doi:10.13442/j.gnss.1008-9268.2018.04.014

基于 Spark 的 GNSS 网基线向量并行化处理

白帆[1],孙宁[2]
(1.92941 部队,辽宁 葫芦岛 125000;
2.辽宁工程技术大学 软件学院,辽宁 葫芦岛 125105)

摘 要:针对传统单机处理大规模复杂 GNSS 观测数据效率低的问题,引入了 Spark 大数据集群,利用子网划分思想并调用 GAMIT 软件对 GNSS 网基线向量进行解算,实现了并行化计算。实验结果表明,在保证解算精度与整体解算在同一量级上的同时,提高了执行效率,性能优于整体解算,较好地满足了大规模复杂 GNSS 数据处理需求。

关键词:GNSS;基线解算;Spark;并行化计算
中图分类号:P228.4;TP39 **文献标志码**:A **文章编号**:1008-9268(2018)04-0077-04

0 引 言

随着 CORS 建立得越来越多,全球卫星导航系统(GNSS)观测的数据量也越来越大,因此,数据的存储及计算面临着巨大的挑战[1]。由于计算机和网络技术的迅猛发展,大规模海量数据的分布式存储以及计算也不断发展起来,并在现实社会中得到了广泛的应用。Spark 是专为大规模数据处理而设计的快速通用计算引擎,Spark 的出现改善了 Hadoop 反复在磁盘上进行读写操作的缺陷,提高了计算大规模海量数据的计算效率。现如今,已经有很多企业引入 Spark 大数据集群平台以解决计算大规模复杂数据的瓶颈。针对目前全球卫星导航系统 GNSS 数据量大、计算效率低等问题,本文将 Spark 大数据集群引入到 GNSS 网的基线解算中,对预处理后的观测文件进行并行化计算和存储,通过实验验证了利用子网划分思想结合 Spark 大数据集群平台解算 GNSS 网基线向量的效率。

1 相关平台

1.1 Hadoop

Hadoop 是一个分布式系统基础架构,用户可以充分利用集群的计算优势进行高速运算和存储。Hadoop 最核心的设计是 HDFS 和 MapReduce。HDFS 为海量的数据提供了存储,MapReduce 为海量的数据提供了计算[2]。Hadoop 采用多进程模型,如图 1 所示。

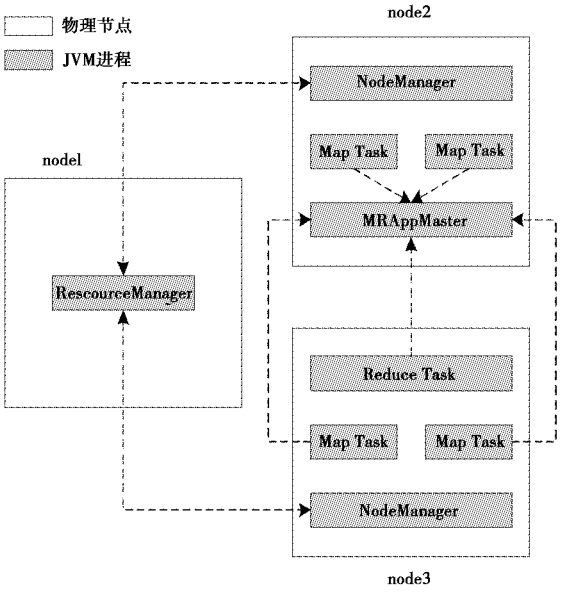

图 1 Hadoop 多进程模型图

Hadoop 的每个 Task 运行在 JVM 进程中,且 Task 不被复用。MapReduce 不构建可重用资源池,Task 动态申请资源且执行结束后立即释放资源。启动 MapReduce 中 Task 进程的速度慢,造成了不必要的启动时间消耗,不适合运行低延迟作业。

收稿日期:2018-07-04
通信作者:白帆 E-mail:40713858@qq.com

1.2 Spark 平台

Spark 是专为大规模数据处理而设计的快速通用的计算引擎，可用来构建大型的、低延迟的数据分析应用程序。Spark 可将中间输出结果保存在内存中，从而不再反复读写 HDFS。实际上，Spark 是对 Hadoop 的补充，Spark 通过名为 Mesos 的第三方集群框架可以在 Hadoop 文件系统中并行运行。Scala 用作 Spark 应用程序框架，Scala 可轻松地操作分布式数据集。Spark 采用多线程模型，如图 2 所示。

Spark 的每个 Executor 运行在 JVM 进程中，Executor 中可运行多个 ShuffleMapTask 或 ReduceTask，Task 则是在 Executor 的一个线程中运行，而且 Task 是可共享的，在 Executor 中加载一次文件或者数据后可一直被 Task 复用，直至程序执行结束后释放资源，避免了任务重复申请资源所造成的时间花费。Spark 可以建立可重用资源池来运行全部的 ShuffleMapTask 和 ReduceTask。Spark 启动任务速度快，可大大降低运行时间。

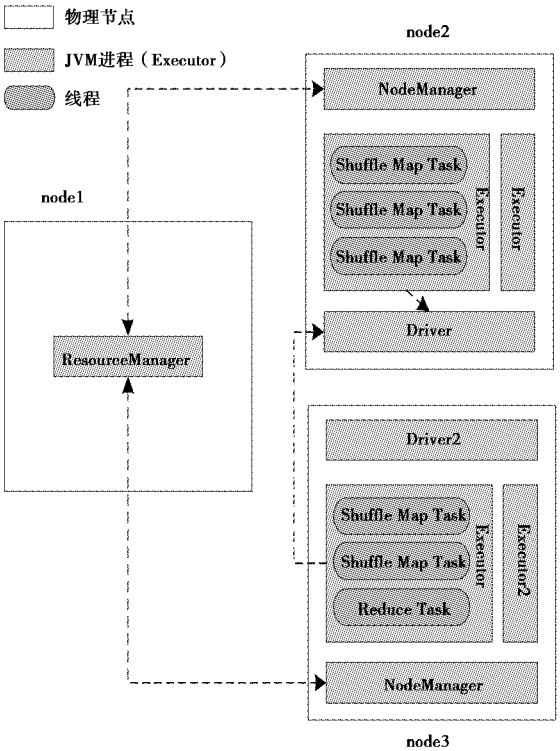

图 2 Spark 采用多线程模型

1.3 GAMIT 软件

GAMIT 软件是目前高精度 GPS 基线解算软件中最为经济、使用最广泛的一个软件。GAMIT 软件是由美国麻省理工学院（MIT）和斯克里普斯海洋研究所（SIO）联合研制的 GPS 数据处理软件[6]。当它采用精密星历和高精度起算点时，其处理长基线和连续时段静态定位相对精度可达 10e-8～10e-9 数量级，处理短基线的精度可达 1～3 mm. 本文使用的 GAMIT 版本为 GAMIT10.6.

2 基于 Spark 的基线解算并行化处理

2.1 利用 Spark 实现并行化计算

Hadoop 分布式文件系统（HDFS）是一种广泛使用的文件系统，而 Spark 支持读写很多种文件系统，可以使用任何我们想要的文件格式，Spark 能够很好地使用 HDFS。HDFS 被设计为可以在廉价的硬件上工作，有弹性地对应节点失败，同时提供高吞吐量。Spark 和 HDFS 可以部署在同一批机器上，这样 Spark 可以利用数据分布来尽量避免一些网络开销。在 Spark 中使用 HDFS 只需要将输入输出路径指定为 hdfs://master:post/path.

YARN 是在 Hadoop 中引入的集群管理器，它可以让多种数据处理构架运行在一个共享的资源池上，并且通常安装在与 Hadoop 文件系统（简称 HDFS）相同的物理节点上。在这样配置的 YARN 集群上运行 Spark 是很有意义的，它可以让 Spark 在存储数据的物理节点上运行，以快速访问 HDFS 中的数据。基于 HDFS 的观测数据分布式文件系统结构如图 3 所示[4]。

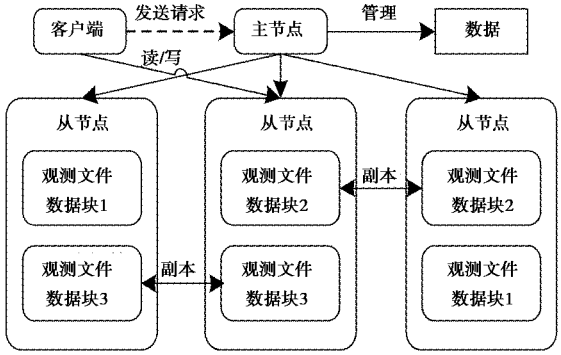

图 3 基于 HDFS 的观测数据分布式文件系统结构

在 Spark 中使用 YARN 主要伪代码如下所示：

第一步，设置 Hadoop 配置目录的环境变量 hadoop_conf_dir。

第二步，使用 Spark-submit 提交应用作业：

exporthadoop_conf_dir="…";

Spark-submit——master yarn yourapp。

spark 读取 HDFS 中的文件和写入数据到 HDFS 中，主要伪代码如下所示：

```
def main(args: Array[String]): Unit={
valconf=new SparkConf()
conf.set("spark.master", "local")
conf.set("spark.app.name", "spark demo")
valsc=new SparkContext(conf);
    //读取 HDFS 数据
valtextFileRdd=sc.textFile("hdfs://路径")
valfRdd = textFileRdd.flatMap {_.split(" ")}
valmrdd=fRdd.map {(_, 1)}
valrbkrdd=mrdd.reduceByKey(_+_)
    //写入数据到 HDFS 系统
rbkrdd.saveAsTextFile("hdfs://路径")
}
```

2.2　基于 Spark 的基线解算过程

第一步，由子网划分创建索引文件，其中子网编号设置为文件名，测站点名设置为文件内容，为了便于对文件内容进行操作，各名之间用空格相隔开。

第二步，客户端上传文件至 HDFS(文件包括：观测文件、精密星历文件、广播星历文件、索引文件)。

第三步，建立 GAMIT 软件执行所需的工程文件目录。

第四步，解析索引文件后获取相应的子网的全部测站名，从 HDFS 中复制相应文件到对应的文件夹中。

第五步，调用 sh-setup 链接外部表文件 tables，再调用 sh-gamit 进行基线解算。

需注意的是，在此过程中，一个索引文件对应一个子网进行处理，Spark 分配 Shuffle Map Task 执行索引文件。基于 Spark 进行基线解算过程如图 4 所示。

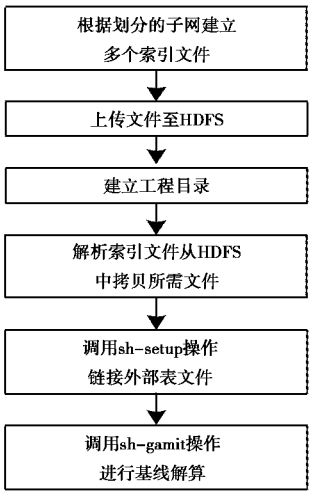

图 4　基于 Spark 进行基线解算过程

3　实验结果及分析

3.1　实验数据及环境

实验数据：采用中国及周边地区 41 个 IGS 连续运行跟踪站 2018 年年积日第 58 天的 30 s 采样间隔的观测值数据文件，如图 5 所示。

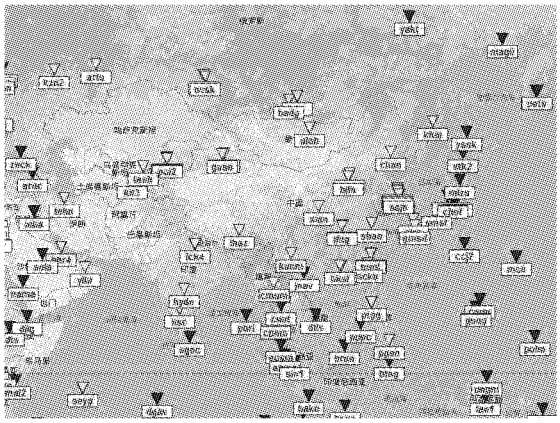

图 5　IGS 连续运行跟踪站分布图

实验环境：所有的实验都是在实验室搭建的 Spark 平台上运行的。平台由 10 个节点组成，物理机配置为 Intel Core i5 处理器，主频 2.30 GHz，双核，内存 2G. 软件配置：CentOS 6.7，JDK 1.8.0，Hadoop 2.6.0，Spark 1.6.0，Scala 2.12.1.

3.2　实验结果分析

采用划分子网的思想，其中单节点对整网进行基线解算，有几个节点就划分几个子网，实验结果如表 1 所示。

表 1　不同数量子网个数和计算节点数下的解算时间

计算节点数(个)	子网个数(个)	时间/s	加速比
单机	整网计算	3 897	1
4	4	1 392	2.80
8	8	538	7.24
10	10	407	9.57

由表 1 可知，随着子网个数和计算节点数的增

加，运行时间在逐步减少，这说明计算效率也在增加。Spark 将大数据处理任务分为 n 块并行处理，其计算能力优势随着计算任务复杂程度增加而扩大。在同样处理 41 个 IGS 连续运行跟踪站数据工程中，从运行消耗加速比来看，计算节点越多时间越短，计算效率越高。Spark 大数据平台会比 Hadoop 速度提高约 20 倍[4]，但实际中因环境及应用的不同等原因理论速度通常无法达到，因此，需要做到尽可能合理地分配计算资源，使效率最大化。

4 结束语

本文针对传统的单机方法无法应对现今对大规模复杂的 GNSS 观测数据的处理需求，利用 Spark 大数据平台和子网划分思想实现了基于 Spark 的基线解算过程的分布式处理，大大提高了计算的效率。可以在实际工程中合理有效地分配计算资源，在保证解算精度的前提下，降低执行时间，提高对海量 GNSS 数据处理能力。

参考文献

[1] 李林阳，崔阳，陈正生，等. 大规模 GNSS 网分布式存储与解算方法[J]. 测绘科学技术学报，2016(5)：464-469.

[2] 吕志平，陈正生，崔阳，等. 大型 CORS 网基线向量的分布式处理[J]. 测绘科学技术学报，2013，30(4)：433-438.

[3] KARAU H，KONWINSKI A，WENDEL P，等. Spark 快速大数据分析[M]. 王道远. 译. 北京：人民邮电出版社，2015：51-59.

[4] 杨国庆，岳东杰，陈浩，等. 基于 Hadoop 的 GNSS 网基线向量的分布式处理[J]. 全球定位系统，2017，42(4)：66-69.

[5] 高彦杰. Spark 大数据处理：技术、应用与性能优化[M]. 北京：机械工业出版社，2014：1-9.

[6] 姜卫平，赵倩，刘鸿飞，等. 子网划分在大规模 GNSS 基准站网数据处理中的应用[J]. 武汉大学学报(信息科学版)，2011，36(4)：389-391＋505.

[7] 王凯，曹建成，王乃生，等. Hadoop 支持下的地理信息大数据处理技术初探[J]. 测绘通报，2015(10)：114-117.

[8] 陈正生，吕志平，崔阳，等. 大规模 GNSS 数据的分布式处理与实现[J]. 武汉大学学报(信息科学版)，2015，40(3)：384-389.

[9] 赵小娟，陈韬. 基于 Hadoop 搭建云 GIS 平台的探索与研究[J]. 广东科技，2014，23(20)：140-141.

[10] 薛慧艳，独知行，李胜春，等. 基于 GAMIT 的 IGS 跟踪站网基线解算[J]. 全球定位系统，2012，37(1)：. 32-34.

作者简介

白帆 (1979－)，男，工程师，主要从事水面靶标和 GPS 导航定位技术的应用。

孙宁 (1979－)，女，讲师，主要研究方向为计算机软件及理论。

Parallel Processing of GNSS Network's Baseline Vectors Based on Spark

BAI Fan[1], SUN Ning[2]

(1. *Unit* 95, *No.* 92941 *Troops of PLA*, *Huludao* 125000, *China*; 2. *College of Software*, *Liaoning Technical University*, *Huludao* 125105, *China*)

Abstract: In order to solve the problem that the traditional single machine processing large-scale and complex GNSS observation data are low efficiency, the Spark large data cluster is introduced, the subnet partition idea and the GAMIT software are used to calculate the GNSS network's baseline vector, and the parallel computing is realized. The experimental results show that the efficiency is improved and the performance is better than the whole solution, and the accuracy is in the same order, which satisfies the demand of large-scale and complex GNSS data processing.

Keywords: GNSS; baseline solution; Spark; parallelization calculation