Hive建orc存储的表

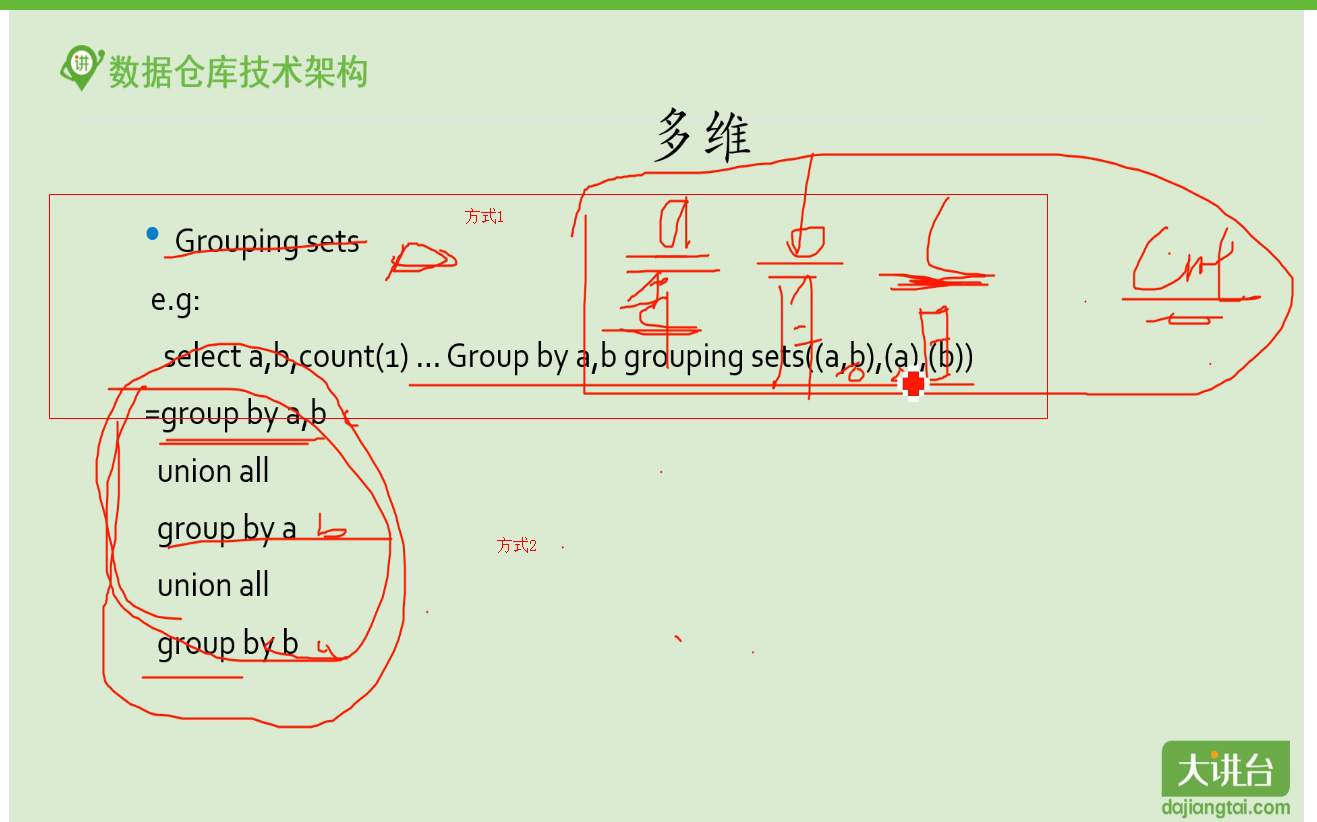


当要按多个“维度”进行索引时：比如年、月、日， 如果我想计算：按（年，月）、年、月

的3种粒度的统计时。可以用下列的grouping sets。 要比UNION3个group by方式快很多。如下：

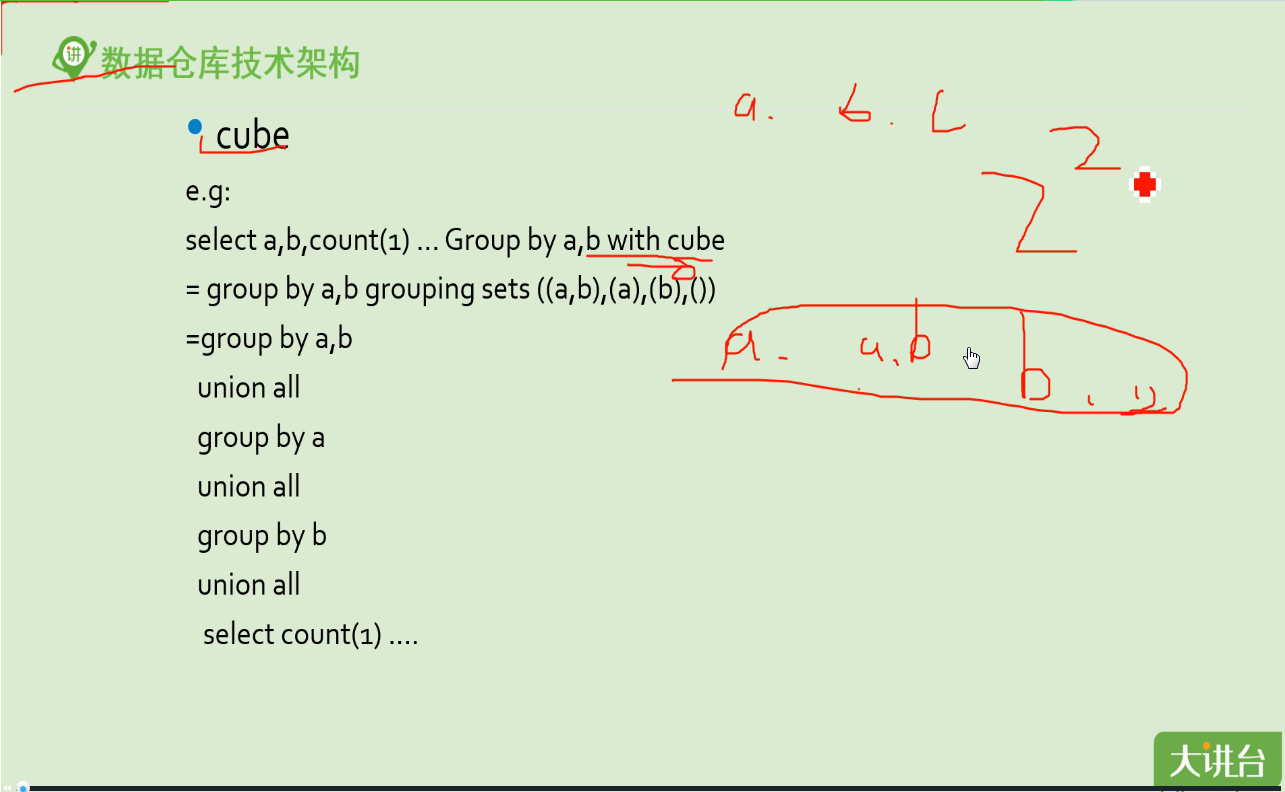
当然cube也是一种方式： N个维度，则有2\*\*N种粒度。

多维的奇淫技巧

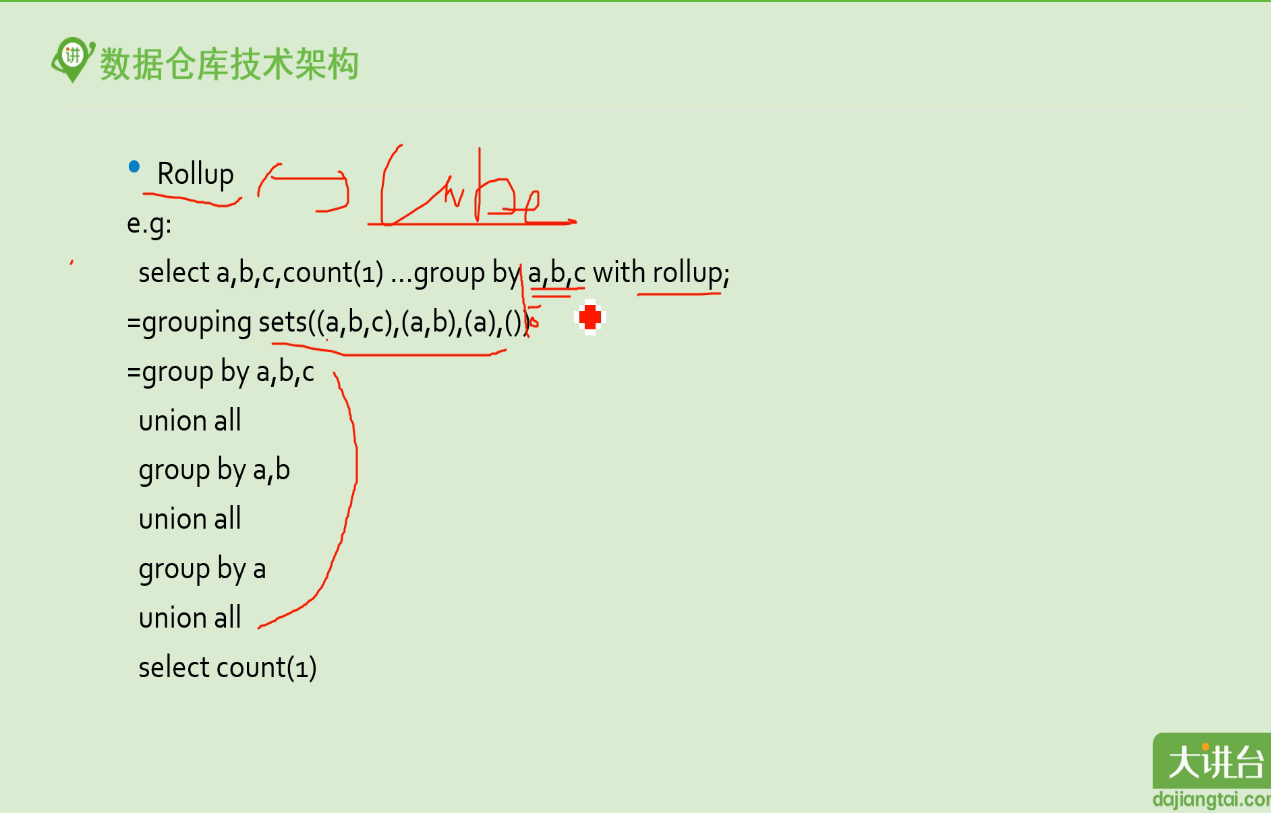


或者使用cube表

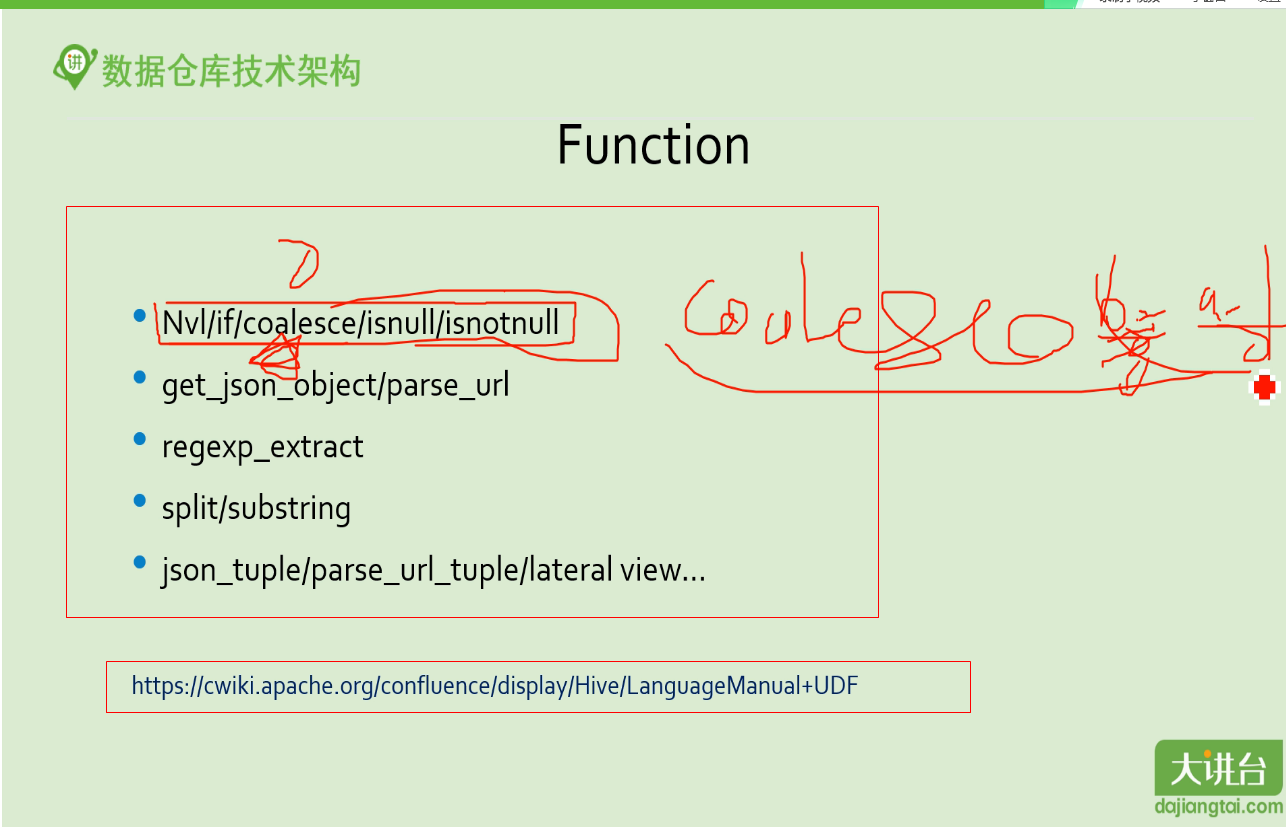
Cube



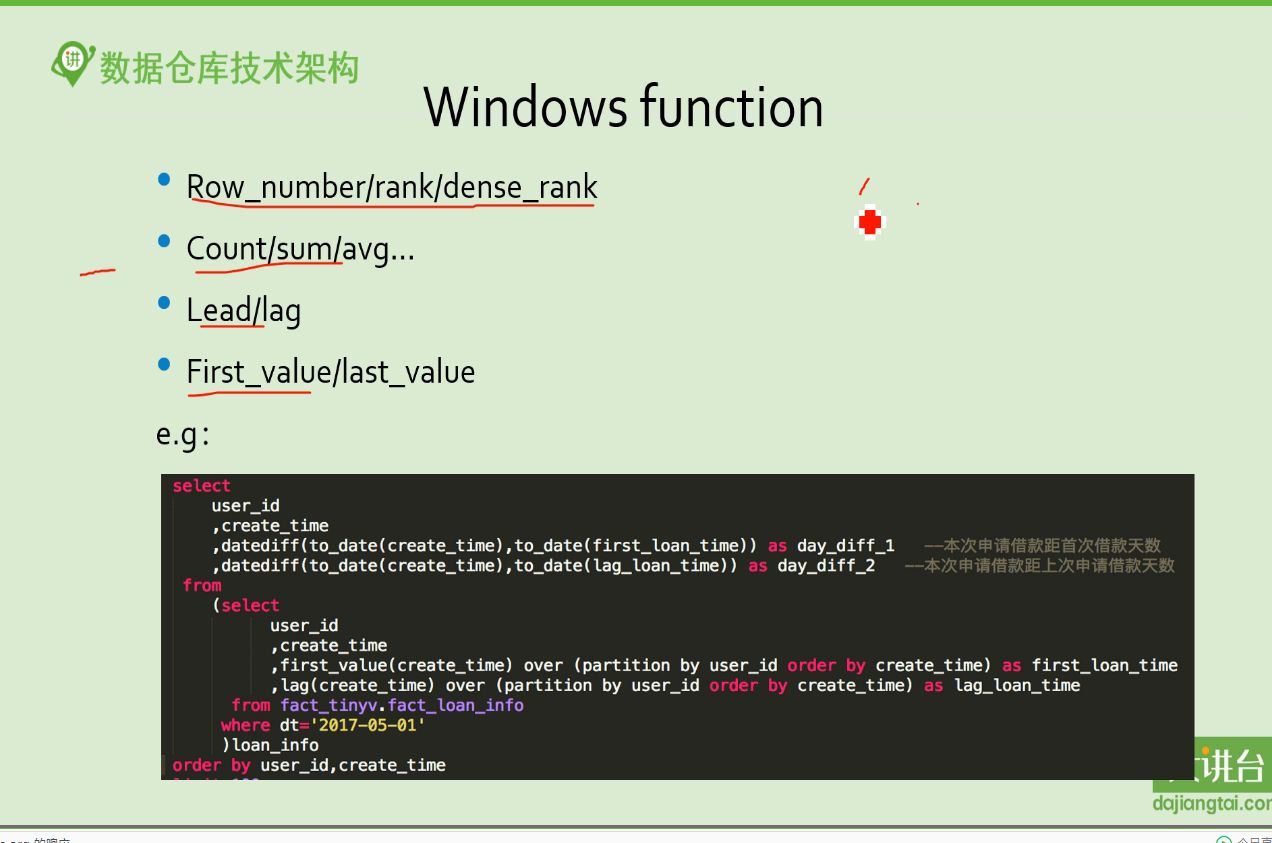
RollUp方式



Hive的其他内置函数：————好好看下



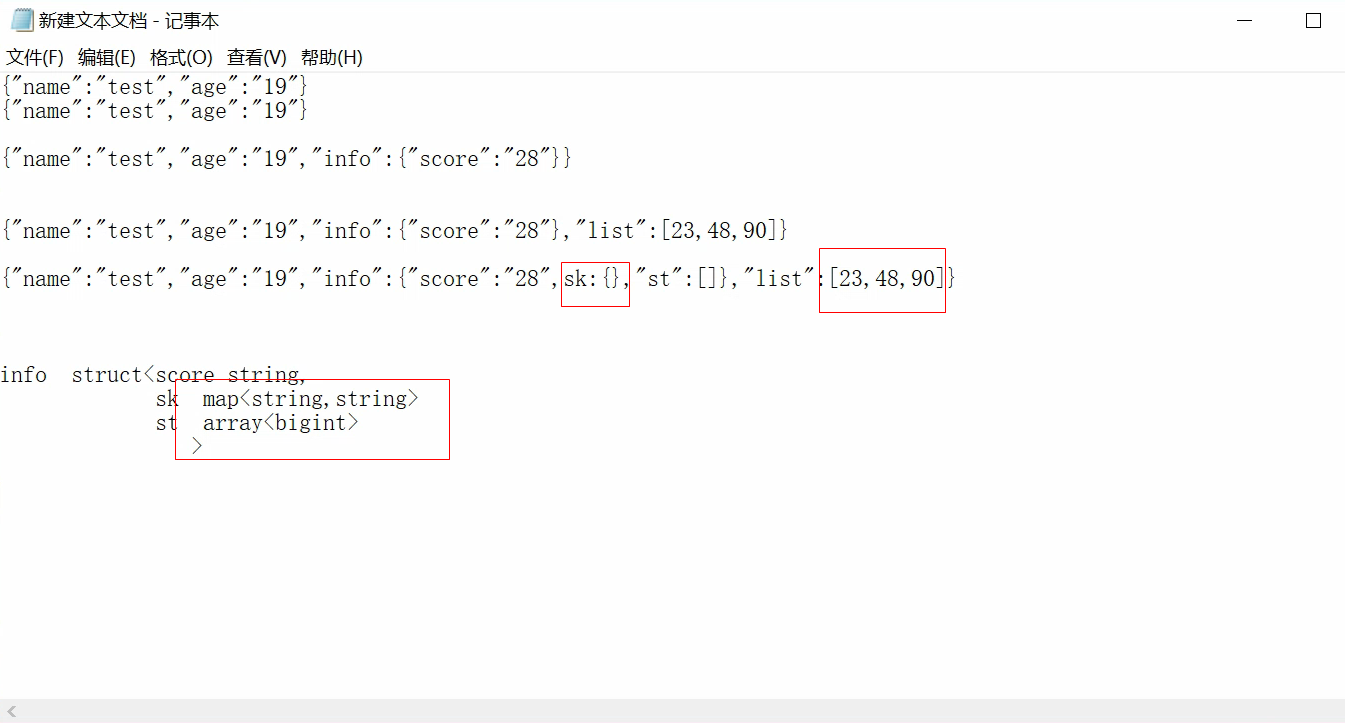
窗口函数————————滚动统计



Hive能很好的使用json。 等数据使用json格式时，能很好的与hive的3种基本容器格式结合——————下面的json里就包含hive中的struct/map/array容器

上面的3种类型，通常用于嵌套。 Json语法中的JsonObject和JsonArray就能完成这3种。

所以json牛，能作为各种数据交换格式。



Hive在建表，接收json文件时，$HIVE\_HOME/lib类里面没有。 但是在$HIVE\_HOME/hcatalog/share/hcatalog/hive-hcatalog-core-0.13.1.jar里面有对应的jar包

场景:去重

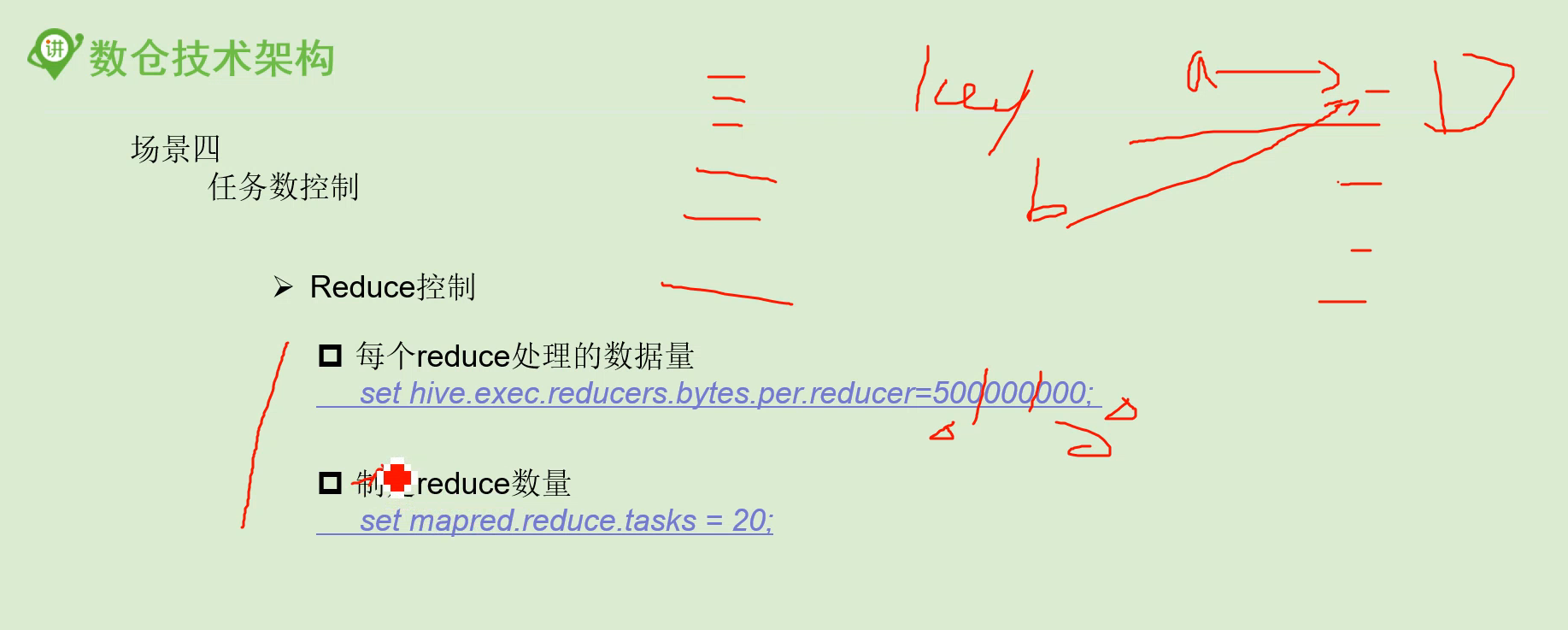


开启并行：

”



Reduce参数调优：————可以防止reduce的数据倾斜



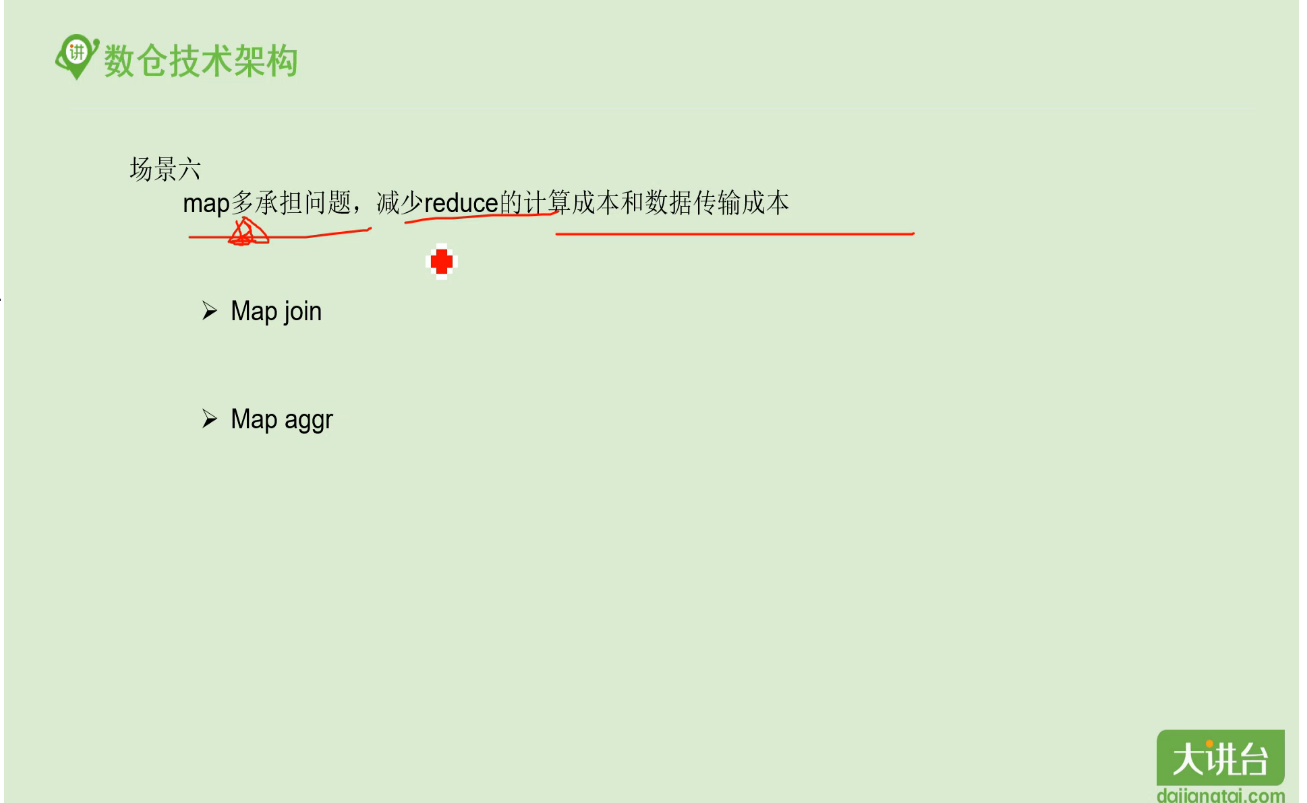
在map结束后合并小文件，或在map reduce结束后合并：

这是为了避免，map或reduce产生太多小文件，导致下一阶段的map reduce的碎片过多，map 和reduce的task任务过多，从而性能下降（yarn的资源调度，jvm启动等），namenode的性能也下降



Order by limit是利用的冒泡排序

将部分map或reduce均可执行的任务，尽量在map端完成，比如map端聚合——————好处:1.map并行度大，2减少中间的shuffle



针对spark和hadoop集群监控的开源软件：

Dr.elelphant

由领英开源