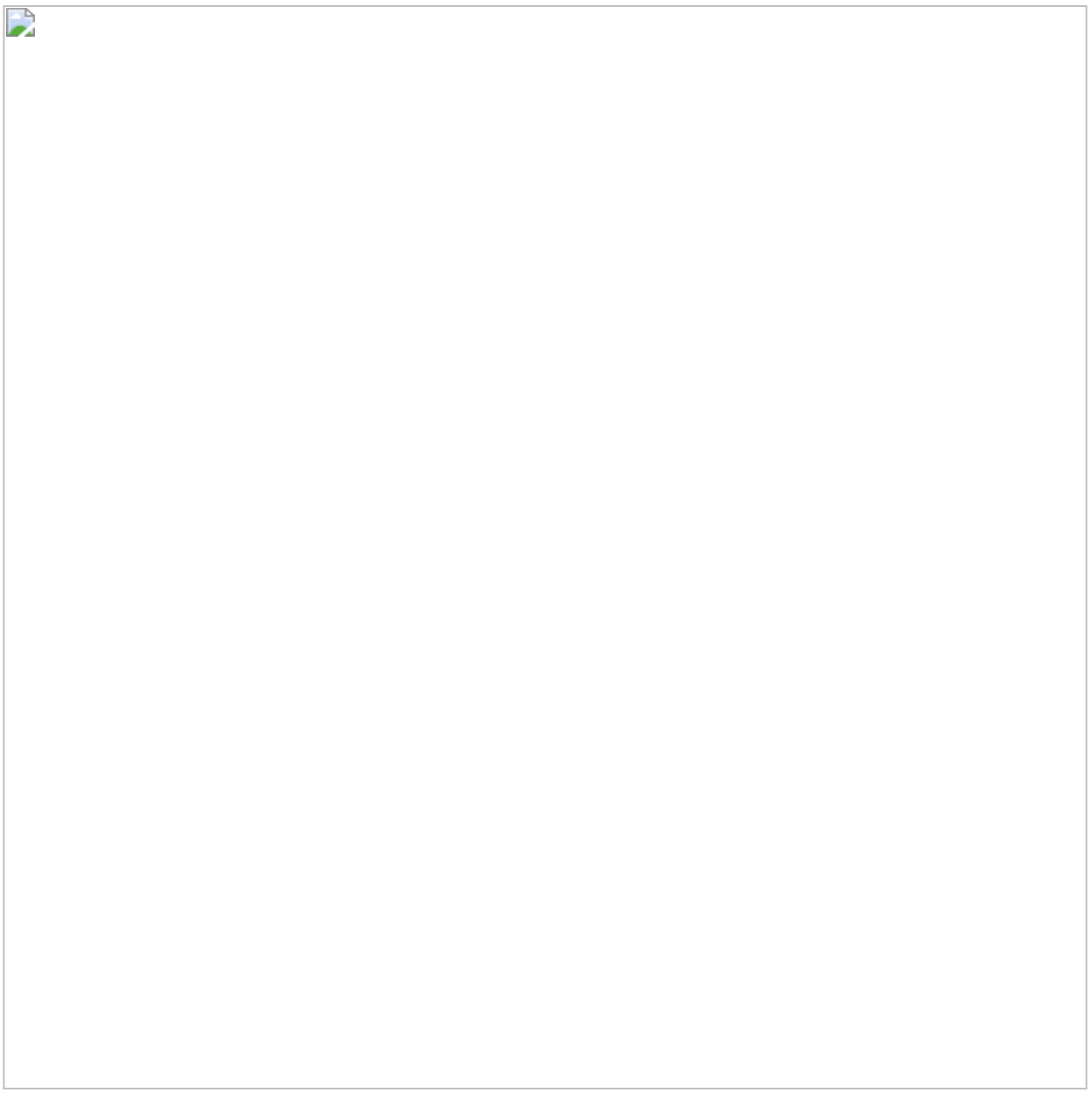
谈谈spark中对RDD的认识。

RDD：基于内存的集群计算容错抽象。关键应该是“容错”　　RDD是数据流模型的

> 弹性分布式数据集（RDD）有：
> Spark RDD
> Spark Streaming RDD
> Spark SQL RDD
> MLLib RDD
> GraphX RDD

RDD与分布式共享内存：



与DSM相比，RDD模型有两个好处。第一，对于RDD中的批量操作，运行时将根据数据存放的位置来调度任务，从而提高性能。第二，对于基于扫描的操作，如果内存不足以缓存整个RDD，就进行部分缓存。把内存放不下的分区存储到磁盘上，此时性能与现有的数据流系统差不多

RDD适用于具有批量转换需求的应用