天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

Spark初级课程

讲师：平常心

Spark SQL

Spark SQL是Apache Spark的一个工作模块，用于结构化数据处理。Spark SQL的核心入口是SQLContext对象，子类有HiveContext，其中HiveContext默认实现的是HiveQL,通过SQLContext创建DataFrame 或者DataSet，执行sql语句或者API操作。

集成:

和spark程序无缝混合，Spark SQL可以查询Spark 程序中的结构化数据，使用SQL或者通过Java,Scala,Python和R的DataFrame API操作。

统一的数据访问:

使用同样的方式连接到数据源，DataFrame和SQL来访问各种数据源，如Hive,Avro,Parquet,ORC,JSON和JDBC。

兼容Hive:

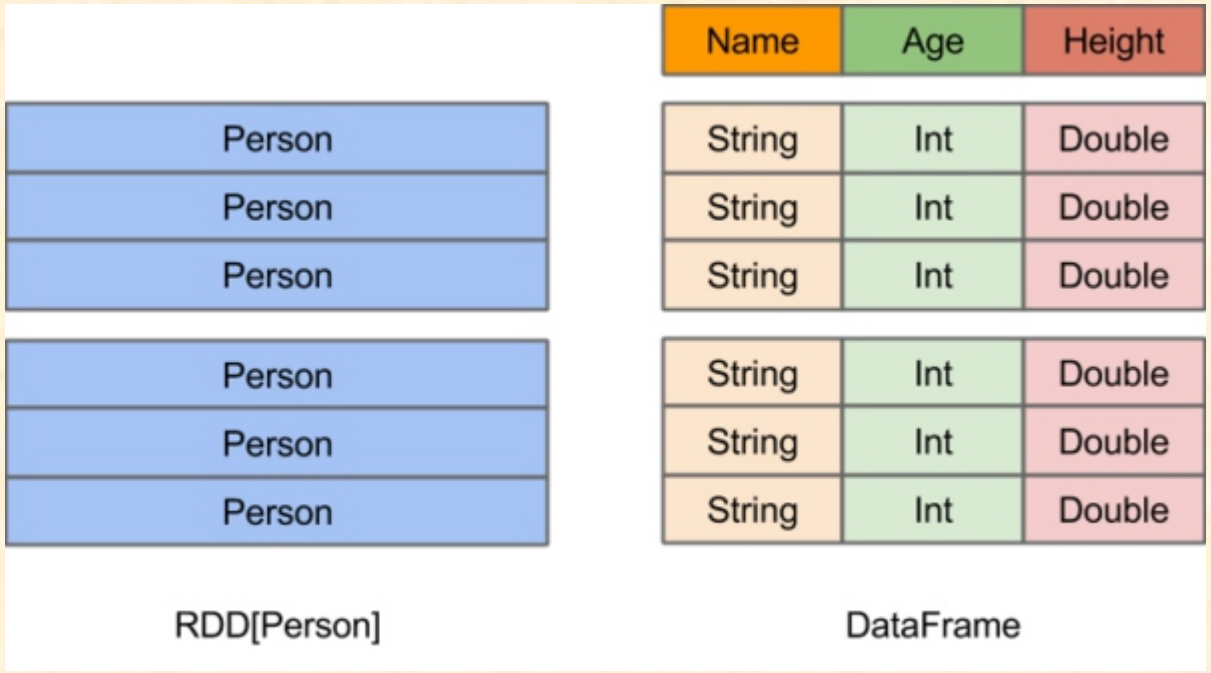
在数据源上直接运行未修改的Hive查询，Spark SQL重用Hive端和元数据，兼容现有的Hive数据，查询和UDF，和Hive并行安装。

标准连接:

通过JDBC或者ODBC连接，采用服务模式为商业工具提供行业标准JDBC和ODBC连接。

DataFrame

DataFrame 就是列形式组织的分布式数据集合。概念上等同于关系型数据库中的表类似，包含列名称和列的数据类型，但是在底层更加丰富的优化。DateFrame可以通过很多数据源进行构建，包括结构化数据文件，Hive中的表，外部数据库或者已经存在的RDD。



DataSet

DataSet在1.6版本还是实验的数据集合。提供了RDD（强类型，使用强大的lambda函数的功能）的好处，并具有Spark SQL优化的执行引擎的优点。可以从JVM对象构建数据集，然后使用功能转换（map，flatMap，filter等）进行操作。

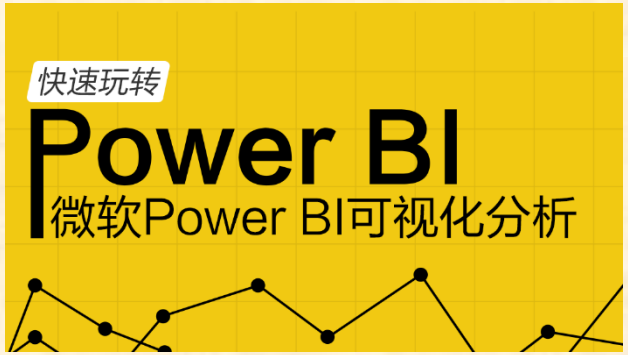
DataFrame由“数据集”行表示，DataFrame只是Dataset [Row]的一个类型别名。通常将将Scala / Java数据集的Rows称为DataFrames。

DataFrame和DataSet的创建和操作

1.通过json文件创建
2.通过parquet文件创建
3.通过sql的表创建
4.通过jdbc创建
...


SparkSQL

更多商业智能BI和大数据精品视频尽在 www.hellobi.com



BI、商业智能
数据挖掘 大数据
数据分析
R Python
机器学习
Tableau
QLIKVIEW
Hive Hadoop
BIWORK
BAO胖子 seng
贝克汉姆
曹浩

天善智能
TIANSHAN SOFT