用户画像及其应用

项目规划说明书

xxxx 年 xx 月

xxx 大数据部门

文档记录：

摘要记录：

关键字	
编　号	用户画像及应用项目-设计说明书
关　联	

变更记录：

变更者	版本	变更日期	批准人	审批日期	变更内容
Xxx	V1.0	2016-10-10			初稿

1. 引言

1.1　项目名称

xxx 用户画像及其应用.

1.2　项目背景及概要

在互联网逐步步入大数据时代后，不可避免的给企业和用户行为带来一系列改变与重塑；其中最大的变化莫过于，用户的一切行为在企业面前是“可视化”的.随着大数据技术的深入研究与应用，企业的专注点日益聚焦于怎样利用大数据来为精细化运营及精准营销服务，进而深入挖掘潜在的商业价值.于是，用户画像的概念也就应运而生.

用户画像可以使产品的服务对象更加聚焦，更加的专注.本项目分别从用户人口属性、订单消费、行为属性、用户偏好、疾病问诊信息、客户满意度六个角度构建用户画像模型；基于PG（关系型数据库）和大数据平台采集分析，分别从用户类别、渠道内容、行为特征及业务场景等多方面进行数据标签配置，实现模型与应用场景数据共享，采用千人千面等方法进行UI数据可视化展现，实现精细化运营及精确营销服务.

1.3　项目目标

全业务运营下，用户画像及应用基于 PG（关系型数据库）和大数据平台采集分析，把用户特征标签封装成数据接口服务，实时推送到一线，使信息数据变成生产力，项目实现目标如下：

一、用户画像模型封装

（1）基于 PG（关系型数据库）和大数据平台（hive、impala）

包含基础标签与分析类知识标签，实现用户特征全貌刻画；

（2）多种封装角度

分用户类别、渠道内容、业务场景进行封装配置.

二、接口数据实时推送

实现用户画像数据实时更新至运营及营销统一视图（WeMeta、WeData、

WeSearch 等）中进行展现，并实时反馈运营及营销信息问题，保证数据应用的时效性.

三、展现 UI 封装

依托用户画像，将推荐信息配置应用端进行可视化展现，集中活动运营，实现千人千面的运营效果.

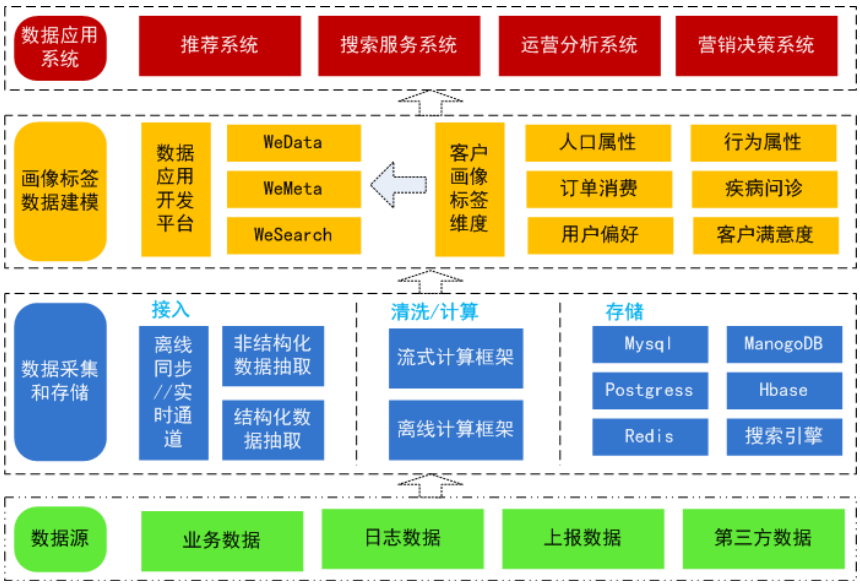
1.4 项目适用范围

- 运营决策人员：对运营的关健问题进行决策.
- 运营分析人员：从事市场竞争分析、用户需求分析、业务分析工作，主要负责用户需求的发现和目标确定，并配合运营策划和评估的实施.
- 运营策划人员：从事运营和实施方案设计，根据用户需求生成创意，将创意转化为策略，并制订实施方案.
- 数据分析人员：负责数据挖掘和数据分析支撑的全体IT支撑人员.
- 其他开发人员.

2. 系统功能及模型架构

2.1 系统功能架构

用户画像及应用项目包括底层数据源采集和存储、画像标签模型构建、数据模型应用三个层级，系统功能架构如下：



2.2 模型架构

画像标签模型分析主要分原始数据统计分析、统计标签建模分析、模型标签预测分析三块，具体如下：

3. 需求设计

3.1 用户画像模型

[需求说明]：用户画像模型是结合用户基本属性分析，对互联网行为特征进行描述，包括用户登录、搜索、关注、消费等各方面数据，对用户的疾病问诊、行为喜好变化、消费订单等全过程的记录，以标签方式展示每个用户的个性化特征，画像是系统分析结果的总结，是系统数据挖掘的起始.

[业务要素]：用户画像模型按照数据内容模块分为：用户人口属性、行为属性、资产消费、疾病问诊、用户偏好、客户满意度六大类标签.-----后续需要新增一些活动、业务类的标签；

[核心算法描述]：核心算法包括聚类分析、分类算法、时间序列分析、RFM模型、推荐系统算法、关联分析等.

1） 聚类分析

聚类分析将看似无序的对象进行分组、归类，以达到更好地理解研究对象的目的. 聚类结果要求组内对象相似性较高，组间对象相似性较低. 在用户研究中，很多问题可以借助聚类分析来解决，比如用户活跃度行为聚类、用户消费

情况聚类等.

2) 分类算法

分类是按照某种标准给用户贴标签，再根据标签来区分归类，分类是事先定义好类别，类别数不变.根据用户群的文化观念，订单消费、行为习惯等不同细分新的类别，企业根据用户的不同制定品牌推广战略和营销策略，将资源针对目标用户集中使用.

3) 时间序列分析

时间序列分析是一种动态的数据统计方法.该方法基于随机过程理论和数理统计学方法，研究随机数据序列所遵从的统计规律，以用于解决实际问题.比如用户的周期性行为分析、因子回归分析建模等.

4) RFM 模型

RFM 模型较为动态地显示一个用户的全部轮廓，R 表示用户购买的时间有多远，F 表示用户在时间内购买的次数，M 表示用户在时间内购买的金额，加权得到 RFM 得分.

5) 推荐系统算法

利用用户的一些行为，通过一些算法（协同过滤、LFM、打分模型、关联分析等），推测出用户可能喜欢的东西.推荐讲究准确性，提高用户-医生（医院）-内容（订单、知识等）等组合的匹配度，提升服务质量.

6) 关联分析

关联分析就是在关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构，挖掘潜在的行为和消费关联特征.

3.1.1 人口属性标签

[需求说明]：人口属性标签是用户的基本信息，这些信息往往是用户注册及使用产品时记录的信息，例如：年龄、性别、注册时间、婚姻状况、身高体重等.通过人口属性刻画，到达对用户初步认知的目的.

[业务要素]：人口属性标签大部分可从数据仓库直接获取，部分数据（生理）可在体检、疾病处方等非结构数据中进行提炼.

标签所属分类	标签名	标签解释	标签 eg
人口属性标签	性别	身份证标识的性别	1 男 2 女 其他未知
	年龄/分层	对平台用户年龄进行分群分析	新生婴儿 0~28 天 婴儿（28 天~1 年）、 幼儿（1~4 年）、 儿童（5~13） 少年（14~18）、 青年（19~44）、 中年（45~59）、 老年（60 以上）
	电话号码所在区域/分层	一二三四线城市 or 城乡标识	城市/农村
	是否临时账户	为第三方账号登录，没有进行过账号验证绑定的账号	是/否
	注册时间	用户的注册日期，格式 yyyy-mm-dd	2016/9/25 11:10
	新老用户标识	基于用户注册时间及订单业务情况建模分析	新/老用户
	教育程度	用户的学历信息	研究生/本科/大专/高中/初中及以下
	身高	健康档案中的个人信息	175cm
	体重	健康档案中的个人信息	65kg
	职业类型	用户从事职业的分类	政企机关/白领/销售等
	收入水平	根据各因子模型预测用户收入等级	高/中/低
	星座	根据用户生日进行分群	白羊/金牛/双子/狮子等 12 星座
	婚姻状况	用户结婚与否	是/否
	生育状态	用户生育情况	未生育/备孕/怀孕/已生育
	是否有老人	联系人识别是否有 60 以上	是/否
	是否有小孩	联系人识别是否有小于 10 岁以下	是/否
	是否二胎	是否有 2 个小于 10 岁以下	是/否

3.1.2 行为属性标签

[需求说明]：行为属性标签是基于用户使用产品过程中产生的信息，包括登录行为，挂号、问诊、协议处方、保险等订单以及平台点击、浏览、关注、搜索、评价等互联网行为数据，通过基础统计分析了解用户的行为周期、习惯偏好、关注内容等.

[业务要素]：行为属性标签主要通过各类订单以及前端的埋点数据的基础统计分析获取，详细内容及口径如下：

标签所属分类	标签名	标签解释	标签 eg
行为属性标签	最近一次登录时间	取最后一次登录的时间	2016/9/25 11:10
	用户成长值	近期用户健康币成长速度/会员等级升级速度	成长期、平稳期、降档
	登录活跃度标识/指数	一周内是否有 2 次登录记录或一个月内是否有 5 次登录记录/通过模型计算活跃指数	是/否；或指数
	近一个月挂号次数	30 天内累计挂号次数	3
	近一个月付费问诊次数	30 天内累计付费问诊次数	4
	近一个月免费问诊次数	30 天内累计免费问诊次数	
	近一个月协议处方次数	30 天内累计协议处方次数	5
	近一个月保险购买次数	30 天内累计协议处方次数	6
	近一个月关注医生/科室/医院数	30 天内点击关注次数	10
	近一个月评价数	30 天内挂号加问诊的评价数量	30
	近一个月浏览文章数	最近 30 天浏览文章总数	12
	近一个月浏览时长	最近 30 天浏览内容停留总时长	10.5h
	页面浏览层级	用户浏览页面的层级深度	首页/深入点击

	是否点击查看健康币	是否点击查看健康币子页面关注健康币及升级情况	是/否
	近一个月搜索次数	最近 30 天搜索点击总数	21
	近 30 天 banner 点击数	最近 30 天 APP 首页 banner 的点击次数	23
	行为类型	登录平台主要目的是挂号问诊还是浏览搜索内容，了解健康疾病知识	业务类型/咨询内容

3.1.3 疾病问诊标签

[需求说明]：疾病问诊标签是基于用户挂号、问诊、处方数据提取用户（用户联系人）的疾病及问诊相关信息，并相应提取用户搜索、浏览、关注、点击等互联网行为相关的疾病问诊标签，通过数据分析与挖掘预测用户疾病问诊的潜在业务需求.

[业务要素]：疾病问诊标签主要通过分析各类挂号问诊订单以及疾病关注信息数据，提取用户疾病及问诊需求的业务标签，详细内容及口径如下：

标签所属分类	标签名	标签解释	标签 eg
疾病问诊标签	最近患病	最近一次问诊/挂号/处方所患疾病名称	高血压
	家族病史	用户健康档案中的家族病史登记	糖尿病/心脏病/…
	药物过敏史	用户健康档案中的药物过敏登记	青霉素/地卡因/…
	食物和接触物过敏史	用户健康档案中的食物和接触物过敏登记	芒果/牛奶/…
	个人习惯	用户健康档案中的个人习惯登记	久坐/饮酒/…
	疑似疑难杂症标识	就诊日在最近 14 天内的并且有 2 个以上不同专家号的有效预约	是/否
	就诊人数	联系人数量	20
	最近一次关注的医生星级	最近一次(30 天内)添加关注医生的级别	主任医师……
	最近一次关注的医院类别	最近一次(30 天内)添加关注医院的类别	三级甲等……
	最近一次关注的科室	最近一次(30 天内)关注的科室名称	儿科
	最近一次就诊医生星级	最近一次(30 天内)就诊医生的级别	主任医师……
	最近一次就诊医院类别	最近一次(30 天内)就诊医院的类别	三级甲等……

	最近一个月浏览最多文章所属疾病	最近一个月浏览文章所属最多的疾病名称	鼻窦炎
	最近一个月浏览最多文章所属科室	最近一个月浏览文章疾病所属最多的科室名称	五官科
	最近一个月搜索最多词汇/文章所属疾病	最近一个月搜素量中所属最多的疾病名称	老年痴呆
	历史挂号问诊疾病最多科室	历史挂号问诊订单中对应最多的科室	普内科

3.1.4 订单消费标签

[需求说明]：订单消费标签是用户基于平台产品使用过程中进行购买或消费，通过分析各业务订单及消费数据，挖掘用户的消费特征，以便针对性服务.

[业务要素]：订单消费标签主要从业务分类及消费金额等数据角度进行轻量统计汇总，详细内容及口径如下：

标签所属分类	标签名	标签解释	标签 eg
订单消费标签	累计消费金额	累计成功付费，不含优惠和退款的金额(含问诊/挂号/保险/处方)	999rmb
	近一个月消费金额/消费等级分层	最近一个月累计成功付费，不含优惠和退款的金额(含问诊/挂号/保险/处方）/根据所有用户消费金额分析分层	777rmb
	近一个月挂号消费金额	近一个月累计成功挂号付费金额，不含优惠及退款的金额	666rmb
	近一个月问诊消费金额	近一个月累计成功问诊付费金额，不含优惠及退款的金额	555rmb
	近一个月协议处方消费金额	近一个月累计成功协议处方付费金额，不含优惠及退款的金额	444rmb
	近一个月保险消费金额	近一个月累计成功购买保险付费金额，不含优惠及退款的金额	333rmb
	优惠券是否主动领取	账户是否有过主动领取优惠券的记录	是/否
	优惠券是否消费	是否在支付业务中使用过优惠券	是/否
	是否价值敏感用户	对于订单购买行为是否大多有优惠及折扣，优惠依赖用户	是/否

	账户余额	当前健康账户余额	222rmb
	健康币余额	会员通过平台操作行为累计的"积分"	500
	健康币使用抵扣优惠金额	累计使用健康币抵扣订单消费优惠总金额	120rmb

3.1.5 用户偏好标签

[需求说明]：用户偏好标签是用户基于平台产品使用的一种喜好特征或者习惯性，重点分析用户常用渠道、问诊类型、就医偏好、用户加关注内容.

[业务要素]：用户偏好标签从用户的终端类型、问诊方式、历史就诊医生类型、就诊医院类型、用户点击关注信息分析用户各模块的标签特征，具体如下：

标签所属分类	标签名	标签解释	标签 eg
用户偏好标签	登录终端类型	用户登录记录统计分析汇总，从 PC（web）/APP（移动 H5）/第三方渠道登录平台使用产品	PC/APP/第三方
	常用问诊类型	近 1 年问诊过程采用最多的问诊方式	图文/视频/电话/急速/其他
	名医偏好	预约挂号专家都为三甲医院副主任以上医生并且在线问诊都为副主任以上医生则记为“是”，反之为“否”	是/否
	中西医偏好	挂号和咨询的医生 50%以上为中医医生则为“中”，无中医医生则为“西”，其他情况为“无”	中/西/无
	科室偏好	分析用户关注文章、医生等内容所属科室，汇总疾病标签中各关注科室的建模进行权重指标计算	儿科
	医院就医偏好	基本以平台免费咨询，且挂号去医院就医为主；平台付费行为基本没有	是/否
	就医地域偏好	挂号的选择地统计分析，得到用户常去的医院所在地	上海
	访问时段偏好	将一天的时间分区间统计用户的访问时长，统计用户最常登录和访问时间	21:00~23:00

3.1.6 客户满意度标签

[需求说明]：客户满意度标签是用户在使用产品过程中的情绪体现，主要从用户在使用产品后的反馈情况以及用户的流失风险进行综合评估.

[业务要素]：客户满意度标签从用户历史是否有投诉信息、主动评价包括差评数据以及多因子建模评估流失风险，具体标签如下。：

标签所属分类	标签名	标签解释	标签 eg
用户满意度标签	历史是否有投诉工单	历史工单是否有投诉或建议级别的工单	是/否
	是否流失	90 天内未登陆	是/否
	流失风险指数	通过用户近期的登录行为、浏览搜索等互联网点击行为、挂号问诊订单信息等多因子建模，计算用户流失风险指数	0.5
	最近一个月评价数	最近一个月挂号加问诊的评价数量	30
	最近一个月挂号差评数	最近一个月挂号的差评数量	3
	最近一个月问诊差评数	最近一个月问诊的差评数量	0

3.2 接口封装

[需求说明]：用户画像接口旨在解决用户画像数据与各业务渠道应用的传输问题，使用户画像标签能够在各渠道应用时个性化展现，并且保证数据运营及营销推荐数据实时更新，数据可每日更新，避免数据不准确和重复交叉应用。

[功能说明]：Hive 数据仓库封装用户画像模型宽表，每日同步至 Postgress 数据库，各业务及运营可通过直接访问 PG 数据库或数据文件下发的方式，访问画像模型数据宽表；也可通过 WeMeta、WeData 以及 WeSearch 平台配置用户分群规则提取相关的用户标签，实时反馈运营及营销接触数据问题，整合画像模型并更新；配置分析及应用平台可视化展现推荐标签库，以实现权限管控需求。

3.3 UI 设计

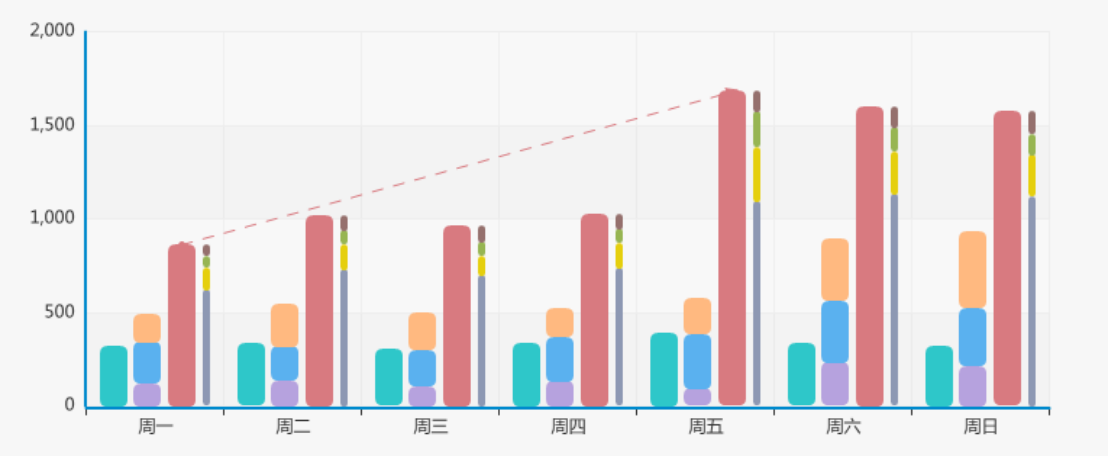
[需求说明]：数字化运营及精准营销的可视化展现，是基于用户画像数据，实现千人千面的展现效果，使运营及营销人员有更好的用户认识，带来更佳的用

户服务质量。

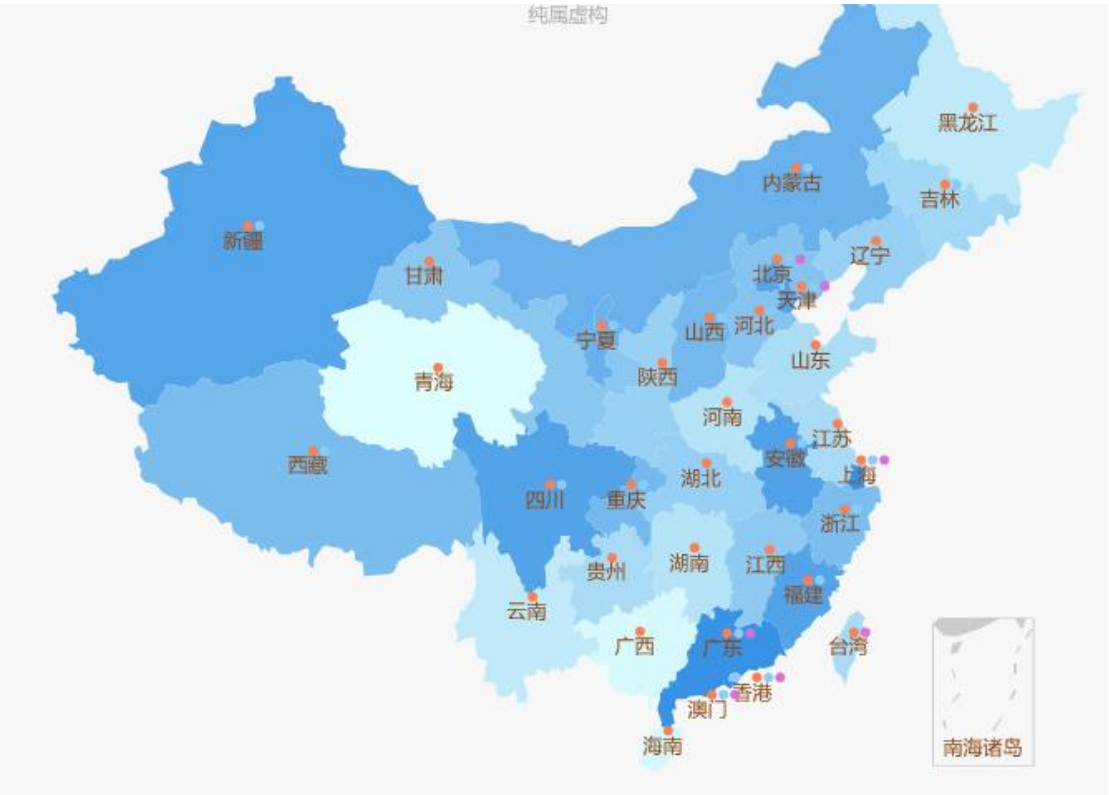
[功能说明]：展现 UI 信息包括：人口属性、行为属性、疾病问诊、订单消费、用户偏好以及客户满意度标签等，同时基于用户汇总实现更多的用户分群统计分析，具体展现样例如下：

- 画像数据展现图

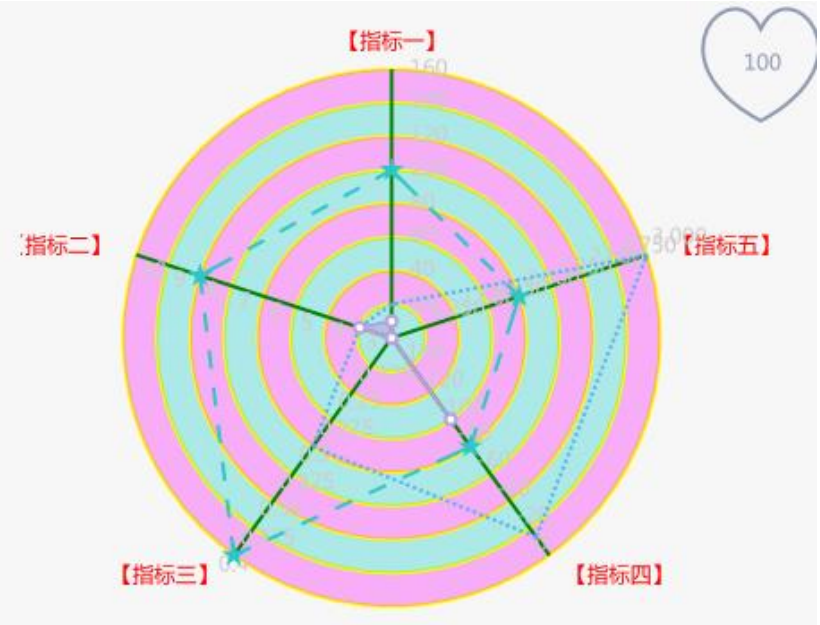
- 业务标签统计柱形图



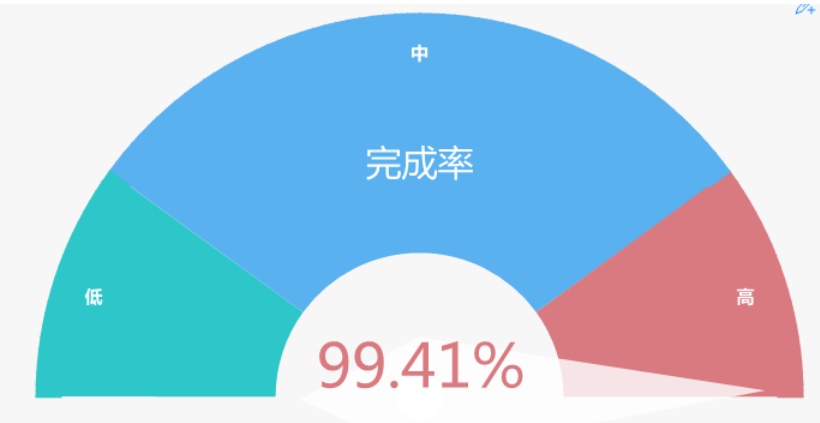
- 用户人群分部热力图



- 用户标签特征雷达图



- 指标情况仪表盘



- 用户关注及搜索疾病标签词云图

3.4　场景应用及流程

待补充

4. 运行环境

4.1 网络与硬件设备

包括数据库服务器、应用服务器配置、网络环境等

4.2 软件平台

Web 服务器环境、数据库操作系统、数据挖掘软件工具等