Data Cleaning and Quality

BERCS1102016-V001000

0:00 / 9:03 1.0x

Start of transcript. Skip to the end.

SPEAKER: Data cleaning helps deal with several issues-- for example, if you have missing data, if I have one data set that has humidity information and another data set that I'm trying to integrate with that does not have that information.

Download video Download transcript .srt

Data Cleaning

(1/1 point)
Data cleaning helps to deal with:

- [x] Unit mismatch ✔
- [x] Entity resolution ✔
- [ ] Relational models
- [x] Samples corrupted by a distortion process ✔
- [ ] Old data

✔

Note: Make sure you select all of the correct options—there may be more than one!

EXPLANATION

Data cleaning helps with unit mismatch (e.g., pounds versus grams), entity resolution (are two records in two different datasets the same item), and the effects of distortion on data.

Data Quality Problems

(1/1 point)
Which of the following are data quality problems?

- Conversions in complex pipelines can mess up data
- Combining multiple datasets can result in errrors
- Data degrades in accuracy or loses value over time
- All of the above ✔

EXPLANATION

All of these are data quality problems that you can encounter. For example, in the lab exercises, you have seen how errors in transformations can result in incorrect answers. Similarly combining multiple datasets can introduce errors. Even data at rest can lose accuracy/value over time. For example, consider geospatial mapping data. As new houses are added, or roads are built or changed, the mapping data loses accuracy and value - the data is static but the world is not.

Data quality is provided through data cleaning, also known as data wrangling. There are many tools for cleaning data, including:

- Stanford's Data Wrangler is an interactive tool for data cleaning and transformation.
- Google's Open Refine tool is a free, open source power tool for working with messy data and improving it.
- Talend provides several open source data quality tools.

These are the slides from Ted Johnson's SIGMOD 2003 Tutorial.
