

2021

China Data Management Solutions Market Report

2021年中国数据管理解决方案市场报告

2021中国データ管理ソリューション市場レポート

Tags: Data lakehouse, Data lake, Data Warehouse, Serverless, Machine learning

(Summary Version)

Any content provided in the report (including but not limited to data, text, charts, images, etc.) is the exclusive and highly confidential document of LeadLeo Research Institute (unless the source is otherwise indicated in the report). Without the prior written permission of LeadLeo Research Institute, no one is allowed to copy, reproduce, disseminate, publish, quote, adapt or compile the contents of this report in any way. If any behaviour violating the above agreement occurs, LeadLeo Research Institute reserves the right to take legal measures and hold relevant personnel responsible. LeadLeo Research Institute uses "LeadLeo Research Institute" or "LeadLeo" trade name or trademark in all business activities conducted by LeadLeo Research Institute. LeadLeo Research Institute neither has other branches other than the aforementioned name nor does it authorize or employ any other third party to carry out business activities on behalf of LeadLeo Research Institute.

Instruction

Frost & Sullivan hereby releases the annual report "China Data Management Solutions Market Report 2021" as part of the China Data Management Series Report. The purpose of this report is to sort out the development trends of data warehouse, data lake, and intelligent lake warehouse products and technologies. Based on the current development situation of China data management market, this report provides insight into the characteristics of users, market stock space and incremental space, and determines the position of various competitors in the field of data management solutions based on the market development prospect.

Frost & Sullivan and LeadLeo Research Institute conducted downstream user experience surveys on core products in the data management solutions field. Respondents are of different sizes and in different segments in each of its industry that includes finance, consumption, pan-entertainment, telecommunications, energy, transportation, manufacturing, government and other fields.

Trends in data management solutions 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

Technology Trends

As two separate data management paradigms, data warehouse and data lake both have mature technology accumulation. In long-term practice they co-exist in a hybrid architecture of lake + warehouse: data lake is used for extraction and processing of original data, while relying on data warehouses for publishing in the data pipeline.

Driven by the needs of users, data lake and data warehouse providers expand the original paradigm to the limits of its scope, and gradually form two paths of "data lakehouse", namely "warehouse on lake" and "warehouse to lake". Although in the underlying logic, lake-warehouse integration is still a binary system, but it can greatly help users to encapsulate a big data paradigm more closely with their needs on the basis of their original IT basis, or directly mount the lake-warehouse integration system with fully hosted services.

Market Analysis

The demand for professionals with 1-5 years of work experience is the highest in the talent market. Data analysts and data scientists have better average salary and salary increase. The demand structure for data management talents varies from industry to industry, with significant demand for data development engineers in IT and Internet industries, and significant demand for data analysts in retail and e-commerce industries.

Security and stability, full functionality, compatibility, cost reduction and efficiency, performance, and expansion limits are the six demand dimensions concerned by users of data management solutions. Machine learning scenarios, open source engine compatibility, and business continuity are the demand keywords emphasized by interviewed users.

From an enterprise perspective, it is easy to fall into the trap of hidden costs and unmet needs without digging into the details of products and services, since products from different providers look similar. Solution selection needs to focus on pricing structure, multi-cloud deployment, artificial intelligence, universal adaptation and other dimensions to comprehensively judge the product and service solutions and quotations from different vendors.

Iterative changes in big data technology

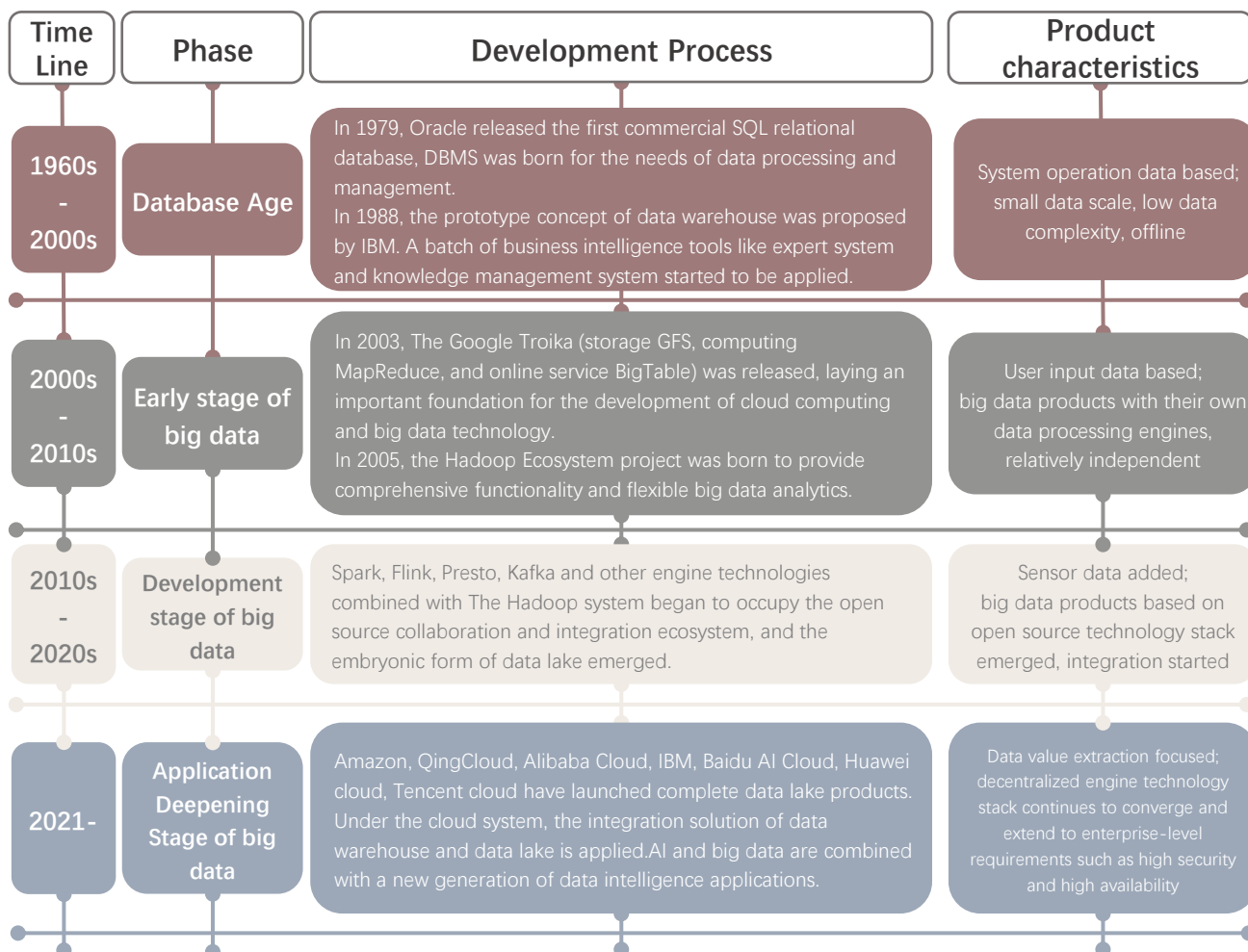
- In the big data industry, reducing storage cost, improving computing speed, multi-dimensional analysis and processing of data, and empowering enterprises to leverage the value of data are the keys to achieving profitability in the big data industry and the root cause of the booming big data technology

Big data technology

The literal understanding of Big Data is massive Data, but this perspective is abstract. In the age of network information, the objective significance of big data is not its huge data scale, but how to store and process data professionally, and dig and extract the required knowledge value from it.

Technological breakthroughs usually come from the actual market demand for products. The continuous development of the Internet, cloud, AI and the integration of big data technology meet business needs. In the big data industry, reducing storage cost, improving computing speed, multi-dimensional analysis and processing of data, and empowering enterprises to leverage the value of data are the keys to achieving profitability in the big data industry and the root cause of the booming big data technology.

Iterative changes in big data technology

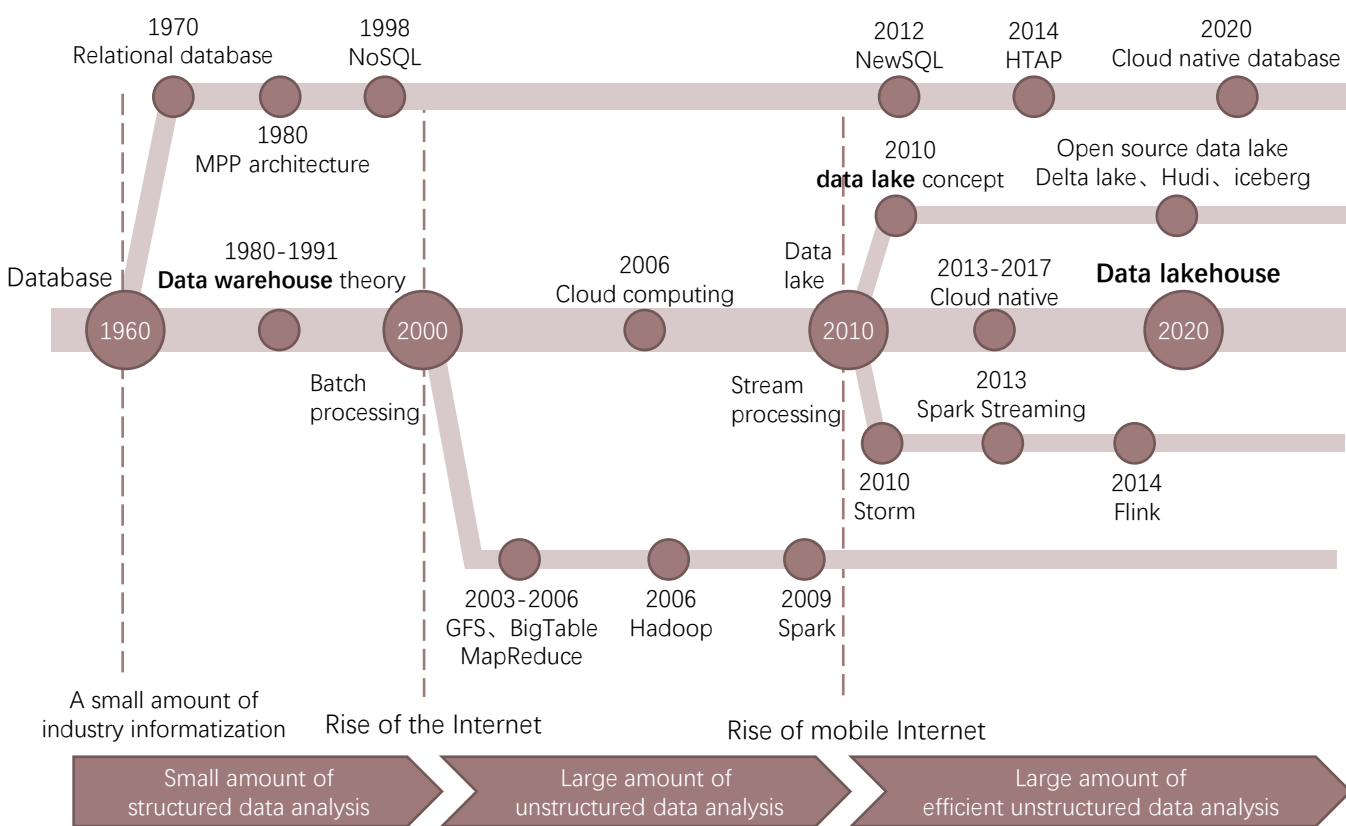


Source: Big Data Technical Standards Promotion Committee(CCSA TC601), Leadleo

Lake warehouse integration

- The lake warehouse integration further eliminates the selection difficulties for users, providing them with a data management platform that combines the structure and governance benefits of a data warehouse with the scalability of a data lake and the convenience it provides for machine learning

Classification of the technological evolution of data management platforms



Source: CAICT, LeadLeo

□ Lake warehouse integration trend

The connotation of big data technology is evolving with the development of traditional information technology and data application, and the core of big data technology system is always the basic technology of storage, calculation and processing for massive data.

During more than 60 years of development of big data technology, data application has experienced the vigorous development and demand transformation of Internet and mobile Internet. The traditional strengths of database and data warehouse based on transaction analysis processing are still the mainstay of current information technology, but they are difficult to match in the face of increasing data complexity requirements and massive elastic data scale.

The breakthrough of distributed architecture and the rise of cloud computing laid the foundation of the concept of data lake. The lake warehouse integration further eliminates the selection difficulties for users, providing them with a data management platform that combines the structure and governance benefits of a data warehouse with the scalability of a data lake and the convenience it provides for machine learning.

Data Warehouse, Data Lake and Data Lakehouse

	Data Warehouse	Data Lake	Data Lakehouse
Architecture			
Definition	<p>A data warehouse is a topic-oriented, integrated, relatively stable collection of data that reflects historical change and is used to support management.</p>	<p>A data lake is a type of system or storage that stores data in a natural/original format, usually object blocks or files. A data lake is typically a single store of complete data in an enterprise, including copies of original data generated by the original system and transformed data generated for various tasks such as reporting, visualization, advanced analysis, and machine learning.</p>	<p>The integration of lake and warehouse combines the semantic flexibility of data lake with the production optimization and delivery of data warehouse. It is a converged infrastructure environment that supports the entire process from original data to refined data, and ultimately provides optimized data for consumption.</p>
Characteristics	<ul style="list-style-type: none"> • Built-in storage system, data is provided in an abstract manner without exposing the file system • Data needs to be cleaned and transformed, usually by ETL/ELT • Emphasizes modeling and data management for business intelligence decisions 	<ul style="list-style-type: none"> • Unified storage system • Storing raw data • Rich computing model/paradigm • Not relevant to cloudification 	<ul style="list-style-type: none"> • Concurrent reads and writes of data • Data management mechanism • Direct access to the original data • Separation of computing and storage • Standardized data formats • Structured and unstructured data • End-to-end stream processing
Advantages	<ul style="list-style-type: none"> • Deep understanding of data, storage and computation for optimization • Data life cycle management, perfect bloodline system • Fine-grained data management and governance • Perfect metadata management ability, easy to build enterprise-level data center 	<ul style="list-style-type: none"> • Collect and ingest all data sources to obtain the entire data set without islands • Support ETL (extract - transpose - load) for real-time and high-speed data streams • Scalability and flexibility • Advanced analytics with artificial intelligence • Low-cost, relatively inexpensive PC servers can be used to build big data clusters 	<ul style="list-style-type: none"> • Stronger data governance capabilities • Richer variety of data • Better data security system • More flexible scale-up applications • Easier data and work migration • More unified data management system
Disadvantages	<ul style="list-style-type: none"> • Unable to process unstructured data • Poor expandability • Long path from data source to data warehouse, low concurrency, creates data islands 	<ul style="list-style-type: none"> • Shortcomings in the technical processing of transaction consistency • Challenge to data governance 	<ul style="list-style-type: none"> • Imperfect load management function of warehouse service

Source: LeadLeo

□ Implementation approach of data lakehouse

As two separate data management paradigms, data warehouse and data lake both have mature technology accumulation. In long-term practice they co-exist in a hybrid architecture of lake + warehouse: data lake is used for extraction and processing of original data, while relying on data warehouses for publishing in the data pipeline.

According to user feedback, the hybrid architecture of lake + warehouse has difficulties in data redundancy under the coexistence of Hadoop and MPP, low timeliness, consistency guarantee, operation and maintenance caused by ETL between the two systems.

Driven by the needs of users, data lake and data warehouse providers expand the original paradigm to the limits of its scope, and gradually form two paths of "data lakehouse", namely "warehouse on lake" and "warehouse to lake". Although in the underlying logic, lake-warehouse integration is still a binary system, but it can greatly help users to encapsulate a big data paradigm more closely with their needs on the basis of their original IT basis, or directly mount the lake-warehouse integration system with fully hosted services.

Implementation path of data lakehouse

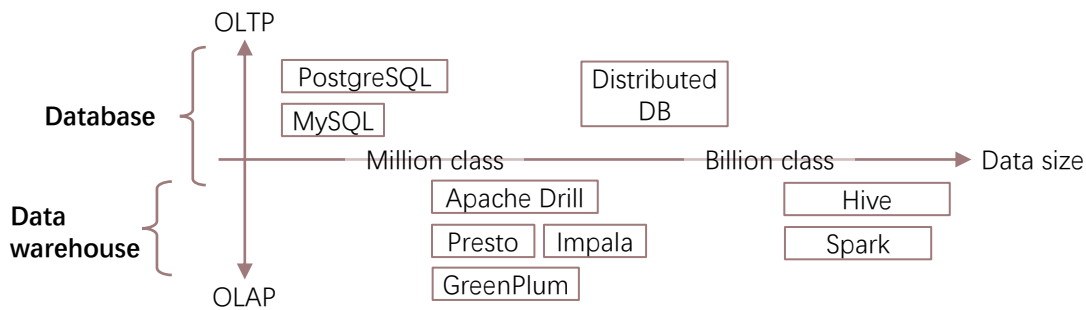
	Warehouse on lake	Warehouse to lake
overview	Data lake architecture based on public cloud, or open source Hadoop ecological components DeltaLake, Hudi, Iceberg as the middle layer of data storage to realize the unified storage of heterogeneous data of multiple sources, integrate computing engine with unified call interface, and realized the lakehouse architecture with upper and lower structure.	The pluggable architecture is used to open the boundary between the data warehouse and the unified storage of the data lake through the open interface. The data is shared at the bottom of the storage layer, and the storage and computing are completely separated. The data is imported into the compute node cluster cache for processing.
Capacity		
products	Huawei Cloud FusionInsight, DEEPEXI FastData, Transwarp lakehouse, Amazon intelligent lakehouse, Kingsoft Cloud KCDE、Delta Lake、Big Lake、Azure Synapse Analytics	Oushu Data Cloud、Maxcompute Snowflake、Redshift、AWS RedshiftSpectrum

Source: Tietoenvry, Snowflake, LeadLeo

Data Warehouse - OLAP Analysis Engine

- Different from database, data warehouse is not a pure technology, the core is to form an architecture for data integration

Load characteristics of database and data warehouse



Source: LeadLeo

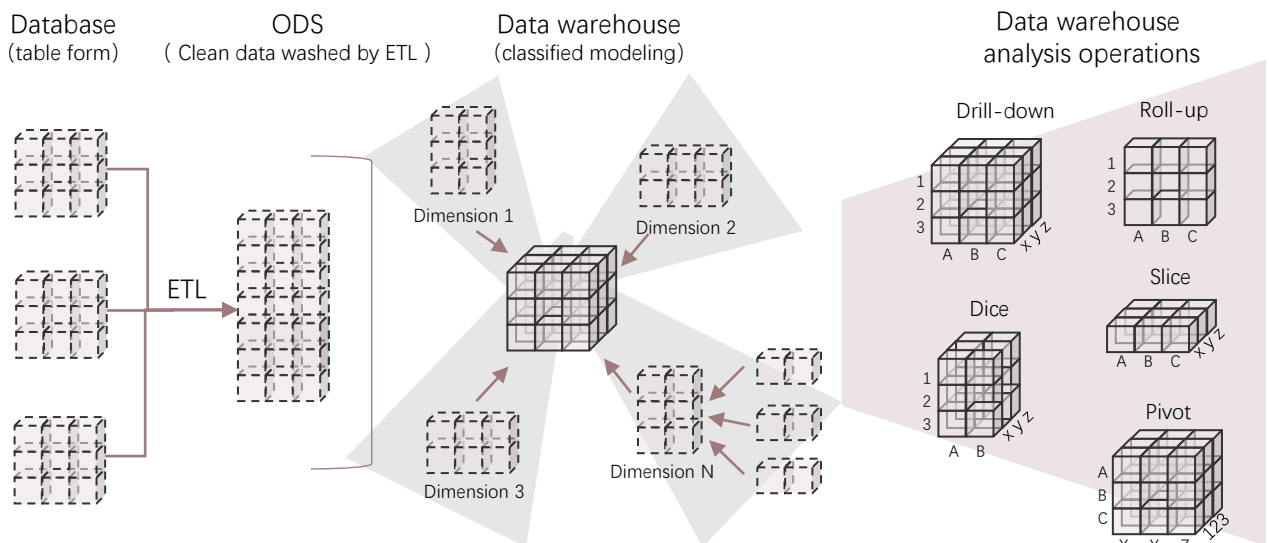
Database and data warehouse

Database and data warehouse are physical design based on traditional relational database theory. But different from database, data warehouse is not a pure technology, the core is to form an architecture for data integration.

Databases focus on OLTP while data warehouses focus on OLAP. Data warehouse is the traditional relational database (such as SQL Server, Oracle, etc.), and it can be turned into a very good data warehouse entity after strict data model design or parameter adjustment. While pure data warehouse such as Terradata, SybaseIQ is not suitable for OLTP system.

OLAP and OLTP are merging into HTAP. The enhancement of AP analysis capability by databases will gradually blur the boundary between databases and data warehouses.

Data warehouse building process



Source: CSDN, LeadLeo

Different implementations of OLAP engines

	Multidimensional OLAP (MOLAP)	Relational OLAP (ROLAP)	Hybrid OLAP (HOLAP)
Architecture			
Definition	<p>Based on native logical models that directly support multidimensional data and operations. Data is physically stored in multidimensional arrays and accessed using location techniques.</p>	<p>Store multidimensional data for analysis in a relational database. This approach relies on SQL to implement the slicing and chunking functions of traditional OLAP, which are essentially equivalent to adding a "WHERE" clause to an SQL statement.</p>	<p>Bridge the technical gap between the two products by allowing the use of both multidimensional databases (MDDDB) and relational databases (RDBMS) as data stores.</p>
Characteristics	<ul style="list-style-type: none"> • Achieve from physical level • Data is pre-computed and stored • Storage designed and optimized for OLAP • Multidimensional indexing and caching are supported 	<ul style="list-style-type: none"> • Do not use pre-computed cubes • No redundant data is imported • Use the existing relational database technology 	<ul style="list-style-type: none"> • Provides fast access to all aggregation levels • The OLAP server only stores aggregation information, and the detailed records are retained in relational database
Advantages	<ul style="list-style-type: none"> • Provide fast access to all aggregation levels • The OLAP server only stores aggregation information, and the detailed records are retained in a relational database 	<ul style="list-style-type: none"> • Easy to manage • Small storage space consumption, no dimensional limit • Queries can be implemented through SQL 	<ul style="list-style-type: none"> • Duplicate copies of detailed records are not kept, balancing disk space requirements • Optimize query performance under given usage scenarios
Disadvantages	<ul style="list-style-type: none"> • Pre-computation is resource-consuming, dimension limited and inflexible • Low data loading speed • Lack of standard data access interface • Difficulties in maintenance 	<ul style="list-style-type: none"> • Slow response • Depend on the database to perform calculations, proprietary capabilities are limited 	<ul style="list-style-type: none"> • Supports both MOLAP and ROLAP, complex architecture • Lack of flexibility
Products	<ul style="list-style-type: none"> • Druid、Kylin、Doris • ESENSOFT ABI 	<ul style="list-style-type: none"> • Amazon Redshift、Dlink、GaussDB(DWS)、OushuDB、KDW • Presto、Impala、GreenPlum、Clickhouse、Elasticsearch、Hive、Spark SQL、Flink SQL 	<ul style="list-style-type: none"> • Kylin、Hulu Sophon • Inspur cloud IEMR
Scenarios	<ul style="list-style-type: none"> • Fixed query scenarios that require high query performance: • Advertising report analysis 	<ul style="list-style-type: none"> • Scenarios with variable query modes and high query flexibility requirements: • Analysis products commonly used by data analysts 	<ul style="list-style-type: none"> • When querying aggregated data, use MOLAP • When querying detailed data, use ROLAP.

Source: CSDN, LeadLeo

Data Warehouse - Execution Model and Architecture

- The performance of the data warehouse itself and ETL depends on communication, I/O capabilities, and hardware performance, while the execution architecture determines the supporting capacity of the data warehouse

Three different execution architectures of data warehouse

	Scatter/Gather	MapReduce (Hadoop)	Massively Parallel Processing (MPP)
Architecture			
Definition	Implement a simple I/O operation on multiple buffers, such as reading data from a channel to multiple buffers, or writing data from multiple buffers to a channel.	Reliability is achieved by distributing large-scale operations on data sets to every node in the network; Each node periodically returns the work it has done and the latest status.	Using shared-nothing architecture, each node uses separate resources and has the best operating environment. Pipelined execution without waiting, data memory storage, no disk I/O.
Characteristics	<ul style="list-style-type: none"> Single node aggregation Equal to a Map and Reduce trip in MapReduce 	<ul style="list-style-type: none"> A parallel programming model for handling massive amounts of data in Hadoop Waiting gap in between tasks due to data transmission and disk I/O 	<ul style="list-style-type: none"> Shared Nothing architecture Distributed parallel execution Distributed storage of data (localization) Transverse linear extension
Advantages	<ul style="list-style-type: none"> Maximize performance benefits from local I/O 	<ul style="list-style-type: none"> Easy to have good scalability in the case of infinite computing resources and no correlation of data Low cost, can be extended with low-end servers 	<ul style="list-style-type: none"> Emphasis on real-time data calculation, greater I/O capability Column storage is used to save storage space Ease of use and scalability
Disadvantages	<ul style="list-style-type: none"> Operations such as large table Join and high cardinality aggregation cannot be completed 	<ul style="list-style-type: none"> Limited by resource allocation, data correlation and other factors The interface is not compatible with SQL and has weak support for complex queries Generate a large number of temporary files 	<ul style="list-style-type: none"> Do not support unstructured data processing, such as log analysis and text analysis Scalability is not as good as architectures such as MR, and performance bottlenecks determine the nodes with the worst performance The intermediate result needs to be recalculated when the node is down, the probability of SQL retry is high
Products	<ul style="list-style-type: none"> Elasticsearch、Druid、Kylin 	<ul style="list-style-type: none"> Hive、Spark SQL、Hadoop IEMR、Inceptor、KMR KDC 	<ul style="list-style-type: none"> Amazon Redshift、KCDE GaussDB(DWS)、HetuEngine Presto、Impala、Doris、Clickhouse、Greenplum、Flink SQL、Asterdata

Source: Doris, CSDN, LeadLeo

Comparative analysis of execution architecture

	Platform openness	SQL standard	Operational difficulty	Scalability	Cost	Management cost	Data size	Data structure
Traditional	Low	High	Mid	Low	High	Mid	TB level	Structured
Hadoop	High	Low	Hard	High	Low	High	PB level	Unstructured/semi-structured/structured
MPP	Low	High	Mid	Mid	Mid	Mid	Partly PB level	Structured

Source: Apache, LeadLeo

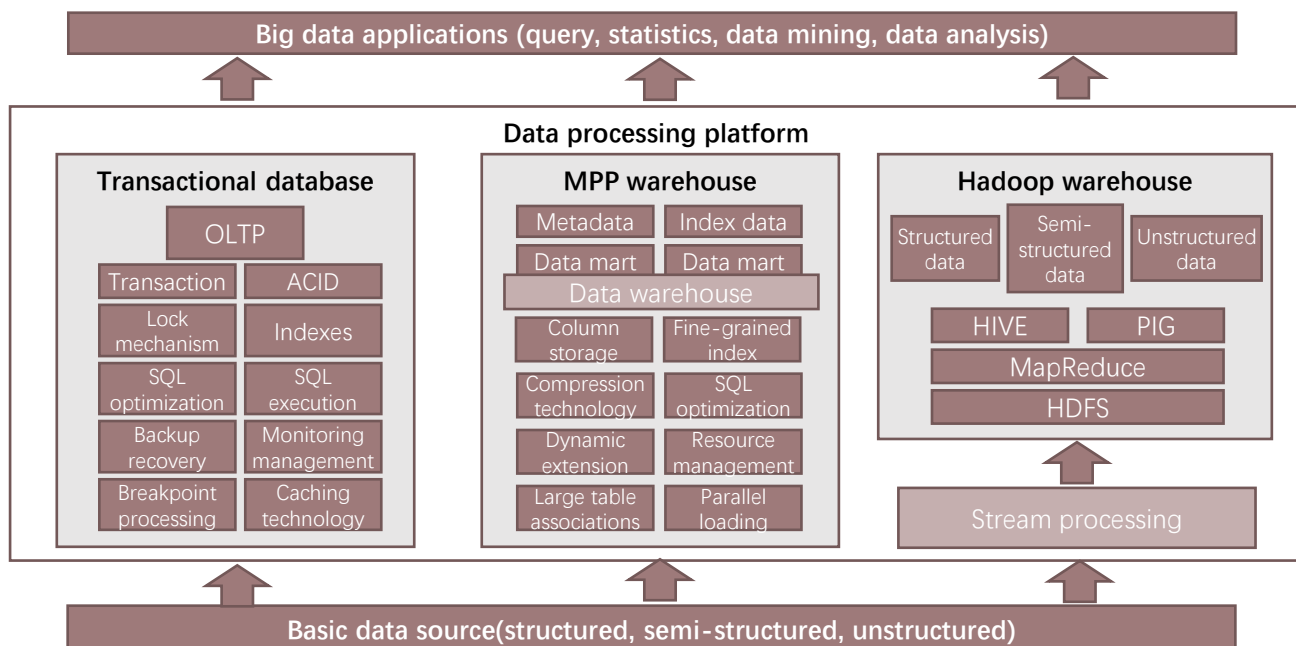
❑ MPP-Hadoop architecture

Hadoop architecture (MapReduce) is suitable for massive data storage query, batch data ETL, and unstructured data analysis. MPP architecture is suitable to replace the big data processing under the existing relational data structure, in order to conduct multi-dimensional data analysis and data mart.

Under the hybrid structure, MPP processes structured data with high-quality, and provide SQL and transaction support, while Hadoop implements semi-structured and unstructured data processing. Through this hybrid approach, the demand for efficient processing of structured, semi-structured and unstructured data is automatically met, solving the difficulties of slow loading, low data query efficiency and difficulty in integrating multiple heterogeneous data sources for analysis under massive data of traditional data warehouses. This approach of breaking down the boundaries between data warehouses has become a mainstream architectural approach. However, in the process of lake warehouse integration, more emerging architectures are being developed and verified. There might be a new generation of architectures that will replace the MPP-Hadoop architecture to become a better architecture solution in the future.

Products: GaussDB(DWS), OushuDB, Dlink, Petabase, KCDE

MPP-Hadoop framework



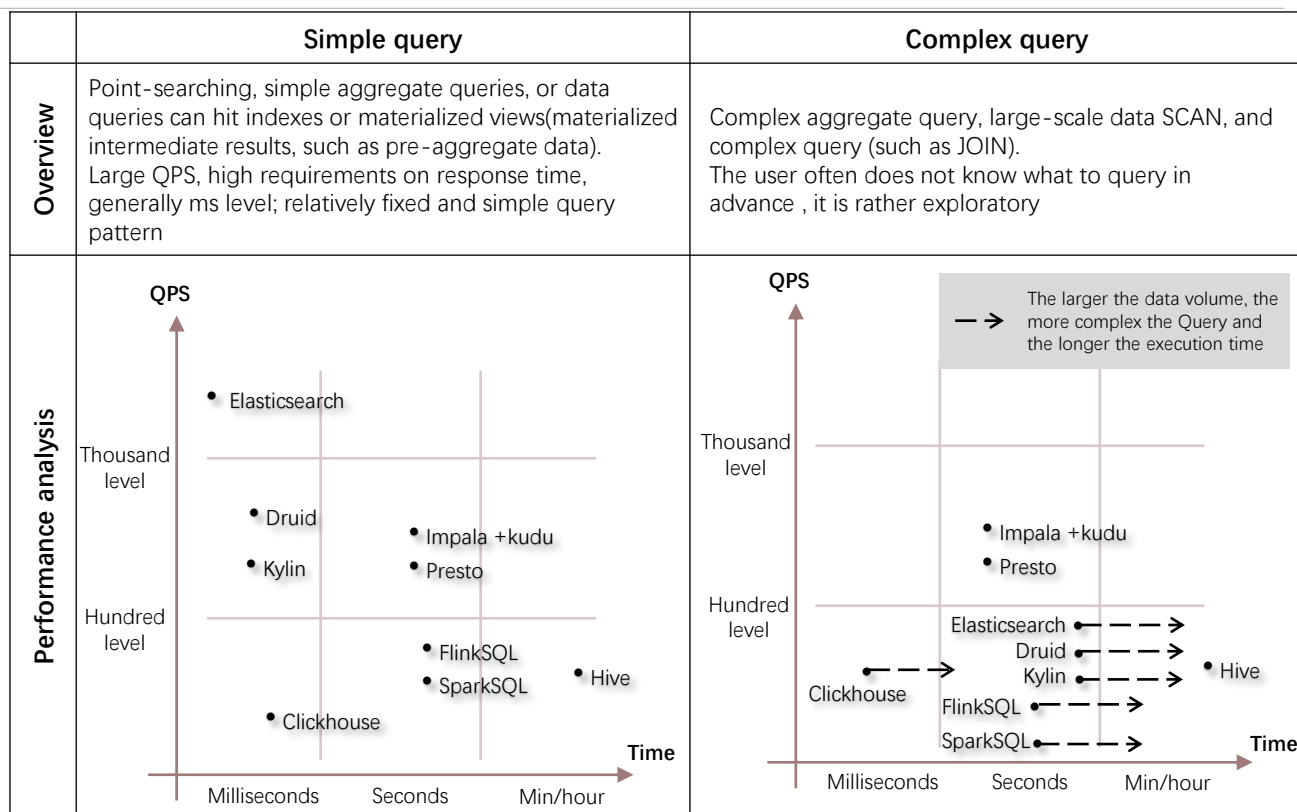
Source: CSDN, LeadLeo

Data Warehouse - Open Source Component Comparison

- The data warehouse can be classified according to the modeling mode or the architecture mode. According to real-time, Hadoop warehouse completes offline analysis through batch processing, and MPP data warehouse completes real-time analysis through stream processing

According to modeling mode, data warehouse can be divided into MOLAP, ROLAP and HOLAP. According to architecture mode, it can be divided into Hadoop and MPP. According to real-time, Hadoop warehouse completes offline analysis through batch processing, and MPP data warehouse completes real-time analysis through stream processing. For vendor selection, there are many open source OLAP engine components available to optimize data warehouse performance based on demand.

Comparison between simple query and complex query scenarios



Source: CSDN, LeadLeo

Open source OLAP engine performance comparison

	Hive	Impala	Presto	SparkSQL	HAWQ	Clickhouse	Greenplum
Associated query of multiple tables	1	5	4	3	4	3	3
Single table query	1	3	4	3	3	5	3
System load	4	2	2	2	2	2	2
Connected data source richness	1	3	5	3	3	1	1
Supported data formats	5	4	5	5	5	3	3
Standard SQL support	4	4	4	4	5	3	5
Ease of use of the system	5	5	5	4	3	5	5
Community activity	5	4	5	5	3	2	4
Customized function development cycle	5	4	5	4	4	1	4

Source: Analysys, LeadLeo

The scale is five, the higher the score, the better the performance

The evaluation results are from 2019, but the relative performance change is not significant, it still has reference value for manufacturer selection

Data lake architecture

- Data lake completes the integration of offline and real-time computing starting from Lambda architecture, and Kappa architecture unified data caliber to simplify data redundancy. The IOTA architecture further accelerates data lake efficiency by eliminating ETL through edge delivery and unified data model

Three architecture for real-time data processing of data lake

	Lambda architecture	Kappa architecture	IOTA architecture
Architecture			
Definition	<p>The goal is to design an architecture that meet the key characteristics of real-time big data systems including high fault tolerance, low latency, and scalability. Integrate offline and real-time computing, and a series of architectural principles like immutability, read-write separation, and complexity isolation.</p>	<p>In view of the shortcomings of Lambda architecture, such as the need to maintain two sets of programs, the core idea of Kappa architecture is to solve the problem of full data processing by improving the streaming computing system, so that real-time computing and batch processing use the same set of codes.</p>	<p>The standard data model is set, and all computing processes are dispersed in the process of data generation, calculation and query through edge computing technology, so as to improve the overall budget efficiency and meet the needs of real-time computing. Various impromptu queries can be used to query the underlying data.</p>
Advantages	<ul style="list-style-type: none"> • High fault tolerance and robustness • Low latency • Horizontal scalability • Easy to debug and maintain • Process large-scale historical data 	<ul style="list-style-type: none"> • Full calculation is carried out if necessary, small calculation cost • With only one framework, it is easy to develop, test, and maintain 	<ul style="list-style-type: none"> • De-ETL, starting from the edge of the calculation, improve the efficiency of the overall data analysis • Query events in the last few seconds without waiting for ETL or Streaming
Disadvantages	<ul style="list-style-type: none"> • Real-time and batch results are inconsistent and conflicting • Double computing + double service, the operation and maintenance of two systems is difficult • The intermediate data exists and the computation cost is high • T+1 Offline times out 	<ul style="list-style-type: none"> • Limited historical data processing capability and small batch processing throughput • The data format collected is not uniform and the development cycle of stream processor is long • New and old instance results need to be stored, which costs a lot • The universality of applicable scenarios is not high 	<ul style="list-style-type: none"> • The data buffer depends on Hbase and the scanning performance is slow
Products	<ul style="list-style-type: none"> • Cassandra+Spark+Pulsar • E-MapReduce、KDE 	<ul style="list-style-type: none"> • Flink+Iceberg、DataHub • DEEPEXI Dlink、DLF、KCDE 	<ul style="list-style-type: none"> • Yiguanmiaosuan、Inspur cloud IEMR

Source: CSDN, LeadLeo

Other data lake architectures include Omega architecture from OUSHU Technology, which consists of a stream processing system and a real-time data warehouse. It combines the advantages of Lambda and Kappa for processing streaming data, increasing the capability of real-time on-demand intelligence and offline on-demand intelligence data processing, as well as the ability to efficiently process real-time snapshots of changeable data.

Logical data lake

- Logical data lake can realize collaborative analysis and interactive query across lakes, warehouses, domains, clouds and business systems, which solves the problems of low performance and data copy caused by traditional scattered construction in collaborative analysis

Logical data lake

Logical data lake can realize collaborative analysis and interactive query across lakes, warehouses, domains, clouds and business systems, which solves the problems of low performance and data copy caused by traditional scattered construction in collaborative analysis.

Logical data lake and physical data lake

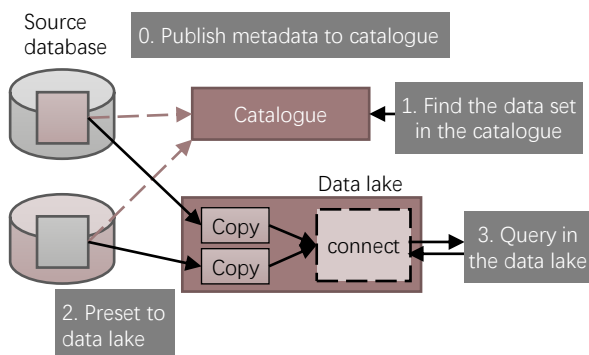
Compared with physical lake, which achieves the performance of storage and computation separation and independent expansion based on open source components (HUDi, Iceberg, Delta, etc.) +OSS, logical lake has less investment and is more suitable for enterprises with mature IP layer.

Although the technical threshold is high, the physical lake can form the core technology asset of the enterprise, with higher performance upper limit and more advantages in lightweight deployment.

Advantages of logical virtualization:

- Using data virtualization to transform physical data lake into more practical logical data lake can overcome the development difficulties of centralized data storage faced by traditional data lakes.
- Based on the high level of data virtualization technology structure, users can get the same experience as all data is centrally stored in a data repository.
- The development of different data lakes has different functional emphases. Through logical database, one can have the experience of multi-functional data lake within one data lake.
- Data virtualization simplifies the migration of data lakes to the cloud and makes cloud native data lakes transparent to most applications and reports.

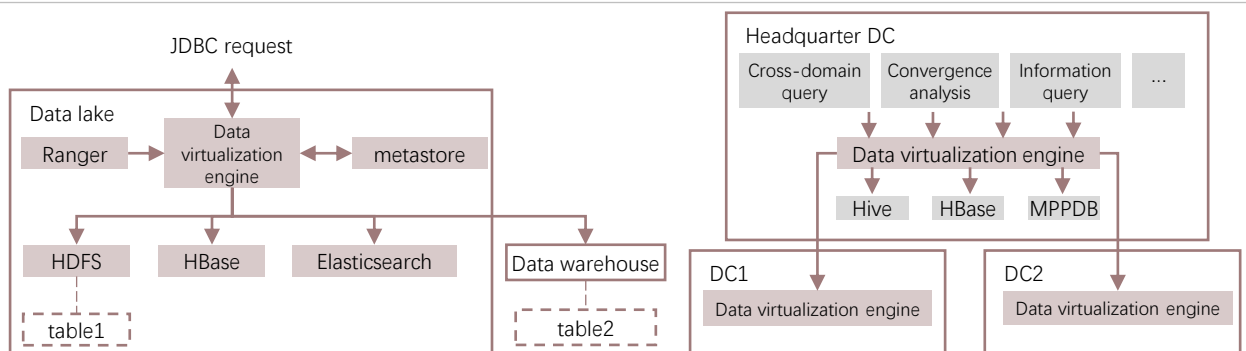
The principle of logical data lake



Source: O'REILLY, LeadLeo

Representatives of logical Data Lake manufacturers include: HetuEngine from Huawei MRS Cloud native Data Lake, Azure Data Lake Storage Gen2 (ADLS Gen2) from Microsoft, Artic from NetEase, etc.

Cross-source and cross-domain architecture advantages of logical data lake

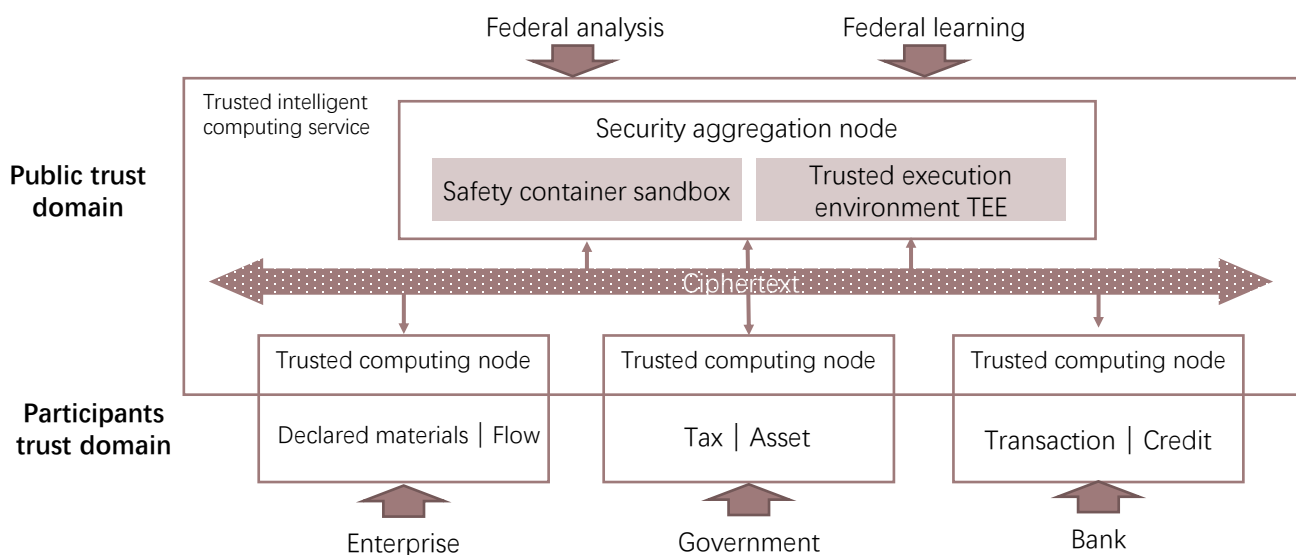


Source: Huawei Cloud HetuEngine, LeadLeo

Trusted intelligent computing

- One of the core objectives of trust is to ensure the integrity of the system and application, so as to determine that the system or software is running in the trusted state expected by the design objective. Trusted computing services enable the trusted flow and computation of data

Trusted computing



Source: Huawei Cloud, LeadLeo

□ Data security and data flow requirements

At present, the information system used by the government and enterprises generally exists the phenomenon of data isolation: due to the consideration of data protection, organizational management mechanism, information system design and other aspects, there are restrictions on data sharing and circulation between different departments or institutions.

With the introduction of "Data Security Law" and "Personal Information Protection Law", it highlights the importance of realizing the circulation of data elements under the premise of satisfying data security and data privacy .

□ Trusted computing

One of the core objectives of "trusted" is to ensure the integrity of the system and application, so as to determine that the system or software is running in the trusted state expected by the design objective. Trusted computing service can realize the trusted circulation and calculation of data, such as controlling the original detailed data in the trust domain of the party to which it belongs. At the same time, it realizes the federal calculation of multi-party data through mutual trust union, thus uniting the data scattered in different organizations and converting them into valuable information or models to realize the circulation of data across databases and nodes.

□ Trusted intelligent computing service

Trusted computing services are a set of theoretical frameworks and technical systems that require the integration of multiple technical domains.

Big data vendors and products that provide such services include TICS from Huawei Cloud, Nitro Enclaves from Amazon, C3S from Ali Cloud, CSPC from Tencent Cloud.

Data Lakehouse + machine learning

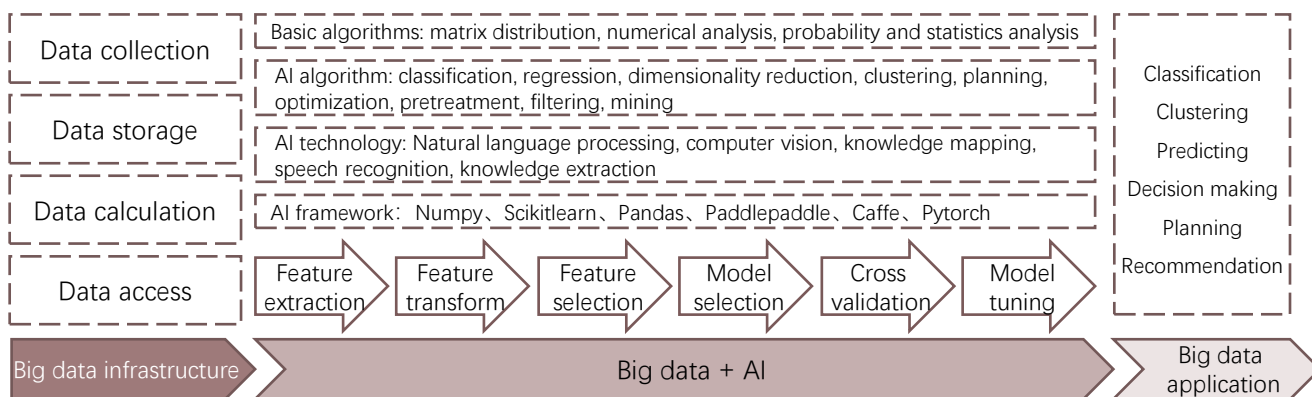
- With the popularity of data intelligence service awareness, it is especially critical for vendors to seamlessly integrate data analytics services with machine learning services to provide smarter and easier-to-use product services for users such as data developers and analysts who do not have an AI algorithm background

Data intelligence

Data base, data warehouse, data lake and lakehouse are data infrastructure. Data value can only be translated by using data analysis tools and driving decisions wisely. Artificial intelligence and machine learning capabilities are important features that give lakehouse the ability to innovate in its services.

Data intelligence is based on big data, processing, analyzing and mining massive amounts of data through AI. It extract information and knowledge from data, and seek solutions to existing problems and achieve predictions by building models to help decision-making.

Data intelligence concept



Source: CAAI, DataYuan, LeadLeo

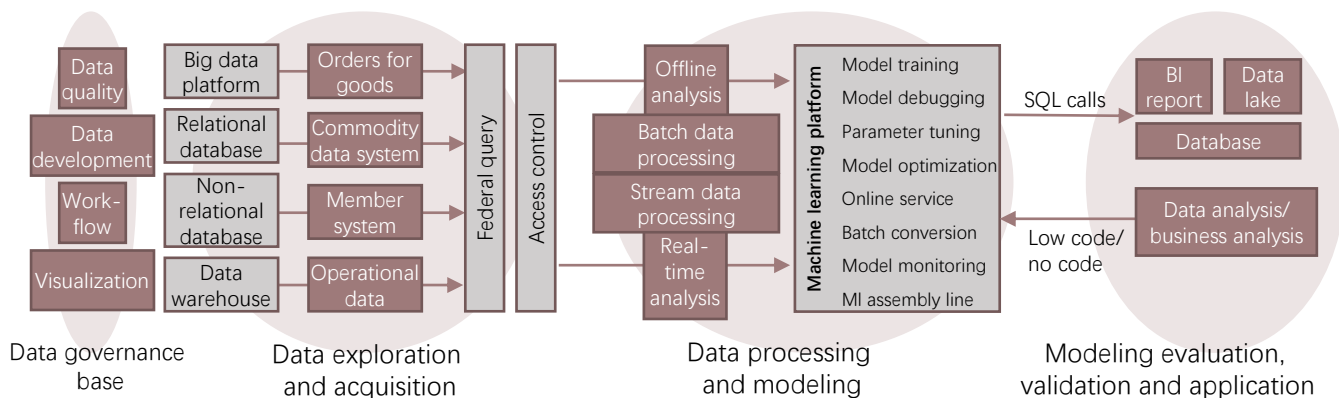
Big data + AI

In the past, BI was the main application scenario of data warehouse as statistical analysis computing, and AI analysis of predictive computing was the mainstream application of data lake. As lake warehouse integration matures, AI+BI dual mode will become an important load form of big data calculation and analysis.

With the development of big data, as well as the integration of offline and real-time processing, and data storage and data analysis, the breakthrough of performance bottleneck of big data system provides huge data service and application potential. Accordingly, with the popularity of data intelligence service awareness, it is especially critical for vendors to seamlessly integrate data analytics services with machine learning services to provide smarter and easier-to-use product services for users such as data developers and analysts who do not have an AI algorithm background, for example:

- Generality: Machine learning model inference can be carried out directly through SQL.
- Ease of use: provide simple tools to realize business, use existing data to realize machine learning model training.
- Transparency: visual data prepared for low-code data cleaning transformation.
- Intelligent O&M: AIOps capabilities applied to the daily operation and maintenance of data platforms.

Data intelligence integration process



Source: Amazon Web Services, LeadLeo

□ Deep integration between machine learning and big data platform

The speed of data processing and automation of the intergrated machine learning big data platform will increase by a generation.

In order to realize the integration of machine learning and big data, the following requirements should be met according to relevant papers:

1. Isolation mechanism: there should be no mutual interference between AI and big data.
2. Code seamlessly: native code that enables big data platforms to support machine learning.
3. Integrated framework: Data integrated engine would be introduced into data processing layer, enabling layer and application layer to deeply fuse data processing layer and enabling layer.

In order to improve the production efficiency of machine learning, the following requirements need to be met:

1. Full lifecycle platformization: it would cover end-to-end capabilities from data preparation, model building, model development to model production.
2. Preset machine learning algorithms and frameworks: users can use them directly without having to build them themselves ;
3. Quick resource startup: The system uses a unified computing cluster for underlying resources on demand.

Machine learning platform products: SageMaker from Amazon, ModelArts from Huawei Cloud, BML from Baidu Cloud, PAI from Alibaba Cloud, Ti-One from Tencent Cloud, Sophon from Transwarp, DataSense from Deepexi, ABI from Esensoft, LittleBoy from Oushu, KingAI from Kingsoft, etc.

Serverless lakehouse integration

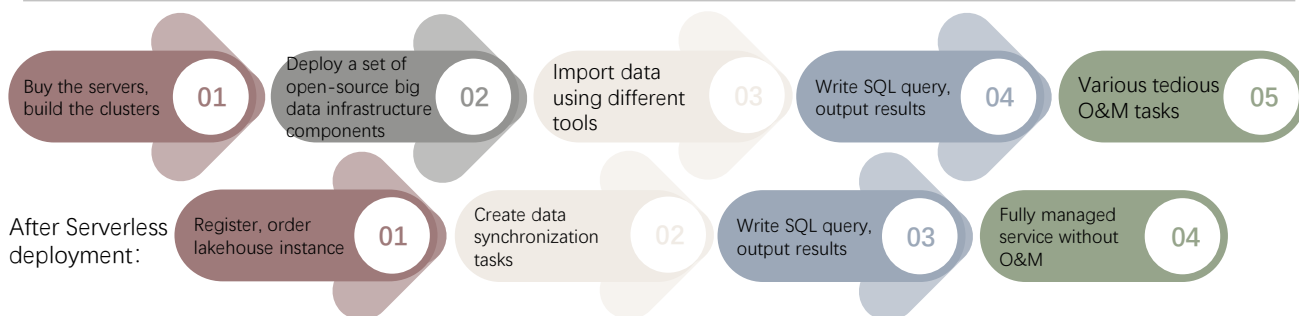
- Serverless lakehouse integration refers to data storage, data query engine, data warehouse, data processing framework, and data catalog products that all support serverless deployment

Serverless deployment

Serverless deployment provides services through FaaS+BaaS, allowing users to develop, run, and manage applications without building and operating a complex infrastructure.

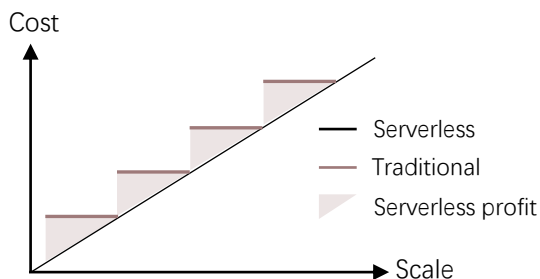
Serverless lakehouse integration refers to data storage, data query engine, data warehouse, data processing framework, and data catalog products that all support serverless deployment

Process of using lakehouse with or without Serverless deployment



Source: China Mobile Cloud Centre, LeadLeo

Serverless cost-saving advantages



Source: Huawei Cloud, LeadLeo

Advantages of Serverless Lakehouse

1. Simplified process of using: it provide users with more easy-to-use experience by adapting the Serverless Lakehouse architecture. Fully managed without O&M approach also helps users focus on the business itself, rather than technical logic, which is in line with the concept of cloud-native.
2. Cost Optimization: Serverless deployments can provide on-demand billing without the need to pay for waiting, allowing for more efficient resource utilization. It is more cost-effective for enterprises whose usage varies greatly over time.

Serverless Lakehouse architecture products

Amazon Cloud realizes Serverless Lakehouse through Redshift+EMR+MSK+Glue+Athena+Amazon Lake Formation with Serverless capability.

Huawei Cloud realizes Serverless deployed big data system through Stack+DLI Serverless+FusionInsight MRS+DWS.

DLA of Ali Cloud creates Maxcompute, an integrated architecture of cloud native+Serverless+database and big data, through core components Lakehouse, Serverless Spark and Serverless SQL.

Other Serverless Lakehouse products include Databricks Serverless SQL, Azure Synapse Analytics Serverless, Mobile Cloud Lakehouse, etc.

Summary of future development trends

- Data management solution vendors need to focus on user experience and continue to develop product technologies from dimensions like data warehouse, data lake, lake warehouse solutions, IaaS, etc.

Future trends of data management solutions



□ User experience is the key of lake warehouse integration

In the context of market users demanding higher flexibility for data warehouses and higher growth for data lakes, the concept of "lakehouse" is a common perception of future big data architectures among industry vendors and users.





Though it has significant advantages at the conceptual level, lakehouse still faces many problems in actual production due to the immaturity of technology or service. Potential users remain cautious due to concerns about user experience and stability, or uncertainty about the input and output value of replacing an existing mature and stable system.

Manufacturers need to focus on user experience and continue to deepen the product technology from multi-dimensional perspectives.

Data management user profiles

- Data management solution team includes four main functions: data analysis, data management, GRC, and business line. Among them, data analyst, data scientist, data management engineer and data development engineer are the main roles of data management solution services, which require different technology stacks

Classification and roles of data management team

 <p>Data analysis team</p>	<p>Data scientist: manage data, build model Data analyst: collect, process and perform statistical data analysis Data development engineer: transform data models into analytical applications Software engineer: embed the analyzer in the operating system</p>	 <p>Data management team</p>	<p>Data management engineer: Optimize data quality and prepare ETL operations Catalog data and perform metadata management Balance data protection and data privacy</p>
 <p>GRC team</p>	<p>Data governance expert: Establish data governance and security policies Ensure data privacy and security throughout the chain Compile requirements for retention, archiving and disposal, and ensure data compliance with policies and regulations</p>	 <p>Business line team</p>	<p>Business decision-making level includes: Chief Marketing Officer, Chief Financial Officer, Chief Human Resources Officer, Chief Data Officer Extract specific data analysis results and feasible decision opinions from the system</p>

Source: IBM, LeadLeo

Overview of technology stacks required for data management roles

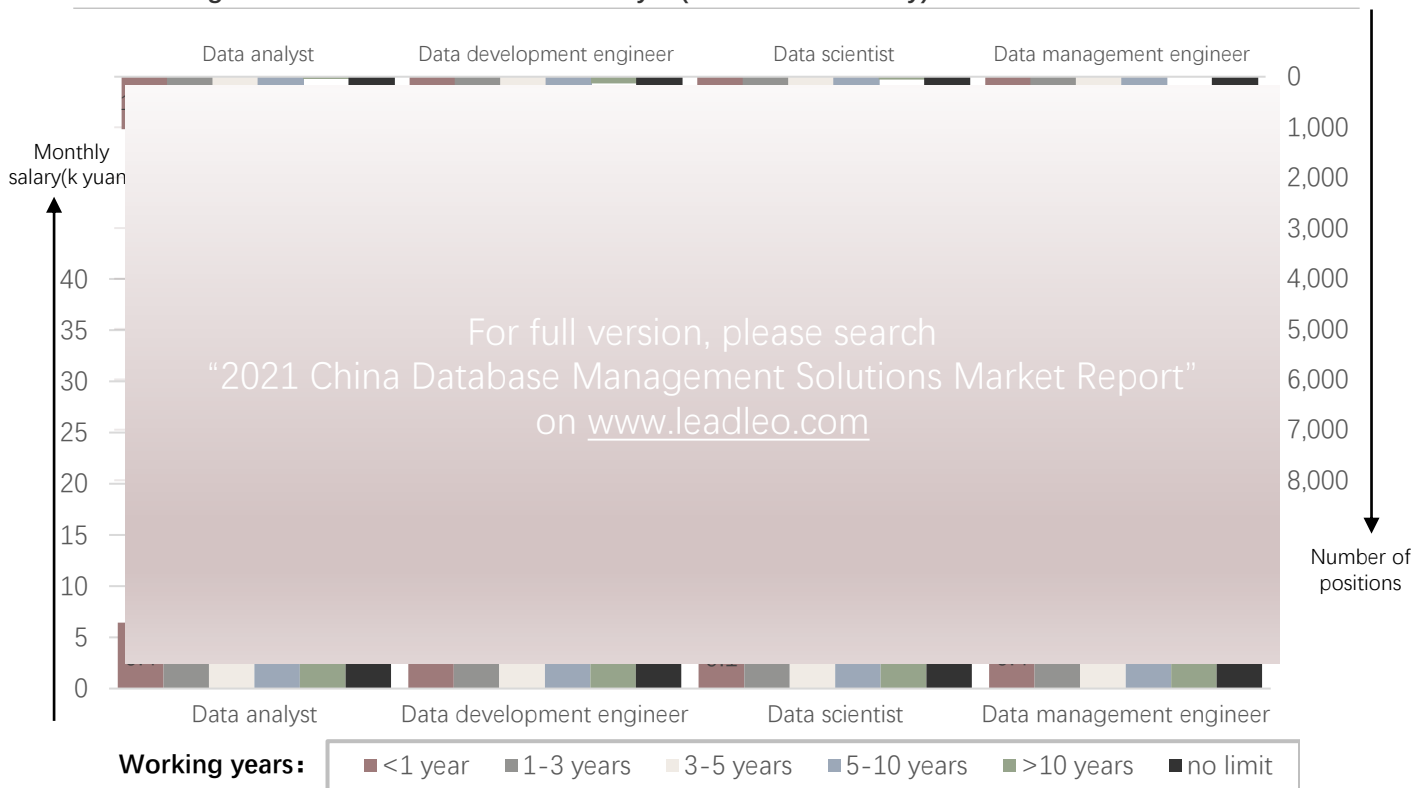
<p>Data analyst</p> <p>Axure Visio SEM</p> <p>Power Pandas SEO</p>	<p>SAS Stata R</p>	<p>SPSS Matlab Tableau</p>	<p>Excel Statistical Modeling</p>	<p>Event Tracking system Pytorch Tensorflow</p>	<p>Data scientist Research and analysis Algorithm design</p>
<p>Web crawler</p>	<p>Python Scala Shell Spark Oracle</p>	<p>Latex Data warehouse Hive Database Kafka</p>	<p>Data mining Machine learning ETL Java Hadoop Flink C++</p>	<p>Socket MapReduce SpringBoot SpringMVC Presto</p>	<p>Hbase Kettle HDFS SpringCloud React Streaming Spring</p>
<p>Django Flask GO</p> <p>Data management engineer</p>	<p>Zookeeper Server Yarn Sqoop Redis</p>	<p>PostgreSQL NoSQL DataX Informatica FineReport</p>	<p>C MongoDB Flume GIT</p>	<p>C# Dubbo Kylin Impala</p>	<p>Druid PHP js CSS3 Vue HTML JavaScript MVC Hibernate HTML5</p> <p>Data development engineer</p>

Source: Boss zhipin, Liepin, 51job, take first- and second-tier cities as sample cities, data analyst, data development engineer, data scientist and data management engineer as keyword. Retrieval time: 2022.05, hot word frequency analysis for technical stack requirements, disposed by Frost&Sullivan.

Data management related talent demand analysis

- The demand for professionals with 1-5 years of work experience is the highest in the talent market. Data analysts and data scientists have better average salary and salary increase. The demand structure for data management talents varies from industry to industry, with significant demand for data development engineers in IT and Internet industries, and significant demand for data analysts in retail and e-commerce industries

Data management related talent demand analysis(Position and Salary)



Number of positions for data management related talent in different industries

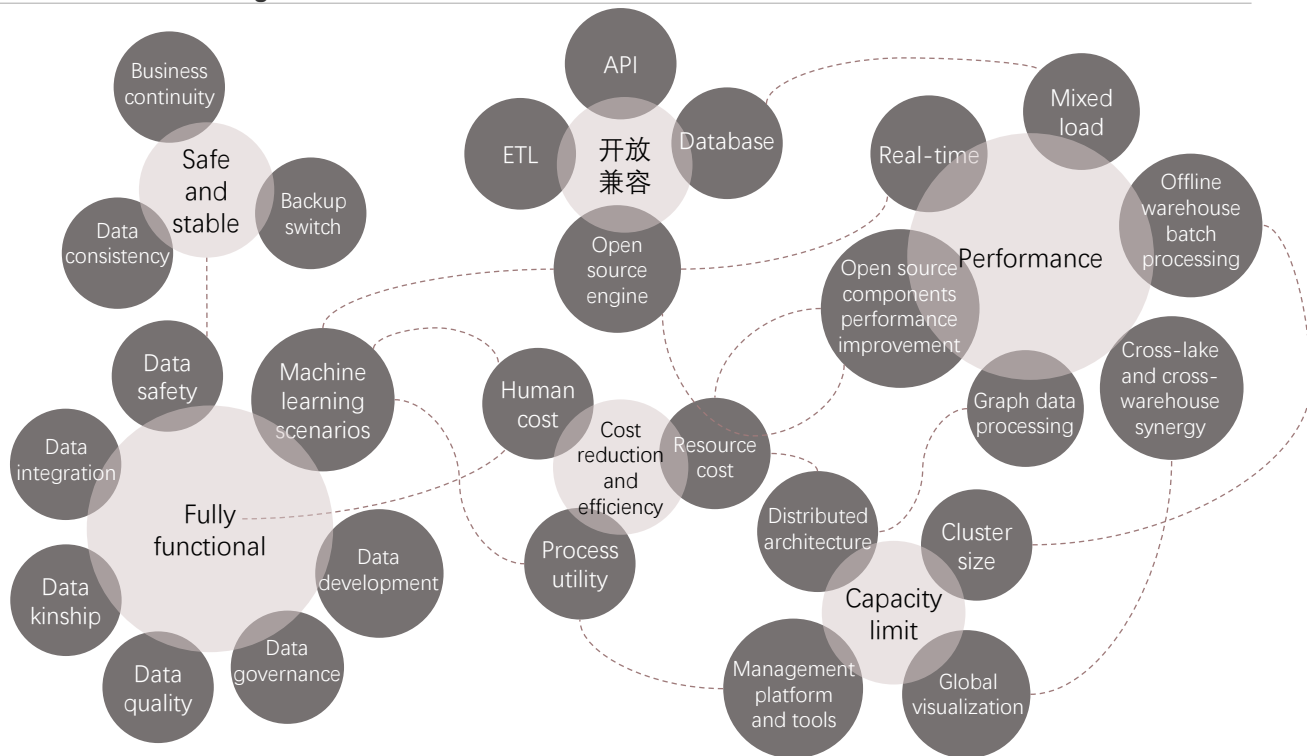


Source: Boss zhipin, Liepin, 51Job, take first- and second-tier cities as sample cities, data analyst, data development engineer, data scientist and data management engineer as keyword. Retrieval time: 2022.05, disposed by Frost&Sullivan.

Data management solution user needs

- Security and stability, full functionality, compatibility, cost reduction and efficiency, performance, and expansion limits are the six demand dimensions concerned by users of data management solutions. Machine learning scenarios, open source engine compatibility, and business continuity are the demand keywords emphasized by interviewed users

Overview of data management solution user needs dimensions



Data management user needs in different industries for different data service scenarios



Source: Frost&Sullivan, LeadLeo

Application scenario dimension-enterprise landscape

- Based on user experience, practical experience in the same application scenario is more representative than that in the same industry. By examining the breadth of industry field, granularity and depth of business capability of data management solution vendors in the scenarios, the ability of their products and services to meet vertical demand is analyzed and generates the map

Data service application scenarios atlas of data management solutions

Representative vendors in different data service application scenarios

Marketing management

- | | | |
|---|--|---|
| <ul style="list-style-type: none"> Customer portrait and customer label functions Multi-channel marketing services Real-time marketing services AI intelligent marketing services | <ul style="list-style-type: none"> 360° comprehensive customer data view Unmanned supermarket Precise advertising Automatic intelligent layout of website or APP pages | <ul style="list-style-type: none"> Customized digital marketing Digital marketing platform Data sharing Coupon utilization rate and repurchase rate promotion |
|---|--|---|

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

Risk management

- | | | |
|--|---|---|
| <ul style="list-style-type: none"> Introduce risk differentiation services Pre-screening from a risk perspective Early warning through risk cluster analysis Risk content identification | <ul style="list-style-type: none"> Anti-fraud, anti-money laundering Repayment risk assessment and credit assessment Enterprise risk assessment, risk control reasoning Regulatory submitting | <ul style="list-style-type: none"> Compliance log analysis, internal control compliance Base station risk management Equipment inspection and analysis Risk analysis of equipment and manufacturing process |
|--|---|---|

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

Customer operation

- | | | |
|--|---|---|
| <ul style="list-style-type: none"> Operational activities optimization Enterprise spectaculars, real-time spectaculars | <ul style="list-style-type: none"> Customer behavior log analysis Digital operation of event tracking library | <ul style="list-style-type: none"> Customer data platform Customer management information service |
|--|---|---|

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

Business analysis

- | | | |
|--|---|---|
| <ul style="list-style-type: none"> Store operation analysis Product quality management data lake | <ul style="list-style-type: none"> Consolidated management of subsidiary financial statements Profit and loss pre-query | <ul style="list-style-type: none"> Product atlas real-time indicator Real-time data index reporting |
|--|---|---|

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

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
Source: Frost&Sullivan, Leadleo

Data service application scenarios atlas of data management solutions (continued)

Representative vendors in different data service application scenarios

User portrait

- Analysis of user natural attribute data
- Multi-user association analysis
- Retail user metrics and profiles
- Consumer life cycle label portrait
- Population funnel analysis
- User trajectory analysis

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

Content understanding & recommendation

- Audio and video content distribution, advertising
- Personalized product reordering and customized direct selling
- Efficient recommendation model training for full data
- Text extraction and understanding, image recognition
- AI bill review
- Face recognition and analysis

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

Process optimization

- Inventory forecasting and analysis optimization
- Supply chain resource integration analysis and decision-making
- Channel management process optimization
- Distribution route planning and optimization
- Intelligent factory production line optimization analysis
- Online change control, automatic scheduling control

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

Data science

- Predictive analysis of protein structure data
- Antiviral drug development
- Big data analysis of ship voyage geography
- Advertising and streaming media efficiency analysis
- Optimization of energy delivery
- Intelligent allocation of stands

For full version, please search "2021 China Database Management Solutions Market Report" on www.leadleo.com

Circulation & integration

- All in one network
- All cards in one
- Internet of Things data platform

Other scenarios

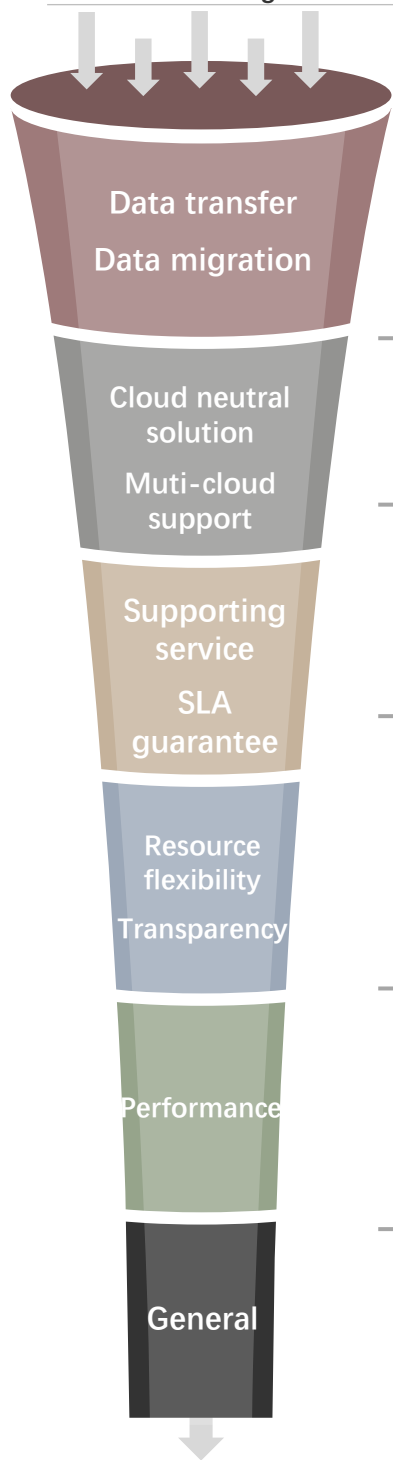
- Livelihood issues cause analysis
- Search engine
- Instant messaging data management
- Financial system data analysis
- Unmanned model training
- Stock trading history training

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
Source: Frost&Sullivan, Leadleo

Cloud data management solution selection essentials

- From an enterprise perspective, it is easy to fall into the trap of hidden costs and unmet needs without digging into the details of products and services, since products from different providers look similar. Solution selection needs to focus on pricing structure, multi-cloud deployment, artificial intelligence, universal adaptation and other dimensions to comprehensively judge the product and service solutions and quotations from different vendors

Cloud data management solution supplier funnel model
















- Most enterprises are implementing the cloudification solution of "producing data under the cloud and managing data on the cloud", which makes the data transmission between the cloud and the local IDC become the most common operational requirements, but also the most easily ignored part.
- Enterprises need to be aware that in the selection process the cloud data management supplier offers a lower price in the early phase of cloud deployment, but charges a higher fee in data transmission.
- In the long run, the cost of data transmission will be a huge cost to cloudification solutions, limiting on-cloud flexibility. At the same time, it hinders data migration and results in binding to a single cloud vendor. Therefore, the balance between deployment cost and transmission cost in pricing structure should be considered in the face of supplier's pricing scheme.
- Understand the supplier's migration tools and migration support services. Service downtime caused by service system switchover and deployment should not exceed 5 minutes.
- Find out if data management products support synchronization and operations across multiple clouds.
- Asking directly if the supplier's contribution to open source technology can also help reveal how cloud neutral that supplier's solution is, in order to avoid being locked in.
- Most cloud suppliers offer discounts to first-purchase and renewal customers. Know if the discounts apply to multi-cloud solutions and avoid the discount traps of single cloud binding.
- Many cloud data management solutions are fully managed services, meaning the solution supplier is responsible for administrative tasks such as deployment, updates, and maintenance. However, professional support services are usually value-added items that increase the total cost of ownership. Enterprises need to know the charging standards of value-added services such as on-site operation and maintenance, online fault response, safety assurance during critical periods, training, etc.
- The SLA guarantee allows the provider to compensate for the loss of availability in the event of a service interruption. The SLA requirement for the supplier should be at least 99.99%.
- It is necessary to expand computing and storage resources separately. Enterprise users can increase computing capacity based on peak demand and then reduce it to achieve more efficient usage and pricing.
- Whether the supplier can provide on-demand billing is also a key item. Compared with Round-the-clock service, the system can only calculate the cost when the system is actually running, analyzing or querying, which can save massive funds in idle resources and give full play to the elastic advantages of the cloud environment.
- Billing transparency cannot be ignored, and the enterprise should require detailed use case resources and duration for the supplier's monthly billing.
- POC testing should validate the performance or query speed claimed by the supplier for their product. Check that the test conditions are similar to those that exist in the data environment. If not, a more representative comparison should be sought.
- From workload management capabilities to concurrent scaling, suppliers should have a variety of solutions that handle high concurrency requirements in different ways without a significant drop in query speed or analysis performance.
- Is the provider developing AIOps, and using machine learning to help the query understand which path to take for the solution.
- To help data scientists and developers get started immediately and not waste time learning proprietary code, your data management solution must support popular data science and machine learning languages, such as Python, Go, Ruby, PHP, Java, Node.js, Sequelize, and Jupyter Notebook.
- Based on a common code base, data virtualization, or product architecture, enterprise users should have easy access to data deployed locally and in the cloud, whether from the same vendor, competing vendors, or open source solutions.

Source: IBM, LeadLeo

Data Management Solutions Product and Vendor Atlas

- Data management solutions vendors are divided into cloud vendor, operator cloud, big data vendor and open source, and the corresponding data warehouse, data lake and data lakehouse of each vendor are also listed

Overseas data management solutions vendors and representative products

Category	Vendor	Data warehouse	Data lake	Data lakehouse
Cloud vendors	 amazon webservices™	Amazon Redshift	S3+Lake Formation	AWS Intelligent Lakehouse, Redshift Spectrum
	 Microsoft Azure	Azure Synapse Analytics	Datalake Analytics	Azure Synapse Analytics
	 Google Cloud Platform	Google BigQuery Mesa	✓	Dataplex
	 IBM	DB2 Warehouse Netezza	Spectrum Scale IBM DataStage	Cloud Pak
	 EMC² where information lives®	EMC GreenPlum	EMC Cloudpool Scale out NAS Isilon	-
Big data vendors	 SAP SAP Commerce Cloud	SAP Data Warehouse Cloud	SAP HANA Cloud	-
	 TERADATA.	Teradata AsterData	Teradata Vantage	Teradata Vantage
	 databricks	-	Delta Lake (open source)	Lakehouse Platform
	 snowflake	Data Cloud	-	✓
	 CLOUDERA	CDH	✓	Cloudera Data Platform (CDP)
	 ORACLE®	Autonomous Data Warehouse	Oracle Data Flow	OCI
Open source	 The Apache Software Foundation http://www.apache.org/	Hive Hadoop	Hudi Iceberg	-
	 Greenplum	Greenplum DW	-	-

Source: Enterprise websites, LeadLeo

Domestic data management solutions vendors and representative products

Category	Vendor	Data warehouse	Data lake	Data lakehouse
Cloud vendors	 HUAWEI	GaussDB(DWS)	MRS、DGC	FusionInsight
	 阿里云	AnalyticDB Hologres	DLF、DLA	Maxcompute
	 金山云	KDW、KDC	KS3、KQES、KDC、 KDE、KMR	KCDE
	 腾讯云	CDW (PG、Clickhouse、 Doris)	EMR、DLC、DLF	Cloud-native intelligent data lake
	 百度智能云	Palo Doris (open source)	EasyDAP	Cloud-native lakehouse architecture
	 京东云	JDW DCS	✓ (+Delta)	JMR_BD
	 浪潮云	DW+(Greenplum/Udpg)	IDLF	Big data storage and analysis IEMR
Operator cloud	 移动云	DWS	DLI、DGC	Cloud-native big data analysis LakeHouse
	 天翼云	DWS	Data Lake Insight	-
Big data vertical vendors	 火山引擎	ByteHouse	EMR	LAS
	 星环科技	Inceptor ArgoDB	Inceptor TDC	TDH
	 SequoiaDB 巨杉数据库	-	-	SequoiaDB - DP
	 OUSHU 偶数	Oushu Database	-	Oushu Data Cloud
	 滴普科技 DEEPEXI	Dlink	Dlink	FastData
	 TAPDATA	-	-	Tapdata Enterprise
	 ESENTOFT 亿信华辰	Ensensoft ABI	-	“Ruizhi” data governance platform
	 GBASE®	GBase GCDW	-	GBase 8a mpp cluster
	 网易数帆	-	Arctic	-
	 HashData	HashData	-	-

Source: Enterprise websites, LeadLeo

Research Director

Livia Li

☎ 13149946576

✉ livia.li@frostchina.com

Principal Analyst

Jackey Hu

☎ 18576027961

✉ jackey.hu@frostchina.com

🌐 www.frostchina.com ; www.leadleo.com

📺 <https://space.bilibili.com/647223552>

📱 <https://weibo.com/u/7303360042>

©Frost & Sullivan (China)

©Leadleo Research Institute

