1、Spark SQL 汇总数据集

1） dmg_hzl_panshi_jiucuo_log（面积纠错）：统计面积纠错的房屋信息，主要包含审核前后的房屋面积及等级、经纪人提交的面积，分别对应字段 before_area、after_area、before_level、after_level，area 为经纪人提交的面积。

	create_at	created_dt	city_id	comm_id	house_id	base_house_id	state	jiucuo_type	before_area	before_level	before_sourcefrom	after_area	after_level	reviewsourcefrom	area
2	2023-09-04 19:05:21.0	2023-09-04	192	100587864	102660213	97329789	审核通过	先发后审-面积纠错	99.0	A	42	53.46	A	20	53.46
3	2024-06-05 10:26:28.0	2024-06-05	105	100434172	227001860	127011589	审核通过	先发后审-面积纠错	36.0	A	42	38.69	A	21	38.69
4	2024-03-25 10:37:45.0	2024-03-25	120	100505906	229232742	128266199	审核通过	先发后审-面积纠错	113.96	A	40	130.59	A	21	130
5	2024-03-31 17:07:28.0	2024-03-31	110	100449156	8107095	5033763	审核通过	先发后审-面积纠错	97.61	A	40	89.0	A	21	89
6	2023-09-27 11:32:24.0	2023-09-27	116	100393729	288666368	156348492	审核通过	先发后审-面积纠错	150.0	A	18	118.0	A	21	118
7	2023-12-14 16:25:52.0	2023-12-14	185	100507847	99679096	95568466	审核通过	先发后审-面积纠错	100.0	A	18	91.0	A	21	91
8	2024-02-20 14:13:55.0	2024-02-20	100	100403728	200894675	104495607	审核失败	先审后发-面积纠错	null	null	null	null	null	null	317
9	2023-10-11 10:01:58.0	2023-10-11	129	100472354	51058834	22166311	审核通过	先发后审-面积纠错	100.68	A	40	100.68	A	21	89
10	2023-09-03 15:46:42.0	2023-09-03	106	100421018	42118603	49142484	审核失败	先发后审-面积纠错	252.67	A	40	null	null	null	93.15
11	2023-08-26 16:42:08.0	2023-08-26	214	100650209	519246014	339613220	审核失败	先发后审-面积纠错	86.0	A	18	null	null	null	83.8
12	2023-09-01 18:20:57.0	2023-09-01	109	100416805	4469362	6032210	审核通过	先发后审-面积纠错	87.25	A	40	71.77	A	20	71.77
13	2024-04-11 10:40:34.0	2024-04-11	245	100632391	293990564	215286026	审核通过	先发后审-面积纠错	null	NONE	null	69.0	A	21	69
14	2023-12-14 12:54:38.0	2023-12-14	285	102063761	511503007	323926849	审核通过	先发后审-面积纠错	null	NONE	null	110.0	A	20	110
15	2024-01-30 12:23:24.0	2024-01-30	219	100619741	458467465	261108685	审核失败	先发后审-面积纠错	118.0	A	41	null	null	null	111.17
16	2024-03-10 13:14:34.0	2024-03-10	114	102032633	332242263	232126464	审核通过	先发后审-面积纠错	115.75	A	40	94.4	A	20	94.4

2） dmg_hzl_panshi_jiucuo_source：统计需要面积纠错的房屋（等级为 A 的？）

```
---纠错前来源经纪人/司
set kyuubi.server.mysql.parse.enabled=false;
drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_jiucuo_source;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_jiucuo_source
as
select a.*,b.sourcefrom,b.area as area_source,b.extrano,b.extradescribe,b.count,b.score,
case when b.sourcefrom in (1,2,4,11,24) and coalesce(b.extrano,'')<>'' then concat('c_',b.extrano)
    when b.sourcefrom in (18) and coalesce(b.extrano,'')<>'' then concat('b_',b.extrano)
    else '' end as extrano_new
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_jiucuo_log a
join hdp_anjuke_dw_db.dw_panshi_house_area_source_log b
on a.created_dt=date_add(b.cal_dt,1) and a.base_house_id=b.house_id
where a.before_level='A' and abs(a.before_area-b.area)<=1
and a.state='审核通过' and abs(after_area-before_area)>1/*有部分房号纠错前后面积差异在1平米以内*/
and b.cal_dt>=date_sub(${dealDate},320) and b.cal_dt<=${dealDate} and b.sourcefrom in(1,2,3,4,12,18,19,24,41,42,54)
;
```

	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	house_id	base_house_id	state	jiucuo_type	before_area	before_level	before_sourcefrom	after_area	after_level	reviewsourcefrom	area	sourcefrom	area_source	extrano
	100381207	96527770	审核通过	先发后审-面积纠错	155.0	A	18	160.23	A	20	160.23	41	155.0	
	239791216	138287471	审核通过	先发后审-面积纠错	86.0	A	18	75.22	A	21	75.22	41	86.0	
	524888353	342754343	审核通过	先发后审-面积纠错	89.0	A	18	75.0	A	21	75	18	89.0	201630570
	505999178	325725236	审核通过	先发后审-面积纠错	92.0	A	41	184.0	A	21	184	41	92.0	
	461126845	263387604	审核通过	先发后审-面积纠错	110.0	A	18	65.36	A	20	65.36	18	110.0	202337805
	28118221	36261774	审核通过	先发后审-面积纠错	47.0	A	41	70.0	A	21	70	41	47.0	
	215863680	113512513	审核通过	先发后审-面积纠错	134.0	A	42	143.93	A	21	143.93	18	134.0	285808
	29056975	37123545	审核通过	先发后审-面积纠错	89.0	A	18	87.35	A	20	87.35	1	90.0	252425
	505999178	325725236	审核通过	先发后审-面积纠错	92.0	A	41	184.0	A	21	184	18	92.0	6359741

3） dmg_hzl_panshi_house_level_a：统计 A 类面积的房屋信息

	A	B	C	D	E	F
1	sourcefrom	extrano	extradescribe	house_id	count	extrano_new
2	41			95982096	1	
3	41			163433521	1	
4	18	2766163	陈磊	251056445	3	b_2766163
5	4	yijingshanghaifangdi_d86_saas2_0b981	壹京（上海）房地产经纪有限公司	271851369	1	c_yijingshanghaifangdi_d86_saas2_0b981
6	1	7984	安徽宇州房地产营销策划有限公司	31544209	1	c_7984
7	1	17465	天津市津房置换有限责任公司	54070659	1	c_17465
8	41			126589687	1	
9	1	224840	成都雅合房地产经纪有限公司	104371023	1	c_224840
10	4	chengdouhehezhiyuanf_c20_saas2_1bf2a	成都合和致远房地产经纪有限公司	254094705	1	c_chengdouhehezhiyuanf_c20_saas2_1bf2a
11	4	damaifangdichan_24e_saas2_377d8f3bfc	大麦房地产	111952983	1	c_damaifangdichan_24e_saas2_377d8f3bfc
12	18	7198186	账号已注销	139717308	2	b_7198186
13	18	6838935	万炎炎	87120604	12	b_6838935

```
---经纪人/司当前构建的A类面积房号量
drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_house_level_a;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_house_level_a
as
select b.sourcefrom,b.extrano,b.extradescribe,b.house_id,b.count,
case when b.sourcefrom in (1,2,4,11,24) and coalesce(b.extrano,'')<>'' then concat('c_',b.extrano)
    when b.sourcefrom in (18) and coalesce(b.extrano,'')<>'' then concat('b_',b.extrano)
    else '' end as extrano_new
from hdp_anjuke_dw_db.dw_panshi_area_status a
join hdp_anjuke_dw_db_temp.dw_panshi_house_area_source_log b
on a.id=b.house_id
where a.cal_dt=${dealDate}
and a.status=1 and a.house_area_level=10 and abs(house_area_v2-area)<=1 and b.sourcefrom in(1,2,3,4,12,18,19,24,41,42,54)
;
```

dmg_hzl_panshi_house_level_a_sts:A 类面积房屋数量
dmg_hzl_panshi_jiucuo_source_sts:统计面积来源房屋数量

```
drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_jiucuo_source_sts;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_jiucuo_source_sts as
select sourcefrom,extrano_new,count(distinct house_id) as fang_cnt,sum(count) as total_cnt
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_jiucuo_source
group by sourcefrom,extrano_new
grouping sets(
    (sourcefrom),
    (extrano_new)
)
;


drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_house_level_a_sts;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_house_level_a_sts
as
select sourcefrom,extrano_new,count(distinct house_id) as fang_cnt,sum(count) as total_cnt
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_house_level_a
group by sourcefrom,extrano_new
grouping sets(
    (sourcefrom),
    (extrano_new)
)
```

3）dmg_hzl_panshi_neiqu_kekao：记录房屋面积、面积等级级面积来源是否可靠等信息

可靠的判断条件：房屋面积与真实面积相差 1 平米且房东房本（人工审核）、纠错房本（人工审核）、租房房本（人工审核）、核实产证为可靠来源

```
MAX(CASE WHEN ABS(a.house_area_v2 - b.area) <= 1 AND
b.sourcefrom IN (16, 20, 44) THEN 1 ELSE 0 END) AS is_kekao
```

	C	D	E	F	G	H	I	J	K	L
1	companyname	room_panshi_id	base_house_id	property_uuid	property_no	square	house_area_v2	house_area_level	house_area_source	is_kekao
2	天津诺家	59935716	59094086	a651bda0f1c8413b8d90b39bcd296272	81106BDFC2CBD	149.0	149.05	10	2	0
3	N+平台（郑州）	443586652	281087941	7a1f0c44e45848bfac789546b9044ef0	812208C878A5F	132.0	132.0	10	10	0
4	武汉合和致远房地产经纪有限公司	67008727	64756929	dc2163cc19d94f21bc9d15208d19cca6	80229DC93F921	110.43	110.43	10	10	0
5	58N+	23767809	32239829	0940348ca929446eae483d27cd8ab54f	81010D513A164	59.0	59.17	10	1	1
6	天津诺家	10404845	15116692	2e55f7ba96f34b868ef2bb5abd428471	8032678DC26A7	244.36	244.19	10	1	1
7	糯家（惠州）信息科技有限公司	249888932	138463909	ad93a43a3ba343dda813fad2f306131a	80212D2C466C4	81.0	82.0	10	2	0
8	安居客N+珠中	15377882	8628615	726c54fecd544f4fb7606c5cf04aa2d8	80725B396A6E9	93.04	91.19	10	2	0
9	武汉合和致远房地产经纪有限公司	67509469	65260230	71f0c63347b146ea876367988b26f03c	807079BCD8883	78.02	78.02	10	10	0
10	天津诺家	43116060	50090798	687ae9a4aac5433092d42911b60f4d59	80802DC15EAA0	125.0	126.0	10	10	0
11	天津诺家	59988698	59147074	2a4f1a4245c149fba2644d66151c76a2	806215B4EBDBE	84.99	84.99	10	2	0
12	天津诺家	66326401	64022826	0309081eeb7340bfaf2b0aa1be6b47fc	80826FBCD26B9	122.37	122.0	10	10	0

4）dmg_hzl_panshi_neiqu_company_yz_sts:统计内渠面积一致且来源可靠的比率（基于表 3 dmg_hzl_panshi_neiqu_kekao）

```
---内渠可置信度，基于可靠来源的面积比对
drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_company_yz_sts;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_company_yz_sts
as
select company_uuid,companyname,
count(distinct base_house_id) as nq_num,
count(distinct case when abs(square-house_area_v2)<=1 then base_house_id end) as nq_yz_num,
count(distinct case when abs(square-house_area_v2)<=1 then base_house_id end)/count(distinct base_house_id) as nq_yz_rate
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_kekao
where is_kekao=1 and house_area_v2>0
group by company_uuid,companyname
;
```

	A	B	C	D	E
1	company_uuid	companyname	nq_num	nq_yz_num	nq_yz_rate
2	nanchanggerenyonghuk_a43_saas2_1c2e4	巧房通	7501	6829	0.9104119450739901
3	nuojiahuizhouxinxike_8d4_saas2_23b14	糯家（惠州）信息科技有限公司	1325	1283	0.9683018867924529
4	suzhouhehezhiyuanpin_bb9_saas2_72ba6	苏州合和茂远房产经纪有限公司	6837	6222	0.910048266783677
5	hehezhiyuan_cd7_saas2_9d71bc8ea53644	长沙安居客N+	4496	4181	0.9299377224199288
6	zhuhaibangfang_companye13583dfc	安居客N+珠中	2910	2702	0.9285223367697595
7	nuojiadongguanxinxik_9e9_saas2_a7861	糯家（东莞）信息科技有限公司	1404	1314	0.9358974358974359
8	anjunuojia_cd0_saas2_d8b6f7945987492	安居诺家	1026	989	0.9639376218323586
9	nuojiaguangzhouxinxi_0ad_saas2_78d2e	糯家（广州）信息科技有限公司	926	864	0.9330453563714903
10	nanjinghehezhiyuan_192_saas2_e72adad	南京合和致远	8700	8125	0.9339080459770115
11	chengdouhehezhiyuanf_c20_saas2_1bf2a	成都合和致远房地产经纪有限公司	5940	5675	0.9553872053872053
12	taiyuangerenyonghuzh_151_saas2_2f2f4	太原合和致远房地产经纪有限公司	2609	2412	0.9244921425833653
13	wuhannuojiafangdicha_saas2_e27b0c8ca	武汉合和致远房地产经纪有限公司	11415	11121	0.9742444152431012
14	nuojiafoshanxinxikej_fcb_saas2_f789a	糯家（佛山）信息科技有限公司	661	606	0.9167927382753404
15	nuojiashenzhenxinxik_33f_saas2_ea212	糯家（深圳）信息科技有限公司	676	634	0.9378698224852071
16	tianjinnuojiafangdic_04e_saas2_f69ab	天津诺家	15388	14216	0.923836755913699

5）dmg_hzl_panshi_neiqu_compare_yz：统计人/司来源的内渠房屋信息（为计算人/司内渠房屋来源面积一致率作准备）

```
---不同经纪人/司内渠比对一致率
drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_compare_yz;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_compare_yz
as
select a.*,b.area,b.sourcefrom,b.extrano,
case when cast(square as string)=cast(area as string) and (a.company_uuid=extrano or a.companyname=extradescribe) then 1 else 0 end as is_nq,
case when b.sourcefrom in (1,2,4,11,24) and coalesce(b.extrano,'')<>'' then concat('c_',b.extrano)
    when b.sourcefrom in (18) and coalesce(b.extrano,'')<>'' then concat('b_',b.extrano)
    else '' end as extrano_new
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_kekao a
join hdp_anjuke_dw_db_temp.dw_panshi_house_area_source_log b
on a.base_house_id=b.house_id
where a.is_kekao=0 or house_area_v2=0 or abs(square-house_area_v2)<=1
;
```

	G	H	I	J	K	L	M	N	O	P	Q
1	property_no	square	house_area_v2	house_area_level	house_area_source	is_kekao	area	sourcefrom	extrano	is_nq	extrano_new
2	80904EE3E6A57	73.8	73.8	10	10	0	73.8	41		0	
3	48D QDCYQS000027622	91.0	91.0	10	10	0	91.0	41		0	
4	81116616C33FC	127.12	127.12	10	1	1	127.12	4	henanyuhefangdichany_70b_saas2_522ff	0	c_henanyuhefangdichany_70b_saas2_522ff
5	80315786D1E75	157.0	156.54	10	2	0	156.0	1	17419	0	c_17419
6	806046756E2EC	113.36	113.36	10	5	0	113.36	41		0	
7	80407FD8DE33E	130.0	130.0	10	10	0	140.0	41		0	
8	804110C03EF0C	105.19	105.19	10	2	0	105.19	17	20210725CIJTI	0	
9	80502FC877106	125.86	125.86	10	5	0	125.86	41		0	
10	81018594F5714	99.84	99.84	10	2	0	99.84	4	taiyuangerenyonghuzh_151_saas2_2f2f4	1	c_taiyuangerenyonghuzh_151_saas2_2f2f4
11	80420370FA4C5	71.0	71.0	10	10	0	71.0	41		0	
12	811207FA255E2	89.0	89.0	10	10	0	89.0	41		0	
13	8110286AA82DB	113.0	114.0	10	10	0	114.0	1	26969	0	c_26969
14	80903B3274EF1	118.05	117.82	10	2	0	118.05	4	hehezhiyuanxian_cca_saas2_053c5fa1c2	1	c_hehezhiyuanxian_cca_saas2_053c5fa1c2

6)dmg_hzl_panshi_neiqu_house_yz_sts:统计内渠每个房屋面积一致率

```
drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_house_yz_sts;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_house_yz_sts
as
select a.base_house_id as house_id,
avg(nq_num) as nq_num,
avg(nq_yz_num) as nq_yz_num,
avg(nq_yz_num)/avg(nq_num) as nq_yz_rate
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_kekao a
join hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_company_yz_sts c
on a.company_uuid=c.company_uuid and a.companyname=c.companyname
group by base_house_id
;
```

	A	B	C	D
1	house_id	nq_num	nq_yz_num	nq_yz_rate
3	130972409	8700.0	8125.0	0.9339080459770115
4	64952361	15388.0	14216.0	0.923836755913699
5	105470292	9535.0	9053.0	0.9494493969585737
6	23304195	9760.0	8355.0	0.8560450819672131
7	115917976	9760.0	8355.0	0.8560450819672131
8	276146038	11415.0	11121.0	0.9742444152431012
9	98167379	11415.0	11121.0	0.9742444152431012
10	254413721	6837.0	6222.0	0.910048266783677
11	287802441	9760.0	8355.0	0.8560450819672131
12	84856898	9760.0	8355.0	0.8560450819672131
13	138160602	16952.0	16010.0	0.94443133553563
14	33169725	1026.0	989.0	0.9639376218323586
15	5136648	926.0	864.0	0.9330453563714903
16	44940448	8700.0	8125.0	0.9339080459770115

7)dmg_hzl_panshi_neiqu_compare_yz_sts: 统计人/司内渠房屋来源面积一致率

```
drop table if exists hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_compare_yz_sts;
create table hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_compare_yz_sts
as
select sourcefrom,extrano_new,
count(distinct base_house_id) as fang_cnt,
count(distinct case when abs(square-area)<=1 then base_house_id end) as fang_yz_cnt,
count(distinct case when abs(square-area)<=1 then base_house_id end)/count(distinct base_house_id) as yizhi_rate
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_neiqu_compare_yz
where is_nq=0
group by sourcefrom,extrano_new
grouping sets(
    (sourcefrom),
    (extrano_new)
)
;
```

	A	B	C	D	E
1	sourcefrom	extrano_new	fang_cnt	fang_yz_cnt	yizhi_rate
2	null	b_3564802	1	1	1.0
3	null	b_202259097	2	0	0.0
4	null	b_1560868	1	0	0.0
5	null	b_202307053	3	2	0.6666666666666666
6	null	b_100856513	1	0	0.0
7	null	b_2777014	1	1	1.0
8	null	b_6135307	3	3	1.0
9	null	b_6094790	1	1	1.0
10	null	b_1905727	5	5	1.0
11	null	b_203156111	1	1	1.0
12	null	b_202359317	2	2	1.0
13	null	b_202351037	1	1	1.0
14	null	b_100351005	1	0	0.0
15	null	b_202282355	2	2	1.0
16	null	b_5462082	1	1	1.0

8）dmg_hzl_panshi_same_house_same_area：统计某个来源的面积与其他来源面积相同的数量，值越大，说明该来源的面积越可靠。

	A	B	C	D	E	F	G	H	I	J
1	house_id	area	sourcefrom	extrano_new	s0_sac	s0_nsac	s0_rsac	s1_sac	s1_nsac	s1_rsac
2	71036224	31.02	1	c_27630	6	1	0.8571428571428571	2	0	1.0
3	18210136	56.0	4	c_nanjingyunjufangchan_787_saas2_d1779	12	0	1.0	11	0	1.0
4	134939501	88.0	42	o_947ab259-a56a-44a9-9d9b-41676940a1b6	0	1	0.0	0	1	0.0
5	235728258	103.19	4	c_suqiandefufangchan_d77_saas2_64ef46c	2	0	1.0	0	0	0.0
6	133247102	139.02	4	c_shimaobinjiangzushou_36b_saas2_e6f44	3	0	1.0	0	0	0.0
7	137129844	80.17	1	c_318939	3	0	1.0	1	0	1.0
8	26846191	168.0	1	c_31575	4	0	1.0	3	0	1.0
9	125867029	120.5	1	c_10299	34	0	1.0	32	0	1.0
10	156355369	95.0	4	c_haerbinshishangfujia_422_saas2_a230b	2	0	1.0	0	0	0.0
11	106928837	203.0	41	o_656ece42-d9a8-4654-b51b-d5bc48d3015f	1	0	1.0	0	0	0.0
12	150729357	61.65	41	o_67438b8b-c8a5-4334-9580-5557282188b5	2	0	1.0	0	0	0.0
13	142002363	90.0	3	o_baa38530-8523-4570-b7ad-1d1e9a337883	0	1	0.0	0	0	0.0
14	158473746	80.4	4	c_yangchengdichan_28c_saas2_4b6b14bbb7	3	1	0.75	0	0	0.0
15	135309956	133.0	1	c_6378	2	1	0.6666666666666666	2	0	1.0
16	262564956	42.85	41	o_bed891ce-41f1-45d2-bfea-cdced7ee2138	1	1	0.5	0	0	0.0

```
drop table if EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_same_house_same_area;
create TABLE if not EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_same_house_same_area as
select t1.house_id,t1.area,t1.sourcefrom,t1.extrano_new,
count(case when abs(t1.area-t2.area)<=1 then t1.house_id end) as s0_sac,--相同面积的房屋数量
count(case when abs(t1.area-t2.area)>1 then t1.house_id end) as s0_nsac,--不同面积数量
count(case when abs(t1.area-t2.area)<=1 then t1.house_id end)/greatest(count(t1.house_id),1) as s0_rsac,--同面积的占比
count(case when t2.sourcefrom=1 and abs(t1.area-t2.area)<=1 then t1.house_id end) as s1_sac,--来源为1的同面积数量
count(case when t2.sourcefrom=1 and abs(t1.area-t2.area)>1 then t1.house_id end) as s1_nsac,
count(case when t2.sourcefrom=1 and abs(t1.area-t2.area)<=1 then t1.house_id end)/greatest(count(case when t2.sourcefrom=1 then
count(case when t2.sourcefrom=2 and abs(t1.area-t2.area)<=1 then t1.house_id end) as s2_sac,
count(case when t2.sourcefrom=2 and abs(t1.area-t2.area)>1 then t1.house_id end) as s2_nsac,
count(case when t2.sourcefrom=2 and abs(t1.area-t2.area)<=1 then t1.house_id end)/greatest(count(case when t2.sourcefrom=2 then
count(case when t2.sourcefrom=3 and abs(t1.area-t2.area)<=1 then t1.house_id end) as s3_sac,
count(case when t2.sourcefrom=3 and abs(t1.area-t2.area)>1 then t1.house_id end) as s3_nsac,
count(case when t2.sourcefrom=3 and abs(t1.area-t2.area)<=1 then t1.house_id end)/greatest(count(case when t2.sourcefrom=3 then
count(case when t2.sourcefrom=4 and abs(t1.area-t2.area)<=1 then t1.house_id end) as s4_sac,
count(case when t2.sourcefrom=4 and abs(t1.area-t2.area)>1 then t1.house_id end) as s4_nsac,
count(case when t2.sourcefrom=4 and abs(t1.area-t2.area)<=1 then t1.house_id end)/greatest(count(case when t2.sourcefrom=4 then
count(case when t2.sourcefrom=5 and abs(t1.area-t2.area)<=1 then t1.house_id end) as s5_sac,
count(case when t2.sourcefrom=5 and abs(t1.area-t2.area)>1 then t1.house_id end) as s5_nsac,
count(case when t2.sourcefrom=5 and abs(t1.area-t2.area)<=1 then t1.house_id end)/greatest(count(case when t2.sourcefrom=5 then
count(case when t2.sourcefrom=6 and abs(t1.area-t2.area)<=1 then t1.house_id end) as s6_sac
```

9）dmg_hzl_panshi_same_unit_same_area:统计同个单元某个来源面积与其他来源面积相同的数量及单元相同 and 房号相同（即上下层关系）的某个来源面积与其他来源面积相同的数量

```
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=41 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=41 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=41 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=42 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=42 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=42 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=43 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=43 and abs(t1.area-t2.area)
count(case when t1.floor_new<>'' and t2.floor_new<>'' and t1.floor_new=t2.floor_new and t2.sourcefrom=43 and abs(t1.area-t2.area)
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_house_area t1
join hdp_anjuke_dw_db_temp.dmg_hzl_panshi_house_area t2
on t1.building_id=t2.building_id and t1.units_id=t2.units_id
where t1.extrano_new<>t2.extrano_new
and t1.sourcefrom in(1,2,3,4,12,18,19,24,41,42,54)
and t2.sourcefrom in(1,3,4,5,7,9,10,11,12,13,17,18,19,21,22,23,41,42,43)
and t1.house_id<>t2.house_id and t1.building_id>0 and t1.units_id>0 and t2.building_id>0 and t2.units_id>0
group by t1.house_id,t1.area,t1.sourcefrom,t1.extrano_new
;
```

10）dmg_hzl_panshi_area_real:统计房屋真实面积

	A	B	C	D	E	F	G	H
1	house_id	house_area_v2	house_area_level	building_id	community_id	units_id	floor_id	house_no
2	346520263	108.99	10	6630970	1774448	16696328	129447512	31313031
3	147821125	63.0	10	2119080	122931	5640077	45389913	323033
4	129905680	89.75	10	1851163	143970	4958058	39772384	33323032
5	261222014	58.18	10	5162258	215149	13375959	103470389	313031
6	136257004	79.8	10	1949141	156167	5179584	41641928	31353032
7	103271922	130.85	10	459581	118226	4183585	32480484	393035
8	43126809	113.1	10	619088	98061	1791425	13066934	373031
9	271362451	116.58	10	5309701	580184	13789292	106524432	333034
10	375480561	78.77	10	7350123	1212888	17812446	137572984	31393036
11	129898683	86.78	10	1851072	144870	4957841	39769760	32353032
12	151624486	80.79	10	2175927	197813	5800583	46704493	343032
13	293942528	81.28	10	5710423	214965	14731861	113789614	353034

11）dmg_hzl_panshi_area_samples：汇总所有表格信息，建立样本数据

```
drop table if EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_samples;
create TABLE if not EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_samples as
select t1.addat,t1.updateat,t1.cmc_disp_local_id as city_id,t1.house_id,t1.sourcefrom,t1.extrano,t1.extradescribe,t2.house_area_v2,t1.area,t1.count,
coalesce(t3.fang_cnt,0) as source_jiucuo_fang_cnt,--某面积来源下需纠错的房屋数量
case when t1.extrano_new rlike '^b_|^c_' then coalesce(t4.fang_cnt,0) else -999 end as jiucuo_fang_cnt,--人/司纠错房号数
coalesce(t5.fang_cnt,0) as source_levela_fang_cnt,--A类来源面积房屋数量
case when t1.extrano_new rlike '^b_|^c_' then coalesce(t6.fang_cnt,0) else -999 end as levela_fang_cnt,--人/司A面积房号数
case when coalesce(t5.fang_cnt,0)=0 then 0 else coalesce(t3.fang_cnt,0)/coalesce(t5.fang_cnt,0) end as source_jiucuo_rate,--来源纠错率
case when t1.extrano_new rlike '^b_|^c_' and coalesce(t6.fang_cnt,0)=0 then 0
    when t1.extrano_new rlike '^b_|^c_' then coalesce(t4.fang_cnt,0)/coalesce(t6.fang_cnt,0)
    else -999 end as jiucuo_rate,--人/司纠错率
coalesce(t7.fang_cnt,0) as source_nq_fang_cnt,--来源内渠房号数
coalesce(t7.fang_yz_cnt,0) as source_nq_fang_yz_cnt,--来源内渠面积一致房号数
coalesce(t7.yizhi_rate,0) as source_nq_yizhi_rate,--来源内渠面积一致房号占比
case when t1.extrano_new rlike '^b_|^c_' then coalesce(t8.fang_cnt,0) else -999 end as nq_fang_cnt,--人/司内渠房号数
case when t1.extrano_new rlike '^b_|^c_' then coalesce(t8.fang_yz_cnt,0) else -999 end as nq_fang_yz_cnt,--人/司内渠面积一致房屋数
case when t1.extrano_new rlike '^b_|^c_' then coalesce(t8.yizhi_rate,0) else -999 end as nq_yizhi_rate,--人/司内渠面积一致房屋数占比
coalesce(t9.nq_num,0) as nq_num,--内渠房屋数
coalesce(t9.nq_yz_num,0) as nq_yz_num,--内渠面积一致的房屋数
coalesce(t9.nq_yz_rate,0) as nq_yz_rate,--内渠面积一致房屋数占比
concat(
    coalesce(t10.s0_sac,0),',',
    coalesce(t10.s0_nsac,0),',',
    coalesce(t10.s0_rsac,0),',',
```

```
case when abs(t1.area-t2.house_area_v2)<=1 then 1 else 0 end as y,
-- 训练样本标识
```

	P	Q	R	S	T	U	V	W	X	Y	
1	jiucuo_rate	source_nq_fang_cnt	source_nq_fang_yz_cnt	source_nq_yizhi_rate	nq_fang_cnt	nq_fang_yz_cnt	nq_yizhi_rate	nq_num	nq_yz_num	nq_yz_rate	area_cnt_info
2	7.5111864461243O8E-4	268868	248110	0.9227948286891708	49	22	0.4489795918367347	0.0	0.0	0.0	1,0,1.0,1,0,1.0,0,0,(
3	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
4	0.0	127021	99274	0.7815558057329103	0	0	0.0	0.0	0.0	0.0	1,0,1.0,0,0,0.0,0,0,(
5	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
6	-999.0	42392	16223	0.38269013021324777	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
7	0.0	127021	99274	0.7815558057329103	0	0	0.0	0.0	0.0	0.0	0,1,0.0,0,0,0.0,0,0,(
8	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
9	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
10	-999.0	42392	16223	0.38269013021324777	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
11	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
12	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
13	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	1,0,1.0,0,0,0.0,0,0,(
14	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(
15	-999.0	372997	280433	0.7518371461432666	-999	-999	-999.0	0.0	0.0	0.0	0,0,0.0,0,0,0.0,0,0,(

训练样本：

```
set hive.exec.compress.output=false;
drop table if EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_train_samples;
create TABLE if not EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_train_samples as
select city_id,house_id,sourcefrom,
source_jiucuo_fang_cnt,jiucuo_fang_cnt,source_levela_fang_cnt,levela_fang_cnt,source_jiucuo_rate,
jiucuo_rate,source_nq_fang_cnt,source_nq_fang_yz_cnt,source_nq_yizhi_rate,nq_fang_cnt,
nq_fang_yz_cnt,nq_yizhi_rate,nq_num,nq_yz_num,nq_yz_rate,area_cnt_info,y
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_samples
where train_dataset=1
and sourcefrom in (1, 3, 4, 12, 18, 19, 41, 42, 54);
```

预测样本：

```
set hive.exec.compress.output=false;
drop table if EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_predict_samples;
create TABLE if not EXISTS hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_predict_samples as
select city_id,house_id,sourcefrom,
source_jiucuo_fang_cnt,jiucuo_fang_cnt,source_levela_fang_cnt,levela_fang_cnt,source_jiucuo_rate,
jiucuo_rate,source_nq_fang_cnt,source_nq_fang_yz_cnt,source_nq_yizhi_rate,nq_fang_cnt,
nq_fang_yz_cnt,nq_yizhi_rate,nq_num,nq_yz_num,nq_yz_rate,area_cnt_info,unique_id
from hdp_anjuke_dw_db_temp.dmg_hzl_panshi_area_samples
where zhixiao>0
;
```

2、使用数据集对 xgboost 训练

1)从 Hadoop 分布式系统将数据集复制到本地

```
'hdfs dfs -copyToLocal
/home/hdp_anjuke_bi/warehouse/hdp_anjuke_dw_db_temp.db/dmg_hzl_panshi
_area_train_samples /code/huzuoliang/panshi/'
```

2）读取并处理数据

1.根据字段将数据分为训练集 X 和标签 Y；

```
X = data[t_cols_name]
y = data[t_col_target_name]
```

	source_jiucuo_fang_cnt	jiucuo_fang_cnt	source_levela_fang_c
0	6353	200	6996486
1	11269	248	7693711
2	32347	-999	13539910
3	34967	1	5054345
4	11269	156	7693711
5	6353	23	6996486
6	11269	1	7693711
7	11269	248	7693711
8	6353	108	6996486
9	2229	-999	1137823
10	11269	248	7693711
11	6353	144	6996486
12	6353	217	6996486
13	6353	14	6996486
14	11269	16	7693711
15	34967	0	5054345
16	6353	62	6996486

	y
0	1
1	1
2	1
3	0
4	1
5	1
6	1
7	1
8	1
9	0
10	1
11	0
12	1
13	1
14	1

2.这里将 sourcefrom 和 cityid 转换为 one-hot 向量的形式

```
ohe_sourcefrom = ohe_transform
ohe_cityid = ohe_transform
```

	sourcefrom_1	sourcefrom_3	sourcefrom_4	sourcefrom_1
0	1.00000	0.00000	0.00000	0.00000
1	0.00000	0.00000	1.00000	0.00000
2	0.00000	0.00000	0.00000	0.00000
3	0.00000	0.00000	0.00000	0.00000
4	0.00000	0.00000	1.00000	0.00000
5	1.00000	0.00000	0.00000	0.00000
6	0.00000	0.00000	1.00000	0.00000
7	0.00000	0.00000	1.00000	0.00000
8	1.00000	0.00000	0.00000	0.00000
9	0.00000	0.00000	0.00000	0.00000
10	0.00000	0.00000	1.00000	0.00000

3. 删除 X 中的 sourcefrom 和 cityid 两列，再以 one-hot 的形式拼接回去

```
X = X.drop("city_id", axis=1)
X = X.drop("sourcefrom", axis=1)
X = pd.concat([X, ohe_sourcefrom, ohe_cityid], axis=1)
```

3）加载模型进行训练

```
clf = xgb.XGBClassifier(tree_method="hist",  clf: XGBClassi
                        missing=-999,
                        n_estimators=500,
                        learning_rate=0.1,
                        max_depth=7,
                        subsample=0.8,
                        colsample_bytree=0.75,
                        enable_categorical=True,
                        scale_pos_weight=0.25
                        # callbacks=[early_stop]
                        )
```

4）将训练完成的模型文件复制到 Hadoop 分布式系统中

```
exec_command('hdfs dfs -copyFromLocal -f ' + model_name + ' /home/hdp_anjuke_bi/resultdata/hzl/model/panshi')
```

PanShiAreaAcceptPredictV2

1）将模型复制到本地，然后加载模型，将模型广播到 Spark 集群。

```
val inputPath = args(0)
val modelPath = args(1)
val partitionNum = args(2).toInt
val outputPath = args(3)
val sourceFromStr = args(4)
val cityIdStr = args(5)
//将模型文件复制到本地文件夹中，并返回本地文件位置
val modelLocalPath = DmFileUtils.copyToLocal(modelPath, "/tmp/xgboost")
//声明了一个不可变的变量 xgb，其类型为 Booster，为加载的xgboost模型
val xgb: Booster = XGBoost.loadModel(modelLocalPath)
// ？？？？？广播给整个Spark集群，在Spark集群中分布式地使用XGBoost模型进行预测或其他机器学习任务，
val xgboostBC = sc.broadcast(xgb)
```

2）对要预测的数据集进行预处理

使得数据与训练时的数据格式保持一致

```
val result = sc.textFile(inputPath).repartition(partitionNum).map {
  line =>

    //        city_id,house_id,sourcefrom,
    //      source_jiucuo_fang_cnt,jiucuo_fang_cnt,source_levela_fang_cnt,levela_fang_cnt,source_jiu
    //      jiucuo_rate,source_nq_fang_cnt,source_nq_fang_yz_cnt,source_nq_yizhi_rate,nq_fang_cnt,
    //      nq_fang_yz_cnt,nq_yizhi_rate,nq_num,nq_yz_num,nq_yz_rate,area_cnt_info,unique_id

    val elems = line.split("\001", -1)
    val cityId = "city_id_" + elems(0)
    val cityFeatures = Array.fill[Float](cityListLen)(0.0f)
    if (cityList.contains(cityId)) {
      cityFeatures(cityList.indexOf(cityId)) = 1.0f
    }
    val sourcefrom = "sourcefrom_" + elems(2)
    val sourcefromFeatures = Array.fill[Float](soursourcefromListLen)(0.0f)
    if (sourcefromList.contains(sourcefrom)) {
      sourcefromFeatures(sourcefromList.indexOf(sourcefrom)) = 1.0f
    }
    val featuresA = (3 ≤ until < 18).map(i => elems(i).toFloat)
    //  s0_sac 字段以后的特征以字符串的形式保存?
    val featuresB = elems(18).split(",").toSeq.map(e => e.toFloat)
    val features = featuresA ++ featuresB ++ sourcefromFeatures ++ cityFeatures
      最后的字段是标签值?
    val sourceFlag = elems.last

    (sourceFlag, features)
```

对数据进行分组，相当于 batch_size=100;接着将数据转为数据矩阵进行预测。

```
}.mapPartitions {
  partitionData =>
      相当于batch_size=100?
    partitionData.grouped(100).flatMap {
      gData =>
        val gSourceFlag = gData.map(_._1)
        val gDmData = gData.flatMap(_._2).toArray
        val nRow = gData.length
        val nCol = gData.head._2.length
          xgboost4j接收的训练数据必须转为DMatrix类，数据矩阵
        val dm = new DMatrix(gDmData, nRow, nCol, -999.0f)
        val proba = xgboostBC.value.predict(dm, false, 0)

        val batchResult = gSourceFlag.zip(proba).map(kv => kv._1 + "\001" + kv._2.head)
        batchResult
    }
}
```

3）将预测结果保存到指定路径 outputPath.

```
result.repartition(50).saveAsTextFile(outputPath)
```