# Spark SQL

...

# What is Spark SQL?

# Spark SQL

Support for DataFrames

- Spark's interface for working with structured and semi-structured data
- Spark SQL makes it both easier and more efficient to load and query such data

# Structure of Data

- What is structured data?
  - Any data that has a schema — a known set of fields for each record
- What is semi-structured data?
  - Any data that has a dynamic schema - the set/hierarchy of fields is open to change
- What is unstructured data?
  - Any data that cannot be described with a schema

# Spark SQL Features

Loading data

- It can load data from a variety of structured sources
  - JSON
  - Hive
  - ...

# Spark SQL Features

Doing SQL

- It lets you query the data using SQL
  - inside a Spark program
  - from external tools that connect to Spark SQL through standard database connectors (JDBC/ODBC)
    - such as Tableau

# Spark SQL Features

Abstractions for data manipulation

- Spark SQL provides rich integration between SQL and regular Python/Java/Scala code, including
  - the ability to join RDDs and SQL tables
  - expose custom functions in SQL
  - many jobs become easier to write using this combination

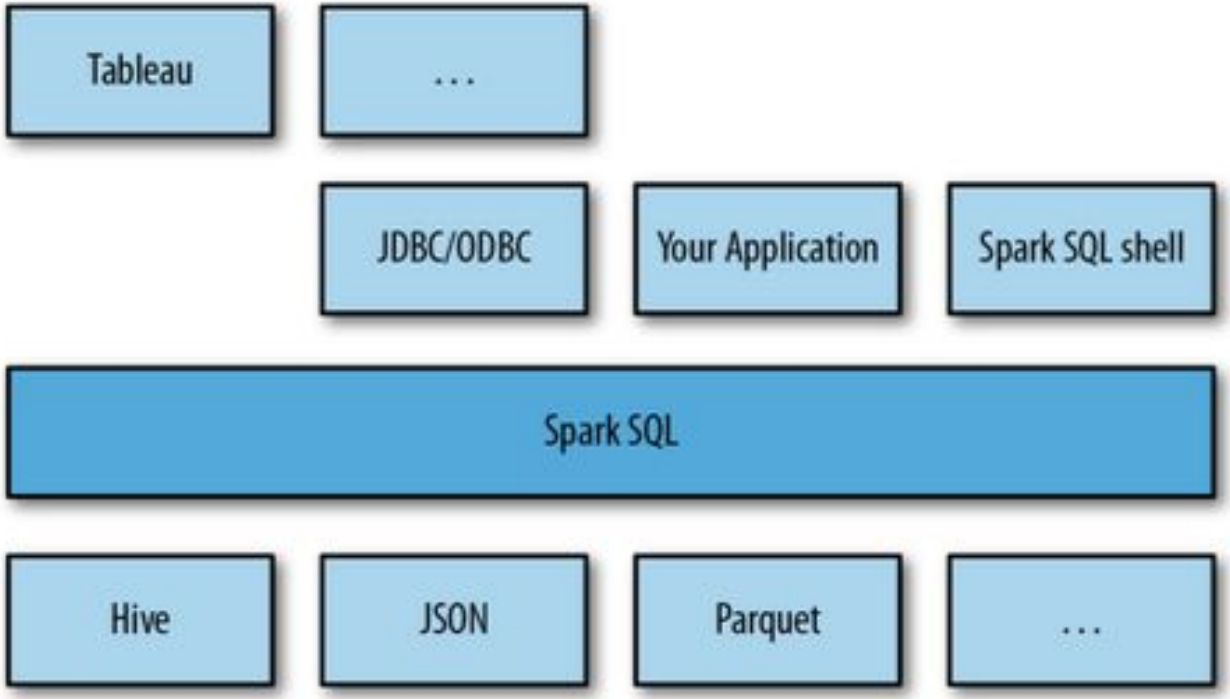
Tableau
...
JDBC/ODBC
Your Application
Spark SQL shell
Spark SQL
Hive
JSON
Parquet
...

# So how does Spark do all these wonderful things?

# SchemaRDD

- Spark SQL provides a special type of RDD called SchemaRDD
- A SchemaRDD is an RDD of Row objects, each representing a record
- A SchemaRDD also knows the schema (i.e., data fields) of its rows

# SchemaRDD

Features

- While SchemaRDDs look like regular RDDs, internally they store data in a more efficient manner, taking advantage of their schema
- They provide new operations not available on RDDs, such as the ability to run SQL queries

# So how do they compare to DataFrames in other languages?

# Creating SchemaRDD

- SchemaRDDs can be created from
  - external data sources
  - the results of queries
  - regular RDDs

# Getting Started

# The lib

- Spark Core itself doesn’t include Spark SQL
- Spark SQL can be built with or without Apache Hive, the Hadoop SQL engine
- If you download Spark in binary form, it should already be built with Hive support

# SQLContext

- HiveContext provides access to HiveQL and other Hive-dependent functionality
- SQLContext provides a subset of the Spark SQL support that does not depend on Hive

# SparkSession

- Starting with pySpark