FROST & SULLIVAN 沙利文 头豹 LeadLeo

2021
China Distributed Database Market Report

2021中国分布式数据库市场报告
2021中国の分散データベース市場レポート

Tags: Cloud-native, multimodel, distributed, open source, application scenarios

(Summary Version)

LeadLeo Research Institute
Frost & Sullivan (China)

Instruction

Frost & Sullivan hereby releases the annual report "China Distributed Database Market Report 2021" as part of the China Database Series Report. The purpose of this report is to analyze the development status, product characteristics and technology trends of distributed database market in China, and identify the competition situation in the market of distributed database in China, and reflect the differentiated competitive advantages of the leading brands in this market segment.

Frost & Sullivan and LeadLeo Research Institute conducted downstream user experience surveys on core products in the distributed database field. Respondents are of different sizes and in different segments in each of its industry that includes internet, media, telecommunications, transportation, government and other fields.

The performance of distributed database products in the financial field is detailed in the "China Financial Distributed Database Market Report" series.

Trends in distributed database presented in this market report also reflect trends in the database industry as a whole. The report's final judgment on market ranking and leadership echelon are only applicable to the industry development cycle of this year.

Abstract

Overview of Distributed Database Industry

The new generation of distributed database meets the core requirements of enterprise users with advantageous features such as ease of use, scalability, fast update iteration, and relatively low cost investment.

The boom in software applications has created a multi-scenario, multi-ecology, multi-user market environment needed for database technology development. GitHub expects China to become the world's largest source of developers by 2030. 2021 is the most active year for investment and financing in China's database track, further catalyzing the rapid growth of China's database market.

Development of Distributed Database Technology

In present choice of database distributed technology route, the primary goal is to solve the problem of data capacity expansion. The mainstream solutions are database and table middleware, native distributed, etc. Different technical routes and products have their own advantages and disadvantages.

The booming development of cloud computing has prompted various IT applications to shift to the cloud, and the unique flexibility of on-demand services and the low cost of on-demand billing or billing by configuration of cloud services are deeply matched with database users. It is particularly important to build database services on the cloud and design a cloud-native database with the basic cloud first and then fully adapted to cloud characteristics.

Development of distributed database market

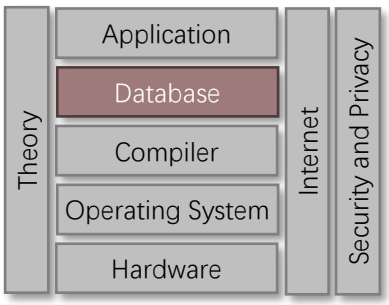
The development of distributed database technology should meet the needs of the times and the market, and return to the rigid needs of database users. The current distributed database needs to reach the level of centralized architecture products in various dimensions so as to play its performance and cost advantages in various scenarios and penetrate into various industries.

In terms of architecture selection, single database, single database sub-database sub-table + distributed middleware or native distributed database, all have their most advantageous application scenarios. Under the trend of distributed database, enterprises should choose distribution rationally.

Definition and classification of databases

- As the infrastructure of most information systems, databases exert hardware computing power downward and enable upper-layer applications upward. Various database products meet different business needs respectively. The speed, ease of use, stability, scalability, and cost of a database are all critical to an enterprise's basic business and growth resilience.

Computer Science Panorama



Source: Tsinghua university, Leadleo

❑ The importance of databases

As the infrastructure of most information systems, database exert hardware computing power and enable upper-layer applications. Various database products meet different business needs respectively. The speed, ease of use, stability, scalability, and cost of a database are all critical to an enterprise's basic business and growth resilience.

If databases never existed, programmers would have to deal with massive data and unreliable computer systems. But on the basis of database, programmers do not need to redesign complex system processes to ensure transactional data processing. Instead, they only need simple operations like C-R-U-D, which greatly reduce the complexity of data storage and processing.

Right after the birth of database, applications development exploded and became an important phrase in the history of computing.

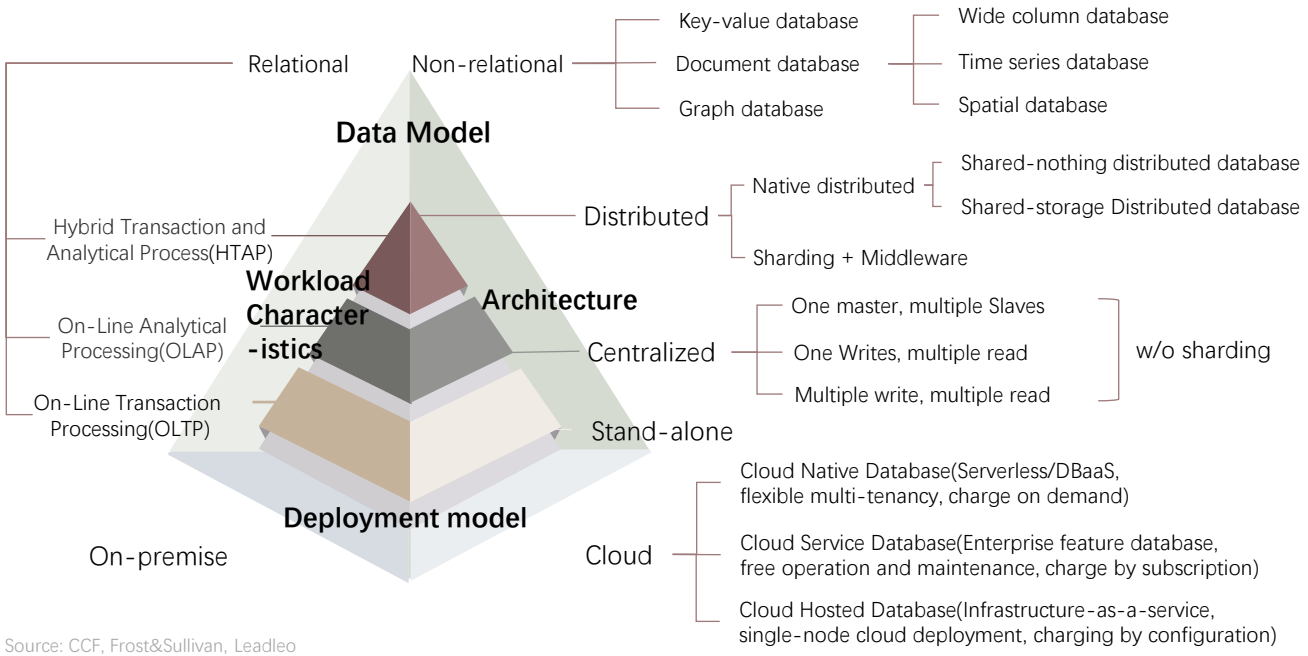
❑ Definition and classification of database

Database System: basic software that organizes, stores, and manages data according to a specific data structure.

Distributed Database: a logically unified database formed by connecting physically dispersed database units through network.

This report focuses on the cutting edge of the database industry from a distributed architecture perspective.

Classification of databases

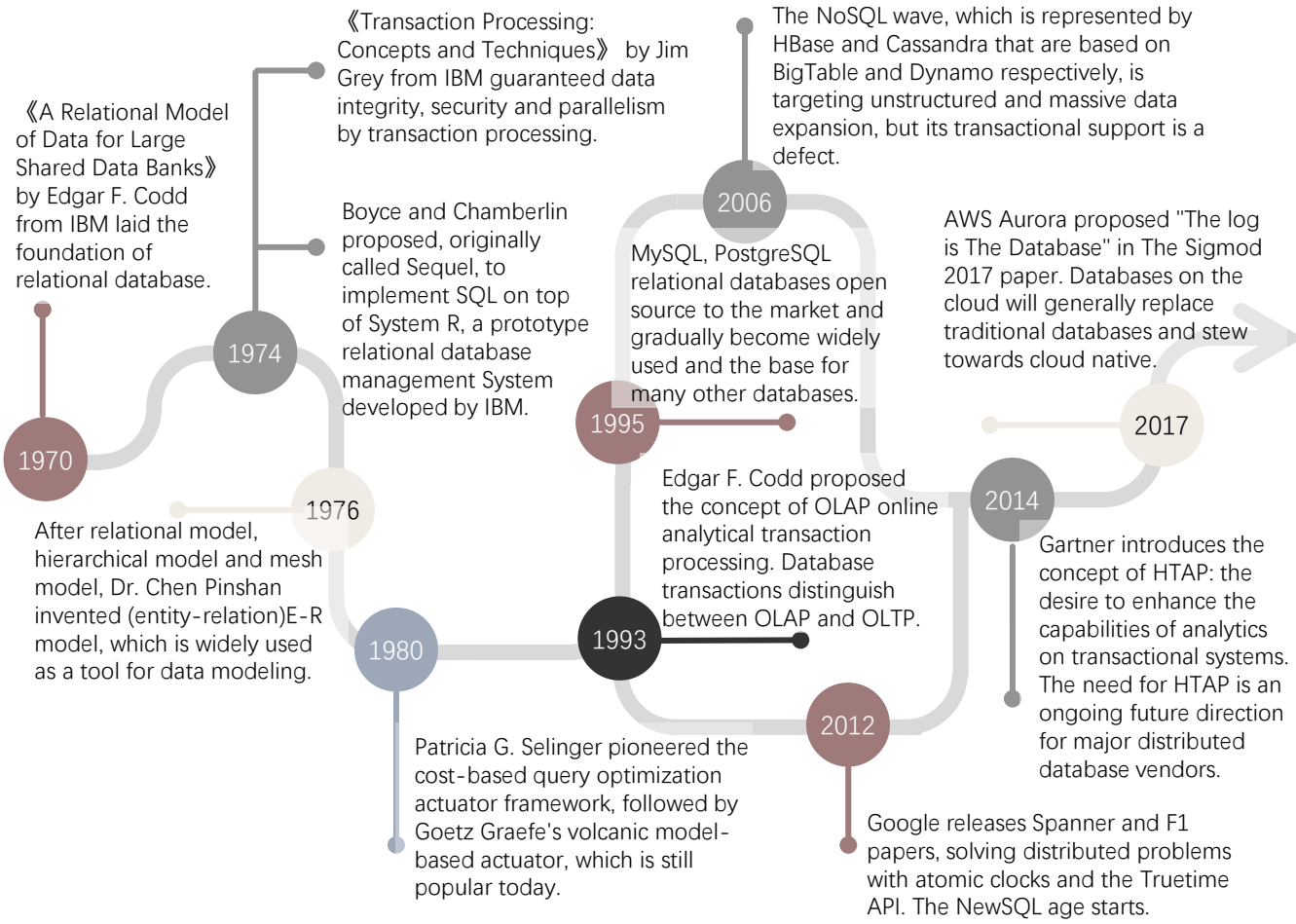


Source: CCF, Frost&Sullivan, Leadleo

Distributed database concept and technology development evolution

- **Since 1970, the database field has been led by the academic and research end. It entered the stage of rapid development of commercialization after 1995, and continued to drive the development of application ecology. After 2015, the database has developed to the direction of distributed architecture and cloud in the development of Internet infrastructure.**

Evolution of the classical architecture of distributed databases



Source: Frost&Sullivan, Leadleo

❑ The evolution of database

Database development has gone through half a century, experienced academic driven, commercialization landing, paper industry realization, enterprise application demand driven and other technical development stages.

From the beginning of the hierarchical model, mesh model, relational model, to object model, object-relational model, semi-structured, etc., the data model has been the core and theoretical foundation of the database. Solid theoretical support and better logical independence will still be the foundation of database in the future.

After commercialization, Oracle led the market with MySQL, Microsoft's SQL Server and other relational databases for many years. From SQL, NoSQL to NewSQL, and even HTAP, business capabilities are driven in iterations.

At present, cloud + distribution has become the only solution to the extreme needs of enterprises, and has created the current outbreak of the database industry. In the current and ongoing industry cycle, advanced products and technologies need to revolve around the market to become the most important competitive advantage.

Construction of distributed database industry support system

- The development of distributed database in China has obtained demographic dividend. Technological innovation needs an advanced academic research system, industry-research integration needs close industrial exchanges, and industry penetration needs a talent training system that keeps up with the needs of the times.

Database industry support system in China

Academic Organization of Database

Top conferences in the field of databases

VLDB、SIGMOD、ICDE

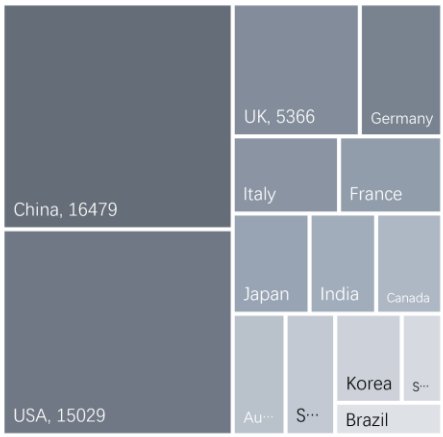
Colleges and Universities

THU, UCAS, HUST, ZJU, etc.

Enterprises

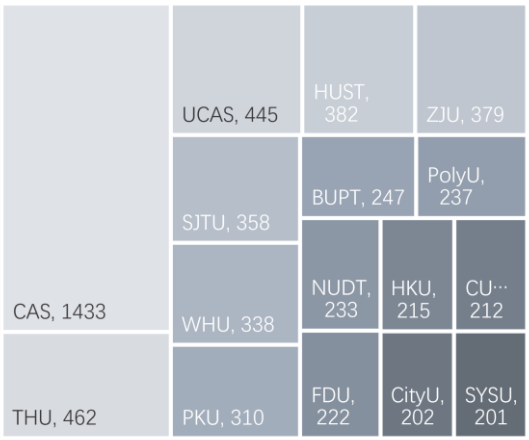
Alibaba, Huawei, Tencent, Baidu, Pingkai Stars, Dameng, DDA Jincang, etc.

Number of papers in distributed databases by country



Retrieve countries and regions of Distributed Database papers based on Web of Science Core Collection.
Source: Web of Science, Frost&Sullivan, Leadleo

Number of papers in distributed databases of academic institutions in China



Retrieve institutions of papers with Distributed Database topic based on Web of Science core collection, select China region.
Source: Web of Science, Frost&Sullivan, Leadleo

Database Industry Support Organization

Spontaneous user organization by database technology enthusiasts

ACDU for DBA, ACOUG for Oracle users, ACMUG for MySQL users, PostgreSQL Branch of China Open Source Software Promotion Alliance for PostgreSQL users

Official technical community for specific database products discussion

Ali Cloud developer Community, Huawei Cloud openGauss community, PingCAP AskTUG community, PostgreSQL Chinese community, Aikesheng open Source community, mobile cloud developer community

Research organizations with official backgrounds

China Computer Federation Database Technical Committee, China Communications Standards Association Big Data Technical Standard Promotion Committee (CCSA TC601)

Third Party Technology Community

ITPUB、MoDB、DBAplus

Database Personnel Training System

Textbook | Professional Courses | Teaching service | Competition | Test certification | Innovation center | Ecological talent market

Database vendor

PingCAP University, Oceanbase College, AUCP of Ali Yun University, OpengAuss community of Tencent Cloud Noah Plan, Giant Sequoia University, etc.

Professional training institutions

ENMO Edu, New Century college, Pangu cloud class, etc.

Colleges and Universities

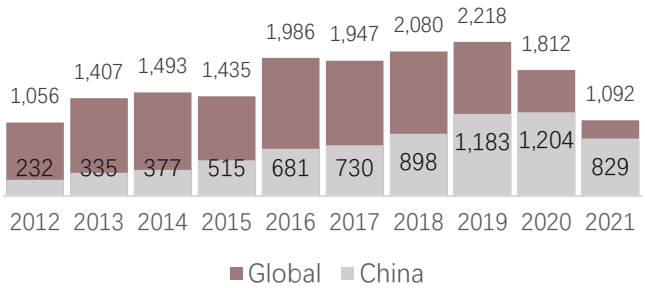
NPU, THU, WHU, RUC, etc.

Source: CAICT, CCF Technical Committee, Company official website, Frost&Sullivan, Leadleo

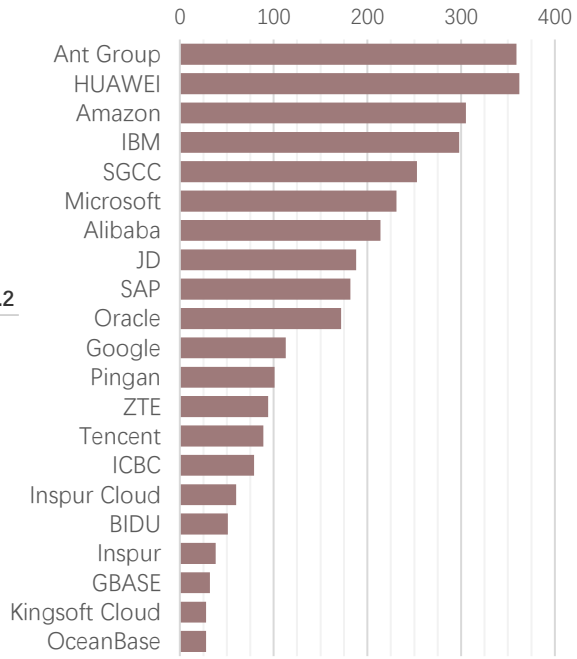
Patent analysis of distributed database vendors

- From the perspective of patent application data, the number of patent applications related to distributed database in China climbed from 22% in 2012 to 76% in 2021. China has become the technological innovation center of distributed database in the world.

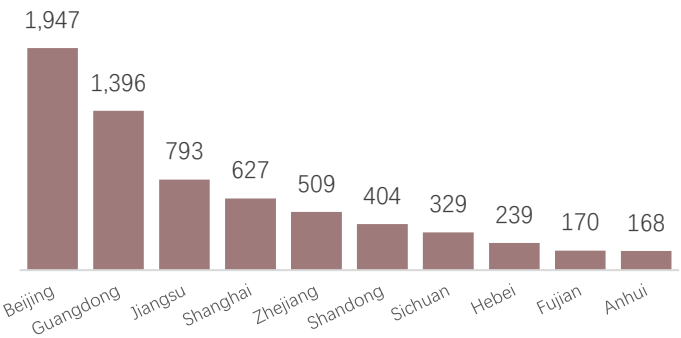
Global and Chinese distributed database patent applications, 2012-2021



Part of enterprises with distributed database patent application, 2012-2022.2



Distributed database patent application by Chinese provinces, 2012-2022.2



Sunburst Chart of distributed database patent topics keywords

From the perspective of patent application data, this report observes the global innovation status of Chinese distributed database market and the innovation vitality of Chinese database enterprises.

Global and Chinese patent application analysis

The number of patent applications related to distributed database in China climbed from 22% in 2012 to 76% in 2021. China has become the technological innovation center of distributed database in the world.

Beijing, Guangdong and Jiangsu are the provinces with a large number of advanced distributed database manufacturers, and the cumulative number of distributed database patent applications leads the whole country.

Keywords of distributed database patent topics

The innovation of distributed database is built on the accumulation of research and application of database technology. The key sections of distributed database including distributed computing, load balancing, control system, distributed storage and other sub-topics have been highlighted.

Retrieved the title or summary for Distributed Database or Distributed Database, 2022.2.
Source: Patsnap, Frost&Sullivan, Leadleo

2021 major events about database in china

- Reviewing the big events in the database industry in 2021, the key words in policy are localization, data security and open source, and the key words in enterprise products are distributed, HTAP, cloud native and open source. Driven by policies and markets, China's database industry is flourishing.

2021 Domestic database and major policy events

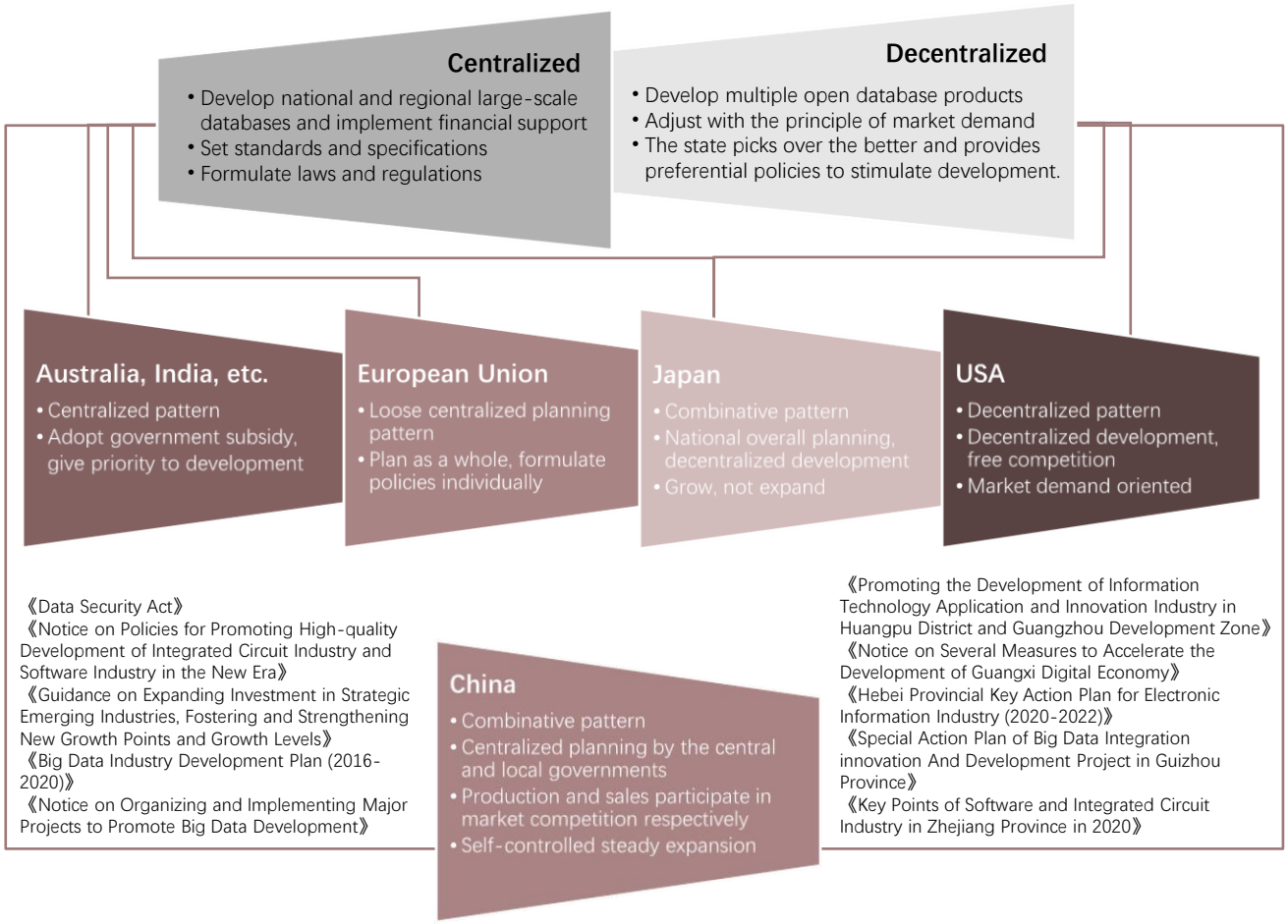
		Review of product, technology and policy development
Dec	20th	OceanBase passed CESI's first open source project maturity assessment.
Oct	20th	Ali Cloud announced the official open source cloud native distributed database PolarDB-X source code in 2021 cloud conference.
Sep	17th	TiDB community was one of the first to pass the Trusted Open Source Community Assessment and was awarded the OSCAR Peak open source project and open source community.
	1st	The Data Security Law was officially implemented. China Academy of Information and Communications Technology (CAICT) and more than 30 organizations officially launched the Data Security Initiative (DSI), which is committed to building a healthy and standardized Data Security ecosystem, helping enterprises understand regulatory requirements and comprehensively improving their Data Security capabilities.
Jul	9th	SGCC officially released the power industry graph database product "GridGraph" with independent intellectual property rights.
	8th	Ali Cloud RDS database brand upgrade, launched cloud native enterprise autonomous database.
Jun	20th	SIGMOD, recognized as the world's top three database conference, is held in Xi 'an, which is the return of SIGMOD conference after 14 years. (SIGMOD was first held in Beijing, China in 2007)
	10th	The Data Security Law of the People's Republic of China was adopted at the 29th session of the Standing Committee of the 13th National People's Congress and came into force on September 1, 2021. It has been clarified and strengthened in the form of legal text, which provides a legal basis for data as a new factor of production to promote innovation and economic development, and will escort the safe development of digital economy in the next stage.
	1st	Ant Group's self-developed database OceanBase announced open source, opening nearly 3 million lines of source code, using Magnolia protocol, code hosting in Gitee, mirroring in GitHub, and the establishment of OceanBase open source community
May	18th	Tencent Cloud released tdSQL-A, the first self-developed distributed analytical database, to meet the demand for real-time analysis of massive data. This is Tencent cloud database's first new release after the brand upgrade.
	11th	Inspur launches open source ZNBase 2021 development plan, which is a NewSQL distributed database.
Apr	25th	PingCAP released TiDB 5.0 for enterprise-level core scenarios, making significant improvements in performance, stability, and ease of use. By introducing MPP architecture, it becomes a distributed database with full HTAP capability.
Mar	19th	The Central Government Procurement Website issued the Transaction Announcement of 2021 Database Software Agreement Supply and Purchase Project of Central State Organs, and 21 database manufacturers were shortlisted, among which, except Oracle and Microsoft SQL Server, all domestic databases accounted for 90%.
	12th	Xinhua News Agency was authorized to broadcast The Outline Of The 14th Five-year Plan Of The People's Republic Of China For National Economic And Social Development And The Vision Of 2035. What is noteworthy is that "open source" has been explicitly included in the five-year plan for national economic and social development for the first time.
Feb	24th	Huawei cloud has released the GaussDB cloud database (for openGauss) for commercial use. GaussDB (for openGauss) is an enterprise-level distributed database developed by Huawei based on openGauss research ecology.

Source: InfoQ, Leadleo

Policy Analysis

- China's database industry policy-making adopts a combination of decentralized and centralized model, and specific policies are the means and measures to achieve macro policy objectives, including property rights protection policy, demand guidance policy, security and confidentiality policy, encouragement of development policy, management policy, international cooperation and exchange policy, talent policy, etc.

National or regional database industry construction and development guidelines



Sources: HRBPS, Peking University, NPC Standing Committee, NDRC, the State Council, MIIT, Frost&Sullivan, Leadleo

❑ China database industry policy analysis

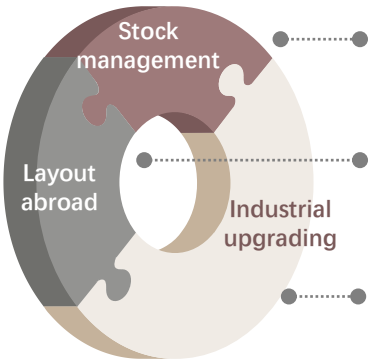
With the accelerated construction of the digital economy, the government attaches more importance to the development of the data industry, and the multi-level policy system of the data industry is gradually improved. China's database industry policy formulation adopts a combination of decentralized and centralized model. Specific policies are means and measures to achieve macro policy objectives, including property rights protection, demand guidance, security and confidentiality, talent policies, etc., and the development of unified industry standards and norms.

In the centralized planning of digital government, digital city and digital transformation of state-owned enterprises, the implementation of financial support is preferred. Develop diversified data products, encourage the separation of production and sales to participate in market competition, adjust with the principle of market demand, and gradually realize independent control and expansion of output.

The wave of digital transformation and business development needs

- At present, Chinese enterprises are facing the needs and challenges in three aspects: stock operation, overseas layout and industrial upgrading. The new generation of distributed database meets the core requirements of enterprise users with the advantages of ease of use, scalability, rapid update and iteration, and relatively low cost investment.

Current business development needs and challenges faced by enterprises



	• Enterprises' problems	• Practical problems
1	• Flow management and user growth enter a bottleneck • Convergence of competitiveness and weak innovation	• Users, channels, experts, products, IP, partners ..... • Not locking key resources of digitalization
2	• Oversea development of domestic enterprises facing obstacles in flow, technology, etc.	• Lack of standardized global business layout • Cross communication barriers • Difficult to meet localization business security and compliance
3	• Lack of experience in industry digital transformation • Decision makers are not clear about return on investment	• No clear and thorough strategic planning • Lack of digital talents • Difficult to build a digital operation system

Sources: Yuanchuan Research Institute, Huaguan Shuzhi, Leadleo

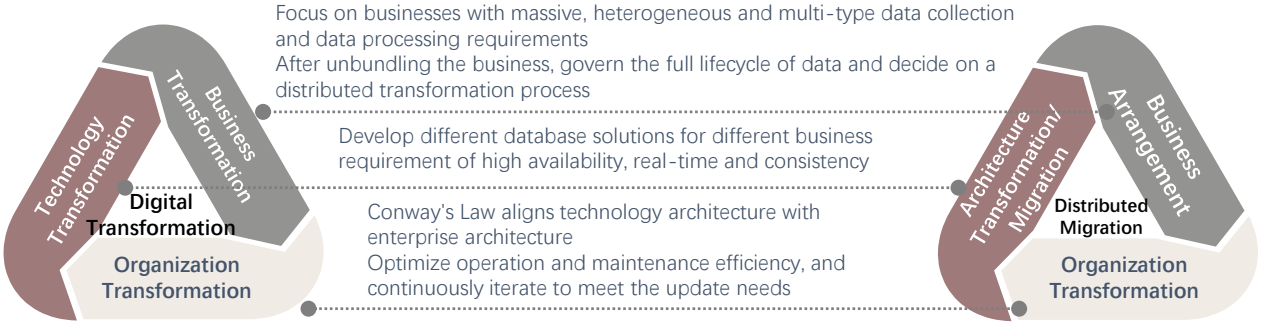
❑ Digital transformation of Chinese enterprises

At present, Chinese enterprises are facing the needs and challenges in three aspects: stock operation, overseas layout and industrial upgrading. Chinese enterprises are rapidly moving into a new era of digital transformation, and the improvement of digital capability is the key.

In the process of enterprise digital transformation, more and more businesses turn to digital, online and intelligent, followed by the exponential growth of storage and computing demand. Traditional commercial database has been unable to meet the needs of enterprises to reduce costs and increase efficiency, and it is also weak when facing rapid change and continuous growth of the business.

The new generation of distributed database addresses the core needs of enterprise users with an architecture that satisfies the concepts of decoupling storage and calculation, functional reuse, and configurability, integrating mature traditional database technology with innovative technology, while featuring high availability, high scalability, rapid update iteration, and relatively low cost investment.

Digital Transformation and Distributed Database Migration



Sources: Tencent Cloud, Leadleo

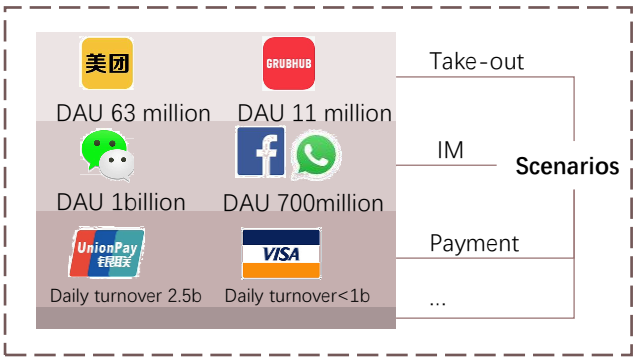
China Opportunities for Distributed Databases

- The prosperity of software application has created the market environment of multi-scene, multi-ecology and multi-user for the development of database technology. GitHub expects China to become the world's largest source of developers by 2030. 2021 is the most active year for investment and financing of Database track in China, which further catalyzes the rapid growth of Database market in China.

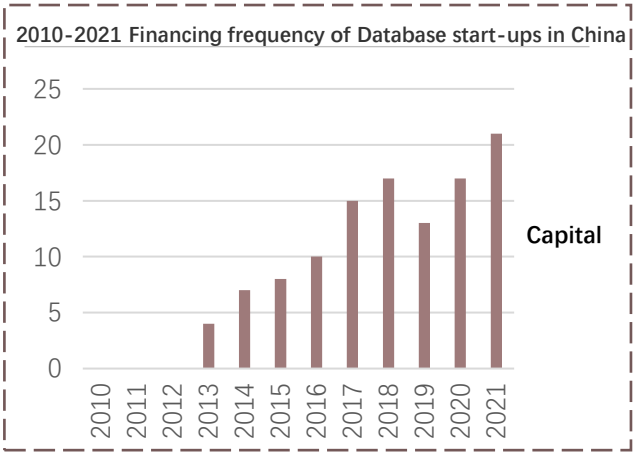
❑ Distributed database development environment in China

China's population base, post-urbanization population density and highly developed economic behavior constitute the attributes of a massive, highly concurrent data environment, and the development of distributed databases in China has achieved a traffic dividend.

Scenario, talent and capital environment of Chinese database market



GitHub predicts contributor country distribution in 2030



▪ Scenario benefits

The flow environment of Internet and mobile Internet has contributed to the rapid development of Information technology in China in the past ten years. The prosperity of software application has created a multi-scene, multi-ecological and multi-user market environment required by the development of database technology, which has given database manufacturers sufficient market environment of R&D-practice-trial-error.

Distributed database track flourishes in the environment of massive and highly concurrent data. There are not only traditional centralized database enterprises, but also cloud enterprises, start-ups and cross-boundary ICT enterprises.

▪ Talent benefits

According to Github's 2021 report, the US is the world's largest developer source at 22.7%, but that's down from 30.4% in 2015. With 7.55 million developers, Or 9.76%, China ranks second globally and is catching up fast. GitHub expects the situation to reverse in 2030, with China becoming the world's largest source of developers.

▪ Capital heat

Snowflake went public on the New York Stock Exchange in September 2020, leading a digital infrastructure investment boom.

According to incomplete statistics, as many as 20 companies received a new round of financing in 2021, and completed more than 14 rounds of 10 million or even hundreds of millions of financing. 2021 is the most active year for investment and financing of Database track in China. Sequoia, Hillhouse, Tencent, Matrix, Yunqi, Future Capital and other investors are deeply concerned and invested in database track. The continuous capital injection to database enterprises will further catalyzes the rapid growth of China's database market.

Sources: Future Capital, GitHub, CAICT, 36Kr, Frost&Sullivan, Leadleo

China Database Product Atlas

- Chinese database vendors are divided into traditional database vendors, emerging database vendors, cloud vendors, and ICT crossover vendors, which provide different centralized database and distributed database products.

China database vendors and representative database products

Traditional

达梦数据库	DM8
人大金仓 Kingbase	KSOne KingbaseES KingbaseAnalyticsDB
GBASE®	Gbase 8a Gbase 8s Gbase 8c Gbase XDM
GreatDB 万里数据库	GreatDB Cluster
中国航天 神舟通用	OSCAR

Cloud

腾讯云	TDSQL TencentDB TcaplusDB
阿里云	PolarDB Lindorm AnalyticDB
HUAWEI	GaussDB OpenGauss TaurusDB
金山云	Dragonbase KingDB KRDS
百度智能云	GaiaDB Palo
京东云	StarDB
天翼云	TeleDB

Emerging

PingCAP	TiDB
SequoiaDB 巨杉数据库	SequoiaDB
TRANSWARP 星环科技	ArgoDB KunDB
热璞科技 HOTPU	HotDB
极数云舟	ArkDB
OCEANBASE	Oceanbase
易鲸捷	EsgynDB
BeagleData 天云数据	Hubble
云和恩墨 ENMOTECH	MogDB
ACTION 爱可生	RDS Shard

ICT crossover

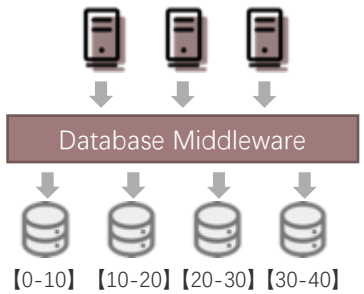
ZTE中兴	GoldenDB
inspur 浪潮	ZNBase K-DB
H3C 新华三集团	SeaSQL

来源：各公司官网头豹研究院

Distributed database technical route classification

- **Data capacity expansion problem is the primary goal of distributed datababse. The mainstream solutions are sharding, shared-storage and shared-nothing. Each route and product has its advantages and disadvantages**

Distributed database architectures



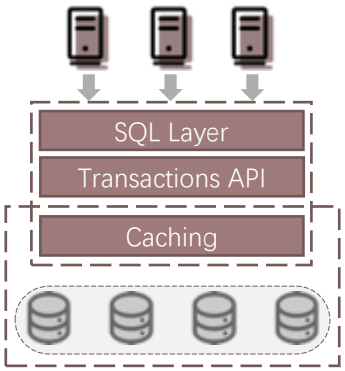
❑ Sharding-Sphere + Middleware

Solution： The lower single-node database provides storage and execution capabilities, and an intermediate layer is encapsulated on top of multiple single-node database to supplement the distributed in different database nodes with uniform data slicing rules and providing SQL parsing, request forwarding and result merging capabilities.

Advantages： Existing open-source databases are of mature and stable features, with high performance、low cost、stability、low user threshold. (low upper limit of capacity but high lower limit)

Disadvantages： Sharding solution is of high cost and its underlying architecture is not natively distributed. Calability bottlenecks exist with limited middleware communication and monolithic database functionality

Products： GoldenDB、TDSQL MySQL版、GreatDB、HotDB、MogDB、GaiaDB-X、openGauss



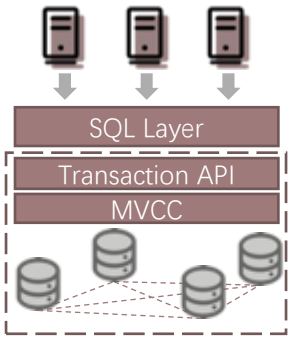
❑ Shared-storage Distributed Database

Solution： Independent compute-nodes share a storage cluster (Shared-storage architecture) . The underlying data storage is of distributed high-performance storage that can be dynamically scaled up and down. With a storage and compute separation architecture, the distributed database can be optimized according to network and storage layers to ensure high availability and high performance.

Advantage： Excellent transaction performance, fastest read and write response, and maximum write capacity limit

Disadvantage： Low architectural adaptability, reliance on shared storage systems, and low portability

Products： AWS Aurora、PolarDB、TDSQL-C、SequoiaDB-MySQL、GaussDB for MySQL、ArkDB



❑ Shared-nothing Distributed Database

Solution： Each node has independent computation and storage functions and does not share data between nodes (Shared-nothing architecture). Storage and compute separation architecture is adopted to achieve smooth scalability. Each node of the distributed cluster is an independent node, and the availability of multiple copies is guaranteed by consensus algorithms such as multi-paxos or multi-raft.

Advantage： Highly decoupled architecture, high compatibility, high portability and deployability, strong consistency and high availability

Disadvantage： With high hardware requirements, ow multi-write performance due to distributed transaction locking mechanism

Products： TiDB、Oceanbase、Google Spanner、Cockroach、Hubble

Source: Frost&Sullivan, Leadleo

Database security and encryption technology

- **Data encryption is an effective means to prevent the data in the database from being lost in storage and transmission, which includes table encryption, transmission encryption, transparent encryption, full encryption, etc.**

Data Security in database

Protection Phase	Threats	Solutions
In transfer	Phishing, Phishing attacks, Replay	HTTPS、SSL、TSL
Under operation	Permission elevation, data tampering, denial	Tamper-proof database, multi-party secure
Outside system	Privacy breach, elevation of privileges	Dynamic data desensitization, security and privacy protection
In calculation	Stack overflow information leakage	Full encryption database
at rest	Drag and drop, information extraction	Transparent encryption, data storage encryption, backup encryption

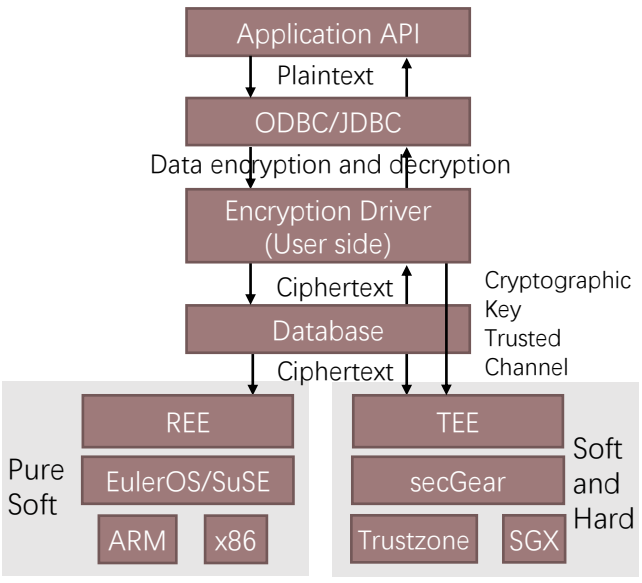
Source: Tsinghua University , CCF, Frost&Sullivan, Leadleo

❑ Data Security and Data Encryption

The security of the database is a top priority. On the basis of infrastructure security, we ensure data link security and data storage security. Database trusted computing follows the principle of "reliable, controllable and visible" and covers the complete data access process of "prior authentication - protection - audit afterwards.

Database security refers to the protection of the database to prevent data leakage, change or damage caused by illegal use. Data encryption is an effective means to prevent data loss in storage and transmission in the database, including table encryption, transmission encryption, transparent encryption, full encryption, etc.

Full encryption



Source: Tsinghua University 、 CCF， Frost&Sullivan, Leadleo

❑ Full encryption data processing

Cryptographic data processing uses technologies such as full homomorphic encryption to store encrypted data in order to achieve the highest possible capacity for processing encrypted data while ensuring data security. We ensure the security of data throughout its life cycle from transmission, computation to storage.

- Pure Soft: emphasize cryptographic index, cryptographic algorithm in cryptographic kernel. Keys do not leave the user environment; full life-cycle cryptomorphic processing; different algorithms for different models; support for common SQL queries.
- Soft and Hard: emphasize the granularity of plaintext and ciphertext isolation and plaintext and ciphertext exchange algorithms in a TEE-executable environment. Rely on trusted security hardware, hard and soft co-processing, and support for full-featured SQL queries.

Vendors: Huawei Cloud, Tencent Cloud, Alibaba Cloud, Gbase, Esgyn, Transwarp, Beagledata, PingCAP, etc.

Heterogeneous multi-model database

- The development of multimodel cannot be separated from the maturation of the single-mode database technology, which sinks the single-mode capability to the vertical engine to become the built-in capability of multimodality, and focuses on the tilt in the processing efficiency of different models

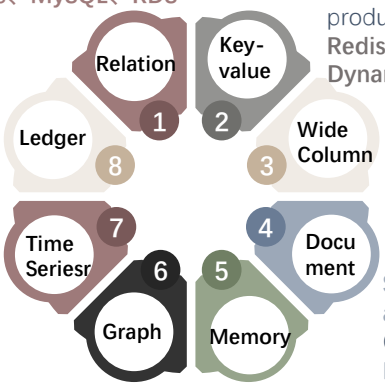
Characteristics of different data models, application scenarios and typical databases

Referential integrity, ACID transactions and write mode
ERP、CRM、Financial System
Oracle、MySQL、RDS

Complete, immutable and verifiable history of application changes
Records, supply chain, medical, registration, financial systems
Azure SQL、Amazon QLDB

Collect, store and process data in chronological order
IoT applications, practice tracking
InfluxDB、Kdb+、TiDB、Lindorm、GuassDB、TDSQL、Hubble、Gbase、GaussDB、EsgynDB、Oceanbase、ChronusDB

Create and find relationships between data quickly and easily
Fraud detection, social networks, recommendation engines
Neo4J、AllegroGraph、Amazon Neptune、TDSQL、Gbase、Oceanbase、Hubble



High throughput, low latency read/write, unlimited scalability
Real-time bidding, shopping cart, social, product catalog, customer preferences
Redis、Apache Ignite、Riak、RocksDB、DynamoDB、TiKV、Hubble、Gbase、KunDB

Store large amounts of data with nearly unlimited scalability
Industrial equipment maintenance, fleet management and route optimization
Cassandra、Hbase、Scylla、Amazon Keyspaces、Lindorm、TDSQL

Store documents and quickly access and query any property
Content Management, Personalization, Mobile
MongoDB、CouchDB、Amazon TimeStream、TiDB、Lindorm、Gbase、EsgynDB、GaussDB

One-click query with subtle delays
Caching, session storage, leaderboards, geospatial services, and real-time analytics
SQLite、H2、Oracle TimesTen、Amazon ElastiCache
ArgoDB、Tair、GaussDB、Hekaton

Source: Amazon Web Services, Frost&Sullivan, Leadleo

❑ Data Model

Many non-relational data structures have emerged in the NoSQL wave, and the database technologies corresponding to different data models have different theoretical cores and architectural ideas to meet different data scenarios and provide different ways of data management and processing.

❑ Specialized model database

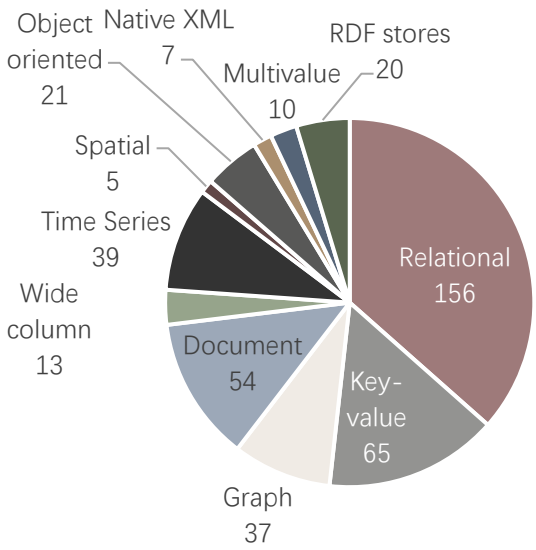
The representative of the specialized model database route is Amazon Web Services, who emphasizes the ultimate scalability and stability brought by the specialized model database, and takes "Purposed bult, Not all in database" as the architectural concept in the database selection practice to build the best scenario for the database users.

❑ Multi-models databases

Multi-model database is to support multiple data models by extending SQL on the basis of relational model database to achieve multiple models in one database. Thus reducing the complexity of management, operation and maintenance, and development of different data models and making it easy to use.

But the multi-model solution also has its disadvantages. In the same data type scenario, multi-model database compared to specialized model database, its storage costs and query performance are deficient. Hence, database selection needs to be decided based on the user's scenario.

Number of databases for different models



Source: DB-Engine, Leadleo
The number of databases in each category is shown, some databases belong to more than one category.

❑ Development status of multimodel database

Oracle, MySQL, SQL Server, PostgreSQL are all relational based to support multimode, and MongoDB, Redis, etc. are also compatible to other types.

Multi-model has become the mainstream in database deveolopment, but the development of multi-model cannot be separated from the maturation of specialized database technology, which contributes specialized capability to the vertical engine as built-in capability of multi-model database. But different databases are inclined to focus on the processing efficiency of different models.

According to the conclusion of CCF, multi-model database should be a new database system that natively supports various data models, has a unified access interface, can automatically data transformation of each model, model evolution and avoid data redundancy.

❑ Development of multi-model database

From the user's perspective, it supports multiple models in one database at the same time to handle more heterogeneous data that do not require high performance with a simpler database architecture, which greatly improves the ease of use, operation and maintenance efficiency, and reduce storage cost. The unified SQL access interface for different data types greatly optimizes the database experience.

With the diversification of application data requirements and the maturation of specialized database technology, users often need to face the analysis of heterogeneous data. Each application needs to develop a data middle layer to interface with multiple databases to handle problems such as model conversion, data distribution, data synchronization, query merging, etc.

When the large data volume is in relational type and the analysis frequency of other data types is not high, a heterogeneous multi-model database system that can provide unified storage, unified access and ensure correct data for the upper layer of business logic becomes a common demand. In addition, HTAP is also an extension concept of this demand.

❑ Unified multi-data type access interface

A multi-model database contains the storage of relational, key-value, document and other data models, and it needs to adopt a unified SQL access interface to provide unified storage and unified access for the upper-level business logic, which will greatly optimize the database experience.

Products: TiDB、Lindorm、GaussDB、TDSQL、Oceanbase、Hubble、Gbase、ArgoDB

❑ Automated management of data transformation of each model

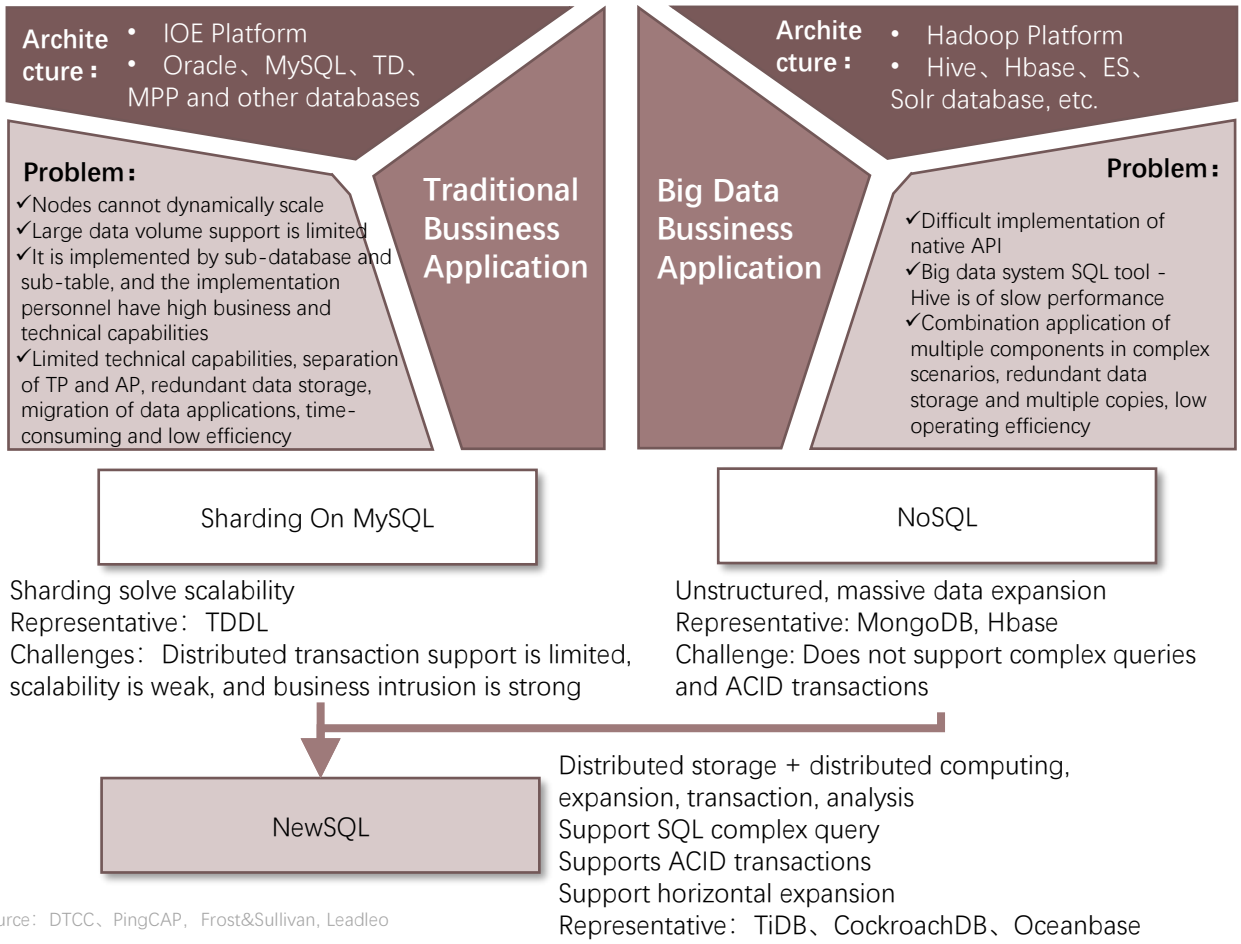
With the diversification of data sources, databases are needed to automate the integration, transformation and management of different types of data to facilitate the utilization of data value analysis.

Products: Lindorm、TDSQL、Oceanbase、ArgoDB

HTAP

- **With the increasing data sources and complexity of business systems, the need for mixed loads is becoming more and more common, and database technology is being oriented towards multi-source heterogeneity, high real-time concurrency, and multiple SQL standard interfaces**

Load demand and architecture evolution



Source: DTCC、PingCAP, Frost&Sullivan, Leadleo

❑ **Business mixed load demand becomes the norm**

Whether it is traditional business application demands relying on the IOE architecture to expand capacity through sharding or big data business application requirements relying on the Hadoop platform architecture, both have accumulated many problems and are difficult to solve. Due to the limited capabilities in operation, maintenance and use, neither can meet the demands of the new age.

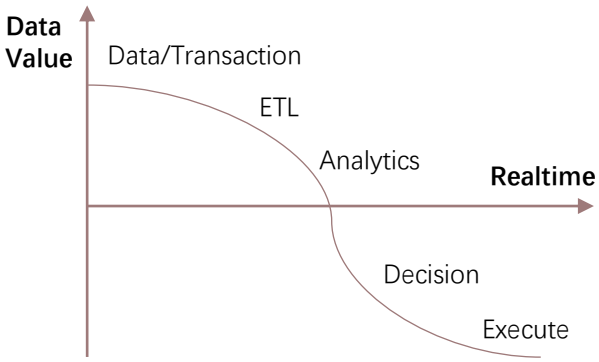
With the increasing data sources and complexity of business systems, the need for mixed loads is becoming more and more common, and database technology is being oriented towards **multi-source heterogeneity, high real-time concurrency, and multiple SQL standard interfaces.** The demand for mixed workloads is becoming more and more common, and users want to use data in different ways (such as offline and online) through a unified access interface (such as SQL).

In 2012, Google Spanner and Google F1 creatively proposed the NewSQL form that combines the ACID guarantee of transactional databases with the scalability and high performance of NoSQL, which inspired many database companies to develop distributed architectures with mixed load support capabilities.

❑ Application Scenarios of HTAP

HTAP ensures a certain real-time performance and can also fully improve the response speed, throughput, concurrent access, transaction size, data access and index scale, bringing innovation and improvement of business and architecture for the following two scenarios:

The value of data decreases with realtime



Source: DTCC, Leadleo

1. Data intensive business

The analysis capability is embedded into the traditional OLTP business system. Data intensive businesses such as IoTs, medical treatment, risk control and personalized recommendation marketing can complete real-time analysis on the transaction side without affecting the performance and data consistency of the transaction.

2. Real time data as the core service

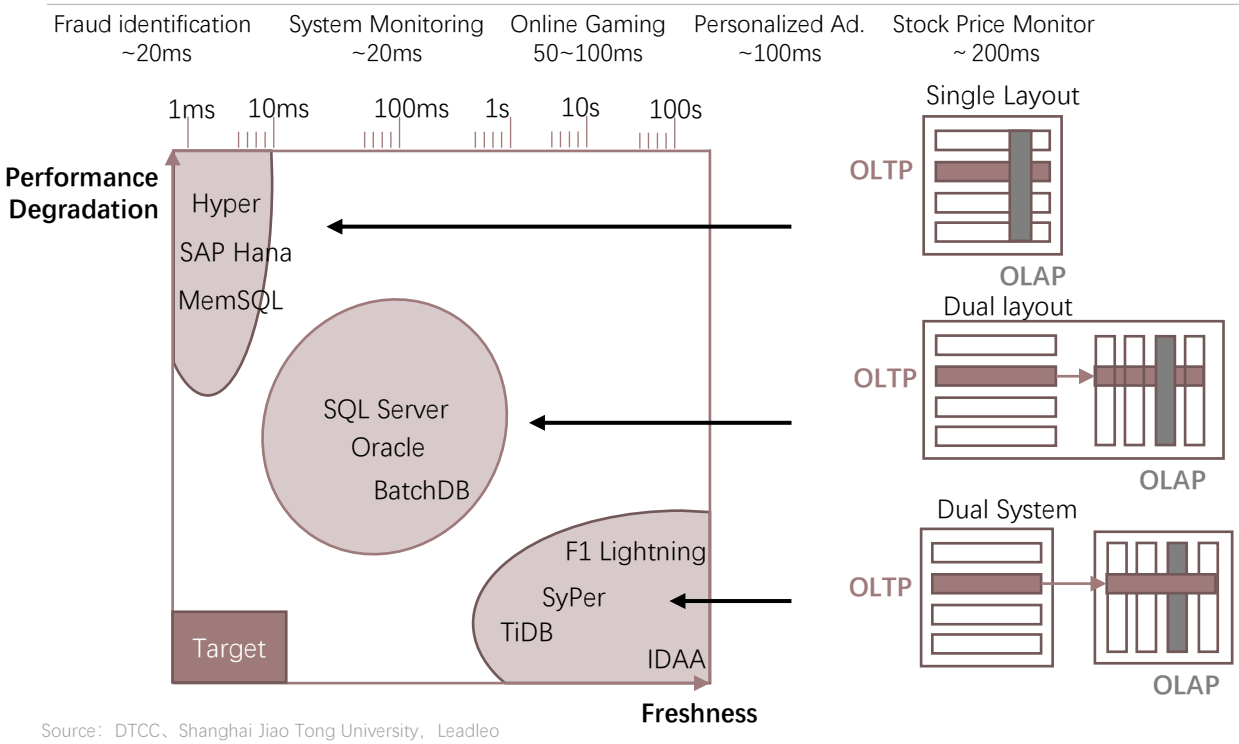
In the existing data platform, the data center platform with "use" as the core and "management" as the basis will become the key innovation and upgrading of enterprise digital planning and implementation. Let all enterprise users freely choose and apply data assets and realize data dividends in real time.

Applicable scenarios and capabilities of OLTP, OLAP and HTAP

	Applicable scenario		Database Capability			
	Scenario	Workload	Transaction consistency	Application adaptability	Data Size	Analytical ability
OLTP	online transaction lightweight analysis	high concurrent Small data volume	Single database consistency Multi-databases consistency of application layer	High	Medium	Low
	online transaction Simple transaction High concurrency	single-Query/Write Limited association and analysis	Final consistency of application layer	Low Intrusion exist	Large	Low
	online transaction batch processing real-time analysis mixed workload	single-Query/Write Limited association and analysis	Srong consistency	High Transparent to application	Large	Low
OLAP	batch processing complex analysis Non-realtime query	complex analysis/query	Low consistency	Low Intrusion exist	Large	High
HTAP	online transaction batch processing real-time analysis mixed workload lightweight analysis	high concurrent small data volume transactional W/R complex analysis	Srong consistency	High Transparent to application	Large	High

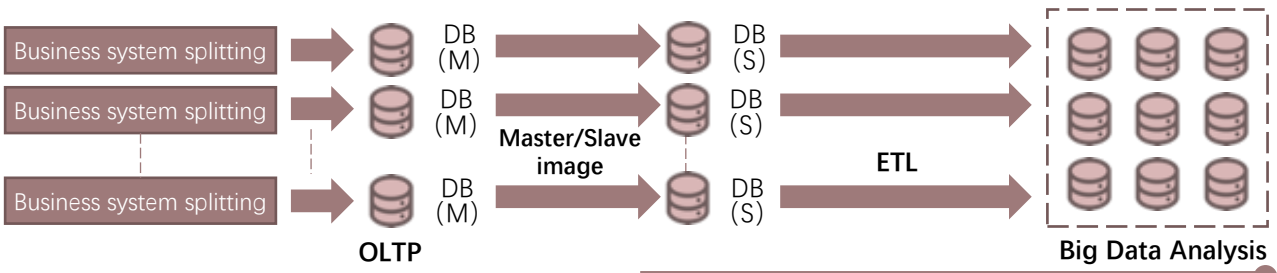
Source: CCIA, Frost&Sullivan, Leadleo

Performance and real-time performance of different HTAP schemes



Source: DTCC、Shanghai Jiao Tong University, Leadleo

Implementation of HTAP

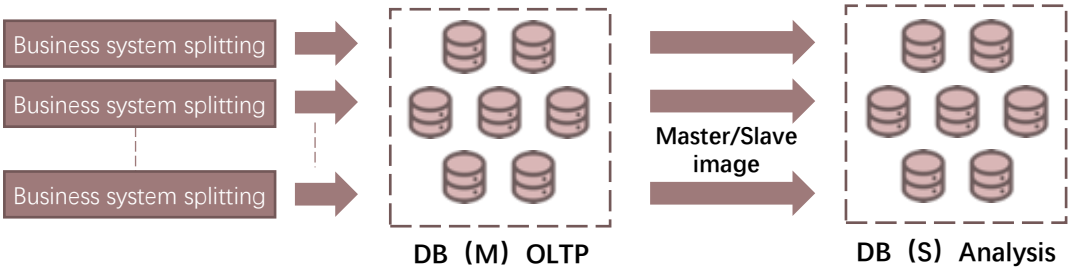


Splitting + Sharding

1.Complex scheme, need to redesign and plan business system
2.Multiple databases, transaction ACID lacking
3.Lack of global index / constraint and unique id need to be implemented
4.The database operation is carried out on each sub database, high cost of data management
5.Single segmentation and access dimension
6.Heterogeneous data warehouse is required for data analysis / decision-making, and it is difficult to synchronize data in real time

1.Smiple scheme, no need to redesign the business system
2.Single database, transaction ACID guaranteed
3.With global index / constraint, the database supports unique ID, etc
4.Database operations are performed on a single database without additional management costs
5.Data can be accessed from any dimension
6.The data analysis / decision support database and the transaction database are isomorphic and synchronized in real time

Native distributed HTAP



Source: Oceanbase, Leadleo

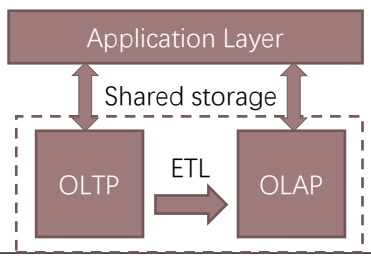
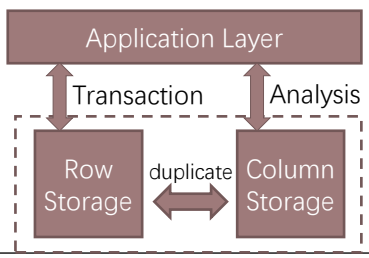
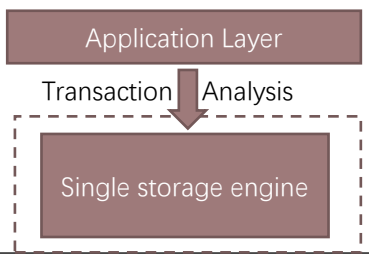
❑ HTAP - hybrid transaction and analysis processing

HTAP describes the gap elimination between OLTP and OLAP, so that a distributed database system can be applied to both transactional and analytical database scenarios, so as to meet the needs of real-time business decision-making.

HTAP allows data to enter the analysis scenario immediately after it is generated, but the biggest problem is how to better put the two mutually exclusive workloads of OLTP and OLAP on one system, and attain low resource interference, high data visibility and short delay.

Currently, HTAP has two schemes: separate architecture and unified architecture. Separate architecture is the mainstream scheme. The integration of cloud native architecture environment and HTAP system is in the tendency to derive new HTAP product solutions and technical features.

HTAP Architecture

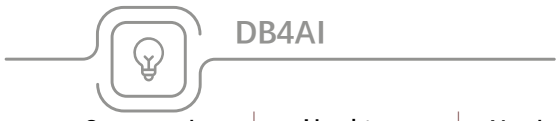
	Not HTAP	Separate system HTAP	Single system HTAP
	OLTP+OLAP in different system	Row storage + column storage separation storage engine	Single storage engine
Architecture			
Definition	OLTP and OLAP are loosely coupled, OLTP results are synchronized to OLAP through ETL, the underlying shared storage shortens the data synchronization time, and hybrid processing is implemented at the application layer to present HTAP capabilities as a whole.	Based on the distributed architecture, the row storage engine handles transaction as OLTP, and the column storage engine analyzes as OLAP, replicates data between engines following a consensus protocol, and present HTAP at the database layer.	Use a single storage engine to support both OLTP transaction processing and OLAP analysis, and implement HTAP at the lowest level.
Products	• SAP	• TiDB、PolarDB、Oceanbase、GaussDB、TDSQL、F1	• Hive、Impala、Kudu、Hyper、MemSQL
Adv.	• Preliminary integration of AP and TP	• elastic expansion on demand • Mature resource isolation technology, high performance	• Completely integrate TP and AP with low latency • High data visibility
Disad.	• High operation and maintenance cost • Synchronization delay, transaction analysis delay	• data synchronization delay exist • Poor data visibility	• immature technology • Poor row and column isolation

Source: DTCC, Leadleo

AI Native Database

- **In needs of artificial intelligence for data management, the future database must redefine and design itself from the perspective of artificial intelligence, from the aspects of data model, data operation model, execution optimization engine, etc.**

AI and DB



Manual ↓ Auto	Empirical ↓ Data-Driven	Heuristics ↓ Intelligent	Separated ↓ Out of box	Hard to use ↓ Easy to use	Hard to optimize ↓ Easy to opt
• Self-optimizing • Self-tuning	• Self-monitoring • Self-healing	• Self-design • Self-security	• AI+DB hybrid model • AI+DB hybrid inference	• Declarative AI • Data governance	• AI optimization • Model management

Improve database self-optimization, self-operation and maintenance, self-monitoring, self-diagnosis through AI

Lowering the threshold for AI use through built-in AI in the database,

Source: Tsinghua University、CCF, Leadleo

The evolution of database governance models

1970s-1990s Self-Adaptive Database

Collects the performance indicators when the application reads data, and then finds an optimal strategy to improve access performance based on the cost model.Focuses on index selection, automatic database partitioning, and data slicing.

Reduce the data management burden on application developers with a DBMS.

1990s-2000s Self-Tuning Database

Tuning tools help DBAs choose optimized indexes, materialized views and partitioned tables according to the estimated workload of executing a specific query and comparing different query execution strategies in the same environment.

Up to DBA to decide to optimize operations, relying on human experience and decision making.

2000s-2010s Cloud Database

With the scale and complexity of cloud platforms, autonomous systems become especially important. Cloud service providers provide customized tools to control deployment, as well as black-box orchestration tools to easily modify the number of machines in the cluster.

Automate operations on the cloud platform.

2018- Self-Driving Database

Automatically decide operation such as modifying physical design, modifying parameters and modifying physical resources with the capability of self-learnning according to the feedback of the operation and the decision-making process.

Automatically select optimization methods, decide when to optimize, and continue to learn.

Source: CMU、Andy Pavlo、MODB, Frost&Sullivan, Leadleo

❑ DB4AI

The database supports built-in AI algorithms, and realizes AI optimization in the database from various perspectives such as data management, query processing, and query optimization. However, the current mainstream databases are designed and optimized with traditional SQL language operations, which are not compatible with the complex operations of AI and can be efficiently supported.

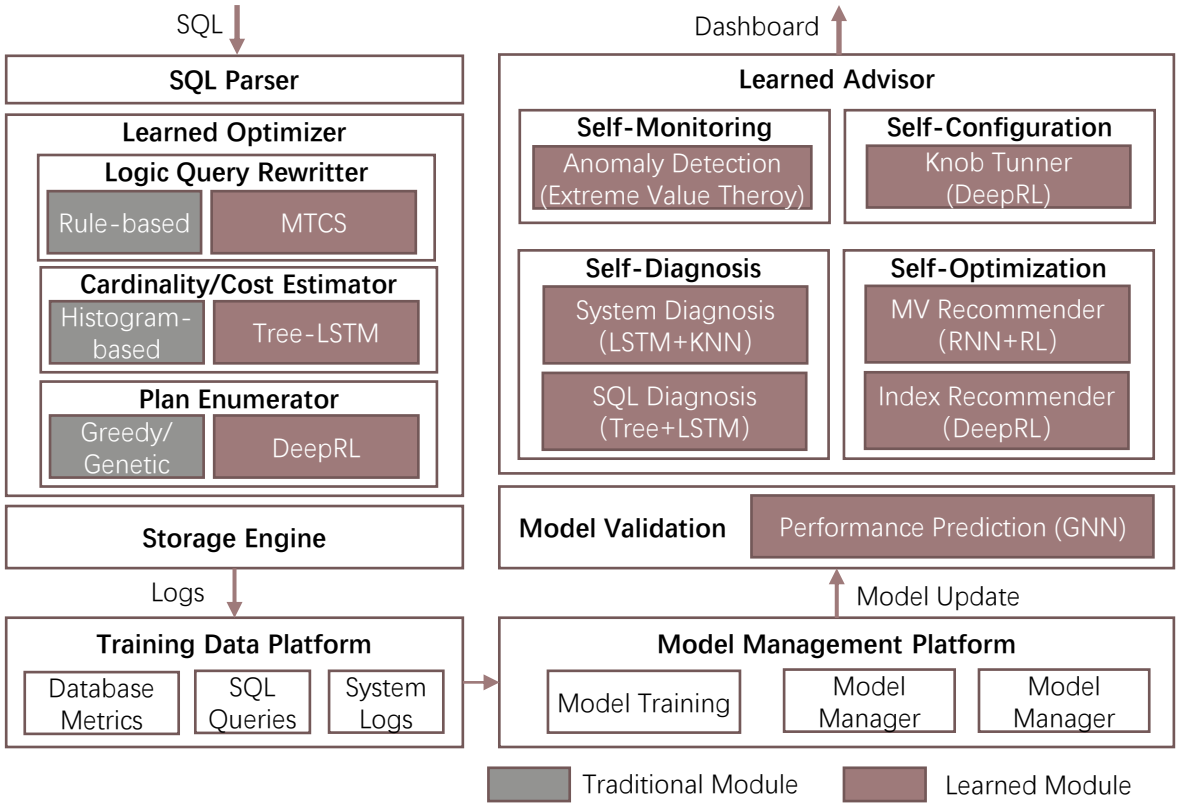
To fully meet the requirements of AI for data management, future database technology must redefine and design databases from the perspective of AI, and promote theoretical and practical innovations in data management from the perspective of data models, data operation models, and execution optimization engines.

❑ AI4DB

Database governance is an important way to ensure database security and control. With the development of business informatization, the scale and complexity of data faced by databases have grown exponentially. Traditional database optimization tools based on experience can no longer meet high-performance requirements such as load tuning. A learning-based database optimization tool is needed: AI4DB.

The database governance model urgently needs operation automation based on cloud platform, automatic parameter adjustment and optimization based on AI, data-driven self-monitoring and self-operation and maintenance, intelligent self-diagnosis and self-design to reduce or even cancel the dependence on DBA, making the database more efficient, smart and better adaptability to different scenarios.

Autonomous Database System



Source: Tsinghua University、VLDB、opengauss，Leadleo

- **Parameter Optimization：** By combining AI technologies such as deep reinforcement learning and global search algorithms, the optimal database parameter configuration can be obtained without manual intervention.

Vendors: Huawei Cloud, Tencent Cloud, Oceanbase, Baidu AI Cloud、Esgyn, Transwarp, Beagledata, Alibaba Cloud, etc.

- **Slow SQL Discovery：** The SQL statement execution time prediction tool uses a template method to predict the execution time of SQL statements based on the logical similarity of statements and historical execution records without obtaining the execution plan of the SQL statement.

Vendors: Huawei Cloud, PingCAP, Tencent Cloud, Oceanbase, Baidu AI Cloud, Esgyn, Transwarp, Beagledata, Gbase, Alibaba Cloud, etc.

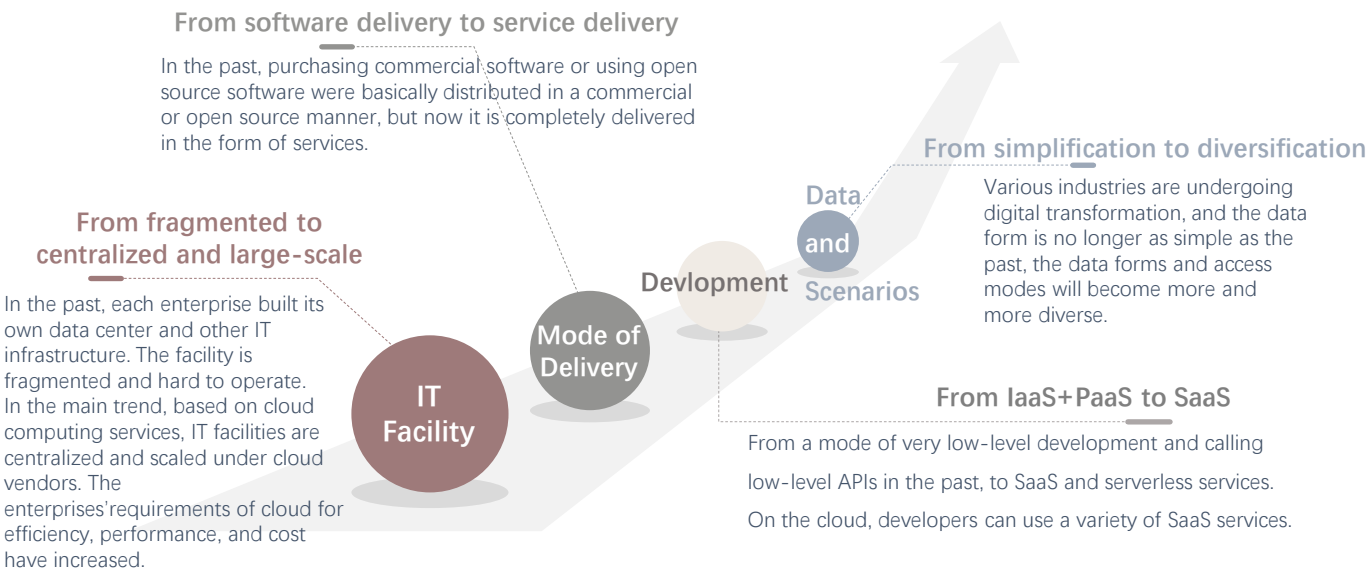
- **Index Recommendation：** It contains three sub-functions: single-query index recommendation, virtual index, and workload-level index recommendation. Machine learning algorithms are used to predict and classify which query plan to use, and a neural network-based cost model is used to alleviate the problems caused by traditional models.

Vendors: Huawei Cloud, Tencent Cloud, Beagledata, Transwarp, Baidu AI Cloud etc,

Cloud Native Database

- The flexibility and cost efficency that cloud services offer are deeply matched with database users'demand. It is particularly important to build database services on the cloud and design a cloud-native database that takes the basic cloud first and adapts to the characteristics of the cloud

The trend of database development in the cloud era



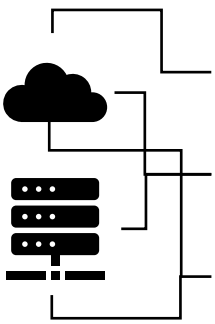
Source: Tencent Cloud、GDCA, Leadleo

❑ **Cloud computing Era**

Cloud computing, fog computing, and edge computing together constitute the infrastructure of the cloud era. Cloud computing is centralized computing; fog computing is a distributed computing method with a hierarchical and networked structure; edge computing relies on individual nodes that do not form a network.

The current cloud databases are all based on cloud computing and are in continuous development. Cloud computing, fog computing, and edge computing will place differentiated requirements on databases for data storage, management, computing, and exchange. In the future, the types and forms of databases will continue to evolve to suit different types of applications.

Data Solution for Database Migration to Cloud



Option 1:
Migrate the database to the cloud as a whole

- Database and Application Compatibility
- Transmission performance

Option 2:
Data is under the cloud, management is on the cloud

- Switch
- Data Management Security and Efficiency
- Intelligent operation and maintenance

Option3:
Production data is in the cloud, secondary data is in the cloud

- Production data: generated by the business
- Secondary data: backup, testing, offline analysis

Source: DTCC, Leadleo

Cloud deployment database

	Cloud-hosted Database	Cloud-service Database	Cloud-native Database
Characteristics	• Infrastructure as a Service Deployment • Stand-alone cloud deployment • Migration implementation	• Platform as a Service Deployment • Enterprise Feature Database • Free of O&M	• Database as a Service (DBaaS) • Serverless • Flexible Multitenancy
Definition	Deploy the traditional database, which originally in the physical IDC, on the cloud host to call cloud service provider's computing and storage resources. Users need to have an O&M team and DBA to be responsible for the availability, security, and performance of the database system to maintain the enterprise-level capabilities of the database.	Users do not need to pay attention to the specific deployment method of the database. Database vendors provide enterprise-level features including high availability, data security, and online capacity scaling. Users can access directly through the interface. An O&M and DBA team is unnecessary. Database vendors can additionally provide expert services including data model design, SQL statement optimization, and performance stress testing.	On the basis of the dynamic "resource pool", the internal computing and storage of the database are separated, and the storage management is placed in the lower-level shared storage, thereby solving the delay problem caused by data synchronization, and at the same time increasing the horizontal scalability. The cloud native model further reducing users' cost of owning a database.
Pay	• Charge by equipment	• Charge by subscription	• Charge on demand
Adv.	• Direct migration to the cloud, easy to implement • Strong isolation	• Free O&M costs, high ROI • User experience improvement • High security and high performance	• Extreme flexibility, high resource efficiency • High availability,security,performance • Intelligent autonomy, continuous iterative update
Disad.	• low resource consolidation density • Low service and high maintenance costs • low availability	• fail dangers of load optimization • Weak elasticity, average resource utilization	• High cost of upfront architecture
DBs	• Oracle • RDS、Gbase	• TiDB、GaussDB、TDSQL、PolarDB-X、Oceanbase	• Aurora、Socrates、PolarDB • TiDB Cloud、TDSQL-C、 • GaussDB(for opengauss) 、

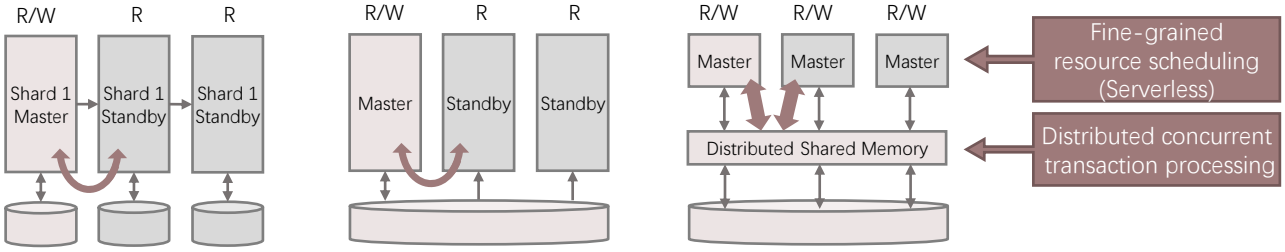
Source: Frost&Sullivan, Leadleo

❑ Database trends from cloud hosting to cloud native

Database migration to the cloud initially used infrastructure as a service (IaaS) to directly host traditional databases on the cloud. Relational database service (RDS) is such a product. For solutions like RDS, performance and transactions need to be compromised in the process of migrating to the cloud, and there are problems such as low resource utilization, high maintenance costs, and low availability.

Therefore, compared with migrating databases to the cloud, It is particularly important to build database services on the cloud and design a cloud-native database that takes the basic cloud first and adapts to the characteristics of the cloud.

Evolution of cloud-native database architecture



Traditional Database

Cloud Native 1.0
Computing-Storage Separation

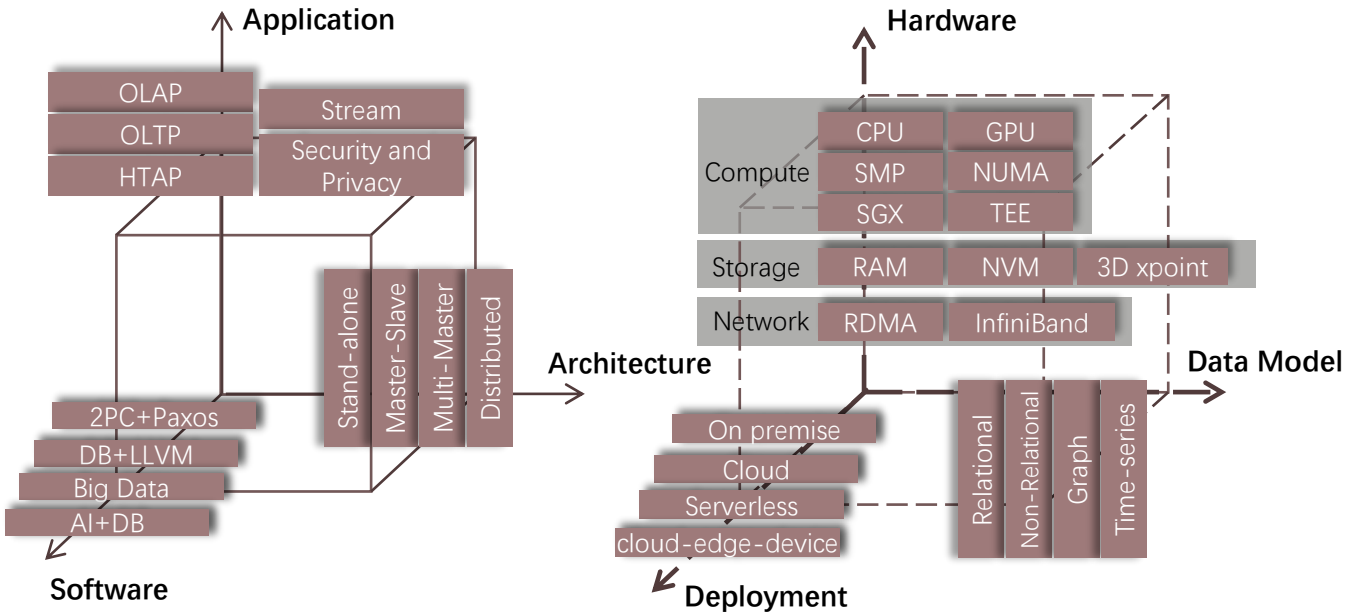
Cloud Native 2.0
Multi Write/Read,
Computing-Memory-Storage Separation

Source: Tsinghua University, Leadleo

future development trend of the database

- Distributed database technology has entered the mature stage of commercial applications, but distributed architecture is only one of the many dimensions of the database. From a long-term perspective, what kinds of development will the database technology behave?

Dismantling of the development dimension of database technology



Source: Tsinghua University, Frost&Sullivan, Leadleo

Database Development Trends in the Next Fifty Years

Now	Future
The high concurrency and low latency requirements of application systems in the development of the Internet make database solution scenarios divided into OLTP, OLAP, and HTAP.	The driving force for the database development still comes from the storage scale, performance requirements and hardware. E.g. distributed databases solve the problem of scalability,
The relational model and SQL are still the mainstream models and interaction languages for databases. NoSQL began to advance to NewSQL by supporting SQL and ACID transactions	SQL will be implemented by the application framework in the future, and technical personnel will access data through natural language, reducing the threshold between code language and data.
The database is still a computing system that requires human intervention, but is optimized through containerized deployment, automated tuning, and cloud-based resource allocation	DBAs will disappear, and databases will gradually become autonomous. Database is dynamically adjusted according to the workload, where ML+DB will develop.
Multimodal, storing video and other media data through arrays and matrices. Now there are data for different data models, such as graphs, spatial data, etc.	The IoT requires data synchronization between devices. Heterogeneous multimodal databases directly achieve differentiated synchronization, which equiring a unifiedl protocol
As for hardware, the emergence of large-capacity memory greatly improves the performance in OLTP scenarios; The distributed database replaces the mainframes with X86 reducing infrastructure costs	Storage and network are the two most urgent components of the database base, such as persistent memory NVM and high-speed network RDMA

Source: CMU Andy Pavlo, Frost&Sullivan, Leadleo

Distributed database architecture selection

- **In terms of architecture selection, each architecture has its most advantageous application scenarios. Under the trend of distributed database, enterprises should choose distribution according to the actual requirements in detailed scenarios.**

❑ Distributed database architecture

Database architecture is a way to build data processing components in order to adapt to external requirements. The distributed database architecture is a software architecture technology that takes transaction consistency and availability as the core, and integrates storage, computing, scalability, ease of use and a series of features to meet user needs.

❑ Complexity of database system architecture

What functions are provided, how tos use, and how modules collaborate are all to be considered when designing a database architecture. In the database system, there are all kinds of explicit or implicit dependence between each module of the software. With the growth of the business system, the increasing number of modules also makes the complex relationship between modules grow geometrically, so that the database itself becomes a highly coupled complex system.

In present choice of database distributed technology route, the primary goal is to solve the problem of data capacity expansion. The mainstream solutions are sharding with middleware, native distributed, etc.

❑ Architecture within the scope of business

In terms of database selection, different technical routes and products have their own advantages and disadvantages. According to different practical needs, there are completely different database adaptation transformation scheme. It is unrealistic to discuss the database architecture out of the scope of business.

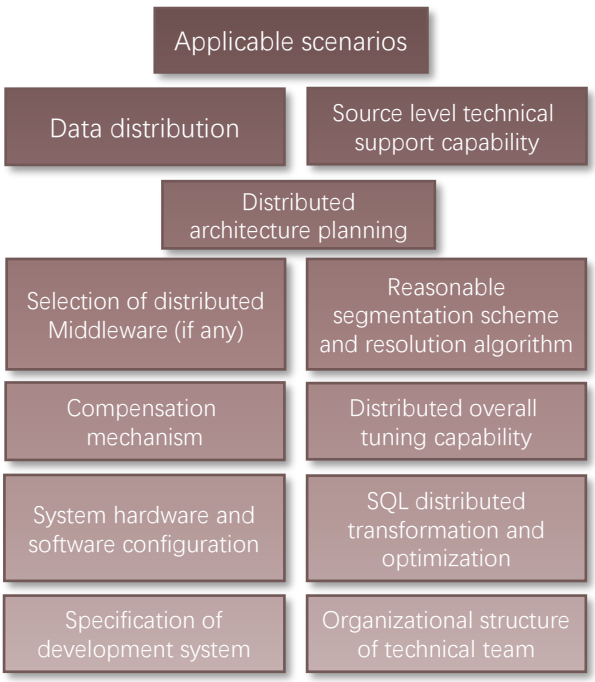
None of the technical routes or system products is perfect in software engineering. ACID characteristics in distributed system require balance and trade-offs among various characteristics of the system. The most suitable database product needs actual business scenarios as the most important basis.

❑ Business requirements drive database innovation

The demand side of the database selection should avoid the misconception of "solving all problems with one database". It has nothing to do with the database technology advancement, there will still be a coexistence of a variety of technical routes in the future. Database selection needs to fully understand the actual business scenarios, follow the objective laws of software architecture, and then choose to optimize and expand the existing centralized system or transform to the distributed architecture.

In terms of architecture selection, standalone-database, sharding+middleware or native distributed database, all have their most advantageous application scenarios. Under the trend of distributed database, enterprises should choose distributed architecture according to the actual requirements in detailed scenarios.

Architecture design considerations

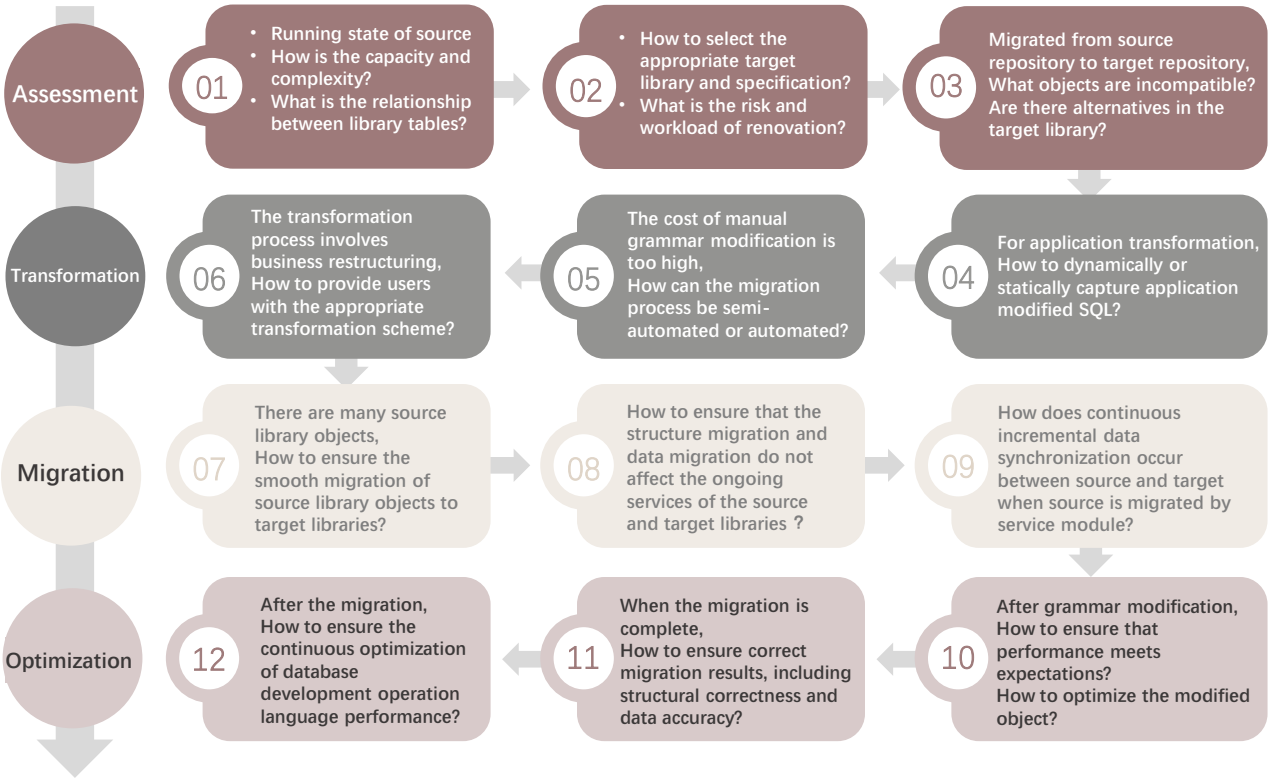


Sources: Frost&Sullivan, Leadleo

Data migration tools

- Database migration is a complex project. In order to improve migration efficiency, database vendors have launched migration tools and integrated data migration task management platform for their own databases. But tools and platforms only aid in the migration process.

Problems faced by traditional database migration



Sources: DTCC2020, Leadleo

Data migration tools or platform of different vendors

Vendors	Data migration tools or platform
PingCAP	TiDB Data Migration TiCDC、TiDB Lightning
Alibaba Cloud	Data Transmission Service，DTS
Oceanbase	Oceanbase Migration Service(OMS)
GBASE	GBase Migration Toolkit
Tencent Cloud	Data Transmission Service，DTS
Huawei Cloud	Database and application migration (UGO) Data Replication Service (DRS)
Amazon Web Service	AWS Database Migration Service

Sources: companys' official websites, Leadleo

❑ Data Migration

Database migration is a complex project, which needs to follow a series of processes like evaluation, transformation, migration and optimization, including infrastructure, application development, business process and so on. The migration process usually lasts for a long time.

❑ Data migration tools

In order to improve migration efficiency, database vendors have launched migration tools and integrated data migration task management platform for their own databases. But tools and platforms only aid in the migration process. By managing and scheduling data synchronization tasks, executing specific data synchronization tasks and controlling the command line of DM cluster, componentize the automatically managed migration operations to form a data migration tool that is convenient for users to use by themselves.

Database vendor consultant support

- **Database selection needs to consider the manufacturer's expert team composed of delivery engineers, operation and maintenance engineers, DBA and R&D to provide customers with pre-sales, mid-sales and post-sales consulting services, technical guidance and solution construction. The service capability of the manufacturer is also an important evaluation index for user selection.**

Database vendor consultant support service process reference

1. Consulting services: deeply communicate with users through technical consultants and industry application experts, help users sort out their needs and form targeted solutions. 2. PoC: Provide PoC planning and capability verification scheme to users according to their requirements for product capability verification, and support relevant PoC product capability verification of user organizations.	Pre-sales	1. Customer demand survey, environmental monitoring, planning and output 《Database product project implementation plan》	Value-added service & on-site service: 1. Daily counseling 2. System normalization service 3. Develop support services 4. Architecture support services 5. Data security services 6. Emergency services
Provide users with professional services including implementation plan planning, various implementation technical manuals, remote and on-site implementation services, implementation guarantee for product launching, technical delivery and knowledge transfer after successful implementation, etc.	Mid-sales	2. On-site deployment, implementation and function debugging of database products. 3. Training on the use and daily maintenance of components 4. Acceptance test, including functional test, performance test and high availability test; output XX Project - Acceptance Report of Database Product Deployment	Other value-added services: 1.Advanced DBA Services 2.Database Training Service 3.Upgrading and Expanding Services 4.System and architecture optimization services 5.Service reinsurance service 6.Health inspection 7.Security reinforcement
Provide a series of comprehensive support services, including emergency response, daily question answering, upgrade and change support, inspection and diagnosis, troubleshooting, emergency response, major fault disposal and root cause analysis, on-site support during important periods and other operation and maintenance support services	After-sales	Standard post-sales service: 1. Hotline: 24/7 2. Remote problem handling: 24/7 3. Defect repair support: provide on-site services 4. Checking support 5. emergency response 6. Update and upgrade authorization	Value-added after-sales service & on-site operation and maintenance services: 1. On-site duty: 5*8h 2. Work order management 3. Defect repair support 4. inspection support 5. Emergency response: 24/7 6. update and upgrade authorization 7. database monitoring 8. backup recovery 9. safety audit

Database selection needs to consider the manufacturer's expert team composed of delivery engineers, operation and maintenance engineers, DBA and R&D to provide customers with pre-sales, mid-sales and post-sales consulting services, technical guidance and solution construction. The service capability of the manufacturer is also an important evaluation index for user selection.

Sources: Tencent Cloud, PingCAP, Frost&Sullivan, Leadleo

Open source situation

- The development of open source concept has experienced the free developer movement represented by Linux and the community driven by technology stack represented by Hadoop. The open source ecosystem of database has stepped into the stage of community collaboration driven by extreme scenarios.

The evolution of open source innovation

1 Free developer movement

PC Era

From the traditional software development cathedral mode to market mode: parallel, point-to-point, dynamic multi-person collaborative development mode.

Developers usually only rely on the Internet to communicate. In this seemingly chaotic and disorderly development environment, software with high quality, efficiency and vitality was produced unexpectedly, such as Linux world-class operating system.

2 Community driven by technology stack

Mobile Internet Era

Business requirements have increased the speed of technology stack updates to a new level.

In the development process of distributed system, due to the separation of services, more distributed requirements and technologies to cope with these requirements are forced to be derived, and technology stack communities such as Hadoop, Spark and Spring Cloud are gradually formed.

For example, Hadoop also includes Hive, MapReduce, HDFS and other technologies to achieve different functions.

3 Community driven by extreme scenarios

Cloud native and muti-cloud era

Business and technology are closer to each other, faster iteration, better growth rates, and more representative of the future will be competitive trump cards. The "scenario-driven development" is executed on top of a popular open source software, the product "touch" comes from the accurate grasp of the universal needs, high popularity represents a high user base and multiple scenarios to reach. After rounds of user feedback, bug modifications, and iterative suggestions, a positive iteration cycle is achieved, covering a wider range of application scenarios.

Sources: PingcAP, Leadleo

Open source situation of domestic database vendors

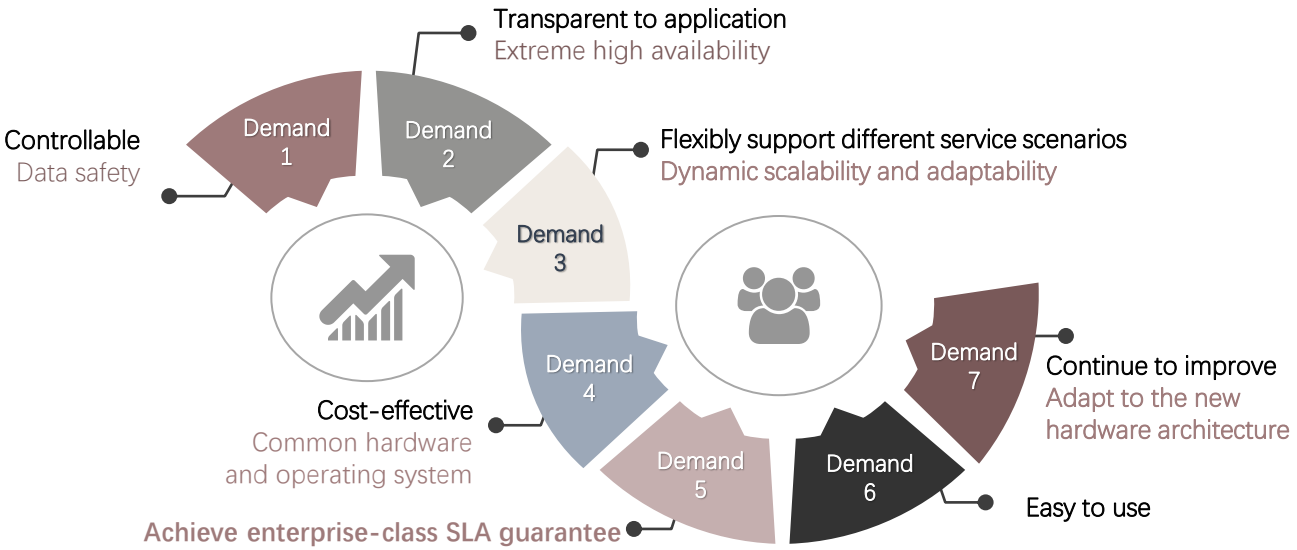
Open source project	Contributor	Time	Community version	Solved issues	Community contributor	Number of users
TiDB	PingCAP	2015	Enterprise version	8.3k	1.6k	2k
OpenGauss-server	Huawei Cloud	2019	Database kernel	3k	2.7k	500+
Tbase	Tencent Cloud	2019	Database kernel	60k	30+	50+
Oceanbase	Oceanbase	2021	Enterprise version	640+	100+	2.3k
PolarDB for PostgreSQL	Alibaba Cloud	2021	Enterprise version	30+	53	N/A
PolarDB-X for MySQL	Alibaba Cloud	2021	High availability version	19	13	N/A
SequoiaDB	SequoiaDB	2014	Enterprise version	68	N/A	N/A

The statistical period is up to 2022/3
Sources: Github, Gitee, Frost & Sullivan, Leadleo

Distributed database user portrait

- The development of distributed database technology should meet the needs of the times and the market, and return to the rigid needs of database users. The current distributed database needs to reach the level of centralized architecture products in various dimensions so as to play its performance and cost advantages in various scenarios and penetrate into various industries.

Seven elements of database users' rigid demand



Sources: ESGYN, Leadleo

❑ **Development of distributed database technology should be closely refer to the scenarios**

The development of distributed database technology should meet the needs of the times and the market, and return to the rigid needs of database users. The current distributed database needs to reach the level of centralized architecture products in various dimensions so as to play its performance and cost advantages in various scenarios and penetrate into various industries.

Demand factors of database users in different industry

	IT regulatory environment	Data business complexity	Core data business characteristics	Cost sensitivity	IT capacity reserve status
Internet	weak	strong	massive, high concurrency, flexibility	strong	strong
Telecom	strong	strong	strong transaction, high concurrency	weak	strong
Entertainment	strong	medium	elastic expansion, massive, low latency	strong	medium
Traffic	strong	strong	high concurrency, high availability, low latency	weak	medium
Logistics	medium	strong	massive, high concurrency, high frequency	medium	strong
Government	strong	medium	strong transaction, high availability, correlation analysis	medium	weak
Medical	medium	medium	low latency, correlation analysis	weak	weak
Manufacture	medium	medium	time variable analysis, high frequency	medium	weak

Sources: CAICT, Leadleo

Application landing scene enterprise map

- Chinese distributed database vendors show a differentiated layout, and the chart screens out Chinese database vendors that provide distributed databases and derivative services to enterprises and institutions in the Internet, telecommunications, transportation, logistics, e-government and other industries.

China Database Industry Application Atlas

Represent vendors in distributed databases application scenarios

Internet Industry

Online mall, VIP membership system, mini program, order system, electronic contract management, real-time risk control, SaaS platform, background data management, batch stream integrated +ML high-concurrency scenario service, intelligent recommendation system

E-commerce	Instant messaging(IM)	Web services
阿里云	腾讯云	腾讯云
OCEANBASE	HUAWEI	百度智能云
腾讯云	阿里云	PingCAP

Telecom Industry

Telecom industry real-time data analysis, subscriber relationship writing, reconciliation platform, phone bill query analysis system, core business migration

GBASE® 易鲸捷 腾讯云 BeagleData 天云数据

Entertainment Industry

Game system, core payment system, user center, advertising system, online education system, smart campus system, curriculum system, public opinion monitoring system

Game	Audio and video	Online education
腾讯云	HUAWEI	阿里云
PingCAP	阿里云	腾讯云

Note: the order and size of the logos above have no practical significance and do not involve ranking, only show some of the industry representative enterprises
Sources: Frost & Sullivan, Leadleo

China Database Industry Application Atlas(continued)

Represent vendors in distributed databases application scenarios

Transportation Industry

Airport customer management system, logistics system, intelligent transportation system, charging management system, car rental system, ticket booking system, railway freight information application platform, railway information operation and maintenance control platform, civil aviation information center comprehensive analysis platform

GBASE® HUAWEI 阿里云 TRANSWARP 星环科技

Logistics Industry

Order and flow system, detailed historical data query, operation analysis and decision making, freight system, storage system, logistics system

PingCAP HUAWEI 阿里云 腾讯云

E-government and Public Service Industry

Government service system, enterprise-related government service platform, mobile office platform, digital finance system, epidemic prevention and control system, public security support system, data hub system, intelligent application system, information management system

HUAWEI GBASE® 易鲸捷 BeagleData 天云数据 腾讯云 ZTE中兴

Medical Industry

Drug target prediction, gene mapping construction, medical security information platform, National Center for Disease Control and Prevention medical Immune system

BeagleData 天云数据 腾讯云 亚马逊云科技

Energy Industry

Power user change relationship identification, State Grid full-service data center, energy analysis platform, oilfield knowledge map

OCEANBASE HUAWEI GBASE®

Manufacture Industry

Supply chain and logistics system, back-end service of algorithmic platform, electronic distribution order model service

易鲸捷 PingCAP 腾讯云

Note: the order and size of the logos above have no practical significance and do not involve ranking, only show some of the industry representative enterprises
Sources: Frost & Sullivan, Leadleo

www.leadleo.com

https://space.bilibili.com/647223552

https://weibo.com/u/7303360042