1 hive表关联查询，如何解决数据倾斜的问题?

|  |
| --- |
| 倾斜原因：  map输出数据按key Hash的分配到reduce中，由于key分布不均匀、业务数据本身的特、建表时考虑不周、等原因造成的reduce 上的数据量差异过大。  1)、key分布不均匀;  2)、业务数据本身的特性;  3)、建表时考虑不周;  4)、某些SQL语句本身就有数据倾斜;  如何避免：对于key为空产生的数据倾斜，可以对其赋予一个随机值。  解决方案  1>.参数调节：  hive.map.aggr = true  hive.groupby.skewindata=true  有数据倾斜的时候进行负载均衡，当选项设定位true,生成的查询计划会有两个MR Job。第一个MR Job中，Map的输出结果集合会随机分布到Reduce中，每个Reduce做部分聚合操作，并输出结果，这样处理的结果是相同的Group By Key有可能被分发到不同的Reduce中，从而达到负载均衡的目的；第二个MR Job再根据预处理的数据结果按照Group By Key 分布到 Reduce 中（这个过程可以保证相同的 Group By Key 被分布到同一个Reduce中），最后完成最终的聚合操作。  2>.SQL 语句调节：  1)、选用join key分布最均匀的表作为驱动表。做好列裁剪和filter操作，以达到两表做join 的时候，数据量相对变小的效果。  2)、大小表Join：  使用map join让小的维度表（1000 条以下的记录条数）先进内存。在map端完成reduce.  4)、大表Join大表：  把空值的key变成一个字符串加上随机数，把倾斜的数据分到不同的reduce上，由于null 值关联不上，处理后并不影响最终结果。  5)、count distinct大量相同特殊值:  count distinct 时，将值为空的情况单独处理，如果是计算count distinct，可以不用处理，直接过滤，在最后结果中加1。如果还有其他计算，需要进行group by，可以先将值为空的记录单独处理，再和其他计算结果进行union。 |

2. 请谈一下hive的特点是什么？hive和RDBMS有什么异同？

|  |
| --- |
| hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供完整的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。其优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。 |

3. 请把下一语句用hive方式实现？

|  |
| --- |
| SELECT a.key,a.value  FROM a  WHERE a.key not in (SELECT b.key FROM b)  答案：  select a.key,a.value from a where a.key not exists (select b.key from b) |

4. Multi-group by 是hive的一个非常好的特性，请举例说明？

|  |
| --- |
| from A  insert overwrite table B   select A.a, count(distinct A.b) group by A.a  insert overwrite table C    select A.c, count(distinct A.b) group by A.c |

5. 请说明hive中 Sort By，Order By，Cluster By，Distrbute By各代表什么意思。

|  |
| --- |
| order by：会对输入做全局排序，因此只有一个reducer（多个reducer无法保证全局有序）。只有一个reducer，会导致当输入规模较大时，需要较长的计算时间。  sort by：不是全局排序，其在数据进入reducer前完成排序。  distribute by：按照指定的字段对数据进行划分输出到不同的reduce中。  cluster by：除了具有 distribute by 的功能外还兼具 sort by 的功能。 |

6.简要描述数据库中的 null，说出null在hive底层如何存储，并解释selecta.\* from t1 a left outer join t2 b on a.id=b.id where b.id is null; 语句的含义

|  |
| --- |
| null与任何值运算的结果都是null, 可以使用is null、is not null函数指定在其值为null情况下的取值。  null在hive底层默认是用'\N'来存储的，可以通过alter table test SET SERDEPROPERTIES('serialization.null.format' = 'a');来修改。  查询出t1表中与t2表中id相等的所有信息。 |

7.写出hive中split、coalesce及collect\_list函数的用法（可举例）。

|  |
| --- |
| Split将字符串转化为数组。  split('a,b,c,d' , ',') ==> ["a","b","c","d"]  COALESCE(T v1, T v2, …) 返回参数中的第一个非空值；如果所有值都为 NULL，那么返回NULL。  collect\_list列出该字段所有的值，不去重  select collect\_list(id) from table; |

8.写出将 text.txt 文件放入 hive 中 test 表‘2016-10-10’ 分区的语句，test 的分区字段是 l\_date。

|  |
| --- |
| LOAD DATA LOCAL INPATH '/your/path/test.txt' OVERWRITE INTO TABLE test PARTITION (l\_date='2016-10-10') |