



游戏大数据平台研究与实践

苏 洋¹, 刘晓军², 唐 勇¹, 黄 洋¹

(1. 炫彩互动网络科技有限公司 南京 210029;

2. 中国电信股份有限公司广东研究院 广州 510630)

摘 要:立足于游戏平台数据快速处理需求,构建基于分布式+关系数据混合多元的大数据处理架构,针对高价值密度的结构化数据采用关系数据处理;大量无序的非结构化数据采用分布式机制处理。两种类型数据实现有序化后,统一交予传统关系数据库构建关系化模型并展示,既能实现关键经营分析数据的高效处理,又能满足无序的日志数据规模化处理需求。通过剖析其总体架构、主要功能和关键技术,从而为其他业务平台大数据分析建设和改造提供借鉴和参考。

关键词:大数据;游戏平台;数据挖掘

doi: 10.3969/j.issn.1000-0801.2014.10.004

Research and Practice of Big Data Platform of Games

Su Yang¹, Liu Xiaojun², Tang Yong¹, Huang Yang¹

(1. Dazzle Interactive Network Technologies, Nanjing 210029, China;

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

Abstract: Based on the game platform data processing needs, the diversified mix of large data processing architecture was built based on distributed + relational data, for high-value-density structured data using relational data processing; lots of disorderly unstructured data processing was using a distributed mechanism. Two types of data to achieve the orderly were unified handed over to traditional relational database model to build relationships and showed that could be achieved through efficient handling of business analysis data, but also to meet the scale of disorderly log data processing requirements. The overall architecture, the main features and key technologies were analyzed, so as reference was provided for other business data analysis platform for large construction and renovation.

Key words: big data, game platform, data mining

1 引言

半个世纪以来,随着计算机技术全面融入社会生活,信息爆炸已经积累到了一个开始引发变革的程度。针对用户规模超2亿户、月活跃用户数超3000万户、在线游戏内容达到数万款的游戏平台而言,每天新增数据量超过100GB。为了实现精细化运营,需要跟踪所有游戏产品的市场效果,以便及时调整运营手段;同时,需要随时跟踪用

户最新状况,以新增、付费、活跃和流失等不同纬度判断总体健康状态,以便及时进行用户激励、挽留等市场策略,促进游戏业务的可持续、健康发展。

传统的数据处理机制,难以在十几分钟内有效全面分析平台数据。如以采样模式,难以完成用户、产品、开发商和渠道等多种维度数据汇总,无法有效指导游戏平台和游戏内容产品的总体运营。

为了满足游戏平台海量数据从无序到有序,从碎片化



到规律化的转化,需要构建基于分布式+关系数据混合多元处理架构的大数据处理机制。针对游戏平台无序和海量的基础采集数据源,采用分布式机制处理,将大量无序数据转化成为有序数据;数据规范化后,为了提升处理速度和展现便捷性,采用传统关系型数据库构建关系化模型并展示。

2 平台架构设计

互联网常见的大数据分析平台,例如 Google、淘宝等,支撑网络搜索和电子商务等业务,采用并行计算+分布式存储架构,汇聚海量服务器计算能力,提升处理时效和响应规模,满足瞬间上千万次服务请求。该类平台处理的都是海量的网站信息、图片等非结构化数据,具有适合统一分布式分片处理和存储等特点。

针对游戏平台而言,其数据量没有达到 Google、淘宝等大数据分析平台的海量化,而且具有相当部分的用户鉴权/支付等核心结构化数据,同时投入数据分析的服务器数量有限,因此在大数据处理和分析过程中,必须同时兼顾针对大量非结构化数据分布式处理能力及结构化数据处理的时效性。

游戏大数据平台采用多元化数据处理技术,兼容分布式计算,以便提升非结构化数据的处理规模和结构化数据处理能力,实现快速的处理效率,将总体架构可分为4层:第1层是数据源采集,第2层是数据存储,第3层是数据处理,第4层是数据可视化与数据服务,如图1所示。

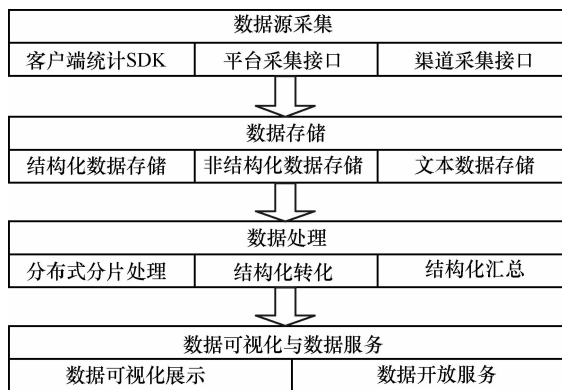


图1 功能总体架构

(1) 数据源采集

主要采用3种方式获取原始数据,包括游戏平台侧、终端侧和销售渠道侧。通过内置的数据采集模块或工具,实现对原始数据的采集。

- 客户端统计 SDK 采集用户行为,将用户终端侧用

户数据日志上传至游戏服务端日志接收服务端。

- 平台收集业务数据:由于用户的某种行为(如下载、短信代收费等),直接触发了游戏平台的对应业务接口,接口直接生成日志文件,放在指定目录。
- 渠道相关数据采集:采用内置渠道广告位、链接以及新闻等内置代码,收集有关用户点击、消费和下载等相关数据。

(2) 数据存储

将3种渠道的数据根据其数据形态的不同,分成结构化、非结构化、文本等方式存储到游戏平台数据资源池中。

- 结构化数据:用户、内容和渠道基本属性相关数据,具有二维表结构逻辑表达特性的主要数据。
- 非结构化数据:各种分时代的碎片化数据,例如用户登录、退出、消费和下载等相关数据。
- 文本数据:受限于原始数据表现形式,例如游戏内容运行日志等,采用文本方式进行存储。

(3) 数据处理

数据处理包括分布式分片处理、结构化转化和结构化汇总等部分。

- 分布式分片处理:将无序和混乱的用户登录/付费/下载/退出等数据,按照一定的时间逻辑,进行数据分片处理,分配给不同的服务器进行分布式处理。
- 结构化转化:将分布式分片处理后的数据,按照时间、内容、渠道、区域等维度转化成为结构化数据,即增添结构化标签。
- 结构化汇总:按照不同的处理要求,将结构化数据汇总,并计算出延伸数据,如流失率、活跃率、付费率和 ARPU 数值等数据。

(4) 数据可视化与数据服务

以图表形式将大数据分析出来的数据进行可视化展示,面向内容商、渠道商提供相关的数据服务。

游戏大数据平台的主要关键点包括:构建既满足经营分析关键数据精确处理要求,又具有海量零散数据分析能力的创新处理机制;针对海量的游戏平台数据,基于分布式计算的存储与处理;为满足自适应数据需求,将数据结果进行可视化展示等。

3 数据处理机制

3.1 数据源

游戏业务,尤其是手机游戏(以下简称手游)业务所产

生的数据与其他业务存在较大差别,即碎片化和时效性更加突出,导致数据处理必须及时、高效。

- 数据时效性突出: 手游内容和用户游玩生命周期较其他网络业务更短,一般仅能持续几个月,而热门的网络应用会生存几年以上,且游戏内容更新换代快,因此必须分析游戏内容生命周期,并对业务数据实时处理。
- 数据碎片化突出: 不同于传统业务数据具有一定持续性和规律性的特点,手游不论是在用户游玩时间,还是持续时长方面都极具随机性,其碎片化特性十分突出,用户或游戏内容的相关数据必须实时采集和处理,这为海量数据处理效率带来了极大的挑战。

同时,对于用户付费、购买和平台/游戏收入等关键经营分析数据,依然保持结构化数据的特点,构成了整体游戏平台非结构数据碎片化、零散性和结构化数据有序化、逻辑化并存的格局。

游戏大数据分析平台的数据源,是直接通过客户端、游戏业务平台和相关渠道获取的所有原始数据,主要包括常量数据和变量数据两大类。常量数据是指在一段时间内保持不变的游戏内容编号、内容提供商编号、渠道编号等,主要是以数据形式展示内容、厂商和渠道,为其相关数据及其分析结果逻辑对应奠定基础;变量数据主要包括用户数据、内容数据和渠道数据等。

游戏大数据平台原始数据源以变量数据为主,具体包括以下 3 种。

- 用户数据: 分时段的登录用户、充值用户、IP 地址情况、网站访问情况、内容下载情况等数据。
- 内容数据: 内容下载、登录、消费、退出等。
- 渠道数据: 渠道的实时流量情况,包括渠道带来的用户点击、下载、消费数据。

游戏大数据平台变量数据中,用户、内容和渠道 3 类数据并非是相互孤立的,而是相互关联的,只不过分析维度不同,即整体数据源资源池只有一个。

3.2 多元化大数据处理

对于游戏用户、游戏内容和渠道 3 种不同碎片化数据,如用户本身日常登录/登出、页面浏览与点击等运营数据,游戏内容下载、卸载、用户实时行为及渠道带来流量的实时变化等,一般以零散的日志形式存在,需要采用 Flume 技术进行海量采集和 Hadoop 分布式模式进行数据处理。

对于用户充值/付费、游戏收入和平台总收入等经营分析的关键数据,需要满足实时精确化查询和统计要求,

而且本身就有传统结构化数据属性,适合采用 Oracle 等关系数据进行处理。

为了满足游戏大数据平台的上述要求,需要采用创新的多元化混搭模式的数据处理机制,如图 2 所示。

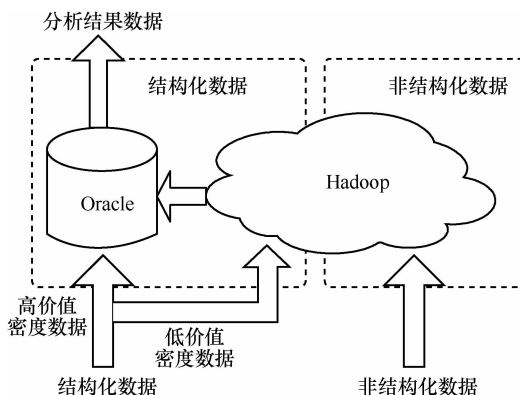


图 2 多元化混搭大数据处理机制

- 数据分类处理: 非结构化数据,如大量日志,交予 Hadoop 分布式处理;结构化数据中高价值密度数据,如收入等经营分析关键数据,交予 Oracle 处理;结构化数据中低价值密度数据,如游戏中单个道具购买等大量低价值海量数据,交予 Hadoop 处理,以便提高效率。
- 数据交互模式: 大量无序数据和低价值密度海量数据经 Hadoop 分布式处理后,转化成为结构化数据,并由 Hive 技术映射成为 SQL 查询的数据表,最终交予 Oracle 进行最终的逻辑运算,并输出用户行为、游戏内容生命周期等分析结果数据。

通过上述的多元化混合制数据分析和处理后,既能实现全面数据分析,又能满足关键经营分析数据的高效处理和输出,能够满足游戏平台精细化运营分析和处理的需求。

4 分布式计算及实践

4.1 分布式计算体系

游戏大数据平台总体分布式计算体系,是以 HDFS 为基础,采用 Hadoop 架构和 HBase 实现,总体分布式软件体系如图 3 所示。

- HDFS: 构建分布式文件系统。以时间为准线,例如以每小时为单元形成简单数据模型,并对整体文件进行分解和分块,以便提交不同服务器进行数据处理。

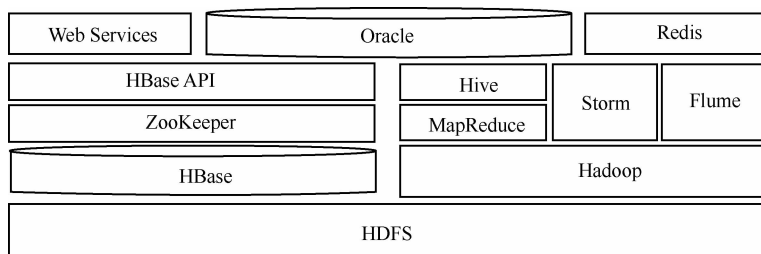


图3 分布式软件体系

- HBase:实现分布式存储系统。利用 HBase 技术实现基于 PC 服务器的大规模结构化存储集群。
- Hadoop:构建分布式处理架构。针对各分块数据,实现基于不同服务器的并行计算,加快处理速度和提高处理效率。
- Storm:降低实时处理复杂性。利用其默认支持 Clojure、Java、Ruby 和 Python 等语言多样性,提升处理效率。
- Flume:主要用于海量日志数据采集。采用分布式模式,对海量日志进行采集、聚合和传输等。
- ZooKeeper:提供一致性服务。主要与 Flume 结合,实现海量日志数据节点的高效、可靠的协同工作。
- MapReduce、HBase API 编程:采用相关 MapReduce 编程架构和 HBase API,实现分布式数据处理编程。
- Hive:数据结构化处理。采用该基于 Hadoop 的数据仓库工具,可以将完成结构化处理的数据文件映射为一张数据库表,并提供简单的 SQL 查询功能。
- Oracle:直接处理收入等重要的经营分析结构化数据以及分布式处理完成的结构化数据,如活跃度、

流失率、付费率和 ARPU 数值等。

4.2 分布式部署与实践

平台的大数据分布式处理,是基于 Hadoop 的体系结构,包括 HDFS 和 HBase 分布式存储系统两部分组成。为了提升文件和数据处理效率,采用 HBase 构建于 HDFS 之上的模式部署。

分布式文件体系方面,采用主从(master/slave)结构模型,如图4所示。192.168.80.31 服务器作为 HDFS 唯一主节点的 NameNode(名称节点),192.168.80.32、192.168.80.33、192.168.80.34 和 192.168.80.35 作为 4 个 DataNode(数据节点),构建分布式文件系统集群。其中 192.168.80.31 负责对游戏平台的文件、目录等元数据信息进行管理和维护,以提供统一的文件名字空间,将整体文件按照对应时间戳进行数据分块,分别交予 DataNode 读写和处理;其他 4 台服务器作为 DataNode,处理文件系统客户端的文件读写,并在 NameNode 的统一调度下进行数据库的创建、删除和复制工作。

分布式存储系统方面,采用 HBase 分布式数据库架构,以 HDFS 为基础,具备“向下提供存储,向上提供运算”

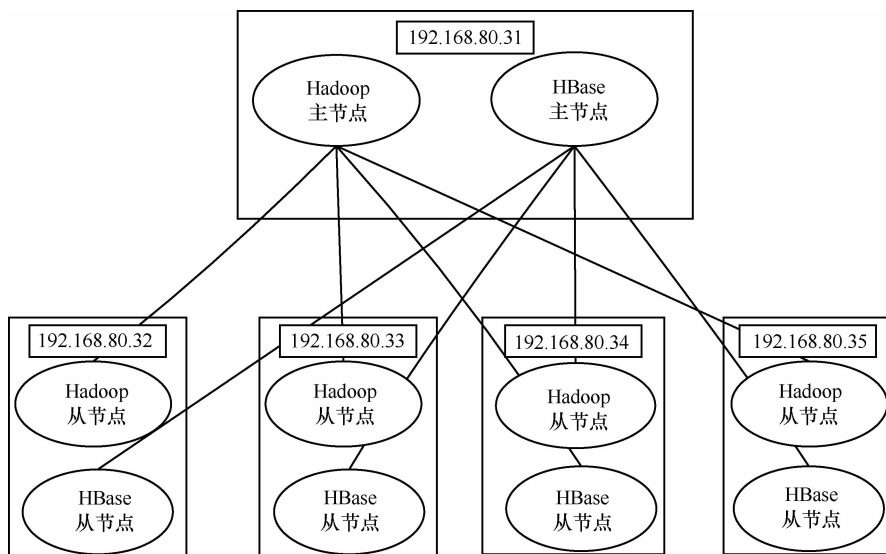


图4 分布式部署

的特点,在 HDFS 主节点 192.168.80.31 之上构建 Hmaster 服务器,管理所有的 HRegion 服务器,进行数据处理的任务分配和总体调度。HBase 主服务器本身并不存储 HBase 中的任何数据,HBase 逻辑上的表被划分成多个区域(region),然后存储到 HRegion 服务器群中。

在 HDFS 4 个从节点上构建 HRegion 服务器,按照 Hmaster 统一调度,HRegion 服务器主要负责响应数据处理的 I/O 请求,向 HDFS 中读写数据,是 HBase 中最核心的模块。HRegion 服务器内部管理了一系列 HRegion 对象,每个 HRegion 对应了表中的一个 region,HRegion 中由多个 HStore 组成。每个 HStore 对应了表中的一个集中存储单元(column family)的存储。

5 可视化展示

游戏大数据平台处理完成后,为了方便指导精细化运营,需要进行图表化直观展示。根据处理数据类型,分为

实时运营数据和非实时运营数据两种不同展示方式。

- 实时运营数据,是指当前一段时间(如 1 h)内,实时付费情况、客户端下载量和重点游戏内容的运营数据监控等;该套实时运营数据可视化系统,相比较其他可视化产品具有终端控制功能,即可使用手持终端进行开关、刷新、切换,尤其方便现场演示。
- 非实时运营数据,是指一段时间内,数据的总体趋势或者规律,如近 3 个月/1 个月/前 1 天的活跃用户、留存率、下载量、产品不同频道的转化率等;该套非实时数据统计分析平台相比业界主流的可视化系统,具有开发简易、动态配置报表、行列可拖拽的优势。

实时数据,一般根据实时运营要求,及时展示最新的平台运营情况。采用 Flume 技术,实时采集生产环境的日志,通过 Storm 流计算框架进行实时统计,得出当前的实时收入、实时下载量等数据,如图 5 所示。

如图 6 所示,为了满足对不同运营情况的实时跟踪和

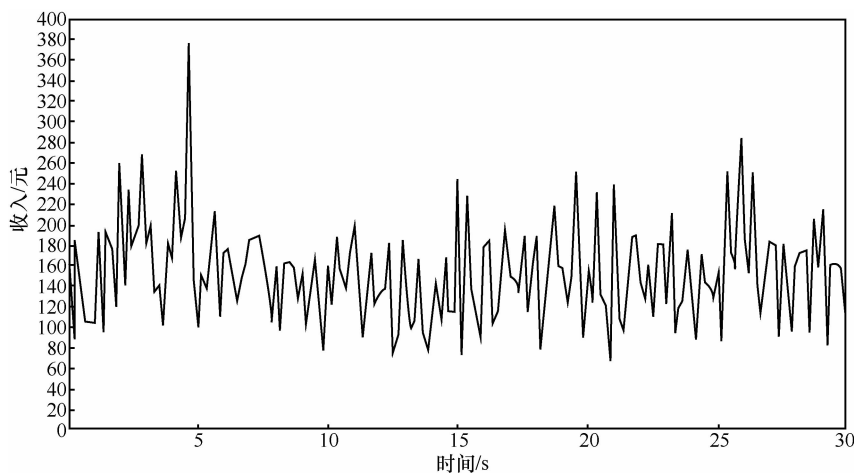


图 5 实时数据监控展现效果

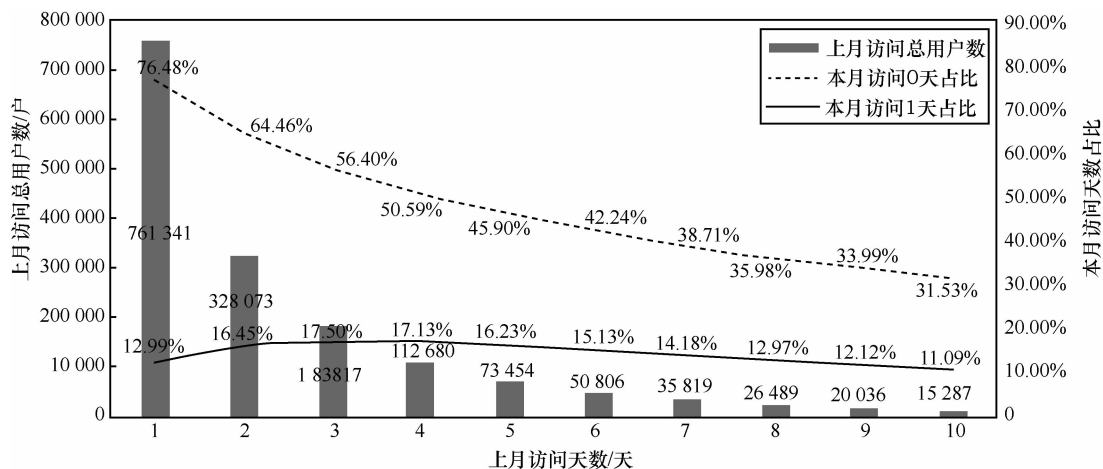


图 6 非实时数据监控展现效果



查询,设计了 SSAS 多维数据集处理技术,实现收入、付费和访问等多维度的灵活组合,提供数据立体化组合、查询和展示,满足个性化实时数据获取需求。

非实时运营数据,主要通过查询 Oracle 分析数据结果数据表,并依靠 FlowBreeze 等 Excel 插件技术,转化成为相应图表直观展示。在非实时运营数据转化过程,通过分析数据需求属性,自动选择柱状图、线形图和饼图等不同的图标格式,展示分析和处理效果。

6 结束语

本文介绍了创新采用分布式+关系数据混合处理架构,构建游戏大数据平台,其充分利用了分布式处理海量无序数据和关系型数据库处理结构化数据效能的双重优势,在实践过程获得很好的应用效果,其中实时数据采集达到秒级时延,采用 HDFS 文件价格数据存储持久性可达 99.999 9%,而且针对冷数据整体累计备份达 40 TB,有效支撑了业务的精细化运营和用户精确画像,构建了以数据为基础的业务发展模型,为其他业务平台大数据分析提供了良好的参考和借鉴。

参考文献

- 1 Boja C, Pocovnicu A, Batagan L. Distributed parallel architecture for "big data". Informatica Economica Journal, 2012, 16(2)

- 2 周江,王伟平,孟丹等. 面向大数据分析的分布式文件系统关键技术. 计算机研究与发展, 2014, 51(2)
- 3 宋杰,郭朝鹏,王智等. 大数据分析的分布式 MOLAP 技术. 软件学报, 2014(4):731~752
- 4 赵辉,杨树强,陈志坤等. 基于 MapReduce 模型的范围查询分析优化技术研究. 计算机研究与发展, 2014(3): 606~617
- 5 智慧来. 面向异构数据分析的形式概念分析扩展模型. 电子学报, 2013, 41(12)

[作者简介]



苏洋,男,炫彩互动网络科技有限公司工程师,主要研究方向为大数据、移动互联网等。

刘晓军,男,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、云计算等。

唐勇,男,博士,炫彩互动网络科技有限公司高级工程师,主要研究方向为大数据、手机游戏等。

黄洋,男,炫彩互动网络科技有限公司工程师,主要研究方向为手机游戏、大数据等。

(收稿日期:2014-09-25)

·简讯·

2014 年“宽带中国”示范城市出炉

据工业和信息化部网站 2014 年 10 月 9 日消息,工业和信息化部与国家发展和改革委员会两部委日前联合发布了 2014 年“宽带中国”示范城市(城市群)名单,确定 39 个城市(城市群)为 2014 年度“宽带中国”示范城市(城市群)。

通告称,为落实《国务院关于印发“宽带中国”战略及实施方案的通知》(国发[2013]31 号),依据《工业和信息化部办公厅发展改革委办公厅关于开展创建“宽带中国”示范城市(城市群)工作的通知》(工信厅联通[2014]5 号),经城市申报、各省预审和专家综合评审,确定 39 个城市(城

市群)为 2014 年度“宽带中国”示范城市(城市群)。

2014 年度“宽带中国”示范城市(城市群)名单如下。

直辖市及城市群:北京市、天津市、上海市、长株潭城市群。

其他城市及省直管县:石家庄市、大连市、本溪市、延边朝鲜族自治州、哈尔滨市、大庆市、南京市、苏州市、镇江市、昆山市、金华市、芜湖市、安庆市、福州市(含平潭)、厦门市、泉州市、南昌市、上饶市、青岛市、淄博市、威海市、临沂市、郑州市、洛阳市、武汉市、广州市、深圳市、中山市、成都市、攀枝花市、阿坝藏族羌族自治州、贵阳市、银川市、吴忠市、阿拉尔市。

(来源:新华网)