HBase使用CSS 计算存储设备性能测试报告

1. 测试总结
2. CPU GZIP在大HBaseJVM（100G）配置下大量写入数据的时候会经常失败导致任务，测试无法完成，使用较小的HBaseJVM（1G）配置可以完成测试，但是性能很差，所以在生成环境中不建议用CPU GZIP压缩。
3. 单节点分别安装单CSS SSD和4x SSD对比测试表明，使用CSS SSD的读写性能略高于CPU LZO，随着单条数据大小增大，优势更加明显，且CSS压缩可以再节省30%-40%的空间。
4. 在CSS SSD+HDD组成异构存储环境，可以充分发挥CSS SSD设备的写入优势，读取性能也能达到平均水准，CSS压缩也可以在HDD上照常使用，压缩率能保持正常水平。
5. 测试目的
   1. 目的：

ScaleFlux提供的CSSSSD产品为计算存储设备，除使用PCIE插槽提高SSD吞吐量外，在产品硬件集成GZ压缩能力，对比普通SATA接口SSD，除了在带宽方面的提升外，能够实现使用GZ压缩（更好的压缩率）情况下，不损失读写性能，本次测试目的如下：

第一，验证使用ScaleFlux的CSSSSD产品情况下，对比普通SATASSD的性能优势。

第二，在第一目的达成，确认CSSSSD产品优势情况下，采用CSSSSD+HDD硬盘异构存储方式部署HBase的读写性能

第三，验证使用CSSSSD做HBase BucketCache缓存对集群读取的影响。

1. 测试详情
   1. 测试1：普通读写测试

测试集群配置：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | Hadoop | 测试工具 | 发压机硬件配置 |
| 集群1：NN\*2  DN、RS\*3 | CPU：32C  内存：256G  硬盘：4\*SATASSD(800G) | CENTOS7.2 | Jdk1.7.067 | Hadoop2.7.3  Hbase1.2.6 | Standard YCSB  ScaleFlux  YCSB | CPU：32C  内存：256G  （非数据节点） |
| 集群2：NN\*2  DN、RS\*3 | CPU：32C  内存：256G  硬盘：1\*CSSSSD(3.2T) | CENTOS7.2 | Jdk1.7.067 | Hadoop2.7.3  Hbase1.2.6 | Standard YCSB  ScaleFlux  YCSB | CPU：32C  内存：256G  （非数据节点） |
|  |  |  |  |  |  |  |

1. 考虑到PCIE通道速率一般4倍于SATA接口，这里使用4块SSD硬盘对等测试
2. 厂商提供的YCSB 工具可以相对明显的体现数据的压缩率，这里为了保证公平，也同时采用标准版YCSB进行测试，
3. Hadoop的JVM相关配置使用京东256G服务器线上环境配置
4. 测试过程中为极限体现硬盘读取能力，关闭HBaseL1和L2缓存
5. 测试过程中集群1分别使用GZ和LZO两种压缩方式来进行对比，集群2则使用厂商的CSSGZ压缩方式
   1. 测试准备

测试表：预切分300个region,每个RS各100region；

测试铺底1亿条数据；

铺底数据和写入数据记录compaction结束后data目录大小以体现压缩情况，随机读和Read/update混合测试不记录大小

Standard YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群名 | Load(数据铺底) | | Read（随机读1000W数据) | Insert(写入1亿条数据) | | Read/update各50%混合（读写1000W条数据） |
| TPS(k/s) | Size(G) | TPS(k/s) | TPS(k/s) | Size(G) | TPS(k/s) |
| 集群1:GZ | 139 | 78.3 | 21 | 失败 |  | 39 |
| 集群1:LZO | 204 | 95.5 | 21.4 | 209 | 118.8 | 45 |
| 集群2:CSS | 212 | 80.2 | 18.8 | 203 | 82.3 | 33 |

ScaleFlux YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群名 | Load(数据铺底) | | Read（随机读1000W数据) | Insert(写入1亿条数据) | | Read/update各50%混合（读写1000W条数据） |
| TPS(k/s) | Size(G) | TPS(k/s) | TPS(k/s) | Size(G) | TPS(k/s) |
| 集群1:GZ | 97 | 43.3 | 26 | 失败 |  | 48 |
| 集群1:LZO | 169 | 67.5 | 10.55 | 168 | 82.5 | 32 |
| 集群2:CSS | 182.9 | 48.6 | 15 | 166 | 56.3 | 37 |

测试结论：

1. 在两种测试工具下，CSS设备写入优势明显高于GZ，略高于LZO，其中GZ多次写入失败导致无结果
2. 在两种测试工具下，关闭缓存后，三种压缩读取性能均下降明显，CSS设备随机读性能略弱于GZ，其中使用ScaleFlux产生的单挑数据记录较大的情况下，LZO读性能下降较明显
3. 压缩率上，GZ最优，CSS有10%的降低，LZO最次。
   1. 测试2：异构存储读写测试

测试集群配置：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | Hadoop | 测试工具 | 发压机硬件配置 |
| 集群2：NN\*2  DN、RS\*3 | CPU：32C  内存：256G  硬盘：1\*CSSSSD(3.2T)+2\*HDD | CENTOS7.2 | Jdk1.7.067 | Hadoop2.7.3  Hbase1.2.6 | Standard YCSB  ScaleFlux  YCSB | CPU：32C  内存：256G  （非数据节点） |
|  |  |  |  |  |  |  |

1. 考虑到在HDFS层采用ONE\_SSD方式配置存储策略，故HDD磁盘数量与CSSSSD为2:1
2. Hadoop的JVM相关配置使用京东256G服务器线上环境配置
3. 测试过程中为极限体现硬盘读取能力，关闭HBaseL1和L2缓存
   1. 测试准备1

测试表：预切分300个region,每个RS各100region；

测试铺底1亿条数据；

铺底数据和写入数据记录compaction结束后data目录大小以体现压缩情况，随机读和Read/update混合测试不记录大小

HBase库的hadoopjar包中hdfs和common版本升级至2.7.3

配置HDFS的/hbase/data目录为ONE\_SSD模式（一个副本写入SSD设备）

异构模式1：配置HBase的WAL写入方式为ALL\_SSD方式（全部副本写入SSD设备）

Standard YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群名 | Load(数据铺底) | | Read（随机读1000W数据) | Insert(写入1亿条数据) | | Read/update各50%混合（读写1000W条数据） |
| TPS(k/s) | Size(G) | TPS(k/s) | TPS(k/s) | Size(G) | TPS(k/s) |
| 集群2:CSS | 212 | 80.2 | 18.8 | 203 | 82.3 | 33 |
| 集群2:CSS（异构1） | 236 | 78.1 | 19.9 | 216 | 84.6 | 35 |

ScaleFlux YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群名 | Load(数据铺底) | | Read（随机读1000W数据） | Insert(写入1亿条数据) | | Read/update各50%混合（读写1000W条数据） |
| TPS(k/s) | Size(G) | TPS(k/s) | TPS(k/s) | Size(G) | TPS(k/s) |
| 集群2:CSS | 182.9 | 48.6 | 15 | 166 | 56.3 | 37 |
| 集群2:CSS（  异构1） | 207 | 46.4 | 16 | 194 | 53 | 19.8 |

测试结论：

1. 在两种测试工具下，异构模式1对比普通全SSD模式，写入性能差别不大，读取性能差别不明显,但在ScaleFluxYCSB测试中，读写混合场景中异构模式1性能下降较多
   1. 测试准备2

测试表：预切分300个region,每个RS各100region；

测试铺底1亿条数据；

铺底数据和写入数据记录compaction结束后data目录大小以体现压缩情况，随机读和Read/update混合测试不记录大小

HBase库的hadoopjar包中hdfs和common版本升级至2.7.3

异构模式2：

配置HBase的WAL写入方式为ALLSSD方式（所有副本写入SSD设备）

配置HDFS的/hbase/data目录存储策略为ONE\_SSD(一个副本写入SSD设备)

Standard YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群名 | Load(数据铺底) | | Read（随机读1000W数据) | Insert(写入1亿条数据) | | Read/update各50%混合（读写1000W条数据） |
| TPS(k/s) | Size(G) | TPS(k/s) | TPS(k/s) | Size(G) | TPS(k/s) |
| 集群2:CSS | 212 | 80.2 | 18.8 | 203 | 82.3 | 33 |
| 集群2:CSS（异构2） | 225 | 80.3 | 20.9 | 202 | 80.9 | 37 |

ScaleFlux YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群名 | Load(数据铺底) | | Read（随机读1000W数据） | Insert(写入1亿条数据) | | Read/update各50%混合（读写1000W条数据） |
| TPS(k/s) | Size(G) | TPS(k/s) | TPS(k/s) | Size(G) | TPS(k/s) |
| 集群2:CSS | 182.9 | 48.6 | 15 | 166 | 56.3 | 37 |
| 集群2:CSS（异构2） | 199 | 48.6 | 18 | 180 | 53 | 53 |

测试结论：

1. 在两种测试工具下，异构模式2对比普通全SSD模式，写入性能差别不大，读取性能差别依然不明显,但在ScaleFluxYCSB测试中，读写混合场景中，异构模式2性能有较大提高
   1. 测试3：CSS缓存随机读测试

测试集群配置：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | Hadoop | 测试工具 | 发压机硬件配置 |
| 集群1：NN\*2  DN、RS\*3 | CPU：32C  内存：256G  硬盘：4\*SATASSD(800G) | CENTOS7.2 | Jdk1.7.067 | Hadoop2.7.3  Hbase1.2.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |
| 集群2：NN\*2  DN、RS\*3 | CPU：32C  内存：256G  硬盘：1\*CSSSSD(3.2T) | CENTOS7.2 | Jdk1.7.067 | Hadoop2.7.3  Hbase1.2.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |
|  |  |  |  |  |  |  |

1. 开启HBase读取缓存，在集群1中设置bucketCache为offHeap，在集群2中分两次设置bucketCache为offHeap和CSSSSD
2. 为了追求最佳读取性能，本次测试只使用standard YCSB进行测试
   1. 测试准备

测试表：预切分300个region,每个RS各100region；

测试铺底10亿条数据；

随机读取5亿条数据，并各进行四次读取，以保证缓存充分使用

Standard YCSB测试结果：

|  |  |  |  |
| --- | --- | --- | --- |
| 集群名 | Read（随机读5Y数据) | BlockCacheHits次数 | BucketCacheHit次数 |
| TPS(k/s) | （亿次/节点） | （亿次/节点） |
| 集群1:offHeap | 157 | 10 | 6 |
| 集群2:offHeap | 171 | 18 | 7 |
| 集群2:CSS | 172 | 16 | 7 |

测试结论：

1. 缓存对HBase读取性能有较大影响，使用CSS做L2缓存后读取性能没有特别明显的变化

测试尚存在的问题:

1. 在100% select和50% select 50% update的测试中，没有看见产生很大的IO，所以这时候不是存储的瓶颈,后续可以考虑加大测试压力
2. 可以考虑开放L1和L2 cache后继续进行读写测试，利于对设备产生更大压力
3. 考虑到CSS存储设备在GZ压缩上的优势,可以考虑设计有更多Compaction事物的场景，验证对CPU负载的影响
4. 考虑进一步验证使用CSS作为L2缓存对复杂读取场景的影响