CENTRO UNIVERSITÁRIO UNIMETROCAMP WYDEN -PR
Graduate Program in Data Science and Analytical Intelligence

Regilene Mariano

How to make a Data Dictionary

Paraná
2024

Table of contents

List of tables

Introduction

A data dictionary is an essential tool for companies that make data-driven decisions. It provides a detailed description of data elements, their characteristics, and their relationships. It can also include technical metadata such as data object names and definitions, properties, business rules for validation, reference data, handling of missing data, and more. The metadata repository, built within the Data Warehouse/Data Marts environment, is a crucial asset for both the BI team and the organization's users, as it maintains essential information about the data contained in the environment, enabling data identification. This article examines the components of a data dictionary, its benefits, the setup process, and best practices.

1. What is a Data Dictionary?

Kimball and Ross (2013) state that metadata is analogous to the encyclopedia of the Data Warehouse and BI. Therefore, the analyst must be diligent in populating and maintaining the metadata repository. According to Khurana (2024), CEO and Data Analytics expert at Atlan, a Data Dictionary is a collection of metadata such as object names, data types, sizes, classifications, and relationships with other data assets. The primary goal of a data dictionary is to help data teams understand and manage these data assets effectively. In the view of The Data Governance T.C.Willian (2023), the Data Dictionary is more technical, providing detailed specifications about data elements and attributes used in databases. Meaning the data dictionary is a repository or document that contains detailed information about the data elements. It serves as a reference guide for data professionals, database administrators, developers, and other stakeholders involved in managing and using data.

2.What is a data dictionary used for?

A data dictionary is used by data administrators, analysts, and engineers to understand and trust data assets. It helps in the creation of transparent and consistent data throughout the organization and it serves several purposes:

Documentation: It documents the definitions, formats, and relationships of data elements, helping users understand how data is organized and used.

Consistency: It ensures consistency in data usage by providing standardized definitions and formats for data elements across different systems and applications.

Reference: It acts as a reference guide for data professionals, database administrators, developers, and other stakeholders, facilitating their work in managing, analyzing, and using data.

Data Management: It aids in effective data management by describing data structures, constraints, and rules, which helps in data integration, quality control, and troubleshooting.

Communication: It improves communication between different teams and stakeholders by providing a common understanding of data elements and their meanings.

3.Components of a data dictionary

According to Monteiro and Podgorski (2024), the components of a Data Dictionary can be:

- A listing of data objects (names and definitions);
- Detailed properties of data elements (data type, size, nullability, optionality, indexes);
- Entity-relationship (ER) and other system-level diagrams;
- Reference data (classification and descriptive domains);
- Missing data and quality-indicator codes;
- Business rules, such as for validation of a schema or data quality;

The data dictionary should also include:

- Data source (data warehouse, data lakes, databases, applications);
- Date and time when the property was created or changed;
- Descriptive statistics that go beyond missing values, such as min-max values and histogram distribution;
- Owners and editors of data sets that contain these variables;
- SQL queries attached to the data asset;

- Social metadata associated with each data asset - stored as tags, notes, and chat transcripts;

4.How to create a Data Dictionary?

Khurana (2023) recommends that Data Dictionary should contains:

- Name, definition, and description of each variable
- Measurement units
- A range of accepted values, along with minimum and maximum values.

Start with basic information about the data:

Table	Relationship	Name of relationship	Description
tbl_customer	tbl_payment	A customer can make multiple payments	Stores information about customers
tbl_product	tbl_payment	A product can be associated with multiple payments	Stores information about products
tbl_payment	tbl_customer	Each payment is made by a customer specific	Stores information about payments and references customers and products
	tbl_product	Each payment is associated with a specific products	

Tabela1: Basic information about data

Describe the attributes of each table:

CUSTOMER TABLE				
Field Name	**Data Type**	**Length**	**Constraint**	**Description**
idcustomer	INT		IDENTITY PRIMARY KEY NN	Automatically generated identification for each customer
name	VARCHAR	100		Name of the customer
cpf	VARCHAR	11	UNIQUE NN	Unique identification number (Brazilian CPF)
age	INT			Age of the customer
birth_data	DATA		NN	Birth date of the custome
email	VARCHAR		NN	Email address of the customer
phone	VARCHAR		NN	Number phone of the customer

Tabela2: customer table - Attributest of data

PRODUCT TABLE				
Field Name	**Data Type**	**Length**	**Constraint**	**Description**
idproduct	INT		IDENTITY PRIMARY KEY NN	Automatically generated identification for each product
product_name	VARCHAR	100	NN	Name of the product
description	TEXT	200	NN	Description of the product
product_type	VARCHAR	100	NN	Type/category of the product
minimum_value	DECIMAL	10,2	NN	Minimum value of the product
maximum_value	DECIMAL	10,2	NN	Maximum value of the product
maximum_installments	VARCHAR	10	NN	Maximum number of installments allowed
special_conditions	TEXT	200	NN	Special conditions related to the product

Tabela13: Table of Product

PAYMENT TABLE				
Field Name	**Data Type**	**Length**	**Constraint**	**Description**
idpayment	INT		IDENTITY PRIMARY KEY NN	Automatically generated identification for each payment
installment_num	INT		NN	Number of the installment
installment_amt	DECIMAL	10,2	NN	Amount of the installment
due_date	DATA		NN	Due date for the payment
payment_date	DATA		NN	Date when the payment was made
payment_status	VARCHAR	100	NN	Status of the payment
customer_id	INT		FK, NN	REFERENCES customer(idcustomer)
product_id	INT		FK, NN	REFERENCES product(idproduct)

Tabela4: Table of payment

Describe the relationships between the tables:

Relationship	Table	Description
Um Customer can make multiple payments	**tbl_custome**r and tbl_payment	Each customer can make multiple payments
A product can be associated with multiple payments	tbl_product and tbl_payment	Each product can be associated with multiple payments.
Each payment is made by a specific customer	tbl_payment and tbl_customer	Each payment is linked to a specific customer via a foreign key
Each payment is associated with a specific product	tbl_payment and tbl_product	Each payment is linked to a specific product via a foreign key.

Tabela5: Relationship of tables

Conclusion

As we can see, a data dictionary is essential for understanding the design, structure, and data flow of a database.It serves as a comprehensive reference that provides detailed descriptions of data elements, their relationships, and definitions. By facilitating a shared vocabulary among data users, it enhances communication and collaboration. It also helps identify and correct errors and inconsistencies, ensuring data quality and integrity. Furthermore, a data dictionary streamlines data discovery, supports reliable analytics and reporting, and assists in compliance audits by managing data quality and security. It aids in enforcing database standards and makes it easier to onboard new team members, ultimately contributing to more efficient and effective data management.

REFERENCES

BARBIERI, C. Governança de dados: práticas, conceitos e novos caminhos. Rio de Janeiro: Alta Books, 2020.

KIMBALL, R. The Data Warehouse toolkit — técnicas para construção de Data Warehouses dimensionais. 1. ed. Rio de Janeiro: Makron Books, 1998.

KIMBALL, R.; ROSS, M. The Data Warehouse toolkit — the definitive guide to dimensional modeling. 3. ed. Indianapolis: John Wiley Sons, 2013.

WILLIAN, T.C. Data glossary data dictionary. Medium, 2023. <https://medium.com/@william.tc/data-glossary-data-dictionary-data-catalog-54b2d398b12a >

KHURANA, A. What is a dictionary. 2024 - 05-30. <https://atlan.com/what-is-a-data-dictionary > Access in 07 july.

MONTEIRO, V. G. S. Arquitetura de Data Warehouse e Data Marts. Rio de Janeiro: YDUQS, 2020.