The Design and Implementation of Big Data Platform for Telecom Operators

Jing Tan(✉)

Shenyang Artillery Academy, Shenyang, China
tanj_tanjing@163.com

Abstract. This paper introduces the background of developing big data platform for telecom operators, and the benefits for telecom operators to using big data platform in some aspects that may improve forward and back changings. This document also presents a method of build big data platform using Hadoop according to the particularities of the data and systems in the telecom operators, and the implementation of this method in the internet department of a province-level telecom operator company.

Keywords: Big data · Telecom operators · Hadoop

1 Introduction

1.1 Background

Big data technology is broadly used in variety industries and companies providing technical support for marketing strategy. For instance, T-Mobile a Germany telecom operator use big data to integrate social media data, CRM and billing data that reduced customer churn rate into half in one season [1], and Walmart discern meaningful big data insights for the millions of customers to enjoy a personalized shopping experience with customers' shopping behavior data from on-and-off line [2].

Big data can be described by four characteristics [3] as follows:

(1) **Volume:** The size of data increases from TB to ZB with the growth of internet, mobile phone and sensors [4].
(2) **Variety:** Different with the structure data, types of unstructured data increases rapidly, like audio streams, video streams, images and geographic data.
(3) **Velocity:** The data is generated and processed fast to meet the demands and challenges of the companies' development and growth.
(4) **Veracity:** The quality of captured data can vary greatly. Accurate analysis depends on the veracity of source data.

Data generated from telecom operators also has these characteristics, take a middle class province of China Unicom as example, in 2012, internet access records reached 1 billion per day, and the quantity of these data is 9T per month [5]. Now, telecom operators start establishing big data platform and mining user profile to support business sales.

J. Wan et al. (Eds.): Industrial IoT 2016, LNICST 173, pp. 3–11, 2016.
DOI: 10.1007/978-3-319-44350-8_1

1.2 Hadoop Introduction

Hadoop is a framework for distributed processing and Analysis of large data sets across clusters of computers using simple programing models [6]. Hadoop is originally present from Apache Nutch a sub-projects of Apache Lucene which is start from 2002 [7]. In 2004, Google published a paper entitled "MapReduce: Simplified Data Processing on Large Clusters" in OSDI (Operating System Design and Implementation) [8] which purpose MapReduce the most important modules in Hadoop for the first time. In the early 2008, Hadoop became Apache top-level project which including many sub-projects, such as Hive, HBase, Pig which are already graduated to be top-level projects in 2010. The core components of Hadoop framework consist of HDFS and MapReduce. HDFS (Hadoop Distributed File System) [9] is a distributed file system that provides high-throughput access to application data, in the meantime, MapReduce is a framework for job scheduling and cluster resource management system for parallel processing of large data sets.

Hadoop has five benefits [10] as follows:

(1) High reliability: Single-Point or Multi-Point failure cannot interrupt Hadoop's service.
(2) High scalability: Hadoop allocated and computed in the Hadoop cluster that could be easily scale to thousand node.
(3) High performance: Hadoop can move the data among the datanodes dramatically to guarantee the equilibrium of each nodes that would be fast in processing speed.
(4) High fault tolerance: Hadoop can automatically save the data in several copies, and can voluntarily relocate the jobs that are failure.
(5) Low-cost: Compare with database machine, business data warehouse and other data mart, Hadoop is an open source software that would substantially reduce the software cost in projects.

With the advantages of Hadoop, many companies chose Hadoop as framework to build up big data platform, including IBM, Adobe, LinkedIn, Facebook [11, 12], so Hadoop could be the choice for telecom operators.

2 The Importance of Developing Big Data Platform for Telecom Operators

Along with mobile network's development, the amount of mobile data increased a lot. However, the revenue of telecom operators does not increase as well. Moreover, the traditional income keep going down as it occupied by mobile data's generator (the third party business in substitution type). Even worse, Telecom operators is going to play as a channel. Therefore, how to take advantage of "channel" role, getting data resource from "channel", controlling another core-competitiveness outside of networking resource is the top question for Telecom operators, in developing mobile network business.

2.1 Improving Business Innovation Ability

Base on analysis of large amount of data, understanding customer's requirement, and then lead to business improvement. After business online, keep tracking and analyzing customer's behavior, such as how to find it, and ordering and usage, as well as any existing problems. These data is the foundation of making strategy for business improvement, enhance business's practicability and convenience, improving business quality and customer experience. Take network optimization as an example, we can use big data technique to analyze network traffic, and trend. Then modify resource configuration in short time, meanwhile, analyzing network log, improve the whole of network, and keep improve network quality and capacity, as well as customer's networking experience.

2.2 Improving the Efficiency of Marketing Promotion

Nowadays, Telecom operators still focus on fixed package in the aspect of traffic operation business, still using fixed pattern for setting package, instead of on user's demand. Base on analysis of user's requirement and characteristic of behavior, we can filter out the target user, matching right product, determining the good time for showing and selling for customer. Moreover, we can combine channel characters and channel execution, developing precision marketing that is based on requirement subdivision and users' precise positioning. Then enhance the standardization of customer's resource management, matching customer's requirement and product features, and finally raising customer's satisfaction and marketing efficiency.

2.3 Exploring New Business Mode

Exploring new business mode includes enhancement of traditional forward charging, as well as developing new mode of back charging.

(1) **Enhancement of Forward Charging:** By improving the ability of business innovation and smart marketing, Telecom operators' ability in forward charging will be improve. Base on that, Telecom operators is able to provide personalized service, and targeted products and services for different level users. Then raising product's value and enhance the ability of forward charging.
(2) **Exploring New Mode of Back Charging:** Refer to internet business mode; telecom operators could have variety of Back-charging business mode.
    (a) **Smart Marketing:** The profit model for Telecom operators is using big data technique to provide smart marketing and precise matching product requirement, combine with easy channel system. All of these can help business partner achieve sales targets rapidly, and then business partner will pay corresponding commission for Telecom operators.
    (b) **Consultant Services:** In the process of developing product, marketing planning and product optimization, Telecom operators provides comprehensive consultant service, which is based on data analysis for business partners, to improve the product's competitive and operating efficiency. Related consultant service is one of profit points for Telecom operators.

(c) **Precision Advertising:** Precision advertising is most valuable mode in back-charging model. Telecom operators has huge number of user groups that are all potential advertising audience. In the meanwhile, diversified media which telecom operators occupied covers multi-aspect of advertising audience become valid carrier of advertisement. More important is by controlling all aspect information of advertising audience; it is easy to achieve targeted advertising and effectiveness, and will be more attractive for advertisers.

2.4 Improving Influence of Industry Chain

Deep processing data, providing information service, and create more opportunity that is new for companies without violate user's privacy. Therefore, big data technology will help telecom operators transform from web service provider to information provider. The competition of mobile network is the competition of data scale and quality, instead of number of users, or product itself. The key action of improving influence of industry chain is trying to get more high quality data and controlling more key nodes of getting data.

3 The Big Data Platform Architecture for Telecom Operators

Big Data Platform Architecture: big data platform includes three main parts: Data Collection layer, Big Data layer and Data Sharing layer which is shown in Fig. 1. Data Sources provide the data that used for store and analysis.

3.1 Data Sources

Main data comes from three channels: user network accessing interface signaling data, internal system data, and internet spider data.

(1) **User network accessing interface signaling data** come from GB port, IUPS port, GN port, LTE port and WLAN port. All of these data is web pages' session via different network, including user-browsing website's IP address, time etc.
(2) **Internal system data** comes from internal operation system, such as BOSS (Business & Operation Support System), CRM (Customer Relationship Management), TAMS (Telecom Marketing & Analysis System). BOSS system consists of network management, system management, billing system, business, and finance, and customer service. BOSS system provides networking data, billing data, and customer data etc. CRM system is able to provide marketing data and user data. TMAS provides business data and customer consumption data etc.
(3) **Network spider data** mainly use spider to extract network information, and then provide data foundation that used to analysis customer's behavior of surfing on internet.

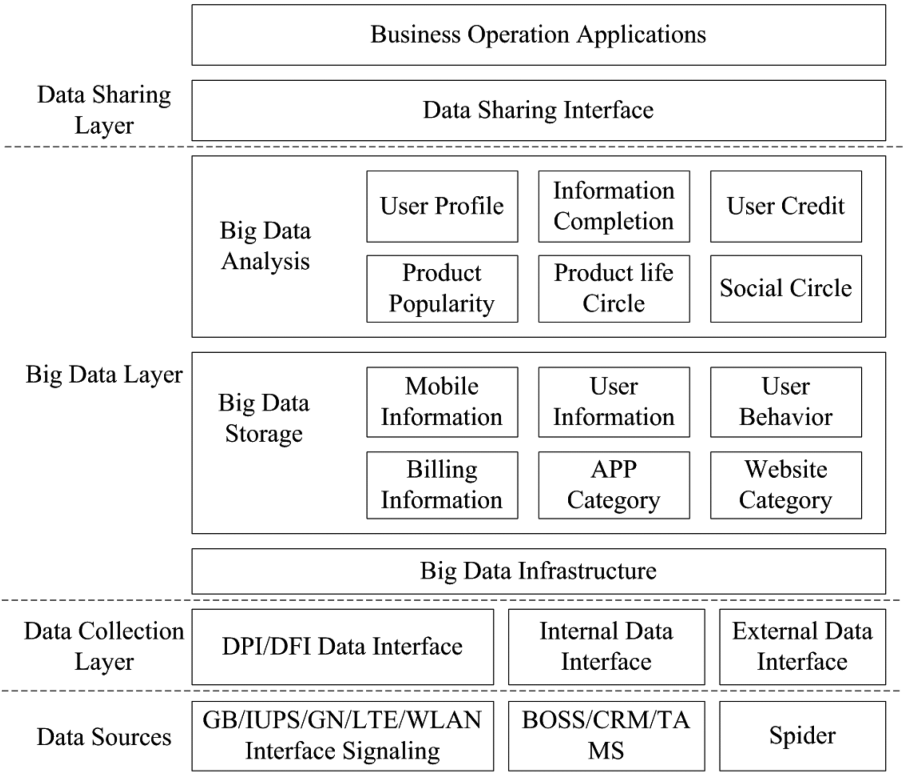

Fig. 1. The Architecture of Big data platform for telecom operators.

3.2 Data Collection Layer

Against three different data sources from Data Sources, use different way to collect data, include DPI/DFI data collection interface, internal data collection interface and external data collection interface.

(1) **DPI/DFI Data Collection Interface:** Collect user network interface signaling data, this system get the IP datagram from OBD (Optical Branching Device) which is connected in backbone network and analysis to the web session, so DPI data collection interface is handling user's web session records.
(2) **Internal Data Collection Interface:** Collect telecom operators' internal system data. This part already builds up relational database system, such as Oracle. Thus internal data collection interface is using JDBC or ODBC to get data from relational database system.
(3) **External Data Collection Interface:** Collect spider data. Network spider has sorted out the data that extract from internet, and store it in file system or database. Thus, external data collection interface is accessing file system or database system.

3.3 Big Data Layer

Big data layer provide big data infrastructure, big data storage and big data analysis, store the data from data collection layer into big data layer.

(1) **Big Data Infrastructure:** Big data distributed cluster base on Hadoop, providing foundation for big data storage and analysis, supporting high speed and high availability data storage and processing.
(2) **Big Data Storage:** Choose proper way to store the data based on data's features and application. Mainly use HDFS and HBASE. For DPI/DFI collected signaling data, original signaling data stores in file system. For the purpose of detail query, use external file connected to Hive. Processed daily data or monthly data stores in HBASE, IMEI are the key. Stored data includes Mobile Information, User Information, User Behavior, Billing Information, APP category, Website Category.
(3) **Big Data Analysis:** Analyzing and digging the data from Big Data Storage, getting new data that is supporting business extension. Data analysis includes:
    (a) **User Profile:** Labeling user's personality, according to gender, age, address, and consumer power, hobby etc., which is supporting smart marketing and precision advertising.
    (b) **Information Completion:** When user registering personal information, many data is incomplete or wrong, so we can use Data mining technique to complete or correct the personal data.
    (c) **User Credit:** Analyzing user's consumption and other basic data, getting user's credit evaluation, provides to Bank or Credit Information Company.
    (d) **Product Popularity:** Analyzing Telecoms' product popularity, supporting marketing.
    (e) **Product Life Circle:** Analyzing Telecoms' product life circle, understanding this product's operating.
    (f) **Social Circle:** Analyzing user's social circle, supporting smart marketing.

3.4 Data Sharing Layer

Data sharing layer adopt unified data accessing interface, open it to telecoms internal use, or open secure interface for external company to personal to use.

4 The Implementation of Big Data Platform

This architecture of big data platform for telecom operators is implemented in the internet department of a middle-class province for collecting and storing user internet accessing data, acquiring users' internet accessing behaviors. Combining with the users' demographic data, we use this platform to analysis the popularity and characteristic of the music, reading and game products of this department, and present marketing strategies for these products.

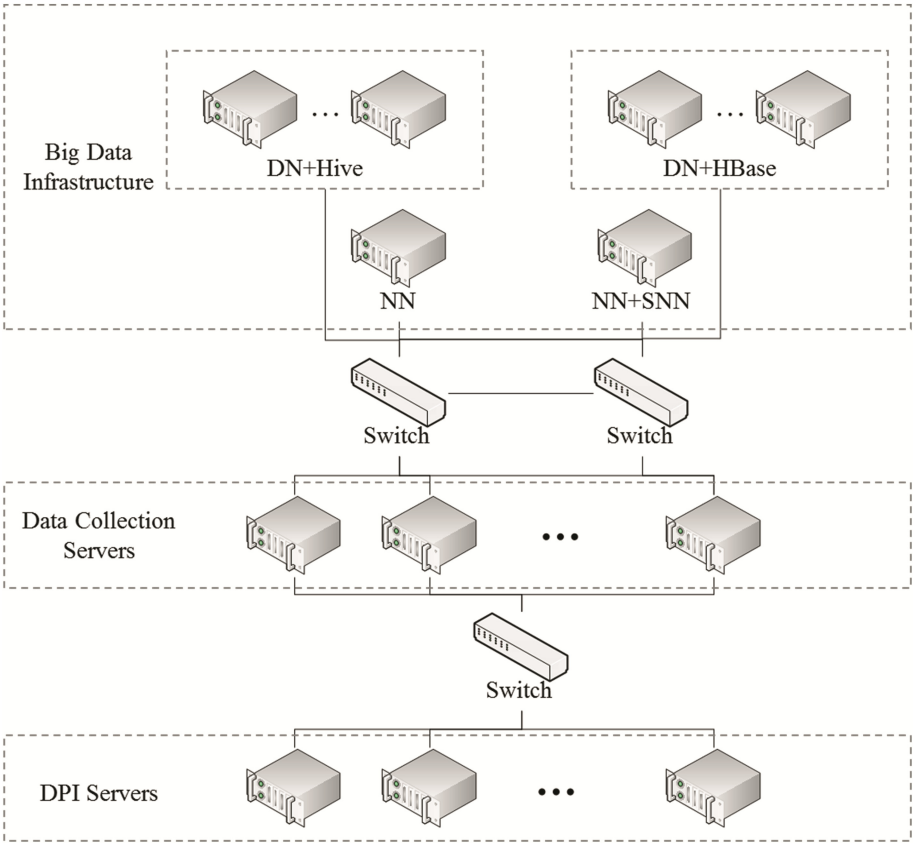

Fig. 2. The implementation of big data platform architecture in a province telecom operator.

Figure 2 shows the implementation of big data platform architecture in a province telecom operator. It includes two main parts, data collection servers and Big Data Infrastructure. In addition, DPI servers are very significant in telecom operators' network for gathering user internet accessing data but not a component in big data platform.

(1) **DPI Server:** DPI server is the data source of user internet data. DPI servers gather user signaling data from backbone network via OBD, and convert the user signaling data to user session data.
(2) **Data Collection Server:** It gets user session data from DPI server, and then sends to Big Data Infrastructure. Since the amount of session data is huge, it almost 300 thousand records per second, so this part consists of 6 servers, and each server need to process 50 thousand records per second on average.
(3) **Big Data Infrastructure:** This is a Hadoop cluster, which consists of servers including NN (NameNode) and DN (DataNode). NameNode is file system naming space in Hadoop that maintain the whole file system tree and all the related files and directories. DataNode is the file system's working node, it store and indexing

data base on dispatching with client or NameNode, also sending block list for NameNode periodically. Big Data Infrastructure is responsible for store the user Internet accessing session data and is summed up in minute, hour and day. Big Data Infrastructure is consist of three components:

(a) **NN and NN+SNN (Secondary NameNode):** NN is NameNode in Hadoop ecosystem. NN+SNN is backup of NN that is using HA, and used as SNN.
(b) **DN+Hive:** It is DataNode in Hadoop, and deploy Hive on it; it stores users' session data, and support session query, and multi-division's query and integration. This part deployed nine servers and each server supports four network interfaces.
(c) **DN+HBase:** Deploy both DataNode and HBase. It stores users' session details records, as well as integrated data in minute, hour and day. Therefore, it able to do detail query and analysis for integrated data. This part deploys nine servers and each server supports four network interfaces.

According to above description, Data Collection Servers can upload data into DN +Hive and DN+HBase concurrently. Each DN+Hive and DN+HBase server is setting two internet domains. One is for NN or NN+SNN servers to access and dispatch the data, the other is used for data collection servers to upload data to DN, so for avoiding confusion in the physic network and guaranteeing the upload speed, DN+Hive and DN +HBase is divided into two separate network segments which is connected by one switch separately. The fact is this kind of architecture is able to fulfill the requirement of concurrency that is about 300 thousand internet accessing session records per second.

5 Conclusion

This article introduced the design and implementation of big data platform for telecom operators. The Framework of this platform is based on Hadoop. We construct the big data platform that is special for Telecom operators. It is collecting users' internet accessing signaling data, internal system data and web spider data, improving the speeding of query data and data mining for users who are interested in music, reading and games, and providing guidelines for marketing strategy. As a result, this platform gain very good achievement.

Nowadays, telecom operator's traditional business, such as voice and SMS, this part's income is keep going down. Thus, Telecom operators is seeking new opportunities, many Telecom operators realized the value of big data, and already constructing big data platform. However, as the data is distributed in different BUs, it is very difficult to integrity data. In addition, different BU's business need is different, so it caused repeating construction of big data platform, and skill level also not the same. However, big data still quite important, it will bring more opportunities for Telecom Operators.

References

1. How to Use Big Data to Stop Customer. http://www.cio.com/article/2391831/big-data/how-to-use-big-data-to-stop-customer-churn.html
2. How Big Data Analysis helped increase Walmart's Sales turnover? http://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmart-s-sales-turnover/109
3. Big data. https://en.wikipedia.org/wiki/Big_data#cite_note-INDIN2014-25
4. Segaran, T., Hammerbacher, J.: Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media, Sebastopol (2009)
5. Wang, Z.J.: The application of hadoop in the telecom industry. Technical report, One China One World (2012)
6. Apache Hadoop. http://hadoop.apache.org/
7. Dean, J., Ghemawat S.: MapReduce: simplified data processing on large clusters. In: OSDI, vol. 51, no. 1, pp. 147–152 (2004)
8. Nagel, S.: Web crawling with Apache Nutch. In: ApacheCon EU 2014, Budapest (2014)
9. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium Mass Storage Systems and Technologies, pp. 1–10. IEEE (2010)
10. Olson, M.: Hadoop: scalable, flexible data storage and analysis. IQT Q. **1**, 14–18 (2010)
11. Borthakur, D., Gray, J., Sarma, J.S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., et al.: Apache hadoop goes realtime at Facebook. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM (2010)
12. Sumbaly, R., Kreps, J., Shah, S.: The big data ecosystem at linkedIn. In: Ross, K.A., Srivastava, D., Papadias, D., (eds.) SIGMOD Conference, pp. 1125–1134. ACM (2013)