Spark stream的核心原理是很么？

Q:先接收实时流流的数据，并根据一定的时间间隔分成一批批的数据，每一批数据对应一个rdd实例，所以整个流数据可以看成很多个rdd数据，本质还是spark 核心处理。这个里面抛开spark核心那么需要理解就是时间间隔以及在spark streaming的spark核心。

Q：从rdd的角度，有一个叫做DStream需要了解，这是一个封装了时间和rdd的Map，可以说是spark流处理的rdd，相对应与DAG，spark streaming也有类似的概念就是DStreamGraph，基本原理跟spark核心大致是一致的，因为spark归根揭底就是spark core之上的封装与应用。

A:批处理间隔，滑动间隔，窗口间隔

Q:只有在涉及到窗口操作的时候才会考虑到滑动间隔和窗口间隔。首先窗口操作并不是什么新鲜的东西，他也是响应某些需求而产生的（工具flink似乎是专门做窗口操作）比如网站记录，每隔一小时计算最近两个小时的pv量，还有一种业务场景，我们可以理解为每个多长时间统计多长时间内的数据。这个里面有一个关系是窗口间隔和滑动间隔需要是批处理间隔的倍数。为什么呢？因为流处理是batch触发的，不是他的倍数怎么触发窗口计算。比如，批处理间隔时间5

S，而滑动间隔6秒，很明显这个6秒这个时刻是达不到的！其实可以看下窗口处理函数的源代码。