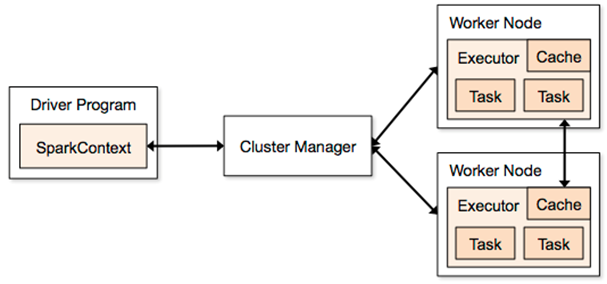
**Spark基本架构**



**基本术语**

**RDD：弹性分布式数据集，一个只读的、可分区的分布式数据集**

**Driver：**负责所有任务的调度

**SparkContext：**是应用程度调度Spark的总入口。

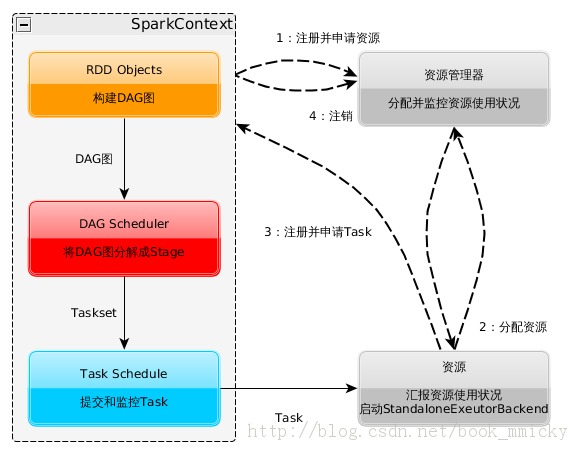
**Cluster Manager：**在集群上获取资源的外部服务（Standalone、Mesos、Yarn）

**Worker Node：**集群中可以运行APP代码的节点

**基本运行流程**

1. **启动Spark Application(SparkContext)**
2. **SparkContext连接cluster Manager**
3. **申请executor资源**
4. **将程序发送到executor上**
5. **SparkContext分解DAG为Stage，将Taskset发送给Task Scheduler**
6. **Task Scheduler将Task发放给Executor运行**
7. **Task在Executor上执行完毕，释放所有资源**

**Spark运行模式**

****