E-mail: csa@iscas.ac.cn http://www.c-s-a.org.cn Tel: +86-10-62661041

基于 Apache Spark 的地震观测数据噪声功率谱 计算^①

郭 凯^{1,2,3},黎建辉¹,温亮明^{1,2},韩振华⁴

¹(中国科学院 计算机网络信息中心,北京 100190) ²(中国科学院大学,北京 100049) ³(中国地震台网中心,北京 100045) ⁴(太原理工大学,太原 030024) 通讯作者:黎建辉, E-mail: lijh@cnic.cn



GL

摘 要: 为解决单机环境下海量地震观测数据计算和分析效率低下的问题, 提出一种基于分布式架构的地震观测数据的存储、计算和分析处理方法, 选择噪声功率谱复杂计算过程的应用场景进行实现. 基于 Hadoop 在海量数据处理上的性能优势, 在分布式文件存储系统 HDFS 上进行地震观测数据的存储和调度, 研究测震数据噪声功率谱的质量评估方法在 Spark 分布式计算架构上的实现, 采用弹性数据集 Spark RDD 将计算任务自动分配到计算节点, 解析存储在 HDFS 中的测震波形数据, 计算结果采用 RowKey 方式放入分布式数据库 HBase 中, 实现了长周期地震噪声功率谱结果的存储和提取. 计算结果表明, 基于 Spark 分布式架构的该方法可以支撑 TB 级海量数据的处理, 并且具有较高的处理效率, 可应用于海量地震观测数据的分析计算.

关键词: 地震观测数据; 噪声功率谱; Spark; Hadoop; 分布式

引用格式: 郭凯,黎建辉,温亮明,韩振华.基于 Apache Spark 的地震观测数据噪声功率谱计算.计算机系统应用,2021,30(8):126-132. http://www.c-s-a.org.cn/1003-3254/8084.html

Noise Power Spectrum Calculation Method of Seismic Data Based on Apache Spark

GUO Kai^{1,2,3}, LI Jian-Hui¹, WEN Liang-Ming^{1,2}, HAN Zhen-Hua⁴

¹(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

²(University of Chinese Academy of Sciences, Beijing 100049, China)

³(China Seismic Network Center, Beijing 100045, China)

⁴(Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: To solve the problem of inefficient calculation and analysis of massive seismic data in a single machine environment, we propose a distributed architecture based method for storage, calculation, and analysis of seismic data and select the complex calculation process of a noise power spectrum as the application scenario for implementation. In light of Hadoop's performance advantage in massive data processing, the storage and scheduling of seismic data are carried out on the Hadoop Distributed File System (HDFS). The implementation of the quality evaluation method for the noise power spectrum of seismic data in Spark distributed computing architecture is studied. The elastic dataset Spark RDD is used to automatically allocate the tasks to the computing nodes, and the seismic waveform data stored in HDFS is analyzed. In addition, the calculation results are input into the distributed database HBase in the RowKey mode, realizing the storage and extraction of the power spectra of long-period seismic noise. The calculation results show that the method based on

收稿时间: 2020-11-11; 修改时间: 2020-12-12, 2021-01-22; 采用时间: 2021-01-26; csa 在线出版时间: 2021-07-31

① 基金项目:国家重点研发计划 (2018YFC1504500);中国地震局监测、预报、科研三结合课题 (3JH-20200207)

Foundation item: National Key Research and Development Program of China (2018YFC1504500); Research Project for Monitoring, Forecasting, and Scientific, China Earthquake Administration (3JH-20200207)

¹²⁶ 软件技术•算法 Software Technique•Algorithm

Spark distributed architecture can support the efficient processing of massive data at the TB level in volume, which can be applied to the analysis and calculation of massive seismic data.

Key words: seismic data; noise power spectrum; Spark; Hadoop; distributed

在随着国内地震观测台站的持续布局,我国的地 震观测台网不断加密,截至2018年底,中国地震台网 中心实时汇集的测震台站数量已经达到了1107个,地 震观测数据每年汇集的量级已经达到数十TB级别.如 此海量级数据体量对数据的汇集和分析带来了巨大的 挑战,采用传统单机模式进行的计算和分析在时效性 上已经无法满足要求.

以数据质量评估为例, 地震观测台站记录系统瞬态变化、仪器毛刺、数据记录的阶跃、尖峰以及由于 系统故障引起的信号失真和重大环境的干扰等, 都是 影响数据记录质量的重要因素, 国际上主要采用噪声 功率谱 PSD^[1]和功率谱密度概率密度函数 PDF 等方 法进行台站的噪声水平监测^[2]. 目前美国地震学研究联 合会 (IRIS) 数据管理中心 (DMC) 将相关方法集成形 成了单机运行的软件 PQLX 和相关脚本程序, 国内相 关研究开发大多都采用 Matlab 等单机程序开发编制^[3]. 这些处理方法计算效率受限于单机处理能力, 在面向 海量地震观测数据计算时均存在磁盘 IO 瓶颈, 主要原 因在于其处理保存的数据量和计算结果非常有限, 一 般仅能处理半年到一年之内全国监测台网的数据, 不 利于对地震观测台站进行长周期的数据质量评估和挖 掘分析.

随着大数据技术的高速发展,以 Hadoop 和 Spark 为代表的开源大数据平台在处理海量数据的 IO 并发 和处理速度方面都体现了巨大优势,它们为存储和处 理大数据提供了动态、弹性和可伸缩的数据存储和分 析解决方案,在密集科学数据处理和分析挖掘应用中 取得了较好的效果^[4],如生物学领域的密集计算^[5]、海 量行人轨迹数据的挖掘分析^[6]、海量日志文件的分析 等.在地震学领域也开展了相关研究^[7],如基于 MapReduce 实现地震数据的成像^[8],基于 Apache Spark 实现地震 事件的快速分类等^[9].尽管相关研究工作已经表明大数 据技术在海量地震观测数据处理上具备可行性和应用 性能优势,但成果的重心偏向于海量数据计算速度上 的提升,而对数据的全链条处理和分析较少关注.本文 试图从海量测震波形数据的分布式汇集、存储到分布 式计算架构构建开展数据全链条研究,选择在地震观 测数据质量评估中需要密集计算的噪声功率谱方法进 行具体实现.

1 地震观测数据分布式处理框架

1.1 地震观测数据的分布式汇集和归档

中国地震台网中心实时汇集的地震观测数据采用 国际标准的 MiniSeed 格式,每一个文件中存储了一个 台站一个分项的 24 个小时观测数据,数据汇集在磁盘 阵列、网络附属存储 (NAS) 等存储介质中,全国 1107 个台站的具体分布如图 1 所示.



按照 NAS 的存储性能和网络带宽限制, 一般每秒 可处理的数据吞吐量在几十 MB 左右, 显然在面对 TB 级别的业务计算场景时, 该速度已经无法满足要求. Hadoop 是一种专门用于批处理的分布式系统基础架 构, 其通过配合使用多个组件来实现批处理业务: ① 分 布式存储层 HDFS, 协调集群节点间的存储和复制, 作 用于数据来源, 保障整个集群中发生故障时数据的冗 余, 用于存储中间态的处理结果和计算的最终结果, 完 成海量数据的存储; ② 资源调度层 YARN, 协调并管 理底层资源和调度作业的运行. Hadoop 广阔的生态系

Software Technique•Algorithm 软件技术•算法 127

统和与其他框架/引擎之间良好的兼容性与集成能力使 其成为多种工作负载处理平台的底层基础.我们选取 了 2013 年 1 月~2020 年 10 月的地震观测波形数据进 行迁移,目前已迁移约数据体量约为 70 TB,通过万兆 网络将这些数据传输到分布式文件系统 HDFS 中, 按照 1:3 的比例对数据进行副本设置以确保归档数据的 安全性,即将一份数据分布存储在 3 个数据节点上 (如图 2 所示),归档数据按照年/月/日/台站_测项的路径进行存储.

			File inform	nation - AH.BAS.00.BHE	E.2020010	1000002.mseed			
Brows	e Dir	ectory	Download						
klala/historyOat	12020/2020	01/20200101	Block into	rmation - Block 0					Gal
Permission	Owner	Group	Block ID:	1106711469	0 1 14006438	79799			
-rwse-xr-x	hadoop	supergroup	Generatio	in Stamp: 32991399	0.1-14000430	raraa		0.8HE.30200101000001.miened	
-rwar-xr-w	hadoop	antesBronth	Size: 125	93152				0 BHON 2020030 1000000 mileon	
-(WXE-X1-#	hadoop	supergroup	Availabilit	ý.				0.8HZ.20200101000004,mited	
-IWRI-XI-K	hadoop	supergroup	 coll stor 	iector-master1 rage-datanode2				0.8HE 20200101000002 minore	
-PM007-XT-8	hadoop	supergroup	• sto	rage-datanode1				C.BHN.2020010(00000.mieed	
-IWAT-ND-R	hadoop	supergroup						0.8H2.20200101000002 meetod	
-rwxt-xr-x	hadoop	aupergroup						0 BHE 20205101000001 meeed	
-IWKI-XI-X	hedoop	supergroup					Close	0.6HN.20200101000003.miced	
-IWAT-XI-R	hadoop	supergroup					_	0.8H17.20200101000003 maeeed	
-fwxt-xt-x	hadoop	supergroup	10.99 MB	2020年1月10日 6:41:06	3	256 MB	AH.82H.0	00 BHE 20200101000002 miseed	
-fwxt-xt-x	hadoop	supergroup	11,24 MB	2020년1月10日 6:41:06	3	258 MB	AHBZHU	00.8HN 20200101000004 inseed	
-FWX7-XF-X	hadoop	supergroup	10.47 MB	2020年1月10日 8:41:06	3	200 MG	AH DZH.	00.8HZ.20200101000001.maeoo	
-fwxt-xt-x	hadoop	supergroup	11,71 MB	2020年1月10日 8:41:06	3	250 MB	AH BZY I	00 SHE 20200101000002 manua	
-PW007-XT-30	nadoop	supergroup	11.65 MB	2020年1月10日 8:41:07	3	256 MB	ALLEY	00.5HN 20200 (01000001.maeed	
	netopo	aeoeraraub	9.79 MB	2020年1日10日8-41-07	3	256 MB			

图 2 地震观测数据的分布式存储归档

1.2 噪声功率谱的分布式计算实现

目前主流的开源分布式计算系统主要有 MapReduce、 Spark、Flink、Storm 等^[10], MapReduce 在数据处理过 程中需要在硬盘进行读取和写入,降低了运行速度,而 随着硬件性能的不断提升,基于内存计算的分布式计 算框架相对 MapReduce 有了明显提升. 针对海量地震 观测数据进行高并发的计算与分析,建立一套具有高 可靠、高处理性能、可在线弹性伸缩、可不间断接收 任务的处理模型,由于本文要实现的数据密集计算场景 一般以小时为单位进行计算,所以在计算架构上不考 虑应用与实时数据解析计算相关的 Storm 和 Flink 系 统. 同时由于 PSD 算法包含了复杂的迭代过程, 相对 于 Hadoop 的 MapReduce 而言, Spark 在这方面有着较 明显的优势^[11].综合以上多种因素,我们的数据处理模 型采用分布式内存计算平台 Spark,聚焦海量地震观测 数据的分布式噪声功率谱计算进行模型设计和实现, 如图3所示.

图 3 中,处理模型从上至下共分为 3 个层次.最上 层为数据源层,主要来源于地震历史观测数据文件,这

128 软件技术•算法 Software Technique•Algorithm

些文件存储在台网中心的 NAS 存储服务器上,通过 HDFS 的标准接口将 NAS 系统上的数据推送到 HDFS 文件系统中.



图 3 地震观测数据的分布式处理模型设计

中间层为数据处理层,主要完成数据计算和数据 处理过程,提供基于地震观测数据的预处理、台站仪 器基础数据以及分布式架构的 PSD 数值计算. Spark 的主节点在接收到前台传递需要处理计算的参数后 (日期、台站以及通道)分配处理任务所需的资源后, 将任务分发到各个工作节点,由各个节点完成计算过 程并将结果反馈给主节点. 当处理层在接收到 PSD 计 算请求时,算法调度模块根据请求参数中的时间范 围、台网以及台站等参数查询出需要计算的每个文件 的具体位置,同时将文件位置以及任务 ID 推送至计算 模块的 sever 端; 计算模块在接收到计算请求后, 通过 参数中的路径直接从 HDFS 上获取需要处理的 mseed 文件,并对每个 mseed 文件进行计算. 这里 PSD 的计 算模块主要采用美国地震学研究联合会 (IRIS) 数据管 理中心 (DMC) 提供的 PSD 计算程序, 数据计算完成 后,计算模块将计算结果生成在一个临时文件中(频率 和 PSD 数据值),并通过 sever 将计算结果和临时文件 的路径通知算法调度模块,算法调度模块根据临时文 件路径去解析临时文件并与查询参数组成的 RowKey 一起存入 HBase 数据表中.

最下层为数据存储层,主要提供处理层中各类计 算结果的存储,包括 PSD 值和 PDF 值.由于全国台站 每个分项的观测数据计算频率为每小时一次,因此需 要将这些大量的结果数据存储在分布式数据库 HBase 中,通过 HBase 接口进行存储和访问处理.当大量的计 算结果到 HBase 数据库时,由于 HBase 表会被划分为 1,…,*n* 个 Region,被托管在 RegionServer 中. Region 两 个重要的属性: start-key 与 end-key 表示这个 Region 维护的 RowKey 范围,当我们要读/写数据时,如果 RowKey 落在某个 start-end key 范围内,那么就会定位 到目标 region 并且读/写到相关的数据.鉴于几点,在 设计表将计算结果按照台站首写字母进行排序存放在 一个 region 里面,使得数据检索更加快捷方便,尤其在 scan 某个通道的数据设置 start-key 和 end-key 时,只需 在某个 Region 里面检索就可以了;各个 Region 分散在 集群的各个位置,增加了数据读写的并发量.这里数据 处理模块按照小时对地震观测数据的 PSD 值进行计 算,采用 RowKey 格式为"台网代码_台站代码_测项代 码_年_月_日_小时", colum 为"value",储存值为以某个 小时的 PSD 数值.

2 实验验证

2.1 测试环境构建

为了验证以上分析处理框架的有效性,我们搭建 了测试系统进行实验验证.设计的系统总体软硬件部 署如图 4 所示,采用 14 台高性能服务器搭建了具体的 系统集群环境 (如图 5 所示).14 台服务器的具体部署 情况为:2 台作为管理节点,8 台存储型服务器作为存 储节点,4 台服务器作为计算节点.各类型服务器的基 本配置信息如表1、表2所示.所有服务器均为机架式 服务器,配置12 个可热拔插的硬盘接口,其中存储服 务节点挂载12 块4 TB SATA 硬盘,计算节点配置 4 块 600 GB 高速 SSD 硬盘.



图 4 计算节点数据处理流程

Software Technique•Algorithm 软件技术•算法 129



图 5 软硬件部署图

表1	存储型服务器基本配置信息
技术指标	参数
服务器外观	机架式,2路服务器
处理器类型	Intel Xeon E5-2650 V2 系列处理器
内存配置容量	128 GB
内置硬盘	48 TB, 4 T Sata*12
-	

表 2 计算型服务器基本配置信息

技术指标	参数	
服务器外观	机架式,2路服务器	
处理器类型	Intel Xeon E5-2650 V2 系列处理器	
内存配置容量	96 GB	
内置硬盘	2.4 TB, 600 GB SSD * 6	15

2.2 性能测试

基于上述搭建的环境,我们以不同数据量开展了 计算测试.为了对系统整体计算性能进行充分测试,按 照4个计算节点,以不同数据集计算时间来进行评估. 如图6所示, Data Number 为计算任务包含的单个台 站24小时记录的连续波形数据数量,每一条数据量约 12 MB. 计算结果显示,集群在数据处理能力上处理时 间与任务量呈现很好的线性关系,处理速度并不会因 为数据量的极速增加而变慢,处理完成1个月全国汇 集的地震观测数据,数据量约1 TB的计算约需要 60个小时,而基于单节点,采用 IRIS DMC 的 PSD 程 序开展同样计算需要处理的时间可能会超过了10天, 且在海量数据的处理和计算结果存储上会受到单机计

130 软件技术•算法 Software Technique•Algorithm

算节点性能的影响.通过整体评估以及已有研究工作^[12], 单机节点计算与本文采用的方法计算速度比值约为 1:*N*, *N*为计算节点的数量,即集群可以直接通过增加计算 节点提升集群处理速度,具有很好的扩展性,面向数据 密级型的业务计算和存储分析具有很好的性能表现.

在计算结果的存储和分析上,我们按照 RowKey 的方式进行计算结果的存储,如图 7(a)所示,一条数据 表示一个台站单个分项某个小时的 PSD 值,图 7(b)为 对存储计算结果的数量统计,目前存储条数已经超过 了 2000 万条.而如此海量计算结果的可以通过 HBase Shell 或者 HappyBase 等多种方式进行高效的检索和 获取,在大数据分析和处理上相对单机节点具有明显 优势.时间分布结果如图 8 所示.

3 结论与展望

本文采用目前主流的 Hadoop 和 Spark 技术, 基于 海量地震观测数据的业务应用场景, 设计了基于 HDFS 的地震观测数据分布式归档和基于 Spark 分布式计算 架构, 适用于数据密级业务场景的计算和分析. 从开展 的不同数量级计算任务来看, 本文所提出的方法在原 始程序数据处理流程上优化了数据的接入和处理环节, 计算速度不会因数据量的大幅度提升而下降, 解决了 单机节点处理海量数据的 IO 瓶颈和计算瓶颈, 大规模 提升了计算能力和可处理数据能力; 在普适性方面, 本 文设计的方法更适用于海量数据的计算和管理,随着 数据量的增加不需要频繁建立新表和选择存储环境, 单机环境下计算和处理存储的数据量有限,数据量较 大时需要投入更多人力处理环节.

Datano	de Inform	ation									M	aster store	ge-master	ŧ.										
in operation	s										Re	gion Servi	ers											
2.4		Calcard Na		East of the	Fe 272.2		1.1.1	Decederal	10.10		24	Interest	Rapoite	24.00'20	Corpedia	10								
	ensient with their					1275		****	1.4	377	See	NTRATE:				lart time			- 24	Request	s Per See	pend	Num 19	NOR NO
10.00	1 1005 102 1003	12	wat to s	2 # TT P.T	T 46 T	1,915		- and share	0.5			ALC: 144417 1007	0.1008321-00	14			a so cort			100			100	
							21.4			20	112	different sectors			- 2				1					
			-							1.22	100	NOT THE PARTY OF	0.1566372333			Auj 24 -4 9	साल्झ	nen		110			120	
	a ape so a resource									100	1830	apa-teanada) ini	011 155625220	1241		or Aug 24 14 59	940 CST	of CH1	- 29	110			120	
			web	max	1. 14.11	19931				14	nkz	ape-teneosterid II	ACAU 4546252	2600		er Aug 24 14 54	120-051	1600	33	15			628	
consideration.	NORMATING AND	a 15		AN LOS SAU	0.011	10210	290.0	o and the second	0000	0.0475			TAL BARRAGE	1.774		to Area of C.S.D.	and the	1/11					12.14	
Iver de unit		× 13			-		2.1		1	213		ay, ar anne.											140	
											#35	ays datanois? (fi	00.159625222	6425		55 Aug 21 - 4 3	san car	2020	1	128			125	
Commence of the											680	apa-daranoda'i 174	PR 1596257 R	1114		or Aug 24 14 5	93PCST	non	83	204			125	
Distantistic	2.0.14										ale	spe-translet its	(A), 1558(25)21-2	730		er August 1454	100.01	25,631	81	136			427	
											1.0	46							- 33				1000	
	inda. bi	a as had, dt oaks		I take	a de alter as i			Barry Section (1)	and a state of the		10.								- 13				1090	
Spark Mas	iter at spark://caci	(a) dation-master	HC	DFS 徇	 拿理身	雨					Sport no	ain Tapo Scop	locret lasks	(b) H	Base	管理	.界面	đ				unperitti	000240
Sola Los Spark Mas Spark Mas (Sola Spark Mas (Sola Spark Mas (Sola Spark Mas (Sola Spark Mas) (Sola Spark Mas (Sola Spark Mas) (Sola Spark Mas	iber at spark;//Cacc me:	(a) dation-master	HC	DFS 行	 穿理界	雨					Spail 200 Spark Jobs	aan Sapo Sengo N	locanat banka	(b) H	Base 1	管理	界面	đ				unter data	0.002949
Via Lo. Spark Mass Unit Spark Mass Consistence super 1/01 1 sector functional are 0 for 1 miles are 0 for 1 miles Areany in Consistence Areany in Consis	ther at spark:@caci tota.com	(a)	HC	DFS 行	管理身	雨					Sport 211 Spark Jobs Decard Total Jobs (21)	ala Sapo Senja N	locrat Luniu	(b) H	Base 1	管理	界面	1				unpartit	0000340
Sub Los Spark Mass Construction Super Vite Service Status, Character Super Status, Character Super Status, Character Super Status, Character Super Status, Character Super Status, Character Super Status, Character Status, Charact	ther at spack://caco mm.cac.esti mm	(a)	HD	DFS 行	管理界	雨					Sport 2017 Spark Jobs Decard Total Jobs (21) Sciences Role 1	ata Sapo Senye N	locanat Lundo	(b) H	Base 1	管理	界面	đ				unpañasti	ware we
Kult Los 2015 Spark Man Constant of the Const Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State Second State of the Second State of the Second State of the Second State Second State of the Second	iter at spark;@caeco minimi	(a) dation master	HC 7077	DFS 行	管理身	雨		Manuary			Sport 211 Spark Jobs Decad Tel Jobs (71) Scholey Refs Colley Ade (- Sport Trees	ata Sapo Senye N	locanut Landa	(b) H	Base 1	管理	界面	đ				ith ^r eason	000234
Num Los Spark Mass Constantino mantel (2017 Constantino mantel (2017 Constantino mantel (2017) Constantino Constan	nter at spark://cacc mer	(a)	HC :7077	DFS 1	管理界	面		Mannay Addition	TT (self.		Sport Jobs Spark Jobs Decad Too Jobs (Job Saling R Jo Saling R Jo Saling R Jo Saling R Jo Saling R Jo Saling R Jos Saling R Job	ata Saya Singa N	Locarut Luniu	(b) H	Base 1	管理	界面	đ				internation of the second s	0.00240
Num Levi Spark Mass University and American University of the Color State of the Color St	ther at spach: (Coace Min come only) int int int int int int int int int int	(a)	HC: 7077	DFS 1	管理界	T T T T T T T T T T T T T T T T T T T	an Al Jones Al Jones Al Jones	Monary 41.5.6.0 50.5.00.0	TT 1940 T 1940 T 1940		Sport 211 Spark Jobs Decad Totel Jobs Schwarg Rei Company Comp	ata Dago Siraya N Deutoriuskel	bernet baaks	(b) H	Base 1	管理	界面	đ				itt ^a ngeu	scarste
Num Los Spark Mass Spark Mass (1997) Spark Mass	ther at spark;/Rases	(a)	HC :7077	DFS 1	管理界		er Bilandi Bilandi Bilandi Bilandi	Monary 41.5 da (ta 41.5 da (ta 41.5 da (ta 41.5 da (ta 41.5 da (ta)	10 court A 6 court 10 court 10 court 10 court 10 court		Spork 211 Spark Jobs Decad Tole Johns (21) Schalege Red. Colle Johnson Calification	All Sept Server	locanat Landa	(b) H	Base 1	管理	界面	đ				internation of the second s	0.000340
V.A. Low	nher at spark;//caco	(a) dation-master	1077	DFS 1	管理 界		Mi Mi Sandi Mi Sandi Mi Sandi Mi Sandi	Managy Mill Galp Mill Galp Mill Galp Mill Galp Mill Galp	10 court A 0 court B 0 court B 0 court B 0 court B 0 court B 0 court B 0 court		Spark Jobs Bernal Territoria Contago C	ate Sago Scope N Pol Events Weld Revent state	locanet Lande	(b) H	Base 1	管理	界面	đ				internation of the second s	sanse
V.B. Low Sparse Mass The Spa	ther at spack (ficaeco management) n n n n n n n n n n n n n n n n n n	(a) dation-master	HC: 7077	DFS 1	管理界		en 2 Januar 2 Januar 2 Januar 2 Januar	Monary 40.00 p 81.00 p 81.00 p 81.00 p 90 p 90 p 90 p 90 p 90 p 90 p 90 p	r U sautt D D sautt D D Sautt D D Sautt D D Sautt N D D D D D D D D D D D D D D D D D D D		Sport vice Sport Jobs Bread Holder: () Job Scharge Ref. () Scharge Ref. () Sch	30 Dage Strap 10	laanat laaka	(b) H	Base 1	管理	界面	đ				interferigen	NV BECCHI
V. K. Low Construction of the second	ther of systeh; (Coper and your week) and and and and and and and and and and	(a) dation-master	HC: 7077	DFS 1	管理		en 2 Fander 2 Fander 2 Fander 2 Fander	Monary Al Cala I No 2 Di No 2 Di No 2 Di No 2 Di No	r D seet. 7 D seet. 7 D seet. 7 D Seet. 8 O Res		Sport John Sport John Sport John Sport John Sport Sport John Sport	ate Sepo Scop N Subar Male Seate Scop Seate Scop	locrast basis	(b) H	Base	管理	界面	đ				ith ^r anner	
ka konstantiation and a second s	ther at spark;//Caaco Million and Statements Million Million Million Marco Million Marco Million Marco Million Marco Million Marco Million Mil	(a) distion-master	HE 7077	DFS 1	管理身		RE B Lands B Lands B Lands B Lands	Manuary 41.5 Ga (1 43.5 Ga (1 43.5 Ga (1 43.5 Ga (1 45.5 Ga (1 45.	r U soot U Saad St U Saad ST U Saad		Sport John Spart John	All Topo Scope	laenst baaks	(b) H	Base 1	管理	界面	đ				it for francess	
V.N. Ex- 2 (1) Spark Has conserve the set of 101 (1) Spark Has conserve the set of 101 (1) Spark Has (1)	ter al spark://cacc IIII	(a) dation-master	HE (7077	DFS 1	管理身 。 · · · · ·		00 01 Januar 01 Januar 01 Januar 01 Januar 01 Januar 11 Januar 11 Januar	Manany Al-Calo S AL-Calo S AL-Calo S Mara Al- Mara Man Mara Man Man Man Man Man Man Man Man Man Man	r 0 (sed) t 0 (sed) t 0 (sed) t 0 (sed) t 0 (sed) n Orr Same T second	tite See	Sport van Sport Jobs Souder Back Souder	at Dev Sorp	isonat isala settuat	(b) H	Base	管理	界面	đ				internet.	sor se
kai kai kai kai kai kai kai kai kai	ther all spark; (Eace BIR (see week) BIR (see week)	(a) Mation-master Memy p	HC :7077	DFS 1	管理身 。 ···································		ng 8 Harat 8 Harat 8 Harat 8 Harat 9 Harat 1 Harat 1 Harat 1 Harat 1 Harat	Abovery ALLAND MARINE Marine M	10 (1900) 10 (1900) 10 (1900) 11 (1900) 11 (1900) 11 (1900) 11 (1900) 11 (1900) 11 (1900) 11 (1900)	dan Dantan Se Ch	Sport 211 Sport Jobs Bacasi Totaleter 71 Same John Same	an Tapo Scop Duta Subb Duta Subb Duta Subb Duta Subb	iournet ionate entry (2005)	(b) H	Base	管理	界面	đ				interest of the second se	
V.b. Ex- Spark Mass Comparison of the second second second second second sec	her al speck:/ficer DR /max-mill mill CET / DE /	(a) dation-master	HC: 7077	DFS 1	学理身 ····································		98 8 Hanti 7 H	Manany Al 1640 Al 1640	7.0 (1998) 10.0 (1	Constant Size City Size	Sport 211 Sport Jobs Bannan Heiniger (7) and Salar and Salar and S	Max Dage Dage Dage Dage Dage	Lournet Louise #172(247) 2114 1	8 58	b) H	Base 1	管理	界面		208	10	83 ×	2049.000	b
k.k. Low Comparison of the second s	her al spark.if.ca.c Bit (una ministration of the second	(a) diation-master	4000 20077		学理身 ************************************		98 8 (sec) 8 (sec) 8 (sec) 9 (sec) 9 (sec) 9 (sec) 1 (Manary Al Sida () Maria () Ma	78 (1995) 10 (1996) 10 (19	ere Second Secon	Sport 2010 Sport Jobs The Jobs The Jobs The Jobs Sciences The Job Constants The Sport The Sp	In State Stop	bernet baske with 2007 Mail 1	8 528	b) H	Base 1	管理	界百		728	1400 PS M (degra	KD X	50000 ¹ 015	80
Value Environmental Section 2014 Value Section 2014 Val	ther at spark.(Reac: DR //www.ell) ther file ther	(a) dation-master	HC: 7077	DFS 1	学理身 " · · · · · · · · · · · · · · · · · ·		99 9 June - 9 June -	Anney At Call & At Call &	10 (1968) 10 (1969) 10 (1969) 10 (1969) 10 (1969) 19 (1969) 19 (1969) 19 (1969) 19 (1969) 19 (1969) 19 (1969) 19 (1969) 19 (1969)	000 100 100 100 100 100 100 100 100	Sport 215 2 Sport Jobs Dereil der Shart Jobs Dereil der Shart Jobs Dereil der Shart Jobs Dereil der Shart Jobs Dereil der Shart Jobs Dereil der Shart Jobs Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil der Dereil dere	No Seale Sea	bernet basin +rtc.ec. Int 2	1 53	b) H	Base 1	管理 ³¹	37		228	#0 Pilótan	22 1	Second Second	81
N.B Los 2 Spack, Maia 2 Spack, Maia 3 Spack, Maia 4 Spack, Maia 4 Spack, Maia 4 Spack, Maia 5 Spack, Maia 6 Spack, Maia 7 Spack, Maia 7 Spack, Maia 8 Spack 8 Spack<	her at spark.if.coc IRE / Inter weil IRE / Int	(a) diation-master	HIC (7077 480% 43.75 47.75 77.75 77.75 77.75 77.75 77.75 77.75 77.75 77.75 777	DFS 1	学理身 ************************************		100 11 June 11	Marry All Galo State State State State State Marries M	10 cont 10 con	0000 200 200 200 200 200 200 200 200 20	Sport 211 Sport Jobs December 11 Socialization (1) Socialization Construction Const	The Depu Scope The Depu Scope The Deput Scope The Depu	locrost looks effGAD HA 1	1 59	b) H	Base 1	管理	784		738	NO P M Otary	83 X	50000 ⁴ 05	81

图 6 建立的分布式数据库集群和 Spark 计算集群环境



(a) PSD 计算结果

★##:::::::::::::::::::::::::::::::::::	of 10.5.190.1 - root@storage-master1/hom	/hadoop/hbase-1.1.5/bir	n - Xsh <mark>ell</mark> 6 (Free	9 <u>2</u> 93	o ×
Control Contro Control Control Control Control Control Control Control Control C	文件(f) 编辑(D) 自石(V) 上具(T) 选项卡(B) 岁L	W) HILLIN			
Ach//doct ^{MMMM} 0910.5.190.1.23 TotOStorage:mster1/ht.	二日····································	• 🥑 🗊 😂 🔒 🖮	P		
BetShildelog, Amberthink,	ssh.//root*****@10.5.190.1:22				
I rootBatorage materi/n. # 2 rootBatorage materi/n. # haser incini):065:db. count '%cismic pad', INTEWAL >> 1006080. 1006080. 1006080. current icani: 1006060. normal: 2000080. root All 151.007.01.20.01.20.12.10. 1006080. current icani: 2000080. root All 151.007.01.20.12.10.21.10.10.10.10.10.10.10.10.10.10.10.10.10	▶ 要求加出的交话、点击左侧的前头接闭。				
bbsc[main]:065(36: count 'scient[ped', INTERVAL -> 1000000 Current count: 2000000, row: AL[S0], NU; 2017, 30, 09, 16 Current count: 2000000, row: AL[S0], NU; 2017, 30, 20, 22, 17 Current count: 2000000, row: DL WF 102, 2017, 30, 20, 20; Current count: 2000000, row: DL WF 102, 2017, 30, 20, 20; Current count: 5000000, row: COLJ DH V, 2017, 2017, 201, 20, 12, 19 Current count: 5000000, row: FJ FD 00; 601; 2018, 30, 01, 30, 90 Current count: 5000000, row: FJ FD 00; 601; 2018, 30, 10, 90 Current count: 5000000, row: FJ FD 00; 601; 2018, 30, 10, 90 Current count: 5000000, row: FJ VTMP, 502, 2018, 30, 32, 17 Current count: 5000000, row: GD VDI BH V, 2014, 30, 35, 87 Current count: 10000000, row: GD VDI BH V, 2014, 30, 35, 87 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 326, 97 Current count: 10000000, row: GD VDI BH V, 2028, 30, 30 Current count: 10000000, row: GD VDI BH V, 2028, 30, 30 Current count: 10000000, row: GD VDI BH V, 2017, 30, 75, 16 Current count: 10000000, row: GD VDI BH V, 2017, 30, 75, 16 Current count: 20000000, row: GD VDI BH V, 2017, 30, 75, 16 Current count: 20000000, row: H, MAP_BH2, 2018, 30, 76, 76, 16 Current count: 20000000, row: H, MAP_BH2, 2018, 30, 76, 76, 18 Current count: 20000000, row: H, MAP_BH2, 2018, 30, 76, 76, 16 Current count: 20000000, row: H, MAP_BH2, 2018, 30, 75, 16 Current co	●1 root@storage-master1:/h ● 2 root8	torage-master1:-			
	Nbase(nain):800:86 count 'scientic ps Current count: 1000630, ros: All_Si Current count: 2000630, ros: All_Si Current count: 2000630, ros: Cl_CL Current count: 2000680, ros: Cl_CL Current count: 2000680, ros: Cl_CL Current count: 2000680, ros: Cl_CL Current count: 2000680, ros: Cl_CL Current count: 1000680, ros: Cl_CL Current count: 10006800, ros: Cl_CL Current	$\begin{array}{c} 1011 \mbox{Field} \rightarrow 0.0000 \mbox{Field} \mbox{Field} \rightarrow 0.0000 \mbox{Field} \mbox{Field} \rightarrow 0.0000 \mbox{Field} \rightarrow 0.00000 \mbox{Field} \rightarrow 0.00000 \mbox{Field} \rightarrow 0.00000 \mbox{Field} \rightarrow 0.00000 \mbox{Field} \rightarrow 0.000000 \mbox{Field} \rightarrow 0.000000000\mbox{Field} \rightarrow 0.0000000000000000000000000000000000$			

(b) HBase 存储数据表

图 7 本文设计的 HBase 存储的 PSD 计算结果和 HBase 存储数据表



图 8 集群处理不同数量级计算任务的时间分布

参考文献

- Peterson JR. Observations and modeling of seismic background noise. U. S. Geological Survey, 1993.
- 2 McNamara DE, Buland RP. Ambient noise levels in the continental United States. Bulletin of the Seismological Society of America, 2004, 94(4): 1517–1527. [doi: 10.1785/ 012003001]
- 3 谢江涛,林丽萍,谌亮,等.地震台站台基噪声功率谱概率 密度函数 Matlab 实现.地震地磁观测与研究,2018,39(2): 84-89. [doi: 10.3969/j.issn.1003-3246.2018.02.012]
- 4 Petrov M, Butakov N, Nasonov D, et al. Adaptive performance model for dynamic scaling Apache Spark

Software Technique•Algorithm 软件技术•算法 131

Streaming. Procedia Computer Science, 2018, 136: 109–117. [doi: 10.1016/j.procs.2018.08.243]

- 5 Cattaneo G, Giancarlo R, Petrillo UF, *et al.* MapReduce in computational biology via hadoop and spark. Encyclopedia of Bioinformatics and Computational Biology, 2019, 1: 221–229.
- 6 Karim L, Boulmakoul A, Mandar M, et al. A new pedestrians' intuitionistic fuzzy risk exposure indicator and big data trajectories analytics on Spark-Hadoop ecosystem. Procedia Computer Science, 2020, 170: 137–144. [doi: 10. 1016/j.procs.2020.03.018]
- 7 Tian Y, Liu C, Yan HH. Accelerate large-scale seismic data Kirchhoff time migration in spark. Proceedings of 2018 4th International Conference on Information Management. Oxford, UK. 2018. 41–45.
- 8 Rizvandi NB, Boloori AJ, Kamyabpour N, *et al.* MapReduce implementation of Prestack Kirchhoff Time Migration (PKTM) on seismic data. Proceedings of the 2011 12th

International Conference on Parallel and Distributed Computing. Gwangju, Republic of Korea. 2012. 86–91.

- 9 Popov SE, Zamaraev RY. A fast algorithm for classifying seismic events using distributed computations in Apache Spark framework. Programming and Computer Software, 2020, 46(1): 35–48. [doi: 10.1134/S0361768820010053]
- 10 Chintapalli S, Dagit D, Evans B, et al. Benchmarking streaming computation engines: Storm, Flink and Spark streaming. Proceedings of 2016 IEEE International Parallel and Distributed Processing Symposium Workshops. Chicago, IL, USA. 2016. 1789–1792.
- 11 胡俊, 胡贤德, 程家兴. 基于 Spark 的大数据混合计算模型. 计算机系统应用, 2015, 24(4): 214-218. [doi: 10.3969/j.issn. 1003-3254.2015.04.038]
- 12 郭凯,黄金刚,彭克银,等.大数据技术在海量测震数据中的研究应用. 地震研究, 2017, 40(2): 317-323. [doi: 10.3969/j. issn.1000-0666.2017.02.020]

