More about data integrity and compliance

This reading illustrates the importance of data integrity using an example of a global company's data. Definitions of terms that are relevant to data integrity will be provided at the end.

Scenario: calendar dates for a global company

Calendar dates are represented in a lot of different short forms. Depending on where you live, a different format might be used.

- In some countries,**12/10/20** (DD/MM/YY) stands for October 12, 2020.
- In other countries, the national standard is YYYY-MM-DD so October 12, 2020 becomes **2020-10-12**.
- In the United States, (MM/DD/YY) is the accepted format so October 12, 2020 is going to be **10/12/20**.

Now, think about what would happen if you were working as a data analyst for a global company and didn't check date formats. Well, your data integrity would probably be questionable. Any analysis of the data would be inaccurate. Imagine ordering extra inventory for December when it was actually needed in October!

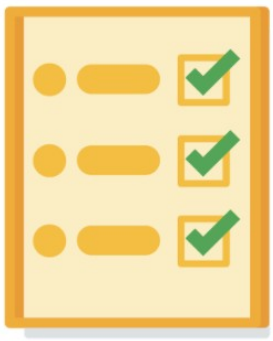
A good analysis depends on the integrity of the data, and data integrity usually depends on using a common format. So it is important to double-check how dates are formatted to make sure what you think is December 10, 2020 isn't really October 12, 2020, and vice versa.

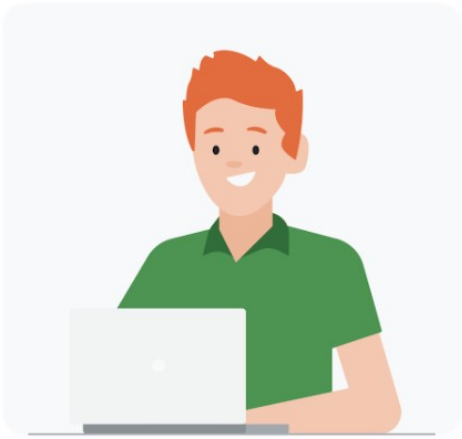
Here are some other things to watch out for:

- **Data replication compromising data integrity:** Continuing with the example, imagine you ask your international counterparts to verify dates and stick to one format. One analyst copies a large dataset to check the dates. But because of memory issues, only part of the dataset is actually copied. The analyst would be verifying and standardizing incomplete data. That partial dataset would be certified as compliant but the full dataset would still contain dates that weren't verified. Two versions of a dataset can introduce inconsistent results. A final audit of results would be essential to reveal what happened and correct all dates.
- **Data transfer compromising data integrity:** Another analyst checks the dates in a spreadsheet and chooses to import the validated and standardized data back to the database. But suppose the date field from the spreadsheet was incorrectly classified as a text field during the data import (transfer) process. Now some of the dates in the database are stored as text strings. At this point, the data needs to be cleaned to restore its integrity.
- **Data manipulation compromising data integrity:** When checking dates, another analyst notices what appears to be a duplicate record in the database and removes it. But it turns out that the analyst removed a unique record for a company's subsidiary and not a duplicate record for the company. Your dataset is now missing data and the data must be restored for completeness.

Conclusion

Fortunately, with a standard date format and compliance by all people and systems that work with the data, data integrity can be maintained. But no matter where your data comes from, always be sure to check that it is valid, complete, and clean before you begin any analysis.





Reference: Data constraints and examples

As you progress in your data journey, you'll come across many types of data constraints (or criteria that determine validity). The table below offers definitions and examples of data constraint terms you might come across.

Data constraint	Definition	Examples
Data type	Values must be of a certain type: date, number, percentage, Boolean, etc.	If the data type is a date, a single number like 30 would fail the constraint and be invalid
Data range	Values must fall between predefined maximum and minimum values	If the data range is 10-20, a value of 30 would fail the constraint and be invalid
Mandatory	Values can’t be left blank or empty	If age is mandatory, that value must be filled in
Unique	Values can’t have a duplicate	Two people can’t have the same mobile phone number within the same service area
Regular expression (regex) patterns	Values must match a prescribed pattern	A phone number must match ###-###-#### (no other characters allowed)

Data constraint	Definition	Examples
Cross-field validation	Certain conditions for multiple fields must be satisfied	Values are percentages and values from multiple fields must add up to 100%
Primary-key	(Databases only) value must be unique per column	A database table can’t have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program.
Set-membership	(Databases only) values for a column must come from a set of discrete values	Value for a column must be set to Yes, No, or Not Applicable
Foreign-key	(Databases only) values for a column must be unique values coming from a column in another table	In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table
Accuracy	The degree to which the data conforms to the actual entity being measured or described	If values for zip codes are validated by street location, the accuracy of the data goes up.
Completeness	The degree to which the data contains all desired components or measures	If data for personal profiles required hair and eye color, and both are collected, the data is complete.
Consistency	The degree to which the data is repeatable from different points of entry or collection	If a customer has the same address in the sales and repair databases, the data is consistent.