

引用格式:杨慧炯,韩燕丽,郭芸俊. 基于分布式数据交换的高校 DAAS 平台研究与设计[J]. 微电子学与计算机,2019,36(8): 101-104. [YANG H J, HAN Y L, GUO Y J. Research and design of university DAAS platform based on distributed data exchange [J]. Microelectronics & Computer, 2019, 36(8): 101-104.]

基于分布式数据交换的高校 DAAS 平台研究与设计

杨慧炯,韩燕丽,郭芸俊

(太原工业学院 计算机工程系,山西 太原 030008)

摘 要: 文章在分析高校信息化发展及数字校园、智慧校园建设过程中数据共享和交换技术存在的主要问题基础上,结合目前高校信息化过程中教学、科研及管理实际业务需求,提出了以分布式数据服务、分布式数据交换和集中任务调度为核心的数据共享和交换 DAAS 平台整体架构,并对其关键技术进行了详细论述. 通过该平台可以较好地解决目前高校集中式数据共享交换造成负载压力急增,服务效率下降的问题.

关键词: 分布式数据服务;分布式数据交换;集中任务调度;数字校园

中图分类号: TP315

文献标识码: A

文章编号: 1000-7180(2019)08-0101-04

Research and design of university DAAS platform based on distributed data exchange

YANG Hui-jiong, HAN Yan-li, GUO Yun-jun

(Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan 030008, China)

Abstract: Based on the analysis of the main problems of data sharing and exchanging technology in the development of university informationization and the construction of digital campus and smart campus, the overall structure of DAAS platform for data sharing and exchanging is proposed, combining the actual business needs of teaching, research and management in the process of current university informationization. The platform is centering on distributed data services, distributed data exchanging and centralized task scheduling. And then, the key technologies of the platform are discussed in detail. Through the platform, the problem of rapid load pressure and low service efficiency caused by centralized data sharing and exchanging in universities can be better solved.

Key words: distributed data services; distributed data exchange; centralized task scheduling; digital campus

1 引言

随着高校信息化发展以及近年来数字校园向智慧校园的发展,高校信息化由传统以部门业务应用为核心转向以用户业务流程和业务服务为核心,基于 SOA 的建设理念,统一数据服务总线的数据共享交换模式逐步得到了较好的应用^[1-5]. 但是随着业务需求的不断增加,业务服务对象由从前的各单位部门变为全校范围内的师生. 更多的

应用、更多的用户导致业务共享与交换的压力呈指数上升,对数据服务架构性能、稳定性、扩展性、可控性提出了前所未有的要求,而数据服务共享由于缺乏统一管控从而导致数据多头采集、重复采集,服务通讯性能和稳定性缺乏保障等问题日益突出^[6]. 因而有必要通过对数据服务架构进行优化,打造一个高效稳定而统一可控的数据共享交换生态平台.

2 系统整体架构

2.1 需解决的主要问题

随着高校在信息化基础设施、信息系统、教育技术以及用户服务方面取得的长足进步,各种应用的逐步深入,管理精细化程度的逐步加深,信息化服务过程中暴露出各类问题的核心主要体现在数据报文和数据文件的共享与交换两个方面.

(1)数据报文主要体现在:

- 1)缺乏统一管理
- 2)接口服务负载压力大
- 3)交换性能缺乏监控

(2)数据文件主要体现在:

- 1)缺乏有效共享
- 2)存在安全隐患
- 3)数据质量不高

2.2 系统架构

针对上述问题,需要通过设计统一开放的数据共享和交换管理平台,以统一的数据模型、数据质量标准 and 接口规范,实现数据的自动提取、转换、发送、校验以及审核、监控. 同时在通过平台架构的整体优化设计,达到不同业务数据交换的高效、稳定和安全.

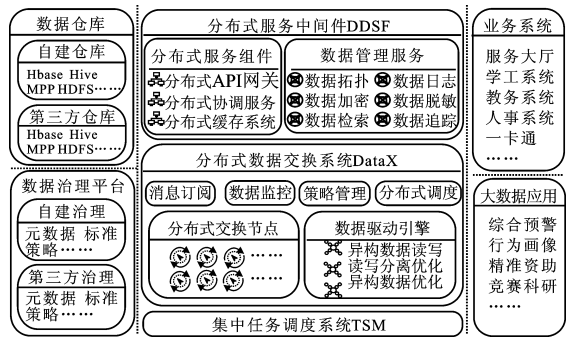


图 1 分布式数据共享交换 DAAS 平台整体架构

如图 1 所示,分布式高校数据共享交换 DAAS 平台由分布式数据交换中间件 DDSF、分布式数据交换系统 DataX、集中任务调度系统 TSM 以及数据仓库构成.

分布式服务中间件 DDSF 基于分布式微服务框架进行设计,结合教育行业数据服务的特点,采用业内领先的服务治理框架 Dubbo,对校内 RPC 通讯网络进行统一梳理与管控,打造稳定的对内服务网关.

分布式交换系统 DataX 结合先进的分布式交换框架进行设计,可以完成多种同构、异构数据之间

的数据提取、转换、过滤,以及传输等操作,以实现多数据源集成的业务需求. 通过 DataX,同时可实现数据调度处理的动态部署和可视化管理.

集中任务调度系统 TSM 是贯穿服务与交换的任务管控系统,是整个数据共享交换去中心化的管理核心,同时也是对数据报文与数据文件的交换进行统一有效监控的核心. 通过将分布式数据交换系统进行任务统一调度,为用户的系统管理提供有力保障.

数据仓库主要包含数据仓储的建设及数据检索及管理两大内容,以 Hadoop 数据仓库或者 MPP 数据仓库为存储工具构建海量可扩展的存储仓库存储介质,提供分布式、高并发性的海量数据存储及访问,并提供数据的集中管理及检索.

3 关键技术

3.1 数据报文传输

通过分布式服务节点将所有数据的调度申请与处理统一进行管控与承载,并通过分布式服务系统为各项业务的数据调用需求提供统一的服务接口,数据使用者在取得数据授权前提下,调用数据接口获取数据服务,系统记录接口调用记录,形成接口调用分析表.

数据报文传输可以提供多种“Request-Response”模式以及多个线程模型的序列化信息交换. 支持 Java RMI、Dubbo、hessian、http、WebService (含 Restful)、thrift、redis 等多种 RPC 协议,并以 NIO 模式进行抽象封装.

3.2 服务治理

采用业内主流 RPC 服务治理框架,包含注册中心、集群容错、负载均衡、服务监控及调度等.

3.2.1 服务交互及监控

如图 2 所示,在一个分布式服务系统 DDSF 中,服务提供者 Provider 向注册中心 Registry 注册服务;服务消费者通过 Registry 订阅自己需要的服务. 当注册中心 Registry 接收到 Consumer 发送的订阅请求后,向 Consumer 返回 Provider 注册的地址列表. Consumer 通过负载均衡算法,从 Provider 返回的地址列表中选择一个服务地址进行调用.

服务的整个注册和调用过程,通过监控中心 Monitor 进行监控,监控中心可以将数据的访问、数据种类、数据量、更新情况、数据活跃度等进行统计,并以报表和图表的形式展示. 同时还可根据数据服务目录,以图表等可视化方式,直观统计各个服务的

调用次数、调用时段分布、调用应用分布、调用用户分布等调用情况。

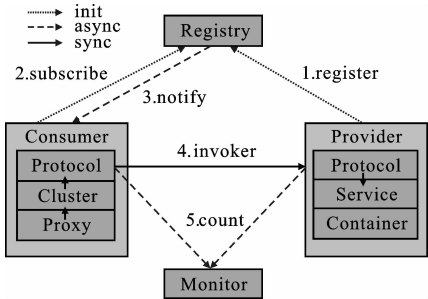


图 2 数据服务交互通讯流程

3. 2. 2 注册中心机制

如图 3 所示,采用 Apache Hadoop 的 Zookeeper 作为服务注册中心,通过监控中心 Monitor 对服务提供者 Provider 和服务消费者 Consumer 以及服务的注册和调用进行监控管理. 为了使得服务消费方能够动态地查找到服务提供方,注册中心采用透明地址的目录服务方式。

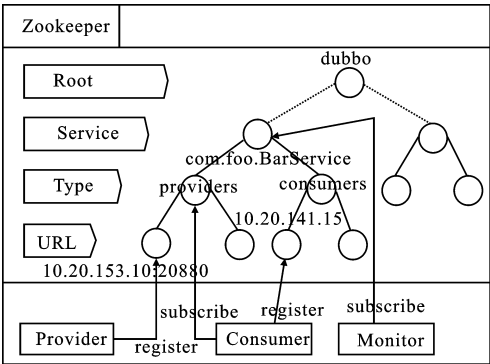


图 3 Zookeeper 注册中心机制

当需要注册一个新的服务时,服务提供方 Provider 向服务目录(Service)写入自己的 URL 地址,其类型为 providers. 当服务消费方 Consumer 需要调用某个服务时,则订阅服务目录(Service)下这个服务的地址(即该服务提供方注册时写入的 URL 地址),同时向服务目录(Service)写入自己的 URL 地址,其类型为 consumers. 在整个目录服务过程中,监控中心 Monitor 则订阅服务目录(Service)下包括服务提供方和消费方的所有 URL 地址,并对所有服务的注册和消费进行监管。

3. 2. 3 集群容错机制

集群容错机制如图 4 所示. 其中 Invoker 是服务提供方 Provider 的一个可调用 Service 抽象,Provider 的地址和服务调用接口信息全部封装在 Invoker 中。

多个服务提供方 Provider 的抽象(即多个 Invoker)以 List 方式封装在 Directory 中,并且随着 Invoker 的动态变化而变化,当注册中心的 Provider 发生推送变更时,相关 Invoker 的值及 Directory 中 Invoker 列表的值都会发生动态变化。

通过 Cluster,将 Directory 中的多个 Invoker 按照一定的容错逻辑封装成一个虚拟 Invoker,并对上层服务开放. Cluster 可采用包括 Failover、Failfast、Failsafe 等多种容错机制,当调用失败是,则按照某个容错机制调用另一个虚拟 Invoker。

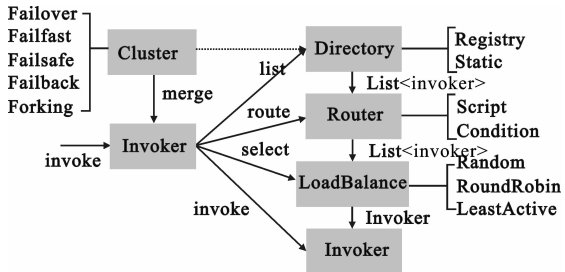


图 4 集群容错机制

服务调用时,首先通过 Router,按照一定的路由规则从 Directory 中选出 Invoker 子集,接下来 LoadBalance 按照负载均衡算法选出一个具体的 Invoker,通过 Cluster 封装完成服务调用. 当调用失败时,则通过容错机制重新进行选择。

3. 3 分布式数据交换

分布式数据交换系统 DataX 支持的数据库包括各主流数据库(如 Oracle、SQLServer、MySQL 等)、地理空间数据(如卫星影像、矢量数据)、常规文件(word、excel、pdf)等各种格式,可以实时或定时源系统采集数据,在数据从原始库流向标准库的过程中,监控数据量、数据内容、数据唯一性、正确性的变化情况,形成数据质量报告和数据质量告警。

分布式数据交换系统 DataX 整体技术架构及数据交换流程如图 5 所示,采用 Framework+plugin 的架构,通过 Reader 和 Writer 插件,以分布式线程模式完成数据同步作业任务。

DataX 接收到一个数据同步作业任务之后,启动一个独立的进程来完成这个数据同步任务. 每个数据同步任务称为一个 Job,一个 Job 启动之后,根据不同数据源将被切分为多个子任务 Task 来并发执行. 为了高效地并发执行每一个 Task,Job 模块通过 Scheduler 模块,按照并发量将子任务 Task 重新组合成任务组 TaskGroup,由 TaskGroup 负责启动其组内的每一个 Task,按照 Reader→Channel→

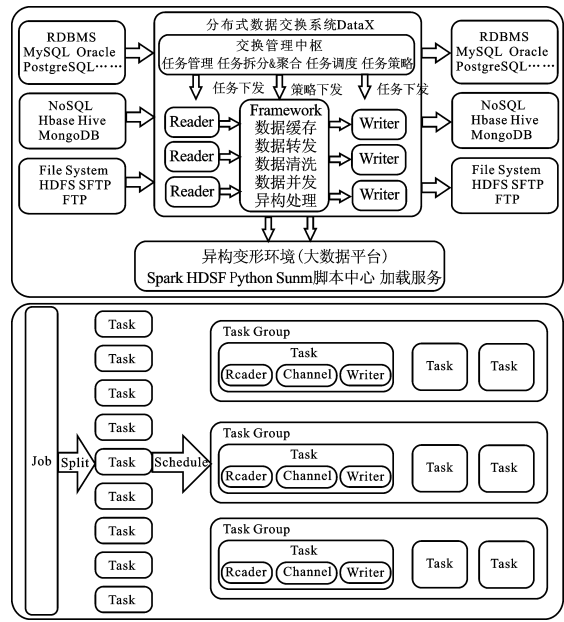


图 5 分布式数据交换系统技术架构及数据交换流程

Writer 的方式完成数据同步交换作业任务. Job 模块则监控并等待多个 TaskGroup 模块的任务完成之后退出.

3.4 异构文件传输

在 DataX 系统的数据交换中, 往往会出现针对结构化文件的异构处理与非结构化文件的异构处理, DataX 系统分别进行相关优化与解析.

3.4.1 结构化文件的异构处理

对于像 mysql 等结构化文件, 在 DataX 系统 Framework+plugin 架构的 Reader 端将转义成统一格式的数据流, 传到 Framework 中进行聚合、清洗, 并以统一格式的数据流进行输出. 在 Writer 端看来 Framework 是输出了统一格式的数据流, 可以将数据写到任意格式的数据库中(如 hbase), 并对数据格式进行优化.

3.4.2 非结构化文件的异构处理

对于非结构化文件的处理, 采用对非结构化文件进行结构化变形的方式. 目前主流的非结构化数据大多数为 Word 文档、Txt 文档、Syslog 日志文档等. 以上非结构化数据又可分为两种不同的类型, 一种为有规律可循的, 如 Syslog 日志文档, 每条纪录都会按照一定的规律进行划分, 每个字段的含义相对固定. 对于这种类型的数据, 在项目实施前期数据传输前, 一般使用大数据异构变形环境对数据进行

结构化处理, 对特征字典进行识别, 并做成结构化索引进行传输. 当数据结构稳定后, 在 Reader 或者 Writer 定制相关结构化变形插件, 方便今后重复调用.

另外一种非结构化文件为没有固定规律的文件, 如 word、txt 等, 针对该类文件在 Reader 读取前, 需要使用专门的语义分析引擎来界定索引制定的规则, 然后将规则交给检索引擎, 让其建立专门的索引列表, 转成结构化文件进行传输, 当解析率达到预期时, 固化解析脚本方便 Reader 进行调用.

4 结束语

本文根据高校信息系统种类繁多, 业务数据类型及业务流程耦合性强的特点, 采用分布式数据共享和交换技术, 在基于 SOA 理念的基础上, 设计了高校数据共享和交换 DAAS 平台整体架构. 基于分布式数据共享交换服务的理念, 是目前智慧校园建设中高校数据治理及数据共享交换较为先进可行的技术手段.

参考文献:

[1] PAUL LOSCH. The digital campus primer[EB]. Sun Microsystems Inc, 2003.

[2] ZHAO J, WANG X F, TANG P H. The Design of Digital Campus Shared Data Center Based on ETL[C]. International Conference on Education Technology and Information System (ICETIS 2013). 2013.

[3] 王威, 蒋东兴, 刘启新. 支持信息集成的校园信息门户的研究与设计[J]. 计算机工程与设计. 2006, 27(20): 3741-3744.

[4] 朱震, 姚奇富. 基于 SOA 的数字校园解决方案[J]. 计算机工程. 2009, 35(7): 267-269.

[5] 王宇, 吴炜鑫, 王兴伟. “互联网+”下高校信息化建设模式的探索与研究[J]. 计算机应用与软件. 2016, 33(11): 41-45.

[6] 蒋东兴, 付小龙, 袁芳, 等. 大数据背景下的高校智慧校园建设探讨[J]. 华东师范大学学报(自然科学版), 2015(3): 56-57.

作者简介:

杨慧炯 男, (1972-), 硕士, 副教授. 研究方向为数据可视化、高校信息化. E-mail: yanghj@tit.edu.cn.