DOI 10.15302/J-SSCAE-2022.04.014

Next-generation Database Benchmark for Financial Scenarios

Jing Yinan[1, 2], Zhang Hanbing[2], Li Zhixin[2], Wang Xiaoyang[1, 2], Wu Jie[1, 2], Chai Hongfeng[1, 2]

1. Institute of Financial Technology, Fudan University, Shanghai 200433, China
2. School of Computer Science, Fudan University, Shanghai 200433, China

Abstract: Banks, which are major financial entities in China, have high performance and security requirements for databases and data service solutions. With the progression of data application services in banking, the data types and business scenarios have become more diverse, and it is difficult for users to make optimal choices among a wide diversity of database products and data service solutions. In combination with the data application demands of the financial industry, this study comprehensively analyzes the current status of database applications in banking, particularly the status and challenges of database localization in recent years, using literature review and theoretical analysis. In addition, we systematically investigate the database benchmarks of China and other countries and further examine the necessity and importance of constructing next-generation database benchmarks for financial scenarios. Current database benchmarks have many shortcomings and face various challenges in dealing with database testing in financial scenarios owing to complex business logic, diverse data patterns, and high security requirements. Therefore, to build a next-generation database benchmark that can meet the requirements of financial scenarios, we propose several suggestions to address these challenges, which involves workloads, data schemes, metrics, and technical architecture.

Keywords: financial industry; bank; financial data; database; benchmark

1 Introduction

As the economy grows steadily and the economic and financial systems progress, the financial industry has become increasingly important to China's national economy [1]. By the end of 2021, the total assets of China's financial industry amounted to RMB 381.95 trillion, an increase of 8.1% year-on-year. In the financial industry, banking had total assets of RMB 344.76 trillion, an increase of 7.8% year-on-year, occupying a major position among all types of financial institutions (banks, securities, and insurance) [2]. With the advent of the mobile Internet era, mobile phone transactions are gradually increasing, and a large amount of data and information is being generated within the banking system every moment; thus, managing these data safely and effectively has become an important demand for banking business systems [3]. The *FinTech Development Plan (2022–2025)* released by the People's Bank of China also states that the overall level and core competitiveness of financial technology needs to be improved by leaps and bounds, data capacity needs to be comprehensively strengthened, and a large number of green data centers need to be built [4].

As a basic software that stores data and provides computing functions, database management systems and data service solutions are critical to the banking business system. They exert hardware computing power downward and support upper-level applications upward. In the 1990s, banks in China used foreign database products, such as Oracle and DB2, to achieve cross-branch deposits and withdrawal business functions (pass-through deposit and exchange).

Received date: June 19, 2022; **revised date:** July 14, 2022

Corresponding author: Jing Yinan, associate professor from School of Computer Science, Fudan University. Email: jingyn@fudan.edu.cn

Funding program: CAE Advisory Project "Strategic Research on Financial Data Security Governance Intelligentization" (2022-XY-12); National Natural Science Foundation of China (92046024, 92146002)

Chinese version: Strategic Study of CAE 2022, 24 (4): 121-132

Cited item: Jing Yinan et al. Next-generation Database Benchmark for Financial Scenarios . *Strategic Study of CAE*, https://doi.org/10.15302/J-SSCAE-2022.04.014

In recent years, influenced by business needs and the international situation, many financial institutions, mainly banks, have gradually transitioned from foreign commercial databases to domestic distributed databases [5] to improve their business systems' performance while avoiding the potential supply chain risks associated with using foreign commercial database products. With China's emphasis on financial data security, replacing databases in the financial industry with domestic databases has been the trend.

Database benchmarks are a significant reference for user choice as a basis for database performance evaluation. These benchmarks can provide a fair and objective test for various database products and data service solutions in a given scenario. However, the existing database benchmarks face many challenges when dealing with database testing in financial scenarios. On one hand, the financial industry (e.g., banking) has more complex business logic and multiple patterns of data. However, existing benchmarks [6,7] lack the ability to comprehensively and accurately test database and data service solutions in this complex environment. On the other hand, it has a higher data security requirement. As financial data security is related to the country's livelihood, compared with existing benchmarks, the benchmarks in financial scenarios need to have stronger and more comprehensive reliability and security testing capabilities to assist in the security governance of financial data and provide assurance for the digital transformation of finance. In addition, as the financial industry continues to replace its databases with domestic distributed databases, benchmarks in financial scenarios need to be able to evaluate the portability, compatibility, and adaptability of database products and data service solutions to distributed architectures, which lack existing benchmarks. Therefore, there is an urgent need to build a database benchmark that meets the needs of financial business development to provide a unified evaluation and measurement of database products and data service solutions, help financial practitioners make more accurate choices, and guide the development of data service vendors.

This study first elaborates the current development of database applications in banking, provides an in-depth analysis of the new requirements and development trends in bank databases in the new era, and introduces the replacement of domestic databases in banks and the challenges they face. Second, it analyzes the main database benchmarks in China and abroad, and the necessity and importance of building a next-generation database benchmark for financial scenarios, considering the data application development needs of the financial industry.

2 The current status of database application in China's banking

2.1 Development and evolution of database applications in banking

In the information age, database systems, as the basic software for storing and managing data, play a key role in banks' financial systems and are directly related to their stability. As shown in Fig. 1, over 40 years of bank informatization development, the construction of databases in the financial industry has gone through the era of standalone manual bookkeeping, interconnection of business data between branches and nodes, large centralization of data in the head office, service-oriented architecture (SOA), and the current and future era of distributed microservices. From the beginning of the interconnection era (1990s), foreign database products were gradually applied to banks, opening the door to the construction of databases in China's financial industry. Later, with the continuous development of China's database industry, some domestic databases were used in many financial institutions, including large state-owned, joint-stock, and city commercial banks, around 2017 and have shown excellent performance [5]. Currently, domestic database vendors can be broadly divided into three types of enterprises: Internet enterprises represented by Alibaba Group Holdings Limited; traditional database companies represented by Beijing Kingbase Technology Inc.; and comprehensive information technology service enterprises represented by Huawei Technologies Company. As of June 2021, there are 135 domestic database products in China [8]. However, the market share of foreign database products, such as Oracle and DB2, exceeds 80% in the RMB 20 billion banking database software market in 2020 [9], which also shows that China's domestic databases have a broad space for development.

2.2 Requirements of banking databases in the new era

Currently, the financial industry has four new requirements for database applications [8]. First, with the rapid development of mobile Internet, the volume of data generated in financial business systems is growing, which places higher demands on data storage and management capabilities of database systems. Second, the implementation of inclusive finance requires database systems to have stronger disaster recovery capabilities to ensure business continuity. Third, with the popularity of electronic payments, database systems need stronger performance to cope with the system pressure caused by high concurrent business and user volume. Fourth, to prevent potential supply chain risks, there is a demand for database localization at the technical level to avoid threats to financial data security.

DOI 10.15302/J-SSCAE-2022.04.014

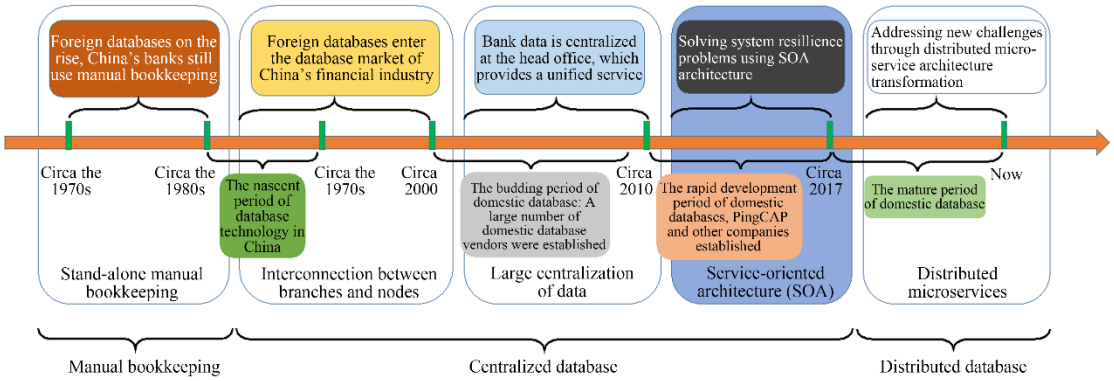

Fig. 1. Evolution of database applications in the banking industry.

Overall, to adapt to the digital transformation and upgrade of banks and meet business development needs, the database of banks in the new era should not only meet the basic elements of databases, such as atomicity, consistency, isolation, and persistence (ACID), but also have characteristics that are different from traditional ones to keep up with or even lead the times. These new requirements include the following: (1) scalability, database systems need to have the ability to expand data storage, access, computing, and other aspects, especially the ability to expand horizontally; (2) autonomy, autonomous control is the premise of information security, and is required for China's financial services development. Therefore, the financial industry, especially the banking business system in the database system should be autonomous and controllable; (3) massive, with the popularity of electronic payments, the database system needs to have the ability to support massive data storage and calculation; (4) real-time, the database system needs to have the ability to process user business in real time under high concurrency environment; (5) high availability, the database system necessitates sufficient disaster recovery ability to provide round-the-clock uninterrupted service to guarantee the stable operation of upper-level business; (6) security, the database system requires sufficient security to provide a guarantee for financial data security; (7) migratability, the database system needs to be able to migrate the business data stored in foreign databases (e.g. Oracle and DB2) perfectly and ensure data integrity and availability.

2.3 New trends in the development of banking databases in the new era

To cope with the performance pressure brought about by business scenarios that require massive data storage and calculation (such as mobile phone payments, loan risk calculation, and bank card theft investigation), ensure the autonomy and control of China's financial information systems to prevent them from being negatively affected by international unilateralism and trade protectionism, three trends have emerged in the use of databases in the financial sector, especially in banks.

2.3.1 Distributed database transformation

As businesses continue to grow, the amount of data to be processed in financial business systems increases dramatically, while existing centralized databases face data processing bottlenecks, and the ability to expand through hardware upgrades is costly and capped. Therefore, to meet the increasing performance demands in financial business systems, it is imperative to transform the existing centralized database into a distributed database that can enhance system performance by adding storage and computing nodes. The *FinTech Development Plan (2019–2021)* [10] issued by the People's Bank of China highlights that the research and application of distributed databases should be strengthened to ensure their sound application in the financial industry. In 2020, three financial industry standards, including the *Financial Application Specification of Distributed Database Technology—Technical Architecture (JR/T 0203-2020)* [11], were released and implemented to guide the application of distributed databases in the financial industry.

2.3.2 Database localization replacement

With the promotion of national strategies, such as national policy-led IT application innovation, network power, information security, and big data, China's demand for data exploitation is gradually increasing, and there is more focus on data security. As the basic software that carries data storage and calculation functions, large-scale use of domestic database products in the financial industry is inevitable to ensure the security of financial data. In addition,

owing to the current international situation, there are many risks in using foreign commercial database products in the financial system. Simultaneously, several cases have shown that domestic databases have better performance levels in financial business systems, which has also boosted financial institutions' confidence in choosing domestic database products.

2.3.3 Not only structured query language (NoSQL) and multimodal database applications

With the rapid growth of data volume in business scenarios that require data analysis, such as loan risk calculation, large transaction judgment, and credit card skimming warning, using traditional relational databases to analyze the huge amount of data can no longer meet the processing speed requirements in these scenarios. To improve the efficiency of data analysis and protect people's assets, the use of not only structured query language (NoSQL) and multimodal databases is needed to store and manage financial data in financial business systems. For example, in a loan risk calculation scenario in which multiple layers of transaction paths need to be analyzed, converting relational data (such as accounts and transaction records) into graph data, and using graph databases (such as Neo4j) to store and analyze these data can provide faster results than relational databases. In addition, some business systems require multimodal databases that can simultaneously support centralized storage, querying, and processing of multiple types of data to meet the system's need for the unified management of structured, semi-structured, and unstructured data. Therefore, to cope with the performance pressure caused by the huge amount of data and to meet the speed requirements for data analysis in various business scenarios, many NoSQL and multimodal databases will be used in financial business systems in the future.

2.4 Current status and challenges of domestic database replacement in banking

Owing to the business needs of banks and the international situation, there are many potential risks when banks use foreign database products, such as Oracle and DB2. Consequently, many financial institutions, including large state-owned, joint-stock, and urban commercial banks, have gradually transitioned from foreign commercial databases to domestic databases [5]. Major state-owned commercial banks have not yet replaced their databases with domestic databases in their major core business systems because they are not able to migrate data from the databases they are using to new databases, while ensuring data integrity and business continuity. However, in some business systems that require massive data processing or do not involve core business, domestic databases with good performance have gradually been introduced, such as TiDB from Beijing Pingkai Xingchen Technology Development Co., Ltd., GBase from Tianjin Nanda General Data Technology Co., Ltd., and OceanBase from Beijing OceanBase Technology Co., Ltd. Joint-stock banks not only adopt domestic databases in the fields of data analysis and model development, but also for traditional transaction-based systems. For example, China CITIC Bank has adopted GoldenDB (a distributed database developed by ZTE Corporation) in the transaction-based business [12]; Bank of Communications has independently developed CBase, a distributed database based on the OceanBase database, and moved the bank's debit card data from the DB2 database to CBase database, resulting in significant performance improvement [13]. Urban commercial banks have fewer customers and involve far less data than large banks. In addition to domestic database replacement in peripheral and emerging businesses, replacement in some core systems is commonplace. For example, the Bank of Nanjing uses the OceanBase database to build a complete core banking system. Overall, domestic databases are autonomous, scalable, high-performing, and highly available, and can meet the needs of online, high-frequency, multi-dimensional, and high-convergence scenarios, helping financial institutions solve technical bottlenecks. The widespread application of distributed, cloud computing, and hybrid deployment architectures in domestic databases can also significantly improve cost control and optimization to achieve cost reduction and efficiency enhancement, which is conducive for the digital transformation of finance.

Currently, there are some challenges in the process of replacing bank databases with domestic ones. First is architectural transformation. The financial industry uses many centralized databases and has rich experience in operation and maintenance. The rapid and effective implementation of the transition from centralized to distributed databases has become a challenge faced by financial institutions. In addition, the financial industry, particularly banks, has extremely stringent database requirements, requiring distributed databases to have high stability and security standards [14]. Second, is data migration. The financial industry has been using foreign database products (such as Oracle and DB2) for a long time; therefore, a large amount of business data is stored in these databases. Migrating data from the database in use to a new database in a highly efficient and low-cost manner to ensure data integrity and uninterrupted business systems has become a major challenge for financial institutions. Third is database product selection. A wide range of database products and data service solutions are available, and there are no standard evaluation indicators for their performance in financial business scenarios. Existing typical database

DOI 10.15302/J-SSCAE-2022.04.014

benchmarks, such as TPC-C [15], TPC-H [6], and TPC-DS [7] provide test tools and documentation for evaluating database products in different scenarios. However, the business logic in financial application scenarios differs from that of business simulated by existing benchmarks. Using these benchmarks to directly evaluate database products in financial scenarios is not a good idea. Therefore, there are many shortcomings in using existing benchmarks to evaluate database products in financial scenarios. There is an urgent need for a financial database benchmark to uniformly evaluate database performance and data service solutions in various financial scenarios.

3 Current status of research on database benchmarks

3.1 Evolution of database benchmarks

For more than half a century, database technology has evolved from simple hierarchical and network models to relational, time series, and graph databases, which are more suitable for existing application scenarios [16] and have injected new energy into the database field. However, with the wide range of database products available, it is difficult to choose a solution that meets user requirements. For example, for data analysis in social networks, graph databases such as Neo4j can better characterize data relationships than traditional relational databases, whereas for massive data analysis, distributed databases are more effective at handling large data volumes and highly concurrent requests. Owing to the complexity of application scenarios and the variety of product choices, it is essential to design database benchmarks. In the database field, a reliable benchmark can fairly and objectively reflect the performance level of different database systems under the same test standard and provide a reference for users to choose database products and data service solutions.

An effective database benchmark consists of data schema, workload, and metrics. The data schema largely determines the application scenario for the benchmark and describes the data composition and structure according to the target application scenario. Most existing benchmarks are designed to generate test datasets that are as close to the real data as possible by setting similar data size, structure, distribution, and correlation [17]. Workloads are used to measure the specific performance of a database system in a given scenario and are the core of the benchmark. Depending on the application domain, computational paradigm, intensity type, and latency requirements, workloads in different benchmarks must have different characteristics [18]. Metrics are often used to describe the performance of the object under evaluation. Commonly used metrics include cache hit rate, throughput, query latency, resource utilization, and system scalability.

Fig. 2 describes the main developments in database benchmarks. The Wisconsin Benchmark [19] was developed in 1983 to resolve the debate over the merits of a large number of database systems. It consisted of three tables and 32 Structured Query Language (SQL) sentences, with total runtime as the performance metric. This benchmark provided a good way for manufacturers of relational database management systems to evaluate the performance of their systems and was a pioneer in database benchmarks. The Wisconsin benchmark also led to the establishment of the Transaction Processing Committee (TPC). The most commonly used database benchmarks are mainly published by TPC, including TPC-C, TPC-H, TPC-DS, among others. For example, the recent performance debate between Databricks and Snowflake has been evaluated by using TPC-DS [20–22]. In addition, as data volumes continue to grow and new application scenarios emerge, a number of new benchmarks have been developed, such as TPCx-BB [23] for big data applications, TS-Benchmark [24] for time series databases, and Graph 500 [25] for graph databases. In summary, the development of database technology has brought about the need for new benchmarks, and database benchmarks have prompted database vendors to further improve existing issues and continuously enhance the usability of their systems, both of which promote each other and develop together.

3.2 Benchmarks for relational databases

Relational databases can be divided into online transaction processing (OLTP), online analytical processing (OLAP), and hybrid transaction/analytical processing (HTAP) databases according to their roles. The performance of these databases was evaluated using the following benchmarks.

3.2.1 Benchmarks for OLTP databases

OLTP databases are usually used for transactional purposes and have ACID characteristics to ensure that transactions are correct and reliable during the writing or updating of data. In this context, a transaction is a logical process that consists of a series of database operations.

TPC-C: TPC-C is a benchmark for transaction processing systems that simulates warehouse order management scenarios [15]. The TPC-C contains nine tables and five transaction types, providing comprehensive coverage of

warehouse order management operations. In addition, the TPC-C uses the throughput of the transaction as a metric to describe the number of transactions processed per minute and the price/performance to measure the system's energy consumption.

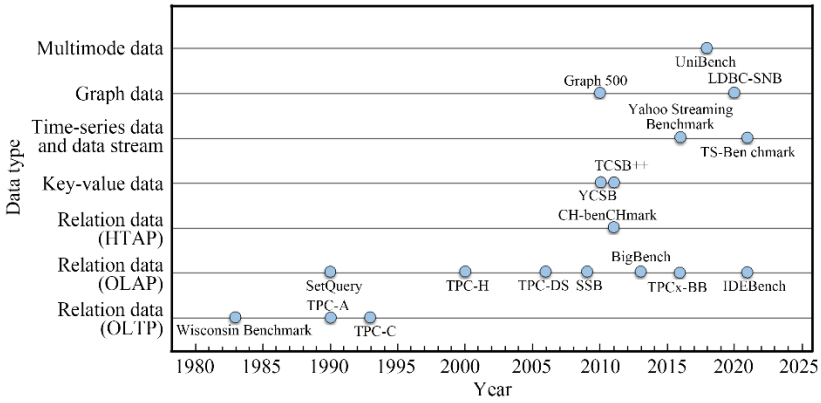

Fig. 2. Evolution of database benchmarks.

3.2.2 Benchmarks for OLAP databases

To improve the efficiency of analysis and decision making, a complex analysis of large-scale data is generally supported through the creation of data warehouses. Therefore, OLAP databases focus more on complex queries and aggregate analysis with an emphasis on decision support.

TPC-H: TPC-H is designed for data analysis systems and simulates realistic business procurement applications [6]. The benchmark is based on the 3NF data model and includes eight data tables (one fact table and seven dimension tables), each with a uniform data distribution. The TPC-H workload contains 22 ad-hoc queries and data update operations that simulate the typical business logic of business procurement applications. TPC-H uses the number of queries executed per hour (throughput) and the price/performance as metrics to evaluate the database performance. However, a uniform data distribution in TPC-H is ideal. It has been shown that TPC-H is not suitable for cardinality estimation-related measurements [26]. Furthermore, as data and business systems become more complex, the data model and workload in TPC-H are too simple.

TPC-DS: TPC-DS is a benchmark for evaluating decision-support systems [7]. The TPC organization introduced it to replace TPC-H. TPC-DS is designed using the snowflake model and contains seven fact tables and 17-dimension tables, with 99 queries and 12 data maintenance operations in the workload. Compared to TPC-H, TPC-DS uses more data tables, and the data distribution is skewed, which is closer to real data. The workload is also more complex and includes various query types, such as ad hoc queries, reporting queries, iterative queries, and data mining, simulating real business logic from multiple perspectives. Most test cases have high I/O loads and CPU computing requirements, allowing for a more comprehensive evaluation of the database system. TPC-DS is a recognized benchmark for its wide range of applications and ability to reflect the actual state of the system.

Star schema benchmark (SSB): SSB was proposed by researchers at the University of Massachusetts Boston to evaluate the performance of decision-support systems [27]. The SSB workload contains 13 queries, which replace some of the complex query cases in the TPC-H workload with more structured OLAP queries and divides them into four query patterns based on query type and query selectivity to provide more applicable functional and selectivity coverage tests.

TPCx-BB: TPCx-BB was formerly known as BigBench [28]. It was accepted by TPC and named as TPCx-BB in 2016 [23]. TPCx-BB is an end-to-end big data benchmark that provides 30 test cases simulating the processing, storage, and analysis scenarios of big data to measure the hardware and software performance of big data systems. It also includes semi-structured web log data and unstructured product review data that are more relevant to real application scenarios. Furthermore, TPCx-BB is designed to measure the performance of big data systems in one business dimension and three technical dimensions, matching the characteristics of business applications in big data.

IDEBench: With the rapid growth in data volumes, interactive data exploration in large datasets has become a typical analytical application and research hotspot. Researchers have proposed techniques to improve query processing, such as approximate query processing, based on sampling theory. However, the workloads and metrics of popular benchmarks (e.g., TPC-DS and SSB) are designed for traditional business application scenarios, and do not capture the unique nature of interactive data exploration. Therefore, IDEBench has been proposed as a

DOI 10.15302/J-SSCAE-2022.04.014

benchmark for interactive data exploration [29]. IDEBench generates test data based on real datasets, scales user-defined datasets to arbitrary sizes, and maintains their original distributions. It divides IDE queries into multiple types, mimics dependencies between user exploration behaviors by defining linking relationships between queries and generates workloads as sequences of temporally ordered aggregated queries.

3.2.3 Benchmarks for HTAP databases

The HTAP databases are hybrid relational databases that provide both OLTP and OLAP [16]. Some HTAP databases support OLTP and OLAP queries using row-column mixing or row-column conversion techniques.

CH-benCHmark: CH-benCHmark [30] combines two benchmarks: TPC-C and TPC-H. It expands the data schema of TPC-C with three data tables from TPC-H and contains 12 data tables. In addition, CH-benCHmark combines the TPC-C workload with the rewritten TPC-H workload to build a hybrid workload that contains both OLAP and OLTP queries. To measure the performance of the databases under a hybrid workload, CH-benCHmark uses the ratio between the number of transactions and queries executed per hour as the metric.

3.3 Benchmarks for NoSQL databases

Additionally, new database benchmarks are being developed with the proliferation of the NoSQL database. In this study, we analyzed four typical database benchmarks for NoSQL databases.

3.3.1 Benchmarks for key-value databases and wide-column databases

Key-value databases and wide-column databases are two common types of NoSQL databases used to accelerate the efficiency of data analysis in systems. Specifically, the key-value database is a type of NoSQL database that stores, retrieves, and manages a collection of key-value pairs, similar to mapping or a dictionary, where each key is unique within the collection, and the value of the key contains various fields or data. A wide-column database is a type of NoSQL database that uses tables, rows, and columns for data storage. It is optimized for the fast retrieval of columns and is often used for online analytical processing.

Yahoo! Cloud Serving Benchmark (YCSB): YCSB is an open-source benchmark for comparing the performances of NoSQL databases. It aims to facilitate the performance comparisons of next-generation cloud data service systems [31]. Currently, YCSB defines a core set of benchmarks for four widely used systems (Cassandra, HBase, PNUTS, and MySQL) and reports the test results. Additionally, YCSB is scalable and supports the definition of new workloads.

YCSB++ [32] extends YCSB to include parallelism testing, weak consistency testing, and block upload testing. YCSB++ provides distributed synchronization between multiple test clients and can measure the optimization of consistency, batch loading, and batch writing. In addition, YCSB++ can measure the performance overhead of additional functions in the database (e.g., access control) and collect resource-monitoring information for each cluster node.

3.3.2 Benchmarks for time series databases and streaming databases

TS-Benchmark: Time-series data are widely used in the Internet of Things, digital finance, and weather forecasting. With the emergence of time series databases, such as InfluxDB and IoTDB, a benchmark for time-series databases has emerged, TS-Benchmark [24]. TS-Benchmark primarily applies a scenario of device monitoring for wind turbines and can systematically evaluate the performance of time-series databases under three types of workload modes: data loading, streaming data injection, and historical data access.

Yahoo Streaming Benchmark: With the development of big data and cloud computing, the number of applications based on streaming data has proliferated, giving rise to many new streaming computing engines (e.g., Storm, Flink, and Spark Streaming). Researchers have proposed the Yahoo Streaming Benchmark to facilitate users to compare different streaming engines and choose the right product, which evaluates streaming engines and has been widely accepted by the industry [33]. The Yahoo Streaming Benchmark constructs a full data pipeline using Kafka and Redis and simulates a real advertising analysis scenario to design the stream processing workload.

3.3.3 Benchmarks for graph databases

Graph 500: The computational performance of supercomputers is measured in Graph 500 [25]. This benchmark covers graph-related scenarios (e.g., cyber security, medical informatics, and social media) and focuses on three problems: concurrent search, single-source shortest path, and maximum independent set.

LDBC-SNB: LDBC-SNB is a benchmark for simulating social network scenarios [34]. It is designed to generate data schemas based on the characteristics of social networks and evaluate graph databases with two typical

workloads, which can cover the business of social networks and is generally applicable and representative. One is an interactive workload that includes tasks, such as neighbor queries and data insertion at a given node. The other is a business intelligence workload that includes complex queries, such as aggregation functions and multi-graph joins, which involve most of the nodes in the graph.

3.3.4 Benchmarks for multi-model databases

UniBench: Unlike databases built around a single data model, multi-model databases can support many different data models simultaneously. UniBench [35] is a benchmark designed to evaluate multi-model databases, consisting of a set of mixed data models that simulate social business applications with a data generator that provides a variety of data formats, including XML, key-value, text, and graphs. The workload in UniBench consists of a set of multi-model read-only and read-write transactions. It involves at least two data models that cover many aspects of multi-model data management.

3.4 New issues facing existing benchmarks in financial scenarios

Database products are widely used in financial business scenarios, especially in banks' public and private business lines. As a result, financial databases are a hot spot of competition for database vendors. Currently, many database products and solutions are available with varying business coverage and service performance, making it difficult for banks to make an appropriate choice. Although a growing number of database benchmarks has significantly contributed to the advancement and development of database technology (e.g., TPC-C and TPC-H), a significant gap remains in the evaluation of financial database products and data service solutions. Furthermore, with the development of data application services in the financial industry, the types of data and business scenarios involved have become more diverse, and existing benchmarks face many problems in financial data services. Therefore, database benchmarks for financial scenarios must be developed and improved to promote the flourishing of financial database technology.

3.4.1 Business scenario adaptation

In terms of business scenarios, the business logic (workload) in financial application scenarios is different from that of business application scenarios simulated by existing benchmarks. Hence, there are many shortcomings in directly using existing benchmarks to evaluate database products and data-service solutions in financial scenarios. For example, by summarizing the business logic of six common financial scenarios, namely transfer, deposit, withdrawal, account inquiry, payroll, and asset inventory [36], researchers found that multiple access operations and security validation are required for these operations. However, existing benchmarks (e.g., TPC-H and TPCx-BB) are far from adequate for effectively evaluating the usability of database products and data service solutions in financial scenarios.

3.4.2 Diversifying data models

Currently, the data service platform needs to process data using different models (e.g., relational, graph, and streaming data) simultaneously. However, existing benchmarks are usually only for a single data model, which is a unimodal, isolated, and closed benchmark that cannot reflect the real application field of financial business. As shown in Fig. 3, for the loan risk calculation, the bank needs to check whether there are N customers (N>3) transferring funds to the same account number (the maximum number of transaction layers is three) during the two months after the loan is issued. Such queries are handled more efficiently using graph databases than relational databases. In financial scenarios, as the demand for analyzing non-relational data increases, database products and service solutions need to cover a wider range of data models, but existing benchmarks fall short.

3.4.3 Test data generation

Existing benchmarks typically generate test data by simulating the data structure, size, distribution, and relevance of real data to provide data support for database evaluation. However, this rule-based data generation method makes it difficult to discover potential features in real data, and it cannot generate test data that are highly similar to real data. Furthermore, the financial industry attaches significant importance to data security. The synthetic dataset must mask some data features related to national security and user privacy while simulating the relevant features of real data. However, it takes a lot of human and material resources to design data generation rules by artificially summarizing data characteristics, which inevitably leads to errors and omissions. As a result, it is difficult to generate test data that matches financial data characteristics while ensuring financial data security.

DOI 10.15302/J-SSCAE-2022.04.014

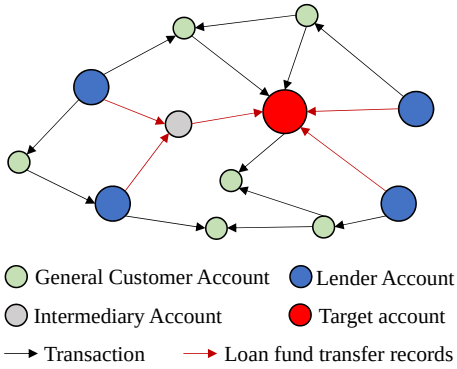

Fig. 3. An example of loan risk calculation.

3.4.4 Compatibility and portability evaluation

Existing benchmarks typically evaluate the performance of database products in a given application scenario. However, they cannot measure the compatibility and portability of the two database products. However, this metric is urgently required by the financial industry. As the localization of databases in the financial industry continues, financial practitioners need a metric to objectively and accurately measure the compatibility and portability of database products to help them choose the most cost-effective product from a wide range of database products and data-service solutions. This product can ensure business continuity and data integrity, while minimizing the negative impact of database replacement on their businesses.

4 Next-generation database benchmarks for financial scenarios

The financial industry has placed new demands on benchmarks. Given the characteristics of the financial data service field, financial practitioners need a set of benchmarks that can comprehensively cover diverse financial business scenarios. Using this benchmark to conduct a uniform and standardized evaluation of database products and data service solutions provided by database vendors, financial practitioners can choose a database solution that suits their actual business needs. As one of the first benchmarks to evaluate database performance in financial scenarios, the China Academy of Information and Communications Technologyhas released DataBench-T [37], a performance testing tool for OLTP databases, and DataBench-A, a performance testing tool for OLAP databases. Using these two benchmarks, they developed and designed a series of data schemas, workloads, and metrics to meet the specific requirements of the financial industry. However, there are still some problems with these benchmarks. For example, DataBench-T and DataBench-A cannot measure the HTAP and multi-model databases. Therefore, to comprehensively evaluate the performance indicators of various database products and data service solutions in financial scenarios, the next generation of financial database benchmarks must be adapted to financial business scenarios in the data model, workload, metrics, and technical architecture. Specifically, we need to further study the benchmarks in the following areas.

4.1 Ability to evaluate HTAP

The financial business area is mixed with many OLTP and OLAP workloads. On one hand, because of the specificity of the financial business, it is necessary to use resource isolation approaches to segregate different workloads. On the other hand, it is necessary to consider certain metrics, such as data freshness, to support both OLTP and OLAP workloads simultaneously. Therefore, when designing workloads in the next-generation financial database benchmark, it is necessary to consider HTAP characteristics of and select appropriate metrics to evaluate different types of workloads uniformly, while simulating real business scenarios.

4.2 Ability to evaluate multi-model databases

With the innovation of financial services, there is a large amount of non-relational data in financial business systems. Therefore, when designing the data schema for the new generation of financial database benchmarks, it is necessary to consider both relational and non-relational data and evaluate multi-model databases. Furthermore, we need to leave ample room for expansion in the benchmark to cater for future developments in the financial business.

4.3 Ability to evaluate distributed architecture adaptability

The transformation of distributed databases in the financial sector, especially the banking sector, is in full swing. Distributed systems are different from stand-alone and centralized systems; the system structure is more complex, and there are more influencing factors. Therefore, when designing the next-generation financial database benchmark, we must focus on how to evaluate the adaptability of database products and data service solutions to the distributed architecture.

4.4 Ability to generate synthetic financial data

To guarantee evaluation accuracy, the benchmarks need to generate high-fidelity test data by simulating real-life application scenarios. However, owing to strict compliance constraints, difficult privacy protection, and high risk of leakage, it is not feasible to use real financial data for database testing. To solve this problem, existing benchmarks use synthetic data to evaluate the databases. The traditional data-generation method is rule-based (e.g., TPC-H). This method makes it difficult to define complex rules, and the data generated are prone to the loss of key features and poor data simulation. Consequently, the data generated by the rule-based data generation method cannot simulate real business scenarios, resulting in inaccurate test results. Fortunately, recent research shows that by using deep learning methods to automatically learn data features from real data to generate synthetic data, real data features can be better simulated, and noise information can be selectively added to some data features to ensure that the final generated data will not cause a privacy leakage [38]. Therefore, the next-generation financial database benchmark must develop AI-based synthetic financial data generation methods to generate highly simulated financial data with a security guarantee.

4.5 Ability to provide more comprehensive reliability and security evaluating

The financial industry has higher reliability and security requirements for its database products. For example, data storage systems should be highly reliable and provide backup and recovery capabilities, and financial databases should have stronger disaster recovery capabilities based on business characteristics. However, existing benchmarks lack the ability to evaluate whether a database product meets the reliability and security requirements of a financial business. Therefore, the next-generation financial database benchmark must improve the ability to evaluate the reliability and security of a database.

4.6 Ability to evaluate the read-and-write performance of a database

To cope with the pressures of highly concurrent businesses, the Internet financial data services platform has higher performance requirements for the read-and-write capabilities of the database. The pressure comes from the huge volume of users (similar to e-commerce applications) and multiple data consistency verifications in the financial business, which are used to safeguard individuals' property. This is very different from the application scenarios modeled using existing benchmarks. Therefore, the next-generation financial database benchmark must enhance the ability to evaluate the read-and-write performance of the database.

4.7 Ability to evaluate the compatibility and portability of a database

In the process of replacing a database with a domestic one, the compatibility and portability of the new database with the existing one must be considered to ensure uninterrupted business during the replacement process and data integrity in the system after the database has been replaced. However, existing benchmarks cannot evaluate the compatibility and portability of the database. Therefore, the next-generation financial database benchmark must increase the ability to evaluate the compatibility and portability of the database.

5 Conclusion

With the rapid expansion of financial services and increased emphasis on data security in China, banks' financial industries have gradually begun to replace databases with domestic ones to cope with more complex business requirements and ensure that their business systems are autonomous and controllable. However, when faced with many database products, existing database benchmarks cannot meet the requirements of the financial industry for database products and data service solutions. Therefore, there is an urgent need for a benchmark that can be adapted to financial business scenarios to help financial practitioners make more wise choices and guide the healthy development of data-service vendors. In this study, we first summarized three challenges faced in the bank's database localization process and analyzed four major problems faced by existing database benchmarks in financial scenarios.

DOI 10.15302/J-SSCAE-2022.04.014

Thereafter, we discussed the importance of building a next-generation database benchmark for financial scenarios and outlined future research directions in the benchmark building.

In future studies, it is necessary to analyze the requirements of database products and data service solution evaluation in financial scenarios from workload, data schema, metrics, synthetic data generation, and evaluate platform architecture. To comprehensively evaluate financial database products and data service solutions in a financial scenario, it is necessary to build a next-generation database benchmark that can meet the evaluation requirements of the financial scenario. This benchmark can provide a reference for relevant national management departments and financial practitioners in database selection and help improve the level of financial technology in China.

References

[1] Lin Y F, Fu C H, Ren X M. How Financial Innovation Promotes High Quality Development: A Perspective of New Structural Economics [J]. Finance Forum, 2019, 24(11): 3-13. Chinese.

[2] The People's Bank of China. Total assets of financial institutions reached 381.95 trillion yuan at end-2021 [EB/OL]. (2022-03-15) [2022-06-22]. http://www.pbc.gov.cn/goutongjiaoliu/113456/113469/ 4507972/index.html. Chinese.

[3] Jazzyear. China Fintech Report Series [R/OL]. Online: Jazzyear, 2020. https://www.jazzyear.com. Chinese.

[4] The People's Bank of China. The People's Bank of China issued the Fintech Development Plan for 2022-2025 [EB/OL]. (2022-01-04) [2022-06-06]. http://www.pbc. gov. cn/goutongjiaoliu/113456/113469/4438627/index.html. Chinese.

[5] Hu L M. Application and Prospect of Distributed Database in Financial Industry [J]. FinTech Time, 2020(05): 25-33. Chinese.

[6] Poess M and Floyd C. New TPC Benchmarks for Decision Support and Web Commerce [J]. ACM Special Interest Group on Management of Data Record, 2000, 29(4): 64-71.

[7] Nambiar R O and Poess M. The Making of TPC-DS [C]. Seoul: Proceedings of the 32nd International Conference on Very Large Data Bases, 2006: 1049-1058.

[8] China Academic of Information and Communications Technology. Database Development Research Report (2021) [R]. Beijing: China Academic of Information and Communications Technology, 2021. Chinese.

[9] ITpub Technology Stack. Thirty Years of Turbulence: The Development and Changes of Bank Databases [EB/OL]. (2021-04-02) [2022-06-06]. https://z.itpub.net/article/det-ail/CE307F44933F633B8EB297FE3CF7379E. Chinese.

[10] The People's Bank of China. The People's Bank of China issued the Fintech Development Plan for 2019-2022 [EB/OL]. (2019-08-22) [2022-06-06]. http://www.pbc.gov. cn/goutongjiaoliu/113456/113469/3878634/index.html. Chinese.

[11] China Financial Standardization Technical Committee. Three Financial Industry Standards Including "Financial Application Specification of Distributed Database Technology Technical Architecture" were Officially Released [EB/OL]. (2020-12-25) [2022-06-06]. https://www.cfstc. org/jinbiaowei/2929436/2978097/index.html. Chinese.

[12] Wang F P. Pursue Excellence and Forge Ahead——China CITIC Bank GoldenDB Distributed Database Transformation Practice [J]. Financial Computerizing, 2020, (02): 76-78. Chinese.

[13] Li Z N. Distributed database financial applications advance steadily and orderly [J]. Financial Computerizing, 2020(12): 34-35. Chinese.

[14] Dai G W. Build A "New Ecology" and Explore the Development of Distributed Databases in The Financial Industry [J]. Financial Computer of China, 2021, (07): 85-86. Chinese.

[15] Leutenegger S T and Dias D M. A modeling study of the TPC-C benchmark [C]. Washington: Proceedings of the 1993 ACM International Conference on Management of Data, 1993: 22-31.

[16] CCF Technical Committee on Database, Tsinghua University, modb.pro. Research on Classification and Evaluation of Database System [EB/OL]. (2021-12-22) [2022-06-06]. https://www.modb.pro/doc/52857. Chinese.

[17] Jin Z Q, Qian W N, Zhou M Q, et al. Benchmarking Data Management Systems: From Traditional Database to Emergent Big Data [J]. Chinese Journal of Computers, 2015, 38(01): 18-34. Chinese.

[18] Yan Y B, Zhu W Q, Yang T, et al. Overview on Benchmark Test of Big Data System [J]. Network New Media Technology, 2018, 7(03): 6-13. Chinese.

[19] Bitton D, DeWitt D J, Turbyfill C. Benchmarking database systems-A systematic approach [R]. University of Wisconsin-Madison Department of Computer Sciences, 1983.

[20] Xin R, Mokhtar M. Databricks Sets Official Data Warehousing Performance Record [EB/OL]. (2021-11-02) [2022-06-06]. https://databricks.com/blog/2021/11/02/ databricks-sets-official-data-warehousing-performance-record.html.

[21] Dageville B, Cruanes T. Industry Benchmarks and Competing with Integrity [EB/OL]. (2021-11-12) [2022-06-06]. https://www.snowflake.com/blog/industry-bench-marks-and-competing-with-integrity/.

[22] Mokhtar M, Tavakoli-Shiraji A, Xin R, Zaharia M. Snowflake Claims Similar Price/Performance to Data-bricks, but Not So Fast! [EB/OL]. (2021-11-15) [2022-06-06]. https://databricks.com/blog/2021/11/15/snowflake-claims-similar-price-performance-to-databricks-but-not-so-fast.html.

[23] Cao P, Gowda B, Lakshmi S, et al. From BigBench to TPCx-BB: Standardization of a Big Data Benchmark [C]. New Delhi: 8th TPC Technology Conference, 2016: 24-44.
[24] Hao Y, Qin X, Chen Y, et al. TS-Benchmark: A Benchmark for Time Series Databases [C/OL]. Chania: 37th IEEE International Conference on Data Engineering, 2021: 588-599. https://doi.org/10.1109/ICDE51399.2021. 00057.
[25] Murphy R C, Wheeler K B, Barrett B W, et al. Introducing the graph 500 [J]. Cray Users Group (CUG), 2010, 19: 45-74.
[26] Dreseler M, Boissier M, Rabl T, et al. Quantifying TPC-H choke points and their optimizations [J]. Proceedings of the VLDB Endowment, 2020, 13(8): 1206-1220.
[27] O'Neil P E, O'Neil E J, Chen X, et al. The star schema benchmark and augmented fact table indexing [C]. Lyon: First TPC Technology Conference, 2009: 237-252.
[28] Ghazal A, Rabl T, Hu M, et al. Bigbench: Towards an industry standard benchmark for big data analytics [C]. New York: Proceedings of the 2013 ACM International Conference on Management of Data, 2013: 1197-1208.
[29] Eichmann P, Zgraggen E, Binnig C, et al. IDEBench: A benchmark for interactive data exploration [C/OL]. Portland: Proceedings of the 2020 ACM International Conference on Management of Data, 2020: 1555-1569. https://doi.org/10.1145/3318464.3380574.
[30] Funke F, Kemper A, Krompass S, et al. Metrics for Measuring the Performance of the Mixed Workload CH-benCHmark [C]. Seattle: Third TPC Technology Conference, 2011: 10-30.
[31] Cooper B F, Silberstein A, Tam E, et al. Benchmarking cloud serving systems with YCSB [C]. Indianapolis: Proceedings of the 1st ACM Symposium on Cloud Computing, 2010: 143-154.
[32] Patil S, Polte M, Ren K, et al. YCSB++: benchmarking and performance debugging advanced features in scalable table stores [C]. Cascais: ACM Symposium on Cloud Computing in conjunction with SOSP 2011, 2011: 9.
[33] Chintapalli S, Dagit D, Evans B, et al. Benchmarking streaming computation engines: Storm, flink and spark streaming [C]. Chicago: 2016 IEEE International Parallel and Distributed Processing Symposium Workshops, 2016: 1789-1792.
[34] Angles R, Antal J B, Averbuch A, et al. The LDBC social network benchmark [J/OL]. CoRR, 2020, abs/2001.02299. http://arxiv.org/abs/2001.02299.
[35] Zhang C, Lu J H, Xu P F, et al. UniBench: A Benchmark for Multi-model Database Management Systems [C]. Riode Janeiro: 10th TPC Technology Conference, 2018: 7-23.
[36] Tian J F, Jiang C Y. Database Performance Evaluation Tool Under Financial Scenario [J]. Information and Communications Technology and Policy, 2020, 46(4): 85-90. Chinese.
[37] Jiang C, Tian J, Ma P. Databench-T: A Transactional Database Benchmark for Financial Scenarios [C]. Shen-yang: 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications, 2021: 1418-1421.
[38] Liew S P, Takahashi T, Ueno M. PEARL: Data Synthesis via Private Embeddings and Adversarial Reconstruction Learning [C/OL]. Online: The Tenth International Conference on Learning Representations, 2022. https://openreview.net/pdf?id=M6M8BEmd6dq.