

Analysis of IPTV User Behaviors with MapReduce

Joohee Kim*, Chankyou Hwang*, Eunkyong Paik*, Youngseok Lee **

* KT, 17 Woomeon-dong, Seocho-gu, Seoul, Korea

** Chungnam National University, Daejeon, Korea

{joohee.kim, chankyou.hwang, eun.paik}@kt.com, lee@cnu.ac.kr

Abstract— IPTV increases opportunities of earning profits for ISPs, and opens new content markets to Content Provider (CP) and Program Provider (PP). In order to expand the markets and maximize profits, IPTV providers need to create more appealing programs and better broadcasting schedules by understanding the usage patterns and user behaviours. They can gather log data from the set-top box (STB) and network devices. However, due to huge amounts of log files in large-scale IPTV system, it is not easy to perform deep analysis for various metrics on time. The volume of log data is expected to increase rapidly because of the explosive growth of users. For the efficient management of analysis jobs for diverse metrics under a large set of log data, in this paper, we propose MapReduce-based IPTV log analysis methods on the cloud computing platform, called *Hadoop*. We carried out experiments with a small testbed and a large one with Amazon service, and found that the MapReduce-based approach outperforms the DBMS by 13 times. From the MapReduce analysis, we also present the characteristics of hourly usage pattern, channel variation, and regional features of IPTV users.

Keywords— MapReduce, Hadoop, cloud computing, IPTV, log, user behavior.

I. INTRODUCTION

Due to the high-speed broadband network, IPTV as well as VoIP have been merged to the Internet service and deployed throughout the world. In Korea, it is reported that more than two million people use IPTV from three national Internet Service Providers (ISPs) in 2009. IPTV subscribers actively use the PP program, VoD contents, and interactive services (e.g. newspaper, weather, retail sales and games), whereas traditional TV users passively watch the TV programs scheduled by PP. IPTV changes the user pattern of watching TV. For IPTV providers, it is essential to understand the users' TV viewing habits and find the characteristics of user behavior, because users tend to select an IPTV provider according to qualities and quantities of the interesting contents. Fortunately, IPTV providers can collect and analyze log data from STBs and network devices to extract useful information of user behavior.

Many operation and management supporting systems on a high-performance central DB server are deployed in IPTV providers. Typically, providers run business intelligence applications on a central server to analyze IPTV log data. However, in a large-scale IPTV system, it is not easy to process large log files for various analysis metrics on a single

server. It not only strongly requires high performance servers, expensive software licenses and sophisticated user-defined programs, but also needs high operational budget to maintain and manage servers. Due to the popular cloud computing platform, we could mitigate the cost of installment, operation, upgrade, and repair of servers. In addition, we could also easily build a scalable analysis system handling a large data set with elastic resources of cloud computing.

Recently, there are many deployments of distributed parallel computing on a cloud computing platform such as Amazon Elastic Compute Cloud (EC2), Google AppEngine, and Microsoft Azure. Google has first developed a distributed parallel processing programming model, MapReduce [1] and a distributed filesystem, Google FileSystem (GFS). Yahoo! is leading an open-source MapReduce software platform, Hadoop [2]. Since Hadoop is useful for large data processing and automatic failover, it has become popular for analyzing business intelligence from massive data.

This paper proposes MapReduce-based IPTV user behavior analysis methods on Hadoop. Our MapReduce-based analysis programs could characterize several properties of aggregate viewing sessions, content popularity, user patterns across regions, and individual user's viewing sessions. From the experiments with the local testbed as well as Amazon EC2, we found that the MapReduce-based approach could significantly reduce the computation time for analyzing large log data, compared with running SQL queries on a single DB server.

The remainder of this paper is organized as follows. In Section 2, we describe the related work. Our MapReduce-based analysis method is explained in Section 3, and its experimental results are presented in Section 4. Finally Section 5 concludes this paper.

II. RELATED WORK

There are a few studies on IPTV user behavior analysis in commercial IPTV systems. Yu et al. [3] analyzed a large VoD system in China Telecom to understand VoD user activities and the content access patterns. Cha et al. [4] presented a study on IPTV live channels. Cha used the log data from network devices, and analyzed characteristics of channel session span, channel popularity dynamics, and channel zapping pattern. Qiu et al. [5] analyzed on/off user trend and channel popularity using STB logs, and devised a channel popularity model. In this paper, we develop a MapReduce-based analysis method and analyze the user log data of 1.7 million households obtained from STBs and the service

platform. We examine user behavior not only for live channels but also for VoD and interactive services.

Recently, many MapReduce programs have been developed by Google, Yahoo!, Amazon, Facebook, Salesforce, and Datameer to analyze “big” data of text files or web logs. Lee et al. [6] developed an Internet flow analysis method with MapReduce, and Chen et al. [7] proposed a log data analysis method with Hadoop for snort log data. Performance comparison between MapReduce and parallel DBMS has been performed in [8] arguing that tasks such as selection, aggregation, and join are better in parallel DBMS, although the process to load data in parallel DBMS took much longer than Hadoop. On the other hand, Dean [9] asserted that MapReduce is a highly effective and efficient tool for large-scale fault-tolerant data analysis. In this work, we propose an IPTV user behavior analysis method in large-scale IPTV system with Hadoop, and present the performance evaluation results of Hadoop and RDBMS.

III. MAPREDUCE-BASED IPTV LOG ANALYSIS

A. Overview

We briefly introduce the overview of IPTV services and its system architecture of KT. The IPTV service consists of live channels, VoD, and interactive programs. KT provides two types of VoD services for different bandwidth group of subscribers. One is the download & play service (called D&P VoD) for low-speed network subscribers such as Digital Subscriber Line (DSL). The other is the real-time streaming service (called VoD) for the subscribers using above 20 Mbps speed such as VDSL and FTTH.

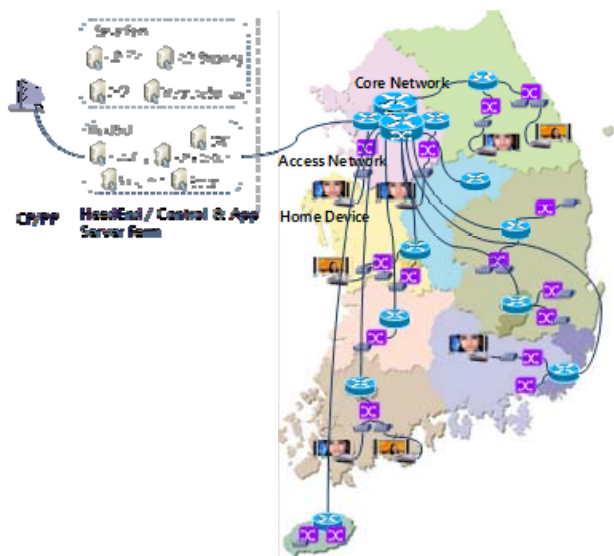


Figure 1. IPTV architecture of KT

Figure 1 shows the IPTV system architecture of KT that is divided into 4 parts: HeadEnd, server farms, access/core networks, and home devices. HeadEnd acquires contents and applications from CP/PP, creates program schedules, and distributes contents to server farms. Various servers and equipments such as DRM (Digital Rights Management) server,

accounting server, and content/application management server, and encoders are operated. In server farms, a DHCP server provides IP addresses to authenticated STBs. Content server farms are separated and located in different regions at the edge of the core network in order to improve the efficiency of network usage and the performance of popular contents delivery. Backbone and access networks support IP multicasting to carry live channel services. Home devices are STBs provided by KT. From the STBs and HeadEnd servers, we collect the service logs for 1.7 million subscribers for a month.

We designed and implemented multiple MapReduce-based analysis programs to compute hourly online user behaviors across genre/service type, content popularity, user pattern across regions, and watching time per user on genre/service type. “Service type” means the kinds of IPTV services such as live channel, VoD, and interactive service. Each service type has different genre sets. Live channels can be grouped by the genre such as free (national free-to-air broadcasted channels), cine/series (moves and series), edu (education), ani/kids (animation and children’s programs), docu (documentaries), adv (advertisement channel). VoD has similar genre grouping with live channels (free, edu, docu, cine/series, ani/kids, adult and etc.). Interactive services can be divided into five kinds of genre: information, entertainment, education, communication, and commerce.

In the MapReduce programming model, computation with map and reduce functions takes a set of input key/value pairs, and produces a set of output key/value pairs. A map written by users takes an input pair and produces intermediate key/value pairs. The Hadoop MapReduce library will shuffle and sort the intermediate values according to the same key. A reduce will merge the intermediate values for smaller values. To implement the statistics analysis programs for each behavior, we have to parse the format of input files, determine appropriate input key/value pairs for each MapReduce program and organize the procedure for multiple MapReduce jobs.

On the Hadoop framework, we have developed 24 MapReduce jobs that deal with sessions (channel, VoD, interactive service, and D&P VoD), service type (hourly viewing list, hourly summation), genre (hourly viewing list, hourly summation), duration (per service type, per genre), ranking, and regional statistics.

B. Hourly online user population

The analysis of hourly online user population provides two types of breakdown results. One is service type-breakdown, and the other genre-breakdown. When we tally online users hourly by service type share, which sums up the user count for service type, the key/value pair for the final MapReduce program will be ((Service_Type, Time), Count). When we analyze online users by genre sharing, the key/value pair will be ((Service_Type, Genre, Time), Count).

Figure 2 shows the procedure of the MapReduce-based program that performs hourly online user analysis per service type. The whole procedure can be divided into three parts. The

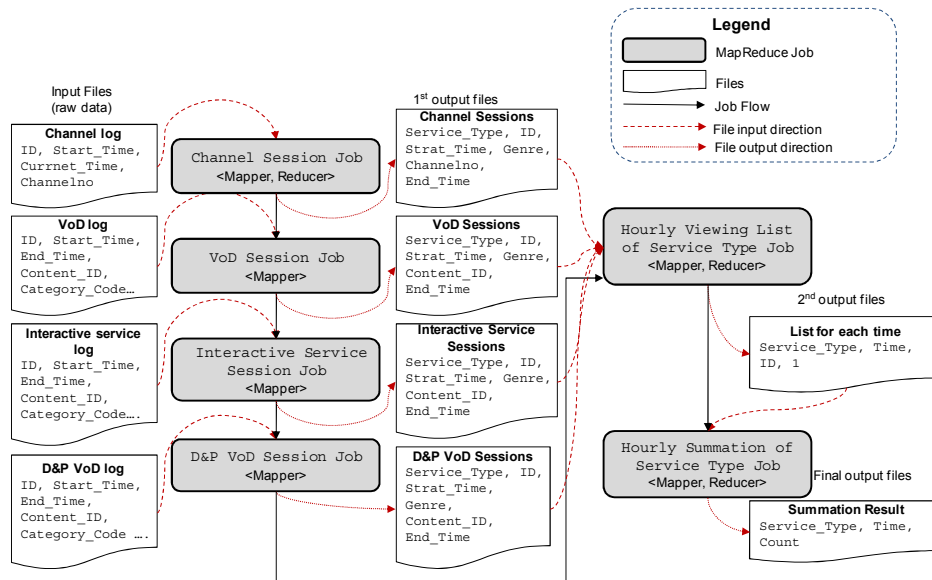


Figure 2. A flow diagram for computing hourly online users across service type

first part is to extract user viewing sessions and genres from each service log. A user-viewing session is defined by a starting time when a user plays a content and an ending time when the user stops watching. In the first part, there are four MapReduce jobs to execute: Channel Session Job, VoD Session Job, Interactive Service Session Job, and D&P VoD Session Job. Channel Session Job consists of a map and a reduce tasks, whereas no reduce functions are necessary for other session jobs, because other logs regarding VoD, interactive service, and D&P VoD include user watching session data in a line. Through four jobs, we can derive all of user viewing sessions in the same output format. The second part is to extract the hourly viewing list from every user session. For

this purpose, Hourly Viewing List of Service Type Job in Fig. 3 processes output files of jobs in the first phase. The third part is to summarize the count of users viewing the same service type at the same time. The output file in the third part is the final output file that shows the statistics results for online users across service type per each hour.

C. Content popularity

Content popularity analysis is to extract content-ranking statistics. The MapReduce program sums up the row count for each content. The key/value pair is ((Service_Type, ContentID), Count). Before executing this MapReduce program, we run the Channel Session Job as a

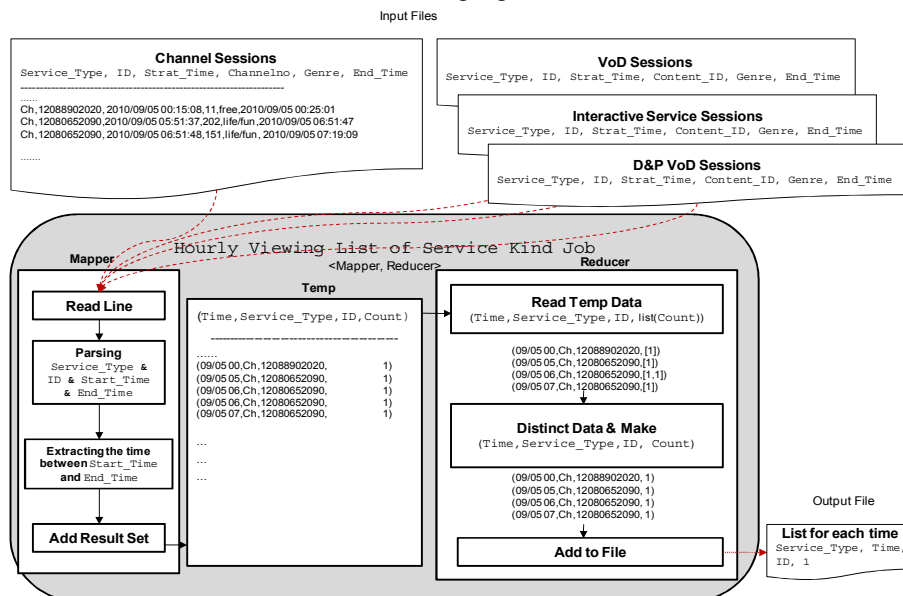


Figure 3. The procedure of Hourly Viewing List of Service Type Job

previous stage. The Channel Session Job generates the channel sessions file as the output from the input files of service logs. From the output files, we can find favorite contents and the characteristics of content popularity.

D. User pattern across regions

We have also implemented two MapReduce-based statistics programs for analyzing user pattern across different regions. One is to analyze daily online user population across regions. It is also performed for service type-breakdown or genre-breakdown. The other is to analyze users' viewing times across regions. In order to match a user to a region, we have to perform the "join" job, which joins user's region to service session data. The Join Job receives channel sessions file and other service log files (VoD, Interactive Service, and D&P VoD) that have the session data in a row and user address files, and then performs reduce-side join to find the sessions with address code output files. Finally, we sum up the row count for service type, genre, and region through Summation of Genre across Regions Job.

The key/value pair in this analysis is ((Service_Type, Genre, Region), Count). On other hand, key/value pair in service type breakdown analysis is ((Service_Type, Region), Count). Analysis of daily online users across regions has one more job to perform than that of viewing times across regions. The job named Daily Viewing List of Service Type Job similar to Hourly Viewing List of Service Type Job extracts the daily viewing list.

E. Individual user analysis

For analyzing the individual user's attention span on genre or service type, we have implemented another MapReduce-based statistics program. The procedure of this analysis program is as follows. First, we extract the session data from the logs in the same way like the first part of online user analysis. Second, we process the session data as input files and sum up the user's span on service type or genre in a day. The key/value pair is ((ID, Service_Type), Time) for service type-breakdown and ((ID, Service_Type, Genre), Time) for genre-breakdown. At last, we find the average of the time on service type or genre for an individual user. The input file is the output of the second step described above. The key/value pair is the same as that of the second step.

IV. EXPERIMENTAL RESULTS

A. Experiment environment

For the performance evaluation of IPTV log analysis with MapReduce, we built up two Hadoop testbeds. One is a small Hadoop testbed consisting of a master node and 10 slave nodes. Each node has octo-core 2.93 GHz Intel i7 CPU, 16 GB memory, and 1 TB hard disk. All Hadoop nodes are connected with 1 Gbps Ethernet cards. The other consists of 100 Amazon EC2 virtual nodes. Each node has 2 virtual core 2.66 GHz CPU, 7.5 GB memory, and 500 GB hard disk. We

used Hadoop 0.20.3 version with the HDFS block size of 64 MB and the replication factor of three. We have collected IPTV log data of 1.7 million subscribers for a month (Sept. 2010) of 68.7 GB: 1-day data corresponds to 2.2 GB, and 1-week 17 GB.

For the comparison, we configured two DBMS servers to execute SQL commands. The low-performance server has four dual-core 1.9 GHz CPUs, 16 GB memory, 272 GB internal hard disk and 320 GB external hard disk. The high-performance one has 12 dual-core 4.2 GHz CPUs, 32 GB memory, 548 GB internal hard disk and 3TB external hard disk. Two servers connected with 100 Mbps Ethernet cards run RDBMS called Tiberio.

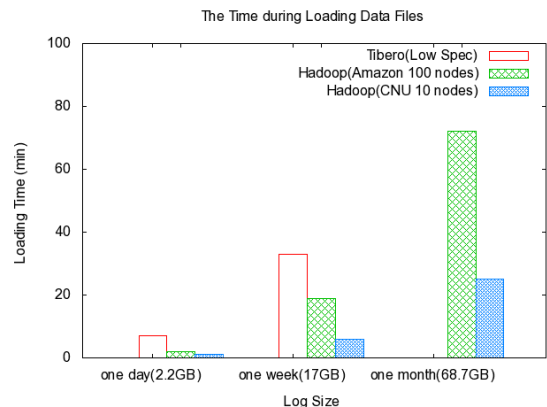


Figure 4. Comparison of data loading time on Hadoop and DBMS

B. Performance evaluation

1) Log file loading time: Figure 4 shows the result of data loading time on Tiberio (Low spec¹) and Hadoop. On Tiberio, it took 7 minutes to load one-day log data and 33 minutes for one-week log data excluding index generation time. We did not measure the loading time of one-month log data because it took too long. On the other hand, it took only 1 minute to load data on CNU 10 Hadoop nodes (2 minutes on Amazon 100 Hadoop nodes). For one-week log, it took 6 minutes on CNU 10 nodes (19 minutes on Amazon 100 nodes). For one-month log, it took 25 minutes on CNU 10 nodes, whereas 72 minutes on Amazon 100 nodes. Compared with Tiberio, the data loading time on Hadoop has decreased by 44 ~ 88%. This result confirms that Hadoop is faster than DBMS excluding the index generation time in [8].

2) Computation time of hourly online user population: We compare the job completion time for computing the online user population per hour in Figure 5. For one-week log data, Amazon EC2 of 100 nodes took 11 minutes whereas Tiberio (High spec) spent 142.2 minutes. Amazon outperformed Tiberio by 13 times for one-week data, whereas 1.1 times for one-day log data. For one-month log, we failed to run the

¹ Since Tiberio (High spec) for commercial services had no enough local disk to load all log files at one time, we did not perform loading experiment on Tiberio (High spec). Instead, we processed the channel log data directly from its storage and loaded small log files several times.

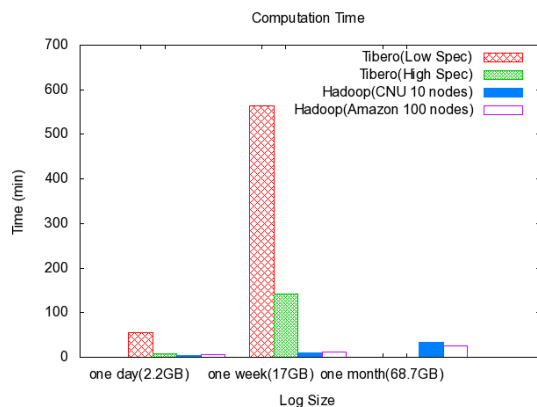


Figure 5. Hourly online user statistics computation time

analysis job on RDBMS at Tiberio. CNU 10 nodes took 34 minutes; 25 minutes on Amazon. Comparing the CNU testbed with Amazon, it is observed that the performance improvement is not significant considering the number of nodes in Amazon EC2. From Fig. 6 shows the computation time on different number of Hadoop worker nodes. We can see that the MapReduce job on 10 nodes finishes 10.3 times faster than on a single Hadoop node for one-month log analysis. As the file size becomes large, the scalability of parallel computing powered by Hadoop is obviously observed.

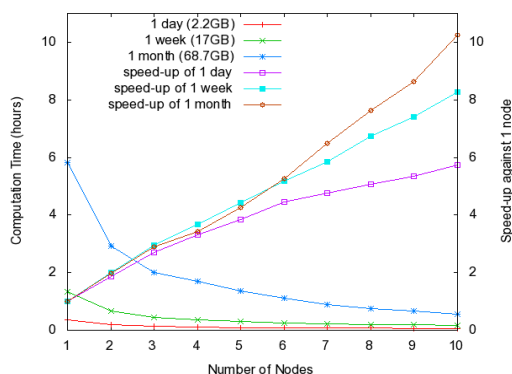


Figure 6. Computation time of hourly online user statistics under varying Hadoop worker nodes

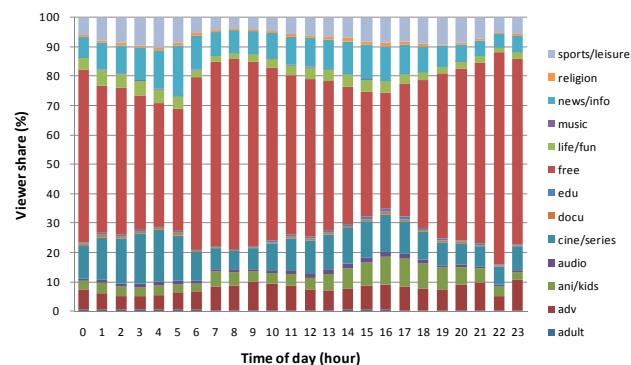


Figure 8. Hourly viewer share across channel genre over a day (Sept. 8 2010)

C. Statistics

1) IPTV user population: From the statistics experiments with 24 MapReduce programs, we present interesting IPTV usage pattern results. Figure 7 shows online user population across service type for a week. The graph shows a diurnal pattern of using the service. We can see that the channel graph begins to rise from about 7 AM and has a peak of 10 PM in weekday. On the weekend, the number of online users ramps up from early morning, particularly increases in the daytime. Among four service types, 2/3 of online users are using the channel service. In Fig. 8, we show the percentage of viewer share per each channel genre every hour for a representative day. Free genre accounts for over 50% of TV viewing instances. Cine/series and news/info genres are also popular. From 3 PM to 6 PM, ani/kids genre viewers become rising due to the increase of viewing children programs. As the result of VoD services, free genre accounts for 50% ~ 70% of all VoD viewers. This is caused by users who missed the programs on time. We can also find the fraction of ani/kids genre increases around 8 AM and 7 PM. For interactive services, entertainment genre (e.g. games and karaoke) is the most popular, followed by information (e.g. newspaper and weather) and commerce (e.g. on-line banking services and stock). Around 9 AM and 11 PM users preferred to use commerce services.

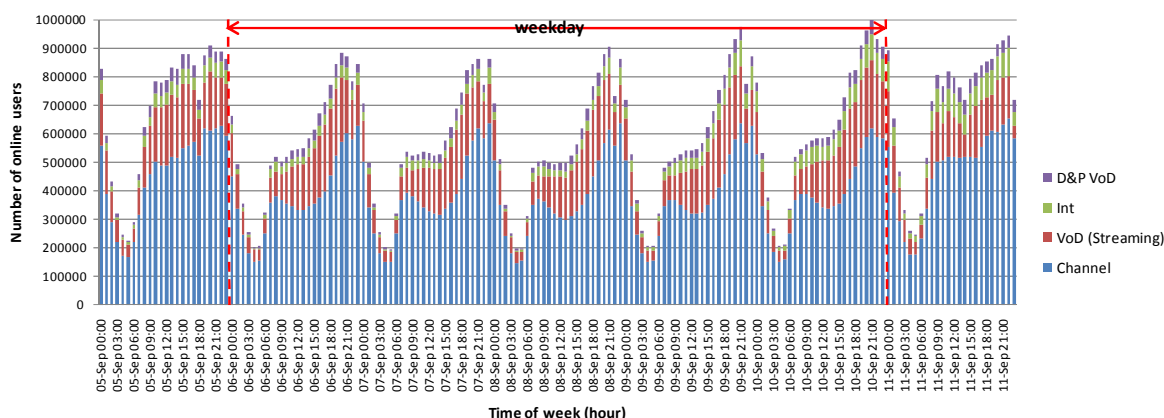


Figure 7. Number of hourly viewer population across service type over a week period (Sept. 5 2010 ~ Sept. 11 2010)

2) Contents popularity: Table 1 explains the result of channel content popularity analysis. Over the different periods of a day, a week, and a month, we computed top 10 channels. From the result we found that channel ranking changes but top 10 channels are stable. An interesting observation is that the home-shopping channels 8 and 10 providing advertisements of merchandises are ranked high. This result implies the importance of channel number allocation, because people tend to push up or down the channel buttons on the remote controller.

Table 1. Channel ranks over the different period²

Ranking	Day	Week	Month
1	7 (KBS2)	7 (KBS2)	7 (KBS2)
2	6 (SBS)	6 (SBS)	6 (SBS)
3	4 (Guide Channel)	11 (MBC)	11 (MBC)
4	11 (MBC)	4 (Guide Channel)	4 (Guide Channel)
5	9 (KBS1)	9 (KBS1)	9 (KBS1)
6	102 (OCN)	102 (OCN)	102 (OCN)
7	103 (SuperAction)	103 (SuperAction)	103 (SuperAction)
8	24 (YTN)	8 (Lotte Homeshopping)	8 (Lotte Homeshopping)
9	8 (Lotte Homeshopping)	24 (YTN)	24 (YTN)
10	10 (Hyundai Homeshopping)	10 (Hyundai Homeshopping)	10 (Hyundai Homeshopping)

3) User pattern across regions: We analyzed log data for a month(Oct. 2010) to get the average VoD viewer share per region. As the result, Geonggi province whose population is 20% of Korea takes about 30% viewer share, whereas Seoul, the capital of Korea with the 20% of Korean population, takes 18.7% viewer share. The interesting point is that Gangwon's viewer share reached 6.6 % even though its population is only 3 %. From this analysis, we can know real IPTV users for a province or a city, and regional characteristics of viewing contents.

4) Individual user analysis: From this analysis we can get the individual user preference of services. For example, a single user's preference for channel service is as follows: during a day the user likes free (48.4 %), news/info (22 %), and religion genres (17.7 %) rather than others. ISPs can use the result to schedule the preferred programs for this user, and provide customized contents recommendation. In addition, ISPs could optimize network performance with traffic engineering using content caching, if the user preference is known

V. CONCLUSIONS

In this paper, we presented a MapReduce-based IPTV log analysis method that could dissect IPTV user behavior. We achieved that the data loading time could be improved by 88% compared with DBMS, and the hourly online user analysis computation time could be improved by 98% compared with DBMS. Based on the MapReduce-based analysis jobs, we derived various characteristics of IPTV users and channels that are important for providers. From this study, we could confirm that MapReduce is an effective and efficient framework for processing large-scale data analysis. For our future work, we are planning to develop in-depth business analysis programs and design a scalable MapReduce-based IPTV log analysis system which could be deployed as a commercial service.

REFERENCES

- [1] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Cluster," USENIX OSDI, 2004.
- [2] Hadoop, <http://hadoop.apache.org/>.
- [3] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding User Behavior in Large-Scale Video-on-demand Systems," ACM Eurosys, 2006.
- [4] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain, "Watching Television Over an IP Network," ACM IMC, 2008.
- [5] T. Qiu, Z. Ge, S. Lee, J. Wang, J. Xu, and Q. Zhao, "Modeling User Activities in a Large IPTV System," ACM IMC, 2009.
- [6] Y. Lee, W. Kang, and H. Son, "An Internet Traffic Analysis Method with MapReduce," IEEE/IFIP CloudMan Workshop, 2010.
- [7] W. Chen and J. Wang, "Building a Cloud Computing Analysis System for Intrusion Detection System," CloudSlam, 2009.
- [8] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D.J. DeWitt, S. Madden, and M. Stonebraker, "A Comparison of Approaches to Large-Scale Data Analysis," ACM SIGMOD, 2009.
- [9] J. Dean and S. Ghemawat, "MapReduce: A Flexible Data Processing Tool," ACM Communications of the ACM, vol. 53 no. 1, pp. 72-77, Jan. 2010.

² KBS1, KBS2, SBS, and MBC are free-to-air broadcasted channels. OCN and SuperAction are grouped as cine/series genre. Lotte HomeShooping and Hyundai Homeshopping are advertisement channels for retail sales. YTN is a news channel.