Hash join 性能测试报告

修订历史

版本	修订日期	修订描述	作者	备注
Cedar	2017-09-28	Hash Join性能测试报告	胡爽	无

测试环境

使用6台虚拟机组成的集群作为测试环境，每台虚拟机的配置相同，包括4核1.2GHz主频CPU、100GB内存、3000GB磁盘，虚拟机上安装了CentOS release 6.5系统，相互之间通过千兆以太网连接。其中三台虚拟机环境，每台部署一个数据库集群，包括UpdateServer, RootServer,ChunkServer以及MergeServer。剩下三台每个环境搭建一个MergeServer。

测试案例一

1.案例描述

测试join的深度对性能的影响。

2.测试方法

- 2张500w的表inner join、3张500w的表inner join、4张500w的表inner join、5张500w的表inner join、6张500w的表inner join。
- 2张1000w的表inner join、3张1000w的表inner join、4张1000w的表inner join、5张1000w的表inner join、6张1000w的表inner join

3.测试结果

500w数据量的测试结果如下图：

	2	3	4	5	6
mg join	31.21	46.45	68.65	79.44	103.26
bl join	38.33	57.37	79.67	89.98	114.29
hash join	32.18	35.38	39.28	45.84	49.92

1000w数据量的测试结果如下图：

	2	3	4	5	6
mg join	98.69	165.21	210.27	254.87	318.87
bl join	100.12	167.28	197.58	289.74	368.28
hash join	97.87	112.12	113.28	120.34	122.47

测试结果：随着join深度增加，Merge Join和Bloomfilter Join所费时间大幅度上升，而Hash Join所耗时间缓慢上升，从图一二可以看出性能最好。

测试案例二

1.案例描述

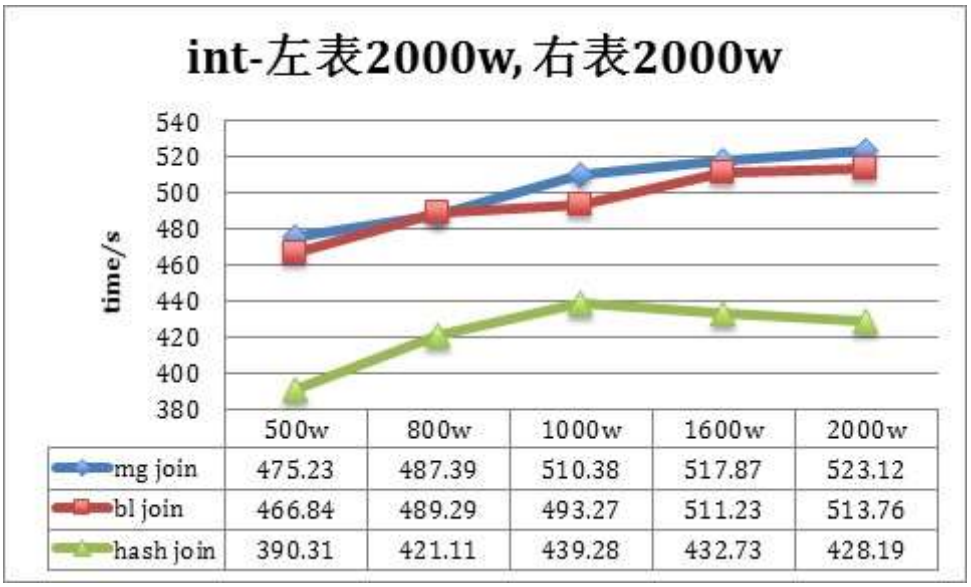
测试Hash Join在选择率为10%的情况下，非重复值个数对性能的影响（连接属性为int类型）。

2.测试方法

左右表都为2000万条数据，左表的非重复值个数为（500w,800w,1000w,1600w,2000w）。

3.测试结果

测试结果如下图：



	500w	800w	1000w	1600w	2000w
mg join	475.23	487.39	510.38	517.87	523.12
bl join	466.84	489.29	493.27	511.23	513.76
hash join	390.31	421.11	439.28	432.73	428.19

测试结果：Join表数目为2000w时，随着左表非重复数的增加，Hash Join所耗时间远低于merge join和bloom filter join，从图三看出hash join性能远好于Bloom filter Join 和Merge Join。

测试案例三

1.案例描述

测试Hash Join在非重复值个数为500万下，选择率对性能的影响（连接属性为int类型）。

2.测试方法

左表1000万条数据（非重复数500万，选择率0.1%，1%，10%，30%，60%，90%）

3.测试结果

	0.10%	1%	10%	30%	60%	90%
mg join	120.12	113.23	124.12	139.89	134.34	147.86
bl join	132.38	130.43	138.67	131.88	142.19	154.39
hash join	103.77	105.87	109.34	107.93	110.48	129.75

在非重复值个数为500w时，Join左右表数目为1000万时，Hash Join所耗时间低于merge join和bloom filter join，从图四看出hash join性能远好于Bloom filter Join 和Merge Join

测试案例四

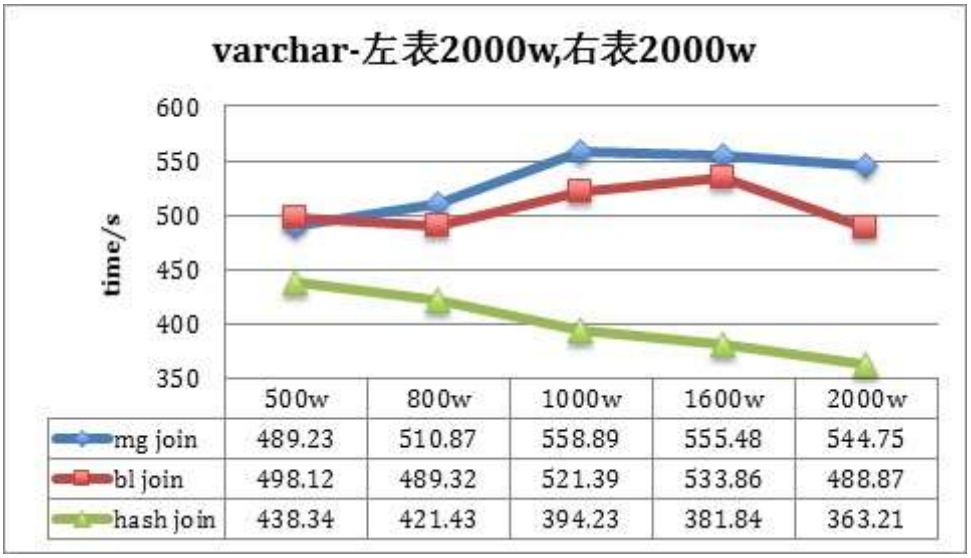
1.案例描述

测试Hash Join在选择率为10%的情况下，非重复值个数对性能的影响（连接属性为varchar类型）。

2.测试方法

左右表2000万条数据（非重复数500w,800w,1000w,1600w,2000w，连接属性为varchar类型）

3.测试结果



	500w	800w	1000w	1600w	2000w
mg join	489.23	510.87	558.89	555.48	544.75
bl join	498.12	489.32	521.39	533.86	488.87
hash join	438.34	421.43	394.23	381.84	363.21

在将左表的非重复值个数增加到（500w,800w,1000w,1600w,2000w）时，Join表数目为2000w时，Hash Join性能远好于Bloom filter Join 和Merge Join。即从图看出，当连接属性为varchar类型时，相较于Bloom filter Join 和Merge Join，Hash Join性能更优。

测试案例五

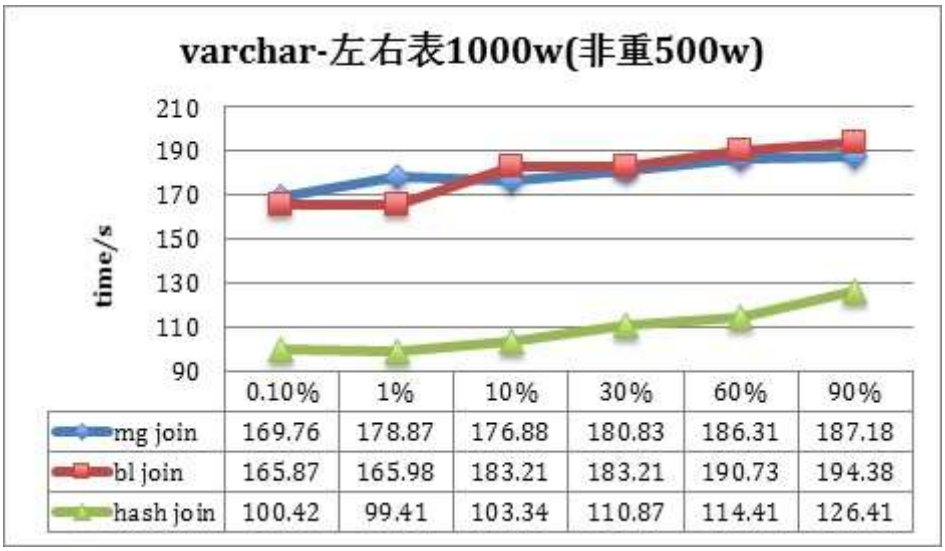
1.案例描述

测试Hash Join在非重复值个数为500万下，选择率对性能的影响（连接属性为varchar类型）。

2.测试方法

左右表1000万条数据（非重复数500万，选择率0.1%，1%，10%，30%，60%，90%）

3.测试结果



	0.10%	1%	10%	30%	60%	90%
mg join	169.76	178.87	176.88	180.83	186.31	187.18
bl join	165.87	165.98	183.21	183.21	190.73	194.38
hash join	100.42	99.41	103.34	110.87	114.41	126.41

连接属性为varchar类型时，随着选择率的升高，相较于Bloom filter Join 和Merge Join，Hash Join性能更优。从图六看出，Hash Join性能远好于Bloom filter Join 和Merge Join。

测试结论

由上述实验分析可知，在大数据量的情况下，hash Join与merge Join，bloom filter join相比性能有明显的优势。