MapReduce和YARN技术原理

www.huawei.com

HUAWEI

目标

- 学完本课程后，您将能够:
  - 熟悉MapReduce和YARN是什么
  - 掌握MapReduce使用的场景及其原理
  - 掌握MapReduce和YARN功能与架构
  - 熟悉YARN的新特性

目录

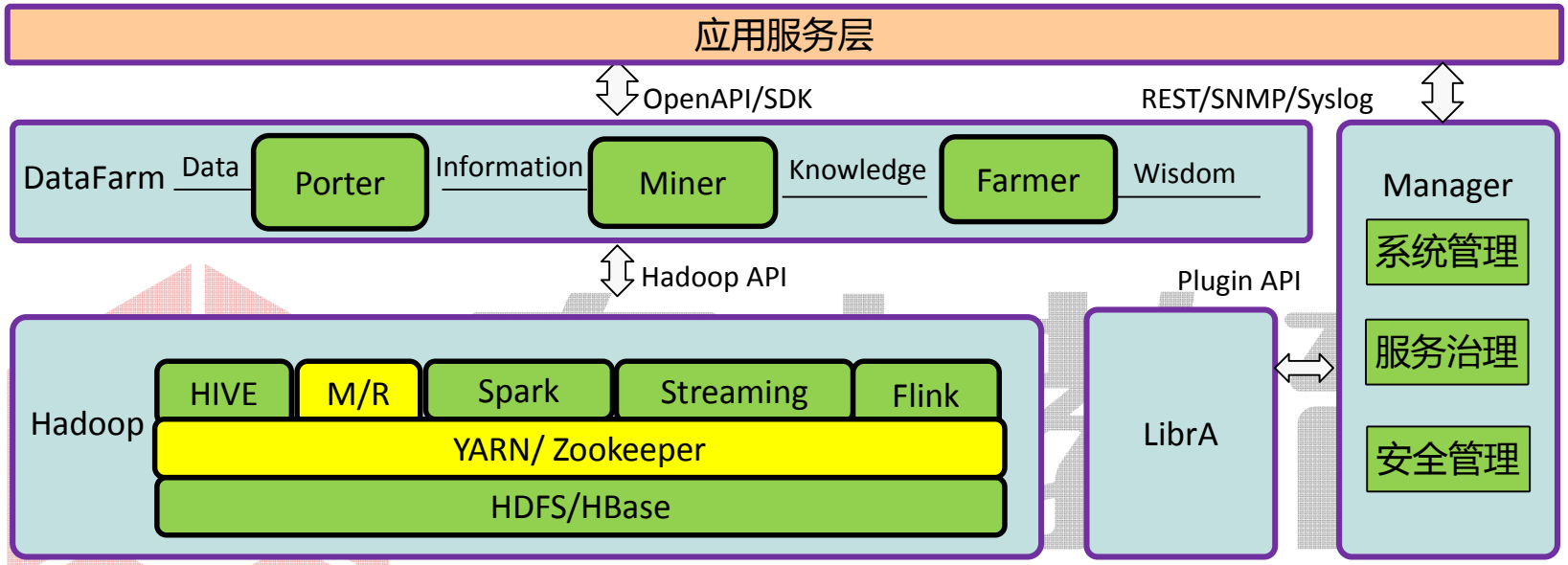
MapReduce概述

- MapReduce基于Google发布的MapReduce论文设计开发，用于大规模数据集（大于1TB）的并行计算，具有如下特点：
  - 易于编程：程序员仅需描述做什么，具体怎么做交由系统的执行框架处理。
  - 良好的扩展性：可通过添加节点以扩展集群能力。
  - 高容错性：通过计算迁移或数据迁移等策略提高集群的可用性与容错性。

YARN概述

- Apache Hadoop YARN (Yet Another Resource Negotiator，另一种资源协调者) 是一种新的 Hadoop 资源管理器，它是一个通用资源管理系统，可为上层应用提供统一的资源管理和调度，它的引入为集群在利用率、资源统一管理和数据共享等方面带来了巨大好处。

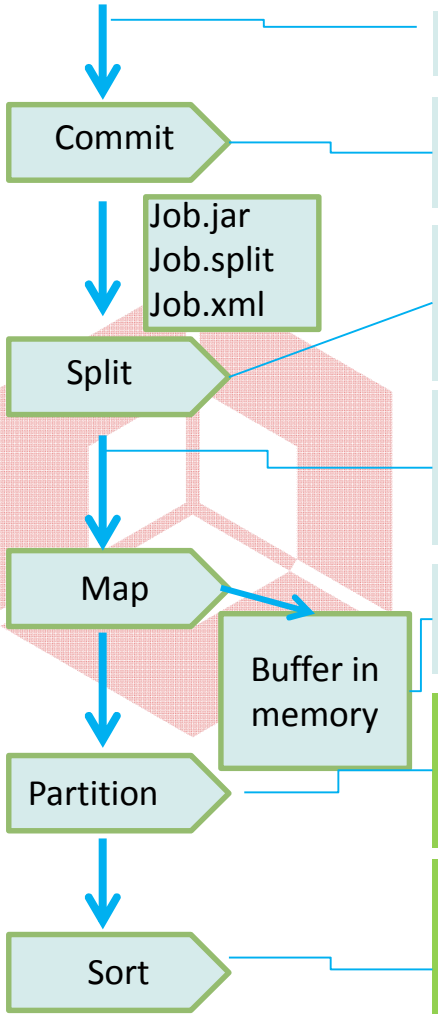
YARN在FusionInsight产品的位置



YARN是Hadoop2.0中的资源管理系统，它是一个通用的资源管理模块，可为各类应用程序提供资源管理和调度功能。

目录

MapReduce过程详解(1)



在启动MapReduce之前，确保待处理的文件放在HDFS上面。

MapReduce应用将请求提交给RM，由RM创建对应的Job，一个应用对应一个Job（jobID形如，job_201431281420_0001）。

Job提交前，先将待处理的文件进行分片（Split）。MR框架默认将一个块（Block）作为一个分片。客户端应用可以重定义块与分片的映射关系。

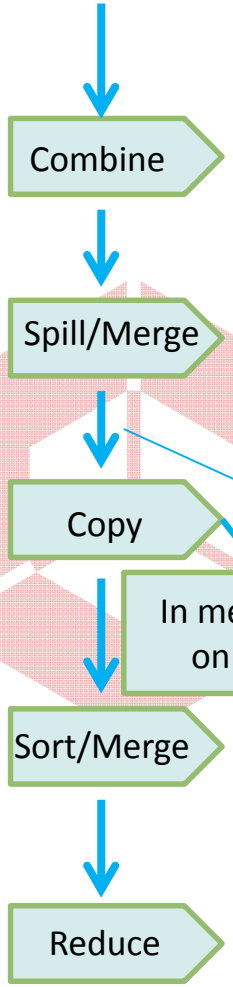
Job提交给RM，RM根据NM的负载在NM集群中挑选合适的节点调度AM，AM负责Job任务的初始化并向RM申请资源，由RM调度合适的NM启动Container, Container来执行Task。

Map的输出放入一个环形内存缓冲区，当缓冲区数据溢出时，需将缓冲区中的数据写入到本地磁盘，写入本地磁盘之前通常需要做如下处理：

1.分区（Partition）---默认采用Hash算法进行分区，MR框架根据Reduce Task个数来确定分区个数。具备相同Key值的记录最终被送到相同的Reduce Task来处理。

2.排序（Sort）---将Map输出的记录排序，例如将('Hi','1'),('Hello','1')重新排序为('Hello','1'), ('Hi','1')。

MapReduce过程详解(2)



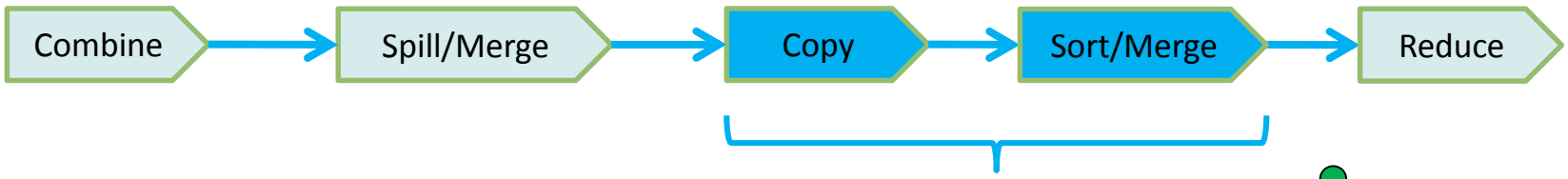
3.组合(Combine)---这个动作MR框架默认是可选的。例如将('Hi','1'), ('Hi','1'),('Hello','1'), (Hello','1')进行合并操作为('Hi','2'), ('Hello','2')。

4.合并(Spill)---Map Task在处理后会产生很多的溢出文件(spill file)，这时需将多个溢出文件进行合并处理，生成一个经过分区和排序的Spill File（MOF:MapOutFile）。为减少写入磁盘的数据量，MR支持对MOF进行压缩后再写入。

通常在Map Task任务完成MOF输出进度到3%时启动Reduce，从各个Map Task获取MOF文件。前面提到Reduce Task个数由客户端决定，Reduce Task个数决定MOF文件分区数。因此Map Task输出的MOF文件都能找到相对应的Redcue Task来处理。

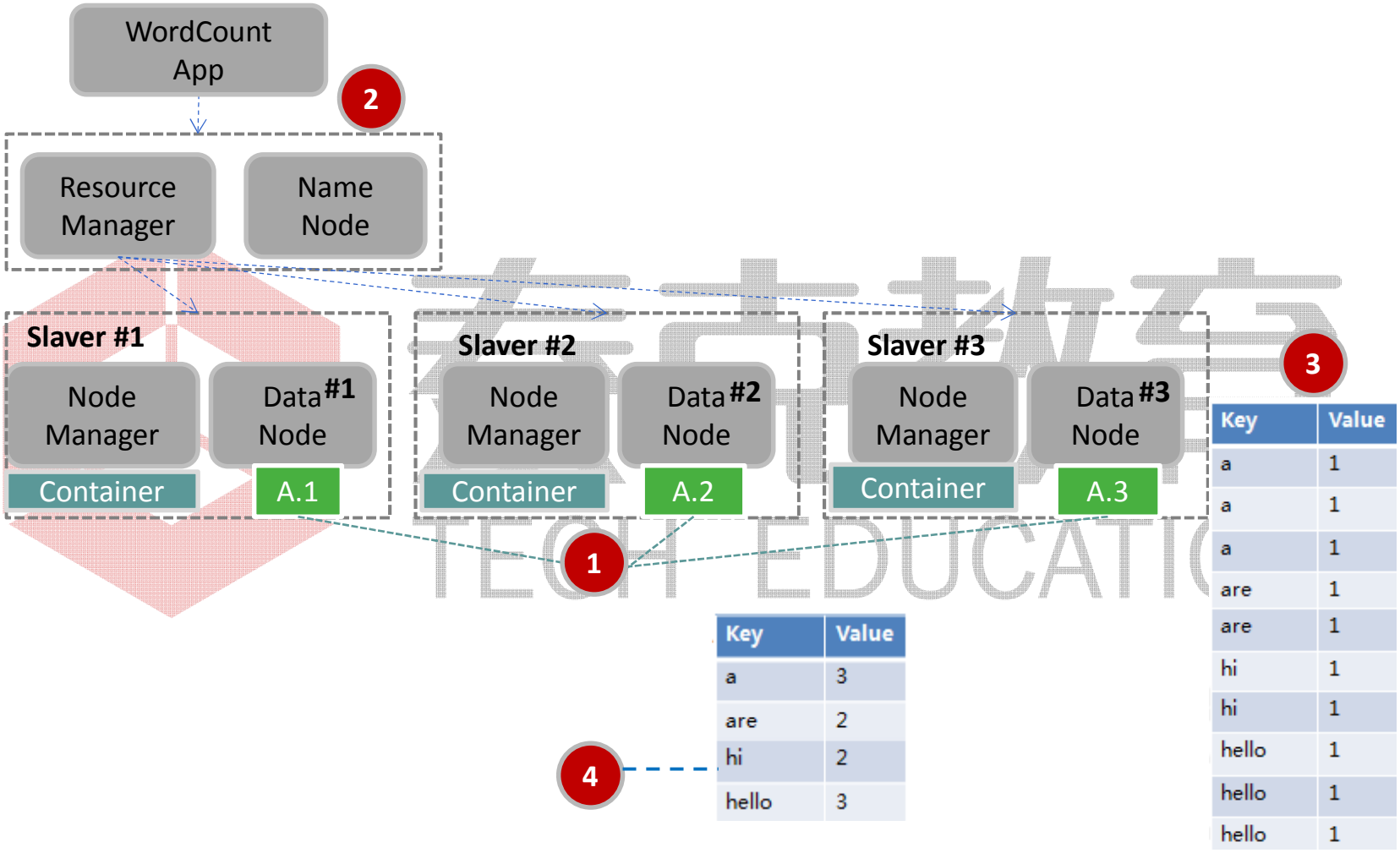
前面提到的MOF文件是经过排序处理的。当Reduce Task接收的数据量不大时，则直接存放在内存缓冲区中，随着缓冲区文件的增多，MR后台线程将它们合并成一个更大的有序文件，这个动作是Reduce阶段的Merge操作，过程中会产生许多中间文件，最后一次合并的结果直接输出到用户自定义的Reduce函数。

Shuffle机制



Shuffle的定义：Map阶段和Reduce阶段之间传递中间数据的过程，包括Reduce Task从各个Map Task获取MOF文件的过程，以及对MOF的排序与合并处理。

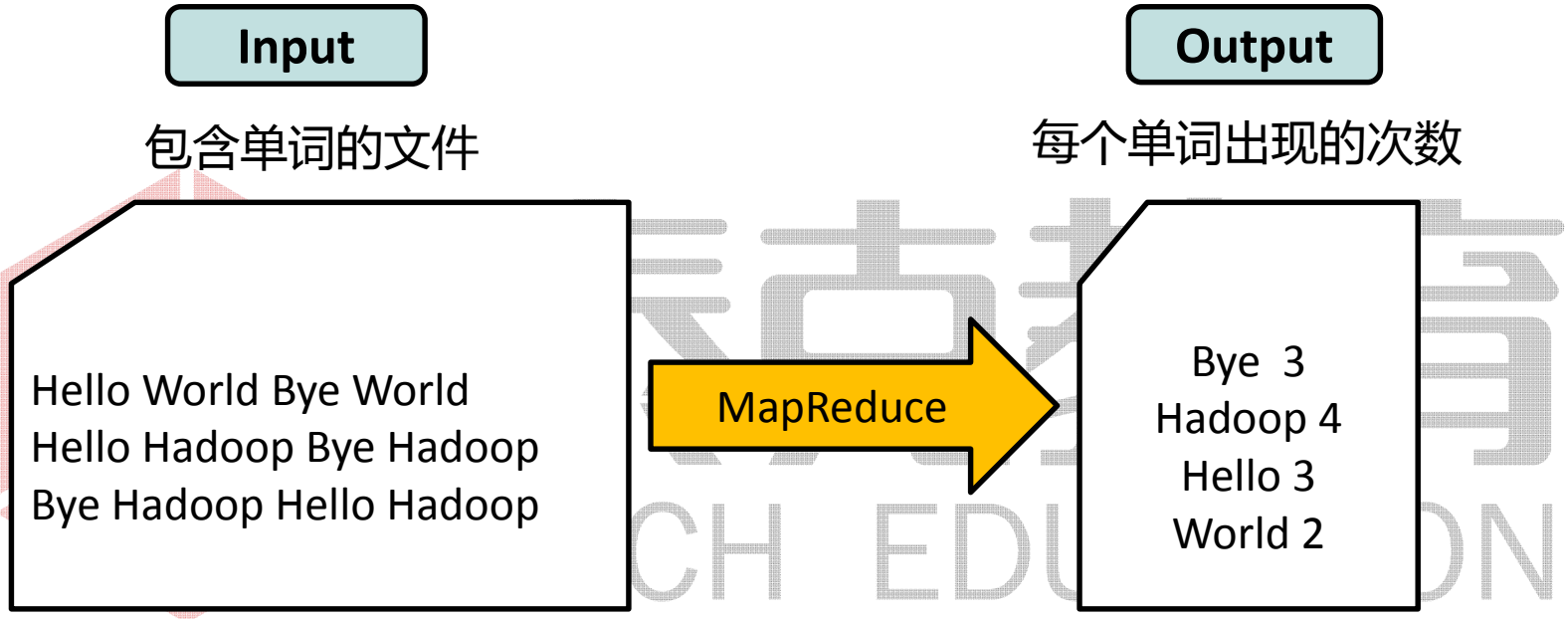
典型程序WordCount举例



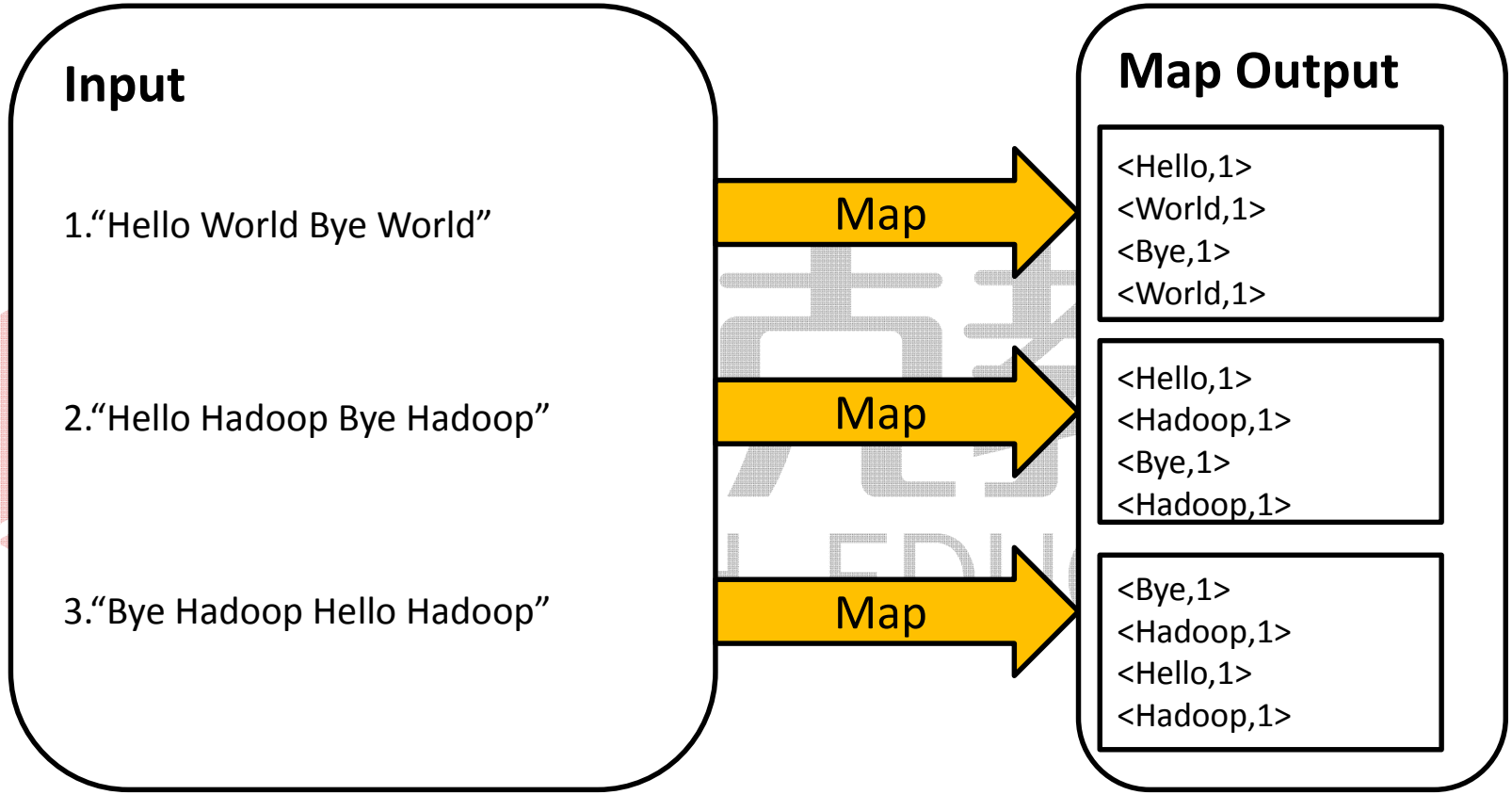
Key	Value
a	3
are	2
hi	2
hello	3

Key	Value
a	1
a	1
a	1
are	1
are	1
hi	1
hi	1
hello	1
hello	1
hello	1

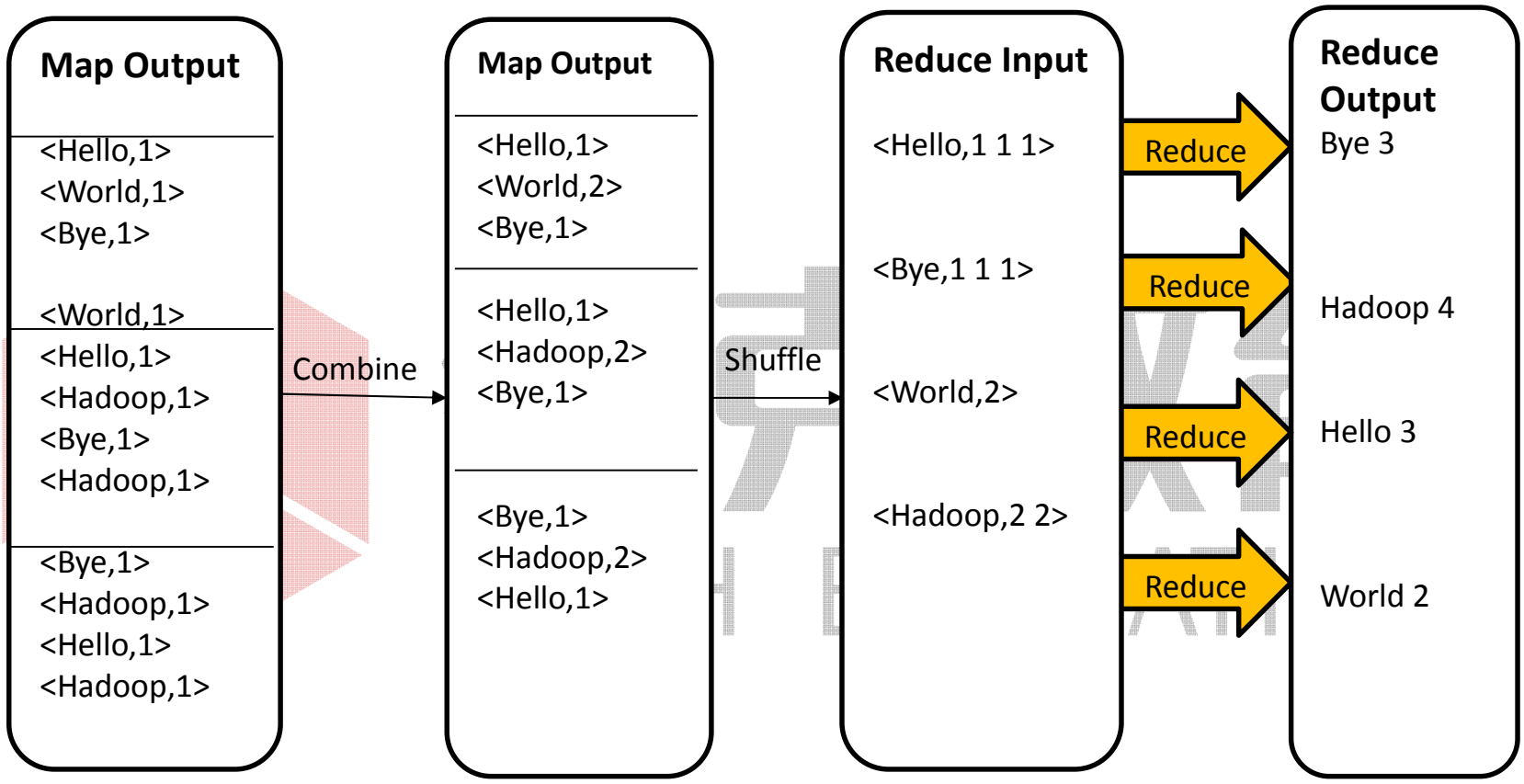
WordCount程序功能



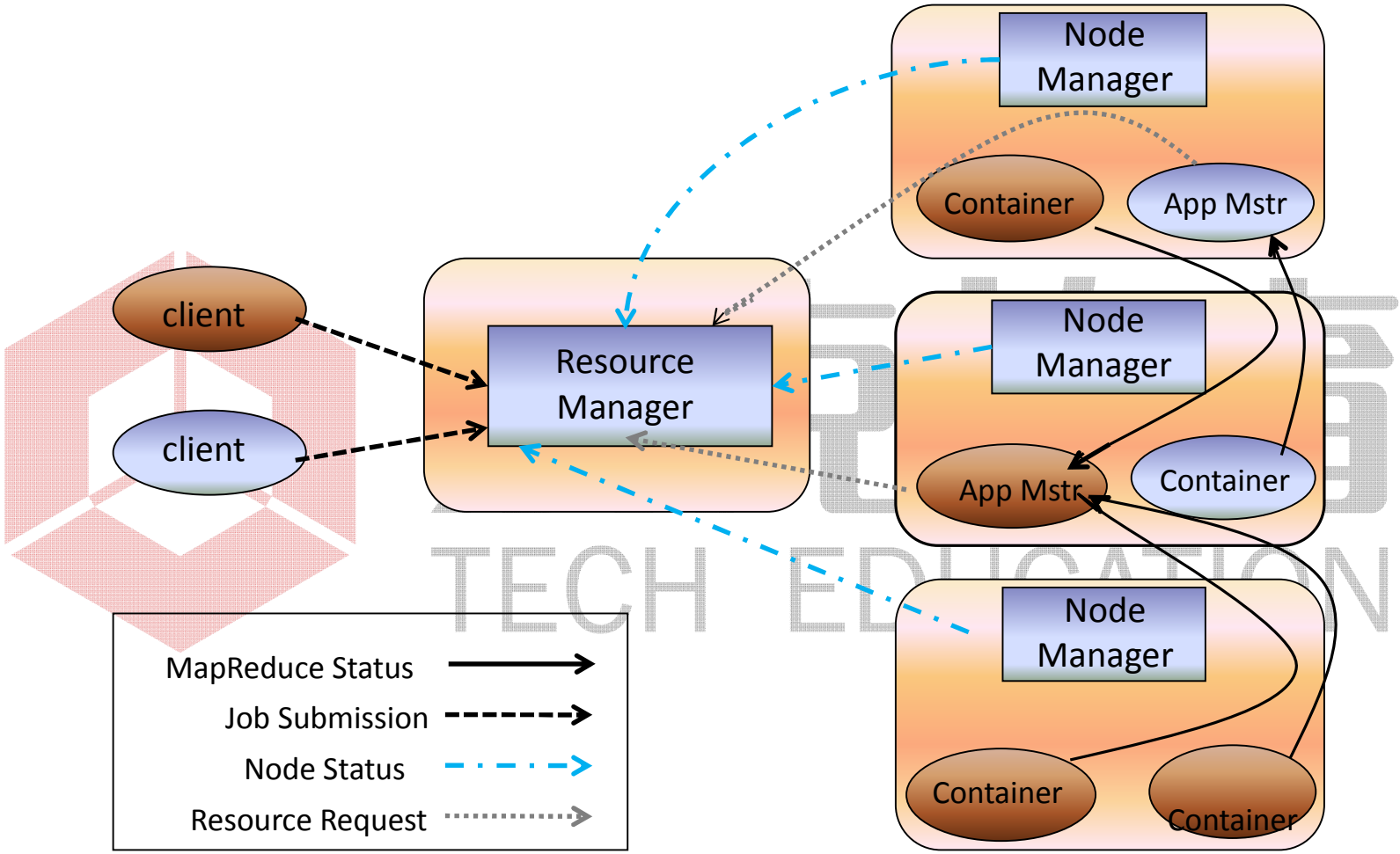
WordCount的Map过程



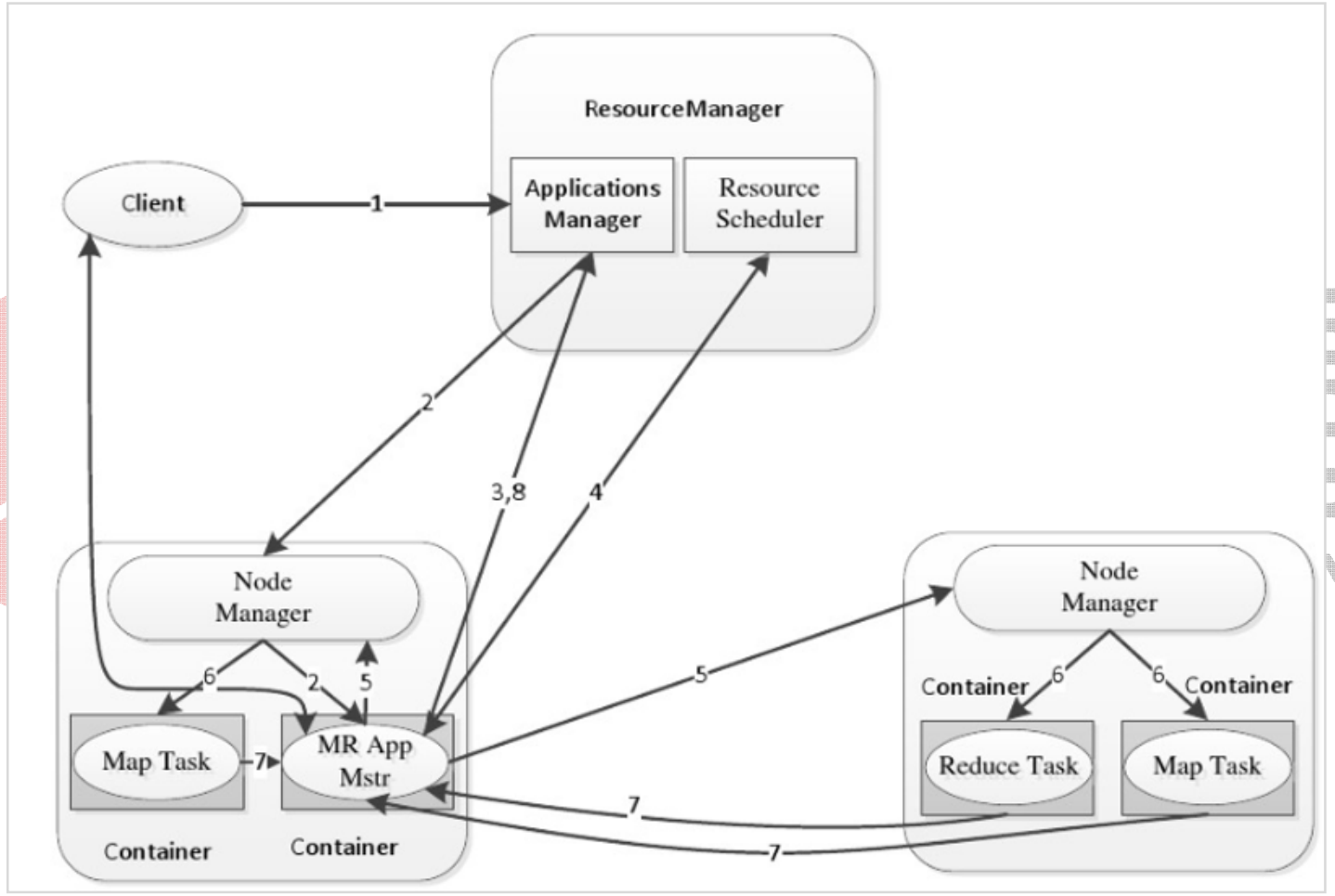
WordCount的Reduce过程



YARN的组件架构



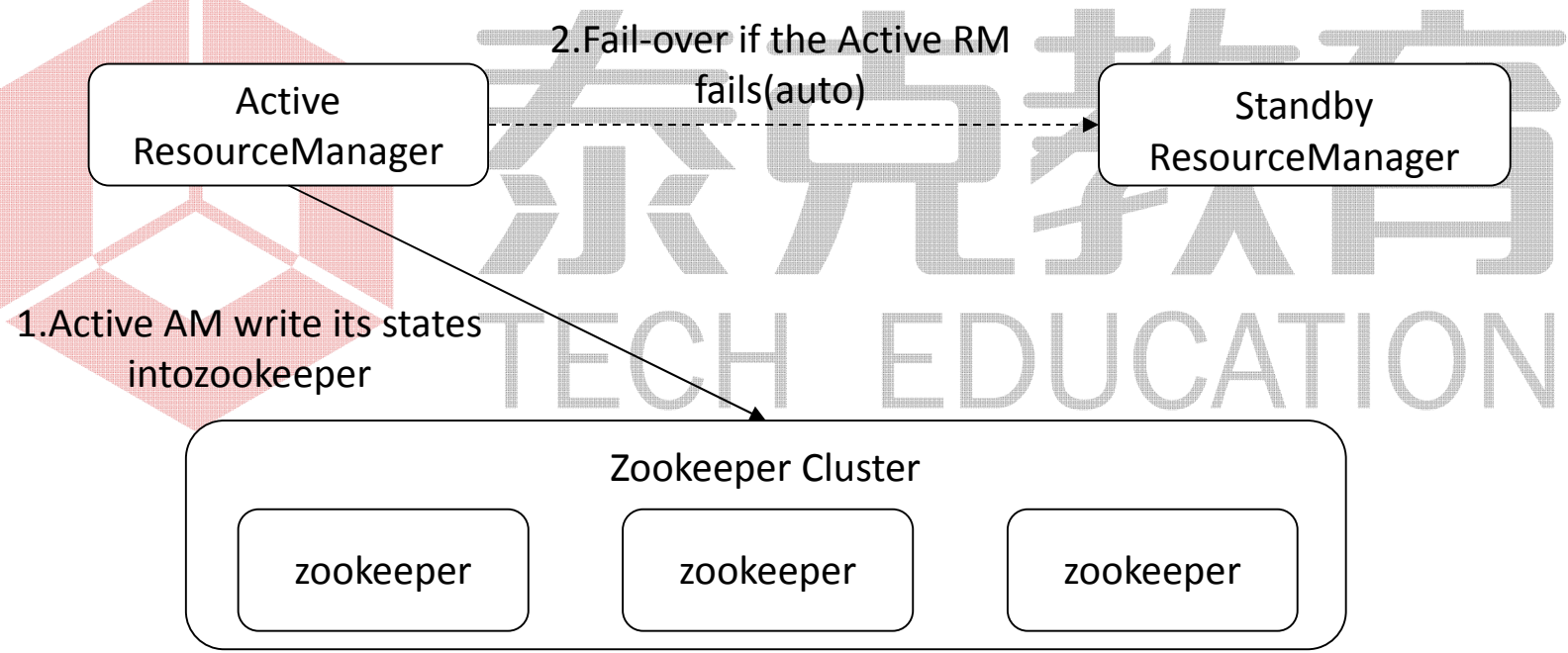
MapReduce On YARN任务调度流程



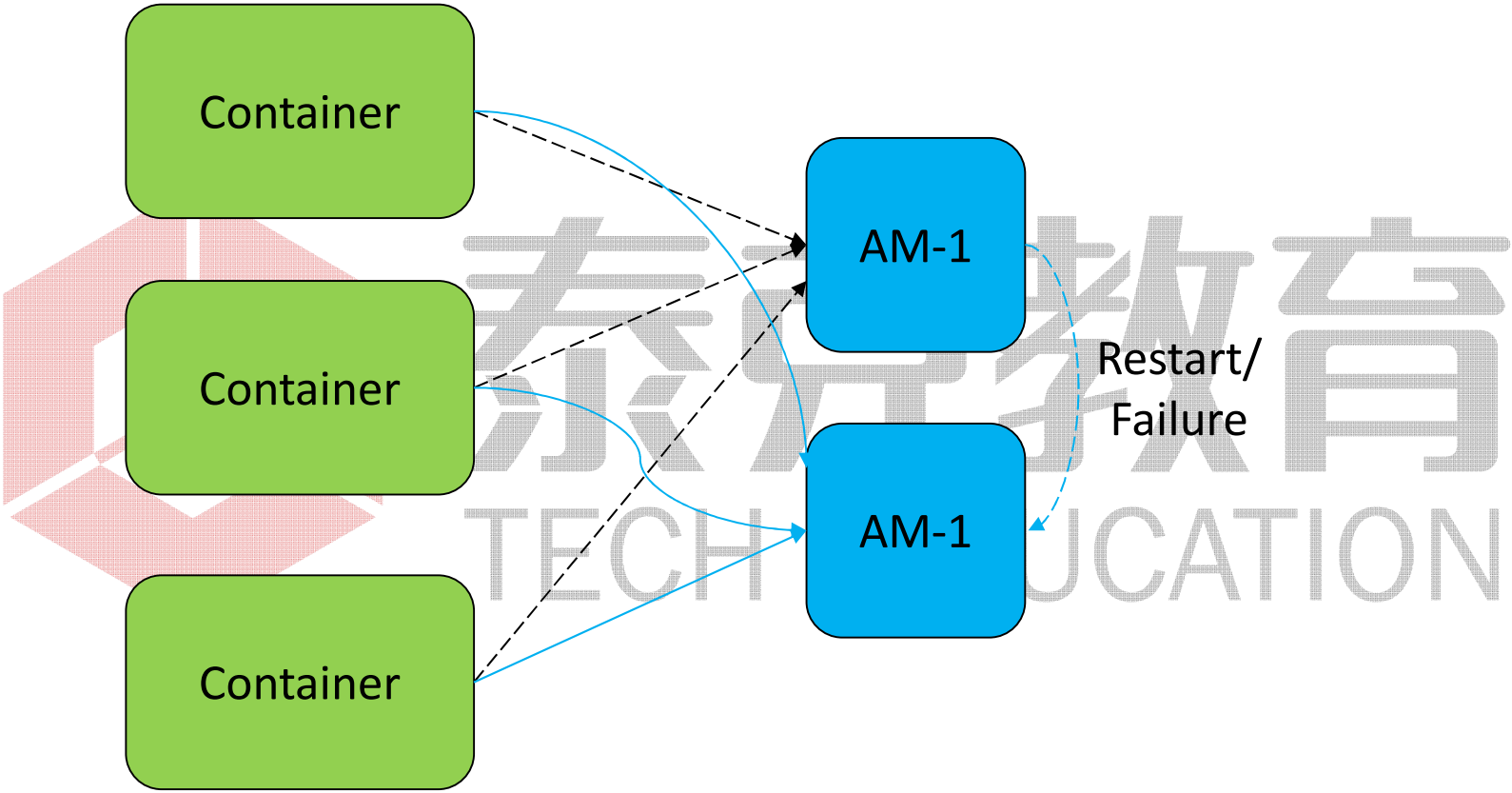
HUAWEI

YARN HA方案

- YARN中的ResourceManager负责整个集群的资源管理和任务调度，YARN高可用性方案通过引入冗余的ResourceManager节点的方式，解决了ResourceManager 单点故障问题。



YARN APPMaster容错机制

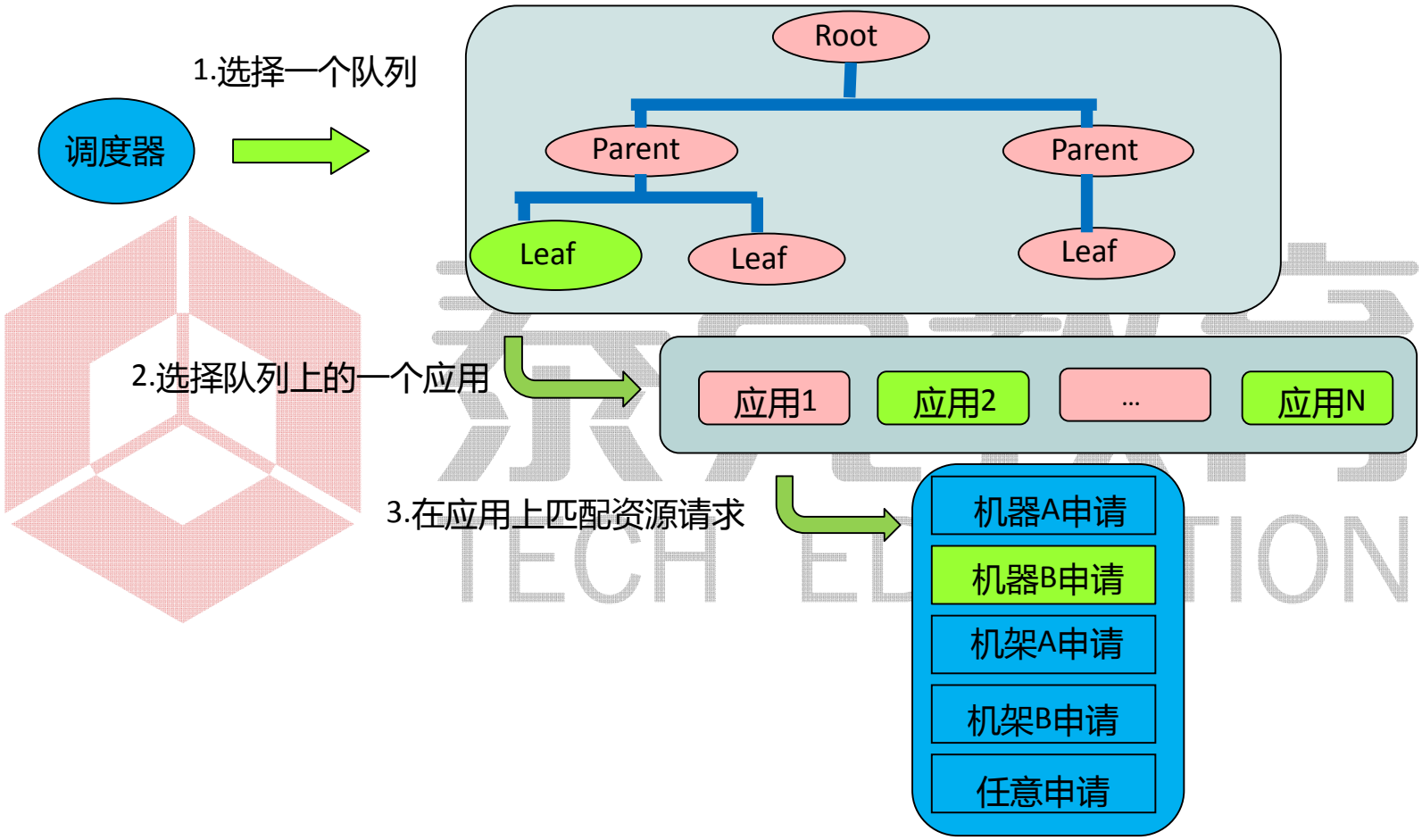


目录

资源管理

- 当前YARN支持内存和CPU两种资源类型的管理和分配。
- 每个NodeManager可分配的内存和CPU的数量可以通过配置选项设置（可在YARN服务配置页面配置）。
  - Yarn.nodemanager.resource.memory-mb
  - Yarn.nodemanager.vmem-pmem-ratio
  - Yarn.nodemanager.resource.cpu-vcore

资源分配模型



容量调度器的介绍

- 容量调度器使得Hadoop应用能够共享的、多用户的、操作简便的运行在集群上，同时最大化集群的吞吐量和利用率。
- 容量调度器以队列为单位划分资源，每个队列都有资源使用的下限和上限。每个用户可以设定资源使用上限。管理员可以约束单个队列、用户或作业的资源使用。支持作业优先级，但不支持资源抢占。

容量调度器的特点

- **容量保证**：管理员可为每个队列设置资源最低保证和资源使用上限，所有提交到该队列的应用程序共享这些资源。
- **灵活性**：如果一个队列中的资源有剩余，可以暂时共享给那些需要资源的队列，当该队列有新的应用程序提交，则其他队列释放的资源会归还给该队列。
- **支持优先级**：队列支持任务优先级调度（默认是FIFO）。
- **多重租赁**：支持多用户共享集群和多应用程序同时运行。为防止单个应用程序、用户或者队列独占集群资源，管理员可为之增加多重约束。
- **动态更新配置文件**：管理员可根据需要动态修改配置参数，以实现在线集群管理。

容量调度器的任务选择

- 调度时，首先按以下策略选择一个合适队列：
  - 资源利用量最低的队列优先，比如同级的两个队列Q1和Q2，它们的容量均为30，而Q1已使用10，Q2已使用12，则会优先将资源分配给Q1。
  - 最小队列层级优先，例如：QueueA与QueueB.childQueueB，则QueueA优先。
  - 资源回收请求队列优先。
- 然后按以下策略选择该队列中一个任务：
  - 按照任务优先级和提交时间顺序选择，同时考虑用户资源量限制和内存限制。

队列资源限制（1）

- 队列的创建是在多租户页面，当创建一个租户关联YARN服务时，会创建同名的队列。比如先创建QueueA,QueueB两个租户，即对应YARN两个队列。

队列资源限制（2）

- 队列的资源容量（百分比），有default、QueueA、QueueB三个队列，每个队列都有一个[队列名].capacity配置：
  - Default队列容量为整个集群资源的20%。
  - QueueA队列容量为整个集群资源的10%。
  - QueueB队列容量为整个集群资源的10%，后台有一个影子队列root-default使队列之和达到100%。

资源分配

租户名（队列）	资源容量
QueueA(root.QueueA)	10%
QueueB(root.QueueB)	10%
TestParent(root.TestParent)	10%
testchild(root.TestParent.testchild)	10%

队列资源限制（3）

- **共享空闲资源**
  - 由于存在资源共享，因此一个队列使用的资源可能超过其容量（例如QueueA.capacity），而最大资源使用量可通过参数限制。
  - 如果某个队列任务较少，可将剩余资源共享给其他队列，例如QueueA的maximum-capacity配置为100，假设当前只有QueueA在运行任务，理论上QueueA可以占用整个集群100%的资源。

用户限制和任务限制

- 用户限制和任务限制的参数可通过“租户管理”>“动态资源计划”>“队列配置”进行配置。

用户限制(1)

- 每个用户最低资源保障（百分比）：
  - 任何时刻，一个队列中每个用户可使用的资源量均有一定的限制，当一个队列中同时运行多个用户的任务时，每个用户的可使用资源量在一个最小值与最大值之间浮动，其中，最大值取决于正在运行的任务数目，而最小值则由minimum-user-limit-percent决定。
  - 例如，设置队列A的这个值为25，即Yarn.scheduler.capacity.root.QueueA.minimum-user-limit-percent=25，那么随着提任务的用户增加，队列资源的调整如下：

第1个用户提交任务到QueueA	会获得QueueA的100%资源。
第2个用户提交任务到QueueA	每个用户会最多获得50%的资源。
第3个用户提交任务到QueueA	每个用户会最多获得33.33%的资源。
第4个用户提交任务到QueueA	每个用户会最多获得25%的资源。
第5个用户提交任务到QueueA	为了保障每个用户最低能获得25%的资源，第5个用户将无法再获取到QueueA的资源，必须等待资源的释放。

用户限制(2)

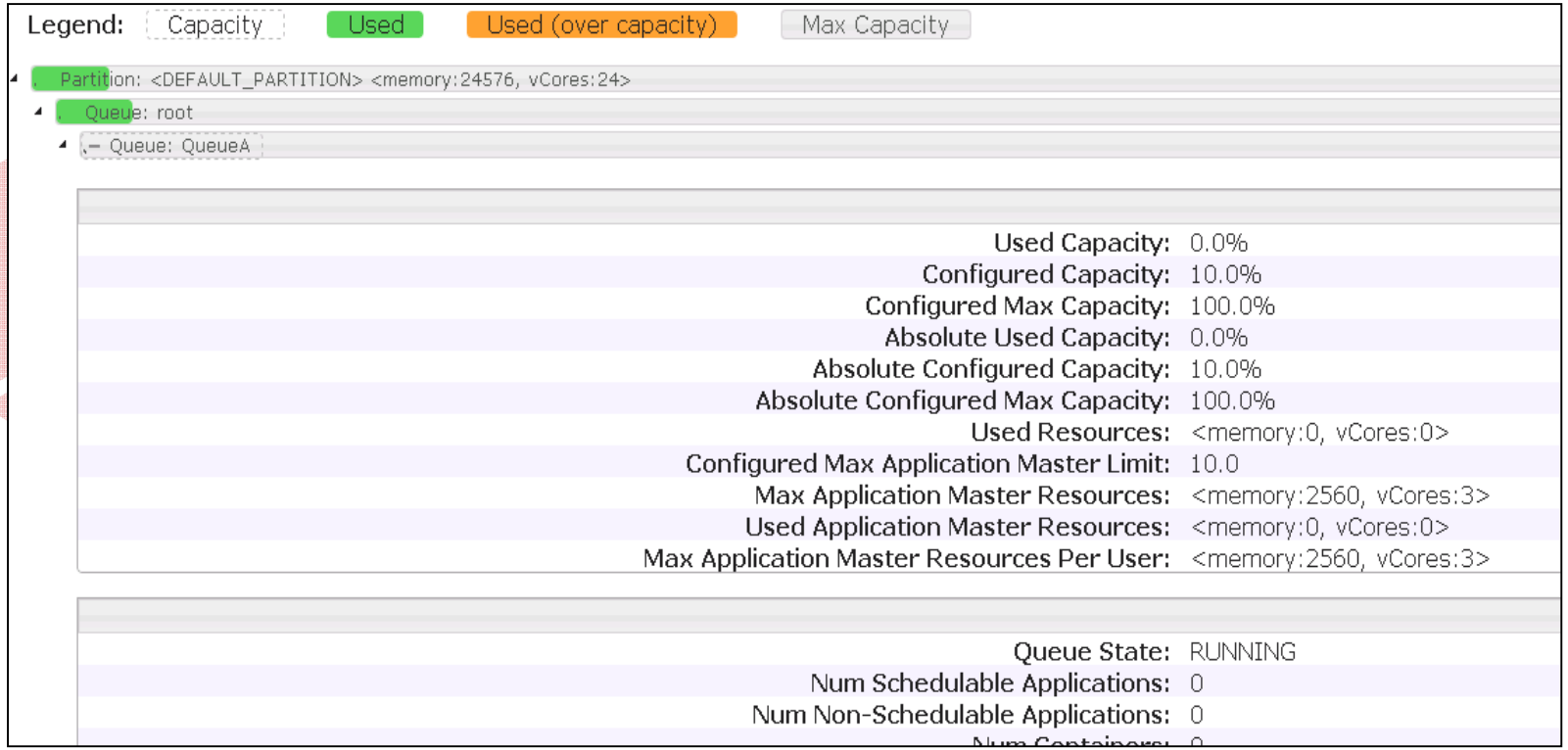
- 每个用户最多可使用的资源量（所在队列容量的倍数）：
  - queue容量的倍数，用来设置一个user可以获取更多的资源。Yarn.scheduler.capacity.root.QueueD.user-limit-factor=1。默认值为1，表示一个user获取的资源容量不能超过queue配置的capacity，无论集群有多少空闲资源，最多不超过maximum-capacity。

任务限制

- 最大活跃任务数：
  - 整个集群中允许的最大活跃任务数，包括运行或挂起状态的所有任务，当提交的任务申请数据达到限制以后，新提交的任务将会被拒绝。默认值10000。
- 每个队列最大任务数：
  - 对于每个队列，可以提交的最大任务数，以QueueA为例，可以在队列配置页面配置，默认是1000，即此队列允许最多1000个活跃任务。
- 每个用户可以提交的最大任务数：
  - 这个数值依赖每个队列最大任务数。根据上面的数据， QueueA最多可以提交1000个任务，那么对于每个用户而言，可以向QueueA提交的最大任务数为1000*用户最低资源保障率（假设25%）* 用户可使用队列资源的倍数(假设1)。

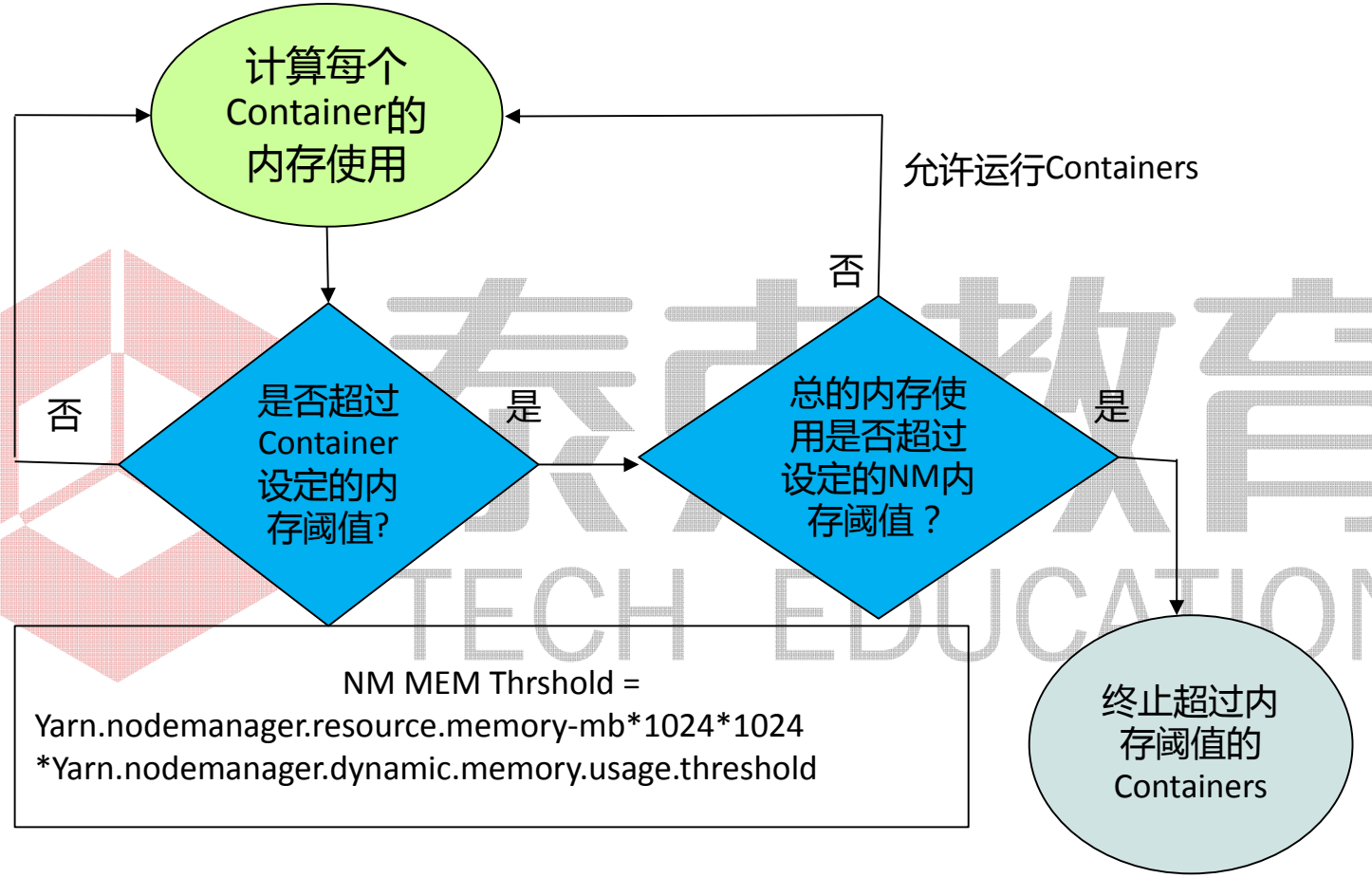
查看队列信息

- 队列的信息可以通过YARN webUI进行查看，进入方法是“服务管理”>“YARN”>“ResouceManager（主）”>“Scheduler”。

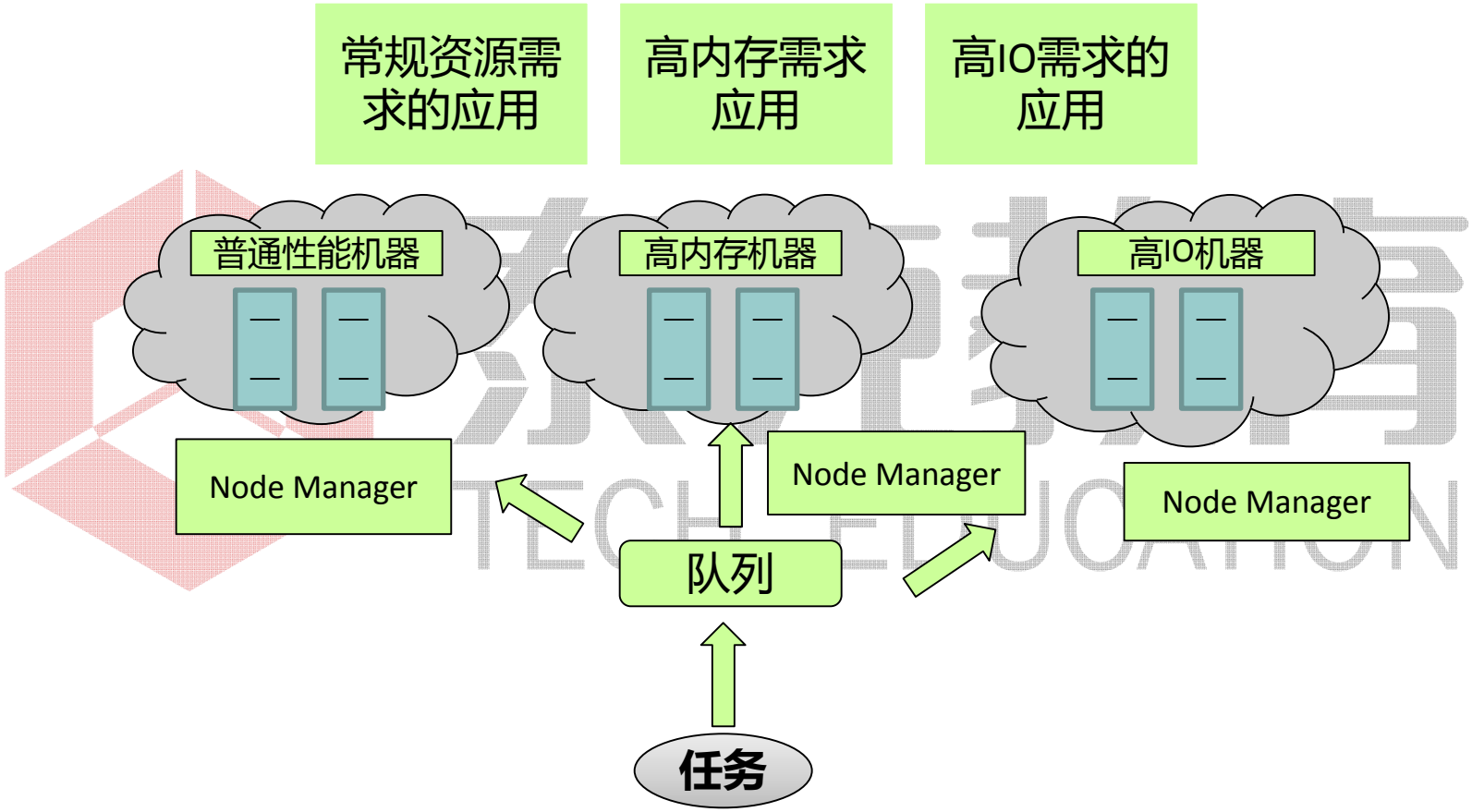


目录

增强特性 – YARN动态内存管理



增强特性 – YARN基于标签调度

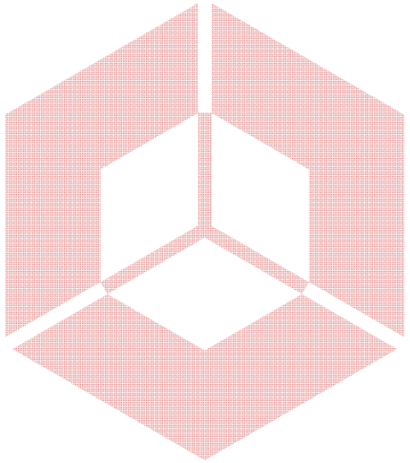


本章总结

- 本章首先讲述了MR和YARN的应用场景和基本架构，然后讲解了YARN资源管理与任务调度的原理，最后介绍了华为FusionInsight HD中YARN的增强特性。

思考题

1. 请简述MapReduce的工作原理。
2. 请简述YARN的工作原理。



思考题

1. 下面哪些是MapReduce的特点？（　　）

   A. 易于编程

   B. 良好的扩展性

   C. 实时计算

   D. 高容错性

思考题

2. YARN中资源抽象用什么表示？（　　）

A. 内存

B. CPU

C. Container

D. 磁盘空间

思考题

3. 下面哪个是MapReduce适合做的？（　　）

A. 迭代计算

B. 离线计算

C. 实时交互计算

D. 流式计算

思考题

4. 容量调度器有哪些特点？（　　）

A. 容量保证

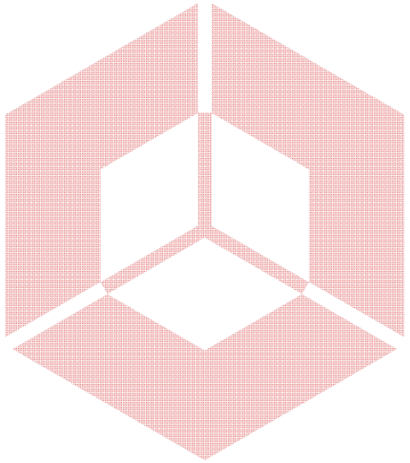
B. 灵活性

C. 多重租赁

D. 动态更新配置文件

更多信息

- 下载培训资料：
  - http://support.huawei.com/learning/trainFaceDetailAction?lang=zh&pbiPath=term1000025185&courseId=Node1000009072
- eLearning课程：
  - http://support.huawei.com/learning/nodeQueryAction!loadTrainProjectInfo?lang=zh&pbiPath=term1000025185&courseId=Node1000009421&navId=MW000001
- 考试大纲：
  - http://support.huawei.com/learning/Certificate!toExamOutlineDetail?lang=zh&nodeId=Node1000003516
- 模拟考试：
  - http://support.huawei.com/learning/Certificate!toSimExamDetail?lang=zh&nodeId=Node1000004285
- 认证流程：
  - http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_40



谢谢

www.huawei.com