

HIVE

HIVE

- 一、课前准备
- 二、课堂主题
- 三、课堂目标
- 四、知识要点
 - 1.Hive基本概念
 - 2.数据仓库与数据库的区别
 - 3.Hive数据类型
 - 4.创建数据库
 - 5.使用数据库
 - 6.删除数据库
 - 7.创建数据表
 - 8.管理表（内部表）、外部表、分区表
 - 9.修改表
 - 10.删除表
 - 11.插入数据
 - 12.导出数据
- 五、拓展点、未来计划、行业趋势（5分钟）
- 六、总结（5分钟）
- 七、作业
- 八、互动问答
- 九、题库 - 本堂课知识点

一、课前准备

1. vmware虚拟机软件*1
2. centos7虚拟机*3
3. hadoop三节点集群
4. 其中一节点安装hive

二、课堂主题

本节课主要讲解hive的基本概念，数据仓库与数据库的区别，hive的数据类型，DDL操作（建库，删库，建表，内部表、外部表、分区表，删表，更改表，导入、导出数据）

三、课堂目标

1. 了解数据仓库与数据库的区别
2. 掌握hive有哪些数据类型
3. 会使用hive，能完成基础的DDL操作

四、知识要点

1.Hive基本概念

The Apache Hive™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Structure can be projected onto data already in storage. A command line tool and JDBC driver are provided to connect users to Hive.

基于 Hadoop 的一个数据仓库工具：

hive本身不提供数据存储功能，使用HDFS做数据存储，

hive也不分布式计算框架，hive的核心工作就是把sql语句翻译成MR程序

hive也不提供资源调度系统，也是默认由Hadoop当中YARN集群来调度

可以将结构化的数据映射为一张数据库表，并提供 HQL(Hive SQL)查询功能

2.数据仓库与数据库的区别

1. 数据库只存放在当前值，数据仓库存放历史值；
2. 数据库内数据是动态变化的，只要有业务发生，数据就会被更新，而数据仓库则是静态的历史数

据，只能定期添加、刷新；

3. 数据库中的数据结构比较复杂，有各种结构以适合业务处理系统的需要，而数据仓库中的数据结构则相对简单；

4. 数据库中数据访问频率较高，但访问量较少，而数据仓库的访问频率低但访问量却很高；

5. 数据库中数据的目标是面向业务处理人员的，为业务处理人员提供信息处理的支持，而数据仓库则是面向高层管理人员的，为其提供决策支持；

6. 数据库在访问数据时要求响应速度快，其响应时间一般在几秒内，而数据仓库的响应时间则可长达数几小时

3.Hive数据类型

基本数据类型

数据类型	所占字节	支持版本
TINYINT	1字节有符号整数，从-128到127	
SMALLINT	2字节有符号整数，从-32,768至32,767	
INT / INTEGER	4字节有符号整数，从-2,147,483,648到2,147,483,647	
BIGINT	8字节符号整数，从-9,223,372,036,854,775,808至9,223,372,036,854,775,80	
FLOAT	4字节单精度浮点数	
DOUBLE	8字节双精度浮点数	
DATE		仅从 Hive 0.12.0开始提供
STRING		

复杂数据类型

数据类型	描述	示例
ARRAY	一组有序字段。字段的类型必须相同	Array(1,2)
MAP	一组无序的键/值对。键的类型必须是原子的，值可以是任何类型，同一个映射的键的类型必须相同，值得类型也必须相同	Map('a',1,'b',2)
STRUCT	一组命名的字段。字段类型可以不同	Struct('a',1,1,0)

4.创建数据库

语法：CREATE (DATABASE | SCHEMA) [IF NOT EXISTS] database_name

[COMMENT database_comment]

[LOCATION hdfs_path]

[WITH DBPROPERTIES (property_name=property_value, ...)];

例子：

CREATE DATABASE myhive;

5.使用数据库

语法：USE database_name;

USE DEFAULT;

例子：

1.查看数据库

SHOW DATABASE ;

2.使用数据库

USE myhive ;

6.删除数据库

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

例子：DROP DATABASE IF EXISTS myhive CASCADE;

7.创建数据表

直接建表法

```
CREATE TABLE t1(id INT  
  
 ,name STRING  
  
 ,hobby ARRAY<STRING>  
  
 ,add MAP<STRING,STRING>  
  
 ROW FORMAT DELIMITED  
  
 FIELDS TERMINATED BY ','  
  
 COLLECTION ITEMS TERMINATED BY '-'  
  
 MAP KEYS TERMINATED BY ':'  
  
 ;
```

查询建表法

例子：CREATE TABLE t2 AS SELECT id,name FROM t1 ;

like建表法

CREATE TABLE empty_key_value_store

LIKE key_value_store;

例子：CREATE TABLE t3 LIKE t2;

8.管理表（内部表）、外部表、分区表

内部表

```
CREATE TABLE t1(  
  
    id      INT  
  
    ,name   STRING  
  
    ,hobby  ARRAY<STRING>  
  
    ,add    MAP<STRING,STRING>  
  
)  
  
ROW FORMAT DELIMITED  
  
FIELDS TERMINATED BY ','  
  
COLLECTION ITEMS TERMINATED BY '-'  
  
MAP KEYS TERMINATED BY ':'  
  
;
```

外部表

```
CREATE EXTERNAL TABLE t5(  
  
    id      INT  
  
    ,name   STRING  
  
)  
  
ROW FORMAT DELIMITED  
  
FIELDS TERMINATED BY ','  
  
;
```

分区表

```
create table logs ( ts int ,line string ) partitioned by(dt string,country string);
```

```
load data local inpath '/home/hadoop/file1' into table logs partition(dt='2001-01-01',country='GB');
```

```
load data local inpath '/home/hadoop/file2' into table logs partition(dt='2001-01-01',country='GB');
```

9.修改表

修改表名

```
ALTER TABLE old_name RENAME TO new_name;
```

修改列名

```
ALTER TABLE name CHANGE column_name new_name new_type
```

增加列

```
ALTER TABLE emp ADD COLUMNS(age INT);
```

10.删除表

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

例子：DROP TABLE IF EXISTS t4;

11.插入数据

```
LOAD DATA [LOCAL] INPATH'filepath'[OVERWRITE] INTO TABLE tablename [PARTITION ( partcol1 = val1 ,  
partcol2 = val2 ... ) ][INPUTFORMAT'inputformat'SERDE'serde']
```

1.从本地导入到hive表

```
LOAD DATA LOCAL INPATH '/home/hadoop/haha.txt' INTO TABLE t1;
```

2.从hdfs导入到hive表

```
LOAD DATA INPATH '/test' INTO table t5;
```

12.导出数据

保存到本地

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/hadoop/' SELECT * FROM t1;
```

保存到hdfs

```
INSERT OVERWRITE DIRECTORY '/user/hadoop/abc' SLELECT * FROM he;
```

五、拓展点、未来计划、行业趋势（5分钟）

<https://hive.apache.org>

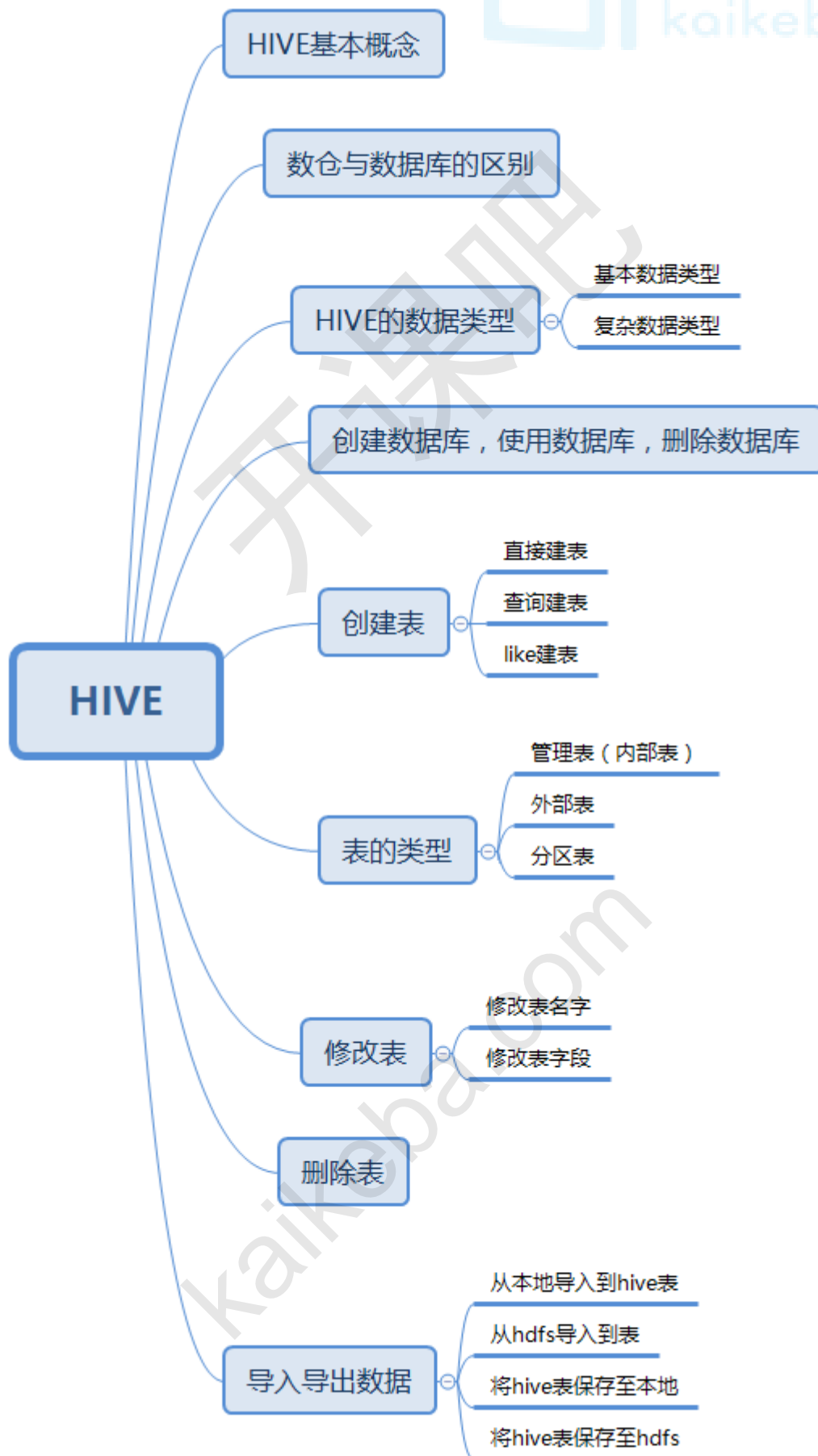
命名规则

库：dws

表：库名_主题域_主题_主体_后缀

字段：满足业务需求，类型合理，需要注释

六、总结（5分钟）



七、作业

- 1.在hadoop集群上任选一节点安装hive
- 2.建库
- 3.并在所建库中建3张表（内部表，外部表，分区表）
- 4.将本地数据导入任一张表中
- 5.将hdfs数据导入任一张表中
- 6.将hive任一表中数据保存至本地
- 7.将hive任一表中数据保存至hdfs

八、互动问答

九、题库 - 本堂课知识点

说明：针对本堂课，设置5-10个题目，来检查学生的掌握情况；