Bookmark

Structured Data and the Structured Query Language

BERCS1052016-V001200

Download video Download transcript .srt

Here are links to descriptions of the large databases discussed in this lecture video segment:

- US Internal Revenue Service: 150 Terabytes
- Australian Bureau of Stats: 250 Terabytes
- AT&T call records: 312 Terabytes
- eBay database: 1.4 Petabytes
- Yahoo click data: 2 Petabytes

Here are some resources for the Structured Query Language and Spark SQL:

- edX SQL course
- Wikipedia SQL page
- SQL Tutorial 1
- SQL Tutorial 2
- SQL Tutorial 3
- SQL Tutorial 4
- Short SQL course

- Web guide to SQL
- Wikipedia page on SQL Join clauses
- Visual explanation of SQL Joins
- SQL Join Tutorial
- Spark SQL & DataFrames
- Spark SQL, DataFrames and Datasets Guide
- Spark SQL compatibility with Apache Hive

Note that a key difference between SQL and Spark SQL is that Spark SQL does not support DELETE. In SQL, DELETE allows you to delete rows from a table. Remember that DataFrames are immutable which means they cannot be changed once you create them. Instead of modifying a DataFrame, you must create a new DataFrame from that DataFrame. You might think that this would make DataFrames very expensive and cause them to take up a lot of memory, but Spark very efficiently handles the implementation of creating new DataFrames from existing ones.

Relational Databases

(1/1 point)
What kind of data is stored in a relational database?

- Unstructured data
- Semi-structured data

- (●) Structured data ✔
- ( ) Random data

EXPLANATION

Relational databases are used to store structured data. We can use DataFrames to store unstructured, semi-structured, and structured data.

Database Advantages

(1/1 point)
Which of the following are NOT advantages of databases:

- [ ] They have a well-defined structure
- [ ] They use indicies for high performance
- [x] They have a rigid structure ✔
- [ ] They guarantee data consistency

- [x] They have poor support for sparse data ✔

✔

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

EXPLANATION

The rigid structure of databases means that they do not work well with sparse, semi-structured, or unstructured data.