## 2019-11-29

* 起因：在新项目“国家治理”中，需要取用新矛盾、高质量、交通爬虫、民心民情（舆情）多系统的指标数据；发现了gov\_id对不齐的情况。

具体如：

①【交通爬虫】中，有gov\_id=3230（山西省|大同市|云冈区）的数据，而这个gov\_id在我本地的df\_2861\_gaode\_geo\_new.csv中是缺失的；

②而我本地df\_2861\_gaode\_geo\_new.csv中，存在gov\_id=232（山西省|大同市|矿区）和gov\_id=233（山西省|大同市|南郊区），在【交通爬虫】中也是缺失的；

③产品端的网页版地图边界及行政区划配置，也与我本地有所不同，是否和【交通爬虫】的版本相同，也未可知。

* 为了保证数据的准确性，爬虫、前端、后端都用同一份行政区划配置，是最理想的情况，否则便会影响最终评分；

现在产生了差异，就需要明白为什么？边界在哪儿？能否调整？如何调整？

* 基于上述现象，展开数据追查，过程各种纠结、牵扯众多，好在结论较明晰。
* 基本结论：

① 存在多个版本的行政区划配置文件，n>=3；

所有配置文件都来自于同一个数据库：

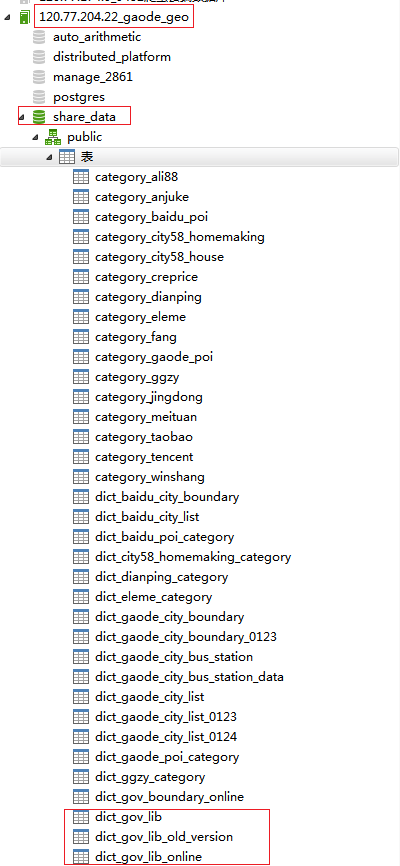
连接信息：

Host = '120.77.204.22'

Port = '5999'

User = 'readonly'

Pwd = '123456'



② 如何理解上图中的三张配置表呢？

1) dict\_gov\_lib\_online

这张表是根据高德返回的地域接口实时变化的，有可能一天一变（如果行政区划有改动的话）。【爬虫端】主要用这张表。

不能让前端和后端直接对齐这张表的原因：

前端需要地图边界数据（**以gov\_id.json的形式存储在服务器上，取来渲染**——边界数据为何不能实时从数据库获取呢？最初设计问题，轮不到我关心。。。），因为是**静态文件**，所以没办法实时对齐；

后端的数据需要和前端地图保持一致，如果后端对齐了这张表，而前端行政区划没对齐，会出现**“已撤销”**的区域，仍在地图上，点击后却无数据，严重或会导致程序报错；而后端**“新增”**的区域，前端不取，则问题不大。

1. dict\_gov\_lib\_old\_version

XMD产品主要用表。相当于1)的某个切片数据，是较老的版本，和目前我本地的df\_2861\_gaode\_geo\_new.csv大致相同（说大致是因为，后面又在本地文件，临时补了一两个区域）

1. dict\_gov\_lib

高质量发展产品主要用表（郑哥主要用表，交通的数据也对齐这张表的配置）。相当于1)的某个切片，数据比1)更新。

但某些gov\_type的划分，和1)中有明显差别。可关注3)中 gov\_type=12的高新区，在1)中都归属于gov\_type=6.

至于差异的原因，可能是郑哥有自己的独特设计和用处吧。。。

民情相关网页版产品错乱的情况：

* **前端的边界静态文件**，由兴林哥生成，当时他取的是dict\_gov\_lib\_online，是他操作时，最新的行政区划的边界数据；（后面时间推移，也不一定和dict\_gov\_lib\_online一致了）
* **后端舆情（微博）**相关板块的数据：

都以df\_2861\_gaode\_geo\_new.csv为准，也就是dict\_gov\_lib\_old\_version的初步衍生版

（衍生的情况，包括在文件中手动添加“南京市|江北新区”，以及把下辖无区县的几个市，gov\_type由1改为31，如“甘肃省|嘉峪关市”、“海南省|儋州市”……）

zk\_social是ES上的微博库，核实发现，微博爬虫端，也用的df\_2861\_gaode\_geo\_new.csv，没有根据高德的变化，新增行政区划。

* 由于【国家治理】项目需要用到别的系统的数据，如业态数据等；

核实发现，**高质量发展（政府绩效指数）**：dict\_gov\_lib

**高德爬虫（交通、路网、业态）**：dict\_gov\_lib\_online

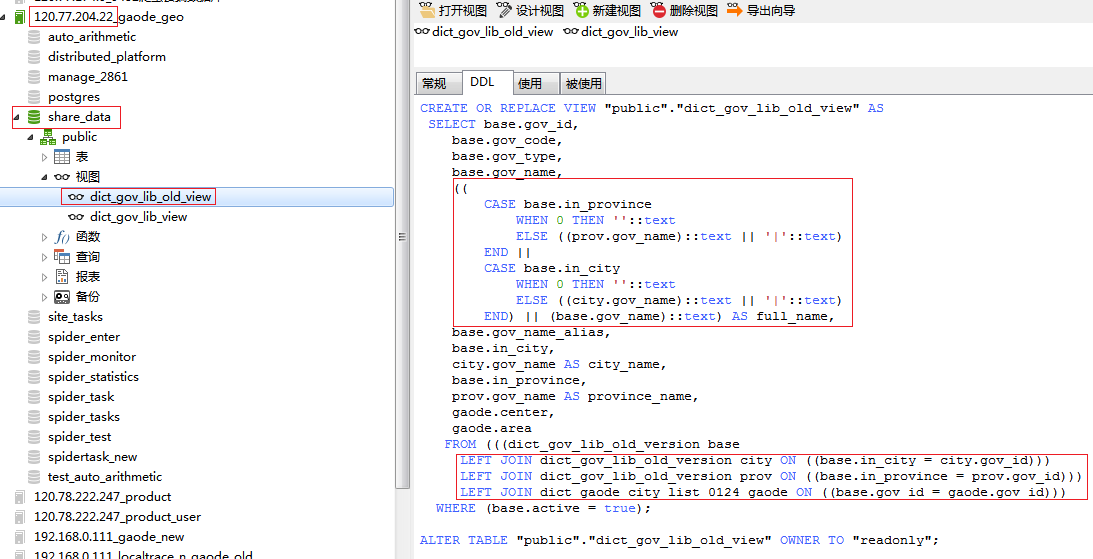
**李科人口**：dict\_gov\_lib\_online

**舆情数据**：dict\_gov\_lib\_old\_version

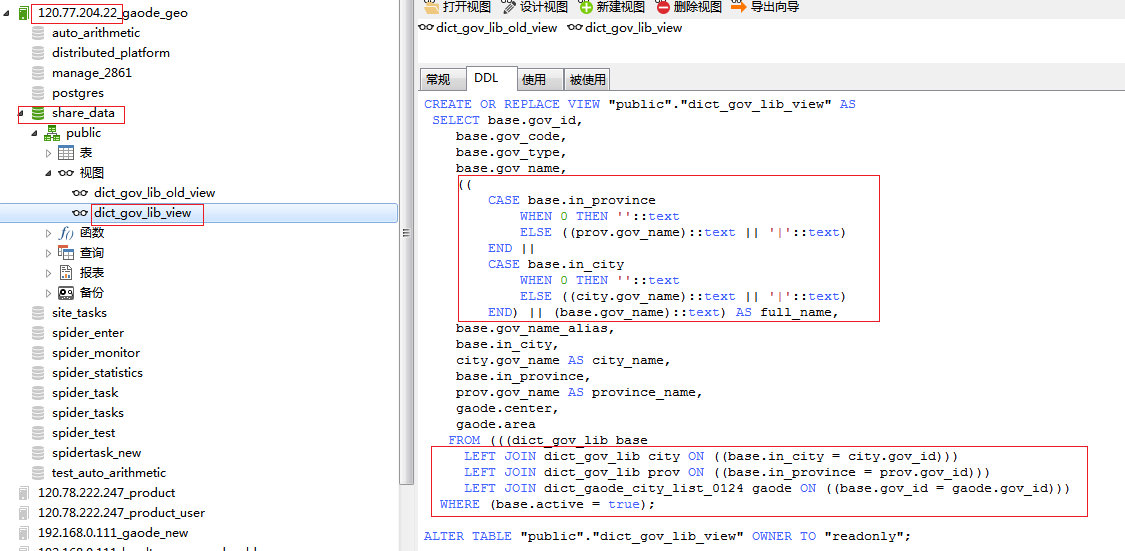
IMG_256

③ 高质量发展指数（gov\_index）的gov\_utils中，用的视图取配置（基本上是基于dict\_gov\_lib表做的视图）——【视图】主要是为了拼接full\_name，因为上图中的三张表中，没有full\_name（“北京市|东城区”）字段，仅有gov\_name（“东城区”）

- Old\_version: 来源表**dict\_gov\_lib\_old\_version**



- 来源表**dic\_gov\_lib**



④ 现在要糅合多系统、基于不同配置文件的数据，怎么办？

1) 最省事：还是以本地文件df\_2861\_gaode\_geo\_new.csv为准，其他指标，多的gov就清洗掉；本地中，已撤销的gov，和前端地图对不上的，如有报错，在代码中打补丁跳过。

2) 最科学：根据当前最新的dict\_gov\_lib\_online，手动修正一版df\_2861\_gaode\_geo\_new.csv，调整新增，去掉撤销；新增区划，（舆情）数据缺失的，就中位值补齐。后续再提醒【微博爬虫】，新增区域爬取。 √

~~3) 最保险：根据当前最新的，dict\_gov\_lib\_online，手动修正一版df\_2861\_gaode\_geo\_new.csv，调整新增，不去掉撤销；保证最全量的数据。~~

（不仅有撤销的情况，还有调整行政区划归属的情况）

⑤ 本地文件df\_2861\_gaode\_geo\_new.csv 与 最新的dict\_gov\_lib\_online差异比对：

- 新增的行政单位\_20191129



- 已撤销的行政单位\_20191129

