Spark的RDD、DataFrame、DataSet、DataStream区别?

可回答：1）Spark RDD和DataSet的关系；2）Spark中RDD、DataFrame介绍

参考答案：

RDD

RDD是一种弹性分布式数据集，是一种只读分区数据。它是spark的基础数据结构，具有内存计算能力、数据容错性以及数据不可修改特性。

DataFrame

DataFrame是一种以RDD为基础的分布式数据集，类似于传统数据库中的二维表格。DataFrame与RDD的主要区别在于，前者带有schema元信息，即DataFrame所表示的二维表数据集的每一列都带有名称和类型。

DataSet

Dataset是DataFrame的扩展，它提供了类型安全，面向对象的编程接口。也就是说DataFrame是Dataset的一种特殊形式。

1）是Dataframe API的一个扩展，是Spark最新的数据抽象。它提供了RDD的优势（强类型，使用强大的lambda函数的能力）以及Spark SQL优化执行引擎的优点。

2）用户友好的API风格，既具有类型安全检查也具有Dataframe的查询优化特性。

3）Dataset支持编解码器，当需要访问非堆上的数据时可以避免反序列化整个对象，提高了效率。

4）样例类被用来在Dataset中定义数据的结构信息，样例类中每个属性的名称直接映射到DataSet中的字段名称。

5） Dataframe是Dataset的特列，DataFrame=Dataset[Row]，所以可以通过as方法将Dataframe转换为Dataset。Row是一个类型，跟Car、Person这些的类型一样，所有的表结构信息我都用Row来表示。

6）DataSet是强类型的。比如可以有Dataset[Car]，Dataset[Person]。

7）DataFrame只是知道字段，但是不知道字段的类型，所以在执行这些操作的时候是没办法在编译的时候检查是否类型失败的，比如你可以对一个String进行减法操作，在执行的时候才报错，而DataSet不仅仅知道字段，而且知道字段类型，所以有更严格的错误检查。就跟JSON对象和类对象之间的类比。

DataStream

DStream是随时间推移而收到的数据的序列。在内部，每个时间区间收到的数据都作为RDD存在，而DStream是由这些RDD所组成的序列（因此得名“离散化”）。所以简单来讲，DStream就是对RDD在实时数据处理场景的一种封装。

蓦然 送你一张星球优惠券

「旧时光大数据」

立减

¥ 40

新人立减券

2023/12/31 12:00 后失效

知识星球

长按扫码领取优惠