

文章编号:1009-9603(2022)01-0121-07

DOI:10.13673/j.cnki.cn37-1359/te.2022.01.015

# 基于Hadoop分布式文件系统的地震 勘探大数据样本采集及存储优化

杨河山<sup>1</sup>,张世明<sup>2</sup>,曹小鹏<sup>1</sup>,李春雷<sup>1</sup>,姜兴兴<sup>1</sup>

(1.中国石化胜利油田分公司勘探开发研究院,山东东营257015;

2.中国石化胜利油田分公司胜利采油厂,山东东营257051)

**摘要:**随着油气勘探开发智能化应用越来越成熟、应用场景越来越丰富,大规模应用日益临近,样本的分布式存储、高效采集及并行计算已成为油气勘探开发智能化应用的迫切需求。地震勘探的智能化是油气勘探开发智能化的重要组成部分。针对地震勘探数据具有的单一文件数据量大、非结构化的特点,在分析地震勘探大数据样本采集需求的基础上,提出基于Hadoop分布式文件系统(HDFS)的大文件分割和合并的解决方案,并对地震勘探数据生成3个不同维度的冗余存储,以提升地震勘探样本的采集效率。测试结果表明,基于HDFS的三倍冗余存储方案在数据量迅速增大的情况下,可以有效地提高地震勘探大数据样本的采集效率,从而满足地震勘探智能化应用需求。

**关键词:**HDFS;地震勘探;大数据;样本采集;存储优化

中图分类号:TE319

文献标识码:A

## HDFS-based collection and storage optimization of seismic exploration big data samples

YANG Heshan<sup>1</sup>, ZHANG Shiming<sup>2</sup>, CAO Xiaopeng<sup>1</sup>, LI Chunlei<sup>1</sup>, JIANG Xingxing<sup>1</sup>

(1.Exploration and Development Research Institute, Shengli Oilfield Company, SINOPEC, Dongying City, Shandong Province, 257015, China; 2.Shengli Oil Production Plant, Shengli Oilfield Company, SINOPEC, Dongying City, Shandong Province, 257051, China)

**Abstract:** As the intelligent application of oil and gas exploration and development matures and application scenarios increase, large-scale application is drawing nearer. As a result, the distributed storage, efficient collection, and parallel computing of samples have become urgent requirements of the intelligent application of oil and gas exploration and development. The intelligent application of seismic exploration is an important part of that of oil and gas exploration and development. In view of the large amount of single file data in and the unstructured characteristic of seismic exploration data, this paper analyzes the collection requirements for seismic exploration big data samples, proposes a solution of large file segmentation and merging based on the Hadoop distributed file system (HDFS), and implements redundant storage of seismic exploration data in three dimensions to improve the efficiency of seismic exploration sample collection. The experimental results show that the HDFS-based triple redundant storage solution can effectively improve the efficiency in collecting seismic exploration big data samples under rapid growth in data amount and therefore meet the requirements for intelligent application of seismic exploration.

**Key words:** HDFS; seismic exploration; big data; sample collection; storage optimization

油气勘探开发涉及学科多,工程量巨大,地震、地质、测井、油藏等多专业数据贯穿着油气勘探开

发的整个过程。大数据技术作为新一代信息技术,已在经济、社会多个领域取得了显著的应用效果,

收稿日期:2021-11-20。

作者简介:杨河山(1981—),男,山东德州人,高级工程师,硕士,从事勘探开发大数据技术算法研究与数据处理工作。E-mail: yangheshan751.slyt@sinopec.com。

基金项目:中国石化科技攻关项目“老油田开发大数据应用系统集成与示范应用”(P20071-4)。

随着大数据技术向油气勘探开发领域的渗透,为油气勘探开发工作的发展带来了前所未有的机遇<sup>[1-4]</sup>。

地震勘探在油气勘探开发方面均具有重要的作用。在勘探阶段,地震勘探是最重要的手段,通过与测井数据结合,主要用于探明地下地质构造并进行储层参数的检测;在开发阶段,开发地震为油藏描述、剩余油分布、注水注气等开发方案提供丰富的信息<sup>[5-7]</sup>。

大数据、人工智能技术与油气勘探开发的深度融合,将在油气田勘探、开发等多个环节发挥重要作用。油气勘探开发智能化是降本增效、提高竞争力的有效途径。地震勘探数据是油气勘探开发智能化的重要数据源,具有单一文件数据量大(GB级)、非结构化的特点,同时地震勘探数据应用场景丰富<sup>[8-9]</sup>。如何利用大数据技术对其进行有效管理、快速访问和计算是目前面临的技术难题<sup>[10]</sup>。

样本数据是智能化应用的基础,地震勘探样本一般以小图片或地震属性数据为表现形式,这些样本主要通过对其源地震勘探数据进行读取、计算及绘图生成。在目前的地震相关智能化研究及应用中,样本采集及存储主要为单一应用场景的项目研究服务<sup>[11]</sup>,未涉及在分布式环境中进行大规模地震勘探样本的采集、存储及数据访问。随着大数据技术在油气勘探领域应用的逐步深入,样本的分布式存储、采集,多用户、多应用场景并行计算,已成为油气智能勘探应用的迫切需求。笔者利用大数据技术,考虑地震勘探数据的特点以及智能化应用场景的需求,开展地震勘探大数据样本采集及存储优化研究,提高大数据样本的适用性、有效性及高效访问,为油气勘探大数据智能化的应用奠定坚实基础<sup>[12]</sup>。

### 1 地震勘探大数据样本分析

#### 1.1 数据特点分析

地震勘探数据作为油气勘探领域的主要数据信息,具有多种国际通用的格式,包括标准的SEG-D和SEG-Y等。不同的地震勘探专业软件应用的数据格式也有所不同,但是不同的地震勘探数据格式的主要结构大体是一致的。按照地震勘探数据信息类型的差异,可将地震勘探数据文件分为三大部分,即文件头、数据道(包括道头和道数据)和文件尾表。综合各种地震勘探数据文件记录格式,地震勘探数据文件统一的结构如图1所示。



图1 地震勘探数据文件结构  
Fig.1 Seismic exploration data file structure

地震勘探数据通常以SEG-Y格式存储,文件以卷头开始,卷头存储元数据包括采样率、采样时间、格式码等,然后对各道数据进行顺序排列,每道数据包含道头和道数据两部分。地震勘探数据中“道”是基本单位,道头主要记录该道数据的参数,相当于道的元数据;道数据采用一维数组存储,数组元素则有多种数据类型,一般为IBM浮点类型,而IBM浮点类型需要通过字节层次的计算才能转换为目前计算机服务器中常用的IEEE浮点类型,降低了数据的读取性能。

地震勘探数据存储的是三维地理空间上的地震相关特征值,这个特征值一般是地震振幅,也可以是其他的地震属性值,因此可以将一个地震勘探数据看作一个三维数组。为了后续的表达方便,设这3个维度分别为X,Y,Z;其中Z为深度维,X和Y分别代表平面上的2个维度(图2)。在对地震勘探数据的访问中,一般以单道或切片进行读取。单道读取表示确定平面某点(即X和Y为定值),读取Z方向的值,即读取1个一维数组;切片读取表示确定三维中任一维度,读取其他2个维度的数据,即读取1个二维数组,表示读取Y方向(Y为定值)的1个切片。

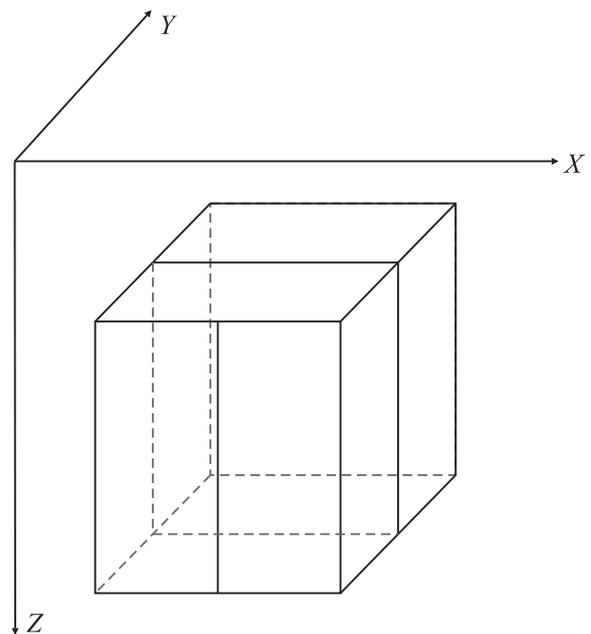


图2 地震勘探数据的三维分布及常用读取方式示意  
Fig.2 Three dimensional distribution and common access methods of seismic exploration data

### 1.2 采集需求分析

通过对2017年—2019年SEG(勘探地球物理学家协会)年会中人工智能方面的论文进行统计发现,目前地震勘探智能化应用主要集中于地震勘探数据处理与解释方面,主要应用场景有地震构造解释(含断层识别、层位追踪、岩丘等地质体追踪)、构造建模、地震相识别、储层参数预测、噪声压制与信号增强、地震波场重建、地震反演、地震速度拾取、初至拾取、地震勘探数据重建与插值、地震属性分析、微地震勘探数据分析、综合解释等方面<sup>[1]</sup>。经分析,这些论文中除少部分是通过三维地质模型正演形成训练样本集外,绝大部分是通过地震勘探数据进行属性或图片提取来生成训练样本集和预测样本集,其样本的主要表征方式包括数字类型和图片类型。

数字类型样本主要是地震勘探数据及其衍生的各种振幅、波形、几何类属性等类型的数据。如表1所示,其中 $X, Y, Z$ 为样本的位置信息,  $value\_1$ 至 $value\_n$ 为样本的特征值。属性主要通过对各道数据的计算形成,因此该类样本的生成需要频繁地访问道数据。

图片类型样本则是在一定范围、一定绘图参数下,分别垂直于 $X, Y, Z$ 方向对三维地震勘探数据进行切片,生成的小图片集合。图3为图片类型样本示例,其中垂直 $X$ 和 $Y$ 方向剖面切片的图片,两者采用不同的绘图模式,图3a为彩色变密度模式,图3b为波形填充模式,图3c为垂直 $Z$ 方向生成的平面切片图片,对每个方向的切片均按等间距切割为 $4 \times 4$ 的小图片集合。

### 1.3 存储优化的必要性

综上所述,地震勘探智能化应用中样本数据的

表1 数字类型样本示例

Table1 Example of digital type sample

样本编号	$X$	$Y$	$Z$	$value\_1$	$value\_2$	...	$value\_n$
1	1 568	3 013	1 000	-0.408	-0.688	...	0.740
2	1 571	2 835	1 002	2.613	-1.656	...	-2.043
3	1 573	2 835	1 004	6.267	-1.758	...	-0.909
4	1 573	3 021	1 006	0.518	-0.634	...	0.445
5	1 576	3 012	1 008	-0.801	-0.560	...	0.461
6	1 580	3 013	1 010	-1.332	-0.335	...	-0.454
7	1 581	2 835	1 012	3.567	-1.307	...	0.721
8	1 583	3 021	1 014	-0.645	-0.525	...	-0.438
9	1 587	3 021	1 016	-1.257	0.272	...	0.400
10	1 589	3 021	1 018	-1.785	0.087	...	-0.238
11	1 600	2 835	1 020	1.753	-0.816	...	-2.359
12	1 602	3 021	1 022	-4.448	0.784	...	-0.003
13	1 605	3 013	1 024	-3.841	0.555	...	0.110

生成,需要对地震勘探数据进行读取,其读取方式主要表现为:数字类型样本(即属性样本)主要通过频繁读取地震单道数据,并利用地震属性计算公式进行处理得到;图片类型样本则通过分别读取 $X, Y, Z$ 方向的切片数据,并进行绘图得到。

在目前相关的地震勘探智能化应用研究中,主要是针对应用场景的智能化算法及应用进行研究,在样本的采集、存储等方面一般采用单一文件方法,这种方法只适用于单一用户的独占式数据访问。而在实际应用中,同一个地震勘探数据,可能面对多个用户同时的、针对不同类型样本的采集需求;对采集后的样本集,用户还可能并发读取用于不同的应用场景研究,即在地震勘探智能化应用时,样本的采集与存储面临的难题是如何解决大规模的并发数据访问及并行计算的效率问题。因此,

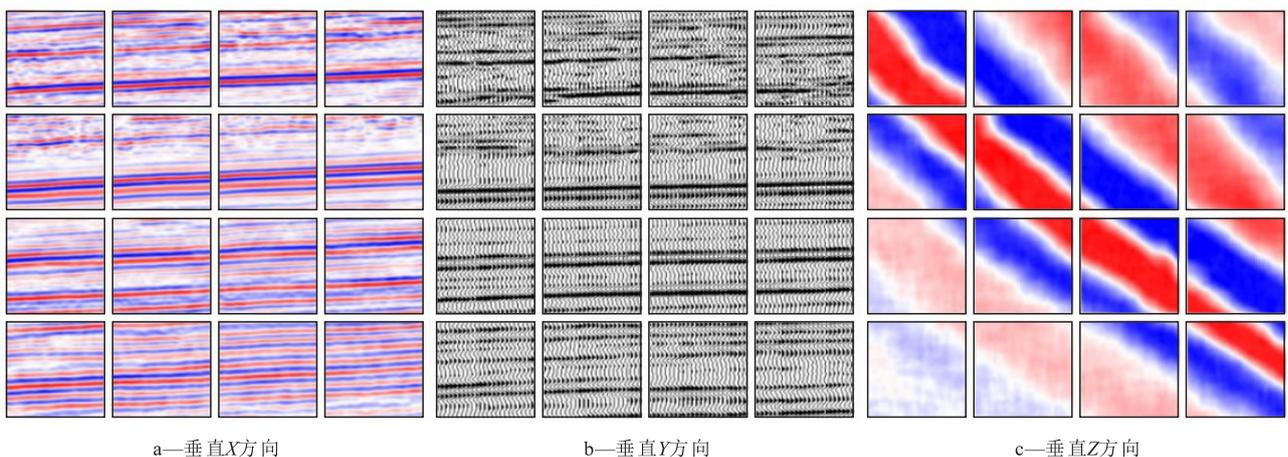


图3 图片类型样本示例

Fig.3 Example of image type sample

在大数据环境下,开展地震勘探数据的存储优化技术研究,以提高样本采集及访问的效率,是油气勘探大数据智能化应用的重要基础。

## 2 地震勘探大数据样本分布式采集、存储方法

大规模的并发数据访问及并行计算主要借助于可扩展的分布式数据库或分布式文件系统。分布式数据库是利用网络将物理上分散的多个数据库单元连接起来,通过统一的数据库管理系统组成一个逻辑上统一的数据库,其中的每一个数据库单元一般采用传统的关系型数据库。这种分布式关系型数据库擅长处理结构化数据,对非结构化数据处理起来效率较低。地震勘探数据为非结构化数据,同时还具有单个数据文件大的特点。因此地震勘探数据并不适合采用分布式数据库系统进行存储和分析,采用分布式文件系统存储和处理地震勘探数据成了必然选择。

### 2.1 Hadoop 分布式文件系统

Hadoop 是由 Apache 软件基金会开发的分布式系统框架<sup>[13]</sup>,中外很多的大数据平台均是在 Hadoop 原生版本基础上进行优化、封装产生的发行版本,包括 Cloudera, Hortonworks, 华为 FusionInsight, 大快 Dkhadoop 等。由于在可靠性、可伸缩性和成本上的优势,Hadoop 已成为目前主流的分布式存储与计算框架<sup>[14-15]</sup>。

HDFS 和 MapReduce 并行计算组件是 Hadoop 生态系统的两大核心组件,前者通过将数据、尤其是非结构化数据按均匀的冗余方式存储于集群的各节点计算机,实现完备的分布式文件存储服务,保证数据存储的可靠性;后者是一种并行计算的编程模型,适用于大规模数据的分布式并行计算,其首先通过 Map 端将分布于各节点的数据进行计算,再通过 Reduce 端将各节点数据进行汇总,从而实现集群式并行计算。

HDFS 分布式文件存储使用主从式结构,集群中有名称节点、次名称节点、数据节点和客户端4种类型的节点。一般 HDFS 集群有多个数据节点,可以通过名称节点增加或减少数据节点,增强集群的伸缩性,可以通过增加节点和副本数扩展 HDFS 并行访问的效率<sup>[16-17]</sup>。

### 2.2 基于 HDFS 的地震勘探大数据样本存储优化方案设计

地震勘探数据具有数据文件大、信息量多、访

问频繁等特性;HDFS 框架主要是针对大数据存储及分布式计算设计的,大文件顺序读写快、一次写入多次读取、效率高是其主要优势。因此 HDFS 很适合处理地震勘探数据这种大容量、非结构化数据,符合地震勘探大数据样本采集的应用需求。

根据前述的地震勘探大数据样本采集需求,三维地震勘探数据在分布式平台上进行大数据样本采集主要具有如下特点:①数据类型的属性样本计算,需要频繁访问道数据。②剖面或平面图片样本生成,需要频繁访问 X, Y, Z 三个方向的切片数据,每个切片包括多个地震道。根据上述数据特点,可以将一道数据作为数据存储的基本单位。

众所周知,对于在顺序文件中提取部分数据的数据访问,连续数据的读取可以极大地提高访问效率;反之,非连续数据的读取,包括在同一块中非连续数据读取和跨块数据读取再组合成最终数据,则会大大降低数据访问效率。

考虑一个几 GB 到几百 GB 的地震勘探数据,根据特点①,地震单道数据的数据量一般在几 KB 到几百 KB, HDFS 中基本数据块的大小一般设置为几十 MB 到几百 MB,因此对于单道数据的访问来说,不存在非连续数据读取的问题,在用 HDFS 存储地震勘探数据时,不需要对这种访问方式进行额外的优化设计,在需要提高访问效率时,只需要利用 HDFS 自身功能增加副本即可。而对于特点②,对一个几 GB 到几百 GB 的地震勘探数据,其切片的数据量一般在几 MB 到几十 MB 甚至上百 MB,需要频繁读取 X, Y, Z 三个方向的切片,单一文件无论如何分块都只能满足一个方向切片的连续读取。因此提高切片访问效率不能只依赖于 HDFS 的数据分块及增加副本功能。

考虑上述数据访问特点,为提高不同方向切片数据访问的效率,如图4所示,设计基于 HDFS 的地震勘探大数据样本存储优化方案。该方案在文件级而不是数据块级将数据增加至三倍冗余,这三倍冗余分别为按照 X, Y, Z 方向进行数据排列,以提高

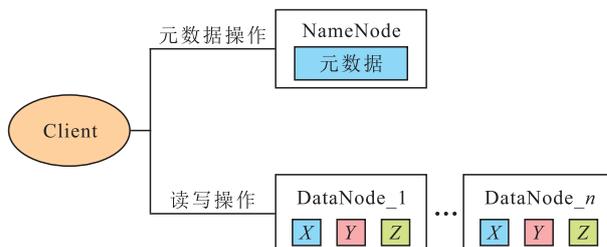


图4 基于HDFS的地震勘探大数据样本存储优化方案  
Fig.4 Optimization design scheme of seismic exploration big data sample storage based on HDFS

不同方向的切片数据访问效率。

在此基础上还可以利用HDFS分块规则及副本策略对上述的各冗余进行分块存储。具体存储规则如下:①将地震勘探数据转换为IEEE浮点类型存储,以避免后续读取数据时IBM浮点类型转换IEEE浮点类型导致的性能开销。②以道数据为基本存储单位,以满足按道进行属性数据计算时的快速访问需求。③将地震道数据分别按 $X, Y, Z$ 三个方向递增的方式进行排列,形成3个冗余文件数据,在此基础上根据HDFS默认的分块策略分别对文件进行分块写入,不能整除的部分用零值补全。④将分块策略存储于元数据中,在读取时利用元数据进行快速访问。

对上述第3项规则,举例说明如下:设样例文件 $X, Y, Z$ 方向的数据量分别为 $n_x, n_y, n_z$ ;HDFS分块数据量为 $N, X$ 方向的数据存储按下面的规则进行。

$X$ 方向每一个切片的数据量为 $m_x$ ,其中:

$$m_x = n_y \times n_z \times 4 \quad (1)$$

HDFS每一分块可以存储的 $X$ 方向的切片数为 $B_x$ ,则:

$$B_x = \text{floor} \left( \frac{N \times 1024 \times 1024}{m_x} \right) \quad (2)$$

$X$ 方向的数据分块数为:

$$\text{block}_x = \text{ceiling} \left( \frac{n_x}{B_x} \right) \quad (3)$$

$Y$ 和 $Z$ 方向的分块规则与 $X$ 方向类似,不再赘述。

总体而言,针对地震勘探大数据样本采集时对不同方向切片的频繁访问问题,本文设计了三倍文件级的冗余存储方案,并设计了具体的存储规则,以满足地震勘探大数据样本采集时的大规模并发访问需求。

在文件访问上,分布式文件系统支持无缝的数据访问,即终端用户不需要知道文件的数据块具体存储在哪个数据节点上,用户先访问管理节点,再由管理节点定位到具体的数据节点。由于上述方案是在文件级别上增加的冗余,不能够直接用HDFS进行无缝的数据访问,因此在HDFS的基础上编写了一个高层的数据调度模块,将地震勘探数据写入HDFS系统时,分别按 $X, Y, Z$ 方向对数据进行优化排列,形成3个数据文件,将文件信息写入元数据;在读取地震勘探数据时,则根据元数据按方向定位到文件,再由HDFS管理节点定位到数据块。

## 2.3 基于HDFS的地震勘探大数据样本采集模块研发

根据地震勘探大数据样本采集需求及存储优化方案,研发了地震勘探大数据样本采集模块,主要包含数字类型样本采集和图片类型样本采集2个功能,实现了2种类型地震勘探大数据样本的采集(图5),并集成在胜利油田勘探开发研究院大数据应用平台上,供专业人员使用。

a—图片类型样本采集界面

b—数字类型样本采集界面

图5 不同类型样本采集界面

Fig.5 Different types of sample collection interface

## 3 应用效果分析

测试采用的集群为4节点Hadoop集群,1个名称节点处理器(NameNode),3个数据节点处理器(DataNode)。其中NameNode节点处理器Intel Xeon E3-1240 v6 3.7 GHz,内存为16 GB,硬盘为4 TB;DataNode节点处理器I7-8700 3.2 GHz,内存为8

GB, 硬盘为2 TB, 集群各节点处于同一百兆局域网。HDFS版本为2.7.0, 采用默认的HDFS副本策略。测试采用的客户端为5台, 采用相同的配置: i5-10500 3.1 GHz, 内存为16 GB, 硬盘为2 TB。

测试数据选择文件大小分别为2.38, 7.90, 20.32 GB的3个地震勘探数据文件, 前2个文件基本代表了常见地震勘探数据的文件大小, 随着地震高密度采集技术的逐步使用, 超过20 GB的文件会越来越常见。其 $X, Y, Z$ 三个维度的切片数量如表2所示。

表2 3个测试地震勘探数据文件信息  
Table2 Information of 3 test seismic exploration data files

文件编号	文件大小/GB	不同方向切片数量/个		
		X方向	Y方向	Z方向
文件1	2.38	420	300	5 001
文件2	7.90	700	500	6 001
文件3	20.32	1 000	900	6 001

设计2种存储方案分别进行读取测试。存储方案①:按 $X$ 方向优化存储地震勘探数据体;存储方案②:分别按 $X, Y, Z$ 方向优化存储地震勘探数据体。这2种存储方案均采用128 MB的数据分块大小和三倍副本。

测试方式为对3个测试文件、2种存储方案, 利用5台客户端电脑分别对 $X, Y, Z$ 三个方向进行切片提取, 每台客户端电脑同时运行6个切片读取模块, 用来模拟30个客户进行并发访问。每个模块分别在 $X, Y, Z$ 方向随机读取200个不同数值的切片, 对每个方向的切片统计其平均读取时间。测试结果如表3所示, 在对存储方案①进行切片读取时, 未优

表3 不同存储方案数据访问性能测试结果

Table3 Data access performance test results of different storage schemes

文件编号	切片读取方向	随机切片平均读取时间/s		遍历所有切片时间/s	
		存储方案①	存储方案②	存储方案①	存储方案②
文件1	X	0.67	0.68	279.83	284.00
	Y	3.10	0.96	930.00	287.38
	Z	0.51	0.12	2 537.40	578.69
文件2	X	1.42	1.42	995.91	995.65
	Y	5.83	1.95	2 915.00	973.91
	Z	1.24	0.36	7 441.24	2 164.10
文件3	X	2.38	2.43	2 383.52	2 433.99
	Y	8.40	2.79	7 564.06	2 506.85
	Z	2.79	0.86	16 722.32	5 152.83

化的 $Y$ 和 $Z$ 方向的切片读取时间与优化后的存储方案②具有3倍以上的差异, 表明针对不同读取方向对地震勘探数据进行优化存储, 能够显著提高地震勘探大数据样本采集的数据访问效率。

## 4 结束语

地震勘探数据作为油气勘探开发中的重要数据, 在智能化应用中具有更丰富的应用场景。在油气勘探开发智能化逐步进入大规模应用时, 如何在分布式环境中提高大规模并发的地震勘探大数据样本采集效率, 对油气勘探开发智能化的推广应用具有重要意义。

本文分析了油气勘探开发智能应用中对地震勘探大数据样本的采集需求, 总结了目前地震勘探智能化应用场景中的地震勘探大数据样本类型, 针对地震勘探数据大容量、非结构化的特点, 分析归纳出不同类型地震勘探大数据样本采集时的数据读取特征, 在此基础上基于HDFS设计了地震勘探数据冗余存储及访问方案, 并开发了相应的数据存取模块和样本采集模块。测试结果表明冗余存储方案比常规方案提高了3倍以上的读取效率, 同时极大地提高了样本采集效率, 为油气勘探开发智能化应用奠定了基础。

### 符号解释

- $block_x$ ——地震勘探数据 $X$ 方向的数据分块数, 个;  
 $B_x$ ——HDFS每一分块可以存储的 $X$ 方向的切片数, 个;  
 $ceiling$ ——向上取整, 取整导致的最后一个分块字节不足部分统一用零值补全;  
 $floor$ ——向下取整, 取整导致的分块字节不足部分统一用零值补全;  
 $m_x$ ——地震勘探数据 $X$ 方向每一个切片的数据量, 字节;  
 $n$ ——特征序号;  
 $n_x$ ——地震勘探数据 $X$ 方向数据量, 字节;  
 $n_y$ ——地震勘探数据 $Y$ 方向数据量, 字节;  
 $n_z$ ——地震勘探数据 $Z$ 方向数据量, 字节;  
 $N$ ——HDFS分块数据量, MB;  
 $value_n$ ——数字类型样本第 $n$ 个特征的取值;  
 $X, Y, Z$ ——三维坐标轴的 $X, Y, Z$ 方向。

### 参考文献

- [1] 李阳, 廉培庆, 薛兆杰, 等. 大数据及人工智能在油气田开发中的应用现状及展望[J]. 中国石油大学学报: 自然科学版, 2020, 44(4): 1-11.

- LI Yang, LIAN Peiqing, XUE Zhaojie, et al. Application status and prospect of big data and artificial intelligence in oil and gas field development [J]. Journal of China University of Petroleum: Edition of Natural Science, 2020, 44(4): 1-11.
- [2] 蔡江辉, 杨雨晴. 大数据分析 & 处理综述 [J]. 太原科技大学学报, 2020, 41(6): 417-424.  
CAI Jianghui, YANG Yuqing. Overview of big data analysis and processing [J]. Journal of Taiyuan University of Science and Technology, 2020, 41(6): 417-424.
- [3] 刘伟, 闫娜. 人工智能在石油工程领域应用及影响 [J]. 石油科技论坛, 2018, 37(4): 32-40.  
LIU Wei, YAN Na. Application and influence of artificial intelligence in petroleum engineering area [J]. Oil Forum, 2018, 37(4): 32-40.
- [4] 宋明水. 济阳坳陷页岩油勘探实践与现状 [J]. 油气地质与采收率, 2019, 26(1): 1-12.  
SONG Mingshui. Practice and current status of shale oil exploration in Jiyang Depression [J]. Petroleum Geology and Recovery Efficiency, 2019, 26(1): 1-12.
- [5] 赵改善. 石油物探智能化发展之路: 从自动化到智能化 [J]. 石油物探, 2019, 58(6): 791-810.  
ZHAO Gaishan. Road to intelligent petroleum geophysical exploration: From automatic to intelligent [J]. Geophysical Prospecting for Petroleum, 2019, 58(6): 791-810.
- [6] 杨午阳, 魏新建, 何欣. 应用地球物理+AI的智能化物探技术发展策略 [J]. 石油科技论坛, 2019, 38(5): 40-47.  
YANG Wuyang, WEI Xinjian, HE Xin. Development plan for intelligent geophysical prospecting technology of applied geophysical+AI [J]. Oil Forum, 2019, 38(5): 40-47.
- [7] 林伯韬, 郭建成. 人工智能在石油工业中的应用现状探讨 [J]. 石油科学通报, 2019, 4(4): 403-413.  
LIN Botao, GUO Jiancheng. Discussion on current application of artificial intelligence in petroleum industry [J]. Petroleum Science Bulletin, 2019, 4(4): 403-413.
- [8] 宋林伟, 王小善, 许海涛, 等. 梦想云推动地震资料处理解释一体化应用 [J]. 中国石油勘探, 2020, 25(5): 43-49.  
SONG Linwei, WANG Xiaoshan, XU Haitao, et al. Application and practice of integrated seismic data processing and interpretation driven by E & P Dream Cloud [J]. China Petroleum Exploration, 2020, 25(5): 43-49.
- [9] 杨平, 詹仕凡, 李明, 等. 基于梦想云的人工智能地震解释模式研究与实践 [J]. 中国石油勘探, 2020, 25(5): 89-96.  
YANG Ping, ZHAN Shifan, LI Ming, et al. Research and practice on artificial intelligence seismic interpretation mode based on E & P Dream Cloud [J]. China Petroleum Exploration, 2020, 25(5): 89-96.
- [10] 刘坚, 马文娟, 李盛乐, 等. 基于大数据技术的地震科学数据集成共享研究 [J]. 中国科技资源导刊, 2017, 49(5): 74-81.  
LIU Jian, MA Wenjuan, LI Shengle, et al. Integration and sharing research on the earthquake science data based on the technology of big data [J]. China Science & Technology Resources Review, 2017, 49(5): 74-81.
- [11] 匡立春, 刘合, 任义丽, 等. 人工智能在石油勘探开发领域的应用现状与发展趋势 [J]. 石油勘探与开发, 2021, 48(1): 1-11.  
KUANG Lichun, LIU He, REN Yili, et al. Application and development trend of artificial intelligence in petroleum exploration and development [J]. Petroleum Exploration and Development, 2021, 48(1): 1-11.
- [12] 李剑峰. 智慧石化建设: 从信息化到智能化 [J]. 石油科技论坛, 2020, 39(1): 34-42.  
LI Jianfeng. Construction of intelligent petrochemical industry: from information to intelligence [J]. Oil Forum, 2020, 39(1): 34-42.
- [13] 田秀霞, 周耀君, 毕忠勤, 等. 基于Hadoop架构的分布式计算和存储技术及其应用 [J]. 上海电力学院学报, 2011, 27(1): 70-74.  
TIAN Xiuxia, ZHOU Yaojun, BI Zhongqin, et al. The technology and application of distributed computing and storage based on Hadoop architecture [J]. Journal of Shanghai University of Electric Power, 2012, 27(1): 70-74.
- [14] 范素娟, 田军锋. 基于Hadoop的云计算平台研究与实现 [J]. 计算机技术与发展, 2016, 26(7): 127-132.  
FAN Sujuan, TIAN Junfeng. Research and implementation of cloud computing platform based on Hadoop [J]. Computer Technology and Development, 2016, 26(7): 127-132.
- [15] 李三淼, 李龙澍. Hadoop中处理小文件的四种方法的性能分析 [J]. 计算机工程与应用, 2016, 52(9): 44-49.  
LI Sanmiao, LI Longshu. Performance analysis of four methods for handling small files in Hadoop [J]. Computer Engineering and Applications, 2016, 52(9): 44-49.
- [16] 王娟, 石艳丽, 姚立峰, 等. 地震勘探数据存储技术研究与应用 [J]. 石油地球物理勘探, 2008, 43(增刊1): 163-167, 193.  
WANG Juan, SHI Yanli, YAO Lifeng, et al. Study and application of seismic prospecting data storage technique [J]. Oil Geophysical Prospecting, 2008, 43(Supplement 1): 163-167, 193.
- [17] 文必龙, 冯翔, 左春雪, 等. 地震资料分布式存取的效率优化设计 [J]. 计算机与数字工程, 2014, 42(8): 1 386-1 389, 1 439.  
WEN Bilong, FENG Xiang, ZUO Chunxue, et al. Efficiency optimization design of seismic data distributed access [J]. Computer & Digital Engineering, 2014, 42(8): 1 386-1 389, 1 439.