



分布式数据质量管理体系在电力企业的实践和应用

李远宁, 刘森, 张诗军, 陈丰, 王志英

(中国南方电网有限责任公司信息部, 广东 广州 510623)

摘要: 随着企业信息化水平和企业精细化管理要求的不断提高, 企业对数据管理的需求也随之增强, 如何提高企业数据质量更是需要重点解决的问题。针对电力企业数据质量管理面临的挑战, 创新提出了分布式数据质量管理解决方案。针对集中式数据质量系统的性能瓶颈, 在研究数据质量系统特点并借鉴国内外对大数据的解决方案后, 提出了基于 Hadoop 分布式处理框架的解决方案。利用 Hadoop 集群, 可以把缺陷数据从 Oracle 中抽离, 分散存储在集群里多台服务器上, 以有效提高磁盘 I/O 性能和数据分析性能。

关键词: 数据质量管理; 分布式; Hadoop

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016104

Practice and application of distributed data quality management system in power enterprise

LI Yuanning, LIU Sen, ZHANG Shijun, CHEN Feng, WANG Zhiying

Information Department of China Southern Power Grid Co., Ltd., Guangzhou 510623, China

Abstract: As the improvement of the enterprise's informationalization level and the increasing management requirement of enterprise refinement, the demand of data management of enterprise is becoming greater and greater, how to improve the data quality of the enterprise is the key problem needed to be solved. Aiming at the challenges of data quality management that the power enterprise faces, some solutions for distributed data quality management were proposed. After researching the system features of data quality, some foreign and domestic cases of big data were analyzed as reference, and a solution based on Hadoop distributed processing framework was given to solve the performance bottleneck of centralized data quality system. Hadoop clustering could dissociate defect data from Oracle and the data would be stored separately on multiple servers of the clustering, which could improve the I/O performance and data analysis performance of the magnetic disk effectively.

Key words: data quality management, distributed, Hadoop

1 引言

随着企业信息化水平和精细化管理要求的不断提高, 数据已成为企业的重要资源和核心资产, 深刻影响企业的业务管理模式。企业数据质量水平, 将关系到企业数据化

管理、决策的效率和成效。

为促进数据质量水平的提升, 电力行业开展了数据质量管理工作, 并完成了数据质量管理系统的建设与推广, 生产、营销、财务、人力资源等业务域的数据质量得到了显著提升, 但也面临如下一些挑战:



- 涉及范围广,涉及各分省公司的营销、生产、人力资源、财务等业务领域;
- 校验规则繁多,各业务域都有相应的数据质量校验规则,涵盖非空、长度、格式、一致性、准确性等类型的规则,共计4 000多条;
- 数据量大,经统计,全网月增量数据为8亿笔,问题数据为3 000万笔;
- 数据校验耗时长,以营销域为例,当校验数据量达到10亿笔或以上规模时,原基于集中式数据存储和计算架构的数据质量管理系统由于数据读写和数据统计操作瓶颈,完成数据质量校验和问题分析需耗费70 h以上的系统运行时间;
- 需要在短时间内执行大量的校验规则,对磁盘读写性能、CPU和内存性能要求较高。

2 传统的数据质量管理模式

传统集中式数据质量管理系统可以规范化管理校验规则、调度规则的执行时间、统一管理数据质量报告,大大提高了数据质量校验的效率,并规范了数据质量的管理。数据量在千万级以内时,可以稳定高效地完成数据质量校验。

集中式数据质量管理系统(以下简称集中式系统)主要包括两大模块,如图1所示。

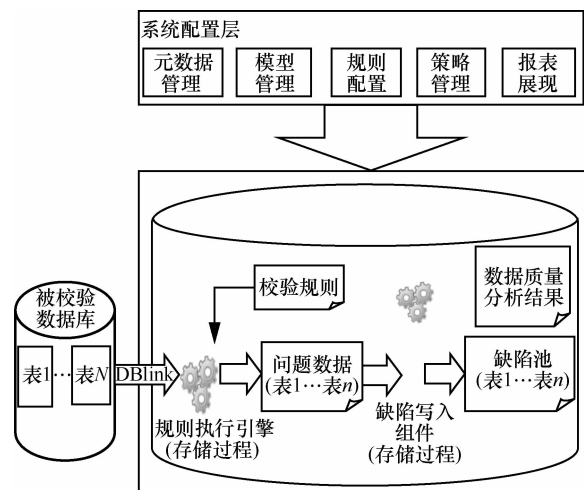


图1 集中式数据质量管理系统架构

(1) 系统配置层

用于配置校验规则、执行策略、元数据等信息,查看数据质量分析结果。

(2) 系统执行层

负责规则的执行和问题数据的存储。质量校验引擎采

用 Oracle 存储过程实现,通过 database link(DBlink)链接到“被校验库”执行预先配置好的校验规则,统计并抽取查询出来的问题数据。问题数据存储在集中式系统数据库里进行进一步的过滤、合并和分析,最终生成数据质量报告。

集中式系统使用关系型数据库存储和分析数据,当面对海量数据时数据库服务器的性能往往会成为瓶颈。只能通过提高 CPU 处理速度和磁盘 I/O 速度来提高数据库性能,但这意味着系统成本的增加,并且随着数据的不断增长,数据库服务器很快又会成为性能瓶颈。图2展示了不使用索引和使用索引时,SQL 执行时间随着数据量增长的变化情况。可以看出,数据量在500 GB 时已出现瓶颈。

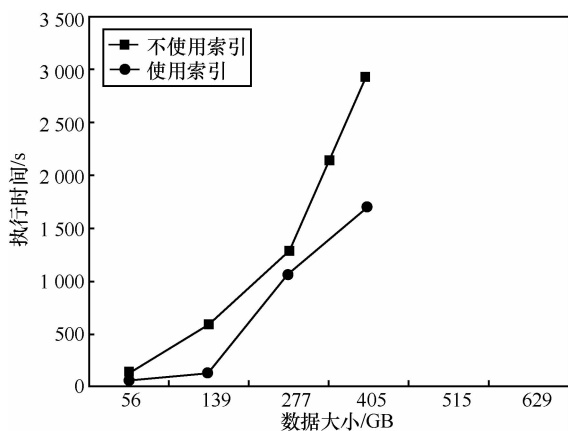


图2 SQL 查询性能

由此可见,传统的集中式系统仍存在不足,需要引入新的解决方案,以提升企业在数据质量管理方面的能力。

3 数据质量管理的实践

3.1 设计思路

针对集中式数据质量系统的性能瓶颈,本文在研究数据质量系统特点并借鉴国内外对大数据的解决方案后,提出了基于 Hadoop 分布式处理框架的解决方案。利用 Hadoop 集群,可以把缺陷数据从 Oracle 中抽离,分散存储在集群里的多台服务器上,从而可以有效提高磁盘 I/O 性能 and 数据分析性能。该方案具有以下优点:

- 基于 x86 服务器,可使用较廉价的服务器组建集群;
- 具有良好的可扩展性,当业务增长,需要处理更多的数据时,可以水平扩展和增加更多的节点;
- 集群的每个节点都可用于计算和存储数据,可应对 PB 级别数据的存储和分析。

3.2 分布式数据质量管理系统

分布式数据质量系统采用 J2EE 架构开发,可逻辑划分为用户交互层、数据处理层和数据存储层,如图 3 所示。

3.2.1 用户交互层

用户交互层即用户界面,通过界面可以进行系统配置、规则管理、报告导出等操作。主要功能包括以下几方面。

- 元数据管理:管理被校验库的元数据,包括表名、字段名、字段类型等信息。
- 模型管理:管理被校验库表之间的关联关系。
- 规则配置:管理质量校验规则,包括规则名称、规则描述以及校验脚本。
- 策略管理:管理校验规则的执行时间和执行参数。
- 报表管理:管理数据质量报告,包括生成、上报、查看等功能。
- 平台管理:管理组织机构、用户信息、系统日志等。

用户交互层是用户与系统交互的界面,在进行系统设计时充分考虑了界面的可用性和便利性,结合 AJAX 技术降低系统的响应时间,提升了用户使用体验;为简化校验

规则的配置工作,系统可根据元数据自动生成校验规则的功能,并采用 AJAX 技术预加载元数据,提高用户操作流畅度,大大提高规则配置的效率;还提供规则参数的功能。利用该功能可以把规则中相似的部分抽取出来,配置成若干个规则参数,因此可以减少规则的数量,降低规则变更的响应时间。

3.2.2 数据处理层

数据处理层是整个系统的核心,负责规则执行、缺陷数据查询分析等任务,主要功能包括以下几方面。

- 执行策略任务调度:负责控制执行策略的任务调度,按周期定时启动执行策略。
- 缺陷数据分析:负责缺陷明细的分析,并生成数据质量报告。
- 规则执行引擎:负责执行规则脚本,记录执行日志。
- 缺陷明细查询和导出:为用户交互层提供从 Hadoop 集群里查询数据的接口。

作为数据处理层的两大核心组件之一,规则执行引擎是数据质量管理系统的核心所在,直接影响了校验的效

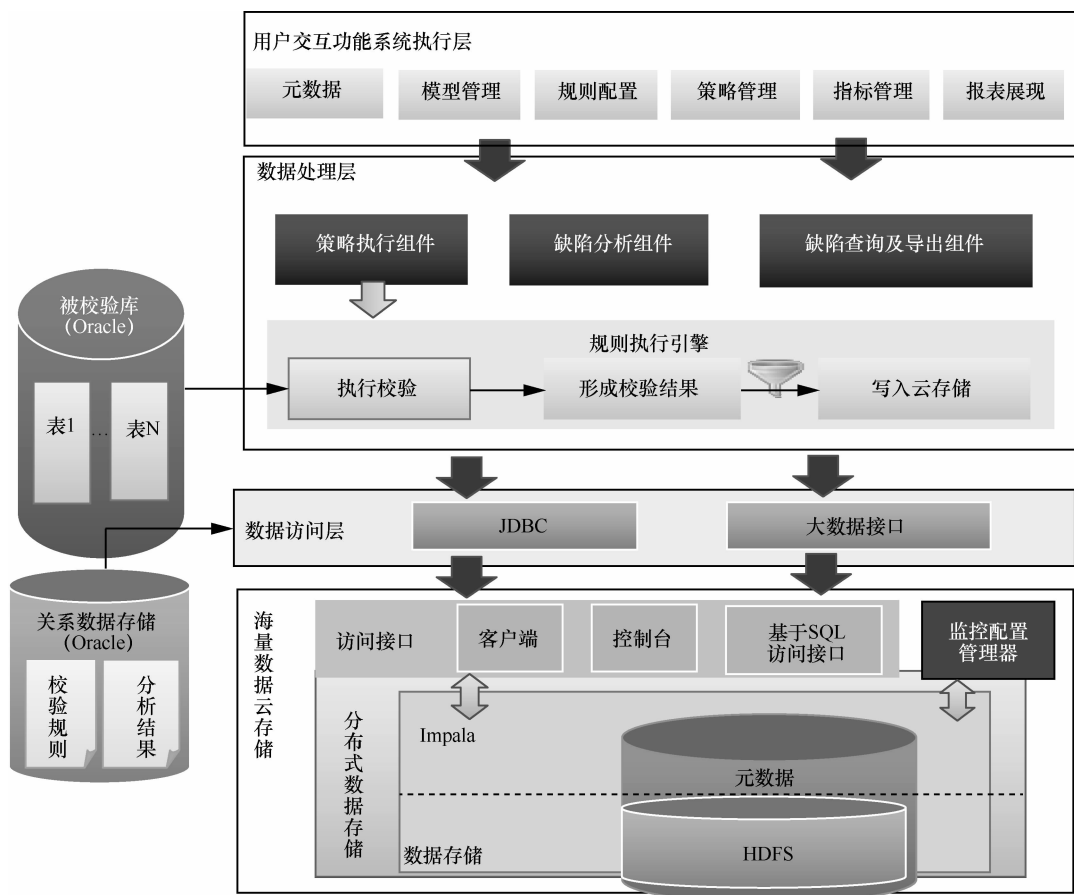


图3 分布式数据质量技术架构



率。在设计规则执行引擎时主要考虑以下几点因素。

性能是首先要考虑的问题。本系统的规则执行引擎采用开源 ETL 工具 Kettle 作为缺陷数据抽取引擎, Kettle 可以把一个表的数据“切片”进行多线程分段抽取, 因此可以比较高效地把问题数据抽取到质量管理体系里。另外, 执行引擎还采用缓存机制, 执行规则时会优先从缓存中获取结果以提高执行效率。

其次, 稳定性是另一个重要因素。在数据质量管理体系里面, 执行引擎里包含了大量的容错处理机制。例如, 在规则执行失败后, 执行引擎可以判断失败的原因并决定是否需要重试, 如果遇到不可恢复的异常, 则直接把规则标记为失败并继续执行下一条规则。

最后, 还要考虑对数据源的适应性。为了扩大数据质量管理体系的使用范围, 执行引擎需要兼容多种主流的数据库。目前已兼容 MySQL、SQL Server 和 Oracle, 日后会兼容更多类型的数据库。

缺陷数据分析引擎是数据处理层的另一个核心组件。缺陷明细数据采集到 Hadoop 后, 该引擎对数据进行分析, 生成并汇总结果, 最终生成数据质量报告。为了加快数据分析的效率, 系统把分析任务细分成多个更小规模的任务, 当所有任务完成后再汇总生成最终的结果。

3.2.3 数据存储层

数据存储层由统一数据访问接口以及数据存储(含系统数据和缺陷明细数据)组成。其中, 系统数据库采用传统的关系型数据库 Oracle, 用于存放系统配置、执行日志、规则信息和缺陷数据分析结果; 而缺陷数据明细存储采用 Hadoop 分布式处理框架。

(1) 统一数据访问接口

系统使用 Hadoop 集群存储缺陷明细数据, 为了能简单、快捷地分析数据, 基于 Impala 框架封装了通用数据访问接口, 兼容 Oracle 和分布式架构下的数据访问, 能在 Hadoop 集群上运行本地 SQL, 可以为存储在 HDFS 或 HBase 中的 Hadoop 数据提供快速、交互式的 SQL 查询, 查询效率比基于 MapReduce 的 Hive 有数量级的提升。

(2) 资源监控

由于集群用到多台服务器, 因此如何便捷、直观地监控每台服务器当前的状态, 成为系统实施与维护首先要考虑的问题。数据质量管理体系使用 Cloudera Manager(免费版)作为 Hadoop 集群管理和监控的工具, 并进行了适当优化, 提供了一个 B/S 结构的管理界面, 用户可以实时查看

到集群里每一台服务器的健康状况、资源使用率、任务执行情况等信息。

3.3 主要创新点

(1) 利用分布式技术解决传统关系型数据库的性能瓶颈

数据质量管理体系具有数据量大、数据增长快以及分析任务繁重等特点。传统关系型数据库在应对这样的大数据量场景时, 数据分析性能迅速下降, 已难以满足实际的应用需求。本研究提出使用 Hadoop 分布式架构重构系统中数据量较大的模块, 关系型数据库只用于存储分析结果。本系统采用了基于 Impala 框架封装的数据分析引擎, 开发人员可以使用熟悉的 SQL 语言进行数据分析, 以降低 Hadoop 集群的应用门槛。利用 Hadoop 集群可以轻易实现数据的水平分布, 从而大大提高系统的可用性和数据分析性能。

(2) 使用多线程并行技术提高数据质量校验引擎的性能

为了能最大限度地利用分布式系统的硬件资源, 提出了基于多线程的校验规则执行引擎。执行引擎能根据规则数量和当前系统负载情况动态创建线程池, 其中包括问题数据查找和问题数据写入两种。在进行数据质量校验的过程中, 系统根据问题数据写入和问题数据查找的情况自动创建或销毁线程, 同时应用“数据切片”技术把数据按照线程的数量平均切分, 配送到不同的线程以提升数据处理效率。利用多线程技术可以充分利用服务器的硬件资源, 极大地提升性能。

(3) 基于元数据自动生成数据质量校验规则

数据质量管理体系需要管理成千上万的校验规则, 在提高校验性能的同时也需要提高校验规则配置的效率。元数据描述了数据库表关系以及数据表字段的详细信息, 如字段类型、长度、数据来源、加密等级、数据质量要求等。本文创新提出了基于元数据的数据质量校验规则自动生成功能, 用户只需在界面选择需要进行校验的表和字段, 系统即可根据元数据中的数据质量要求自动生成校验规则, 提高了规则配置的准确性, 并且大大节省了人工成本, 提高了规则配置效率。目前, 营销、生产、财务、人力资源等业务领域约 3 255 条校验规则均通过此方法生成。

4 系统实践

4.1 性能对比测试

实际测试表明, 分布式系统具有优秀的水平扩展性,

随着数据量的增长,Hadoop 集群的执行时间几乎呈线性增长。应用集中式架构的平均数据读取速度为 150 MB/s,读取 1 TB 的数据需要约 2 h;Oracle 的处理时间则呈指数级增加,如图 4 所示。可以看出,当数据量达到 500 GB 时,在 Oracle 上执行 SQL 已出现性能瓶颈。

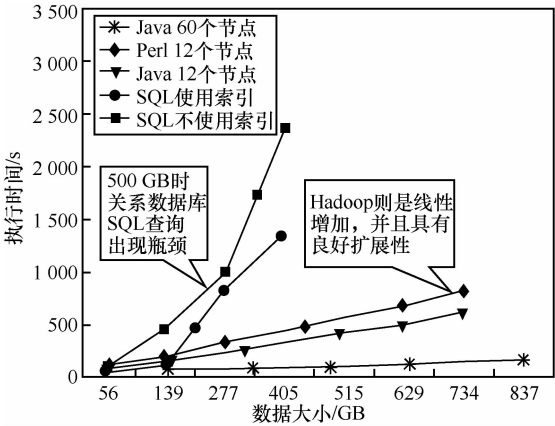


图 4 关系型数据库与 Hadoop 集群执行时间对比

4.2 应用情况

目前分布式数据质量管理体系已完成开发,并在广西电网公司、广东电网公司试运行,根据现场收集的运行数据,应用分布式数据质量管理体系,最高可实现数据质量校验效率比原集中式系统提升 15 倍左右。

服务器配置信息见表 1。

表 1 服务器配置信息

数据库	广西电网公司	广东电网公司
应用服务器 (Weblogic)	1 台 Dell R910	1 台 Dell R910
系统数据库 (Oracle)	与应用服务器共用一台机器	与应用服务器共用一台机器
分布式数据库	4 台 Dell R910	4 台 Dell R910

数据质量校验情况见表 2。

从表 2 可以看出,分布式系统的性能比集中式系统有数量级的提升。由于客观因素的限制,广东电网公司的被校验库和 Hadoop 集群部署在不同的子网络,受到网络和

表 2 数据质量校验性能对比

性能	广西电网公司	广东电网公司
总记录数/亿条	3.65	16
问题记录数/万条	1 714	3 000
集中式系统校验时间/h	32	37.5
分布式系统校验时间	1 h 35 min	9 h 20 min

防火墙的影响,校验时间要长于广西电网公司。

5 结束语

本文针对电力企业数据质量管理工作的现状和难点,提出了分布式数据质量管理体系解决方案,通过开展分布式存储和计算关键技术研究,初步搭建大数据存储及计算平台,并针对数据质量管理子系统进行了技术架构升级和试点应用,原需 45 h 的数据质量校验工作时间缩短至 3 h 以内,有效提升了系统处理和分析效率,同时为中国南方电网公司数据中心架构升级工作储备了技术基础。

参考文献:

[1] 田秀霞,周耀军. 基于 Hadoop 架构的分布式计算和存储技术及其应用[J]. 上海电力学院学报, 2011, 27(1): 70-75.
TIAN X X, ZHOU Y J. The technology and application of distributed computing and storage based on Hadoop architecture[J]. Journal of Shanghai University of Electric Power, 2011, 27(1): 70-75.

[2] BIRMAN K P, GANESH L, RENESSE R. Running smart grid control software on cloud computing architectures[C]//Workshop on Computational Needs for the Next Generation Electric Grid, April 19-20, 2011, Cornell University, Ithaca. [S.l.s.n.], 2011: 1-28.

[3] 刘鹏. 云计算[M]. 北京:电子工业出版社, 2010.
LIU P. Cloud computing[M]. Beijing: Publishing House of Electronics Industry, 2010.

[4] REESE G. Cloud application architectures: building applications and infrastructure in the cloud[M]. New York: O'Reilly Media, 2009.

[5] 辛军,陈康,郑纬民. 虚拟化集群管理技术研究[J]. 计算机科学与探索, 2010(4): 325-327.
XIN J, CHEN K, ZHENG W M. Studies on virtualization of cluster resource management technology[J]. Journal of Frontiers of Computer Science and Technology, 2010(4): 325-327.

[6] HDFS scalability with multiple NameSpaces [EB/OL]. [2015-09-20]. <http://issues.apache.org/jira/browse/HDFS-1052>.

[7] WHITE T. Hadoop: the definitive guide[M]. New York: O'Reilly Media, 2009.

[8] Hadoop apache project [EB/OL]. [2015-09-20]. <http://hadoop.apache.org>.

[9] GHEMAWAT S, GOBIOFF H, LEUNG S T. The Google file system [C]//SOSP, October 19-22, 2003, Bolton Landing, New York, USA. New York: ACM Press, 2003.

[10] 陈远,罗琳. 信息系统中的数据质量问题研究 [J]. 中国图书馆学报,2004(1):48-50.
CHEN Y, LUO L. Research on data quality in information system[J]. Journal of Library Science in China, 2004(1) :48-50.



[11] 胡金林, 梅士员. 基于元数据扩展的空间数据质量管理方法 [J]. 现代测绘, 2004, 27(3):21-24.

HU J L, MEI S Y. The extended metadata method of spatial data quality management [J]. Modern Surveying and Mapping, 2004, 27(3): 21-24.

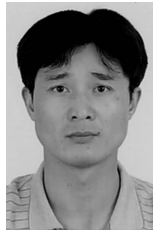
[作者简介]



李远宁(1981-),男,博士,中国南方电网有限责任公司信息部高级工程师,主要从事大数据分析及应用工作。



刘森(1983-),男,博士,中国南方电网有限责任公司信息部工程师,主要从事大数据分析及应用工作。



张诗军(1973-),男,中国南方电网有限责任公司信息部高级工程师,主要从事数据管理、管理信息化工作。



陈丰(1973-),男,中国南方电网有限责任公司信息部工程师,主要从事管理信息化、架构设计工作。



王志英(1962-),男,中国南方电网有限责任公司信息部教授级高级工程师,主要从事管理信息化、架构设计工作。