高校智慧校园建设中数据同步共享机制的研究与实现

周伟　杜伟

（常州信息职业技术学院　江苏常州　213164）

摘要：随着新一代信息化技术与智慧校园建设不断融合，国内大部分高校初步完成了教学、科研、财务、学工、门户、一卡通等业务管理系统的建设。同时随着业务系统数据的不断积累，业务系统间数据的同步和共享成为智慧校园建设中亟待解决的问题。本文分析了高校数据现状和需求，结合阿里云私有云平台，使用数梦工场数据治理工具，提出了一套高校数据同步共享机制，实现了高校业务系统之间数据的同步与共享。

关键词：数据同步；数据共享；云平台；全量同步；实时同步

随着数字化校园和智慧校园不断建设，同时高校信息化环境中累积的数据也在逐渐增加，形成了一个较全面的大数据环境。但如何对高校大数据进行有效的共享以及交换管理，是如今高校服务体系建设所不能忽视的问题。

一、研究现状

近年高校智慧校园建设持续展开，初步建成了以教学、科研、学工、财务、图书馆、一卡通、OA、门户以及各类微应用为主体的信息系统，支撑了学校各业务部门的业务运转。但是学校信息化普遍还存在如下几个问题：

➢数据不一致

由于各系统建设时期不一致、厂家不一致、技术不一致等原因，很多系统各自维护了学校公共基础数据，例如：教师数据、学生数据、组织架构数据等。这给后期数据维护带来极大的工作量，且很难确保数据准确性。

➢数据强耦合

智慧校园前期建设过程中，涉及到系统间数据对接，多为业务系统之间数据的强耦合对接，例如：人事系统提供教师基本信息接口，各业务对接数据。随着同步业务的增加，同步关系难以梳理，同步过程难以监控，同步设置难以配置等情况。

➢数据开发可配置不高

前期得数据开发都为厂家定制开发，数据结构固定，无法满足全部应用场景，亟需可视化、可操作性的数据开发平台。

综上所述亟需建设一个的数据中心平台来解决各业务系统间数据同步共享问题。

二、研究内容

本文旨在通过顶层设计、数据规划、业务梳理，通过搭建数据中心，有序改造现有业务系统的对接方式，规范数据流向，形成“一数一源、同源共享”的数据中心架构，彻底解决数据同步共享问题，为教学和管理的智能化提供标准的数据支撑。

基于上述研究内容，本文提出如下图 1 分布式云计算大数据平台体系架构，

图 1　分布式云计算大数据平台体系架构图

云计算层，基于现有计算、存储和网络资源所建设的分布式云计算大数据支撑平台，提供弹性计算服务、开放存储服务、备份服务、数据处理、中间件服务、虚拟网络服务和大数据服务，实现资源的集中化、规模化管理，实现对各类异构软硬件基础资源的兼容和动态流转调度，并将信息汇集、资源共享。

中台层，包含业务中台和数据中台。通过业务中台实现对各类系统的基础应用支撑，包括应用的统一管理，统一身份认证、统一消息网关、任务网关、支付网关、生物识别等；通过数据中台建设，实现对数据管理、数据集成、数据开放和数据应用的基本系统支撑和数据标准建设落地。

三、研究方案

3.1 分布式大数据平台

分布式云和大数据操作系统是在数据中心的大规模 Linux 集群之上构建的一套综合性的软件系统，将多台服务器联成一台“超级计算机”，并且将这台超级计算机的存储资源和计算资源，以服务的方式支撑用户或者应用系统访问。分布式云操作系统为上层的云服务提供存储、计算和调度等方面的底层支持，主要模块包括协调服务、远程过程调用、安全管理、资源管理、分布式文件系统、任务调度、集群部署和集群监控模块，为打造云计算平台和大数据计算提供支撑。

图 2　分布式云计算大数据平台体系架构图

本文采用阿里云飞天私有化平台（下文简称小飞天）做为云计算大数据平台，小飞天是阿里云自主研发、进行私有化部署的通用计算操作系统。它可以将机房基础服务器连成一台超级计算机，以在线公共服务的方式提供计算能力。小飞天内核跑在数据中心里面，它负责统一管理数据中心内的通用服务器集群，调度集群的计算、存储资源，支撑分布式应用的部署和执行，并自动进行故障恢复和数据冗余。

3.2 数据中心平台

校本数据中心平台是数据采集、数据治理、共享交换以及大数据分析的核心产品和技术。本文采用数梦工厂平台做为校本数据中心平台，主要包括如下部分。

3.2.1 数据采集

具备多种格式数据采集获取能力，能够对接数据库、XML、日志、文件等各种数据源，能够支持对数据进行批量/同步/实时/流式持续的采集，并能够对数据采集进行细粒度多周期的调度、更新和管理，能够监控和管理数据采集流程。针对海量大数据，必须能够支持超大带宽的、高并发、多任务的高速数据采集和加载。采用 Databridge 和 DataGate 构建数据采集的 ETL 环境，Databridge 提供可视化的交互配置、配置管理、数据采集任务依赖管理等功能。需要具备的数据采集整合平台应包括如下能力：

多数据源支持：支持所有主要数据存储系统的全量快照和增量流式的数据传输、同步、交换。并支持多种关系数据库和 NOSQL 的数据导入。主要支持数据源包括：关系型数据库，支持 Oracle、MySQL、SQL Server、DB2 等传统关系型数据；分布式存储，支持大数据平台所配置的所有分布式存储系统；非结构化数据，支持结构化的平面文件（csv/txt 等），支持非结构化的文件（文档/图像/音频/视频等）采集和传输，支持半结构化文件（网站日志/XML/JSON 等）的采集和同步。API 接口，支持从 HTTP、FTP、QUEUE 等接口获取数据源。

弹性伸缩的传输通道：可动态分配、释放同步的传输单元通道，按需调控资源、按需使用。

大规模并行数据采集：根据用户流控需要自动启动多线程乃至于多进程并行传输海量数据，强大的传输引擎支持无限扩展的吞吐能力。全新的分布式模型，吞吐量支持水平扩展，能够提供 GB/TB 级数据吞吐能力。

可靠故障处理和恢复：采用高容错和异常处理的架构，提供故障智能检测、自动传输恢复，屏蔽不可靠的异常因素，保证数据传输的稳定性和健壮性。支持各种数据类型的转换；能精确识别脏数据，进行过滤、采集、展示，提供可靠的脏数据处理，准确把控数据质量；提供作业全链路的流量、数据量、脏数据探测和运行时汇报。

3.2.2 数据标准化

校本数据中心平台必须支持数据标准化的全过程，按照统一的数据格式进行标准化，包含以下内容：标准文档库的管理，包括国标、部标、行标、企标的数据元录入、查询、编辑、状态管理；标准文档收录、查阅、状态管理；限定词、同义词、术语等信息库管理，包括收录、查阅、关联显示、状态管理；数据集成关联；标准化库表建设；基于数据标准的质量稽核；可基于标准数据元关联显示所有相关的数据资产。平台采用数据治理工具 DataRiver，实现数据的清洗、元数据、数据标准化、血缘关系、数据地图、数据质量等数据全生命周期的实验。

3.2.3 数据质量管理

大数据运用的前提是高质量的数据，而大数据的规模庞大，更需要数据质量管理工具进行严格管控。根据预设的规则来检测数据中的质量问题，检测规则可自主配置，系统提供默认的规则模板，用户也可以自主编写规则表达式。数据质量监控与调度系统强耦合，发现脏数据可实现事中拦截，避免错误的数据流入下游应用。

数据发生变化的时候，则会触发数据质量的校验逻辑，对数据进行校验，帮助用户避免脏数据的产生和质量不高的数据对整体数据的污染，同时需要保留所有规则的历史检验结果，以方便用户对数据的质量进行分析和定级。

数据质量需要提供配置规则、按照各种粒度查看历史校验结果和数据质量报警等等能力。当检测规则校验不通过时，可通过在线展示、邮件、短信、手机等途径发出告警。检测规则的校验结果被永久性保留下来，以便日后分析和规则优化。系统也应提供页面展示报告，针对单个节点或整体项目的数据质量进行统计分析。支持图表格式，支持查询。

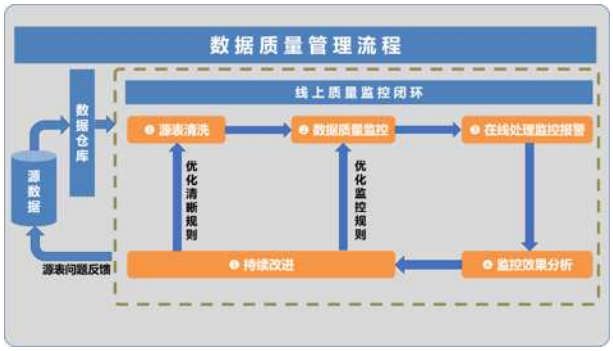

图 3　数据质量管理流程图

数据质量监控能力应与调度系统强耦合，在调度系统执行完成一个节点后，立即启动针对该节点的数据质量监控，监控规则支持拦截模式/非拦截模式，在拦截模式下，一旦检测规则校验不通过，则调度任务状态被置成失败，从而避免错误的数据流入下游应用。

3.2.4 数据共享交换

数据共享交换是目前大数据应用的必要条件。本平台采用一套 DataMall 产品构建，包括“数据资源目录子系统”和“数据资源交换子系统”，考虑到实际应用复杂的数据交换需求，总结出四种不同的交换模式：直接交换，共享交换，安全交换和数据 API 服务。通过 4 种数据交换模式覆盖各种场景的数据交换。

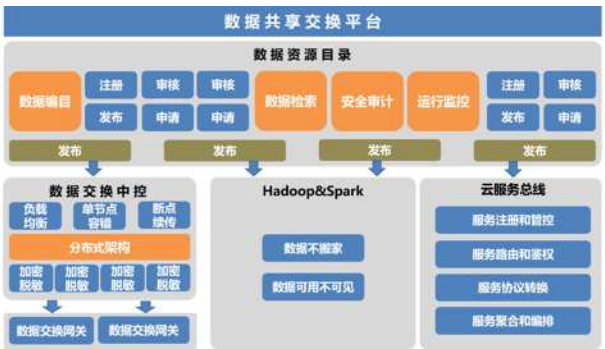

图 4　数据共享交换平台体系图

四、总结

本文通过结合云计算等技术，实现信息基础设施的资源虚拟化，计算能力共享和动态分配；通过标准 ETL 工具，实现数据同步的可视化配置和同步任务实施，通过制定数据标准和接口规范，汇聚、清洗全校应用数据资源；再次基础上最终实现数据的多种方式交换。

课题项目：本文为常州信息职业技术学院双高建设研究课题项目《校级数据标准的构建和应用》（项目编号：CCITSG202008）课题成果。

作者简介：周伟（1986.09～），男（汉族），江苏省南通市人，硕士研究生，工程师，常州信息职业技术学院信息中心软件科科员，研究方向：移动互联应用技术。

杜伟（1982.05～），男（汉族），江苏省常州市人，硕士研究生，高级工程师，常州信息职业技术学院信息中心软件科科长，研究方向：智慧校园、计算机应用技术。