Hive练习2 数据表操作

编程要求

请根据右侧命令行内的提示，在`Begin - End`区域内进行`sql`语句代码补充，具体任务如下：

`student`表结构：

INFO	TYPE	COMMENT
Sno	INT	student sno
name	STRING	student name
age	INT	student age
sex	STRING	student sex
score	`STRUCT`	student score

- 创建数据库`test2`
- 在`test2`中创建表`student`，表结构如上所示
- 使用`LIKE`关键字创建一个与`student`表结构相同的表`student_info`
- 删除表`student`

按照以上要求填写命令。每个要求对应一条命令，共`4`条命令，以`;`隔开。

答案

```
#********* Begin *********#
echo "

CREATE DATABASE test2;
CREATE TABLE IF NOT EXISTS test2.student(
    Sno INT COMMENT 'student sno',
    name STRING COMMENT 'student name',
    age INT COMMENT 'student age',
    sex STRING COMMENT 'student sex',
    score STRUCT<Chinese:FLOAT, Math:FLOAT, English:FLOAT> COMMENT 'student
score');
CREATE TABLE IF NOT EXISTS test2.student_info LIKE student;
DROP TABLE IF EXISTS test2.student;
"
#********* End *********#
```

题目解析

Create 创建表

创建表的语法为：

```
CREATE [TEMPOPARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.] table_name
   [(col_name data_type [COMMENT col_comment], ...)]
   [COMMENT table_comment]
   [PARTITIONED BY (col_name data_type [COMMENT   col_comment], ...)]
   [CLUSTERED BY (col_name, col_name, ...) [SORTED BY (col_name [ASC|DESC],
...)] INTO num_buckets BUCKETS]
    [SKEWED BY (col_name,col_name,…) ON ([(col_value,col_value,…),
…|col_value,col_value,…]) [STORED AS DIRECTORIES] ]
    [
        [ROW FORMAT DELIMITED [FIFLDS TERMINATED BY char [ESCAPED BY char]]
[COLLECTION ITEMS TERMINATED BY char] [MAP KEYS TERMINATED BY char] [LINES
TERMINATED BY char] [NULL DEFINED AS char]
        | SERDE serde_name [WITH SERDEPROPERTIES
(property_name=property_value,property_name=property_value,…)]
      ]
      [STORED AS file_format]
      | STORED BY 'storage.handler.class.name' [WITH SERDEPROPERTIES (…)]
    ]
    [LOCATION hdfs_path]
    [TBLPROPERTIES (property_name=property_value,…)]
    [AS select_statement];
```

参数说明如下：

- `TEMPOPARY`：创建临时表，若未指定，则默认创建的是普通表
- `EXTERNAL`：创建外部表，若未指定，则默认创建的是内部表
- `IF NOT EXISTS`：若表不存在才创建，若未指定，当目标表存在时，创建操作抛出异常
- `db_name.`：前缀，指定表所属于的数据库。若未指定且当前数据库非`db_name`，则使用`default`数据库
- `COMMENT`：添加注释说明，注释内容位于单引号内
- `PARTITIONED BY`：针对存储有大量数据集的表，根据表内容所具有的某些共同特征定义一个标签，将这类数据存储在该标签所标识的位置，可以提高表内容的查询速度。`PARTITIONED BY`中的列名为伪列或标记列，不能与表中的实体列名相同，否则 hive 表创建操作报错
- `CLUSTERED BY`：根据列之间的相关性指定列聚类在相同桶中（`BUCKETS`），可以对表内容按某一列进行升序（`ASC`）或降序（`DESC`）排序（`SORTED BY`关键字）
- `SKEWED BY`：用于过滤掉特定列`col_name`中包含值`col_value`（`ON(col_value,…)`关键字指定的值）的记录，并单独存储在指定目录（`STORED AS DIRECTORIES`）下的单独文件中
- `ROW FORMAT`：指定 hive 表行对象（`ROW Object`）数据与 HDFS 数据之间进行传输的转换方式（`HDFS files -> Deserializer ->Row object`以及`Row object ->Serializer ->HDFS files`），以及数据文件内容与表行记录各列的对应。在创建表时可以指定数据列分隔符（`FIFLDS TERMINATED BY 子句`）、对特殊字符进行转义的特殊字符（`ESCAPED BY 子句`）、符合数据类型值分隔符（`COLLECTION ITEMS TERMINATED BY 子句`）、`MAP key-value`类型分隔符（`MAP KEYS TERMINATED BY`）、数据记录行分隔符（`LINES TERMINATED BY`）、定义`NULL`字符（`NULL DEFINED AS`），同时可以指定自定义的`SerDE`（`Serializer`和`Deserializer`，序列化和反序列化），也可以指定默认的`SerDE`。如果`ROW FORMAT`未指定或指定为`ROW FORMAT DELIMITED`，将使用内部默认`SerDe`
- `STORED AS`：指定 hive 表数据在 HDFS 上的存储方式。`file_format`值包括`TEXTFILE`（普通文本文件，默认方式）、`SEQUENCEFILE`（压缩模式）、`ORC`（`ORC`文件格式）和`AVRO`（`AVRO`文件格式）
- `STORED BY`：创建一个非本地表，如创建一个 HBase 表

- `LOCATION`：指定表数据在 HDFS 上的存储位置。若未指定，`db_name`数据库将会储存在`${hive.metastore.warehouse.dir}`定义位置的`db_name`目录下
- `TBLPROPERTIES`：为所创建的表设置属性（如创建时间和创建者，默认为当前用户和当前系统时间）
- `AS select_statement`：使用`select`子句创建一个复制表（包括select子句返回的表模式和表数据）

1. 在之前创建的`shopping`数据库创建一张商品信息表（`items_info`）：

```
CREATE TABLE IF NOT EXISTS shopping.items_info(
name STRING COMMENT 'item name',
price FLOAT COMMENT 'item price',
category STRING COMMENT 'item category',
brand STRING COMMENT 'item brand',
type STRING COMMENT 'item type',
stock INT COMMENT 'item stock',
address STRUCT<street:STRING, city:STRING, state:STRING, zip:INT> COMMENT 'item
sales address')
COMMENT 'goods information table'
TBLPROPERTIES ('creator'='Xiaoming','date'='2019-01-01');
```

2. 查看`items_info`表结构:

```
use shopping;  #切换到数据库shopping中
desc items_info;
```

复制表

按照已存在的表或视图定义一个相同结构的表或视图（使用`LIKE`关键字，只复制表定义，不复制表数据）。

1. 复制刚才创建的表`items_info`起名为`items_info2`。

```
CREATE TABLE IF NOT EXISTS items_info2LIKE items_info;
```

Drop 删除表

删除表的语法为:

```
DROP TABLE [IF EXISTS] table_name;
```

- `[IF EXISTS]`：关键字可选;
- 若未指定且表`table_name`不存在时，`Hive`返回错误。

1. 删除刚才复制的表`items_info2`:

```
DROP TABLE IF EXISTS items_info2;
```

Truncate 截断表

大家众所周知，当我们在自己的电脑上删除某一个文件，它并没有彻底删除而是进入了回收站，你要在回收站中再将其删除才算彻底清除。截断表就相当于直接将数据从电脑上删除，而不会放入回收站。

截断表（删除表中所有行）的语法为：

```
TRUNCATE TABLE table_name [PARTITION partition_spec];partition_spec:
(partition_column=partition_col_value,partition_column=partition_col_value,…)
```