任务 4-6 HiveQL相关操作

幻灯片78：任务4-6 HiveQL相关操作（标题页）

最后一个任务是HiveQL的相关操作。

这部分会总结（zǒng jié）前面学过的知识，还有补充一些新内容，比如索引。

幻灯片79-85：4.6.1 数据库DDL操作总结

首先我们来总结一下数据库的DDL操作。

DDL就是Data Definition Language，数据定义语言。

在Hive中，DATABASE和SCHEMA是一样的，可以互换（hù huàn）使用。

默认的数据库叫 default，数据存在 /user/hive/warehouse 下面。

用户自己创建的数据库存在 /user/hive/warehouse/database_name.db 下面。

数据库操作总结：

CREATE DATABASE——创建数据库

DESCRIBE DATABASE——查看数据库信息

USE DATABASE——切换数据库

DROP DATABASE——删除数据库（默认RESTRICT，加CASCADE可以删除非空数据库）

ALTER DATABASE——修改数据库属性

幻灯片86-87：4.6.2 数据表DDL操作总结

表的DDL操作是核心（hé xīn）中的核心。

表的定义（dìng yì）是否成功，直接影响数据能不能正确映射（yìng shè）。

实际工作中，如果建表有问题，通常（tōng cháng）直接删了重建，不去修改。

因为Hive建表后加载数据很快。

查看表信息：

DESCRIBE 或 DESC——查看基本信息

加 EXTENDED——显示所有元数据（用Thrift序列化格式）

加 FORMATTED——以表格形式显示元数据，更好读

幻灯片88-93：4.6.3 Hive Partition(分区）DDL操作(索引操作)

现在来讲一个新内容——索引（suǒ yǐn）。

索引是什么?

索引就像书的目录一样。你想找某个内容，不用翻遍（fān biàn）整本书，看目录就知道在哪一页了。

索引的作用：通过索引列查询数据时，可以避免全表扫描，提高查询速度。

但是，创建索引会消耗（xiāo hào）额外的资源和磁盘（cí pán）空间。

所以索引是一种以空间换时间的方式。

索引的原理：

Hive的索引其实是一张索引表。

索引表里存了：索引列的值、对应的HDFS文件路径、在文件中的偏移量（piān yí liàng）。

查询时先找索引表，知道数据在哪个文件的哪个位置，直接去读那部分数据。

创建索引的语法：

```
CREATE INDEX index_name
ON TABLE base_table_name (col_name, ...)
AS index_type
[WITH DEFERRED REBUILD]
[IN TABLE index_table_name];
```

查看索引：

```
SHOW [FORMATTED] INDEXES ON table_with_index [(FROM|IN) db_name];
```

重建（chóng jiàn）索引：

```
ALTER INDEX index_name ON table_name REBUILD;
```

删除索引：

```
DROP INDEX [IF EXISTS] index_name ON table_name;
```

幻灯片94-98：4.6.3 索引任务实施

我们来实际操作一下索引。

（1）为表 managed_table_new 的 staff_name 列创建索引：

【实习命令 / 실습 명령】

```
CREATE INDEX index_staff_name
ON TABLE hive_database.managed_table_new (staff_name)
AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
WITH DEFERRED REBUILD
IN TABLE index_name_table
TBLPROPERTIES ('create'='hive_database')
COMMENT 'index comment';
```

（2）查看索引表：

【实习命令 / 실습 명령】

```
SHOW TABLES;
DESC FORMATTED index_name_table;
```

（3）查看索引信息：

【实习命令 / 실습 명령】

```
SHOW INDEXES ON managed_table_new FROM hive_database;
```

（4）重建索引：

【实习命令 / 실습 명령】

```
ALTER INDEX index_staff_name ON hive_database.managed_table_new REBUILD;
```

（5）删除索引：

【实习命令 / 실습 명령】

```
DROP INDEX IF EXISTS index_staff_name ON managed_table_new;
```

注意！Hive建了索引之后，查询时不会自动使用索引，需要开启（kāi qǐ）配置：

【实习命令 / 실습 명령】

```
SET hive.input.format=org.apache.hadoop.hive.ql.io.HiveInputFormat;
SET hive.optimize.index.filter=true;
SET hive.optimize.index.filter.compact.minsize=0;
```

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！