计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(4):77−81 [doi: 10.15888/j.cnki.csa.007847]

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

基于 Spark 的航空信息服务平台①

颜廷龙, 李　瑛, 王凤芹

(海军航空大学, 烟台 264001)
通讯作者: 颜廷龙, E-mail: 1724332258@qq.com

摘　要: 针对大数据时代下, 海军航空部队存在的种种数据治理问题, 设计了一种基于 Spark 的航空信息服务平台, 平台实现了航空数据的存储, 分析与挖掘等功能. 平台采用 4 层体系架构, 使用了 HDFS 分布式文件存储框架和 Hive 数据仓库工具实现了数据的存储和管理. 最后, 通过仿真实验, 比较在不同数据量下航空信息服务平台与传统航空数据仓库的性能优劣. 通过海军航空信息服务平台建设, 可以有效为海军航空部队实训提供数据支撑, 为平台使用者提供辅助决策.

关键词: 航空数据; Spark; 大数据; 数据管理; 数据仓库

引用格式: 颜廷龙,李瑛,王凤芹.基于 Spark 的航空信息服务平台.计算机系统应用,2021,30(4):77–81. http://www.c-s-a.org.cn/1003-3254/7847.html

Aviation Public Information Service Platform Based on Spark

YAN Ting-Long, LI Ying, WANG Feng-Qin

(Naval Aviation University, Yantai 264001, China)

Abstract: In response to the data governance problems about naval aviation in the era of big data, a Spark-based aviation information service platform is designed, enabling data storage, analysis, and mining. The platform has a four-tier architecture, with the Hadoop Distributed File System (HDFS) and Hive for data storage and management. Finally, the performances of the aviation information service platform and the traditional aviation data warehouse are compared through simulation experiments regarding different data volumes. The naval aviation information service platform can serve as a strong data support for naval aviation training and assist users for decision-making.

Key words: aeronautical data; Spark; big data; data management; data warehouse

随着当今世界的发展, 各行各业产生了大量的数据, 已经远超常规数据处理手段的处理能力, 海量数据处理面临重大挑战[1]. 作为海量数据处理的有效手段[2], 大数据处理技术发展已日趋成熟[3], 并在生活服务的各个领域中广泛发展[4]. 官方也制定了相关政策, 推动大数据技术的发展[5].

现代化战机装有众多传感器, 能够记录数百个飞行参数. 目前, 海军航空兵部队, 操课, 训练, 演习都产生了大量的飞行数据. 除了这些飞行数据, 维修保障基地, 军械部门也产生了大量数据. 如果可以将这些数据集成分析, 可以有效提高飞机飞行安全, 提高海军航空兵部队的科学化管理和决策能力.

针对大数据技术在航空系统的应用, 目前不少专家已经进行了研究. 在航空数据分析领域, Singh 和 Kaushik 介绍了使用大数据基础架构分析航空大数据的方法, 并应用大数据工具为机务维修保障提出辅助决策[6]. 陈金等基于大数据技术设计了一个飞机健康管理的平台[7]. Li 等针对目前机务维修保障存在的维护效率低等问题, 提出了一种飞机健康管理的大数据体系架构[8]. Rehm 等将高维数据可视化技术应用到航空

① 收稿时间: 2020-08-04; 修改时间: 2020-08-28; 采用时间: 2020-09-01; csa 在线出版时间: 2021-03-30

大数据和天气大数据, 解决了航空大数据和天气大数据的数据分析问题[9].

虽然在各大航空公司, 大数据技术已经得到广泛应用, 但部队的航空数据管理上还尚不成熟. 其业务管理模式, 数据来源, 平台性能需求等方面都与传统的航空大数据平台都有显著不同. 这就要求针对部队的实际业务需求设计航空数据平台.

Apache Spark 是用于大规模数据处理的统一分析引擎. 它提供多种语言的 API 接口. 它还包含丰富的上层基础应用. 相比于 Hadoop, 使用方便, 运行速度快, 适用场景更广泛.

航空信息服务平台设计是在针对航空数据的优化采集和分析基础上, 面向航空兵部队的应用需要, 实现航空大数据平台的开发和设计. 本文提出了基于 Spark 的航空信息服务平台设计方案, 首先进行平台的总体框架设计, 然后设计了主要的功能模块, 最后在实验环境下实现平台的设计与开发, 进行仿真实验, 验证平台在处理大数据方面的性能优势.

1 平台需求分析

1.1 航空信息的来源与性质

海军航空兵部队的航空信息涵盖飞行训练过程中产生的所有与飞机飞行相关的数据, 主要包括: (1) 飞参数据; (2) 地空数据链 (ACARS); (3) 作战指挥数据; (4) 任务计划数据等, 数据量随着海航部队飞机的增多和时间的积累逐步增大.

航空服务信息具有多源异构的属性, 包括多来源、多性质、多层次、关联性等特点.

目前, 海军航空部队传统的数据存储方案是各个机务部门将数据分散存储到多个系统中. 这样不便于工作人员的管理和维护, 因此迫切需要一个可以汇总存储多个系统的航空信息信息服务平台.

1.2 平台功能需求分析

根据对航空信息的来源的特点, 本文设计了一个基于大数据的航空信息服务平台, 使用 Hadoop 分布式文件系统 (HDFS) 完成航空信息的存储, 基于 Spark 和相关数据挖掘算法实现航空信息的快速处理, 数据挖掘[10]. 分析部队的实际需求, 平台必须满足下列要求:

(1) 分布式. 根据航空信息的数据量大, 来源广等特点, 平台需采用分布式文件系统存储, 在实现数据的高效存储和高容错性.

(2) 并行化. 平台必须支持运算并行化来提升计算速度, 具有良好的数据处理能力. 并且后续仅通过简单的节点增加就可以带来计算速度的提升, 以便于相关管理人员日后的维护和管理.

(3) 扩展性强. 平台应模块化设计, 由于海军航空部队的业务需求复杂, 需要平台可根据业务需求的变化, 更新数据分析模块, 满足定制化需求.

(4) 可用性强[11]. 平台以使用者为中心, 设计能符合使用者的习惯与需求, 简单易用.

2 平台总体结构设计

本平台基于 Spark 进行搭建将航空信息服务平台为四层架构, 自顶向下采用接口连接相邻层, 数据的获取和存储是最底层, 分别为数据源层和数据存储层. 再上层为计算分析层, 主要提供计算框架和数据处理功能. 交互应用层, 为平台的用户提供交互界面[12], 平台总体框架图如图 1 所示.

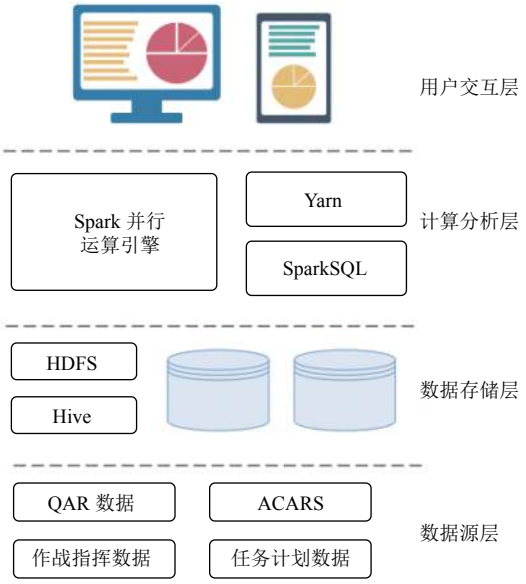

图 1 平台总体框架图

2.1 数据源层设计

数据源层的主要功能是数据的获取, 即各个航空单位存储的航空数据, 包含需要从原有关系型数据库导入的数据, 和新产生的数据, 数据具有多源异构的特性. 数据源层还会进行数据预处理工作, 通常包含清洗、集成、变换以及归约. 目的是对重复数据的清洗和对缺少数据的填补; 消除数据的冗余性; 将数据的存储形式改变以更适合数据处理[13]; 预处理后的数据经过 ETL 处理, 采用 Sqoop 将处理后的源数据高效的存

储到数据存储层的数据仓库中.

2.2 数据存储层设计

数据存储层采用 Hadoop 的分布式存储框架 HDFS, 将航空数据以文件形式存储获取到的航空数据, HDFS 采用主从架构, 由一个 NameNode 和多个 DataNode 组成. NameNode 主要管理文件块的 Namespace 和 Block 管理, 维护着系统文件树的元数据和各个文件所在的 DataNode 位置信息. DataNode 存储和读取具体文件, 并定时地向 NameNode 发送心跳信息. HDFS 将文件分块存储在各个 DataNode 上, 默认的 Block 块大小为 128 MB. 为提高平台业务运算效率, 采用 Hadoop 生态下的开源工具 Hive 构建航空信息数据仓库, 管理元数据.

2.3 计算分析层设计

计算分析层的作用是对存储的航空数据进行数据分析, 实现各种业务需求. 包括针对航空数据的数据挖掘, 执行 SparkSQL 命令等. 其中航空数据挖掘基于 SparkMllib 库和利用 Spark 编程接口自定义的聚类算法完成.

计算分析层的基本工作流程如图 2 所示, 首先, 构建运行环境, 创建一个 SparkContext, 并且向资源管理器 Yarn 申请 Executor 资源, 并启动相应资源. 然后 SparkContext 依据 RDD 的依赖关系构建 DAG 图, 同时创建一个 DAGScheduler 对象依据作业和任务的依赖, 制定调度逻辑, 将 DAG 图分解成 Stage, 因为 Stage 之间存在依赖关系, 只有前面的 Stage 运算完, 后面的才开始运算. 最后, 将完成的 Stage 发送给 TaskScheduler, 再由 TaskScheduler 将 Task 发送给 Executor 运行, 运行结束后释放计算资源[14].

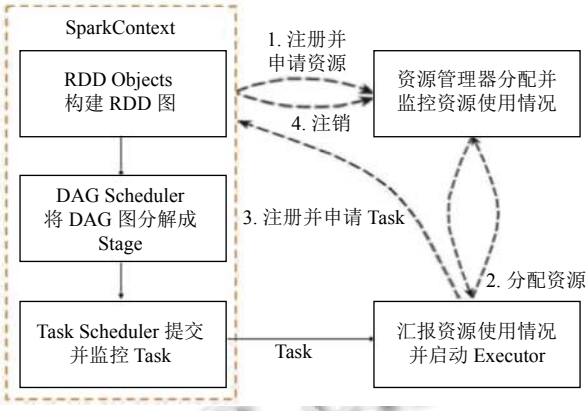

图 2 计算分析层基本工作流程

2.4 用户交互层设计

用户交互层主要功能是为用户提供良好的使用界面, 并包含数据查询, 数据分析, 数据可视化等功能, 并将航空信息信息直观地展示出来, 提高海军航空部队的训练效率和管理效能, 平台的数据可以通过图表, 直观展示出原始飞参数据, 平台的分析结果等.

3 主要功能子模块设计

航空信息服务平台的功能模块主要划分为数据存储模块、数据分析模块和信息查询模块 3 个功能子模块, 如图 3 所示. 数据存储模块的功能保障数据的存储和资源管理. 数据分析模块主要实现将数据挖掘算法写入到 Spark 中, 也可以使用 Spark 自带的 Mllib 机器学习库进行分析现对航空信息的数据挖掘. 信息查询模块主要实现对平台基础数据和分析数据查询, 还有相关数据的上传下载.

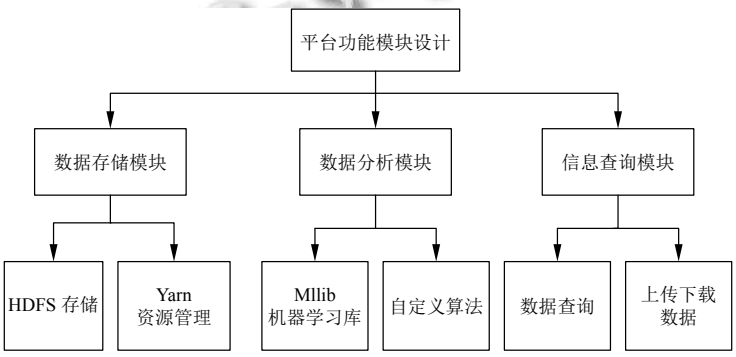

图 3 平台主要功能模块设计

3.1 数据存储模块设计

数据存储模块的功能是将从多个数据源获取的航空数据传入到 HDFS 中. 主要包括经过预处理后的航空数据以及平台生成的数据, 以特定的格式存储到 HDFS 中. 同时, 平台针对每一份数据都包含多份备份, 提高了平台的容错能力. 避免出现数据丢失的情况, 此

外, 还应用 Yarn 进行统一的资源管理和调度.

3.2 数据分析模块设计

数据分析模块主要是应用 Spark 的内存计算引擎, 实现针对航空信息的数据挖掘处理. 利用 Spark 自带的编程接口和相关组件实现数据分析模块的调入. 平台也根据飞参数据具有多元时间序列数据的特点, 自定义了数据分析算法, 可根据飞参数据进行飞机的飞行动作识别和划分.

3.3 信息查询模块设计

开发了基于航空服务信息服务平台的 Web 应用, 可对相关数据上传下载, 相关数据的可视化展示, 应用 SparkSQL 实现数据查询功能, 为用户提供了统一的数据源访问接口.

4 实验结果与分析

为了测试航空信息服务平台的性能, 本文将航空信息服务平台 (AISP) 与基于 SQL Server 的传统航空数据仓库 (TDW) 针对航空数据的计算与存储进行对照实验.

4.1 实验环境

本文测试航空信息服务平台采用的实验集群由 1 台 master 节点和 8 台 slave 节点组成, 集群的节点配置参见表 1.

表 1 计算机节点配置

配置	配置参数
服务器型号	戴尔易安信PowerEdge R330
CPU	Xeon E3-1220 v5
网卡	EB-SFP10G599-SR2
内存	8 GB
Hadoop版本	Hadoop
Spark版本	Spark 2.1
Jdk版本	Jdk1.7

4.2 集群性能对比

实验数据采用海军某场站存储的飞参数据, 数据量分别为 2.4 MB, 20 MB, 200 MB, 500 MB, 1000 MB 和 2000 MB, 共 5 组数据, 分别进行数据查找和数据预处理测试.

第 1 组实验测试使用同一条 SQL 语句查找符合条件的数据, 实验次数为 5 次, 查找时间取平均值, 测试用 SQL 语句为“select type,count(*) as count from test group by type order by count desc;”. SQL 语句执行效率对比如图 4 所示.

第 2 组实验比较航空信息服务平台针对不同数量级的实验数据预处理的情况, 采用插值法拟合空缺的数据, 实验结果如图 5 所示.

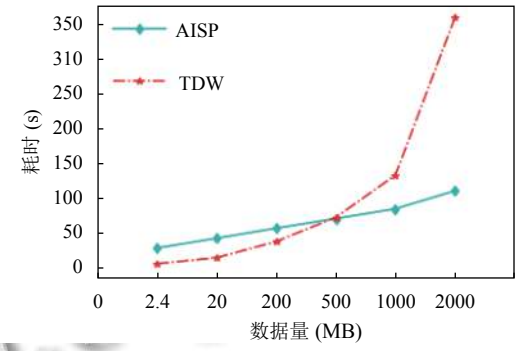

图 4 SQL 语句执行情况

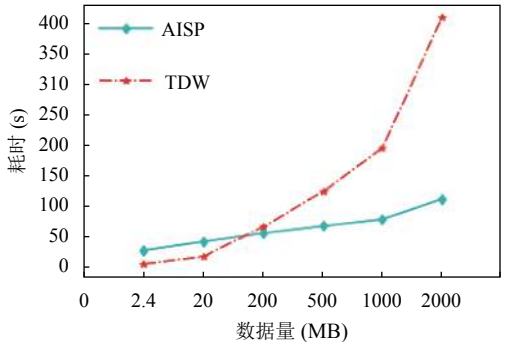

图 5 预处理时间使用情况

由实验结果可以看出, 当数据量较小的时候, 基于 SQL Server 的传统航空数据仓库数据处理速度优于基于 Spark 的航空信息服务平台, 但是当数据量达到 500 MB 时, 基于 Spark 的航空信息服务平台更具优势. 通过简要分析, 基于 Spark 的航空信息服务平台在接收到数据处理任务时, 需要进行初始化, 节点通信, 资源调度等, 这些将耗费一定的时间和资源, 由此可知, 在数据量小的时候, 大数据平台很多资源都浪费在系统资源开销上, 效率反不如单机系统. 但大数据在面对海量数据的优势依旧明显. 另外, 大数据平台可以通过增加集群数量实现性能的扩展, 更能适应海军航空兵数据量高速增长的需要.

5 结束语

本文主要实现了基于 Spark 的航空信息服务平台的设计, 采用 Hadoop 的分布式存储框架 HDFS 以文件形式存储获取到的航空数据, 和开源工具 Hive 构建航

空信息数据仓库, 并详细介绍了平台 3 个主要功能模块的设计. 另外, 本文实验对比了航空信息服务平台与传统航空数据仓库之间的性能对比. 最后, 实验结果表明, 航空信息服务平台在计算大规模数据上具有明显优势. 但目前航空信息服务平台的业务还不完善. 未来的主要工作一是增加针对其他航空信息的业务应用探索. 二是完善针对海军各型飞机的 QAR 数据译码工作.

参考文献

1 王浩. 大数据时代下的思维方式变革 [硕士学位论文]. 上海: 东华大学, 2015.

2 卫凤林, 董建, 张群. 《工业大数据白皮书 (2017 版)》解读. 信息技术与标准化, 2017, (4): 13–17.

3 禹艳. 美国的大数据国家战略研究 [硕士学位论文]. 长春: 吉林大学, 2017. 4–11.

4 宋京. 政府大数据建设推进机制研究. 电子世界, 2018, (11): 105.

5 陈金, 党帅, 吴波. 民机运行大数据分析平台整体架构研究. 计算机测量与控制, 2018, 26(1): 281–283, 288.

6 Li SJ, Zhang GG, Wang J. Civil aircraft health management research based on big data and deep learning technologies. 2017 IEEE International Conference on Prognostics and Health Management. Dallas, TX, USA. 2017. 154–159.

7 Rehm F, Klawonn F, Russ G, *et al*. Modern data visualization for air traffic management. NAFIPS 2007–2007 Annual Meeting of the North American Fuzzy Information Processing Society. San Diego, CA, USA. 2007. 19–24.

8 涂新莉, 刘波, 林伟伟. 大数据研究综述. 计算机应用研究, 2014, 31(6): 1612–1616, 1623. [doi: 10.3969/j.issn.1001-3695.2014.06.003]

9 姜吉宁. 基于 Spark 和 Hive 的新型种质资源数据仓库的设计和实现 [硕士学位论文]. 合肥: 中国科学技术大学, 2018.

10 马振宇, 张威, 吴纬, 等. 软件可靠性验证测试研究方法综述. 兵器装备工程学报, 2019, 40(7): 118–123. [doi: 10.11809/bqzbgcxb2019.07.024]

11 姜吉宁, 王儒敬, 魏圆圆, 等. 基于大数据的新型种质资源数据仓库的设计. 仪表技术, 2018, (10): 6–8.

12 张锐. 基于 Hive 数据仓库的物流大数据平台的研究与设计. 电子设计工程, 2017, 25(9): 31–35.

13 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. 航空学报, 2018, 39(12): 022311.

14 马彬彬. 简要分析大数据的发展现状与挑战. 科技资讯, 2016, 14(10): 142, 144.