计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):224−228 [doi: 10.15888/j.cnki.csa.006541]

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

基于 ETL 的政务云气象数据仓库构建①

许皓皓[1], 廉　亮[2], 姚浩立[1]

[1](宁波市气象网络与装备保障中心, 宁波 315012)
[2](宁波市气象服务中心, 宁波 315012)

摘　要: 针对气象网站等应用系统向地方政务云迁移过程中缺乏基础数据的现状, 从功能性, 开发成本, 灵活性方面考虑选用 ETL 工具, 基于 Kettle 软件对气象数据 ETL 流程进行建模, 使用 Quartz 开发作业调度系统实现 ETL 流程的自动化运行, 在政务云搭建 SQL Server 数据库集群, 构建了政务云气象数据仓库. 该数据仓库实现了异构环境气象数据在政务云的实时同步和存储, 为气象应用系统在政务云的全面部署提供了数据支持, 也为气象部门参与电子政务数据交换和共享打下基础.

关键词: 政务云; 气象数据仓库; 数据同步; ETL; Kettle

引用格式: 许皓皓,廉亮,姚浩立.基于 ETL 的政务云气象数据仓库构建.计算机系统应用,2018,27(9):224–228. http://www.c-s-a.org.cn/1003-3254/6541.html

Establishment of Meteorological Data Warehouse Based on ETL Tools

XU Hao-Hao[1], LIAN Liang[2], YAO Hao-Li[1]

[1](Ningbo Meteorological Network and Equipment Support Center, Ningbo 315012, China)
[2](Ningbo Meteorological Service Center, Ningbo 315012, China)

Abstract: Aiming to solve current status of data deficiencies in the process of meteorological websites and other applications migrating to local government cloud, choosing ETL tools by considering functionality, development cost, and flexibility, a job scheduling system based on Quartz framework was developed to automatize the meteorological data ETL processes which are modeled by using Kettle software, an SQL Server database cluster was built, overall established a meteorological data warehouse in government cloud. This data warehouse fulfills the purpose of multiple-source meteorological data synchronizing and storing to government cloud in real-time, provides data support for those meteorological application systems has been or will be deployed in government cloud, also lays the foundation for meteorological department involving in E-government data sharing and exchanging.

Key words: government cloud; meteorological data warehouse; data synchronization; ETL; Kettle

引言

随着电子政务这种信息化环境下的新型政务模式的不断深入发展, 采用云计算模式建设的各级地方政府政务云应运而生, 减轻了各级政府机构信息化基础设施的建设和运维成本, 也为解决信息孤岛, 实现部门间信息共享提供了新的可能, 给智慧城市 建设和电子政务转型发展提供了全新的路径和解决方案[1–3]. 气象部门应用系统在向政务云迁移部署过程中, 存在基础气象数据缺乏, 政务云网络安全策略限制导致应用系统开发模式单一等诸多问题, 制约了气象服务类应用系统的开发方式, 以及进一步在政务云部署应用的效果, 也无法满足公众对气象服务的需求, 因此在政务云上构建一套稳定可靠的气象数据仓库非常有必要性. 各地气象部门在气象数据仓库和气象电子政务建设和

① 收稿时间: 2018-02-03; 采用时间: 2018-03-07; csa 在线出版时间: 2018-08-16

应用领域开展了大量的研究. 薛胜军等[4]基于 Hadoop 建立气象信息数据仓库, 实现了海量气象数据文件的分布式存储、元数据管理以及气象数据的查询; 王红霞等[5]将数据仓库技术应用于气象数据, 建立气象数据仓库, 利用联机分析的快速数据统计和数据挖掘的自动知识发现技术, 提取知识点, 为气象服务等领域提供决策支持; 梁文生等[6]根据电子政务的特点及其在气象部门中的应用, 对电子政务系统的安全性及其实现进行探讨并提出解决方案.

政务云气象数据仓库建设需要解决异构气象数据集成、数据处理和作业调度策略自定义, 数据仓库运维和故障排查等一系列难题, 使用开源 ETL 工具的开发模式在功能性、开发成本、灵活性方面具有明显优势. 基于上述背景, 本研究基于 Kettle 对气象数据 ETL 流程进行建模, 并使用 Quartz 开源作业调度框架开发作业调度系统实现 ETL 流程的自动化运行, 构建了地方政务云气象数据仓库, 为气象网站等各类气象应用系统提供了基础气象数据, 补齐了气象应用系统大规模向政务云迁移部署的数据短板, 在合理利用电子政务资源, 减轻气象部门信息系统运维压力, 以及节能减排方面也拥有良好的实用性和经济效益.

1 需求分析和数据同步方案选择

1.1 需求分析

受限于政务云网络访问策略限制, 气象专网数据只能单向传送至政务云, 而部署在政务云的应用系统无法反向访问气象专网的数据, 因此只能在气象专网将数据二次转换成 XML 等中间格式, 再通过 FTP 等方式推送至政务云提供给应用系统访问, 这种模式严重制约了应用程序的开发方式, 给气象应用系统向政务云迁移和部署制造了巨大阻力, 气象数据融入政务数据共享服务体系参与跨部门共享也因此无法实现. 为解决这一问题提出建设政务云气象数据仓库的解决方案, 其目的是为部署在政务云的各类气象应用系统提供及时、高效的数据服务, 在建设过程中要解决异构气象数据集成、气象数据处理和同步、数据仓库载体搭建等问题, 同时提供一套高效实用的数据仓库监控和故障分析解决方案, 此外还要对数据仓库进行优化, 以保证其性能和稳定性. 详细需求描述如下:

1) 异构数据集成: 气象数据由观测和业务生产系统源源不断产生, 存储在 SQL Server、Oracle、MySQL 等多种数据源, 数据仓库建设首先要解决这些异构气象数据的集成问题, 实现异构数据源和数据仓库载体的无缝对接.

2) 数据自动化提取、处理和同步: 基于研究环境气象数据特性, 对各类气象数据进行增量数据提取, 加工处理后通过灵活的作业调度策略自动化同步至政务云数据库落地, 这是数据仓库建设需要解决的核心问题.

3) 数据仓库载体搭建: 搭建一套高可用高性能的企业级数据库系统, 支持分布式扩展和实时副本同步, 满足高吞吐量数据集中存储和读取的性能要求, 为数据仓库提供载体.

4) 数据仓库性能优化: 对数据仓库设计、ETL 流程各环节进行优化, 提升数据仓库性能和稳定性.

5) 数据仓库监控和故障分析: 支持对数据仓库中各类数据 ETL 流程运行状态进行监控, 提供完备的日志管理功能, 为故障排查提供精准化信息.

1.2 数据同步方案选择

政务云气象数据仓库建设首先要解决气象数据如何同步至政务云这一问题, 通过研究和比较 Oracle Golden Gate、SQL Server 发布订阅、Kettle ETL 工具集三种主流的数据库同步解决方案, 并且在业务环境进行了测试, 得出 Kettle ETL 工具集在功能性、开发成本、灵活性方面具有明显优势. 三种解决方案技术特点比较见表 1.

表 1 三种解决方案技术特点比较

解决方案	技术特点
Oracle Golden Gate	企业级数据库同步解决方案, 支持异构环境数据同步, 和 Oracle 数据库结合紧密, 软件套件购买成本高, 部署和学习难度较大
SQL Server 发布订阅	企业级数据库同步解决方案, 只支持 SQL Server 数据库, 提供图形化管理和监控界面, 配置使用较为方便, 但无法解决异构数据平台集成问题
Kettle ETL 工具集	开源数据库 ETL 工具, 支持 Oracle、SQL Server 等主流数据库, 提供图形化建模和测试工具, 定制开发灵活性较高, 提供监控和日志工具, 支持集群方式部署

ETL 是企业数据仓库建设中实现异构数据集成的一种技术手段, 即数据抽取 (Extract)、转换 (Transform)、装载 (Load) 的过程, 目的是将分散、零乱、标准不统一的数据整合到一起, 在数据仓库的构建中, ETL 贯穿于项目始终, 是整个数据仓库的生命线[7–10]. Kettle 是构建数据集成解决方案的一款开源 ETL 工具, 采用 Java 语言编写, 其官方正式名称是 PDI (Pentaho Data Integration), 可以运行在 Windows、Linux、Unix 等操

作系统. Kettle 提供丰富的应用对接方式和数据输出接口, 可以方便的和各类开发框架和应用系统对接, 为数据仓库建设全流程提供完备的解决方案.

2 基于 ETL 的数据仓库设计与构建

2.1 源数据分析

数据仓库构建前期, 需要对元数据环境进行详细分析, 着力分析元数据源类型、数据表类型和数量、数据更新特性、数据表结构设计和数据更新比对项等内容, 根据分析结论为下一步对数据 ETL 流程建模和作业调度策略设置提供依据. 本研究环境数据分析结论见表 2.

2.2 数据仓库架构和流程设计

基于 ETL 流程的建模和调度来实现气象数据的整合和同步, 解决了异构气象数据集成、增量数据抽取、数据清洗和转换、数据载入这些数据仓库构建各环节涉及的关键问题. 引入开源 ETL 软件 Kettle 来对各类气象数据 ETL 全流程建模, 基于 Quartz 开源作业调度框架开发作业调度系统实现 ETL 流程的自动化, 搭建 SQL Server 企业级集群作为数据仓库载体. 政务云气象数据仓库设计流程如图 1 所示.

表 2 数据环境分析

分析项	分析结论
数据源类型	生产环境共有 SQL Server 和 Oracle 两种类型数据源, 其中 SQL Server 采用单服务器方式部署, Oracle 数据库采用两台服务器搭建 RAC 集群方式部署
数据类型和数量	数据类型主要分为探测和预报 2 大类共 60 余种, 存储在 SQL Server 和 Oracle 两种数据库平台, 此外还有重要天气和基础信息类数据分散存储
数据更新特性	预报、探测类数据更新频率、更新时间、数据量差异较大, 需要对 ETL 流程正对性建模、配置不同的作业调度策略, 重要天气数据时效性要求极高
数据表结构设计	数据表均设置了主键或联合主键, 可使用 Int 类型自增量字段用于数据更新比对; 探测类数据表可通过入库时间, 或者字符型观测日期和时间等字段进行联合比对, 少数数据表需要对字段或联合字段进行字符处理后才可用于比对提取增量数据

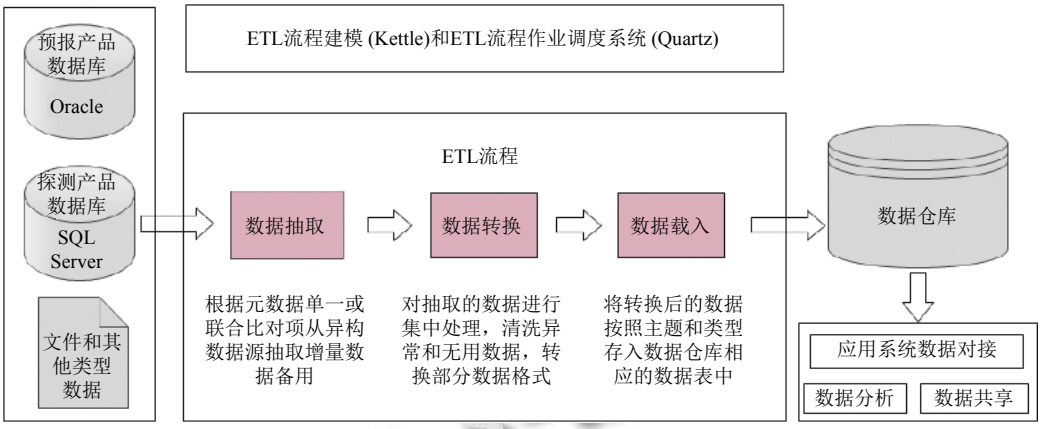

图 1 数据仓库设计流程图

2.3 气象数据 ETL 流程建模

本研究基于 Kettle 来实现 ETL 流程的建模. Kettle 支持丰富的数据输入输出数据源, 提供图形界面和可视化建模控件, 可以通过拖拽控件的方式方便地定义数据传输的拓扑, 使用 Kettle 对 ETL 流程建模流程描述如下:

1) 新建一个新的 transformation 模型, 选择存储位置并命名, 模型创建完毕后将保存为为 ktr 文件, 该文件可直接在 Kettle 运行环境执行或通过程序调用;

2) 在 Kettle 主对象树界面进行数据库连接配置, 数据库配置需填写连接类型、连接方式和连接参数信息, 测试正常后按数据库名称保存, 该步骤需要将本次 ETL 流程涉及的数据库全部配置完毕;

3) 通过拖拽方式新增 Kettle 可视化对象, 配置完毕保存为 ETL 步骤, 核心步骤主要包括: ① 获取源数据表比对项值; ② 获取数据仓库数据表比对项值; ③ 通过比对项联合查询获取增量数据; ④ 对数据进行清洗和转换操作; ⑤ 将最终数据存入数据仓库.

4) 通过 Hops 节点连接模块, 创建连接将之前保存的 ETL 关键步骤有效连接, 形成完整的 ETL 流程拓扑, 拓扑创建完毕后运行 transformation 模型, Kettle 会显示本次 ETL 过程的执行状态、执行时间、数据量和日志等各类可视化信息, 至此 ETL 模型创建完毕.

2.4 ETL 流程自动化和作业调度策略设置

通过 Kettle 对 ETL 流程建模为数据仓库建设提供了基础, 但是仍然没有解决 ETL 流程自动化运行问题. 基于 Quartz 开源作业调度框架开发数据仓库作业调度系统, 实现了各类气象数据 ETL 流程模型的自动化, 打通了数据仓库建设的最后一个环节. 数据仓库作业调度系统基于轻量级的开源 Java 开发框架 Spring MVC 开发, 系统提供灵活的作业调度规则, 可实现类 Unix 系统下 Cron 作业调度器的功能, 支持图形界面配置, 同时提供作业调度状态监控和故障分析等功能.

由于气象数据种类丰富、各类数据更新的频次和数据量不尽相同, 因此需要详细的分析和测试不同类型数据的特性, 制定相应的作业调度策略, 否则很可能会导致数据更新延迟和数据溢出等问题, 影响数据仓库更新速度和数据服务质量. 对主要气象数据的特性分析和作业调度规则设置情况见表 3.

表 3 数据特性分析和作业调度规则设置

数据类型	数据特性分析	作业调度策略设置
预报类	更新频次低、更新时间有规律、数据量不大	根据规律性更新时间设置定时运行策略, 因为数据量不大, 部分预报数据可设置多个调度策略
探测类	更新频率高(1分钟、5分钟、10分钟), 中尺度自动站等资料存量和更新数据量较大, 对同步效率和时效要求较高	单时次更新数据量大的探测数据, 采用定时运行策略, 避免对数据库运行产生较大负担; 数据量小的探测资料采用循环运行策略, 保证时效性
重要天气类	数据更新没有规律性, 数据量也不大, 但是对时效性要求非常严苛, 例如灾害天气预警数据	采用循环运行调度策略, 不考虑数据库运行时间等损耗, 保证数据更新的及时性
基础信息类	数据相对比较固定, 更新频率低, 例如站点信息数据	采用一天或一周更新一次的保守策略, 或者采用手工运行策略, 尽量减少数据库负担

3 关键技术研究

3.1 数据仓库载体搭建

从性能、稳定性、维护成本等方面考虑选择 SQL Server 数据库集群作为数据仓库的数据存储平台. 本研究搭建的 SQL Server 集群由一台域控服务器、2 台集群节点服务器、SQL Server 故障转移群集专用群集资源组组成, 其中故障转移群集专用群集资源组包含网络名称、IP 地址、数据共享盘、MSDTC 共享盘、仲裁盘等内容. 集群内服务器全部采用虚拟服务器, 操作系统按照微软官方推荐全部安装 Windows Server 2008 企业版, 因为该版操作系统大大简化了 Windows 故障转移集群的管理维护. SQL Server 集群共有两个数据库节点, 正常运行时, 只有一个节点上的 SQL Server 实例进程在运行, 此节点称为活动节点 (Active Node), 而另外一个节点则称为被动节点 (Passive Node). 集群的虚拟网络名字总是映射到当前活动节点的 IP 上, 保障数据库服务不中断. SQL Server 集群内部的状态信息会实时记载到集群日志和 Windows 事件浏览器中, 一旦集群发生了异常, 可以通过研究这些信息了解系统状态变化的全过程并针对性的处理.

SQL Server 集群是一套成熟的企业级数据库集群解决方案. 集群提供了诸如节点之间心跳检测、故障转移策略管理等特性, 可以在单台集群硬件和软件故障时将故障资源从一个集群节点转移到另一个节点, 实现数据库双活功能. SQL Server 集群支持横向扩展, 也可以通过虚拟服务器配置提升实现纵向扩展, 满足数据仓库高可用和高性能需求, 也为未来数据仓库的扩容和性能提升提供了保障.

3.2 数据仓库优化

数据仓库优化主要解决的是性能和稳定性两个问题, 并且贯穿在数据仓库的设计和运行过程中, 在此过程中对 ETL 流程 SQL 语句、ETL 流程作业调度策略、Kettle 软件运行和设置等内容和环节进行了调整和优化, 提升了数据仓库系统的稳定性和运行效率, 优化内容和优化方式描述如下:

1) SQL 语句优化: 对 ETL 流程的 SQL 语句进行优化, 比如控制联合查询数量, 避免在索引列上使用函数或计算等操作, 同时注意正确的创建和使用索引.

2) 作业调度策略优化: 对数据库存储过程、ETL 流程执行计划进行调整, 避开繁忙时间段, 防止长时间运行或重复操作造成死锁等问题.

3) Kettle 软件优化: Kettle 软件基于 Java 语言开发, 通过对软件运行方式和相关设置进行调整优化, 可以大幅提升 ETL 流程的运行效率, 主要策略包含: ① 优化 JVM 运行内存大小; ② 使用大内存方式启动 Kettle; ③ 运行缓存设置尽量大; ④ 调整抽取和载入过程中记录集合内记录数量; ⑤ 优化增量更新策略尽量缩小输入数据集大小; ⑥ 优先使用数据库连接池方式连接.

3.3 数据仓库监控和日志管理

数据仓库运行过程中, 由于气象数据具有数据源多样, 作业执行周期短、数据量大、同步规则复杂等特征, 极易发生各类故障, 因此对数据仓库 ETL 过程的监控和日志管理功能非常有必要. 本研究在设计数据仓库作业调度系统时集成了监控和日志管理功能. 监控和日志管理基于 Kettle 强大的日志输出功能开发, Kettle 提供没有日志 (Nothing)、错误日志 (Error)、最小日志 (Minimal)、基本日志 (Basic)、详细日志 (Detailed)、调试日志 (Debug)、行级日志 (Row level)7 种不同级别的日志输出, 日志记录详细程度依次递增. 在使用 Kettle 进行复杂的 ETL 操作时输出日志会非常多且杂乱, 可读性较差, 因此需要在日志管理模块开发过程中充分利用 Kettle 分级日志的功能, 不同场景匹配不同的分级日志, 同时提取常见异常信息关键字建立故障信息指标库, 通过字符匹配和文本截取等技术手段, 精确提取故障日志信息显示, 避免运维人员在故障定位过程中, 耗费的大量的时间阅读无效日志. 监控模块开发依赖日志管理模块, 提取每种数据每次 ETL 过程的运行状态信息在作业调度系统上直观的显示, 方便运维人员及时发现运行异常, 并结合日志管理模块输出的错误日志进行处置. 需要特别关注的是在数据仓库 ETL 运行过程中对 Kettle 日志输出进行监控和抓取会影响系统的整体性能, 因此选择日志输出级别应非常审慎, 并且在程序获取日志完毕后要清空日志缓冲区, 避免缓冲区堆积或溢出造成的系统性能异常.

4 应用情况

气象数据仓库的建成为各类气象应用系统向政务云的迁移和部署提供了完备的基础气象数据服务. 目前数据仓库除了服务于市级气象门户网站以外, 还为港口航运服务平台、决策气象 APP 等系统提供数据支撑, 此外市级突发事件预警信息发布平台等项目的设计和开发也基于政务云气象数据仓库开展, 未来气象数据仓库将服务于更多部署在政务云的气象服务和应用系统.

政务云气象数据仓库也为气象部门参与电子政务数据交换和共享打下基础. 由宁波市政府推动的“E 宁波”移动智慧社管信息系统已与 2017 年上半年完成了和政务云气象数据仓库的对接, 首期对接项目包含短期预报、天气预警、实况观测等资料. “E 宁波”作为宁波市政务网格化管理的统一信息工作平台, 将社会治理的每项工作都能渗透落实到网格中, 及时解决群众最关心、最直接的利益诉求. 此外依托于政务云气象数据仓库, 气象部门已经完成市大数据管理局首批政务数据归集工作, 气象数据将服务于更多社会治理领域.

5 结语

政务云气象数据仓库实现了预报、预警、探测等 60 余种气象资料在政务云的落地, 既服务于部署在政务云的气象应用系统, 也为气象部门参与政务数据交换提供条件, 具有很高的实用价值. 气象数据仓库的建设是一个持续性的工作, 根据中国气象局气象信息化行动方案, 未来可以在气象数据仓库部署标准化气象数据服务接口, 改变以数据库为中心的传统开发模式, 为各类应用系统提供标准化的对接方式. 也可以充分利用政务云已建成的 Hadoop 大数据分析与处理平台, 开展气象大数据分析工作, 研究气象大数据和其他行业大数据的关联性, 应用于防灾减灾和社会治理等领域, 让气象数据发挥更大的现实价值.

参考文献

1 汪向东. 我国电子政务的进展、现状及发展趋势. 电子政务, 2009, (7): 44–68.
2 宁家骏. 《国家电子政务“十二五”规划》之解读. 电子政务, 2012, (5): 43–49.
3 王益民. 2014 中国城市电子政务发展水平调查报告. 电子政务, 2014, (12): 2–13.
4 薛胜军, 刘寅. 基于 Hadoop 的气象信息数据仓库建立与测试. 计算机测量与控制, 2012, (4): 926–928, 932.
5 王红霞, 朱喜林, 马季兰, 等. 气象数据仓库建立及数据统计与挖掘. 太原理工大学学报, 2006, (S1): 101–103.
6 梁文生, 龚智勇, 李建勇. 气象电子政务系统的特点及安全措施. 气象与环境科学, 2007, (S1): 203–204.
7 薛蕾, 姚燕, 巢丽娟. 气象通信数据多维模型构建及 OLAP 应用初探. 气象科技, 2013, 41(4): 644–647. [doi: 10.3969/j.issn.1671-6345.2013.04.010]
8 徐俊刚, 裴莹. 数据 ETL 研究综述. 计算机科学, 2011, (4): 15–20. [doi: 10.3969/j.issn.1002-137X.2011.04.003]
9 刘峰, 蔡明高, 于波, 等. 数控机床传感器数据分析中 ETL 系统改进. 计算机系统应用, 2017, 26(9): 93–97. [doi: 10.15888/j.cnki.csa.005968]
10 张春瑞, 赵成坤, 吴川, 等. 银行管理信息系统服务器国产化过程中的 ETL 应用迁移. 计算机系统应用, 2015, 24(10): 264–270. [doi: 10.3969/j.issn.1003-3254.2015.10.045]