廖婷婷,支亚京,李进讷,等. 气象数据在多种类数据库上查询统计能力初探[J]. 中低纬山地气象,2023,47(3):108 - 112.

气象数据在多种类数据库上查询统计能力初探

廖婷婷[1],支亚京[1],李进讷[1],虞雪莹[2],汪　华[1]

(1. 贵州省气象信息中心,贵州　贵阳　550002;2. 贵州省贵阳市乌当区气象局,贵州　贵阳　550018)

摘　要:气象结构化数据包含地面小时数据、地面分钟数据等多种数据种类,在分析和调用中需要利用时间属性和空间属性。随着数据量日益增大,高并发、复杂统计条件下的查询与统计对数据库的效率要求十分严苛。在此背景下,传统的关系型数据库逐渐难以满足实时应用需求。该文使用不同类型的分布式数据库和关系型数据库,结合气象业务使用场景,探索多种类型下数据库的使用性能,研究不同场景的性能差异与不同数据库架构框架之间的关系,以提高数据服务实时响应能力。

关键词:分布式数据库;关系型数据库;气象数据

中图分类号:TP311.52　**文献标识码**:B

Preliminary Study on the Query and Statistics Ability of Meteorological Data on Various Databases

LIAO Tingting[1], ZHI Yajing[1], LI Jinne[1], YU Xueying[2], WANG Hua[1]

(1. Guizhou Meteorological Information Center, Guiyang 550002, China;
2. Meteorological Bureau of Wudang District, Guiyang City, Guizhou Province, Guiyang 550018, China)

Abstract: The meteorological structured data includes a variety of data types, such as ground hourly data, ground minute data, etc. Time attributes and spatial attributes need to be used in the analysis and call. With the increasing amount of data, queries and statistics under the conditions of high concurrency and complex statistics have strict requirements on the efficiency of the database. In this context, the traditional relational database is gradually difficult to meet the needs of real - time applications. The paper uses different types of distributed databases and relational databases, combined with meteorological business use scenarios, to explore the performance of databases under various types, and study the relationship between the performance differences of different scenarios and different database architecture frameworks to improve the real - time response capability of data services.

Key words: distributed database; relational database; meteorological data

0　引言

在日益剧增的数据量和实际应用场景下,气象结构化数据在传统数据库中的存储处理、数据读取等方面存在负载饱和、读写性能不理想等问题[1-3]。气象结构化数据在调用和更新上具有空间和时间特性,需要通过统计算法得到特定的统计结果,这对数据库及服务器的性能需求特别严苛[4-6],其中

收稿日期:2022 - 09 - 02

第一作者简介:廖婷婷(1992—),女,硕士,工程师,主要从事气象信息系统研发、气象数据服务等工作,E - mail:1563939569@qq.com。

资助项目:贵州省科技支撑计划项目([2017]2819 号):基于大数据分布式技术的信息数据管理技术研究;贵州气象局重要业务科研项目(黔气标合 ZY[2020]09 号):基于大数据的多源气象数据存储共享及支撑技术研究。

地面观测小时值其表的字段有 260 个左右,光贵州省存储的站点数就多达 2700 余个,累计数据量已达到亿数量级,在实际查询时,需要空间范围上的统计计算特性,表与表之间存在联动查询和存储更新等操作,使得应用在调用数据和存储更新时需要花费更多的等待时间,甚至出现数据库无法返回计算结果的情况。

本文通过对比关系型数据库 Oracle、关系型数据库 MySQL、基于 Kudu 的分布式数据库集群、基于融合架构的分布式数据库易鲸捷(EsgynDB)在存储及应用上的性能,探索多种类数据库框架下的使用性能,研究数据库架构更适应气象结构化数据的具体应用场景,在实际调用数据库中可以通过查询条件定制查询不同数据库,提高数据服务实时响应能力。

1　数据库现状

关系数据库使用最直观的行列来存储,根据定义好的表格结构直接新增数据内容,常用的关系型包括 Oracle 和 MySQL。和 Oracle 相比,MySQL 是开源的,且安装所用资源较少,这样就增加了速度并提高了灵活性,数据库中的表对应一个或者多个数据库目录下的文件,并取表存储时的存储引擎[7-8]。而一个 Oracle 数据库包含一个或者多个表空间,表空间对应数据在磁盘上的物理存储。气象结构化数据在关系型数据库中具有稳定性高的特性,数据管理也很简单直观。

主流的分布式数据库系统包括基于 Hadoop 的分布式数据库、大规模并行分析(Analytical Massively Parallel Processing ,MPP)数据库、融合架构分布式数据库。Kudu 作为基于 Hadoop 的分布式数据库,能较好地适应气象结构化数据特点,比 HBase 批处理快,适用于联机分析处理(On - Line Analytic Processing,OLAP)的分析场景,而且比 HDFS 随机读写能力强,适用于实时写入或者更新的场景[9-10]。贵州省气象信息中心搭建的大数据云平台采用通过 Impala 集群实现基于 Kudu 的结构化数据长序列的灵活查询支撑[11],优化数据查询执行效率。

MPP 数据库是针对分析工作负载进行了优化的数据库,采用列式存储使复杂的分析查询可以更快、更有效地处理[12]。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 2 类大数据平台的优点[13]。将事务交易与数据分析结合在同一个海量并发大数据库上,仅用 SQL 即能获得高级、可靠、线性拓展的数据库服务。可根据应用场景搭建多种类型的数据仓库,通过数据虚拟化技术实现数据的统一访问。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的同时开展在线数据分析的能力,采用 Share - Nothing 分布式计算架构,最大程度地利用硬件资源提升计算效率。

2　测试方法和结果对比

2.1　方案设计

数据库在硬件条件上统一部署于 4 个物理节点,内存 256G,CPU 为 16 核,千兆网卡。测试数据为中国地面逐小时探测资料。模拟用户应用场景下的检索和统计习惯,设置多场景进行单发和并发测试。按照贵州省气象信息中心总结的各业务单位的常用的业务场景统计(表 1),可以分为检索场景 4 个(表 1 中序号 1 ~ 4),统计场景 4 个(包括极值计算和求和计算,表 1 中序号 5 ~ 8)。

表 1　业务场景内容

Tab. **1**　Content of business scenario

序号	测试场景	具体检索条件	时间范围/d
1	按时间范围的简单条件查询	检索某时间段、某站点的小时降水量	1、10、30
2	按时间范围及区域范围的组合查询	检索某时间段、贵州区域地面自动站全要素数据	1、10、30
3	按时间范围及指定站点的组合查询	检索某时间段、84 个贵州省国家站点的温压湿风要素值资料	1、10、30
4	按时间范围及空间范围的组合查询	检索某时间段、某经纬度范围检索地面全要素数据	1、10、30
5	按时间范围的简单条件极值计算	统计某时间段、某个站点的小时降水量极值	1、10、30、90、180、365
6	按时间及空间范围的极值计算	统计某时间段、某经纬度范围的小时降水量极值	1、10、30、90、180、365
7	按时间范围的简单条件求和计算	统计某时间段、某个站点的小时降水量求和	1、10、30、90、180、365
8	按时间及空间范围的求和计算	统计某时间段、某经纬度范围的小时降水量求和	1、10、30、90、180、365

本文针对业务场景进行单发、并发和应用测试,具体内容和实现方法如下:

(1)单发场景测试:以单条语句串行执行 SQL 方式进行功能测试,使用 Jdbc 接口的 Trafci 工具。

(2)并发场景测试:根据实际使用场景,选取单一场景的部分测试内容进行测试,分别开展 50、100

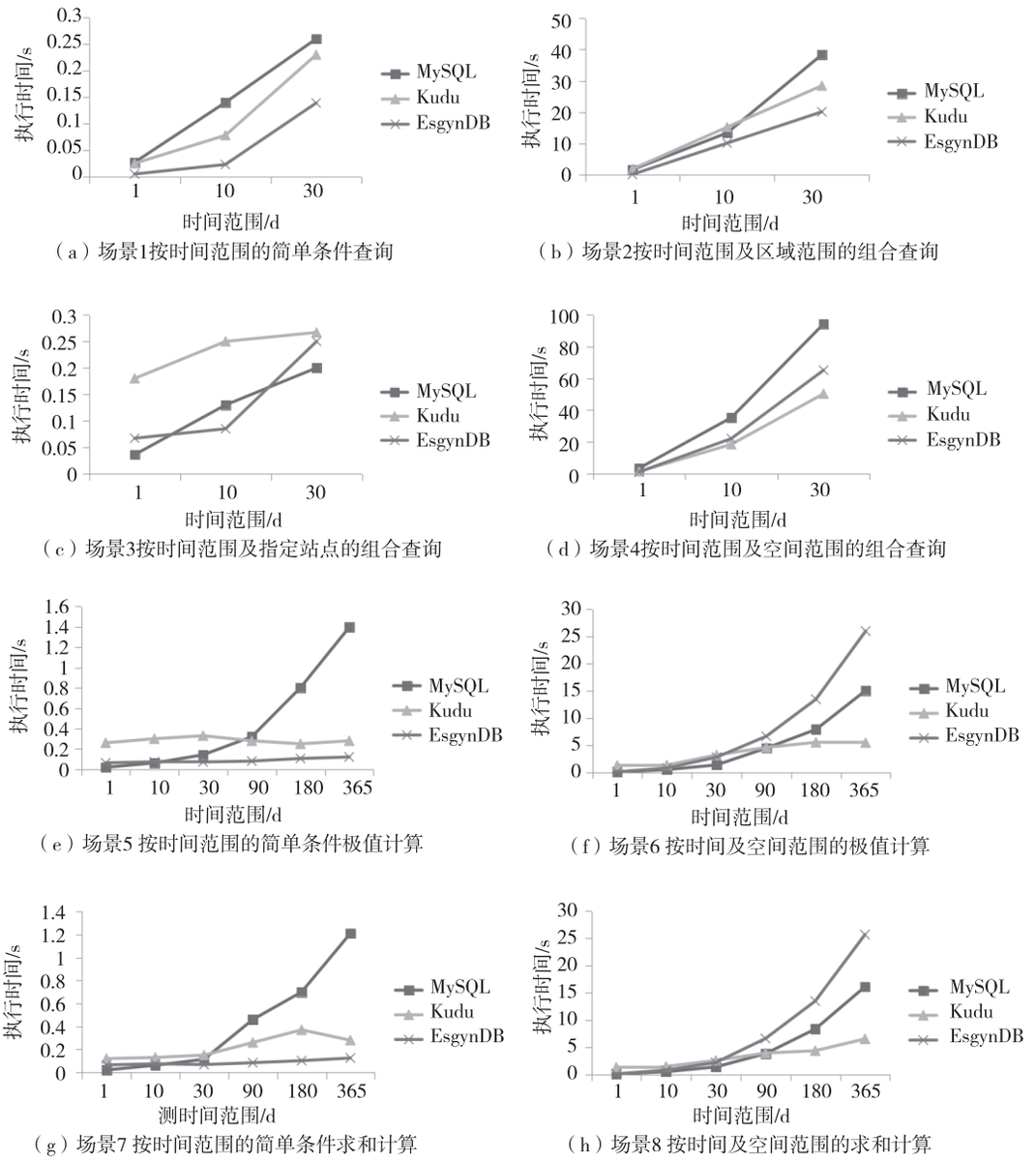
条检索语句并发执行场景下的功能验证，使用 Jdbc 接口的 Jmeter 工具。

（3）资源占用情况：对比几类数据库集群在测试过程中的网络 IO、内存占用、CPU 使用率的情况。

（4）与其他研究者的分布式数据库进行对比实验，使用他人测试场景，测试本文章数据库的适应能力。

2.2　测试结果

2.2.1　单发场景测试　图 1 为检索查询的测试结果和统计查询的测试结果。测试过程中 Oracle 数据库与其他数据库测试结果差异明显，与 MySQL 数据库相比，测试时间达到其 10 倍以下占比 31%，10～100 倍占比 23%，100 倍以上占比 30%，超时未返回结果占比 16%，由于巨大的结果差异，所以 Oracle 的测试结果未列在图中进行比较。



（a）场景1按时间范围的简单条件查询　（b）场景2按时间范围及区域范围的组合查询

（c）场景3按时间范围及指定站点的组合查询　（d）场景4按时间范围及空间范围的组合查询

（e）场景5 按时间范围的简单条件极值计算　（f）场景6 按时间及空间范围的极值计算

（g）场景7 按时间范围的简单条件求和计算　（h）场景8 按时间及空间范围的求和计算

图 1　单发检索查询（a～d）以及单发统计查询（e～h）测试结果比较

Fig. 1　Comparison of single search query （a～d）ovnd single－shot statistical query （e～h） test results

如图 1a～1d 所示，在检索场景当中，分布式数据库相比关系型数据库优势明显，其中 Esgyn 的效果普遍优于 MySQL 和 Kudu，在场景 4 中虽然没有 Kudu 表现优异但是差距也在 10 s 以内。在图 1e～1h 中，统计场景的结果在不同数据库中有所差异，场景 5 和场景 7 都是按照时间范围的单站的统计检索，他们的测试结果中 2 个分布式数据库结果很接近且明显优于 MySQL 数据库，在场景 6 和 8 中，加

入了空间尺度即经纬度作为检索要素,测试结果中,当时间范围在 90 d 以前 3 个数据库表现很接近,Kudu 比较 2 个数据库用时略微高一点,90 d 以后 Kudu 表现最优,其次是 MySQL 和 Esgyn,由此可见在长时间序列和大尺度范围下 Kudu 优势明显。

2.2.2　并发场景测试　通过单发测试的结果,发现检索场景测试中,场景 1、3 和 4 结果比较统一,且由于场景 3 和 4 涉及到的数据量庞大,所以截取场景 1 和 2 进行并发测试,便于比较不同数据量测试下的并发结果。统计场景中选取场景 5 和 6 进行测试,便于测试空间尺度和时间尺度对测试结果的影响。由于并发测试数据量陡增,由此在统计测试中只选取了 30 d 时间范围。测试结果如图 2 所示,其中横坐标为“测试语句时间范围/并发次数”,如“10/50”代表 SQL 语句时间范围 10 d,并发执行 50 次的测试结果,纵坐标为语句执行花费的时间。

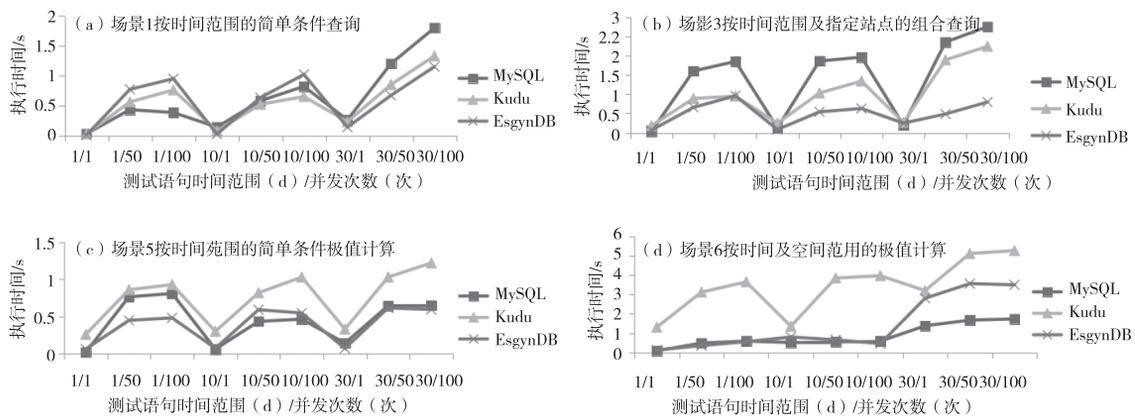


图 2　并发测试结果比较

Fig. 2　Comparison of concurrent test results

由实验结果可以看出,在场景 1 中 3 个数据库检索时间很接近,不同并发数得到的结果也没有拉开测试结果的差距,但是在场景 3 中,由于加入了 84 个站点,且多了 4 个气象要素之后,检索的数据量剧增,可以明显看出场景 3 的测试结果中 Esgyn 的优势明显,其次是 Kudu 和 MySQL。在统计测试结果中,场景 5 和 6 测试结果趋势也是很接近,随着并发次数的增多,Kudu 的优势却不像单发测试那样明显,反而表现不如 MySQL 和 Esgyn,其中 MySQL 表现出了较好的性能。由此可以得出在并发测试中,Esgyn 在检索场景下的性能较高,MySQL 在统计场景下的性能最高。

2.2.3　资源占用情况　通过统计测试过程中,不同数据库集群的资源占用情况,得到如表 2 的统计结果,可以看出关系型数据库的瞬时最大资源占用很容易达到较大的水平,而分布式数据库则将硬件资源维持在较低水平,其中 Esgyn 相对于 Kudu 数据库硬件占用更低,可能是因为 Esgyn 采用 Share – Nothing 分布式计算架构,每一个节点都是独立的、自给的,在系统中不存在单点竞争,最大程度利用硬件资源提升计算效率。

表 2　数据库资源占用情况

Tab. **2**　Database resource usage

数据库类型	网络 IO/(MB · s^{-1})			内存占用/GB			CPU 使用率/%		
	最大	最小	平均	最大	最小	平均	最大	最小	平均
Oracle	301	65	130	153	5	88	90	2.8	28
MySQL	220	43	102	78	4	35	87	1.5	23
Kudu	186	55	80	51	5	21	66	2.4	16
EsgynDB	197	45	83	32	3	13	41	2.1	10

2.2.4　对比实验　对比其他研究中不同类型数据库,分别对比 Spark 数据库和虚谷数据库。Spark 并行计算框架基于 MapReduce 的开源并行分布式计算框架,虚谷数据库为气象大数据云平台采用的用于存储结构化数据的数据库。结果如表 3 和表 4 所示。可以看出本文所使用的数据库相比 Spark 数据库性能都较高,与虚谷的对比各有优势,但是差距不大,本文的数据库可以较好地适应不同场景。

表 3　与 Spark 数据库对比结果

Tab. **3**　Comparison results with Spark database

站数	统计要素	时间跨度/d	并发次数/次	测试用时/s			
				MySQL	Kudu	EsgynDB	Spark[14]
1	降水量极值	365	1	1.395	0.28	0.121	3
40	降水量极值	365	1	15.06	5.47	25.987	50
100	降水量极值	365	1	50.34	18.23	43.5	160

表 4　与虚谷数据库对比结果

Tab. **4**　Comparison results with Xugu database

站数	检索要素	时间跨度/h	并发次数/次	测试用时/s			
				MySQL	Kudu	EsgynDB	虚谷[15]
2700	降水量	1	100	0.27	0.25	0.85	0.67
1	全要素	1	100	0.36	0.38	0.067	0.19

2.3　测试结论

通过统计单发和并发测试结果，可以看出无论单发还是并发，Esgyn 在检索场景上有明显优势。在单发统计测试中，随着时间与空间范围增大 Kudu 表现优异，但是在并发统计测试中 MySQL 表现出色。其原因可能是 MySQL 集群的优势，分布式是以缩短单个任务的执行时间来提升效率的，而集群则是通过提高单位时间内执行的任务数来提升效率[16]，当出现许多运算量大、高频次的检索时，MySQL 按照集群任务平均分配任务执行，分布式集群着眼于分布任务和减少单独任务的执行时间，这给任务造成了一定的挤压。

3　小结

随着地面气象数据在应用中的使用更加广泛和复杂，用户需要高质量的数据检索和统计反馈。本文根据实际情况将测试分为不同的业务场景，分别对几种数据库进行了性能验证，发现融合架构分布式数据库在检索场景下优势明显，而统计场景下，MySQL 数据库在并发测试中相比分布式数据库性能更好，这有助于针对不同业务场景切换数据库提供更好的用户体验。后期还需要对非结构化类型数据的性能进行进一步对比，以验证不同数据库在非结构化数据上的使用性能。

参考文献

[1] 黄志，苏传程，苏晓红. 大数据环境下 Spark 性能优化分析研究与应用[J]. 气象科技，2022，50(1)：51－58.

[2] 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.

[3] 李莉，张慧卿，陈传振，等. 自动站资料在 MS SQL SERVER 数据库中的 CIMISS 数据定制[J]. 中低纬山地气象，2020，44(4)：72－76.

[4] 淡嘉，郑昊，徐诚，等. 四川省气象预警决策发布系统负载均衡实现与性能优化[J]. 中低纬山地气象，2021，45(6)：106－110.

[5] 柳刚. 分布式技术与数据库应用于计算机技术领域解析[J]. 煤炭技术，2013，32(7)：198－199.

[6] 李从英，金石声，王彪，等. 使用 SymmetricDS 软件同步 CIMISS 核心库数据[J]. 中低纬山地气象，2020，44(1)：71－75.

[7] 李荣国，王见. MySQL 数据库在自动测试系统中的应用[J]. 计算机应用，2011，31(增刊 2)：169－171.

[8] 康文杰，王勇，俸皓. 云平台中 MySQL 数据库高可用性的设计与实现[J]. 计算机工程与设计，2018，39(1)：296－301.

[9] 顾飞杨，孔莹. 基于 Kudu 的大数据平台实时业务处理能力提升方案[J]. 电信科学，2019，35(10)：151－156.

[10] 蒋春平，黄煜骁，周晓君. 基于 Kudu 的实时业务应用场景解决方案[J]. 电信科学，2020，36(增刊)：268－275.

[11] 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)：531－565.

[12] 陈达伦，陈荣国，谢炯. 基于 MPP 架构的并行空间数据库原型系统的设计与实现[J]. 地球信息科学学报，2016，18(2)：151－159.

[13] 汪华，李波，王彪，等. 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[J]. 中低纬山地气象，2020，44(5)：93－96.

[14] 黄志，詹利群，任晓炜，等. Hadoop 环境下基于 SparkSQL 海量自动站数据查询统计初探[J]. 气象科技，2019，47(5)：768－772.

[15] 宋智，徐晓莉，张常亮，等. 应用分布式存储技术优化省级 CIMISS 数据服务能力[J]. 气象科技，2019，47(3)：433－438.

[16] 王素丽. 在云平台中高可用性数据库 MySQL 的设计与实现[J]. 计算机与数字工程，2020，48(7)：1633－1637.