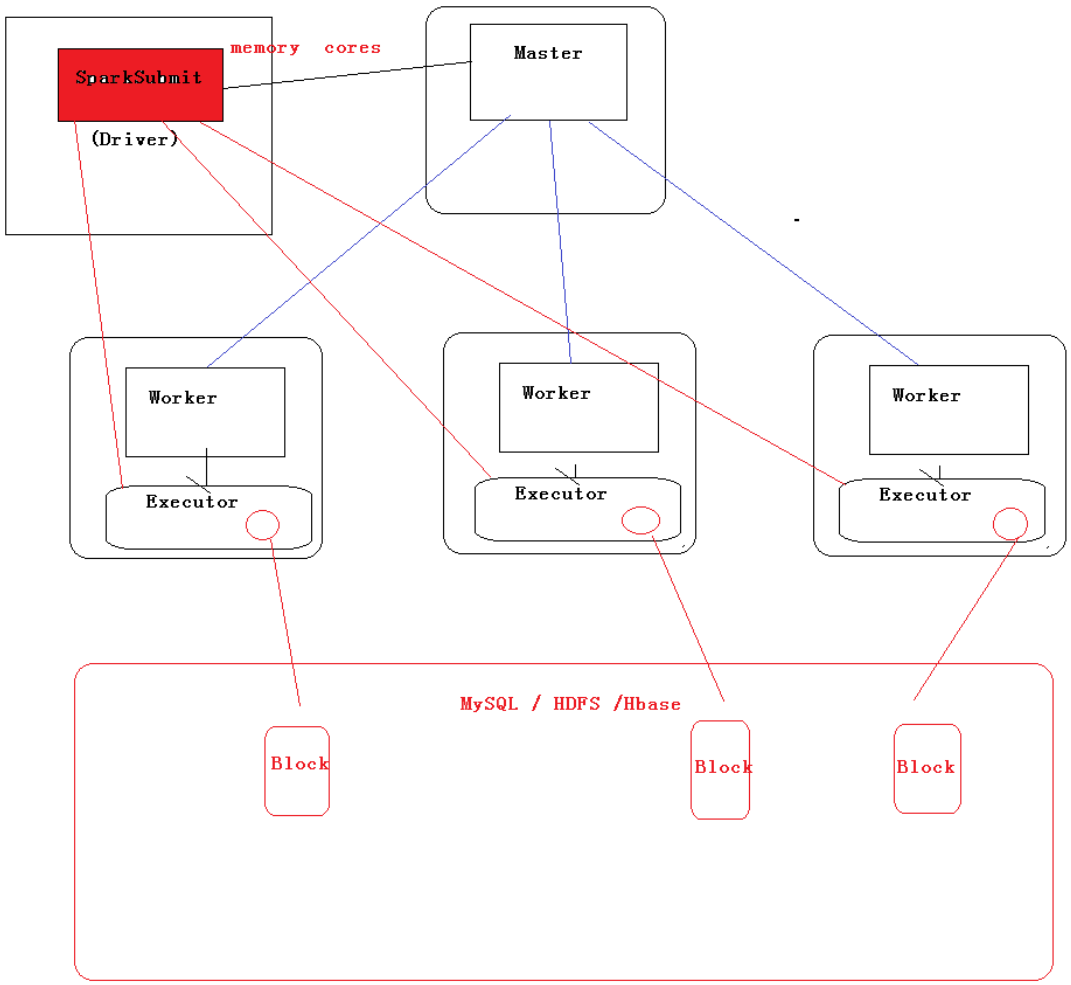
1. Spark 任务执行过程

Spark 程序在 Executor 进程中执行，在 Driver 端由 sparksubmit 进程提交任务，且 driver 端不存在多线程问题。



说明：

（1）集群启动时，Worker 节点向 Master 节点汇报自身资源情况；

（2）sparksubmit 进程启动时创建 SparkContext 对象；

（3）创建 SparkContext 对象时 sparksubmit 进程与 Master 进程通信申请资源；

（4）Master 根据 memory、executor 筛选并分配资源，通知 Worker 启动 Executor 子进程；

（5）Executor 进程接收任务并执行，而 Driver 则监控 Executor 上任务执行状态；（Driver 相当于 YARN 中的 AppMaster 负责监控任务执行）

（6）所有算子都是在 Driver 端执行，即 Driver 记录了 RDD 的依赖关系；但 Action 算子会触发任务提交，触发 Action 后调用 runJob()方法执行任务，程序同步阻塞等待返回结果后继续执行。

（7）触发 Action 时 Driver 端将准备好的任务以 TaskSet（Stage）形式提交到集群，即将 RDD 上保存的业务逻辑发送到 Executor；

（8）调用 saveAsTextFile()算子将数据写入 hdfs 或外部存储介质，在 Executor 计算完成后并行写入。Driver 端写数据到外部存储介质可能带来网络带宽、速度慢、内存溢出等问题。