大数据在房产领域的实践

蔡白银

LIANJIA 链家.com

自我介绍

- **蔡白银**

毕业于北京大学，在大数据数据挖掘领域有多年的经验, 目前就任链家网大数据架构师，负责链家网大数据体系的建设，运用大数据挖掘大数据价值助力房产领域的O2O，提升房屋买卖体验，使买卖房屋不再难

自我介绍

- **蔡白银**

毕业于北京大学，在大数据数据挖掘领域有多年的经验, 目前就认链家网大数据架构师，负责链家网大数据体系的建设，运用大数据挖掘大数据价值助力房产领域的O2O，提升房屋买卖体验，使买卖房屋不再难

提纲

- 蜀道难难于上青天
- 行困难而正确之事
- 往事可鉴未来可追

蜀道难难于上青天

- 客少、物少————————数据来源少
- 买卖行为少周期长————行为数据稀少
- 线下行为重容易分流—线上线下难打通
- 业务复杂性—-—————分析挖掘无坦途

提纲

- 蜀道难难于上青天
- 行困难而正确之事
- 往事可鉴未来可追

提升服务品质的环节

- 房源真实无虚假
- 合适的房屋给合适的人
- 房屋买卖不再难
- 缩短周期见效率
- 减少资源浪费

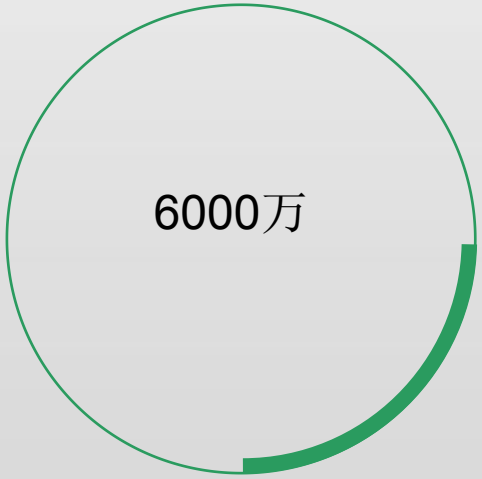
新浪财经 生活 > 正文

演员方子哥买房遭骗393万

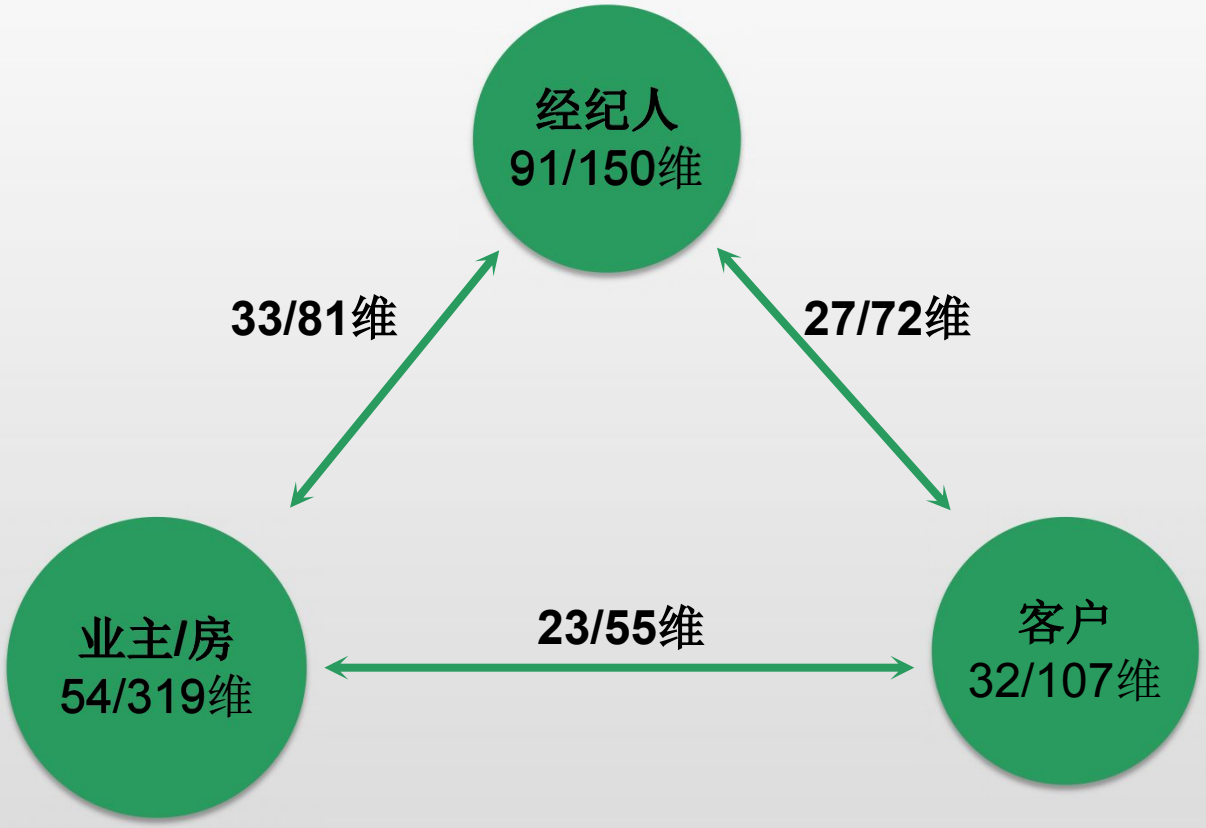
2015年09月24日 02:40 北京晨报

2012年6月，方子哥的妻子颜女士为儿子看中了位于朝阳区千鹤家园的一套房屋，后通过[redacted]约定以420万元的价格购买，并支付居间服务费92400元以及向担保公司支付17600元。此后，颜女士陆续向卖家李佳霖个人支付了393万元。她没有想到，李佳霖隐瞒已将房屋委托他人出售并办理最高额抵押贷款的事实，使用伪造的房屋所有权证骗取房款。后李佳霖因诈骗罪、合同诈骗罪被判处无期徒刑。

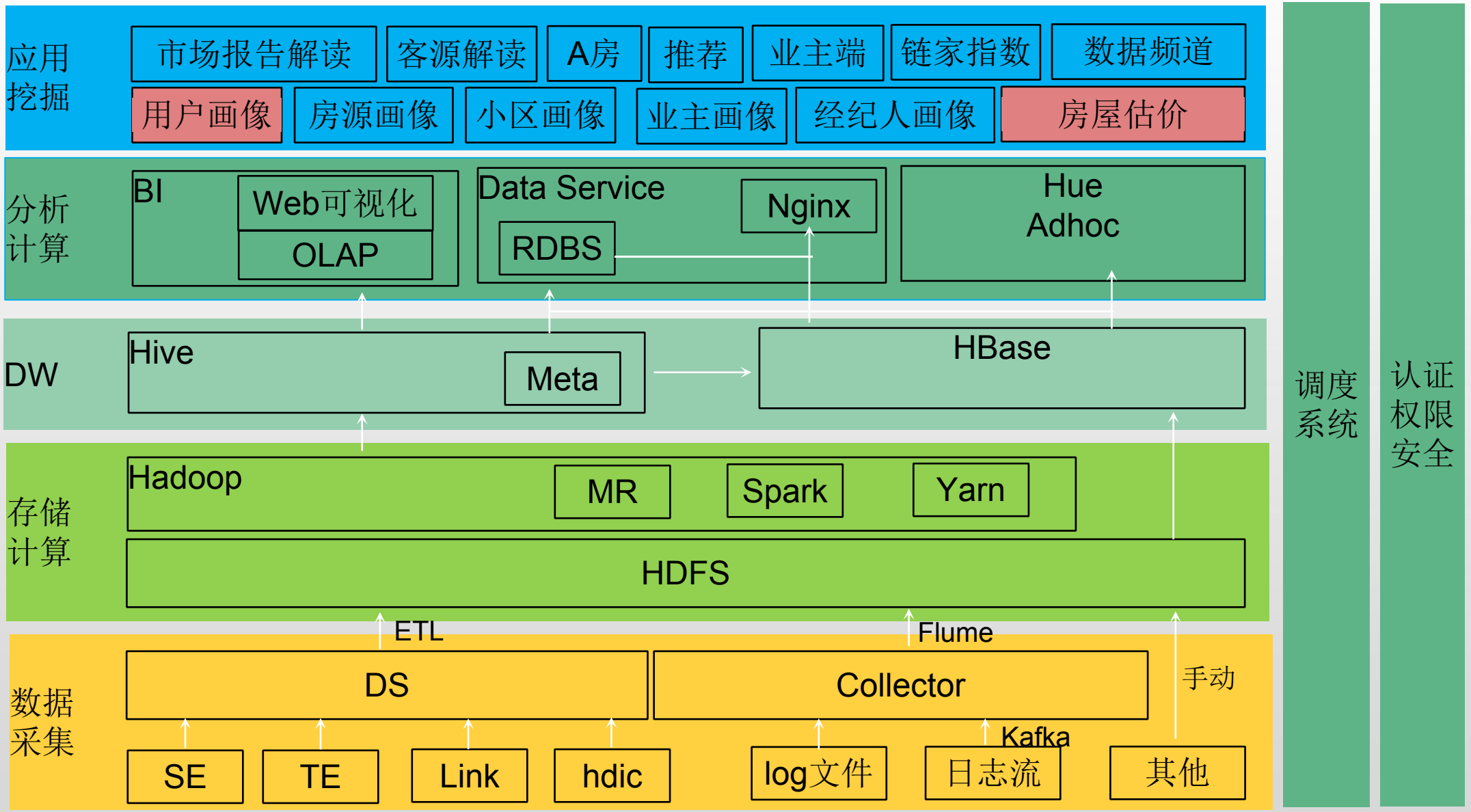
效果概述



效果概述



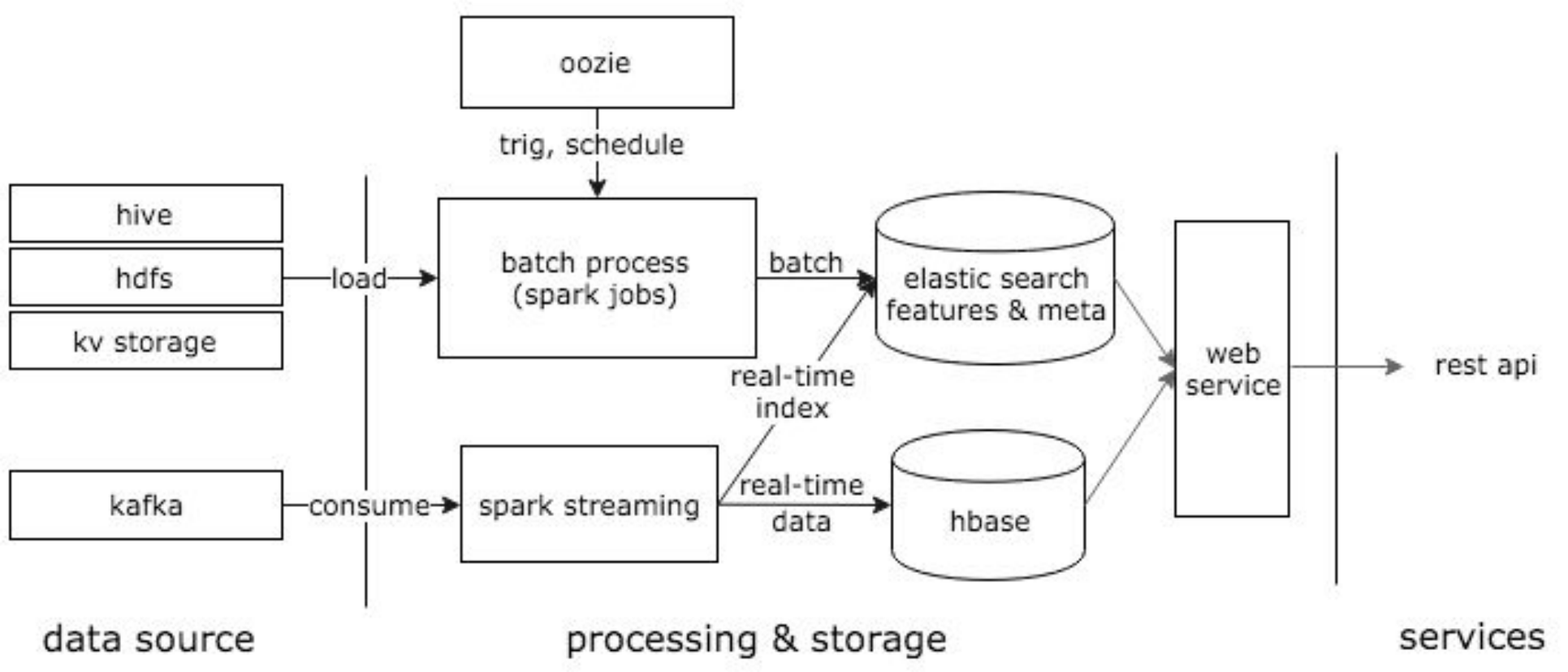
技术架构


应用挖掘
市场报告解读
客源解读
A房
推荐
业主端
链家指数
数据频道
用户画像
房源画像
小区画像
业主画像
经纪人画像
房屋估价
分析计算
BI
Web可视化
OLAP
Data Service
Nginx
RDBS
Hue
Adhoc
DW
Hive
Meta
HBase
存储计算
Hadoop
MR
Spark
Yarn
HDFS
ETL
Flume
数据采集
DS
Collector
手动
SE
TE
Link
hdic
log文件
Kafka
日志流
其他
调度系统
认证权限安全

用户画像

- elasticsearch，hbase，spark等成熟的开源数据存储、处理系统上
- elasticsearch存储、索引融合层全量数据，线上用户行为数据全量索引以及热数据
- hbase存储线上用户行为数据
- spark完成批量和流式数据处理，包括线下全量/增量数据导入，线上日志流处理并传送至elasticsearch集群。

用户画像

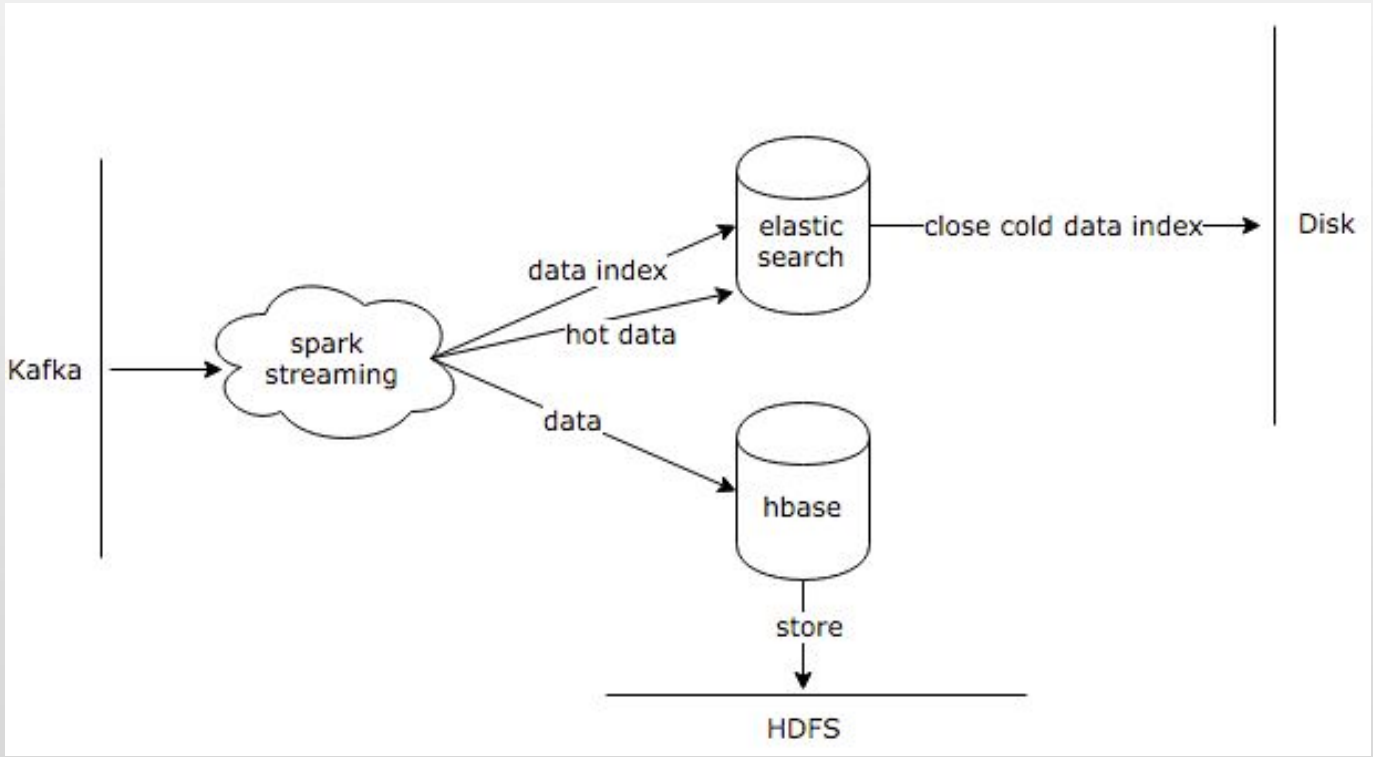

oozie
trig, schedule
hive
hdfs
kv storage
load
batch process
(spark jobs)
batch
elastic search
features & meta
real-time
index
web
service
rest api
kafka
consume
spark streaming
real-time
data
hbase
data source
processing & storage
services

用户画像

KV查询：如通过手机号查询客源的一切数据

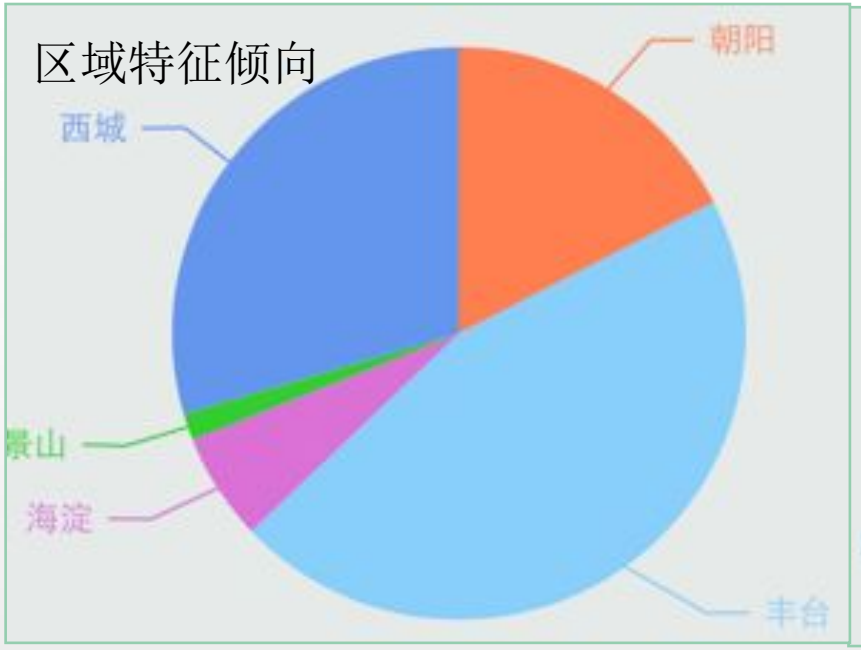
数据筛选：如筛选西山商圈，近三个月新增房源的小区名、挂牌价和房屋状态，要求房屋必须是精装修或大于3居室

OLAP查询: 如查询海淀区2015年不同月份客源带看次数的分布

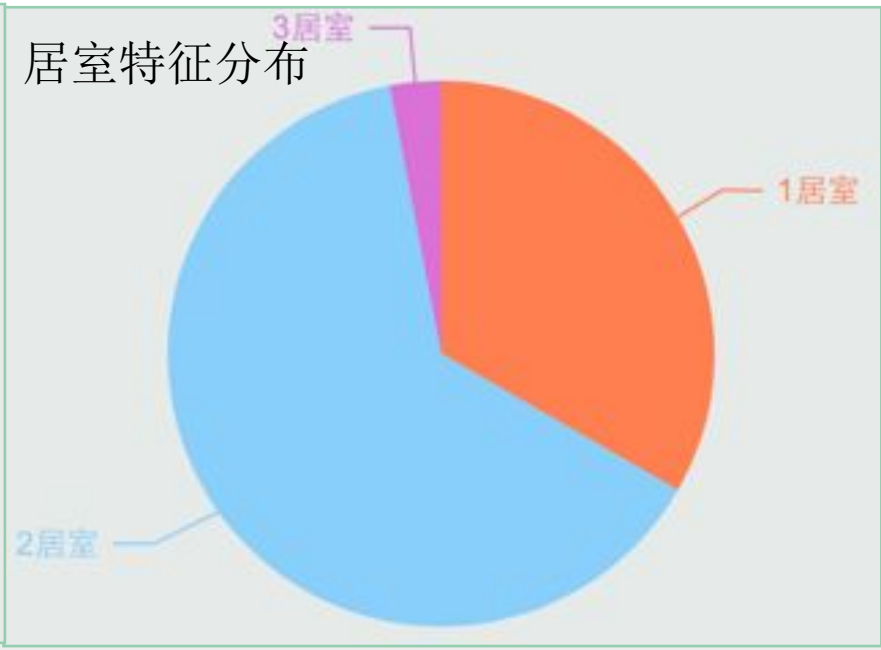


用户画像

区域特征倾向

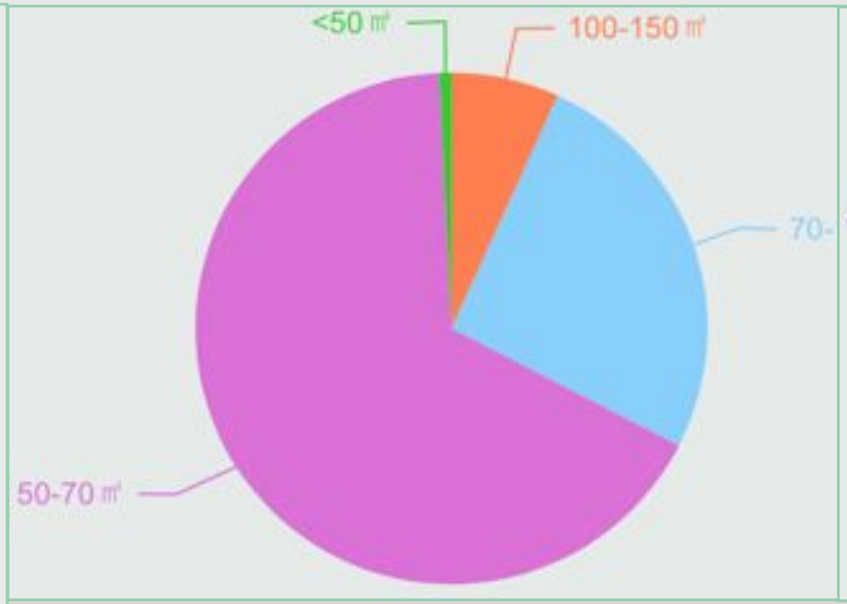


居室特征分布

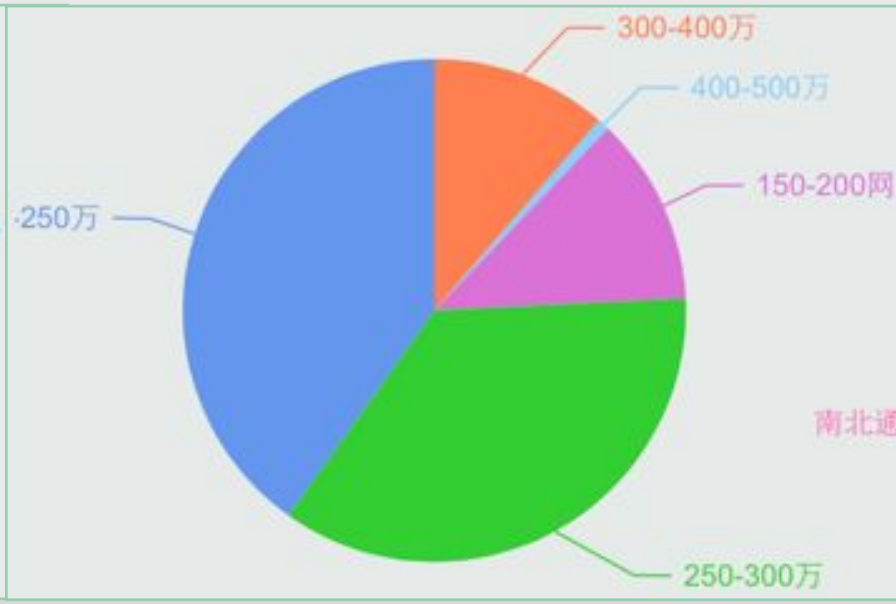


用户特征倾向

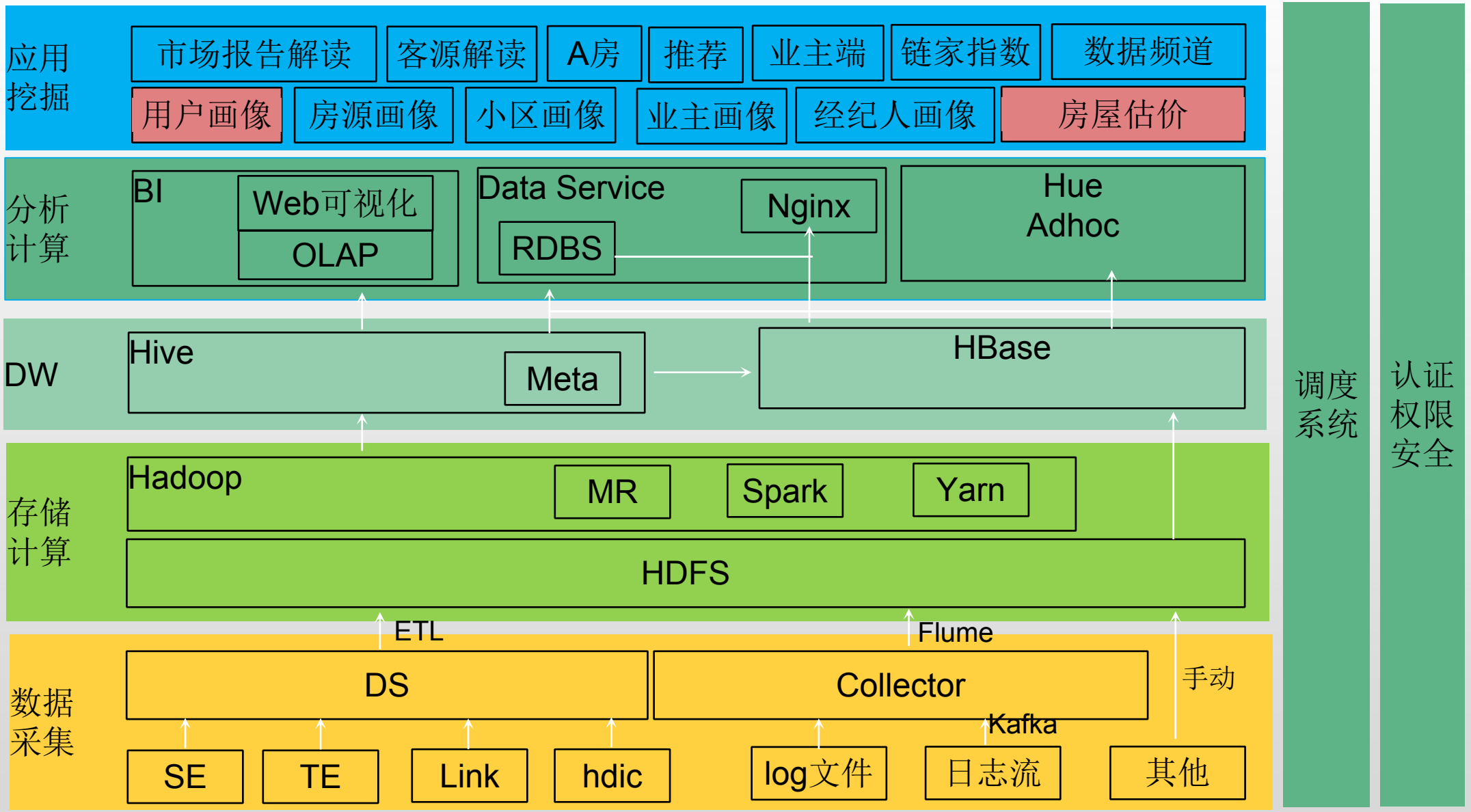
面积特征倾向



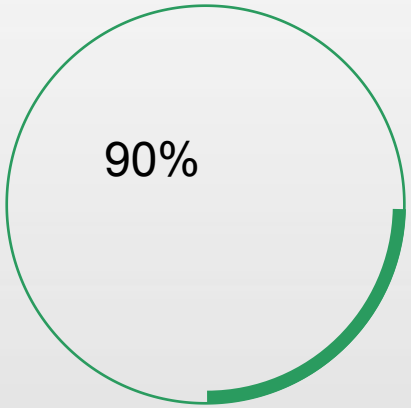
价格特征分布

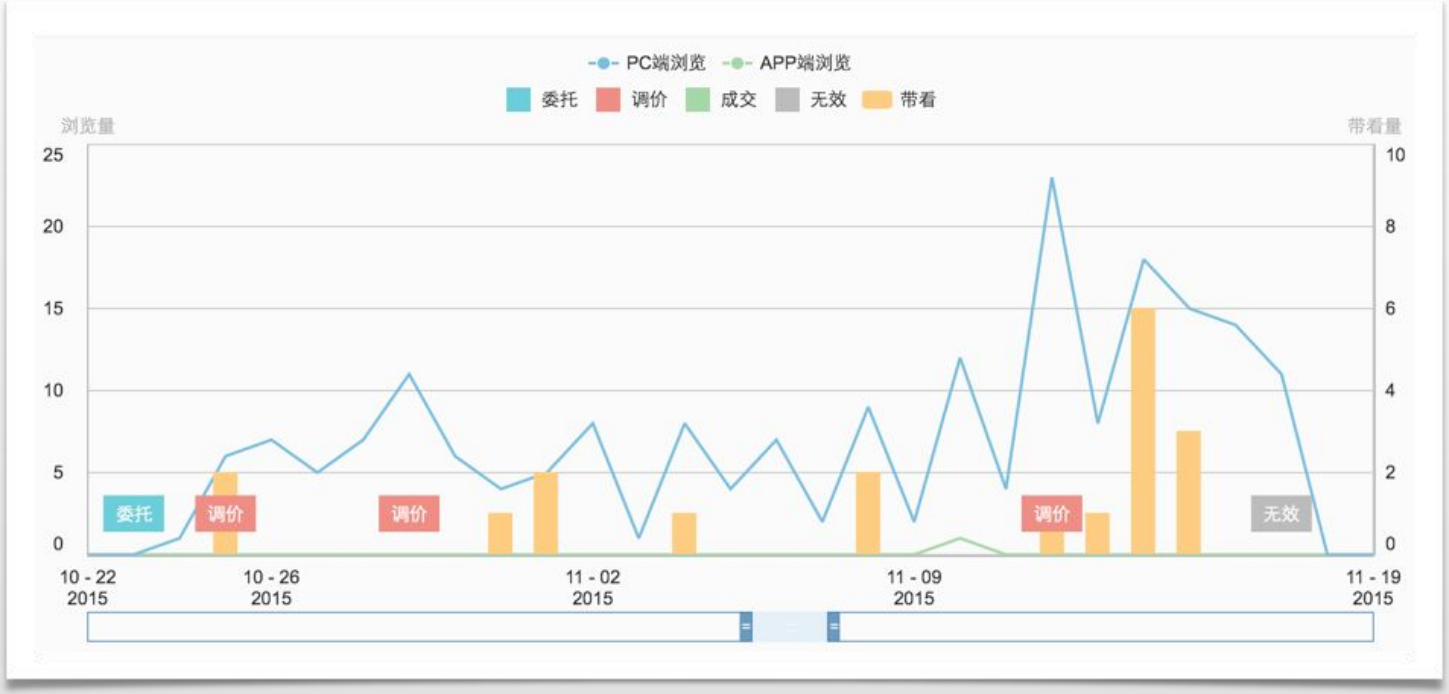


行困难而正确之事

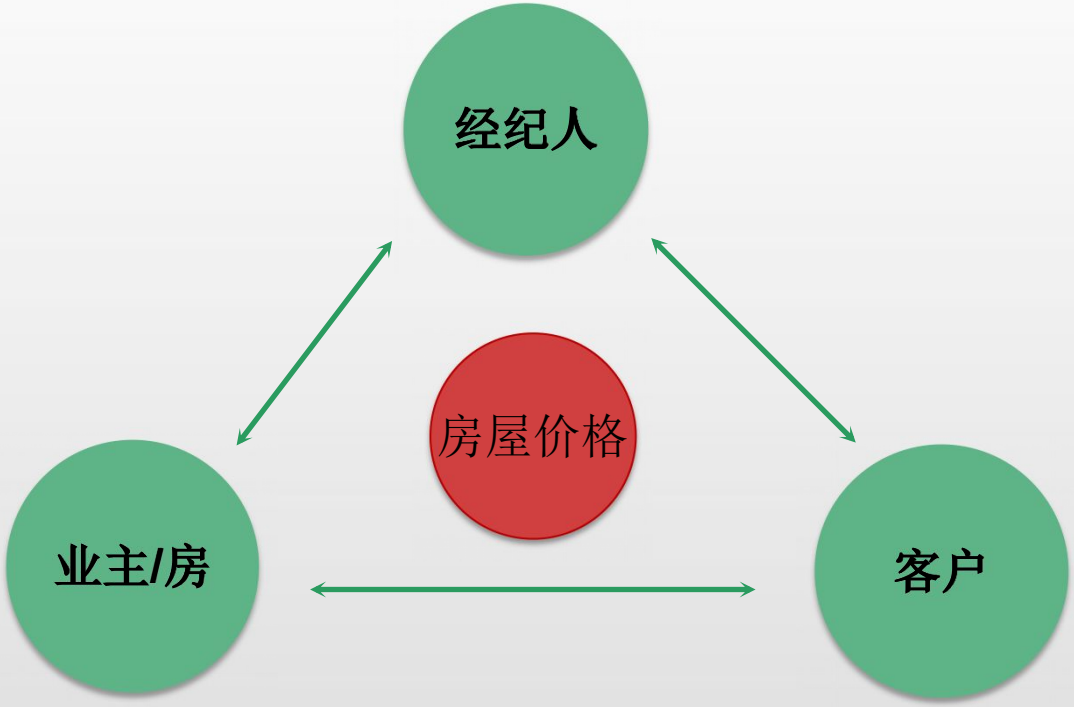

应用挖掘
市场报告解读
客源解读
A房
推荐
业主端
链家指数
数据频道
用户画像
房源画像
小区画像
业主画像
经纪人画像
房屋估价
分析计算
BI
Web可视化
OLAP
Data Service
RDBS
Nginx
Hue
Adhoc
DW
Hive
Meta
HBase
存储计算
Hadoop
MR
Spark
Yarn
HDFS
ETL
Flume
数据采集
DS
Collector
手动
Kafka
SE
TE
Link
hdic
log文件
日志流
其他
调度系统
认证权限安全

房屋估价


90%


PC端浏览
APP端浏览
委托
调价
成交
无效
带看
浏览量
带看量
25
20
15
10
5
0
10
8
6
4
2
0
委托
调价
调价
调价
无效
10 - 22
2015
10 - 26
2015
11 - 02
2015
11 - 09
2015
11 - 19
2015

房屋估价

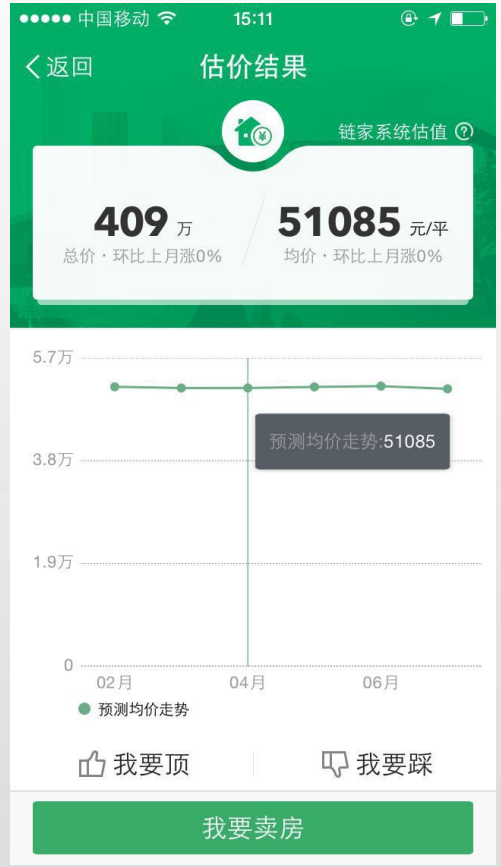


房屋估价


链家
lianjia.com


中国移动 15:12
北京
输入小区名或商圈名称
地图找房 优质学区 房产指南 查成交
购房亿元保单
亿元保障为您购房保驾护航
买房助手
链家买房全攻略
科
小区
房价
估价
计算器
房价-二手市场行情
全市均价
44341 元/平
昨日成交量
169 套
首页
消息
我

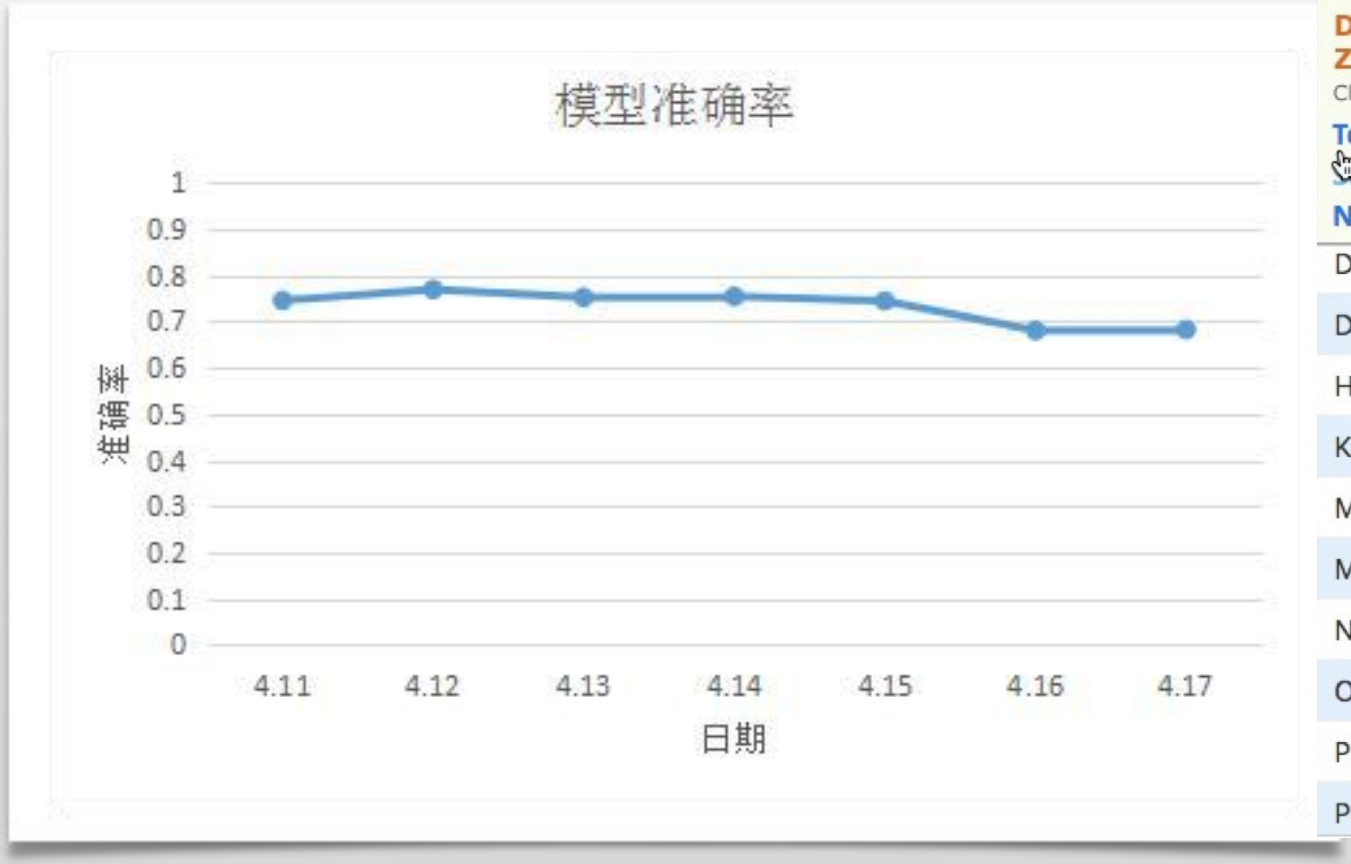

中国移动 15:12
返回
估价
清空
小区名称
远洋山水
楼栋号
4号楼
户型
2室1厅1卫
朝向
南北
面积
80 平方米
楼层
3 层
11 总层数
建筑年代
2000 年
户型结构
平层（一层结构）
房源特色(选填)


中国移动 15:11
返回
估价结果
链家系统估值
409 万
总价・环比上月涨0%
51085 元/平
均价・环比上月涨0%
5.7万
3.8万
1.9万
0
预测均价走势:51085
02月
04月
06月
预测均价走势
我要顶
我要踩
我要卖房

房屋估价

$$diff = \frac{|real\_money - predict\_money|}{real\_money}$$

估价准确：diff <= 5%　　$$right\_precent = \frac{right\_amount}{total\_amount}$$



Data Coverage and Zestimate Accuracy Table

Choose a location type below to change data:

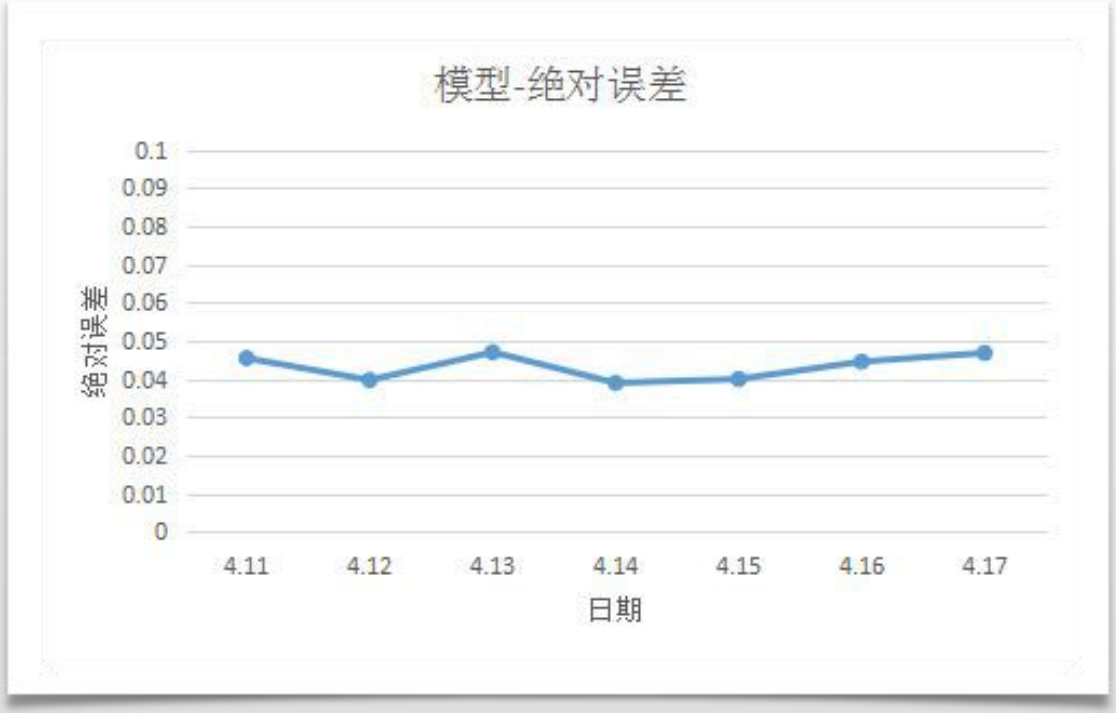
Top Metro Areas

States/Counties*

National

	Zestimate Accuracy	Homes on Zillow	Homes With Zestimates	Within 5% of Sale Price	Within 10% of Sale Price	Within 20% of Sale Price	Median Error
Denver, CO	★★★★	957.7K	887.3K	40.5%	66.9%	88.8%	6.5%
Detroit, MI	★★★	1.8M	1.7M	36.3%	60.5%	81.4%	7.5%
Houston, TX	★	2.2M	1.9M	--	--	--	--
Kansas City, MO	★	756.0K	692.9K	--	--	--	--
Miami-Fort Lauderdale, FL	★★	2.5M	2.4M	31.1%	54.3%	79.4%	8.9%
Minneapolis-St Paul, MN	★★★	1.2M	1.1M	35.8%	60.5%	84.6%	7.6%
New York, NY	★★★	5.3M	4.9M	32.4%	55.3%	77.0%	8.6%
Orlando, FL	★★★	875.5K	803.0K	37.1%	61.9%	83.0%	7.2%
Philadelphia, PA	★★★	2.1M	2.0M	35.5%	58.1%	79.1%	7.8%
Phoenix, AZ	★★★★	1.7M	1.5M	43.0%	68.0%	87.8%	6.2%

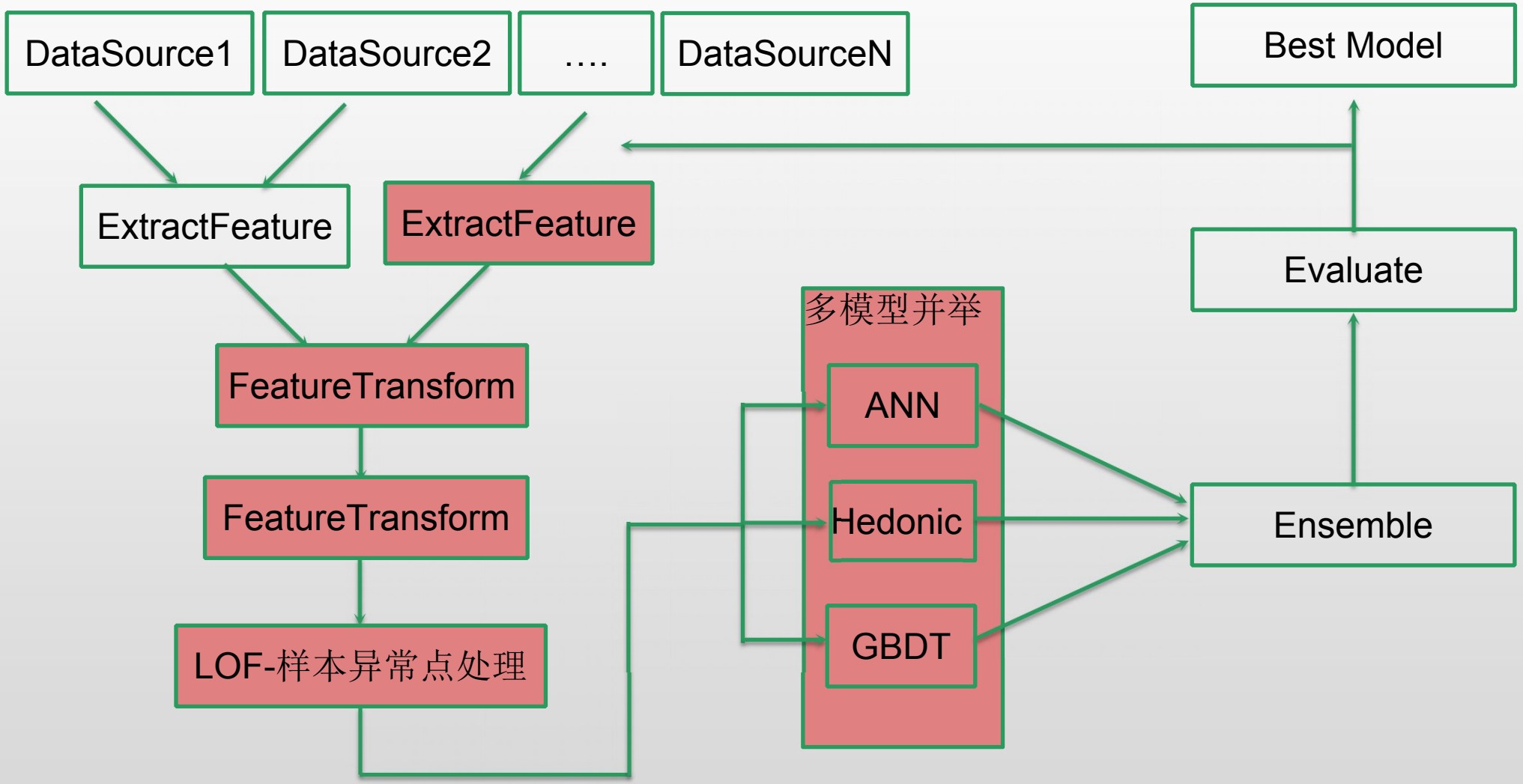
房屋估价


模型-绝对误差
绝对误差
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
4.11
4.12
4.13
4.14
4.15
4.16
4.17
日期

房屋估价


准确率
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
201503
201504
201505
201506
201507
201508
201509
201510
201511
201512
201601
201602
201603

房屋估价


DataSource1
DataSource2
….
DataSourceN
Best Model
ExtractFeature
ExtractFeature
Evaluate
多模型并举
FeatureTransform
ANN
FeatureTransform
Hedonic
Ensemble
LOF-样本异常点处理
GBDT

房屋估价

- LOF(Local Outlier Factor)算法是一种机遇密度的异常检测算法，
- 通过计算每个实例相对于其邻居的孤立情况来判断这个实例是否为离群点
- 为每一个每个实例计算一个异常分数，这个分数称为实例的局部离群因子（LOF）
- 较高的LOF值指示这个实例可能是异常的，较低的LOF值指示这个实例可能是正常的

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}$$

$$lrd_k(p) = \left[\frac{\sum_{o \in N_k(p)} reach - dist_k(p,o)}{|N_k(p)|} \right]^{-1}$$

提纲

- 蜀道难难于上青天
- 行困难而正确之事
- 往事可鉴未来可追

往事可鉴未来可追

链家金融

智能家电

链家装修

家政服务

亿万房产O2O服务平台打造住的入口

3D看房

VR看房

社区服务

海外置业


Thanks
LIANJIA.com 链家