

Metadata (Management) for Data Warehousing

An Overview of Process, Advantages and Types



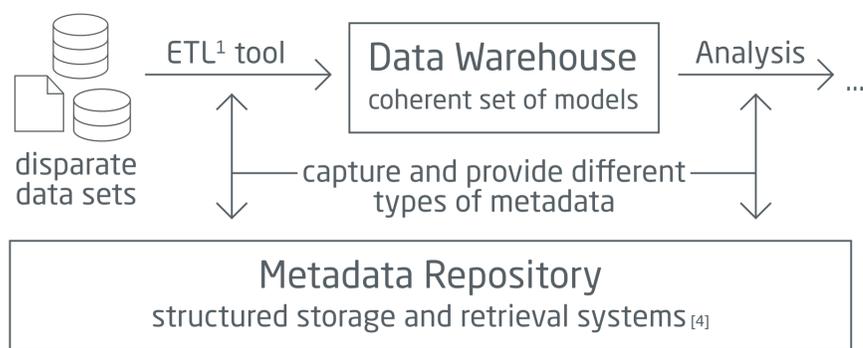
Abstract

Metadata management is one of the **key success factors** in data warehousing. Arbitrary analyses on data warehouse systems with enormous amounts of processed data (*Snowflake* has customers with data in petabyte ranges [1], *SAP HANA* collected > 100 TB of development metadata for several years [2]) require metadata. This metadata needs to be **collected continuously while creation, operation and use of the data warehouse**. There are different types of metadata **to improve work for both, the end user and the developer**. *Due to different definitions and categorizations of metadata, the author tried to summarize and generalize various approaches.*

Metadata

... is any information supporting the **administration and effective exploitation of data warehouses** [4] including the resulting information supply. Therefore, metadata improves the understanding of data elements and their relations. [5]

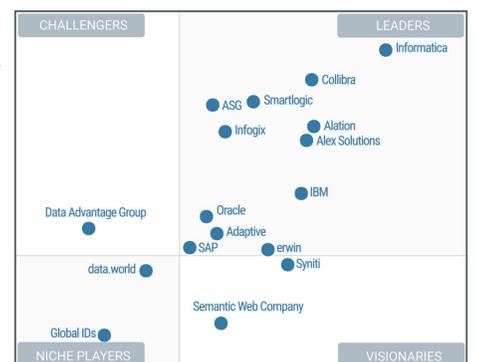
Typical Data Warehouse Process [3]



Metadata Management

... are all processes related to generation/administration/use of metadata. [5]

- metadata can be modeled and managed the same way other data is managed – it is also data [3]
- complex market due to various scope/capability (vendors on the right) [8]
- *Snowflake* uses the Key-Value-Store *FoundationDB* as metadata store [9]



Excerpt from *Magic Quadrant for Metadata Management Solutions* [8]

Types of Metadata [4,6] and Examples [7]

Operational Metadata

- operational reporting and statistics (access logs, timestamps, etc.)
- record manipulation information (last access, last change, etc.)
- data practitioners

ETL¹ JOBS ■ TRANSFORMATIONS

mostly used by end users

mostly used by database and software systems

Structural Metadata

- table, data element and record structure information

DATA MODELS ■ RECORD MODELS

Business Metadata

- meaning of data in business sense [7]
- enables business processes and business analytics
- merges *technical metadata* with business meanings of the data, reporting directives, etc.

BUSINESS DEF. ■ LOGICAL DATA MAPPINGS

Technical Metadata

- information necessary for application development and execution

INTERACTION TYPES ■ TABLE INDECES

Advantages of Metadata [4,5]

- ↑ increasing effectivity of extracting information
 - improve interaction with the data warehouse
 - advance data analysis (locate and interpret relevant data)
- + enabling of data unification
 - serve defined (or computed) specific data transformations
 - e.g.: marketing defines customer as account owner, but for the sales department it is a person with purchases
- + pruning, time travel and zero-copy cloning² [1]
 - invented by *Snowflake*, saving metadata for every micro-partition of the stored data on each data update
- ↓ lowering efforts for development, maintenance and administration
 - store reusable abstractions and configurations
 - consistent and integrated documentation
 - automate various administration processes through metadata-driven ETL¹ tool [6]

Notes

- 1 Extract, Transform, Load data from multiple sources into target data store incl. data cleansing (detect and correct/remove corrupt records).
 2 Pruning: determine only affected files when selecting; **time travel**: select data store state at specific time; **zero-copy cloning**: consolidate info about records from multiple files without copying.

References

- [1] Hentschel, Martin; Heibel, Max. "File Metadata Management in Snowflake." *Lecture Series on Practical Data Engineering 2019/20*, 12.11.2019.
 [2] Böhm, Dr. Alexander. "SAP HANA. Software Development Process." *Lecture Series on Practical Data Engineering 2019/20*, 10.12.2019.
 [3] Loshin, David. *Business Intelligence. The Savvy Manager's Guide*, Second Edition, 2013.
 [4] Vaduva, Anca; Dittrich, Klaus. *Metadata Management for Data Warehousing: Between Vision and Reality*, 2001.
 [5] Melchert, Florian; Auth, Gunnar; Herrmann, Clemens. *Integriertes Metadatenmanagement für das Data Warehousing*, 2002.
 [6] Kimball, Ralph; Ross, Margy. *The Data Warehouse Toolkit*, Third Edition, 2013.
 [7] Kimball, Ralph; Caserta, Joe. *The Data Warehouse ETL Toolkit*, 2004.
 [8] Simoni, Guido De; Beyer, Mark; Jain, Ankush. "Magic Quadrant for Metadata Management Solutions." *Gartner*, Published on 16.10.2019.
 [9] "Metadata. Data Warehousing Glossar." *Snowflake*. Last accessed on 27.01.2020. www.snowflake.com/data-warehousing-glossary/metadata.

All graphics created by the author.

Author

Vincent Xeno Rahn
 Bachelor Student

Lecture Series on Practical Data Engineering
 Digital Engineering | University of Potsdam
 January 2020