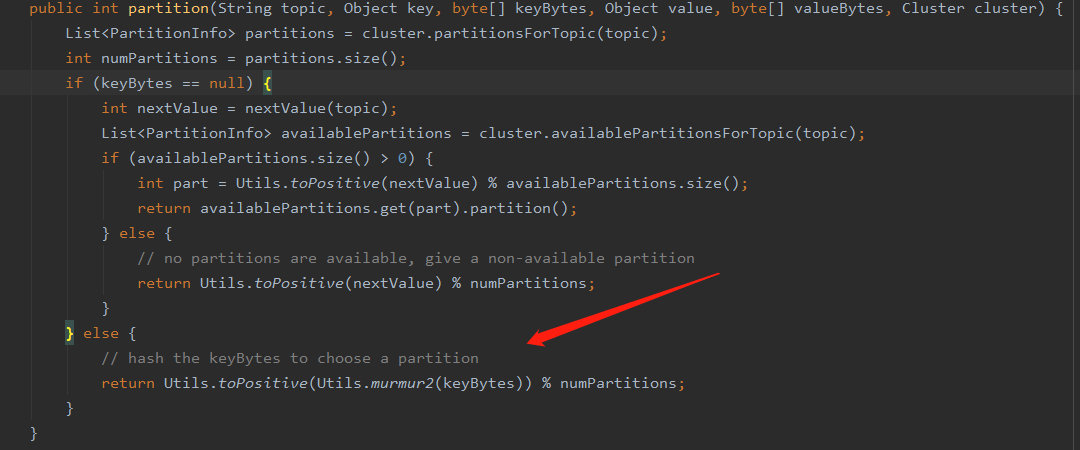
1. 关于分区



由上图箭头指示，再向kafka发送数据时，指定了key会根据key的hash值 % 分区个数，选择分区，后续不修改topic的partition个数就能保证同一个key 进入 同一个分区。

1. 关于Spark Streaming 读取kafka

读取kafka有两种方式一种reciver模式，一种direct模式，direct模式读取kafka性能更高且更安全， direct模式读取kafka数据是一个kafka的partition 对应 一个 Spark 的task，这样就保证了同一个key的数据再同一个task运行。

建议再保证不了数据不偏移的情况下 尽量kafka的partition个数 > spark 分配的core个数。数据不偏移两者相等。

1. 关于程序

这是读取hdfs文件数据，根据车牌号进行分区

主要逻辑是 发送者的key 和 value， key 和value根据应用场景修改。

参数

1. Hdfs文件路径
2. Kafka broker 的ip和端口
3. 写入的主题名字
4. 一次发送个数
5. 发送次数的时间间隔（ms）（比如一次发送10条，这里间隔为500，就是1s发送20条数据 ）
6. 启动的producer个数， 注意这里rdd的foreachPartition算子， producer封装在这个算子里， producer 并行个数 与 Spark 分配的 core相关。