Hive的cluster by 、 sort by、 distribute by 、 order by 区别?

可回答：1）Hive的排序函数；2）Hive的排序，以及各自的区别；3）四个by的区别?

参考答案：

共有四种排序：Order By，Sort By，Distribute By，Cluster By

1、Order By：全局排序

- 对输入的数据做排序，故此只有一个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；

1）使用 ORDER BY 子句排序

ASC（ascend）: 升序（默认）

DESC（descend）: 降序

2）ORDER BY 子句在SELECT语句的结尾

3）案例

查询员工信息按工资升序排列

```
select * from emp order by sal;
```

2、Sort By：非全局排序

- 在数据进入reducer前完成排序；
- 当mapred.reduce.tasks > 1时，只能保证每个reducer的输出有序，不保证全局有序；

3、Distribute By：分区排序

- 按照指定的字段对数据进行划分输出到不同的reduce中，通常是为了进行后续的聚集操作；
- 常和sort by一起使用，并且distribute by必须在sort by前面；

4、Cluster By

相当于distribute by+sort by，只能默认升序，不能使用倒序。