Data Quality Definitions and Measures

There are many definitions of data quality, in general, data quality is the assessment of how much the data is usable and fits for the following DM task

Having accurate and business-ready data is an absolutely integral component to ensure that DM application users (companies) do not experience the negative impacts that can accompany “bad” or “dirty” data.

There are five components that will ensure data quality; completeness, consistency, accuracy, validity, and timeliness. When each of these components are properly executed, it will result in high-quality data.

1. **Accuracy:** Data collected is correct, relevant and accurately represents what it should. Accuracy is more challenging to remedy than data completeness and consistency. The Inaccuracy may be resulted from Incorrect Logging.

   Incorrect logging

   Note : “Inaccuracy is not Noise” : In DM Process, people use the term “Noise” for exceptional behavior - not for incorrect logging. If DM algorithm is said to be able to deal with noise, then it can identify (Mine) low-frequent behavior with the help of the main process flow only. It is impossible for discovery algorithms to distinguish incorrect logging from exceptional events.
   What incorrect logging means is that the *recorded data is wrong*.

   Here are two true stories of incorrect data:

   - Example : 1

     - In an ERP system, data entries from invoice documents had been scanned automatically. However, because of a mistake in the scanning procedure the invoice ID was interpreted as the invoice date for some of the cases. As a result, activities with a timestamp of the year 2020 appeared in the log data.

   - Example 2:

     - In a process improvement project in a hospital the **data showed low utilization rates** for two specific wards. Actually the hospital closed 2 wards but had to re-open them again shortly afterwards, which resulted in the “missing information” for those two wards. The problem was also due to the manual entries by the staff one day later so the entry date were not matching to the closing and reopening dates of the wards.

- The other inaccuracy may be resulted due to inconsistency in collection process. Eg. One sensor is sensing at the end of time cycle and another is sensing at the begning of the time cycle

2. **Completeness:** Ensuring there are no gaps in the data from what was supposed to be collected and what was actually collected. The Noncompleteness may be result from Insufficient logging

Insufficient logging

While incorrect logging is about wrong data, insufficient logging is about missing data.

Typical problems with missing data are:

- Fields in the database of the information system are simply overwritten. So, old entries are lost and the database only provides information about the current status, but not the overall history of what happened in the past.

- Some systems employ "batch logging" procedures, where, for example, activities are logged once a day (all at once). This way, all changes in-between are lost as well as the ordering of what happened when cannot be reconstructed anymore.

    - Typical OLAP and data mining techniques do not require the whole history of a process, and therefore data warehouses often do not contain all the data that is needed for process mining.

- Another problem is that, ironically, by logging too much data sometimes there is not enough data. I have heard of more than one SAP or enterprise service bus system that does not keep logs longer than one month for the sheer amount of data that would accumulate otherwise. But processes often run longer than one month and, therefore, logs from a larger timeframe would be needed.

- Finally, for specific types of analysis additional data is required. For example, to calculate execution times for activities both start and completion timestamps must be available in the data. For an organizational analysis, the person or the department that performed an activity should be included in the log extract, and so forth.

3. **Consistency:** The types of data must align with the expected versions of the data being collected. Inconsistency means Violation of semantic rules defined over the database.

   One of the biggest challenges can be to find the right information and to understand what it means.

   In fact, figuring out the semantics of existing IT logs can be anything between really easy and incredibly complicated. It largely depends on how distant the logs are from the actual business logic. For example, the performed business process steps may be recorded directly with their activity name, or you might need a mapping between some kind of cryptic action code and the actual business activity.

   Sol1: Normally menu driven Data Entry will help preserving semantics

   Sol 2: Predefining log time for extracting data

4. **Validity:** Validity is derived from the process instead of the final result. When there is a need to fix invalid data, there is an issue with the process rather than the results (Invalid data will affect results in absolute meaning). This makes it a little trickier to resolve. Any change in the storage-structure (domain, range, normalization etc) creates the need to validate the data before DM process. Co-relationship plays major role in maintaining validity..

   Correlation

   Because process mining is based on the *history* of a process, the individual process instances need to be reconstructed from the log data. Correlation is about stitching everything together in the correct way:

- Business processes often span multiple IT systems, and usually each IT system has its own local IDs. One needs to correlate these local process IDs to combine log fragments from the different systems (local ID from system No. 1 and local ID from system No. 2) in order to get a full picture of the process from start to end.

- Even within the same system correlation may be necessary. For example, in an ERP purchase-to-pay process purchase orders are identified by purchase order IDs and later on the invoices are characterized by invoice IDs. To get an end-to-end process perspective, the corresponding purchase order IDs and invoice IDs need to be matched.

- Sometimes, there are hierarchical processes and then activity instances need to be distinguished to correlate lower-level events that belong to these (activity) sub processes.

5. **Timeliness:** The data should be received at the expected time in order for the information to be utilized efficiently. Anything slower becomes an inadequate source of information. With real time data and analytics, companies are better equipped to make more effective and informed decisions. There is a pressing need to eliminate the lag time between when a survey is completed in the field and when it is received. Timing is Important to maintain the timeliness.

Precisely because process mining evaluates the history of performed process instances, the timing is very important for ordering the events within each sequence. If the timestamps are wrong or not precise enough, then it is difficult to create the correct order of events in the history.

Some of the problems seen with timestamps are:

- Timestamp resolution is too low. For example, only the date of a performed activity (but not the time) is recorded. But even if the time is recorded, it may be necessary to record it at least with millisecond accuracy if many events follow each other in automated systems.

- Different timestamp granularities on different systems. For example, the timestamps in one system may be rounded to minutes. Another system (which is also executing a part of the process) records events with 1-second resolution. When put together, the order of some of the events may be wrong due to the granularity difference.

- Different clocks on different systems. If multiple computers record data, then these computers can have different system clocks. In the merged log, these time differences then create problems, since they destroy the correct order of events.

Ideally, timestamps should be precise, not be rounded up or down, and synchronized (if there are multiple systems). If there are differences, it may help to work with offsets. If too many events have the same timestamp, one can try to use the original sequence of events.