09 | 为什么我们管Yarn叫作资源调度框架?

2018-11-17 李智慧

从0开始学大数据 进入课程 >

讲述：李智慧
时长 12:07 大小 5.56M

我们知道，Hadoop 主要是由三部分组成，除了前面我讲过的分布式文件系统 HDFS、分布式计算框架 MapReduce，还有一个是分布式集群资源调度框架 Yarn。但是 Yarn 并不是随 Hadoop 的推出一开始就有的，Yarn 作为分布式集群的资源调度框架，它的出现伴随着 Hadoop 的发展，使 Hadoop 从一个单一的大数据计算引擎，成为一个集存储、计算、资源管理为一体的完整大数据平台，进而发展出自己的生态体系，成为大数据的代名词。

所以在我们开始聊 Yarn 的实现原理前，有必要看看 Yarn 发展的过程，这对你理解 Yarn 的原理以及为什么被称为资源调度框架很有帮助。

先回忆一下我们学习的 MapReduce 的架构，在 MapReduce 应用程序的启动过程中，最重要的就是要把 MapReduce 程序分发到大数据集群的服务器上，在 Hadoop 1 中，这个过程主要是通过 TaskTracker 和 JobTracker 通信来完成。

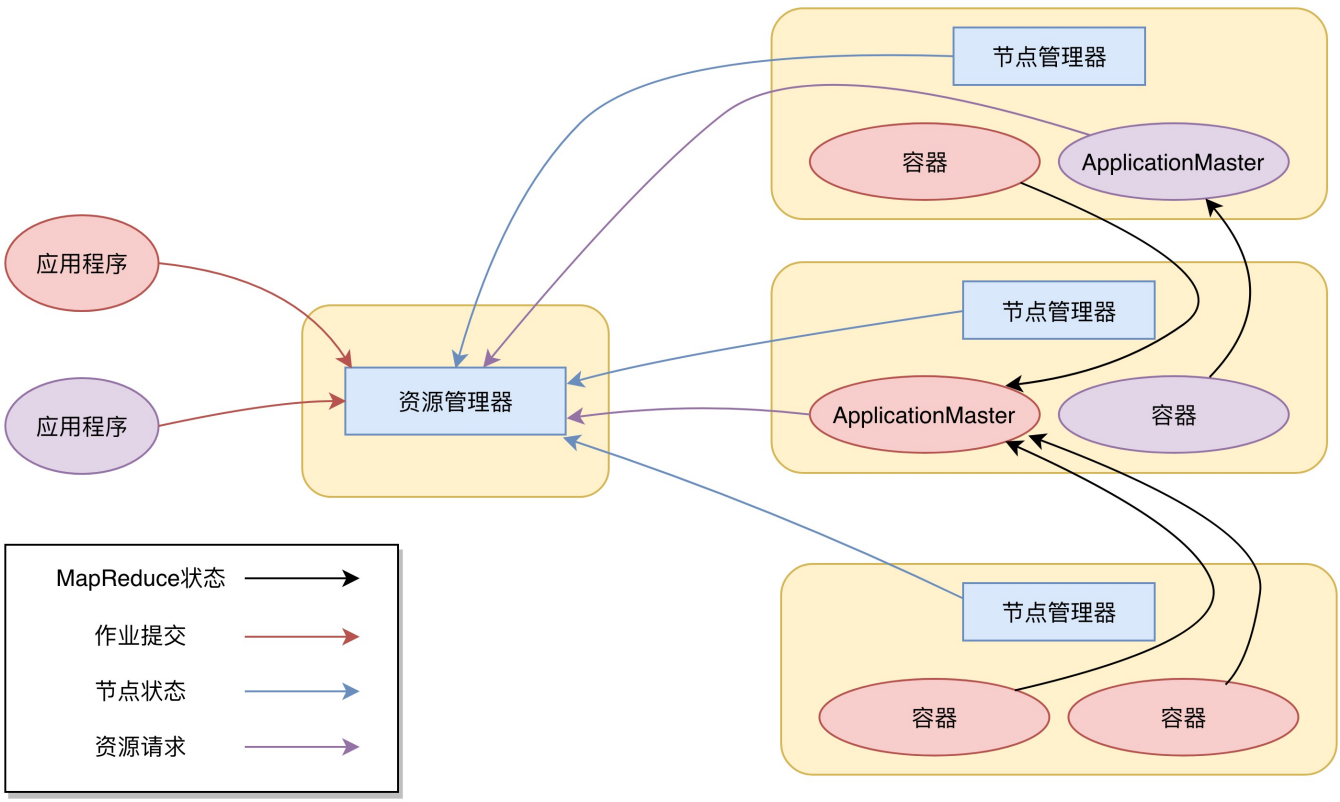
这个方案有什么缺点吗?

这种架构方案的主要缺点是，**服务器集群资源调度管理和 MapReduce 执行过程耦合在一起，如果想在当前集群中运行其他计算任务，比如 Spark 或者 Storm，就无法统一使用集群中的资源了。**

在 Hadoop 早期的时候，大数据技术就只有 Hadoop 一家，这个缺点并不明显。但随着大数据技术的发展，各种新的计算框架不断出现，我们不可能为每一种计算框架部署一个服务器集群，而且就算能部署新集群，数据还是在原来集群的 HDFS 上。所以我们需要把 MapReduce 的资源管理和计算框架分开，这也是 Hadoop 2 最主要的变化，就是将 Yarn 从 MapReduce 中分离出来，成为一个独立的资源调度框架。

Yarn 是“Yet Another Resource Negotiator”的缩写，字面意思就是“另一种资源调度器”。事实上，在 Hadoop 社区决定将资源管理从 Hadoop 1 中分离出来，独立开发 Yarn 的时候，业界已经有一些大数据资源管理产品了，比如 Mesos 等，所以 Yarn 的开发者索性管自己的产品叫“另一种资源调度器”。这种命名方法并不鲜见，曾经名噪一时的 Java 项目编译工具 Ant 就是“Another Neat Tool”的缩写，意思是“另一种整理工具”。

下图是 Yarn 的架构。



从图上看，Yarn 包括两个部分：一个是资源管理器（Resource Manager），一个是节点管理器（Node Manager）。这也是 Yarn 的两种主要进程：ResourceManager 进程负责整个集群的资源调度管理，通常部署在独立的服务器上；NodeManager 进程负责具体服务器上的资源和任务管理，在集群的每一台计算服务器上都会启动，基本上跟 HDFS 的 DataNode 进程一起出现。

具体说来，资源管理器又包括两个主要组件：调度器和应用程序管理器。

调度器其实就是一个资源分配算法，根据应用程序（Client）提交的资源申请和当前服务器集群的资源状况进行资源分配。Yarn 内置了几种资源调度算法，包括 Fair Scheduler、Capacity Scheduler 等，你也可以开发自己的资源调度算法供 Yarn 调用。

Yarn 进行资源分配的单位是容器（Container），每个容器包含了一定量的内存、CPU 等计算资源，默认配置下，每个容器包含一个 CPU 核心。容器由 NodeManager 进程启动和管理，NodeManger 进程会监控本节点上容器的运行状况并向 ResourceManger 进程汇报。

应用程序管理器负责应用程序的提交、监控应用程序运行状态等。应用程序启动后需要在集群中运行一个 ApplicationMaster，ApplicationMaster 也需要运行在容器里面。每个应用程序启动后都会先启动自己的 ApplicationMaster，由 ApplicationMaster 根据应用程序的资源需求进一步向 ResourceManager 进程申请容器资源，得到容器以后就会分发自己的应用程序代码到容器上启动，进而开始分布式计算。

我们以一个 MapReduce 程序为例，来看一下 Yarn 的整个工作流程。

1. 我们向 Yarn 提交应用程序，包括 MapReduce ApplicationMaster、我们的 MapReduce 程序，以及 MapReduce Application 启动命令。

2.ResourceManager 进程和 NodeManager 进程通信，根据集群资源，为用户程序分配第一个容器，并将 MapReduce ApplicationMaster 分发到这个容器上面，并在容器里面启动 MapReduce ApplicationMaster。

3.MapReduce ApplicationMaster 启动后立即向 ResourceManager 进程注册，并为自己的应用程序申请容器资源。

4.MapReduce ApplicationMaster 申请到需要的容器后，立即和相应的 NodeManager 进程通信，将用户 MapReduce 程序分发到 NodeManager 进程所在服务器，并在容器中运行，运行的就是 Map 或者 Reduce 任务。

5.Map 或者 Reduce 任务在运行期和 MapReduce ApplicationMaster 通信，汇报自己的运行状态，如果运行结束，MapReduce ApplicationMaster 向 ResourceManager 进程注销并释放所有的容器资源。

MapReduce 如果想在 Yarn 上运行，就需要开发遵循 Yarn 规范的 MapReduce ApplicationMaster，相应地，其他大数据计算框架也可以开发遵循 Yarn 规范的 ApplicationMaster，这样在一个 Yarn 集群中就可以同时并发执行各种不同的大数据计算框架，实现资源的统一调度管理。

细心的你可能会发现，我在今天文章开头的时候提到 Hadoop 的三个主要组成部分的时候，管 HDFS 叫分布式文件**系统**，管 MapReduce 叫分布式计算**框架**，管 Yarn 叫分布式集群资源调度**框架**。

为什么 HDFS 是系统，而 MapReduce 和 Yarn 则是框架?

框架在架构设计上遵循一个重要的设计原则叫“**依赖倒转原则**”，依赖倒转原则是**高层模块不能依赖低层模块，它们应该共同依赖一个抽象，这个抽象由高层模块定义，由低层模块实现。**

所谓高层模块和低层模块的划分，简单说来就是在调用链上，处于前面的是高层，后面的是低层。我们以典型的 Java Web 应用举例，用户请求在到达服务器以后，最先处理用户请求的是 Java Web 容器，比如 Tomcat、Jetty 这些，通过监听 80 端口，把 HTTP 二进制流封装成 Request 对象；然后是 Spring MVC 框架，把 Request 对象里的用户参数提取出来，根据请求的 URL 分发给相应的 Model 对象处理；再然后就是我们的应用程序，负责处理用户请求，具体来看，还会分成服务层、数据持久层等。

在这个例子中，Tomcat 相对于 Spring MVC 就是高层模块，Spring MVC 相对于我们的应用程序也算是高层模块。我们看到虽然 Tomcat 会调用 Spring MVC，因为 Tomcat 要把 Request 交给 Spring MVC 处理，但是 Tomcat 并没有依赖 Spring MVC，Tomcat 的代码里不可能有任何一行关于 Spring MVC 的代码。

那么，Tomcat 如何做到不依赖 Spring MVC，却可以调用 Spring MVC？如果你不了解框架的一般设计方法，这里还是会感到有点小小的神奇是不是？

秘诀就是 Tomcat 和 Spring MVC 都依赖 J2EE 规范，Spring MVC 实现了 J2EE 规范的 HttpServlet 抽象类，即 DispatcherServlet，并配置在 web.xml 中。这样，Tomcat 就可以调用 DispatcherServlet 处理用户发来的请求。

同样 Spring MVC 也不需要依赖我们写的 Java 代码，而是通过依赖 Spring MVC 的配置文件或者 Annotation 这样的抽象，来调用我们的 Java 代码。

所以，Tomcat 或者 Spring MVC 都可以称作是框架，它们都遵循依赖倒转原则。

现在我们再回到 MapReduce 和 Yarn。实现 MapReduce 编程接口、遵循 MapReduce 编程规范就可以被 MapReduce 框架调用，在分布式集群中计算大规模数据；实现了 Yarn 的接口规范，比如 Hadoop 2 的 MapReduce，就可以被 Yarn 调度管理，统一安排服务器资源。所以说，MapReduce 和 Yarn 都是框架。

相反地，HDFS 就不是框架，使用 HDFS 就是直接调用 HDFS 提供的 API 接口，HDFS 作为底层模块被直接依赖。

小结

Yarn 作为一个大数据资源调度框架，调度的是大数据计算引擎本身。它不像 MapReduce 或 Spark 编程，每个大数据应用开发者都需要根据需求开发自己的 MapReduce 程序或者 Spark 程序。而现在主流的大数据计算引擎所使用的 Yarn 模块，也早已被这些计算引擎的开发者做出来供我们使用了。作为普通的大数据开发者，我们几乎没有机会编写 Yarn 的相关程序。但是，这是否意味着只有大数据计算引擎的开发者需要基于 Yarn 开发，才需要理解 Yarn 的实现原理呢？

恰恰相反，我认为理解 Yarn 的工作原理和架构，对于正确使用大数据技术，理解大数据的工作原理，是非常重要的。在云计算的时代，一切资源都是动态管理的，理解这种动态管理的原理对于理解云计算也非常重要。Yarn 作为一个大数据平台的资源管理框架，简化了应用场景，对于帮助我们理解云计算的资源管理很有帮助。

思考题

Web 应用程序的服务层 Service 和数据持久层 DAO 也是上下层模块关系，你设计的 Service 层是否按照框架的一般架构方法，遵循依赖倒转原则？

欢迎你写下自己的思考或疑问，与我和其他同学一起讨论。

上一篇　08 | MapReduce如何让数据完成一次旅行？

下一篇　10 | 模块答疑：我们能从Hadoop学到什么？

精选留言 (40)

写留言

落叶飞逝的...

2018-11-17

16

实际项目开发中，要做到依赖倒置的方法，一般就是抽象出相应的接口的方法，不依赖具体。面向接口编程。

作者回复: 是的，但是更重要的是接口是高层需求的抽象，还是底层实现的抽象。这是依赖倒置的关键，面向接口本身并不能保证依赖倒置原则，否则和接口隔离原则没有区别。

Zach_
2018-11-19

5

老师、我看了一下，还是不知道MR程序是怎么分发的，提问如下：应用程序给ResourceManager提交了MR应用程序、ResourceManager给MR应用程序分配了首节点、并在分配的首节点上分配了MapReduce ApplicationMaster、以及分配了MapReduce ApplicationMaster的容器，MapReduce Application 启动后和其他节点通信，会分发MapReduce应用程序。可是这个时候ApplicationMaster并没有MR应用程...
展开

小千
2018-11-22

3

sql语言是不是也是依赖倒转原则？不同的数据库都要支持sql语言规范，（很多）sql语句语句都可以在不同的数据库执行。

> 作者回复: 依赖倒转一般是指两个实现之间的依赖关系倒转。
> 这里上下文的两个实现应该分别是应用程序和数据库，应用程序依赖SQL，数据库实现SQL。
>
> 但是，SQL作为规范是数据库制定的规范，是底层规范，而不是应用程序制定的，所以这种情况一般不认为是依赖倒转。

老男孩
2018-11-19

3

突然明白了，这么多年都是错误的观点。我之前的所谓分层展现层，服务层，持久层其实都是上层依赖下层的抽象，不是依赖倒置。

> 作者回复: 是的👍

Mcnulty
2018-11-19

3

前文中写道
3.JobTacker 根据作业调度策略创建 JobInProcess 树，每个作业都会有一个自己的

JobInProcess 树。
6. 如果 TaskTracker 有空闲的计算资源（有空闲 CPU 核心），JobTracker 就会给它分配任务。...
展开

作者回复: 我觉得你已经描述很清楚了，资源管理和执行过程耦合，再感受一下～

小辉辉
2018-11-18
3

老师讲得通俗易懂，没接触大数据之前，一直以为大数据是很高深的东西。经过几讲的了解之后，通过从原理出发，上手就很容易了。

纯洁的憎恶
2018-11-18
3

MapReduce框架遵循把程序发送到数据存储位置运行的原则。而资源调度框架的任务是动态调配计算资源（内存+cpu），那么就很有可能出现本地数据需要发送到其他节点计算的情况，于是就会有网络传输大量数据的现象，这是否与程序在数据存储节点运行的初衷相悖呢？我这么理解对么？

作者回复: 有可能。
不用yarn也会有这个问题。
网络效率这几年提升很快，这个问题不严重。

Li Shundu...
2018-11-17
3

请问Yarn里的容器和docker这一类容器有什么关系吗？
展开

作者回复: 没有关系。两种容器设计思路差不多，docker更通用。

hua168
2018-11-17
3

看完几期感觉没有什么能难得住大神你的，回答问题在您那里感觉都很简单……我一般学习是先找视频看一下，照着截图，练习，然后去官网看一下说明文档，看更新了哪些知识。照视频学习又要截图，感觉很慢，很费时，看官方文档又很难深入，能否请教一下自学如果能深入，是我方法不对吗？有很多问题官网都没答案的啊，google不少也搜索不出来……运维类学的东西很多，精通感觉比较难……把原理东西，理解好，慢慢锻炼能不能…

展开

作者回复: 理解原理后倒推它应该是什么样，训练自己从设计者角度分析问题，而不是一味被动学习。

也是这个专栏想达到的目的。

蜡笔小新

2018-11-24

2

老师请教一下，MapReduce ApplicationMaster怎么计算出需要多少资源的呢?

作者回复: 根据数据量和分片大小计算，相除就可以。

梁中华

2018-11-19

2

提个建议，可以多放一些留言出来，鼓励大家多留言，老师多互动，这样热闹一些，学习氛围更浓一些。

三木子

2018-11-19

2

强烈建议老师加餐一篇你对架构设计理解的文章。^^

展开

作者回复: 也许我们将来可以再开个架构的专栏😁

hunterlodg...

2018-11-19

2

老师前面说"每个应用程序启动后都会先启动自己的 ApplicationMaster"，后面具体mapReduce例子里又是先启动ApplicationMaster的，这不矛盾吗？

作者回复: 应用程序指client，最先启动，通常不在Hadoop集群中。client提交作业后，是am在Hadoop中启动，对hadoop而言可以认为是am最先启动。

生活在别处
2019-02-19

1

老师，资源调度和计算调度的区别是什么？

展开

作者回复: 资源调度如Yarn，管理的是集群中的计算资源，如CPU、内存的分配和回收。
计算调度应该是计算任务调度，如map和reduce的任务或者spark的任务，应该在哪个container启动，启动前后顺序管理等。

Wiggle Wi...
2019-02-01

1

在service中使用的是dao的接口，接口规定了一类dao所要实现的功能，比如crud操作。crud操作的实现由具体的dao类实现，具体实现可能会因为数据库的改变而改变，但任何改变都不会影响service

展开

tom
2019-01-17

1

依赖倒转原则还是不太清楚，service层抽象通用的crud操作，在crud方法中调用基类dao的方法，这个符合依赖倒转原则吗？

作者回复: 不符合，还是上层依赖下层。

Jowin
2018-12-01

1

请教老师，关于mapreduce和yarn的结合，是不是mapreduce ApplicationMaster 向资源管理器申请计算资源时可以指定目标节点（数据分片所在节点），而如果系统资源能够满足，就会把mapreduce计算任务分发到指定的服务器上。如果资源不允许，比如目标节点非常繁忙，这时部分mapreduce计算任务可能会分配另外的服务器（数据分片不在本地）？也就是说，yarn对资源调度是尽力而为，不保值一定满足ApplicationMaster的要...

展开

作者回复: 是的

宝宝疯

2018-11-27

1

理解原理后倒推它应该是什么样，训练自己从设计者角度分析问题，而不是一味被动学习。
回答很精彩啊

展开

scorpiozj

2018-11-19

1

理解透彻业务，然后抽象出接口；需要多实践才行。

另外，假设配置相同的情形下，资源管理器通常可以管理多少个节点管理器

展开

hunterlodg...

2018-11-19

1

赞老师的延伸内容，受益良多！谢谢！