一页纸精华 | MapReduce

原创 2016-02-18 牛家浩 中兴大数据

>>>> 这是 **中兴大数据** 第**210**篇原创文章

要入门大数据，最好的办法就是理清Hadoop的生态系统。中兴大数据公众号将推出“一页纸精华”栏目，将用最精炼的语言，陆续为你介绍Hadoop生态系统的各个组件。本期为你介绍Hadoop分布式计算框架MapReduce。

MapReduce是为了解决传统HPC框架在面对海量数据时扩展困难而产生的。

MapReduce致力于解决大规模数据处理的问题，利用局部性原理将整个问题分而治之。MapReduce集群由普通PC机构成，为无共享式架构。在处理之前，将数据集分布至各个节点。处理时，每个节点就近读取本地存储的数据处理(Map)，将处理后的数据进行合并(Combine)、排序(Shuffle and Sort)后再分发(至Reduce节点)，避免了大量数据的传输，提高了处理效率。无共享式架构的另一个好处是配合复制(Replication)策略，集群可以具有良好的容错性，一部分节点的宕机对集群的正常工作不会造成影响。

MapReduce提供了一种并行计算的模型，

其优点是：

- **易于编程**：将所有并行程序均需要关注的设计细节抽象成公共模块并交由系统实现，而用户只需专注于自己的应用程序逻辑实现，这样简化了分布式程序设计且提高了开发效率。
- **良好的扩展性**：通过添加机器以达到线性扩展集群能力的目的。
- **高容错性**：在分布式环境下，随着集群规模的增加，集群中的故障率（这里的“故障”包括磁盘损坏、机器宕机、节点间通信失败等硬件故障和坏数据或者用户程序Bug产生的软件故障）会显著增加，进而导致任务失败和数据丢失的可能性增加。 Hadoop通过计算迁移或者数据迁移等策略提高集群的可用性与容错性。

其不足在于：

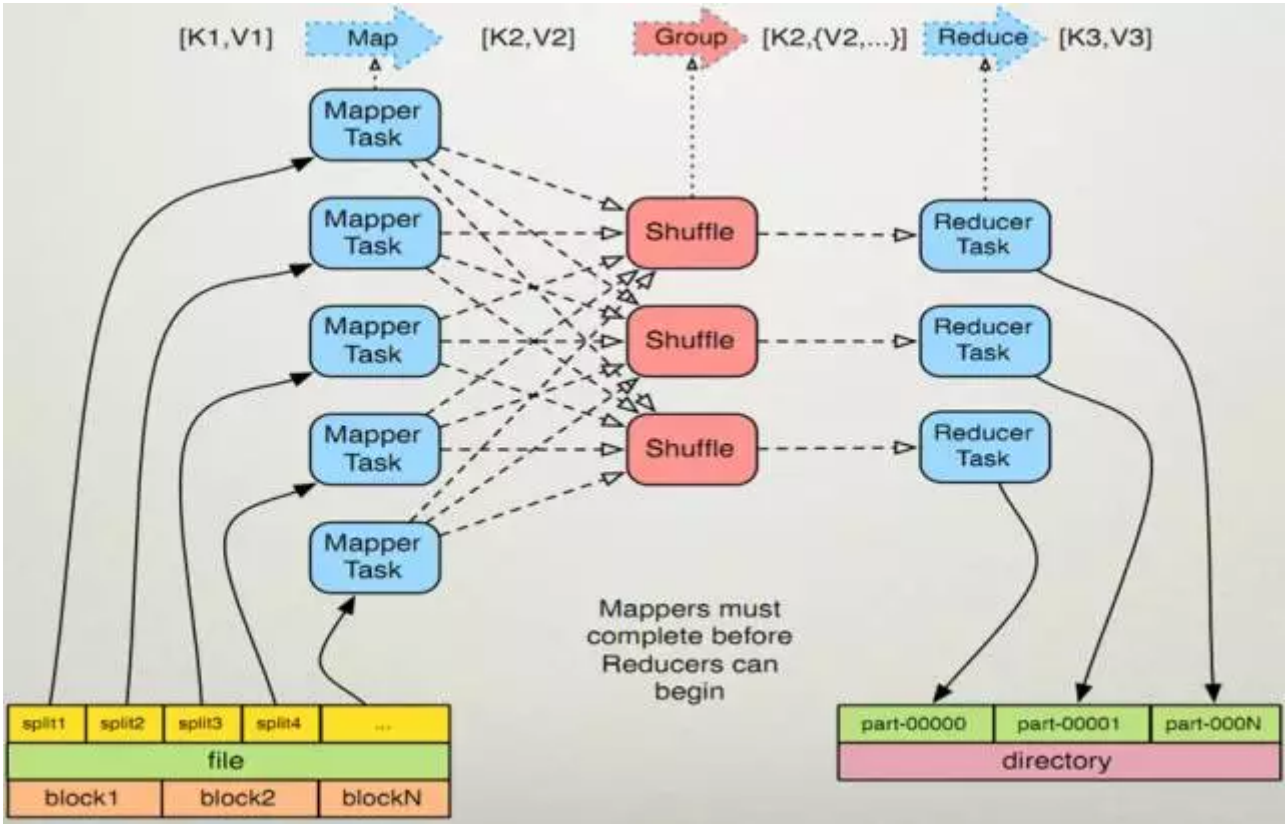
- **延时较高**：不适应实时应用的需求。
- **对随机访问的处理能力不足**： 其是一种线性的编程模型。适用于顺序处理数据。

MapReduce作为一个分布式计算框架，主要由三部分组成：

1. **编程模型**：为用户提供了非常易用的编程接口，用户只需要考虑如何使用MapReduce模型描述问题，实现几个简单的hook函数即可实现一个分布式程序；
2. **数据处理引擎**：由MapTask和ReduceTask组成，分别负责Map阶段逻辑和Reduce阶段逻辑的处理；
3. **运行时环境**：用以执行MapReduce程序，并行程序执行的诸多细节，如分发、合并、同步、监测等功能均交由执行框架负责，用户无须关心这些细节。

MapReduce可编程组件

MapReduce提供了5个可编程组件，如下图所示，实际上可编程组件全部属于回调接口。当用户按照约定实现这几个接口后，MapReduce运行时环境会自动调用以实现用户定制的效果。 `



MapReduce可编程组件

1. **InputFormat**：主要用于描述输入数据的格式，其按照某个策略将输入数据切分成若干个Split，并为Mapper提供输入数据，将Split解析成一个个Key/Value对。
2. **Mapper**：对Split传入的key1/value1对进行处理，产生新的键值key2/value2对。 即Map：(k1，v1) → (k2，v2) 。
3. **Partitioner**：作用是对Mapper产生的中间结果进行分区，以便将Key有耦合关系的数据交给同一个Reducer处理，它直接影响Reduce阶段的负载均衡。
4. **Reducer**：以Map的输出作为输入，对其进行排序和分组，再进行处理产生新的数据集。即Reducer：(k2，list(v2)) → (k3， v3)。
5. **OutputFormat**：主要用于描述输出数据的格式，它能够将用户提供的Key/Value对写入特定格式的文件中。

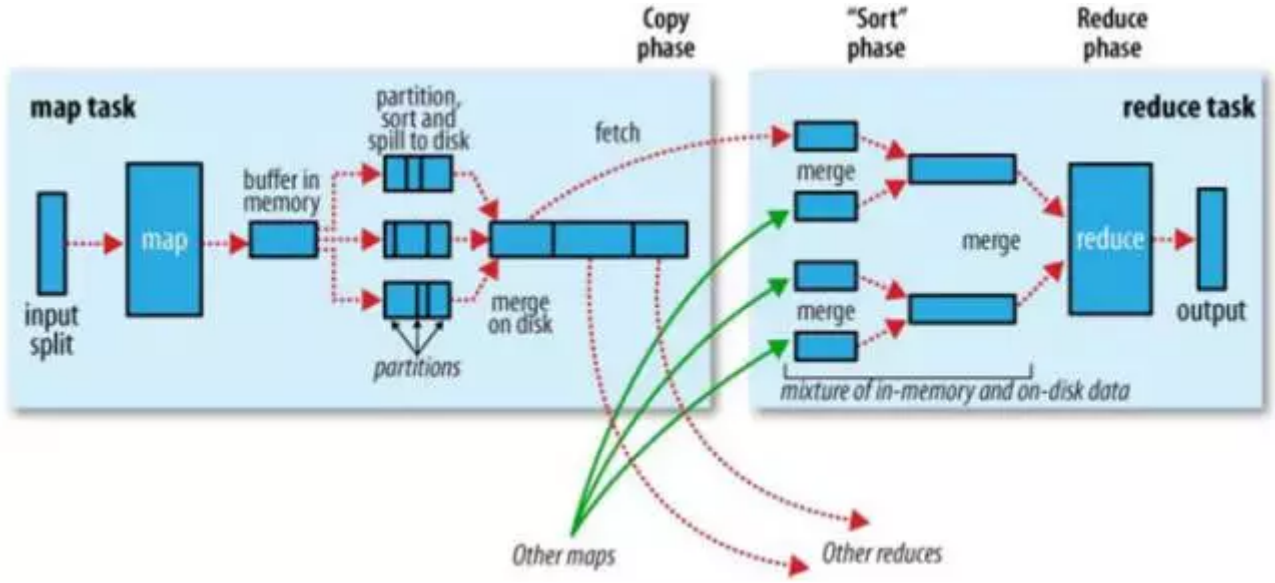
编程流程的运行流程如下：

1. 作业提交后 InputFormat按照既定策略将输入数据切分成若干个Split；
2. 各Map任务节点上根据分配的Split元信息获取相应数据并将其迭代解析成一个个key1/value1对；
3. 迭代的key1/value1对由Mapper处理为新的key2/value2对；

4. 新的key2/value2对先进行排序，然后由Partitioner将有耦合关系的数据分到同一个Reducer上进行处理，中间数据存入本地磁盘；
5. 各Reduce任务节点根据到已有的Map节点上远程获取数据(只获取属于该Reduce的数据，该过程称为Shuffle)；
6. 对数据进行排序，并进行分组(将相同Key的数据分为一组)；
7. 迭代Key/Value对，并由Reducer合并处理为新的key3/value3对；
8. 新的key3/value3对由OutputFormat保存到输出文件中。

MapReduce数据处理引擎

在MapReduce计算框架中， 一个Job被划分成Map和Reduce两个计算阶段，它们分别由多个Map Task和Reduce Task组成。 这两种服务构成了MapReduce数据处理引擎。如下图所示：



MapReduce处理流程图

MapTask的整体计算流程共分为5个阶段：

1. **Read阶段**：MapTask通过用户编写的RecordReader，从输入InputSplit中解析出一个个Key/Value；
2. **Map阶段**：将解析出的Key/Value交给用户编写的Map函数处理，并产生一系列新的Key/Value；
3. **Collect阶段**： Map函数生成的Key/Value通过调用Partitioner进行分片,并写入一个环形内存缓冲区中；

4. **Spill阶段**：即"溢写"，当环形缓冲区满后，MapReduce会将数据写到本地磁盘上, 生成一个临时文件；
5. **Combine阶段**：所有数据处理完成后，MapTask对所有临时文件进行一次合并, 以确保最终只会生成一个数据文件。

Reduce Task的整体计算流程共分为5个阶段：

1. **Shuffle阶段**：Reduce Task从各个Map Task上远程拷贝一片数据，并针对某一片数据，如其大小超过一定阈值则写到磁盘上，否则直接放到内存中；
2. **Merge阶段**：在远程拷贝数据的同时，Reduce Task启动了三个后台线程对内存和磁盘上的文件进行合并，以防止内存使用过多或磁盘上文件过多；
3. **Sort阶段**：采用了基于排序的策略将Key相同的数据聚在一起.由于各个Map Task已经实现对自己的处理结果进行了局部排序，因此Reduce Task只需对所有数据进行一次归并排序即可；
4. **Reduce阶段**：将每组数据依次交给用户编写的Reduce函数处理；
5. **Write阶段**：Reduce函数将计算结果写到HDFS上。

MapReduce版本对比

MapReduce 主要分为两个大版本 MRv1和 MRv2。之所以出现MRv2， 是因为MRv1具有如下的局限性：

- **扩展性差**：在MRvl中，JobTracker同时兼备了资源管理和作业控制两个功能，这成为系统的一个最大瓶颈，严重制约了 Hadoop集群扩展性(业界总结出MRv1只能支持4000节点主机的上限)；
- **可靠性差**： MRvl采用了 Master/Slave结构。其中，Master存在单点故障问题，一旦它出现故障将导致整个集群不可用；
- **资源利用率低**：MRvl采用了基于槽位的资源分配模型，槽位是一种粗粒度的资源划分单位，通常一个任务不会用完槽位对应的资源，且其他任务也无法使用这些空闲资源。此外，Hadoop将槽位分为Map Slot和Reduce Slot两种，且不允许它们之间共享，常常会导致一种槽位资源紧张而另外一种闲置(比如一个作业刚刚提交时，只会运行Map Task，此时Reduce Slot闲置）；
- **无法支持多种计算框架**：随着互联网高速发展，MapReduce这种基于磁盘的离线计算框架已经不能满足应用要求，从而出现了一些新的计算框架，包括内存计算框架、流式计算框架和迭代式计算框架等，而MRvl不能支持多种计算框架并存。

新旧MR的对比如下表所示：

	MRvl	MRv2
编程模型	新旧 API	新旧 API
数据处理引擎	MapTask / ReduceTask	MapTask / ReduceTask (重构优化)
运行时环境	由（一个）JobTracker 和（若干）TaskTracker 构成：JobTracker 负责资源管理和所有作业的控制，而 TaskTracker 负责接收来自 JobTracker 的命令并执行它。	YARN（由 ResourceManager 和 NodeManager 构成）和 MRAppMaster 构成：YARN 提供一个资源管理和调度的平台，而 MRAppMaster 作为运行在 YARN 资源管理平台上的一个应用，仅负责一个作业的管理。

MRV1与MRV2对比表

简言之，MRvl仅是一个独立的离线计算框架， 而MRv2则是运行于YARN之上的MapReduce应用， 每个作业都有一个应用ApplicationMaster。

中兴通讯大数据团队广招贤才啦！

招聘信息

招聘岗位： WEB软件开发工程师

任职要求：

1. 全日制重点本科及以上学历，硕士两年以上的工作经验，本科三年以上相关经验。
2. 熟练掌握JAVA语言。
3. 熟练掌握WEB开发技术（HTML/CSS/JAVASCRIPT等），了解较前沿的WEB技术（如HTML5等）。
4. 能够使用主流的WEB框架和组件（Struts,Ajax,jQuery,YUI等）进行设计和开发。
5. 良好的沟通协作能力，能阅读一般的英文技术文章。
6. 有WEB应用项目、中大型网站开发经验者优先。

主要职责：

1. 负责智慧城市等政企应用软件的开发、设计工作。
2. 参与和用户的沟通、交流等工作。

简历请投至：xie.haibo@zte.com.cn