Spark:Yarn-cluster和Yarn-client区别与联系

《Spark on YARN集群模式作业运行全过程分析》
《Spark on YARN客户端模式作业运行全过程分析》
《Spark:Yarn-cluster和Yarn-client区别与联系》
《Spark和Hadoop作业之间的区别》
《Spark Standalone模式作业运行全过程分析》（未发布）

我们都知道Spark支持在yarn上运行，但是Spark on yarn有分为两种模式yarn-cluster和yarn-client，它们究竟有什么区别与联系？阅读完本文，你将了解。

Spark支持可插拔的集群管理模式(Standalone、Mesos以及YARN)，集群管理负责启动executor进程，编写Spark application 的人根本不需要知道Spark用的是什么集群管理。Spark支持的三种集群模式，这三种集群模式都由两个组件组成:master和slave。Master服务(YARN ResourceManager,Mesos master和Spark standalone master)决定哪些application可以运行，什么时候运行以及哪里去运行。而slave服务(YARN NodeManager, Mesos slave和Spark standalone slave)实际上运行executor进程。

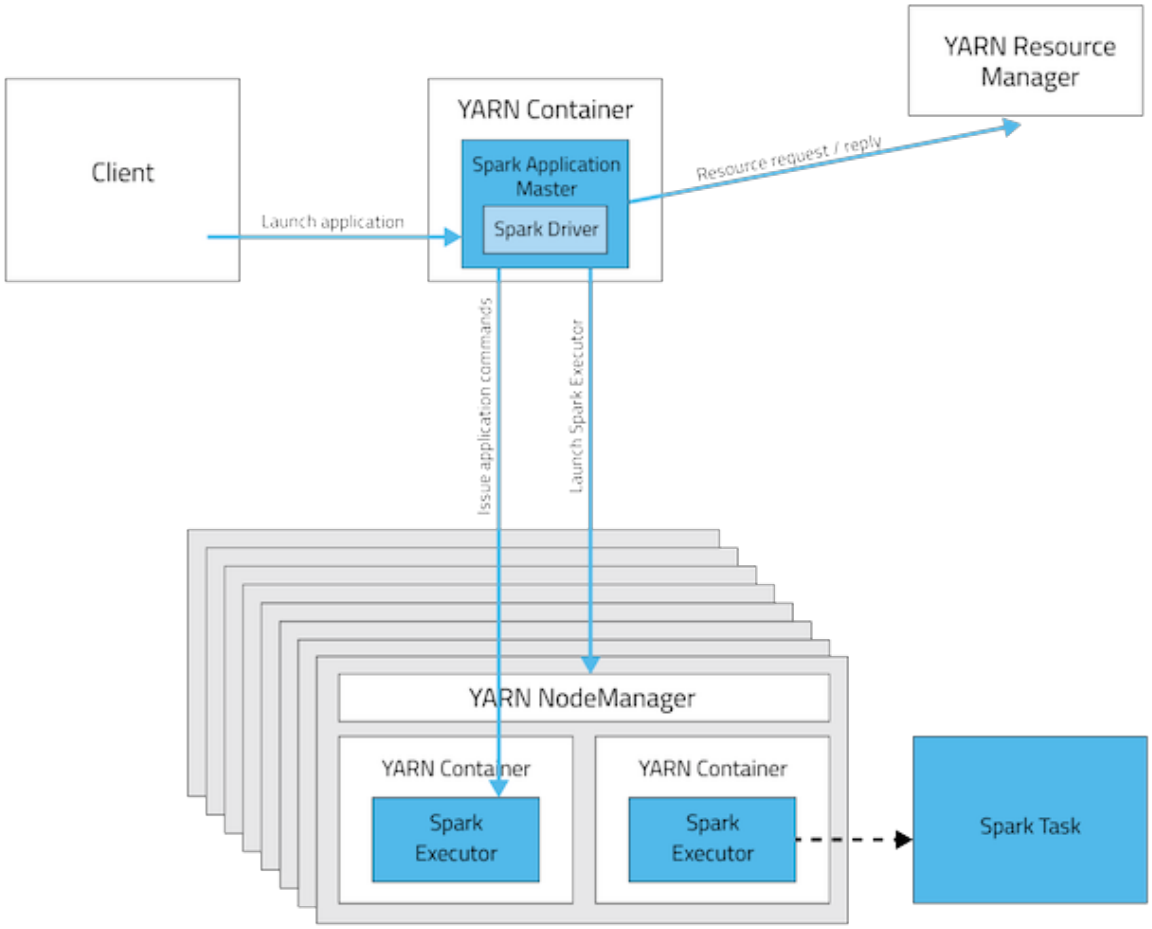
当在YARN上运行Spark作业，每个Spark executor作为一个YARN容器(container)运行。Spark可以使得多个Tasks在同一个容器(container)里面运行。这是个很大的优点。

注意这里和Hadoop的MapReduce作业不一样，MapReduce作业为每个Task开启不同的JVM来运行。虽然说MapReduce可以通过参数来配置。详见mapreduce.job.jvm.numtasks。关于这个参数的介绍已经超过本篇文章的介绍。

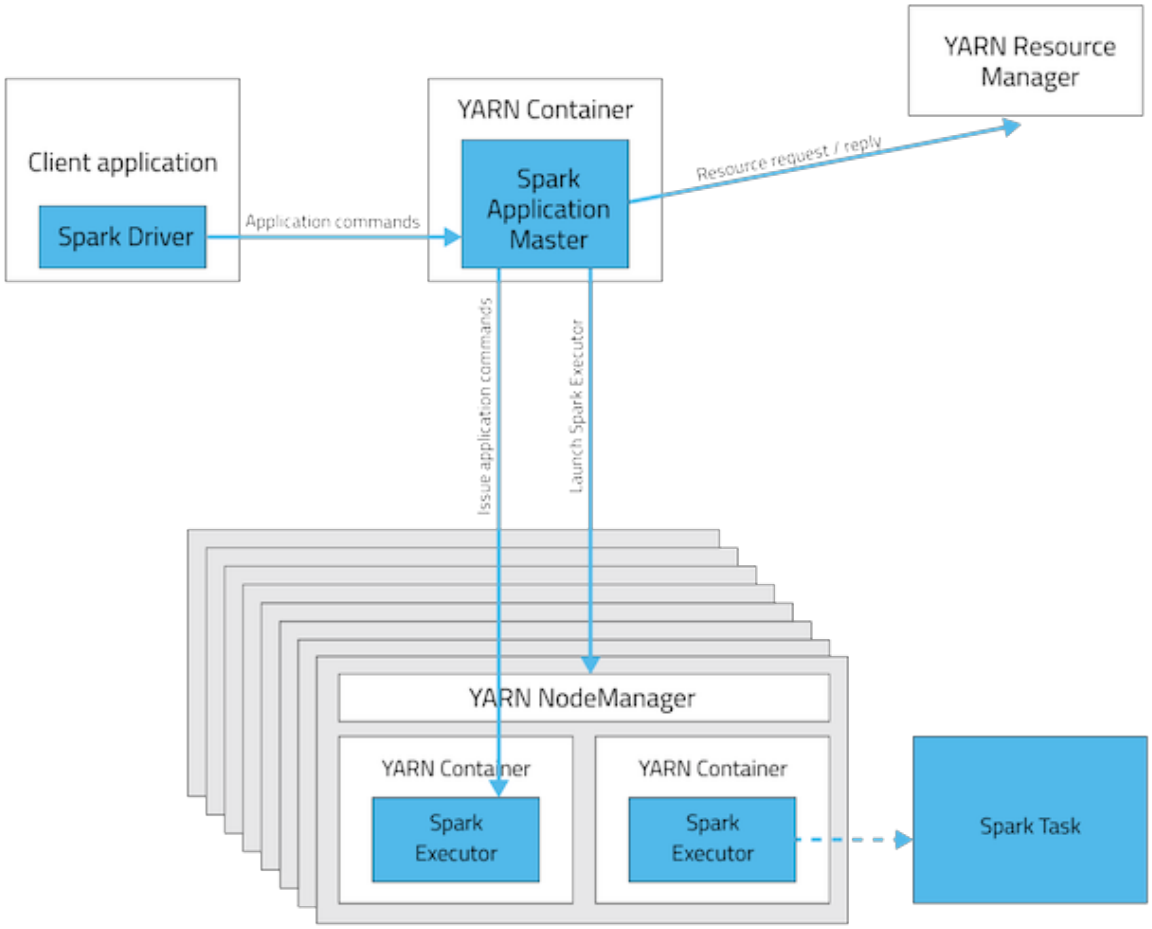
从广义上讲，yarn-cluster适用于生产环境；而yarn-client适用于交互和调试，也就是希望快速地看到application的输出。

在我们介绍yarn-cluster和yarn-client的深层次的区别之前，我们先明白一个概念：Application Master。在YARN中，每个Application实例都有一个Application Master进程，它是Application启动的第一个容器。它负责和ResourceManager打交道，并请求资源。获取资源之后告诉NodeManager为其启动container。

从深层次的含义讲，yarn-cluster和yarn-client模式的区别其实就是Application Master进程的区别，yarn-cluster模式下，driver运行在AM(Application Master)中，它负责向YARN申请资源，并监督作业的运行状况。当用户提交了作业之后，就可以关掉Client，作业会继续在YARN上运行。然而yarn-cluster模式不适合运行交互类型的作业。而yarn-client模式下，Application Master仅仅向YARN请求executor，client会和请求的container通信来调度他们工作，也就是说Client不能离开。看下下面的两幅图应该会明白（上图是yarn-cluster模式，下图是yarn-client模式）：



如果想及时了解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop



如果想及时了解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

从源码实现，调度器等方面看，请参照本博客的《Spark on YARN集群模式作业运行全过程分析》和《Spark on YARN客户端模式作业运行全过程分析》的介绍。

本博客文章除特别声明，全部都是原创！
原创文章版权归过往记忆大数据（过往记忆）所有，未经许可不得转载。
本文链接: 【】（）