# 广电用户大数据系统存贮需求分析

何 煌

（广东创新科技职业学院 广东东莞 523960）

摘要:通过大数据应用，把握广电用户群体的特征和收视行为习惯模式，了解客户的实际特征和实际需求，并提供个性化、精准化和智能化的推荐服务。为用户提供一种更直接、更方便、更个性化的用户体验，以此挽留客户、减少客户的流失。本文通过对媒体内容和用户服务两个方面的需求分析，结合大数据平台架构和数据治理流程，介绍了广电大数据平台的典型应用以及数据持续滚动的作用机制，同时对如何推动广电产业数字化进行了思考。

关键词: 大数据 大数据存贮分布式系统

## 1 引言

据2019年《中国广电有线网络技术及年度发展报告》数据，截至 2019 年第 3 季度末，全国的有线电视台用户数量为 2.1 亿，相对于 2018 年小幅减少。期中实际用户数为 1.9 亿，与 18 年同比减少 0.1 亿户，降幅达 5%；数字电视实际用户数占广电实际用户数比例为 90.5%，较去年同比降低了 0.4%。2019 年第三季度缴费用户下降为 1.5 亿户。

## 2 需求分析

新一代信息技术和互联网的迅猛发展，深刻改变有线网络行业，同时也带来前所未有的巨大挑战和机遇。这几年来，包括云、大、物、智等技术应用不断深入广电行业，新一代 5G 网络也全国各地推广开来，用户对于广电网络服务差异化、多样化、个性化的需求越来越迫切。与国外领先的网络运营商和国内 3 大电信运营商相比，广电业务和技术发展较为迟缓，差距不断拉大，时刻面临着用户量持续减少等前所未有的生存和发展压力，挑战严峻。随着互联网技术的快速发展和应用扩展，国家正式推进三网融合政策，3 大网络通过技术升级改造，使得各自功能趋于雷同，业务范围越来越相近，通赤互联互通、使得各业务资源共享，同时为终端用户提供数据、语音、电视等多样的服务。

新媒体业务的飞速发展，对传统媒体造成了巨大冲击，广电行业依靠资源稀缺形成的优势已经失去，在复杂激烈的竞争环境中，使得广电的客户流失问题变得异常突出。如何减少客户流失、挽留客户并挖掘客户潜在需求，是广电公司目前急需解决的问题。

在以往传统电视广播年代，广电公司“不知道用户是谁，不知道客户在何方，更不知道客户想收看什么”，故不能很好地把握用户需求。随着有线数字电视的不断推广与普及，广电公司具备了获取用户身份信息数据、实时收视数据的能力。通过网络终端设备和后台系统采集用户基本信息数据、用户收视数据、用户订单数据、用户账单数据等信息，广电行业已慢慢形成一个具有人口统计特征数据、终端用记内容使用记录、终端客户行为痕迹记录、终端搜索与需求信息记录、客户消费行为记录、客户社交交互与建议记录数据等真实巨量数据。利用此用户信息数据库，广电公司可以根据用户的特点，从人群、时间、地点、产品和付费方式等维度分析挖掘用户数据，对用户进行全面的画像。例如，从人群维度分析明确用户的年龄特征，如少儿、青少年、中年或老年等，以及分析收视语言是外语、普通话、粤语等；从时间维度分析用户每天观看电视的时长或用户观看某一电视节目的时长；从地点维度分析明确用户的收视常在地；从产品维度分析用户喜欢观看的电视频道或节目类型，如点播频道，回看频道或直播频道等，节目类型如体育、电视剧、购物、少儿等；从付费方式维度分析用户是收费用户还是免费用户。通过大数据应用，把握广电用户群体的特征和收视行为习惯模式，了解客户的实际特征和实际需求，并提供个性化、精准化和智能化的推荐服务。为用户提供一种更直接、更方便、更个性化的用户体验，以此挽留客户、减少客户的流失。

## 3 系统架构选型

信息化和经济全球化相互促进，互联网已经融入社会生活方方面面，深刻改变了人们的生产和生活方式。我国正处在这个大潮之中，受到的影响越来越深。

大数据关键技术包括大数据存储、处理和应用等多个方面。根据其处理的过程，又可拆分为采集数据、预处理数据、存储数据、分析与挖掘数据、具体应用五大环节。在广电有线网络中中，大数据技术发挥着重大作用，延伸到运营生产、客户服务、管理运营等多项业务。

一者，在充分借助公司地缘优势、数据优势，各地的分公司可通过推动与数据后台的接口，共享广电有线网络历年积累的行业大数据。二者，针对广电家庭客户喜好习惯的分析，了解客户存在所有样本个体，再使用记录个体在不同时间行为记录研究。利用最新的大数据技术，与外部数据和应用数据进行关联，提供业务和个性化服务推荐的后台支持。最后，在产品分类标签和客

户用户画像分类为基础，从分类信息研判客户爱好，和对潜在行为的分析预估，再以人工智能算法进行喜好内容客户推荐。

通过收视行为分析，用户活跃度分析，对客户服务进行分级定义，挖掘分析用户相关数据，对用户数据进行标签化，建立一个用户画像模型，并提供标签的增加和删除。以此为基础，建立分类模型，预测用户是否值得挽留，并将预测结果作为用户画像的一个标签。通过数据分析建立客户服务分析模型，一方面可以给用户提供更好的服务，另一方面可以进行客户流失预测，从而支撑用户挽留工作，最终提高用户使用粘度，为广电业务开展和拓展提供有力支撑。

在对广电公司客户相关的海量数据进行分析前，需要考虑采用何种存储技术保存数据，以便后续的数据查询和分析。因此，需要先了解大数据的几类主流的存储技术。

在大数据应用中，对海量数据采集、清洗后，需要确定可以将数据长期进行保存的存储方式，同时也应考虑一种组织管理数据的方案以便业务上的查询使用，最后也需要权衡是否需要使用内存存储和处理方式从而提高性能。目前大数据存储解决方案较多，既有商用的 AWS S3 和 EMC 系列产品，也有开源的 HDFS、Swift 和 Alluxio 等。

AWSS3 存储形态能够方便地进行横向扩展，以适应大量用户高并发访问的场景，但是不支持随机位置读写操作，只能整个文件统一操作。HDFS 是一种易于扩展的分布式文件系统，基于“移动计算比移动数据更经济”的设计理念，可使大量普通 PC 上进行搭建，节约不必要的投资，并具备可靠数据容错能力，有效减少运营维护成本。HBase 更适合海量数据随机读、写的业务场景，适合存储海量稀疏数据。EMC 则提供了支持 PB~ZB 级各类数据存储的高端产品和解决方案，具有极佳的数据保护安全性，但由于是商用产品，故应用成本较高。Swift 支持多租户模式，可靠地存储数量非常多的大小不一的文件，但并针对大型文件作优化处理。Alluxio 是以内存为中心的虚拟分布式存储系统，核心思想是将存储与计算分离，使 Spark 等框架更专注于计算，从而达到更高的执行效率。

广电公司用户数据主要特点是用户量巨大，相关信息文件也非常大，基础数据一经写入，不会再频繁修改，故选用 Hadoop 开源框架的 HDFS 分布式文件系统作为数据存储平台更为合适。

3.1 大数据存贮工具 Hive 介绍

Hive 是基于 Hadoop 的一个数据仓库工具，其优势是学习门槛较低，一般的数据人员便可以结构化脚本语句实现快速 MapReduce 统计，而不必使用 Java 语言云开发相应的程序，让 MapReduce 的使用变得更加简单。故 Hive 是数据仓库应用当中十分适合进行数据统计分析工具。

在 Hadoop 中 MapReduce 的主要工作原理是将计算任务切分成多个小单元降低成本同时提高扩展性。但是使用 MapReduce 人员要求较高，须掌握 Java 等语言面向 MapReduce API 编程接口进行编程，才能较为熟练地处理。此外，数据存储在 Hadoop 的 HDFS 中，并没有存贮 Schema 信息。如果数据从传统的关系型数据库迁移至 Hadoop HDFS 进行应用，会产生 Schema 信息的缺失，这时候 Hive 就成为了传统数据架构和 Hadoop MapReduce 之间的桥梁。

3.2Hive 原理架构

Hadoop 生态系统包含了若干用于协助 Hadoop 的不同的子项目（工具）模块，如 Sqoop、Pig 和 Hive。以下为它们的主要用途。Sqoop：用于在 HDFS 和 RDBMS 之间来回导入和导出数据。Pig：用于开发 MapReduce 操作的脚本程序语言的平台。Hive：用于开发 SQL 类型脚本用于做 MapReduce 操作的平台。

同时在 Hadoop 生态圈中，有多种方法可以执行 MapReduce 作业。传统的方法是使用 Java MapReduce 程序结构化、半结构化和非结构化数据。针对 MapReduce 的脚本的方式，可使用 Pig 处理结构化和半结构化数据，而 Hive 查询语言则采用 Hive 为 MapReduce 的处理结构化数据。

Hive 定义了简单的类 SQL 查询语言，使用类似脚本语言来查询数据。同时，这个它允许熟悉 MapReduce 的开发人员，通过自定义的 Mapper 和 Reducer 处理内建的 Mapper 和 Reducer 无法完成的复杂的分析工作。

Hive 的体系架构设计遵从主从架构的设计模式，架构如图 1 所示。

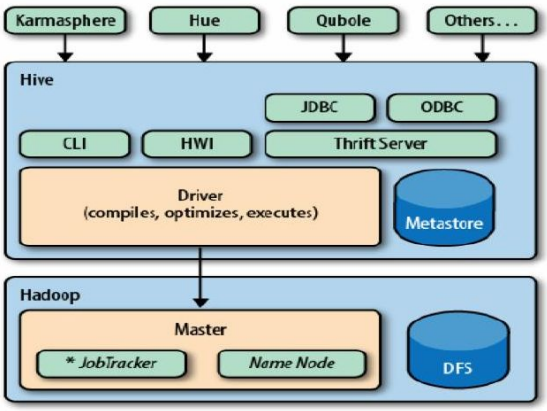


图 1Hive 架构

3.3 Hive 访问接口

Hive 提供了 3 种访问方式，命令行方式（Command Line Interface，CLI），浏览器方式（Hive Web Interface，

HWI）以及 Thrift Server 客户端连接方式。具体介绍如下。

CLI 是 Command Line Interface 缩写，即控制台命令行接口，是基础的连接方式，使用“hive”命令连接。CLI 启动时会同时启动一个 Hive 副本，相当于“hive –service cli”。启用 CLI 只需要在命令行下执行“$HIVE_HOME/bin/hive”命令即可。在命令行状态下执行“hive –H”可以查看 CLI 选项。

Thrift Server 提供 JDBC（Java Database Connectivity，JDBC）和 ODBC（Open Database Connectivity，ODBC）接入的能力，用于进行可扩展跨语言服务的开发，Hive 集成了该服务，可以让不同的编程语言调用 Hive 的接口。对于 Java 程序访问 Hive 提供了 JDBC 驱动，对于其它应用程序，Hive 提供了 ODBC 驱动。

Hive Web Interface 是 Hive 命令行接口（CLI）的一个 Web 接入方式。HWI 的特点是相对于命令行方式界面友好，适合不太熟悉 Linux 命令行操作方式的数据分人员。通过浏览器打开“http://主机 IP:9999/hwi/”网址可访问 Hive 服务，HWI 默认端口为 9999。

3.4 MetaStore

元数据服务组件，这个组件存储 Hive 的元数据，Hive 的元数据存储在关系数据库里，Hive 支持的关系数据库有 Derby、MySQL。

Hive 中的数据分为两部分，一部分是真实数据，一般存放在 HDFS 中；另一部分是真实数据的元数据，单独存储在关系型数据库中，如 Derby、MySQL 等。元数据用于存储 Hive 中的数据库、表、表模式、目录、分区、索引以及命名空间等信息，是对真实数据的描述。元数据会不断更新变化，所以不适合存储在 HDFS 中。实现任何对 Hive 真实数据的访问均须首先访问元数据。元数据对于 Hive 十分重要，因此 Hive 把 MetaStore 服务独立出来，从而防止 Hive 服务和 MetaStore 服务耦合，以保证 Hive 运行的健壮性。

在默认情况下，Hive 会使用内置的 Derby 数据库，其只提供有限的单进程存储服务。此时，Hive 不能执行 2 个并发的 HiveCLI 实例，通常被应用于开发、测试环境中。对于生产环境，需要使用 MySQL 关系型数据库。

3.5 Driver

Driver 是 Hive 的核心组件，也是整个 Hive 的核心，该组件包括 Parser（解释器：将 Hive SQL 转换为抽象语法树），Compiler（编译器，将语法树编译为逻辑执行计划），Optimizer（优化器，对逻辑执行计划进行优化，成为更优逻辑计划），Executor（执行器，将逻辑计划切成对应引擎的可执行物理计划，调用底层执行框架执行），由此看出，Driver 的主要功能是将用户编写的 HQL 语句进行解析、编译优化，生成执行逻辑计划，并提交给 Hadoop 集群进行处理。

3.6 Hive 的特点

Hive 让普通数据库用户从现有的，基于关系型数据库和结构化查询语句的数据基础架构转移到 Hadoop 上，对于大量的 SQL 用户而言，Hive 提供了一个被称为 Hive 查询语言（HQL）查询和存储在 Hadoop HDFS 上的数据，这减少了开发人员的学习成本，使得开发人员可以使用一种熟悉的语言操作和分析存储在 Hadoop HDFS 上的数据。

Hive 具有如下特点。

（1） HQL 与 SQL 有着相似的语法，大大提高了开发效率。

（2） Hive 支持运行在不同的计算框架上，包括 YARN、Tez、Spark、Flink 等。

（3） Hive 支持 HDFS 与 HBase 上的 ad–hoc。

（4） Hive 支持用户自定义的函数、脚本等。

（5） Hive 支持 Java 数据库连接 JDBC 与开放数据库连接驱动 ODBC，建立了自身与 ETL、BI 工具的通道。

Hive 还具有以下优势。

（1） 可扩展：Hive 可以自由扩展集群的规模，一般情况下无须重启服务。

（2） 可延展：Hive 支持用户自定义函数，用户可根据自己的需求来编写自定义函数。

（3） 可容错：Hive 良好的容错性使得节点出现问题时 SQL 仍可完成执行。

简而言之，当使用 Hive 时，操作接口采用类 SQL 语法，提高了快速开发的能力，避免了编写复杂的 MapReduce 任务，减少了开发人员的学习成本，而且扩展很方便。

3.7 Hive 的适用场景

大数据集的批量作业，是 Hive 工具的最佳使用场合，最典型应用就是网络日志分析。除了底层封装了 Hadoop，Hive 还使用类似于 SQL 的 HiveQL 语言实现数据查询，它本质是是一种数据仓库处理工具。故 Hive 的数据会存储在 Hadoop 兼容的文件系统：Amazon S3、HDFS 等等。在加载数据过程中，Hive 不会对数据进行任何改动，只会将数据移动到 HDFS 中预先设定的 Hive 目录下，这说明它并不支持对数据添加改写，数据都是在加载时确定的。

3.7 Hive 与传统数据库的区别

在 Hadoop 诞生前，大部分的数据仓库应用程序都还是基于关系型数据库实现的，而数据仓库应用程序则是建立在数据仓库上的数据应用，包括报表展示、即席查询、数据分析、数据挖掘等。数据仓库数据源自数据库

（下转第 104 页）

（上接第 92 页）
而又不同于数据库，主要区别在于数据仓库适合联机分析处理（On-Line Analytical Processing，OLAP），通常是对某些主题的历史数据进行分析；而数据库适合联机事务处理（On-Line Transaction Processing，OLTP），通常是在数据库联机时对业务数据进行添加、删除、修改、查询等操作。Hive 被设计成数据仓库，其早期版本或新版本在缺省情况（系统默认状态）下并不支持事务，一般来说并不适合 OLTP。

Hive 与传统关系型数据库（Relational Database Management System，RDBMS）有很多相同的地方，包括查询语言与数据存储模型等。Hive 的 SQL 方言一般被称为 HiveQL，简称 HQL。HQL 并不完全遵循 SQL92 标准，如 HQL 只支持在 From 子句中使用子查询，并且子查询必须有名字。最重要的是，HQL 须在 Hadoop 上执行，而非传统的数据库。在存储模型方面，数据库、表都是相同的概念，但 Hive 中增加了分区和分桶的概念。

Hive 与 RDBMS 也有其他不同的地方，如在 RDBMS 中，表的 Schema 是在数据加载时就已确定，若不符合 Schema 则会加载失败；而 Hive 在加载过程中不对数据进行任何验证，只是简单地将数据复制或移动到表对应的目录下。这也是 Hive 能够支持大规模数据的基础之一。事务、索引以及更新是 RDBMS 非常重要的特性，鉴于 Hive 的设计初衷，这些特性在开始之初就不是 Hive 设计目标。与 RDBMS 相比，Hive 具有更多支持的存储类型，包括纯文本文件、HBase 中的文件；另外 Hive 还利用关系数据库，将元数据保存在其中，从而缩减了在查询执行语义的检查时间；除此之外存储在 Hadoop 文件系统中的数据，也可以被 Hive 直接使用；除了内置了大量用户函数：如操作字符串、时间处理；还支持其他的数据挖掘工具，支持 UDF 函数，提供类 SQL 的查询方式，并将其转换为 MapReduce 任务在 Hadoop 集群上执行。

## 4 结语

本文阐述了广电大数据用户画像需求背景，由此对当前市场上常见的几种大数据存储和分析技术进行介绍，着重介绍 Hadoop 与 Hive 集成的大数据存储和分析架构技术，从 Hive 的发展、原理架构、主要特点、与传统 DBMS 的对比等方面对 Hive 进行了深入探讨。

**参考文献：**

［1］中国广电有线网络技术及年度发展报告(2019)2020.10.全国互联网与音视频广播发展研讨会暨中国数字广播电视与网络发展年会(2020 年特辑)

［2］卡普廖洛等. Hive 编程指南[M].曹坤,译. 北京：人民邮电出版社,2013.

［3］孙帅,王美佳. Hive 编程技术与应用[M]. 北京：水利水电出版，2018.

作者简介：何煌，1973 年 8 月出生，男，汉族，户籍：广东省广州市，软件工程硕士、软件专业讲师、主要研究方向：大数据技术、软件开发等。