1. Hive 的 sort by 和 order by 的区别

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.
因此，如果用sort by进行排序，并且设置mapred.reduce.tasks>1， 则sort by只保证每个reducer的输出有序，不保证全局有序。

2. Hbase 和 hive

有什么区别hive 与 hbase 的底层存储是什么？hive是产生的原因是什么？habase是为了弥补hadoop的什么缺陷？

答案

共同点：

1. hbase与hive都是架构在hadoop之上的。都是用hadoop作为底层存储

区别：

2. Hive是建立在Hadoop之上为了减少MapReducejobs编写工作的批处理系统，HBase是为了支持弥补Hadoop对实时操作的缺陷的项目 。

3. 想象你在操作RMDB数据库，如果是全表扫描，就用Hive+Hadoop,如果是索引访问，就用HBase+Hadoop 。

4. Hive query就是MapReduce jobs可以从5分钟到数小时不止，HBase是非常高效的，肯定比Hive高效的多。

5. Hive本身不存储和计算数据，它完全依赖于HDFS和MapReduce，Hive中的表纯逻辑。

6. hive借用hadoop的MapReduce来完成一些hive中的命令的执行

7. hbase是物理表，不是逻辑表，提供一个超大的内存hash表，搜索引擎通过它来存储索引，方便查询操作。

8. hbase是列存储。

9. hdfs作为底层存储，hdfs是存放文件的系统，而Hbase负责组织文件。

10. hive需要用到hdfs存储文件，需要用到MapReduce计算框架。

3. 数据仓库hive中，启动hive服务器的命令有哪些？分别代表什么意思？内部表与外部表有啥区别？分区与分桶，指的是什么？

命令：

hive --service metastore　　启动元数据

hive：本地运行hive命令

hiveserver2：远程服务，开放默认端口 10000

内部表：内部表删除表时，数据也会被删除，

外部表：外部表在创建时需要加external，删除表时，表中的数据仍然会存储在hadoop中，不会丢失

分区：分文件夹：分目录，把一个大的数据集根据业务需要分割成小的数据集

分桶：分数据：分桶是将数据集分解成更容易管理的若干部分

原文链接：

https://blog.csdn.net/pingsha_luoyan/article/details/97750251

4. hive中集合数据类型什么？有什么作用？什么情况下，hive需要使用集合类型?

数据类型：
6个基本类型：整数，布尔类型，浮点数，字符，时间类型。字节数组
2个集合数据类型： struct，map，array