Spark SQL

МОДУЛЬ 5

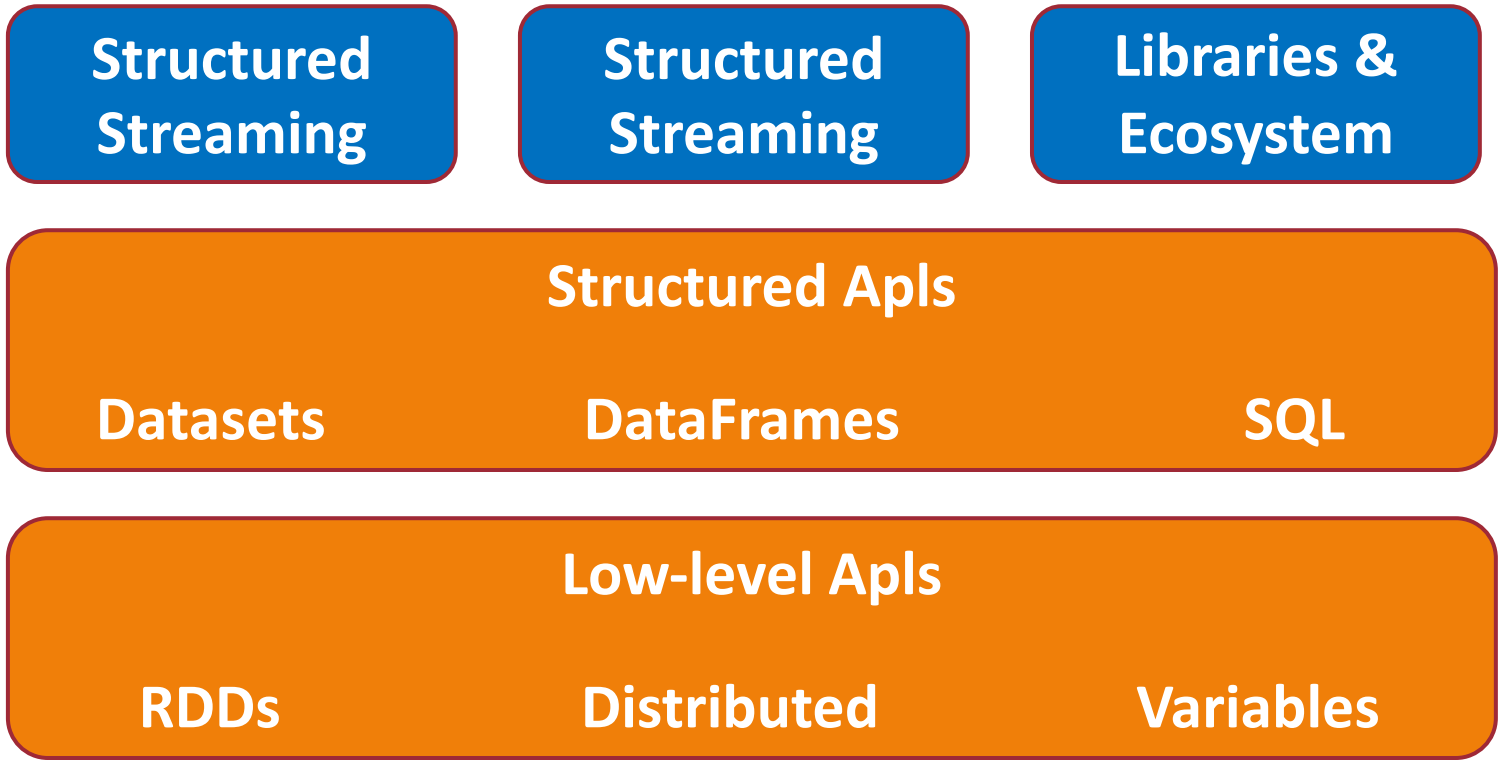
ШКОЛА БОЛЬШИХ ДАННЫХ

Spark SQL – helicopter view

независимая реализация ANSI SQL интерфейса к данным

- ✓ работа с данными с помощью **SQL**
- ✓ создание **таблиц** и **view**
- ✓ поддержка **метаданных** (собственных или в Hive)
- ✓ SQL компилируется в манипуляции с **RDD** (как и Structured API)

Что входит в Spark



Предоставляемые возможности

- **Spark SQL cli**: интерактивная работа
- **API** для работы с данными на SQL
- **Thrift JDBC/ODBC server** для внешнего доступа

Поддержка мета данных

- **“Catalog”**: объект для работы с метаданными
- **“Warehouse”**: директория, в которой создаются таблицы
  - ✓ как правило, совмещена с Hive
- Поддержка **“Hive QL”** (включая DDL)
- Манипуляции с **database, table, view**

Spark SQL API

- метод “**sql**” возвращает dataframe
- объект “**Catalog**” для работы с метаданными
- функции “**sql.functions**” на уровне колонок

SQL vs Dataframe

- находятся **на одном уровне**
- компилируются в RDD
- полная **взаимозаменяемость**
  - ✓ SQL -> DF: SQL порождает dataframe
  - ✓ DF -> SQL: dataframe можно сделать view

Complex data types

Поддерживаются

- **“struct”**: иерархические данные
- **“array”**: списки
- **“map”**: ключ-значение

Вопросы?

ШКОЛА БОЛЬШИХ ДАННЫХ