Design and Implementation of Bank Wind Control Anti-fraud Project Based on Big Data Technology

Yu Jianhao
Shanghai Huateng Software Systems Co., Ltd.

Abstract: With the advent of the big data era, the construction process of financial big data product system is also gradually accelerating. This paper discusses a case of financial big data products created by banks—"credit risk control anti-fraud project". The author first clarifies the requirements and background of the project construction, then shares the framework and application architecture design scheme, highlights the specific design and implementation methods of the system, finally summarizes the bottlenecks encountered in the project construction process and solutions, and points out the characteristics of the system in practical application.

1. Introduction
The Ministry of Industry and Information Technology's "13th Five-Year Plan for the Development of Big Data Industry (2016-2020)" involves "accelerating the research and development of key technologies for big data, fostering a safe and controllable big data product system, and innovating big data technology service models [1]". The development of the big data industry focuses on two aspects: one is to promote data openness and sharing, and the other is to create an independent industrial ecosystem in which data, technology, applications and security are coordinated to develop [2]. The financial big data industry, especially the banking industry, has a unique advantage in building a big data product system. It has a large number of customers. By integrating transaction information, operation information, customer attribute information, equipment information and other data in the trading day, it stores them in the ODS (Operational Data Store), and fully exploits the data value behind them by using big data analysis tools, it can create a series of big data products, such as accurate marketing, fraud identification, operation optimization and so on.

2. Overall Overview of the Project
The bank wind control anti-fraud project is an intelligent credit wind control application system built by the bank based on the existing equipment information, credit card information and retail credit application information, integrating various data of trading days. Through the construction of this project, the bank expects to break the traditional static and qualitative wind control system and establish a new set of automatic, intelligent, full-link, full-process financial wind control system.

The Bank Wind Control Anti-fraud Project proposes an anti-fraud solution for large financial data based on customer portrait data through mobile banking, personal online banking and counter channels. The purpose of this project is to strengthen the bank's credit risk control ability, improve its ability to identify malicious fraud, and further ensure the safety of customers' funds.

The big data platform is supported by big data technology and application. Through the combination of source system data and a certain dimension, a credit risk control anti-fraud system with credit application information, user information and equipment information is connected. The system
is designed to identify abnormal data, including false applications, internal and external collusion fraud, intermediary packaging fraud, gang fraud, etc.

In the process of project construction, according to actual business requirements, the bank does not need to access data in real time, but makes offline batch processing statistics on the transaction information of the previous trading day, stores the statistical results and provides an external query interface. Hive supports large-scale data storage and analysis and has good scalability [4]. It relies on HDFS for data storage and uses MapReduce, a distributed parallel computing framework, to process data. It is a reliable and intuitive data analysis tool [5]. Therefore, in the design process, developers used Hive as an offline batch processing tool based on Hadoop top-level data warehouse [6]. In order to meet the many requirements of staff search, developers choose Elastic Search, an enterprise search engine based on RESTful web interface, and deploy it in cluster mode. Elastic Search cluster can integrate storage and real-time search, and provide a safe, reliable and fast search solution for business queries.

3. Project Framework Design
Firstly, we will look at the operational framework of the banking system. As shown in Figure 1, the operation of the bank is based on data intelligence layer and technology support layer. When customers interact with the bank, their interactive information will be recorded in real time and integrated with the components of channel integration layer. The integrated data will enter the business service layer. Classified storage is carried out. When there is a need for data analysis, they are all large data components based on data intelligence layer and technology support layer for data processing, and then provide data support for management decision-making layer. Specifically, after banks obtain and collect credit application data through various channels such as mobile banking and personal online banking, they use the channel integration layer to integrate the public financial platform and submit it to the business service layer. The business service layer mainly undertakes most of the demand for credit application. This layer provides big data analysis and processing services for the anti-fraud system based on the technology support layer and the data intelligence layer. The processed data arrives at the counter gateway, and the retail credit application of the management decision layer obtains the response result by sending inquiry requests to the counter gateway.
The application architecture diagram of the anti-fraud system is shown in Figure 2. It is divided into four modules: retail credit application, cabinet gateway, query pre-management platform and offline data processing platform. When the retail credit application initiates an inquiry, the query request is forwarded by the counter gateway and arrives at the pre-query management platform. After receiving the request, the anti-fraud rules interface of the platform will query the Elasticity Search Service Center, which will query and return the Elasticity Search data in the offline computing platform. The result data will be fed back layer by layer and finally displayed to the searcher. In the offline data processing platform, Elastic Search saves synchronous and self-assertive data scheduling platform to perform Hadoop offline batch tasks regularly every day and generate Hive result tables.
4. System Development and Implementation

4.1. Design of Data Access Structure
The data access channels of the anti-fraud system include mobile banking, UnionPay front, direct selling bank, corporate online banking and other transaction channels. It is imported into Hive table using the data import platform developed within the bank. There are three types of table structures accessed. (1) flow chart; (2) snapshot table; (3) Complete scale. Pipeline table is Hive partition table, partition according to each day, data in partition is incremental data of the previous day, snapshot table is Hive partition table, partition every day, but the data in each partition contains the total data of the previous day, the full scale is not a partition table, when it is connected to Hive regularly every day. The library time is the full amount of data from the previous day. Banks use full scale for data access with less change such as equipment information, snapshot table for more important data access such as
customer application, data analysis and processing can be done only by dividing the date of one day before the transaction date.

4.2. Business Process Design

The business process design of the anti-fraud system is shown in Fig. 3. the source system is retail credit application and paperless credit card system. The data of the source system is stored in ODS. ODS regularly sends data to NAS (Network Attached Storage) in the early morning of each day. These data are accessed to specific tables in Hive through the data import platform, and Hive QL statements of processing business logic are encapsulated into SQL scripts and mounted on the big data scheduling platform for regular execution. Fusion Insight HD, a large data dispatching platform, performs off-line batch processing tasks on a daily basis, synchronizes the processing results to Elastic Search for storage, and finally submits them to the real-time interface of the pre-query management platform for invocation and query.

![Business process design of anti-fraud system](image)

4.3. Data Processing Implementation

In order to establish and perfect a unified and complete bank risk prevention and control system, the anti-fraud system should have the following modules or functions: anti-fraud business processing function, anti-fraud behavior analysis function, anti-fraud visual data report module and anti-fraud rule module. The design of anti-fraud rule module is the core of the whole system, which determines the system's ability to identify anomalies and prevent fraud. In credit risk control, fraudsters often use viruses to steal user accounts, try passwords in high frequency for a short time or pretend to be real natural people to defraud, and some fraudsters gather into gangs or disguise as intermediaries for high frequency. Fraud, and some fraudsters collude with bank staff inside and outside to commit fraud. These frauds seriously endanger the security of customers' funds and pose a huge threat to the internal systems of banks. In response to the above fraudulent behaviors, the anti-fraud system obtains all the data of the three types of source table (flow meter, snapshot table, full scale) through the Hive platform, and the developer designs the Hive QL statement to match the credit application device with the customer information. The device details information and customer attributes (such as customer
spouse information) are stereo-associated to generate a sparse table with integrated multi-information, and then logical processing based on this table to extract abnormal data and historical data or real normal data. Screening and comparison. The whole process is divided into two parts. Sparse tables are designed as partition tables. In addition to the original data, each partition keeps the incremental data of the day's transaction. The result table is an ordinary Hive table. The latest data coverage calculated every day is written into the table. When the large data dispatching platform completes the off-line batch processing task, downstream Elastic Search actively synchronizes the result table data and adds it. The index is stored in the Elastic Search cluster and provides the query interface. The anti-fraud model deploys the high-frequency database collision or camouflage fraud in depth, and encapsulates the whole link of the application process, effectively resisting the interference of the operation outside the application link, so as to build a solid risk prevention and control system.

4.4. Data Security Design
Because the big data platform stores the bank's user data, sensitive information of users may be included in its data analysis process. Big Data Platform adopts encryption mechanism for sensitive information in the process of entry, transmission and use, and the whole process of user information is not visible to the outside. And a certain method has been adopted to avoid data tampering. In the data storage process, the big data platform adopts a data backup mechanism to protect user data security.

4.5. Bottlenecks in Development and Solutions
In the process of data development, the author encounters the situation of data skew, and summarizes the solutions as follows.

First, filter out the invalid value or NULL value in the intermediate table data and then perform the multi-table join operation. This method can avoid the case where the join operation key value is tilted. Secondly, reasonably set the number of maps and reduce to avoid data skew caused by too many or too few. Finally, the Hive QL code is actively optimized. A table with a large amount of data is associated with a table with a large amount of data. In the previous stage of join operation, the generated data will be stored in the buffer of the reducer as an intermediate result. Therefore, the amount of data in the intermediate result should be kept as small as possible, and the table with the largest amount of data should be placed in the final association to reduce memory overhead and improve operation efficiency.

5. Application of Anti-Fraud System
The anti-fraud system discussed in this paper is mainly used in bank credit wind control management, and a new risk defense system is established for bank credit business before, during and after loan. In the process of running, it first summarizes the basic data in the credit application, and then analyzes and processes the data based on the anti-fraud rule model, and finally generates the result data or the visual report, which provides the decision support for the business personnel and management. In the application, the system has the following characteristics.

The anti-fraud system can traverse all the outlets across the region in the application, and the data is highly consistent in real time in the whole network. The main purpose of the design is to deal with the risk behaviors such as large-value transfer and password modification after login in different places. The system can alert or suspend for manual verification against these risk behaviors. The biggest advantage of the cross-regional operation of the system is to realize the maximum coverage of fraud prevention and the interconnection and cooperation of data.

Credit applications can be made through mobile phones, web pages or business systems. The anti-fraud system eliminates the difference in access caliber when it is built. It can seamlessly connect with the above channels, realizing cross-platform communication, and each application device has only a unique and non-repeatable number for life.

The anti-fraud system records all credit applications, and through the specification and comparison with historical information, it can find out the old list and real-name users with bad historical records,
as well as risk customers who have had high-frequency transactions.

The anti-fraud system uses automated data access software and automated task scheduling tools from data collection to data processing and analysis, and then to data storage and data feedback. The whole link is automated.

6. Conclusion

In view of the current trend of credit application fraud in the form of collectivization and specialization, an anti-fraud system based on big data technology emerges as the times require. This system is an innovative and innovative attempt of the bank in the credit data control system under the era of big data. It subtly utilizes the advantages of big data components in data analysis and processing, and proposes a kind of value mining for massive user data of banks. A new intelligent wind control model with multiple data integration and deep defense.

References

[1] Interpretation of the Ministry of Industry and Information Technology "Big Data Industry Development Plan (2016-2020)" [J]. China Information Security, 2017, (5): 59-60.
[2] Zhang Feng. (2016) Seize the new opportunities of big data development to promote the healthy development of big data [J]. China Informatization, (7): 9-11.
[3] Luo Jing. (2016) Design and implementation of web-based bank credit management system [D]. Xiamen: Xiamen University.
[4] Yao Di. (2011) Research on Key Technologies of Object Storage Security [D]. Wuhan: Huazhong University of Science and Technology.
[5] Wang Hongbin. (2015) Research and Application of web Log Mining Technology Based on Distributed Computing Platform [D]. Jinan: Shandong University.
[6] Kang Xiaodong. (2004) Data Mining Technology Based on Data Warehouse [M]. Beijing: Mechanical Industry Press, 34-34.