http://www.avionicstech.com.cn avionicstech@careri.com
DOI : 10.12175/j.issn.1006-141X.2020.02.03

基于大数据的航空数据采集与处理系统研究与设计

殷华杰，张 彦，王 凯
（中国航空无线电电子研究所，上海 200241）

［摘 要］提供了一种基于大数据的航空作战数据采集与处理系统架构技术，包括：多源异构数据采集层，用于获取原始数据并发送给数据治理层；数据治理层，用于接收和对所述原始数据执行解析、清洗转换形成干净数据，并进行预处理形成主题数据；数据管理层，用于对数据生命周期、元数据、数据库、文件等进行管理及统计分析；数据服务层，用于提供数据智能缓存、事件分析、数据发布、数据挖掘等服务；数据应用层，用于接收所述主题数据进行任务复盘和回放，以及基于元数据进行数据地图、商业智能（BI：Business Intelligence）等有价值的数据分析应用。该方法能够有效提高航空数据管理的智能化水平，挖掘数据价值。

［关键词］数据采集；清洗转换；数据生命周期；元数据；数据挖掘；数据地图
［中图分类号］TP311.13 ［文献标识码］A ［文章编号］1006-141X(2020)02-0011-05

Research and Design of the Data Acquisition and Processing System Based on Big Data

YIN Hua-jie, ZHANG Yan, WANG Kai
（China National Aeronautical Radio Electronics Research Institute, Shanghai 200241, China）

Abstract: An architecture technology of data acquisition and processing system is provided for aerial operations based on large data, including: multi-source heterogeneous data acquisition layer for acquiring raw data and sending it to data processing layer; data processing layer for receiving and parsing the raw data, cleaning and transforming to form clean data. The data management layer is used to manage and analyze data life cycle, metadata, database, files, etc. The data service layer is used to provide data intelligent caching, event analysis, data publishing, data mining and other services. The data application layer is used to receive the said topics. Data are rewritten and replayed by tasks, and valuable data applications such as data map and BI analysis based on metadata. The method described can improve the intelligence level of aviation data management and mine the value of data.

Key words: data acquisition; cleaning and conversion; data lifecycle; metadata; data mining; data map

目前大数据技术发展突飞猛进，但在航空作战领域的应用还比较少。随着数据收集手段的不断丰富及完善，越来越多的行业数据被积累下来，数据规模已增长到传统软件无法承载的海量级别。目前

收稿日期 : 2019-11-26
引用格式 : 殷华杰，张彦，王凯 . 基于大数据的航空数据采集与处理系统研究与设计 [J]. 航空电子技术 , 2020,51(2): 11-15.

的航空领域机载数据管理系统已形成较为成熟的系统架构，但是数据卸载到地面后，如何管理庞大的航空数据仍是一个亟待解决的问题，包括解决大规模数据的高效存储、检索问题，并通过数据挖掘从中得到有效信息，以支持越来越复杂的上层业务需求。本文描述的基于大数据的航空作战数据采集与处理系统架构技术，通过改造传统数据采集与处理技术，可有效解决上述问题，推动航空作战数据管理的进一步发展。

本文首先介绍了大数据在航空领域的研究现状，然后详细说明了基于大数据的航空作战数据采集与处理方法，并进行系统化结构化，提炼形成航空作战大数据的平台架构。

1 研究现状

航空系统生来拥有大数据基因，作为众多系统及零部件构成的复杂产品，现代化飞机装载的传感器能够记录数以千计的飞行参数，每次飞行仅仅是发动机就能产生 1 TB（Terabyte）数据[1]。大数据技术在航空作战领域具有广阔的应用前景。

民航领域的大数据应用要早于军用领域，美联航把“收集、探测、行动”定义为新的数据收集三部曲，并加入 150 多个影响旅客消费的变量以实时评估其可能动向[2]。美联航曾表示航空系统在数据收集上一直做得很好，但在数据利用上却并不擅长[1]。国内航空公司在营销、客服等业务领域开展诸多尝试，如春秋航空的社交网络大数据营销、南航的飞行数据分析及客户分析等[2]。C919 飞机的总设计师吴光辉认为，迄今为止航空大数据仍未得到很好的挖掘利用[1]。

军用航空领域的大数据应用尚处于起步阶段，主要用于装备保障建设中，如分散的数据中心如何实现相互之间的数据传输、共享，充分发挥各数据中心的作用，提高装备保障数据整体利用率。军队装备保障信息化重点工作也从信息系统建设和武器装备信息化改造方面逐渐转移到对海量装备保障数据进行数据采集（包括抽取、清洗和整理等）和数据管控（包括转化、存储和动态组织等）上来[3]。

本文在传统航空作战数据采集与处理方法的基础上，引入大数据技术，提出了一种航空作战的数据采集、分析、应用的系统架构技术，它具有以下特点：（1）多张机载数据卡并行卸载；（2）多源异构数据采集与集成，包括关系型数据库、文件、流数据等数据源采集；（3）基于元数据的数据管理方式，对数据的状态、位置、时间及其他特性进行描述，以提供数据的精确理解、定位和其他应用方面的信息；（4）数据全生命周期管理，包括各类数据基本的增删改查、数据目录、数据标准化、数据血缘、基于主题的数据共享等数据管理功能；（5）基于授权访问机制保障数据安全。

2 航空数据采集与处理系统功能架构设计

基于大数据的航空数据采集与处理系统架构技术，实现了从数据源进行数据集成形成数据资产、数据治理形成数据仓库、数据服务化形成数据集市、数据共享发布给数据应用、应用处理后的有价值数据回流到关系型数据库的全生命周期过程。

航空数据采集与处理的数据处理流程图如图 1 所示，功能结构图如图 2 所示。

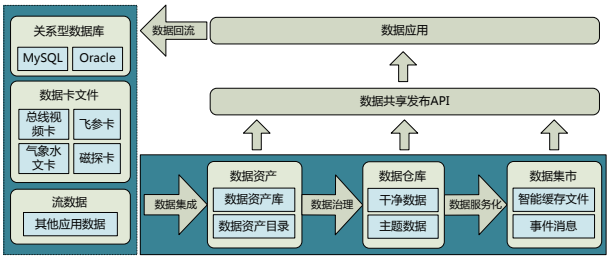

图 1 航空数据采集与处理方法流程图

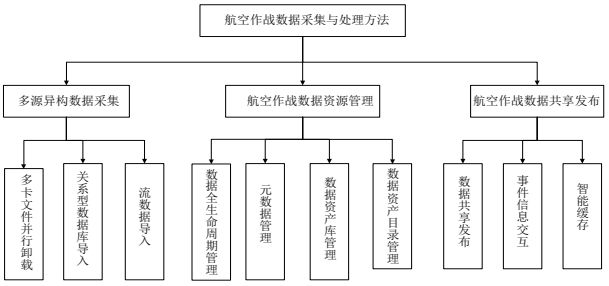

图 2 航空数据采集与处理方法功能结构图

2.1 多源异构数据采集

航空数据采集支持与常见数据源如关系型数据库和文件进行数据导入导出，以及流式数据导入。

2.1.1 多卡文件并行卸载

多卡数据卸载采用可视化操作界面，包括数据卡插拔识别、多数据卡目录浏览、目标目录浏览、数据并行卸载等功能。

（1）数据卡插拔识别：用户配置数据卡识别标识，通过检测 USB 插拔识别数据卡，遍历目录结构，判断卡中文件夹名称前缀是否包含数据卡识别标识来

区分数据卡，并在浏览窗口中显示卡内容；

（2）目录浏览：数据卡目录支持单选、多选、目录跳转、打开文件，不同数据卡窗口间切换时进行目录结构刷新。用户将当前任务元数据信息录入到数据库，并建立相应卸载文件夹，并显示目录结构；

（3）数据并行卸载：可以选择多张数据卡或目标目录进行卸载，支持多线程实现多卡并行卸载。

2.1.2 关系型数据库导入

关系型数据库作为数据源时，可以通过 Sqoop 数据导入工具导入，这是款开源工具，主要用于 Hadoop 大数据平台与关系型数据库间的数据传递，实现 MySQL 数据库数据导入到分布式文件系统、数据仓库和列式存储数据库。

2.1.3 流数据导入

其他应用程序输出的流数据可以通过 Kafka/Flume 流数据导入工具导入到分布式存储中。它是一个基于发布 - 订阅模型的消息系统，应用程序的实时数据均以话题形式发布，它将数据生产者和消费者分离，编写自己的生产者和消费者代码实现数据的导入通过。

2.2 航空数据资源管理

数据资源管理软件采用浏览器服务器架构（BS: Browser Server），用户通过浏览器页面轻松地进行操作，主要包括数据全生命周期管理、元数据管理、数据资产库管理、数据资产目录管理、配置管理、运维管理、日志管理等功能。

2.2.1 数据全生命周期管理

数据可管理的前提是标准化和规范化，针对大数据环境建设，通过数据资源管理工具对数据进行全生命周期管理，按照不同数据来应用数据管理技术。从数据源导入原始数据，通过解析模块将原始数据解析为标准格式文件并入库，通过清洗转换模块进行数据标准化形成符合规范的干净数据，并对所述干净数据进行预处理形成主题数据，并缓存为主题数据文件，通过数据共享发布接口，发布给上层应用，外部有价值的数据回流到系统中形成增值数据。

2.2.2 元数据管理

元数据是对数据以及对这些数据进行操作的应用和处理过程的描述性信息，可分为技术性元数据和商业性元数据 [5]。元数据管理包括以下几方面：

（1）用户录入元数据信息，包含任务编号、时间段、人员、数据卡编号、任务描述、数据状态等信息；

（2）自动创建目录、数据卡目录、数据类型目录，数据处理过程中会更新相应的数据状态信息，包括：原始数据、标准数据、主题数据、增值数据的状态、生成时间、路径信息；

（3）库表结构的元数据管理，对所有业务数据表、字段的含义进行确定性解释，存储于相应的中英文对照元数据表中。数据表按照“数据类型、数据种类、数据表”的树形目录结构进行中文字段显示；

（4）文件元数据信息表管理数据资产目录中所有文件的路径、大小、时间、创建 / 上传者、下载次数等信息，下载次数的下级描述信息包括当前文件的下载人员、下载时间等信息。文件元数据信息会根据数据的上传、下载、删除、复制、剪切等操作进行自动化更新，减轻数据运行维护人员的管理压力；

（5）在数据资源元数据描述中明确了数据资源共享应用的权限，接口服务层和应用层严格遵循数据资源元数据属性定义按照授权应用数据资源，有效保证数据资源共享的可控性；

（6）通过完善的元数据管理接口，提供任务统计、数据血缘、上下游层级关系等有价值的数据应用。

2.2.3 数据资产库管理

从数据源集成数据到存储结构后形成了数据资产。数据资产库管理包括以下几方面：

（1）各数据类型的标准数据表、干净数据表、主题数据表、增值数据表的管理。清洗转换模块执行标准数据表的导入，清洗后形成干净数据表，并进行预处理生成主题数据表，并更新相应数据状态；

（2）元数据信息表包括：任务元数据、数据状态元数据、库表结构元数据、文件元数据等表的管理；

（3）各数据表支持模糊检索，任务元数据表支持多关键字组合的高级检索，方便用户的操作。

2.2.4 数据资产目录管理

数据资产目录管理包括以下几方面：

（1）在用户配置的存储目录中，自动建立如下路径：根目录、任务目录、数据类型目录、数据种类目录。任务目录以“回合信息 _ 任务编号 _ 任务批次 _ 开始时间 _ 结束时间”格式命名；数据类型目录以各个数据卡名称命名；数据种类目录包括原始数据、标准数据、主题数据、增值数据；

（2）数据资产目录支持浏览、目录跳转、新建目录、上传、下载、删除、重命名、复制、剪切、检索、文件元数据信息查看等操作；

（3）针对每个任务目录，数据资源管理软件可以一键调用解析模块、清洗转换模块完成数据解析、标准数据入库、清洗形成干净数据、预处理形成主题数据等功能，减少用户的操作。

2.3 航空数据共享发布

数据共享发布模块提供数据共享访问统一接口，数据传输支持用户数据报协议（UDP：User Datagram Protocol）、文件传输协议（FTP：File Transfer Protocol）等协议，数据格式支持原始数据文件、标准数据包格式，包括数据共享发布、事件信息交互、智能缓存等功能。

2.3.1 共享发布功能

实现任务数据的共享发布功能，接收上层应用软件的主题数据请求，根据元数据信息表描述的相应主题数据文件路径，从相应数据资产目录中提取主题数据，通过调用文件传输协议（FTP：File Transfer Protocol）将主题数据文件传输至上层应用软件，并通过用户数据报协议（UDP：User Datagram Protocol）发送通知消息。

2.3.2 事件信息交互功能

通过用户数据报协议（UDP：User Datagram Protocol）实现上层应用软件同步控制信息交互功能，并接收应用同步控制信息，并将处理过后的应用同步控制信息发送给航空业务复盘软件，以及通过 Windows 消息机制发送至音视频回放软件，包括控制航空业务复盘软件和音视频回放软件的启动、数据预加载、显示、隐藏、播放、停止等功能。

2.3.3 数据智能缓存

用于对上层应用如任务复盘需要的主题数据，预先从主题数据表中缓存到本地形成标准格式文件。应用请求数据后可直接通过数据共享发布模块获取相应数据。

3 航空大数据平台

3.1 系统架构

在分析海量数据场景下，由于单台服务器的处理能力有限，数据分析人员通常采用分布式计算模式，但分布式的计算模型对数据分析人员提出了较高的要求，且不易维护，使用分布式模型，数据分析人员不仅需要了解业务需求，同时还需要熟悉底层计算模型。航空大数据平台提供了可视化的一站式开发、运维管理平台，能够更快速地解决海量数据计算问题，减轻数据运维人员管理压力。大数据平台架构如图 3 所示。

3.1.2 数据治理层

MySQL 是传统关系型数据库，一些重要的数据经过大数据处理后，还要回流到MySQL进行存储，以缩短调用的延迟，也达到双备份目的，确保数据安全。

Hadoop 分布式文件系统（HDFS：Hadoop Distributed File System）是基于 Hadoop 大数据平台的分布式、可扩展的分布式文件系统。

Hadoop 数据库（HBase：Hadoop Database）是基于 Hadoop 的非关系型列式存储数据库，主要适用于海量明细数据的随机实时查询。

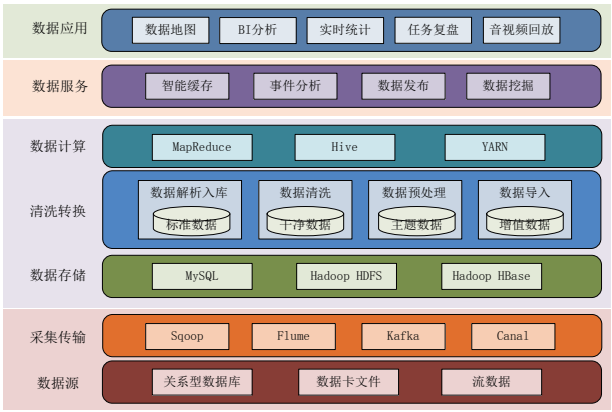

图 3 航空大数据平台架构图

MapReduce 是一个编程模型和计算框架，它通过编程模型进行编程开发，然后将程序通过这个计算框架分发到 Hadoop 大数据平台的集群中运行。支持关系代数运算和矩阵运算，大数据领域的计算需求基本可以通过 MapReduce 编程来实现。

Hive 数据仓库工具是基于 Hadoop 大数据平台的数据仓库，通过元数据来描述结构化文本数据，实现数据的提取、转化和加载，将结构化查询语言转化为 MapReduce 任务在后台运行，适用于批量离线数据的计算。

另一种资源协调者（YARN：Yet Another Resource Negotiator）是 Hadoop 大数据平台的集群资源管理系统，用于资源的抽象和调度，最初是为了改善 MapReduce 的实现，但它具有足够的通用性，也支持其他的分布式计算模式。

3.1.3 数据服务层

数据智能缓存：用于对上层应用如任务复盘需要的主题数据，预先从主题数据表中缓存到本地形成标准格式文件。应用请求数据后可直接通过数据

共享发布模块获取相应数据。

事件分析接口：通过数据共享发布模块接收应用的事件信息并进行分析，将处理过后的应用事件反馈信息通过共享发布模块发送给上层应用。

数据发布接口：用于发布主题数据给上层应用，根据元数据信息表描述的相应主题数据文件路径，从相应缓存目录中提取数据，通过调共享发布模块发送给上层应用，并发送通知消息。

数据挖掘：Mahout 数据挖掘工具提供了很多分类、聚类和回归挖掘的算法，并且通过 Hadoop 大数据平台将算法有效地扩展到分布式系统中，分布式的协调处理将时间消耗尽量控制在线性范围。R 语言数据挖掘工具采用脚本语言更容易理解，提供丰富的范例，支持数据分析、作图等功能，提供交互式环境和可视化工具 [4]。

3.1.4 数据应用层

数据地图：通过元数据提供的接口，实现数据血缘、影响分析等有价值的数据应用。

商业智能（BI：Business Intelligence）分析：提供业务数据报表，可视化业务统计分析，如一年或一月内的任务信息统计、各类业务数据占用情况等。

实时统计：实现实时监控集群主机的资源占用（如处理器、内存、磁盘占用等）、接口调用情况，在达到临界值时进行告警和日志事件记录。

任务复盘：通过共享发布模块发布主题数据到复盘应用进行任务复盘。

音视频回放：通过共享发布模块发布数据到音视频回放软件进行回放。

4　结束语

针对航空保障场景的大数据应用案例较少的现状，本文在传统航空数据采集与处理方法的基础上，引入大数据技术，提出了一种航空数据采集与处理系统架构技术，向用户提供数据集成、数据分析、算法开发、数据应用等可视化的一站式开发、运行维护管理平台，能够更快速地解决海量数据计算问题，减轻数据运维人员管理压力，且通过用户授权机制保障数据安全，有效解决了分散的、规模庞大的航空数据在地面的数据管理问题，为大数据技术在航空领域的应用提供了一种可行的方式。

参 考 文 献

[1] 孔祥芬 , 蔡峻青 , 张利寒 , 等 . 大数据在航空系统的研究现状与发展趋势 [J]. 航空学报 , 2018, 39(12): 8-23.

[2] 芮祥麟 . 大数据在航空业的应用 [J]. 软件和集成电路 , 2015(2): 66-67.

[3] 马双涛 , 王柏华 , 陈乃阔 , 等 . 军队装备保障信息化 “ 大数据之路 ” 初探 [J]. 价值工程 , 2016, 35(19): 202-204.

[4] 黄申 . 大数据架构商业之路 : 从业务需求到技术方案 [M]. 北京：机械工业出版社 , 2016.

[5] 曹蓟光 , 王申康 . 元数据管理策略的比较研究 [J]. 计算机应用 , 2001, 21(2):3-5.