文章编号：2096-1618(2018)05-0531-05

贵州省气象大数据平台架构设计

郭 茜，王 彪，汪 华，金石声
(贵州省气象信息中心，贵州 贵阳 550001)

摘要：气象大数据建设是气象信息化和气象现代化的重要内容之一。近年来，随着气象数据量暴涨，现有的气象设备与信息技术手段已很难满足气象业务需求。贵州省气象局内存在各个业务系统林立，部门内数据分散、气象数据收集缺乏全面性和系统性，“信息孤岛”的现象严重，数据整合受到不同系统和软件开发平台的限制，服务器利用率低下，CPU、内存、磁盘空间等资源得不到有效利用，数据存储存在单点故障等问题。针对以上问题，贵州省气象信息中心提出气象大数据平台整体架构的设计，帮助提高气象预报预测的准确率，使数据存储管理和服务实现集约高效和数据共享。对气象大数据平台建设中气象数据采集、数据存储和数据处理进行了概括，介绍了气象信息系统的现状，从完善顶层设计入手，对集群数据库方案选择进行对比，设计出合理、高效的气象大数据平台，实现气象大数据行业内部与外部的融合与共享。

关 键 词：气象大数据平台；Impala 数据库集群；GreenPlum 数据库集群
中图分类号：TP311.13 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2018.05.008

0 引言

随着信息技术、计算机技术和互联网技术的高速发展，社会各类数据爆炸性增长的速度令人瞠目结舌，社会对大数据的需求也随之增高，步入了大数据时代。气象数据有着大数据明显的4个特征：数据量巨大、数据类型繁多、速度快时效高、价值密度低，其中包含了结构化与非结构化等不同格式的数据[1]。

2015年9月中国政府制定了《促进大数据发展行动纲要》，提出把大数据作为推动经济转型发展的新动力，促进生产组织方式的集约和创新，同时明确了气象数据要接入政府信息共享交换平台、政府统一数据开放平台。行动纲要将大数据产业指定为国家战略产业并将贵州省确定为发展大数据产业的示范区，贵州省气象局响应省政府号召，大力推动气象大数据平台建设。目前，气象部门运用气象数据主要是为开展防灾减灾和服务农业生产，只有利用大数据技术将气象数据与其他行业数据的融合与挖掘，改变当下“信息孤岛”的现状，才能更好实现气象数据的价值，更好服务民生，服务地方经济发展。

主要介绍气象数据的特点、存储和应用，Impala数据库的特点，Impala数据库与GreenPlum(GP)数据库的对比，以及贵州省气象大数据平台架构。

收稿日期：2018-04-27
基金项目：贵州省科技技术资助项目(黔科合支撑[2017]2819)

1 气象数据概括

1.1 气象数据的特点

目前，气象信息系统通过国际通信系统、国内通信系统每日实时收集来自国内外、部门间交换的资料，包括：地面、高空、辐射、海洋、农业气象和生态气象、大气成分、卫星、雷达、气象灾害、数值预报产品、历史气候代用、科学试验和考察、气象服务产品和其他资料等共14大类。其中，气象服务的产品分为结构化数据和非结构化数据，除了遥测、传感设备产生的观测数据，大量参与气象服务和共享信息的数据都以视频、音频、图片、图像、文档、文本等多种形式存储，数据资源总量持续增长，符合“大数据”的4V特点。这些数据长期存储于气象各部门的平台上未能加以合理利用。

1.2 气象数据的现状

贵州省目前气象资料的存储现状是多种数据库系统并存，一种业务独占一个数据库，缺少统一规划和顶层设计，进一步开发与维护比较困难。资料管理与数据共享存在诸多不便。虽然经过多年发展，中国气象数据存储与服务能力不断增强，构建了基于CIMISS的集约化数据环境，CIMISS已经在一定程度上实现了对基础数据访问的业务需求，可以在一定程度上满足业

务基本需求,但面向智慧气象和大数据服务的现实需求,信息系统仍存在数据供应不足、数据交换传输不灵活、数据存储服务不够高效、业务系统烟囱不减、基础设施资源供应不足等困难。当前的气象业务需要从不同的角度、不同维度,抽取数据为决策服务提供支持。数据产品质量有待进一步完善,缺少有效的数据质量控制与质量评估手段,数据产品的生成与质量监视不同步,无法实时保障数据质量。当数据质量出现问题时,缺少有效的协同修正机制。此时,建立行业大数据、互联网数据、社会化气象观测数据的收集机制就变得迫在眉睫。

2 气象大数据平台

贵州省气象局基于云计算、大数据等新的信息技术,构建气象大数据平台。加强对气象部门内外数据的汇聚,优化数据交换、入库、加工和服务流程,开放信息系统的平台、数据和算法资源,促进实现气象应用和创新众创的发展模式,全面提升气象业务和服务水平。

2.1 系统架构

贵州气象大数据平台包括统一的元数据管理、分布式 ETL 管理、调度管理,实现大数据采集、处理、存储、对外共享,并具备一定的数据质量管理、系统自监控、系统自运维能力。气象大数据平台的构建基于 Hadoop 集群、Impala 集群和 Oracle 集群构建,总体架构分为数据源层、大数据平台层和应用层,系统架构如图 1 所示。为了保障数据高效稳定运行,气象大数据平台对数据资源进行全流程规范化管理,快速汇聚、交换、质控和入库。

图 1 贵州气象大数据平台系统架构图

大数据平台的数据源层包括所有与气象业务相关的数据源,通过对气象数据的分类,将其划分成数据块,包括 CIMISS 结构化数据、CIMISS 非结构化数据(卫星、雷达、数值预报等)、气象行业外部数据(交通、水利、航空、农业、保险等)、互联网数据和空间数据。

大数据平台层是基于接口机集群,对数据进行加工处理,生产丰富和高质量的数据产品并分类进行存储;利用接口机集群里的分布式 ETL 工具对原始数据进行数据抽取、清洗转换、文件合并,然后按照预先定义好的数据仓库模型,将数据加载到数据仓库中。大数据平台层是整个大数据平台架构的核心组成部分,一个完整的大数据平台应该提供离线计算、即席查询、实时计算、实时查询这几个方面的功能,只依赖于一种集群是无法良好地完成以上几个方面的,所以如何将它们组合好、利用好很重要。Impala 集群主要接收结构化数据和基础数据,由 Kudu,Alluxio 和 Impala 部分组成。Hadoop 集群主要存储非结构化数据和部分结构化数据,由 HDFS,Alluxio,HBase 和 Spark 构成;Oracle 集群主要存储专题应用数据。在大数据平台的上层部署 Solr 集群的分布式搜索,通过利用 Solr 的集中式配置、自动容错和实时搜索的功能,可以部署大规模处理索引量很大、搜索请求并发很高的数据场景和应用。Solr 集群主要存储索引数据[2]。在 Solr 集群的上层部署了 Music 气象数据统一服务接口,该接口基于下层部署的大数据平台,气象业务人员与气象科研人员可通过该接口获取到统一、标准、丰富的数据,并且

通过后端开发,针对不同的数据格式和数据种类在 Music 平台上进行部署,做到丰富气象数据的统一展示。

大数据平台的应用层位于整体架构的最顶层,主要用在基于大数据平台开发的业务应用,从大数据平台层中提取出来的数据,将应用于多项气象应用,能够高效满足即席查询、日常统计查询、多种众创应用的专题应用开发。目前贵州省气象局正在进行气象万千、防灾减灾联动支撑平台、气象与消费、气象与旅游等应用专题研究,并且已经取得了一定的成果。

2.2　气象数据存储

气象行业的数据存储现状比较复杂,在大数据处理中很多分析需要传统数据和文件分析同时进行。因此,贵州省气象大数据平台基于 Cloudera 框架搭建,综合利用传统数据库集群,Nosql 等多种数据库并存组合,为海量级的数据提供了一个可伸缩、综合、稳定的数据计算框架,包括 Spark,Impala,支持对 HDFS 的高性能 SQL 查询引擎,Hive 数据仓库,可方便业务人员管理快速增长庞大的气象数据,并且有很好的防护机制保证数据的安全性,这种混合型的架构需要大量的 ETL 过程来进行数据的转换和存储。贵州省气象大数据平台友好界面能够方便高效地管理该平台。以下介绍在 Cloudera 框架中采用的几种集群机制。集群部署结构如图 2 所示。

图 2　贵州气象大数据平台集群部署结构图

2.2.1　Hadoop 集群

Hadoop 能够维护大数据处理框架多个数据副本,确保能够针对失败的节点重新分布处理,支持非结构数据存储、查询及结构化数据的关联、处理。对于存储在 Hadoop 集群中的气象数据,设立一定的规则进行规范的存储[3]。

(1)结构化数据采集后存放到 Kudu 分布式存储系统中(包含审核过的数据及未审核的数据),其中未审核数据的更新操作基于 Kudu 实现单记录的增删改操作[4]。

(2)结构化数据的汇总通过基于 Kudu 上的 Spark 实现,对未审核数据每天进行重新汇总,并通过 Merger 的方式更新到 Oracle 数据库集群中。

(3)非结构化数据存储在 HDFS 与 HBase 中,其中小于10 M数据直接存储与 Hbase,大于10 M数据直接存储于 HDFS 中;其非结构化数据的索引数据存储在 Oracle 数据库集群。

2.2.2　Impala 集群

Impala 用于处理存储在 Hadoop 集群中大规模并行处理(MPP)的 SQL 查询引擎,是访问 Hadoop 分布式文件系统中数据最快的方法。在 Impala 中能够使 ETL 耗时周期缩短,实现快速的数据发现和写入[5]。根据 Impala 数据集群的特性,设立一些规则来存储这些气象数据。

(1)通过 Impala 实现基于 Kudu 的结构化数据长序列的灵活查询支撑。

(2)采用 Impala 存储结构化汇总数据,并实时将计算结果更新到 Oracle 数据库集群中。

(3)结构化数据采集后存放到 Kudu 分布式存储系统中(包含审核过的数据及未审核的数据),其中未审核数据的更新操作基于 Kudu 实现单记录的增删改操作。

2.2.3 Oracle 集群

使用 Oracle 集群提供高可靠性和可伸缩性的均衡负载能力,并且实现透明的应用程序故障切换、消除了单点故障的问题。在大数据平台架构中,Oracle 集群主要负责存储以下几类数据。

(1)存储采集结构化原始数据,并支撑对未审核数据的增删改操作。

(2)存储基于原始数据的按时间维度、空间维度等各类维度的汇总统计。

(3)存储面向应用专题的分析结论类数据。

(4)存储非结构化数据的索引数据信息。

大数据可通过多种方式来存储、获取、处理和分析,处理并存储大数据时,会涉及到更多维度,比如治理、安全性和策略。选择一种架构并构建合适的大数据解决方案极具挑战,要考虑非常多的因素。

2.3 Impala 集群与 GreenPlum 集群对比选型

在选择大数据平台的数据库时,最初选择了 GreenPlum(GP)数据库集群,但经过多次对比测试后,发现 Impala 数据库集群更符合气象业务的场景,并通过实验对 Impala 数据库集群与 GP 数据库集群进行对比选型。GP 数据库与 Hadoop 基本是同一时期出现的,是海量数据的一种新的计算方式。面对爆发式数据增长,它的架构易于扩展,并且在 CPU 计算和I/O吞吐上能较好满足海量数据的计算需求;从应用编程接口上讲,它支持 ODBC 和 JDBC,能较好的支持常用 SQL,且移植性较好。但 GP 数据库虽然能进行动态扩容,但在扩展过程中 master 服务会停止,以创建新节点时做元数据复制,对实时业务和数据库性能会造成一定的影响,扩容成本高,所以 GP 数据库集群的架构不适合经常调整节点的数量[6]。

相比之下,Impala 的最大特点就是快速,为存储在 HDFS 和 HBase 中的数据提供了一个实时 SQL 查询接口。Impala 使用类似 GP 的 MPP(并行运算)方案,可以创建更高效的执行计划,特别在多表 join 查询的时候。Impala 通过服务进程来调度作业,把 join 的数据都加载到内存中进行操作,省去了将中间结果写入磁盘读取I/O的过程和启动 MapReduce 任务的开销,大大提高了数据访问的速度。按目前贵州省气象信息业务需求,将 Impala 集群与 GP 集群进行对比,发现 Impala 集群更适合业务场景,如表 1 所示。

表 1 Impala 数据库集群与 GP 数据库集群对比

序号	比较项	Impala	GP
1	数据更新	基于 Kudu 支持	不支持
2	SQL 支持	较好支持常用 SQL	能够较好的支持常用 SQL
3	数据汇总	汇总效率高	汇总效率高
4	查询效率	查询效率高	查询效率高
5	集群扩容	支持在线动态扩容	不支持在线动态扩容
6	扩容成本	对硬件要求低,扩容成本低	对硬件要求高,扩容成本高
7	数据安全	多副本方式,确保数据安全	多副本方式,确保数据安全

3 结束语

如今,气象信息数据的体量正进一步增长,如何对爆发式增长的数据进行科学地处理和有价值地研究已成为气象行业内研究人员重点关注的问题。气象大数据的服务与应用可加快气象部门信息共享,行业信息化进程。快速、高效的气象数据处理依赖于良好的大数据平台性能。研究通过对集群数据库方案选择进行对比,并且根据气象数据的特点和业务特性设计出合理、高效的气象大数据平台架构,满足了业务管理数据和各业务系统数据规范传送,标准化整理和存储,建立统一数据仓库的需求,为行业内各个应用系统提供规范的数据交换服务。贵州省气象大数据平台为贵州特色研究专题提供了数据支撑,同时也为气象行业大数据服务奠定了基础,提供了一个技术参照。通过设计气象大数据平台,选择有效支撑气象大数据服务的技术方案,使得贵州气象数据处理的质量得到进一步提升,为专业气象服务以及相关的软件开发提供了良好的数据基础。数据质量对大数据应用起着至关重要的作用,在未来,贵州省大数据平台还应继续对大数据质量管理的技术挑战做进一步的研究。

参考文献:

[1] Wilson L,Goh T T,Wang W Y C. Big Data Management Challenges in a Meteorological Organiza-

tion [J]. International Journal of E-Adoption, 2017,4(2):1-14.
[2] 林子雨,赖永炫,林琛,等. 云数据库研究 [J]. 软件学报,2012,23(5):1148-1166.
[3] 章国材. 气象云建设的研究与思考[J]. 气象与环境科学,2015,38(4):1-11.
[4] 华丽,陈澄. 云计算环境下气象大数据服务应用[J]. 农业与技术, 2017(20): 231-231.
[5] 邓贤峰,桑菁华. 基于大数据的智慧城市环境气候图[J]. 上海城市管理,2014(4): 33-36.
[6] 李永生,刘修伟,杨玉红. 气象大数据跨平台分析与应用技术研究[J]. 电脑知识与技术,2013, 9(31): 6943-6947.

Architecture Design of Guizhou Meteorological Big Data Platform

GUO Xi, WANG Biao, WANG Hua, JING Shi-sheng
(Department of Meteorological Information Center of Guizhou Province, Guiyang 550001, China)

Abstract: The construction of meteorological big data is one of the important contents of the meteorological information and the modernization. In recent years, with the rise of meteorological data suddenly and sharply, the existing meteorological equipment and information technology methods are difficult to meet the needs of meteorological services. There are various business systems in Guizhou Meteorological Bureau. Data scattered in departments, data collection lack of comprehensiveness and systematisms, "information isolated island" is serious and the data integration is limited to different systems and software development platforms, server utilization rate becomes low, and there sources of CPU, memory, disk space can't be used effectively, and data storage exists single point of failure. In view of the above problems, Guizhou Meteorological Information Center proposed the design of the overall framework of the meteorological big data platform, which helps to improve the accuracy, intensive, efficient and data sharing of the forecast, and make the data storage management and service achieve intensive and high efficient data sharing. This paper summarizes meteorological data collection, data storage and data processing in the construction of meteorological big data platform, and it introduces the present situation of the meteorological information system. From the improvement of the top-level design, compare the selection of the cluster database, and designed a reasonable and efficient meteorological big data platform, in order to realize the integration and sharing of meteorological big data industry.

Keywords: meteorological big data platform; Impala cluster; GreenPlum