# HIVE安装

## 1、下载并解压

|  |
| --- |
| hive.apache.org  注意版本。1.2.1(?)之后，之前就是基于HDFS, 就是基于spark了 |
|  |

## 2、配置连接数据库

|  |
| --- |
| 新建$HIVE\_HOME/conf/hive-site.xml  配置内容  <configuration>  <property>  <name>javax.jdo.option.ConnectionURL</name>  <value>jdbc:mysql://localhost:3306/hive?createDatabaseIfNotExist=true</value>  </property>  <property>  <name>javax.dao.option.ConnectionDriverName</name>  <value>com.mysql.jdbc.driver</value>  </property>  <property>  <name>javax.jdo.option.ConnectionUserName</name>  <value>root</value>  </property>  <property>  <name>javax.jdo.option.ConnectionPassword</name>  <value>123456</value>  </property>  </configuration> |
|  |

## 2、将映射的数据库的驱动放入hive的lib文件夹中

## 3、配置环境变量（便于使用）

|  |
| --- |
| vim /etc/profile  新增  #hive  export HIVE\_HOME=/opt/module/hive-1.2.1  export PATH=$PATH:$HIVE\_HOME/bin  使配置生效  source /etc/profile |
|  |

## 4、启动交互程序（使用一）

|  |
| --- |
| hive启动的前提，必需运行在hdfs上，所以必需保证安装有hadoop，配置好了hdfs  使用命令hive,进入交互模式。  此模式下，可以使用sql进行操作  set hive.cli.print.current.db=true; 设置在hive交货界面显示当前操作的库  此种交互模式下，只对当前会话有效  可以在当前用户的家目录中，新建.hiverc , 写入以上命令。hive启动时就会去读这个文件 |
|  |

## 5、启动服务与客户商交与程序（使用二）

|  |
| --- |
| 1、启动服务  bin/hiveserver2  查看是否启动好  netstat –nlp 查找是否在监听10000端口  静默启动服务  nohup hiveserver2 1>/dev/null 2>&1 $  nohup 用户退出，也运行服务  1>/dev/null 将标准输出清出  2>&1 错误输出也清出  2、启动客户端  bin./beeline  3、连接  !connect jdbc:hive2://192.168.111.101:10000  4、输入hadoop的用户，回车。不用输入密码，回车  或者直接连  beeline –u jdbc:hive2://hadoop101:10000 –n root |
|  |

## 6、使用脚本（使用三）

|  |
| --- |
| 1、方式一  hive –e “ sql ”  以上可以直接以命令形式使用。（不需要启动服务 或者 启动交互界面）  所以可以新建shell脚本，  !/bin/bash  hive –e “sql1”  hive –e “sql2”  2、方式二  可以将大量sql写在一个文本文件里面  使用 hive –f file 命令执行 |

# 2、操作

## 1、建表

|  |
| --- |
| 1、内部表：表目录是按照hive的规范来部署的，位于hive仓库目录/user/hive/warehouse中  create table test(id string ,name string)  row format delimited fields  terminated by ',' #指定分割符  2、外部表：表目录由建目录由用户自己指定  #新建外部表  create external table test(id string ,name string)  row format delimited fields  terminated by ',' #指定分割符  location “/access/log”  3、分区表  create table t\_pv\_log (ip string ,url string,access\_time string)  partitioned by (day string)  row format delimited fields terminated by ',' #指定分割符  导入数据时，才会新建目录  load data local inpath '/opt/module/hadoop-2.7.2/hiveFile' into table t\_pv\_log partition(day='20170915');  分区字段也属于表的字段  create table table\_name like tab1 新建一个我tab1不一样的表  #将新建一个我tab1不一样的表，并将查询的数据存入其中  create table table\_name  as  select \* from tab1 where [condigion] |

## 2、删表

|  |
| --- |
| drop table table\_name  内部表hive会将hdfs中的表目录删除，还会删除数据库中的表定义  外部表hive会将 还会删除数据库中的表定义 |

## 3、导入数据

|  |
| --- |
| 1、方式一  手动用hdfs命令，将文件放入表目录中  2、在hive的交互shell中，用hive的命令来导入数据 到表目录  load data local inpath '/opt/module/hadoop-2.7.2/hiveFile' into table t\_pv\_log  3、用hive命令导入hdfs中的数据 目录 |
| load data inpath '/opt/module/hadoop-2.7.2/hiveFile' into table t\_pv\_log |

## 4、查询

|  |
| --- |
| 同数据库 |

## 5、数据类型

|  |
| --- |
| 1、数字类型  TinyInt （1byte -128--127）  Smallint（2byte）  Int/integer（4byte）  BigInt （8byte）  Float（4byte）  Double（8byte）  2、时间  date  timestamp  3、字符串类型  string  varchar  char  4、布尔  boolean  5、复合类型  1）Array数组类型  create table tname(ID INT ,ACTORS ARRAYS<STRING>)  row format delimited fields terminated by ‘,’ #指定列分割符  collection items terminated by “:” #指定集合元素分割符  取数 select actors[0] from tname; 取数组的第一个元素  查数组中包含吴刚的 select \* from tname where array\_contains(actors ,”吴刚”)；  查数组个数 select size(actors) from tname;  2）map  Create table family(id int,name string ,family\_mem map<string,string>,age int)  row format delimited fields terminated by ‘,’ #指定列分割符  collection items terminated by “#” #指定集合元素分割符  map keys terminated by ‘:’ #指定map的key以什么结束  查map中的某一个key的傎 select id,name,family\_mem[“fater”] from t\_family; |
| 查mep的所有key select maps\_keys(family\_mem) as relations from …..  查map的所有value select map\_values(family\_men)  查map的数量 size（family\_men） |
| 查map 中key包含关键字的数据 array\_contains(map\_keys(family\_men),’brother’) |
| Value 也类同上一条  3）struct  映射 Info struct<age:int,sex:string,addr:string>  查询结构体的成员 select info.age from table |
|  |

# 3、函数

## 1、常用函数

|  |
| --- |
| 1、字符  1）substr（str,strart,len） 下标从1开始  2）substring(str,start)  3）length（str）  4) concat（str1,str2 ...） 拼接  5) split(str,spliter) 分割成一个数组  6）upper lower  2、cast （“5“ as int） case（”2017-03-03“ as date）  3、数学运算  1）round(number)  2) ceil floor abs greatest least |
| 4、时间  1）unix\_timestamp()  2)from unix\_timestamp();  3)to\_date() |

## 2、表生成函数

|  |
| --- |
| 1、explod () 行转列。将所有查出数据的第一列都当作一行  2、lateral view  select id, name,tmp.sub  from  t\_stu\_subject  lateral view  explode(subjects) tem as sub;  Lateral view explode(column) 将每行中的subjects中的每个元素与 每行拼接生成 多行数据。 |

## 3、集合函数

|  |
| --- |
| array.contains(array,val)  sort\_array(array)  size(collection)  map\_keys(map)  map\_values(map) |

## 4、条件控制函数

|  |
| --- |
| case when … then ...else end 同sql  if（condition，trueVal, falseVal） |

## 5、rownumber() over() 分组topn

|  |
| --- |
| 1、先分组，每组的前几个  select \* from  (select \* ,rownumber() over(partion by sex order by age desc) as rn from t\_table) tmp  where tmp.rn<3 |
|  |

## 6、窗口分析函数 sum() over（）

|  |
| --- |
| 1、需求  查询用户每月花费 及 至当月累计花费总额  2、分析  先查询按月分组 累计每月的的花费  再，查询结果按用户自关联，过滤条件过滤（左表.月份<右表.月份）  最后，按右表.月分分组求和。  3、使用窗口分析函数  先查询按月分组 累计每月的的花费  再，sum(每月的花费总额) over(group by 用户 order by 月份 between unbounded preceding and current row),可直接实现累计至当月的花费。  Over（）中的解释  Unbounded preceding : 向前无限界  Current row 到当前行 |

## 7、json解析函数

|  |
| --- |
| json\_tuple(json，k1,k2,k3) as (别名1，别名2，别名3) |

## 8、自定义函数

|  |
| --- |
| 1、定义json解析函数  2、流程  1）用java写一个程序，继承hive包中中的UDF类，重写evaluate(),实现想要的功能，  2）将java程序导入出jar 包 |
| 3）用hive命令行中的将jar包添加到classpath  add jar filePath  4）在hive命令中创建一个函数叫myjson，关联自己的java类:  Create temporary function myjson as ‘所写类的全路径名’ |

# 4、练习

|  |
| --- |
|  |
|  |
| 1、建表映射日志数据  2、建表保存日活数据  3、分别导入每天的数据  load data local inpath file\_path into table t\_name partition(day=201705) |
| 4、指标统计  日活： |