项目四 Hive数据定义



《Hive数据仓库应用教程》

目 录 / CONTENTS

1 任务4-1 HiveQL操作数据库

2 任务4-2 HiveQL操作数据表

3 任务4-3 HiveQL操作分区表

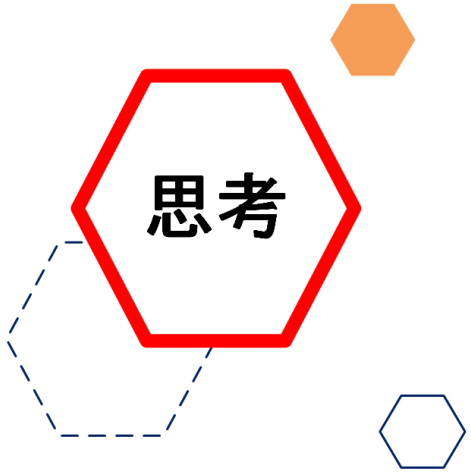
4 任务4-4 HiveQL操作分桶表

5 任务4-5 HiveQL操作视图

6 任务4-6 HiveQL相关操作

任务 4-3 HiveQL操作分区表

4.3.1 任务描述



随着系统运行时间的增加，表的数据量会越来越大，而Hive查询数据通常是使用全表扫描，这样会导致对大量不必要数据的扫描，从而降低查询效率，

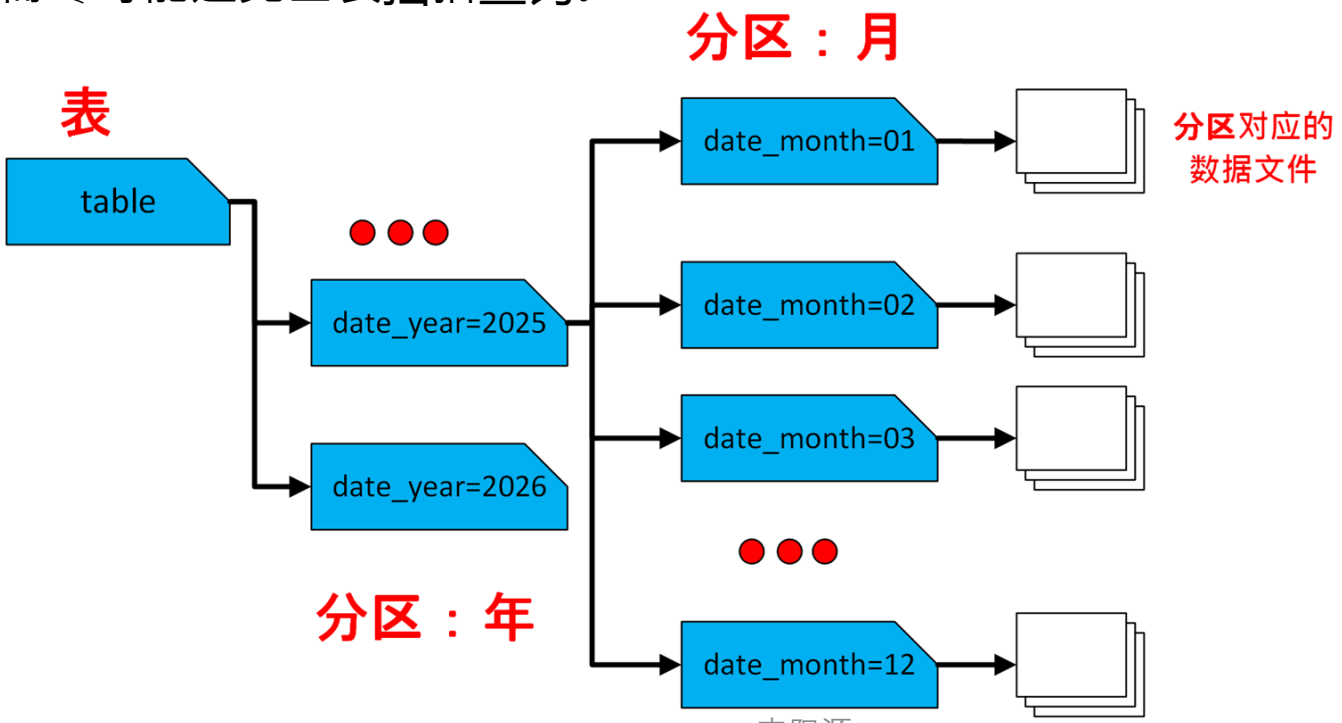
为了解决这一问题，Hive引进了分区技术，分区主要是将表的整体数据根据业务需求，划分成多个子目录来存储，每个子目录对应一个分区。

通过扫描分区表中指定分区的数据，避免Hive全表扫描，从而提升Hive查询数据的效率。

概念

- 当Hive表对应的数据量大、文件数量多时，为了避免查询时全表扫描数据，Hive支持根据指定的字段对表进行分区。分区的字段可以是日期、地域、种类等具有标识意义的字段。

- 例如，将一整年的数据按照月份划分为12个月（即12个分区），后续就可以直接查询指定月份分区的数据，从而尽可能避免全表扫描查询。



Hive Partitioned Tables 分区表： 完整语法树

```
CREATE [TEMPORARY] [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]
[ROW FORMAT DELIMITED|SERDE serde_name WITH SERDEPROPERTIES (property_name=property_value,...)]
[STORED AS file_format]
[LOCATION hdfs_path]
[TBLPROPERTIES (property_name=property_value, ...)];
```

Hive Partitioned Tables 分区表： 创建

- 分区表语法

注意：分区字段不能是表中已经存在的字段，因为分区字段最终也会以虚拟字段的形式显示在表结构上。

```
--分区表建表语法
CREATE TABLE table_name (
    column1 data_type,
    column2 data_type,
    ....)
PARTITIONED BY (partition1 data_type, partition2 data_type,...);
```

1．查询分区表

分区表属于Hive数据表的一种，有关查询分区表分区信息语法格式如下：

```
SHOW PARTITIONS [db_name.]table_name [PARTITION(partition_column = partition_col_value, partition_column = partition_col_value, ...)];
```

- SHOW PARTITIONS：表示查看分区表分区信息的语句；
- db_name.：可选，用于指定数据库名称，若不指定则使用当前数据库；
- table_name：用于指定分区表名称；
- PARTITION(…)：可选，查询指定分区信息。

2．添加分区表

添加分区是在分区表中根据分区字段添加实际分区，语法格式如下：

```
ALTER TABLE table_name ADD [IF NOT EXISTS] PARTITION
(partition_column = partition_col_value,
partition_column = partition_col_value, ...)
[LOCATION 'location']…;
```

- ALTER TABLE：表示修改数据表结构信息的语句；
- ADD [IF NOT EXISTS] PARTITION：用于添加分区；
- partition_column：用于指定分区字段；
- partition_col_value：用于指定分区字段的值，即实际分区；

LOCATION 'location'：可选，用于指定分区在HDFS上的存储位置。

3．重命名分区表

重命名分区是根据分区表的分区字段修改分区表的实际分区，
重命名分区语法格式如下：

```
ALTER TABLE table_name PARTITION
(partition_column = partition_col_value,
partition_column = partition_col_value, ...)
RENAME TO PARTITION (partition_column = partition_col_value,
partition_column = partition_col_value, ...);
```

4．删除分区

删除分区是根据分区表的分区字段删除分区表的实际分区，语法格式如下：

```
ALTER TABLE table_name DROP [IF EXISTS]
PARTITION (partition_column = partition_col_value,
partition_column = partition_col_value, ...) [PURGE];
```

5．修改分区存储位置和文件格式

（1）修改分区表中分区在HDFS存储位置，具体语法格式如下：

```
ALTER TABLE PARTITION (partition_column = partition_col_value,
partition_column = partition_col_value, ...) SET LOCATION "new location";
```

（2）修改分区表中分区的文件存储格式：

```
ALTER TABLE table_name PARTITION (partition_column = partition_col_value,
partition_column = partition_col_value, ...) SET FILEFORMAT file_format;
```

4.3.3 任务实施

1．创建分区表

在数据库hive_database中创建分区表partitioned_table，语句如下：

```
CREATE DATABASE hive_database;

CREATE TABLE IF NOT EXISTS
hive_database.partitioned_table(
username STRING COMMENT "This is username",
age INT COMMENT "This is user age")
PARTITIONED BY (
province STRING COMMENT "User live in province",
city STRING COMMENT "User live in city")
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS textfile
TBLPROPERTIES("comment"="This is a partitioned table");
```

注意：

（1）分区表中的分区字段名称不能与分区表的列名重名。

（2）分区字段在创建分区表时指定,一旦分区表创建完成,后续则无法修改或者添加分区字段的。

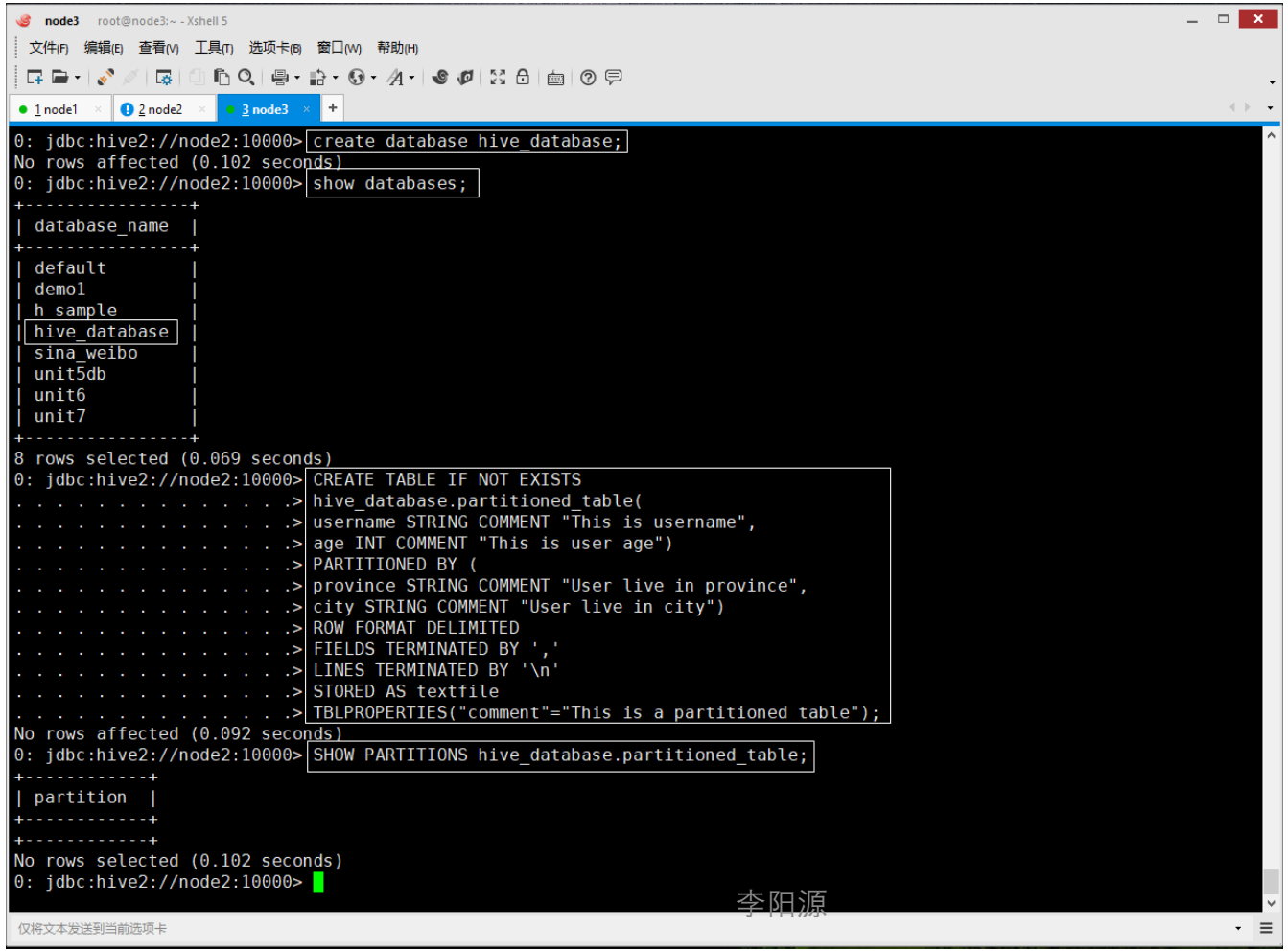
4.3.3 任务实施

2．查询分区表

查询数据库hive_database中分区表partitioned_table的分区信息，语句如下：

```
SHOW PARTITIONS hive_database.partitioned_table;
```

执行结果如图4-16所示。



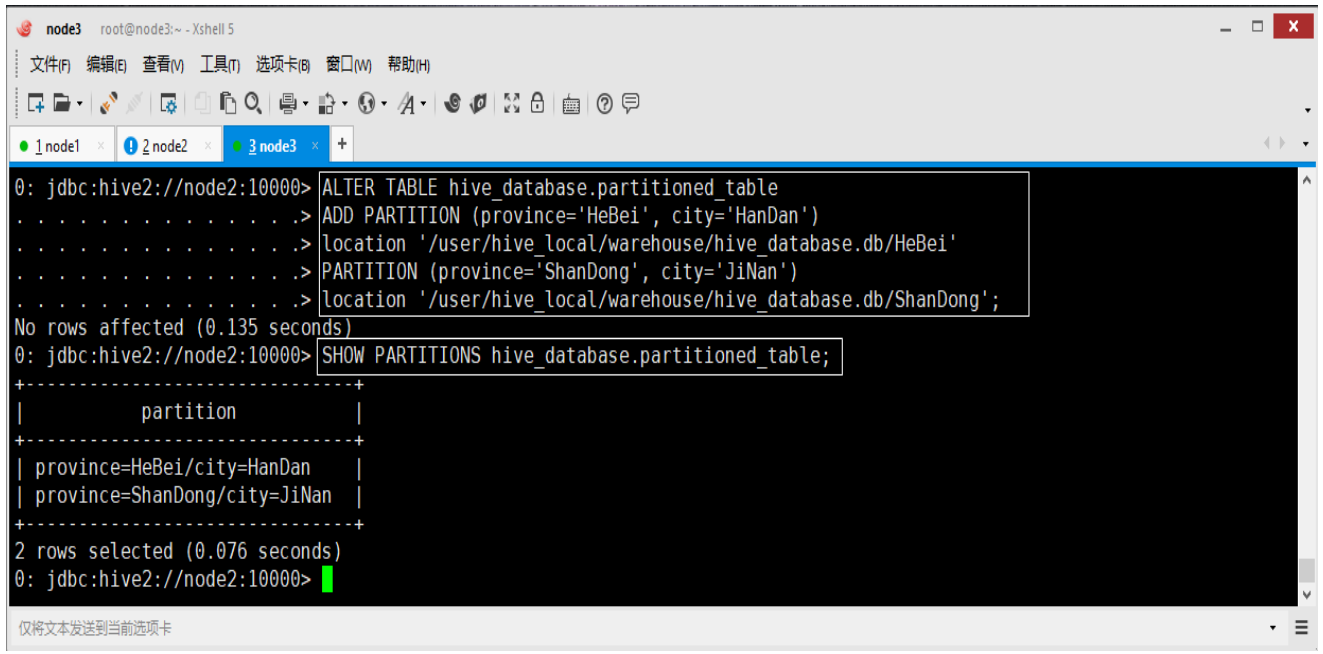
4.3.3 任务实施

3．添加分区

向数据库hive_database的分区表partitioned_table添加分区，语句如下：

```
ALTER TABLE hive_database.partitioned_table
ADD PARTITION (province='HeBei', city='HanDan')
location '/user/hive/warehouse/hive_database.db/HeBei'
PARTITION (province='ShanDong', city='JiNan')
location '/user/hive/warehouse/hive_database.db/ShanDong';
```

执行结果如下图所示：



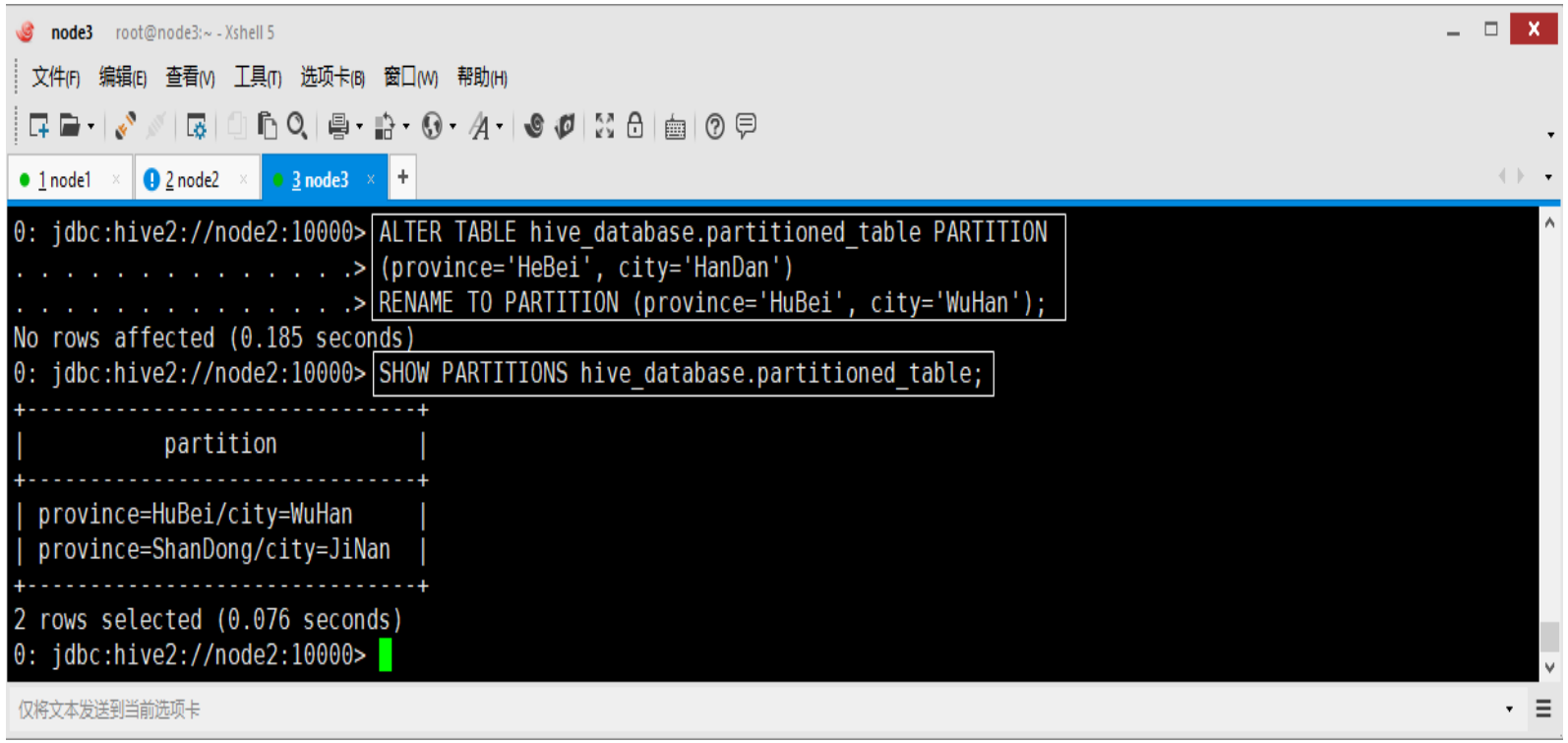
4.3.3 任务实施

4．重命名分区

重命名数据库hive_database中分区表partitioned_table的分区，语句如下：

```
ALTER TABLE hive_database.partitioned_table PARTITION
(province='HeBei', city='HanDan')
RENAME TO PARTITION (province='HuBei', city='WuHan');
```

执行结果如图4-18所示。



5．移动分区

在虚拟机中使用客户端工具，远程连接虚拟机服务操作Hive，移动分区的具体步骤如下：

（1）创建分区表

在数据库hive_database中，创建与分区表partitioned_table表结构一致的分区表partitioned_table1，语句如下：

```
CREATE  TABLE IF NOT EXISTS
hive_database.partitioned_table1(
username STRING COMMENT "This is username",
age INT COMMENT "This is user age"
)
PARTITIONED BY (
province STRING COMMENT "User live in province",
city STRING COMMENT "User live in city"
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS textfile
TBLPROPERTIES("comment"="This is a partitioned table1");
```

4.3.3 任务实施

（2）移动分区

将分区表partitioned_table的二级分区province=HuBei/city=WuHan移动到分区表partitioned_table1，语句如下：

```
ALTER TABLE hive_database.partitioned_table1 EXCHANGE
PARTITION (province='HuBei', city='WuHan') WITH TABLE partitioned_table;
```

（3）查看分区

```
SHOW PARTITIONS hive_database.partitioned_table;
SHOW PARTITIONS hive_database.partitioned_table1;
```

```
0: jdbc:hive2://localhost:10000> SHOW PARTITIONS partitioned_table;
+-------------------------------+
|           partition           |
+-------------------------------+
| province=ShanDong/city=JiNan  |
+-------------------------------+
0: jdbc:hive2://localhost:10000> SHOW PARTITIONS partitioned_table1;
+-------------------------------+
|           partition           |
+-------------------------------+
| province=HeBei/city=HanDan    |
| province=HuBei/city=WuHan     |
| province=ShanDong/city=JiNan  |
+-------------------------------+
```

6．删除分区

删除数据库hive_database中分区表partitioned_table1的分区，语句如下：

```
ALTER TABLE hive_database.partitioned_table1 DROP IF EXISTS
PARTITION (province='HuBei', city='WuHan');
```

```
0: jdbc:hive2://localhost:10000> ALTER TABLE hive_database.partitioned_table1 DROP IF EXISTS
0: jdbc:hive2://localhost:10000> PARTITION (province='HuBei', city='WuHan');
0: jdbc:hive2://localhost:10000> SHOW PARTITIONS hive_database.partitioned_table1;
+-----------------------------+
|           partition         |
+-----------------------------+
| province=HeBei/city=HanDan  |
| province=ShanDong/city=JiNan|
+-----------------------------+
```

7. 分区表的注意事项

一、分区表不是建表的必需语法规则，是一种优化手段表，可选；

二、分区字段不能是表中已有的字段，不能重复；

三、分区字段是虚拟字段，其数据并不存储在底层文件中；

四、分区字段值的确定来自于用户手动指定数据（静态分区）或根据查询结果位置自动推断（动态分区）；

五、Hive支持多重分区，也就是说在分区的基础上继续分区，划分更加细粒度。

感谢您的观看



《Hive数据仓库应用教程》