DAT 202 - Assignment 1

Colin Bowers - May 29, 2024

Identify if each data set in the assignment is structured, semi-structured or unstructured.

Each of these datasets are structured. They exist in a fixed, tabular schema with predefined columns and rows. Each row represents an individual record and each column has a definite type (e.g. text, numeric) that describes an attribute of each record.

Explain the relationship between the Technical and Business metadata

Technical metadata includes details about the structure, storage, and management of data, such as data types, constraints, database schemas, table structures, and column definitions. Business metadata provides context and meaning to the data from a business perspective including business definitions, data usage, business rules, calculation methods and data ownership.

The relationship between the two is that business metadata gives context to the technical metadata, ensuring that the data is understood and used correctly for business purposes. Both are essential for effective data governance and usage.

Explain why metadata is so important to understand the data for this use case?

Metadata is crucial for understanding and effectively using these sets of data. Here are some examples of where a misinterpretation of how to use it could have a negative business impact.

In the sales table, the *saleamount* field represents the total amount of a sale transaction. Without clear metadata, a business user might confuse this with other monetary fields such as *rate* or *quote*, leading to errors in revenue calculations, sales performance analysis, and forecasting.

Furthermore, understanding that the industry field is a type of sub-category of the *sector* field and each value “belongs to” a particular sector value. Misinterpreting these fields could result in incorrect categorization of clients and sales data, impacting market strategy, client targeting, and competitive analysis.

Explain how maintaining this metadata repository will assist in providing better quality data for analysis.

Maintaining a metadata repository offers several benefits that contribute to better data quality and usability. Firstly, it improves data quality by enforcing data standards, providing clear definitions, and ensuring consistency across datasets.

Secondly, a metadata repository enhances data discoverability, making it easier for users to find and understand available data assets, leading to better-informed data usage. It also improves any analysis on the data by reducing the complexity and the cognitive load required.

Finally, it facilitates better data integration by helping to understand relationships between different datasets, reducing duplication, and ensuring smoother data integration processes.

For the case study prepare the following using excel (the Metadata tool is not available) a) Business Glossary b) Technical Metadata

Business Glossary

Term	Definition	Owner	Source
Client ID	Unique identifier for a client/company	Data team	Clients
Sector	Category of business activity that the company belongs to	Marketing team	Clients, Sales
Industry	Specific industry within a sector that the company belongs to	Marketing team	Clients, Sales
Company Name	Name of the company that made a purchase or having a market capitalization	Accounts team	Clients, Sales
Customer Name	Name of the employee of the company that authorized the purchase		
Sale Amount	Total dollar amount of a sale transaction (in US Dollars)	Accounts team	Sales
Market Capitalization	Total market value of a company's outstanding shares	Finance team	Clients
Zip Code	Postal code used for geographic location	Accounts team	Sales

Technical Glossary

Note: This list is incomplete; only showing the fields related to this assignment.

Table	Column	Type	Null	Key	Notes
clients	client_id	int	No	Primary	
	name	text	No	Unique	Linked to sales.companyname
	symbol	text	No	Unique	
	last_sale	float	No		
	market_cap	text	No		
	cap_float	float	No		
	ipo_year	int	No		
	sector	text	No		Same value range as sales.sector
	industry	text	No		Same value range as sales.industry
	url	text	No		
sales	sector	text	No		Same value range as clients.sector
	industry	text	No		Same value range as clients.industry
	companyname	text	No	Foreign	Linked to clients.name
	zipcode	int	No		
	saleamount	float	No		