文章编号：2095-2163(2020)12-0160-05 中图分类号：TN929.5 文献标志码：A

基于 Hadoop 架构的电信离线数据综合处理的设计与实现

张丽华[1]，马家龙[2]，程晓旭[1]，邹雨轩[1]，刘博宁[1]，贾美娟[1]
（1 大庆师范学院 计算机科学与信息技术学院，黑龙江 大庆 163712；2 上海安顺信息技术股份有限公司，上海 201101）

摘 要：本文研究的系统搭建在 Hadoop 平台上，通过 Flume-Kafka 技术实现对每日数百亿的数据进行数据清洗、数据分析以及数据挖掘等。完成数据消费后，对分析后的数据进行入库操作，通过 Webserver 技术实现建立仿真的 BI 前端系统，按照手机号、时间、通话时长等维度进行数据展示。为电信运营商从多个角度定义用户，形成用户肖像，为决策系统的建立提供数据支持。

关键词：Hadoop；Flume-Kafka；数据挖掘；Webserver；BI

Design and implementation of Telecom Offline data integrated processing based on Hadoop Architecture

ZHANG Lihua[1], MA Jialong[2], CHENG Xiaoxu [1], ZOU Yuxuan[1], LIU Boning[1], JIA Meijuan[1]
(1 School of Computer Science and Information Technology, Daqing Normal University, Daqing Heilongjiang 163712, China; 2 Shanghai Anshun Information Technology Co., Ltd. Shanghai 201101, China)

【Abstract】 The system studied in this paper is built on the Hadoop platform. Through Flume-Kafka technology, it can carry out data cleaning, data analysis and data mining of tens of billions of data every day. After data consumption is completed, the analyzed data is put into the database, and a simulated BI front-end system is established through Webserver technology, and data is displayed according to the dimensions of mobile phone number, time, call duration, etc.

To define users from multiple angles for telecom operators, to form user portrait, to provide data support for the establishment of decision system.

【Key words】 Hadoop; Flume-Kafka; Data mining; Webserver; BI

0 引 言

随着云计算的出现和深入发展，电信行业客户各类结构数据信息量的与日俱增，如何提升数据的有效率并能对其进行更深层次的挖掘，是急需待解决的问题。鉴于开源的 Hadoop 云平台自身在云端处理数据能力所展现的优越性，如：能充分利用分布式存储机制[1]和 MapReduce 编程模型，对已迁移到该云平台上原有数据仓库的数据做存储，并利用通用的设备对其实现并行处理任务、并行计算等，这些信息极有利于深度挖掘电信的社交网络，在低成本的情况下能极大提升数据处理能力[2]。

本文首先对 Hadoop 技术进行了简单介绍，然后对搭建在 Hadoop 平台上的 Flume-Kafka 高可用离线数据采集方案进行设计，重点介绍了本系统的 MapReduce 核心算法。系统使用 MapReduce 技术并行处理数据，通过 Sqoop 组件读取 HDFS 中计算过的数据，并将其读入 MySQL 数据库中保存，最后使用 Webserver 技术实现数据可视化。经测试，系统运行正常有效，表明基于 Hadoop 的数据实时离线处理平台能够满足电信运营商的要求，可以为企业获取有价值的数据。

1 “大数据”时代的计算机信息处理技术

在当前大数据计算机科学管理框架中，获取、加工、存储，实现大数据信息的相关结合并保障其安全性尤为关键，对大数据进行全方位的综合分析有赖于计算机信息处理技术，主要包括信息采集、存储、信息加工传输、信息安全等方面内容[3]。

信息采集及传播技术主要是通过监控数据源，从网络中获取海量的数据信息中筛选提炼。再将一些具有含金量、价值高的信息导入计算机数据库前，对其进行合理化的控制及监管，除此外，也有针对性的对数据进行传播，提升传播率。目前，广泛使用的数据搜索引擎作为信息采集的关键技术，能够有条理的对数据信息进行检索，便于人们快速获取所需

基金项目：大庆市指导性科技计划项目(zd-2019-69)；大庆师范学院科学研究基金资助项目(19ZR10)。

作者简介：张丽华(1980-)，女，硕士，讲师，主要研究方向：计算机网络及云计算；马家龙(1997-)，男，学士，BI 工程师，主要研究方向：网络安全；程晓旭(1965-)，女，硕士，副教授，主要研究方向：移动互联网技术。

收稿日期：2020-10-18

信息。

信息存储技术，主要是指跨越时间保存信息的技术，能在信息加工完成后对加工的信息做有效存储，是实现数据库对加工后的信息管理和处理的一个最优选择。分布式数据存储技术是目前信息处理中最常使用的技术，它带来的优势就是快速高效。

另外，在力求信息存储能力提升的情况下，云存储技术逐渐走入人们的生活视野，更成为打破传统常规的（如 U 盘、移动硬盘、存储卡等）重要存储信息方式。它的优势体现在灵活多变，在功能层面上不依赖于物理硬盘，而是基于平台服务，当前流行使用的云存储工具，如百度云、360 云盘等，不但在设备层面上展现出优越性，在功能、安全、访问速度以及数据所有权等方面也有了新的突破。

信息加工主要是结合用户提出的特定条件，对采集的数据信息从已存储的数据库中提取出所需信息，并通过归纳与整理对提取信息的准确性加以提升。此外，利用信息传输技术，可以将加工后的信息反馈给用户的客户端并进行传输。

数据的安全性，尤其是对于重要信息和数据而言，想要确保对数据信息真实性和完整性的安全控制，除了要加强用户和技术人员的安全意识，及时对相关数据进行检查，改善硬件的安全性能，更要制定满足“大数据”时代信息发展标准，积极开展安全技术的研发工作，构建信息技术安全体系。

2　技术分析及开发环境

2.1　技术分析

本研究提出了基于提高数据存储、平台稳定性及最大化提升安全、高效性能为目的，通过搭建一个分布式的数据采集系统，实现在大数据网络互联大背景平台下对海量数据进行采集、存储、计算和可视化等数据操作处理。

另外，也提出了相应的网络故障检测技术，并为其节点设计流量数据采集监控机制框架，确保数据的完整、准确。为了解决在 Hadoop 平台上电信运营商所面临的各种结构混合型的数据这种异构环境，该系统提出了一种改善在云计算环境中优化如苹果等性能节点的动态存储分配算法[4-5]。除此外，还对 MapReduce 等核心算法进行了优化操作。

2.2　数据处理平台搭建的集群和硬件环境

本系统基于的 Hadoop 平台，由 VMware Workstation 的三台虚拟机来搭建，包含一个主节点，以及三个从节点，这三个节点的系统均为 Centos6.5，涉及的集群环境图分别见表 1、表 2。

表 1　集群环境

Tab. 1　Cluster environment

框架	CDH 版本	Apache 版本
Hadoop	Cdh5.3.6-2.5.0	Apache2.7.6
Zookeeper	Cdh5.3.6-3.4.5	Apache3.4.12
HBase	Cdh5.3.6-0.98	Apache1.4.5
Hive	Cdh5.3.6-0.13	Apache1.2.2
Flume	Cdh5.3.6-1.5.0	Apache1.7.0
Kafka	2.10-0.8.2.1	2.11-0.11.0.3

表 2　硬件环境

Tab. 2　Hardware environment

	Node102	Node103	Node104
内存	2G	1G	1G
CPU	2 核	1 核	1 核
硬盘	20G	20G	20G

3　电信大数据分析平台的搭建与实现

3.1　平台的架构

本系统的整体架构可分为数据采集层、数据存储计算层、数据分析展示层等 3 个层次，如图 1 所示。数据采集层作为整个系统的数据源，通过编写 Java 代码的形式，生成虚拟的通话数据，将实时的数据通过 Flume 采集到 Kafka，最后提供给 HBase 消费；数据存储计算层是整个系统的核心部分，主要将采集过来的数据永久化的存储到 HBase 数据库中。通过 MapReduce 进行分布式计算，然后将解析后的结果导入 MySQL 数据库中，准备进行数据可视化；数据分析展示层主要是通过聚类分析算法，将结果按照分析的不同数据以柱状图或折线图等形式展现出来，并给出一些决策性的建议。

3.2　数据采集层

3.2.1　数据生产

本系统的数据全部由 Java 编程模拟产生，通过创建 Java 集合类存放模拟的电话号码和联系人。并随机选取两个手机号码当作“主叫”与“被叫”，产出相应字段数据。可随机生成通话建立时间及通话时长，再将产出的一条数据拼接封装到一个字符串中。最后，使用 I/O 操作将产出的一条通话数据写入到本地文件中。

3.2.2　Flume 数据收集模块

Flume 是一种针对分布式数据采集、传输等可靠性能较高的工具，通过其能实现数据到数据流的控制管理。本系统利用传输管道（Channel），在 NN 结点上对 Flume 进行配置时，通过将由数据源（Source）产生的程序收集（exec）相应模拟日志数据，再实现将其上传到安装目录中的 conf 文件夹后，系统配置成功，即可启动 Flume 模拟并传输数据到 Kafka 由其数据消费。

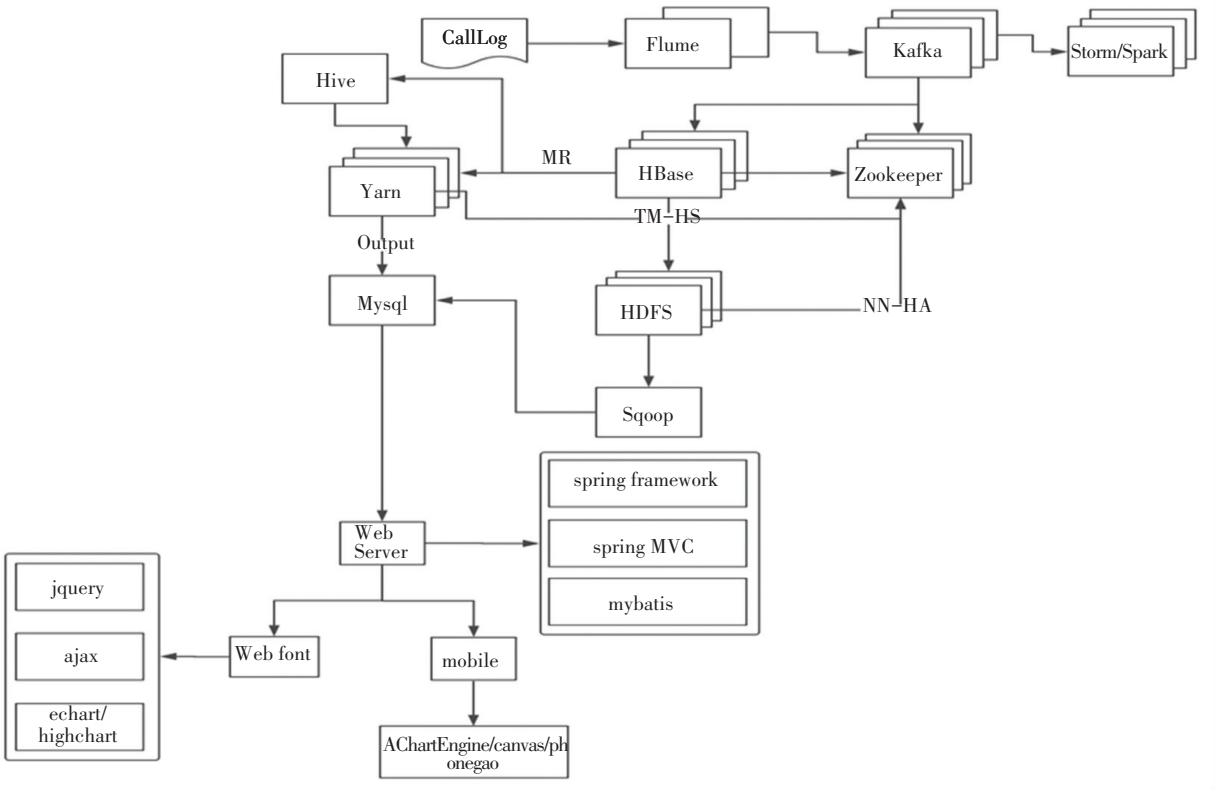

图 1 系统整体架构图

Fig. 1 Overall system architecture diagram

3.2.3 Kafka 缓存模块

Kafka 是一种能实现消息大规模吞吐量,并支持数据线上/线下操作的发布-订阅模式的系统。将 Flume 数据源收集(exec)的模拟数据存入到已创建的日志主题中,并为之指定主题分区数和副本数。Kafka 消费者负责分别将数据永久化或集群到 HBase 和 Flume 中。需强调的是,要将 Sink 组件的接收类型指定为 Kafka Sink,并为 Kafka 设置如主题、服务器地址以及端口号等。

3.2.4 高可用数据采集方案的设计

鉴于 HBase 的安全性和海量数据的存储能力,本系统选择 HBase 来作为 Kafka 的消费者。需要将实时数据通过 Flume 采集到 Kafka 提供给 HBase 消费。再编写操作 HBase 的代码,用于消费数据,将产生的数据实时存储在 HBase 中。

消费存储模块流程图如下图 2 所示。

3.3 数据存储计算层

3.3.1 数据存储

实现话单查询实时快捷化,精确定位话单故障点,及时受理客户投诉,并能对重点区域进行保护等,一直都是电信急于解决的问题。主要原因在于 IOE 结构无法扩展的局限性以及话单庞大数据量。鉴于满足对数据量的写入及读取高速化这种需求,在大数据平台上,可采用 HBase 这个分布式列式数据库[6]。

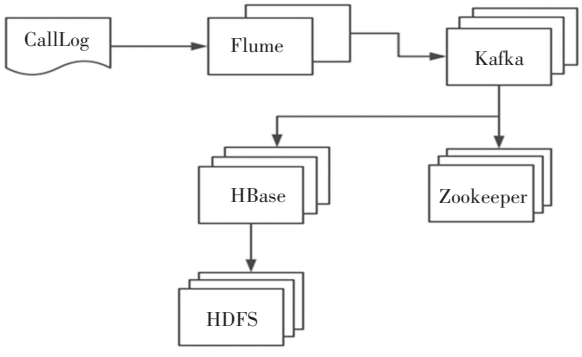

图 2 消费存储模块流程图

Fig. 2 Flow chart of consumption storage module

如 2G 话单,由 HBase 存储原理可知,每一列数据都需要按顺序依次录入包括 Row Key 的信息、时间戳的信息、列族的信息,然后再是列的信息,最后是值。考虑到重复信息对 I/O 无谓的消耗过于庞大,尤其是 Row Key、时间戳、列族等信息。因此,在优化时也要顾及此方面[7]。

对于用户的需求,重点要从用户拨通手机号码以及通话时长查询排查后,抽取话单的呼叫成功或

是种类字段信息等进行过滤，不影响查询过滤条件下，大幅度减少总字段数。经过测试发现，入库性能较之前提高了 10 倍之多[8]。

3.3.2　数据计算

在计算过程中，本系统不一定会采取一个业务指标对应一个 MapReduce-job 的方式。如果环境条件允许，会采取一个 MapReduce 分析多个业务指标的方式来进行任务，统计出想要的结果，然后将计算的数据结果保存在 MySQL 中，以方便进行数据分析和 Web 展示。再根据需求目标，按照时间范围（年月日），结合 MapReduce，统计出所属时间范围内所有手机号码的通话次数总和以及通话时长总和，具体统计如下：

（1）维度，即某个角度、某个视角；

（2）通过 Mapper 将数据按照不同维度聚合给 Reducer；

（3）通过 Reducer 拿到按照各个维度聚合过来的数据，进行汇总、输出；

（4）根据业务需求，将 Reducer 的输出通过 Output format 把数据输出到 MySQL 中；

（5）已知目标，那么需要结合目标思考已有数据是否能够支撑目标实现；

（6）根据目标数据结构，构建 MySQL 表结构，建表；

（7）思考代码需要涉及到哪些功能模块，建立不同功能模块对应的包结构；

（8）描述数据，一定是基于某个维度（视角）的，因此，需构建维度类；

（9）自定义 Output Format 用于对接 MySQL，使数据输出；

（10）创建相关工具类。

3.4　数据分析展示层

本层主要提供电信数据信息搜索查询、统计分析、报表显示等功能；服务层利用 Web 页面向用户提供服务。在完成配置后，撰写数据可视化代码。首先，测试数据通顺以及完整性，写一个联系人的测试用例。测试通过后，通过输入手机号码以及时间参数，查询指定维度的数据，并以图表展示。

3.5　定时任务

新的数据每天都会产生，都需要更新离线的分析结果，此时可以用各种各样的定时任务调度工具来完成此操作[9]。

4　平台测试与调优

本系统对数据的解析意在说明数据探索的重要性。现实的真实数据经常会出现各种不规整和异常的情况，本系统在对其解析时必须进行清洗和过滤等工作，这不但能发现数据在完整性和质量上的问题，而且对最终算法处理的结果也具有非常重要的意义，因此，在本系统中还是存在数据处理手段的安全性问题，以及 Hadoop 任务的运转效率等问题。由于 Hadoop 自身的一些特点，它只适合用于将 Linux 作为操作系统的生产环境。在实际应用场景中，对 Linux 系统的内核参数进行优化，也能够在一定程度上提升 Hadoop 任务的运转效率。因此考虑到大数据平台的特性做了具体调整，多图表测试展示图如图 3 所示。

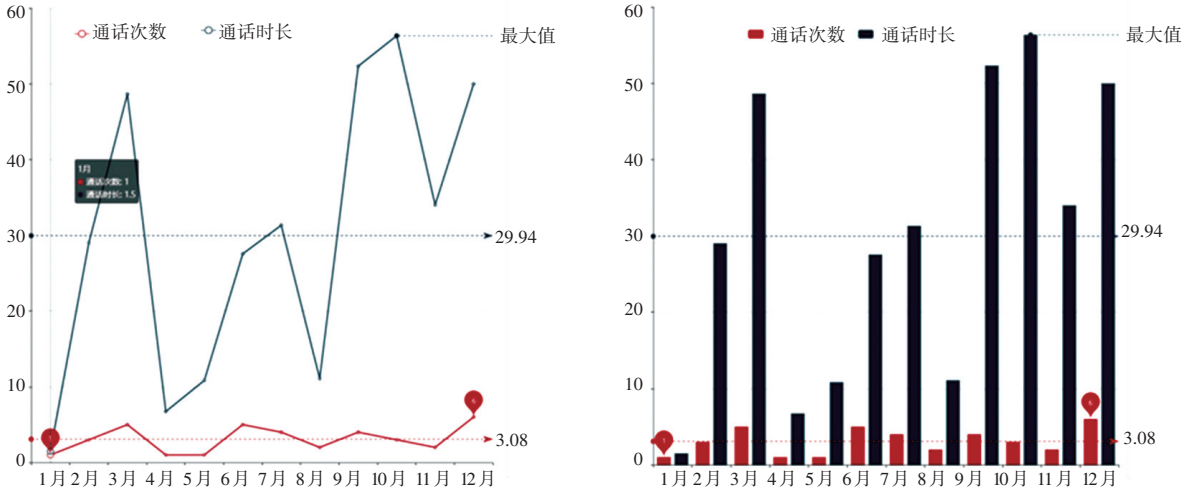

图 3　多图表展示图

Fig. 3　Multi-chart presentation diagram

（下转第 169 页）