

文章编号: 2096-1472(2016)-01-48-02

Hadoop平台下的地震波形数据存储与应用规划

王丹宁, 柴旭超, 王文青

(中国地震局第二监测中心, 陕西 西安 710054)

摘 要: 地震波形数据的存储与应用是国家地震数据灾备中心的重要业务之一。本文主要针对海量地震波形数据基于传统关系型数据库和文件系统的存储方式所存在的数据存储离散、查询效率低下等问题, 从大数据平台的角度, 提出基于Hadoop的地震波形数据存储解决方案, 着重阐述了业务需求、功能设计和实现原理。希望为国家地震数据灾备中心存储平台的建设提供有益的借鉴。

关键词: 大数据; 地震波形数据; Hadoop; Hbase; 数据解析

中图分类号: TP392 **文献标识码:** A

Seismic Waveform Data Storage and Application Programming under Hadoop Platform

WANG Danning, CHAI Xuchao, WANG Wenqing

(Second Monitoring and Application Center, CEA, Xi'an 710054, China)

Abstract: The storage and application of seismic waveform data is one of the important business in National Earthquake Data Disaster Recovery Center. The paper mainly aimed at some questions, which include data storage discrete, low querying efficiency, etc., exist in massive seismic waveform data storage base on traditional relational database and file system, proposed the seismic waveform data storage solution based on Hadoop from the perspective of big data platform, and emphatically describes the business requirements, functional design and principle. We hope to provide useful reference for the construction of the National Earthquake Data Disaster Recovery Center's data storage platform.

Keywords: big data; earthquake waveform data; hadoop; Hbase; data analysis

1 引言(Introduction)

地震科学数据作为中国地震局最重要的核心资源之一, 对监测预报、震灾预防、应急救援三大业务提供强大的底层数据支持。多年来, 地震科学数据经过不断的采集、观测、传输、存储, 形成了极大的数据规模, 相应地, 地震局各级机构分别构建了自己的数据存储和应用系统。2008年, 中国地震局第二监测中心拟建设国家地震数据灾备中心, 2013年, 项目开始实质运行, 至今, 已经进入软硬件测试和存储方案设计阶段, 将来汇入灾备中心的地震数据如何进行存储和应用, 成为亟待解决的问题。

2 地震数据存储和应用现状(The status of earthquake data storage and application)

地震数据可以从狭义和广义两方面进行认识。狭义上, 地震科学数据按照其获取途径可以划分为观测数据、探测数据、调查数据、实验数据和专题数据^[1]; 按照观测手段可以划分为测震、前兆等数据。比如, 专业上习惯称测震数据为地震波形数据, 为了规范地震波形数据的存档和交换, 方便地震研究人员使用, 由FDSN、IRIS和USGS共同发布的地震数据交换标准(The Standard for the Exchange of Earthquake Data, 简称SEED)便成为国内地震行业地震波形数据的标准数据格式^[2]。广义上, 特别从时下流行的大数据的

角度来看, 包括观测数据、文档、照片、视频、地图等结构化和非结构化数据都可以涵盖于地震数据的范畴之内。

现有地震数据的存储, 我们可以笼统归纳为两大方式。其一, 类似前兆业务数据等可以结构化的数据, 都存储于以Oracle、Mysql为代表的关系型数据库中, 所以业务软件和应用都基于关系型数据库进行开发, 这也是大多数行业和数据部门多年来采用的主流存储方式。其二, 对于不能结构化的数据, 特别是文档、照片、视频等典型的非结构化数据和虽能结构化但受限于现实应用而不能存入关系表中的SEED数据, 大都以数字化文件形式存储于地震系统内大大小小的传统存储平台和介质上, 比如磁盘阵列和光盘。

数据存储决定数据应用。又以测震业务为例, 由于测震业务现有流程和关系型数据库的限制, 测震数据中SEED或MiniSeed数据以文件形式存放于磁盘中, 其他辅助型数据存放于Oracle或Mysql数据库中。其实际应用首先慢于纯数据库应用; 其次受限于数据的庞大体积, 导致时间尺度不够宽广, 历史数据应用不够充分; 最后, 由于测震数据总数据量达到200TB左右, 国内并没由任何唯一数据库平台可容纳全部数据, 对于全量分析等高级应用的可能也只好停留在想象阶段。

3 大数据的发展对测震数据应用的启发(The inspiration of the development of big data to the application of seismic waveform data)

研究机构Gartner赋予大数据如下定义：“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。仅地震行业测震数据现存储量便达到200TB，可以称之为海量；测震数据的采样频率多为100Hz，也包含有200或300Hz数据，而强震业务的采样频率更达到了500Hz，其未来的数据增长率会越来越高；测震数据作为传感器采样数据，与同样作为传感器采样数据的前兆数据进行长时间尺度的联合分析，更结合多样的地震监测数据进行比对，其应用前景可能非常远大。而这样的“地震大数据”急需新的处理模式来挖掘深藏于其内部的关于地震形成机理乃至地球内部结构的秘密，从而发挥其对地震预报的高决策力和洞察力。

又根据维基百科的定义，大数据是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。就测震数据应用来说，短时间尺度、有限区域内的数据处理和分析所花费时间是可以承受的，但长时间尺度、广阔区域乃至全量分析在传统平台处理所需要的时间就令人绝望了，这也是还无研究人员进行如此操作的原因。

总之，测震数据的半结构化特性，其对地震预报的可能的巨大价值，其和前兆等地震数据的比对分析前景，都促使我们对其底层的存储模式进行大数据平台方向的探索。

4 Hadoop平台简介(Brief introduction of hadoop platform)

4.1 Hadoop生态圈

Hadoop是Apache Lucene创始人Doug Cutting创建的项目，来源于谷歌公司两篇关于GFS和MapReduce的重要论文。其形成的目的在于利用通用服务器集群进行分布式计算和海量数据处理^[3]。

Hadoop生态圈不仅包括分布式文件系统HDFS和分布式数据处理模型MapReduce，还包括一系列围绕其宗旨构建的相关项目，Hadoop主要项目及简单架构如图1所示。

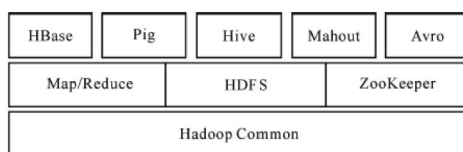


图1 Hadoop主要项目

Fig.1 Hadoop main project

图中，Hadoop Common是一组分布式文件系统和通用

I/O的组件和接口。在其之上，MapReduce为分布式数据处理模型，HDFS为商用机集群上的分布式文件系统。最上层的Hbase是一个分布式、按列存储的数据库，使用HDFS作为底层存储^[4-9]，类似于Oracle和Windows文件系统之间的关系。

4.2 灾备中心技术平台需求

Hadoop生态圈项目涵盖数据存储、查询、分析、管理等全部环节，其对地震波形数据的安全存储、查询定位、可视化展现等是一种有益的尝试。具体于国家地震数据灾备中心的应用来说，如何利用Hadoop项目存储好地震波形数据，从而为未来的数据分析和可视化打好基础，是方案组首先要考虑的问题。

5 Hadoop在测震数据存储方面的应用(Application of hadoop in the measurement of seismic data storage)

5.1 业务需求

国家地震数据灾备中心地震波形数据的存储需求主要分为两类：满足中国地震台网中心即源数据端的数据备份恢复需求和如何满足灾备端未来的数据应用。起初，项目组考虑地震波形数据集的数据量比较庞大，为满足全部数据的存储空间需求，设计存储可用容量为500TB，HDFS平台复制因子为3，即实际存储容量应达到1.5PB。

数据来源方面，地震波形数据大致以两种方式汇入灾备机房，即传感器数据流直接汇入和历史数据文件导入。可以说数据流汇入对应实时数据备份，历史数据导入对应历史数据备份。对应Hadoop技术平台需求，为最大限度满足灾备安全性能，应该选择HDFS分布式文件系统存储历史和实时数据；但是，数据以文件形式存储，会影响未来数据的分析和可视化应用，而分布式数据库可以更好的适应此需求，所以，Hbase分布式数据库便成为需求平台之一。当然，如何平衡数据灾备安全性和数据应用之间的关系，也是需要慎重考虑的。

5.2 功能设计和实现原理

根据国家地震数据灾备中心对于地震波形数据灾备的需求，其功能设计列举如下：

- (1)实时数据流接入、解析、导入Hbase数据库。
- (2)历史数据迁移、解析、导入Hbase数据库。
- (3)存储空间满足情况下，HDFS内以SEED文件形式存储一份地震波形数据，以满足灾备安全需求。
- (4)以源数据端要求进行数据级容灾。

(下转第31页)