



# 基于云计算的移动互联网大数据用户行为分析引擎设计<sup>\*</sup>

陶彩霞, 谢晓军, 陈 康, 郭利荣, 刘 春

(中国电信股份有限公司广东研究院 广州 510630)

**摘 要:**随着移动互联网的迅速发展,运营商面临的竞争日益激烈,流量经营势在必行,基于用户行为分析的精确营销是重要手段。但在大数据时代,随着移动互联网业务和用户数量的快速增长,传统的架构难以适应海量数据挖掘的要求。本文提出了一种基于云计算的移动互联网大数据用户行为分析引擎解决方案,包括系统总体架构设计、大数据入库与预处理组件、大数据用户行为分析模型等关键模块的设计,最后分析了系统测试效果。

**关键词:**云计算;大数据;移动互联网;用户行为分析;MapReduce;应用平台;DPI

**doi:** 10.3969/j.issn.1000-0801.2013.03.006

## Design of Mobile Internet Big Data User Behavior Analysis Engine Based on Cloud Computing

Tao Caixia, Xie Xiaojun, Chen Kang, Guo Lirong, Liu Chun

(Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

**Abstract:** Along with the rapid development of mobile internet, telecom operators face an increasingly competition. So traffic operations are imperative for telecom operators, and precise marketing based on user behavior analysis is an important means. But in the era of big data, with the rapid growth of the mobile internet service and the number of users, the traditional architecture is difficult to adapt to the requirements of mass data mining. A design of mobile internet big data user behavior analysis engine based on cloud computing was provided, including the overall system architecture design solution, the big data loading components and the big data user behavior analysis model components. Additionally, the test results of system performance were analyzed.

**Key words:** cloud computing, big data, mobile internet, user behavior analysis, MapReduce, application platform, DPI

### 1 引言

从 2010 年开始,国内移动互联网进入快速发展阶段,领先的电信运营商开始布局移动互联网,同时,互联网公司全面介入,终端厂商也基于应用商店模式快速加入移动互联网领域。根据艾瑞咨询发布的数据,全球移动互联网用户数正呈爆发式增长,2014 年全球移动互联网用户数有望达到 14 亿。然而在数据流量快速增长的同时,电信运

营商却出现数据业务收入增速放缓的困境,面临被管道化的威胁。为应对移动互联网带来的挑战,中国电信提出了从话务量经营转向流量经营的战略目标,从传统的注重用户规模转变为注重流量发展。运营商拥有庞大的用户,同时具有对终端及用户上网通道的掌控能力,使得在用户行为分析方面具有很好的数据基础,深入分析用户流量行为特征和规律,发现用户潜在流量使用需求,是提升流量规模和价值、提高流量经营水平的有效手段。

<sup>\*</sup> 国家“863”计划基金资助项目(No.2012AA12A203)



然而随着移动互联网的迅速发展,用户行为分析面临着新的挑战:一是移动互联网新业务、新产品“短、平、快”的特征,要求运营商支持快速变化的营销活动;二是随着移动互联网业务及终端、传感器技术发展带来的数据量的急剧膨胀,需要分析和处理的数据规模从GB级迈向TB级甚至PB级。传统的数据分析架构已经不能适应这种海量数据处理和快速、深度挖掘的需求,迫切需要引入大规模并行处理技术和分布式架构,构建基于云计算的移动互联网用户行为分析引擎系统,以应对移动互联网大数据时代的挑战。

## 2 基于云计算的系统总体设计方案

### 2.1 系统总体技术架构

本文设计的移动互联网用户行为分析引擎通过云计算技术实现分布式并发的大规模计算能力,构建移动互联网端到端的大数据挖掘分析系统,实现对DPI和应用平台用户上网行为的偏好分析,提供个性化推荐服务,打通从

数据采集、分析到服务提供、营销执行的全过程。

系统通过FTP服务器获取数据,在接口层采用分布式计算与批量处理相结合的方式,将大数据存入Hbase数据库中,支持海量数据和非结构化数据的存储,数据入库之后利用Hive进行整合层和汇总层的ETL处理,再基于MapReduce计算框架设计大数据分析模型,最后通过Hive数据库将结果导入前端展现数据库。在数据处理层,利用Hbase、Hive的优势进行海量数据的存储和处理,考虑到前端展现要求的灵活性,采用关系型数据库MySQL作为前端展现。系统总体技术架构如图1所示。

### 2.2 系统总体拓扑和功能分布

系统的总体拓扑如图2所示,系统由一台服务器作为Hadoop平台和Hbase的主节点服务器,其他服务器为Hadoop平台和Hbase的从节点服务器,从节点服务器的数量可根据系统处理需求动态扩展。主节点服务器主要负责从节点服务器任务和流量的分配,并对从节点服务器的执

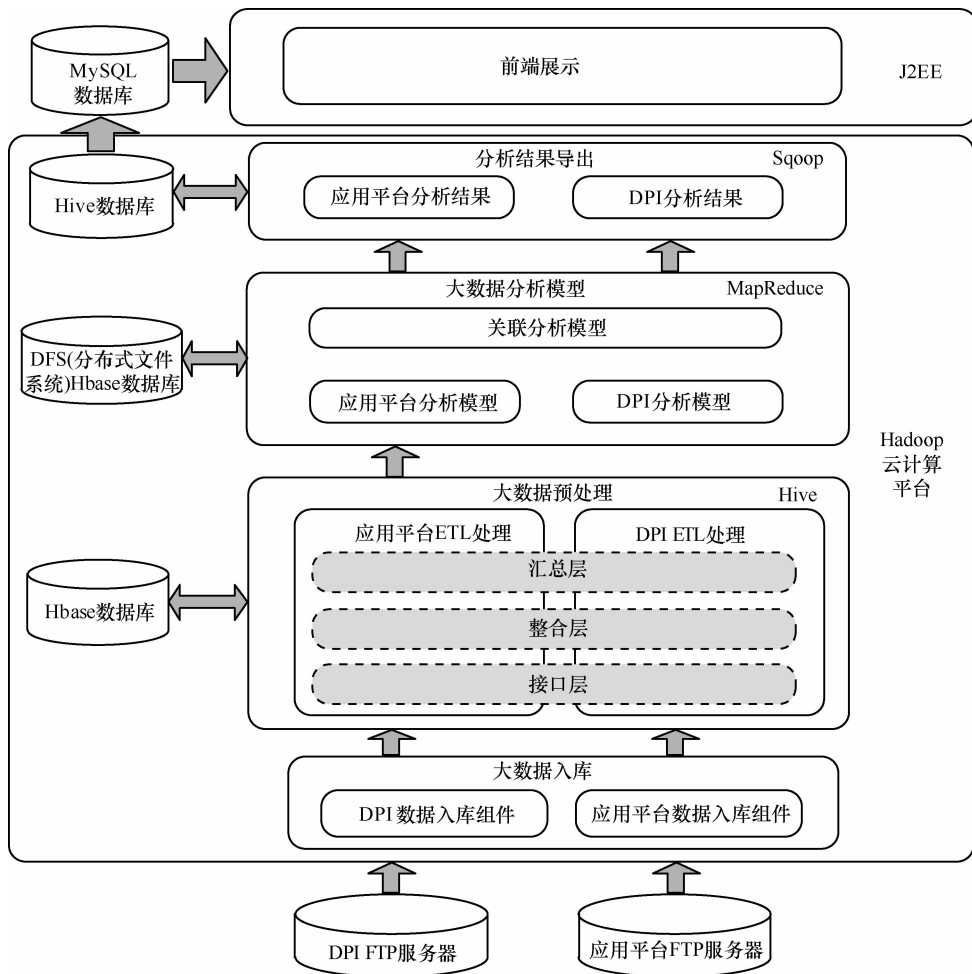


图1 基于云计算的移动互联网大数据用户行为分析引擎总体技术架构

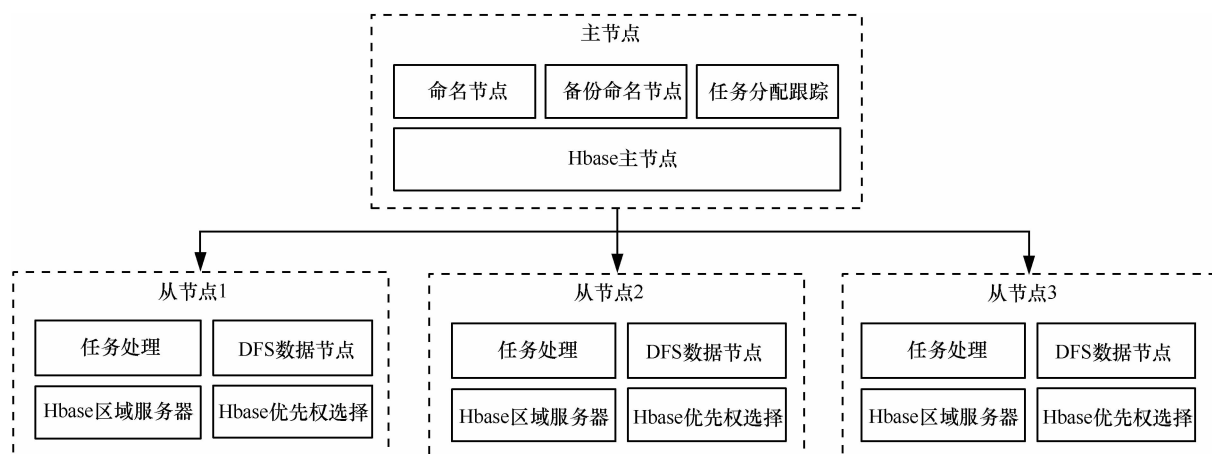


图2 系统总体拓扑

行状态进行监控,多台从节点服务器在主节点服务器的控制下执行具体的任务。

主节点服务器的软件功能架构如图3所示,各模块具体介绍如下。

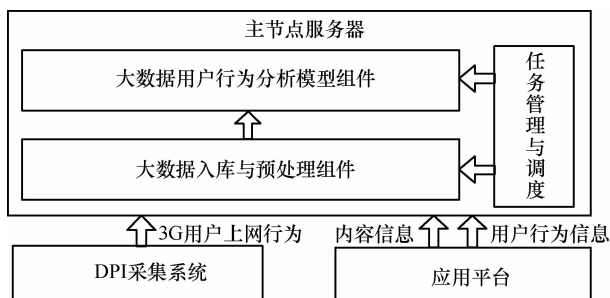


图3 主节点服务器功能架构

#### (1)任务管理与调度模块

集中式的任务调度控制台,提供任务的创建、调整和删除等功能,通过业务类型选择、执行周期设置等,定义应用的处理逻辑;自动控制数据抽取、数据整理到数据建模、模型运行、结果输出等过程,根据任务设置的激活处理条件,自动加载任务运行,系统提供任务的暂停、恢复以及优先级管理功能。

#### (2)大数据入库与预处理组件

将DPI用户的上网行为、应用平台的用户行为和内容信息等大数据,及时导入用户行为分析引擎系统,作为数据分析和模型挖掘的数据源。

#### (3)大数据用户行为分析模型组件

基于汇聚到系统中的海量移动互联网用户行为数据,利用MapReduce计算框架构建用户行为分析模型资源池,快速分析用户的偏好、社会关系信息,且支持多类业务实

现精准的内容推荐。

从节点服务器的软件结构与主节点服务器基本相同,区别主要在于从节点服务器不需要部署任务管理和调度模块。

### 3 大数据入库组件设计

移动互联网用户行为分析引擎的数据来源主要有两类:应用平台数据和DPI数据。两类数据源的特点不同:应用平台的数据主要集中在一个访问行为表上,每天一个文件,每个文件的大小为GB级;而DPI数据的特点是大量的小文件,每个文件大小在10MB以内,但文件来源频率快,一般2min就有好几个文件,一个省份累计1天的数据量可达1TB。

针对上述不同的数据源特点,系统采用不同的技术方案,具体介绍如下。

#### (1)应用平台数据入库

应用平台采用每天批量入库一次的方式,文件大小为GB级。大文件的入库,适合Hadoop平台的使用场景,考虑到使用MapReduce分布式处理的Hbase入库方式效率不高,系统对该方式进行了优化,采用MapReduce分布式处理结合批量入库的方式。

由于Hadoop通常使用的TextInputFormat类,在map中读取到的是文件的一行记录。因此,系统使用NLineInputFormat类实现在MapReduce中的批量入库。通过使用NLineInputFormat类,每个分片有N行记录,通过参数的配置,每次可读取文件的N行记录,那么可以在map中直接执行批量入库的操作,效率相对较高。

#### (2)DPI数据入库

由于DPI的行为数据是大量来源频率很快的小文件,



在 Hadoop 平台下处理小文件采取的措施通常如下。

- 利用 SequenceFiles 将小文件打包上传,可从源头避免小文件产生,但无论是 Hadoop shell 还是 MapReduce,都不能进行灵活读取。
- 使用 HAR 将 HDFS 中的小文件打包归档(从 HDFS),可减少既有 HDFS 中的小文件数量,但 HAR 文件读取性能差。
- Hadoop append 可直接追加数据到相同文件中,但每个小文件的大小不同,同时考虑每天的 DPI 日志有峰值和低谷,对文件数量的控制和处理来说有一定的麻烦。
- Flume、FlumeNG、Scribe,可通过中间层汇聚数据的办法减少小文件数量,但 FlumeNG 和 Scribe 都不能很好地传输压缩文件。

通过以上分析可以看到,上述 4 种方式均存在一定的缺点,因此针对 DPI 数据的特征,采用 Hadoop 平台的 CombineFileInputFormat 类方式,即通过继承 CombineFile InputFormat,实现 CreateRecordReader,同时设置数据分片的大小,通过这种方式实现 DPI 大数据的入库。

#### 4 大数据用户行为分析模型组件设计

大数据用户行为分析模型组件提供多个在 Hadoop 分布式平台上运行的分析模型,其功能结构及其与其他组件的关系如图 4 所示。

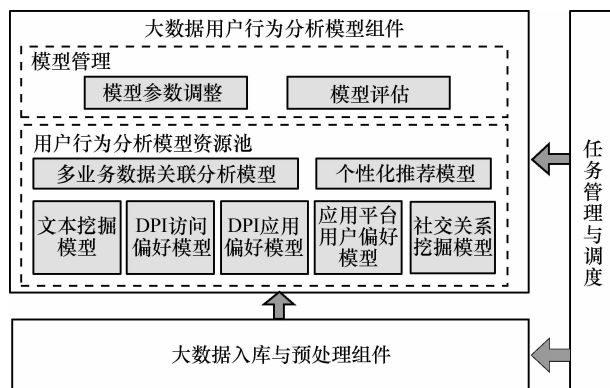


图 4 大数据用户行为分析模型组件

本组件主要由以下几个模块组成。

- 模型参数调整:提供对模型算法中的变量设定、参数调整、样本空间规模设置等功能。
- 模型评估:提供创建模型校验任务,将实际数据与模型计算结果进行比对,输出模型校验指标,进行

模型校验和模型有效性评价。

- 多业务数据关联分析模型:对用户的互联网行为和爱游戏业务平台的行为进行关联分析,判断 DPI 用户上网行为偏好与在应用平台上的行为偏好是否存在关联关系,采用关联算法找出其中存在的规则,并将规则固化到系统中,从而有助于交叉营销。
- 个性化推荐模型:以协同过滤技术和内容推荐技术为主,采用混合推荐技术,综合考虑来自产品内容和用户两个维度的影响,按照综合相似度向用户推荐相应的信息,实现用户动态推荐算法。
- 文本挖掘模型:对文本内容(如网页)通过预处理去除噪声(如网页导航栏、页首、页尾、广告等不相关内容),提取出文本主体部分,根据文本(网页)分类标准构造标注语料库,通过分类训练算法进行模型训练和机器学习,建立文本(网页)人工智能分类模型。
- DPI 访问偏好模型:基于网页内容分类,通过用户访问网页分析,计算用户 Web 访问兴趣偏好。
- DPI 应用偏好模型:基于 DPI 采集数据,通过应用知识库识别应用,计算用户应用兴趣偏好。
- 应用平台用户偏好模型:依据用户在应用平台上的各种操作行为,找出用户对应用平台各种内容的偏好规律。
- 社交关系挖掘模型:社交关系挖掘包括用户社交图谱和兴趣图谱的构建。社交图谱通过用户的位置轨迹进行挖掘分析,建立用户之间的好友等社交关系;兴趣图谱基于用户偏好模型,计算用户偏好的相似度,得到与用户兴趣最相近的邻居集合,建立用户之间的相同兴趣爱好关系。

以上模型的建模过程中很多用到了 MapReduce 计算框架。在 MapReduce 计算框架中,每个 MapReduce 作业主要分为两个阶段:map 阶段和 reduce 阶段,分别用 map 函数和 reduce 函数实现。map 函数对一个<key,value>形式的输入进行计算,产生一个<key,value>形式的中间输出,而 Hadoop 负责把 map 输出的相同 key 值的 value 集合到一起,传递给 reduce 函数,reduce 函数对每个 key 值对应的 value 集合进行处理,也以<key,value>的形式输出。以个性化推荐模型为例,基于内容推荐的算法分为 3 个 job 完成,每个 job 的操作如下。

(1)job1:计算用户对内容的偏好度

map 函数:从 Hbase 中读取用户行为数据,组合相关的

数据。key 为用户 ID,value 为用户浏览过的内容信息(如游戏、视频等)。

reduce 函数:获取每个用户所有的行为信息,计算用户对内容的偏好度。key 为用户 ID+内容 ID,value 为内容偏好度。

(2)job2:计算用户对内容属性的偏好度

map 函数:读取内容偏好度信息传给 reduce 函数。key 为用户 ID,value 为内容 ID+内容偏好度。

reduce 函数:计算每个用户的内容属性偏好度。

(3)job3:计算基于内容的推荐列表

map 函数:获取用户内容属性偏好度和用户内容偏好度,key 为用户 ID,value 为属性偏好度+内容偏好度。

reduce 函数:计算推荐列表。

通过这种方式,系统把数据密集型的大数据用户行为分析模型运算任务分配到各个计算节点分布式运行,从而大大提高模型运算性能。

## 5 系统测试与结果分析

为了验证本文提出的基于云计算的移动互联网大数据用户行为分析引擎设计方案,在实验室用 PC 服务器搭建了 4 个节点的 Hadoop 分布式平台,其中一个为主节点,3 个为从节点,以大数据入库为测试场景,与单机运行环境进行对比,测试了 5 组不同数据规模情况下的系统运算时间,测试结果如图 5 所示。

从测试结果可以看出,单机批量入库方式的性能很低,对于 3 GB、6 GB、9 GB、12 GB、15 GB 的 5 组数据规模,其系统运算时间分别为 32 min、60 min、95 min、121 min、151 min,系统运算时间随着数据量的增长几乎呈线性增长,对于海量的大数据来说,很难在合理的时间内快速完成

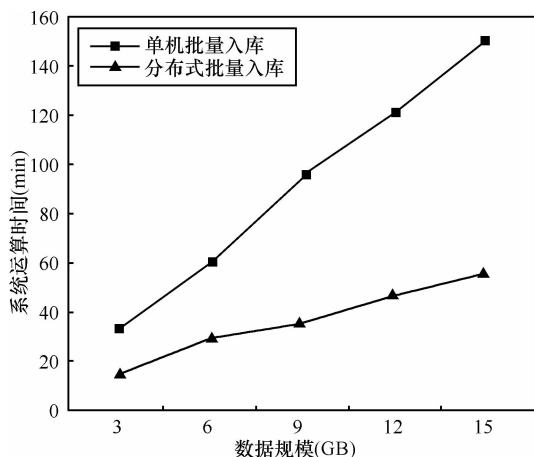


图 5 单机与分布式批量入库测试性能对比

数据的入库操作;而采用云计算分布式批量入库方式,系统处理时间大大缩短,对于 3 GB、6 GB、9 GB、12 GB、15 GB 的 5 组数据规模来说,系统运算时间分别为 14 min、29 min、35 min、47 min、55 min。从两种方式的测试结果对比可以看出,随着数据规模的增大,云计算分布式批量入库方式与单机批量入库方式的处理时间差距越大,优势越明显。

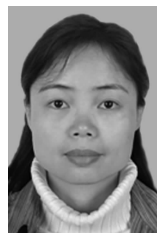
## 6 结束语

本文针对移动互联网大数据时代给运营商带来的挑战,提出了一种基于云计算的移动互联网大数据用户行为分析引擎设计方案。通过实验验证,该方案能有效地应对移动互联网新业务、新产品“短、平、快”的特征和数据规模急剧膨胀等问题,能够在有效的时间内完成大数据处理和分析任务。

## 参考文献

- 1 陈康,郑纬民.云计算:系统实例与研究现状.软件学报,2009,20(5)
- 2 冯铭,王保进,蔡建宇.基于云计算的可重构移动互联网用户行为分析系统的设计.计算机科学,2011(8)
- 3 Anand Rajaraman, Jeffrey David Ullman. Mining of Massive Datasets. Cambridge University Press, 2011

### [作者简介]



陶彩霞,女,硕士,中国电信股份有限公司广东研究院高级工程师,主要研究方向为大数据、移动互联网、云计算。

谢晓军,男,硕士,中国电信股份有限公司广东研究院高级工程师、技术总监,主要研究方向为大数据、电信支撑系统、移动互联网、云计算。

陈康,男,硕士,中国电信股份有限公司广东研究院主任,主要研究方向为大数据、云计算、电信支撑系统、移动互联网。

郭利荣,男,学士,中国电信股份有限公司广东研究院工程师,主要研究方向为大数据。

刘春,男,硕士,中国电信股份有限公司广东研究院工程师,主要研究方向为大数据、移动互联网。

(收稿日期:2013-03-04)