分布式地震数据文件系统

刘永江[1],邵 庆[2],彭淑罗[3]
(1. 中海油研究总院 技术研发中心,北京 100027;
2. 东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318;
3. 中国石油华北油田公司数据中心,河北 任丘 062550)

摘 要:在地震资料处理系统中,对 TB 级的海量地震数据文件的存取存在严重的 I/O 瓶颈。针对地震数据文件操作的特点,提出了一个基于集群环境的分布式地震数据文件系统(DSFS)。该系统由数据库节点、地震数据文件节点、计算节点组成。设计了面向地震道集数据的分布式文件架构,该架构以地震道头信息索引数据库为中心,将一个大的地震数据文件分解为多个可独立操作的子文件,建立基于道头字的索引数据表,根据索引表可快速定位道数据所在的子文件及数据块。设计了一组 DSFS 文件操作和道集数据读写操作,提供了针对虚拟文件按道头字值进行数据查询和输入输出,屏蔽了分布式文件的细节,并提出了文件分布及并行存取策略。DSFS 在地震资料处理系统得到应用并具有很高的数据存取效率。

关键词:地震数据;道集;并行计算;虚拟文件;分布式文件系统
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2015)11-0209-04
doi:10.3969/j.issn.1673-629X.2015.11.042

Distributed Seismic Data File System

LIU Yong-jiang[1], SHAO Qing[2], PENG Shu-luo[3]
(1. Technology R&D Center, CNOOC Research Institute, Beijing 100027, China;
2. School of Computer & Information Technology, Northeast Petroleum University, Daqing 163318, China;
3. Data Center of PetroChina Huabei Oilfield Company, Renqiu 062550, China)

Abstract: The serious I/O bottleneck was found in seismic data process system to access large seismic data file which size reaches TB grade. Aiming at the futures of accessing seismic data files, a distributed seismic data file system (DSFS) based on cluster computer is put forward, that is composed of database node, seismic data file node and computing node. A distributed file architecture facing seismic gather is designed, taking seismic trace indexing database as the center, spliting a large seismic data file into sub files that can be access seismic data separately, a seismic trace indexing table is designed to locate the sub file and data block that a trace data is stored. A set of operators on DSFS file and accessing gather data are designed, providing the data query and input and output data by trace key values in large file, the distributed file details is screened. The strategy for distributed file and parallel I/O is presented. The DSFS has been used in seismic data process system, and has very high I/O efficiency.

Key words: seismic data; gather; parallel computing; virtual file; distributed file system

0 引 言

随着地球物理勘探技术的发展,地震数据采集的精度越来越高,数据的容量越来越大,一个三维地震数据体的容量通常达到 TB 级[1-2]。海量的地震数据文件给地震资料处理系统带来了巨大的挑战,同时在数据的 I/O 效率方面也带来了以下问题[3-4]:

(1)数据文件操作消耗时间和磁盘资源,如文件合并、复制等。

(2)文件读写慢:数据文件在磁盘上是按照文件分配链表的方式存在的,在文件读写过程中,要查询链表对地震道数据定位。测试表明,当数据文件很大时,频繁的链表操作会明显降低数据读写速度。

(3)并行写冲突:地震数据处理一般采用多节点多进程并行作业方式,当多个作业对同一数据文件输出数据时,就会出现写冲突,导致作业排队,影响数据存取效率。

收稿日期:2015-01-06 **修回日期**:2015-04-10 **网络出版时间**:2015-11-04
基金项目:国家科技重大专项(2011ZX05023-005-012)
作者简介:刘永江(1965-),男,硕士,高级工程师,从事地震资料处理技术方面的研究。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20151104.0952.048.html

当前主要从 3 个方面解决上述问题:硬件层、文件系统层、应用层。

硬件层解决方式[5]:通过磁盘阵列实现多节点共享磁盘文件,因为磁盘阵列是多个可并行读写的盘片,因此可支持多个应用端并发读写文件。这种方式支持的并发节点数有限。如果在单节点上启动多个进程并行对同一文件读写,其内部采用队列机制,本质上是串行文件读写。

文件系统层解决方式:提供分布式文件系统,如采用 Hadoop 的 HDFS 技术[6],将文件存储分布到多个节点、多个文件块,通过统一的文件目录,实现分布式文件操作,但 HDFS 只允许一个写入操作,并且只能在文件末尾进行写入,无法实现并行数据写入[5]。HDFS 以固定的数据块将文件碎片化,以提高并行读的速度,但地震数据的特点是以道或道集为单位进行数据存取,其大小不固定,必然出现大量跨块操作,再将数据传输到单一节点进行整合,增加了网络开销,导致出现数据 I/O 瓶颈[7]。

应用层解决方式:文献[8-10]专门针对地震数据提出了虚拟地震数据文件系统,但其主要是针对文件合并等数据管理操作进行的,数据文件必须存储在各节点可共享的磁盘阵列上,没有解决分布式、并行存取的问题。

针对上述问题,文中提出一种适合集群环境的分布式地震数据文件系统(Distributed Seismic File System,DSFS)。DSFS 由索引数据库节点和计算节点组成,通过地震道信息索引数据库建立分布式地震数据文件,提出了一套包括合并、删除、移动、复制等功能的分布式文件管理系统;以此为基础,建立了分布式、并行数据存取机制。

1 地震数据操作的特点

在地震资料处理过程中,对地震数据文件操作具有以下特点:

(1)地震数据有标准的格式。

国际上标准的地震数据文件格式包括 SEG-D[11]、SEG-Y[12],中海油制定了 DHT 格式[4]。这些格式的基本共同点是地震数据文件由文件头、数据体和文件尾组成。

(2)需要进行系列文件管理操作。

这些操作包括格式转换、文件合并、文件拆分等[3]。在文件的拆分与合并过程中,要求每一个文件都可以被独立进行处理,因此需要保持完整的文件格式结构。由于数据容量大,对数据文件进行合并、导入、导出等操作需要花费大量的时间和磁盘空间。

(3)以道或道集为单位进行数据读写。

地震数据存取的基本单元是道数据,复杂的算法以道集为数据存取单元。一个道集是一组关键字相同道数据集合,如基于炮号的道集就是将所有炮号相同的道数据组成一个道集。

从以上特点可以看出,用单一文件来存储海量的地震数据文件难以适应实际地震资料处理的需要,而利用 Hadoop 等通用的分布式文件系统虽然可以提高数据存取效率[13-14],但在针对地震资料处理过程中爆发式的数据 I/O 请求,其仍不能满足海量地震数据存取的需求。有必要针对这些特点,研究具有大容量、高 I/O 吞吐率和高扩展能力专门的数据存取机制,以最大限度地提高海量地震数据的存取效率。

2 DSFS 体系架构

结合地震数据操作的特点,地震数据存取机制应该具有以下能力:

(1)可按道头关键字值快速进行数据检索和道集数据抽取。

(2)可方便、快速进行文件合并、拆分操作,其中子文件具有标准的格式,可以被独立进行处理。

(3)可以在多个节点、多个进程中对数据进行并行读写。

按上述目标,文中提出了基于集群环境的分布式地震数据文件系统(DSFS)。如图 1 所示,DSFS 由索引数据库节点、数据存储节点、计算节点组成,各节点间通过集群内部的万兆级高速网络进行连接。

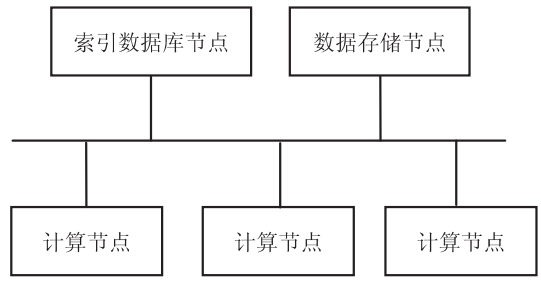

图 1 DSFS 的架构

索引数据库节点负责维护全局的工区信息、地震数据文件信息、道数据索引信息。

数据存储节点存储所有的地震数据文件,各计算节点都以统一的目录访问数据存储节点。在实际配置中,数据存储节点与计算节点并不区分开,其好处是减少数据在节点间的传输。数据存储节点上运行一个数据读进程和一个数据写进程,负责计算节点发出的数据读写请求。

各计算节点通过专门的数据 I/O 接口读写数据。I/O 接口根据道集筛选条件从索引数据库获取地震道所在的数据存储节点及相应的文件、道号,然后向数据存储节点发地震道数据读取请求。

3　DSFS 地震道索引数据库结构

地震道索引数据库用于存储地震工区相关的地震数据文件基本信息,以及每个地震数据文件的道头信息,并提供数据索引服务。其 ER(Entity Relationship,实体关系)主体结构见图 2。

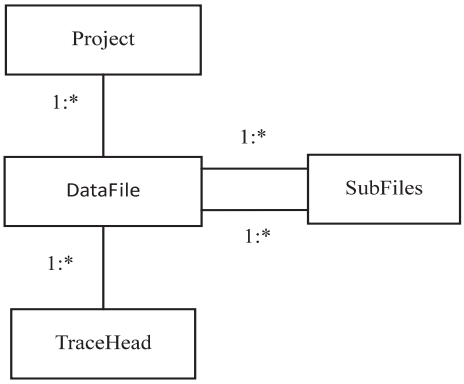

图 2　地震道索引数据库 ER 图

表 Project 定义工区相关的信息。地震勘探是以工区为作业单元,在一个工区中进行地震施工、数据采集,通过数据处理、资料解释,形成对该工区地下地质结构的认识。

表 DataFile 定义地震数据文件的相关信息,包括文件所在节点名或 IP 地址、文件路径、地震数据文件的文件头中的主要信息。一个工区对应多个地震数据文件 DataFile 记录。

表 TraceHead 定义道头字信息,道头字中的每一个关键字都对应一个 TraceHead 字段,每一道对应一条 TraceHead 记录。除道头字外,还记录道所在文件的文件号 fileno,每个道在文件中的序号 id,第 1 道的 id 值为 0。通过 fileno,就可将一个地震数据体的道存储在不同文件中。每一个地震数据文件对应多个 TraceHead 记录,三维地震数据中一个地震数据文件的 TraceHead 可达千万条。为了提高道头字检索速度,每一个文件建立一个 TraceHead 表。

表 SubFiles 定义分布式地震数据文件的组成信息。当 DataFile 表中一个文件的路径为空时,表示该文件是一个分布式地震数据文件,即该文件(称为主文件)被划分为多个独立的地震数据文件(称为子文件),每个子文件符合地震数据格式规范,可独立进行操作,各个子文件分布在不同的节点主机上。主文件在磁盘上并不存在,只在表 DataFile 中有一条记录。表 SubFiles 记录一个主文件由哪些子文件组成,并对每个子文件指定一个编号 fileno,文件编号从 0 开始。

4　DSFS 数据操作

以 DSFS 的索引数据库为基础,DSFS 除提供文件创建、文件合并、添加子文件、文件复制、文件删除等基本文件操作功能,最重要的是提供基于道集的数据存取操作。

4.1　抽取道集数据

抽取道集数据即从地震数据文件中符合条件的道数据读取出来,形成一个数据集。道集可以用一个三元组来表示:gather(select,Key_order,IDs)。

其中,select 是道集合选择条件。select 一般用 1 个或多个格式为"key=value 数值"的键值对逻辑表达式进行描述,key 为道头字。二维地震数据只有 1 组键值对,如:"炮号=4";三维地震数据有 2 组键值对,如"炮号=4","通道号=1"。

Key_order 为一个与 select 中键值对中不一样的关键字名称,用来指定道集内道的排列方式,即按该关键字值从小到大按升序排序。

IDs 为道集中所有道的标识集。

根据道集的三元组定义,可以构建从索引数据库中获取一个道集标识的查询 SQL 语句:

Select fileno,id fromTraceHead where key1 = value1 and key2 = value2 order by Key_order

获取道集中每一道的 fileno 和 id 后,根据文件号获取相应文件所在的 IP 地址,向该 IP 地址发出道数据读取请求,将文件号 fileno 和道号 id 发送给 IP 对应的数据读取进程。数据读取进程根据文件号各道号读取道数据,并将数据返回给请求者。由于子文件的文件长度较小,随机文件定位和读取数据的速度很快。

4.2　写入道集数据

地震数据处理模块在完成计算后,生成新的道集,需要写入到地震数据文件中。有 2 种道集数据写入方式:

一种是只向 TraceHead 表中写入道头信息,不进行物理数据的写入。将 SEGD、SEGY 等从野外采集的原始数据加载到系统中时,通过扫描原始数据的道头信息,对每一个道头在 TraceHead 表中创建一条相应的记录。

另一种是向 TraceHead 表中写入道头信息的同时,物理数据文件中写入道数据。从处理模块输出道数据时采用这种方式。

5　DSFS 的文件分布及存取策略

地震资料处理采用作业流方式,见图 3。计算节点上运行 1 个数据输入进程、1 个数据输出进程、若干个作业进程。

一次地震资料处理,由若干个处理作业完成。每个作业会在若干个节点上启动图 3 的作业流程。每次作业会有若干个输入文件和一个输出文件。

为了保证数据存取的高效,按如下方式对 DSFS

进行并行分布存取：

(1)按作业进程输出文件。

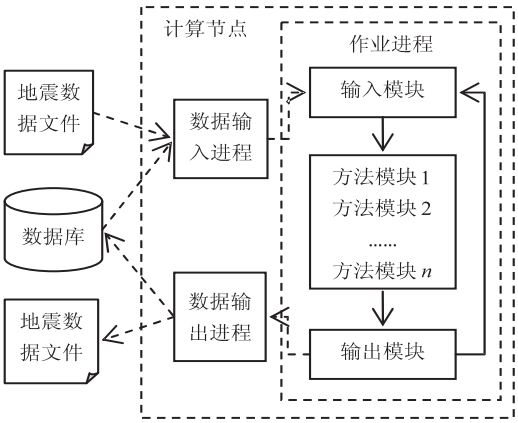

图 3 地震资料处理流程

每个作业的输出文件是一个主文件，该主文件由各个作业进程分别输出子文件组成。

每一个进程输出一个独立的子文件，直接存储在进程所在节点上，由于是本地输出，因此有很高的输出速度。各进程并行运行，因此数据输出也是并行的。这是 DSFS 并行数据输出的关键。

各子进程输出子文件时，在数据库中进行子文件统一编号和记录，因此在数据库中，形成主文件的 SubFiles 组成记录。

(2)统一调度，按数据存储节点分配计算任务。

一个道集的数据可能来自多个节点的文件。如果一个节点存储的该道集的道数最多，则将此道集的处理任务分配给该节点。这样输入该道集数据时，从其他节点获取的数据量最小，网络传输量最低，抽取道集速度最快。

6 结束语

DSFS 针对地震资料处理的特点，实现数据文件的虚拟化，使数据可以分布在集群中的不同节点。提供了一套完整的地震数据文件操作和数据读写的功能，为地震资料处理作业流程提供了分布式、并行数据输入输出。与基于 Hadoop 的地震数据文件系统相比，DSFS在数据分块、任务调度、数据输出方面都考虑了地震数据格式和存取的特点，因此具有更高的输入输出速度。DSFS 已经在地震资料处理系统平台得到应用，具有很高的数据 I/O 效率。

参考文献：

[1] 陈生昌，王汉闯，陈 林. 三维 VSP 数据高效偏移成像的超道集方法[J]. 地球物理学报，2012，55(1)：232-237.

[2] Pang Tinghua, Lu Wenkai, Ma Yongjun. Adaptive multiple subtraction using a constrained L_1-norm method with lateral continuity[J]. Applied Geophysics, 2009, 6(3): 241-247.

[3] 赵 满. 地震数据并行访问策略的研究[D]. 大庆：东北石油大学，2012.

[4] 文必龙，胡学庆，刘永江，等. 海量数据跨盘存储机制的设计与实现[J]. 郑州轻工业学院学报：自然科学版，2010，25(3)：46-48.

[5] 张得震. 基于 Hadoop 的分布式文件系统优化技术研究[D]. 兰州：兰州交通大学，2013.

[6] 许春玲，张广泉. 分布式文件系统 Hadoop HDFS 与传统文件系统 Linux FS 的比较与分析[J]. 苏州大学学报：工科版，2010，30(4)：5-9.

[7] 冯 翔. 基于 hadoop 的地震数据分布式存储策略的研究[D]. 大庆：东北石油大学，2014.

[8] 文必龙，赵 满，刘永江，等. 虚拟地震数据文件并行访问策略[J]. 计算机系统应用，2013，22(4)：211-215.

[9] Xiao Bo, Wen Bilong. Unified format definition for bulk data [C]//Proceedings of 2011 international conference on electronic and mechanical engineering and information technology. Harbin: IEEE, 2011: 2571-2575.

[10] Liu Yongjiang, Wen Bilong. Unified format definition for seismic data[C]//Proceedings of 2011 international conference on system design and data processing. [s. l.]: [s. n.], 2011: 205-208.

[11] SEG Technical Standards Committee. SEG field tape standards [S]. [s. l.]: Society of Exploration Geophysicist, 2012.

[12] SEG Technical Standards Committee. Data exchange[S]. [s. l.]: Society of Exploration Geophysicist, 2002.

[13] 吴 昊. 基于 HDFS 的分布式文件系统数据冗余技术研究[D]. 西安：西安电子科技大学，2011.

[14] 李 林. 基于 hadoop 的海量图片存储模型的分析和设计[D]. 杭州：杭州电子科技大学，2011.

分布式地震数据文件系统

作者： 刘永江， 邵庆， 彭淑罗， LIU Yong-jiang， SHAO Qing， PENG Shu-luo

作者单位： 刘永江,LIU Yong-jiang(中海油研究总院 技术研发中心,北京,100027)， 邵庆,SHAO Qing(东北石油大学 计算机与信息技术学院,黑龙江 大庆,163318)， 彭淑罗,PENG Shu-luo(中国石油华北油田公司数据中心,河北 任丘,062550)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2015,25(11)