文章编号:1671-4598(2021)07-0175-05 DOI:10.16526/j.cnki.11—4762/tp.2021.07.035 中图分类号:TP183 文献标识码:A

民机运行支持数据湖设计与实现

马 驰
(上海飞机客户服务有限公司,上海 200241)

摘要:随着大数据技术的发展,如何存储和处理各类民机运行支持数据,为企业用户提供所需数据,已成为主制造商企业进行数字化转型、洞察企业盈利和增值的关键因素之一;当前,主制造商数据采用传统数据库或数据仓库的模式无法满足航空运行数据指数级增长的需求;同时,数据管理标准、格式差异较大,不同用户无法快速获取有用信息,造成“企业数据孤岛”;鉴于此,文章提出了一种基于Lambda的运行支持数据湖系统设计方法;首先介绍了数据湖的概念和特点;然后,介绍了基于Lambda的运行支持数据湖系统的架构设计方案;结合运行支持数据湖的服务需求对系统功能进行设计;基于Angular、Spring Boot等开发了民机运行支持数据湖系统,为主制造商开展集中式数据管理、挖掘数据应用价值,实现企业数字化转型提供支撑。

关键词:民机运行支持;数据湖;Lambda

Design and Implementation of Civil Aircraft Operation Support Data Lake

MA Chi
(Shanghai Aircraft Customer Service Co., Ltd., Shanghai 200241, China)

Abstract: With the development of big data technology, how to store and process various types of civil aircraft operation support data to meet the data needs of enterprise users has become one of the key factors for major manufacturers to carry out digital transformation and gain insight into corporate profits and value-added. The traditional database or data warehouse model for main manufacturer data cannot meet the exponential growth of aviation operation data. Data management standards and formats differ greatly, and different users cannot quickly obtain useful information, resulting in " enterprise data islands". This paper conducts a Lambda-based operation support data lake system design method. Firstly, the concept and characteristics of the data lake are introduced. Then, the architecture design ideas and technical implementation methods of the Lambda-based operation support data lake system are introduced. The system functions are designed in accordance with the service requirements of operating and supporting the data lake. Based on Angular, Spring Boot, etc., a civil aircraft operation support data lake system has been developed to provide support for manufacturers to carry out centralized data management, tap data application value, and realize the digital transformation of enterprises.

Keywords: civil aircraft operation support; data lake; Lambda

0 引言

随着数字化、大数据技术的发展,数据已成为驱动航空企业特别是民机主制造商企业创新、盈利、增值的关键要素之一。民机主制造商数据量、数据来源和数据格式日益增多[1]。企业需要一个能够存储各类原始数据的大型仓库,用于处理各类数据,以满足不同业务对数据的存储、处理、分析及传输需求,为挖掘数据价值提供高性能的服务支撑[2]。国外航空主制造商GE构建以Postgre+mongoDB+Redis+Blob(S3)的数据湖存储组合技术的大规模存储技术的民航数据湖生态系统,以满足不同数据集探索发现、分析、数据服务和报告及可视化服务的需求[3]。目前,国内民机主制造商大多采用传统数仓模式进行数据管理,数据标准格式差异较大、应用端数据获取不规范,导致数据质量不足,数据分析应用困难,造成数据孤岛和冗余现象[4-5]。

本文提出基于Lambda模型的民机运行数据湖系统设计方法。数据收集、管理、业务分析用户可在系统中完成数据获取、清洗、标准化转换以及定制、自助的数据服务等功能,形成数据资产目录,挖掘数据应用价值,助力民机主制造商企业和数据应用决策,实现以客户/服务为中心的数字化转型。

1 数据湖的概念及特点

1.1 数据湖概念

数据湖是一个可以存储任何形式(包括结构化数据、半结构化数据和非结构化数据)、任意规模的原始数据仓库[6]。结合用户使用需求决定对哪些数据湖原始数据进行结构化处理。通过对数据的查询、处理、分析消费,帮助企业用户快速挖掘数据有用信息。

对于民机主制造商而言,数据包括生产类数据、运行类数据、经营类数据、管理类数据、外部综合类数据五类。

生产类数据如PLM数据、EBOM/MBOM等,需同飞

收稿日期:2020-11-03; 修回日期:2021-01-08。
作者简介:马 驰(1978-),男,山东平度人,高级工程师,主要从事航线数据管理、分析及数字化产品研发等方向的研究。
引用格式:马 驰.民机运行支持数据湖设计与实现[J].计算机测量与控制,2021,29(7):175-179.

机构型管理数据保持一致，此类数据变更频率不高，属于相对静态数据；

运行类数据主要以时间线进行分类、分层管理，分为飞行类数据、维修类数据、燃油效率类数据，如机上 FDR/DAR/QAR 数据、ACARS、EFB 数据以及地面例行工卡/非例行工卡等。此类数据需要实时或近实时更新，数据变更频率较高，属于相对动态数据；

经营类数据，如 ERP 数据、SCM 数据、CRM 数据，此类数据支撑运行经济性分析和企业经营，是经营决策的重要依据；

管理类数据包括各类运行支持计划、进度、问题改进措施、解决方案等转化形成的数据以及组织、人员等管理要素数据；

外部综合类数据，如气象数据、地理数据、油价数据、空管航线数据、航旅流量数据等，主要来源于第三方数据源。本文研究内容主要针对民机主制造商运行类数据。

民机运行支持数据湖旨在收集、存储飞机自交付、运营过程产生的气象、地理、机场、航班运行、维修等原始数据。数据形式包括数据库存储的结构化数据、EXCEL、XML 等半结构化数据以及 PDF、WORD、音视频等非结构化数据。民机主制造商用户可对数据进行结构化处理，便于数据分析人员进行数据分析。同时，为了增强用户使用分析体验，民机运行支持数据湖还为用户提供可定制、自助式服务、数据中台和标准化、简单易用的数据服务，使得数据分析人员更专注于数据、算法和业务，加快民机主制造商产品迭代，增强用户的购机体验，形成民机主制造商数据生态。

1.2 数据湖的特点

依据数据湖的普遍定义，数据湖具有以下几个特点：

1）类型多样。数据湖可存放各种类型的数据，包括结构化数据、PDF、WORD 等非结构化数据、EXCEL、XML 等半结构化数据。数据湖可处理所有类型的数据。数据类型依赖于数据源系统的原始数格式。对于民机运行支持数据湖而言，包括运控系统数据、维修类工卡数据、飞行记录本、驾驶舱音视频数据等；

2）原始记录。数据湖收集各类原始数据，并保留数据最原始的特征，为数据的加工和消费提供丰富的可能。对于民机运行支持数湖而言，是否有数据转换、清洗、加工等处理需求，所有数据入湖必须存储原始数据；

3）海量存储和计算能力。数据湖拥有强大的计算能力，用来处理和分析所有类型的数据。用户也可根据需求将处理后的数据存储成各种类型数据文件格式。所有分析后产生的数据均会被存储起来供用户使用。

除了具备上述特征外，民机运行支持数据湖还具备可定制、自助式的数据服务、数据中台服务的特点。

1）可定制、自助式的数据服务。对于民机运行支持数据湖而言，可靠性工程师需要航司日报数据、基础数据、发动机运行数据等进行可靠性分析；维修工程师需要利用 QAR 数据、ACARS 数据、CMS 数据、维修工卡、EO、AMM 手册等进行维修状态监控和维修预测；飞行运行支持工程师则需要 QAR 数据、ACARS 数据、飞行履历本、气象地理数据等开展飞行品质分析服务。不同类型用户，数据分析需求也不同。因此，数据湖需要为用户提供可定制和自助式等多种服务模式，增强数据服务的弹性。各类工程师可根据自身需求，采取数据订阅或自制视图的方式形成服务于自己日常工作的数据集，进而开展建模、监控、分析工作；

为避免数据集的冗余、更高效地实现数据的高效消费，系统需支持数据发布和分享，帮助其他业务相关用户快速获取数据集。系统也为数据工程师提供在后台不同数据视图应用情况分析功能，帮助用户发现更深层次、更多维度的数据关系，组合或构建更高效的服务新视图，或对数据湖集成、存储、消费环节进行优化，以提高整体运行效率。

2）数据中台服务。随着视图数据的增多，必定会产生一些重复度高、使用频率高的数据服务需求，数据工程师可整合相关需求，开发标准化的数据服务（Data Service），形成轻量级服务对外公开数据，例如 QAR 译码服务等。由于数据湖中大量业务数据以原始形态被提供给消费者，存在数据形态斑驳，质量参差不齐，数据延迟等问题，导致分析人员需要花费大量时间进行数据整备工作，数据中台可将同质化服务需求整合后，提供标准化服务接口，迅速拉近数据与分析业务应用之间的距离，使数据分析人员可以更加专注在分析建模和应用开发本身，从而缩短项目周期，加快产品迭代速度。

2 运行支持数据湖设计方案

由于航空主制造商有大量的存量系统和不断新建的新系统，因此很难从零开始构建一个全新的数据湖以承担企业数据中心的职责，传统企业构建数据湖首先应该对数据及数据关系进行分类，定义统一的企业模型，调整现有数据流程，并以增量的方式逐步构建数据湖。

数据湖的多种数据处理方式，大致可以分为批处理和（近）实时数据处理，这两种场景在航空企业普遍存在。为同时发挥流处理和批处理的优势，保证大型数据集执行跨度的可伸缩性、数据负载等，本文提出了基于 Lambda 架构的主制造商运行支持数据湖的构建方法。

民机运行支持数据湖工作流程如图 1 所示。数据流入是构建数据湖的起始。主制造商运行支持数据湖数据源包括飞行类数据、维修类数据、燃油效率类数据，如机上 FDR/DAR/QAR 数据、ACARS、EFB 数据以及地面例行工卡/非例行工卡等。通过自动/半自动方式，从数据源获取批量/流式数据。由于数据来源不同，在使用数据之前，需要对数据进行清洗、熔接、缝合，完成数据的集成。数据清洗是对重复数据、错误数据和缺失数据的剔除和修正过程，以满足后续对数据的操作和数据可用性。数据的集成过程需要借助于 ETL 等工具。集成后的数据进行标准化规范，存储至数据库中，形成数据资产目录和数据图谱。

用户可按需直接提供流式数据服务。当数据更新后，更新数据资产目录和数据图谱。在此基础上，数据工程师和业务工程师开展数据应用消费包括数据分析、大数据处理、自助工具与服务等，便于应用端用户更专注于数据应用，深入挖掘数据价值。

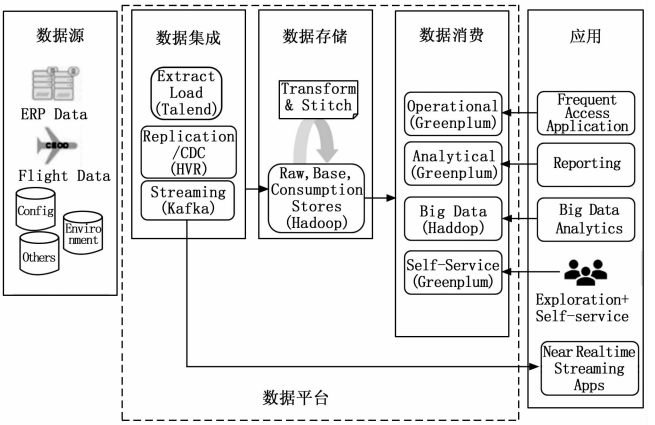

图 1　主制造商运行支持数据湖工作流程

2.1　基于 Lambda 模型的民机运行支持数据湖架构设计

Lambda 架构是一种整合离线计算和实时计算的大数据处理框架。通过批处理和实时处理功能来平衡数据延迟，实现数据容错，具备高容错、低延时和扩展性好等特点。尽管 Lambda 架构将多种大数据组件串联在一起实行一体化管理，但仍会在后续数据治理和开放能力上存在问题和痛点。因此，本文在搭建基于 Lambda 架构的民机运行支持数据湖架构的过程中，提供了多种平台及工具来助力民机运行支持数据湖的构建。

基于 Lambda 架构的民机运行支持数据湖架构分为数据获取层、消息层、批处理层、快速处理层、服务层和数据存储层，如图 2 所示。

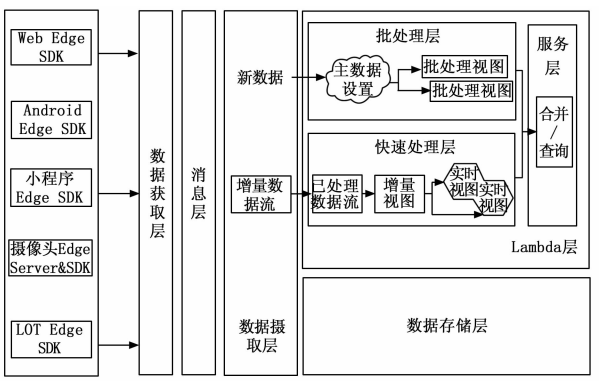

图 2　基于 Lambda 架构的民机运行支持数据架构

1）数据获取层：从多个数据源获取数据，并转换为消息层可处理的消息或事件，转换的目的是最小化传输延迟，且消息层如无法到达，数据获取层需缓存数据以备故障恢复。

民机运行支持数据包括外部源系统实时数据以及人工获取的大量历史数据。使用 SQOOP、Flume 工具完成数据的收集。使用 SQOOP 工具可将源系统关系型数据库数据收集存储至目标位置。使用 Flume 工具可确保数据聚合存储在目标位置。当系统组件、硬件发生故障或网络带宽性能不佳时，Flume 仍能保证系统主要功能运转，而非完全关闭。为了避免大量数据无组织入湖导致出现“数据沼泽”的现象，系统设置多个 Flume 代理，确保数据湖中的数据可以按照不同的维度组织起来，例如，不同航司获取的数据存储在各自单独的目录中。

2）消息层：主要为数据湖架构里的消息中间件，主要作用是让数据湖各层组装件之间解耦，同时保证消息传递安全性。消息层支持队列通信与发布/订阅两种模式，即一对一和一对多消息消费模式。

民机运行支持数据湖采用 Kafka 组件作为中间层实现数据源和数据消费者的解耦。源数据汇聚至 Flume，经由 kafka 消息中间层一方面将原始数据写入 Hadoop 文件系统存储起来，另一方面，基于 kafka 的发布、订阅功能以及高可靠性、低延迟的特性将数据流入数据摄取层。

3）数据摄取层：主要作用是摄取数据用于处理和存储，即将数据快速传递到 Lambda 架构的工作模型中，该层关键功能包括：a）可按需扩展的负载能力；b）容错和故障转移能力；c）多线程多事物并行处理能力；d）快速将所获取数据结构转换为目标数据格式的能力，包括非结构化、半结构化数据转结构化数据。例如飞行数据译码、图像数据转换都在这一层完成。

基于民机运行支持数据的批量性和实时性的特点，摄取层采取 Flink 技术，提供数据并发以及并行化计算的流数据处理引擎，保证数据在大规模运行过程中，出现无序或者延迟加载的情况下可以提供准确的数据处理结果。

4）批处理层：批量处理已提取数据，并转换输出为数据模型，为服务层提供输入。该层主要任务包括：在已摄取的原始数据基础上执行数据清洗、数据处理、数据建模算法；进行机器学习算法或数据科学处理，以产生高质量的数据模型；通过查重、检错等任务提高模型数据质量；具备故障恢复能力。

采用 HDFS 将不同源系统不同类型的原始数据进行持久化存储。采用 Pig 技术用于数据访问和处理。Pig 提供数据流功能，可将 ETL 功能抽象出来，允许用户查检索大型数据集，并进行必要操作。最后根据需求将计算结果存储起来。使用 Hive 技术，基于民机运行过程将数据划分多个主题域构建民机运行支持数据仓库，并提供数据汇总和即席查询。

5）快速处理层：将从数据摄取层获取的数据进行近实时处理，以满足对数据快速、高效、并发场景的需求，该层需确保数据处理、存储、读取能力达到近实时的预期，一般都建立在内存消息传输功能之上。

采用 Flink 实现基于内存计算的近实时数据处理和开箱即用的 windowing 功能，不仅基于事件时间，还可基于计数和会话。当出现数据故障，进行数据恢复过程中，不会造成数据损失。

6）服务层：服务层在 Lambda 架构中负责数据的对外提供，支持各种数据传输协议，是数据消费的接口层，对内从数据存储层消费数据，对外向数据消费者按约定接口提供数据传输，一般包括数据推送（数据导出、数据发布）和数据拉取（数据服务、数据视图）两种方式。如向维修工程师推送关注架机的最新维修工卡记录就是典型的数据推送，而为某个飞行品质 APP 以数据服务接口的方式提供数据更新则是拉取动作。

采用 SpringBoot 快速搭建服务层，并与 Swagger 等服务定义工具，为构建和发布通用 REST 服务。使用 Hive 技术实现数据视图、报表处理和即席数据分析。使用 pig 或 Sqoop 技术组件，通过预先设置的 cron 任务将数据从数仓中导出。

7）数据存储层：存储数据湖所有数据，由于摄取数据操作包括批处理和近实时处理两种，因此存储层要至少支持两种类型的存储模式，一般用 Hadoop 处理串行读写的批量数据，而用 Flume 处理需要随机访问和快速检索的流式数据。数据存储层应同时支持关系数据存储和分布式存储，如数据分析所需的关键参数，应存储在关系型数据库中提高检索效率，而原始数据的存储可采用 Hadoop 分布式解决方案。采用 HDFS 存储所有不同类型的原始数据。基于 Elasticsearch 建立索引数据服务，帮助用户进行快速检索。

综上所述，Lambda 模型中新增数据将同时分配到批处理层和快速处理层，分别形成批处理视图和快速处理视图，查询命令会合并两个视图来生成适当的查询结果。

2.2 运行支持数据湖功能

运行支持数据湖包括数据收集、数据管理、数据中台、系统监控和系统维护以及个人中心组成，如图 3 所示。

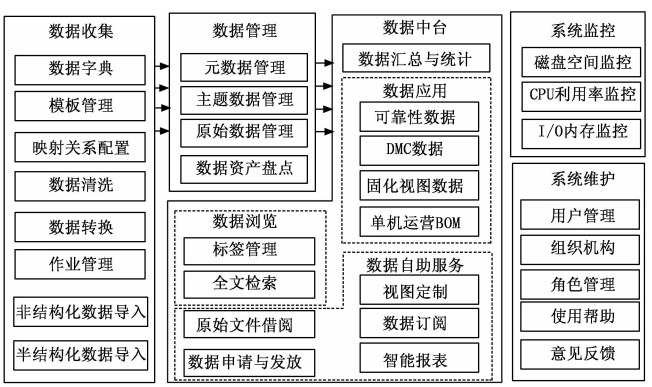

图 3 系统总体功能图

数据收集模块完成不同类型数据的导入，数据的流转配置、数据清洗、标准化转换以及数据模板和数据字典的维护功能。根据数据的类型，系统为用户提供了作业管理、半结构化数据导入、非结构化数据导入功能。作业管理实现系统到系统的数据抽取功能。此外，作业管理也提供数据转换标准化作业的新建、监控、管理功能。半结构化数据导入提供 EXCEL 等数据的导入、解析以及数据映射。非结构化数提供视频、图片、文档等数据的导入、数据映射和表单录入功能。数据清洗提供清洗规则库的配置维护功能和重复数据、错误数据和缺失数据的修正以及修正数据重新映射功能。数据转换提供数据转换规则的配置维护功能和转换问题数据的修正和修正数据的重新入库功能。

数据管理模块完成元数据的管理功能、主题标准化数据管理和原始数据管理功能。元数据提供系统元数据的查看和数据依赖关系的查看功能。主题数据管理提供转换作业标准化后数据的管理功能。标准化数据按照飞机运行过程业务分为基础数据、试验与验证数据、飞行与运行数据和维修与工程数据。原始数据提供不同数据源接入原始数据的管理功能。数据管理功能针对数据管理类用户。

数据中台模块提供数据资产、统计分析、数据自助服务、原始文件借阅、数据发放功能，主要面向数据分析和业务分析人员。用户在数据中台可查看形成的数据资产目录信息，支持对数据打标签。对于没有数据权限的用户，用户可以申请数据发放权限。同时，用户可借阅原始数据文件。针对一些重复度高、使用频率高的数据应用需求，系统提供数据应用支持和统计分析功能。由于不同类型用户的数据分析需求也不同，系统还提供了数据自助服务功能，支持数据分析用户自定义数据报表、自定义视图、数据订阅和自定义数据 API，便于业务分析人员快速进行数据消费，了解数据价值。

系统资源的使用情况关乎整个数据湖系统的稳定性、可靠性。系统监控主要监控 CPU 利用率和 I/O 内存利用率等。当数据湖中 CPU 利用率和和 I/O 内存利用率超限时，管理员可第一时间发现问题进行排查，增强系统用户的使用体验。

系统维护提供系统用户信息管理、组织机构信息维护、系统功能和数据权限管理以及系统使用帮助信息的维护管理功能。

3 软件实现

系统基于 Lambda 架构，采用前后端分离和微服务的原则，采用 Angular，Spring Boot、Echars、Redis 等设计开发，数据库用 PostgreSQL 和 HDFS，借助于 kettle 开源的 ETL 工具完成系统对系统结构化数据的抽取。

民机运行支持数据湖实现流程如图 4 所示。用户需梳理原始数据，结合数据入湖后的数据流向配置数据字典、数据模板、清洗规则、映射关系、转换规则。结合源数据的获取方式（包括文件上传、接口获取、人工录入等）以及数据格式（包括结构化、半结构化、非结构化）从不同的功能入口完成不同类型数据的收集、处理，经过数据标准化转换作业形成标准统一的数据，建立数据资产。通过对数据的聚合、导出、发布实现数据的统计分析、自助服务和数据应用。

4 数据湖应用结果与分析

以某型号某航数据为例，在民机制造商数据收集过程中，需要收集 PDF、EXCEL、WORD 的文件类型数据，将

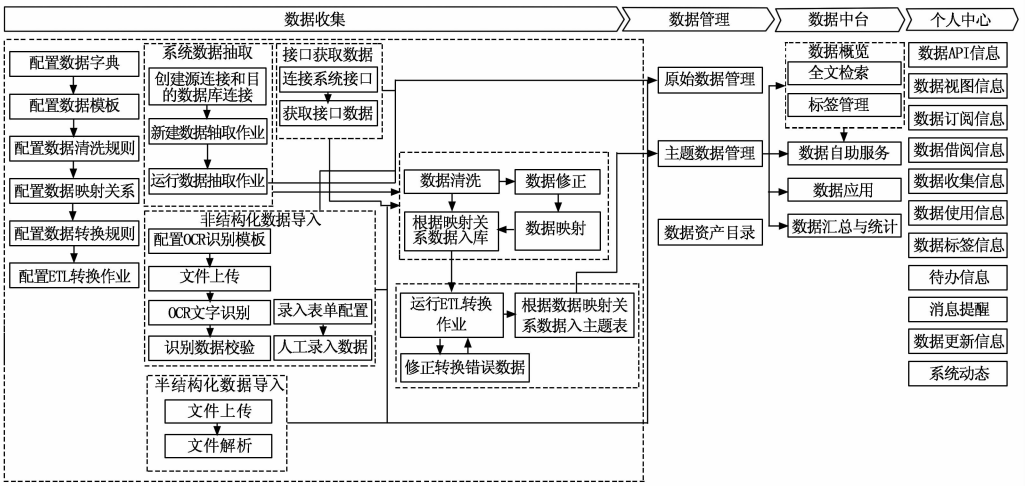

图 4　Lambda 架构数据湖工作机制

其作为民机运行支持数湖的输入。通过对不同类型数据的收集、存储、处理、标准化、整合，大大提高数据管理人员的工作效率。业务分析人员通过对数据中台的数据应用、数据自助服务等功能可快速完成数据的聚合、分析，快速响应用户数据分析的需求，降低了企业成本。数据管理人员通过监控数据的使用情况可快速了解数据价值，清理无价值的数据，避免出现“数据沼泽”。

5　结束语

本文设计实现了一种基于 Lambda 的运行支持数据湖系统，解决了传统数据库、数据仓库无法满足航空数据指数级增长、毫秒级摄取、多维度应用的问题，为主制造商开展集中式数据管理，实现数字化转型提供支撑。同时，基于 Lambda 的运行支持数据湖系统有利于形成以单一架机视图（SAV，single aircraft view）为核心的数据服务，通过将数据科学和机器学习技术，帮助业务部门高效利用数据，挖掘潜在价值，帮助主制造商优化设计制造，制定更灵活、更有针对性的经营策略，为航空公司运行运营工作提供支持。然而，存在不同航空公司对于同一数据定义不同以及各航空公司数据质量参差不齐的情况。随着接入数据来源种类的增多，如何优化运行支持数湖系统[7]，建立更为完善的数据湖安全及将是未来研究需要重点解决的问题。

参考文献：

[1] WALKER，CORAL. Personal Data Lake with Data Gravity Pull [C] //5th IEEE International Conference on Big Data and Cloud Computing，BDCloud 2015，160－173.

[2] HALEVY A Y，KORN F，NOY N F，et al. Whang SE (2016) Managing Google's data lake：an overview of the goods system [J]. IEEE Data Eng Bull 2016，39 (3)：5－14.

[3] 王一扬. GE 的工业数据湖平台 [J]. 新理财，2015 (11)：45－46.

[4] SUN D P. Big data learning resources integration and processing in cloud environments [J]. Journal of Chemical and Pharmaceutical Research，2014，6 (5)：936－943.

[5] 郭文惠. 数据湖技术——一种更好的大数据存储架构 [J]. 电脑知识与技术，2016 (30)：4－6.

[6] NATALIA M，ALEXANDER T. Big Data，Fast Data and Data Lake Concepts [J]. Procedia Computer Science，2016，1 (88)：300－305.

[7] 邱燕娜. 数据湖不能成为数据沼泽 [N]. 中国计算机报，2015－9－28.

(上接第 174 页)

[4] 童　晶. 基于生物电识别的远程遥操作仿人机器人控制系统研究 [D]. 沈阳：东北大学，2015.

[5] ZHANG X，CHEN J，YIN G，et al. An Approach for Human－Robot Interactive Control of Lower Limb Rehabilitation Robot Based on Surface EMG Perception [J]. Journal of Vibration，Measurement & Diagnosis，2018 (4)：866－880.

[6] LEON M，GUTIERREZ J M，LEIJA L，et al. EMG pattern recognition using Support Vector Machines classifier for myoelectric control purposes [C] //Health Care Exchanges，IEEE，2011.

[7] 加拿大推出手势控制臂环挥手指可控制电脑 [J]. 传感器世界，2013 (3)：44.

[8] 郑修军. 纵行神经束内电极的生物相容性及其记录的信号在电子假手控制中应用的实验研究 [D]. 上海：复旦大学，2003.

[9] 谢作述，王从庆. 一种肌电信号采集的神经接口设计 [J]. 计算机测量与控制，2020，28 (9)：168－172.

[10] 杜召辉，刘安东. 基于 Qt 的移动机器人上位机软件设计与实现 [J]. 计算机测量与控制，2018，26 (5)：113－117.

[11] REKHI，ARORA，SINGH，et al. Multi－Class SVM Classification of Surface EMG Signal for Upper Limb Function [C] //International Conference on Bioinformatics & Biomedical Engineering，IEEE，2009.

[12] 龚永富，王少云，雷仲魁，等. 高速、高精度数据采集系统的上位机软件设计 [J]. 电子设计工程，2020，28 (15)：38－42.

[13] 刘白林，段明晔，肖　亮，等. 基于 Linux/QT 的火控系统实时故障诊断方法 [J]. 火力与指挥控制，2011 (10)：176－179.