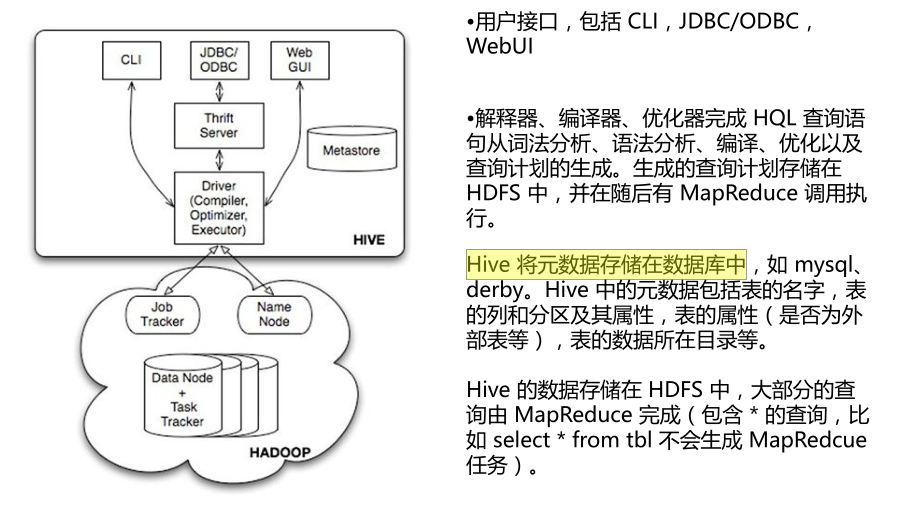
**大数据平台部分实践作业**

# 一、Hive介绍

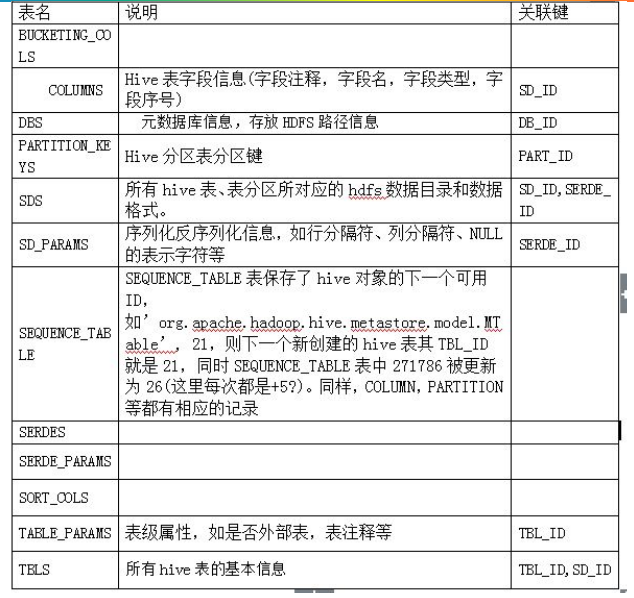
Hive 是基于 Hadoop 的一个数据仓库基础框架，提供了一系列的工具，可以用来进行数据提取、转换、加载（ETL）,这是一种可以存储、查询和分析存储在hadoop的大规模数据机制。可以将结构化的数据文件映射为一张数据库表，并提供类 SQL 查询功能。Hive 的主要功能是将类 SQL 语句转换为 MapReduce 任务运行。

Hive的表其实就是hdfs的目录，按表名把文件夹分开。

## 1.1、Hive的体系结构



## **1.2、元数据库数据字典**



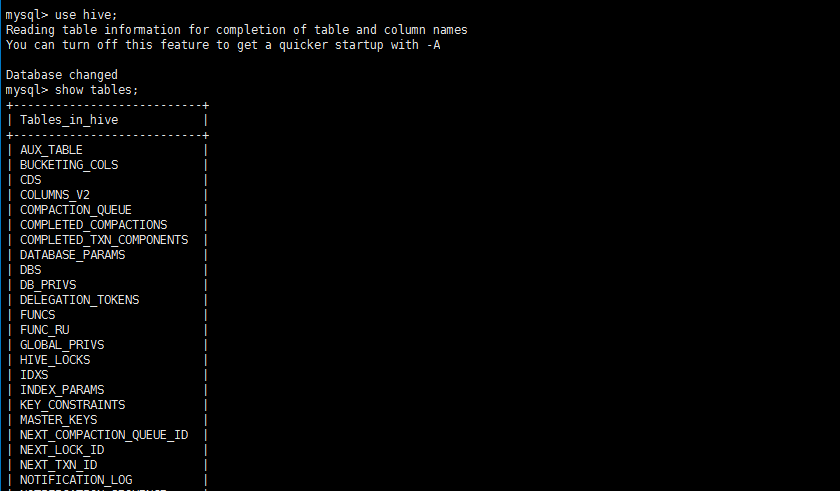
## **1.3、在mysql下观察hive的元数据库**

hive的元数据存储在mysql数据库中,包括表名、表的列、分区及其属性等。

show databases;

use hive;

show tables;



## **1.4、hive表创建表的过程**

解析用户提交hive语句，对其进行解析，分解为表、字段、分区等hive对象

根据解析到的信息构建对应的表、字段、分区等对象，从SEQUENCE\_TABLE中获取构建对象的最新ID，与构建对象信息(名称，类型等)一同通过DAO方法写入到元数据表中去，成功后将SEQUENCE\_TABLE中对应的最新ID+5

## **1.5、内部表外部表**

Hive 创建内部表时，会将数据移动到数据仓库指向的路径；若创建外部表，仅记录数据所在的路径，不对数据的位置做任何改变。在删除表的时候，内部表的元数据和数据会被一起删除，而外部表只删除元数据，不删除数据。这样外部表相对来说更加安全些，数据组织也更加灵活，方便共享源数据。

# Hive创建自己的数据仓库

## **2.1、建立内部表**

1、进入hive创建表结构student

create database cqyd;

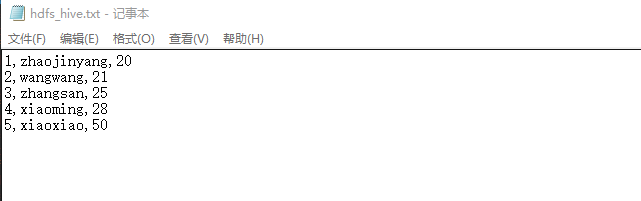
show databases;

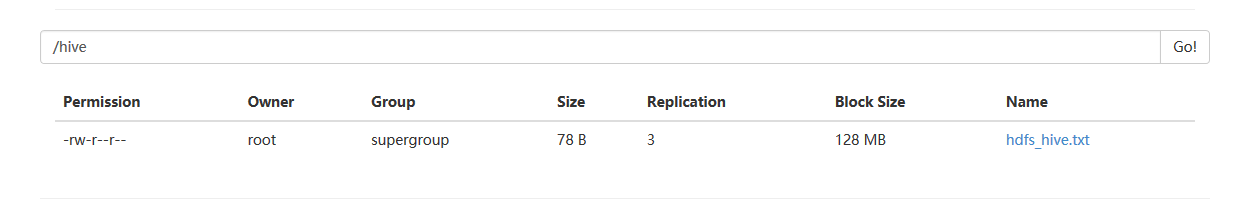
use cqyd;

加载数据方式一：

1. 将student数据放入到hadoop/hive目录下。

hadoop fs -put hdfs\_hive.txt /hive







1. 使用hbase shell命令，创建student表

create table student(id string,name string,age string)

row format delimited

fields terminated by ','

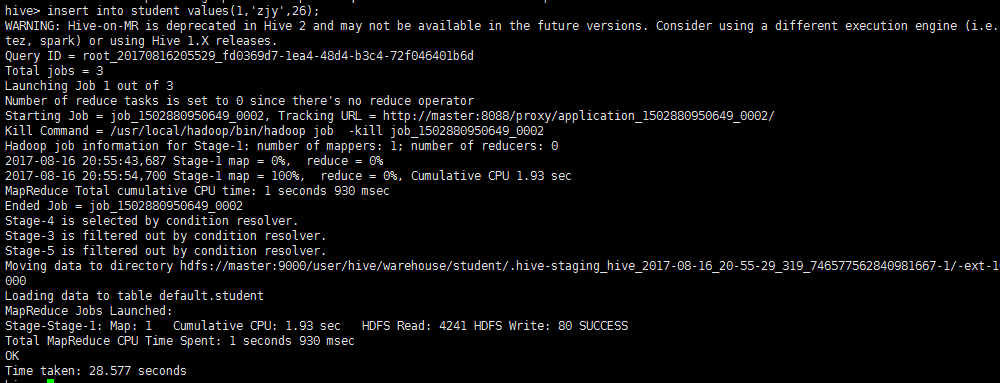
stored as textfile;

4、将hdfs的hive目录下hdfs\_hive.txt加载到hive中，是数据的转移不是复制（内表）

load data inpath '/hive/hdfs\_hive.txt' into table cqyd.student;

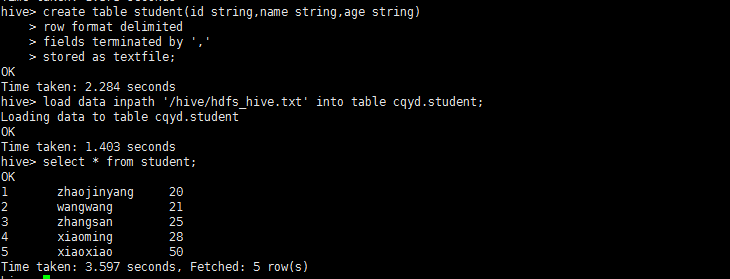
插入数据方法二：

insert into student values(1,'zjy',26);



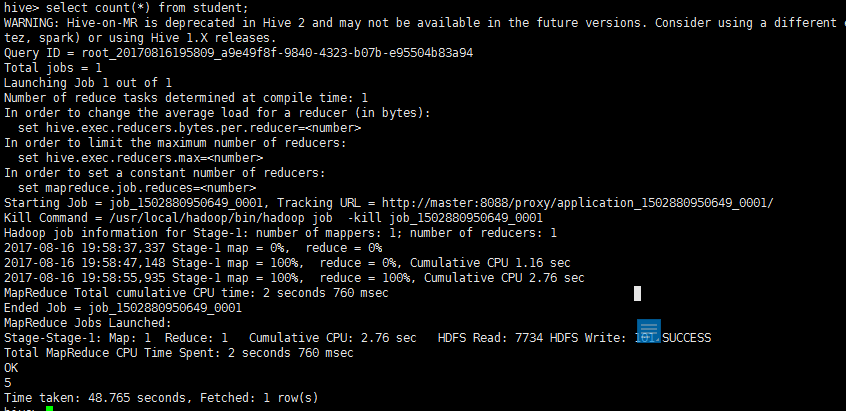
5、使用hbase shell命令查看student表的内容

select \* from student;



6、查询student表中记录数

select count(\*) from student;



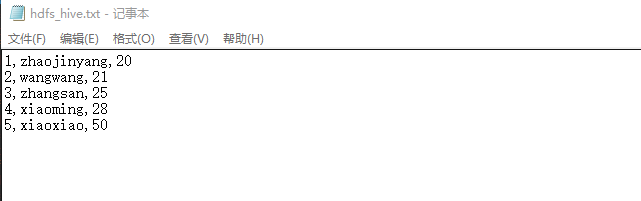
## **2.2、建立外部表**

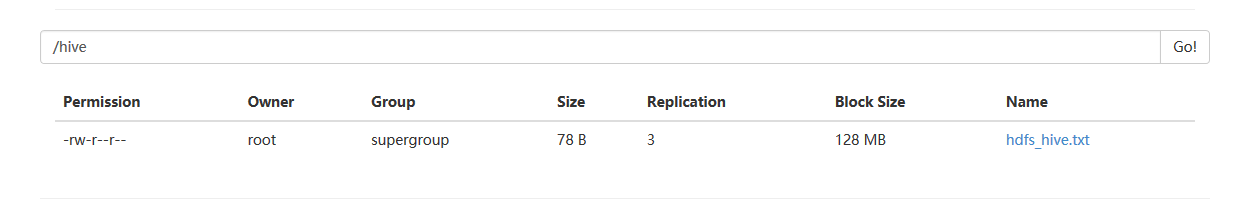
1、进入hive创建表结构student\_exter

加载数据方式一：

1. 将student数据放入到hadoop/hive目录下。

hadoop fs -put hdfs\_hive.txt /hive







1. 使用hbase shell命令，创建student\_exter表

create external table student\_exter(id string,name string,age string)

row format delimited

fields terminated by ','

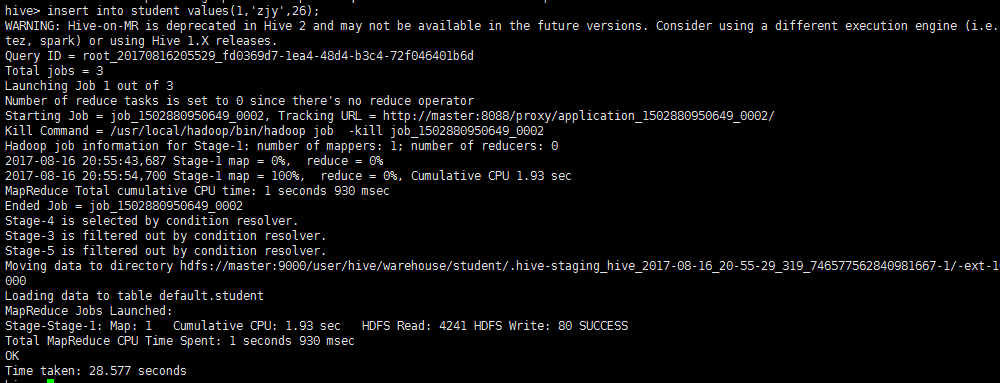
stored as textfile;

4、将hdfs的hive目录下hdfs\_hive.txt加载到hive中

load data inpath '/hive/hdfs\_hive.txt' into table cqyd.student\_exter;

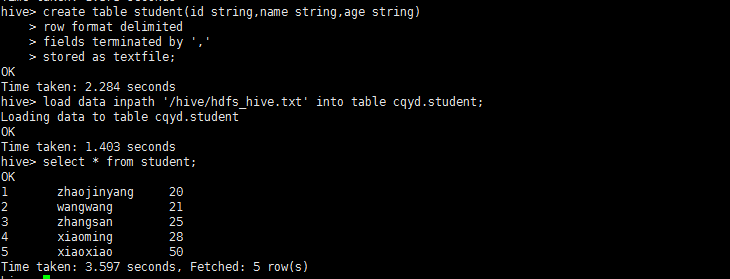
插入数据方法二：

insert into student\_exter values(1,'zjy',26);

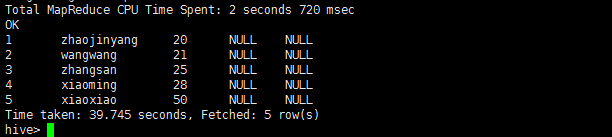


5、使用hbase shell命令查看student表的内容

select \* from student\_exter;



select \* from student\_hive order by age;



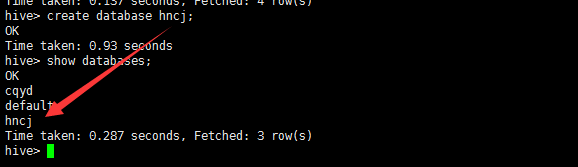
# **Hive在数据仓库上进行增删改查操作**

1. hive创建数据库

create database hncj;

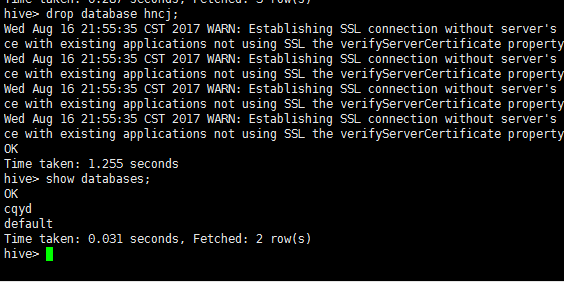
验证数据库

show databases;



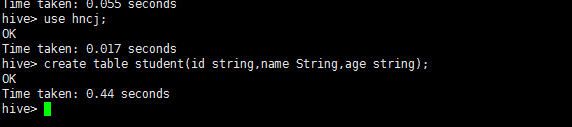
1. hive删除数据库

drop database hncj;



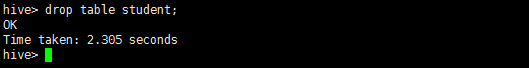
3、hive创建表

create table student(id string,name String,age string);



4、hive删除表

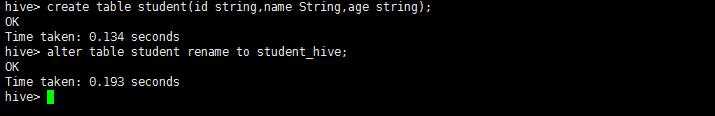
drop table student;



1. hive修改表

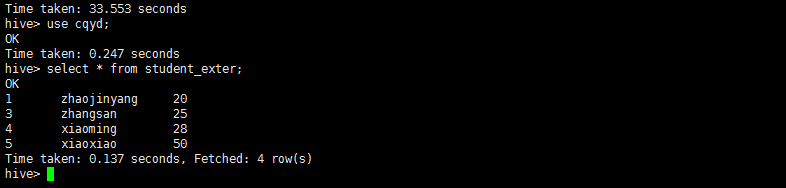
更改表名

alter table student rename to student\_hive;



1. 删除某条数据

insert overwrite table cqyd.student\_exter select id,name,age from cqyd.student\_exter where id !=2;



7、插入数据

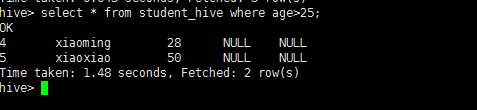
load data inpath '/hive/hdfs\_hive.txt' into table cqyd.student\_exter;

insert into student\_exter values(1,'zjy',26);

1. 查询数据

select \* from student\_exter;

select \* from student\_hive where age>25;

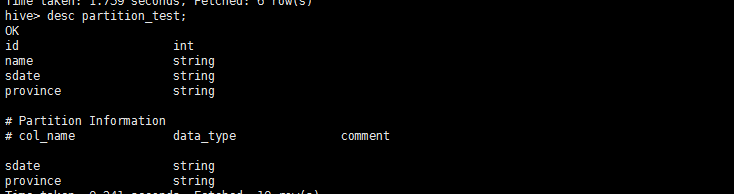


# Hive的Shell操作练习

## **4.1、分区表**

hive中创建分区表没有什么复杂的分区类型(范围分区、列表分区、hash分区、混合分区等)。分区列也不是表中的一个实际的字段，而是一个或者多个伪列。意思是说在表的数据文件中实际上并不保存分区列的信息与数据。使用分区，很容易对数据进行部分查询。

1、分区表  
create table partition\_test  
(id string,  
name string  
)  
partitioned by (  
sdate string,  
province string)  
row format delimited fields terminated by ',';



导入数据临时表

create table partition\_test\_input  
(sdate string,

id string,  
name string,  
province string)  
row format delimited fields terminated by ',';

2、向partition\_test插入sdate='20110728',province='zhejiang'数据

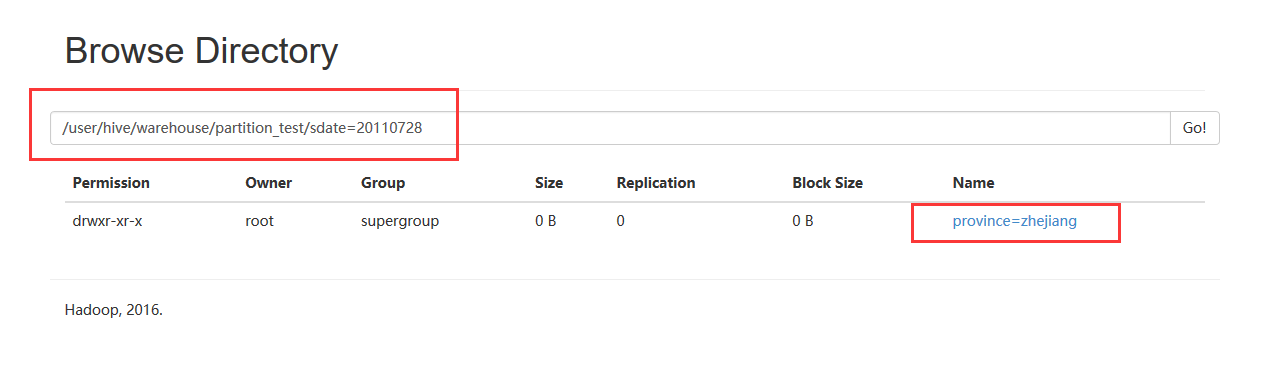
alter table partition\_test add partition (sdate='20110728',province='zhejiang');

IMG_256

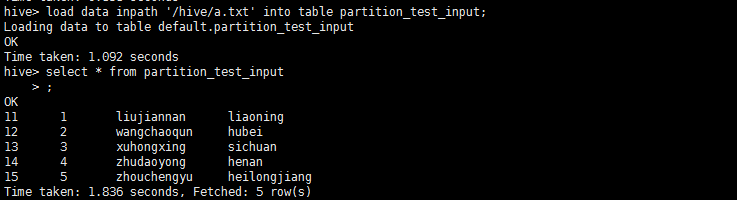
3、在hdfs文件中观察sdate和province位置。

hadoop fs -ls /user/hive/warehouse/partition\_test/sdate=20110728

每一个分区都会有一个独立的文件夹，下面是该分区所有的数据文件。在这个例子中sdate是主层次，province是副层次。



1. 使用一个辅助的非分区表partition\_test\_input准备向partition\_test中插入数据（步骤见2.1）



1. 向partition\_test的分区中插入指定分区数据。

from partition\_test\_input

insert overwrite table partition\_test partition (sdate='1',province='liaoning')

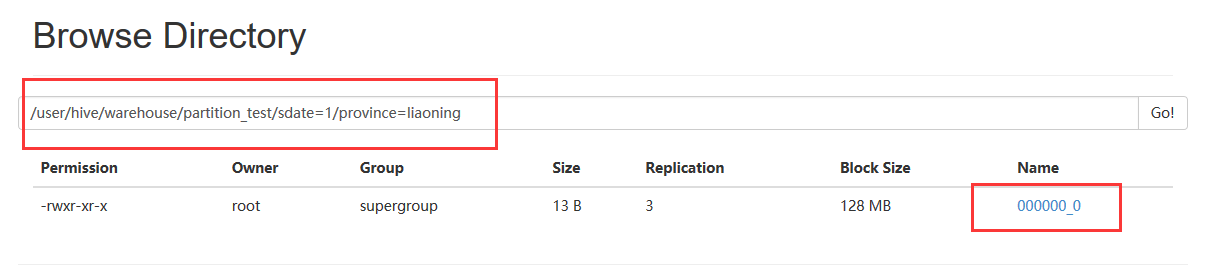
select id,name where sdate='11' and province='liaoning';

IMG_256

1. 在hive中观察表



1. 在hdfs中观察表



IMG_256

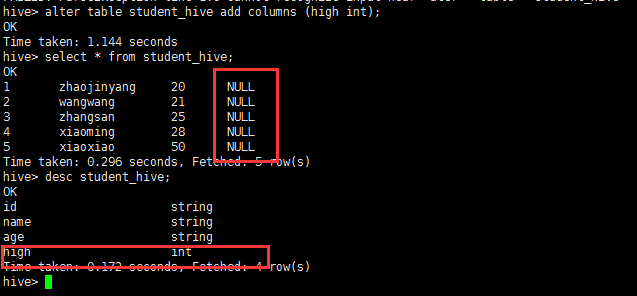
alter使用

在student\_hive表中增加一列

alter table student\_hive add columns (high int);

alter table student\_hive add columns (weight int comment 'a ');

并增加描述



## **4.2、DML对表中数据进行操作**

load data inpath '/hive/hdfs\_hive.txt' into table cqyd.student;

insert into student values(1,'zjy',26);

## **4.3、DLL对表的结构进行操作**

create database hncj;

show databases;

drop database hncj;

create table student(id string,name string,age string)

row format delimited

fields terminated by ','

stored as textfile;

show tables;

drop table student;

desc student;

## **4.4、SQL对hive进行SQL语句的使用**

select \* from student;

alter table student rename to student;

insert overwrite table cqyd.student\_exter select id,name,age from cqyd.student\_exter where id !=2;