Bookmarks

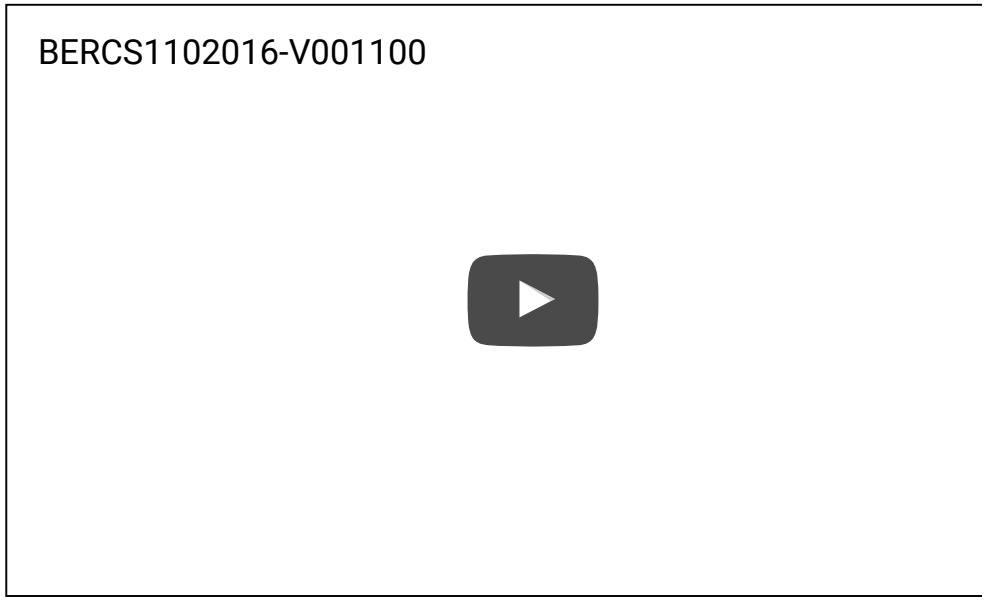
- Week 1 - Big Data and Data Science
- **Week 2 - Performing Data Science**

Lecture 2: Performing Data Science and Preparing Data
Quizzes

Lab 2 - Movie Rating Prediction using Alternating Least Squares
Lab due Sep 13, 2016 at 04:30 IST

Lab 2 Quiz Questions
Quizzes

Data Quality Constraints and Data Integration



Start of transcript. Skip to the end.

SPEAKER: We can use data quality constraints to capture

many data quality problems.

For example, we can use a schema static constraints,

such as NULLS not being allowed or specifying constraints

on field domains or foreign key constraints.

Download video Download transcript .srt

Data Quality Metrics

(1/1 point)

In the CS105x Lab 2, Apache web server log analysis, why did we include a check for log lines that failed to be correctly parsed?

- To count the number of lines that were correctly parsed
- To dynamically measure data quality ✔
- To make the lab hard to complete
- To teach you about regular expressions

EXPLANATION

While the check for log lines did make it more difficult to complete the lab and also taught you about regular expressions, the reason for including the check for failed log lines was so that we could meaure the quality of the data. The lines that failed to parse correctly represented dirty data, and without such a check, we would have siliently ignored those lines.