

Metadata and Metadata Management

Objectives

After completing this reading, you will be able to:

- Define what metadata is
- Describe what metadata management is
- Explain the importance of metadata management
- List popular tools for metadata management

What is metadata?

Metadata is data that provides information about other data.

This is a very broad definition. Here we will consider the concept of metadata within the context of databases, data warehousing, business intelligence systems, and all kinds of data repositories and platforms.

We'll consider the following three main types of metadata:

- Technical metadata
- Process metadata, and
- Business metadata

Technical metadata

Technical metadata is metadata which defines the data structures in data repositories or platforms, primarily from a technical perspective.

For example, technical metadata in a data warehouse includes assets such as:

- Tables that record information about the tables stored in a database, like:
 - each table's name
 - the number of columns and rows each table has
- A data catalog, which is an inventory of tables that contain information, like:
 - the name of each database in the enterprise data warehouse
 - the name of each column present in each database
 - the names of every table that each column is contained in
 - the type of data that each column contains

The technical metadata for relational databases is typically stored in specialized tables in the database called the System Catalog.

Process metadata

Process metadata describes the processes that operate behind business systems such as data warehouses, accounting systems, or customer relationship management tools.

Many important enterprise systems are responsible for collecting and processing data from various sources. Such critical systems need to be monitored for failures and any performance anomalies that arise. Process metadata for such systems includes tracking things like:

- process start and end times
- disk usage
- where data was moved from and to, and
- how many users access the system at any given time

This sort of data is invaluable for troubleshooting and optimizing workflows and ad hoc queries.

Business metadata

Users who want to explore and analyze data within and outside the enterprise are typically interested in *data discovery*. They need to be able to find data which is meaningful and valuable to them and know where that data can be accessed from. These business-minded users are thus interested in business metadata, which is information about the data described in readily interpretable ways, such as:

- how the data is acquired
- what the data is measuring or describing
- the connection between the data and other data sources

Business metadata also serves as documentation for the entire data warehouse system.

Managing metadata

Managing metadata includes developing and administering policies and processes to ensure information can be accessed and integrated from various sources and appropriately shared across the entire enterprise.

Creation of a reliable, user-friendly data catalog is a primary objective of a metadata management model. The data catalog is a core component of a modern metadata management system, serving as the main asset around which metadata management is administered. It serves as the basis by which companies can inventory and efficiently organize their data systems. A modern metadata management model will include a web-based user interface that enables engineers and business users to easily search for and find information on key attributes such as CustomerName or ProductType. This kind of model is central to any Data Governance initiative.

Why is metadata management important?

Good metadata management has many valuable benefits. Having access to a well implemented data catalog greatly enhances data discovery, repeatability, governance, and can also facilitate access to data.

Well managed metadata helps you to understand both the business context associated with the enterprise data and the data lineage, which helps to improve data governance. Data lineage provides information about the origin of the data and how it gets transformed and moved, and thus it facilitates tracing of data errors back to their root cause. Data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data, and data controls are implemented that support business objectives.

The key focus areas of data governance include availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organization.

Popular tools for metadata management

Popular metadata management tools include:

- IBM InfoSphere Information Server
- CA Erwin Data Modeler
- Oracle Warehouse Builder
- SAS Data Integration Server
- Talend Data Fabric
- Alation Data Catalog
- SAP Information Steward
- Microsoft Azure Data Catalog
- IBM Watson Knowledge Catalog
- Oracle Enterprise Metadata Management (OEMM)
- Adaptive Metadata Manager
- Unifi Data Catalog
- data.world
- Informatica Enterprise Data Catalog

Summary

In this reading, you learned that:

- Metadata is data that provides information about other data, and includes three main types: technical, process, and business metadata
- The technical metadata for relational databases is typically stored in specialized tables in the database called the system catalog
- A primary objective of business metadata management modelling is the creation and maintenance of a reliable, user-friendly data catalog
- Having access to a well-implemented data catalog greatly enhances data discovery, repeatability, governance, and can also facilitate access to data
- Metadata management tools from IBM include InfoSphere Information Server and Watson Knowledge Catalog