输入：

–$\boldsymbol{D}_{\mathrm{HDFS}}$：HDFS 存储的数据块文件；

–P：HDFS 文件 RDD partition 的个数；

–Q：输出数据块个数.

计算方法：

$\boldsymbol{D}_{\mathrm{RDD}}$←把 $\boldsymbol{D}_{\mathrm{HDFS}}$ 读入到 P 个 Spark 内存中的 RDD

for 所有 $\boldsymbol{D}_{\mathrm{RDD}}$ 数据块 i

K=Seq(1to n)

随机打乱 K;

end for

根据 K 值生成新 RDD 键值对 RDD(K, V) ;

使用哈希函数 HashPartitioner(Q)分发当前键值对 RDD(K, V)到对应的存储节点；

输出：

–$\boldsymbol{D}_{\mathrm{RSP\text{-}HDFS}}$：将新的 RDD(K, V)回存到 HDFS 存储的数据块文件.

基于 Spark 的随机样本块生成程序伪代码

Random sample partition based on Spark