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 4-6 HiveQL相关操作

整体概述

- 在Hive中，DATABASE的概念和RDBMS中类似，我们称之为数据库，DATABASE和SCHEMA是可互换的，都可以使用。
- 默认的数据库叫做default，存储数据位置位于/user/hive/warehouse下。
- 用户自己创建的数据库存储位置是/user/hive/warehouse/database_name.db下。

create database

- create database用于创建新的数据库

- COMMENT：数据库的注释说明语句
- LOCATION：指定数据库在HDFS存储位置，默认/user/hive/warehouse/dbname.db
- WITH DBPROPERTIES：用于指定一些数据库的属性配置。

```
CREATE (DATABASE|SCHEMA) [IF NOT EXISTS] database_name
[COMMENT database_comment]
[LOCATION hdfs_path]
[WITH DBPROPERTIES (property_name=property_value, ...)];
```

create database

- 例子：创建数据库itcast

```
create database if not exists itcast
comment "this is my first db"
with dbproperties ('createdBy'='YangWeon');
```

➢ **注意：如果需要使用 location 指定路径的时候，最好指向的是一个新创建的空文件夹**

```
-- 正确方法：先在HDFS中创建空文件夹，然后使用
hdfs dfs -mkdir /user/hive/custom_location

CREATE DATABASE mydb
LOCATION '/user/hive/custom_location';

-- 可能会有问题的方法：指定已存在数据的文件夹
CREATE DATABASE mydb
LOCATION '/user/hive/existing_data_folder'; -- ⚠ 危险
```


李阳颖

describe|desc database

- 显示Hive中数据库的名称、注释（如果已设置）及其在文件系统上的位置等信息。

EXTENDED关键字用于显示更多信息。可以将关键字describe简写成desc使用。

```
DESCRIBE DATABASE/SCHEMA [EXTENDED] db_name;
```

```
describe database itcast;
describe database extended itcast;
desc database extended itcast;
```

use database

- 选择特定的数据库

切换当前会话使用哪一个数据库进行操作

```
0: jdbc:hive2://localhost:10000> show databases;
+----------------+
| database_name  |
+----------------+
| default        |
| h_sample       |
| hive_database  |
| itcast         |
+----------------+
0: jdbc:hive2://localhost:10000> use itcast;
```

drop database

- 删除数据库

默认行为是RESTRICT, 这意味着仅在数据库为空时才删除它。
要删除带表的数据库（不为空的数据库）， 我们可以使用CASCAD。

```
DROP (DATABASE|SCHEMA) [IF EXISTS] database_name [RESTRICT|CASCADE];
```

alter database

- 更改写Hive 中的数据库关联的元数据

```
-- 更改数据库属性
ALTER DATABASE|SCHEMA) database_name SET DBPROPERTIES (property_name=property_value, ...);

-- 更改数据库所有者
ALTER DATABASE|SCHEMA) database_name SET OWNER [USER|ROLE] user_or_role;

-- 更改数据库位置
ALTER DATABASE|SCHEMA) database_name SET LOCATION hdfs_path;
```

整体概述

- Hive中针对表的DDL操作可以说是DDL中的核心操作，包括建表、修改表、删除表、描述表元数据信息。
- 其中以建表语句为核心中的核心，详见Hive DDL建表语句。
- 可以说表的定义是否成功直接影响数据能否成功映射，进而影响能否顺利使用Hive开展数据分析。
- 由于Hive建表后加载映射数据很快，实际中如果建表有问题，可以不修改，直接删除重建。

describe|desc table

- 显示Hive中表的元数据信息
  - 如果指定了EXTENDED关键字，则它将以Thrift序列化形式显示的所有元数据。
  - 如果指定了FORMATTED关键字，则它将以表格格式显示元数据。

```
describe formatted itcast.student;
```

Output | Result 3

35 rows

	col_name	data_type	comment
1	# col_name	data_type	comment
2	sno	int	
3	sname	string	
4	sex	string	
5	sage	int	
6	sdept	string	
7		<null>	<null>
8	# Detailed Table Information	<null>	<null>
9	Database:	itcast	<null>
10	OwnerType:	USER	<null>
11	Owner:	root	<null>
12	CreateTime:	Thu Mar 25 11:00:46 CST 2021	<null>
13	LastAccessTime:	UNKNOWN	<null>
14	Retention:	0	<null>
15	Location:	hdfs://node1.itcast.cn:8020/user/hive…	<null>

4.6.3 Hive Partition(分区）DDL操作

在应用系统中，尤其在联机事务处理系统中(OLTP)，对数据查询及处理的速度已成为衡量应用系统成败的标准。

采用索引加快查询数据的速度是广大数据库用户所接受的一种优化方法。

在良好的数据库设计基础上，有效地使用索引可以提高数据库性能。

4.6.3 Hive Partition(分区）DDL操作 (任务准备知识)

1．索引的作用

索引创建在Hive表的指定列，创建索引的列称为索引列，通过索引列执行查询操作时，可以避免全表扫描以及全分区扫描，从而提高查询速度。然而在提高查询速度的同时，Hive会额外消耗资源去创建索引，以及需要更多的磁盘空间存储索引。可以总结为索引是一种以空间换取时间的方式。

2．索引的原理

Hive的索引其实是一张索引表，在表中存储了索引列的值、索引列的值在HDFS对应的数据文件路径以及索引列的值在数据文件中的偏移量。在涉及到索引列的查询时，首先会去索引表中查找索引列的值在HDFS对应的数据文件路径以及索引列的值在数据文件中的偏移量，通过数据文件路径和偏移量去扫描全表的部分数据，从而避免全表扫描。

4.6.3 Hive Partition(分区）DDL操作 (任务准备知识)

3．创建索引

创建索引的语法格式:

```
CREATE INDEX index_name
  ON TABLE base_table_name (col_name, ...)
  AS index_type
  [WITH DEFERRED REBUILD]
  [IN TABLE index_table_name]
  [
     [ ROW FORMAT ...] STORED AS ...
     | STORED BY ...
  ]
  [LOCATION hdfs_path]
  [TBLPROPERTIES (...)]
  [COMMENT "index comment"];
```

说明:

- CREATE INDEX：表示创建索引的语句;
- index_name：指定创建索引时实现的类;
- AS index_type：用于指定索引类型;
- STORED AS：可选，用于指定存储格式;
- STORED BY：可选，用于指定排序规则;
- TBLPROPERTIES：用于指定索引表属性。

4．查看索引

查看索引是指查看Hive中创建索引数据表的索引信息，语法如下：

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

说明：

- INDEX和INDEXES含义相同，可以切换使用，表示用于查看索引信息；
- table_with_index用于指定查看创建索引的数据表。

5．重建索引

重建索引操作，将索引列的值、列的值在HDFS对应的数据文件路径和索引列的值在数据文件中的偏移量，这些数据加载到索引表中，此过程需要通过MapReduce实现。语法格式：

ALTER INDEX index_name ON table_name [PARTITION partition_spec] REBUILD;

说明：

- index_name：表示索引名称；
- table_name：表示索引所在的数据表。

4.6.3 Hive Partition(分区）DDL操作 (任务准备知识)

6．删除索引

删除索引的同时会删除索引对应的索引表，删除索引语法格式如下：

DROP INDEX [IF EXISTS] index_name ON table_name;

说明：

- index_name表示索引名称；
- table_name表示索引所在的数据表。

另外，Hive表建立了索引，但是在查询时是不会自动去使用索引功能的，需要开启相关配置。开启配置后，涉及到索引列的查询就会使用索引功能去优化查询，具体配置如下：

```
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;
```

4.6.3 Hive Partition(分区）DDL操作 (任务准备知识)

1．创建索引

为数据库hive_database的内部表managed_table_new创建索引，语句如下：

```
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";
```

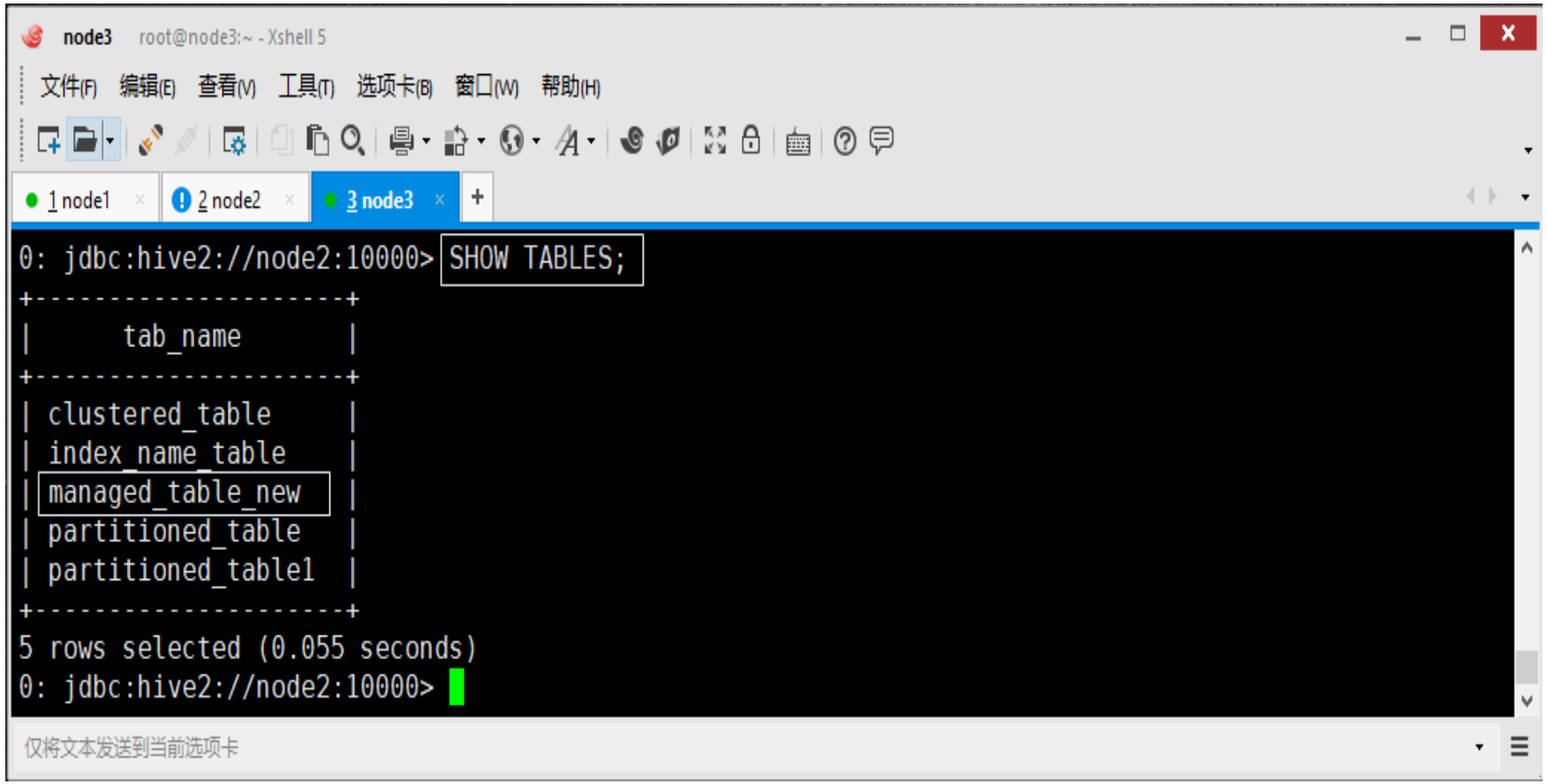
2．查看索引表详细信息

查看当前数据库中的索引表，语句如下：

```
SHOW TABLES;
```

执行结果如图4-24所示。

4.6.3 任务实施

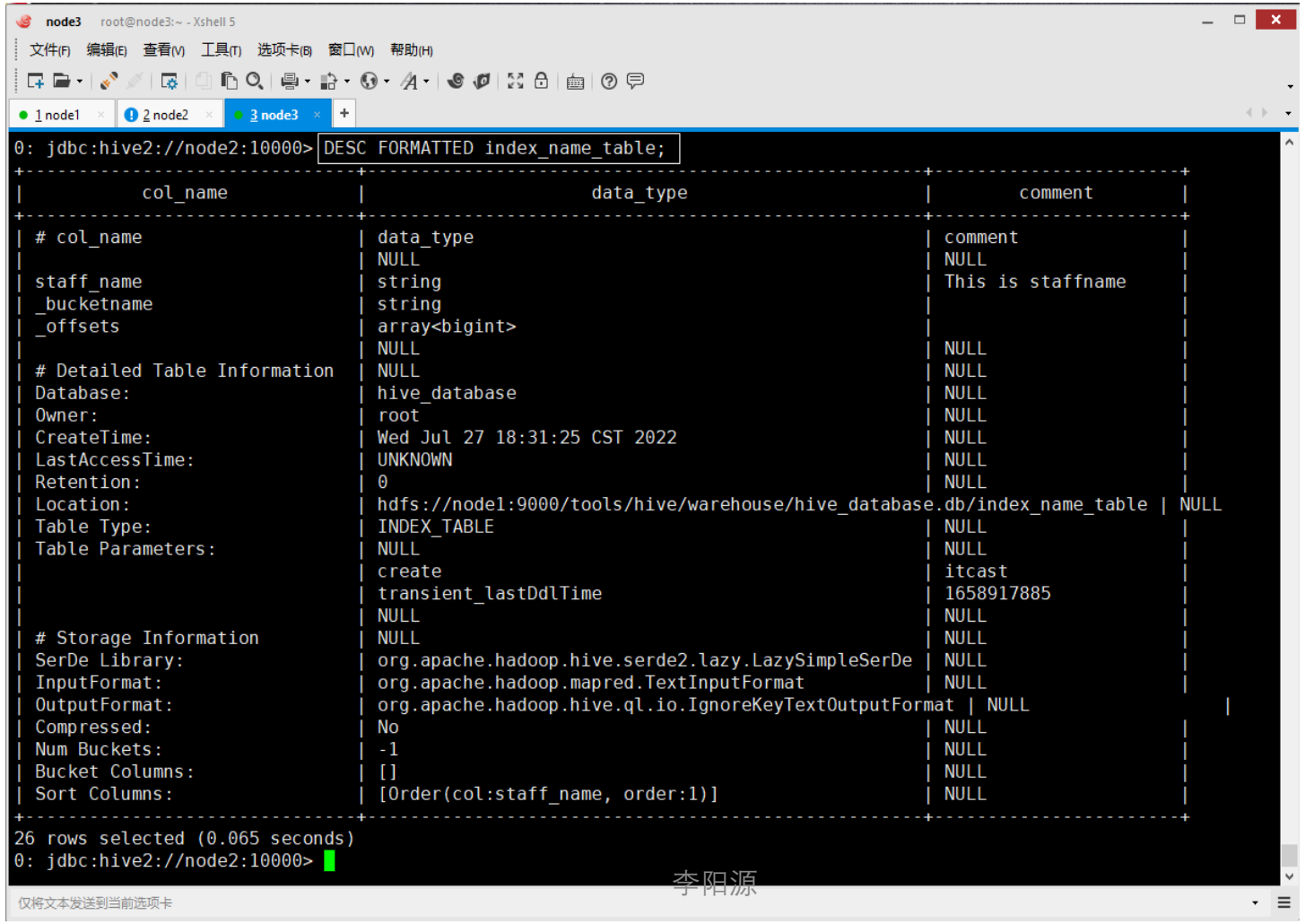

node3 root@node3:~ - Xshell 5
文件(F) 编辑(E) 查看(V) 工具(T) 选项卡(B) 窗口(W) 帮助(H)
1 node1
2 node2
3 node3
0: jdbc:hive2://node2:10000> SHOW TABLES;
tab_name
clustered_table
index_name_table
managed_table_new
partitioned_table
partitioned_table1
5 rows selected (0.055 seconds)
0: jdbc:hive2://node2:10000>
仅将文本发送到当前选项卡

4.6.3 任务实施

查看索引表的详细结构信息，语句如下：

```
DESC FORMATTED index_name_table;
```

执行结果如图4-25所示。



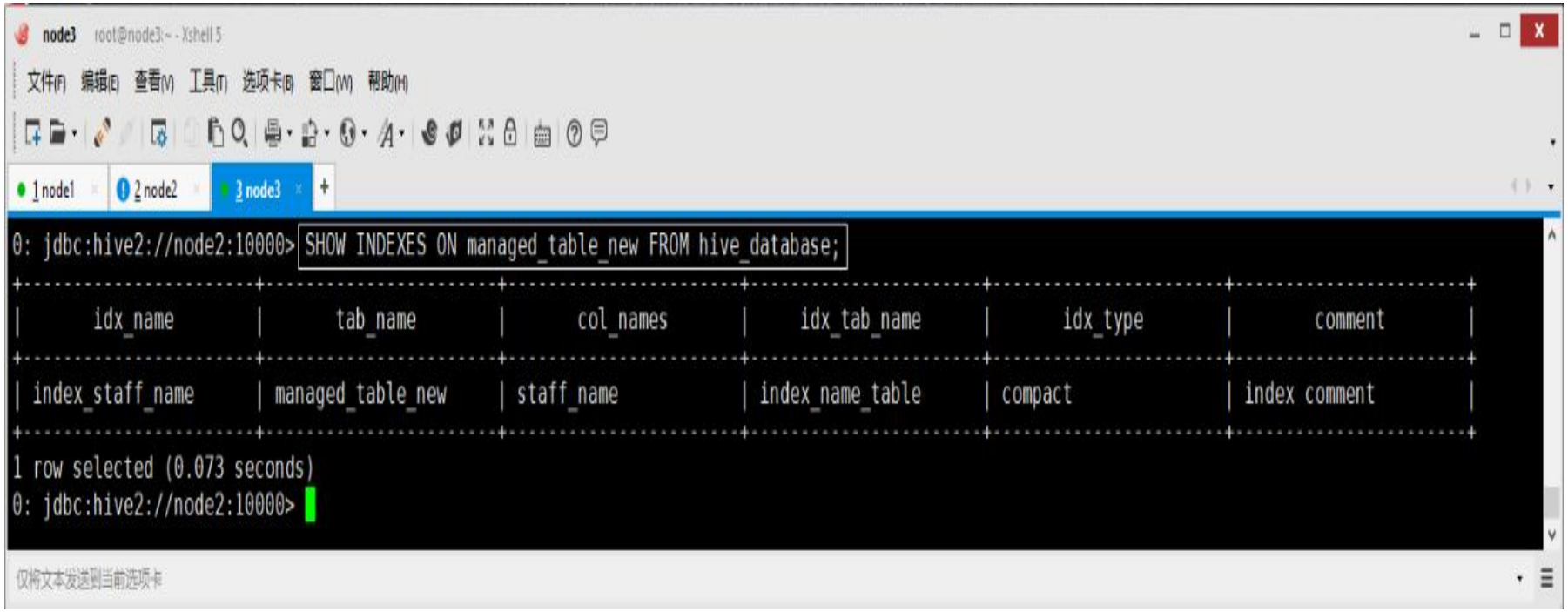
4.6.3 任务实施

3．查看索引

查看数据库hive_database中内部表managed_table_new的索引信息，语句如下：

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

执行结果如图4-26所示。



4.6.3 任务实施

4．重建索引

重建数据库hive_database中内部表managed_table_new的索引index_staff_name，语句如下：

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

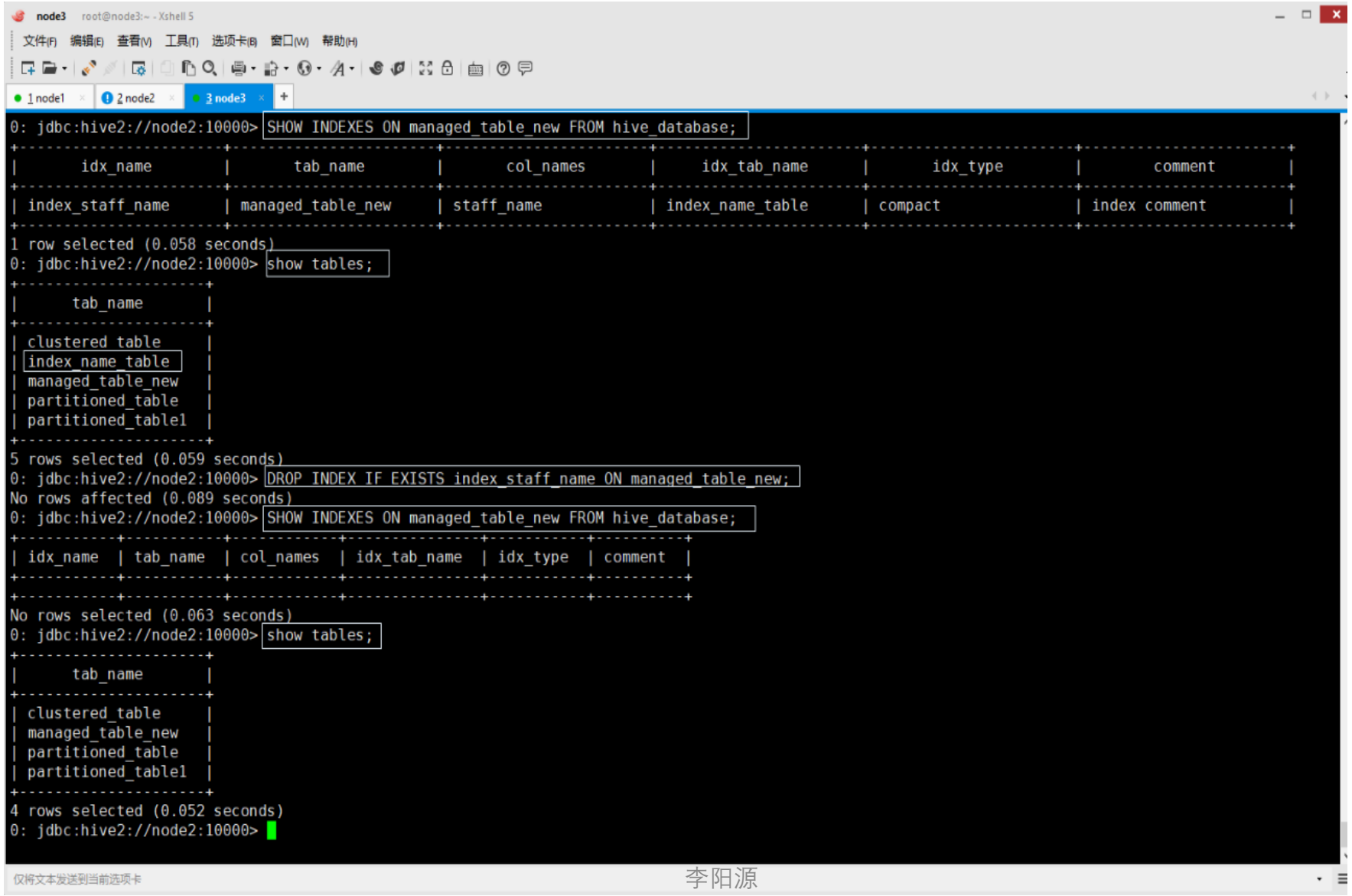
5．删除索引

删除内部表managed_table_new创建的索引index_staff_name，语句如下：

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

执行结果如图4-26所示。

4.6.3 任务实施


node3 root@node3:~ - Xshell 5
文件(F) 编辑(E) 查看(V) 工具(T) 选项卡(B) 窗口(W) 帮助(H)
1 node1
2 node2
3 node3
0: jdbc:hive2://node2:10000> SHOW INDEXES ON managed_table_new FROM hive_database;
| idx_name | tab_name | col_names | idx_tab_name | idx_type | comment |
| index_staff_name | managed_table_new | staff_name | index_name_table | compact | index comment |
1 row selected (0.058 seconds)
0: jdbc:hive2://node2:10000> show tables;
| tab_name |
| clustered_table |
| index_name_table |
| managed_table_new |
| partitioned_table |
| partitioned_table1 |
5 rows selected (0.059 seconds)
0: jdbc:hive2://node2:10000> DROP INDEX IF EXISTS index_staff_name ON managed_table_new;
No rows affected (0.089 seconds)
0: jdbc:hive2://node2:10000> SHOW INDEXES ON managed_table_new FROM hive_database;
| idx_name | tab_name | col_names | idx_tab_name | idx_type | comment |
No rows selected (0.063 seconds)
0: jdbc:hive2://node2:10000> show tables;
| tab_name |
| clustered_table |
| managed_table_new |
| partitioned_table |
| partitioned_table1 |
4 rows selected (0.052 seconds)
0: jdbc:hive2://node2:10000>
仅将文本发送到当前选项卡

项目小结

我们在项目中介绍了Hive查询语言（Hive Query Language简称HiveQL）进行数据定义的方法，首先学习了HiveQL对Hive中的数据库进行创建、查询、显示信息、切换、修改以及删除的操作，HiveQL语句对Hive中的数据表进行创建、查看、修改和删除的操作，HiveQL语句对Hive中的分区表进行创建、查询、添加、重命名、移动和删除的操作，HiveQL语句对Hive中的分桶表进行创建和查看信息的操作，HiveQL语句对Hive中的分桶表进行创建和查看信息的操作，HiveQL语句对Hive中的视图进行创建、查看、修改以及删除的操作，最后介绍了使用HiveQL语句对Hive数据表中的索引进行创建、查看、重建和删除的操作。

【思政小贴士】

在走向“大国工匠”的道路上，必然会经历许多挫折、坎坷和磨难，但是，没有任何一种力量能够阻挡工匠们那颗敬业奉献的心。

因为这样，才能成为行业的佼佼者。

当前，我国正从“生产大国”向“制造大国”转变。

一个制造强国的诞生，必须出现千千万万个“大国工匠”，因此我们每个人都要具有“工匠精神”，只有这样，创新的中国梦才能实现。

在液化天然气船上“缝”钢板的焊接大师张冬伟，其焊接的殷瓦板只有牛皮纸一样薄，手工焊缝长达13公里，如果有一个针眼大小的漏点，都有可能带来致命后果。

面对如此艰辛的任务，他做到了万无一失。

感谢您的观看

