

Spark SQL

Introduction

- ▶ **Spark SQL** is a module for structured data processing, upon which the higher-level structured APIs are built.
- ▶ Lets us query data using JDBC/ODBC connectors from external business intelligence sources such as Tableau, Power BI, Snowflake, or from RDBMS such as MySQL and PostgreSQL.
- ▶ Offers an interactive shell to issue SQL queries on our structured data.
 - ▶ We can interact with the SQL interface via a programming language and the data set will be returned as a **Dataset/DataFrame**.
- ▶ When computing a result the same execution engine is used, independent of which API/language you are using to express the computation. This unification means that developers can easily switch back and forth.

Examples

- ▶ Example 1: Using SQL on DataFrames
- ▶ Example 2: Manipulating databases directly using SQL in Spark
- ▶ Example 3: Java and SQL in Spark