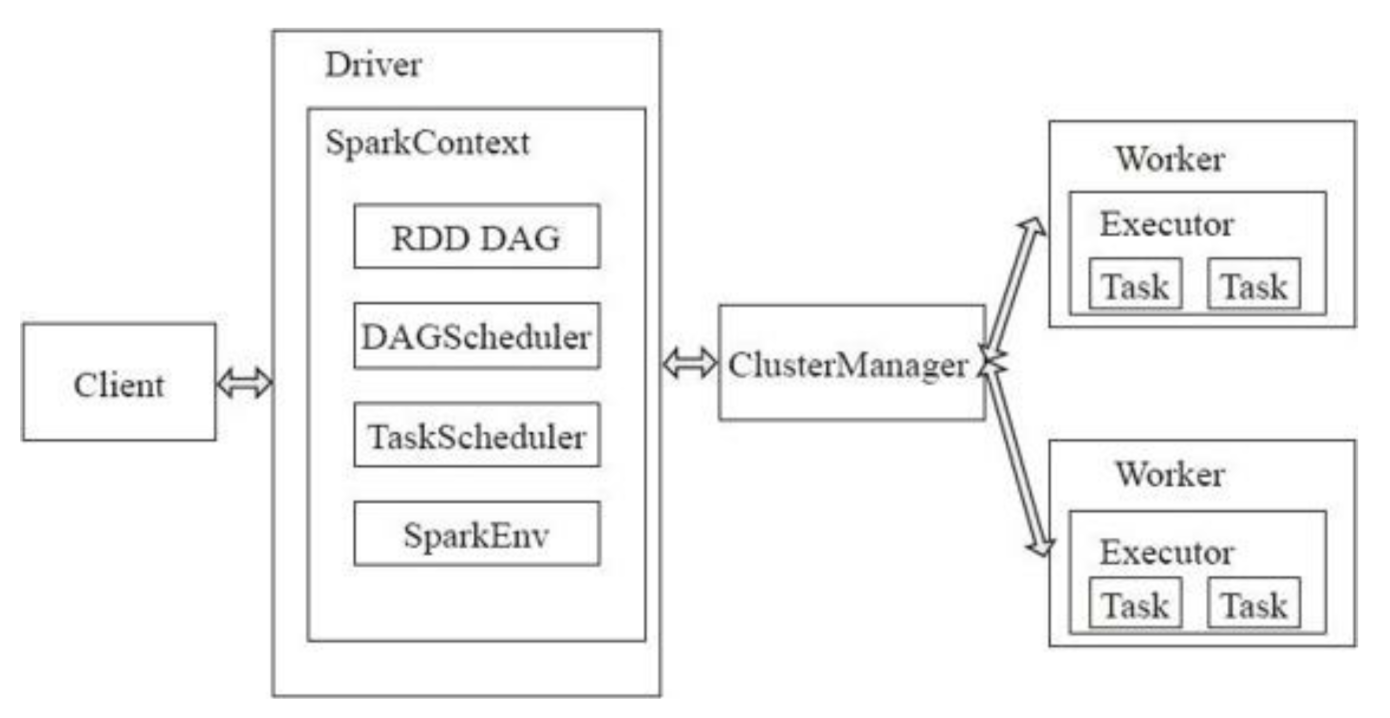
1. 简答题（共5题，每题20分，总100分）



图（1）

1. 根据图（1）解释对应部件在Spark中的作用？

Worker:相当于计算节点，接收主节点命令与进行状态汇报，启动Executor或Driver

Master:作为整个集群的控制器，负责整个集群的正常运行

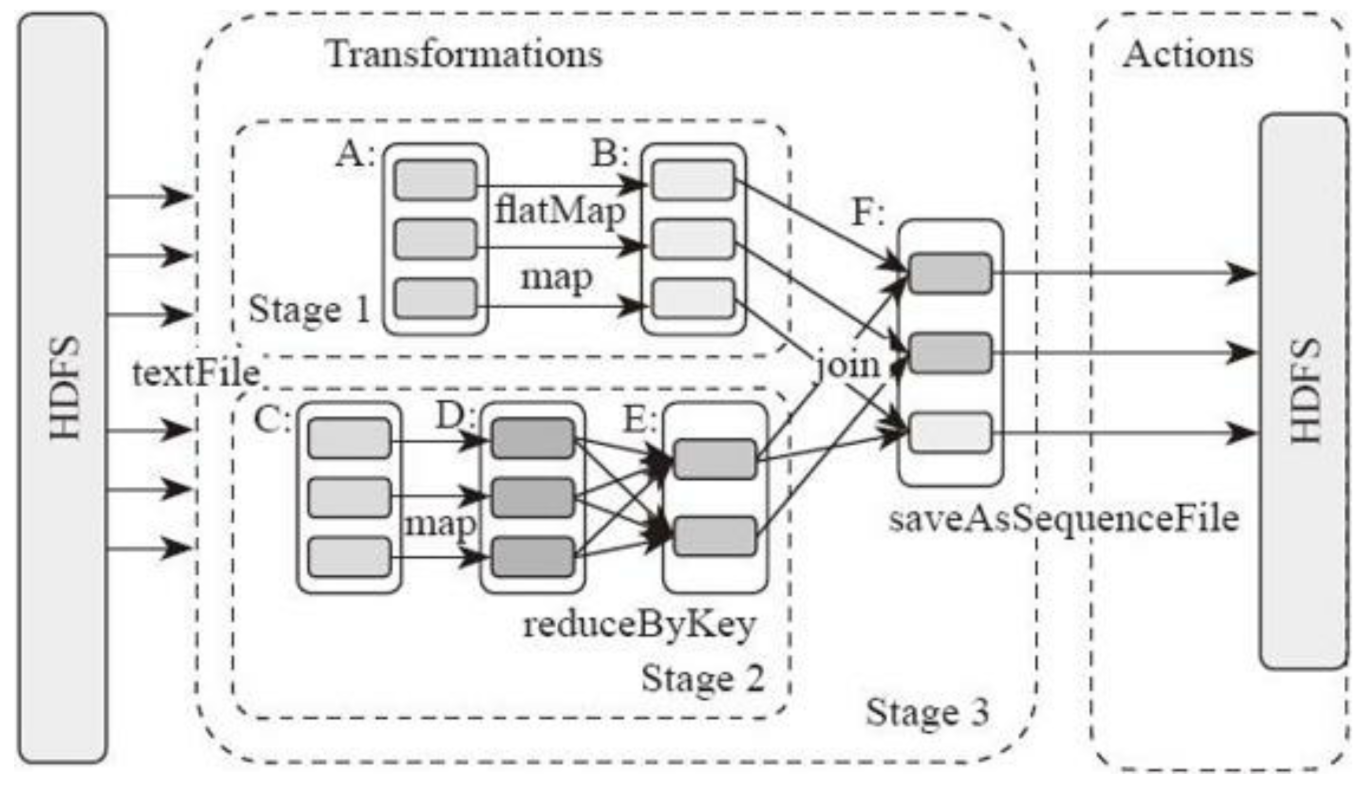
ClusterManager:在Standalone模式中即为Master，在Yarn模式下为资源管理器

Driver:负责控制一个应用的执行

Executor:负责任务的执行

1. 根据图（1）简单描述当client发起命令后，在spark中如何运作？

Spark的整体流程：Client提交应用，Master找到一个Worker启动Driver，Driver向Master或者资源管理器申请资源，之后将应用转化为RDD DAG,再由DAG Schedule将RDD DAG转换为Stage的有向无环图提交给Task Schedule,由TaskSchedule提交任务给Executor执行。在任务执行的过程中，其他组件协同工作，确保整个应用正常执行。



图（2）

1. 根据图（2）简单描述spark的执行逻辑

图(2)A、B、C、D、E、F分别代表不同的RDD,RDD内的方框代表分区。数据从HDFS读取，得到RDD A和RDD C,RDD C上执行map操作，转换为RDD D,RDD B和RDD E进行join操作获得RDD F,在B和E连接转化的过程中会进行Shuffle，最后RDD F通过函数saveAsSequenceFile将数据保存到HDFS中。

1. （1）编写程序并描述创建RDD的方式（15分）



（2）Spark中RDD算子有哪几种（5分）

从大方向来说，Spark 算子大致可以分为以下两类:

1）Transformation 变换/转换算子：这种变换并不触发提交作业，完成作业中间过程处理。

Transformation 操作是延迟计算的，也就是说从一个RDD 转换生成另一个 RDD 的转换操作不是马上执行，需要等到有 Action 操作的时候才会真正触发运算。

2）Action 行动算子：这类算子会触发 SparkContext 提交 Job 作业。

　　Action 算子会触发 Spark 提交作业（Job），并将数据输出 Spark系统。

从小方向来说，Spark 算子大致可以分为以下三类:

1）Value数据类型的Transformation算子，这种变换并不触发提交作业，针对处理的数据项是Value型的数据。  
 2）Key-Value数据类型的Transfromation算子，这种变换并不触发提交作业，针对处理的数据项是Key-Value型的数据对。

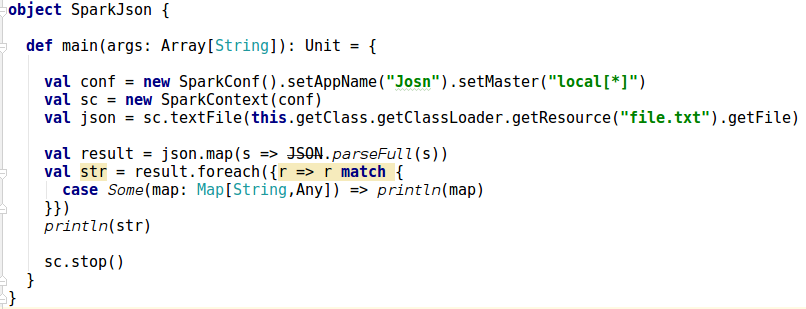
3）Action算子，这类算子会触发SparkContext提交Job作业。

1. 利用Spark map算子和foreach算子将以下的日志中的dst\_ip,src\_ip转成ip类型，并遍历。(提示：这是json类型，利用fastJson对日志进行处理)

{"dst\_as":0,"dst\_ip":1885335600,"sampling\_algorithm":0,"src\_ip":3232243929,"protocol":6,"time\_received":1516410237629}

{"dst\_as":0,"dst\_ip":3395944468,"sampling\_algorithm":0,"src\_ip":3232237789,"protocol":17,"time\_received":1516410237633}

{"dst\_as":0,"dst\_ip":1987799559,"sampling\_algorithm":0,"src\_ip":3232250528,"protocol":6,"time\_received":1516410237629}

只有使用了map和foreach,其他的功能没有。