Design and Application of Bank Big Data Platform Based on Hadoop Technology

Hao Dong[1,a*] Wenquan Han[2b]
[1]Economics and Commerce Sejong University Seoul, South Korea
[2]Department of Economic Management Shandong Vocational College of Science and Technology Weifang, China

Abstract—At present, the design of big data platforms for many domestic banks is still insufficient. This article builds a Hadoop-based bank big data platform based on the big data platform construction experience of a large state-owned commercial bank. By introducing cutting-edge mainstream big data open-source tools, the overall architecture of the big data platform of commercial banks is built. At the same time, this article discusses critical technical solutions such as storage engine, resource management, calculation engine, analysis engine, interactive front end, and data management, task management, and user management to build a bank's big data platform. This article hopes to provide a reference for other banks to make big data platforms.

1 Introduction

The new round of technology represented by mobile Internet, cloud computing, big data, and artificial intelligence is rapidly changing traditional production and management methods. They have a widespread impact on the business model and even the intermediary function of commercial banks. Whether a commercial bank can make good use of big data and accelerate innovation to achieve transformation determines its future sustainable development capabilities. The traditional data analysis platform based on the relational data warehouses cannot meet the needs of current business development. Commercial banks need to use cloud computing and big data as the core technology to upgrade the single platform in the past into a diverse ecosystem[1]. The new big data platform can not only meet the basic processing needs of large data capacity, multiple types, and active circulation, but also support the collection, storage, processing, and analysis of big data, and meet the architecture's availability, scalability, and fault tolerance aspects. Basic guidelines. At the same time, it can meet the basic requirements for data analysis in the original format.

2 Necessity

With the rapid development of online banking, mobile banking, and electronic financial markets, China's banking industry has entered the era of big data. The existing online analysis and processing technology can no longer fully meet the massive demand of data resources for banking expansion, and it is not conducive to the smooth construction of national data centers and the further development of big data technologies. Commercial banks face severe challenges in processing big data. At present, neither the data processing software nor the hardware cost can meet the relevant standards, nor can the expansion performance of big data systems reach the optimal state. With the transformation of big data platforms from database platforms to cloud computing platforms, China's commercial banks' data analysis systems are at the forefront of the transformation of business intelligence into big data platforms. Continuously improving the scientific and rationality of the architecture design of big data platforms has become an issue that designers need to explore. Therefore, improving the structure of the bank's big data platform is of great significance to meet the bank's demand for data processing. At present, the distributed system infrastructure Hadoop has begun to apply[2]. The distributed system infrastructure provides services with the cooperation of multiple inexpensive PCs, with stable performance and high-speed data processing capabilities. Therefore, it has attracted the attention of many e-commerce companies and banks.

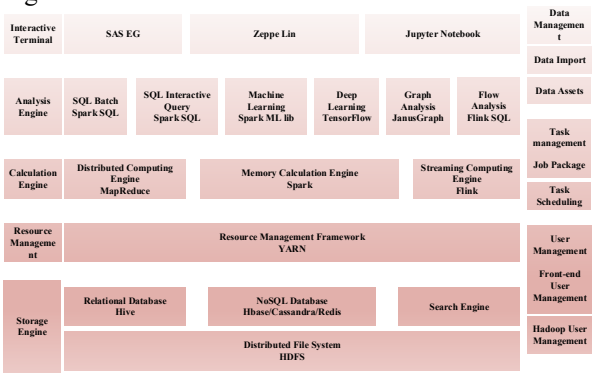
3 Overall Structure

The bank's big data platform is a big data ecosystem for bank data analysts, as well as a laboratory, toolbox, and knowledge base for the whole bank's big data analysis work. The bank introduced Hadoop technology based on a data warehouse to build an "MPP+Hadoop" dual-

[a]475536334@qq.com [b]hwqls@qq.com

engine architecture. Among them, the positioning of Hadoop has five aspects. First, for large-scale structured data, it is necessary to make use of Hadoop's features of distributed computing and high-performance memory computing to improve algorithm efficiency, to meet the needs of massive data exploration and analysis and mining. Second, it is utilized for Hadoop to store and calculate unstructured data such as text and image, and a deep learning algorithm is applied to enable it to analyze and mine and support application scenarios[3]. The third is to use Hadoop to provide streaming data processing function, to meet real-time data acquisition, data analysis, data delivery, and other application scenarios. Fourth, the graph analysis ability of Hadoop is utilized to meet the graph analysis needs of relational network exploration and knowledge graph construction. Fifthly, as the exploration environment of big data frontier tools, algorithms and methods, Hadoop is utilized to explore innovative applications.

During the construction of the big data platform, the bank continuously tracks the latest technological achievements in big data and related fields around Hadoop. Through in-depth research on the technology application in the field of big data, based on traditional SAS data mining tools, Python, R, TensorFlow, Zeppelin, Jupyter Notebook, JanusGraph, and other cutting-edge big data open-source tools are introduced. At present, the bank has formed an enterprise-level Hadoop big data platform, including storage engine, resource management, calculation engine, analysis engine, interactive front-end, data management, task management, user management, and so on. The overall architecture of the bank's big data platform is shown in Figure 1.



- Figure 1 Big data platform architecture design

4 Key Technical Solutions

A. Storage Engine

Enterprise-level Hadoop big data platforms need to provide storage capabilities and solutions for data of different data types and different application types. As we all know, commercial banks' business operations for many years have accumulated not only large amounts of structured data such as customer information, but also massive unstructured data such as text and images. Massive structured and unstructured data requires standardized and orderly management. Therefore, the Hadoop big data platform built by the bank is optimized for data storage. First, the bank divided different data areas according to different data types, including semi-structured data areas, structured data areas, unstructured data areas, and so on. Secondly, the bank selects different storage technologies for different data partitions, thereby realizing the storage needs for different types of data. For semi-structured data areas, the bank uses HDFS components to store semi-structured data such as log files. For the structured data area, the bank uses Hive components to store traditional structured data such as customers, accounts, and transactions. For the unstructured data area, the bank uses HBase's key-value storage structure to store unstructured data such as images[4]. Finally, the bank subdivides each data partition into a basic data layer and an application data layer according to the type of application. The basic data layer is the primary shared data source for big data analysis and mining. The application data layer is a data processing storage area for each data analysis application item. At the same time, the bank conducts data lifecycle management according to the cycle of the project. Besides, data access control of different departments is a key point that needs to be considered in data storage. The bank uses Hive's internal and external appearance mechanism and data partitioning mechanism to implement data splitting according to institutions to achieve the purpose of controlling data access rights.

B. Resource Management

Enterprise-level Hadoop big data platforms need to solve the resource management problems of different types of jobs in different departments and scenarios. Hadoop platform resources are divided into computing resources and storage resources. In computing resource management, YARN, based on MapReduce 1.0, has become universal resource management for Hadoop 2.0. The queuing mechanism can realize the allocation and management of computing resources and the scheduling strategy of different resources. Multi-tenant technology based on YARN can not only meet the allocation and management of computing resources but also meet the quota management of storage resources and the management of HBase service resources. YARN provides a solution for the overall resource management of the Hadoop big data platform.

C. Calculation Engine

Enterprise-level Hadoop big data platforms require distributed high-performance computing functions and real-time computing functions as the basis for upper-level analysis engines. MapReduce, as a distributed computing framework, provides distributed parallel computing capabilities for large-scale data sets. Spark has improved problems with MapReduce. Spark can provide a high-performance memory iterative computing framework, which greatly increases the speed of parallel computing, so it is particularly suitable for complex computing modes that require multiple iterations in data mining scenarios. Besides, Spark has a unique advantage. It can be based on SparkCore and use SparkSQL,

SparkMLlib, and other components for batch processing and mining of data, to provide solutions for the upper-level analysis engine. Flink is a new star in the field of big data processing. For streaming data and batch data, Flink provides a low-latency distributed real-time processing engine. And Flind has better real-time processing performance than Spark[5]. At present, Flink has become a popular technology choice for streaming computing engines in the field of big data. After research and selection testing, the bank chose MapReduce, Spark, and Flink components to build the calculation engine as the basis of the upper-level analysis engine.

D. Analysis Engine

The Hadoop big data platform needs to provide SQL batch processing, SQL interactive query, machine learning engine, deep learning, streaming analysis, graph analysis, and other analysis engines to meet the analysis needs of different scenarios. The bank chose SparkSQL as the engine for SQL batch processing and interactive query, SparkMLlib as the machine learning engine, TensorFlow as the deep learning engine, Flink as the streaming analysis engine, and JanusGraph as the graph analysis engine. SQL batch processing engine and SQL interactive query engine have HiveonMapReduce, HiveonSpark, SparkSQL, and other candidate solutions. Among them, HiveonMapReduce has a relatively high delay compared to HiveonSpark and SparkSQL, and should not be used as the preferred solution for interactive query. HiveonSpark and SparkSQL are SQL solutions based on the Spark engine. From the perspective of community activity and technology development trends, SparkSQL is even more recommended as a SQL engine. SparkMLlib, as Spark's machine learning library, consists of some general learning algorithms and tools, including classification, regression, clustering, collaborative filtering, and dimensionality reduction. It also includes low-level optimization primitives and high-level pipeline APIs. Besides, with the in-depth application of artificial intelligence and data analysis, Python has also received great attention. Therefore, it is recommended as the preferred language for SparkMLlib data analysis mining.

E. Interactive Front End

According to the Big Data Panorama report published annually by Matt Turck, Besides to SASEG, Zeppelin and Jupyter Notebook are mainstream open-source visual interactive analysis tools. They support the Hadoop-based data mining program writing and real-time graphical representation. Zeppelin is a web-based notebook. It provides functions such as data ingestion, data discovery, data analysis, data visualization, and collaboration, and supports multi-user management. Zeppelin's core concept is the interpreter. It provides interpreters in languagessuch as Python, R, Python (ApacheSpark), SparkSQL, Hive, HBase, Flink, etc. It supports users to use their specific programming languagesor data processing methods for data analysis and mining. Jupyter Notebook is also an interactive notebook that supports running multiple programming languages such as Python. It can be used for data cleaning and transformation, numerical simulation, statistical modeling, machine learning, etc. Jupyter Notebook provides multi-user management capabilities through JupyterHub[6]. Overall, Zeppelin and Jupyter NoteBook have their advantages. The bank's solution is to deploy both Zeppelin and Jupyter Notebook as an interactive analysis tool for the Hadoop platform.

F. Data Management

The bank's data management mainly includes data asset management and data import management. Data asset management mainly plans the HDFS storage directory structure and formulates naming specifications such as HBase namespace and Hive library table objects. At the same time, it also needs to provide data asset query functions to achieve effective management of data assets and help users quickly use data. Among them, the HDFS directory structure needs to be unified planned by with the dimensions of the organization and application type to effectively and orderly manage the data files. The data asset query function helps users to see what data is on the platform, and how often and when the data is loaded. Data import management is responsible for importing various types of data required for big data analysis work through the use of data replication technology components and other means from data sources such as enterprise-level data warehouses to the Hadoop platform.

G. Task Management

On the Hadoop big data platform, the automatic operation and management of tasks is a big tentacle to embed the application value of big data into business processes. From the perspective of Software Engineering IPO (Input-Process-Out), task management includes centralized scheduling and monitoring management such as data import, data analysis, and data delivery. Data analysis job refers to a data analysis and mining program written in languages such as SparkSQL, Python (ApacheSpark), etc., and encapsulated by a shell script to enable it to support periodic execution and automated scheduling. Data delivery refers to pushing the model results to the SMS platform, WeChat public account, and various business systems to achieve online delivery of data.

H. User Management

Multi-user management is an important basis for promoting the use of the platform and improving the vitality of the platform. User management of the Hadoop big data platform includes Hadoop user management and interactive end-user management. The bank uses a standard RBAC model for Hadoop user management. It defines the access rights of objects through roles, assigns user access rights to objects by binding roles to users, and uses the "user-department-role-permission" relationship chain to specify user operation rights. In the user dimension, users are created according to the granularity of business departments and business applications. In the role dimension, two roles, ordinary role and tenant role, are bound for each user. Common roles give users permissions to operate data objects such

as HDFS, HBase, and Hive. Tenant roles give users access to tenant services and resources. The goal of interactive end-user management is to unify users and permissions for tools such as SAS, Zeppelin, JupyterNoteBook. It supports users to log in to each tool with only one set of username and password and has consistent permissions. The bank has adopted LDAP technology to provide a unified user authentication system for the entire bank. Through subsequent development or integrated configuration of the above tools, unified management and unified authentication of end-users are achieved. In terms of unified authority, the end-user-operating-system user-Hadoop user relationship chain is used to give end-users access to Hadoop resources.

5 Conclusion

Driven by the "Internet +" and "Digital" national strategies, whether commercial banks can make good use of big data and consolidate their information resource advantages determines their future sustainable development capabilities. Commercial banks need to embrace fintech actively, keep track of the latest developments in the industry and explore cutting-edge technologies, continue to improve the capabilities of big data platforms, and provide a stronger driving force for digital transformation.

Acknowledgment

This paper is the stage achievement of the social science planning project of shandong province in 2019. (Project Name. shandong province digital economy and real economy integration mechanism, evaluation and countermeasure research. Project Number. 19CPYJ90)

References

1. Zhang Boyu, Research on Data Audit Mode of Commercial Banks Based on Hadoop Technology, China Internal Audit, 2019, (5): 33-37.
2. Zhang Dengyao, Big Data Analysis of Commercial Banks Based on Hadoop Distributed File System .Journal of Shandong Agricultural University (Natural Science Edition), 2018,49 (5): 884-888.
3. Han Jian, Architecture Design of Bank Big Data Platform Based on Hadoop Technology, Electronic World, 2017, (22): 162-163.
4. Chen Yan, Research on Online Banking Historical Data Based on Hadoop, Computer Knowledge and Technology, 2017,13 (16): 21-23.
5. Li Deyou, Zhao Libo, Xie Chenguang, Research on Bank Mass Data Storage System Based on Hadoop .Journal of Harbin University of Science and Technology, 2015,20 (4): 60-65.
6. Xin Huaiyi, Research on HDFS Copy Strategy Optimization Based on the Big Data Access Law of Commercial Banks, Software, 2015,36 (11): 74-79.