杨芳,金琪,邢丽平,等. 基于 Cassandra 的实时气象数据分布式存储系统[J]. 陕西气象,2017(4):26-30.
文章编号:1006-4354(2017)04-0026-05

基于 Cassandra 的实时气象数据分布式存储系统

杨 芳,金 琪,邢丽平,杨代才
(湖北省气象信息与技术保障中心,武汉 430074)

摘 要:介绍了 Cassandra 关键技术、实时气象数据存储模型。湖北省运行 MICAPS4 采用基于 Cassandra 的分布式数据存储环境。对分布式存储系统高可用性、读取性能的测试结果表明,系统提高了 MICAPS 客户端对实时气象数据的检索效率,并极大减轻了各级台站的维护压力。

关键词:MICAPS4;气象数据;分布式;存储;Cassandra
中图分类号:P409 **文献标识码**:B

随着气象观测资料种类的增加、观测频次的增多,在气象数据存储技术领域,基于目录树结构的文件系统难以承受文件数量的急剧增长,在管理大量的结构化与非结构化数据方面,也不能较好地满足业务人员对数据进行按序访问的需求。中国气象局推出的 MICAPS(meteorological information comprehensive analysis processing system)4.0 版本,充分发挥了观测资料的作用,满足了不断发展的预报业务和气象服务需求。湖北作为全国气象部门首家省级试点单位,2016 年将 MICAPS4.0 投入测试运行。其后台数据环境引入非结构数据库 Cassandra;采用多维索引数据模型的实时气象数据存储系统,在保证 MICAPS 客户端快速调用(毫秒级别)实时气象数据的同时,可为全省各级气象部门预报业务系统提供基于气象卫星、天气雷达和数值预报、站点实况等数据支撑,实现数据在省级单点更新,极大减轻台站的维护压力;可灵活地添加节点来实现存储系统扩充,而不给系统带来较大的性能抖动;与 Spark 结合,利用 Spark 来实现复杂数据分析与处理功能,如处理集合预报资料等。

1 Cassandra 关键技术点

1.1 系统架构

ApacheCassandra 是一套开源分布式 NoSQL 数据库系统,基于一致性哈希算法的 P2P 环形结构。这种结构各节点功能完全相同,可灵活添加节点来完成系统的扩充或删除节点,且无需大规模转移数据,同时彻底避免系统因单点故障导致的不稳定性;每个节点通过 Gossip 机制进行消息同步;每个数据项都会被复制到 N 个节点(N 是通过参数配置的副本因子),系统利用数据的复制机将存储在各节点上的数据复制到其他节点上,实现了数据的高度可获得性与安全性[1]。

1.2 数据模型

Cassandra 使用宽列存储模型,每行数据记录是以 Key-Value 形式进行存储,其中 Key 为唯一标识。每个 Key-Value 其中的 Value 也称为 Column[2-3],作为一个三元组,包含有 Column Name、Column Value 与 timestamp;每个 CF 由一个 Key 及其对应的若干个 Column 标识组成。一个 keyspace 包含若干个 CF,类似关系型数据库中一个 database 可有多个 table。图 1 为一个 Column 型数据模型。

1.3 CAP 理论

NoSQL 典型遵循由 Eric Brewer 提出的 CAP 理论[4-5],依据此理论,在一个大规模的分布式数据系统中,有三个需求是彼此循环依赖的,一

收稿日期:2016-11-29
作者简介:杨芳(1982—),女,湖北洪湖人,学士,高工,从事气象信息通信研究。
基金项目:湖北省气象局科技发展基金重点项目(2015Z02)

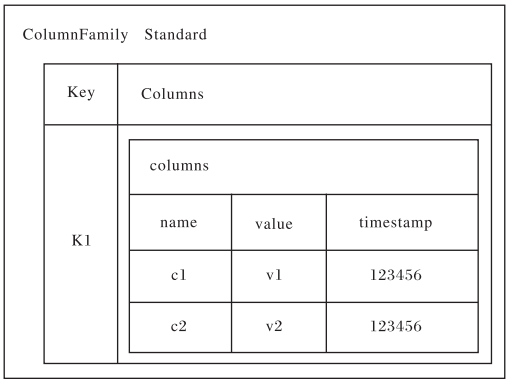

图 1　基于 Key-Value 的数据模型

致性(consistency)、可用性(availability)、分区耐受性(partition tolerance)。一致性为对所有数据库客户端使用同样查询都可得到相同的数据;可用性即所有数据库客户端都可读写数据;分区耐受性是指数据库分散到多个服务器上,即使发生网络故障,仍可提供服务。CAP 理论可简单描述为,一个分布式系统不能同时满足以上三个特性,最多只能同时满足两个。Cassandra 主要支持可用性和分区耐受性。在 Cassandra 中,数据具备最终一致性,集群整体的完全可用性。

1.4　存储机制

Cassandra 依赖本地的文件系统通过内存与磁盘的双重存储机制来保证数据的持久性[5-6]。Cassandra 有三个重要的数据结构,记录于内存的 Memtable,保存在磁盘中的 Commit Log 和 SSTable。Memtable 记录最近的修改,而 SSTable 记录着数据库所承载的绝大部分数据。通常情况下,一个 Cassandra 表会对应着一个 Memtable 和多个 SSTable。Cassandra 接收到客户端发送来的数据,首先将写操作记录到位于磁盘的 CommitLog 中;上述操作成功后,更新位于内存中的 Memtable 数据结构。持续的写入数据,使得 Memtable 逐渐增长,当其数据量到达某个阈值时,Cassandra 的数据迁移被触发,一方面将 Memtable 刷写到本地磁盘上成为永久的 SSTable,另一方面将 CommitLog 中的写入记录移除。对于读操作,客户端先查询 Memtable 中的数据,若无法获取所需信息,则检索本地磁盘。Cassandra 会定期执行压紧 compact 操作,将同一条数据不同的版本进行合并,过时数据也会在此过程中被删除;分层数据压缩,有效减少数据体积及磁盘 I/O。

2　实时气象数据存储模型

预报员在分析天气形势时,需通过 MICAPS4 客户端获取实时地面、高空观探测数据以及气象卫星、天气雷达、模式数据等;以这些数据为主所构成的“气象大数据”,都可归属于非结构化数据,具有多维、维度有序或无序等特点。基于 Key-Value 分布式数据库 Cassandra 在存储海量小文件、多维数据索引等方面有明显的优势[7-8];故采用 Cassandra 作为存储实时气象数据的解决方案。

2.1　系统设置

针对实时气象数据存储系统,用户对该系统读取的性能需求远远高于写入数据。通过对副本数进行合理设置,可分散读取压力。对于 5 节点集群,将副本数设置为 3;Row 分区模式:采用自动分区方式,使不同的 Row Key 均匀分布在各节点上,有利于数据读取压力的分散。

2.2　Cassandra 表设计

作为典型的非结构化数据,气象数据可以由多维索引来确定一个唯一的数据。业务用户常见的操作包括“最新数据”“左右翻页”“上下翻页”等。

2.2.1　数据表　根据不同数据类型建立相应数据表,用于存储数据内容,包括:ECMWF_HR(高分辨率数值预报产品)、SATELLITE(卫星资料)、UPPER_AIR(高空站点资料)、SINGLERADAR(雷达资料)等。以“T639”为例说明数据表结构(表 1)。

建表语句:CREATE TABLE "T639" ("dataPath" text, column1 text, value blob, PRIMARY KEY ("dataPath", column1));

2.2.2　层次表　用于存储所有模式或实况的层次信息,表名为 level;用户在客户端进行上下翻页操作,从 level 表中获取当前层次的上一层或下一层信息;利用层次表与数据表,可检索到不同层次的数据(表 2)。

建表语句:CREATE TABLE level ("dataPath" text, column1 int, value int, PRIMARY

KEY ("dataPath", column1)) ;

2.2.3 最新时刻表 用于存储各类数据的最新时刻信息,表名为 latestdatatime。利用最新时刻表,用户能通过客户端快速查找到最新数据文件名。用户根据完整索引(文件路径与最新数据文件名),例:T639/WIND/500/17030108.000,即可在"数据表"中获取到对应数据(表 3)。

建表语句:CREATE TABLE latestdatatime ("dataPath" text, column1 text, value text, PRIMARY KEY ("dataPath", column1)) ;

表 1 Cassandra **数据表结构**

Cassandra 术语	对应描述	数据类型	示例
ColumnFamily	模式名或实况名	字符串	T639
Row Key	原 samba 服务器上文件路径(从物理量开始)	字符串	WIND/500
Column name	文件名	字符串	17030108.000
Column Value	压缩后的文件内容	二进制数据	0101010010…

表 2 Cassandra **层次表结构**

Cassandra 术语	对应描述	数据类型	示例
ColumnFamily	模式或实况资料的层次	字符串	level
Row Key	文件路径(模式名或者实况名到层次的上一级目录)	字符串	T639/WIND
Column name	层次名	整数	500
Column Value	0(无意义)	整数	0

表 3 Cassandra **最新时刻表结构**

Cassandra 术语	对应描述	数据类型	示例
ColumnFamily	最新时刻表	字符串	latestdatatime
Row Key	文件路径(模式名+物理量+层次)	字符串	T639/WIND/500
Column name	通配符+扩展名	字符串	*08.000
Column Value	当前最新数据名称	字符串	17030108.000

3 存储系统性能测试

3.1 测试环境

选用 5 台相同配置的服务器用来搭建分布式存储系统。服务器操作系统为 Red Hat Enterprise Linux Server release 7.1,处理器参数为 Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10 GHz,主频为 2.1 GHz;内存大小为 256 GB;6 块 4 TB SATA 硬盘;服务器间通过万兆光纤连接。Cassandra 数据库版本为 2.2.5。

3.2 高可用性测试

由 5 个节点所组成分布式存储系统,其结构上具有如下特点。

(1)服务器双网卡绑定,即将两个物理网卡虚拟成一个逻辑网卡;提升服务器之间的传输带宽,实现网卡冗余。

(2)用于集群内部数据交换的两台万兆光纤交换机,采取级联方式,可互为备份。

(3)6 块 SATA 硬盘,其中 2 块做 RAID1,安装操作系统及软件;另外 4 块 4 TB 用作两个 RAID0,用于存储数据。

(4)服务器集群为环形结构,没有 master 节点,各节点功能完全一样。

按照表 4 中内容,对系统的基础设施层(包括网络设备、存储设备等)、软件层(数据库)进行测试,来验证系统的高可用性;从表中结论可知,系统中用于内部数据交换的光纤或网卡、交换机及任一 Cassandra 服务器故障,均不影响 MICAPS4 客户端调取数据。

表 4　Cassandra 集群系统高可用性测试

内容	方法	结论
集群内部数据通信异常	断开任一服务器上用于内部交换的光纤,模拟光纤/网卡故障,检查内部数据通信是否正常	正常
	关闭任一交换机,在数据处理服务器(DPC)上查看日志,检查解码后的数据是否写入 Cassandra 集群	正常
Cassandra 服务器故障	关闭任一台 Cassandra 服务器,模拟数据库或服务器故障,在 MICAPS4.0 客户端检查是否能正常调用资料,资料是否完整	正常

3.3　读取性能测试

通过读取数据的脚本文件(可获取数据字节数信息,表 5 中 ECMWF_HR/TMP/100 目录下数据字节数为 132 642 字节,SATELLITE/FY2E/L1/IR3/EQUAL 下数据字节数为554 944 字节,T639/WIND/100 下数据字节数为1 449 052 字节),模拟单用户及 50 用户、100 用户客户端对同一类型数据进行读取,共分 3 组,即对三种不同类型的数据进行测试,测试性能见表 5,注意测试结果包含网络传输时间。

表 5　三种不同类型的数据在不同压力下传输时间测试　　ms

数据类型	类别	单用户			50 用户并发			100 用户并发		
		最高	最低	平均	最高	最低	平均	最高	最低	平均
ECMWF_HR/TMP/100	获取最新	11.1	2.0	10.60	33.2	1.9	10.70	16.9	1.9	10.70
	读取文件	3.5	2.1	2.50	3.5	2.1	2.45	3.2	2.0	2.32
	左右翻页	39.3	3.8	5.40	8.8	2.7	4.88	8.4	1.0	4.54
ECMWF_HR/RH/100	获取最新	12.4	2.1	10.3	11.1	1.4	10.28	11.0	2.0	10.24
	读取文件	3.5	1.9	2.19	2.8	1.3	1.76	6.6	1.9	2.19
	左右翻页	35.4	3.9	4.65	7.4	2.7	4.55	7.0	2.8	4.54
T639/WIND/100	获取最新	11.1	2.1	10.70	11.0	2.2	10.40	11.3	2.0	10.70
	读取文件	45.8	16.5	20.00	73.6	18.9	20.16	46.9	16.5	20.40
	左右翻页	66.9	31.1	40.20	87.5	33.8	40.10	67.3	30.8	40.30
T639/SPFH/100	获取最新	11.4	2.1	10.75	11.1	2.1	10.70	31.0	2.0	10.75
	读取文件	8.2	2.9	6.77	7.7	3.0	6.85	7.5	3.0	6.77
	左右翻页	44.7	11.0	13.24	44.6	12.3	13.32	39.8	10.5	13.22
SATELLITE/FY2E/L1/IR3/EQUAL	获取最新	11.1	2.5	10.72	34.1	1.6	10.72	34.7	2.2	10.74
	读取文件	9.0	3.0	7.58	9.1	3.1	7.77	39.6	7.1	7.58
	左右翻页	46.6	10.7	12.46	60.7	11.1	12.16	38.5	10.9	11.82

从数据读取的测试结果可以看出:(1)50 用户并发和 100 用户并发客户端同时对同一类型数据进行读取的时间与单用户读取时间相当。以 T639/WIND/100 为例,50 用户并发和 100 用户并发与单用户读取相关数据所花费的平均时间均在 20 ms 左右。(2)在 100 用户并发情况下,从数据库中调取数据所消耗的时间均以 ms 量级为单位(包含网络传输时间),时间远远小于在 samba 服务器上读取数据的时间。(3)数据读取时间和单个数据的字节数近似成正比,即单个数据文件字节数越大,读取数据所花费的时间则越长。

蒋小莉,王凌军,雷延鹏,等. 洛川国家基准气候站二次迁站观测数据对比评估[J]. 陕西气象,2017(4):30-34.
文章编号:1006-4354(2017)04-0030-05

洛川国家基准气候站二次迁站观测数据对比评估

蒋小莉,王凌军,雷延鹏,李生袖
(延安市气象局,陕西延安 716000)

摘 要:对洛川国家基准气候站(53942)第二次迁站后的3号站与第一次迁站后的2号站,2015—2016年气温、降水量、相对湿度、平均风速等观测资料进行差异对比评估,并对两次迁站前后3个站点的观测资料进行显著性T检验。结果表明:3号站与2号站平均气温差异不明显,3号站最高气温高于2号站,最低气温低于2号站,相对湿度差异变化在2%左右,2分钟平均风速3号站小于2号站,风向一致性较差,受地形影响明显。迁站前后3个站点温度、相对湿度资料序列可合并使用;本站气压、降水量仅1号站(第一次迁站前的旧站)与2号站的资料序列可合并使用,但经气压高度差订正后的本站气压3个站点资料序列可合并使用;2分钟平均风速资料不可合并使用。3个站点降水量的序列分析待3号站降水量资料完善及序列增加后再行评估。

关键词:气象站迁移;观测数据;差异;对比评估
中图分类号:P416.2 **文献标识码**:A

洛川国家基准气候站始建于1954年11月1日,到2012年12月31日站址一直未变。建站近60 a来随着县域经济的不断发展,县城规模的不断扩大,使得原本处于"郊外"的洛川国家基准气候站(以下简称1号站),逐渐发展为人口活动密集的县城中心,其周边高大建筑物逐年增加,气象数据逐渐失去了代表性。为了全面改观地面观测环境,2013年1月1日洛川国家基准气候站由洛

收稿日期:2017-02-17

作者简介:蒋小莉(1962—),女,陕西合阳人,高级工程师,从事人工影响天气工作。

4 结语

利用Cassandra分布式数据库搭建的存储环境,提高了实时气象数据存储效率与检索速度,通过统一的数据平台,实现了市县两级运维人员对该系统"零"维护。通过在实际业务环境中进行测试,验证了该分布式数据环境的高可用性;以毫秒级为单位的数据读取时间,能很好地满足业务对数据时效性的需求。

参考文献:

[1] 王意洁,孙伟东,周松,等. 云计算环境下的分布存储关键技术[J]. 软件学报,2012,23(4):971-975.
[2] 申德荣,于戈,王习特,等. 支持大数据管理的NoSQL系统研究综述[J]. 软件学报,2013,24(8):1789-1792.
[3] 秦苻珂. Cassandra应用研究[J]. 电脑知识与技术,2016,12(9):14-16.
[4] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013,50(1):154-157.
[5] Eben Hewitt. Cassandra:The Definitive Guide[M]. 北京:人民邮电出版社,2011.
[6] 覃雄派,王会举,李芙蓉,等. 数据管理技术的新格局[J]. 软件学报,2013,24(2):179-185.
[7] 王若曈,黄向东,张博等. 海量气象数据实时解析与存储系统的设计与实现[J]. 计算机工程与科学,2015,37(11):2045-2054.
[8] 钟雨,黄向东,刘丹,等. 大规模装备检测数据的NoSQL存储方案[J]. 计算机集成制造系统,2013(12):3008-3016.