对比实验报告

1. 简介
   1. 目的

本次对比实验的目的是比较基于区块链的数据共享系统与hadoop在不同数量级的数据下查询的性能。

* 1. 范围

本次对比实验只比较目标的查询性能，不考虑增，删，改的性能。

* 1. 实验对象介绍

基于区块链的数据共享系统本质上是一个分布式数据库，但是结合了区块链技术，将数据表和节点信息存储于智能合约中。

Hadoop实现了一个分布式文件系统（Hadoop Distributed File System），将数据文件分布在多个节点中，并使用map-reduce对数据进行操作。

1. 实验设计
   1. 实验数据

本次实验的数据由3000条体检数据扩充而成。一共扩充成3组数据，规模分别为100万条，500万条，1000万条。

* 1. 实验环境

硬件环境均为i5 cpu，16G内存。

基于区块链的数据共享系统为3个节点，数据平均分布。

hadoop为一个主节点，三个从节点，数据按照默认配置分布。

* 1. 实验方法

使用以下4种sql语句测试两个目标系统的查询性能。

|  |  |  |
| --- | --- | --- |
| 编号 | Sql语句 | 说明 |
| A | Select count(1) from ehr\_data | 查询数据的总条数 |
| B | Select bz10 from ehr\_data | 获取bz10列的数据，每个为5个汉字以内的数据 |
| C | Select bz83 from ehr\_data | 获取bz83列的数据，每个为20个汉字以上的数据 |
| D | Select \* from ehr\_data where bz103=7 | 根据条件查询数据 |

1. 实验结果
   1. Hadoop

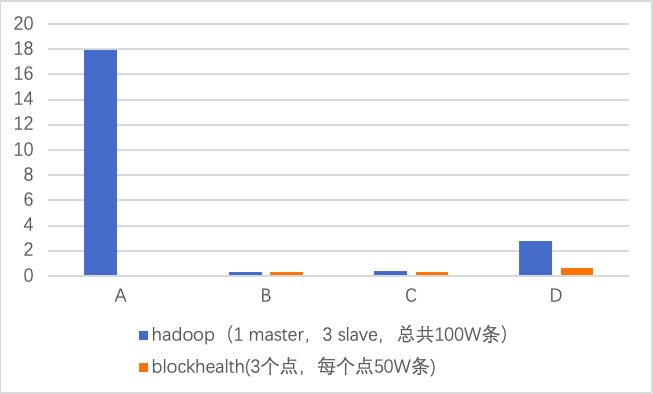
|  |  |  |  |
| --- | --- | --- | --- |
| sql | 耗时/秒 | | |
| 100万 | 500万 | 1000万 |
| Select count(1) from ehr\_data | 17.962 | 26.412 | 57.459 |
| Select bz10 from ehr\_data | 0.33 | 1.489 | 3.482 |
| Select bz83 from ehr\_data | 0.371 | 1.517 | 5.513 |
| Select \* from ehr\_data where bz103=7 | 2.768 | 18.656 | 30.279 |

* 1. 基于区块链的数据共享系统

|  |  |  |  |
| --- | --- | --- | --- |
| sql | 耗时/秒 | | |
| 100万 | 500万 | 1000万 |
| Select count(1) from ehr\_data | 0.058 | 0.839 | 0.397 |
| Select bz10 from ehr\_data | 0.29 | 1.48 | 2.146 |
| Select bz83 from ehr\_data | 0.309 | 1.526 | 2.534 |
| Select \* from ehr\_data where bz103=7 | 0.595 | 2.368 | 4.61 |

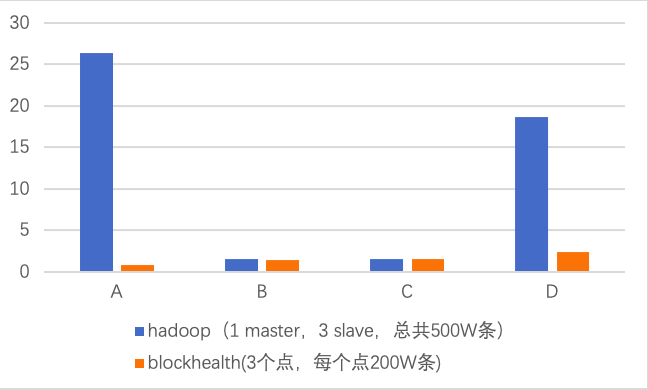
* 1. 结果对比

100万数量级数据对比结果如下：



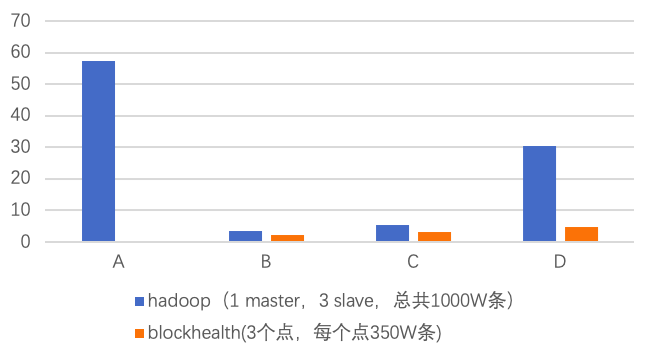
从图中可以看出，hadoop对于count(1)操作效率十分低（图中的A），因为它需要遍历每一条数据才能得到结果。在选取某一列的查询操作中（图中的B和C），hadoop和基于区块链的数据共享系统的性能差不多。在根据条件查询数据的操作中（图中的D），在两边的系统都没有索引的情况下，基于区块链的数据共享系统的性能要优于hadoop。

500万数量级数据对比结果如下：



从图中可以看出，500万数据级的比对结果和100万数量级差不多。

1000万数量级数据对比结果如下：



从图中可以看出，在查询某一列数据的操作中，基于区块链的数据共享系统的性能已经优于hadoop。

1. 实验总结

在本次实验中，基于区块链的数据共享系统的性能在各个数量级的查询性能都要优于hadoop，而hadoop是当下最为热门的大数据分布式存储方案，说明本系统在性能上有较强的竞争力。不过本次实验还存在如下几个缺陷：

1. Hadoop的配置全部为默认配置，没有进行优化，可能导致性能不理想。在后续工作中会尝试对hadoop进行优化，再进行对比实验。
2. 基于区块链的数据共享系统的数据全都是平均分布，并不符合实际情况。而本系统的性能瓶颈在于查询速度最慢的节点，所以实际应用的表现可能会变差。在后续工作中会继续考虑这方面的解决方案。

尽管本系统还存在缺陷，但它通过去中心化和权限验证来保证了数据的安全性，性能也比较优秀，对于数据敏感的行业还是有很大的价值的。