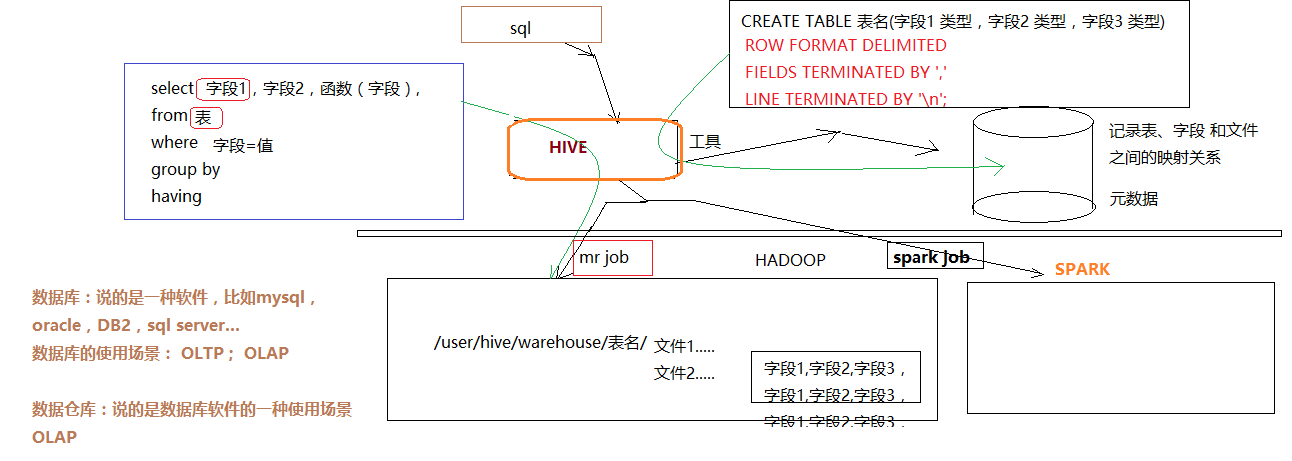
# 什么是hive

## hive基本思想

Hive是基于Hadoop的一个数据仓库工具(离线)，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。



## 为什么使用Hive

* 直接使用hadoop所面临的问题

人员学习成本太高

项目周期要求太短

MapReduce实现复杂查询逻辑开发难度太大

* 为什么要使用Hive

操作接口采用类SQL语法，提供快速开发的能力。

避免了去写MapReduce，减少开发人员的学习成本。

功能扩展很方便。

## Hive的特点

* 可扩展

Hive可以自由的扩展集群的规模，一般情况下不需要重启服务。

* 延展性

Hive支持用户自定义函数，用户可以根据自己的需求来实现自己的函数。

* 容错

良好的容错性，节点出现问题SQL仍可完成执行。

# hive的基本架构



Jobtracker是hadoop1.x中的组件，它的功能相当于：

Resourcemanager+MRAppMaster

TaskTracker 相当于：

Nodemanager + yarnchild

# hive安装

## 最简安装：用内嵌derby作为元数据库

准备工作：安装hive的机器上应该有HADOOP环境（安装目录，HADOOP\_HOME环境变量）

安装：直接解压一个hive安装包即可

此时，安装的这个hive实例使用其内嵌的derby数据库作为记录元数据的数据库

此模式不便于让团队成员之间共享协作

## 标准安装：将mysql作为元数据库

### mysql安装

1. 上传mysql安装包
2. 解压：

[root@mylove ~]# tar -xvf MySQL-5.6.26-1.linux\_glibc2.5.x86\_64.rpm-bundle.tar

1. 安装mysql的server包

[root@mylove ~]# rpm -ivh MySQL-server-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

依赖报错：

缺perl

yum install perl

安装完perl后 ，继续重新安装mysql-server

（可以配置一个本地yum源进行安装：

1、先在vmware中给这台虚拟机连接一个光盘镜像

2、挂在光驱到一个指定目录：mount -t iso9660 -o loop /dev/cdrom /mnt/cdrom

3、将yum的配置文件中baseURL指向/mnt/cdrom

）

[root@mylove ~]# rpm -ivh MySQL-server-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

又出错：包冲突conflict with

移除老版本的冲突包：mysql-libs-5.1.73-3.el6\_5.x86\_64

[root@mylove ~]# rpm -e mysql-libs-5.1.73-3.el6\_5.x86\_64 --nodeps

继续重新安装mysql-server

[root@mylove ~]# rpm -ivh MySQL-server-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

成功后，注意提示：里面有初始密码及如何改密码的信息

初始密码：/root/.mysql\_secret

改密码脚本：/usr/bin/mysql\_secure\_installation

1. 安装mysql的客户端包：

[root@mylove ~]# rpm -ivh MySQL-client-5.6.26-1.linux\_glibc2.5.x86\_64.rpm

1. 启动mysql的服务端：

[root@mylove ~]# service mysql start

Starting MySQL. SUCCESS!

1. 修改root的初始密码：

[root@mylove ~]# /usr/bin/mysql\_secure\_installation 按提示

1. 测试：

用mysql命令行客户端登陆mysql服务器看能否成功

[root@mylove ~]# mysql -uroot -proot

mysql> show databases;

1. 给root用户授予从任何机器上登陆mysql服务器的权限：

mysql> grant all privileges on \*.\* to 'root'@'%' identified by '你的密码' with grant option;

Query OK, 0 rows affected (0.00 sec)

mysql> flush privileges;

Query OK, 0 rows affected (0.00 sec)

注意点：要让mysql可以远程登录访问

最直接测试方法：从windows上用Navicat去连接，能连，则可以，不能连，则要去mysql的机器上用命令行客户端进行授权：

在mysql的机器上,启动命令行客户端：

mysql -uroot -proot

mysql>grant all privileges on \*.\* to 'root'@'%' identified by 'root的密码' with grant option;  
mysql>flush privileges;

### hive的元数据库配置

vi conf/hive-site.xml

|  |
| --- |
| <configuration>  <property>  <name>javax.jdo.option.ConnectionURL</name>  <value>jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true</value>  <description>JDBC connect string for a JDBC metastore</description>  </property>  <property>  <name>javax.jdo.option.ConnectionDriverName</name>  <value>com.mysql.jdbc.Driver</value>  <description>Driver class name for a JDBC metastore</description>  </property>  <property>  <name>javax.jdo.option.ConnectionUserName</name>  <value>root</value>  <description>username to use against metastore database</description>  </property>  <property>  <name>javax.jdo.option.ConnectionPassword</name>  <value>root</value>  <description>password to use against metastore database</description>  </property>  </configuration> |

1. 上传一个mysql的驱动jar包到hive的安装目录的lib中
2. 配置HADOOP\_HOME 和HIVE\_HOME到系统环境变量中：/etc/profile
3. source /etc/profile

5、hive启动测试

然后用命令启动hive交互界面：

[root@hdp20-04 ~]# hive

# hive使用方式

## 最基本使用方式

启动一个hive交互shell

bin/hive

hive>

设置一些基本参数，让hive使用起来更便捷，比如：

1、让提示符显示当前库：

hive>set hive.cli.print.current.db=true;

1. 显示查询结果时显示字段名称：

hive>set hive.cli.print.header=true;

但是这样设置只对当前会话有效，重启hive会话后就失效，解决办法：

在linux的当前用户目录中，编辑一个.hiverc文件，将参数写入其中：

vi .hiverc

|  |
| --- |
| set hive.cli.print.header=true;  set hive.cli.print.current.db=true; |

## 启动hive服务使用

启动hive的服务：

[root@hdp20-04 hive-1.2.1]# bin/hiveserver2 -hiveconf hive.root.logger=DEBUG,console

上述启动，会将这个服务启动在前台，如果要启动在后台，则命令如下：

nohup bin/hiveserver2 1>/dev/null 2>&1 &

**nohup  hive --service hiveserver2 &**

**bin/hiveserver2 &**

启动成功后，可以在别的节点上用beeline去连接

* 方式（1）

[root@hdp20-04 hive-1.2.1]# bin/beeline 回车，进入beeline的命令界面

输入命令连接hiveserver2

beeline> !connect jdbc:hive2://192.168.137.8:10000

（hadoop01是hiveserver2所启动的那台主机名，端口默认是10000）

* 方式（2）

启动时直接连接：

bin/beeline -u jdbc:hive2://master:10000 -n huaqiang

接下来就可以做正常sql查询了

## 脚本化运行

大量的hive查询任务，如果用交互式shell来进行输入的话，显然效率及其低下，因此，生产中更多的是使用脚本化运行机制：

该机制的核心点是：hive可以用一次性命令的方式来执行给定的hql语句

[root@hdp20-04 ~]# hive -e "insert into table t\_dest select \* from t\_src;"

然后，进一步，可以将上述命令写入shell脚本中，以便于脚本化运行hive任务，并控制、调度众多hive任务，示例如下：

vi t\_order\_etl.sh

|  |
| --- |
| #!/bin/bash  hive -e "select \* from db\_order.t\_order"  hive -e "select \* from default.t\_user"  hql="create table default.t\_bash as select \* from db\_order.t\_order"  hive -e "$hql" |

如果要执行的hql语句特别复杂，那么，可以把hql语句写入一个文件：

vi x.hql

|  |
| --- |
| select \* from db\_order.t\_order;  select count(1) from db\_order.t\_user; |

然后，用hive -f /root/x.hql 来执行

# hive建库建表与数据导入

## 建库

hive中有一个默认的库：

库名： default

库目录：hdfs://hdp20-01:9000/user/hive/warehouse

新建库：

create database db\_order;

库建好后，在hdfs中会生成一个库目录：

hdfs://hdp20-01:9000/user/hive/warehouse/db\_order.db

## 建表

### 基本建表语句

use db\_order;

create table t\_order(id string,create\_time string,amount float,uid string);

表建好后，会在所属的库目录中生成一个表目录

/user/hive/warehouse/db\_order.db/t\_order

只是，这样建表的话，hive会认为表数据文件中的字段分隔符为 ^A

正确的建表语句为：

create table t\_order(id string,create\_time string,amount float,uid string)

row format delimited

fields terminated by ',';

这样就指定了，我们的表数据文件中的字段分隔符为 ","

### 删除表

drop table t\_order;

删除表的效果是：

hive会从元数据库中清除关于这个表的信息；

hive还会从hdfs中删除这个表的表目录；

### 内部表与外部表

内部表(MANAGED\_TABLE)：表目录按照hive的规范来部署，位于hive的仓库目录/user/hive/warehouse中

外部表(EXTERNAL\_TABLE)：表目录由建表用户自己指定

create external table t\_access(ip string,url string,access\_time string)

row format delimited

fields terminated by ','

location '/access/log';

外部表和内部表的特性差别：

1. 内部表的目录在hive的仓库目录中 VS 外部表的目录由用户指定
2. drop一个内部表时：hive会清除相关元数据，并删除表数据目录
3. drop一个外部表时：hive只会清除相关元数据；

一个hive的数据仓库，最底层的表，一定是来自于外部系统，为了不影响外部系统的工作逻辑，在hive中可建external表来映射这些外部系统产生的数据目录；

然后，后续的etl操作，产生的各种表建议用managed\_table

### 分区表

**分区表的实质是**：在表目录中为数据文件创建分区子目录，以便于在查询时，MR程序可以针对分区子目录中的数据进行处理，缩减读取数据的范围。

比如，网站每天产生的浏览记录，浏览记录应该建一个表来存放，但是，有时候，我们可能只需要对某一天的浏览记录进行分析

这时，就可以将这个表建为分区表，每天的数据导入其中的一个分区；

当然，每日的分区目录，应该有一个目录名（分区字段）

一个分区字段的实例：

示例如下：

1. 创建带分区的表

|  |
| --- |
| create table t\_access(ip string,url string,access\_time string)  partitioned by(dt string)  row format delimited  fields terminated by ','; |

注意：分区字段不能是表定义中的已存在字段

1. 向分区中导入数据

load data local inpath '/root/access.log.2017-08-04.log' into table t\_access partition(dt='20170804');

load data local inpath '/root/access.log.2017-08-05.log' into table t\_access partition(dt='20170805');

1. 针对分区数据进行查询

*a、统计8月4号的总PV：*

*select count(\*) from t\_access where dt='20170804';*

*实质：就是将分区字段当成表字段来用，就可以使用where子句指定分区了*

*b、统计表中所有数据总的PV：*

*select count(\*) from t\_access;*

*实质：不指定分区条件即可*

多个分区字段示例

建表：

create table t\_partition(id int,name string,age int)

partitioned by(department string,sex string,howold int)

row format delimited fields terminated by ',';

导数据：

load data local inpath '/root/p1.dat' into table t\_partition partition(department='xiangsheng',sex='male',howold=20);

### CTAS建表语法

可以通过已存在表来建表：

1、create table t\_user\_2 like t\_user;

新建的t\_user\_2表结构定义与源表t\_user一致，但是没有数据

1. 在建表的同时插入数据

|  |
| --- |
| create table t\_access\_user  as  select ip,url from t\_access; |

t\_access\_user会根据select查询的字段来建表，同时将查询的结果插入新表中

## 数据导入导出

### 将数据文件导入hive的表

方式1：导入数据的一种方式：

手动用hdfs命令，将文件放入表目录；

方式2：在hive的交互式shell中用hive命令来导入本地数据到表目录

hive>load data local inpath '/root/order.data.2' into table t\_order;

方式3：用hive命令导入hdfs中的数据文件到表目录

hive>load data inpath '/access.log.2017-08-06.log' into table t\_access partition(dt='20170806');

*注意：导本地文件和导HDFS文件的区别：*

*本地文件导入表：复制*

*hdfs文件导入表：移动*

### 将hive表中的数据导出到指定路径的文件

1. 将hive表中的数据导入HDFS的文件

*insert overwrite directory '/root/access-data'*

*row format delimited fields terminated by ','*

*select \* from t\_access;*

1. 将hive表中的数据导入本地磁盘文件

*insert overwrite local directory '/root/access-data'*

*row format delimited fields terminated by ','*

*select \* from t\_access limit 100000;*

## 数据类型

### 数字类型

[TINYINT](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-tinyint) (1-byte signed integer, from -128 to 127)

[SMALLINT](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-smallint) (2-byte signed integer, from -32,768 to 32,767)

[INT](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-int)/INTEGER (4-byte signed integer, from -2,147,483,648 to 2,147,483,647)

[BIGINT](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-bigint) (8-byte signed integer, from -9,223,372,036,854,775,808 to 9,223,372,036,854,775,807)

FLOAT (4-byte single precision floating point number)

DOUBLE (8-byte double precision floating point number)

*示例：*

*create table t\_test(a string ,b int,c bigint,d float,e double,f tinyint,g smallint)*

### 日期时间类型

[TIMESTAMP](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-timestamp) (Note: Only available starting with Hive [0.8.0](https://issues.apache.org/jira/browse/HIVE-2272))

[DATE](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-date) (Note: Only available starting with Hive [0.12.0](https://issues.apache.org/jira/browse/HIVE-4055))

示例，假如有以下数据文件：

|  |
| --- |
| 1,zhangsan,1985-06-30  2,lisi,1986-07-10  3,wangwu,1985-08-09 |

那么，就可以建一个表来对数据进行映射

create table t\_customer(id int,name string,birthday date)

row format delimited fields terminated by ',';

然后导入数据

load data local inpath '/root/customer.dat' into table t\_customer;

然后，就可以正确查询

### 字符串类型

[STRING](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-string)

[VARCHAR](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-varchar) (Note: Only available starting with Hive [0.12.0](https://issues.apache.org/jira/browse/HIVE-4844))

[CHAR](https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types" \l "LanguageManualTypes-char) (Note: Only available starting with Hive [0.13.0](https://issues.apache.org/jira/browse/HIVE-5191))

### 混杂类型

BOOLEAN ： true false | 婚否

BINARY (Note: Only available starting with Hive [0.8.0](https://issues.apache.org/jira/browse/HIVE-2380))

### 复合类型

array数组类型

arrays: ARRAY<data\_type> (Note: negative values and non-constant expressions are allowed as of [Hive 0.14](https://issues.apache.org/jira/browse/HIVE-7325).)

示例：array类型的应用

假如有如下数据需要用hive的表去映射：

|  |
| --- |
| 战狼2,吴京:吴刚:龙母,2017-08-16  三生三世十里桃花,刘亦菲:痒痒,2017-08-20  天下无贼,刘德华:王宝强:刘若英:葛优,2008-08-08  查询所有电影的第一个主演是谁？  查询电影的主演的个数。 |

设想：如果主演信息用一个数组来映射比较方便

建表：

create table t\_movie(moive\_name string,actors array<string>,first\_show date)

row format delimited fields terminated by ','

collection items terminated by ':';

导入数据：

load data local inpath '/root/movie.dat' into table t\_movie;

查询：

select \* from t\_movie;

select moive\_name,actors[0] from t\_movie;

select moive\_name,actors from t\_movie where array\_contains(actors,'吴刚');

select moive\_name,size(actors) from t\_movie;

map类型

maps: MAP<primitive\_type, data\_type> (Note: negative values and non-constant expressions are allowed as of [Hive 0.14](https://issues.apache.org/jira/browse/HIVE-7325).)

1. **假如有以下数据：**

|  |
| --- |
| 1,zhangsan,father:xiaoming#mother:xiaohuang#brother:xiaoxu,28  2,lisi,father:mayun#mother:huangyi#brother:guanyu,22  3,wangwu,father:wangjianlin#mother:ruhua#sister:jingtian,29  4,mayun,father:mayongzhen#mother:angelababy,26 |

可以用一个map类型来对上述数据中的家庭成员进行描述

1. **建表语句：**

create table t\_person(id int,name string,family\_members map<string,string>,age int)

row format delimited fields terminated by ','

collection items terminated by '#'

map keys terminated by ':';

1. **查询**

select \* from t\_person;

*## 取map字段的指定key的值*

select id,name,family\_members['father'] as father from t\_person;

## 取map字段的所有key

select id,name,map\_keys(family\_members) as relation from t\_person;

## 取map字段的所有value

select id,name,map\_values(family\_members) from t\_person;

select id,name,map\_values(family\_members)[0] from t\_person;

*## 综合：查询有brother的用户信息*

|  |
| --- |
| select id,name,father  from  (select id,name,family\_members['brother'] as father from t\_person) tmp  where father is not null; |

struct类型

structs: STRUCT<col\_name : data\_type, ...>

1. **假如有如下数据：**

|  |
| --- |
| 1,zhangsan,18:male:beijing  2,lisi,28:female:shanghai |

其中的用户信息包含：年龄：整数，性别：字符串，地址：字符串

设想用一个字段来描述整个用户信息，可以采用struct

1. **建表：**

*create table t\_person\_struct(id int,name string,info struct<age:int,sex:string,addr:string>)*

*row format delimited fields terminated by ','*

*collection items terminated by ':';*

1. **查询**

*select \* from t\_person\_struct;*

*select id,name,info.age from t\_person\_struct;*

## 修改表定义

仅修改Hive元数据，不会触动表中的数据，用户需要确定实际的数据布局符合元数据的定义。

修改表名：

ALTER TABLE table\_name RENAME TO new\_table\_name

示例：alter table t\_1 rename to t\_x;

修改分区名：

alter table t\_partition partition(department='xiangsheng',sex='male',howold=20) rename to partition(department='1',sex='1',howold=20);

添加分区：

alter table t\_partition add partition (department='2',sex='0',howold=40);

删除分区：

alter table t\_partition drop partition (department='2',sex='2',howold=24);

# hive查询语法

提示：在做小数据量查询测试时，可以让hive将mrjob提交给本地运行器运行，可以在hive会话中设置如下参数：

hive> set hive.exec.mode.local.auto=true;

## 基本查询示例

select \* from t\_access;

select count(\*) from t\_access;

select max(ip) from t\_access;

## 条件查询

select \* from t\_access where access\_time<'2017-08-06 15:30:20'

select \* from t\_access where access\_time<'2017-08-06 16:30:20' and ip>'192.168.33.3';

## join关联查询示例

假如有a.txt文件

|  |
| --- |
| a,1  b,2  c,3  d,4 |

假如有b.txt文件

|  |
| --- |
| a,xx  b,yy  d,zz  e,pp |

进行各种join查询：

1. inner join（join）

|  |
| --- |
| select  a.name as aname,  a.numb as anumb,  b.name as bname,  b.nick as bnick  from t\_a a  join t\_b b  on a.name=b.name |

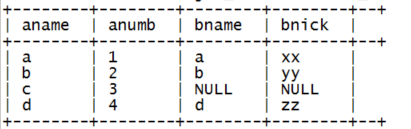
结果：

|  |
| --- |
| +--------+--------+--------+--------+--+  | aname | anumb | bname | bnick |  +--------+--------+--------+--------+--+  | a | 1 | a | xx |  | b | 2 | b | yy |  | d | 4 | d | zz |  +--------+--------+--------+--------+--+ |

1. left outer join（left join）

|  |
| --- |
| select  a.name as aname,  a.numb as anumb,  b.name as bname,  b.nick as bnick  from t\_a a  left outer join t\_b b  on a.name=b.name |

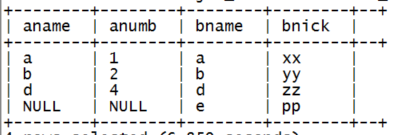
结果：



1. right outer join（right join）

|  |
| --- |
| select  a.name as aname,  a.numb as anumb,  b.name as bname,  b.nick as bnick  from t\_a a  right outer join t\_b b  on a.name=b.name |

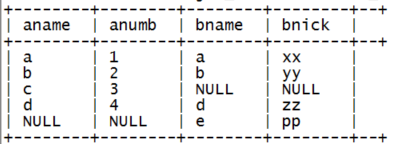
结果：



1. full outer join（full join）

|  |
| --- |
| select  a.name as aname,  a.numb as anumb,  b.name as bname,  b.nick as bnick  from t\_a a  full join t\_b b  on a.name=b.name; |

结果：

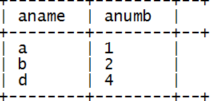


## left semi join

hive中不支持exist/IN子查询，可以用left semi join来实现同样的效果：

|  |
| --- |
| select  a.name as aname,  a.numb as anumb  from t\_a a  left semi join t\_b b  on a.name=b.name; |

结果：



*注意： left semi join的 select子句中，不能有右表的字段*

## group by分组聚合

select dt,count(\*),max(ip) as cnt from t\_access group by dt;

select dt,count(\*),max(ip) as cnt from t\_access group by dt having dt>'20170804';

select

dt,count(\*),max(ip) as cnt

from t\_access

where url='http://www.edu360.cn/job'

group by dt having dt>'20170804';

注意： 一旦有group by子句，那么，在select子句中就不能有 （分组字段，聚合函数） 以外的字段

## 为什么where必须写在group by的前面，为什么group by后面的条件只能用having

因为，where是用于在真正执行查询逻辑之前过滤数据用的

having是对group by聚合之后的结果进行再过滤；

*上述语句的执行逻辑：*

1. *where过滤不满足条件的数据*
2. *用聚合函数和group by进行数据运算聚合，得到聚合结果*
3. *用having条件过滤掉聚合结果中不满足条件的数据*

## 子查询

select id,name,father

from

(select id,name,family\_members['brother'] as father from t\_person) tmp

where father is not null;

# hive函数使用

小技巧：测试函数的用法，可以专门准备一个专门的dual表

~~create table dual(x string);~~

~~insert into table dual values('');~~

***其实：直接用常量来测试函数即可***

select substr("abcdefg",1,3);

hive的所有函数手册：

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF#LanguageManualUDF-Built-inTable-GeneratingFunctions(UDTF)

## 常用内置函数

### 类型转换函数

select cast("5" as int) ~~from dual;~~

select cast("2017-08-03" as date) ;

select cast(current\_timestamp as date);

*示例：*

|  |  |  |
| --- | --- | --- |
| *1* | *1995-05-05 13:30:59* | *1200.3* |
| *2* | *1994-04-05 13:30:59* | *2200* |
| *3* | *1996-06-01 12:20:30* | *80000.5* |

*create table t\_fun(id string,birthday string,salary string)*

*row format delimited fields terminated by ',';*

*select id,cast(birthday as date) as bir,cast(salary as float) from t\_fun;*

### 数学运算函数

select round(5.4) from dual; ## 5

select round(5.1345,3) from dual; ##5.135

select ceil(5.4) from dual; // select ceiling(5.4) from dual; ## 6

select floor(5.4) from dual; ## 5

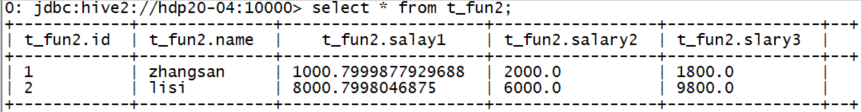
select abs(-5.4) from dual; ## 5.4

select greatest(3,5,6) from dual; ## 6 求多列最大值

select least(3,5,6) from dual;

示例：

有表如下：



select greatest(cast(s1 as double),cast(s2 as double),cast(s3 as double)) from t\_fun2;

结果：

+---------+--+

| \_c0 |

+---------+--+

| 2000.0 |

| 9800.0 |

+---------+--+

select max(age) from t\_person; 聚合函数

select min(age) from t\_person; 聚合函数

### 字符串函数

substr(string, int start) ## 截取子串

substring(string, int start)

*示例：select substr("abcdefg",2) from dual;*

**substr(string, int start, int len)**

substring(string, int start, int len)

*示例：select substr("abcdefg",2,3) from dual;*

**concat**(string A, string B...) ## 拼接字符串

**concat\_ws**(string SEP, string A, string B...)

*示例：select concat("ab","xy") from dual;*

*select concat\_ws(".","192","168","33","44") from dual;*

**length**(string A)

*示例：select length("192.168.33.44") from dual;*

**split(string str, string pat)**

示例：~~select split("192.168.33.44",".") from dual;~~ 错误的，因为.号是正则语法中的特定字符

select split("192.168.33.44","\\.") from dual;

**upper**(string str) ##转大写

### 时间函数

select current\_timestamp;

select current\_date;

## 取当前时间的毫秒数时间戳

select **unix\_timestamp()**;

## unix时间戳转字符串

**from\_unixtime**(bigint unixtime[, string format])

示例：select from\_unixtime(unix\_timestamp());

select from\_unixtime(unix\_timestamp(),"yyyy/MM/dd HH:mm:ss");

## 字符串转unix时间戳

unix\_timestamp(string date, string pattern)

示例： select unix\_timestamp("2017-08-10 17:50:30");

select unix\_timestamp("2017/08/10 17:50:30","yyyy/MM/dd HH:mm:ss");

## 将字符串转成日期date

select **to\_date**("2017-09-17 16:58:32");

### 表生成函数

行转列函数：explode()

假如有以下数据：

|  |
| --- |
| 1,zhangsan,化学:物理:数学:语文  2,lisi,化学:数学:生物:生理:卫生  3,wangwu,化学:语文:英语:体育:生物  查询：考了化学科目的总数。 |

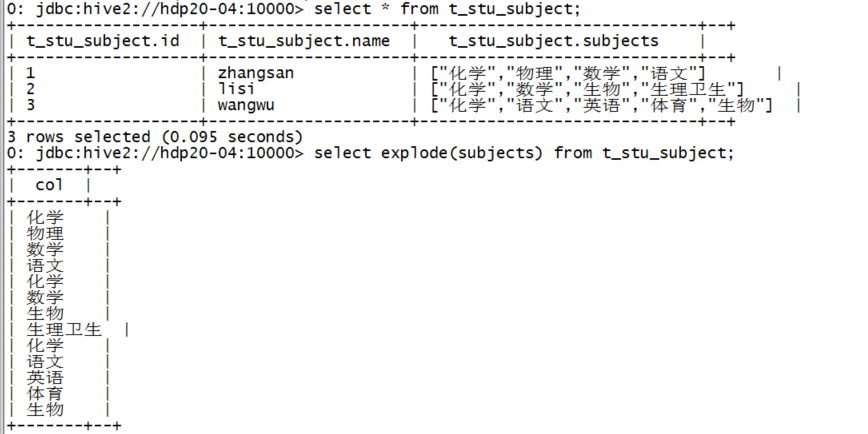
映射成一张表：

create table t\_stu\_subject(id int,name string,subjects array<string>)

row format delimited fields terminated by ','

collection items terminated by ':';

使用explode()对数组字段“炸裂”



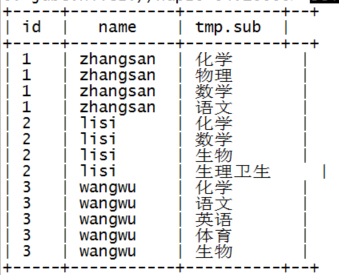
然后，我们利用这个explode的结果，来求去重的课程：

|  |
| --- |
| select distinct tmp.sub  from  (select explode(subjects) as sub from t\_stu\_subject) tmp; |

表生成函数lateral view

select id,name,tmp.sub

from t\_stu\_subject lateral view explode(subjects) tmp as sub;



*理解： lateral view 相当于两个表在join*

*左表：是原表*

*右表：是explode(某个集合字段)之后产生的表*

*而且：这个join只在同一行的数据间进行*

*那样，可以方便做更多的查询：*

*比如，查询选修了生物课的同学*

*select a.id,a.name,a.sub from*

(select id,name,tmp.sub as sub from t\_stu\_subject lateral view explode(subjects) tmp as sub) a

where sub='生物';

### 集合函数

array\_contains(Array<T>, value) 返回boolean值

示例：

*select moive\_name,array\_contains(actors,'吴刚') from t\_movie;*

*select array\_contains(array('a','b','c'),'c') from dual;*

sort\_array(Array<T>) 返回排序后的数组

*示例：*

*select sort\_array(array('c','b','a')) from dual;*

*select 'haha',sort\_array(array('c','b','a')) as xx from (select 0) tmp;*

size(Array<T>) 返回一个int值

*示例：*

*select moive\_name,size(actors) as actor\_number from t\_movie;*

size(Map<K.V>) 返回一个int值

map\_keys(Map<K.V>) 返回一个数组

map\_values(Map<K.V>) 返回一个数组

### 条件控制函数

case when

语法：

CASE [ expression ]

WHEN condition1 THEN result1

WHEN condition2 THEN result2

...

WHEN conditionn THEN resultn

ELSE result

END

示例：

select id,name,

case

when age<28 then 'youngth'

when age>27 and age<40 then 'zhongnian'

else 'old'

end

from t\_user;

IF

*select id,if(age>25,'working','worked') from t\_user;*

*select moive\_name,if(array\_contains(actors,'吴刚'),'好电影','rom t\_movie;*

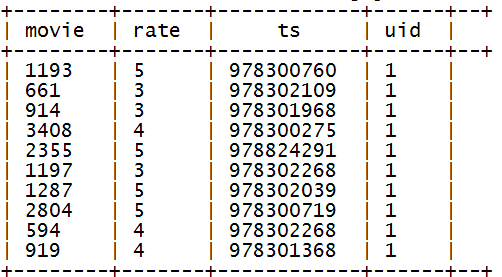
### json解析函数：表生成函数

json\_tuple函数

示例：

select json\_tuple(json,'movie','rate','timeStamp','uid') as(movie,rate,ts,uid) from t\_rating\_json;

产生结果：



create table t\_rating(

info string

);

利用json\_tuple从原始json数据表中，etl出一个详细信息表：

|  |
| --- |
| create table t\_rate  as  select  uid,  movie,  rate,  year(from\_unixtime(cast(ts as bigint))) as year,  month(from\_unixtime(cast(ts as bigint))) as month,  day(from\_unixtime(cast(ts as bigint))) as day,  hour(from\_unixtime(cast(ts as bigint))) as hour,  minute(from\_unixtime(cast(ts as bigint))) as minute,  from\_unixtime(cast(ts as bigint)) as ts  from  (select  json\_tuple(rateinfo,'movie','rate','timeStamp','uid') as(movie,rate,ts,uid)  from t\_json) tmp  ; |

### 分析函数：row\_number() over()——分组TOPN

需求

有如下数据：

|  |
| --- |
| 1,18,a,male  2,19,b,male  3,22,c,female  4,16,d,female  5,30,e,male  6,26,f,female |

需要查询出每种性别中年龄最大的2条数据

实现：

使用row\_number函数，对表中的数据按照性别分组，按照年龄倒序排序并进行标记

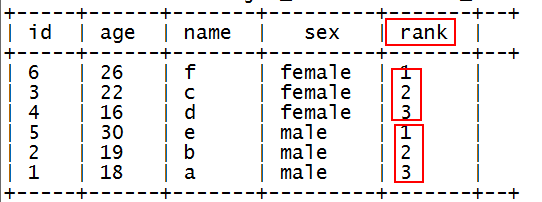
hql代码：

select id,age,name,sex,

row\_number() over(partition by sex order by age desc) as rank

from t\_rownumber

产生结果：



然后，利用上面的结果，查询出rank<=2的即为最终需求

select id,age,name,sex

from

(select id,age,name,sex,

row\_number() over(partition by sex order by age desc) as rank

from t\_rownumber) tmp

where rank<=2;

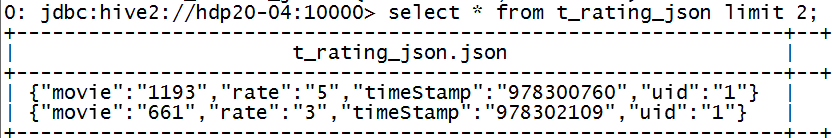
*练习：求出电影评分数据中，每个用户评分最高的topn条数据*

## 自定义函数

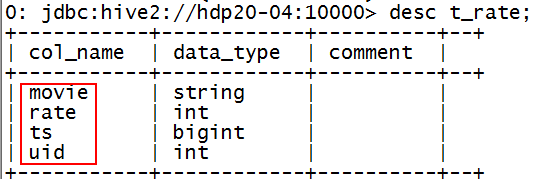
### 需求：

需要对json数据表中的json数据写一个自定义函数，用于传入一个json，返回一个数据值的数组

json原始数据表：



需要做ETL操作，将json数据变成普通表数据，插入另一个表中：



### 实现步骤：

1、开发JAVA的UDF类

|  |
| --- |
| **public** **class** ParseJson **extends** UDF{    // 重载 ：返回值类型 和参数类型及个数，完全由用户自己决定  // 本处需求是：给一个字符串，返回一个数组  **public** String[] evaluate(String json) {    String[] split = json.split("\"");  String[] res = **new** String[]{split[3],split[7],split[11],split[15]};  **return** res;  }  } |

2、打jar包

在eclipse中使用export即可

1. 上传jar包到运行hive所在的linux机器
2. 在hive中创建临时函数：

在hive的提示符中：

*hive> add jar /root/jsonparse.jar;*

然后，在hive的提示符中，创建一个临时函数：

hive>CREATE TEMPORARY FUNCTION jsonp AS 'cn.edu360.hdp.hive.ParseJson';

1. 开发hql语句，利用自定义函数，从原始表中抽取数据插入新表

|  |
| --- |
| insert into table t\_rate  select  split(jsonp(json),',')[0],  cast(split(jsonp(json),',')[1] as int),  cast(split(jsonp(json),',')[2] as bigint),  cast(split(jsonp(json),',')[3] as int)  from  t\_rating\_json; |

*注：临时函数只在一次hive会话中有效，重启会话后就无效*

如果需要经常使用该自定义函数，可以考虑创建永久函数：

拷贝jar包到hive的类路径中：

cp wc.jar apps/hive-1.2.1/lib/

创建了：

create function pfuncx as 'com.doit.hive.udf.UserInfoParser';

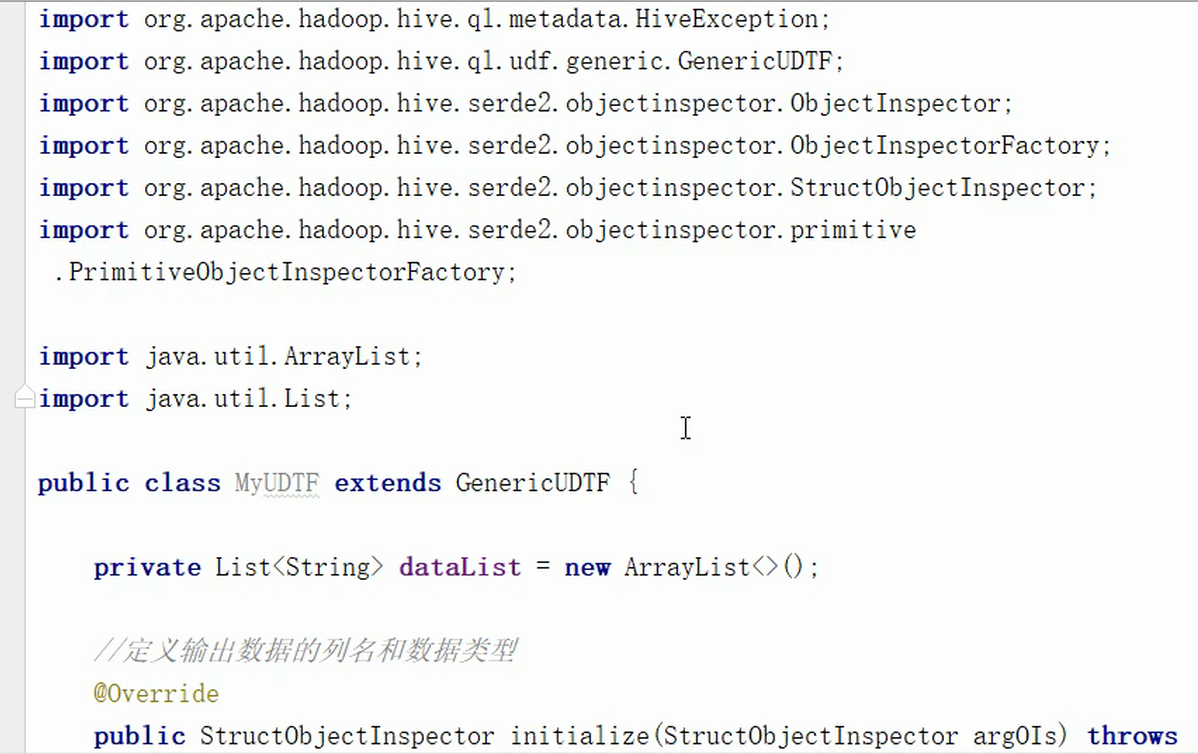
删除函数：

DROP TEMPORARY FUNCTION [IF EXISTS] function\_name

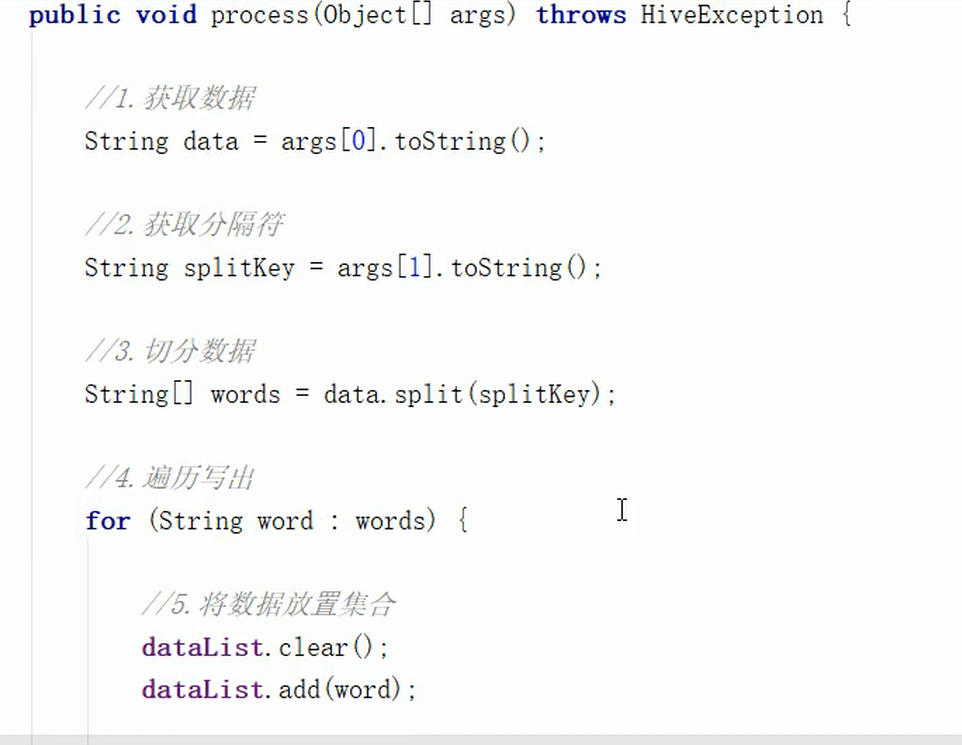
DROP FUNCTION[IF EXISTS] function\_name

7.1 UDTF

步骤：继承GenericUDTF 类，重写3个方法









# 综合查询案例

## 用hql来做wordcount

有以下文本文件：

|  |
| --- |
| hello tom hello jim  hello rose hello tom  tom love rose rose love jim  jim love tom love is what  what is love |

需要用hive做wordcount

-- 建表映射

create table t\_wc(sentence string);

-- 导入数据

load data local inpath '/root/hivetest/xx.txt' into table t\_wc;

hql答案：

|  |
| --- |
| SELECT word  ,count(1) as cnts  FROM (  SELECT explode(split(sentence, ' ')) AS word  FROM t\_wc  ) tmp  GROUP BY word  order by cnts desc  ; |

## 级联报表查询

有如下数据：

select sum(money) from t\_table group by 1,2;

|  |
| --- |
| A,2015-01,5  A,2015-01,15  B,2015-01,5  A,2015-01,8  B,2015-01,25  A,2015-01,5  C,2015-01,10  C,2015-01,20  A,2015-02,4  A,2015-02,6  C,2015-02,30  C,2015-02,10  B,2015-02,10  B,2015-02,5  A,2015-03,14  A,2015-03,6  B,2015-03,20  B,2015-03,25  C,2015-03,10  C,2015-03,20 |

建表映射：

|  |
| --- |
| create table t\_access\_times(username string,month string,counts int)  row format delimited fields terminated by ','; |

需要要开发hql脚本，来统计出如下累计报表：

|  |  |  |  |
| --- | --- | --- | --- |
| 用户 | 月份 | 月总额 | 累计到当月的总额 |
| A | 2015-01 | 33 | 33 |
| A | 2015-02 | 10 | 43 |
| A | 2015-03 | 30 | 73 |
| B | 2015-01 | 30 | 30 |
| B | 2015-02 | 15 | 45 |
|  |  |  |  |

*答案： 参见《复杂sql案例--套路--累计报表.txt》*