实时数据生产源 => 消息中间键(kafka) => 处理

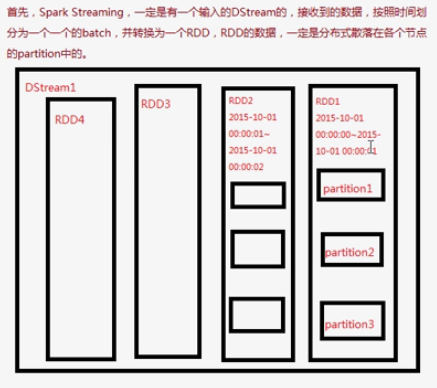
Kafka是作为实时数据处理时的一种缓冲

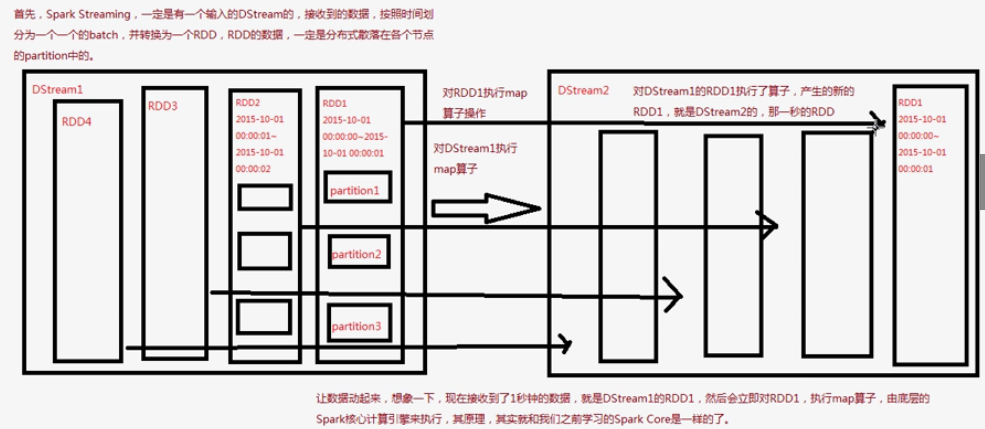
TCPsocket flume Kafka

流处理的常用操作：map、reduce、join、window

处理后的结果可以写道database dashboard hdfs hbase

核心的处理对象Dstream，是discretized stream的简称，是batch of RDD构成的：





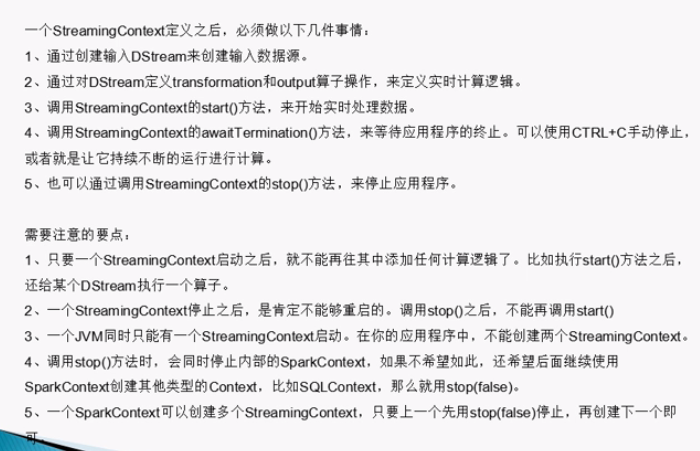
Spark和storm的对比



Storm适用于精准、延迟低的，计算资源相对不足，有些时段高峰明显，需要调节并行度的。仅仅是流处理，不会涉及到SQL交互查询、transformation

Streaming 高吞吐量，并且可能会和一些交互查询、机器学习等结合。

Spark streaming context



代码实操

Scm:stream的context

Scm.socketTextStream(“地址”, 端口) // 能够获取socket数据

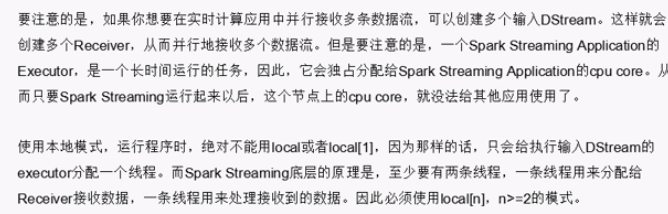
Spark streaming的DStream和Receiver

Dstream是从其他服务（比如nc，socket）等接受数据，除了文件数据源外，所有的数据源都绑定一个receiver。

支持的数据源：

1. 基本数据源：文本、socket、akka
2. 高级数据源：
3. 自定义数据源：

一个应用会独占executor，一个ssc应用至少分配给他receiver数目 + 1个core



KafkaUtil提供了对kafka发送的数据进行接受的API。

upStateByKey

