Computer System Networking and Telecommunications
计算机系统网络和电信，2019,1(2)
ISSN:2661-3719 (Print); 2661-3727 (Online)

Big Data Analysis Technology based on Spark

Rong WANG Zhengde BAO Chenxi LI
School of Computer and Software, Jincheng College, Sichuan University, Chengdu, 611731

Abstract

The rapid development of mobile Internet and Internet of things technology has enriched people's information acquisition methods, increased the number of network information dissemination and improved the speed of information exchange. With the development of distributed technology, the problem of mass data storage and management has been solved by distributed file system. Spark is a kind of data processing and analysis technology for in-memory computing. Its calculation model of RDD can do both MapReduce and iterative computing. In addition to being able to process and analyze large amounts of data, Spark can also be used in areas such as streaming computing and machine learning, creating more possibilities for the development of these areas.

Key Words

Big Data, Spark, Hadoop, Kafka, RDD

DOI:10.18686/jsjxt.v1i2.691

基于 Spark 的大数据分析技术

王 溶 鲍正德 李晨曦
四川大学锦城学院计算机与软件学院，四川成都，611731

摘 要

移动互联网和物联网技术的快速发展丰富了人们的信息获取方式和增加了网络信息传播数量以及提升了信息交流速度。在分布式技术逐渐成熟的今天，海量数据的存储管理难题已经通过分布式文件系统得到良好的解决。Spark 就是在内存计算的一种数据处理和分析的一种技术，它的 RDD 的计算模型可以同时做到 MapReduce 和迭代型计算。除了可以对大量数据进行处理和分析，Spark 还能够用于流式计算和机器学习等领域，为这些领域的发展创建了更多的可能性。

关键词

大数据，Spark，Hadoop，Kafka，RDD

1.引言

在今天这样一个大数据火热盛行的时候，数据每天都在呈指数级增长，因此数据成为了任何组织的根基。为了公司更好的发展，势必会通过数据获取价值来增长他们公司的利益，故这就成了现在大数据不容错过的必须要处理的问题。数十年前的组织很少产生拍字节级（PB）甚至太字节级（TB）的数据，所以当时的数据库不能解决如今大量数据的处理。现今的大部分公司由于产生了数太字节的数据，并且数据在日益增长，因此会更迫切的寻求新的技术以达到他们所产生数据的处理和分析的目的，而这也是大数据现如今这样火的原因。

2.与 Spark 有关的大数据技术

技术持续更替，因此有一些技术和 Spark 技术的某些功能是一样的故它们可以一同运用或 Spark 取代其中的技术。如：Hadoop，Spark 技术能取代 Hadoop 中的 MapReduce 进行复杂的数据算法并且要更高效和 Spark 中的 Spark SQL 可以取代 Hive。

2.1 Hadoop

Hadoop 由 Apache 基金会所开发的分布式系统基础架构是是一个可扩展、可容错用来处理跨越集群的大数据集的系统且是最早流行的开源大数据技术之一。

Hadoop 的优点：可靠性，Hadoop 在计算的过程中会通过维护多个工作的数据副本来解决故障问题，当计算元素和存储失败时，Hadoop 会将失败的节点上的数据重新复制到另一个节点上；高效性，因为是并行处理的工作方式所以速度会比串行的要更快；可伸缩性，它能够处理 PB 级数据也可以处理和分析小量的数据即 KB 级；成本低，Hadoop 是依赖于社区服务的一种开源的大数据技术；方便性，Hadoop 可能在最初设计时就想过要在 Linux 平台上运行，所以它有 Java 语言编写的代码框架并可以用其他语言编写，[5]如 C++。故处理海量数据的应用程序用户可以轻松地进行开发和运行。

开发者希望 Hadoop 能够存储和计算故 HDFS 和 MapReduce 是 Hadoop 框架最核心的设计。前者提供了存储能力，后者提供了计算能力。[1]

2.1.1 HDFS

HDFS 是一个分布式文件系统包含 NameNode 和 DataNode 并且根据工作的原理 NameNode 只有一个而 DataNode 有多个。NameNode 是一个负责管理文件系统名称空间和控制外部客户机的访问的运行在 HDFS 实例中单独机器上的软件并且有规律性地接受 DataNode 发送的心跳信号和块状态报告，如每隔 10 秒，当过了时间 NameNode 没有收到心跳则会认为 DataNode 出现故障死亡则会采取修复措施来复制数据。块状态报告包含数据块属于哪个文件，数据块的编号和修改的时间等。DataNode 存储以文件块的形式的实际的文件内容并且接收 HDFS 客户机的读写请求和 NameNode 的创建、删除、复制块的命令是通过一个交换机将所有系统连接起来的机架的形式组织。

HDFS 读文件的过程：

当客户端到 HDFS 去读取文件时，NameNode 会先对客户端的访问回应组成文件的所有对应文件块数据的位置，然后 DataNode 会根据客户端的读请求顺序来返回数据。这时各个 DataNode 之间是没有联系的。

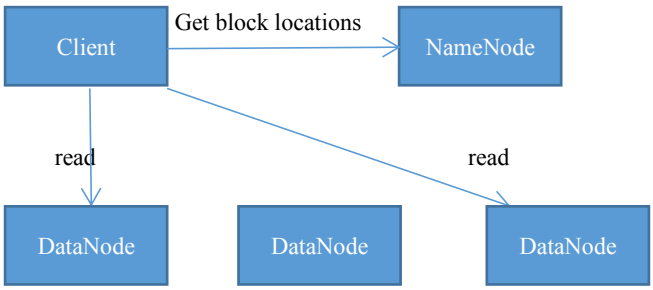

图 1

HDFS 写文件的过程：

当写数据到 HDFS 时，NameNode 会先回应客户端的请求通过检查文件是否已创建或者客户端是否有权限来执行是否创建新文件的操作。在创建好新文件之后，客户端会根据收到的 NameNode 发过来的分配好的 DataNode 的编号依次写入数据但当 DataNode 不够文件的存储时 NameNode 会响应要求再分配。这时各个 DataNode 之间其实是有一个通道的。

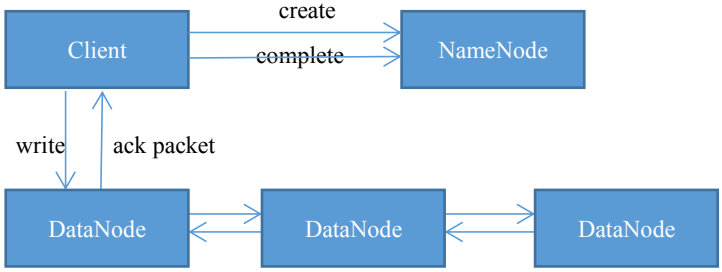

图 2

2.1.2 MapReduce

将文件中的数据作为键值对输入到 map 函数中并根据所写代码的不同的功能输出结果，Hadoop 会自动的对结果进行排序并分组。最后将这些数据传入 reduce 函数。reduce 函数会将这些数据根据性质加起来并输出。以上是 MapReduce 中的 map 和 reduce 函数的数据算法的工作原理。因此 map 和 reduce 函数一起就组成了分布式计算引擎且能够自动在集群中各计算机上调度应用。MapReduce 可以说是 Spark 的前身。

2.2 分布式消息系统：Kafka

Kafka 和 Spark 之间的关系是 Kafka 可以和 Spark 中的 Spark Streaming 组件一起来构建实时数据处理系统。

Kafka 是有高吞吐量和持久性等特性的可以作为发布-订阅式消息的分布式的、可划分、重复的提交日志服务的一种系统。[2]它的关键特性：高吞吐量，一个代理可以处理大量数据的读写；可扩展性，在集群上增加更多的节点来扩大容量；持久性，可以在硬盘上保存消息且不会定期的被删除。Kafka 的架构是显示分布式架构且简单和有很多个缓存代理、生产者和消费者。生产者和消费者这两个角色是用于实现 kafka 注册功能的接口，数据由生产者发出，中间会经过代理，这时代理可能会对数据进行缓存和分发，最后到达消费者。简单明了、超高性能且与编程语言无关的 TCP 协议用于作为 kafka 之间数据的传递。

3.Spark 概述

Spark 是一个集群计算框架且基于内存用于处理、分析大数据。[4]设计者为了让应用程序开发者方便使用所以在其中加入了一套简单的编程接口来对集群节点的 CPU、内存和存储资源进行管理。

3.1 Spark 的主要特点

1）使用方便

当计算复杂的数据算法时只用 map 和 reduce 这两个操作符不能完全的进行计算会有 bug。为了让 Spark 能够处理复杂的数据因此设计者为此专门提供了丰富的 API 和 80 多个用于处理数据的操作符来使得算法变得更加简单且易理解。在使用 Hadoop 处理数据时会用到很多的模块代码，使得编写困难，因此设计者考虑到这点所以消除了模块代码，使得开发者不用再为此为烦恼提升了工作效率。

2）快速

Spark 可以使用基于内存的集群计算和它实现了更先进的执行引擎，这会让 Spark 比 Hadoop 更快速。

Spark 会在计算过程中缓存数据。计算过程中的结果会直接存储在内存中，这样进行下一步的计算时无需从磁盘进行读取，即只需要从硬盘读取数据一次即可。因为减少了 I/O 延迟，所以作业总的执行时间也减少了。

Hadoop 在对数据进行处理时，会对其通过 map 和 reduce 创建有向无环图，当算法太复杂不能通过一个 DAG 完成时会对作业进行划分，再创建 DAG 然后顺序执行。Spark 也是一样将作业转化成 DAG，不同的是，当数据处理算法太复杂时，不会划分多个作业，可以一次执行。因此速度更快。

3）通用

在研发 Spark 时，设计者可能考虑 Spark 会用于机器学习等领域所以其中包含了批处理、交互分析等库，这就相当于一个统一的集成平台。因此 Spark 在处理流水线时可用单一的框架来实现包含多个不同类型任务。

4）可扩展

在集群上增加节点来提升 Spark 的集群的数据处理的能力。

5）容错

在一个集群的数百个节点中，硬盘的损坏可能会导致节点出现系统故障，而且可能性还很高。在 Spark 中，这些问题并不需要太过于担心，它会自动处理节点的故障而这时出现的问题会导致性能的下降却不会使应用无法运行。

3.2 总体架构

一个 Spark 应用包括驱动程序、集群管理员、worker、执行者和任务这 5 个部分[4]（见图 3）。

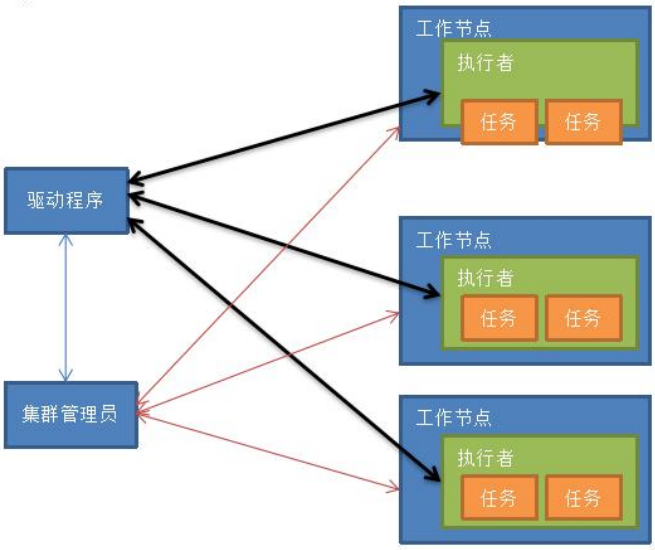

图 3

驱动程序提供在工作节点上执行的数据处理的代码且是一个把 Spark 当成库来使用的应用。集群管理员用于管理节点上的计算资源而且能将不同功能的应用的集群管理从底层调度到和在多个应用之间分享集群资源。在 Spark 中提供 CPU、内存和存储资源的是节点。执行者是在每一个工作节点上创建的 JVM 进程。任务是最小的工作单元并且运行在执行者的线程中。

3.3 RDD

弹性分布式数据集（RDD）是其设计的核心为内存计算、适合计算机集群、高效率容错的可以在 Spark 上进行并行操作的关于分区数据元素的集合且是只读的在 Spark 中的数据结构。

RDD 的特点：不可变性，一个修改 RDD 的操作都会返回一个新的 RDD 且原 RDD 是没有改变的，[3]所以它是不可变的；分片，RDD 表示一组数据的分区是分布在多个集群节点上的分区数据是各个分布式数据源中数据的一个抽象；容错性，设计者考虑了容错的问题能够让 Spark 可以自动处理故障，设置了每一个 RDD 的血统信息用以让 Spark 来恢复数据,接口，为了让很多不同类型的数据源也能进行处理和分析故 RDD 还被设计成了一个为多种数据源提供处理功能的统一接口；强类型，考虑到可以处理不同类型的数据故 RDD 被设计成了一个抽象的类且让其中的一个参数来表示类型；长期驻留内存，Spark 拥有一套支持内存计算的 API 因此作为 Spark 的组件 RDD 也能够使用。[4]

3.3.1 创建 RDD

在上面提到 RDD 是一个无法直接创建实例的抽象类故设计者为了能够创建实例提供了 SparkContext 的类中的一些方法和让 RDD 可以转化成实例得到。

1)parallelize

parallelize 可以用于从本地 Scala 集合创建。

```
val xs=(1 to 100000).toList
val rdd=sc.parallelize(xs)
```

2)textFile

用于从文本文件创建。

```
val rdd=sc.textFile()
```

函数的第一个参数是文件的路径，函数的第二个参数是指定分区的个数（可选），在默认情况下 rdd 的分区数 = max(hdfs 文件的 block 数目, sc.defaultMinPartitions)，取最大值，其中 defaultMinPartitions 默认分区。

3.3.2 RDD 操作

转换和行动是 RDD 的两种操作，前者会创建一个新的 RDD 实例，后者则会将结果返回给驱动程序。

转换：是指通过一些方法来对原有 RDD 实例进行计算来得以创建新 RDD 实例，下面会举一些常用的用于转换的方法。

Map 方法可以创建一个新 RDD 实例是因为它把一个函数作为参数并作用到原有 RDD 的元素上来实现转化创建。

Filter 方法是一个把只返回 ture 和 false 两种值的布尔函数作为参数传进去并将原有的 RDD 的每个元素来实现这个函数用看是否返回 ture 的数据创建一个新的

RDD 实例的高阶函数。当布尔函数返回 ture 时的元素会成为新 RDD 的实例的数据集故可以说新的 RDD 实例的数据集是原有的 RDD 的子集。这时，新 RDD 和原 RDD 有一个包含的关系。

flatMap 方法返回的是一个有限的序列，在这之中是用一个函数作为它的参数输入并按照一定的规律处理原 RDD 中的元素，最后通过扁平化结果可以得到新 RDD 的数据集。

行动：通过函数对元素进行计算然后将结果返回给驱动程序。

常用的有：collect 方法和 count 方法、first 方法：第一个方法返回的是一个的数组，第二个返回元素的个数，第三个将第一个元素返回。

3.4 Spark Streaming

Spark Streaming 是一个 Spark 组件且可以运行在 Spark 上的 Spark 类库，为 Spark 添加了数据流处理的功能即把数据流按照非常小的固定时间间隔分成一批一批的数据，因此是分布式数据流处理框架。每一批数据以 RDD 的形式存储。数据流的处理结果可以被多方使用。可以输出给任何应用，其他应用再进行进一步的处理或直接展示出来。

3.5 Spark SQL

Spark SQL 是 Spark 中的一个模块，可以用我们熟悉的 SQL 语句进行查询数据的功能。它支持多种查询语句，包括 SQL、HiveQL 和集成了查询功能的语言。而且还能进行交互分析。它可以和其他 Spark 库进行无缝集成。

它提供了用于处理结构化数据的高层抽象和 API。比 Spark CoreAPI，Spark SQL 更加易用。

4.结束语

Apache Spark 是加州大学伯克利分校 AMP 实验室所开源的类 Hadoop MapReduce 的通用并行框架且是专门为了处理和分析大数据而设计的快速通用的计算引擎其中它是基于内存来用于处理数据和分析数据。它能更好地适用于数据挖掘与机器学习是由于他可以将中间的输出结果直接保存在内存中而不是保存在硬盘上。Spark 可以实现 MapReduce 的数据处理功能且可以在 Hadoop 文件系统中运行。因为 Spark 技术通用、快速，所以当今越来越多的公司集团采用 Spark 来搭建平台，根据不同的用途应用于各个领域如机器学习和图计算等。

参考文献

[1]黄黎,顾筠.基于 Hadoop 平台的并行化数据分类算法研究[J].制造业自动化,2014,36(14):5-9.
[2]杨宁. 基于 Spark 的云化报表系统的设计与实现[D].南京邮电大学,2016.
[3]萨初日拉. 基于 Spark 平台的数据立方体快速计算方法研究[D].华北电力大学,2016.
[4]郭丽红. 我爱上了 SP AR K[J]. 都市家教月刊, 2012(8):191-192.
[5]申玲艳.MapReduce 计算模式的性能优化设计及其应用[J].信息与电脑(理论版),2016(14):49-50.

作者简介

第一作者：王溶（1998-），女，汉，四川省巴中市通江县杨柏乡，本科，四川大学锦城学院，研究方向：大数据
第二作者（通讯作者）：鲍正德（1989-），男，汉，黑龙江哈尔滨，研究生，四川大学锦城学院，研究方向：电子商务。
第三作者：李晨曦（1998-），男，汉，贵州省贵阳市，本科，四川大学锦城学院，研究方向：大数据技术开发