Dirty data

Dirty data is a term used to describe any type of electronic data that is outdated, incomplete, or otherwise not accurate. Data of this type may be created due to errors in data entry, a failure to update the data on a regular basis, or even the entry of the same data more than once. At times, the incorrect data is nothing more than errors in punctuation in the text of electronic documents. In other instances, dirty data may be information that is intentionally misleading, such as attempts to modify accounting records to present a specific image to investors and others.

Dirty data may also occur due to a failure to update existing records when information changes. For example, if salespeople fail to update customer files when personnel changes occur with a given customer, those files are no longer accurate and are considered dirty. As with correcting spelling and punctuation errors, taking the time to remove outdated information and replace it with current data helps to increase the overall usability of the database.

There are situations where the creation of dirty data is intentional. Companies may choose to omit specific information from a database in order to create a specific perception regarding finances, such as highlighting the amount of generated revenue for a given period, but choosing to not enter data that relates to the amount of collected revenue for the same period. In this type of dirty data, the information that is presented is accurate as far as it goes, but is considered incomplete.

With some types of dirty data, the decision may be to not take the time and effort to make corrections. This is common when the incorrect data does not have any impact on the ability of the business to function properly, or presents no potential for causing any great distress. This means that just about any entity that maintains some type of database probably has at least a little dirty data interspersed with other information that is current and accurate.

Data cleansing

Data cleansing, also known as data scrubbing, is the process of ensuring that a set of data is correct and accurate. During data cleansing, records are checked for accuracy and consistency, and either corrected, or deleted as necessary. Data cleansing can occur within a single set of records, or between multiple sets of data which need to be merged, or which will work together.

At its most simple form, data cleansing involves a person or persons reading through a set of records and verifying their accuracy. Typos and spelling errors are corrected, mislabeled data is properly labeled and filed, and incomplete or missing entries are completed. Data cleansing operations often purge out of date or unrecoverable records, so that they do not take up space and cause inefficient operations.

In more complex operations, data cleansing can be performed by computer programs. These data cleansing programs can check the data with a variety of rules and procedures decided upon by the user. A data cleansing program could be set to delete all records which have not been updated within the last five years, correct any misspelled words, and delete any duplicate copies. A more complex data cleansing program might be able to fill in a missing city based on a correct zip code, or change the prices of all items in a database to Euros instead of US Dollars.

Data cleansing is very important to the efficiency of any data dependent business. If some of the clients within a database do not have accurate phone numbers, your employees cannot easily contact them. If your clients' email addresses are not formatted correctly, an automated email system would be unable to send out the latest coupons and special deals. The job of data cleansing is to insure that the data within a system is correct, so that the system is able to use the data. Inaccurate or incomplete records are not much use to anyone.

Whenever two systems of data need to work together, data cleansing is even more important. If a company has two branches, which might work with many of the same customers, not only does the data in each branch need to be complete and accurate, but the two branches need to have matching data. If a customer updates her phone number with one branch, the data at the other branch needs to be updated with the same information to insure the highest efficiency. Data cleansing works not only to make sure that data is accurate, but also that it is consistent between different records.

Any time you are storing a lot of data, errors are bound to creep into the system. The goal of data cleansing is to minimize these errors, and to make the data as useful and meaningful as possible. Without regular data cleansing, mistakes and errors can add up, leading to less efficient work and more complications down the road.