甲方名称：

多家科研机构、咨询公司、网络公司

甲方需求：

- 采集大众点评网全国所有商户数据；
- 每个商户需要提取如下信息：

  “shop_id"(商户ID，唯一、固定不变)

  “verified"（是否加V）

  "is_closed"（是否停业）

  "name"（商户名称）

  "alias"（别名）

  "province"（省）

  "real_city"（真实所属城市）

  "city"（市）

  "city_pinyin"（市拼音）

  "city_id"

  "area"（区）

  "big_cate"（一级分类）

  "big_cate_id"

  "small_cate"（二级分类）

  "small_cate_id"

  "address"（地址）

  "business_area"（商圈）

  "phone"（电话）

  "hours"（营业时间）

  "avg_price"（均价）

  "stars"（星级）

  "photos"（图片）

  "description"（描述）

  "tags"（标签）

  "latitude"（纬度, 腾讯地图）

  "longitude"（经度, 腾讯地图）

  "navigation"（导航栏、面包屑内容）

  "characteristics"（特色）

  "product_rating"（口味/产品评分）

  "environment_rating"（环境评分）

  "service_rating"（服务评分）

  "all_remarks"（全部点评数）

  “default_remarks"（默认点评数）

  "very_good_remarks"（5星数）

  "good_remarks"（4星数）

  "common_remarks"（3星数）

  "bad_remarks"（2星数）

  "very_bad_remarks"（1星数）

  "recommended_dishes"（推荐菜）

"recommended_products"（推荐产品）

"parking"（停车位信息）

"nearby_shops"（周边商户）

"is_chains"（是否是连锁店）

"take-away"（是否外卖）

"group"（团购信息）

"card"（会员卡）

"latest_comment_date"（最新评论日期）

"history"（历史时间点：加入时间、更新时间、最后评论时间三类信息分号分隔）

- 能够处理各种不同详情页模板（大众点评不同分类下的商户模板不尽相同）。
- 要保证数据完整性，最终只交付数据。

技术难点：

- 大众点评封IP厉害，访问过快会出现验证码；
- 商户列表页有最多50页可见限制；
- 数据量非常大，下载量超过5000万；
- 详情页有多种模板要处理；

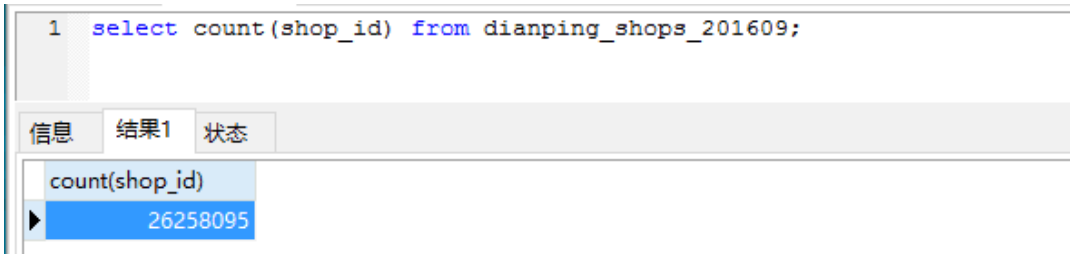
实现方案：

- 使用大量稳定高匿HTTP代理轮换采集；
- 进入各级子类、各级子区域、商圈等列表进行采集，集合遍历店铺ID采集；
- 通过分布式集群采集；
- 分别处理每种模板下的字段提取；

shop_id	verified	is_cl	name	alias	provin	real_c	city	city_pin	city_	area	big_c	big_s	small_cat	smal	address	business_ar
19413782	True	Fals	内蒙羊蝎子火锅城		海南	三亚	三亚	sanya	345	金鸡岭	美食	10	东北菜	g106	金鸡岭路中级人民法院斜对面	
5390342	False	Fals	春光海南特产专营店(解放路步		海南	三亚	三亚	sanya	345	解放路	购物	20	食品茶酒	g184	解放路步行街	
1999912	True	Fals	粥公粥婆(河西店)		海南	三亚	三亚	sanya	345	河西路	美食	10	小吃快餐	g112	河西路美和家园B幢1楼(近佳	
21965218	True	Fals	清芯重庆原味火锅		海南	三亚	三亚	sanya	345	三亚湾	美食	10	火锅	g110	三亚湾路通海四巷02号(海边	
8858815	True	Fals	宽巷子川菜馆		海南	三亚	三亚	sanya	345	海棠湾	美食	10	川菜	g102	后海西村101号(蜈支洲岛大门	
22650690	True	Fals	湘村厨娘		海南	三亚	三亚	sanya	345	三亚湾	美食	10	湘菜	g104	三亚湾路168号海坡村南门万	
22210663	True	Fals	饺子人家		海南	三亚	三亚	sanya	345	三亚湾	美食	10	东北菜	g106	三亚湾外贸路美心公寓105号	
3452236	True	Fals	一块豆腐(迎宾路总店)		海南	三亚	三亚	sanya	345	三亚市区	美食	10	东北菜	g106	迎宾路147号(近中法供水)	
11552994	True	Fals	三湘人家(三亚解放路店)		海南	三亚	三亚	sanya	345	国际购物	美食	10	湘菜	g104	解放四路176号东方海景大酒	
19432682	True	Fals	蜜恋甜品		海南	三亚	三亚	sanya	345	河西区	美食	10	面包甜点	g117	外贸路5号楼11号铺面	
17209339	False	Fals	解放四路旺亳超市		海南	三亚	三亚	sanya	345	河西区	购物	20	超市/便利	g187	解放四路与金鸡岭路交汇处(	
5526112	True	Fals	澳门豆捞(三亚店)		海南	三亚	三亚	sanya	345	大东海区	美食	10	火锅	g110	大东海榆亚大道林达海景酒店	
18522552	True	Fals	正宗湖南土菜馆		海南	三亚	三亚	sanya	345	河东区	美食	10	湘菜	g104	詹洲铁路西巷十二号(益杨宾	
5564018	True	Fals	海豚美式餐厅酒吧		海南	三亚	三亚	sanya	345	大东海区	美食	10	西餐	g116	榆亚路99-8号(近城市酒店)	
557747	True	Fals	亚龙湾三亚喜来登度假酒店	瑶池	海南	三亚	三亚	sanya	345	亚龙湾	美食	10	西餐	g116	亚龙湾三亚喜来登度假酒店(	
3033985	False	Fals	阿牛椰香香椰奶清补凉(旗舰店		海南	三亚	三亚	sanya	345	第一市场	美食	10	面包甜点	g117	新建街140号(第一市场南门对	
5623367	True	Fals	美高梅度假酒店-钓鱼台锦汇餐		海南	三亚	三亚	sanya	345	亚龙湾	美食	10	粤菜	g103	亚龙湾中段美高梅度假酒店内	
3538105	False	Fals	鸿港批发市场		海南	三亚	三亚	sanya	345	河西区	购物	20	更多购物	g131	胜利路(鸿港码头)	
19573898	True	Fals	三亚免税店	海棠	海南	三亚	三亚	sanya	345	海棠湾	购物	20	更多购物	g131	海棠北路118号(天房洲际)	
3891574	True	Fals	京润珍珠明珠广场专柜		海南	三亚	三亚	sanya	345	解放路	购物	20	珠宝饰品	g122	解放路668号明珠购物广场1	
4120494	False	Fals	百佳汇超市		海南	三亚	三亚	sanya	345	河西路	购物	20	超市/便利	g187	西河西路(近第一市场)	
3705499	False	Fals	南国超市		海南	三亚	三亚	sanya	345	三亚湾	购物	20	超市/便利	g187	旅游区迎宾路海月广场对面	
2546122	False	Fals	南国佳品超市(夏日百货店)		海南	三亚	三亚	sanya	345	大东海区	购物	20	超市/便利	g187	海鹤路1号夏日百货1楼(近榆	
6376720	True	Fals	喜乐购超市【新天地美食购物		海南	三亚	三亚	sanya	345	亚龙湾	购物	20	超市/便利	g187	新天地美食娱乐购物广场(近	
1597603	False	Fals	鸿港市场5号小郑水果摊		海南	三亚	三亚	sanya	345	国际购物	购物	20	水果生鲜	g271	胜利路鸿港综合批发市场5号	
6232501	False	Fals	购书中心		海南	三亚	三亚	sanya	345	解放路	购物	20	书店	g127	解放路208号	
2250802	False	Fals	旺豪超市(解放二路店)		海南	三亚	三亚	sanya	345	解放路	购物	20	超市/便利	g187	解放二路69号(信兴电器对面	
19138160	True	Fals	胜利购物广场旺豪超市		海南	三亚	三亚	sanya	345	河西路	购物	20	综合商场	g119	胜利路与新风街交叉口胜利购	
5509523	True	Fals	环球城大酒店		海南	三亚	三亚	sanya	345	亚龙湾	购物	20	综合商场	g119	亚龙湾国家旅游度假区	
4220751	True	Fals	福乐多超市(红旗路口店)		海南	三亚	三亚	sanya	345	国际购物	购物	20	超市/便利	g187	红旗路	
18029847	False	Fals	升升超市		海南	三亚	三亚	sanya	345	三亚湾	购物	20	超市/便利	g187	三亚海虹路鲁能地产一二楼(	
17681168	False	Fals	世纪华联超市		海南	三亚	三亚	sanya	345	亚龙湾	购物	20	超市/便利	g187	亚龙湾路百花谷商业街内	
19126534	False	Fals	冰雪大世界(龙坡村店)		海南	三亚	三亚	sanya	345	三亚旅游	购物	20	更多购物	g131	龙坡村美丽三亚(印象冰雕展)	
16975191	True	Fals	博视眼镜(博视眼镜国际购物中		海南	三亚	三亚	sanya	345	国际购物	购物	20	眼镜店	g128	解放一路1号国际购物中心1楼	

示例数据截图

最终采集店铺（商户）数超过2600万家（如下图所示）。



数据量MySQL查询截图