文章编号:2096 - 5389(2020)05 - 0093 - 04

融合架构的分布式数据库技术在气象大数据平台上的应用实践

汪　华,李　波,王　彪,廖婷婷

(贵州省气象信息中心,贵州　贵阳　550002)

摘　要:随着信息技术的发展和应用,气象发展已融入到政治、经济、民生等各领域的发展中,气象数据呈现数据体量大、实时性要求高、数据种类丰富的大数据特征,数据的应用效果随时间呈断崖式下降的趋势非常明显,是区别其他行业数据应用的明显差异点。传统关系型数据库暴露出长序列数据访问效率低、软硬件架构成本高、性能扩展能力弱等缺陷,已无法满足业务发展的需要,该文通过融合架构的分布式数据库在气象业务中的应用效果分析,从技术原理和实践两个角度,为气象大数据的应用提供技术借鉴。

关键词:分布式数据库;hadoop;融合架构;长序列

中图分类号:TP311.133.1　**文献标识码**:B

Application Practice of Distributed Database Technology with Fusion Architecture on Meteorological Big Data Platform

WANG Hua, LI Bo, WANG Biao, LIAO Tingting

(Guizhou Meteorological Information Center, Guiyang 550002, China)

Abstract: With the development and application of information technology, meteorological development has been integrated into the development of politics, economy, people's livelihood and other fields. Meteorological data shows the characteristics of big data with large volume, high real - time requirements, and rich data types. The application effect of data shows a cliff like downward trend over time, which is a significant difference between data applications in other industries. The traditional relational database exposes the defects of low access efficiency of long sequence data, high cost of software and hardware architecture, weak capability of performance expansion and so on, which can't meet the needs of business development. This paper analyzes the application effect of the distributed database based on the fusion architecture of meteorological business, and provides technical direction reference to the application of meteorological big data from two aspects of technical principle and practice.

Key words: distributed database; hadoop; fusion architecture; long sequence

0　引言

信息化已经进入全面渗透、跨界融合、加速创新、引领发展的新阶段,大数据时代的到来对各行各业都产生了深远的影响,数据已成为国家基础性战略资源,各行业的数据都呈现爆发增长和海量集聚的趋势,人们可以通过对数据进行深度结合发掘超出本部门业务领域的信息,获得具有开创性的新知识或预测信息,甚至开创新的业务领域。气象本身就是信息化程度较高、对数据依赖度较高的行

收稿日期:2019 - 11 - 25

第一作者简介:汪华(1974—),女,副高,主要从事气象信息化建设研究工作,E - mail:wh_mail@ foxmail. com。

通讯作者简介:李波(1972—),女,工程师,主要从事气象信息化业务管理工作,E - mail:1798356823@ qq. com。

资助项目:黔科合支撑[2019]2386 号:基于大数据的气象防灾减灾服务平台构建与应用。

业，全球最具权威的 IT 研究与顾问咨询公司 Gartner 预测行业的大数据平台将面临处理速度、多源类型、负载剧增、价值挖掘等 4 方面的挑战，传统的关系型数据库很难支撑大数据的服务要求，气象大数据应用的发展能够揭示传统技术方式难以展现的关联关系，从省级层面探索不同的主流技术在行业内部的应用将很有借鉴意义。

1　气象大数据应用典型特征

气象观测数据具有数据体量大、实时性要求高、数据多样性、价值高 4 大特征，是非常典型的大数据应用场景，但是传统的计算方式无法对海量的数据进行深入挖掘，数据采集过程从逐时向逐秒发展，传统数据处理技术无法有效应对数据的巨量并发，以贵州省内的区域自动站为例，全省每个收集时次会产生 3 000 多条气象观测数据，一个月就能累计上亿的数据量。

气象数据要取得最大的应用效果，要求实时采集、实时处理、实时计算，典型应用在防灾减灾救灾领域，气象数据价值往往体现在预警效用上，需要随时能调用、查询、计算、分析，而且数据应用的效用随时间呈断崖式下降，这也是气象大数据区别于其它行业数据应用最明显的差异点和价值特征。

2　气象观测数据在线检索分析对大数据平台的要求

2.1　传统关系型数据库的技术瓶颈

在大数据时代，Oracle 等传统关系型数据库很难满足海量气象数据存储和在线分析的需求，主要原因在于传统的关系型数据库是为 OLTP（联机事务处理 on－line transaction processing）场景定制，其技术特征是为满足行业中日常事务和流程处理的需求，典型场景如金融行业的交易和支付系统，在海量数据存储在线分析等应用场景下，传统数据库暴露出效率低、软硬件成本高、扩展能力弱等天然缺陷。

数据规模：传统数据库数据规模，通常以 MB、GB 单位为主，较少达到 TB 级别，大数据平台需要处理的数据规模通常以 GB、TB 和 PB 单位为主。

数据类型：传统数据库需要处理的数据结构类型单一，主要处理结构化的数据，而大数据平台需要连接多个数据来源，多种类型，包含来自 Oracle、MySQL 等主流数据库的结构化数据，还包含文本、日志、音频、视频和图片等半结构化和非结构化的数据。

数据范围：传统数据库往往只包含行业全部数据的部分样本，仅建立部分库表或业务单元的联系，而大数据平台需要处理全量数据，需要全部库表与所有业务内容建立联系并开展数据挖掘。

数据算法：传统数据库以常见的加减乘除为主，而大数据平台需要对数据进行聚合、分类、范围统计、数据挖掘等各种更为复杂的加工处理。

数据应用：传统数据库以支持实时交易和业务流程为主，大数据平台面向已有数据能开展不限时间、空间各维度的分析与对比，可为多种不同的应用灵活地提供数据调用与展示。

因此在进行数据库平台选型的时候，需要依据业务应用的需求来全局、合理地进行规划和分类，对于 OLTP 型操作的业务需求，可以使用传统关系型数据库，对于需要进行海量数据在线分析和数据挖掘等业务需求，则需要采用高效、安全、可扩展的大数据平台。

2.2　主流大数据技术的对比分析

随着海量数据的出现，关系型数据库收到冲击，数据分析开始向分布式数据库和数据仓库发展，在主流的大数据平台层面，Hadoop 开源生态和 MPP（Massively Parallel Processor 大规模并行处理）分布式数据库是当前两大主流的大数据平台技术，两类平台技术具有各自的特征。

2.2.1　Hadoop 开源大数据平台　Hadoop 开源大数据起源于互联网行业，最初是为解决海量的日志等半结构化数据及图片、视频等非结构化数据的存储和计算问题发展而来，主要依托于开源社区进行技术迭代更新，Hadoop 的扩展性好，适合海量数据的批量处理，但在气象观测数据在线检索和分析场景下，Hadoop 存在一些技术缺陷，如 Hive 是 Hadoop 生态数据分析场景主流组件，Hive 底层采用的是 Map－Reduce 计算引擎，其响应时延是分钟和小时级，在实时性要求越来越高的气象业务场景下，Hive 明显延时过高，不能满足气象在线分析检索的要求，同时 Hbase 不支持关联计算，不支持事务，不支持标准 SQL 语法，以上的问题决定了其很难满足海量气象数据在线分析检索场景的要求。Spark 的 RDD 存储格式采用类似数据随机分布机制，也没有索引机制，在处理数据时需要实行全扫描机制和大量数据在节点间的传输和广播，在精确检索、范围检索、海量数据关联分析、聚合运算等方面存在性能问题，Hadoop 生态不支持事务，导致在数据插入和批量更新等场景下可能发生数据丢失和不一致等情况，对于数据一致性和完整性有很高要求的气象行业用户是一个严重问题。Impala 不支持索引，即使小数据量的检索，Impala 也需要对全表做扫描，

导致并发效率较低。同时开源软件普遍存在性能调优复杂,容易出现内存耗净等复杂的问题,需要高级的技术人员进行长期维护,维护成本很高。

表 1　主流 Hadoop 生态技术功能对比

Tab. 1　Comparison of mainstream Hadoop eco technology functions

技术功能	JDBC/ODBC	ANSI SQL	事务	关系型查询	多表关联	修改删除	实时响应	即席查询	报表统计
Hive	不支持	不支持	不支持	支持	不支持	不支持	不支持	不支持	支持
Hbase	不支持	不支持	不支持	不支持	不支持	支持	支持	不支持	不支持
SparkSQL	支持	支持	不支持	支持	支持	不支持	不支持	不支持	支持
Impala	支持	不支持	不支持	支持	支持	不支持	支持	支持	不支持
融合架构分布式数据库	支持	支持	支持	支持	支持	支持	支持	支持	支持

2.2.2　MPP 分布式数据库　MPP 数据库是新型分布式数据库,其技术起源于传统关系型数据库面向大数据时代的改良,主要特点是实时性,缺点是对非结构化数据库的支持不好。MPP 核心原理是将一个大的查询通过分拆为多个子查询,分布到底层执行,最后再合并结果,其核心技术原理是通过多线程并发的暴力扫描来实现高效分析,这种暴力扫描对单个查询来说,动用了整个系统的能力,单个查询的确比较快,但同时带来用力过猛的问题,整个系统能支持的并发受到严重制约。从目前 MPP 数据库的行业应用情况和经验来看,最高支持 30 个以上的并发查询将导致查询效率的急剧下降。MPP 数据库普遍采用存储计算资源紧耦合的设计,导致 MPP 无法和 Hadoop HDFS 分布式存储引擎兼容,必须单独搭建硬件集群进行部署且要求独占计算和存储资源,无法和 Hadoop 生态有效融合,扩展能力存在明显短板,同时集群节点数目不能过多,超过一定数量后集群的计算能力和效率增长率均会明显下降,单个节点软硬件故障会导致平台运行效率严重下降。

表 2　融合架构分布式数据与传统 MPP 数据库的功能对比

Tab. 2　Functional comparison between the fusion architecture distributed data and the traditional MPP database

技术选型对比	架构先进性	并发支持能力	高可用	扩展能力	存储过程	在线扩容	集群容灾
DB2	传统小型机 + 磁盘阵列,严重的性能瓶颈和天花板效应)	一般	基于 IBM 传统高可用技术	差(节点数增加时,磁盘阵列可能成为 I/O 瓶颈)	支持	差(添加节点需要重新启动,升级时间长)	不支持
Vertica	较好(share - nothing 无共享架构,无管理节点)。但是国内 100 节点以上的案例很少	一般(不超过几十个并发)	通过 K - Safe,冗余机制,K 值理论上可以设置 2 以上,但实际根据系统性能,一般只设置为 2)	50 节点以下,节点增多性能下降、木桶效应	不支持存储过程,从传统数据库迁移到 Vertica 时,内部分析逻辑需要重新开发	支持	不支持
GreenPlum	一般(share - nothing 架构,但是有管理节点和主控节点,集群规模增长成为性能和扩展性瓶颈)	差(有 Master 节点,并发支持能力一般)	一般(只支持一个副本,不支持在线替换节点)	一般(一般不超过 64 节点)	差(只支持全量备份)	一般	不支持
融合架构分布式数据库	好(支持 1 000 台以上的大规模集群)。支持强大的 Sql on Hadoop 方案	高(支持上千并发)	好(1 - 5 副本可选)	好(单集群最大支持 1 000 个节点)	好(支持存储过程和多种 olap 分析函数)	强(支持在线扩容和缩容,支持扩展过程中的写操作,最大化在线扩容能力)	支持(提供基于日志复制技术邮集群容灾解决方案)

3　融合架构分布式数据库在气象数据在线分析场景的应用

Gartner 机构 2016 年提出融合型大数据的概念,根据应用场景搭建多种类型的数据仓库,如采用逻辑数据仓库 LDW(Logical Dataware House)模式解决多个数据源以及多种类型数据的综合分析场景,通过数据虚拟化技术实现数据的统一访问,数据联邦实现跨数据源的访问;搭建运营数据仓库实现流数据的分析,支撑实时性要求高的数据加载和分析;采用模型无关数据仓库,支撑深度机器学习模型的建立,采用文本和图形的计算引擎和算法库,解决深度数据挖掘和复杂的机器学习应用问题,根据气象行业数据应用的特点采用融合架构的分布式技术是比较适合的选择。

3.1　融合架构分布式数据库的技术特征

融合架构分布式数据库是行业最新的大数据平台技术,其技术原理融合了 Hadoop 和 MPP 两类大数据平台的优点,代表大数据平台技术最新发展方向。一般的技术架构是底层采用 Hadoop 作为存储引擎,结合高效的分布式 SQL 引擎,具备支持高并发数据写入同时开展在线数据分析的能力,采用 Share – Nothing 分布式计算架构,最大程度利用硬件资源提升计算效率。

3.2　融合架构分布式数据库在实际气象业务场景的测试分析

根据对于融合分布式数据库的技术特征分析,前期集合实际数据和业务场景开展了实地测试,测试共使用 5 台服务器,其中 4 台用于数据库计算节点,1 台部署管理节点,选取的数据主要是 1.6 亿条的自动站数据表,测试采取脚本加人工方式进行,编制脚本进行场景测试,对于无法采用脚本实现的测试案例通过人工判断、执行。

测试 81 条典型 SQL,按照业务需求分为以下 5 大类场景:

从按时间范围的简单条件查询;按时间范围的简单条件极值计算;按时间及其它范围的极值计算;按时间范围及其它范围的组合查询;按时间范围的模糊查询。

测试结果表明,5 个业务场景的全部 81 个 SQL 在分布式数据库中都能够直接成功执行,分布式数据库在 SQL 兼容性和功能性方面可以满足气象局分析应用场景的要求。在 5 个不同的应用场景下的 81 个 SQL 执行时间在毫秒级的占比 85.2%(69/81),执行时间在 1 ~ 5 s 的占比 3.70%(3/81),执行时间在 5 ~ 10 s 的占比 3.7%(3/81),执行时间在 10 ~ 30 s 的占比 7.2%(6/81)。整体测试结果表明 11 个场景功能适配占比 82%,84 种条件查询适配占比 88%,小数据量条件查询响应 1 s 左右,占比 24%,中大数据量查询响应在秒级,性能远高于 Oracle,占比 76%,整体性能表现非常出色。

4　结论

从融合架构分布式数据库的测试和应用来看,能得到以下两点结论:

①系统 SQL 兼容性较好支持标准的 SQL 语句,不需要对应用进行大幅度的修改就能平滑迁移,技术架构能够解决现有商业化的大数据平台存在的技术缺陷,又能较好的与 CIMISS 的数据服务服务 MUSIC 服务接口对接;

②查询数据量小的情况响应时间在毫秒级,略低于 Oracle 或者和 Oracle 性能基本持平;在数据量大的情况下分布式数据库性能远高于 Oracle。76% 的大数据量查询在秒级,性能提升非常明显,能够满足气象数据分析应用场景对数据库平台的要求,数据平台在实际业务场景中还需要考虑应用算法研究,从技术原理和实践两个角度,探索融合大数据技术形成大数据在气象行业的有效应用实践。

参考文献

[1] 朱亮,钟艳雯,贺炜,等. 基于分布式的农业气象大数据平台设计与实现[J]. 湖北农业科学,2019,58(6):128 – 130.
[2] 闫健卓,高凯丽,许红霞,等. 基于虚拟化的水务分布式大数据存储平台设计[J]. 水利信息化,2019(3):17 – 24.
[3] 张海涛,张文娟. 基于大数据的分布式文件系统技术研究[J]. 电子测试,2019(4):82 – 83.
[4] 朱兰英. 分布式大数据管理系统的设计与实现研究[J]. 电脑知识与技术,2019,15(5):25 – 26.
[5] 郭茜,王彪,汪华,等. 贵州省气象大数据平台架构设计[J]. 成都信息工程大学学报,2018,33(5):531 – 535.
[6] 汪华. 虚拟化技术在气象业务系统集约化整合中的应用实践[C]//第 35 届中国气象学会年会 S20 深度信息化:应用支持与智能发展. 中国气象学会,2018:4.
[7] 汪华. 气象监测数据集中管理系统的开发与实现[C]//第 32 届中国气象学会年会 S14 第五届气象服务发展论坛——气象服务与信息化. 中国气象学会,2015:7.
[8] 廖婷婷,王彪,肖卫青,等. Storm 流式技术在地面气象数据处理中的应用[J]. 中低纬山地气象,2019,43(5):78 – 81.