



**PLURALSIGHT**

# Data Management – Key Concepts

# Key Concepts

- Master Data Management (MDM)
- Metadata
- Data Governance
- Data Catalog
- Data Lineage

# Master Data Management (MDM)

- **Definition:** A method used to define and manage an organization's critical data to provide a single point of reference.
- **Purpose:** Ensures consistency, accuracy, stewardship, and accountability of shared data assets.
- **Key Idea:** Think of it as the 'single source of truth' for core data in an organization, like customer or product data.

# Metadata

- **Definition:** Data about data. Metadata provides information about a certain data's content, quality, condition, origin, and other characteristics.
- **Purpose:** Helps in organizing, locating, and understanding data.
- **Key Idea:** Just as a book's summary gives you a quick idea about the book's content, metadata provides insights into what the data is about.

# Data Lineage

- **Definition:** A visual representation that tracks data from its source to its destination, showcasing how it moves and transforms through systems.
- **Purpose:** Provides transparency and clarity on how data is sourced, used, and transformed.
- **Key Idea:** It's like a map for data, showing its journey and transformations from start to finish.

# Master Data Management

# Master Data Management Deep Dive

## What is Master Data?

- **Definition:** The core data that is essential to an organization's operations and decision-making. This includes data on customers, products, suppliers, employees, etc.
- Master Data Management (MDM) is a process that creates a uniform set of data on customers, products, suppliers, and other business entities across different IT systems.
- **WHEN:** Organizations establish the need for MDM when they hold more than one copy of data about a business entity.
- **GOAL:**
- MDM seeks to ensure that an organization does not use multiple (potentially inconsistent) versions of the same master data in different parts of its operations, which can occur in large organizations.
- The goal of MDM is to create a single master record for each person, place, or thing in a business, from across internal and external data sources and applications
- **Example:** A product's name, ID, and specifications in a manufacturing company.



# Master Data Management Deep Dive

## Data System Types – MDM Perspective:

### System of Record (SOR):

- A system of record is the authoritative data source for a particular data element or transaction
- It is the primary system that creates, stores, and maintains a particular data element or transaction

### Examples

- ERP systems (like financials, manufacturing, CRM, HR) often serve as the primary source of data for various business processes.
- Financial systems are the definitive source for financial data.
- EMR systems in healthcare are authoritative for patient medical records.
- Inventory systems for stock levels.
- Asset tracking systems for the status and location of assets.
- Employee and customer records often reside in HR systems or CRMs and are authoritative for those specific entities.

# Master Data Management Deep Dive

## Data System Types – MDM Perspective:

### System of Reference (SOREf):

- A system of reference is a system that provides reference data to other systems
- It is a system that provides a standard set of data that is used by other systems to ensure consistency and accuracy

### Examples:

- Reference data systems typically provide standard sets of data used across the enterprise.
- Standard data sets could be industry benchmarks or standardized datasets used for analytics.
- Taxonomy systems, code sets, and data dictionaries all provide standardized ways of categorizing or defining data.

# Master Data Management Deep Dive

## Data System Types – MDM Perspective:

### System of Engagement (SOE):

- A system of engagement is a system that interacts with customers, employees, or other stakeholders
- It is a system that provides a user interface for users to interact with the organization

### Examples:

**Email systems** capture communication interactions.

**Collaboration systems** like team chat applications or project management tools.

**Social networking systems** for social interactions.

**Learning systems** for educational interactions and progress tracking.

**Mobile applications** that provide specific functionalities and capture user behaviors.

**Self-service systems**, like customer portals where users can manage their accounts.

# Master Data Management Deep Dive

## What is a Golden Record?

- **Definition:** A single, well-defined version of all the data entities in an organizational ecosystem. It's the most accurate, complete, and reliable representation.
- A golden record is a fundamental concept within **Master Data Management (MDM)** and **Data Quality Management (DQM)** that identifies and defines the single version of truth, where truth is understood to be data that is trusted to both be accurate and correct
- **Example:** If there are three different addresses for a customer across three systems, the 'golden' one is the most accurate and up-to-date.

# Master Data Management Deep Dive

## Difficulties and Challenges in MDM

- Data from disparate sources
- Handling data inconsistencies
- Changing data sources and structures over time
- Ensuring data quality and accuracy

# Master Data Management Deep Dive

## Matching

### Probabilistic vs Deterministic Matching

#### Deterministic:

- Uses specific rules to find exact matches.
- **Example:** Two records are a match if their Social Security Numbers and birth dates are the same.

#### Probabilistic:

- Uses statistical algorithms to find likely matches based on a number of attributes.
- **Example:** Two records might be a 90% match if names, addresses, and birth dates are similar but not identical.

## Deduplication

- The process of removing duplicate records and consolidating data.
- **Importance:** Ensures that each piece of master data is represented only once.

# Master Data Management Deep Dive

Master information encompasses all the core attributes and metadata about the master data.

## **Importance of understanding this info:**

- Ensures completeness of data
- Provides context for decision-making
- Avoids data misuse

# Master Data Management Deep Dive

## **Get familiar with Master Data Management (MDM) as a practice and discipline**

- MDM is both a technology and a business discipline.
- It ensures that organizations have consistent, accurate, and trustworthy master data.
- Involves processes, governance, policies, standards, and tools to manage and maintain master data.



# Master Data Management Deep Dive

## Explore MDM Application

MDM solutions help in creating, maintaining, and using master data.

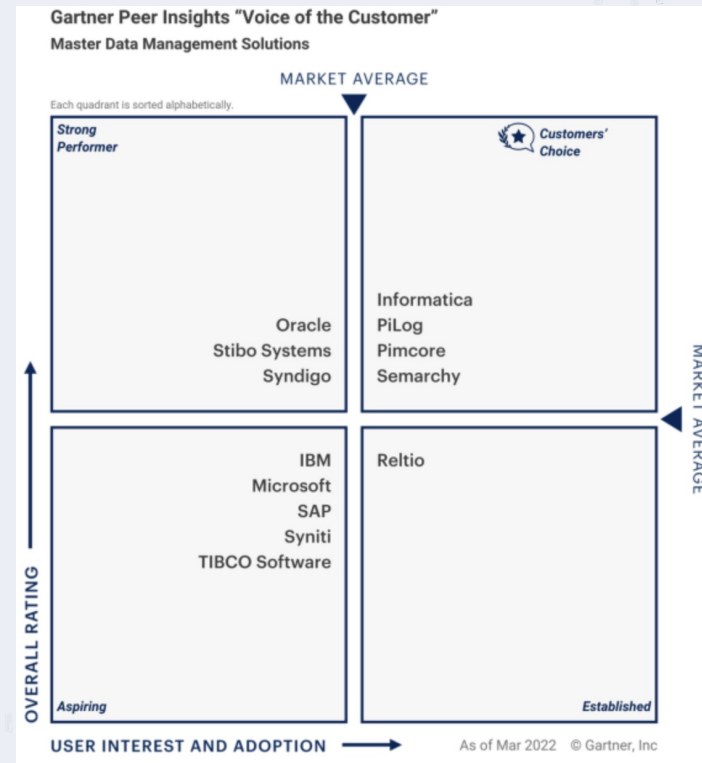
### Applications may include:

- Data modeling
- Data integration
- Data quality tools
- Data governance tools

# Master Data Management Deep Dive

## Popular Tools in the Market

Boomi Master Data Hub  
IBM InfoSphere MDM  
Informatica MDM



# Master Data Management Deep Dive

## Data quality in MDM

- **Ensures reliability:** Stakeholders trust and use the data.
- **Improves efficiency:** Reduces costs associated with bad data (e.g., incorrect shipments).
- **Boosts decision-making:** Accurate data leads to better business decisions.
- **Regulatory compliance:** Ensures organizations meet data standards and regulations.

# Data Governance

# Data Governance Deep Dive

Data governance is the overarching strategy, policies, and processes that ensure the formal management of data assets within an organization.

**Data Governance is a set of principles and practices that ensure high quality through the complete lifecycle of data**

**Effective data governance ensures that data is consistent, trustworthy, and doesn't get misused**

# Data Governance Deep Dive

## Components of a Data Governance Framework

### **Data Governance Council or Committee:**

- This is a cross-functional team responsible for setting the strategic direction for data governance efforts. They make key decisions and provide sponsorship and support for governance initiatives.

### **Data Stewards:**

- Individuals or teams designated to oversee the quality, lifecycle, and usage policies for specific data domains or business areas. They act as liaisons between IT and business units and are accountable for data within their purview.

### **Policies and Standards:**

- Documents that outline the rules, guidelines, and best practices for data acquisition, maintenance, use, and disposal. These can include data quality standards, metadata standards, security policies, and more.

# Data Governance Deep Dive

## Components of a Data Governance Framework

### **Data Governance Processes and Procedures:**

- Defined workflows that outline how data-related issues are identified, reported, and resolved. This also includes processes for change management, data classification, and more.

### **Data Quality Measures:**

- Tools and metrics to continuously monitor and report on the quality of data. This ensures that data remains accurate, consistent, timely, and fit for its intended purpose.

### **Data Catalog & Metadata Management:**

- Systems or tools that help in documenting data definitions, data lineage, relationships, and other metadata. This aids in understanding and finding data assets within the organization.

### **Data Privacy and Security:**

- Measures to ensure that data is accessed, stored, and shared in compliance with regulations and that sensitive data is protected. This includes access controls, encryption, and more.

# Data Governance Deep Dive

## Components of a Data Governance Framework

### **Data Architecture:**

- A blueprint of how data is structured, integrated, and interrelated. It helps in understanding the flow of data and the relationships between different data elements.

### **Training and Communication:**

- Regular training sessions, workshops, and communication channels to educate stakeholders about data governance policies, standards, and best practices. This ensures that everyone in the organization is aligned with the governance objectives.

### **Issue Resolution & Escalation Processes:**

- Mechanisms to address data issues, conflicts, or breaches. It defines how issues are reported, who addresses them, and how they are escalated if necessary.



# Data Governance Deep Dive

## Why Data Governance?

- Ensures data's trustworthiness and reliability.
- Helps in leveraging data as an asset.
- Makes sure data is used appropriately and is protected.

## Ownership and Accountability

- Ensures that there's clear responsibility for data's accuracy, accessibility, consistency, and protection.
- Designated roles (like Data Stewards) ensure that data is managed and used appropriately.

# Data Governance Deep Dive

## What is the Notion of Audit?

Audits involve assessing processes, systems, and data to ensure compliance with governance policies and other regulations.

## Audit Centric View of Governance

- Understand that everyone has a role in governance.
- Even if one isn't the decision-making authority, they still play a vital part.
- They might need to **report** on data processes, **explain** inconsistencies, and **communicate** data challenges and successes.

# Data Governance Deep Dive

## What Data Governance Covers

- **Data Architecture and Modeling:** How data is structured and interrelated.
- **Data Quality Management:** Ensuring data is accurate, reliable, and usable.
- **Data Operations Management:** Handling, processing, and maintenance of data.
- **Data Security Management:** Protecting data from unauthorized access and breaches.
- **Reference and Master Data Management:** Managing core consistent data points.
- **Data Warehousing and Business Intelligence Management:** Storage, retrieval, and analysis of data.
- **Document and Content Management:** Managing unstructured data and documents.
- **Meta Data Management:** Managing data about the data, providing context.
- **Data Development:** Adapting and innovating with data.
- **Database Operations Management:** Ensuring databases run smoothly and efficiently

# Data Governance Deep Dive

## The Ten Facets of Data Governance

- **Vision and Business Case:** A clear vision and business case that delivers business value.
- **People:** A team responsible for managing the data governance program, including executive leadership, project management, line-of-business managers, and data stewards.
- **Tools and Architecture:** Tools and architecture to support the architectural scope of data governance.
- **Policies:** Policies that make up the data governance function, including security, archiving, and other policies.
- **Measurement:** Measuring the level of influence of a data governance initiative and measuring its effectiveness, such as increasing revenue, improving operational efficiency, reducing risk, reducing cost, or improving customer satisfaction.
- **Change Management:** Incentives to workforce, partners, and customers to get better quality data in and potential repercussions if data is not of good quality.
- **Organizational Alignment:** How the organization will work together across silos.
- **Dependent Processes:** Identifying data lifecycles, all processes consuming the data, and processes to store and manage the data.
- **Program Management:** Effective program management skills to build out communication strategy, measurement, and reporting.
- **Communication:** Communication strategy to ensure that all stakeholders are informed and engaged.

# Data Governance Deep Dive

## Value Proposition for Data Governance

- **Risk Management:** Avoid fines and sanctions from non-compliance.
- **Efficiency and Cost Reduction:** Reduce costs associated with bad data.
- **Improved Decision-making:** Reliable data leads to better business decisions.
- **Enhanced Data Usability:** Ensures data is available, accurate, and easy to find/use.
- **Reputation:** Build trust with stakeholders, customers, and partners.

# Data Governance Deep Dive

## Key Areas of Data Governance

- **Policy & Standards:** Defining the rules for how data is used and managed.
- **Data Quality:** Ensuring the accuracy, completeness, reliability, and timeliness of data.
- **Privacy & Security:** Protecting data and ensuring compliance with regulations.
- **Data Lifecycle Management:** Overseeing data from creation to deletion.
- **Data Roles & Responsibilities:** Defining who does what in the data world.
- **Data Architecture:** Structuring data in a way that meets organizational needs.
- **Communication & Training:** Ensuring everyone understands their roles and responsibilities related to data.

# Data Lineage/Cataloging

# Data Lineage & Cataloging Deep Dive

Both tools help organizations understand and manage their data better. While a data catalog focuses on organizing and finding data, data lineage visualizes how data moves and transforms.



# Data Lineage & Cataloging Deep Dive

## What is a Data Catalog?

- **Definition:** A centralized repository that allows users to find, understand, and use the data they need.
- It's essentially a "Google" for your data, making it easier to discover and understand.

## Why is a Data Catalog Fundamental?

- **Ease of Discovery:** With the vast amounts of data organizations handle, a catalog makes it easier to find what's needed.
- **Enhanced Collaboration:** Users can share insights, make annotations, or rate datasets.
- **Ensures Compliance:** Helps in tracking sensitive data and ensuring it's used appropriately.
- **Reduces Redundancy:** Prevents the same data from being pulled or analyzed multiple times.

# Data Lineage & Cataloging Deep Dive

## Key Components of a Data Catalog

- **Metadata Repository:** Stores details about each data set (like data type, last updated, source, etc.).
- **Search Engine:** Allows users to search for data using various criteria.
- **User Interface:** A user-friendly portal where users can browse, search, and annotate data.

# Data Lineage & Cataloging Deep Dive

## Practical Usage of a Data Catalog

- **Data Discovery:** Find new data sets that could be relevant for analysis.
- **Collaboration:** Share notes about a particular data set with colleagues.
- **Governance:** Track where sensitive data is and who is accessing it.
- **Self-Service Analytics:** Empower non-technical users to find and use data on their own.

# Data Lineage & Cataloging Deep Dive

## What is Data Lineage?

- **Definition:** A visual representation that tracks data from its source to its destination, showcasing how it moves and transforms through systems.
- Helps in understanding data's journey – from where it originates, how it's processed, to where it's consumed.

# Data Lineage & Cataloging Deep Dive

## Value and Importance of Data Mapping

- **Integration:** Simplifies the process of integrating data from disparate sources.
- **Data Transformation:** Understand how data is transformed or processed as it moves.
- **Accuracy:** Ensures data remains accurate and consistent when transferred or integrated.
- **Compliance:** Helps in ensuring data is appropriately handled, especially sensitive data.

# Data Lineage & Cataloging Deep Dive

## How to Use the Data Mapping Document

- **Identify Source and Target:** Understand where your data is coming from and where it's going.
- **Understand Transformations:** Check for any operations or changes the data undergoes.
- **Validate Mappings:** Ensure that mappings are correct and that no data is lost or misrepresented.
- **Document Changes:** Always make note of any changes or updates to the mappings for future reference.

# Personal Identifiable Information (PII)

# WHAT IS PII?

**Personal Identifiable Information (PII)** refers to any information that can be used to distinguish or trace an individual's identity. This might be standalone data like a Social Security number or a combination of data points like full name combined with date of birth.



# WHAT IS PII?

## What is Non-Personal Information?

Non-Personal Information pertains to data that cannot identify an individual on its own. This could include aggregated data or information about user behavior that doesn't tie back to an individual directly.

# WHAT IS PII?

## How to Identify PII:

- Recognizing PII involves distinguishing data that can directly or indirectly pinpoint individual identities. Common examples include names, addresses, phone numbers, email addresses, and identification numbers.

## How to associate sensitive data with PII:

- Some PII is more sensitive than other types. For instance, financial information or health records, when associated with a name or another identifier, becomes particularly delicate.

## Primary vs. Secondary:

- Primary PII can identify a person independently (e.g., Social Security numbers). Secondary PII, like a date of birth or place of birth, often needs to be combined with other data to identify an individual.
- **Example:** VIN (Vehicle Identification Number) - On its own, a VIN is not PII. However, when linked with an owner's name or address, it becomes part of a PII dataset.

# WHAT IS PII?

## Handling PII:

Proper PII management ensures individual privacy and regulatory compliance.

## Techniques and Approaches:

- **Data Masking:** Concealing original data with modified content but structurally similar to the original.
- **Encryption:** Transforming data into code to prevent unauthorized access.
- **Tokenization:** Replacing sensitive data with non-sensitive substitute tokens.
- **Access Controls:** Ensuring only authorized individuals can access PII.
- **Regular Audits:** Periodically reviewing and verifying PII handling practices.

# WHAT IS PII?

## **CPRA (California Privacy Rights Act) and PI:**

The CPRA strengthens consumer privacy rights for California residents. It expands on the California Consumer Privacy Act (CCPA) and introduces new requirements for businesses and new privacy rights.

## **CPPA - California Privacy Protection Agency:**

Established under CPRA, the CPPA is an agency dedicated to protecting the privacy rights of consumers by enforcing the CPRA.

# WHAT IS PII?

## PHI, HIPAA, and PCI:

- **PHI (Protected Health Information):** Any information about health status, provision of healthcare, or payment for healthcare that can be linked to an individual.
- **HIPAA (Health Insurance Portability and Accountability Act):** A US law designed to protect PHI and ensure healthcare industry entities secure patient data.
- **PCI (Payment Card Industry) Data Security Standard:** A set of security standards designed to ensure that all companies accept, process, store, or transmit credit card information securely.

# WHAT IS PII?

## **GDPR (What is GDPR and why it is important):**

The General Data Protection Regulation (GDPR) is a regulation in EU law that deals with data protection and privacy for individuals within the European Union. It's crucial due to its widespread implications for businesses that process the data of EU citizens, regardless of where the business is located. Non-compliance can lead to hefty fines. The GDPR emphasizes transparency, security, and accountability by data controllers, while also granting individuals more control over their personal data.