测试说明

1. 测试环境

1.1 云平台虚拟机

虚拟机配置 2.39GHZ 8cpu，16g memory，500g disk，CentOS release 6.5

1.2 Hadoop集群

云平台虚拟机8台，hadoop2.6.3

1.3 Kafka集群

云平台虚拟机3台，Kafka0.9.1

1.4 Spark集群

云平台虚拟机8台，Spark1.6.2

2. 测试数据

基于TPC-H中的lineitem表和order表。测试数据的字段包括lineitem表中的所有字段以及关联的order表中的custkey字段。查询测试数据大小为100GB。

3. 测试结果（具体数据见表格）

3.1 数据装载测试（见sheet ‘Loader Performance’）

3.1.1 kafka producer线程数量对吞吐的影响。kafka topic的partition数量为160，batch size为10000，缓冲区大小为10000。测试中线程数量增加了，吞吐反而下降，主要应该为阻塞队列的影响，后面需要对这部分代码做进一步修改。

3.1.2 kafka batch size对吞吐的影响。

3.1.3 缓冲区大小对吞吐的影响。

3.1.4 kafka topic partition数量与kafka broker数量对吞吐的影响。

3.2 查询测试（见sheet ‘Query Performance’）

3.2.1 查询1

查询语句：

```
SELECT sum(quantity) AS sum__qty, sum(extendedprice) AS sum__base__price, avg(quantity) AS avg__qty, avg(extendedprice) AS avg__price, avg(discount) AS avg__disc, count(*) AS count__order, min(orderkey) AS min__orderkey, max(orderkey) AS max__orderkey from realtime where messagedate 〈 %s and messagedate 〉 %s
```

查询选择率分别为%1, %2, %5, %10，根据时间范围选定。

3.2.2 查询2

查询语句：

```
SELECT sum(quantity) AS sum_qty, sum(extendedprice) AS sum_base_price, avg(quantity) AS avg_qty, avg(extendedprice) AS avg_price, avg(discount) AS avg_disc, count(*) AS count_order, min(orderkey) AS min_orderkey, max(orderkey) AS max_orderkey from realtime where custkey = 1348000 AND messagedate > %s and messagedate < %s
```

查询选择率根据查询1中的时间范围选定，再在此基础上根据某一个fiber id过滤。