Computer Science and Application 计算机科学与应用, 2022, 12(10), 2338-2347
Published Online October 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1210239

Hans汉斯

基于Hadoop作战试验异构数据平台的数据治理研究

周彬彬，王　磊，孙志成，秦易夫，刘小朋

63861部队，吉林　白城

收稿日期：2022年9月20日；录用日期：2022年10月18日；发布日期：2022年10月27日

摘　要

目前构建作战试验异构数据平台用以存储作战试验相关数据，然而装备作战试验的各项数据相对分散杂乱，缺乏对试验数据的有效管理与应用研究。针对平台建设使用过程中出现信息孤岛、数据质量低下、数据来源重复繁杂等不同程度的数据问题，进行数据治理的方法研究。本文分析了作战试验数据特点及治理需求，提出一套基于Hadoop作战试验异构数据平台的数据治理框架，构建了面向作战试验异构数据的HAO治理模型。有效解决大数据平台建设及使用、数据分析及可视化探索等大规模数据使用场景中可能遇到的数据问题，实现了大数据平台数据规范统一管理，极大地提高了数据质量，实现更加高效地发挥和挖掘作战试验数据的价值。

关键词

作战试验，数据治理，数据管理，数据质量

Research on Data Governance Based on the Hadoop of Combat Test Heterogeneous Data Platform

Binbin Zhou, Lei Wang, Zhicheng Sun, Yifu Qin, Xiaopeng Liu

63861 Troop, Baicheng Jilin

Received: Sep. 20th, 2022; accepted: Oct. 18th, 2022; published: Oct. 27th, 2022

Abstract

The platform of heterogeneous combat test data is built to store the data related to combat test. The data of equipment combat test is relatively scattered and disorderly, and the effective man-

文章引用: 周彬彬, 王磊, 孙志成, 秦易夫, 刘小朋. 基于 Hadoop 作战试验异构数据平台的数据治理研究[J]. 计算机科学与应用, 2022, 12(10): 2338-2347. DOI: 10.12677/csa.2022.1210239

agement of test data is lacking. For the data problems of platform construction such as information islands, low data quality, and duplication of data sources, it studies the methods of data governance. This paper analyzes the characteristics and governance requirements of combat test data, proposes a set of data governance framework based on Hadoop combat test heterogeneous data platform, and constructs a HAO governance model. It effectively solves data problems that may be encountered in large-scale data usage scenarios such as construction and use of big data, data analysis and visual exploration, greatly improves data quality, realizes the unified management of data specifications of data platforms, and realizes more efficient play and mining of the value of combat test data.

Keywords

Combat Tests, Data Governance, Data Management, Data Quality

1. 引言

随着作战试验任务的频繁开展，加之被试装备类型各异、种类繁多，测试设备种类多样、手段复杂，随之产生的作战试验数据与日俱增，呈指数型爆炸增长[1]。随着军队信息化建设的逐步发展，数字化高清视频、音频、图像等检测设备在装备作战试验中得到普遍应用[2]，大量的半结构化数据和非结构化数据占据整个作战试验数据的主体部分[3]。多年来，海量的试验数据以纸质文档、电子文档、数据表格、音频视频等形式保存。基于 Hadoop 构建的作战试验异构数据平台，存储作战试验过程中各测试设备采集的数据和被试装备数据，为试验鉴定人员评估分析提供数据基础，实现了作战试验数据的储存管理一体化，但同时也存在许多问题。

作战试验异构数据平台可以将积累的海量试验数据进行有效存储，但是，异构数据存储系统的使用，也导致各类装备试验数据类型和含义驳杂，整个大数据存储平台显得混乱不堪。伴随着系统的长期运转，信息孤岛、数据质量低下、数据来源重复繁杂等不同程度的数据问题日益凸显。大规模试验数据的不断涌入、数据基数的逐渐增大，使得数据预处理的工作量和复杂程度呈指数增长，也导致后续的作战试验综合评估和分析挖掘工作难以有效开展。这些工作仅靠数据管理和试验评估人员是远远不够的，不仅需要花费大量的时间成本，也可能因为缺少规范和流程化的操作带来新的数据问题，进而陷入不停地进行数据处理操作和解决突发问题的恶性循环中。

因此，本文针对作战试验数据特点，分析数据治理需求，提出了提出一套基于 Hadoop 作战试验异构数据平台的数据治理框架，构建了作战试验异构数据 HAO 治理模型，实现了大数据平台数据规范统一管理，有效解决试验异构数据平台建设及使用、数据分析及可视化探索等大规模数据使用场景中可能遇到的数据问题，极大的提高了数据质量，更加高效地发挥和实现作战试验数据的价值，也为武器装备作战试验数据治理方法提供一些思路。

2. 作战试验背景下的数据治理需求分析

2.1. 作战试验数据特性分析

试验鉴定的发展日趋完善以及对常规武器装备的作战试验的理论和技术深入研究，多地域、多种战

场环境下多个科目的作战试验必将是一项长期而艰巨的任务。随着常规武器作战试验的开展，试验数据存量呈指数增长。建设装备作战数据中心能够统一规范管理和有效应用这些宝贵的数据资源，充分发挥试验数据资源对于后续对于指导后续试验的设计和开展、装备的改进升级及后续实际应用、作战和演习演训筹划等活动，具有重要辅助作用[4]。目前常规武器装备作战试验数据从不同的角度来看，呈现出以下特性：

从数据规模来看，数据具有海量特性。只要进行试验时佩戴装备接收信号，就会源源不断产生数据，随着时间推移自然会产生海量的数据。在装备试验、实战演训等活动时候，会持续不断产生数据，随着时间推移自然会快速产生并积累海量的数据。

从数据结构和类型来看，数据具有多源异构特性，且数据类型和格式逐渐增多。试验数据往往来自不同演训科目、不同场景以及不同的装备，具有完全不同的数据结构；数据类型包括位置数据、姿态数据、温湿度数据、气象数据、天时天候数据、满意度、靶标数据、对抗数据、总线数据、电磁环境数据和弹道坐标数据等。

从数据描述角度来看，数据具有低维度特性。传感设备的原始信号数据实际上是一种一维时序数据，每一个单点值不具有实际的描述意义，只有通过统计分析产生具有实际描述意义的特征。

从数据价值来看，孤立的数据具有低价值性。存量庞大、分布广泛和主要信息的低密度性，是作战试验数据的一个显著特点。只有当从大量数据中综合分析，才能体现出大数据的价值性。

2.2. 作战试验数据治理需求分析

数据治理委员会 IBM (IBM DG Council)对于数据治理的定义是，数据治理是一种质量控制规程，用于在管理、使用、改进和保护组织信息的过程中添加新的严谨性和纪律性[5]。由于装备作战试验在是在逼真复杂的战场环境下开展的，同时对于装备的多个特性开展考核，负责不同分工的试验单位采集的数据管理权限尚不明确，数据大多分散在各自独立的数据库中，没有形成系统完整的数据体系。查询已有的数据时，需要多渠道跨系统才可以将数据收集整合，针对试验数据内容的综合查询能力较差，难以实现数据的全面对比。由以上的分析可以看出，在作战试验背景下，试验数据治理面临着严峻的挑战，其主要分为以下三个方面：

1) 数据规模宏大，质量问题严重。

随着武器装备的作战试验进行，各种配套测试设备的同步展开，与之产生的数据量极为庞大，往往一项作战试验能够产生几百 TB 级的数据。这些数据既包括被测装备本身传感器产生的装备过程状态数据，也包含测试项目指标所需的各类测试设备采集的实战考核数据。这些成百上千的设备投入作战试验中，使得试验数据呈海量化、多元化增长。由于测试设备具备独立的采集系统和存储空间，导致作战试验数据中存在大量的冗余数据分布于各项测试设备系统中，且无法同步更新，当冗余数据内容不一致时更是难以决策实际的数据项内容，进而带来了许多严重的数据质量问题。

2) 数据来源广泛，数据结构混杂。

作战试验往往在复杂场景下针对不同装备进行多项演训科目的实战化考核，具有完全不同的数据结构。面向多环境、多场景、不同试验装备、不同人员、不同试验科目的多源异构数据，导致试验数据结构混杂，结构化、半结构化和非结构化数据相互杂糅，具体表现为环境条件数据、GPS 数据、指标信息采集数据、装备过程状态数据、音视频数据等。

3) 数据标准繁杂，共享集成困难。

目前我军作战试验开展仍处于起步发展阶段，数据采集软件各式各样，试验过程中不同的采集软件对于采集的数据格式标准尚不统一，部分配套的数据信息采集软件在处理数据上相对分散，处理数据的

完整性得不到根本保障。缺少配套的数据标准规范，大量数据格式和定义不统一，极少的数据标准也没有得到贯彻执行，数据建设和维护使用无章可循，缺乏统一的技术标准和技术手段，数据资源的集成较为困难，无法实现交互共享。

3. 基于 Hadoop 作战试验数据平台的数据治理框架

传统的数据治理框架大多基于元数据和数据仓库方式存储和管理[6]，然而试验数据指数型增长，异构数据融合的需求增多，现有存储设备的新增和管理维护成本激增，给试验数据分析管理带来了巨大的压力。通过对作战试验背景下的数据治理需求分析可以看出，数据质量管理、数据清洗、数据标准化、数据交换和数据集成共享等过程中存在的问题是平台数据治理时面临的首要挑战。传统的数据存储方式难以满足当前作战试验数据存储和管理需求，并且一旦发生技术故障，数据丢失的风险高，造成无法挽回的损失。每一场作战试验的数据都极其宝贵且重要，一旦丢失某一项试验数据，那么就无法完成对试验装备的作战体系化考核。

因此，本文基于 Hadoop 作战试验数据平台，从装备试验数据的分析挖掘、可视化应用以及作战目标的角度提出数据分层治理框架，该框架自上而下包括治理目标、数据标准化体系、治理域、实施，具有层次化、松耦合、面向开放共享的特色架构。治理框架如图 1 所示。

Figure 1. Operational test data governance framework
图 1. 作战试验数据治理框架

3.1. 目标

吴信东[7]提出大数据治理的目标包括战略一致、风险可控、运营合规和价值创造四个部分，而这对

作战试验数据治理的实施过程同样具有指导作用，所以我们框架的顶层 4 个治理目标是相一致的。其中，战略一致主要要求在治理域中制定战略时要结合试验方案、总体目标和作战使命任务，与数据业务需要相适应。试验数据既是作战试验产品，同样也是部队建设的重要资产。风险可控要求在治理试验数据时，在整个的数据生命周期管理过程中要严格遵守规范，在加上平台试验数据的合规运营，可以提高数据质量，合理控制治理风险，实现作战试验数据价值创造。分析挖掘试验数据背后隐含的重要价值，是数据治理最核心、最关键的目标，同样也是作战使命任务的要求。

3.2. 数据标准化体系

数据标准化体系提供从数据运行的基础环境到数据的采集、装载、转换、处理、管理、应用的全程的数据标准化保障，进而实现数据的统一，提升数据治理效率。数据规则配置模块主要负责管理用户自定义规则，负责把用户的输入规则抽象化，并存储于数据库中，提供规则的增、删、改、查功能，同时负责用户输入的规则校验工作，确保规则合法。

3.3. 治理域

治理域描述了试验数据分析管理部门在开展作战数据治理时主要的治理核心对象，主要包括战略、组织、作战试验数据架构、数据管理、数据生命周期管理、数据验证、数据安全与合规。数据管理具备 ETL 和数据基础管理功能，能够将现有多源数据经 ETL 和数据治理后接入其他应用和从其他应用读取数据的能力；对数据具有清洗、标准化、去重合并处理、资产管理等能力，具备多种类型数据源抽取和加载能力，具备数据清洗、数据校验和转换功能，具备规则管理、状态管理和日志记录功能。数据生命周期管理数据的持久化结构和内存结构，管理数据内容、数据元信息、数据结构或内容预览视图、数据逻辑关系，并对数据的接入、编辑、销毁等可用周期进行管理。数据验证主要负责把规则库中的元数据实例化，生成实例后通过数据连接模块访问数据仓库系统，对数据仓库系统数据进行及时性、完整性和准确性校验，并把运行结果保存到数据库当中。数据安全与合规是平台数据治理的重点，数据安全管理工作的重点是根据数据级别的不同和特点进行安全级别的管理。其方法是对内外数据进行汇聚、挖掘形成不同价值的数据资源，并对外提供各类数据访问服务，分级别设置数据资源访问权限及安全管控策略。

3.4. 目标

作战试验数据治理的实施主要包括数据源、数据接入、数据存储、数据集成共享、数据可视化和数据分析应用 6 个部分。作战试验数据的治理工作要牢牢地与作战任务目标相结合，治理的目的是为了使数据更好的服务应用，这就要求试验数据从采集到汇总，再到接入平台中存储集成和分析等过程有着标准的规划与实施。这 6 个部分包含了平台数据流向的全过程，针对这 6 各方面进行的数据治理，使得在数据生命周期的各个阶段针对不同的问题有着不同的治理侧重点，对试验数据的持续发展和再利用具有重大的意义。

4. 面向作战试验数据的 HAO 治理模型

本文通过改进 HAO 治理模型[5]的分层结构化，支撑分布式作战异构数据平台的数据治理。根据作战试验具体展开实际和平台数据汇集，将传统的 HAO 治理模型的数据接入模块分为数据采集模块和数据存储模块两部分，再加上数据管理的治理模块和数据分析挖掘的数据服务模块，共同构成“采－存－管－用”4 个周期化的层次模型结构。其具体治理模型架构如图 2 所示：

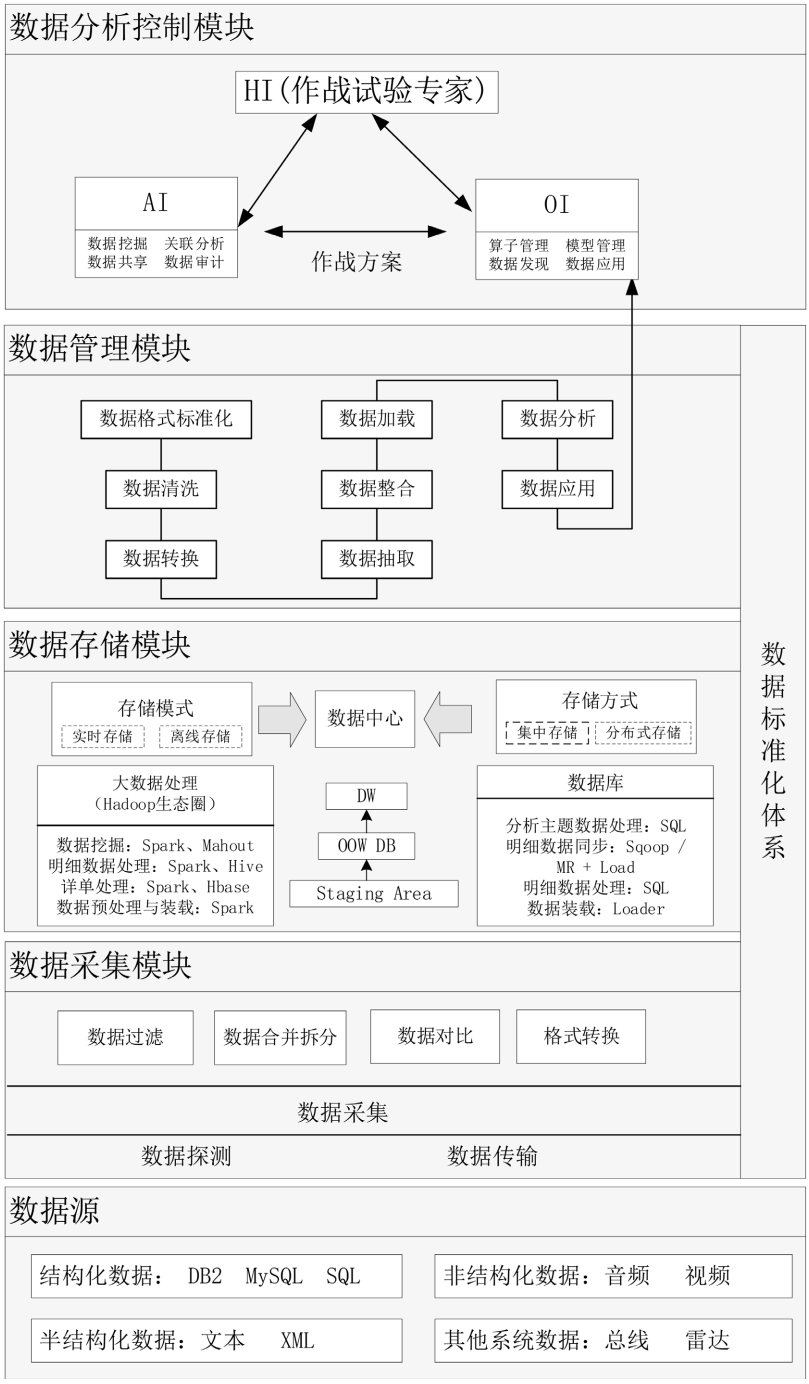

Figure 2. HAO Governance model for operational test data
图 2. 面向作战试验数据的 HAO 治理模型

该模型针对作战试验异构数据平台的业务需求，为提供基础数据服务能力保障，满足对于数据中心功能和性能的要求，主要实现以下目标：

1) 建立完善的数据资源体系，基于国家标准、军队标准和行业标准，兼顾标准之间的兼容性，建设形成一套服务自身实际的管理信息化标准体系，保障数据资源的完整性，同时可以按照用户的个性化应用需求，实现数据的可理解、可比较和可共享；

2) 建立有效的采集、清洗、存储和管理的方式手段，通过可视化图形操作界面，基于元数据的配置，来实现数据探测、数据预处理、数据采集、数据 ETL，为其他系统的数据服务提供数据基础。支持处理逻辑配置，和处理流程设置与管理，使数据处理过程达到可查看、可监督、可调控的全程透明化，提高数据管理工作效率；

3) 建立可视化的元数据定义和关联、映射的管理服务功能，实现简单直接的进行数据模型和数据应用的管理，实现对数据之间的关联、映射关系的管理，方便用户从多个维度的关注、了解和使用数据资源；

4) 提供数据开放共享功能，用户可以通过终端实现灵活便捷的数据检索功能，可以按照权限分配，进行数据服务的创建、审核和发布，使用户可以按需获取数据资源；利用集中的数据共享实现对上层应用的数据共享和支撑，实现海量数据的应用；

5) 建设数据评估框架平台，建设、开发、集成通用数据分析工具，以向导的方式整合各单位设计的新算法、新工具，并以服务的形式提供用户使用，实现对数据资源的关联、重组、分析和挖掘，最大限度地发挥数据资源的建设效益。

4.1. 数据采集模块

“采”——数据采集和数据接入模块，实现数据采集和接入，包含数据判读、数据质量检查以及初步的数据格式转换等，完成数据的初步集中工作，完成试验数据平台接入各类数据源，针对不同数据源的不同特征，应用不同的采集技术。数据源连接模块主要负责与各个数据库进行连接适配，支持常见的关系型数据库如 Mysql、Oracle、SQL Server 等，同时支持大数据相关数据库如：Hive、Impala 等，以及 NoSQL 数据库如 Hbase 等。同时负责数据库的连接池管理，提高访问效率。功能架构图如图 3 所示。

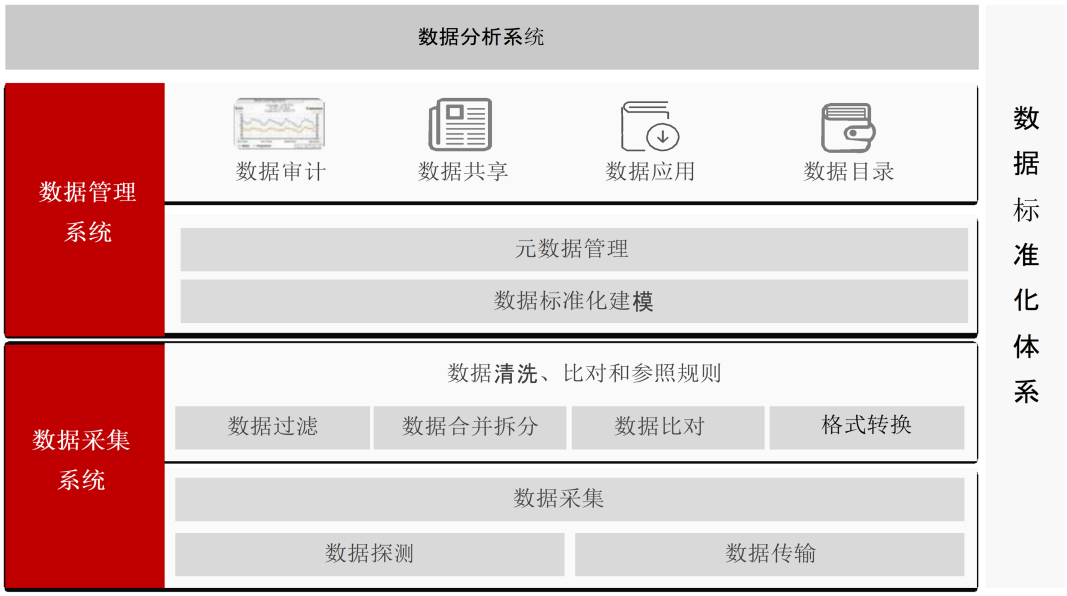

Figure 3. Functional architecture diagram of test big data comprehensive management platform
图 3. 试验大数据综合管理平台功能架构图

数据采集模块支持多种异构的数据存储方式，可提供多样的异构数据采集、数据存储排列组合方案：

1) 支持从不同种类的数据存储之间进行采集、加载，例如：从文件目录或者定制的数据源采集数据，加载到文件或数据库中；

2) 支持同种类数据存储之间细分类别之间进行采集加载，例如：从 NoSQL 数据库采集一张表的数据，加载到 SQL 数据库的指定一张表中；

3) 支持数据按重要性程度加载到磁盘阵列或普通硬盘；

4) 支持数据按照备份策略采用集中存储或分布式存储。

4.2. 数据存储模块

“存”——数据存储模块，实现数据采集存储和数据汇集，同时根据数据的类别和重要性程度，采用合适的存储模式(实时存储或离线存储)，采取适当的存储方式(集中存储或分布式存储)实现对数据的分类保存，并同时完成对数据的安全备份。如图 4 所示。

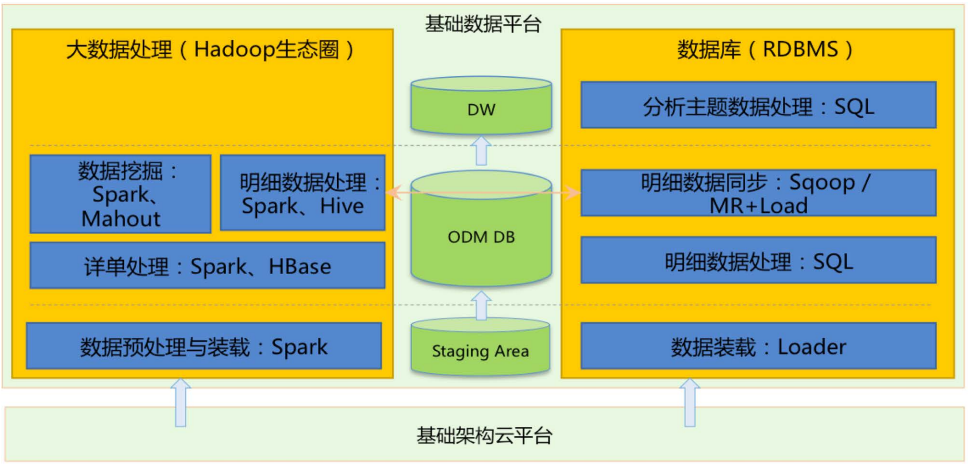

Figure 4. Data storage software architecture diagram
图 4. 数据存储软件架构图

试验数据源种类较多，不同种类的数据在存储后的访问方式差别较大，因此存储模块采用分布式文件系统和关系型数据库综合存储的方式。使用分布式文件存储非结构化数据、大文件(包括图像文件)的数据，这类数据只需要整块连续查询或读取，不需要复杂查询。使用关系型数据库存储需要频繁复杂查询的结构化数据。虽然数据存储在不同的介质中，但是通过上述的存储方式，建立 Hadoop 环境下的基础数据平台，可以实现数据的一致查询和分析。

4.3. 数据管理模块

“管”——数据管理模块，实现在整个过程中的融入式的数据管理，对数据的分类、组织、模型、规则、关系等进行统一管理，实现全程的元数据驱动；实现对于数据的浏览、查询和初步应用；同时采用数据服务的方式对外提供数据共享，将共享数据封装为服务，实现对上层应用和第三方系统的数据共享。

数据管理模块采用分布式架构图体系，按照软件架构设计分层理论将系统划分为服务层、执行层、应用层等层次。如图 5 所示。

4.4. 数据分析挖掘模块

“用”——数据分析挖掘模块，依托数据挖掘技术，设计各种主题(比如历次试验效果趋势分析、作战单元贡献率趋势、故障率趋势分析等)，对常规武器装备作战数据分析挖掘其内在价值。

数据分析挖掘模块提供多种数据分析与计算算法，支持自定义多种计算算法，实现实验算法节点的增删改查、配置、运行、停止、复制以及查看数据等。基于机器学习、深度学习和自然语言处理等先进

技术，对现有海量数据进行分析挖掘，通过数据建模、模型评估、模型发布、模型导入等业务流程，提升大数据应用能力。在现有海量数据基础上，深度发现隐藏的数据价值，对业务提供智能化数据服务。

分析挖掘模块以数据建模为核心功能，提供系统管理、数据管理和可视化、建模可视化、拓展编程、多用户协作/空间共享、模型部署等几大功能，通过前端 WEB 界面的可视化操作，完成轻量级的建模。从技术上分为两大模块，分别是后端和前端。前端提供人机交互界面。WEB 前端通过 RESTful 风格的 API 与后端交互，后端负责了数据管理、算法管理，并提供异构算法流程调度引擎。如图 6 所示。

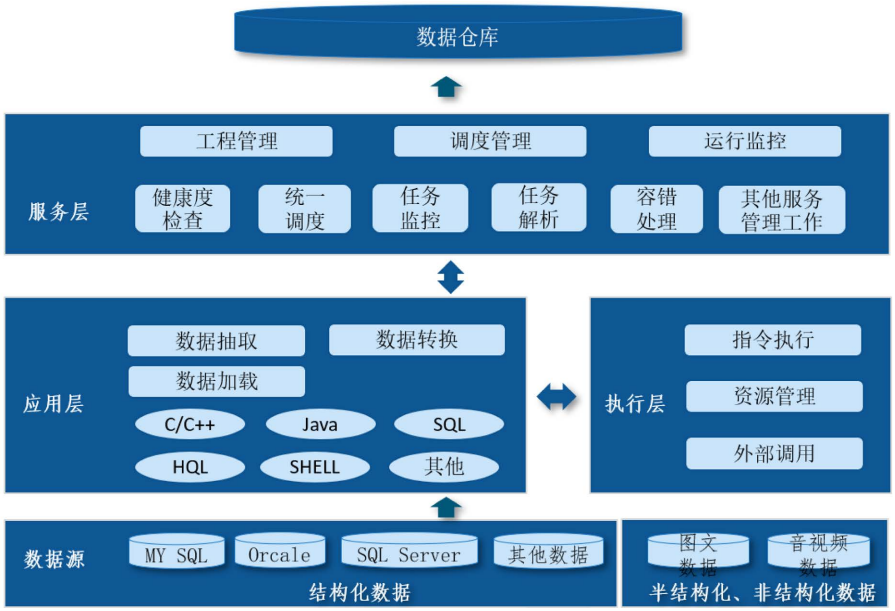

Figure 5. Data management module structure
图 5. 数据管理模块结构

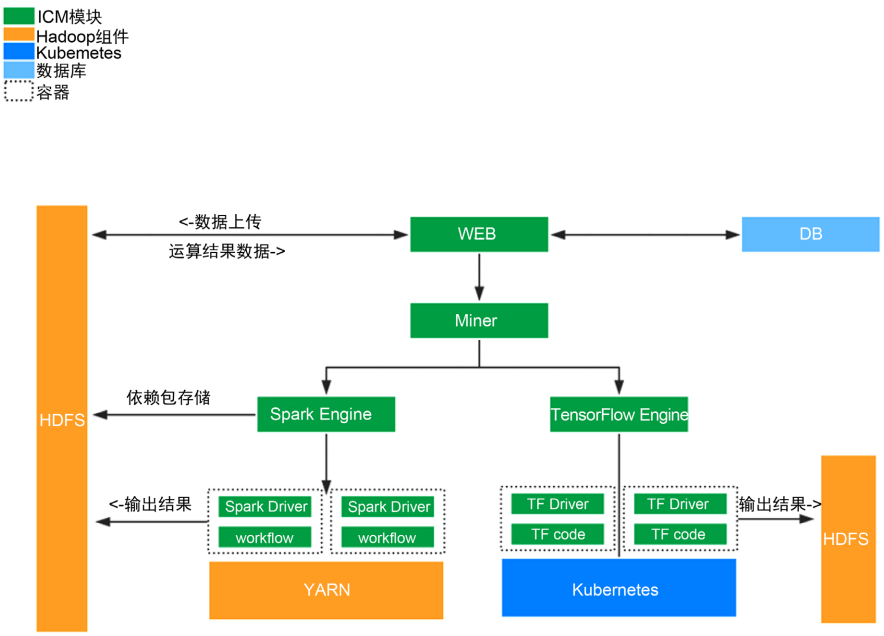

Figure 6. Analyze the technical structure diagram of mining module
图 6. 分析挖掘模块技术结构图

前端是基于 React 框架的组件化单页应用。它遵照 RESTful 标准，异步向后端获取结构化数据，通过高性能渲染引擎生成可动态交互的视图界面。核心工作流编辑器是基于 React+Redux+D3 开发的高性能强交互性应用，通过页面交互改变 URL，React-Router 检测 URL 的变化，自上而下渲染一棵虚拟 DOM 树呈现给用户，期间利用 React 高效的 diff 算法以提升渲染性能。WEB 前端的主要工作就集中在业务无关的高可复用性组件的开发和业务相关的复合组件的实现上。根组件一般会使用 RESTful API 向后端请求数据并更新视图，用户和组件间的交互会改变组件状态，组件根据不同的状态呈现不同的视图。

后端使用 Spring Boot 搭建开发环境结合 Spring MVC、Spring Data、Spring Security、Hibernate、Flyway 等作为开发框架，在此基础上开发具体业务逻辑，前端通过 RESTful 风格的 API 和具体的路由表和后端交互，后端和引擎之间通过 MQ 交互，持久化数据库使用 MySQL。

5. 总结与展望

随着信息化、数字化建设的完善，作战数据采集已经成为试验过程中的重要内容，针对作战试验过程中的主要试验内容、解决的关键问题与方法和记录的关键数据与分析结果，体现了作战试验的主要成果。想要挖掘作战数据中的价值，那么就需要先进行数据治理来提高数据质量，减少分析挖掘的时间，实现作战数据的可持续应用发展。本文通过对作战数据特性分析，围绕作战试验数据治理的需求，提出了基于 Hadoop 作战试验数据平台的数据治理框架，应用面向作战数据的 HAO 治理模型，实现了大数据平台数据规范统一管理，实现更加高效地发挥和挖掘作战试验数据的价值。

参考文献

[1] 军事科学院军事科学信息研究中心. 试验鉴定领域发展报告[M]. 北京: 国防工业出版社, 2019.

[2] 姚鹏飞. 装备试验大数据应用架构研究[J]. 舰船电子工程, 2019, 39(1): 10-13+113.

[3] 吴志凡, 蒋瑞琼, 万亮. “大数据”时代我军信息化建设应对策略刍议[C]//中国指挥与控制学会. 第二届中国指挥控制大会论文集(下). 2014.

[4] 迟明祎, 侯兴明, 周瑜, 孙瑜. 常规武器作战试验数据再利用有关问题研究[J]. 兵工自动化, 2021, 40(5): 8-13.

[5] 张绍华, 潘蓉, 宗宇伟. 大数据治理与服务[M]. 上海: 上海科学技术出版社, 2016: 1-224.

[6] 杨琳, 高洪美, 宋俊典, 张绍华. 大数据环境下的数据治理框架研究及应用[J]. 计算机应用与软件, 2017, 34(4): 65-69.

[7] 吴信东, 董丙冰, 堵新政, 杨威. 数据治理技术[J]. 软件学报, 2019, 30(9): 2830-2856.