# Hive中order by,Sort by,Distribute by和Cluster By介绍

我们在使用Hive的时候经常会使用到order by、Sort by、Distribute by和Cluster By，本文对其含义进行介绍。

## order by

Hive中的order by和数据库中的order by 功能一致，按照某一项或者几项排序输出，可以指定是升序或者是降序排序。它保证全局有序，但是进行order by的时候是将所有的数据全部发送到一个Reduce中，所以在大数据量的情况下可能不能接受，最后这个操作将会产生一个文件。

## sort by

sort by只能保证在同一个reduce中的数据可以按指定字段排序。使用sort by 你可以指定执行的reduce个数（set mapreduce.job.reduce=）这样可以输出更多的数据。对输出的数据再执行归并排序，即可以得到全部结果。需要注意的是，N个Reduce处理的数据范围是可以重叠的，所以最后排序完的N个文件之间数据范围是有重叠的。

## distribute by

按照指定的字段 将数据划分到不同的输出reduce中，这使用在本博客的《Hive：解决Hive创建文件数过多的问题》有详细的介绍。这个可以保证每个Reduce处理的数据范围不重叠，每个分区内的数据是没有排序的。

## cluster By

cluster by 除了具有 distribute by 的功能外还兼具 sort by 的功能。 所以最终的结果是每个Reduce处理的数据范围不重叠，而且每个Reduce内的数据是排序的，而且可以打到全局有序的结果。

## 英文解释

下面是英文对各个词的介绍

1、ORDER BY x: guarantees global ordering, but does this by pushing all data through just one reducer. This is basically unacceptable for large datasets. You end up one sorted file as output.

2、SORT BY x: orders data at each of N reducers, but each reducer can receive overlapping ranges of data. You end up with N or more sorted files with overlapping ranges.

3、DISTRIBUTE BY x: ensures each of N reducers gets non-overlapping ranges of x, but doesn't sort the output of each reducer. You end up with N or unsorted files with non-overlapping ranges.

4、CLUSTER BY x: ensures each of N reducers gets non-overlapping ranges, then sorts by those ranges at the reducers. This gives you global ordering, and is the same as doing (DISTRIBUTE BY x and SORT BY x). You end up with N or more sorted files with non-overlapping ranges.