Spark SQL

Objective

Use Spark SQL *programmatically*

Objective

Use Spark as a "big data" database

SQL

Universal standard for accessing structured data

Abstraction over DataFrames for engineers familiar with databases

Supported in Spark

- progammatically in expressions
- in the Spark shell

Spark SQL

- has the concept of "database", "table", "view"
- allows accessing DataFrames as tables

DataFrame vs table

- identical in terms of storage and partitioning
- DataFrames can be processed programmatically, tables in SQL

Spark Tables

Managed

- Spark is in charge of the metadata + the data
- if you drop the table, you lose the data

External (unmanaged)

- Spark is in charge of the metadata only
- if you drop the table, you keep the data

Spark rocks