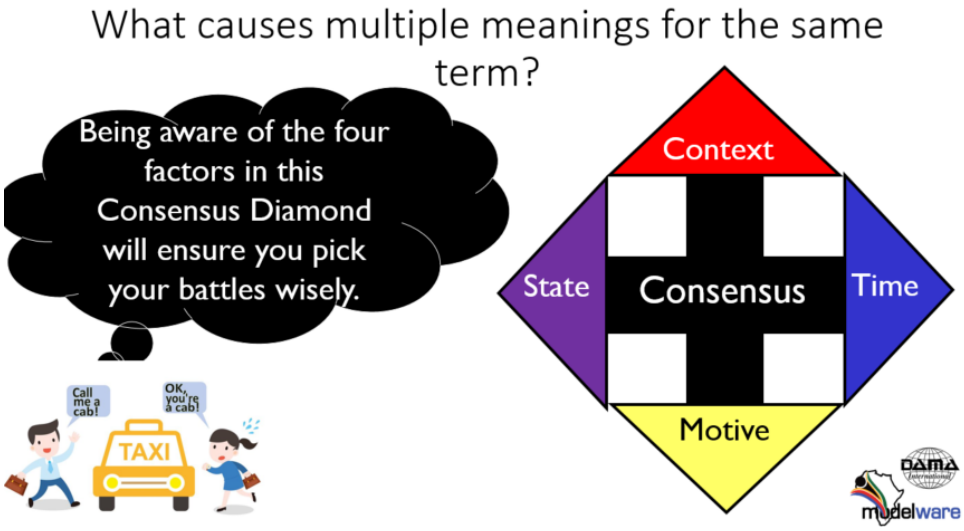


Figure 88 Example Metadata Repository Metamodel

2.3.2 Apply Metadata Standards

Governance monitors for compliance:

- Internal standards such as naming conventions (ISO 11179 for naming conventions)
- External standards such as data exchange formats



2.3.3 Manage Metadata Stores

Control activities should have governance oversight and include:

- Job scheduling and monitoring
- Load statistical analysis
- Backup, recovery, archive, purging
- Configuration modifications
- Performance tuning
- Query statistics analysis
- Query and report generation
- Security management

Quality control activities:

- QA, quality control
- Matching update sets to timeframes
- Missing Metadata reports

Metadata management activities include:

- Loading, scanning, importing and tagging assets
- Source mapping and movement
- Versioning
- User interface management
- Linking data sets Metadata maintenance – for NoSQL provisioning
- Linking data sets to internal data acquisition – custom links and job Metadata
- Licensing for external data sources and feeds
- Data enhancement Metadata e.g. Link to GIS

Training:

- Education and training of users and data stewards
- Management metrics generation and analysis
- Training on the control activities and query and reporting

2.4 Create and Maintain Metadata

Metadata should be planned and created as a product. Profile and inspect for quality. Schedule maintenance.

General principles of Metadata management:

Accountability: Metadata is often produced through existing processes – hold those process owners accountable for quality of Metadata.

Standards: Set, enforce and audit Metadata standards

Improvement: Create a consumer feedback mechanism

2.4.1 Integrate Metadata

As Metadata is gathered from many sources, and integrated into the Metadata repository, challenges arise which require governance.

Two approaches for repository scanning:

- **Proprietary interface:** Collection and loading of Metadata occurs in single step. No format specific file output.
- **Semi-proprietary interface:** Two step process where scanner collects Metadata from a source and outputs it to a format specific data file

Files used during the scanning process:

- **Control file:** Contains the source structure of the data model
- **Reuse file:** Contains the rules for managing reuse of process tools
- **Log file:** Produced during each phase of the process
- **Temporary and backup files:** Used for traceability

2.4.2 Distribute and Deliver Metadata

Metadata is delivered to consumers/applications/tools requiring Metadata feeds:

- Metadata intranet websites
- Reports, glossaries and other documents
- Data warehouses, data marts and BI tools
- Modelling and software development tools
- Messaging and transactions
- Web services and APIs
- External organisation interface solutions

Metadata is exchanged with external organisations using files (flat, XML or JSON) or web services.

2.5 Query, Report and Analyse Metadata

A Metadata repository has a front-end application for search and retrieval as Metadata guides the use of data assets

3 Tools

The Metadata repository is the primary tool used to manage Metadata. It has an integration layer and an interface for manual updates. Metadata repository management tools are also a source of Metadata

4 Techniques

4.1 Data Lineage and Impact Analysis

Metadata about the physical assets provides information about how the data is transformed as it moves between systems. Limited to the scope of the Metadata management system.

- **'As Implemented Lineage'**:
  - Current version of lineage based on programming code.
  - Imported from various tools
- **'As Designed Lineage'**:
  - Lineage described in mapping specification documents.
  - Not extractable by Metadata management system
- **Stitching**: The process whereby the Metadata management system augments the 'As Implemented' data lineage with the 'As Designed'.

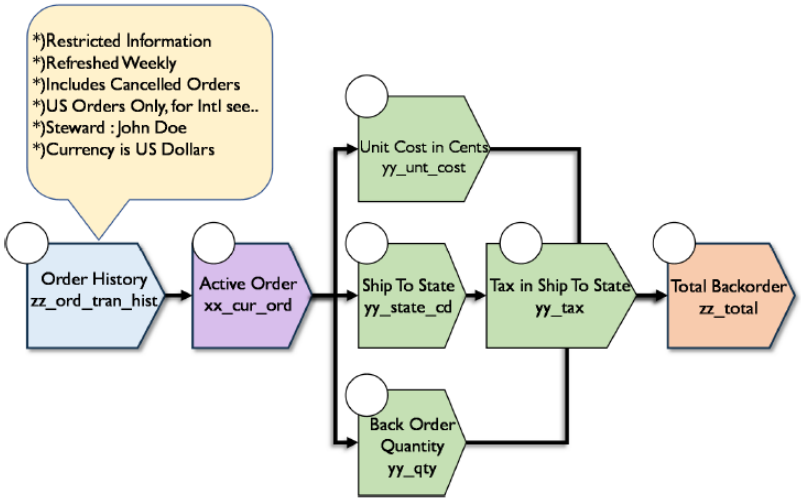


Figure 89 Sample Data Element Lineage Flow Diagram

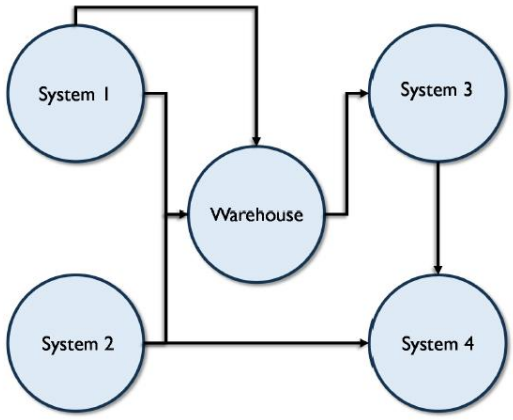


Figure 90 Sample System Lineage Flow Diagram

Successful lineage discovery needs to account for both business and technical focus:

- **Business focus:** Data elements prioritised by business
  - Start at target and trace back to source
  - Gives business understanding what happens to a data element as it moves
  - DQ measurements with lineage can pinpoint where system design impacts quality.
- **Technical focus:**

- Start at source systems
- Identify immediate consumers, then next sets of consumers until all systems are identified.

4.2 Metadata for Big Data Ingest

Metadata tags should be applied to data on ingestion to the data lake. Ingestion engines can profile data as well.

5 Implementation Guidelines

Implement the Metadata environment in incremental steps to minimise risks and facilitate acceptance. Use a relational database platform. Contents should be generic in design and should be integrated so that consumers can see across different data sources. Should house current, planned and historical versions of Metadata.

5.1 Readiness assessment

People should be aware of the risks of not managing Metadata:

- Errors in judgement due to lack of knowledge of the context of data
- Exposure of sensitive data
- Risk that SMEs will leave and take their knowledge of the data with them

A formal assessment of the current maturity of Metadata activities includes:

- Critical data elements
- Available Metadata glossaries
- Lineage
- Data profiling and data quality processes
- MDM maturity

5.2 Organisational and Cultural Change

Metadata efforts often meet with resistance. Needs senior management support and engagement. Business and technical staff work closely in a cross-functional manner.

6 Metadata Governance

Determine the specific requirements for the management of the Metadata lifecycle, and establish governance processes. Formal roles and responsibilities need to be assigned to dedicated resources.

6.1 Process Controls

Governance team responsible for:

- Defining standards
- Managing status changes for Metadata
- Promotion of Metadata
- Training
- Management of business terms

Metadata strategy should be integrated into the SDLC to ensure Metadata is collected and remains current.

6.2 Documentation of Metadata Solutions

A master catalogue of Metadata of the sources and targets currently on scope. It is a 'what-is-where' guide for the user community and includes:

- Metadata implementation status
- Source and target Metadata store
- Schedule information for updates
- Retention and versions kept
- Contents
- Quality statements or warnings
- System of record or other data source statuses
- Tools, architecture and people involved
- Sensitive information and removal or masking for the source

6.3 Metadata Standards and Guidelines

Metadata standards are required in the exchange of data with operational trading partners.

- Use industry-based Metadata standards early
- Tool vendors provide XML, JSON or REST support to exchange their data
- Tools offer import/export capabilities using XML
- Templates and examples
- ISO standards

6.4 Metrics

- **Metadata repository completeness:** Ideal coverage compared to actual coverage
- **Metadata Management Maturity:** Based on the Capability Maturity Model (CMM-DMM) assessment approach
- **Steward representation:** Coverage across enterprise for stewardship
- **Metadata usage:** User uptake measured in logins
- **Business Glossary activity:** Usage, update, resolution of definitions, coverage
- **Metadata documentation quality:** Assess automatically and manually
- **Metadata repository availability:** Uptime, processing time (batch and query)