elastic | 中国开发者大会

Elasticsearch在爱奇艺用户画像系统的应用

杜益凡
爱奇艺

Elasticsearch
在爱奇艺用户画像系统的应用

爱奇艺技术产品中心　　杜益凡

2018.11

爱奇艺

Elasticsearch在爱奇艺

爱奇艺BI

iQIYI 爱奇艺

Elasticsearch在爱奇艺的应用

爱奇艺BI

日志分析

Venus日志收集分析系统

作为数据库

视频、音频检索，AI业务等

搜索引擎

MySQLIO等

数据分析

脸谱、画像系统等

爱奇艺

用户画像

爱奇艺BI

用户行为数据抓取、规则统计、分析预测获得的**用户特征集合**。通常是将用户的社会属性、行为特征、潜在需求进行抽象提炼，形成的动态**用户标签体系**。

爱奇艺

用户画像的用途

爱奇艺BI

爱奇艺

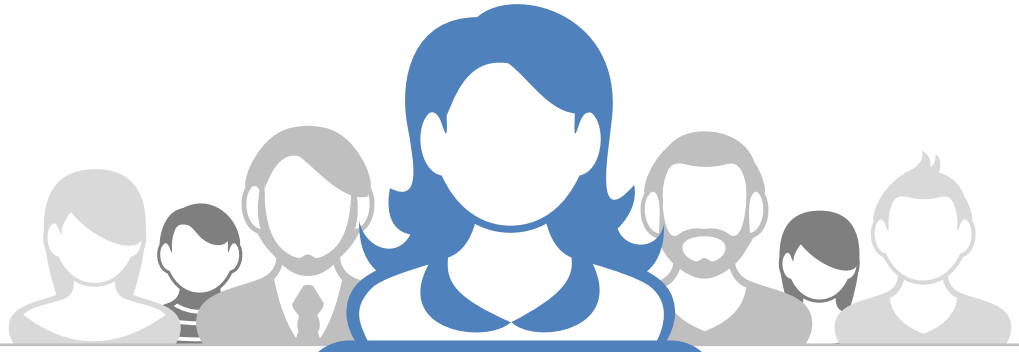
爱奇艺的用户画像产品矩阵

爱奇艺BI

画像数据层	**脸谱 – 用户标签体系库**	用户标签体系、标签质量管理、用户个人画像
画像服务层	**画像分析接口服务**	PC WEB 播放器 – 专辑流量分析 爱奇艺号 – 数据增值服务：自媒体受众画像 爱奇艺年度账单、偶像练习生粉丝画像、会员月刊 精准广告产品、广告DMP
	脸谱服务 – 精准广告人群	
	脸谱服务 – 个性化推荐	奇异果 长视频推荐
	脸谱服务 – RPC服务	商城商品推荐、奇秀主播推荐、爱奇艺文学…
画像应用层	**达芬奇-爱奇艺画像系统**	定制化人群、定制化画像分析
	后羿 – 精准营销系统	定制化人群、快速画像预览服务
	运营分析系统	爱奇艺指数、运营系统、内容消费指数

爱奇艺用户标签体系结构

爱奇艺BI



用户画像

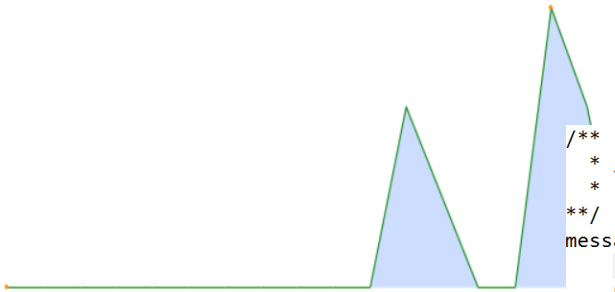
自然属性	设备属性	身份属性	行为偏好
年龄	手机品牌	用户类型	频道偏好
人生阶段	网络情况	会员状态	剧集偏好
性别	手机系统	新用户	明星偏好
.....	.....	.....	.....

iQIYI 爱奇艺

爱奇艺用户标签数据结构

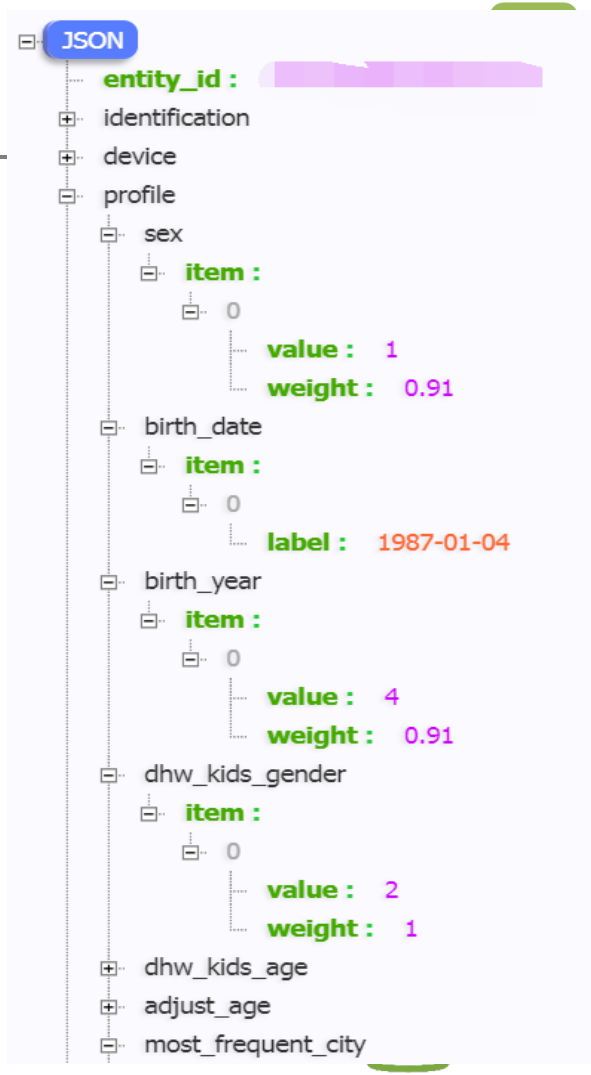
♡ 行为偏好

- 用户忠实度 : 7.50%
- 剧集偏好 : 画皮 、 奇葩说第5季 、 汪汪队立大功 第4季 、 爱情公寓 、 好戏一出 、 梦想改造家第5季 、 凉生我们可不可以不忧伤
- 明星偏好 : 马东 、 高晓松 、 蔡康永 、 李诞 、 薛兆丰 、 papi酱 、 施琰 、 骆新
- 视频tag偏好 : 爱奇艺出品(综艺) 、 奇葩说(综艺) 、 脱口秀(综艺) 、 马东(综艺) 、 高晓松(综艺) 、 蔡康永(综艺) 、 爱情(电影) 、 华语(电影) 、 7-10岁(少儿) 、 内地(电影) 、 ...
- 频道偏好 : 综艺 、 电影 、 少儿 、 片花 、 娱乐
- 观影小时偏好 (General) : 15 、 16 、 23 、 11 、 12



- 观影小时偏好 (工作日) : 15 、 23

```
/**
  * 脸谱schema
  *
**/
message UserModel{
    //脸谱ID
    required int64          entity_id      =  1 ;
    //身份信息
    optional Identification identification =  2 ;
    //设备信息
    repeated Device         device         =  3 ;
    //个人自然属性;
    optional Profile        profile        =  4 ;
    //个人行为信息
    optional Preference     preference     =  5 ;
}
```



爱奇艺用户标签的生产

爱奇艺BI

各产品的用户行为

用户标签

爱奇艺

爱奇艺的用户画像数据流

爱奇艺BI

用户数据 → 标签体系 → 画像系统

爱奇艺及全网海量用户行为数据

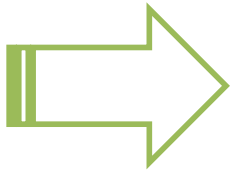
6亿+ 月活用户行为数据
观影、直播、奇搜、会员
文学、ACG、商城 ...

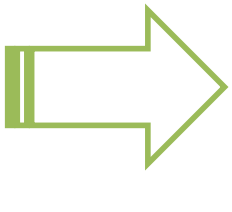
数据挖掘和建模，提炼用户标签

近30亿用户设备，**近3亿** 用户账号
5TB+特征数据，**250维+** 用户标签
涵盖自然属性、设备特征、兴趣偏好及
21个业务线行为标签

基于用户标签，输出定制化画像

支持50+维特征的定制化分析
支持10+垂线业务的交叉画像





爱奇艺的用户标签体系库——脸谱

爱奇艺BI

近30亿
用户设备

近3亿
用户账号

250维+
用户标签

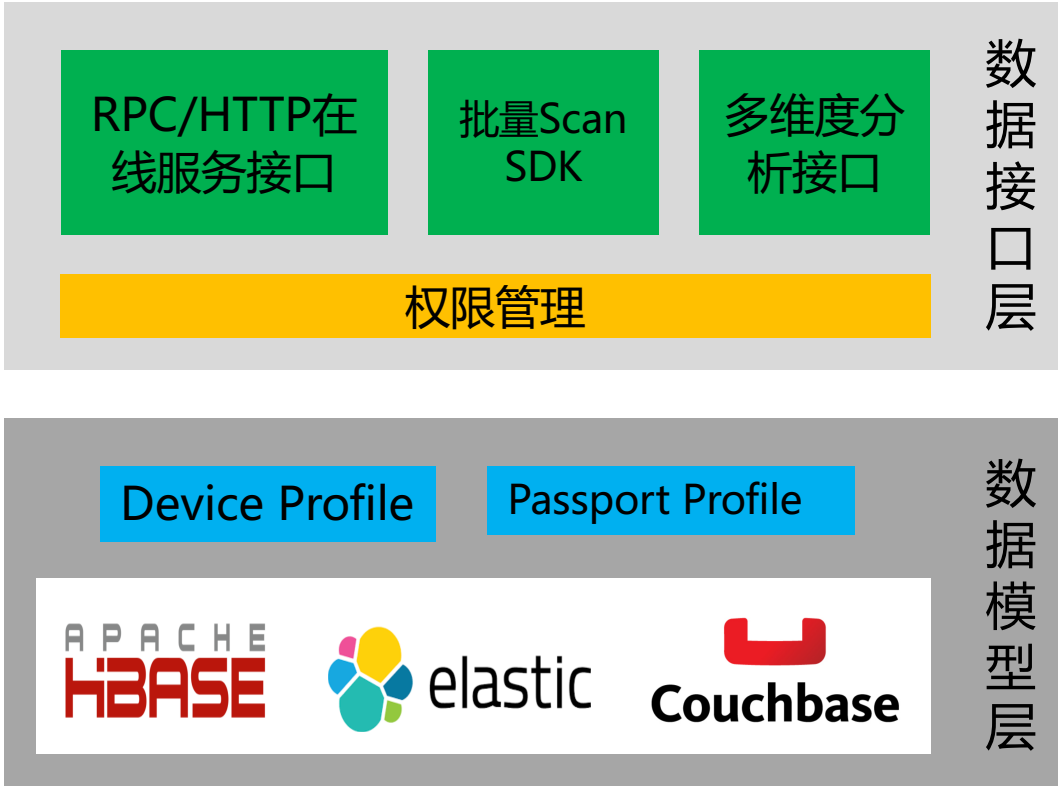
21类
业务线标签

爱奇艺

脸谱系统的架构

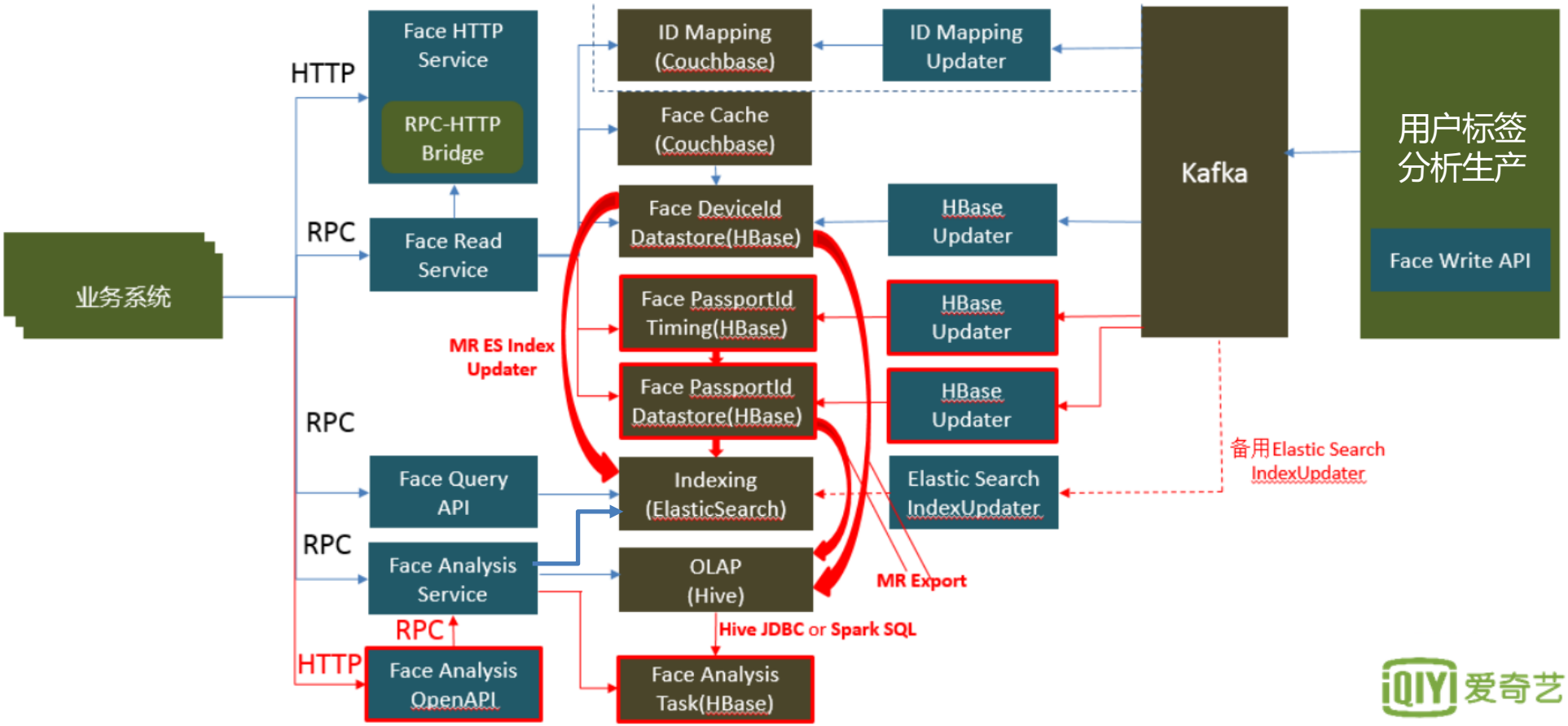
爱奇艺BI

- 数据接口层
  - RPC/HTTP在线服务接口提供毫秒级接口服务用于支持个性化等在线服务
  - 批量Scan接口用于批量获取用户特征数据来进行用户画像
  - 多维度分析接口主要用于人群估量和简单的分布统计
- 数据模型层
  - 设备维度&Passport维度
  - 存储基于Hbase和Elasitcsearch
  - ES用于服务常规业务
Hbase服务底层大数据量分析
  - 使用Couchbase作为实时数据缓存



爱奇艺

用户标签数据写入脸谱


Face HTTP Service
RPC-HTTP Bridge
HTTP
RPC
Face Read Service
业务系统
RPC
Face Query API
RPC
Face Analysis Service
RPC
HTTP
Face Analysis OpenAPI
ID Mapping (Couchbase)
ID Mapping Updater
Face Cache (Couchbase)
Face DeviceId Datastore(HBase)
HBase Updater
Face PassportId Timing(HBase)
HBase Updater
MR ES Index Updater
Face PassportId Datastore(HBase)
HBase Updater
Indexing (ElasticSearch)
Elastic Search IndexUpdater
OLAP (Hive)
MR Export
Hive JDBC or Spark SQL
Face Analysis Task(HBase)
Kafka
用户标签分析生产
Face Write API
备用Elastic Search IndexUpdater

爱奇艺用户画像系统——达芬奇

爱奇艺BI

- 基于爱奇艺全站用户行为数据、脸谱-用户标签体系的**通用人群画像分析平台**，为全公司各业务线提供**极速、精准、定制化**的画像分析服务

爱奇艺

达芬奇画像解决的问题

爱奇艺BI

2线城市女白领们在关心什么？

喜欢看《延禧攻略》的用户特征是什么

喜欢看《中国新说唱》的男性还看什么综艺

奇秀与漫画的用户重合情况

搜吴亦凡的更容易买VIP么？

《奇葩说》第四季的用户喜欢什么商品？

年卡半价吸引的这波会员都有什么特点？

华为和小米用户有什么不同？

纳逗APP的竞品都有谁？

爱奇艺

达芬奇系统的核心功能

爱奇艺BI

配置
定制化
画像人群

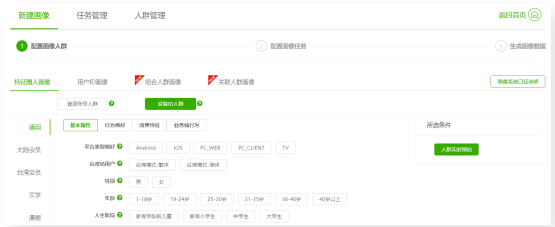
特征人群　组合人群
用户ID人群　关联人群

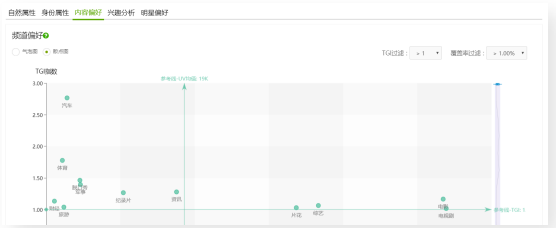
配置
定制化
画像任务

启动时间　输出方式
运行周期　数据范围

输出
可视化
画像数据

特征占比　特征TGI
特征分布





爱奇艺

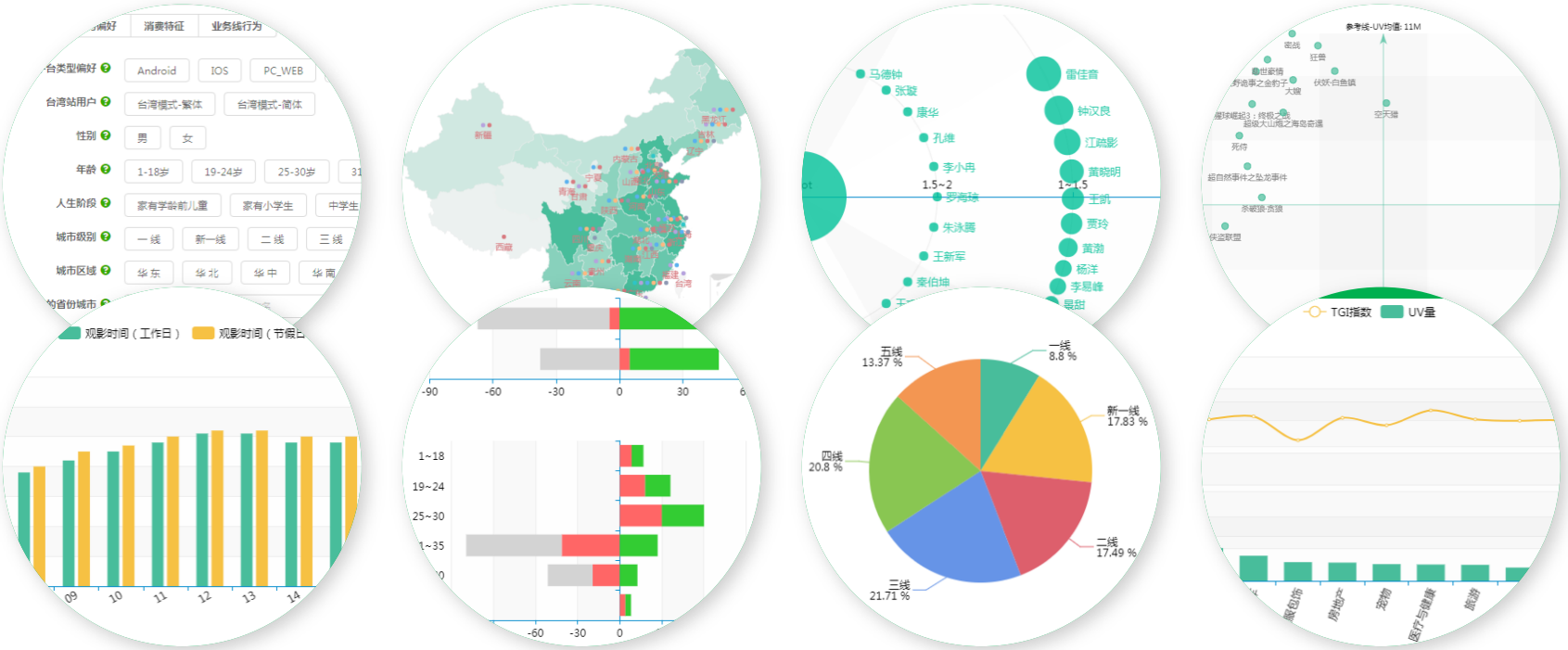
达芬奇系统的核心功能

爱奇艺BI

爱奇艺

达芬奇系统的画像数据结果

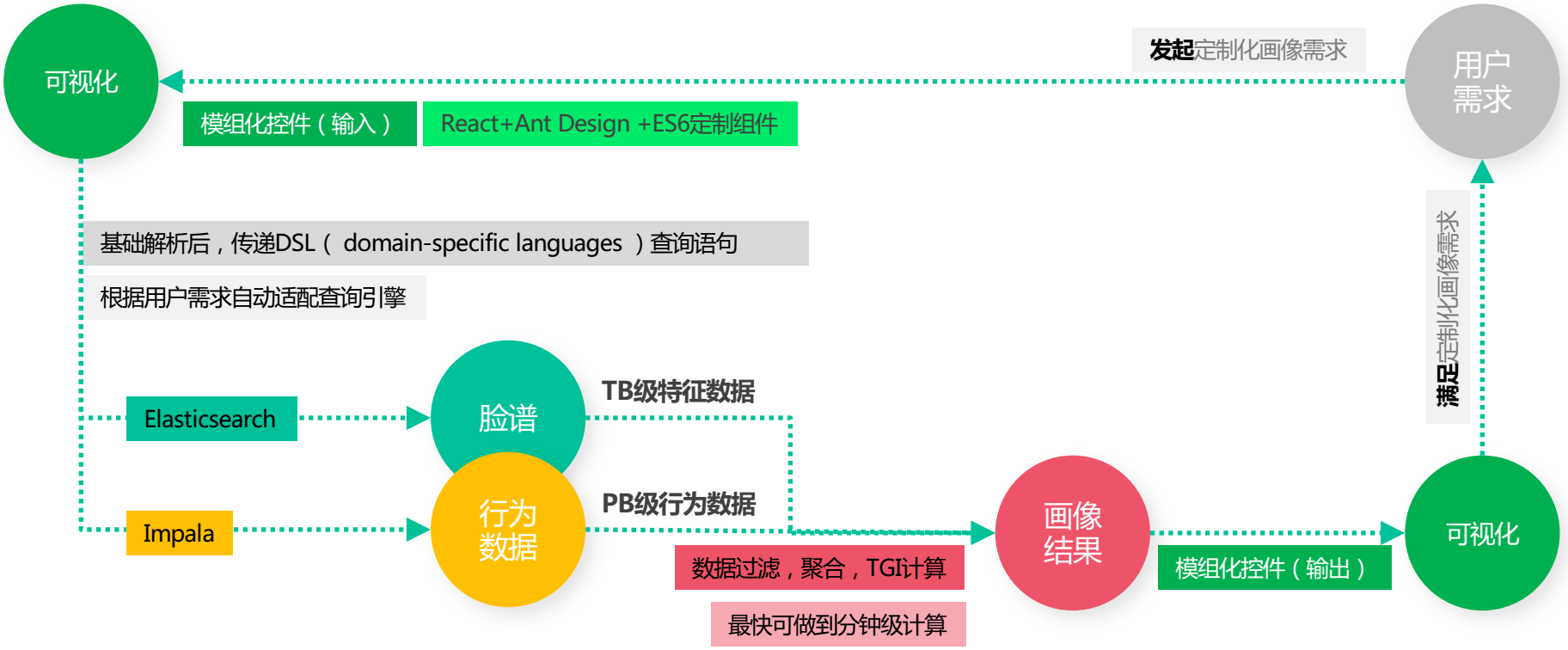
爱奇艺BI



爱奇艺

达芬奇系统的技术逻辑

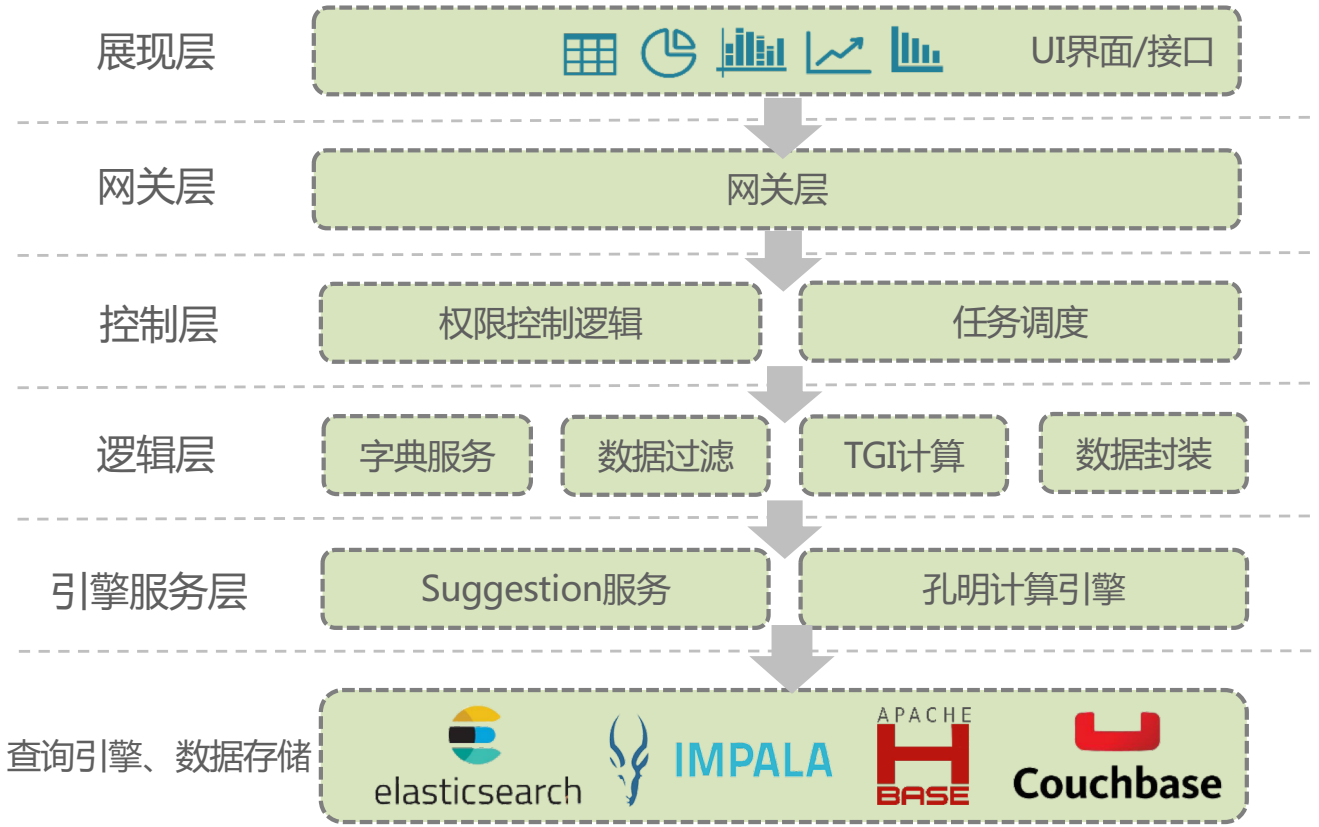
爱奇艺BI



爱奇艺

达芬奇系统的技术架构

爱奇艺BI

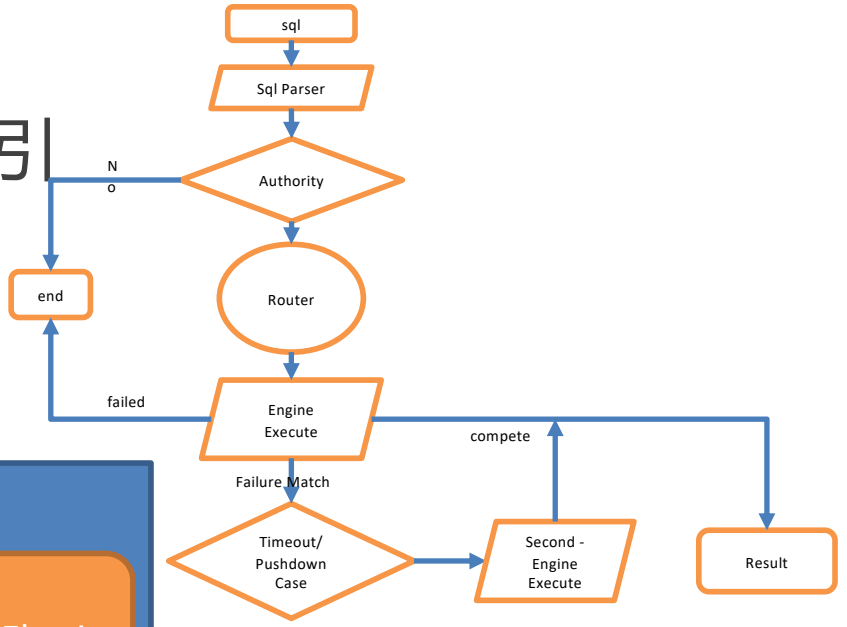


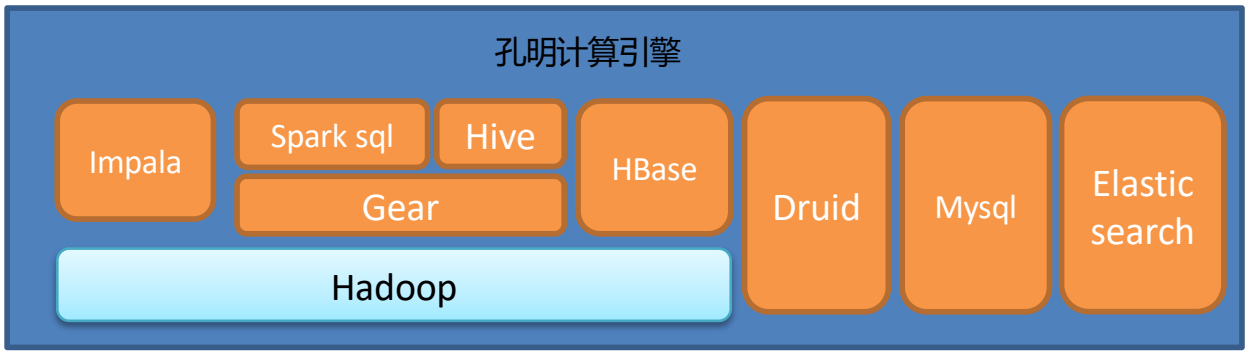
iQIYI 爱奇艺

孔明计算引擎

爱奇艺BI

- 统一查询接口的计算引擎
- 智能路由，选择最合适的实际执行引擎
- 智能下沉，确保查询可靠性





爱奇艺

Elasticsearch在达芬奇系统中的作用

爱奇艺BI

存储

存储特征标签并建立索引,来进行快速检索和Suggestion

分析

使用检索统计功能对人群进行快速分析

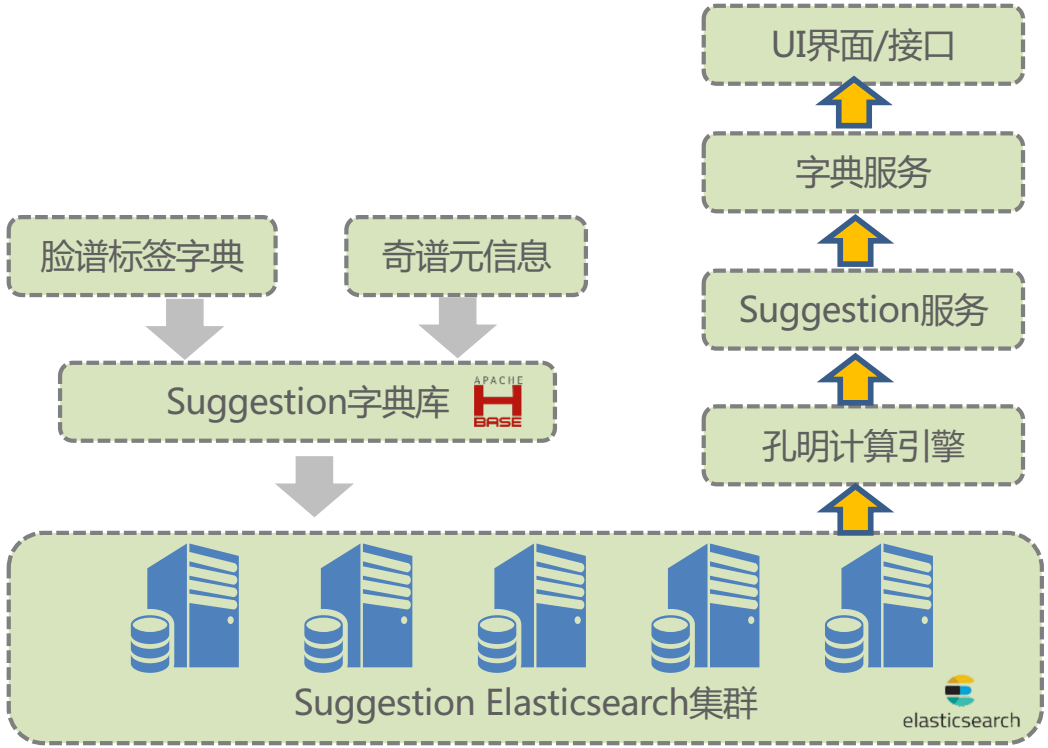
圈定

使用聚合功能根据条件圈定人群

爱奇艺

Suggestion服务

爱奇艺BI



爱奇艺

人群圈定与分析

基本属性：

X 性别　　男

行为偏好：

X 视频偏好　　偏好程度：较高，西虹市首富(电影)

X 明星偏好　　偏好程度：极高，沈腾

人群实时预估

预估UV：

```
// build statement
FaceTermBuilder faceTermBuilder = FaceTermBuilder.newBuilder()
        .feature(UserModel.PROFILE_FIELD_NUMBER)
        .field(Profile.BIRTH_YEAR_FIELD_NUMBER).value(1).build(); // 年龄在0-18岁
FaceTermBuilder faceTermBuilder2 = FaceTermBuilder.newBuilder()
        .feature(UserModel.PROFILE_FIELD_NUMBER)
        .field(Profile.INCOME_FIELD_NUMBER).value(6).build(); // 收入在5001-8000
FaceBoolBuilder faceBoolBuilder = FaceBoolBuilder.newBuilder()
        .must(faceTermBuilder).must(faceTermBuilder2).build(); // and
// 关系，还支持or(should)以及not(mustNot)
FieldAnalysisBuilder fieldAnalysisBuilder = FieldAnalysisBuilder
        .newBuilder().feature(UserModel.PROFILE_FIELD_NUMBER)
        .field(Profile.SEX_FIELD_NUMBER).build(); // 相当于group by 性别
FaceAnalysisBuilder faceTermAnalysisBuilder = FaceTermAnalysisBuilder
        .newBuilder().faceBoolBuilder(faceBoolBuilder)
        .addField(fieldAnalysisBuilder).build(); // 相当于 select sex,
// count(entity_id) as
// count_id from
// qipu_face_table
// where birth_year=1
// and income=6 group
// by sex order by
// count_id;
// RPC request and response
FaceESAnalysisService faceAnalysisService = FaceESAnalysisService
        .newBuilder().dataCenter(DataCenter.JYLT).username("abc")
        .password("abcde").timeoutMs(10000) // 根据复杂程度设置超时时间
        .build();
```

爱奇艺

画像系统所使用的Elasticsearch集群

- Suggestion服务集群+画像服务集群
- 5台+8台物理机
- 系统：CentOS6.4
- 内存：192G
- CPU：E5-2650 V3
- 硬盘：900G SSD
- JAVA版本：1.8
- ES版本：2.3
- 每日数据量：50T

爱奇艺

一些应用实践

- 优先使用Java8，客户端和服务端使用相同版本JVM
- 需要创建大量连接就使用传输客户端
- 只是少数持久的对象连接到集群，客户端节点可以更高效，但是会造成耦合
- 线程池的线程个数根据CPU核数设置，最多两倍，再多也浪费
- 主分片数设置为data节点数的1倍
- 能用short就不用long
- 读写分离场景关闭sniffer


iQIYI
爱奇艺

elastic
中文社区

专业、垂直、纯粹的 Elastic 开源技术交流社区
https://elasticsearch.cn/

elastic