



Data Capture and Data Quality



The key differences between manual data entry and automated data capture including the advantages and disadvantages:-

- Manual data entry is reliant upon humans and can be error prone and inefficient. This can be collected via forms and surveys etc.
- Automated data capture is done by use of technology. This can be collected via sensors, API's and web scraping tools etc. These are more efficient, scalable and contain less errors.

6 elements of data quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Uniqueness
- Validity



ILLUSTRATION: INUENG/GETTY IMAGES; 62024 TECHTARGET, ALL RIGHTS RESERVED

The importance of data quality in data science projects and why it is critical to ensure high data quality at the beginning of the data lifecycle:-

In order for businesses to make change successful they need to have data from verifiable sources that is complete, error free and whole. As you can see from the graphic previously, it shows the 6 elements of data quality that is part of the DQAF - Data Quality Assessment Framework. This needs to be followed to allow any change for the business to reflect the correct needs of that business going forward. If the data that is being relied upon does not follow this framework then it could be disastrous and cause loss and valuable time to reimplement the change that was needed in the first place. This could prove very costly and if the business was for example in sales then the marketing, customer data and preferences of their customer base could be directly affected, sales could slump and business would be stunted. They would most likely lose a lot of their customer base as people do not forgive mistakes easily, which in turn could lead to bad reviews that will turn people away from using their services.



Two key dimensions of data quality:-

- Completeness - The more complete the data the better the output will be after it has been analysed. This needs to be 100% or it is classed as incomplete.
- Consistency - The data needs to be consistent over every system to ensure that when they are being compared they are the same and accurate and should not conflict with each other.

How they affect data analysis:-

- If the data is not complete, then it can skew the results, giving the wrong basis for the way forward. e.g if the product being sold is a financial product and the data collected gives the wrong information, the customer may not choose that product as the historical data has

shown the wrong trend in the investment. This makes it harder for the customer to compare products.

- If the data is not consistent and all systems do not show the correct information, the customers information will not have been updated across the company. This could lead to their order for example going to their old address, or their birthday card (if automatically sent out) could be sent on the wrong date. This could lead to loss of business and a tarnished reputation for the business.



Challenges that arise during real-time data capture, and how this can impact the quality of data:-

- Integration of the vast amounts of data that is coming in by the nano second can be challenging as it needs to be cleaned, transformed and loaded all at the same time. This can affect the business because of the cost and scale of the IT architecture needed is vast in order to process this efficiently.
- The data coming into the system may be of a different format and not be compatible with the other.
- If for example the data is being captured for the stock market then this will need to be very accurate and up to date to ensure the traders can buy and sell at the correct time. If not the investors may lose money. The quality of the data needs to be of a very high standard and be of a high integrity.
- If the data is not accurate or of a low quality then it could lead to incorrect outputs and end up incorrectly guiding the company in the wrong direction.
- The data quality can also be affected by anomalous data and inconsistency in the sources. Some of the data may also be fraudulent.
- Current data is usually favoured over historical analysis as it is not always possible to carry out both and can affect the long term plans of the business.



The relationship between data capture and data quality, and how poor data capture practices affects the entire data science process

- If the data capture is of poor quality then it can wrongly guide people who are going to use the data for their own company.
- The standardisation of forms across all data science companies should be used as i.e. if data formats vary between institutes then the outcome of that data from the forms can be erroneous. This will change the output and affect how this information is used.
- The data captured should be validated so that it ensures the correct or desired output and makes the data far more accurate.
- Data being collected shouldn't have too many data fields as this could confuse the processed data with too much information.
- Make sure the questions that are being asked are correct and suitable for the task you are performing. If not, then it could affect the output of the data.
- Data should be continually audited as it changes over time. If the data is outdated then it can produce the wrong results.