Sample data dictionary and data lineage

As you have been learning in this course, business intelligence professionals have three primary tools to help them ensure conformity from source to destination: schema validation, data dictionaries, and data lineages. In this reading, you're going to explore some examples of data dictionaries and lineages to get a better understanding of how these items work.

Data dictionaries

A data dictionary is a collection of information that describes the content, format, and structure of data objects within a database, as well as their relationships. This can also be referred to as a metadata repository because data dictionaries use metadata to define the use and origin of other pieces of data. Here's an example of a product table that exists within a sales database:

Product Table

Item_ID	Price	Department	Number_of_Sales	Number_in_Stock	Seasonal
47257	$33.00	Gardening	744	598	Yes
39496	$82.00	Home Decor	383	729	Yes
73302	$56.00	Furniture	874	193	No
16507	$100.00	Home Office	310	559	Yes
1232	$125.00	Party Supplies	351	517	No
3412	$45.00	Gardening	901	942	No
54228	$60.00	Party Supplies	139	520	No
66415	$38.00	Home Decor	615	433	Yes
78736	$12.00	Grocery	739	648	No
34369	$28.00	Gardening	555	389	Yes

This table is actually the final target table for data gathered from multiple sources. It's important to ensure consistency from the sources to the destination because this data is coming from different places within the system. This is where the data dictionary comes in:

Data dictionary

Name	Definition	Data Type
Item_ID	ID number assigned to all product items in-store	Integer
Price	Current price of product item	Integer
Department	Which department the product item belongs to	Character
Number_of_Sales	The current number of product items sold	Integer
Number_in_Stock	The current number of product items in stock	Integer
Seasonal	Whether or not the product item is only seasonally available	Boolean

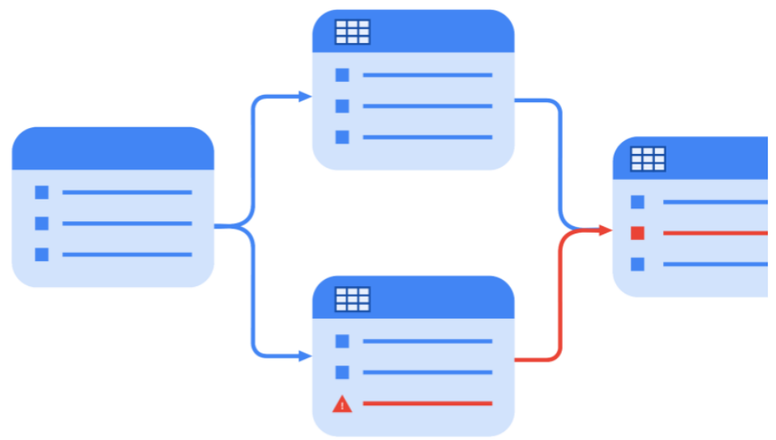
You can use the properties outlined in the dictionary to compare incoming data to the destination table. If any data objects don't match the entries in the dictionary, then the data validation will flag the error before the incorrect data is ingested.

For example, if incoming data that is being delivered to the Department column contains numerical data, you can quickly identify that there has been an error before it gets delivered because the data dictionary states Department data should be character-type.

Data lineages

A data lineage describes the process of identifying the origin of data, where it has moved throughout the system, and how it has transformed over time. This can be really helpful for BI professionals, because when they do encounter an error, they can actually track it to the source using the lineage. Then, they can implement checks to prevent the same issue from occuring again.

For example, imagine your system flagged an error with some incoming data about the number of sales for a particular item. It can be hard to find where this error occurred if you don't know the lineage of that particular piece of data– but by following that data's path through your system, you can figure out where to build a check.



By tracking the sales data through its life cycle in the system, you find that there was an issue with the original database it came from and that data needs to be transformed before it's ingested into later tables.

Key takeaways

Tools such as data dictionaries and data lineages are useful for preventing inconsistencies as data is moved from source systems to its final destination. It is important that users accessing and using that data can be confident that it is correct and consistent. This depends on the data being delivered into the target systems has already been validated. This is key for building trustworthy reports and dashboards as a BI professional!

Mark as completed