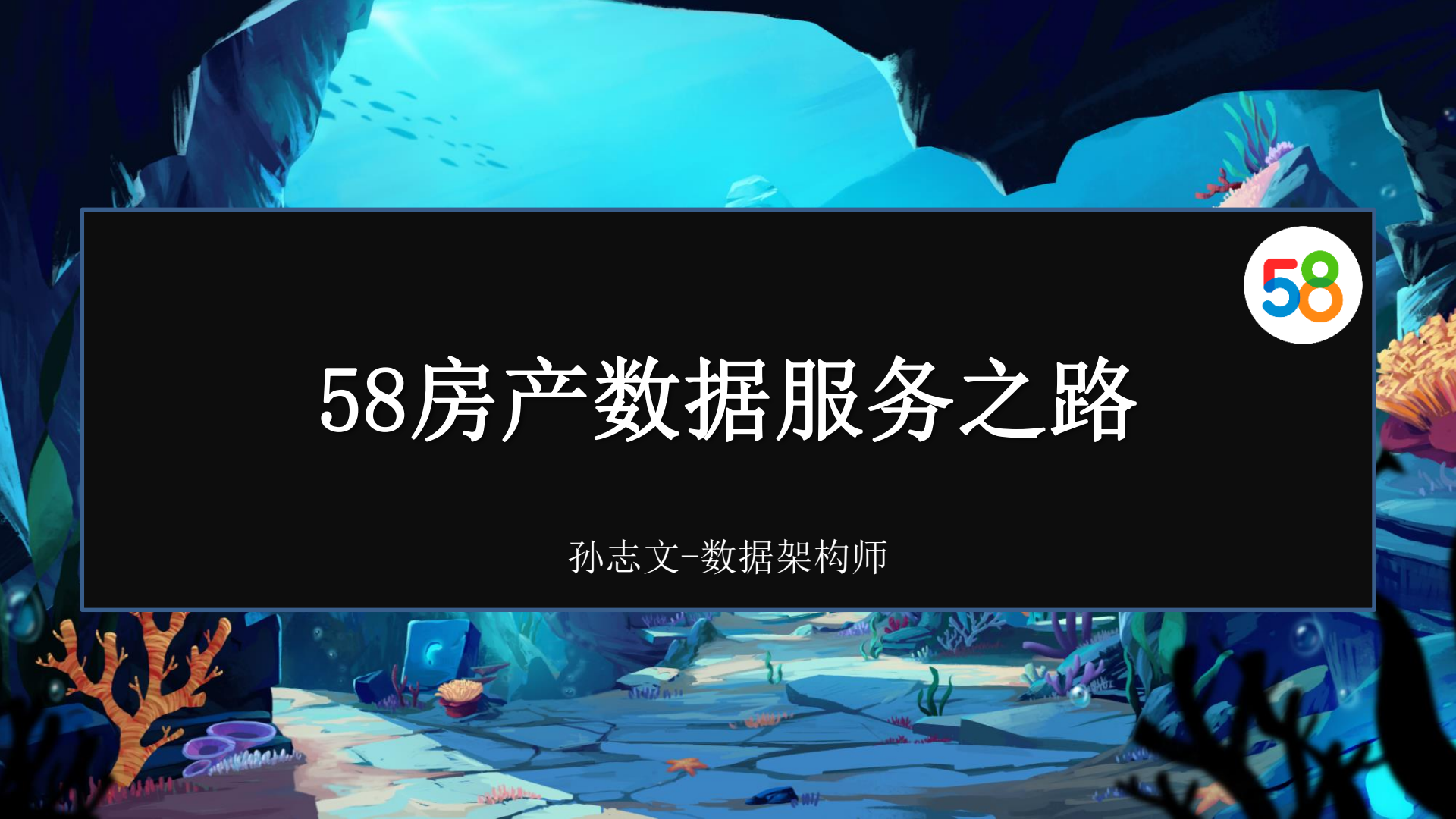

DataFunTalk
58同城大数据应用实践
2020.08.18-19



58房产数据服务之路

孙志文-数据架构师

目录

58房产数据服务的场景

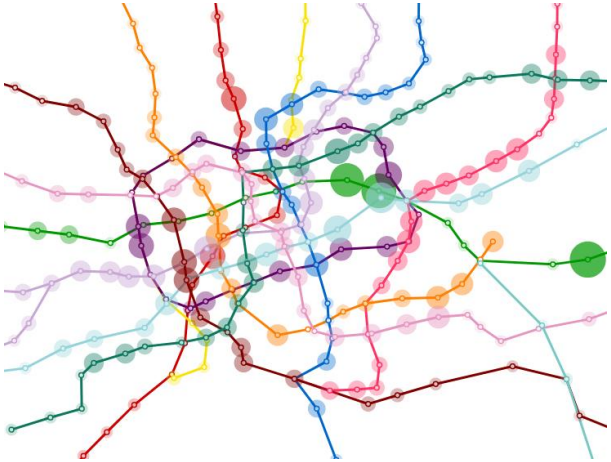
数据植入业务


17:15
中原地产中原地产金桥证大分行
3.9 综合评分
信息基础 4.3
服务水平 3.8
10个月 服务年限
9 服务人数
0 带看次数
网络考核
无忧交易·已加入无忧交易保障
服务范围
服务公司: 上海中原物业顾问有限公司
主营商圈: 浦东-金桥、浦东-高行、浦东-航头
主营小区: 东陆新村六街坊、幸福小镇明丰绿都(公...、仁恒森兰雅苑(二期)
佣金范围: ≤0.5%
增值服务: 专车接送 新房代购 置换服务
用户评价
好评率50%，2人评价
关注
微聊
电话

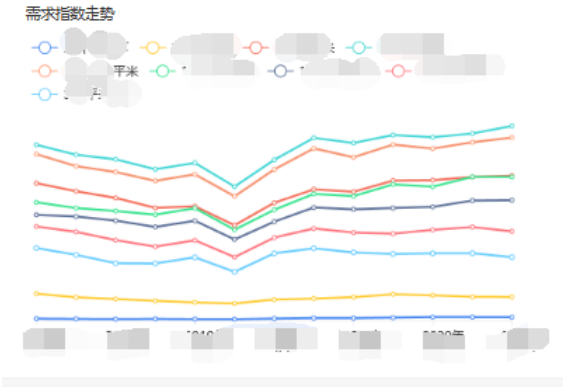

17:04
北极星
安选房源
抢手房源榜
根据浏览、咨询、搜索、成交量等多维度计算人气指数，每日更新
TOP1
提香别墅,,,7月13日新增,,,700平大花园,,,真真意义...
康桥 | 万科提香别墅
2500万 | 4室3厅 | 313m²
康桥人气小区榜
TOP2
地铁口精装大别墅南北通自住 免个税，低于小区其他...
康桥 | 绿宝园五期
2400万 | 4室3厅 | 379.07m²
TOP3
新挂牌临河边套 毛坯省心可随意设计装修 高品质别...
康桥 | 中邦城市艺墅
1700万 | 4室3厅 | 332m²

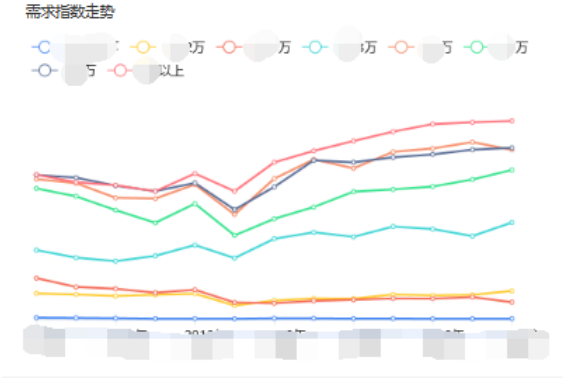

17:03
慧眼
二手房
51513元/m²
8月挂牌均价
0.21%
比上月
2.08%
比去年
新房
56862元/m²
7月成交均价
0.13%
比上月
5.2%
比去年
近30天有670914人估价
我要估价
二手房
新房
价格走势
近3月
近1年
近3年
上海均价(万/m²)
5.15万
5.14万
5.12万
5.11万
5.09万
06-04
06-18
07-02
07-16
07-30
08-13
上海二手房
51513元/m²
关注

数据业务化




9:41 AM
100%
最新动态
活跃客户数
今日
近7日
20
访问58安居客
40
访问微店
客户动态
安颖颖 132****6200 58安居客 新客户
5分钟前在 安居客浏览了房源
查看足迹
微聊
电话
安勤奋 138****8000 58安居客
5分钟前在 安居客浏览了房源
微聊
电话
安勤奋 138****8000 微店 新客户
12分钟前在 58浏览房源
复制昵称
微聊
电话


需求指数走势


需求指数走势

58房产数据架构

数据服务

- 数据化运营：运营报表、Deep dive、实时大屏、任我查、明细下载、销售决策
- 数据植入业务：房源Rank、用户画像、房源推荐、经纪人评分、楼盘榜单、房易通
- 数据业务化：楼盘洞察、公司参谋、经纪人参谋、锁客宝、门店参谋、房价地图

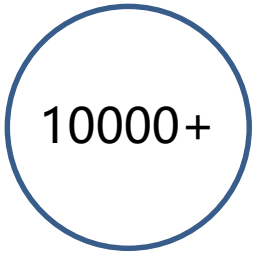
数据仓库

DWS	用户流量	收入	开发商	经纪人	市场渠道	销售行为
DWD	曝光	点击、连接	房源、楼盘	推广计划	会员	套餐、花费
ODS	浏览日志	微聊、电话	房源信息	经纪人信息	消耗明细	广告数据

数据平台

- 数据开发
- 元数据管理
- 数据监控
- 云窗
- 算法平台
- 实时计算平台
- KV数据服务
- Spark
- Hive
- Flink
- Tensorflow
- DB 数据交换
- KV 数据交换

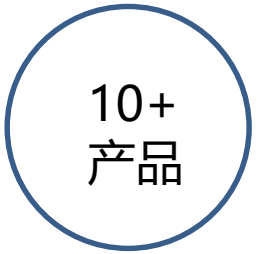
数据服务相关统计



任务数

数据化运营

数据植入业务



数据业务化

数据业务化带来的挑战

数据量增加	数据质量要求更高	时效要求更高
• DB 无法满足	• 质量问题导致客诉	• 8:00 点完成

通用KV数据服务

设计、流程改进

智能机器人

通用KV数据服务


查询
负载均衡
查询条件转换
权限检查
存储
HBase
WTable
导入
HFile 生成
Protobuf 生成
数据过滤
Key 转换
元数据
运维

通用KV数据服务


Hive hdp_anjuke_dw_db_da_ajk_avg_price_zf_room_monthly
集群
58 Hive
数据库
hdp_anjuke_dw_(
表
da_ajk_avg_price
字段 注释 上传 RowKey
1 city_id 城市id
2 comm_id 小区id
3 areacode 区域代码
4 room_id 几室
5 room_desc 几室描述
Hive Where:
month_id='$cal_dt'
例如：cal_dt='$cal_dt', 表示取当天数据
调度 ID:
DWMS/58DP 中调度任务 ID
Rowkey Rules
服务名:
hdp_anjuke_dw_db_da_ajk_avg_price_
KEY
1 comm_id 取模 MD5 无
2 room_id 范围

通用KV数据服务

基本信息

表名	hdp_anjuke_ana_db_si	状态	ONLINE
Hive #1	hdp_anjuke_ana_db.sai.	集群	58 Hive
用途	经纪人后台-三网推荐标签统一	负责人	

分客户端调用量

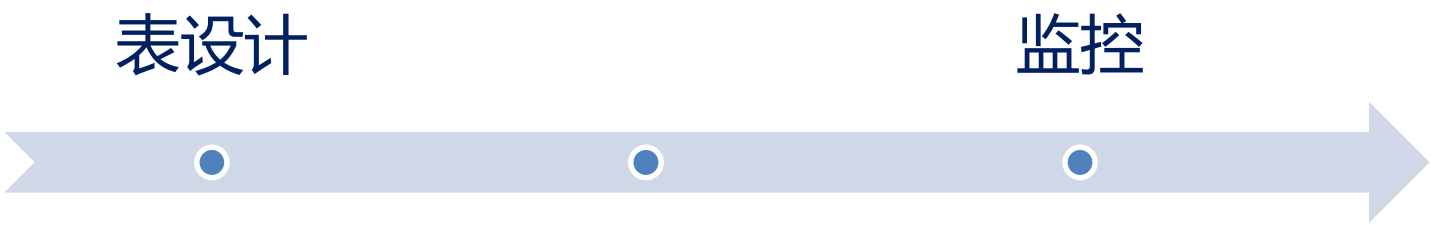
今天调用量

历史调用量

表结构

```
{
        "tableName": "hdp_anjuke_ana_db_
        "columns": {
```

设计、流程改进



1. 中间表设计
2. 应用层设计
3. 维表设计

1. 导出前质量检查
2. 设置数据信号

更全面的监控，提前发现问题

设计、流程改进

表设计

1. 中间层：针对内部的”大宽表”和对外服务的“小表”同时存在，
   中间层的解耦，确保对外数据流程的时效和稳定性
2. 服务层：根据业务特性（最新OR全部）决定是否加入分区键
   在满足业务需求的场景下，尽量减少存储压力
3. 维表： 是否直接使用线上业务的维表？谨慎处理维度退化

设计、流程改进

流程

1. 导出数据到线上服务时，先做质量检查；异常时，人工介入
2. 导出数据完成后，写数据信号，方便线上业务切换到新的数据

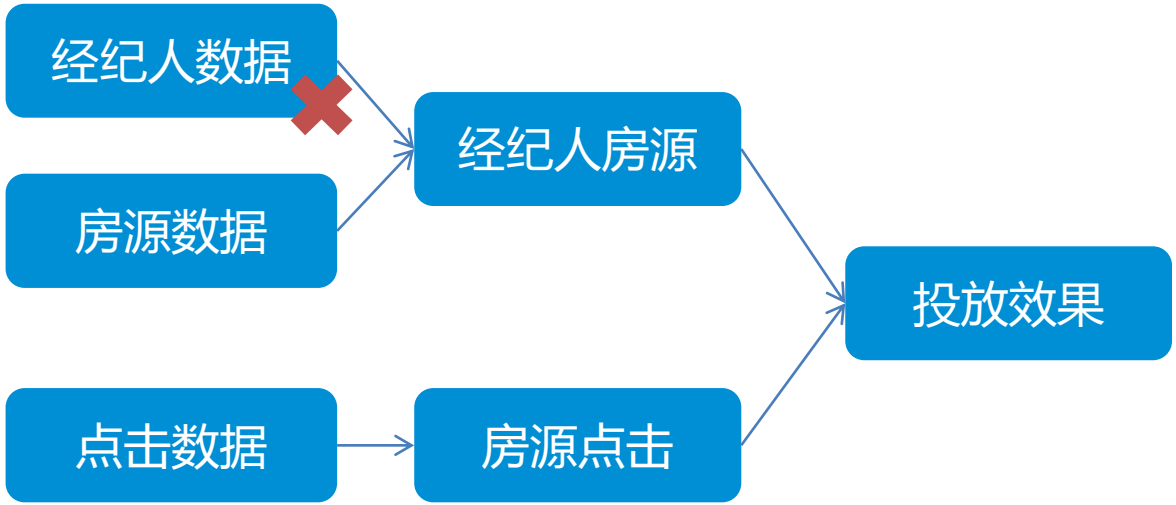
设计、流程改进

监控

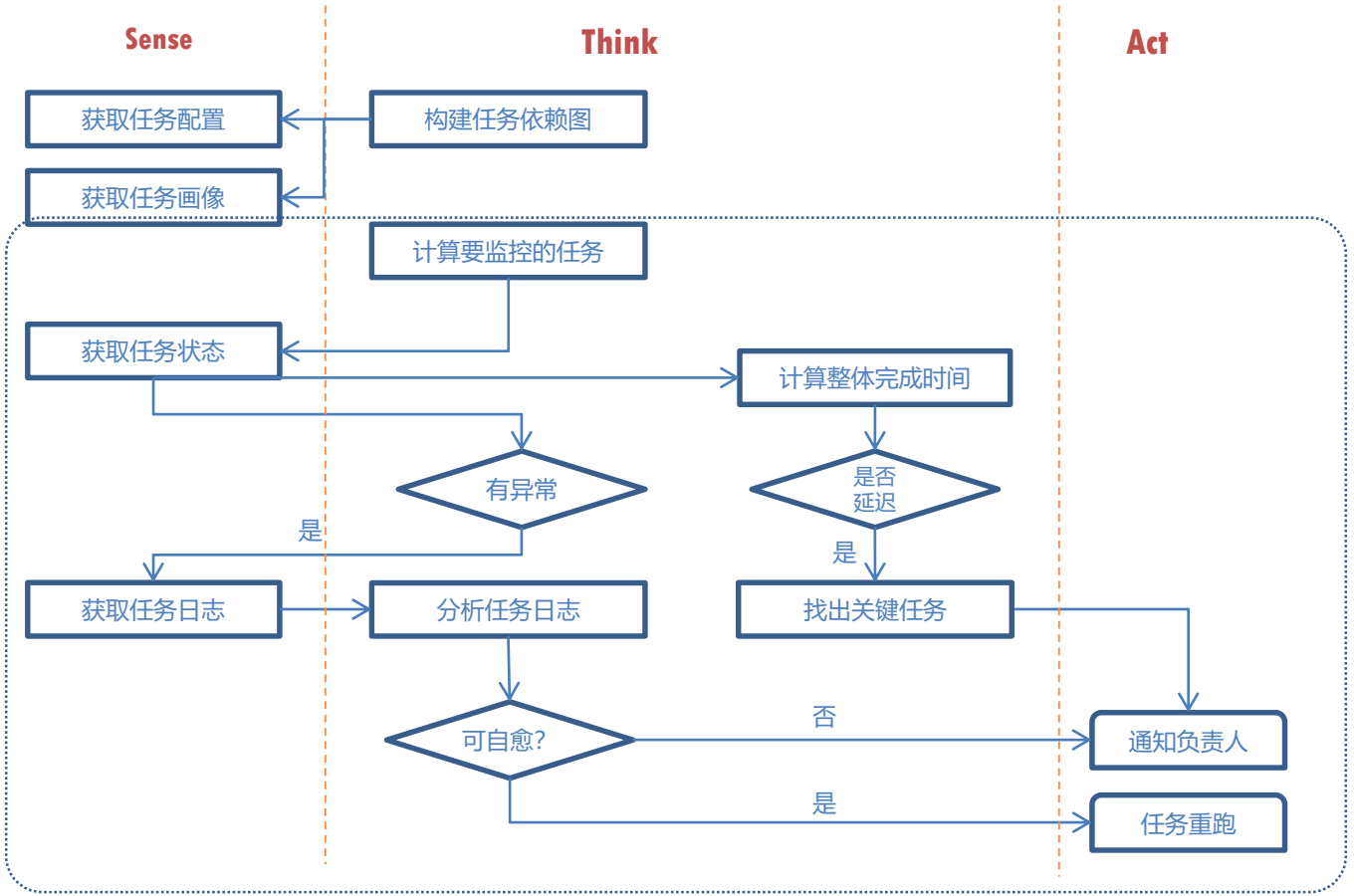
任务监控从点到面

单个任务监控方案的缺陷

1. 监控片面
2. 延迟发现晚
3. 任务诊断难



智能机器人架构

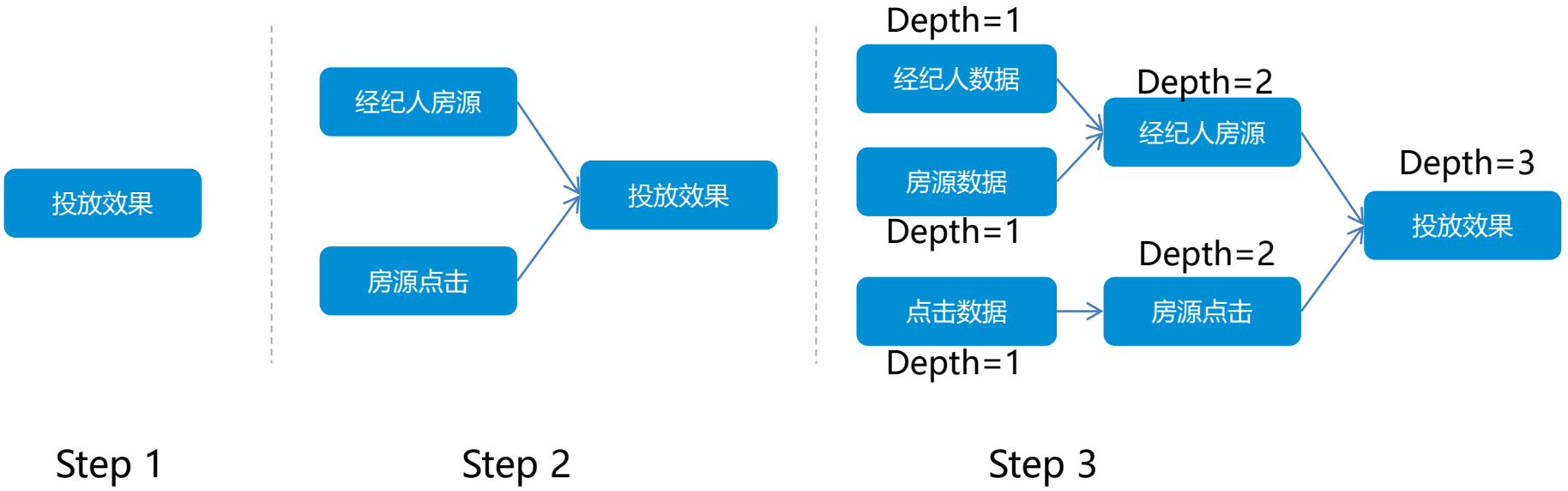

Sense
Think
Act
获取任务配置
构建任务依赖图
获取任务画像
计算要监控的任务
获取任务状态
计算整体完成时间
有异常
是否
延迟
是
是
获取任务日志
分析任务日志
找出关键任务
可自愈?
否
通知负责人
是
任务重跑

智能机器人

See	Act
• 任务依赖关系 • 任务运行状态和日志 • 任务画像 • 集群状态	• 任务重跑 • 通知负责人

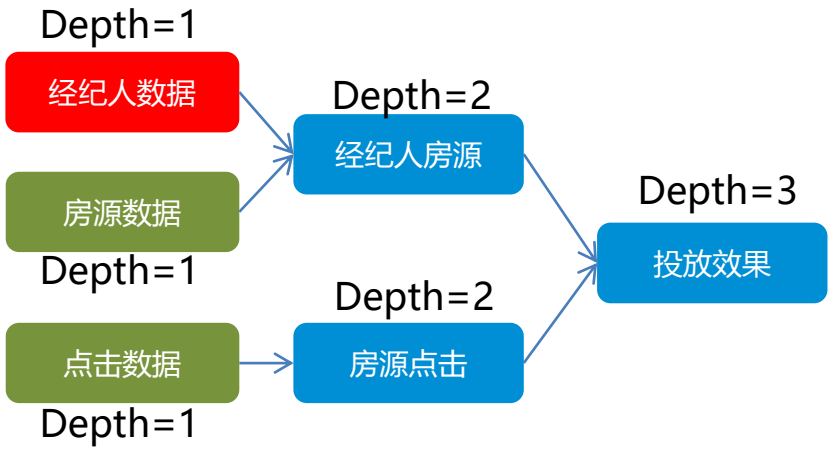
Think

1. 构建 job 依赖图
2. 设置 job 深度



Think

1. 监控当前深度下的未完成任务，若均完成，深度 + 1
2. 若发现异常任务，则采取对应动作



智能机器人

Think

通过日志等信息，将异常分类

不可重试

1. 语法错误
2. 相关配置错误
3. 缺少权限
4. Mapjoin 异常
5. 表不存在

可重试

1. 连接数过多
2. 读取数据超时
3. 拿不到锁

Think

3. 估算所有任务完成时间

4. 若最终完成时间超出设定阈值，找出慢的任务，并告警

$$End\ time = Max(start\ time + execute\ time, current\ time)$$

$$Start\ time = Max(schedule\ time, parent\ job\ end\ time, \ldots)$$

$$Execute\ time = Max(Median(Execute\ time), current\ execute\ time + R)$$

智能机器人效果

常规巡检

- 减轻负责人凌晨压力

全局可控

- 任何影响到核心流程的异常都会被监控

快速定位

- 快速定位异常点，更快恢复流程

70%

减轻执勤压力

98%

准确率

未来规划

1. 实时数仓
2. 数据治理


THANKS !