Yukun Ma

Week #3: Data and Metadata

Page #1

Problem Statement

I will describe data and metadata, evaluate the differences between data and metadata and show how they are manipulated in DBMS.

References

[1] Wikipedia. Data(computing) - wikipedia. https://en.wikipedia.org/wiki/Data_(computing). Accessed March 17, 2018.

[2] Wikipedia. Metadata - wikipedia. https://en.wikipedia.org/wiki/Metadata. Accessed March 17, 2018.

Overview of Data and Metadata

In computer science, data is any sequence of one or more symbols given meaning by specific act(s) of interpretation. And it requires interpretation to become information. [1] While the metadata is ”data that provides information about other data” [2]. And it is well-organized information.

In DBMS[1], data is saved in a table by using data structures like B+ Trees. One can use DML[2] to manipulate(add、delete or update) the data in a database. While for the metadata, one should use DDL[3] to manipulate it.

Critical Thinking

Table 1: The Differences Between Data and Metadata

	Data	Metadata
What is?	a set of symbols	data about data
Informative?	maybe(needs interpretation)	always
Manipulation Language	DML	DDL
Example	column names	scores of students
Amount	huge	smaller

[1] DataBase Management System

[2] Data Manipulation Language (e.g., SELECT * FROM STUDENT: pick all the data from table “STUDENT”)

[3] Data Definition Language (e.g., DROP TABLE STUDENT: delete the table “STUDENT” and all the data in it)

Page #2

Question

By using metadata and other data, can we check the validity of some data? If we can, how to do it?

Method

Describe how you are going to answer your own question stated above.

Analysis and Discussion

I think it is difficult for us to be very confident in the verification of data. But I think we can check the validity of the data at least to a certain degree in most cases.

For example, when dealing with a column with a column name “Precipitation”. We can not confirm that every item in that column is accurate. But we can be certainly sure that every item must be a number although it may be expressed in scientific notation or ends with different units of measure such as “cm” or “mm”.

We can check the validity of data with or without other data. When doing it without other data, we can only use our knowledge and the limits made through metadata to validate the data. Limits can be formats(e.g., initializing with “#” and ending with a number), ranges(e.g., normal human body temperature is between 36 degrees Celsius and 38 degrees Celsius), etc.

While when doing that with other data, things are getting interesting. We can deduce a lot of interesting limits on the data from other data. For example, if one’s current state is Georgia, then the city he is currently in cannot be New York. But which one is not correct? Georgia or New York? Of course, it can be further analyzed. But it needs inference and induction which are not so simple to me and to computer.

I believe it can be solved more elegantly by using deep learning. We can just use some data that we can totally trust to train the neural networks. And then we use the networks to pick the unauthenticated data.