大数据开发套件使用介绍

背景：数据驱动产品，人人都是数据分析师。
目的：方便接触数据，后续数据提取及简单分析都可以各自独立完成。

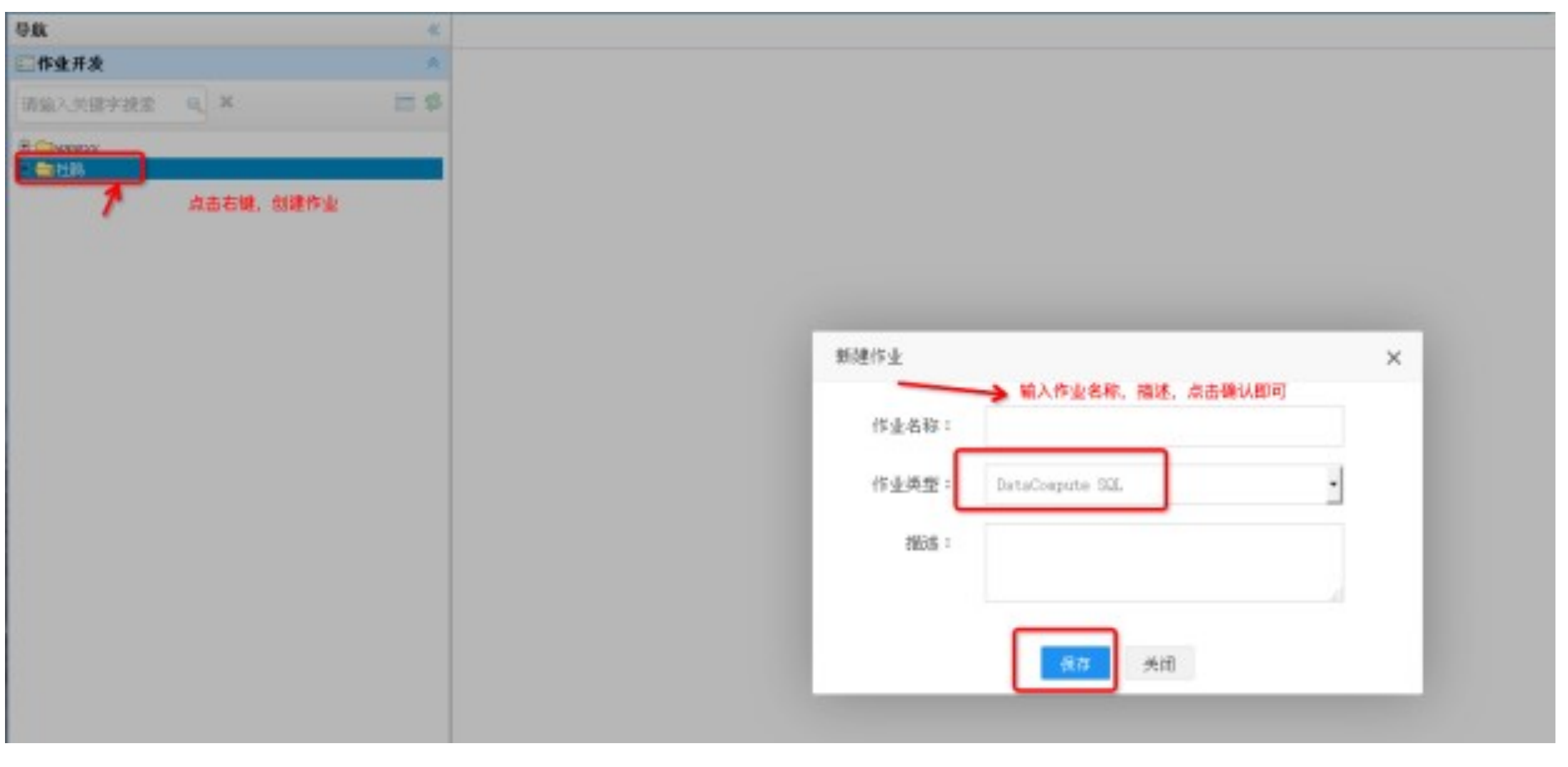
一、如何才能接触数据？需要的权限如下：

1、云桌面：找IT申请
2、大数据开发套件：http://datacompute.tongdun.cn:8088，只能通过云桌面访问，公司OA账号登录即可；
3、数据表：数据管理-->数据查找—>搜索需要的表，点击申请权限，由表管理人员审批同意即可。

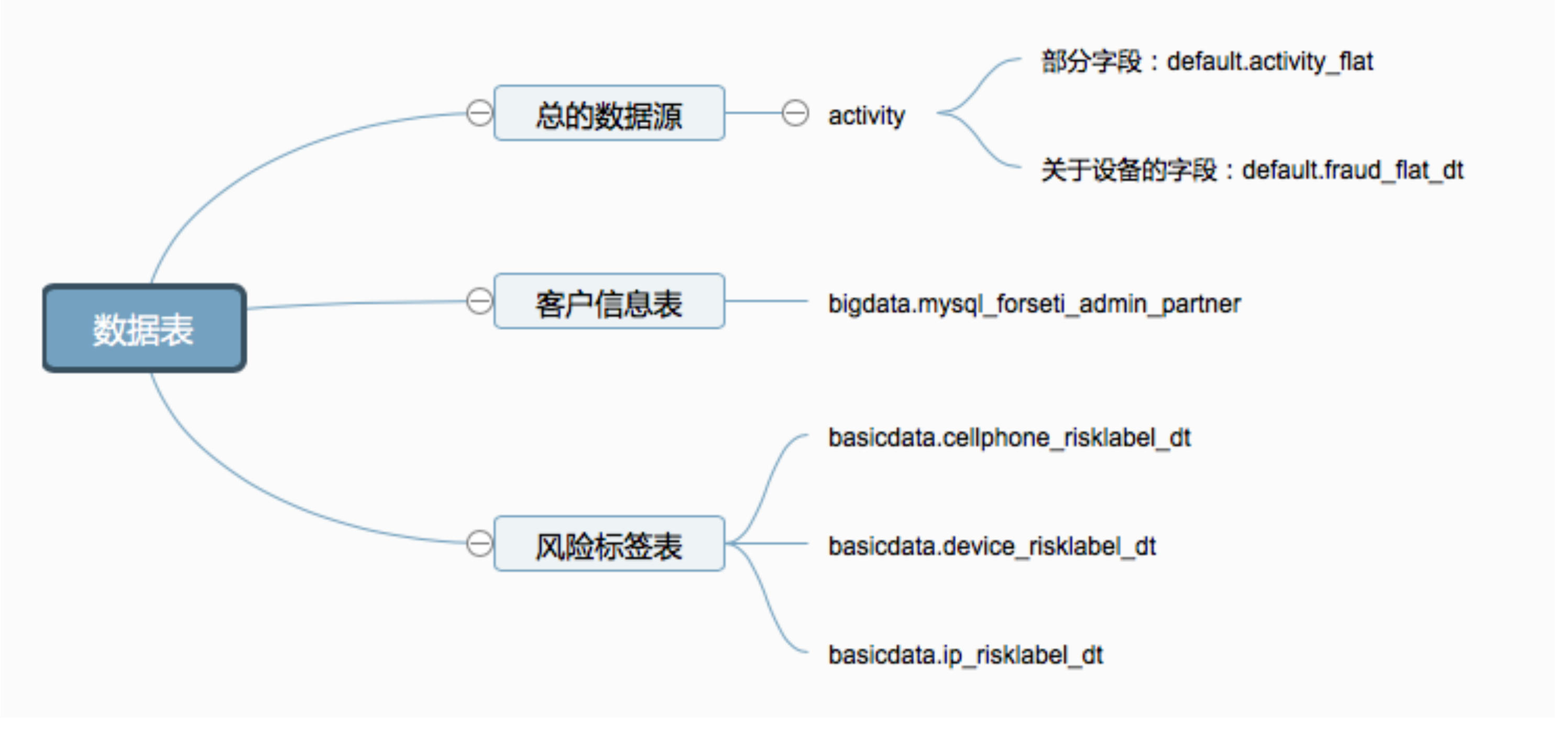
4、查看申请的表或者自己管理表：

二、开始取数：鉴于该平台只能写hive-sql，所以只要掌握sql就可以轻松实现取数。

1、创建取数窗口：数据开发—>作业开发-->新建个人目录—>点击右键，创建作业，选择DataCompute SQL—>开始写sql



2、数据表介绍：



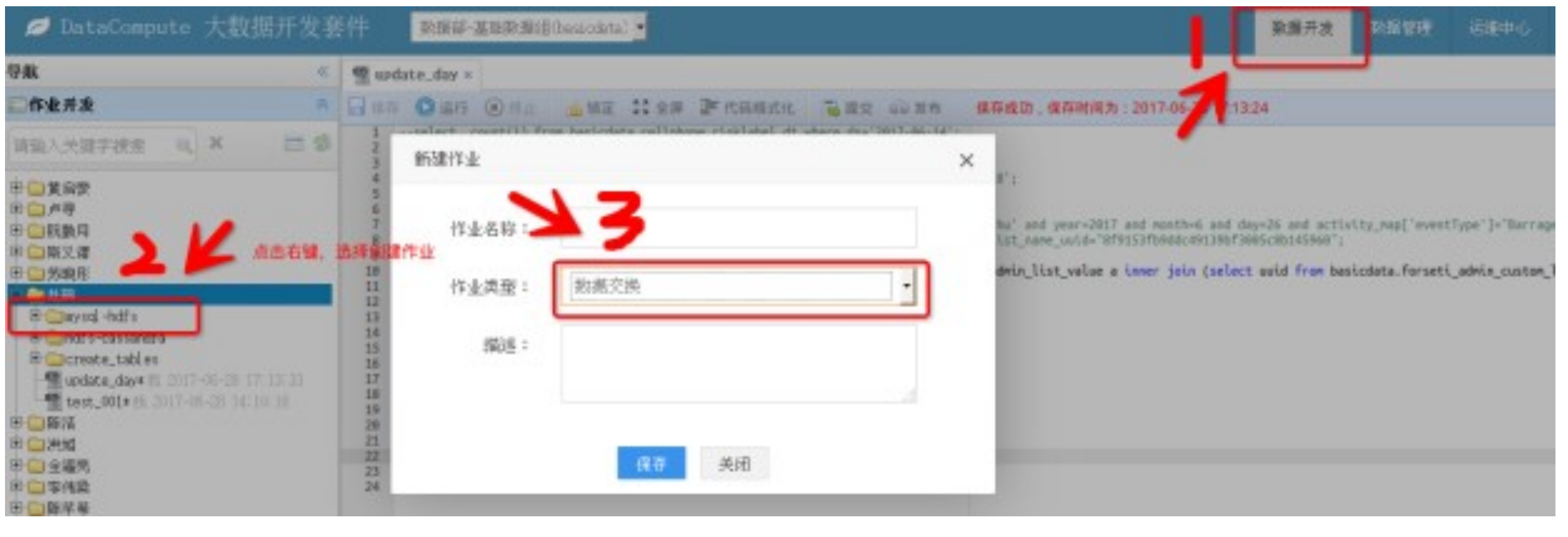
- default.activity:是调用规则引擎的详细表，里面的字段都以map的形式存放，常用的map有：
  - activity_map:客户基本信息，系统字段；
  - device_map:通过设备指纹获取的信息；
  - event_result_map:规则引擎结果信息
  - policy_map:规则详情
- default.activity_flat:基于activity，拉取常用的字段，且对部门关键字段进行了清洗，有手机号、身份证、银行卡等，格式校验不通过的字段会带有前de缀“TDERR”，该表能满足平常80%的需求，所以如果能用改表就尽量用该表，可以大大提升数据获取速度。
- default.fraud_flat_dt:是基于设备的信息的表，里面包含了activity中device_map全部的信息，且额外包含客户等常用的信息。
- bigdata.mysql_forseti_admin_partner:包含客户的基本信息，客户简称\全称、对应的策略、运营、销售人员等。其他的表中都只有partner_code。
- 三张风险标签表：字段，标签，出现次数，出现合作方数

3、sql基本操作：

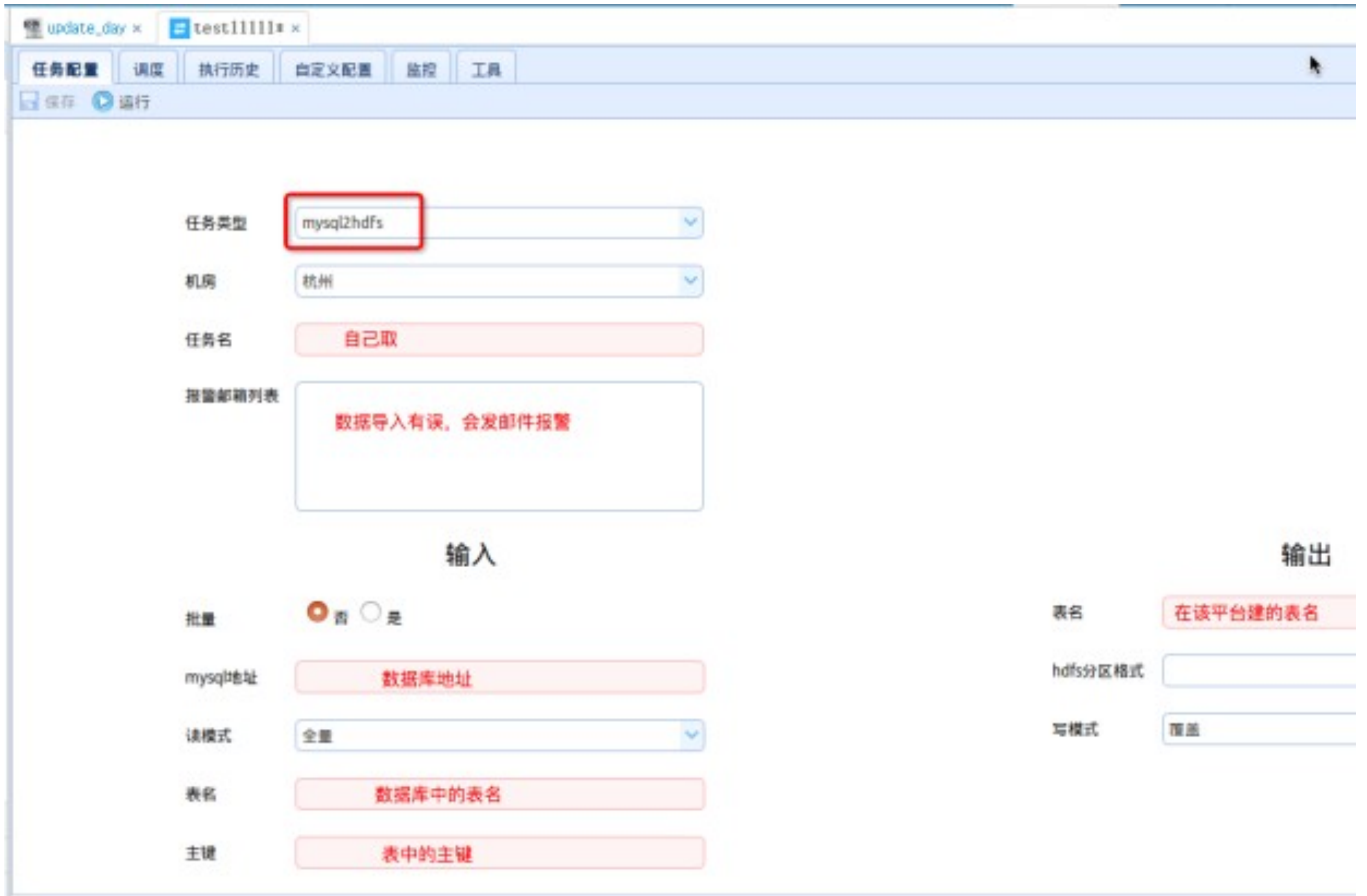
- 查询注意点：
  - 1、不能在where条件中嵌入子查询，必须用关联；
  - 2、提供分区字段，筛选时最好加上；
  - 3、每张表都需要加上改表所在的组，如default.activity_flat
  - 4、创建临时表必须以‘tdl’开头
- hive-sql DDL操作：https://gitlab.fraudmetrix.cn/binsong.li/datacompute-document/wikis/DataComputeSQL
- 查看表结构：**desc** default.activity_flat
- 创建临时表：**create** table **tdl**_s1 **lifecycle** 7 as **select** * **from** activity_flat **where** year=2017 and month=6 and day=1；
- 删除临时表：**drop** table tdl_s1;
- 查询/模糊查询/字段重命名/排序：select partnercode as partner_code from default.activity_flat where year=2017 and month=6 and day=1 and partnercode rlike ‘wa’ **order by** partner_code desc;
- 字段去重：select **distinct** partnercode from default.activity_flat ;select partnercode from default.activity_flat **group by** partnercode;
- 分组计算：select partnercode ,**count**(1) as cnt from default.activity_flat where year=2017 and month=6 and day=26 **group by** partnercode;
- 表关联：**left join /right join /inner join** :select partner_code from default.activity_flat a **left join** forseti_admin_partner b **on** a.partner_code =b.partner_code；
- 参考文档：https://help.aliyun.com/document_detail/34994.html?spm=5176.doc48975.6.615.VK2WdN

4、如果数据表在平台中没有，需要自己拉取，将mysql数据导入hdfs中：

- 创建一张和mysql表结构一样的空表；
- 创建数据交换任务：数据开发—>作业开发—>个人目录，右键创建作业—>数据交换



填写相关信息，保存—>运行—>执行历史中看执行结果。
具体见：http://wiki.tongdun.cn/pages/viewpage.action?pageId=14770522



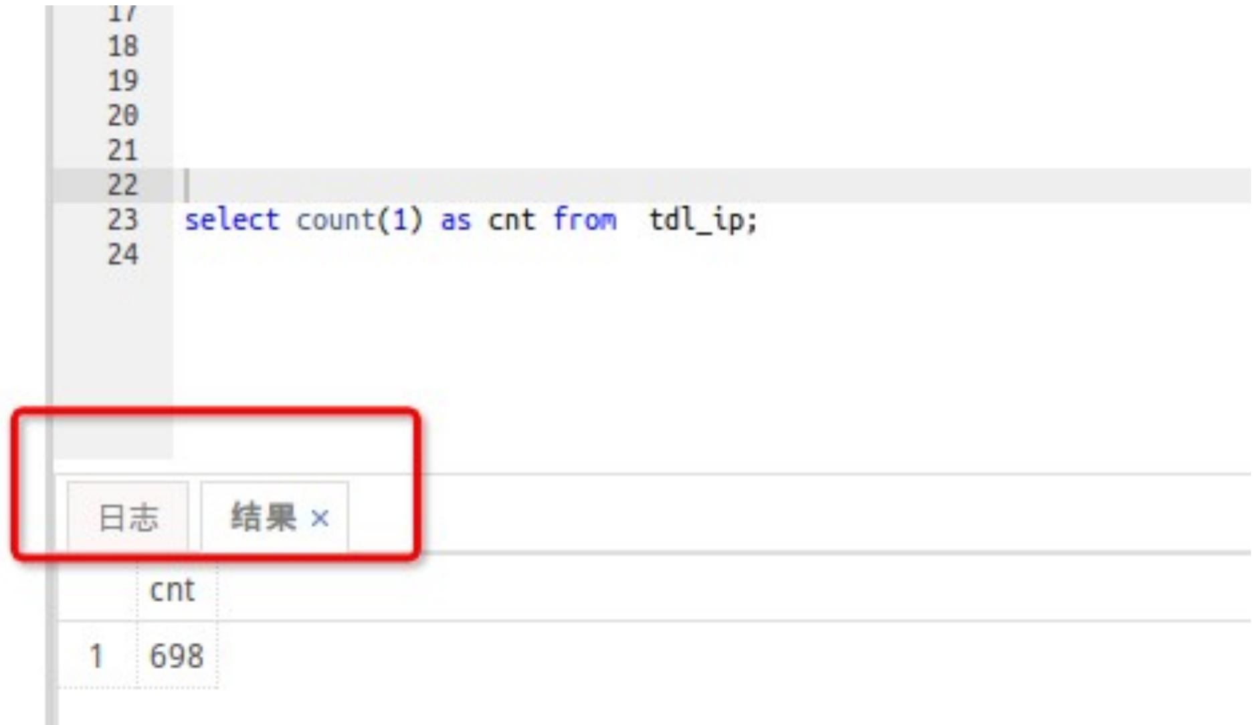
任务类型 mysql2hdfs
机房 杭州
任务名 自己取
报警邮箱列表 数据导入有误，会发邮件报警

输入
批量 ● 否 ○ 是
mysql地址 数据库地址
读模式 全量
表名 数据库中的表名
主键 表中的主键

输出
表名 在该平台建的表名
hdfs分区格式
写模式 覆盖

三、取结果：分下载和非下载

1、非下载：直接在sql窗口运行就可以得到数据结果；

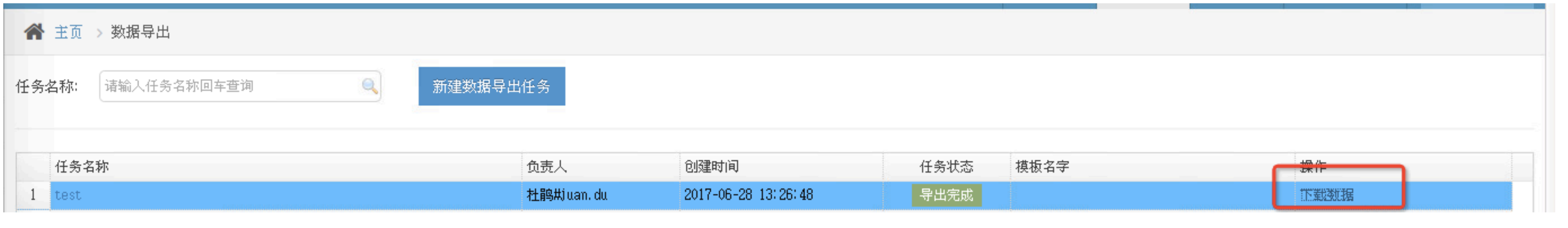


```
select count(1) as cnt from  tdl_ip;
```

```
   cnt
1  698
```

2、下载：数据导出：数据管理—>数据导出—>新建数据导出任务—>数据导出窗口，填写任务名称、sql、文件格式。

写入sql语句，点击执行，计算成功后，会有“下载数据”字段，可以点击下载到云桌面，如下图：



	任务名称	负责人	创建时间	任务状态	模板名字	操作
1	test	杜鹃juan.du	2017-06-28 13:26:48	导出完成		下载数据

四、sql实战分析：该部分也存放在大数据开发套件目录：反欺诈及基础风控组-杜鹃

eg1:

```
#取数字段：sqeId || cellLocation || mnc || mcc\n
#取数口径： (4.19) appType= Android，deviceMap中字段 cellLocation不为空，mnc不为空，mcc不为空，trueIp不为空，版本号(fmVersion >= 3.0.0 ),
locationOfCell为空的数据
```

code：

```
select activity_map['sequenceId'],device_map['cellLocation'],device_map['mnc'],device_map['mcc']
from activity
where lower(event_result_map['appType']='android' and device_map['cellLocation'] is not null
            and device_map['mnc'] is not null and device_map['mcc'] is not null and device_map['trueIp'] is not null
            and device_map['fmVersion']>='3.0.0' and device_map['locationOfCell'] is null
            and year=2017 and month=4 and day=19
```

eg2:

```
#58转转匹配风险标签
#2017.6.20-21号的手机号,是否命中风险标签.
```

code:

```
****step1:提取手机号*****
create table tdl_s1 lifecycle 7 as
select accountmobile
from activity_flat
where partnercode = "58ganji" and year = 2017 and month = 6 and day >= 21 and day<=22
group by accountmobile

****step2:关联风险标签*****
create table tdl_s2 lifecycle 7 as
select a.accountmobile,b.label
from tdl_s1 a left join basicdata.cellphone_risklabel_dt b on a.accountmobile = b.phone
where b.label is not null and ds = '2017-06-22'

****step3:取结果*****
##有标签的手机号数
select count(distinct accountmobile) as cnt from tdl_s2;
#每个标签的命中手机号数
select label,count(distinct accountmobile) as mob_cnt from tdl_s2 group by label;
```

eg3:

```
--IP列表：龙珠IP黑名单_弹幕1级，需要今天早上4点前的数据
--需要这些ip下所有产生过弹幕事件的账号信息，包括以下几个字段：
--账号名，账号昵称，账号等级，发送弹幕次数，弹幕拒绝次数，注册时间（查注册事件），注册手机号（查注册事件）
```

code:

```
****step1:取龙珠IP黑名单_弹幕1级中对应的IP*****
create table tdl_ip lifecycle 7 as
select data_value as ip
from basicdata.forseti_admin_list_value a
inner join
    (select uuid from basicdata.forseti_admin_custom_list where id =71691) b
on a.fk_list_name_uuid=b.uuid;

****step2:取ip对应的账号信息*****
select accountLogin,accountName,ext_user_level,sum(msgContent) as barrage_cnt,sum(reject) as rej_cnt
from
    (select activity_map['accountLogin']
            ,activity_map['accountName']
            ,activity_map['ext_user_level']
            ,activity_map['ipAddress']
            ,case when activity_map['msgContent'] is not null then 1 else 0 end msgContent
            ,case when activity_map['riskStatus']=='Reject' then 1 else 0 end as reject
            from activity
            where activity_map['partnerCode']='longzhu' and activity_map['eventType']='Barrage' and year=2017 and month=6 and
day>=21 and day<=22) a
inner join
    tdl_ip b
on a.ipAddress=b.ip
group by accountLogin,accountName,ext_user_level
```

owner:杜鹃
date:2017.6.28