MongoDB
构建实时推荐系统

锦木信息
唐峰

锦木信息

官方合作伙伴，　　提供MongoDB订阅、咨询、技术支持服务。

Shanghai Jinmu Information Technology Co., Ltd.

Jinmu is a MongoDB reseller and services partner based on Shanghai. They provide NoSQL database, big data software, and consulting services solutions. Jinmu supports the full project lifecycle from initial design to production support. Industry focuses include financial services, transportation and retail. Their mission is to provide top-tier data solutions to customers.

主要客户

太平洋保险 CPIC

国泰君安证券 GUOTAI JUNAN SECURITIES

中国太平 CHINA TAIPING

CHINA FOREIGN EXCHANGE TRADE SYSTEM 中国外汇交易中心 全国银行间同业拆借中心 NATIONAL INTERBANK FUNDING CENTER

天弘基金 TIANHONG ASSET MANAGEMENT

MARY KAY

COSCO SHIPPING

中國東方航空 CHINA EASTERN

Ctrip 携程

议程安排

- **推荐系统介绍**
- 案例分享
- MongoDB

个性化推荐

内容推荐

- 用户行为
- 推荐内容
- 推荐商品

客户画像

- 客户标签
- 精准营销
- 产品分析

关系推荐

- 代理人推荐
- 社交关系
- 商户推荐

实时要求更高

从 **低效 静态**

非实时、静态数据

定时数据同步

数据手工聚合

批量数据处理

到 **实时 动态**

实时、动态数据

流式数据处理

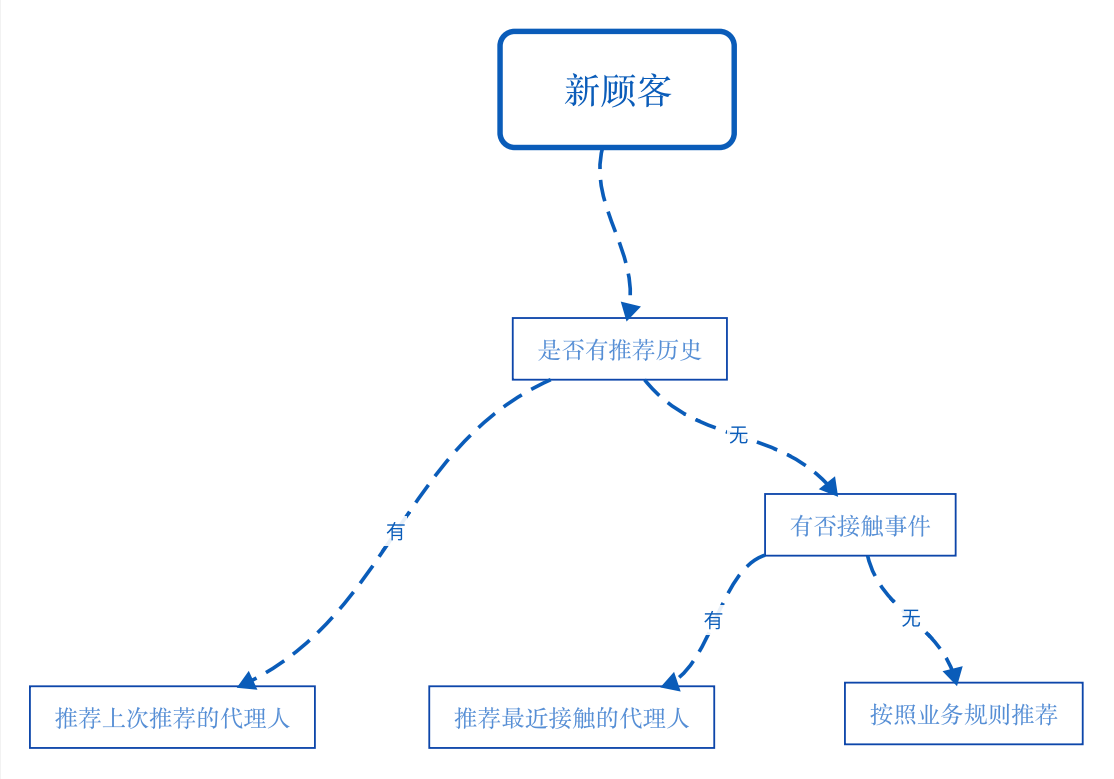
数据动态聚合

基于行为触发处理

议程安排

- 推荐系统介绍
- **案例分享**
- MongoDB

最初需求：新客户推荐



业务推荐规则

– 附近N公里（同城市）

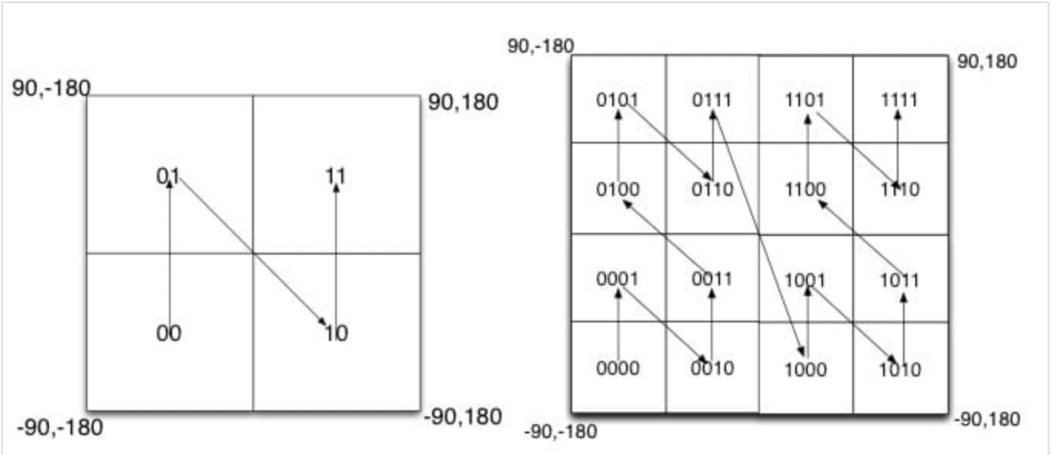
– 代理人（级别、年龄等）

– 推荐次数（升序）

地理位置查询

GeoHash

- 经纬度转化为字符串
- Hash值越长，越精确
- 无法得到距离、边界问题



MongoDB地理位置索引

```
{
    agentID: 0001,
    agentName: “Alice”,
    shipToAddress: {
        province: “Shandong”,
        city: “Qingdao”,
        formatAddress: “Shandong,Qingdao, street..”,
        loc : [-74, 40.74]
        },
    level: “senior”,
    recommendCount: 0,
}

db.agent.createIndex({
        ”shipToAddress.city" : 1,
        "recommendCount" : 1,
        "shipToaddress.loc" : "2dsphere"})

db.agent.find( {
  "shipToAddress.loc" : {
        $geoWithin:{$centerSphere: [[-88,30],10/6378.1]}},
  "shipToAddress.city" : “senior”}).sort({recommendCount :
1}).limit(1)
```

2dsphere&2d索引

- 支持地理位置包含、交叉、附件的查询
- 支持复合索引
- 可以计算距离

业务场景

- 同一个城市
- 10公里范围内
- 历史推荐次数最少的代理人

权重计算

接触事件:

- 关注
- 点赞
- 转发
- 拜访
- ……

```
{    _id: ObjectId(),
     customer: 0001,
     agent: "301671",
     eventType: "Follow|Share|Like|…",
     timeStamp: Date("2016/04/01 …")
}
```

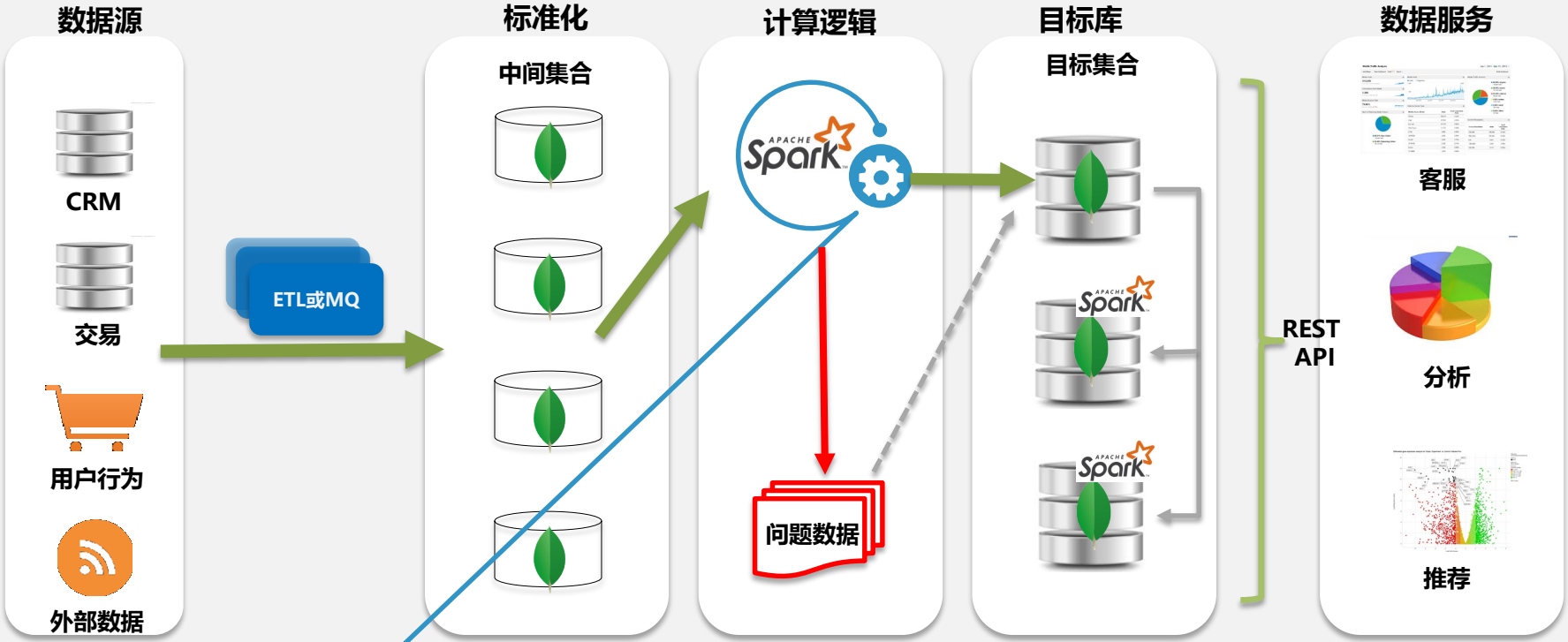
事件	权重
关注	0.1
点赞	0.5
转发	1
拜访	5

```
{    _id: ObjectId(),
     customer: 0001,
     agent: "301671",
     score: 80
}
```

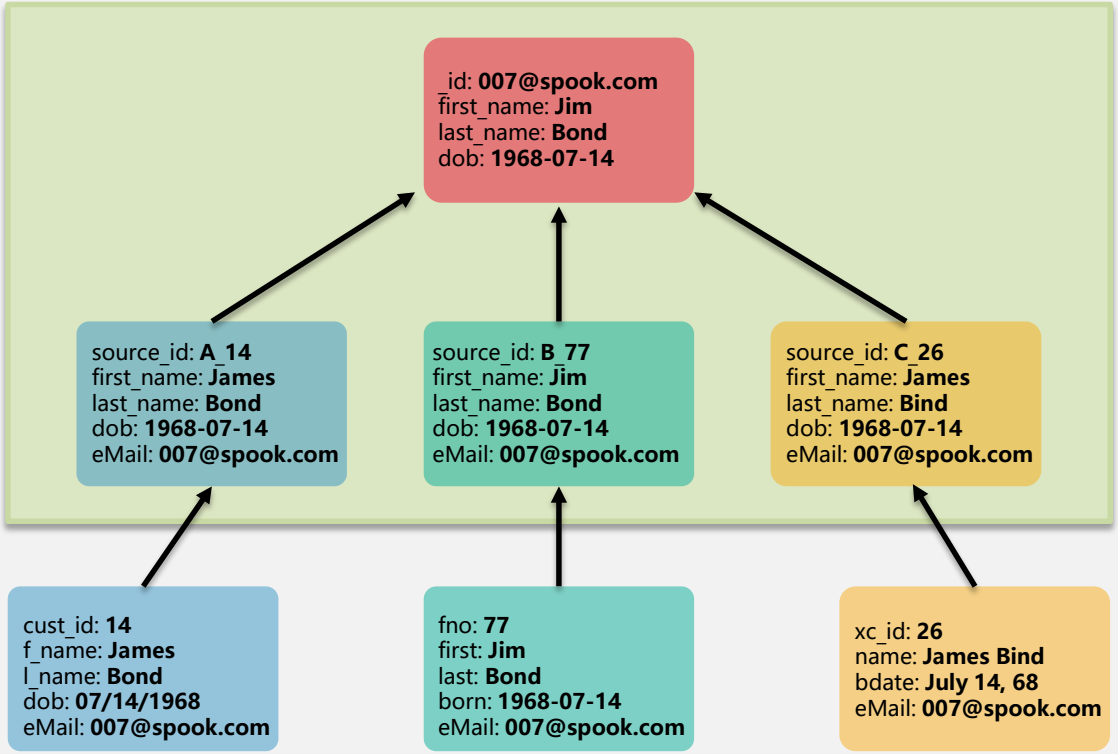
关系推荐

前台应用根据需要筛选并推荐

数据处理架构


数据源
CRM
交易
用户行为
外部数据
ETL或MQ
标准化
中间集合
计算逻辑
APACHE Spark
问题数据
目标库
目标集合
APACHE Spark
APACHE Spark
REST API
数据服务
客服
分析
推荐

数据集成（标准化）



目标表

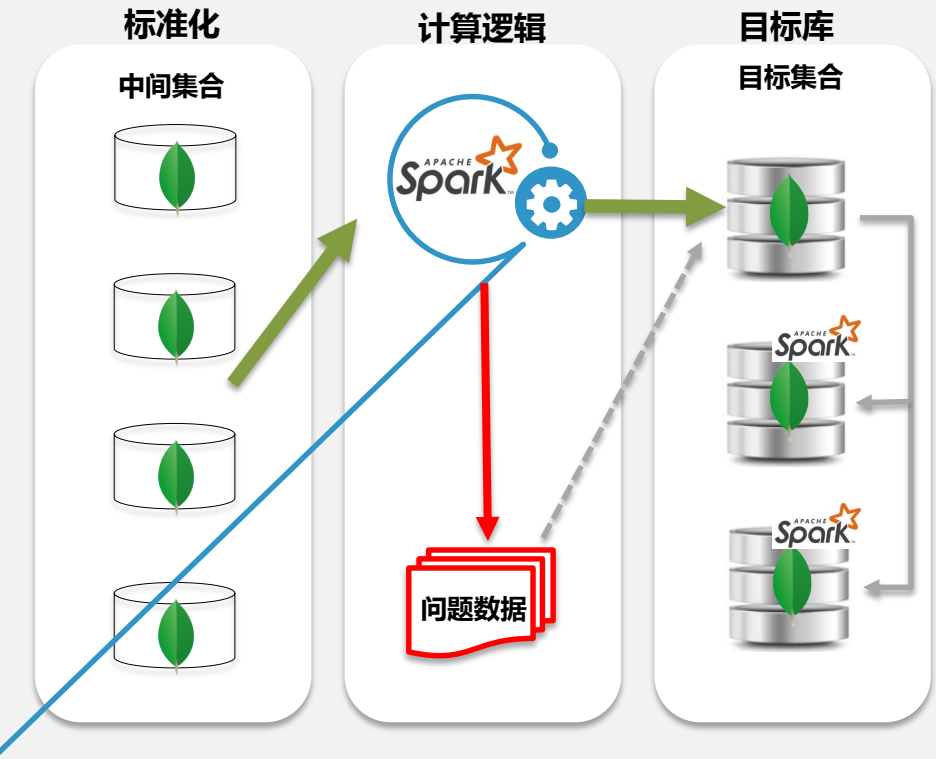
数据合并、测试、调和

中间表

统一字段名、数据类型

数据源

数据集成（计算逻辑）



业务规则

- 使用场景（需要哪些属性）
- 计算规则（权重、时间衰减等）

计算框架

- Spark（Streaming、 MLib、GraphX、 Spark SQL）
- AWS Lambda

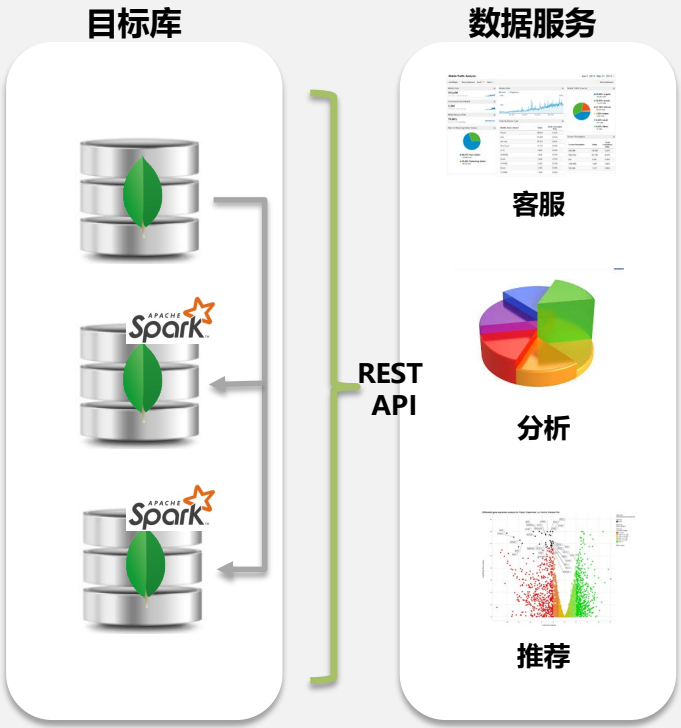
并发控制

- 多渠道并发数据冲突
- redis分布式锁

异常处理

- 数据时间、字段格式等
- 记入问题集合
- 定时任务或手工处理

数据服务



应用改造

–创建API来提供数据服务（例如：RESTful web service）

–重定向应用来使用web service

议程安排

- 推荐系统介绍
- 案例分享
- **MongoDB**

MongoDB优势

- JSON文档，面向对象，开发高效
- 数据结构变更无代价，便于程序快速迭代
- 弹性伸缩，大数据量处理无压力
- 直接与Spark对接

MongoDB注意点

2dsphere&2d索引

- 支持地理位置上包含、交叉、附近类型的查询
- 支持sparse属性（文档没有对应的field，不会更新索引）
- 如果集合里包含多个2dsphere、2d索引，在$geoNear中指定key；如果不指定，默认先使用第一个2d索引，没有再使用第一个2dsphere索引
- 不能是分片片键
- 联合索引，2d必须loc为前导字段，2dsphere则不需要

MongoDB注意点

Aggregation

– pipeline开始阶段，尽早使用$match过滤，并尽量使用索引过滤和排序

– $lookup的效率，**foreignField**不要忘记索引

– SERVER-7568： Aggregation framework favors non-blocking sorts

解决办法：在$match和$sort之间加$project

```
db.event.aggregate([
        {$match: {customerID: 123456}},
        {$sort: { eventTime: -1}},
        ……
        ])

#索引
db.event.createIndex({customerID:1, eventTime: -1})
#OK
db.event.createIndex({eventTime: -1})
#bad
```

更多问题?