任务 4-4 HiveQL操作分桶表

幻灯片60：任务4-4 HiveQL操作分桶表（标题页）

接下来是分桶表。

分桶表也是一种优化查询的技术。

幻灯片61-62：4.4.1 任务描述 —— 分桶表概念

刚才我们学了分区表，但是分区有时候会有问题。

有些分区的数据量很大，有些分区的数据量很小，这就叫数据倾斜（shù jù qīng xié）。

为了解决这个问题，Hive提供了分桶技术。

分桶就是指定表的某一列，用哈希取模（hā xī qǔ mó）的方式，把数据均匀（jūn yún）地分到各个桶文件中。

> 💡 *举个例子：就像你有一堆苹果，分区是按颜色分（红苹果一堆、绿苹果一堆），可能红色的特别多。分桶就是不管颜色，按编号均匀分成5堆，每堆数量差不多。*

幻灯片63-64：4.4.2 任务准备知识 —— 分桶规则

分桶的规则（guī zé）是：

桶编号 = hash_function(分桶字段的值) % 桶的数量

编号相同的数据会被分到同一个桶。

如果分桶字段是int类型，hash值就是它自己。

如果是其他类型，比如String，就用hashcode来计算。

幻灯片65：4.4.2 —— 完整语法树

创建分桶表的关键语法是两个部分：

```
CLUSTERED BY (col_name)  -- 指定根据哪个字段分桶
INTO N BUCKETS           -- 指定分成几个桶
```

注意：分桶字段必须是表中已经存在的字段。

这跟分区表不一样！分区字段不能是已有字段，分桶字段必须是已有字段。

幻灯片66-69：4.4.3 任务实施 —— 分桶表实战

我们用一个美国COVID-19的数据来做练习。

根据state（州）字段分成5个桶：

【实习命令 / 실습 명령】

```
CREATE TABLE t_usa_covid19_bucket(
  count_date STRING, county STRING,
  state STRING, fips INT,
  cases INT, deaths INT)
CLUSTERED BY (state) INTO 5 BUCKETS;
```

还可以在分桶的基础上指定排序规则（pái xù guī zé）。

数据加载到分桶表通常用 INSERT 方式，而不是直接 LOAD。

幻灯片70-72：4.4.4 分桶表使用好处

分桶表有三个好处：

（1）基于分桶字段查询时，可以减少全表扫描。

比如查询New York州的数据，Hive会根据hash计算出New York在哪个桶里，直接去那个桶查就行了。

【实习命令 / 실습 명령】

```
select * from t_usa_covid19_bucket where state="New York";
```

（2）可以提高JOIN操作的效率。

（3）可以提高采样（cǎi yàng）查询的效率。

项目小结

幻灯片101：项目小结

好，同学们，今天的课程到这里就全部结束了！
我们来回顾（huí gù）一下今天学了什么。

任务4-1：学习了HiveQL操作数据库——创建、查询、查看、切换、修改和删除数据库。
任务4-2：学习了HiveQL操作数据表——创建内部表和外部表、查看、修改和删除表。
任务4-3：学习了HiveQL操作分区表——分区的概念、创建、添加、重命名和删除分区。
任务4-4：学习了HiveQL操作分桶表——分桶的概念、创建分桶表、分桶的好处。
任务4-5：学习了HiveQL操作视图——视图的概念、创建、查看、修改和删除视图。
任务4-6：学习了HiveQL相关操作——索引的概念、创建、查看、重建和删除索引。

今天的内容比较多，但是大部分操作的模式（mó shì）是一样的：
CREATE 创建、SHOW 查看列表、DESCRIBE 查看详情、ALTER 修改、DROP 删除。
掌握（zhǎng wò）了这个规律（guī lǜ），就能举一反三（jǔ yī fǎn sān）了。

课后大家一定要在虚拟机（xū nǐ jī）上动手练习这些命令，光看不练是学不会的！
下次课我们会继续学习Hive的数据操作。
同学们，今天辛苦（xīn kǔ）了，下课！谢谢大家！