Journal of Communications and Information Networks, Vol.2, No.3, Sept. 2017
DOI: 10.1007/s41650-017-0010-1

Research paper

Big data in telecommunication operators: data, platform and practices

Zhen Wang[1], Guofu Wei[1], Yaling Zhan[1], Yanhuan Sun[2*]

1. Information Center, China Telecom Co., Ltd. Anhui branch, Hefei 230001, China
2. Key Laboratory of Wireless-Optical Communications, Chinese Academy of Science, University of Science and Technology of China, Hefei 230017, China

* Corresponding author, Email: stracy@mail.ustc.edu.cn

Abstract: In the age of information explosion, big data has brought challenges but also great opportunities that support a wide range of applications for people in all walks of life. Faced with the continuous and intense competition from OTT service providers, traditional telecommunications service providers have been forced to undergo enterprise transformation. Fortunately, these providers have natural and unique advantages in terms of both data sources and data scale, all of which give them a competitive advantage. Multiple foreign mainstream telecom operators have already applied big data for their own growth, from internal business to external applications. Armed with big data, domestic telecom companies are also innovating business models. This paper will introduce three aspects of big data in the telecommunications industry. First, the unique characteristics and advantages of communications industry big data are discussed. Second, the development of the big data platform architecture is introduced in detail, which incorporates five crucial sub-systems. We highlight the data collection and data processing systems. Finally, three internal or external application areas based on big data analysis are discussed, namely basic business, network construction, and intelligent tracing. Our work sheds light on how to deal with big data for telecommunications enterprise development.

Keywords: telecommunication operator, enterprise transformation, big data, platform architecture, practical applications

Citation: Z. Wang, G. F. Wei, Y. L. Zhan, et al. Big data in telecommunication operators: data, platform and practices [J]. Journal of communications and information networks, 2017, 2(3): 78-91.

1 Introduction

With the explosive increase in global data, the term "big data" has been used to describe huge datasets. Doug Laney[1] was the first to mention the 3V's of big data management: volume, velocity, and variety. The data generation capacity has never been more powerful and enormous since the birth of information technology in the early 19th century[2]. The overall mobile data traffic is expected to reach 30.6 exabytes per month by 2020, an eight-fold increase over 2015[3]. In 2013, IBM issued the report "applications of big data to the real world", which stated that the internal data of enterprises were the main sources of big data[4].

Thus, for enterprises, the fundamental challenges of big data applications are exploring the large volumes of data and extracting useful information or knowledge for future decision making[5]. Big data has begun supporting a wide range of potential applications. For instance, Facebook has used social networking data to track user interest patterns and

Manuscript received Dec. 30, 2016; accepted Feb. 8, 2017
This work is supported partially by Key Program of National Natural Science Foundation of China (No. 61631018), the Fundamental Research Funds for the Central Universities and Huawei Technology Innovative Research on Wireless Big Data.

carry out precision marketing, which has yielded profitable results. In 2014, Alibaba launched the "DMP", which enabled businesses to implement different marketing strategies for different people based on the analysis of user information obtained through this product. Applications of big data in other fields include tracking movie box office receipts[6], healthcare system[7,8], customer surveys[9], and user characteristics analysis[10]. All these big data applications are gradually transforming the way we live, work, study, etc.

Faced with continuous and intense competition from OTT service providers, traditional telecommunications service providers must undergo enterprise transformation. Fortunately, these operators have access to rich data sources and huge datasets, which other industries do not have. Large numbers of customers will generate loads of behavioral data every second of the day, including calling, messaging, networking, and other kinds of information. Even when the customer is inactive, location-based data will be generated. Moreover, combined with registration and business information, customer billing data can be obtained.

Consequently, the vast amounts of data that operators have can potentially outpace the ability of existing CDR-based processing to improve our daily lives[11]. Telecommunications data can be used to optimize operations and drive operational business intelligence to realize immediate business opportunities[12]. Multiple foreign mainstream telecom operators have already applied big data for their own development. Orange Business Services for instance, used big data to enhance the accuracy of their churn detection. Spain's Telefonica Dynamic Insights obtained reliable predictions of user behavior by packaging and analyzing data. In 2014, Verizon built data centers in California to implement precision marketing[13]. Domestic operators are also innovating their business models by exploring the use of big data.

This paper provides detailed discussions of three aspects of big data in the telecommunications industry. Section 2 discusses the sources and unique advantages of communications industry big data compared to other industries. Section 3 introduces the framework of the industry's big data platforms in detail, from collection systems to storage systems to application systems. Section 4 details three internal and external applications based on big data. Finally, this paper ends with a summary and directions for future research.

2 Data sources and advantages

In this section, the major sources of wireless big data and the advantages of operators are introduced.

2.1 Data sources

As providers of basic network services, the goal of telecom operators is to provide an information channel between people and equipment, and between different types of equipment[14]. Operators themselves are the producers of big data. Data generated in a communications network is the main source of Internet big data.

Communications data is mainly derived from the following three sources[15].

• Data in IT system: user attributes, business consumption information, and terminal information data collected from CRM, billing systems, and terminal self-registration platforms, respectively. Basic user profiles and characteristics can be described in accordance to these data.

• Data in access network and core network: mobile signaling, DPI, M2M data, etc. These data accumulate in wired/wireless networks whenever clients use voice, SMS, or networking services. The underlying structure of the data is complex, hence targeted analysis and processes are needed for different types of data to achieve scenario-based descriptions of user locations and preferences.

• Data in operators own Internet applications: online business hall data, palm business office data, wing payment data, etc. All data, including user access modes, addresses, times, business preferences, investment and consumption habits are completely

stored in the background of the application, which can be obtained directly.

In terms of "Volume", hundreds of millions of users' behavioral data are already in the petabyte or even terabyte range. In terms of "Variety", communications data covers all businesses, customers, and channels, as well as Internet data, human attributes, position trajectories, and terminal information. In terms of "Velocity", the quality of communication services should meet the real-time requirements of various applications.

2.2 Advantages

In China, three operators have the largest number of users compared to all other industries, i.e., approximately 1.3 billion mobile users and 300 million fixed broadband users[16]. The massive number of users combined with their own industry policies provide the following advantages to operators.

• Authentic user information. Owing to the existing real name system, non-real name users have limited services and are required by law to register. This not only ensures the authenticity of user information, but also guarantees that the data has one-on-one correspondence with a real person.

• Comprehensive and intact information. Unlike Internet companies that can only interact with users through their own App business, operators can access all behavioral data on users in the network all the time, such as when and where they used the service, terminal type, website accessed, products searched, hot topic interests, etc. With enough storage and computing power, we can efficiently and completely uncover all these behaviors. Moreover, with the availability of authentic user information, the complete and accurate descriptions of user profiles and features can be obtained.

• Identifiable and relatable data. User identifications in the operator system include the mobile phone number, ID card number, terminal ID, cookies, and many other types of information. These data can be related to financial, Internet, hotel, transportation usage, and other business-related data to break down the isolated data and develop a real sense of the big data cloud, under the premise that user privacy is guaranteed.

• Continuous and real-time data. Compared to Internet services providers, telecom operators can obtain position tracking data through the cellular network protocol even when users only power-on their devices and have no data connection (Wi-Fi/3G/4G). This mechanism guarantees real-time and continuous data collection, which will be more powerful in real-time applications such as issuing early traffic warnings.

3 Big data platform

In this section, we will introduce the telecommunications operator's internal big data platform in detail. The overall design of the big data platform is based on the principles of data concentration, degree of openness, and cloud computing. It aims to provide secure access, storage, sharing, analysis, applications, and management. It helps construct an enterprise-level and future-oriented data center. Moreover, the platform will create an open and shared public data environment. The above attributes can guarantee the application implementation in all internal departments.

3.1 Overall framework

As is shown in Fig. 1 the platform mainly consists of 5 parts: data collection system, data storage system, data processing system, open mobile system and management system.

The overall framework has a distinct hierarchy and arrangement. The selected technologies and components are mature and stable. On one hand, it can satisfy the data processing requirements in the current data environment. On the other hand, all the included technologies are supposed to be in line with the future direction of big data.

3.2 Data collection system

The data collection system is the basic part of the platform. It provides a variety of data access tools

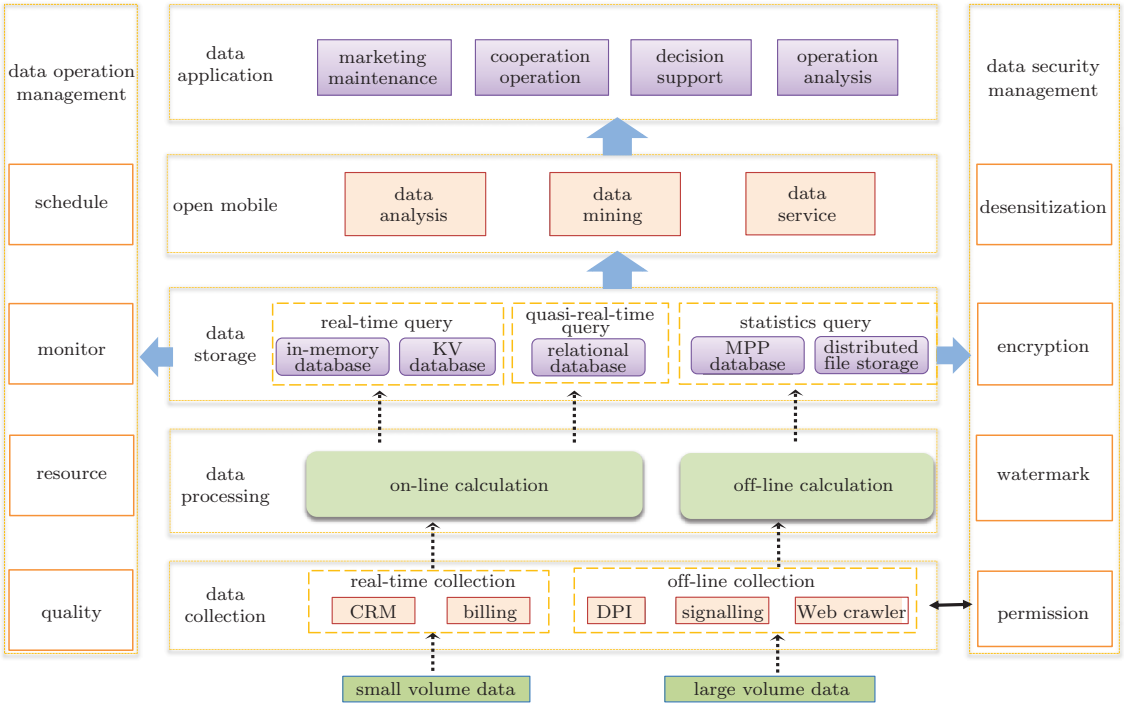

Figure 1 Big data platform framework

and aggregates the critical structured and unstructured system data from all enterprise management departments, front-end and back-end. By combining data from the offline acquisition and real-time acquisition phase, the system can break down the isolated information and aggregate all original data into the unified platform.

3.2.1 Collection interfaces

There are many different systems in the communications enterprise. Thus, interfaces are needed to connect the collection system to other kinds of systems. Some data are stored in files while others are real-time data. There are two kinds of interfaces. The data collection interface collects data from various interior source systems. The service interface manages data sharing and transfer among different intermediate systems. Tab. 1 introduces the different data interfaces.

3.2.2 Collection technologies

In this part, we introduce three common collection technologies.

Synchronization technology based on relational database: Both dblink and OGG are synchronization technologies for Oracle databases. OGG is a comprehensive software package for real-time data integration and replication in heterogeneous IT environments. The product enables high availability solutions, real-time data integration, transactional change data capture; and data replication, transformation, and verification between operational and analytical enterprise systems.

Applied scenarios: Dblink is mainly used for data synchronization between Oracle databases. It is often used in full-scale synchronization. OGG uses the database file synchronization mode. Because of its high efficiency and small influence on the source system, it is currently used in production systems and other time-sensitive applications, such as the synchronization of attributes tables, orders, and lists.

Interaction technology between HDFS and tables based on Sqoop: Sqoop Apache (SQL-to-Hadoop) was designed to help the RDBMS and Hadoop achieve efficient big data exchange. With the help of Sqoop, users can transfer relational database data

Table 1 Data interface

interface type	end system	specification	interface mode
data collection	national platform of internet log	mobile network DPI, mobile network AAA data	file
data collection	DPI platform of network operation	fixed network DPI, source IP, AD subscriber ID, timestamp, request URL, user agent, referrer URL, destination IP, cookie user port, destination port, etc.	file
data collection	DPI platform of network operation	fixed network AAA data, including WLAN authentication and broadband user authentication	quasi real-time
data collection	OIDD platform	OIDD system signaling data	file, real-time
data collection	ODMS	UDB, ISMP, business pilot, WLAN hotspot management platform, TSM platform and other value-added business data	file
data collection	billing system	mobile network billing details (calling and called), SMS billing, flow billing	file
data collection	billing system	fixed network billing details (calling and called)	file
data service	ability product and application platform	all types of original list of external business	file, real-time
data service	provincial IT system	provincial roaming data issued	file

to related systems in Hadoop, such as HBase and Hive. Sqoop can also extract data from the Hadoop system and then export it to the relational database.

Applied scenarios: The development of businesses and applications, especially the impact of big data, has led to the exponential growth of enterprise data. Data formats are becoming increasingly diverse, such as text, video, Web crawler data, and many other structured and unstructured data. The traditional dblink and OGG synchronization technologies have failed to meet the demands of the industry. Hence, the Hadoop open source framework for data processing was introduced. Because of the use of HDFS file storage mode, Sqoop is a good solution to the synchronization problem between the relational database and distributed database file system. Currently, the data stored on the Hadoop platform includes all user information, subsidies, sales, orders, DPI, signaling, and other structured or unstructured data. These data are collected by Sqoop components and will be able to meet the subsequent processing requirements of big data SQL engines such as Impala, Spark, and Hive.

Incremental document collection technology based on Flume: Flume NG is a distributed, reliable, and available system provided by Cloudera. It can efficiently collect, aggregate, and move massive amounts of log data from different sources and store them in a centralized data storage system. It is a lightweight and simple gadget which can easily adapt to various collection methods and balance loads.

Applied scenarios: Flume technology is mainly used for the log collection of each system. The development of cloud application systems, distributed architectures, and increasing node numbers make daily operations and maintenance processes increasingly difficult, such as dispersion, storage pressure, non-standardized log formats, non-unified query channels, and non-automatic push of abnormal information. These problems spurred us to build a log database. The business applications cover the track analysis of operation and maintenance personnel, operations staff, business processes in the business hall, and user Web page access. For example, a clerk reports on the part of the business that is inefficient and provides a specific order number. Then, operations and maintenance personnel, according to the analysis of customer tracks, can identify the time-consuming link, customer waiting time, and pure system operation time. Based on the above steps, we can determine the real reasons for the inefficiency and provide recommendations for the optimization

Table 2 Data collection sources and data scale of a company

business type	frequency	capacity/TB	increment/TB	processing memory	total storage/TB	duration/month
customer, account and user information	day	1.00	0.90	2.00	31.00	1
inventory data integration	day/month	1.20	1.20	2.40	37.20	1
mobile network DPI	day	1.90	1.90	3.80	68.40	36
fixed network DPI, ITV	day	5.00	5.00	10.00	155.00	1
wing payment	month	0.01	0.01	—	0.24	24
port A signaling	day	0.60	0.50	2.00	37.20	2
OSS data	day	0.20	0.10	0.60	74.40	12
income, bill	month	3.00	0.80	2.60	75.00	25
statements	day/month	1.70	0.20	3.40	40.80	24
group data	month	0.30	0.01	—	7.20	24
account	—	14.91	10.62	26.8	526.44	—

and management of the IT system.

One company's current collection system is shown in Tab. 2.

3.3 Data processing system

The data processing system is the core of the platform, providing deep mining and analysis services. Using the distributed storage and parallel computing framework combined with many kinds of computing engines, this system can accomplish fast and distributed computing for structured, semi-structured, and unstructured information resources.

3.3.1 Processing architecture

In order to achieve efficient collaboration in data processing and meet the requirements of different applications, we divided the system into a real-time module and an offline module as shown in Fig. 2.

Real-time scenario: Real-time data, including mobile broadband/product development, terminal sales, package development, 4G flow, and gross income are all displayed by instrument panels, progress bars, trend charts, regional hotspot maps, and other forms. The development status and progress are self-explanatory. Personnel can make timely adjustments to marketing decisions by utilizing the screen display and rolling update.

Off-line scenario: Business development and application complexities apply loads of pressure to storage and processing. Moreover, the ODS and EDW hardware platform basically use minicomputers or integrated machines, which lead to hard management. Fortunately, open source technologies can integrate both structured data (e.g., BSS, OSS, MSS) and unstructured data (e.g., mobile DPI and fixed-network DPI). After the construction of the offline analysis platform, we can observe the daily critical quota. The specific steps are as follows:

1. check the external table data according to certain rules, such as volatility and consistency;
2. check and insert the data into internal tables in interface layer, and do time stamp and partition;
3. store the mild summary and detail data generated by the model calculation in HDFS format;
4. process based on business logic.

3.3.2 Processing level

Tab. 3 shows the data processing level of one provincial telecommunication company.

3.4 Other systems

This section introduces the other three systems, namely the data storage system, open mobile platforms, and management systems.

Functioning as the support of data analysis and sharing, data storage systems can store and query

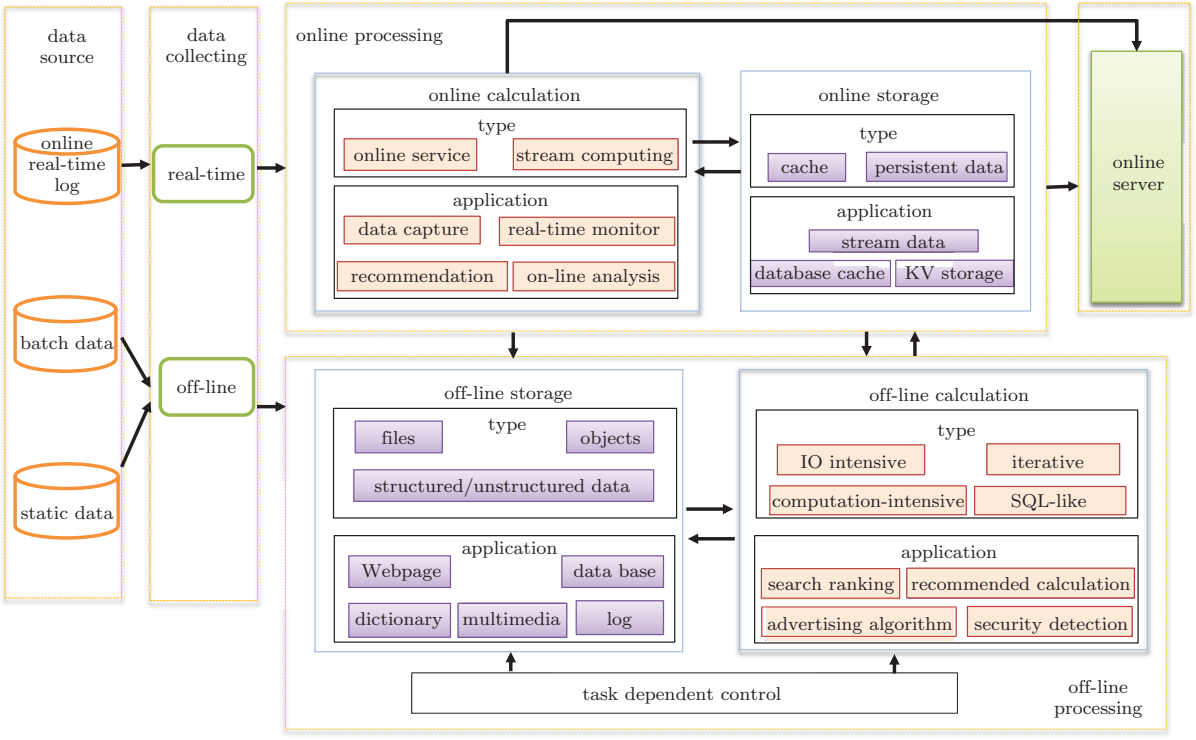

Figure 2 Data processing system frame

Table 3 Data processing scale

data type	data scale	preprocessing		message queue		flow preprocessing	
		normal value	maximum	normal value	maximum	normal value	maximum
information	2.3 million/day	125 pieces/s	500 pieces/s	125 pieces/s	500 pieces/s	125 pieces/s	500 pieces/s
file	85 million/day	1 000 pieces/s	2 000 pieces/s				

structured, semi-structured, and unstructured data. In order to achieve efficient data transfer, there are four layers in the storage system. Interface layer: this layer aims at peripheral data sources and is responsible for data collection and preprocessing. It can manage external data sources, interface types, format requirements, scheduling methods, and supervision of data acquisition and exchange. Integration layer: this layer integrates the isolated business model to establish a set of theme-oriented enterprise data models. Intermediary layer: this layer refines the integration layer information for the purpose of application. It can reduce the degree of coupling between models through the fragmented way of processing and storage, which supports fast and agile data processing and assembly. Summary layer: this layer can provide data analysis, data mining, and extempore query for cross-domain data.

The open mobile platform supports both internal data applications and external business. First, it is a platform for foreign businesses using the multitenant mode. Second, the operator is the platform operator as well as one of the tenants. The platform needs to assign users and permissions to tenants, and provide user-level independent storage space, well allocated computing resources, secure data protection, etc. The multitenant mode needs to make full use of the data analysis capacity and help tenants apply for resources. It can also perform the intelligent management of tenant resources by recycling those with high idle rates and expanding limited resources.

The management system has two parts: data management and security management. The data management module is responsible for process scheduling

and monitoring, generation of the main data and index database, and data resource management. The security management module is responsible for user rights, data access, access control, data desensitization, data encryption, watermarks, and other system management functions.

4 Practice and applications

This section introduces big data analysis-based practices from three perspectives. The first one is their application to normal business. Then, it shows their effects on network optimization. Finally, we will introduce a business in which the telecommunications operator collaborates with the government.

4.1 3G/4G upgrading

4G has become a key business for telecom operators since the release of TDD/FDD LTE licenses. It has a strong influence on future user profiles. By now, the terminal-SIM matching rate is relatively low. The number of matching users for the Anhui province is 3 800 000, which accounts for 37.1% of the total as of July 2016. Thus, using big data to enhance 4G terminal sales is an effective way to validate the present study. The first step is using data mining to identify potential users. The ARPU can be considered followed by target marketing. This way, both the user scale and value can be enhanced.

4.1.1 Dataset

The sample consists of 4 600 000 customers of a provincial company as of April 2016. They used neither 4G terminals nor LTE flow. Over the next three months, the number of 4G terminal upgrades was 334 000 (i.e., the number of positive samples). Because of the large difference between positive and negative samples, we performed some balance measurements. Meanwhile, 70% of the sample was designated as the training set and the remainder as the test set.

In order to ensure the purity of the data, we need to check the data on user information, self-registration, billing, terminal type, etc. to handle abnormal values, outliers, and missing values.

Then we generate derivative variables by combining business rules. Cluster the ARPU and flow into three categories and generate ARPU-rank field (1, 2, 3) and flow-rank field (1, 2, 3) respectively. Calculate other derived fields including the overflow consumption, ARPU and terminal price matching degree.

4.1.2 Model and algorithm

First, filter the valid input variables. The number of input variables follows a short and refined principle. Too many input variables are likely to cause problems, such as interference and over-fitting, which can lead to a decline in the stability of the model. There are two methods to select variables: choosing by business analysis and choosing accordance with the correlation coefficient. When the correlation coefficient between two variables is equal to or greater than 0.6, this indicates a moderate or above linear relationship between the two variables. Here, we only need to keep one variable.

In order to ensure the universality of the model, we need to divide the data into the training set and test set. The model is constructed on the training set. The hit rate, coverage, and applicability are verified on the test set. The availability is guaranteed by the cross validation.

By comparing the indicators obtained by the automatic classifier node in the SPSS Modeler, we find that decision tree algorithm has the best overall performance among the different algorithms shown in Tab. 4.

Table 4 Performance of different classifiers

algorithm	performance	
	precision	recall
decision tree	70%	80%
neural network	71%	50%
logistic regression	69%	55%

Fig. 3 shows the key factors chosen by the feature selection module of SPSS.

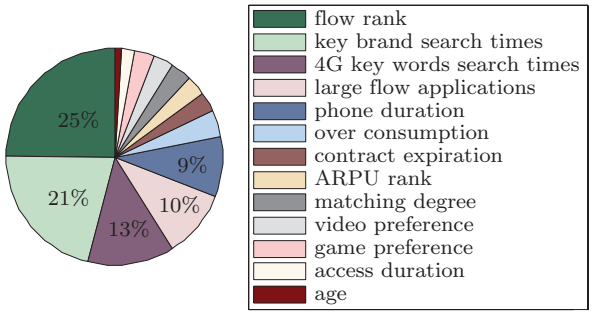

Figure 3 Key impact indicators

According to the key indicators mentioned above, we screened four categories of non-4G terminal and non-LTE flow mobile users at the end of July.

Cluster 1: Preferred large flow applications such as videos and games. The monthly average flow was greater than or equal to 500 MB.

Cluster 2: Highly concerned about 4G terminals. Frequently searched terminal brands and used keywords like 4G.

Cluster 3: Preferred high-speed applications; played games and used other large flow applications frequently. ARPU was greater than or equal to 59 Yuan.

Cluster 4: Preferred certain terminal brands when the contract or replacement period expires.

4.1.3 Results and discussions

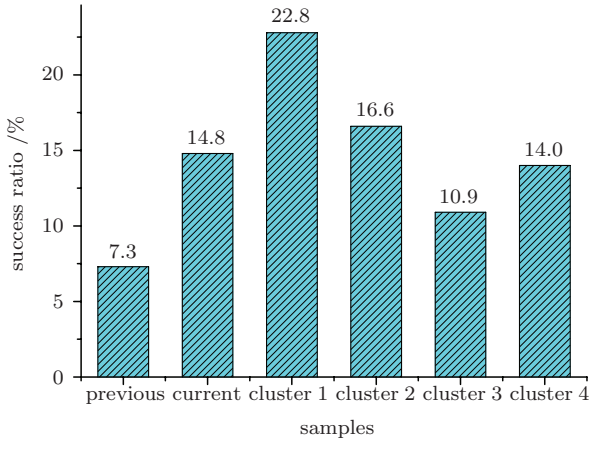

Figure 4 Practical effect of proposed model

From August to September 2016, we divided 1 500 000 users into four clusters and performed target marketing, mainly through CRM pop-ups and phone calls. Overall, 223 000 users replaced their phones. The outcomes of marketing activities were significantly improved and the success ratio was 7% higher than previous campaign, especially for cluster 1, as shown in Fig. 4.

Mining potential clients and target marketing have proven to be quite effective. It may be possible to use additional types of classifiers to enhance algorithm performance and choose the crucial indicators more accurately. Moreover, users can be clustered according to more detailed standards, which may further improve the success ratio.

4.2 Big data driven network construction

4G users will switch to the CDMA Ev-Do network in areas where LTE network coverage is weak or nonexistent. This is commonly known as "cutting-down flow", where users obtain network access from the 3G base station and the packets are forwarded through the 4G core network (connected by the eHRPD in 3G core network). However, the same frequency interference and other complex wireless disruptions may still cause users to cut down to the 3G base station, even though the capacity and bandwidth are sufficient. In order to avoid changing user perceptions about the quality of the 4G service, personnel can optimize the network or plan a new site. By analyzing the traffic thermodynamic diagram, we can determine which areas need expanded LTE network coverage and report these conclusions to the wireless network optimization center or construction center.

4.2.1 Model and algorithm

Resolving the frequent cutting-downs in core business districts will enhance the user experience and balance the LTE network load. We can identify the cutting-down station by implementing range determination, data integration, and thermodynamic diagram analysis.

1. Automatic classification of base station coverage area based on grid holography. Based on holog-

raphy grid GIS platform, we developed the automatic classification of base station coverage area. By GIS mapping and spatial calculation ability, we realized the intelligent recognition of base stations in core business circles, transportation hubs, and other key areas. GIS platform input: local network IDs, base station IDs, macrobase station labels, base station latitudes, and base station longitude. GIS platform outputs: coverage names, coverage types, corresponding base station IDs, and local network IDs.

2. Form flow integration tables. We integrated and cleaned data on bill payments, including 1X, Ev-Do, LTE, and prepaid bills to form a traffic-wide table. The tables contain the traffic flow type, Internet connection modes, up and down flow, time, places visited, base stations, and other features. Combining these tables with the base station tables in GIS platform, we can analyze the whole network traffic efficiently based on different divided zones and time periods.

3. Thermodynamic diagram display. In order to facilitate the visualization, we developed a thermodynamic diagram display function. We selected a part of the province's core business centers to analyze the cutting-down flow. The thermodynamic map can provide more information on the distribution of top cutting-down stations.

As is shown in Fig. 5, the darker color means that there is too much cutting-down flow. Additional optimizations should be performed in these darker regions.

4.2.2 Results and discussions

We cooperated with the Anqing branch to optimize wireless signal coverage in core business centers. In the process, the number of cutting-downs decreased significantly and the amount of network traffic increased. The number of cutting-down times decreased from 27 000 to 11 000 for a single station each month, and the average daily traffic increased from 11.2 GB to 18.6 GB.

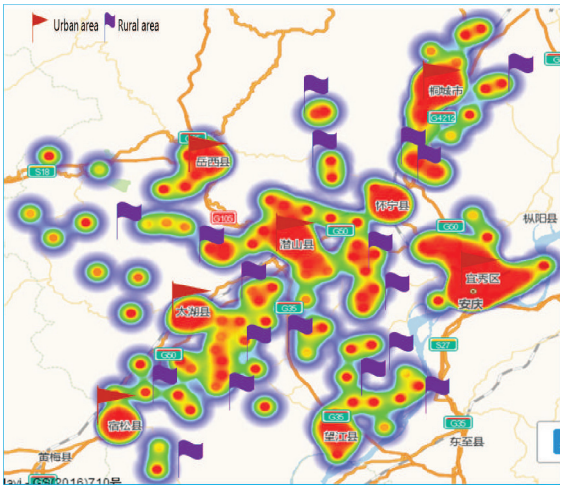

Figure 5 Condition of cutting-down flow

By using big data in network construction, we not only enhance the experiences of users and the quality of service in busy districts, but also can save more resources in low flow regions, which will improve resource utilization. In our future work, we will focus on applying this analysis to the city's WLAN establishment to further optimize the infrastructure.

4.3 Intelligent trace

With the development of user-scale mobile Internet, the total number of active domestic equipment is expected to reach 1.3 billion by June 2016[17]. The widespread popularity of mobile phones will generate massive signaling data. Mobile phones will transmit user location information to base stations, such as calls, SMS, Internet access, periodic events, etc. Based on the above features, there will be a new way to obtain city planning information by analyzing signaling data. We can exploit the massive individual user characteristics which will accurately reflect the overall characteristics, especially spatial movement. Moreover, this scheme is low cost and stable. Just a few installed equipment are needed to enable real-time traffic data collection for a wide area within a short time, with small impact on the network. Compared with traditional traffic survey technology, signaling analysis has many advantages, such as wider coverage, larger sample, lower cost, and long-term

continuous monitoring. This section explores the application of the intelligent tracing system based on the mining of spatial and temporal mobile data.

We aim to build a support environment and interactive interface for the intelligent footprint system. We will also establish a spatial-temporal analysis standard for big data and industry. Moreover, will promote multidimensional interconnections in order to achieve efficient organization, orderly management, reasonable use, and high value when analyzing spatial-temporal big data.

4.3.1 Model and algorithm

1. Real time data acquisition technology based on mobile signaling. Based on real-time monitoring and using specialized signaling acquisition software and hardware, operators can filter and analyze specific signaling processes and obtain information about base stations and signaling. This technology can locate signals from small cells to large regions, leading to personalized services in road monitoring applications. The data recorded include the user IDs, time stamps, positions, and other location information. It updates every 5 min to ensure the accuracy and continuity of user location information.

2. Moving sequence detection. There are many uncertainties and disturbances in the signaling time sequence. For example, a user's signal may suddenly move far away from the trace, which we call "flying-points". Another case is shown in Fig. 6, where the user does not move at all but handovers occur frequently. The reason is the overlapping region, as represented by the red areas. Hence, we need to conduct preprocessing to filter out the abnormal data and obtain the real moving sequence by real-time flow computing technology.

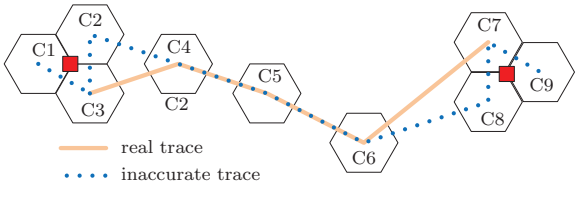

Figure 6 Signaling sequence

3. Key location mining algorithm. In location mining, we chose DBDCAN algorithm which based on density instead of normal K-means algorithm. This is because 1) the number of K-means is difficult to determine in advance; 2) K-means polygons form round clustering shapes easily and is not suitable for squares, rivers, or places with other shapes; 3) K-means is very sensitive to noise data and the trace data is not clear enough. On the other hand, DBSCAN is adaptable and can avoid the above mentioned problems.

4. Spatial-temporal trajectory real-time road matching algorithm. The base station coverage radius is about a few hundred meters and it is often far away from the road. Information is updated frequently, generally in less than 1 min. Faced with high frequency and high error rates, a map-matching algorithm is essential to map the base station location to road-level positions accurately and in real-time. Thus, we based the algorithm on the commonly used probability graph model to conduct road network matching. Because of the characteristics of signaling, some optimization is needed to ensure the accuracy and efficiency of the algorithm. A road test can provide powerful support for this algorithm. First, install the technical analytic device in the car. Then, record the vehicle trajectory and collect the corresponding handover sequence. Finally, using the marked data, such as time, latitude, and longitude to adjust the parameters.

4.3.2 Results and discussions

Based on the above models, we can construct the intelligent traffic analysis platform in cities, which will provide dynamic crowd analysis, real-time data on traffic conditions, traffic behavior analysis, urban planning support, etc. Some specific applications are described below.

The first is traffic demand analysis and road planning. Based on the analysis of 24/7 hours crowd movement, we established the planning OD(Origin Destination) matrix. Then, use tracking modeling we analyzed the main trajectory. According to the established traffic grid and the movement coordi-

nates, we can obtain the real road load demand, which is based on the crowd flow. The OD trajectory at different time periods can generate a full time OD trajectory diagram, as shown in Tabs. 5 and 6.

Table 5 OD matrix of 8:00 am∼9:00 am

O	D									
	J01	J02	J03	J04	S01	S02	S03	S04	S05	SUM
J01	50	3	11	5	19	15	19	38	36	196
J02	5	38	69	1	14	5	19	21	67	239
J03	8	53	82	3	13	10	8	17	59	253
J04	20	1	1	0	9	22	11	44	8	116
S01	3	8	22	1	53	6	7	21	76	197
S02	7	4	6	23	7	45	6	45	19	162
S03	7	7	8	2	14	6	0	17	34	95
S04	36	18	29	32	36	42	7	140	74	414
S05	16	19	51	1	31	8	19	34	117	296
SUM	152	151	279	68	196	159	96	377	490	1 968

Table 6 OD matrix of 18:00 pm∼19:00 pm

O	D									
	J01	J02	J03	J04	S01	S02	S03	S04	S05	SUM
J01	55	8	11	15	17	9	14	34	17	180
J02	9	42	61	4	17	6	14	20	57	230
J03	8	81	115	2	20	8	17	14	45	310
J04	14	2	4	0	3	24	6	38	7	98
S01	23	26	26	12	150	39	20	80	111	487
S02	12	3	6	39	7	109	4	69	18	267
S03	15	19	10	8	16	7	0	26	26	127
S04	59	18	31	36	34	60	14	243	39	534
S05	41	70	112	16	128	39	57	80	277	820
SUM	236	269	376	132	392	301	146	604	597	3 053

These tables show the population flow from the corresponding origin (O) to the corresponding destination (D). J** and S** represent the residential and commercial zones, respectively. In Tab. 5, many people go to the business district in the morning for work. Tab. 6 shows that some of these people go home after work while others spend their leisure time in the business district.

The second application is the display of road conditions. We calculated the speed of each moving sequence and displayed the real-time road conditions by a spatial-temporal real-time road matching algorithm. Testing of the area proved that the results of our algorithm accurately represented the actual conditions. Fig. 7 shows that traffic jams usually occur at the time people go to work or when they go home, at 8:00 and 18:00, respectively. We also observed that the rush hour on weekdays starts approximately 2 h earlier than on weekends. Moreover, the overall condition of the road is slightly better on weekends.

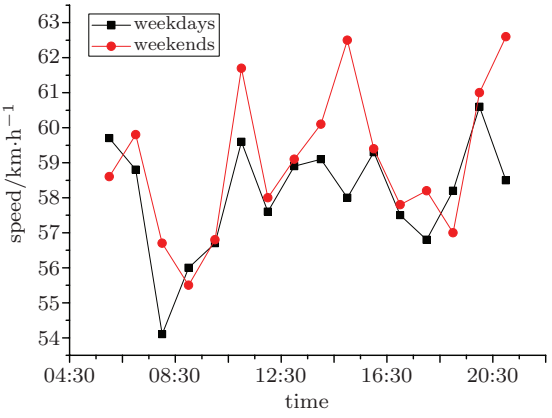

Figure 7 Display of road speed

Based on mobile phone terminal position information, data such as real-time population flow, road congestion, etc., can be obtained with high precision. On the one hand, operators can provide travel recommendations for users. On the other hand, this can provide the basis for traffic management departments to control traffic, such as controlling the time of traffic lights and adjusting the buses routes. Moreover, unexpected events such as gathering crowds can be prevented. The analysis also facilitates the quick response to abnormal situations. There are multiple and significant applications that we must develop in the future.

5 Conclusion and future work

Big data is transforming the way we live and do business in this era of information explosion. It has brought a lot of challenges, but also great opportunities to the communications industry. Faced with continuous and intense competition from OTT ser-

vice providers, the ways in which to exploit big data to achieve enterprise transformation is an important topic. This paper analyzed three aspects of big data: the big data characteristics of the communications industry, big data platform architectures, and big data applications.

In the product development and application process, according to national laws and regulations, we addressed the management of information safety and set up a hierarchical information protection mechanism. From system architecture design to application development, we adopted multiple technologies to protect user information, such as encryption and watermarking.

We will conduct in-depth research in the two areas: internal data application and external cooperation. The first is enhancing the efficiency and effects of the application of big data by establishing production systems for internal customers' online interactions and paying attention to the customer experience by performing interface and process optimization. Through rapid iteration and response of the system function, the front line can become the indispensable tool for sales and service support. The second is to provide better services in terms of the security of user information in external connections. We will further improve the multiple levels of privacy protection technologies during the construction of big data platform.

Appendix

This part introduces some specific abbreviations and their meanings.

Table 7 Abbreviations

OTT	Over The Top
SMS	Short Message Service
BSS	Business Support System
CDR	Call Data Record
OSS	Operation Support System
MSS	Management Support System
EDA	Enterprise Data Architecture
CRM	Customer Relationship Management
DPI	Deep Packet Inspection
OIDD	Open Information Dynamic Data
ODMS	Operation Data Management System
AAA	Authentication, Authorization and Accounting
ISMP	Integrated Services Management Platform
UDB	User Database
M2M	Machine to Machine
TSM	Truted Service Manager
RDBMS	Relational Database Management System
WLAN	Wireless Local Area Network
OGG	Oracle Golden Gate
IT	Information Technology
ARPU	Average Revenue Per User

References

[1] D. Laney. 3D data management: controlling data volume, velocity and variety [J]. META group research note, 2001, 6: 70.

[2] X. D. Wu, X. Q. Zhu, G. Q. Wu, et al. Data mining with big data [J]. IEEE transactions on knowledge and data engineering, 2014, 26(1): 97-107.

[3] C. V. N. Index. Global mobile data traffic forecast update, 2015-2020 white paper [R]. 2016, 4.

[4] M. Chen, S. W. Mao, Y. H. Liu. Big data: a survey [J]. Mobile networks and applications, 2014, 19(2): 171-209.

[5] J. Leskovec, A. Rajaraman, J. D. Ullman. Mining of massive datasets [M]. Cambridge: Cambridge University Press, 2014.

[6] M. Mestyán, T. Yasseri, J. Kertész. Early prediction of movie box office success based on Wikipedia activity big data [J]. PloS one, 2013, 8(8): e71226.

[7] W. Raghupathi, V. Raghupathi. Big data analytics in healthcare: promise and potential [J]. Health information science and systems, 2014, 2(1): 1.

[8] V. Yadav, M. Verma, V. D. Kaushik. Big data analytics for health systems [C]//IEEE International Conference on Green Computing and Internet of Things (ICGCIoT), Delhi, India, 2015: 253-258.

[9] P. J. Li, Y. H. Yan, C. M. Wang, et al. Customer voice sensor: a comprehensive opinion mining system for call center conversation [C]//IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2016: 324-329.

[10] A. C. E. S. Lima, L. N. de Castro. Predicting temperament from Twitter data [C]//The 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Kumamoto, Japan, 2016: 599-604.

[11] W. Fan, A. Bifet. Mining big data: current status, and forecast to the future [J]. ACM SIGKDD explorations newsletter, 2013, 14(2): 1-5.

[12] R. L. Villars, C. W. Olofson, M. Eastwood. Big data: What it is and why you should care [J]. White paper,

IDC, 2011.

[13] X. F. Zheng. Big data application and revelation of foreign telecom operators [J]. Mobile communications, 2015, 39(13): 29-33.

[14] Y. J. Huang, M. Feng, S. Y. Ding, et al. Big data development strategy for telecom operators (in Chinese) [J]. Telecommunication science, 2013, 29(3): 6-11.

[15] K. F. Chen, H. C. Zhou. Research on realization mode of telecom operators' big data resource and its strategy [J]. Mobile communications, 2016, 40(1): 63-67.

[16] Ministry of Industry and Information Technology of the People's Republic of China. Completion of the main indicators of the communications industry by Nov. 2016 [EB/OL]. http://www.miit.gov.cn/n1146312/n1146904/n1648372/c5427058/content.html.

[17] People's Daily Online. Annual report on China's mobile Internet development(2016) [R]. China: People's network, 2016.

About the authors

Zhen Wang was born in Xi'an, Shaanxi. He received the B.S., M.S., and Ph.D. degrees all from University of Science and Technology of China (USTC), Hefei, Anhui, in 2002, 2005, 2009 respectively. He is currently the general manager of Information Center of China Telecom Co., Ltd. Anhui branch. He engaged in wireless big data and big data applications at Samsung Electronics Communications Research Institute, Suwon, Korea, from 2008 to 2010. He published nearly ten papers in core journals and got three patents. His research interests are the applications of big data. (Email: 15305516521@189.cn)

Guofu Wei was born in Guang'an Sichuan. He received the B.S. degree in mathematics from University of Science and Technology of China (USTC), Hefei, Anhui, in 1997, and the Ph.D. degree in computational mathematics from USTC, in 2002. He is now a senior economist at China Telecom Co., Ltd. Anhui branch. His research interests include big data modeling and big data operations. (Email: weiguofu@189.cn)

Yaling Zhan was born in Tongcheng, Anhui. She received the B.S. degree in computer science and technology from Chongqing University of Posts and Telecommunications in 2004. She is now an assistant engineer at China Telecom Co., Ltd. Anhui branch. She deeply participates in the construction of enterprise internal application system. She is also in charge of the target acquisition of precision marketing. Her research interests include the big data management, modeling and operation. (Email: 18955159329@189.cn)

Yanhuan Sun [corresponding author] was born in Yantai, Shandong. She received the B.S. degree in information engineering from Xidian University, Xi'an, China, in 2015. She is currently working toward the M.S. degree in information and communication systems at University of Science and Technology of China (USTC). She received the IEEE Wireless Personal Multimedia Communications 2016 Best Student Paper Award in 2016. Her research interests include machine type communication and big data application. (Email: stracy@mail.ustc.edu.cn)