## 第二周总结及对DataFrame,SQL的分析比较

本周通过对spark 的继续学习,熟悉了RDD的操作,理解了分布式数据集的概念,并能熟练对RDD操作,在进行操作的时候能合理的使用转换和动作,使程序更加的优化,具体可参考源码scoreList2.scala.

除此之外学习了DataFrame和新的模块spark SQL,两着都是对数据进行处理,有相似的地方,也有不同的地方.DataFrame是一个由命名列组成的分布式数据集。它从概念上讲相当于关系数据库里的一张表，或R/Python里的数据框架，但内部有很多优化。而Spark SQL是Spark中一个处理结构化数据的模块。它提供了DataFrame的编程抽象，并能作为一个分布式的SQL查询引擎。DataFrame能从已有的RDD,数据库,或者其它数据源中读取数据,创建起来非常方便.从使用上来说两者对数据的处理上很多效果都相似,不过spark SQL支持SQL查询语句,对有SQL有基础的学者来说更容易上手,操作起来更加的方便.DataFrame对数据的操作也比较成熟,有很多实用的API,除了有Basic DataFrame Functions外,还有可供查询处理的Language Integrated Queries方法,其中一些方法甚至比SQL更好用,对SQL的一些查询语句,其中也能找到替代方法,而且它 的方法和RDD中一些方法相似度很高,理解起来也不难,对于这两者来说我们并不是非的选其中一种,两者都有互相转换的方法,不同的情况下可以选择不同的方式,不过SQL在查询的时候耗时似乎比DataFrame更高一些,我测试的最大一个数据有20M,在查询的时候SQL的时间比DataFrame高出2-3倍,这个也可能是我们数据太小的原因,等以后再用更大的数据来测试下.

对SQL和DataFrame的测试操作可以参考附件源码DataFrameTest.scala和SQLTest2.scala