

面向装备试验数据的融合存储技术研究与应用

曹芳芳, 任慧敏, 上官子粮, 丁派克

(北京航天自动控制研究所, 北京 100076)

✉ caoff@126.com; renhm@12.calt.casc; sgzl2003@163.com; dingpk@12.calt.casc



摘要:针对海量装备试验数据结构复杂、分散存储及检索访问等问题,开展装备试验数据融合存储技术研究。提出了数据源插件技术、分布式采集处理技术、海量数据文件索引技术,实现多源异构试验数据的采集整理、校核验证、入库和多维信息查询,以及图像、视频、文档类数据文件的检入、检出等功能,并保证系统稳定、高效运行,确保数据应用业务可扩展、可发展。该装备试验数据融合存储技术已应用到某装备系统中,为系统试验数据的稳定存储和应用提供了基础支撑,也为多源异构试验数据的融合存储和管理提供了一种解决方案。

关键词:融合存储;试验数据;多源异构数据

中图分类号:TP319 **文献标志码:**A

Research and Application of Fusion Storage Technology for Equipment Test Data

CAO Fangfang, REN Huimin, SHANGGUAN Ziliang, DING Paik

(Beijing Aerospace Automatic Control Institute, Beijing 100076, China)

✉ caoff@126.com; renhm@12.calt.casc; sgzl2003@163.com; dingpk@12.calt.casc

Abstract: Aiming at the problems of complex structure, scattered storage, retrieval and access of massive equipment test data, this paper proposes to conduct research on equipment test data fusion storage technology. Data source plug-in technology, distributed collection and processing technology, and mass data file index technology are proposed to realize functions of collection and collation, verification, storage and multidimensional information query of multi-source and heterogeneous test data, as well as the check-in and check-out of image, video, and document data files, etc., which ensures the stable and efficient operation of the system, so as to make sure that data application business can be expanded and developed. The fusion storage technology has been applied to an equipment system to ensure the stable storage and safe availability of system test data, and it provides a solution for the fusion storage and management of multi-source and heterogeneous test data.

Key words: fusion storage; experimental data; multi-source and heterogeneous data

0 引言(Introduction)

随着信息化技术的发展,融合存储技术在各行各业中被大范围应用。数据融合存储技术包括对各种信息源数据的采集、传输、处理、校核及存储。数据融合存储技术主要应用领域有多源影像复合、机器人和智能仪器系统、战场和无人驾驶飞机、图像分析与理解、目标检测与跟踪、自动目标识别等。

近年来,装备信息化、智能化技术发展迅速,航天产品在设计、生产、试验、交付等各个阶段产生了大规模的试验数据,这些数据分散存储在各系统中,种类庞杂、数据量巨大、存储与管

理困难。由于各种智能设备的信息化水平参差不齐,功能各异,彼此之间相互独立,因此存在试验数据来源多样化、数据格式多样化、数据存储位置越来越离散化等特点,导致严重的信息孤岛现象。

由于试验数据缺少统一的标准、规范,所以数据格式和数据内容差异很大,不同业务场景对数据的消费和应用也不同,对数据的存储与管理 and 共享使用带来很多困难。为消除数据割裂局面,解决现有数据积累、存储管理、检索查询、协同复用等问题,试验数据的融合存储与管理已成为当前迫切要解决的问题^[1]。

1 研究思路(Research ideas)

目前,数据融合存储的思路是把不同来源、格式、特点的数据在逻辑上或物理上有机地集中起来,从而提供全面的数据共享。这些数据管理方法通常只针对特定的数据类型和数据应用场景,数据的读取方式和数据服务都比较单一,无法提供统一的数据接口和数据服务标准。对于航天装备系统而言,需要对试验数据进行重新组织和连接,拥有清晰的数据标准接口和结构,在保证数据一致性和安全的前提下,让数据更易获取,最终打破数据应用间的数据孤岛。

装备试验数据从结构上划分为结构化数据、非结构化数据、半结构化数据。从数据类型上划分为二进制数据流、数据库文件、图像、音视频类等数据文件。对于不同的数据类型,构建数据业务宽表,对接不同数据源;通过数据源插件配置管理的方式,实现对多来源数据的智能匹配和接入;针对海量数据的快速处理要求,基于 Spark(Apache Spark 是专为大规模数据处理而设计的快速通用的计算引擎)的优秀资源调度能力结合集群的计算能力实现在多源异构的采集场景下的分布式数据快速处理^[2];针对多类型数据的检索访问问题,通过倒排索引机制实现海量数据文件的快速检索^[3];通过数据分析接口的设计将数据应用与平台进行剥离,对外释放数据服务接口,通过服务注册的模式完成对数据业务的部署和应用,确保业务的可扩展、可发展。试验数据融合存储技术的总体设计思路如图 1 所示。

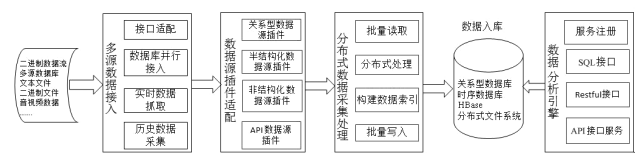


图 1 总体设计思路

Fig. 1 Overall design idea

平台融合多种类型数据库作为多源异构试验数据的存储底座,包括关系型数据库、时序数据库、HBase(Hadoop 数据库)、分布式文件系统,面向不同的数据源智能匹配数据入库方式。将分散的装备实测数据、试验数据、仿真数据、地理环境信息数据等数据信息进行统一的存储、管理和维护。

2 试验数据融合存储技术研究(Research on experimental data fusion storage technology)

对于多来源试验数据的采集接入需求,梳理当前装备存储设备的采集协议和方式,将设备 IP 信息、采集协议、用户名权限等信息以配置文件的方式进行存储和加载,增强代码的复用性。针对各类数据源不同的底层接口,面向用户提供基于流程选择的统一接口,采用多线程方式完成多来源数据的并发采集。对于规格化的试验数据通过适配关系型数据源插件自动提交入库任务,对于非规格化的数据库文件,适配非结构化的数据源插件,提交加载任务。为使用者提供简化、统一的人机交互操作方式,实现一键获取数据、一键下载数据、一键自动加载数据源。

2.1 数据源插件技术研究

针对结构化试验数据,采用关系型数据库存储信息数据、基础信息数据、基础模型数据、仿真试验数据等。针对非结构化数据,采用分布式列数据库和分布式文件系统进行统一的存储^[4-6]。由于装备系统的自主可控要求,试验数据融合存储平台要求兼容国产平台,而国产服务器的计算性能和网络数据传

输速度等比 X86 架构的进口服务器性能差,因此需提供一种高可靠性的分布式数据存储方案应对海量半结构化、非结构化数据的存储需求。本文选择将分布式列数据库 HBase 和 HDFS(Hadoop 分布式文件系统)构成分布式文件存储系统,其中 HBase 是一种高可靠性、高性能、面向列、可伸缩、能实时读写的分布式存储系统,利用 HBase 的高可靠性存储技术可以在普通的 PC Server 上搭建大规模结构化存储集群。

面向上述多种类型的数据库,本文设计了数据源插件技术实现新数据源的注册、发现和适配运行。数据源插件是基于 Java SPI(Service Provider Interface)技术,封装统一的数据源驱动接口,完成新增数据源的开发适配,实现动态新增数据源。数据源插件工作过程示意图如图 2 所示。

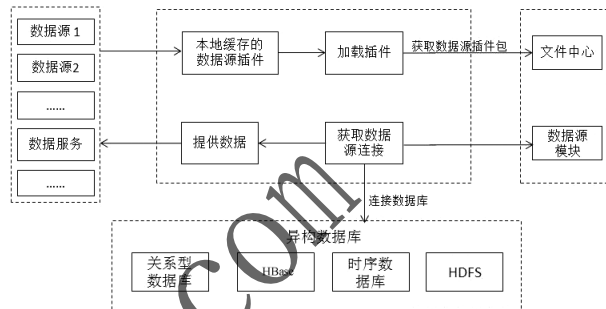


图 2 数据源插件工作过程示意图

Fig. 2 Schematic diagram of working process of data source plug-in

基于控制反转机制,将数据源与数据库驱动包进行隔离,保证新增数据源与已适配的数据源链接方式、驱动包、驱动类名称等兼容。

当多客户端重复请求加载插件时,为避免网络高频占用,采用本地缓存的方式,首次加载插件时先查看服务器本地文件夹下是否有对应的插件,如果没有则通过远程调用请求文件中心的接口,将对应插件下载并缓存到本地。如果本地发现已有对应插件,但为了避免插件的替换热部署导致现场版本不是最新的插件,此时会对比缓存区的插件的 MD5 码和插件数据库中的 MD5 码,保证插件是最新的版本。数据源插件更新流程图如图 3 所示。

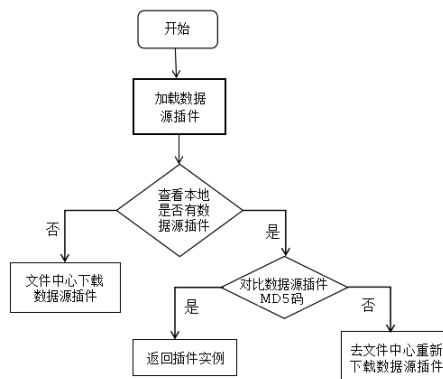


图 3 数据源插件更新流程图

Fig. 3 Flow chart of data source plug-in

面向不同属性的数据源类型分别接入异构数据库,并根据属性配置形成不同的数据索引主题,便于数据的检索。

2.2 分布式采集处理技术

传统的 Java 数据库连接(Java Database Connect, JDBC)数据采集,在读取大数据量时,效率低下,需要将数据全部读取到

内存,可能会出现内存溢出,导致采集失败的情况,严重时会导致程序异常退出。同时,由于它只能通过单节点、单线程的方式进行数据处理,所以对集群的资源利用率偏低,不能充分利用 CPU 资源,已经不能满足多源异构数据源采集的业务场景和性能要求。

Spark 的优秀资源调度能力结合集群的计算能力,更适合多源异构数据源的采集处理场景^[7-8]。本文采用 Spark 技术实现多源异构数据源的采集处理,其采集处理过程如图 4 所示。

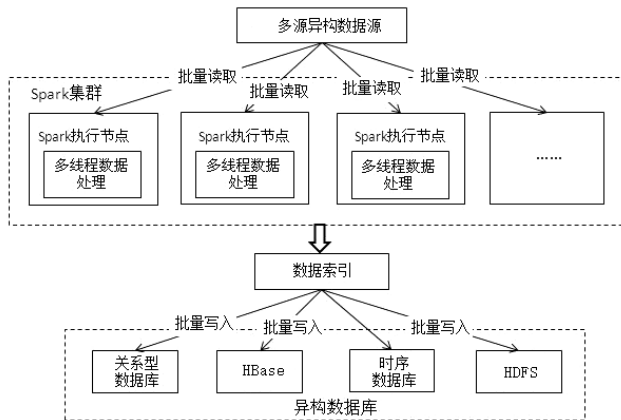


图 4 Spark 采集处理流程图

Fig. 4 Flow chart of Spark collection and processing

利用 Spark 的分布式处理能力,通过多个执行节点批量读取数据和处理任务,有效提高数据的处理能力。Spark 内部会根据数据处理任务,构造有向无环图,并划分成不同阶段,再划分为不同的任务,最后将任务分发到不同的执行节点上执行,实现了大数据量的分而治之。每个执行节点还可以分配多个线程,采用多线程处理数据,充分利用集群的优势将每个节点的 CPU 充分利用起来。

2.3 海量数据文件检索技术研究

常见的 Google、百度等网站的搜索都是根据网页中的关键字生成索引,在搜索时输入关键字,将该关键字匹配到所有网页后返回信息流。传统的数据库全文检索需要扫描整个表,如果数据量大就会出现卡滞、缓慢甚至检索失败的情况,即使建立了索引库,维护成本也很高,对于 insert 和 update 操作都需要重新构建索引^[9]。所以,对于装备非结构化试验数据采用传统的关系型数据库和常规全文搜索方式已无法满足其性能要求。

为了解决装备结构化数据搜索和非结构化数据搜索性能问题,需要采用更适合装备试验数据特点,更加健壮、强大的全文搜索方式。本文引入 Elasticsearch(数据搜索分析引擎)的倒排索引机制加速文件的检索速度。倒排索引是以分词的方式支持快速模糊查询,以分词内存化的方式加速查询速度。适合倒排索引方式的数据特点主要如下:①搜索的数据对象是大量非结构化的文本数据;②文件记录量达到数十万个或数百万个甚至更多;③支持大量基于交互式文本的查询;④需求非常灵活的全文搜索查询;⑤对不同记录类型、非文本数据操作或安全事务处理的需求相对较少的情况。倒排索引的方式与装备海量数据文件的精确和模糊检索场景非常契合。海量数据文件检索原理图如图 5 所示。

当一个查询被触发,所有已知的数据段将按顺序被查询,单词统计模块会对所有段的结果进行聚合,以保证每个词和每个文档的关联都被准确计算,而不需要对数据库中所有的数据

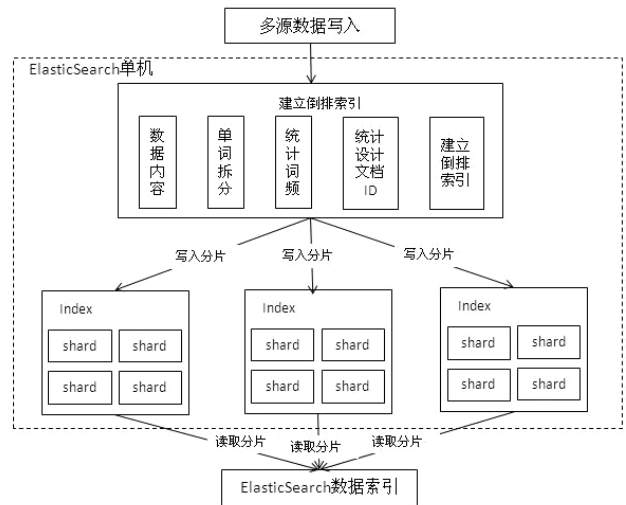


图 5 海量数据文件检索原理图

Fig. 5 Principle diagram of mass data file retrieval

内容进行过滤查询,可有效提升数据的查询速度。当有新的数据写入索引时,单词统计模块会根据业务进行词划分,不同的关键词对应不同的数据 ID,并同步增加词频统计值。采用这种倒排索引可有效降低新文档添加到索引库的难度。

2.4 数据分析接口设计研究

数据分析接口是响应数据接口指令并提供数据业务的逻辑处理过程。通过数据接口封装保证了数据平台与业务之间的剥离,便于业务的拓展和开发。在分析用户的使用场景后,本系统对外提供的数据库接口主要有 JDBC 接口、Restful 接口、SDK 接口。其中,JDBC 接口提供 SQL 语句化的数据检索处理方式;Restful 接口是基于 Http 协议的标准化数据分析接口;SDK 接口是为上层应用服务提供 API 函数接口和基本的数值计算接口。

2.4.1 JDBC 接口的设计与实现

解析用户的数据检索指令,将其标准化为类 SQL 查询语句。通过标准化的类 SQL 查询语句从索引数据库中抽取对应的数据信息,将索引信息与数据内容进行对应,完成数据的提取和反馈。SQL 数据流处理过程示意图如图 6 所示。

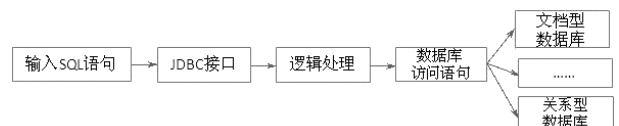


图 6 SQL 数据流处理过程

Fig. 6 The diagram of the SQL data flow process

2.4.2 Restful 接口的设计与实现

基于 Http 协议封装为标准化的 Restful 数据分析接口。标准化的数据分析接口是专门用于结构化数据计算的程序语言,将平台的分析功能通过自然语言的方式进行抽象,标准化为数据分析接口。使平台提供的应用服务具备可扩展性,即不会因为未来应用需求变更而更改底层数据的读取方式,使平台具有高内聚、低耦合的特性。本文设计的数据分析接口示意图如图 7 所示。

图 7 中,“date”是指数据源,即从数据库里读取的数据。“Index1, Index2, ...”是在数据源读取的数据中选择 Index1、Index2 等几列。“Head 100”是指读前 100 条数据。

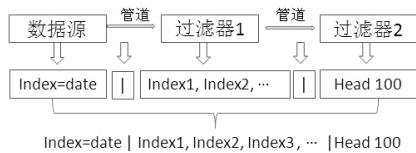


图 7 数据分析接口示意图

Fig. 7 Schematic diagram of data analysis interface

2.4.3 SDK 接口的设计与实现

该接口是为用户提供 Java 的 API 接口,也为上层数据应用提供了高效的读写接口,包括大文件的存取操作和基础数值函数。主要封装的 API 函数接口有建库函数、删库函数、数据表的单条插入、删除和批处理函数、写入文件函数和读取文件函数等。

3 试验数据存储平台的应用 (Application of experimental data storage platform)

基于上述技术设计实现了试验数据存储平台,将该平台应用到航天装备中,针对该型号在系统试验过程中产生的数据流、音视频数据和数据库文件等信息,实现了采集处理、数据入库,以及数据文件的检入、检出等功能。

3.1 平台吞吐能力分析

从平台的数据吞吐能力、线程并发处理能力及数据检索能力几个方面进行对比与分析。Insert 吞吐能力是指每秒可以写多少条记录,Read 吞吐能力是指每秒可以读多少条记录,Scan 吞吐能力是指每秒可以查看多少条记录,延迟时间是指对应操作需要的时长。选择三台 8 核、32 GB 性能的机器搭建集群测试融合管理平台的数据吞吐能力,数据统一选择 UDP (User Datagram Protocol, 用户数据报协议) 实时数据流,每条数据包含 15 个字段,每个字段的字段长度为 32 字节,测试情况如表 1 所示。

表 1 平台数据吞吐能力测试结果

Tab.1 Test results of platform data throughput capacity

数据量/条	Insert		Read		Scan	
	吞吐能力/ (条·s ⁻¹)	延迟时 间/ms	吞吐能力/ (条·s ⁻¹)	延迟时 间/ms	吞吐能力/ (条·s ⁻¹)	延迟时 间/ms
1 亿	146 701	13.240	61 776	3.03	38 799	51.21
2 亿	134 531	13.450	60 454	3.01	35 655	50.81
3 亿	132 492	13.440	589 531	3.00	33 034	50.02
4 亿	131 233	13.450	585 421	3.02	33 560	50.21

从表 1 中的测试结果可以看出,数据量的增大并不影响数据的吞吐能力和读取时间。采用传统的关系型数据库时,写入 1 亿条相同的数据大概需要 2 min,数据读取速度很慢。通过表 1 中的测试结果可以得出本文平台的数据吞吐能力更强、更可靠。

3.2 多数据源接入效果展示效果

选择将不同类型数据库接入试验数据管理平台。数据源插件管理效果与试验数据查询效果如图 8 和图 9 所示。

4 结论 (Conclusion)

聚焦装备复杂试验数据结构、存储冗余管理、访问速度提升等几个方面,本文通过数据源插件技术、分布式采集处理技术、海量数据文件索引技术和数据分析接口的设计研究,完成了试验数据管理平台的设计应用,并在某装备型号上进行了应用验证。通过对比分析得出平台运行稳定、可靠且具备较强的数据吞吐能力的结论,实现了对海量试验数据的快速采集和稳定存储,为装备研制、试验鉴定等提供了平台支撑。该融合存储技

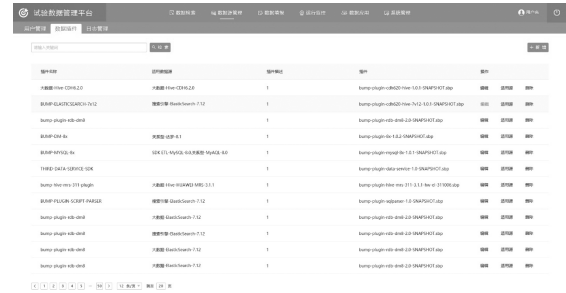


图 8 数据源插件管理效果图

Fig. 8 The diagram of data source plug-in management

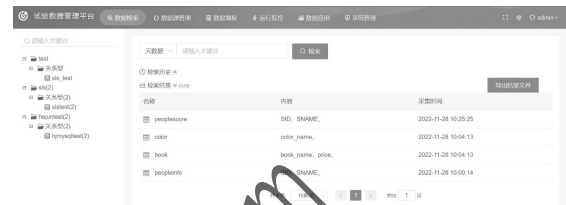


图 9 试验数据查询效果图

Fig. 9 Test data query rendering

术可扩展应用于运载火箭等其他航天产品的数据处理领域,也可用于装备研制及人体健康状态评估分析,为模型校验、目标识别、任务规划提供实测数据,促进装备试验研究的成果应用。

参考文献 (References)

- [1] 丁世来,陈克澎,葛智君,等. 多源试验数据重构与融合存储技术研究[J]. 电子产品可靠性与环境试验,2022,40(1):11-15.
- [2] 金国栋,卞昊穹,陈跃国,等. HDFS 存储和优化技术研究综述[J]. 软件学报,2020,31(1):137-161.
- [3] 王磊,陈磊,张明儒,等. 面向数据库查询的非结构化数据融合存储系统[J]. 电子设计工程,2022,30(24):148-152.
- [4] 刘伟,张留山. 数据库在数据融合系统中的应用[J]. 舰船电子对抗,2004,27(4):30-34.
- [5] 薛尧予,王建林,赵利强. 分布式过程实时数据集成方法及其实现[J]. 计算机工程,2010,36(3):55-57.
- [6] 孙留倩. 基于本体的多源数据融合方法研究与应用[D]. 哈尔滨:哈尔滨工业大学,2021.
- [7] 夏立斌,刘晓宇,孙玮,等. Spark 任务间消息传递方法研究[J]. 计算机工程与应用,2022,58(21):91-97.
- [8] JIANG W, XU L G, HU H B, et al. Improvement design for distributed real-time stream processing systems[J]. Journal of Electronic Science and Technology, 2019,17(1):3-12.
- [9] 马健,张太红,陈燕红. 中文搜索引擎分块倒排索引存储模式[J]. 计算机应用,2013,33(7):2031-2036.

作者简介:

曹芳芳(1984-),女,硕士,高级工程师。研究领域:云计算,大数据技术应用。

任慧敏(1980-),女,硕士,高级工程师。研究领域:嵌入式软件开发。

上官子粮(1985-),男,硕士,高级工程师。研究领域:大数据技术应用。

丁派克(1992-),男,硕士,工程师。研究领域:大数据技术应用。