Design of Flow Big Data System Based on Smart Pipeline Theory

Jianqing Zhang[1], Shuai Li[2,*] and Lilan Liu[3]

[1]Department of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[2]Department of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[3]Department of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
Corresponding Email: lsmichaelly@163.com

Abstract. As telecom operators to build intelligent pipe more and more, analysis and processing of big data technology to deal the huge amounts of data intelligent pipeline generated has become an inevitable trend. Intelligent pipe describes operational data, sales data; operator's pipe flow data make the value for e-commerce business form and business model in mobile e-business environment. Intelligent pipe is the third dimension of 3 D pipeline mobile electronic commerce system. Intelligent operation dimensions make the mobile e-business three-dimensional artifacts. This paper discusses the smart pipeline theory, smart pipeline flow big data system, their system framework and core technology.

Keywords. Smart pipeline, big data, e-commerce, framework design.

1 Smart pipeline theory

Smart pipeline [1] originating from the communications industry, the ultimate goal is to meet customer a variety of flexible ways of access through establishing the intelligent pipeline which can be controlled and the end-to-end difference. Meanwhile corresponding to different business forms and provide innovative business experience. Intelligent pipe is the third dimension theory of the pipeline system superimposed on a one-dimensional, two-dimensional pipe. If the superposition of 2D pipeline, mobile e-business system from a single bearing extensions to the plane, it be open from carriers and operators resources clouded. That 3D pipeline intelligent pipe [2] is for mobile electronic commerce further extension, extended to the three-dimensional model. 3D integration is the ultimate form of mobile electronic commerce theory model, the third dimension emphasized in the era of mobile Internet, the era of big data, operators' pipeline intelligent theory, pipeline intelligent technology and intelligent application of business model, business value. Intelligent pipe generates a lot of users surfing the Internet log files to provide date source for billed for billing and user behavior online query and management analysis. Analyzing business data effectively in the intelligent pipeline, and use the results of the analysis and processing for control and management of pipeline is essential for the successful operation of intelligent pipe to the realization of the rights of the operators. As shown in Figure 1.

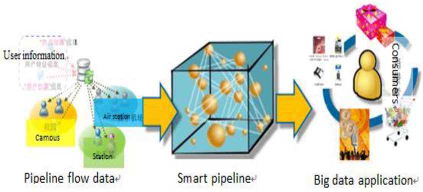

Fig.1. 3 D pipeline under the integration of intelligent mobile e-commerce ecosystem.

2 Big data systems of operating intelligent pipe flow TAS

Smart pipeline flow analysis support system TAS is the subsystem of Intelligent Mobile E-Commerce Cloud Platform (IMECCP). TAS mainly for operators in the network traffic data, traffic data collection and analysis, data mining applied in IMECCP platform.

2.1 Business mode

TAS platform business and application is mainly aimed at the target merchants, operation analysis means for merchants to provide depth, marketing intelligence (BI) capabilities. Its main business model has three categories:

(1) Operational analysis categories: TAS provide merchants basic operational data analysis function, operation analysis module provides free module and charging module, for daily general merchants sales, warehouse, order statistics report to provide free services, as for some merchants customized data report, customized report will be charged according to the needs of businesses.

(2) Smart marketing categories: Smart marketing business is a major core business model of TAS. Merchants can order intelligent marketing service, for example, obtaining merchant track customer behavior data ,smart marketing class can be divided into offline mode smart marketing and online marketing.

(3) Run the auction categories: Operation bidding class is to help merchants analysis industry information, TAS collecting data form various industries and send it to the data center, can provide Shared data within the scope of industry recognition, such as providing the same industry with many merchants business development and the consulting services about an overview of the operating data

2.2 TAS Platform application system

TAS platform application system mainly divided into three parts:

(1) Data gathering: Date gathering is based on the operator data platform to large flow data and gathering all kinds of data sources: Operators, mobile Internet traffic data and customer behavior data and so on. Such as Internet access to the user, use the terminal behavior and the content of the order package actual usage and data integration network quality, etc. Data gathering constructing data system as large operators + mobile Internet data. Completing the structured and unstructured data preprocessing and data storage function.

(2) Data driven：Data driven is the second largest system in TAS. Its role is to enable mobile e-business towards intelligent mobile e-commerce, data driven to gather the data for data processing, big data mining and data analysis, and the result used for driving the process of mobile e-commerce operation, marketing and service of each link. TAS can get the user data from baidu ranked, SEO index, etc. It can provide business decisions propaganda drainage on the strategy. In terms of customer behavior, TAS provides behavior data can drive merchants product operation optimization, such as purchasing, payment process, at the same time can also provide goods operation adjustment data basis. TAS provides segment traffic data, can make the merchants operating staff know the conversion rate of various business or goods.

(3) Data labels: It is the third largest TAS system. TAS after routinization and a large number of intelligent data processing. Customers, merchants and other attributes can be formed TAS data labels. On the one hand, it fixed-line

data-driven data processing results; On the one hand, it open data tag library can be used as a data model is a child module of TAS opened to the public. TAS tag library is characteristic of TAS formed after the big data processing collection classes. Merchants select existing tags for commercial applications by label adaptation. TAS label application is TAS data-driven upgrade version, on the one hand is data driven conceptual characteristic collection, classification, on the other hand is as TAS ability to open module.

3 TAS smart pipeline architecture design

TAS platform is mainly composed of data gathering, data driven and data labels. As shown in Figure 2. Data gathering is to realize the function of data acquisition deployment, data acquisition, data preprocessing, and data adapter; Data driven is business intelligence data mining, intelligent application rules, and other functions; Dateable is the data-driven model, the application and the ability to open the component library.

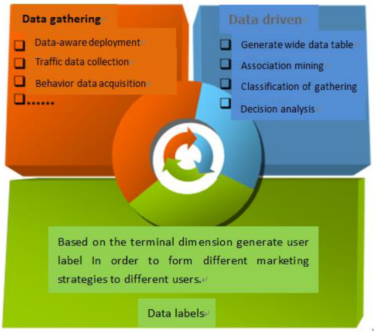
Data gathering
□ Data-aware deployment
□ Traffic data collection
□ Behavior data acquisition
□......
Data driven
□ Generate wide data table
□ Association mining
□ Classification of gathering
□ Decision analysis
Based on the terminal dimension generate user label In order to form different marketing strategies to different users.
Data labels

Fig.2. TAS system three big modules.

3.1 Based on the data gathering pipeline filter architecture

Pipeline filter [3] is one of the classic design of software architecture, in the pipe/filter style of software, architecture is composed of multiple hierarchy structural connections, each level component with one or a set of input and output, when data got into a single component (pipe), component begin internal processing and output data stream. To change the data input data stream, specifications and other processing becomes filtering, the component becomes the filter. Such a style of fitting is become pipe/filter. This is shown in Figure 3.

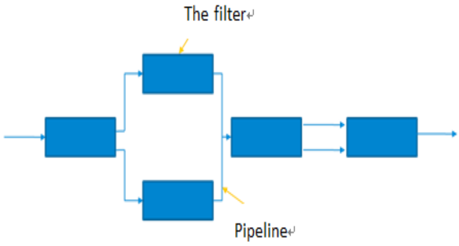

Fig.3. Pipe/filter design architecture.

Pipeline filter architecture design has the following characteristics:
(1) Pipeline filter has a good concealment, high cohesion and low coupling;
(2) Can support data gathering more data sources, the demand of the multi-layer data preprocessing;
(3) Have higher software reuse and reusability, similar to function of the filter can be reuse connection;
Data gathering architecture based on the pipeline filter architecture style as shown in Figure 4.

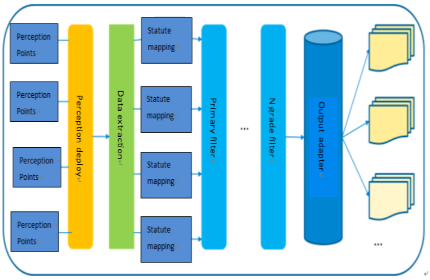

Fig.4. Data gathering architecture.

Perception of deployment: Requires data analysis is used to drive the aware of acquisition for deployment;

Data extraction: Is a set of integrated filter module, complete the basic data packet extraction, such as the data is divided into content, behavior, etc.;

Statute mapping: Refers to the rules on the application of the different data gathering data mapping, mainly according to the data gathering after mapping the drive application of management;

Filter layer: TAS data gathering data processing layer, the same data preprocessing function module at the same level is available for reuse and through the connection between layer and layer thinning data processing;

Output adapter: Output format is based on the outer data applications, data driven data requires;

3.2 Based on the data warehouse architecture and knowledge base

After complete the data gathering, different businesses have different business development needs. Therefore the function of the data driven goals and modules will be different, but the application of data driven commonness exists in the method. Thus cannot blindly to the expansion of data driven application modules. So the driver module for data of TAS, based on the warehouse and the knowledge base of architecture design style, as shown in Figure 5. Warehouse and knowledge base structure is to apply data driven and data-driven rules and unified, the warehouse storage component module has complete data driven capacity. For example, correlation algorithm, clustering, classification, linear regression and so on. It can categorize knowledge stored in the knowledge base and distinguish between the management of the private intelligence of the merchants. The knowledge base was shown as the blackboard.

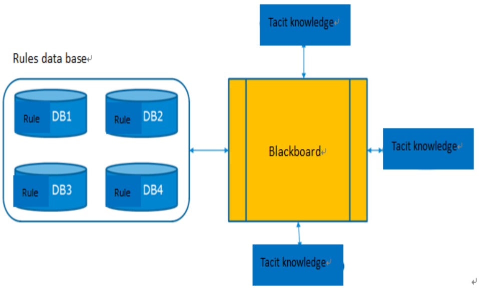

Fig.5. Warehouse and the knowledge base architecture.

Rule knowledge base is used to store different classification or different merchant’s private business intelligence driving rules. The blackboard extracts data from data gathering of tacit knowledge, docking rules according to the corresponding application on the blackboard. The blackboard is a state of driver and application driver module, the data display and driver is on the blackboard to tacit knowledge and the corresponding knowledge base organization to solve the problem of specific business applications.

Warehouse and corresponding TAS data-driven knowledge base architecture features as follows:

(1) Users don't need to pay attention to the blackboard after the concrete application of transparent to the client, simple to use;

(2) Merchants can manage their own application rules, and can reuse from common rule knowledge base;

3.3 Data labels based on the Hadoop

Data labels are the curing and open ability of TAS platform module. Data label is to solve the problems of various merchants during the actual store operations, all kinds of data analysis and data intelligent marketing demand. Data labels can integrate, mining and manage the effective customer information.

Integration: Existing a lot of customer information resources stored separately in different systems, the need to form a unified view of customer information, raise the utilization ratio of information.

Mining: The current use of customer information with index is staying the level of analysis and report. The further analysis of customer information, deepen the characteristics analysis and customer demand is important.

Management: Daily operation of customer did not form the effective management of knowledge. Knowledge utilization is low, the need to strengthen the effective management of customer knowledge.

Data labels need to solve the problem of large data processing ability. Therefore in the project practice, this paper adopt the Hadoop [4] related architecture design technology. Hadoop contains two types of technology: One is the HDFS distributed file storage system; The second is graphs distributed technology. HDFS is used to solve the TAS to deal with many merchants, large data problem of data storage capacity, its technical architecture diagram as shown in Figure 6.

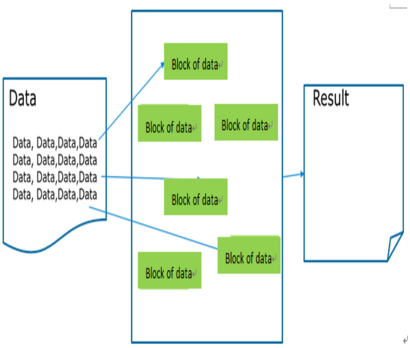

Fig.6. HDFS Schematic diagram.

HDFS file is composed of blocks. Each block has multiple copies, multiple copies in the HDFS architecture can be stored in a building on different computers in the system. The graphs can be split into multiple small pieces for processing shown as Figure7.

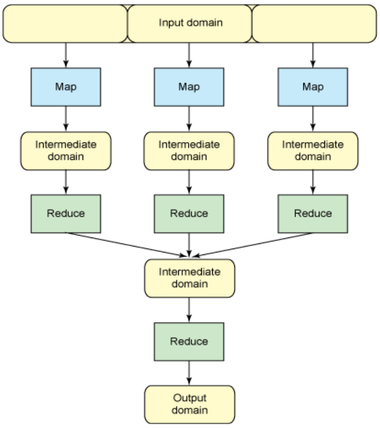

Fig.7. MapReduce Schematic diagram.

HDFS architecture in the aspect of application architecture is transparent for the outer applications. Applications need not care about the data location, as local calls for data calls.HDFS + graphs can solve data label under the TAS big data module, data processing ability, and TAS data label module technology architecture as shown in Figure 8.

4. The system key techniques in the big data

4.1Key techniques in data gathering

(1) Big data access technology
As can be seen from the operators of big data, the traditional billing domain data, how to file as a carrier; Signaling class data, the data stream as the carrier; Broadband Internet data, the data message as the carrier.

A. Grid collection: Grid acquisition [5] is suitable for gathering data in the file as a carrier. Based on the concept of worker bee collaboration, the traditional FTP / SFTP file collection technology was extended and realized the distributed collaborative concurrent acquisition. When data network generate batch files at a time, the traditional sampling need time sequence generated by file, it takes a long time; But grid sampling will split stay files into multiple division (divided by the data source ID, the information such as file name, start, end displacement), different node/application receive the respective division for concurrent collaborative collection. When the data network element at a time to generate large files, the traditional collection depends on the only data download channel, likely to cause incomplete data acquisition or acquisition multiple times, but grid acquisition will split the large file into multiple partitions, different nodes / Application to receive their own division for concurrent collaborative collection, stitching and verification when landing. Collection dispatching and breakpoint aimed at persistence management. A real-time synchronization of multi-node grid collection of calibration, guarantee grid to collect accurate and complete and at the same time security when fault occurs it can quickly move to the new node stability without gap operations

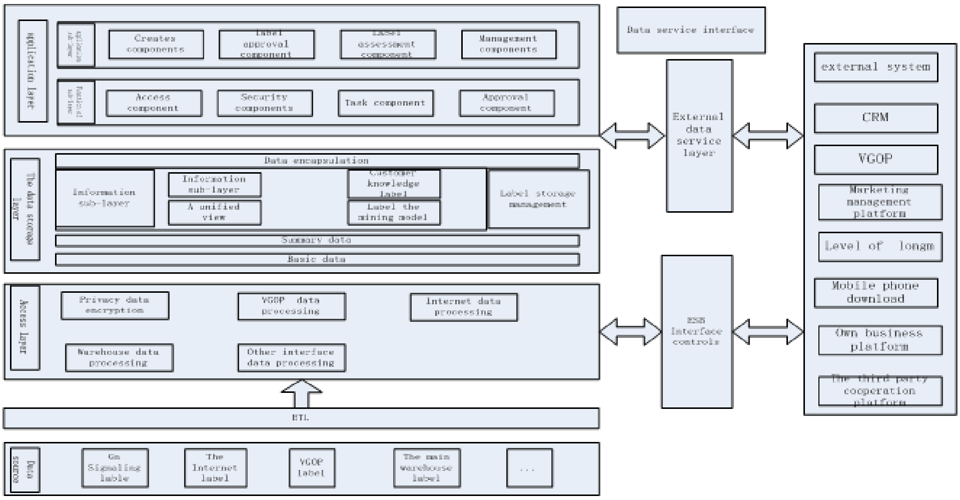

Fig.8.TAS data label module technology architecture.

b. Streaming Collection: Streaming Collection is suitable for aggregating signaling class data with high time efficiency. Streaming Collection need the large data center and data network element prescribed uniform flow interface, large data center send data subscription request to the data network element, waiting for the data network element returned data subscription response. The large data aggregation center analyzing and checking streaming message according to the message format, translating legitimate messages into internal format events, pushing to the data delivery module, and writing the message backup to the distributed file system.

c. Spectral collection: Spectral acquisition is suitable for aggregating broadband interconnect data from the core network. Spectral collection has the ability to capture data directly to ensure data comprehensiveness and timeliness; it protects the value of the data to maximize on the source. For example, the signaling location data information can be split and access from the signaling port A, and the preamble is used to collect the signaling A port signaling by the network access protocol converter. After the signaling data are analyzed and correlated, the synthesized position is updated, Called the message, the main called voice calls and other signaling CDR / TDR.

(2)Big data processing technology

As can be seen from the telecom operators of large data, the traditional billing domain data, there are billing, DPI file data, and also signaling, location information class message flow data. Billing domain and other types of data processing need to be processed in sequence and be calculated accurately with the high timeliness; Signaling, location information and other types of data processing need to quickly deal with the high timeliness. Hadoop, Storm, Spark is the most popular large-scale data processing tools, in the application of telecom operators in large data processing, the need to combine the characteristics of timing and timeliness to select the framework technology. Hadoop uses MapReduce distributed [6] computing framework, developed HDFS distributed file system according to GFS and developed HBase data storage system according to BigTable. Hadoop is often used for off-line complex large data processing.

MapReduce technology framework contains three levels of content: (1) Distributed file system; (2) Parallel programming model; (3) Parallel execution engine. The Google file system runs on a large cluster, which is built using cheap machines, and the data is stored in key / value mode. The entire file system uses metadata centralized management, data block distributed storage model, through data replication (at least 3 copies of each data.) to achieve a high degree of fault tolerance. Data is stored in chunks (64MB or 128MB for 1 block), it can be easily compressed on the data, saving storage space and transmission bandwidth. The MapReduce parallel programming model decomposes the computational process into two main phases, the Map phase and the Reduce phase. The Map function handles the Key / Value pair and generates a series of intermediate Key / Value pairs The Reduce function is used to merge all intermediate key-value pairs with the same Key value and calculate the final result.

4.2 The key algorithm of data mining

Data mining [7] also known as Knowledge Discovery in Database. It is from a large, incomplete, noisy, fuzzy and random big data to extract implicit in people is unknown in advance, has the potential value of the process of information and knowledge. Through association rule mining, can be implied in sea has potential value in the amount of data useful information.

Assuming assemble $\mathrm{I} = \{i_{1,}i_{2,}i_{3,} \ldots i_{m,}\}$, i_k（k=1,2,…,m）represent the item. If $\mathrm{x} \in \mathrm{I}$, X is called item sets. When$|x| = k$, X is called K’s item sets. Affair two-tuples $\mathrm{T} = (\mathrm{tid}, \mathrm{X})$, tid is the uniqueness identification called affair NO. Data assemble$\mathrm{D} = \{i_{1,}i_{2,}i_{3,} \ldots. i_n\}$ consist of t_1, t_2… t_n. Association rules can be described as: like the formula A=>B in which A∈I，B∈I，并且$\mathrm{A} \cap \mathrm{B} = \emptyset$.The support of item sets X is S which is the percentage of D contained X’s number of transactions and All real number remembered as $\mathrm{s(x)} = \mathrm{p(x)} = \frac{\mathrm{sup}(x)}{|D|}$.The degree of confidence of item sets X is C which is the percentage of D contained X∪Y number of transactions and All real number remembered as $\mathrm{c(x)} = \mathrm{p}(X|Y) = \frac{\mathrm{sup}(X\cup Y)}{\mathrm{sup}(X)}$. Minimum support (minsup) and minimum confidence (minconf) are depended on the users. If sup(X)≥minsup, the item sets X is called cumbersome item sets in Which generated association rules all support and confidence in not less than minsup and minconf is called the strong association rules. Commonly used algorithm[8] for mining association rules has breadth-first algorithm、depth-first algorithm、dataset partition algorithm、sampling algorithm and parallel algorithm. Parallel algorithm is used at the same time to perform a collection of the process of the interaction and coordination to complete for a given problem solving. It has the higher efficiency.

Parallel FCM algorithm adopts the master/slave mode. A main process controls a number of process to calculate. In order to be kept in distributed parallel interconnect workstations environment mining fuzzy association rules. Improve the master/slave mode for single program multiple data (SPMD) model and divide the quantitative attributes. Each quantitative attribute values on the record as the initial data set. The PCM algorithm as follows:

FCM.1: In each process divided between the initial data set. Each process for n/s data, in which n is the total number of data records' to start the process.

$$\{x_1, x_2, \ldots, x_n \mid x_{\left(\frac{n}{s}\right)}, \ldots, x_{\frac{2n}{s}} \Big| \ldots \Big| x_{(s-1)(n/s)}, \ldots x_n\}$$

Process1 Process2 Process S

FCM2: Process by the root process (process id 0) initialization center $v_i, i = 1,2 \ldots\ldots c$ and broadcast to all processes.

FCM3: Each process receives the $v_i, i = 1,2 \ldots\ldots c$ and calculate the membership degree

$$u_{ik} = [\sum_{j=1}^{c} (\|x_k - v_{i0}||_A \backslash ||x_k - v_{j0}||_A)^{2(m-1)}]^{-1}$$

Each process j operation on its data subset $\left\{x_k, \mathrm{k} = \frac{(\mathrm{j}-1)\mathrm{n}}{\mathrm{s}} + 1, \ldots j\frac{n}{s}\right\}$. Initializes is end.

FCM4: Each process according to the local data calculation.

$$\alpha_{ij} = \sum_{k=1}^{h} (u_{kj,t-1})^m X_r) \qquad \beta_{ij} = \sum_{k=1}^{h} (u_{k,t-1})^m)$$

h=n/s is the received data subset size for each process.

FCM5: Each process j send results $(\alpha_{ij}, \beta_{ij})$ to the root process, then the root process about computing v_{it} and broadcast to all processes.

$$v_{it} = \frac{\sum_{j=1}^{s} \alpha_{ij}}{\sum_{j=1}^{s} \beta_{ij}} \quad \mathrm{i} = 1,2 \ldots.\mathrm{c}$$

FCM6: Each process receives class center value and calculates the membership degree. Each processes their operations on the data subset.

$$u_{ik\ t} = [\sum_{j=1}^{h} (\|x_k - v_{it}||_A \backslash ||x_k - v_{jt}||_A)^{2(m-1)}]^{-1} \ \forall i, k$$

FCM7: Each process calculation error. Each process j only calculation error, and sent to a process.

$e_j = \sum_{k=1}^{h} \sum_{i=1}^{c} (u_{ik,t-1} - u_{k,t})^2$ For j=1, 2... S

Finally in the root process calculate summary error E_t.

$E_t = (\sum_{j=1}^{s} e_j)^{1/2}$.

If E_t.<ε, notify each process termination procedure.

For each quantitative attribute, the clustering center C and dividing matrix U. Each class represents a fuzzy set. And the elements in the matrix are divided into each values belonging to the degree of fuzzy sets. When using the Euclidean distance, the time complexity of parallel FCM algorithm is $\mathrm{o}\left(\frac{c^2\mathrm{pn}}{\mathrm{s}}\right)$.

5 Conclusion

This paper describes the concept of smart pipeline. In order to apply the network traffic data, traffic data collection and analysis, data mining applied in mobile e-commerce platform，it establish the system of intelligent pipe flow analysis support system TAS Reference. This paper mainly introduces the Business model, the platform body system, and framework design and core technology.

References

1. Huiling Zhao, Xianghui Xu. Connotation and characteristics of intelligent pipe. Telecom science,S1, 2011, pp.1-5

2. Anaraki FB. Elearning and mLearning at assumption university. 2011 International Conference on e-Education, Entertainment and e-Management (ICEEE), 2011, pp.31-34
3. Ping Xu, Jing Liu, Shengju Sang. Introduction to the software architecture style. Science and technology information. No.26, 2008, pp.42-43.
4. Shukui Hao. Introduction to Hadoop HDFS and MapReduce.Design technology of posts and telecommunications. No.7, 2012, pp.37-42.
5. Li,Guojie.The scientific value of big data research.Communications of CCF,2012,8(9):8-15
6. Jeffrey Dean Sanjay Ghemawat，Map Ｒeduce: Simplified Data Processing on Large Clusters．USA: San Frcmcisco，cakfcrnia． 2004:137 －150.
7. Han Jiawei , Kamb er M. The concept of data mining and technology.Beijing: Mechanical industry press, 2001: 149-176.
8. Cheung D W, Han J, Ng V T. A Fast Distr ibuted Algorithmfor Mining Assoeiation Rules //Proceedings of 1996 Inter-nat ional conference on Parallel and Distr ibuted InformationSystem. Miami Beaeh, F L: [s. n.] , 1996: 31-44.