FROST & SULLIVAN 沙利文 头豹 LeadLeo

2021
China Data Management Solutions Market Report

2021年中国数据管理解决方案市场报告
2021年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

LeadLeo Research Institute
Frost & Sullivan (China)

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, Internet, retail, entertainment, telecommunications, energy, logistics, transportation, manufacturing, energy, medical, 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.

All figures, tables and text in this report are based on the surveys from Frost & Sullivan China and LeadLeo Research Institute. All data are rounded to one decimal place.

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.

Contents

- ◆ Data Management Solution Technology Trends ----- 7
  - Iterative changes in big data technology
  - Lake warehouse integration
  - Data Warehouse - OLAP Analysis Engine
  - Data Warehouse - Execution Model and Architecture
  - Data Warehouse - Open Source Component Comparison
  - Data lake architecture
  - Logical data lake
  - Data Lakehouse + machine learning
  - Trusted intelligent computing
  - Serverless lakehouse integration
  - Summary of future development trends
- ◆ Data Management Solutions Market Analysis ----- 23
  - Data management user profiles
  - Data management related talent demand analysis
  - Data management solution user needs
  - Application scenario dimension-enterprise Landscape
  - Cloud data management solution selection essentials
  - Data Management Solutions Product and Vendor Atlas
- ◆ Competitive landscape in China's data management solutions market ----- 32
  - Assessment Scoring
  - Comprehensive Vendors Assessment – Frost Radar
  - Leading Competitors
- ◆ Terms ----- 41
- ◆ Methodology ----- 42
- ◆ Legal Statement ----- 43

Figures & Charts

- Iterative changes in big data technology 07
- Classification of the technological evolution of data management platforms 08
- Data Warehouse, Data Lake and Data Lakehouse 09
- Implementation path of data lakehouse 10
- Load characteristics of database and data warehouse 11
- Data warehouse building process 11
- Different implementations of OLAP engines 12
- Three different execution architectures of data warehouse 13
- Comparative analysis of execution architecture 14
- MPP-Hadoop framework 14
- Comparison between simple query and complex query scenarios 15
- Open source OLAP engine performance comparison 15
- Three architecture for real-time data processing of data lake 16
- The principle of logical data lake 17
- Cross-source and cross-domain architecture advantages of logical data lake 17
- Trusted computing 18
- Data intelligence concept 19
- Data intelligence integration process 20
- Process of using lakehouse with or without Serverless deployment 21
- Serverless cost-saving advantages 21
- Future trends of data management solutions 22
- Classification and roles of data management team 24
- Overview of technology stacks required for data management roles 24
- Data management related talent demand analysis(Position and Salary) 25

Figures & Charts

- Number of positions for data management related talent in different industries 25
- Overview of data management solution user needs dimensions 27
- Data management user needs in different industries for different data service scenarios 27
- Data service application scenarios atlas of data management solutions 28
- Cloud data management solution supplier funnel model 30
- Overseas data management solutions vendors and representative products 31
- Chinadata management solutions vendors and representative products 32
- Innovation Index Evaluation System Indicators 33
- Growth Index Assessment System Indicators 34
- Comprehensive Competitive Performance of China' s Data Management Solutions Market - Frost Radar 35
- Amazon Web Services Lake House Architecture Upgrades 36
- Huawei Cloud FusionInsight Lakehouse Architecture Overview 38
- Alibaba Cloud Maxcompute Lakehouse Architecture Overview 40

01 Data Management Solution Technology Trends

- ❑ Iterative changes in big data technology
- ❑ Lake warehouse integration
- ❑ Data Warehouse - OLAP Analysis Engine
- ❑ Data Warehouse - Execution Model and Architecture
- ❑ Data Warehouse - Open Source Component Comparison
- ❑ Data lake architecture
- ❑ Logical data lake
- ❑ Data Lakehouse + machine learning
- ❑ Trusted intelligent computing
- ❑ Serverless lakehouse integration
- ❑ Summary of future development trends

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

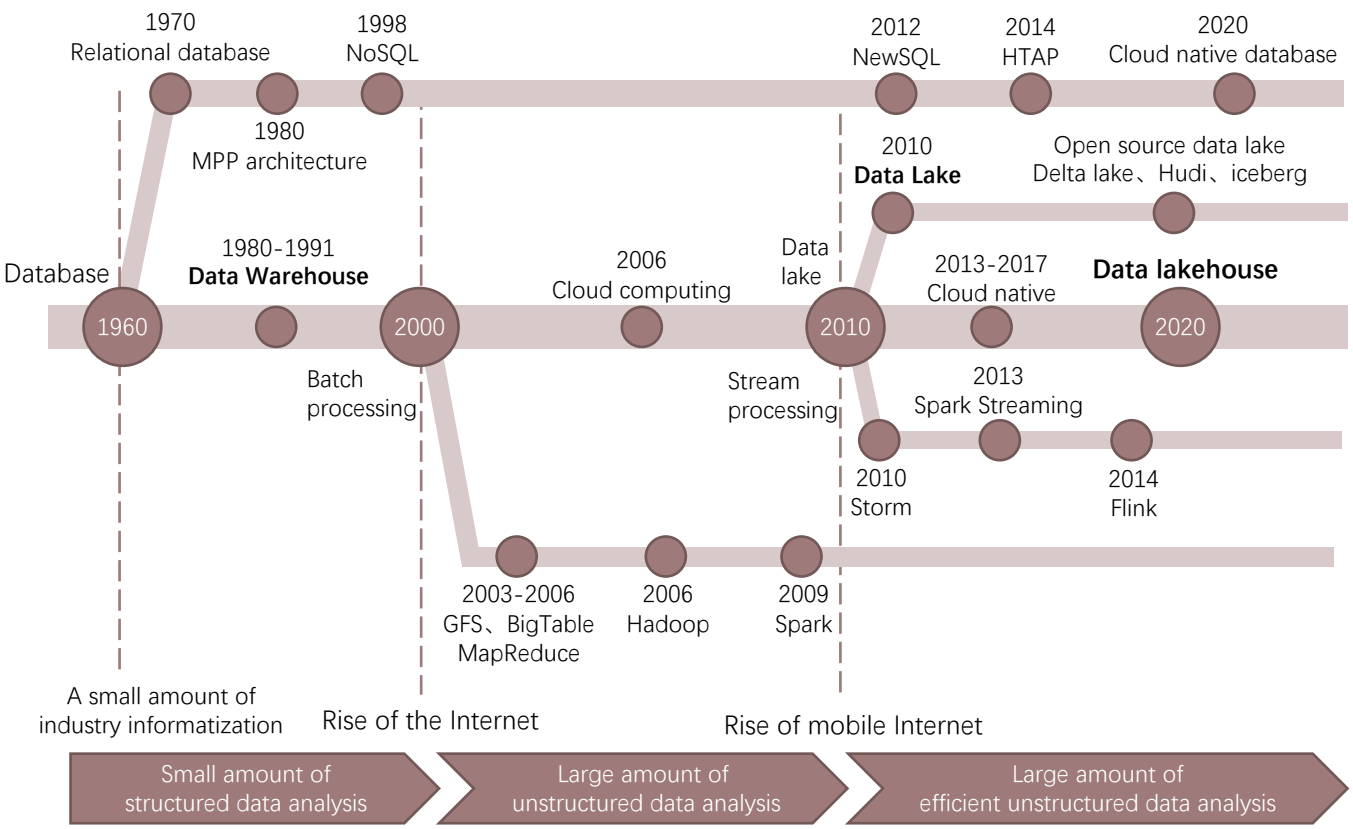
Time Line	Phase	Development Process	Product characteristics
1960s - 2000s	Database Age	In 1979, Oracle released the first commercial SQL relational database, DBMS was born for the needs of data processing and management. In 1988, the prototype concept of data warehouse was proposed by IBM. A batch of business intelligence tools like expert system and knowledge management system started to be applied.	System operation data based; small data scale, low data complexity, offline
2000s - 2010s	Early stage of big data	In 2003, The Google Troika (storage GFS, computing MapReduce, and online service BigTable) was released, laying an important foundation for the development of cloud computing and big data technology. In 2005, the Hadoop Ecosystem project was born to provide comprehensive functionality and flexible big data analytics.	User input data based; big data products with their own data processing engines, relatively independent
2010s - 2020s	Development stage of big data	Spark, Flink, Presto, Kafka and other engine technologies combined with The Hadoop system began to occupy the open source collaboration and integration ecosystem, and the embryonic form of data lake emerged.	Sensor data added; big data products based on open source technology stack emerged, integration started
2021-	Application Deepening Stage of big data	Amazon Wev Services, Alibaba Cloud, IBM, Baidu AI Cloud, Huawei Cloud, Tencent cloud have launched complete data lake products. Under the cloud system, the integration solution of data warehouse and data lake is applied.AI and big data are combined with a new generation of data intelligence applications.	Data value extraction focused; decentralized engine technology stack continues to converge and extend to enterprise-level requirements such as high security and high availability

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

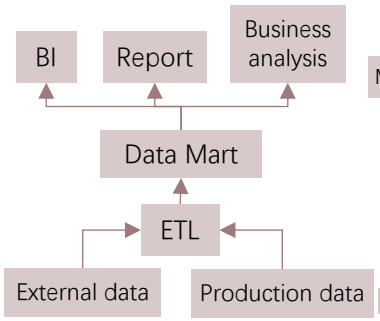
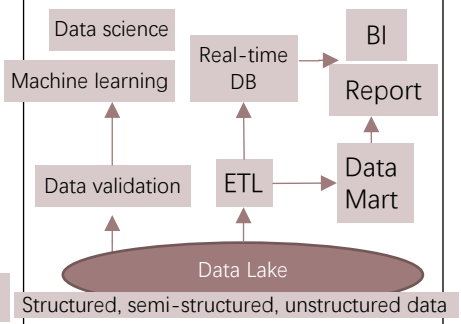
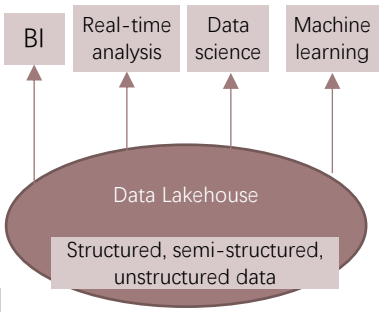
Lake warehouse integration trend

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	A data warehouse is a topic-oriented, integrated, relatively stable collection of data that reflects historical change and is used to support management.	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.	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.
Characteristics	• 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	• Unified storage system • Storing raw data • Rich computing model/paradigm • Not relevant to cloudification	• 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	• 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	• 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	• 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	• Unable to process unstructured data • Poor expandability • Long path from data source to data warehouse, low concurrency, creates data islands	• Shortcomings in the technical processing of transaction consistency • Challenge to data governance	• Imperfect load management function of warehouse service

Source: LeadLeo

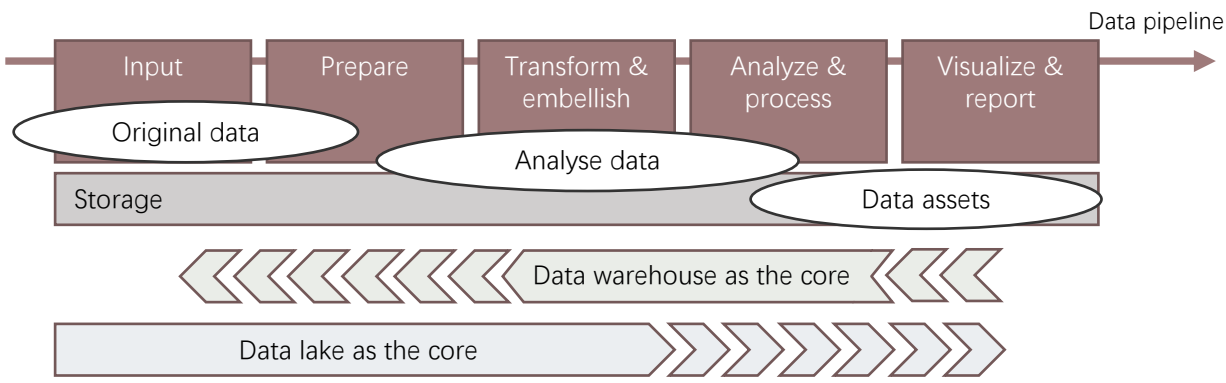
❑ Implementation approach of data lakehouse

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

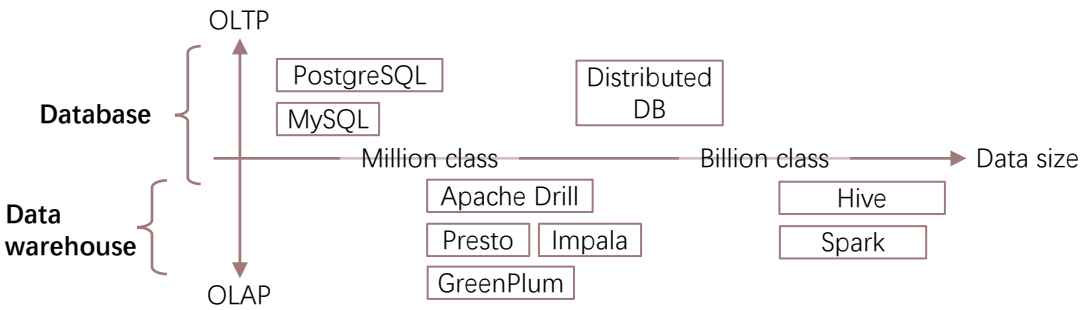
	Warehouse on lake	Warehouse to lake
overview	Data lake architecture based on public cloud, or open source Hadoop ecological components DeltaLake, Hudi, Iceberg as the middle layer of data storage to realize the unified storage of heterogeneous data of multiple sources, integrate computing engine with 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: Tietoevry, 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

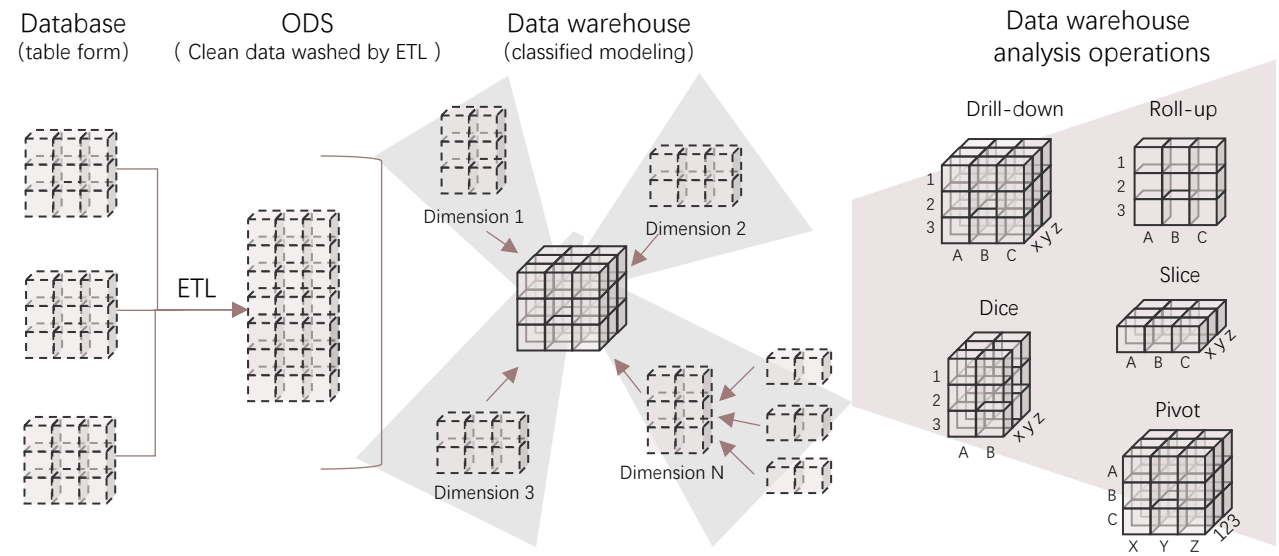
Database and data warehouse

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.

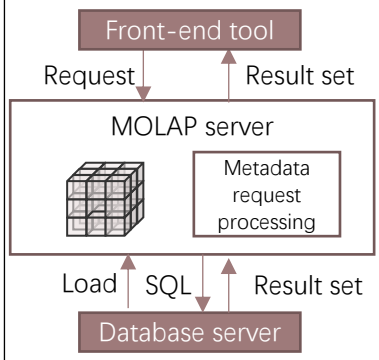
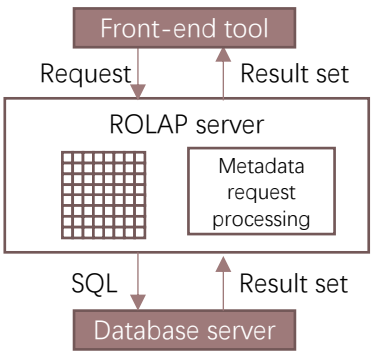
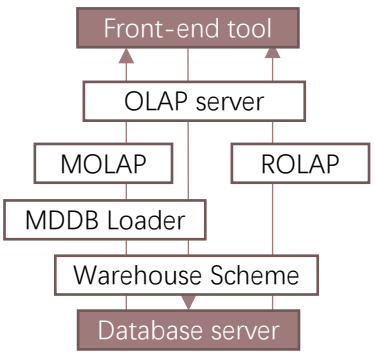
OLAP and OLTP are merging into HTAP. The enhancement of AP analysis capability by databases will gradually blur the boundary between databases and data warehouses.

Data warehouse building process



Source: CSDN, LeadLeo

Different implementations of OLAP engines

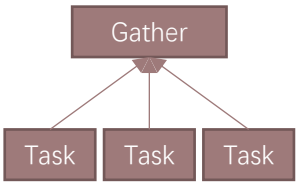
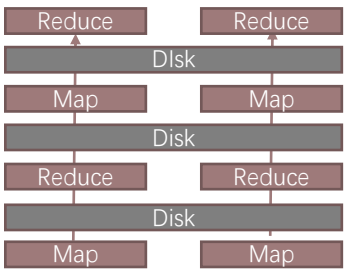
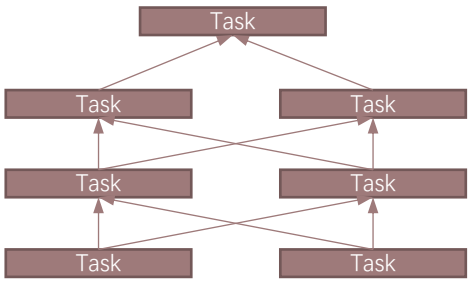
	Multidimensional OLAP (MOLAP)	Relational OLAP (ROLAP)	Hybrid OLAP (HOLAP)
Architecture			
Definition	Based on native logical models that directly support multidimensional data and operations. Data is physically stored in multidimensional arrays and accessed using location techniques.	Store multidimensional 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.	Bridge the technical gap between the two products by allowing the use of both multidimensional databases (MDDB) and relational databases (RDBMS) as data stores.
Characteristics	• Achieve from physical level • Data is pre-computed and stored • Storage designed and optimized for OLAP • Multidimensional indexing and caching are supported	• Do not use pre-computed cubes • No redundant data is imported • Use the existing relational database technology	• Provides fast access to all aggregation levels • The OLAP server only stores aggregation information, and the detailed records are retained in relational database
Advantages	• Provide fast access to all aggregation levels • The OLAP server only stores aggregation information, and the detailed records are retained in a relational database	• Easy to manage • Small storage space consumption, no dimensional limit • Queries can be implemented through SQL	• Duplicate copies of detailed records are not kept, balancing disk space requirements • Optimize query performance under given usage scenarios
Disadvantages	• Pre-computation is resource-consuming, dimension limited and inflexible • Low data loading speed • Lack of standard data access interface • Difficulties in maintenance	• Slow response • Depend on the database to perform calculations, proprietary capabilities are limited	• Supports both MOLAP and ROLAP, complex architecture • Lack of flexibility
Products	• Druid、Kylin、Doris • ESENSOFT ABI	• Amazon Redshift、Dlink、GaussDB(DWS)、OushuDB、KDW • Presto、Impala、GreenPlum、Clickhouse、Elasticsearch、Hive、Spark SQL、Flink SQL	• Kylin、Hulu Sophon • Inspur cloud IEMR
Scenarios	• Fixed query scenarios that require high query performance: • Advertising report analysis	• Scenarios with variable query modes and high query flexibility requirements: • Analysis products commonly used by data analysts	• 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	• Single node aggregation • Equal to a Map and Reduce trip in MapReduce	• 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	• Shared Nothing architecture • Distributed parallel execution • Distributed storage of data (localization) • Transverse linear extension
Advantages	• Maximize performance benefits from local I/O	• 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	• Emphasis on real-time data calculation, greater I/O capability • Column storage is used to save storage space • Ease of use and scalability
Disadvantages	• Operations such as large table Join and high cardinality aggregation cannot be completed	• 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	• 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	• Elasticsearch、Druid、Kylin	• Hive、Spark SQL、Hadoop • IEMR、Inceptor、KMR • KDC	• 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

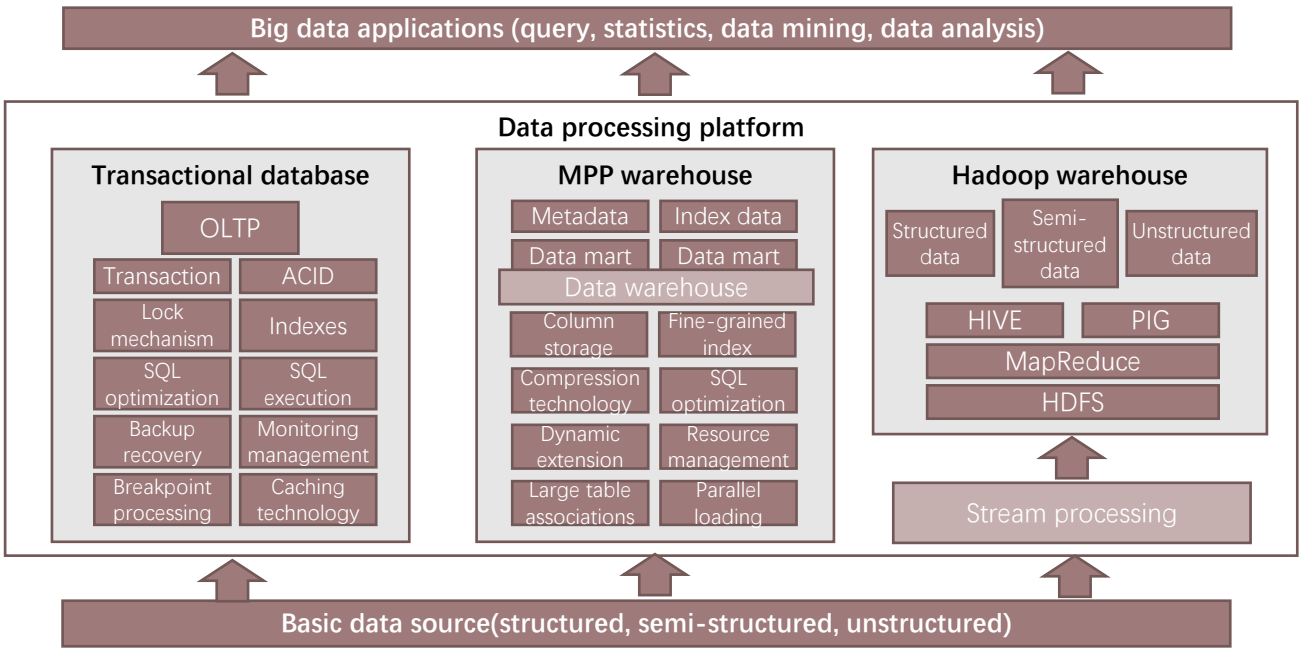
❑ **MPP-Hadoop architecture**

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



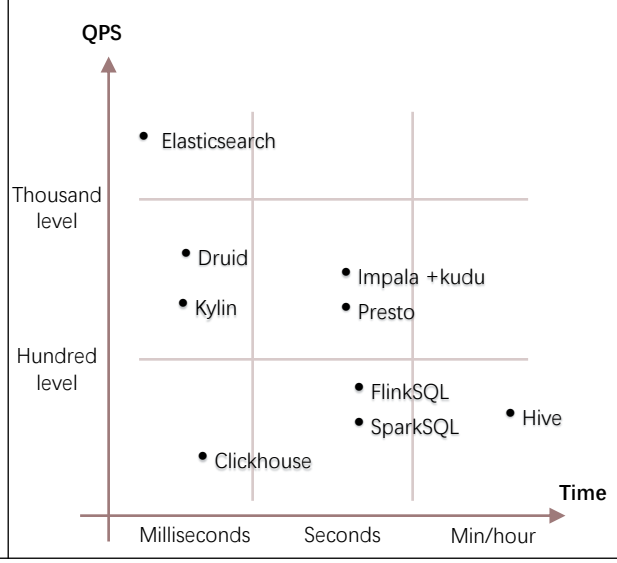
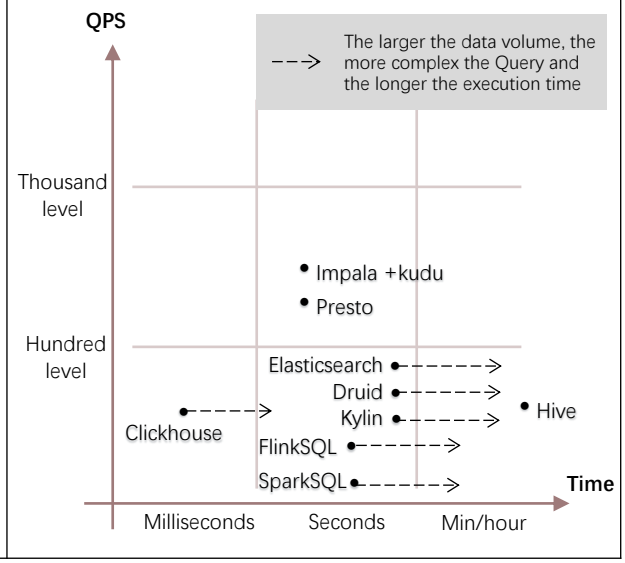
Source: CSDN, LeadLeo

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

	Simple query	Complex query
Overview	Point-searching, simple aggregate queries, or data queries can hit indexes or materialized views(materialized intermediate results, such as pre-aggregate data). Large QPS, high requirements on response time, generally ms level; relatively fixed and simple query pattern	Complex aggregate query, large-scale data SCAN, and complex query (such as JOIN). The user often does not know what to query in advance , it is rather exploratory
Performance analysis		

Source: CSDN, LeadLeo

Open source OLAP engine performance comparison

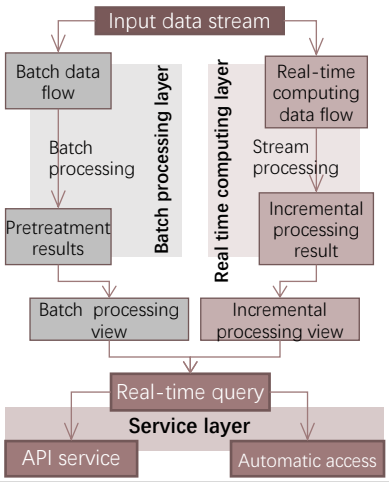
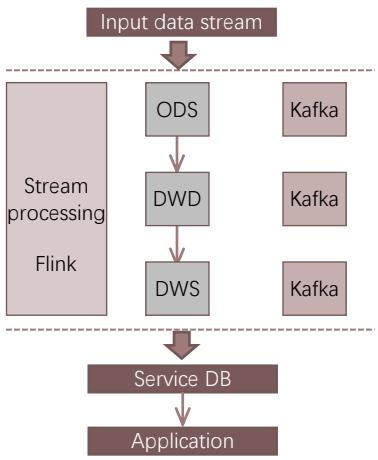
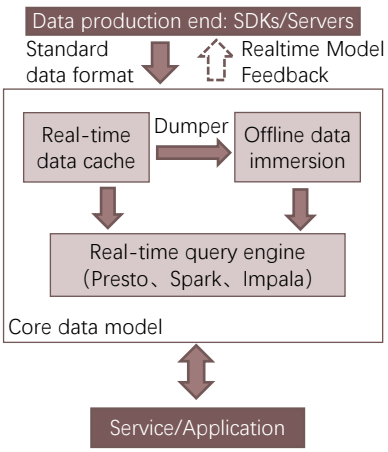
	Hive	Impala	Presto	SparkSQL	HAWQ	Clickhouse	Greenplum
Associated query of multiple tables	1	5	4	3	4	3	3
Single table query	1	3	4	3	3	5	3
System load	4	2	2	2	2	2	2
Connected data source richness	1	3	5	3	3	1	1
Supported data formats	5	4	5	5	5	3	3
Standard SQL support	4	4	4	4	5	3	5
Ease of use of the system	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	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.	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.	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.
Advantages	• High fault tolerance and robustness • Low latency • Horizontal scalability • Easy to debug and maintain • Process large-scale historical data	• Full calculation is carried out if necessary, small calculation cost • With only one framework, it is easy to develop, test, and maintain	• 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	• 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	• 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	• The data buffer depends on Hbase and the scanning performance is slow
Products	• Cassandra+Spark+Pulsar • E-MapReduce、KDE	• Flink+Iceberg、DataHub • DEEPEXI Dlink、DLF、KCDE	• 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.

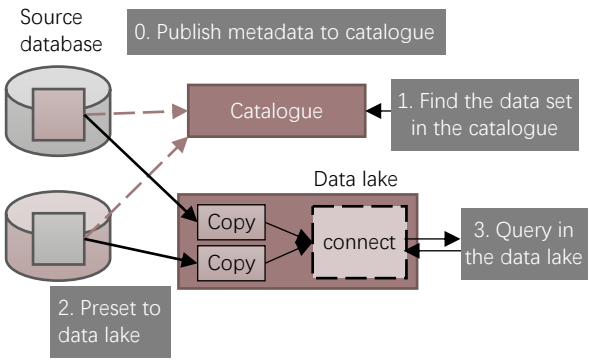
Although the technical threshold is high, the physical lake can form the core technology asset of the enterprise, with higher performance upper limit and more advantages in lightweight deployment.

❑ Advantages of logical virtualization:

1. 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.
2. 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.
3. The development of different data lakes has different functional emphases. Through logical database, one can have the experience of muti-functional data lake within one data lake.
4. Data virtualization simplifies the migration of data lakes to the cloud and makes cloud native data lakes transparent to most applications and reports.

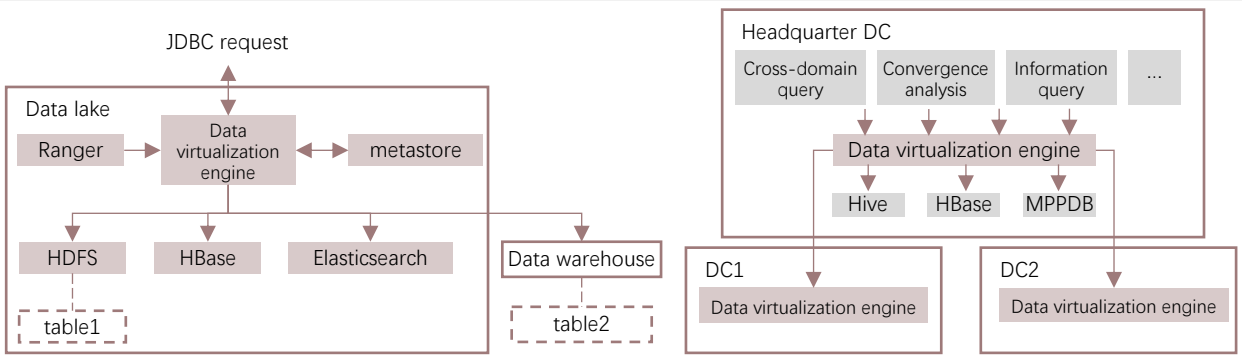
Representatives of logical Data Lake manufacturers include： HetuEngine from Huawei MRS Cloud native Data Lake, Azure Data Lake Storage Gen2 (ADLS Gen2) from Microsoft, Artic from NetEase, etc.

The principle of logical data lake



Source: O'REILLY, LeadLeo

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

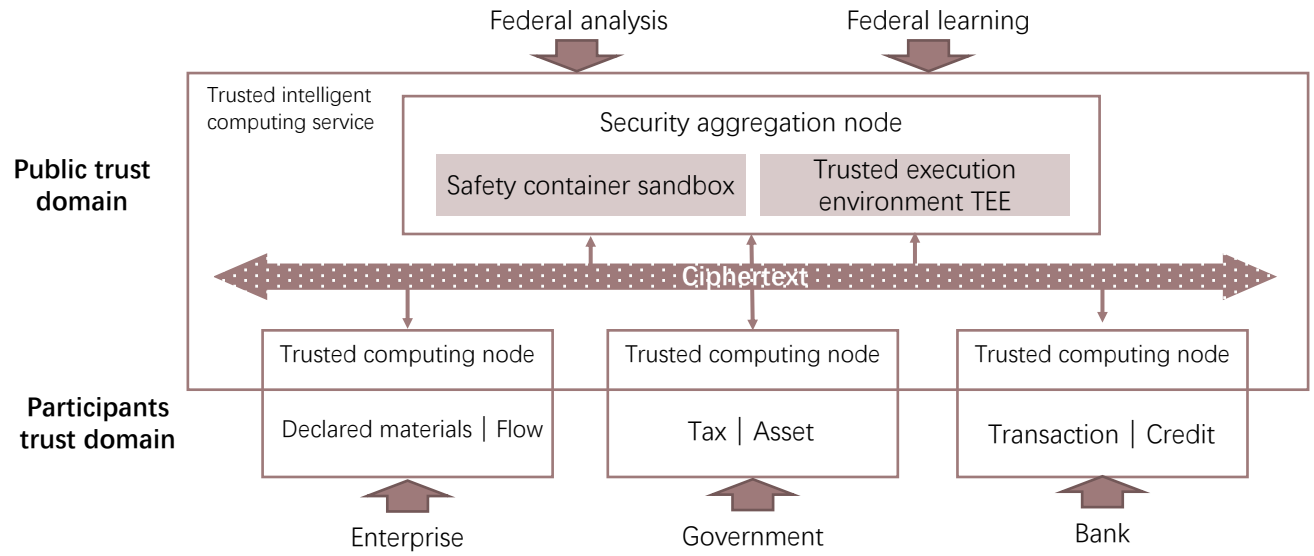


Source: Huawei Cloud HetuEngine, LeadLeo

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

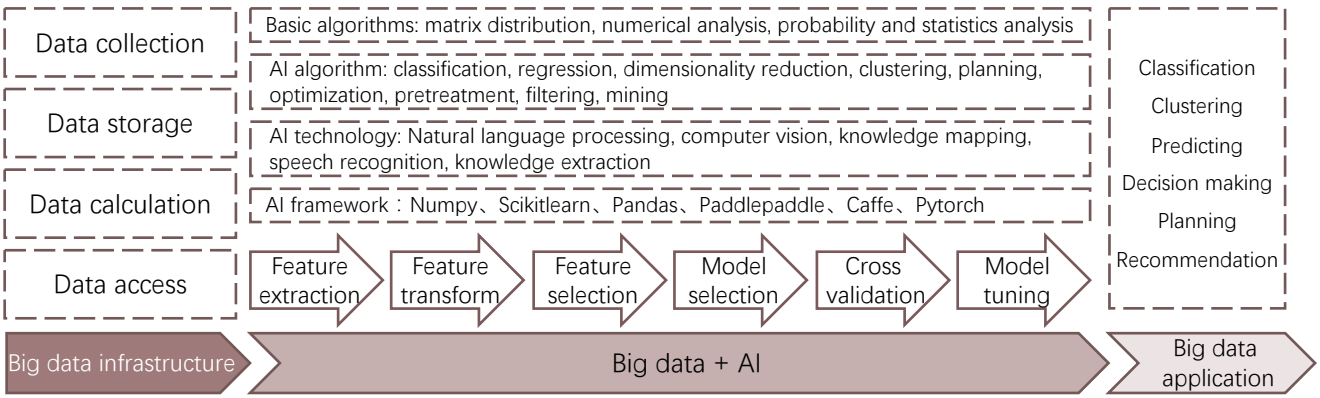
- **With the popularity of data intelligence service awareness, it is especially critical for vendors to seamlessly integrate data analytics services with machine learning services to provide smarter and easier-to-use product services for users such as data developers and analysts who do not have an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

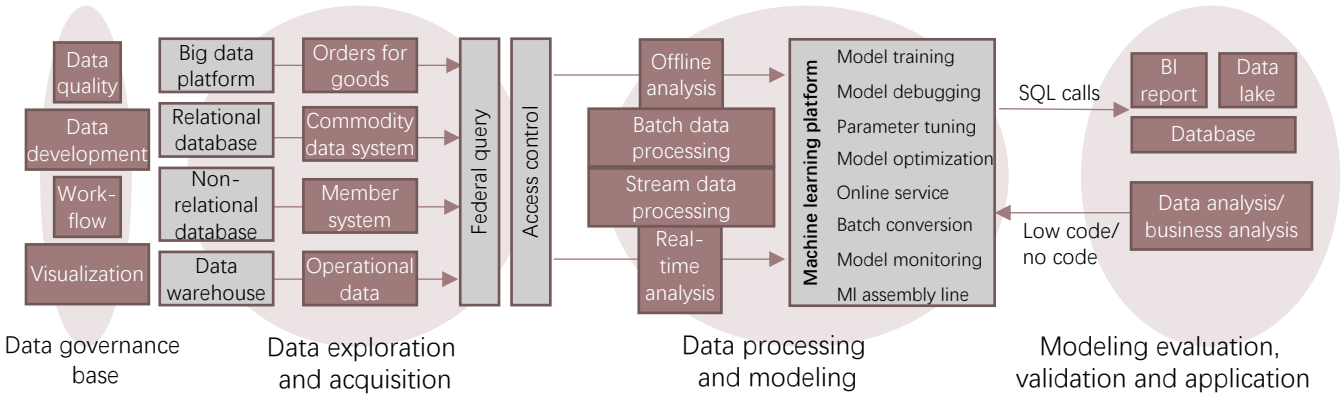
❑ Big data + AI

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:

1. Generality: Machine learning model inference can be carried out directly through SQL.
2. Ease of use: provide simple tools to realize business, use existing data to realize machine learning model training.
3. Transparency: visual data prepared for low-code data cleaning transformation.
4. 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

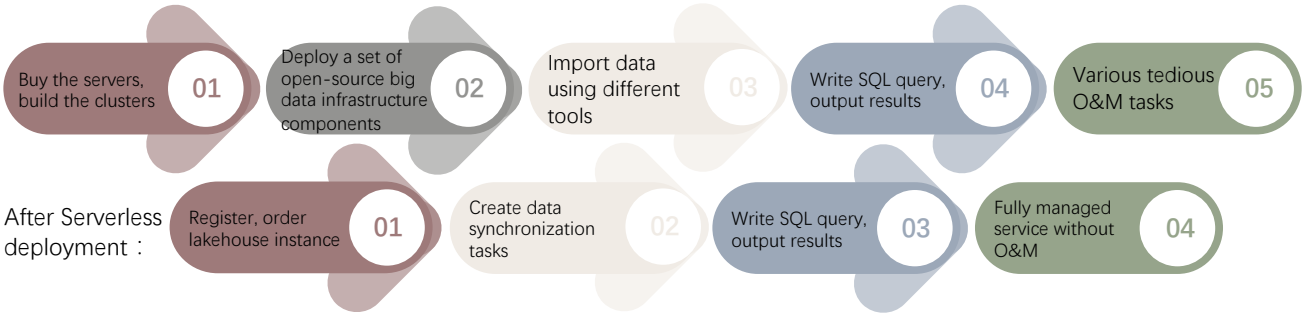
- **Serverless lakehouse integration refers to data storage, data query engine, data warehouse, data processing framework, and data catalog products that all support serverless deployment**

❑ Serverless deployment

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

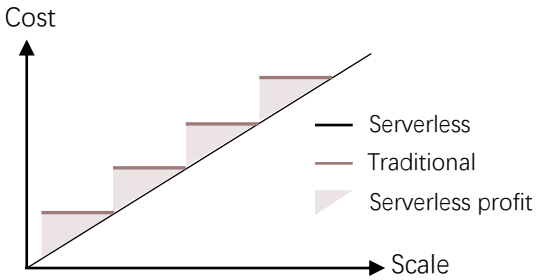
Serverless lakehouse integration refers to data storage, data query engine, data warehouse, data processing framework, and data catalog products that all support serverless deployment

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 Serverlesss 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.

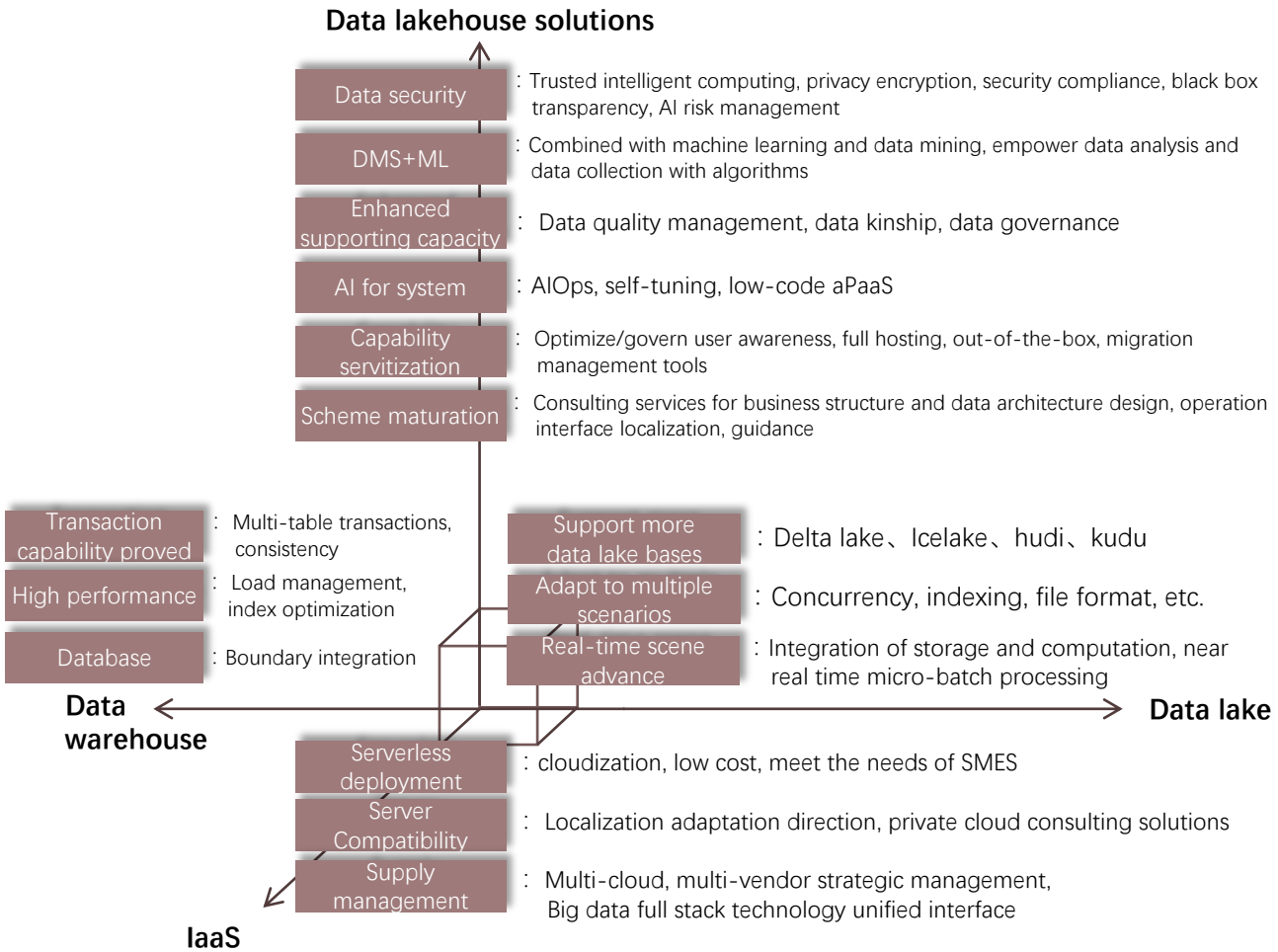
DLA of Ali Cloud creates Maxcompute, an integrated architecture of cloud native+Serverless+database and big data, through core components Lakehouse, Serverless Spark and Serverless SQL.

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



Source: Frost&Sullivan

❑ **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.

02 Data Management Solutions Market Analysis

- Data management user profiles
- Data management related talent demand analysis
- Data management solution user needs
- Application scenario dimension-enterprise map
- Cloud data management solution selection essentials
- Data Management Solutions Product and Vendor Atlas

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

Data analysis team

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

Data management team

Data management engineer：
Optimize data quality and prepare ETL operations
Catalog data and perform metadata management
Balance data protection and data privacy

GRC team

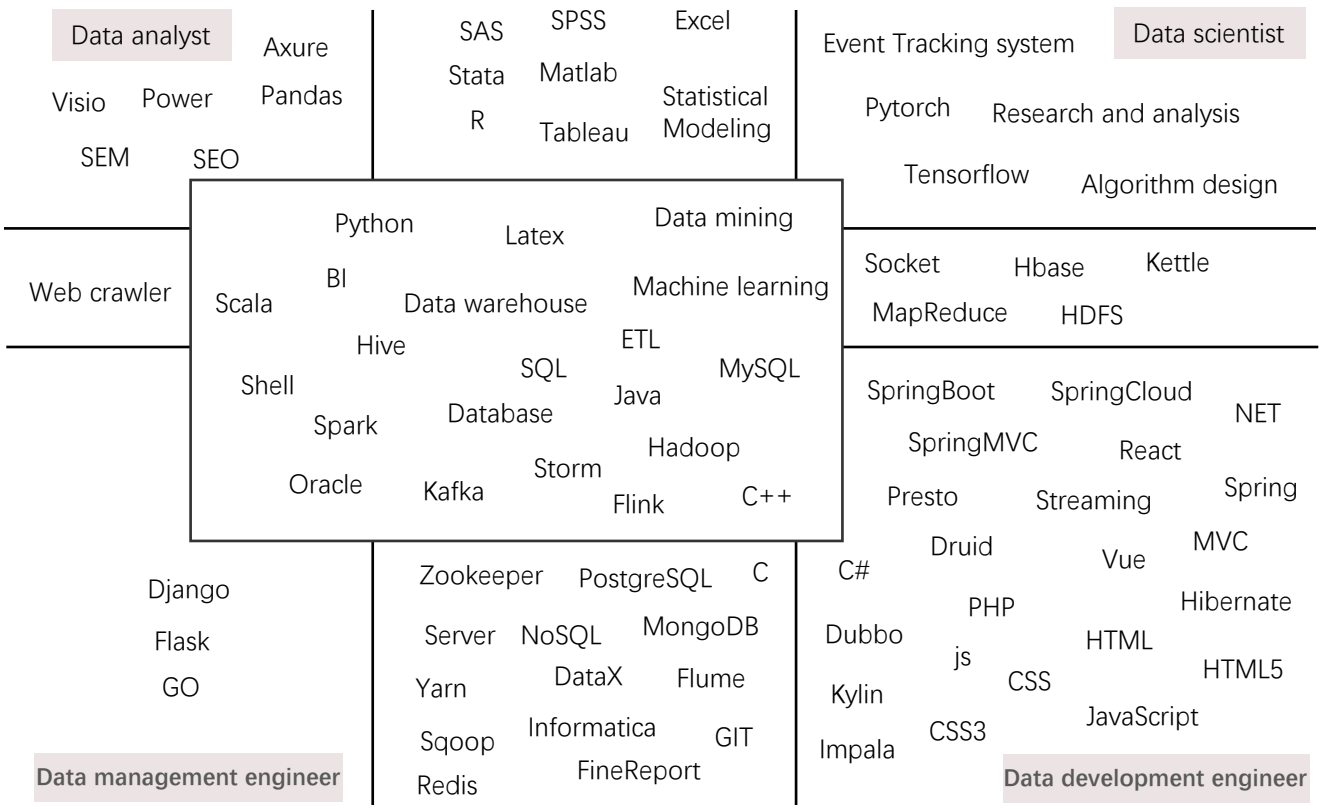
Data governance expert:
Establish data governance and security policies
Ensure data privacy and security throughout the chain
Compile requirements for retention, archiving and disposal, and ensure data compliance with policies and regulations

Business line team

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

Source: IBM, LeadLeo

Overview of technology stacks required for data management roles

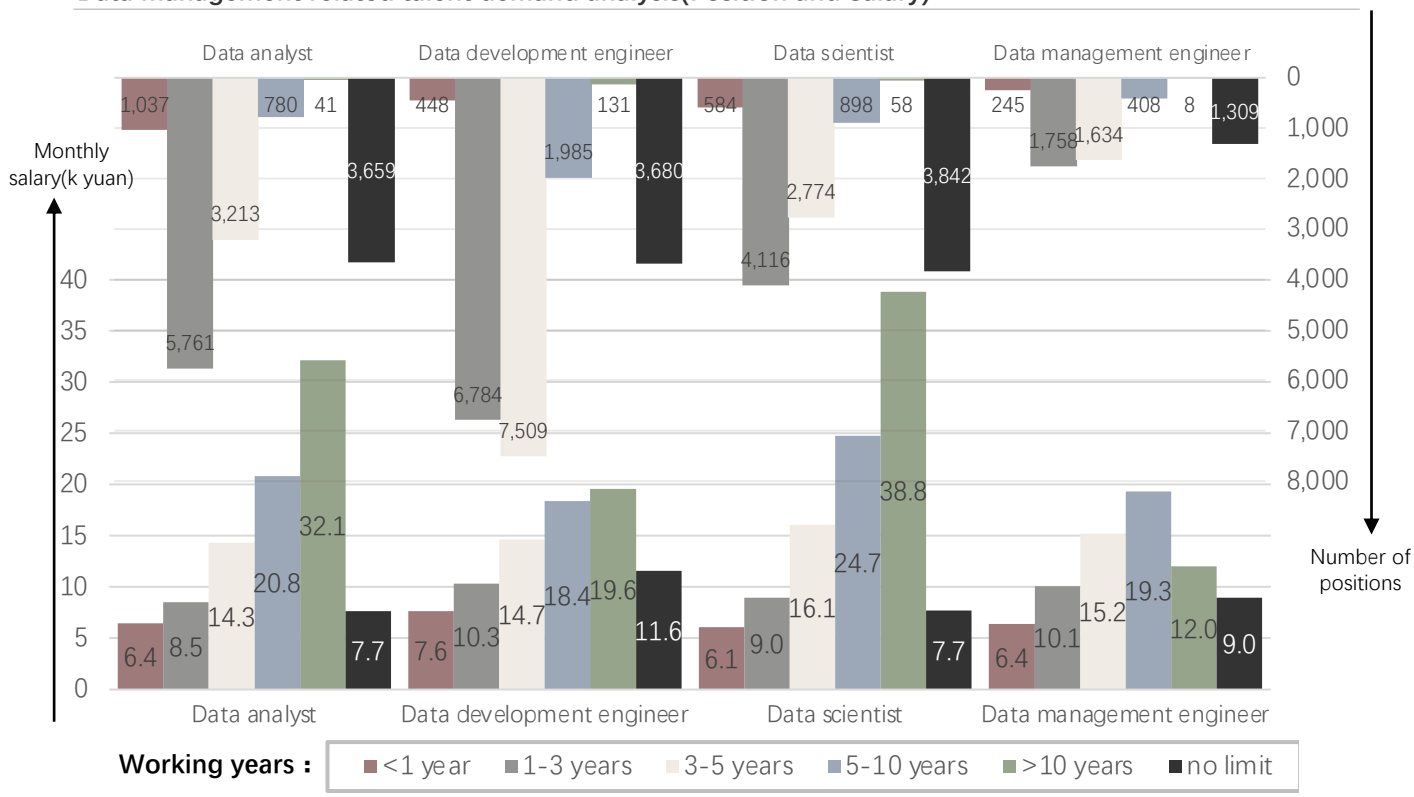


Source：Boss zhipin, Liepin, 51job, take first- and second-tier cities as sample cities, data analyst, data development engineer, data scientist and data management engineer as keyword. Retrieval time: 2022.05, hot word frequency analysis for technical stack requirements, disposed by Frost&Sullivan.

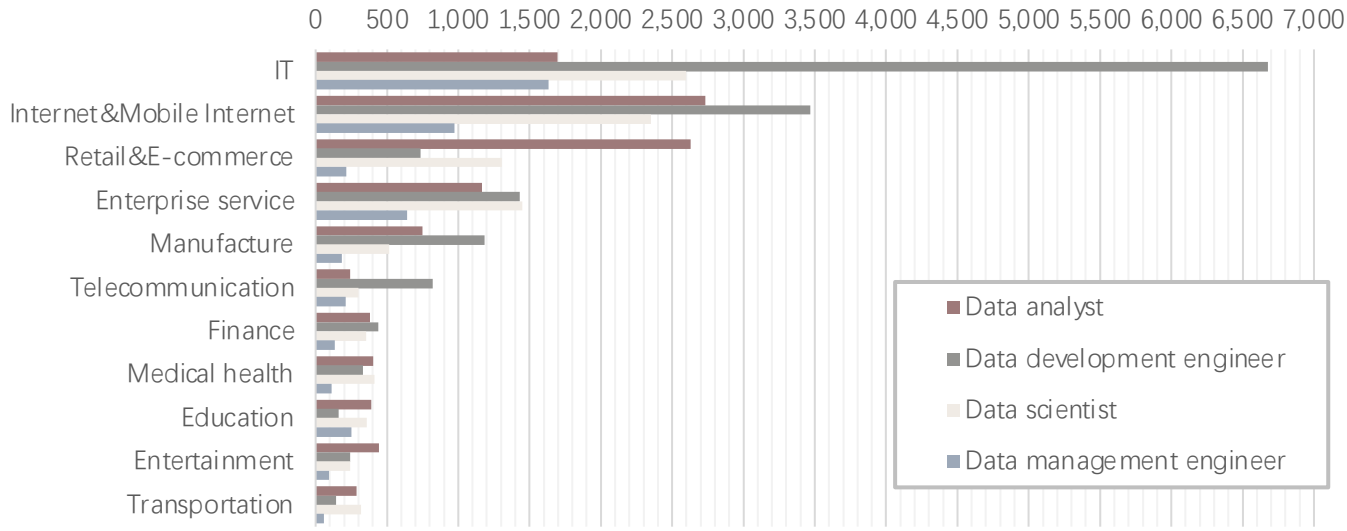
Data management related talent demand analysis

- 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

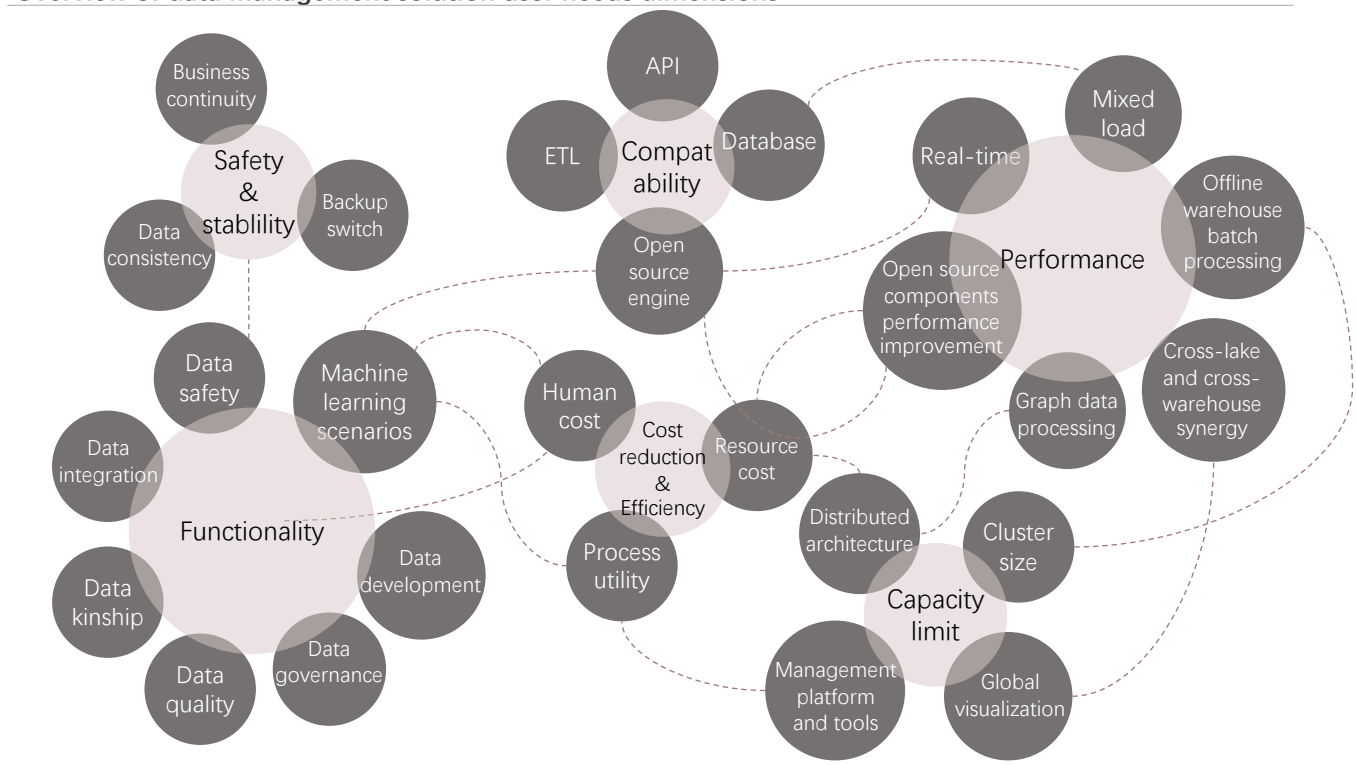


Source: Boss zhipin, Liepin, 51job, take first- and second-tier cities as sample cities, data analyst, data development engineer, data scientist and data management engineer as keyword. Retrieval time: 2022.05, disposed by Frost&Sullivan.

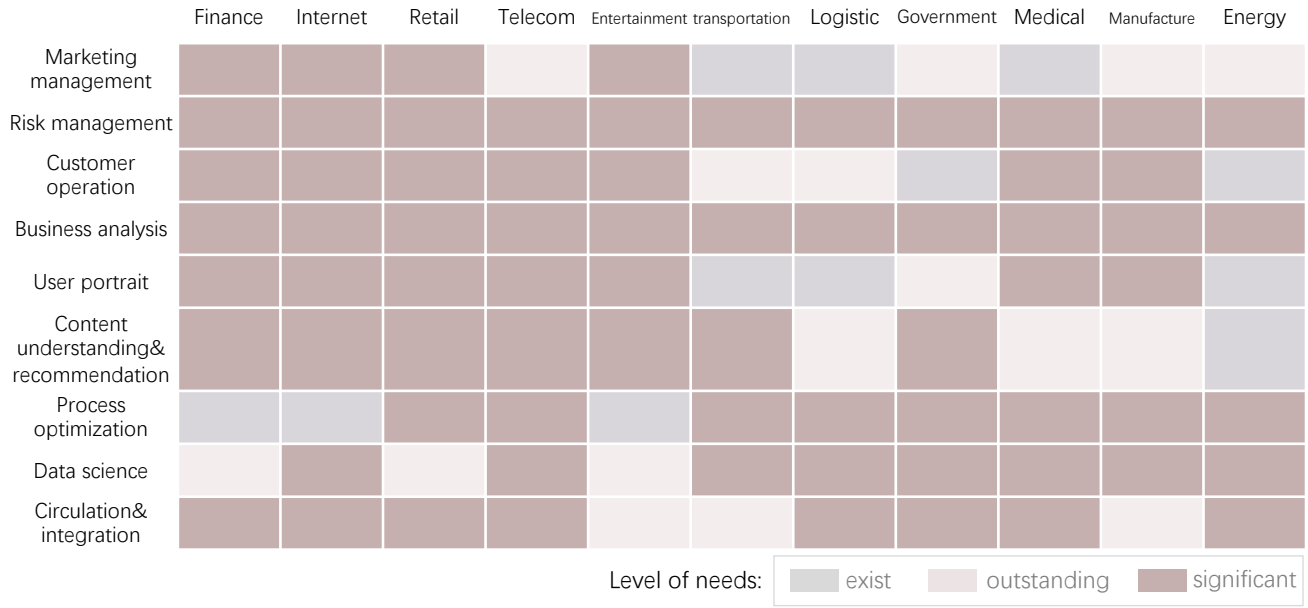
Data management solution user needs

- 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

Marketing management

- Customer portrait and customer label functions
- Multi-channel marketing services
- Real-time marketing services
- AI intelligent marketing services
- 360° comprehensive customer data view
- Unmanned supermarket
- Precise advertising
- Automatic intelligent layout of website or APP pages
- Customized digital marketing
- Digital marketing platform

阿里云 亚马逊云科技 HUAWEI 火山引擎 腾讯云

Risk management

- Introduce risk differentiation services
- Pre-screening from a risk perspective
- Risk content identification; Regulatory submitting
- Anti-fraud, anti-money laundering
- Repayment risk assessment and credit assessment
- Enterprise risk assessment, risk control reasoning
- Compliance log analysis, internal control compliance
- Base station risk management
- Risk analysis of equipment and manufacturing process

HUAWEI 亚马逊云科技 SequoiaDB 巨杉数据库 浪潮云 TRANSWARP 星环科技

Customer operation

- Operational activities optimization
- Enterprise spectaculars, real-time spectaculars
- Customer behavior log analysis
- Digital operation of event tracking library
- Customer data platform
- Customer management information service

阿里云 亚马逊云科技 HUAWEI 火山引擎 金山云

Business analysis

- Store operation analysis
- Product quality management data lake
- Consolidated management of subsidiary financial statements
- Profit and loss pre-query
- Product atlas real-time indicator
- Real-time data index reporting

亚马逊云科技 百度智能云 HUAWEI 金山云 TRANSWARP 星环科技

Note: The logos displayed are sorted by the first letter.
Source: Sullivan, Head Leopard Institute

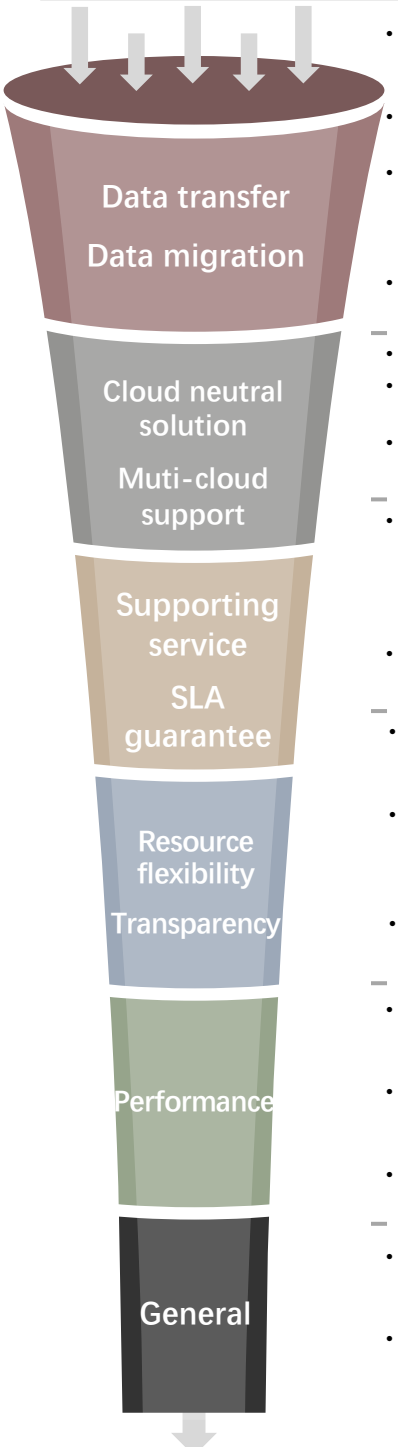
Data service application scenarios atlas of data management solutions (continued)

Note: The logos displayed are sorted by the first letter.
Source: Sullivan, Head Leopard Institute

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 useing 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 lakhouse of each vendor are also listed

Oversea data management solutions vendors and representative products

Category	Vendor	Data warehouse	Data lake	Data lakehouse
Cloud vendors	amazon web services	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
	TRANSWARP 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
	ESENSOFT 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

03 Competitive landscape in China's data management solutions market

- Assessment Scoring
- Comprehensive Vendors Assessment – Frost Radar
- Leading Competitors

Innovation Index Evaluation System Indicators

- This report sets up an innovation index evaluation system to evaluate and analyze data management solutions, with four indicators: data storage module, data preparation module, analysis support module, and data analysis module

Tier 1 indicator	Tier 2 indicator	Key points
Data Storage Module	data storage module	data life cycle storage, scaling, distributed storage operations, storage formats, compressed storage technologies, data processing and warehousing performance acceleration, storage indexes, etc.
Data Preparation Module	data extraction & cleansing	knowledge extraction method, extraction mode, classification dimension, consistency check, etc.
	data conversion & loading & synchronization	transformation function, structured processing of unstructured data, cross-system synchronization capability, etc.
Analysis Support Module	query, statistical analysis and visualization	query function, query acceleration, federal query, visualized display
	machine learning	machine learning algorithm, machine learning process component, high ease of use
	lake warehouse integration	data management integration capabilities, multi-lake and multi-warehouse associated computing, serverless deployment
Data Analysis Module	batch data analysis	work scheduling and optimization, data source operation, structured query language
	stream data analysis	real-time streaming data manipulation, time-window based analysis, high fault tolerance
	stream-batch integration data analysis	data warehouse architecture, data lake architecture, stream-batch processing technology
	online analysis processing (OLAP)	OLAP implementation, Ad Hoc, complex task scheduling, OLAP operation
	graph computing framework	iterative algorithm writing model, graph calculation function, graph query, graph analysis
	storage computing framework	highly abstract operator, fast construction of data processing application, storage computing data processing
	log analysis framework	out-of-the-box scheme, hot and cold data stratification, source code

Growth Index Assessment System Indicators

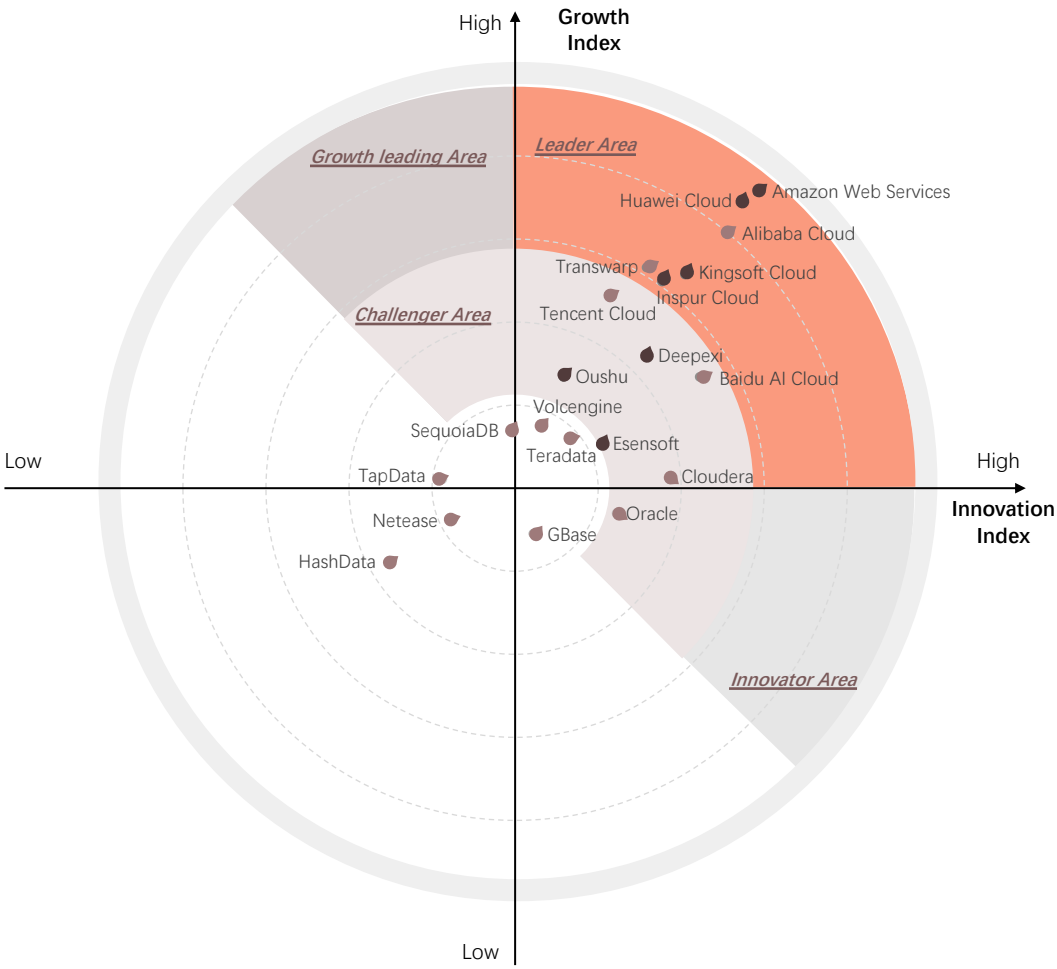
- This report sets up an growth index evaluation system to evaluate and analyze data management solutions, with seven indicators: Process orchestration and management module, compatibility assessment, query and computing performance assessment, disaster recovery construction and security, service support level, open source community and industry chain ecology, data service scenario solutions

Tier 1 indicator	Tier 2 indicator	Key points
Process Orchestration and Management Module	workflow management	visualized process orchestration interface, scheduling trigger mechanism, multi-person collaboration
	maintainability	error diagnosis, fully managed without O&M, multi-dimensional visualized system monitoring alarms
	resource management and data management functions	resource allocation strategy, global resource centralized management, data management function
Compatibility Assessment	data format & interface	formatting, unstructured, application program interface
	cloud compatibility & external compatibility	cloud deployment solutions, data replication and migration across cloud platforms, open source engine
Query and Computing Performance Assessment	query and analyze performance	query delay time, concurrent query number, real-time degree of analysis, data mart
	Highly availability & extensibility	SLA, automatic fault detection, online and offline capacity expansion, multiple copies
Disaster Recovery Construction and Security	backup, restore, and migration	Data backup and recovery, fault recovery and migration
	users, permissions, and logs	Multi-factor authentication, node access control, logging elements
	data security, security protection	Trusted computing services, all-secret data, security protection technology
Service Support Level	service support level	Implementation services, value-added services, expert teams, product documentation
Open Source Community and Industry Chain Ecology	open source situation	open source components, contributors, issues, representative user
	industry chain cooperation	hardware, peers, middleware, internal product lines, universities
Data Service Scenario Solutions	marketing management, risk management, customer operation, business analysis, user portrait, content understanding and recommendation, process optimization, data science and other data service scenarios	industry application - industry breadth of scenario practices scenario service functions - service granularity and depth of scenario practices dominant technology or service pattern - vertical demand service capability

Comprehensive Vendors Assessment – Frost Radar

- China’ s data management solutions market is in a steady growth phase, and competitive players will be divided into echelons based on their performance in two dimensions: innovation capability and growth capability

❑ **Comprehensive Competitive Performance of China’ s Data Management Solutions Market - Frost Radar™**



Note: The circle corresponds to the lowest to highest overall score, and the competitiveness is derived from the “Innovation Index” and “Growth Index”.

Light-coloured markers indicate that the vendor’s performance is based on information and data from third-party research. The position of the marker indicates the current competitive performance, and the markers’ direction indicates the estimation of the future competitive performance trend of each vendor.

The development of China’ s data management solutions market is in a steady growth phase, and the conclusions of this report’s analysis of the comprehensive competitiveness of data management solutions products and services of competing entities apply only to the data management solutions market development at this stage. Sullivan will continue to monitor the data management solutions market and capture competitive trends.

❑ **The vertical coordinate represents the "growth index “:**

It measures the competitiveness of competitive entities in the data management solutions growth dimension, the higher the position, the stronger theinnovative technologies or capabilities of data management solutions such as data storage, data preparation, machine learning analysis support, integration of lake and warehouse, multi-dimensional and multi-frame data analysis, etc.

❑ **The horizontal coordinate represents the "innovation index “:**

It measures the competitiveness of competitive entities in the data management solutions innovation dimension, The more the position is to the right, the stronger the market growth capability and level of data management solutions in terms of compatibility, query & computing performance, disaster recovery security, service support, industry chain ecology, data service scenario solutions, etc.

Leader: Amazon Web Services

- Amazon Web Services is a leader in data management solutions in China, with technology capabilities to break down data silos. build the unified data governance base in the cloud, and transform machine learning from experiment to practice to support agile business innovation. Amazon Web Services provides highly innovative and scalable data management solutions with professional and in-depth technical support.

Breaking down data silos with unified data base

- By creating a unified data base in the cloud to break the bottleneck of data sharing, Amazon SageMaker Studio capabilities are upgraded to provide a unified big data and machine learning development platform.
- Amazon Web Services Lake House Architecture supports SQL for machine learning model inference, graphical implementation of data preparation, and model no-code automatic training, which empower business people to explore machine learning modeling independently and autonomously, unifying technology and business value.

Flexible Deployment and Agile Innovation

- User at any stage with any data size can achieve agile data innovation by Amazon Web Services Lake House Architecture.
- Highly decoupled and componentized tools and services for each process in big data task are easy to use.
- Serverless deployment function covers batch computing and streaming computing scenarios, and greatly reduces the user threshold of deployment planning and operation & maintenance.

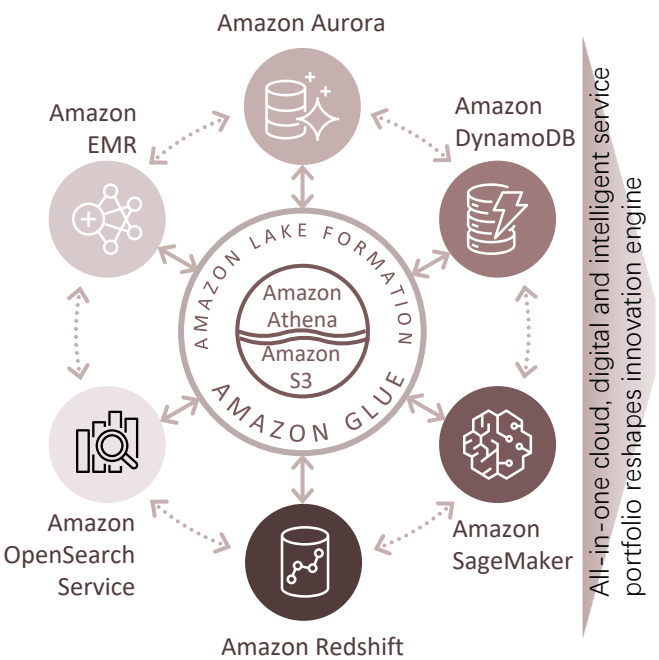
Professional and in-depth technical support services

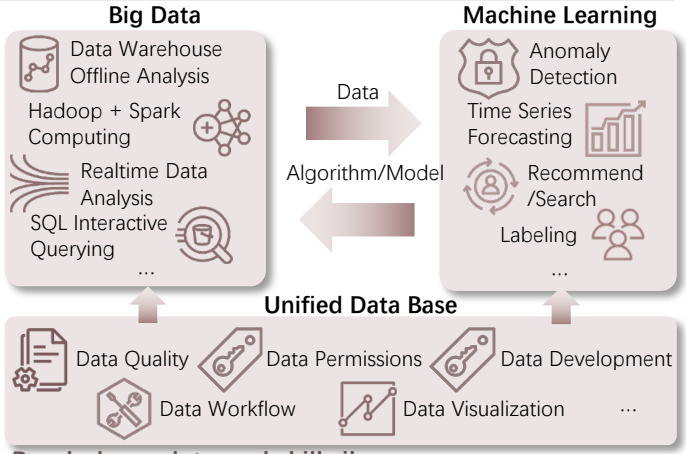
- Amazon Web Services' data engineers, data scientists and machine learning engineers supports users around the world.
- Through application scientists in data labs geared towards rapid algorithm prototyping, machine learning solution labs geared towards production accuracy model guidance, and professional services experts in big data analytics and machine learning who provide end-to-end consulting and delivery, Amazon Web Services help its customers to transform data-driven business and organization from the beginning to the end.

Global multi-industry, multi-scenario business practices on the cloud

- Amazon We Services has global experience in a wide range of industries and scenarios for enterprises of different sizes.
- With global leading technology and service level, Amazon Web Services helps Chinese customers to cultivate locally, helps overseas customers to root in China, and help Chinese enterprises to go abroad successfully.

Amazon Web Services Lake House Architecture Upgrades





Break down data and skill silos

Amazon Web Services Lake House Architecture integrates big data technology and machine learning technology with a unified governance base on the cloud.
Amazon Athena Federated Query enables complete query services without data movement, supports fast on-the-fly queries and complex analytics, and has data persistence with high availability through Amazon S3.
Amazon Lake Formation provides cross-geographic, cross-account metadata storage sharing and permission control capabilities.

Transform Machine Learning from Experiment to Practice

Amazon SageMaker offers over 15 available algorithms in pre-built container images.
SageMaker JumpStart supports one-click deployment and optimization of 150+ open source models.
Amazon SageMaker Data Wrangler includes more than 300 built-in data transformations.

Empower business analysts to explore and unleash innovation

Both Amazon Redshift ML and Amazon Athena ML support launching model training requests as SQL, and Amazon SageMaker Canvas AutoML capabilities provide model training, then returned as SQL.
Amazon SageMaker Canvas visually builds machine learning models without code.
Amazon SageMaker Pipelines build fully automated ML workflows.
Amazon SageMaker Canvas integrates with Amazon SageMaker Studio, making it easier for business analysts to share models and datasets with data scientists.

Serverless deployment

Purpose-built, flexible and scalable serverless analytics services cover both bulk and streaming data.
Amazon Redshift, Amazon EMR, Amazon Kinesis, Amazon Glue, Amazon Athena, Amazon MSK all support serverless deployment.

High Performance at Low Cost

Spark on Amazon EKS delivers a 60% performance boost over Spark on K8s.
Amazon EMR is 1.7x faster than Apache Spark 3.0 at 40% of the cost.
Amazon EMR is 2.6x faster than Presto 0.238, at 80% of the cost.
Amazon EMR on Graviton2 delivers a 15% performance improvement over x86 instances at 75% of the cost.

Amazon Web Services
2021 China Data Management Solutions
Frost Radar Ranking Notes

✓Amazon Web Services Ranked 1st in Growth Index in Frost Radar
✓Amazon Web Services Ranked 1st in Innovation Index in Frost Radar

Amazon Web Services ranked #1 in the Growth Index, scoring highest in the following metrics :

- **Amazon Web Services scored highest in Data Service Scenario solutions :**
  - Amazon Web Services provides mature data services solutions including marketing management, risk management, customer operations, business analytics, user profiling, content understanding and recommendation, process optimization, data science, etc. Amazon Web Services' customer cases cover automotive, manufacturing, finance, retail and consumer goods, healthcare and life sciences, media and entertainment, education, gaming, e-commerce, energy and power industries.
  - Amazon Web Services provides with 200+ full-featured services, over 15 years of experience serving 100 thousands global partner network members. Amazon Web Services' global business system is strongly supported to provide customers with additional resources and added value.

- **Amazon Web Services scores highest in orchestration and management capabilities :**
  - Amazon Web Services provides visual process orchestration operation interface, both support through Amazon Glue workflow or Amazon Step functions for dragging, sliding and other ways to operate the orchestration and revision, also support for engineers who prefer code orchestration using Amazon MWAA for orchestration; through Cloud9 cloud IDE, to achieve real-time multi-person collaboration; it also supports complex multiple process orchestration joint orchestration, greatly reducing the complexity of the code and improving the efficiency of secondary development of users.

Amazon Web Services ranked #1 in the Innovation Index, scoring highest in the following metrics :

- **Amazon Web Services scored highest in lakehouse formation and integration :**
  - Amazon Web Services' LakeHouse architecture offers the benefits of data convergence and unified governance. Amazon Redshift can interact directly with data lakes through Amazon Redshift spectrum and federate queries via SQL; Amazon Redshift can also federate queries with serverless analytics product Amazon Athena without data exchange; Amazon Lake Formation can complete the rapid construction of data lakes and cell-level access rights control through supervised tables, providing users with efficient and agile multi-lake, multi-bin correlation computing capabilities. In addition, the LakeHouse architecture supports serverless deployment.

- **Amazon Web Services scored highest in Machine Learning and Big Data integration:**
  - Amazon Web Services provides a one-stop end-to-end machine learning platform, Amazon SageMaker, with built-in service-oriented components for the entire machine learning process; a variety of algorithms and pre-trained models are pre-built through Amazon SageMaker JumpStart to help users quickly deploy models and run inference; support for launching model training requests from the data warehouse via SQL enables business analysts to operate with low code and a visual interface, without the need to turn to algorithm engineers, deeply empowering business innovation.

- **Amazon Web Services scored highest in Stream Data Computing :**
  - Amazon Web Services offers a variety of in-memory-based computing frameworks, including Amazon Kinesis Data Analysis, which is a real-time streaming computing engine, and Flink & Spark, which can be deployed on Amazon EMR. Standard Spark arithmetic is available through Amazon Glue. DynamicFrame extensions also enable rapid building of distributed semi-structured data processing applications.

Leader: Huawei Cloud

- **Huawei Cloud is a leader in data management solutions in China, providing highly innovative and high-growth data management solutions through technological innovations in lakehouse collaboration, flexible scaling, and data intelligence integration, as well as open source and partner ecology, business empowerment, and talent development systems.**

❑ **Lakehouse collaboration, Flexible scaling, and data intelligence integration**

- HetuEngine one-stop interactive SQL analysis engine to achieve unified access and collaborative analysis of data inside and outside the lake, DLC+OBS to provide global unified data storage at the bottom layer, and flexible combination of multiple engines at the top layer according to business needs.
- FusionInsight MRS+ModelArts to build RTD real-time decision engine. ModelArts support for MLOps.
- Huawei Cloud Stack+MRS realizes mixed deployment of big data on-premises resources and elastic resources, and the overall lake warehouse integrated architecture supports serverlessness, where OBS data storage, data query engine, data warehouse, and data catalog products support serverless deployment.

❑ **Leading open source, Partner ecology.**

- FusionInsight's contribution is Top2 in Hadoop community and Top4 in Spark community; it insists on the open route and has opened CarbonData, OpenLooKeng, IoTDB, and now the community has 28 PMC & Committer. 2,400+ enhancements, including ClickHouse with 10 times higher performance than open source; Superior Super Scheduler with 30 times higher performance than open source.
- 1,000+ industry application solution eco-partners, covering government, finance, operators, Internet, and pan-enterprise sectors, to build landing solutions.

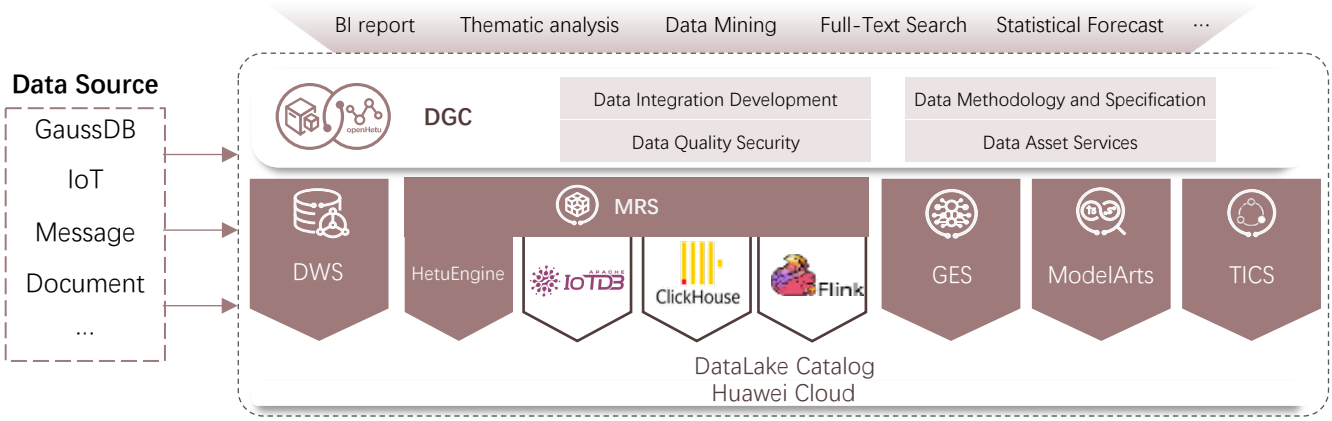
❑ **Business empowerment, industry-education integration**

- FusionInsight provides a three-stage course from data lake top-level architecture design, secondary development, and operation and maintenance, enabling users in depth in all aspects.
- SmartBase project covers 72+ colleges and universities, developing 9 artificial intelligence textbooks, 8 books in the smart computing series, and 5 open source innovation practice courses. Jointly promote the combination of open source and computer education in universities, and attract 12,000+ students from 500+ universities to register for the China Software Open Source Innovation Competition 2021.

❑ **Typical Customer Cases**

- ICBC introduces FusionInsight to build a big data system and enhance the new experience of instant BI for 13,000 data analysts across the bank.
- FusionInsight's Intelligent Data Lake Escorts 10x Growth in User Volume in the Next 2 Years as Dream Rate Migrates to Huawei Cloud.
- Shanghai 12345 uses GES to build a "cognitive hub" for government services, increasing hotline connection rate by 20% and work order processing efficiency by 60%.

Huawei Cloud FusionInsight Lakehouse Architecture Overview



MRS Cloud Native Data Lake
Offline data lake: single cluster 60,000+ nodes, rolling upgrade 0 interruptions.
Real-time data lake: self-researched CDL to achieve full-link real-time, T+0 supply data.
Logical data lake: self-researched river map engine, cross-lake and cross-warehouse analysis 50 times more efficient.

GES graph engine
Large-scale: distributed architecture, supporting 100 billion large-scale graph computation.
High performance: memory parallel computing, 6 seconds jump check.
Easy to use: built-in 30+ algorithms, drag-and-drop editing, visual presentation.

GaussDB(DWS) Cloud Data Warehouse
Ultra-large scale: Built the world's largest financial data warehouse list cluster with 480 nodes.
One-stop analysis: support enterprise data warehouse, data mart and IoT multi-scenarios.
Full-scene deployment: one set of architecture supports multi-cloud deployment with consistent user experience.

DGC Data Lake Governance Center
One-stop: Full-link data governance tools cover the whole life cycle of data, improving efficiency by 7 times +.
Modeling fast: support one-click import template reuse, data modeling shortened from month level to day level.
Compatible and open: support access to 40+ data sources and Huawei Cloud various data services.

ModelArts AI platform
Multi-scenario: support for online, batch, video and edge and other full-scenario AI services AI resources.
Unified management: heterogeneous resources, pooled scheduling, and 100% increase in AI hardware utilization.
Pervasive AI: visualization to complete training, deployment, monitoring, and updating of AI services, significantly reducing the threshold of AI applications.

GaussDB Database
High performance: single node processing capacity up to 1.5 million tpmC, 32 nodes processing capacity up to 15 million tpmC
High scalability: support online elastic scaling of 1,000+ large distributed clusters
High availability: support two locations and three centers, in line with the highest financial regulation; same city cross-AZ RPO=0, RTO<60s

Huawei Cloud
2021 China Data Management Solutions Frost Radar Ranking Notes

✓Huawei Cloud Ranked 2rd in Growth Index in Frost Radar
✓Huawei Cloud Ranked 2rd in Innovation Index in Frost Radar

Huawei Cloud ranked #2 in the Growth Index, scoring highest in the following metrics :

- **Huawei Cloud scored highest in Open Source Community Contribution :**
  - 20+ open source components including ClickHouse, Hudi, Spark, Flink, CDL real-time access, HetuEngine/OpenLooKeng interactive SQL engine, Superior scheduler, CarbonData engine, and the time-series database IoTDB. Huawei Cloud leads the industry in open source community contribution from the perspective of the number of community contributors and the number of resolved Issues.
- **Huawei Cloud scores highest in Domestic industry chain ecology :**
  - Huawei Cloud has 1,000+ ISV independent software/middleware/OS operating system development partners in government, finance, operators, large enterprises, Internet and general industry, developing tools and solutions for migration, operation and maintenance for different application scenarios, respectively.
  - Huawei Cloud is committed to establishing a talent training system that integrates industry and education. Huawei Cloud work closely with Tsinghua University, Shanghai Jiaotong University and Tongji University on big data and AI, and also provide systematic certification of big data and AI talents on Huawei Talent Online, so that Huawei can continuously export big data talents who meet actual production needs for big data-related industries.

Huawei Cloud ranked #2 in the Innovation Index, scoring highest in the following metrics :

- **Huawei Cloud scored highest in Query Function:**
  - Huawei Cloud provides multi-source federation queries through its own HetuEngine, which supports cross-source queries from multiple data sources such as MRS Hive, HBase, ES, and DWS, with features of ease of use, high efficiency and cloud-native. In addition, Huawei Cloud provides SQL on Anywhere capability through GaussDB (DWS), which maps data files on multiple data sources to external tables and associates queries with table data in the cluster, creating a data environment for customers to interact across cluster interconnection.
- **Huawei Cloud scored highest in Batch Data Computing :**
  - Huawei Cloud provides users with efficient job scheduling and optimization through its self-developed Superior Scheduler, which provides patented scheduling optimization algorithms, global views, and abundant scheduling policies. It supports performing aggregation, filtering, sorting, Distinct and other operations across relational databases and big data storage systems, and fully supports DDL, DML, DQL, DCL, TCL, HiveSQL, FlinkSQL and other languages.
- **Huawei Cloud scored highest in Graph Analytics Capability :**
  - Huawei Cloud' s Graph Engine Service (GES), using Huawei's self-developed EYWA kernel, supports BSP and GAS two simultaneous iterative algorithms to write models, built-in 30+ graph analysis algorithms, supports 100 billion points trillion with 6-hop second query, and realizes super large-scale integrated graph analysis and query. GES also Support Gremlin, Cypher query language, provide wizard-style and easy-to-use visual analysis interface.

Leader: Alibaba Cloud

- **Alibaba Cloud is a leader in data management solutions in China, providing highly innovative and high-growth data management solutions through technological innovation in lakehouse integration, data intelligence integration, and other areas, as well as open construction and partner ecology.**

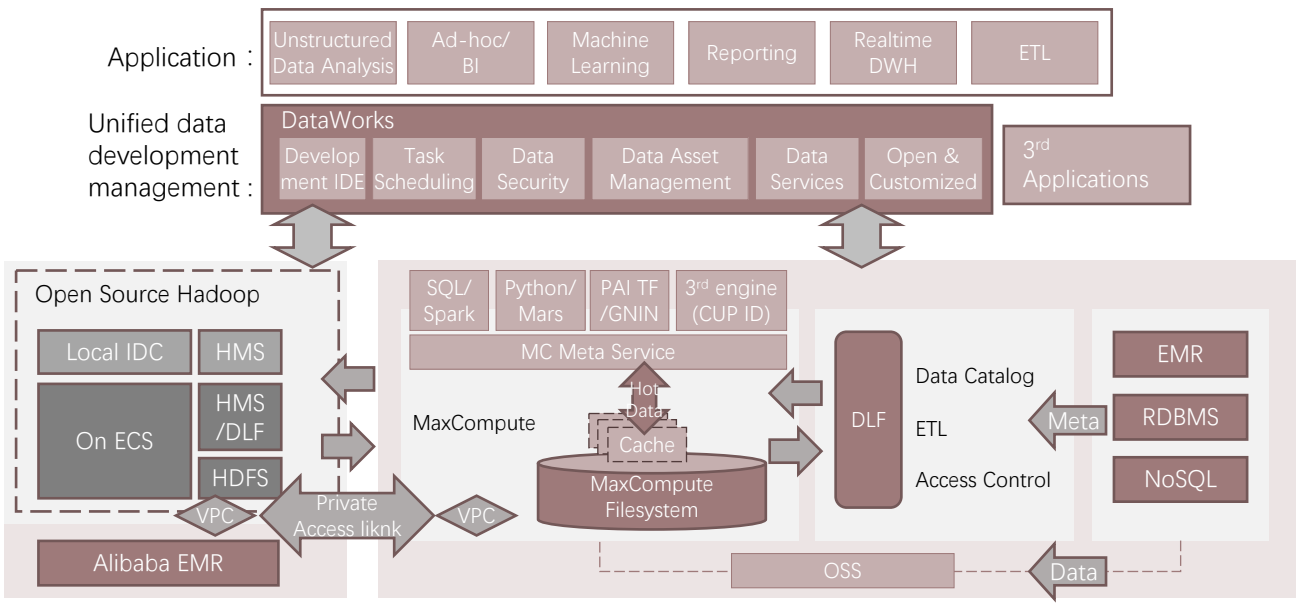
❑ **Lakehouse Integration, Data Intelligence Integration**

- Offline data warehouse MaxCompute and real-time data warehouse Hologres are deeply integrated to provide offline real-time integrated massive cloud data warehouse architecture.
- DataWorks unified engine docking and management, supporting full-domain perspective of data assets for MaxCompute, EMR, CDP, CDH, HDP, open source Hadoop, AnalyticDB and other engines, extracting elements and information from data sources.
- Linkage with data lakes (OSS, EMR-HDFS) to seamlessly process and analyze data lake data without data relocation.
- MaxCompute+PAI, a machine learning platform built on top of MaxCompute's powerful computing resources for one-stop experience, has been widely used in many industries and fields.

❑ **Open construction, Partner ecology**

- Compatible with a variety of development syntax, seamlessly connect to the original development method; support a variety of custom interfaces, can customize the development method and process.
- Support more than a dozen products such as Tableau, QuickBI, PowerBI, FineBI, etc. to quickly complete data analysis and visual presentation. Deep integration with various eco-partners' products such as Ant Group, Divine, and Digital Language, providing a multi-dimensional product portfolio for various big data scenarios.

Alibaba Cloud Maxcompute Lakehouse Architecture Overview



Best practices for building data lakehouses

Maxcompute supports building a data lake with DLF, object storage OSS, and Hadoop, including local server room, virtual machine on cloud, and EMR. The metadata information of data lake is mapped to MaxCompute Project shared data warehouse supporting tool chain.

Open and compatible, fast access

MaxCompute is highly compatible with Hive/Spark, and supports a set of tasks that can be run flexibly and seamlessly in two systems of lakehouses.
Self-developed PrivateAccess network connectivity can be connected to IDC Hadoop, self-built Hadoop on ECS on cloud and EMR Hadoop cluster.
Non-intrusive, no need to retrofit existing business.
Connects to purchased MaxCompute silos in a quick and easy turn-up procedure.

Unified data development and management

Open up the network, storage, and compute layers to achieve DB metadata visibility.
DataWorks unifies various data assets, making them visible in the data map across the entire domain and supporting the extraction of elements and information from data sources.
Cross-data lineage within the same engine, unifying table-level and field-level. Multiple hadoop clusters can be mounted for unified engine docking and management.

Super-scale machine learning

Seamless integration with the machine learning platform PAI, providing powerful machine learning processing capabilities.
Can use the familiar Spark-ML to carry out intelligent analysis; use the Python machine learning tripartite library.
Enables mega datasets for machine learning, deep learning training, and a complete pipeline for high-performance hyperscale sample generation and feature processing.

Terms

- **Data Warehouse**: A strategic collection of all types of data to support the decision making process at all levels of the enterprise.
- **Data Lake**: a centralized repository that allows you to store all your structured and unstructured data at any scale.
- **Hadoop**: is a distributed system architecture that leverages the power of clusters for high-speed computing and storage. The core design framework is HDFS, which provides storage for massive amounts of data, and MapReduce, which provides computation for massive amounts of data.
- **OLTP**, is an online transaction processing system, OLTP system is mainly to record transactions when the current update, insert and delete.
- **OLAP** is an online analytical processing system. It stores the input historical data. It allows users to view different summaries of multidimensional data.
- **Logical Data Lake**: A centralized logical data lake catalog Catlog or data virtualization engine enables collaborative analysis and interactive queries across lakes, silos, domains, multiple clouds, and business systems.
- **Serverless**: Serverless deployment provides services through FaaS+BaaS, allowing users to develop, run and manage applications without building or maintaining a complex infrastructure.
- **Lambda Architecture**: An architecture that meets the key characteristics of real-time Big Data systems, including: high error tolerance, low latency, and scalability. Integrates offline and real-time computing, incorporating a set of architectural principles such as immutability, read/write separation, and complexity isolation.
- **Kappa architecture**: In response to the above 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 stream computing system, making real-time computing and batch processing using the same set of code.
- **Scatter/Gather model**: 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.
- **MapReduce model (Hadoop)**: achieves reliability by distributing large-scale operations on a dataset to each node on the network; each node periodically returns the work it has done and the latest status.
- **Massively Parallel Processing (MPP)**: employs a Shared-nothing architecture, where each node uses individual resources and has an optimal operating environment. Streamlined execution with no waiting, in-memory storage of data, and no disk IO.

Methodology

- Frost & Sullivan has conducted in-depth research on the market changes of 10 major industries and 54 vertical industries in China with more than 500,000 industry research samples accumulated and more than 10,000 independent research and consulting projects completed.
- Rooted on the active economic environment in China, the research institute, starting from data management and big data fields, covers the development of the industry cycle, follows from the enterprises' establishment, development, expansion, IPO and maturation. Research analysts of the institute continuously explore and evaluate the vagaries of the industrial development model, enterprise business and operation model, Interpret the evolution of the industry from a professional perspective.
- Research institute integrates the traditional and new research methods, adopts the use of self-developed algorithms, excavates the logic behind the quantitative data with the big data across industries and diversified research methods, analyses the views behind the qualitative content, describes the present situation of the industry objectively and authentically, predicts the trend of the development of industry prospectively. Every research report includes a complete presentation of the past, present and future of the industry.
- Research institute pays close attention to the latest trends of industry development. The report content and data will be updated and optimized continuously with the development of the industry, technological innovation, changes in the competitive landscape, promulgations of policies and regulations, and in-depth market research.
- Adhering to the purpose of research with originality and tenacity, the research institute analyses the industry from the perspective of strategy and reads the industry from the perspective of execution, so as to provide worthy research reports for the report readers of each industry.

Legal Disclaimer

- The copyright of this report belongs to LeadLeo. Without written permission, no organization or individual may reproduce, reproduce, publish or quote this report in any form. If the report is to be quoted or published with the permission of LeadLeo, it should be used within the permitted scope, and the source should be given as "LeadLeo Research Institute", also the report should not be quoted, deleted or modified in any way contrary to the original intention.
- The analysts in this report are of professional research capabilities and ensure that the data in the report are from legal and compliance channels. The opinions and data analysis are based on the analysts' objective understanding of the industry. This report is not subject to any third party's instruction or influence.
- The views or information contained in this report are for reference only and do not constitute any investment recommendations. This report is issued only as permitted by the relevant laws and is issued only for information purposes and does not constitute any advertisement. If permitted by law, LeadLeo may provide or seek to provide relevant services such as investment, financing or consulting for the enterprises mentioned in the report. The value, price and investment income of the company or investment subject referred to in this report will vary from time to time.
- Some of the information in this report is derived from publicly available sources, and LeadLeo makes no warranties as to the accuracy, completeness or reliability of such information. The information, opinions and speculations contained herein only reflect the judgment of the analysts of leopard at the first date of publication of this report. The descriptions in previous reports should not be taken as the basis for future performance. At different times, the LeadLeo may issue reports and articles that are inconsistent with the information, opinions and conjectures contained herein. LeadLeo does not guarantee that the information contained in this report is kept up to date. At the same time, the information contained in this report may be modified by LeadLeo without notice, and readers should pay their own attention to the corresponding updates or modifications. Any organization or individual shall be responsible for all activities carried out by it using the data, analysis, research, part or all of the contents of this report and shall be liable for any loss or injury caused by such activities.

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