云计算大数据处理分布式数据库

——数据立方与 Hive 对比测试报告

一、目的

最近进行了云创存储的数据立方性能测试，并将其与开源数据仓库 Hive 进行了对比，从而得出在廉价的服务器上两者的性能测试结果。

二、测试内容

本次测试主要从数据查询方面进行对比测试，具体测试内容如下：

1.统计单表记录数

测试项	序号	测试内容	执行 SQL
hive	1	统计查询 800W 条	select count(*)from e_mp_power_curve;
	2	统计查询 1000W 条	select count(*) from a_tmnl_task;
数据立方	1	统计查询 800W 条	select count(*)from e_mp_power_curve;
	2	统计查询 1000W 条	select count(*) from a_tmnl_task;

2.查询单表字段数据

测试项	序号	测试内容	执行 SQL
hive	1	查询 e_mp_power_curve 表	select id,DATA_TYPE,DATA_POINT_FLAG,DATA_WHOLE_FLAG from e_mp_power_curve where id=100001100;
	2	查询 a_tmnl_task	select TMNL_TASK_ID,TERMINAL_ID,TASK_ID from a_tmnl_task where tmnl_task_id=100001000;
数据立方	1	查询 e_mp_power_curve 表	select id,DATA_TYPE,DATA_POINT_FLAG,DATA_WHOLE_FLAG from e_mp_power_curve where id=100001100;
	2	查询 a_tmnl_task	select TMNL_TASK_ID,TERMINAL_ID,TASK_ID from a_tmnl_task where tmnl_task_id=100001000;

3.两表 join 查询

测试项	序号	测试内容
hive	1	A 表 1000W 数据，B 表 800W,两表 join
	2	A 表 1000W 数据，B 表 1000W,两表 join

数据立方	1	查询 e_mp_power_curve 表
	2	查询 a_tmnl_task

4.三表 join 查询

测试项	序号	测试内容
hive	1	A 表 1000W 数据，B 表 800W,C 表 800W 三表 join
数据立方	1	A 表 1000W 数据，B 表 800W,C 表 800W 三表 join

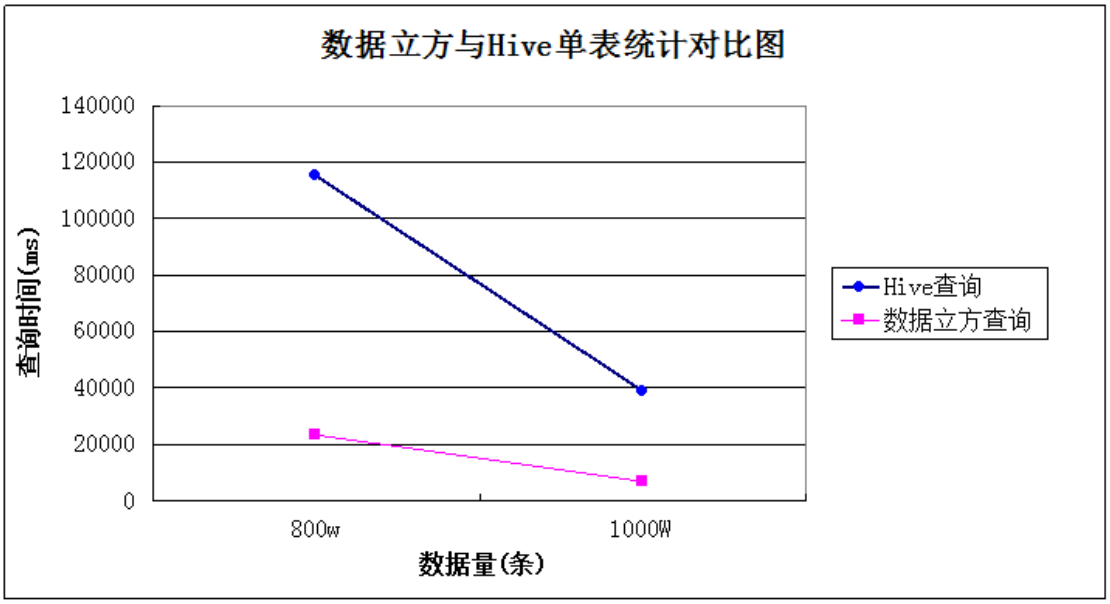
三．测试环境

	Hive	数据立方
控制节点数量(台)	1	1
处理节点数量（台）	9	9
cpu	IntelE5-2620 2.0G/15M/6C	IntelE5-2620 2.0G/15M/6C
内存	32G	32G
网络	千兆以太网	千兆以太网
硬盘	3T×2	3T×2
软件版本	hive-0.9.0-cdh4.1.2	datacube-1.0

四．测试结果

1.单表数据量查询性能对比测试

产品名称	数据量(条)	查询时间(ms)
hive	800W	115114
	1000W	38887
数据立方	800W	23340
	1000W	6910

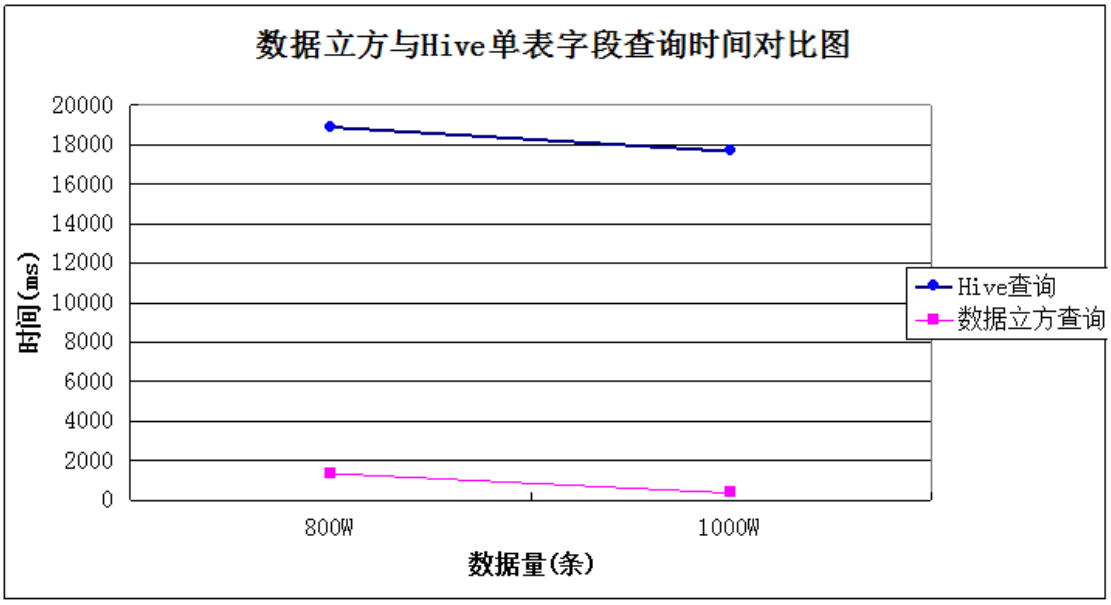


通过相同环境下测试可以看出：数据立方单表统计查询速度是 Hive 的至少 3-5 倍。

2.单表字段查询性能对比测试

产品名称	数据量(条)	查询时间(ms)
hive	800W	17659

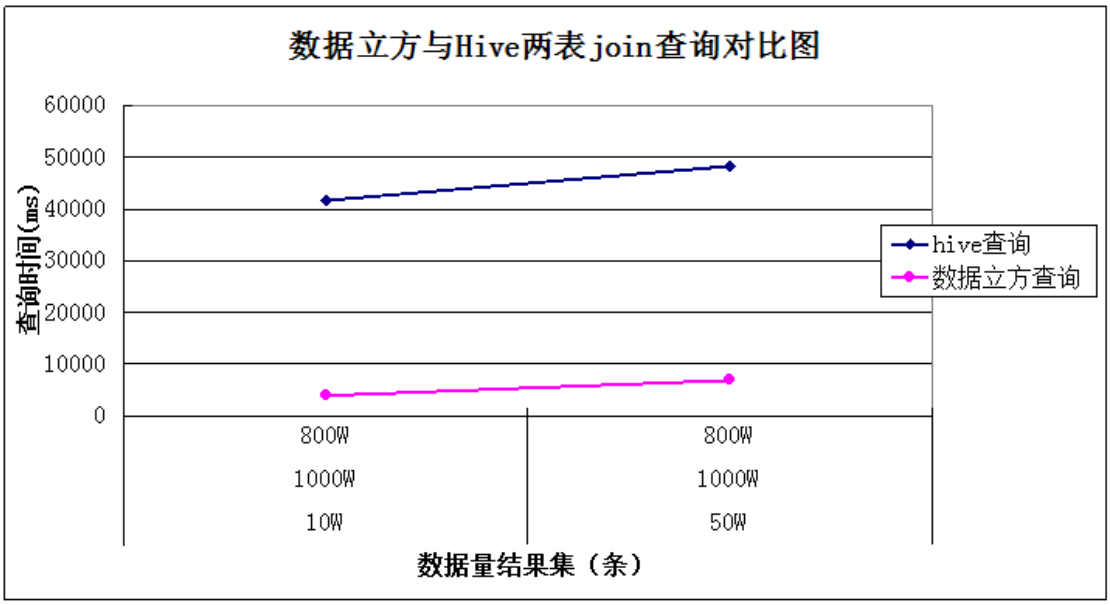
	1000W	18883
数据立方	800W	370
	1000W	1360



从数据图可以看到单表字段查询时数据立方速度是 hive 的 5-10 倍。

3.两表 join 查询性能对比测试

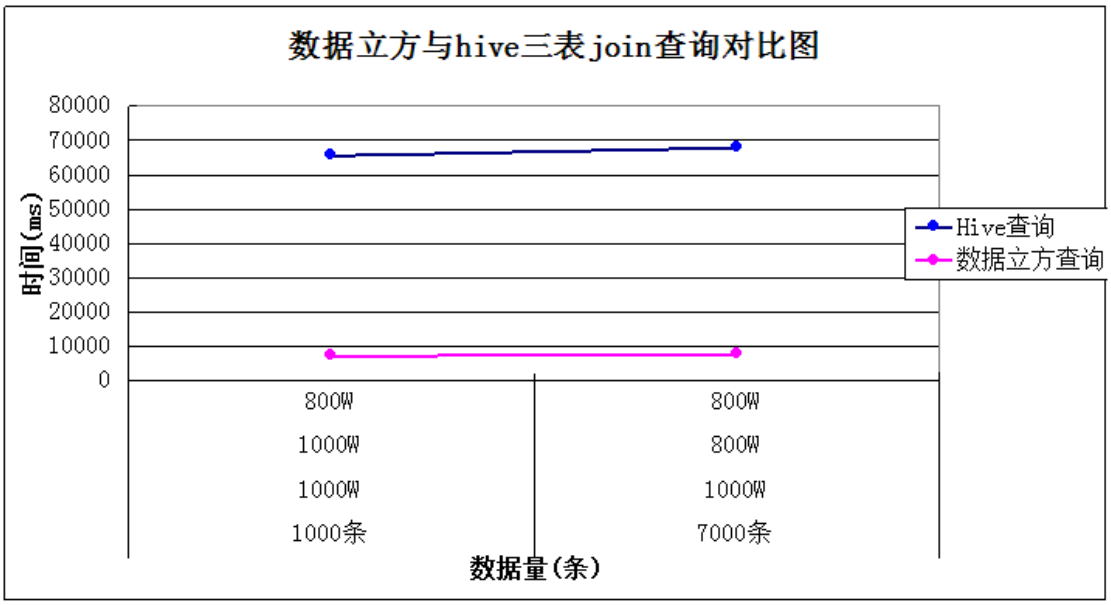
结果集(条)	A 表	B 表	hive 查询时间(ms)	数据立方查询时间(ms)
10W	1000W	800W	41667	3950
50W	1000W	800W	48296	6880



从数据图可以看到两表 join 查询时数据立方速度是 hive 的 6-10 倍以上。

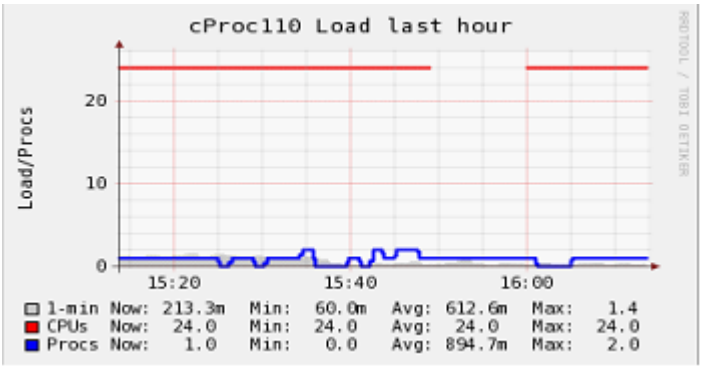
4.三表 join 查询性能对比测试

结果集	A 表	B 表	C 表	数据立方查询时间(ms)	数据立方查询时间(ms)
1000条	1000W	1000W	800W	65550	7170
7000条	1000W	800W	800W	67891	7710

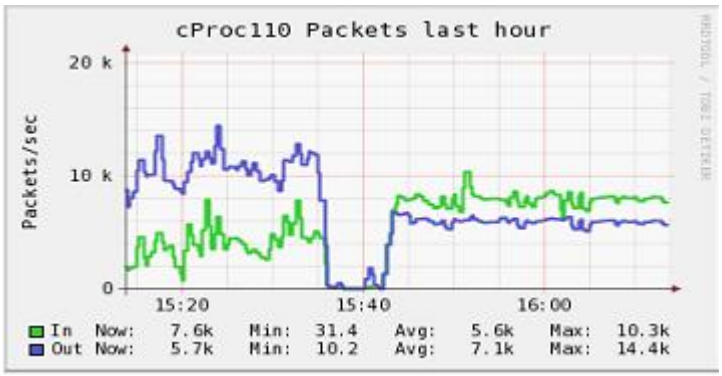


从数据图可以看到三表 join 查询时数据立方速度是 hive 速度的几乎 10 倍。

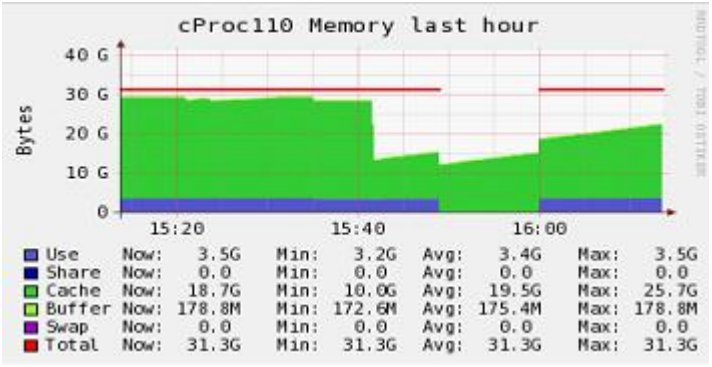
5.GangLia 监控截图：



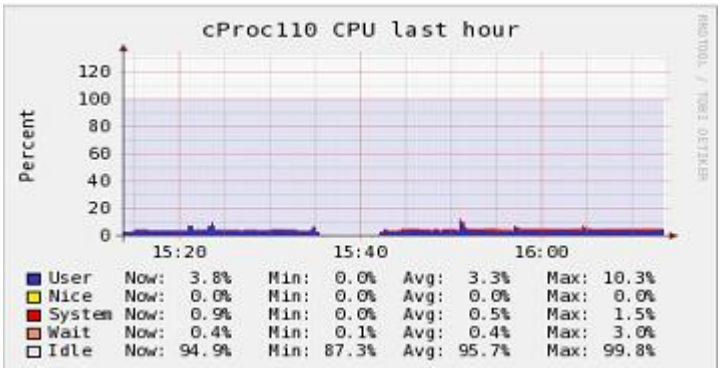
1 个小时内 CPU 负载情况图



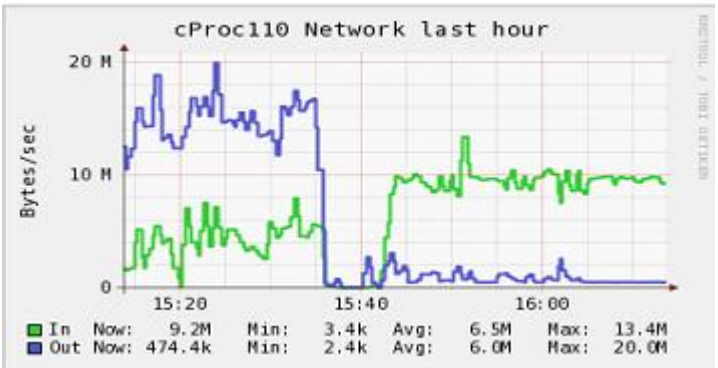
1 个小时内内存使用情况图



1 个小时内 CPU 负载情况图



1 个小时内网络负载情况图



1 个小时内数据包流量图

五.测试总结

本次在一体机上分别对云创存储的数据立方、开源数据仓库 Hive 在不同数据量情况下，进行了单表数据量查询性能对比测试，单表字段查询性能对比测试，两表 join 查询性能对比测试,三表 join 查询性能对比测试,从测试结果来看，数据立方在数据查询方面都优越于开源数据仓库 Hive，数据立方的查询速度一般是 Hive 的 3-10 倍。证明了其优异的性能。具体性能结果请见第四章节测试结果部分。