# MR栅格化帮助文档

目录

1. 总体要求
2. 操作过程
3. 要求拿到客户给的mro数据与指纹库数据进行栅格化操作，分别输出mro栅格化结果（按小时分成文件）和 分析结果的.csv 文件。
4. 操作过程
5. 将mro数据从硬盘拷贝至linux中
6. 创建hive外部表按天和小时分区，最好每个地区一个库。例：

Create database shanghai;

Use shanghai;

CREATE EXTERNAL TABLE `mro`(

`objectid` bigint,

`mmeues1apid` string,

`time\_stamp` string,

`ncellobjectid` bigint,

`ltescrsrp` int,

`ltescrsrq` int,

`ltescsinrul` double,

`ltencrsrp` int,

`ta` int,

`aoa` int,

`LteScPUSCHPRBNum` int,

`LteScPDSCHPRBNum` int,

`ulqci1` double,

`dlqci1` double)

PARTITIONED BY (

`day` string,

`hour` string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

1. 将数据上传到hdfs上，并location到hive表的小时分区中。

例: 比如数据已经上传到linux /home/hadoop/data/shanghai/20180301

在hdfs 上创建分区文件夹

hadoop fs -mkdir /data/shanghai/20180301/00

hadoop fs -mkdir /data/shanghai/20180301/01

以此类推数据有多少小时就创建多少个文件夹。将linux上的数据上传到hdfs 上面所创建的目录中。

例: cd /home/hadoop/data/shanghai/20180301 进入目录

将每个小时的数据分别上传到hdfs分区文件夹中。

例：hadoop fs -put ./2018030100\* /data/shanghai/20180301/00 (具体看文件名而定)

4.数据上传完之后给hive添加上分区。

alter table mro add partition(day='20180301',hour='00') location '/data/shanghai/20170301/00';

以此类推 ...

1. 进hive 里查询mro表的分区表是否添加成功 hive表中是否有数据：

Select \* from mro where day=’20180301’ and hour=’00’ limit 10;

1. mro 数据操作都做好之后，开始生成指纹库操作。修改代码，首先修改表config.properties ，修改各个表的表名和库名等全局变量，指纹库是winprop生成的，确定数据库中的库名是什么，表名是什么，修改之后，跑gengerateTable这个类将sqlserver数据库中的表跑成一层和高层的两个指纹库到hive中。
2. 接下来跑mro和指纹库的分区表：根据mro判断是否生成解决数据倾斜的分区表，例如mro数据量很大就不用生成数据倾斜的分区表了。因为生成表也会占用时间。如果mro数据量非常大 生成一次mro分区表和指纹库分区表就行了。同样是跑gengerateTable这个类，就是再生成指纹库之后同时生成mro和指纹库的分区表。
3. 表都生成完之后，接下来修改脚本里的时间，找到脚本存放位置，修改脚本里跑数据所用的日期

例：将运行日期从20170602 改成20180301

sed -i "s/20170602/20180301/g" `grep 20170602 -rl ./脚本路径` 这样就把所有脚本的时间都改了

1. 进入azkaban 设置时间调度任务，根据数据量的大小设置任务调度的间隔时间。

Azkaban地址https:10.254.188.242:8443 任务已经制作好，只需要点击每个小时的job任务时间调度schedule 即可。注意时间间隔。

1. 栅格化跑完之后需要合并一次小文件，由于为了增快栅格化的速度，栅格化之前有很多公共字段是没有带上的，栅格化后需要和原表进行join，将字段获取回来以便后面的计算。

合并小文件和计算指标都在threed\_mrlocate\_gridavg中，可以一个方法一个方法的跑，这样容易控制。

1. 将计算指标的结果resultMr表输出成csv倒到本地硬盘中。或者根据要求输出到sqlserver中去。
2. 栅格化结果是没有带上经纬度的，如果需要栅格化结果需要重新加上经纬度并按照小时分区
3. 在hive输出csv时，

set hive.exec.compress.output=false 设置输出不压缩

insert overwrite local directory '/home/datangdev/out/test'

row format delimited

fields terminated by ','

select ltescrsrp,longitude,latitude from mdt distribute by 1