深度解析京东个性化推荐系统演进史

中兴大数据 浪尖聊大数据 2018-01-05

作者 | fisherman、Davidxiaozhi

> 本文摘自《决战618：探秘京东技术取胜之道》，两位作者时任京东推荐系统负责人和系统架构师。

在电商领域，推荐的价值在于挖掘用户潜在购买需求，缩短用户到商品的距离，提升用户的购物体验。

京东推荐的演进史是绚丽多彩的。京东的推荐起步于2012年，当时的推荐产品甚至是基于规则匹配做的。整个推荐产品线组合就像一个个松散的原始部落一样，部落与部落之前没有任何工程、算法的交集。2013年，国内大数据时代到来，一方面如果做的事情与大数据不沾边，都显得自己水平不够，另外一方面京东业务在这一年开始飞速发展，所以传统的方式已经跟不上业务的发展了，为此推荐团队专门设计了新的推荐系统。

随着业务的快速发展以及移动互联网的到来，多屏（京东App、京东PC商城、M站、微信手Q等）互通，推荐类型从传统的商品推荐，逐步扩展到其他类型的推荐，如活动、分类、优惠券、楼层、入口图、文章、清单、好货等。个性化推荐业务需求比较强烈，基于大数据和个性化推荐算法，实现向不同用户展示不同内容的效果。

为此，团队于2015年底再次升级推荐系统。2016年618期间，个性化推荐大放异彩，特别是团队开创的“智能卖场”，实现了活动会场的个性化分发，不仅带来GMV的明显提升，也大幅降低了人工成本，大大提高了流量效率和用户体验，从而达到商家和用户双赢，此产品获得了2016年度的集团优秀产品。为了更好地支撑多种个性化场景推荐业务，推荐系统一直在迭代优化升级，未来将朝着“满屏皆智能推荐”的方向发展。

推荐产品

用户从产生购买意向，到经历购买决策，直至最后下单的整个过程，在任何一个购物链路上的节点，推荐产品都能在一定程度上帮助用户决策。

推荐产品发展过程

推荐产品发展历程主要经历了几个阶段（图1），由简单的关联推荐过程到个性化推荐，逐步过渡到场景智能推荐。从相关、相似的产品推荐过渡到多特征、多维度、用户实时行为、结合用户场景进行的全方位智能推荐。

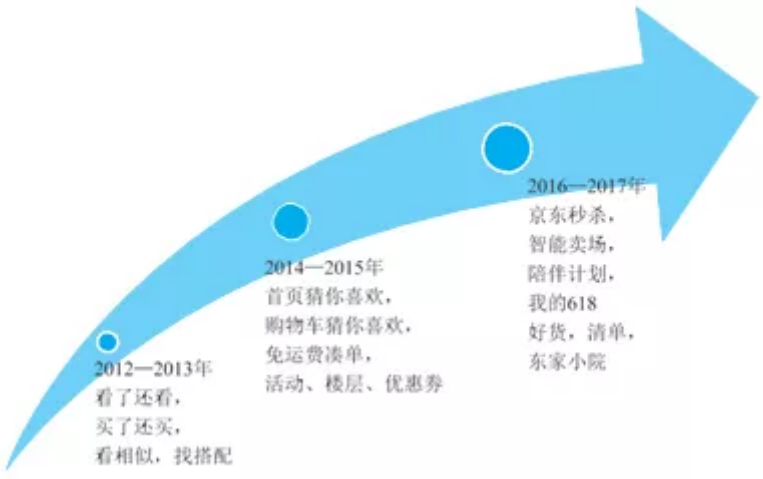

图1　推荐产品发展历程

多屏多类型产品形态

多类型主要指推荐类型覆盖到多种类型，如商品、活动、分类、优惠券、楼层、入口图、文章、清单、好货等。在移动互联时代，多屏场景非常普遍，整合用户在多屏的信息，能使个性化推荐更精准。多屏整合的背后技术是通过前端埋点，用户行为触发埋点事件，通过点击流系统进行多屏的行为信息收集。这些行为数据通过实时流计算平台来计算用户的兴趣偏好，从而根据用户兴趣偏好对推荐结果进行重排序，达到个性化推荐的效果。京东多屏终端如图2所示。

图2　京东多屏终端

推荐系统架构

整体业务架构

推荐系统的目标是通过全方位的精准数据刻画用户的购买意图，推荐用户有购买意愿的商品，给用户最好的体验，提升下单转化率，增强用户黏性。推荐系统的业务架构如图3所示。

图3　推荐系统的业务架构

- 系统架构。对外提供统一的HTTP推荐服务，服务京东所有终端的推荐业务。
- 模型服务。为了提高个性化的效果而开发的一系列公共的个性化服务，用户维度有用户行为服务和用户画像服务，商品维度有商品画像，地域维度有小区画像，特征维度有特征服务。通过这些基础服务，让个性化推荐更简单、更精准。
- 机器学习。算法模型训练阶段，尝试多种机器学习模型，结合离线测评和在线A/B，验证不同场景下的算法模型的效果，提高推荐的转化率。
- 数据平台。数据是推荐的源泉，包括数据收集和数据计算。数据虽然是整体推荐架构的最底层，却是非常重要的，因为数据直接关系到推荐的健康发展和效果提升。

个性化推荐架构

在起步初期，推荐产品比较简单，每个推荐产品都是独立服务实现。新版推荐系统是一个系统性工程，其依赖数据、架构、算法、人机交互等环节的有机结合。新版推荐系统的目标，是通过个性化数据挖掘、机器学习等技术，将“千人一面”变为“千人千面”，提高用户忠诚度和用户体验，提高用户购物决策的质量和效率；提高网站交叉销售能力，缩短用户购物路径，提高流量转化率（CVR）。目前新版推荐系统支持多类型个性化推荐，包括商品、店铺、品牌、活动、优惠券、楼层等。新版个性化推荐系统架构如图4所示。

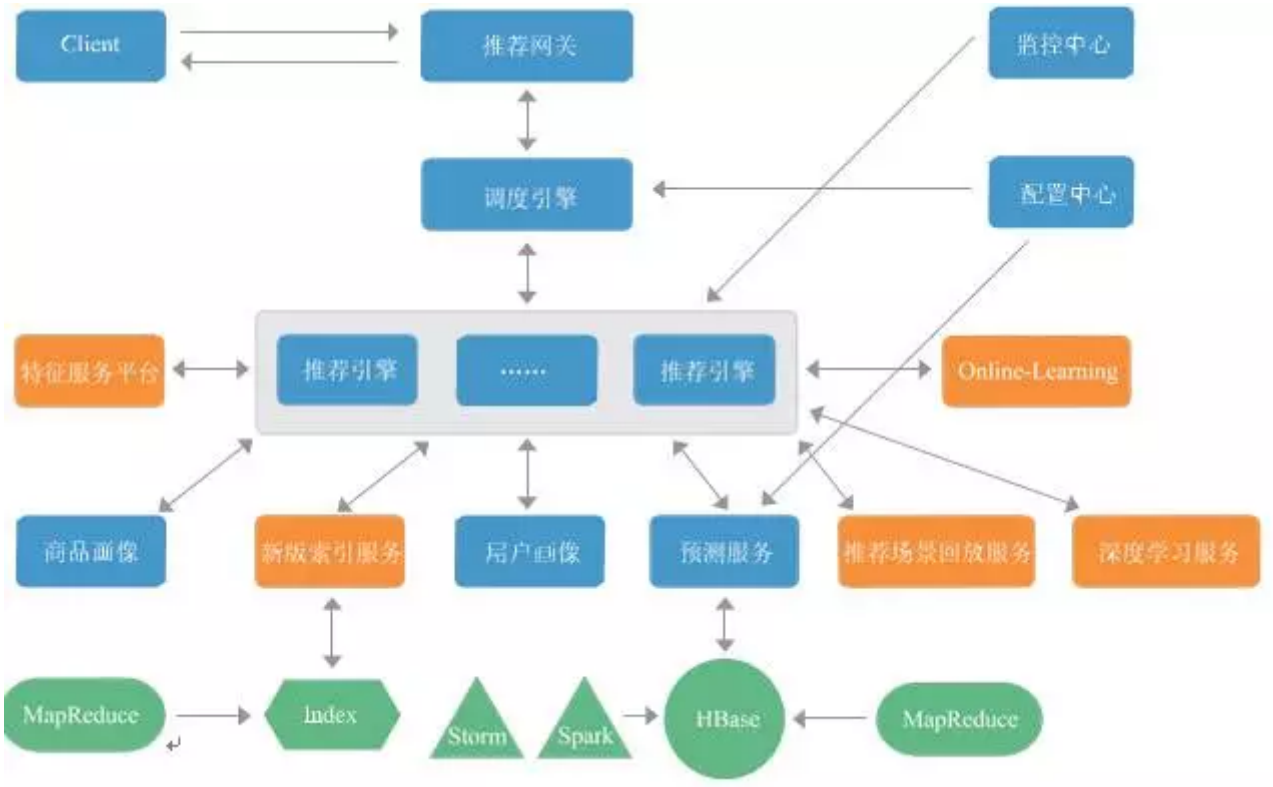

图4　新版个性化推荐系统架构

个性化推荐系统架构图中不同的颜色代表不同的业务处理场景：数据处理部分（最底层绿色模块），包括离线数据预处理、机器学习模型训练，以及在线实时行为的接入、实时特征计算。推荐平台（蓝色模块），主要体现响应用户请求时推荐系统的各服务模块之间的交互关系。推荐系统核心模块：

- 推荐网关。推荐服务的入口，负责推荐请求的合法性检查、请求分发、在线Debug 以及组装请求响应的结果。
- 调度引擎。负责推荐服务按策略调度及流量分发，主要根据配置中心的推荐产品 的实验配置策略进行分流，支持按用户分流、随机分流和按关键参数分流。支持自定义埋 点，收集实时数据；支持应急预案功能，处理紧急情况，秒级生效。
- 推荐引擎。负责推荐在线算法逻辑实现，主要包括召回、过滤、特征计算、排序、 多样化等处理过程。
- 个性化基础服务。目前主要个性化基础服务有用户画像、商品画像、用户行为、 预测服务。用户画像包括用户的长期兴趣、短期兴趣、实时兴趣。兴趣主要有性别、品牌 偏好、品类偏好、购买力等级、自营偏好、尺码颜色偏好、促销敏感度、家庭情况等。商品画像主要包括商品的产品词、修饰词、品牌词、质量分、价格等级、性别、年龄、标签等。用户行为主要获取用户近期行为，包括用户的搜索、点击、关注、加入购车、下单等。预测服务主要是基于用户的历史行为，使用机器学习训练模型，用于调整召回候选集的权重。
- 特征服务平台。负责为个性服务提供特征数据和特征计算，特征服务平台主要针对 特征数据，进行有效的声明、管理，进而达到特征资源的共享，快速支持针对不同的特征进行有效的声明、上线、测试以及A/B实验效果对比。

个性化技术（橙色模块），个性化主要通过特征和算法训练模型来进行重排序，达到精准推荐的目的。特征服务平台主要用于提供大量多维度的特征信息，推荐场景回放技术是指通过用户实时场景特征信息反馈到推荐排序，在线学习（Online-Learning）和深度学习都是大规模特征计算的个性化服务。

个性化推荐系统的主要优势体现为支持多类型推荐和多屏产品形态，支持算法模型A/B实验快速迭代，支持系统架构与算法解耦，支持存储资源与推荐引擎计算的解耦，支持预测召回与推荐引擎计

算的解耦，支持自定义埋点功能；推荐特征数据服务平台化，支持推荐场景回放。

数据平台

京东拥有庞大的用户量和全品类的商品以及多种促销活动，可以根据用户在京东平台上的行为记录积累数据，如浏览、加购物车、关注、搜索、购买、评论等行为数据，以及商品本身的品牌、品类、描述、价格等属性数据的积累，活动、素材等资源的数据积累。这些数据是大规模机器学习的基础，也是更精确地进行个性化推荐的前提。

数据收集

用户行为数据收集流程一般是用户在京东平台（京东App、京东PC网站、微信手Q）上相关操作，都会触发埋点请求点击流系统（专门用于收集行为数据的平台系统）。点击流系统接到请求后，进行实时消息发送（用于实时计算业务消费）和落本地日志（用于离线模型计算），定时自动抽取行为日志到大数据平台中心。算法人员在数据集市上通过机器学习训练模型，这些算法模型应用于推荐服务，推荐服务辅助用户决策，进一步影响用户的购物行为，购物行为数据再发送到点击流，从而达到数据收集闭环。

离线计算

目前离线计算平台涉及的计算内容主要有离线模型、离线特征、用户画像、商品画像、用户行为，离线计算主要在Hadoop上运行MapReduce，也有部分在Spark平台上计算，计算的结果通过公共导数工具导入存储库。团队考虑到业务种类繁多、类型复杂以及存储类型多样，开发了插件化导数工具，降低离线数据开发及维护的成本。数据离线计算架构如图5所示。

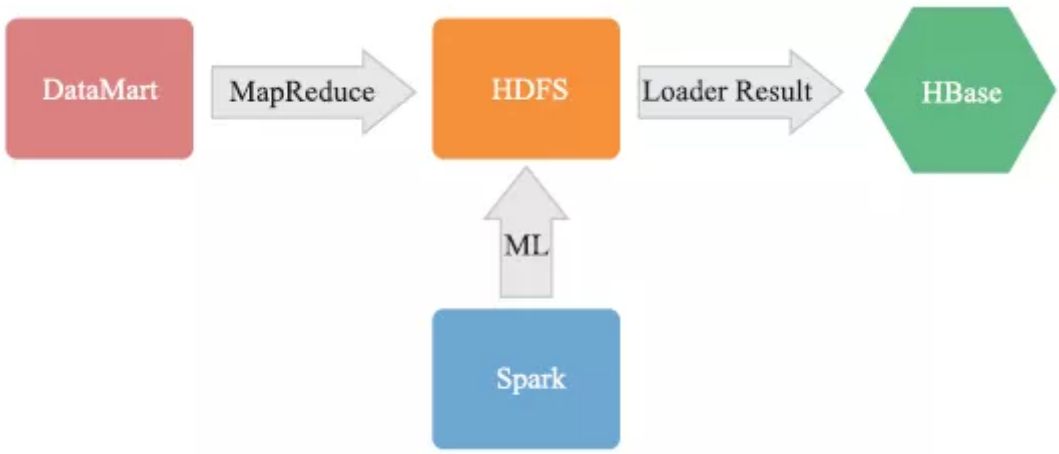

图5　数据离线计算架构

在线计算

目前在线计算的范围主要有用户实时行为、用户实时画像、用户实时反馈、实时交互特征计算等。在线计算是根据业务需求，快速捕捉用户的兴趣和场景特征，从而实时反馈 到用户的推荐结果及排序，给用户专属的个性化体验。在线计算的实现消息主要来源于Kafka集群的消息订阅和JMQ消息订阅，通过Storm集群或Spark集群实时消费，推送到Redis集群和HBase集群存储。数据在线计算框架如图6所示。

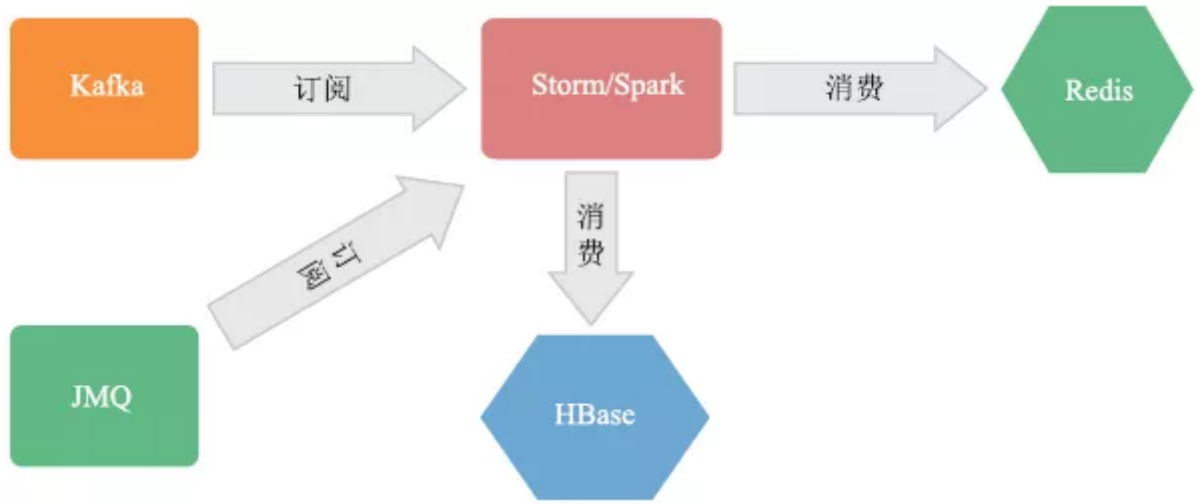

图6　数据在线计算架构

关键技术

推荐系统涉及的技术点比较多，考虑到篇幅有限，这里重点介绍个性化推荐中比较重要的部分。

推荐引擎

个性化推荐系统的核心是推荐引擎，推荐引擎的一般处理过程是召回候选集，进行规 则过滤，使用算法模型打分，模型融合排序，推荐结果多样化展示。主要使用的技术是机器学习模型，结合知识图谱，挖掘商品间的关系，按用户场景，通过高维特征计算和海量召回，大规模排序模型，进行个性化推荐，提升排序效果，给用户极致的购物体验。

推荐引擎处理逻辑主要包括分配任务，执行推荐器，合并召回结果。推荐器负责召回 候选集、业务规则过滤、特征计算、排序等处理。推荐引擎技术架构如图7所示。

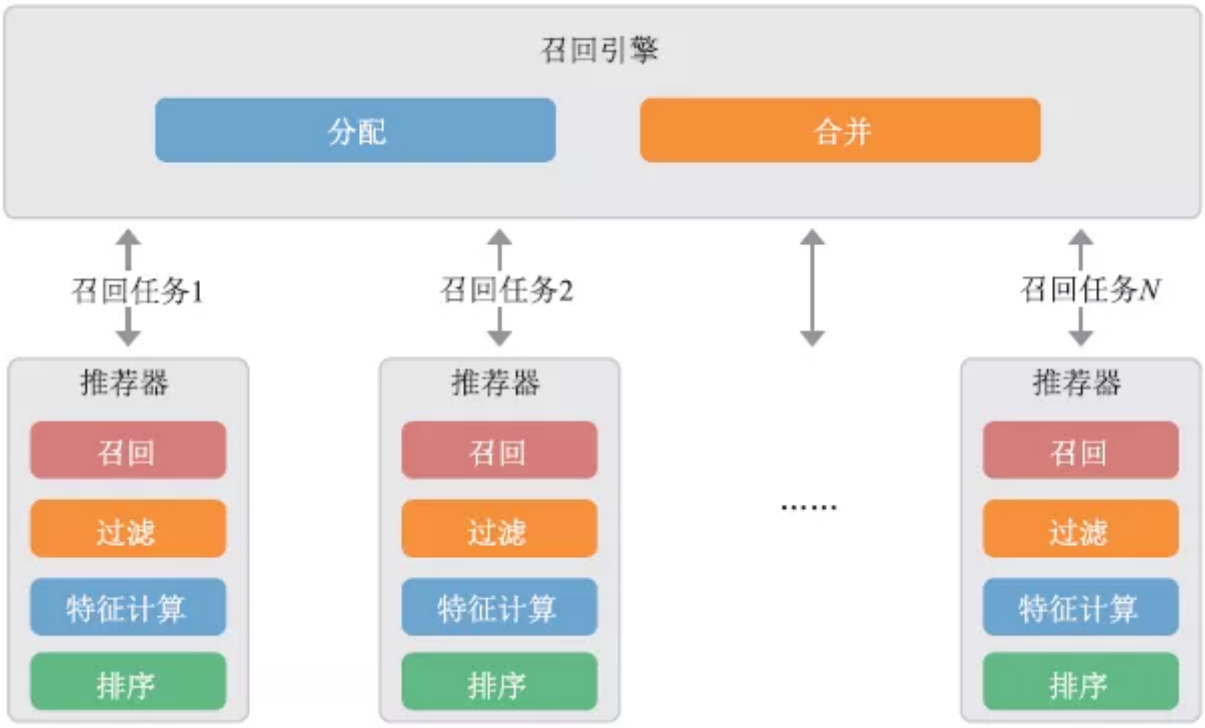

图7　推荐引擎技术架构

分配。根据推荐场景，按召回源进行任务拆分，关键是让分布式任务到达负载均衡。

推荐器。推荐引擎的核心执行组件，获取个性化推荐结果，推荐器的实现如图8所示。

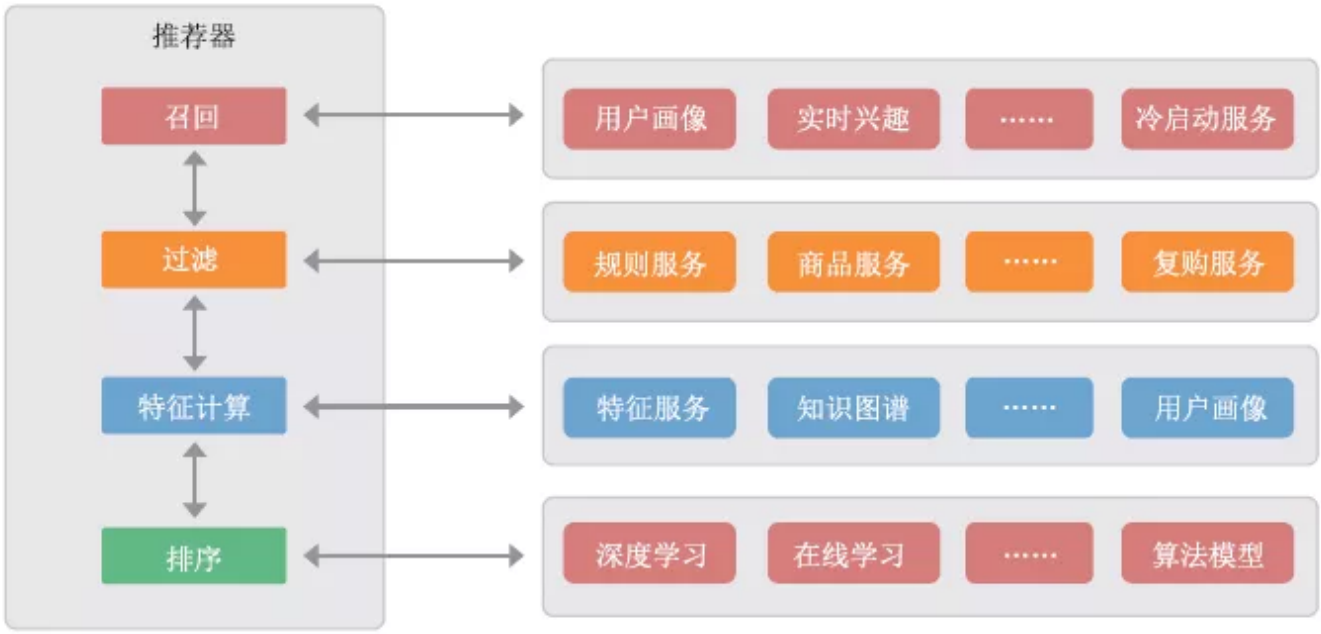

图8　推荐器架构

- 召回阶段。获取候选集，一般从基于用户画像、用户偏好、地域等维度进行召回，如果是新用户的召回资源不够，会使用冷启动服务进行召回。
- 规则过滤阶段。对人工规则、一品多商、子母码、邮差差价等进行过滤。

- 特征计算阶段。结合用户实时行为、用户画像、知识图谱、特征服务，计算出召回的候选集的特征向量。
- 排序阶段。使用算法模型对召回候选集打分，根据召回源和候选集的分值，按一定的策略对候选集进行重新排序。

合并。归并多个推荐器返回的推荐结果，按业务规则进行合并，考虑一定的多样性。举例来说，京东App首页“猜你喜欢”的实现过程如图9所示。首先根据用户画像信息和用户的近期行为及相关反馈信息，选择不同的召回方式，进行业务规则过滤；对满足要求的候选商品集，提取用户特征、商品特征、用户和商品的交叉特征；使用算法模型根据这些特征计算候选商品的得分；根据每个商品的得分对商品进行排序，同时会丰富推荐理由，考虑用户体验，会对最终排好序推荐结果进行微调整，如多样性展示。

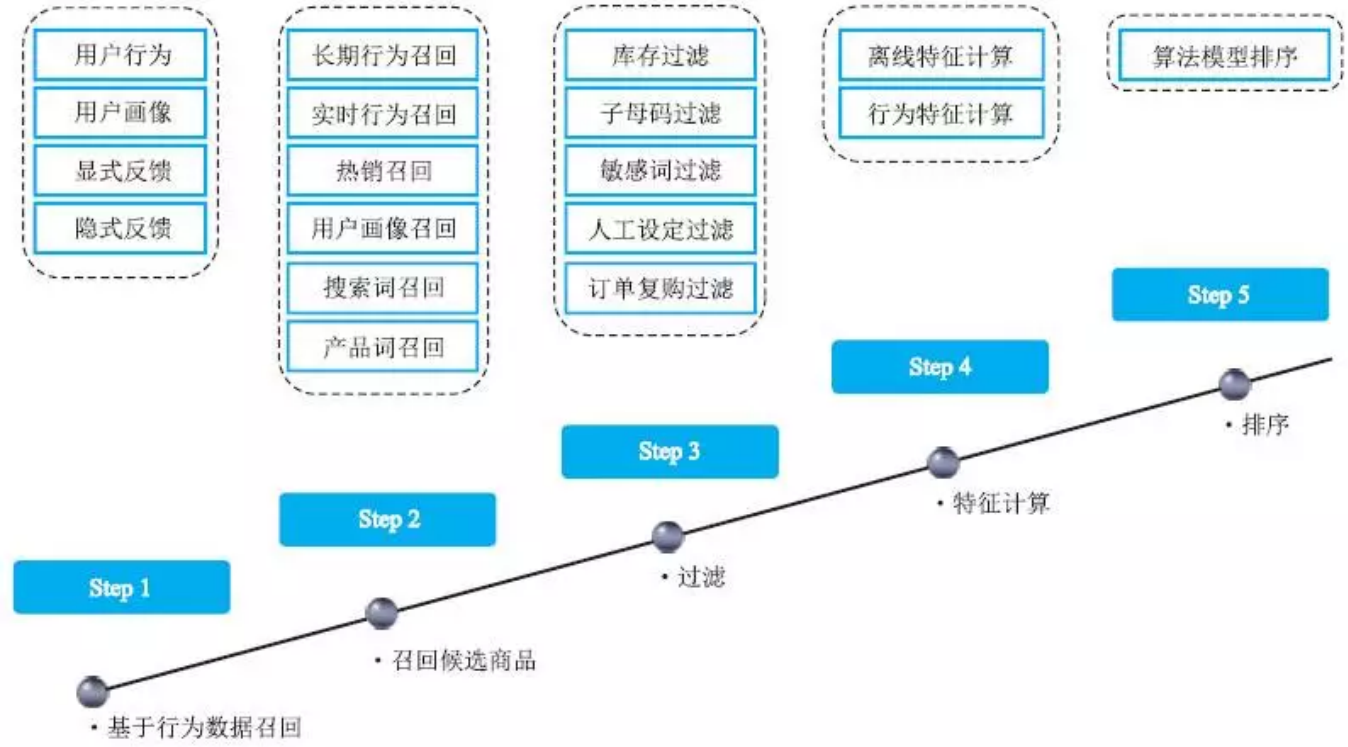

图9　猜你喜欢实现过程图

用户画像

京东大数据有别于其他厂商的地方就是京东拥有最长的价值链和全流程的数据积累。京东数据的特征非常全面，数据链记录着每个用户的每一步操作：从登录到搜索、浏览、选择商品、页面停留时间、评论阅读、是否关注促销，以及加入购物车、下订单、付款、配送方式，最终是否有售后和返

修，整个用户的购物行为完整数据都被记录下来。通过对这些用户行为及相关场景的分析，构建了京东用户画像，如图10所示。

其中不仅有用户的年龄、性别、购物习惯，更有根据其购物行为分析出的大量数据， 例如是否已婚，是否有孩子，对促销是否敏感等。另外，实时用户画像可以秒级分析出用户的购买意图，以及实时兴趣偏好。京东推荐用户画像技术体系如图11所示。

用户画像在京东各终端的推荐产品中都有应用，618推出的智能卖场是用户画像的典型 应用场景。智能卖场的产品包括发现好货、个性化楼层、秒杀、活动、优惠券、分类、标签等。以秒杀为例，推荐结果会根据当前用户的用户画像中的画像模型（性别、年龄、促销敏感度、品类偏好、购买力）进行加权，让用户最感兴趣的商品排在前面。

用户画像也是场景推荐的核心基础。以东家小院为例，根据用户的历史行为汇聚出很多场景标签，按当前用户的画像模型，调整场景标签的排序。如用户选择“包治百病”标签，会按用户画像中的性别、年龄、品类、促销敏感度等画像模型进行推荐商品的重排序。

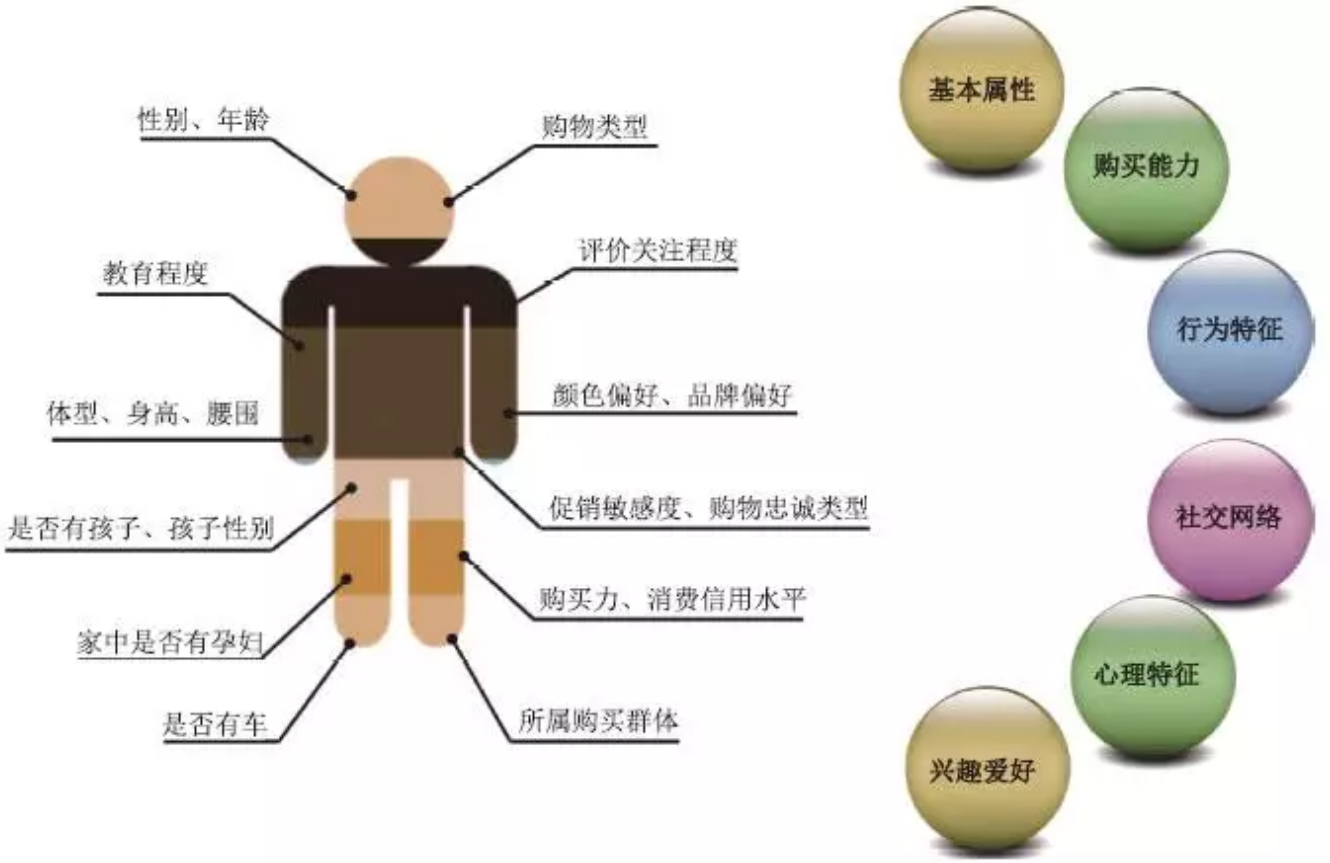

图10　用户画像示意图

图11　京东推荐用户画像技术体系

特征服务平台

特征就是一种属性的描述，特征是个性化推荐的基础，常用的特征分为单边特征和双边特征。单边特征是指对象本身的属性描述，如商品的颜色；双边特征是指两个对象交互程度的描述，如某用户最近一小时浏览的品牌与候选集中品牌的匹配程度。从特征生成的场景来说，分为离线特征和实时特征。离线特征是通过算法模型提前生成，实时特征是通过实时计算的方式生成的。特征的质量直接影响推荐的效果、特征计算的性能，同时影响个性化推荐的处理能力。另外，共享和复用特征可以提高算法的迭代速度并节约人力成本。

特征服务管理平台主要针对特征数据和特征计算，进行有效声明和管理，进而达到特征资源的共享和复用。特征服务平台能快速满足针对制定不同的特征进行有效的声明、上线、测试以及A/B实验效果对比的需求，做到特征的可维护、可说明、可验证。特征服务平台的主要功能如下：离线特征的定制化使用，在线特征的定制化使用，由定制化特征产生新的特征，部分特征、模型在线申明，不同特征效果快速A/B。特征服务平台架构如图12所示。

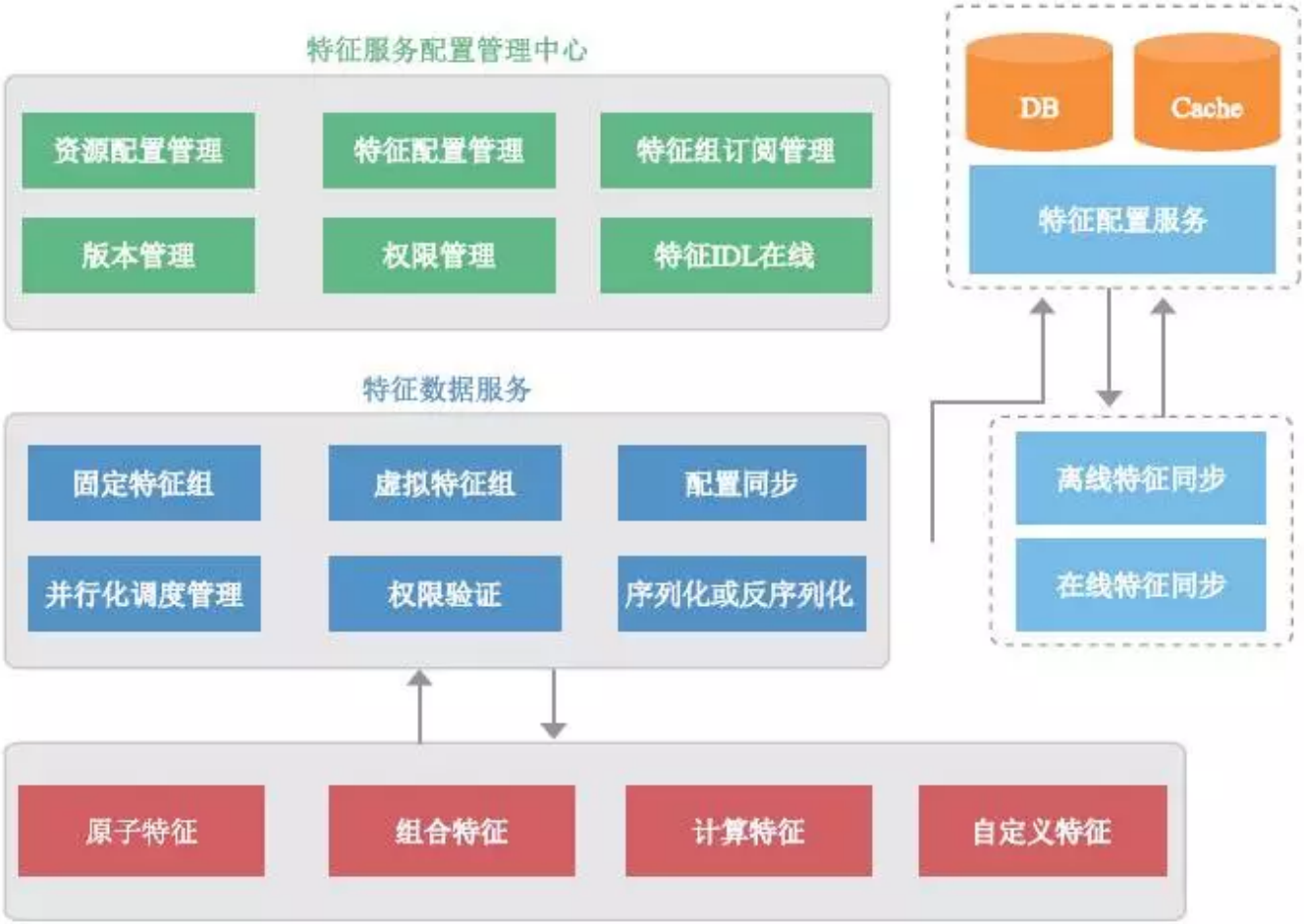

图12　特征服务平台架构

场景特征回放技术

推荐的一般处理逻辑是每次请求会召回一批商品，然后根据用户的行为数据和用户模型计算出每个商品的特征。算法模型会根据每个商品的特征计算出每个商品的得分，最后选出得分最高的几个商品推荐给用户。

线上计算特征这种行为是一次性的，不会被记录下来。因此在线下训练模型的时候，如果想利用上述的特征，就需要在线下机器上再次计算一遍这些特征。遗憾的是，线下计算出来的特征往往不能和线上特征完全相同，这就导致了模型训练的效果较差。场景特征回放示意图如图13所示，推荐业务调用推荐引擎，推荐引擎将场景特征通过特征回放服务记录下来，推送至大数据平台，机器学习根据场景特征数据重新训练算法模型，进而影响推荐引擎中的排序，形成一个场景闭环推荐，达到更准确的个性化推荐。

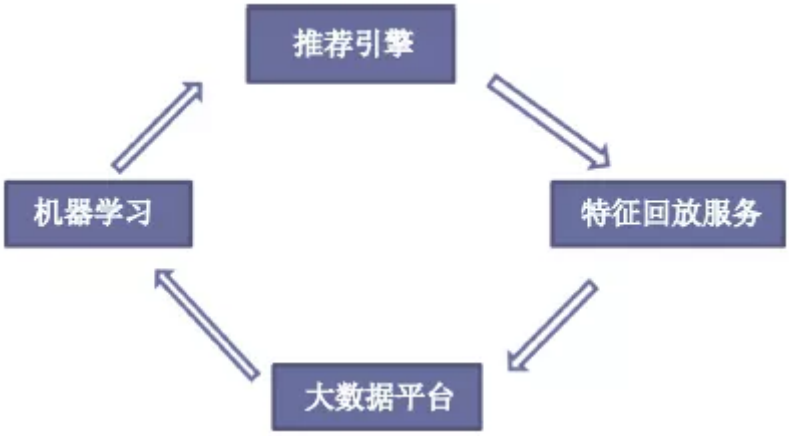

图13　场景特征回放示意图

场景特征回放技术架构如图14所示，场景特征回放技术实现过程如下。线上特征一般是一系列的数值，我们将这些特征按照一定的规则组装成一个字符串，然后将特征使用HTTP的POST方法异步发送到服务端。

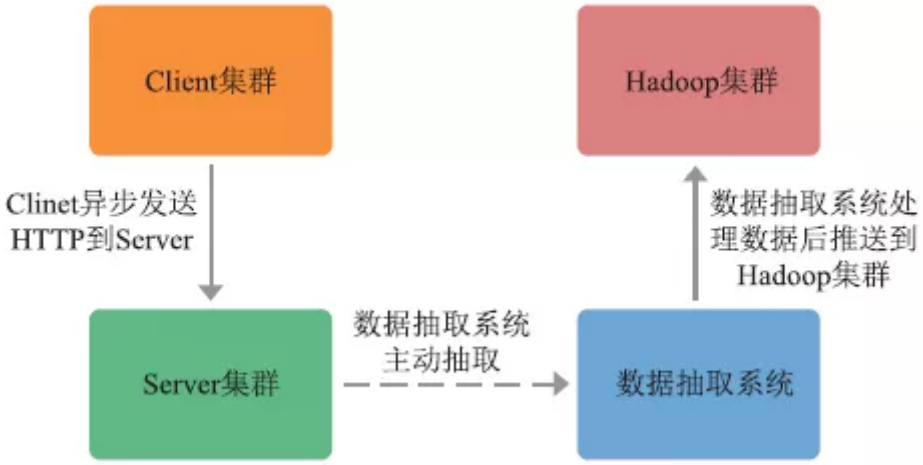

图14　场景特征回放技术架构

服务端使用Openresty接收这些HTTP请求，并把HTTP请求中的特征数据落地到本地磁盘文件中。Openresty是一种高性能的Web服务器，能够承受很高的QPS，并且具有很高的稳定性，它的这两点特性保障了服务的稳定。

数据抽取系统把服务器集群磁盘上的数据抽取到临时仓库。

数据抽取系统对数据进行压缩和过滤处理，然后推送到Hive表中。不同类型的请求会放到不同的分区中，更加方便算法工程师使用这些数据。

个性化推荐系统是一个系统工程，依赖产品、数据、架构、算法、人机交互等进行场景推荐，本节重点从这几个维度阐述了京东的个性化推荐系统。推荐系统随着业务发展和社会生活方式的改变而

进行不断升级，经历了从PC时代到移动互联时代，从关联推荐走向个性化推荐，从纯商品推荐到多类型推荐的转变。个性化推荐系统已经实现了千人千面。诚然，个性化的效果也有待提升，有些体验类的问题也在逐步完善。目前正在进行或有待提高的方面包括：算法方面丰富知识图谱、深度学习广泛应用；推荐系统方面会更好地支持海量召回、高维特征计算、在线学习，推荐更实时，更精准；产品方面已向“满屏皆智能推荐”方向迈进。最后，希望个性化推荐系统能让购物变得简单，变得更人性化、更丰富、更美好。

推荐阅读：

1，推荐系统之用户行为分析

2，案例：Spark基于用户的协同过滤算法

3，请别再问我Spark的MLlib和ML库的区别

--- 密 --- 封 --- 线 --- 内 --- 不 --- 要 --- 答 --- 题 ---