北风网 IBEIFENG.COM

《快学 Spark 2.0》

第 32 讲、Streaming 读 Kafka 开发 WordCount 案例

讲师-Cloudy

1.1 DStream 的两种模式

DStream 的转化操作可以分为无状态（stateless）和有状态（stateful）两种。
• 在无状态转化操作中，每个批次的处理不依赖于之前批次的数据。
如 map()、filter()、reduceByKey() 等，都是无状态转化操作。
场景：写库类操作

• 相对地，有状态转化操作需要使用之前批次的数据或者是中间结果来计算当前批次的数据。有状态转化操作包括基于滑动窗口的转化操作和追踪状态变化的转化操作。
场景：滑动窗口、汇总、去重类操作

无状态的操作没啥特别要说的，后边看案例即可。

1.2 有状态的操作

有状态转化操作需要在你的 StreamingContext 中打开检查点机制来确保容错性，
设置检查点 ssc.checkpoint("hdfs://...") ，存储有状态操作所需要的中间过程数据，
也可以使用本地路径（例如/tmp）取代 HDFS。

对 kafka 来讲，groupid 的作用是什么？

多个作业同时消费同一个 topic 时：
1、每个作业拿到完整数据，计算互不干扰；
2、每个作业拿到一部分数据，相当于进行了负载均衡；

当多个作业 groupid 相同时，属于情况 2；
否则属于情况 1.

setMaster("local[1]")　*//核数至少给 2，如果只是 1 的话，无法进行数据计算。*

控制台会报如下提示：
WARN StreamingContext: spark.master should be set as local[n], n > 1 in local mode if you have receivers to get data, otherwise Spark jobs will not get resources to process the received data.

Cloudy 讲师简介：

15 年工作经验，近 10 年一直从事大数据技术领域，经历国内多家知名互联网企业，现就职国内一知名电商任数据部首席架构师。

主导公司 Spark 项目从无到有的建设，完成公司 Hive 到 Spark 的转变，当前 Spark 已承载数据部 85%的离线作业和 30%实时作业。Spark2.0 第一批研究者。