赵辉 . 地震监测数据的 Hadoop 存储解决方案 [J]. 华南地震 , 2020, 40（3）：70-75. [ZHAO Hui. The Hadoop Storage Scheme of Seismic Monitoring Data [J]. South China journal of seismology, 2020, 40（3）：70-75]

地震监测数据的 Hadoop 存储解决方案

赵 辉
（内蒙古自治区地震局，呼和浩特 010000）

摘要：根据地震监测数据日常存储实际情况，结合目前主流的 Hadoop 技术，提出地震监测数据的大数据存储方案，以解决数据存储设备选购常见问题，并对数据存储环境进行详细配置。以内蒙古自治区地震局测震数据为样本，在虚拟机上搭建伪分布集群方式，对数据进行上传及下载操作，以验证该方案的可行性。结果表明：使用 Hadoop 技术进行地震监测数据存储可靠性更高，且在对历史地震监测数据进行处理时，分析统计手段更多，数据读取速度更快。

关键词：Hadoop；大数据；数据存储；地震监测
中图分类号：P315.73　　**文献标志码：**A　　**文章编号：**1001-8662（2020）03-0070-06
DOI：10.13512/j.hndz.2020.03.010

The Hadoop Storage Scheme of Seismic Monitoring Data

ZHAO Hui
(*Inner Mongolia Autonomous Regon Earthquake Agency*, *Huhhot* 010000, *China*)

Abstract: According to the actual situation of daily storage of seismic monitoring data and the current mainstream Hadoop technology, this paper proposes a big data storage solution for seismic monitoring to solve the common problems of selecting and purchasing data storage equipment, as well as configuring the data storage environment in detail. Taking the seismic data of Inner Mongolia Earthquake Agency as a sample, the paper builds a pseudo distributed cluster on the virtual machine, uploads and downloads the data to verify the feasibility of the scheme. The result shows that the Hadoop technology is more reliable for seismic monitoring data storage, and when processing historical seismic monitoring data, there are more analysis and statistics methods with more faster data reading speed.

Keywords: Hadoop; Big data; Data storage; Deismic monitoring

0 引言

在地震系统内，日常的地震监测数据都是保存在服务器或者相关电脑上，对数据的应用也是将数据保存到本地主机后再进行分析使用，然而随着日常地震监测数据量的增长，常规的单机以及服务器存储方式逐渐将无法满足日常监测数据的存储需求。以测震数据为例，内蒙古自治区地震局共有测震台站 48 个，日均产生的测震数据在 1.4 G 以上，这些数据在存储的过程中又分为“台站卷”与“台网卷”，所以每

收稿日期：2019-11-06
基金项目：内蒙古自治区地震局局长基金项目资助（项目编号：2019ZC25）
作者简介：赵辉（1990-），男，助理工程师，主要从事办公网络维护、数据存储研究工作。
E-mail: 363359122@qq.com

日最终所产生的需要存储的测震数据就在 4.48G 以上，光测震数据的年均存储量就在 1.6T 以上。

因此，为了满足地震监测数据的存储需求，创新性的将目前较为流行 hadoop 技术应用到地震监测数据的存储中，既解决了海量监测数据的存储问题，也增加了监测数据的分析手段。本文将会从物理环境到软件配置以及对数据的简单操作给出相应的实现方法和解决策略，以实现对地震监测数据的大规模存储。

1 关于 Hadoop 和大数据

Hadoop 是一个由 Apache 基金会所开发的分布式系统基础架构，主要解决海量数据的存储和海量数据的分析计算问题，具有高可靠性、高扩展性、高效性、高容错性等特点。广义上来说，Hadoop 通常指一个更广泛的概念—Hadoop 生态圈。

Hadoop 由 HDFS、MapReduce、YARN、Common 等组件构成，这四个组建是其基本生态圈的框架。

1.1 MapReduce 离线计算

MapReduce 是 Hadoop 中的计算工具，主要用于进行离线计算，其将计算过程分为两个阶段。首先 Map 阶段对输入的数据进行并行处理，然后 Reduce 阶段对 Map 阶段所处理数据的结果进行汇总。将 MapReduce 技术应用到对地震监测数据的分析上，便可以实现对历史地震监测数据的大规模并行处理分析。这样在分析过程中，所获得的数据将更为全面准确。

1.2 YARN 资源调度

YARN 是 Hadoop 的资源管理器，主要负责集群的资源管理和调度，它的引入为集群在利用率、资源统一管理和数据共享等方面带来了巨大的好处。

YARN 主要由四部分构成。ResourceManager 是总调度，用来处理请求，资源分配与调度。NodeManager 负责单个节点上的资源管理。ApplicationMaster 负责管理任务 job，申请资源。Container 是对任务运行环境的抽象，封装了 CPU、内存等多维资源以及环境变量、启动命令等，任务运行相关信息（图 1）。

由于 YARN 在整个 Hadoop 中起着最为关键的作用，因此在部署集群的过程中，YARN 是否能部署好，直接关系到整个方案的成败。下文在部署集群时将会着重说明如何部署 YARN。

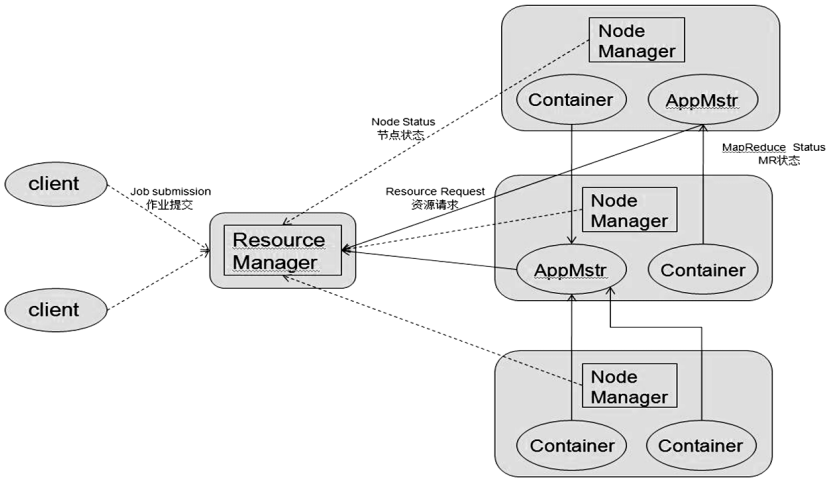

图 1 YARN 资源调度流程图

Fig.1 Resources schedule flow chart of YARN

2 地震监测数据存储方案

以测震数据为例，内蒙古自治区地震局 48 个测震点分布相对稀疏，需要处理的测震数据相对较分散。因此从测震数据的使用需求来看，更加侧重于数据分析，即对历史测震数据进行分析研究。但是，现阶段采用的是将监测数据存储到服务器与 PC 扩展硬盘上，随着数据量的不断增长，存储介质易损坏、恢复难、数据导

入导出慢等问题也日益突出，构建可靠性高的存储分析系统也更为重要。

2.1 监测数据的存储结构

以测震数据为例，目前国家地震台网中心采用基于 JOPENS 的 SSS 流服务器来负责接收和分发波形数据，接收 512 字节的纯数据 SEED 卷（Miniseed 数据），包含有固定头段本分（48 字节）和数据部分，主要记录台站名、起始时间、样本数目、测震数据等，以一个台站一个分项[1]。而对测震数据的分析与应用所使用的软件是 Jopens-msdp5.2, 所以考虑到应用数据时的便捷性，我们在数据的存储过程仍采用这一结构对数据进行存储，数据包为 seed 格式。同理，其他监测数据也均采用目前所使用的格式直接进行存储。

2.2 大数据的数量级

数据的存储结构可以确定后，就需要确定数据的数量级以便选择合适的硬件。内蒙古地震局测震数据存储量年均在 1.6 T 以上，加上分析后所产出的测震数据，年均需要存储量在 6T 以上。另外还有前兆的相关监测与分析数据，年数据规模在 10 T 以上。面对如此规模庞大的数据量需要引入大数据的数量级这一概念来加以探讨。因此以大数据的数量级来看，根据内蒙古地震局的数据规模以及确保系统有一定的数据冗余这两点要求来确定大数据数量级，需要达到 EB 级。所以在选择服务器与存储矩阵时要遵循 EB 级的数量级这一原则。

2.3 物理硬件选择

对集群硬件的选择往往被轻视，认为选择配置较高价格较贵的硬件既可满足一切需要。然而几乎在所有情形下，MapReduce 要么会在从硬盘读取数据时遇到瓶颈，要么在处理数据数时遇到瓶颈，前者称为 IO 受限后者称为 CPU 受限。所以在选择硬件时，一定要根据工作负载进行选择。在这里可以将根据工作负载选择硬件的一般原则总结成一张图（图 2）。

因此，根据对使用需求以及内蒙古自治区地震局现有的监测数据量进行分析，依据工作负载，在性能和经济性上选择硬件的最佳平衡。基于硬件选择的规律，在这里选择了通用的服务器均衡配置，结果见表 1。

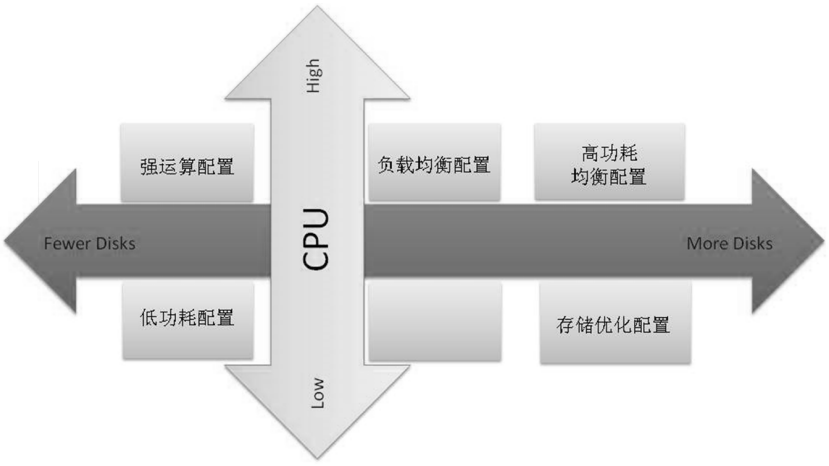

图 2 根据工作负载选择服务器原则

Fig.2 Principles for selecting servers based on workload

表 1 服务器配置方案表

Table 1 Server configuration schema table

服务器	CPU	内存	Disks
1U	4 * hex-core cpus	512G	4 * 8T

2.4 集群规划部署

HDFS 的三个组件均需要部署，YARN 的四个组件其中两个组件需要部署。HDFS 中的 NameNode、SecondaryNameNode 比较占资源，需要部署在一个节点上，DataNode 为实际存放数据的模块，根据需要部署即可。YARN 中需要部署的 ResourceManager 较占资源，需要部署在一个节点上。NodeManager 与 NameNode 对应部署。因此，根据上述原理，我们为集群规划了 3 台服

务器，并按表 2 进行规划部署。

从表 2 中可以看到将比较占用资源的 NameNode、SecondaryNameNode、ResourceManager 部署在了不同的节点上。随着数据量的增多，如果需要对集群进行扩展，我们也可以将这三个较为占资源的服务部署在三个独立的服务器上，其余服务器均作为数据存储服务器，便于日后进行再次扩展。

表 2 集群规划部署表

Table 2 The table of cluster planning and deployment

	Server01	Server02	Server03
HDFS	NameNode DataNode	DataNode	SecondaryNameNode DataNode
YARN	NodeManager	NodeManager ResourceManager	NodeManager

2.5 环境配置

根据上述原则做好集群的搭建后，需要对集群环境进行配置。选择 Linux 系统一般都会选择较为稳定的 CentOS-6.8 版本。同时需要安装 JAVA、ant、maven 和 protobuf 工具为编译 Hadoop 做好准备。在这里之所以要重新配置环境变量以及编译 Hadoop 的源代码，是为了让整个系统与硬件更加适配。所配置的软件与 Hadoop 源码参考版本如下：

jdk-8u181-linux-x64.tar、ant-1.9.14-bin.tar、maven-3.6.1-bin.tar、protobuf-2.5.0.tar、glibc-2.14.1.tar、hadoop-2.7.7-src.tar。

在这里需要需注意的是，CentOS7.1 以前的版本，默认支持的 glibc 最高版本为 glibc-2.12，而 hadoop-2.7.1 以上的源码所需要的 glibc 版本需要在 2.14 以上，如果我们不对 CentOS-6.8 中的 glibc 进行升级的话，在进行源码编译的过程中会报错，从而导致编译失败。因此在这里需要下载 glibc-2.14.1 或以上的版本对原系统中的 glibc 进行升级。

对相关环境做好配置后，再根据具体情况与参数做好 core-site.xml、yarn-site.xml、mapred-site.xml 这 3 个核心配置文件的配置，这样 Hadoop 的存储集群就搭建完毕了。

3 可行性验证

为验证上述方案的可行性，我们通过在电脑上安装 VMware 软件，并搭建伪分布式集群的形式来进行测震数据的上传存储并对已上传的数据进行简单的操作以验证上传数据的可操作性。所使用的电脑配置参数如下表 3 所示。

表 3 电脑配置参数

Table 3 Computer configuration parameters

CPU	内存	操作系统	硬盘
CORE i5 8th Gen	16G	Win10 x64	1T

在 VMware 上安装 Linux 的 CentOS-6.8 版本，并在此系统上搭建伪分布式集群，做好相应的集群环境准备。配置好相应的环境后，安装 Hadoop2.7.7 版本。

3.1 监测数据上传测试

启动 hadoop，创建“tz”（台站卷）与“tw”（台网卷）文件目录，分别存储对应数据。使用“hdfs dfs –put”命令分别上传内蒙古自治区地震局 2018 年 1 月 1 日的台网卷与台站卷的部分测震数据进行测试 (图 3)。

从图中可以看出，名为 20180101.NM.AGL 的台站卷与 2018010100.NM 的台网卷数据已经上传，并提供了可下载数据的对应地址。因此，此次实验的测震数据已成功上传。

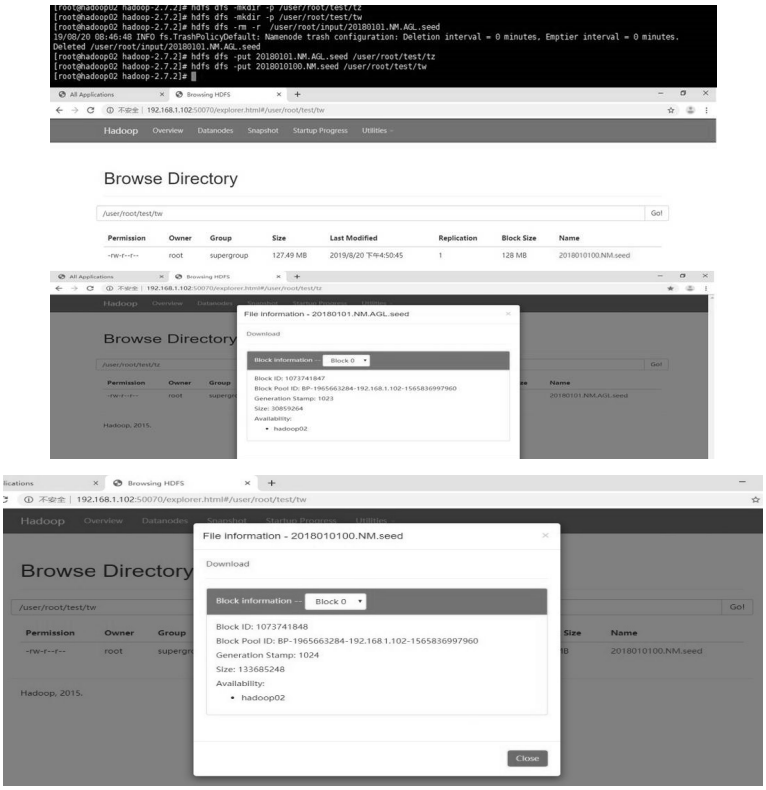

图 3 数据上传测试图

Fig.3 Data upload test chart

3.2 已上传数据可操作性测试

我们继续使用 YARN 运行简单的 MR 程序对已上传的数据进行运算操作，并通过 Hadoop 的 All Applications 进行监控 (图 4、图 5)。

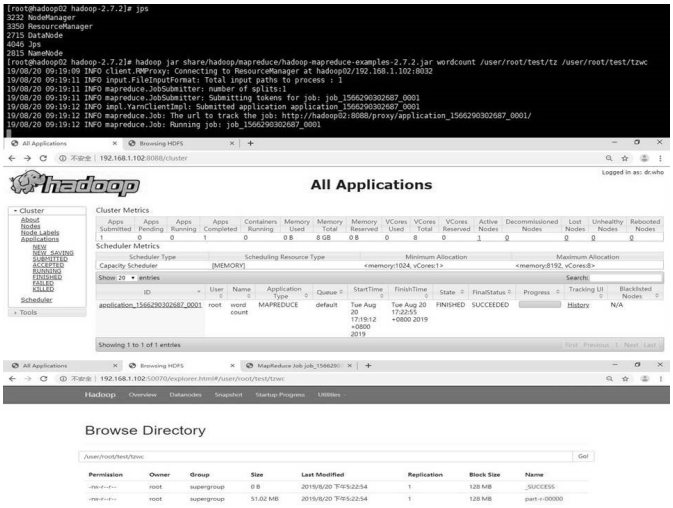

图 4 数据操作图

Fig.4 Data manipulation chart

从图 4 与图 5 可以看出，对台站卷数据 20180101.NM.AGL 与台网卷 2018010100.NM 的 WORDCOUNT 操作已经成功运行，分别用时为 1 分 58 秒与 4 分 39 秒，并在文件夹“tzwc”与“twwc”产出了对应的数据文件。因此，上传至服务器的测震数据文件均可正常使用，并可使用大数据技术对已上传的数据进行操作。

图 5 数据操作记录图

Fig.5 Data operation record chart

4 总结

根据上述的测试结果来看，本文所提出的地震监测数据的 Hadoop 存储解决方案可行。虽然本方案还未应用到内蒙古地震局实际工作中，但使用 Hadoop 技术对地震监测数据进行存储与分析势必将应用到整个地震行业。Hadoop 技术不仅可以使地震监测数据在日常存储上具有更高的可靠性，同时在对历史地震监测数据进行处理时，具有了更多的分析统计手段和更快的数据读取速度。同时在地震监测数据不断增长的情况下，使用 Hadoop 技术对监测数据进行存储日后对服务器进行存储扩展也极为便利。

参考文献：

[1] 周辉，申学林，王文青，等 . 通用测震数据获取软件包的设计与实现 [J]. 地震研究 , 2011, 34(1) : 102-107.

[2] Tom White.Hadoop 权威指南 [M]. 北京：清华大学出版社 , 2018.

[3] 刘彬斌，李柏章，周磊，等 .Hadoop+Spark 大数据技术 [M]. 北京：清华大学出版社 , 2018, 95-182.

[4] Richard B, Christine B.Linux 命令行与 shell 脚本编程大全（第 3 版）[M]. 人民邮电出版社 , 2016.

[5] 王丹宁，柴旭超，王文青 .Hadoop 平台下的地震波形数据存储与应用规划 [J]. 软件工程 , 2016, 19(1)：48-49.

[6] 徐德智，刘扬 , Sarfraz Ahmed. 基于 Hadoop 的 RDF 数据存储及查询优化 [J]. 计算机应用研究 , 2017, 34(2)：477-486.

[7] 王林童，赵腾，张焰，等 . 基于 Hadoop 的风力发电监测大数据存储优化及并行查询方法 [J]. 电测与仪表 , 2018, 55(11)：1-6.

[8] 王海荣，刘珂 . 基于 Hadoop 的海量数据存储系统设计 [J]. 科技通报 , 2014, 30(9)：127-130.

[9] 章静，林捷，杨乐 . 测震专业软件评估平台在地震行业中的应用 [J]. 震灾防御技术 , 2013, 8(3)：326-333.

[10] 路婧 . 地震数据分析中格式转换问题 [J]. 科教导刊 , 2016(1):177.