WeData Data Development Platform

Product Overview

Service Notice

This document provides an overview of the as–is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre–sales consultation and technical after–sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

Product Overview
- Overview
- Strengths
- Product Architecture
- Product Features
- Application Scenario

Product Overview
Overview

Last updated：2024-08-22 09:52:52

WeData (hereinafter referred to as WeData) is a cloud-based, one-stop data development and governance platform. It integrates full-chain DataOps capabilities, including DataInLong, data development, and task operation and maintenance. Additionally, it features a series of data governance and operation capabilities such as data map, data quality, and data security, helping enterprises achieve cost reduction and efficiency improvement, and maximize data value during the data construction and application process.

Positioning

Target Industries and Users

Suitable for industries such as government, finance, pan-Internet, industry, energy, transportation, education, cultural tourism, real estate, retail, medical, and media. The target audience includes but is not limited to:

- Technical personnel engaged in data development, algorithm development, data operation and maintenance.
- Business personnel engaged in data analysis, product operation.
- Management personnel responsible for data security compliance.
- Management personnel in charge of the company's core data assets.

Business Challenges and Pain Points

Since the outbreak of the information technology revolution and the rapid development of mobile internet in recent years, coupled with the continuous evolution and implementation of the Internet+ concept, enterprises in various industries have accumulated more and more data, leading to an urgent need for data processing and data application. However, there are many problems and challenges in this process:

- Complex infrastructure construction: Technologies such as Hadoop and Spark are numerous and complex to construct.
- Weak technical risk resistance: There is a disconnect between development and testing, leading to high error rates in data, numerous data tasks, complex dependencies, and a lack of effective change control.
- Complex data link: Open source projects often only solve specific scenarios and require the combination of multiple open-source projects to build a complete data link.
- Difficult data management: Involves cross-departmental and cross-team collaboration, complex team roles, and high communication costs.
- Difficult data governance implementation: Data quality and data security cannot be guaranteed, making upper-level applications hesitant to use the data.

- Long business build cycle: Data warehouse construction cycles are too long, taking six months to a year; data requirement responses are slow, with delays of two to three days.

Core Capabilities

WeData provides comprehensive product services for data production and consumption. The core service capabilities are as follows:

Collaboration

Based on the collaborative space around the data value chain, it enables better collaboration among different roles in the data team, breaking down silos between teams and shortening the path from raw data to data value.

- DataOps Concept
    - In large-scale task development scenarios, high concurrency allows for online execution of data development and testing.
        - Developers focus on task development and unit testing, avoiding the learning curve of business logic.
        - Orchestration personnel focus on task orchestration and scheduling configuration, with dedicated personnel shortening the implementation cycle.
    - In agile development scenarios, the integration of development and orchestration improves efficiency.
        - The data task development is completed in the process of implementing orchestration business logic.
        - It is possible to test data logic and business logic simultaneously.
- Implementation Process
    - First develop, then orchestrate: Workflow design does not block development work, and developers do not need to understand orchestration logic.
        - After completing the development space, import into the orchestration space for dedicated task orchestration.
        - Suitable for central teams working on large-scale, high-concurrency development tasks.
    - First orchestrate, then develop: Developers understand business logic, design workflows first, then develop.
        - Directly orchestrate tasks and conduct development testing in the orchestration space for more agility.
        - Suitable for teams working on small-scale or incremental tasks in an agile development model.

Efficiency

Based on the DataOps concept with agile iterations, automated processes, and tools to enhance data reliability and speed up data generation and analysis link efficiency.

- Agile and easy to use: Supports incremental code development and release; code auto-completion; visual drag-and-drop process design; online code debugging and log viewing.

- Flexible development: Development modes adapt to multiple scenarios, supporting both first develop then orchestrate and first orchestrate then develop.
- High performance and scalable: High-performance scheduling engine supporting millions of daily task schedules, integrates with multiple engines and supports engine extensions, default supports most engines with JDBC interfaces including EMR, DLC, TBDS, RDS, and more than 20 engines.
- DataOps concept
  - Supports submission, comparison, and rollback capabilities for version management to enable gray release of tasks.
  - Supports incremental release of tasks, events, parameters, and functions instead of traditional cyclical releases.
  - Agile development, rapid iteration to overall shorten the data assetization cycle.
- Implementation Process
  - After the development of data tasks, version submission is required to reflect in the workflow.
  - Different task versions can be quickly debugged within the same workflow.
  - Gray release implemented in different workflows of the same project with different task versions.
  - Incremental release by date in release management, enabling fast iteration.

Integrated

Serving multiple roles in enterprise data management, data production, data application, and data operations, providing an integrated product experience from different perspectives.

- Full-link Production Governance: Provides robust quality and security guarantees for data production and consumption through pre-planning, in-process exception blocking, post-event quality and cost analysis, and data flow security control.
- One-stop Operational Governance: Based on data self-service and the democratic concept, on a secure and stable foundation, data maps, data insights, and sharing make it easier to find, understand, analyze, and share data.

Quality

Data quality control covering pre, during, and post stages, embedded in the DataOps pipeline-style development process, ensuring comprehensive data quality enhancement.

- DataOps Concept
  - Shift from post-event quality scoring to in-process quality monitoring, integrating both code testing and data testing to ensure high-quality data analysis.
  - Shift from post-event standard benchmarking to pre-event standard implementation to ensure data quality and consistency in statistical metrics during data analysis.
- Implementation Process
  - Data tasks/workflow must pass online debugging before submission. Online debugging will automatically trigger corresponding quality monitoring tasks for the data tables.
  - Agile Data Warehouse Modeling Tool supports direct referencing of pre-defined data standards, ensuring standard implementation at the source.

- ○ Tables adhering to data standards support setting a zero-tolerance threshold for dirty data during DataInLong tasks to ensure standard compliance.

Strengths

Last updated：2024-08-22 09:53:32

As a big data development and governance platform, WeData boasts the following advantages:

Based on open-source

WeData supports open-source access and widely supports common big data open-source technologies, such as Hadoop, Hive, Spark, etc. Users with experience in open-source software can easily transfer their experience.

Ease of Use

Through the abstraction of core concepts such as workspaces, data sources, and workflows, and their organic integration in modules like data maps and data quality, users can quickly understand and seamlessly use WeData for data development and governance.

Cost reduction and efficiency improvement

WeData provides many features to help users reduce costs and improve efficiency, such as the data temperature feature in the data map, which assists in identifying infrequently used but high-cost data for cleaning or transferring; the canvas feature in workflow development enables easy organization of workflow task dependencies through drag-and-drop.

High Security and Stability

The data security module provides data access control capabilities, enabling pre-approval, interception during the process, and post-event audit for data access rights; data content control capabilities allow for business data desensitization, establishing the last line of defense for data security.
Robust and powerful high availability, CLB, and timely, multi-channel monitoring and alerts also ensure the stability of service status and task execution.

Rapid realization of big data monetization

The product helps users quickly discover and understand data through integrated operations, solves complex data pipeline development with DataOps, liberates data development productivity, and achieves rapid data R&D and delivery.

Meeting business self-service needs

Data analysts/business personnel can focus more on the business logic itself, combined with the product's self-service data discovery, exploration, and analysis capabilities, to meet the smoother data usage needs of different roles.

Reducing corporate management costs

Data development requires cross-team and multi-role collaboration, but traditional data tool architectures are fragmented and difficult to coordinate. The product provides a tool basis for different roles to perform their duties and collaborate effectively through spatial division.

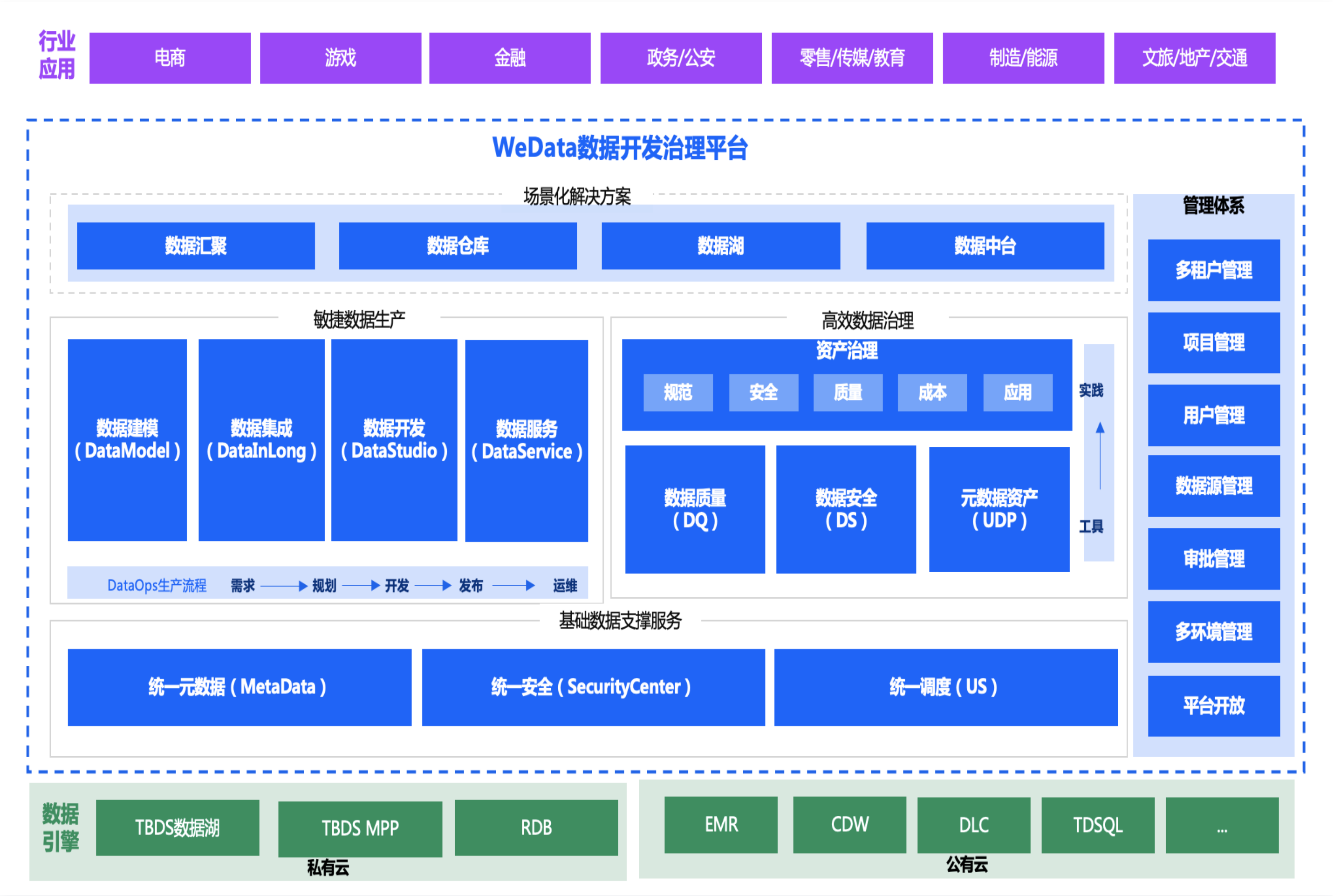
Enhancing enterprise data quality and trustworthiness

The product resolves the issue of "double-skinning" of data through pipeline operations in development, testing, and production spaces, ensuring data compliance, standardization, quality monitoring, and improvement throughout, securing data quality.

Product Architecture

Last updated：2024-08-22 09:54:03

Tencent Cloud WeData Product Solutions and Product Architecture are shown in the figure below:



The overall architecture of WeData consists of an Operational Management System and Development and Operation Tools:

- **The Operational Management System mainly includes**: Multi-tenant Management, Multi-environment Management, and Platform Openness.
  The Operational Management System provides fundamental guarantees for the isolation of data environments and the secure circulation of data. Multi-environment Management enables WeData to support different data engines, including Private Cloud TBDS and most public cloud basic products, such as EMR, Cloud Data Warehouse TCHouse-P, and DLC. Platform Openness allows WeData's capabilities to be exposed to users in the form of Open API, Open Messaging, and plugins, facilitating third-party integration and customized capabilities.
- **Development and Operation Tools include**: DataOps Agile Data Production, Full-link Data Governance, and Integrated Data Operations.
  DataOps Agile Data Production follows the DataOps philosophy, standardizing the data production process and enhancing data production efficiency through agile iterations. Full-link governance

ensures data quality, data security, and cost optimization before, during, and after the entire data production and consumption process, improving data quality and usability. Integrated Operations aim to release data value by enabling quick data discovery, understanding, insights, and application through Data Map, Data Insight, and Data Sharing, thus reducing the barriers to data usage and shortening the path to data value realization.

Product Features

Last updated：2024-08-22 10:03:42

Core features of WeData include the following:

Project Management

Achieve project isolation from the system/tenant level, providing administrators with the ability to manage user (member) permissions, underlying computing engine configuration, and execution resources for users using WeData.

Data Planning

> **Note:**
> Thank you very much for your attention and support for our product. However, the Data Planning feature is not yet open. Thank you again for your patience and understanding. We look forward to sharing our latest feature with you in the near future. Thanks!

Provides holistic data planning design capabilities, including data warehouse layering, logical model design, metric dimension definition, and data standards. This helps enterprises unify data warehouse specification design and standard definition, achieving automated transition from design to development.

- Data Warehouse Specification: Data planning based on the global business object planning and standard definition, with layer design management of models, classified and domain management according to specific business themes, forming a hierarchical structure of business tags.
- Model Design: Definition and design of logical models and entity relationships, including definition, copy, modification, deletion, import/export, version management abilities, and establishing associative mapping with physical models and metric dimensions to achieve automatic synchronization from design to development.
- Standard Management: Includes standard content management and benchmark task management. By designing standard rules and configuring tasks, it standardizes data values, libraries, table structures, table names, metric dimension tags, and other levels.
- Business Definition: Metric/Dimension dictionary, lifecycle definition management of base/derived metrics, dimension criteria (common dimensions, business constraints, time cycles, degenerate dimensions), and establishing associations with models for automatic metric production code generation.

DataInLong

Lightweight operation, visualized process, and open capabilities of DataInLong, supporting high-speed, stable mass data synchronization between heterogeneous data sources in complex network environments.

- Full-scenario Synchronization: Includes real-time synchronization and offline synchronization.

- Multi-type Heterogeneous Data Sources: Supports 30+ data sources, providing star schema support for random read-write matching.
- Transformation
  - Data Level: Perform content transformation on synchronized data, such as data filtering, Join, etc.
  - Field Level: Provide single-field transformation processing, including custom data field, format conversion, date format conversion, etc.
- Task and Data Monitoring
  - Read and Write Metrics: Support real-time statistics of task read-write metrics, including total read-write volume, speed, throughput, and dirty data, etc.
  - Monitoring and Alarm: Support task and resource monitoring, covering multi-channel alarms including SMS, Email, and HTTP.

Data Development

Through rigorous CI/CD process specifications and automation capabilities for test release and operation and maintenance, shorten the path from raw data processing and operation and maintenance to business application data, improving efficiency while ensuring data quality.

- Online Code Development: Support code development, allowing easy drag-and-drop orchestration of task workflows, and also support visual presentation of large-scale task orchestration.
  - Code Development: Support online code development, debugging, and version management for tasks such as HiveSQL, SparkSQL, JDBCSQL, Spark, Shell, MapReduce, PySpark, Python, TBase, DLC SQL, DLCSpark, TCHouse-P, Impala, etc.
  - Task Testing: Support task and workflow testing and version management.
  - Development Assistance: Provide parameter configuration at three levels of granularity: project, workflow, and task, supporting time parameter operations and function parameters.
  - Version Management: Support version management of events, functions, tasks, and parameters.
  - Code Management: Provide unified code management, import, and export.
- Orchestration and Scheduling: Perform flow orchestration and submission scheduling of tasks.
  - Scheduling Method: Support periodic, one-time, and event-triggered scheduling, with periodic scheduling configured in a crontab manner.
  - Dependency Strategy: Support task self-dependency and workflow self-dependency.
  - Cross-cycle Dependency Configuration: Provide cross-cycle dependency configuration and self-definition dependency configuration. The scope of upstream and downstream dependency instances can be selected as needed by self-defining.
  - Batch Orchestration: Offer the ability to batch create tasks and dependencies via Excel, speeding up task dependency orchestration efficiency.
- Release and Operation: Publish completed development tasks to the production environment as needed, and provide unified monitoring and operation for the tasks.
  - Task Release: Support releasing the development outcomes online.

  - Monitoring and Operation: Perform flow orchestration and submission scheduling of tasks.
- Analysis and Exploration: Enhance task collaborative development efficiency through intelligent and user-friendly data development methods, helping users clearly view the task processing steps and significantly improving data ad-hoc exploration efficiency.
  - Online Editing: Provide a visual interactive analysis IDE.
  - Run: Offer visualized execution information.
  - Development Assistance: Provide efficiency tools for development assistance.

Data Quality

Comprehensive data quality auditing capabilities are provided through flexible rule configuration, comprehensive task management, and multi-dimensional quality assessment across all stages of the data lifecycle from ingestion, integration, processing, to consumption.

- Multi-source Data Monitoring: Support monitoring data sources and engine types including EMR Hive, Spark, DLC (public cloud), TCHouse-P, TBDS, Gbase (private cloud), etc., offering the ability to perform full-scale data validation across multiple sources.
- Rich Rule Templates: Currently provides 6 dimensions and 56 industry-standard built-in table-level and field-level rule templates, realizing true out-of-the-box usability and significantly improving quality control workflow efficiency, helping users perceive data changes and issues arising during the ETL process from various dimensions.
- Flexible Quality Control Configuration: Support three rule creation modes — system quality rule templates, self-defined templates, and self-defined SQL. Parameters can be adjusted according to business needs, task execution strategies can be configured, and full-link quality control validation can be easily achieved.
- Global Link Guarantee: Supports two execution methods: associated production scheduling and offline periodic detection, providing pre-, mid-, and post-event full-link data assurance operational capability. Timely alerts and block interceptions prevent dirty data from spreading downstream.
- Multi-dimensional Governance Visualization: The Quality Overview and Quality Report modules provide users with a global perspective, allowing them to fully understand the status of quality tasks, alarm blockage trends, and quality scores across various dimensions. This helps quickly identify and locate issues, and understand quality improvement effects.

Data Security

Provides centralized data security control and a collaborative mechanism to ensure effective data flow under secure conditions.

- Unified Data Security Management: Deeply integrates security policies with bound computational storage engines, unifying data access and simplifying data use processes.
- Permission Approval: Bridges the Ranger Permission Policy System, achieving accountability to individuals and table-level permission control capability. Provides channels for permission application and approval, securely opening data access control capability.

Data Operations

Based on powerful underlying metadata capabilities, provides data directory, lineage analysis, popularity analysis, asset rating, business classification, tag management, and other data asset services, effectively enhancing users' understanding, control, and cooperation abilities with enterprise-level massive data.

- Data Discovery: Unified metadata collection and management.
- Data Overview: Provides an overview of data assets, including basic information such as items, tables, storage volume, and data type coverage, as well as features like data panorama and hot ranking.
- Data Directory: Supports quick search and location of global table-level and field-level data; table details provide comprehensive technical and business information, as well as data lineage, temperature, quality, production and change, preview, and other features.
- Database Table Management: Supports management of global database tables.
- Business Classification: Supports creating and managing topic categories, data warehouse layering, and business tags according to business needs, and bulk classification and layering operations on data tables.

Data Service

Provides capabilities covering the full lifecycle of APIs, including API production, API management, and API market, helping enterprises unify the management of internal and external API services and build a unified data service bus.

- Quick API Production.
- API Management and Operation.
- API Secure Invocation.

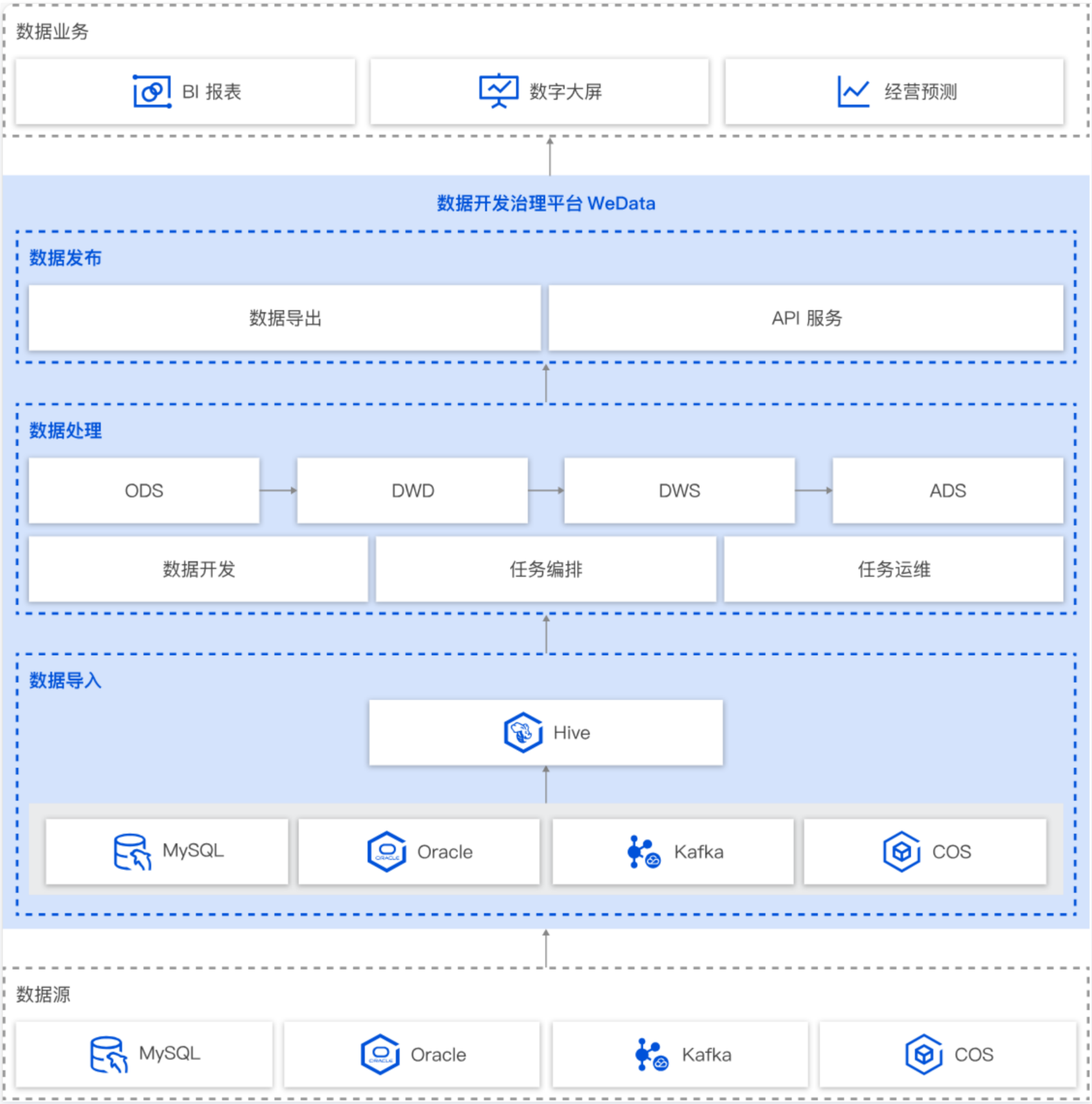
Application Scenario

Last updated：2024-08-22 15:01:59

WeData, as a versatile data tool product, has a wide range of application scenarios. The following describes several typical scenarios.

Enterprise Data Warehouse Construction

Tencent Cloud WeData provides comprehensive data processing features, covering the entire data warehouse construction chain. From heterogeneous data source ingestion to support for data development, task orchestration, and task operation and maintenance through a rich set of big data components, it ensures standardized production across various levels and data domains within the data warehouse, guaranteeing data standardization, integrity, timeliness, and more. Finally, by exporting data or through API Services, the standardized data produced by the data warehouse is applied to various types of enterprise data services, empowering businesses with data.



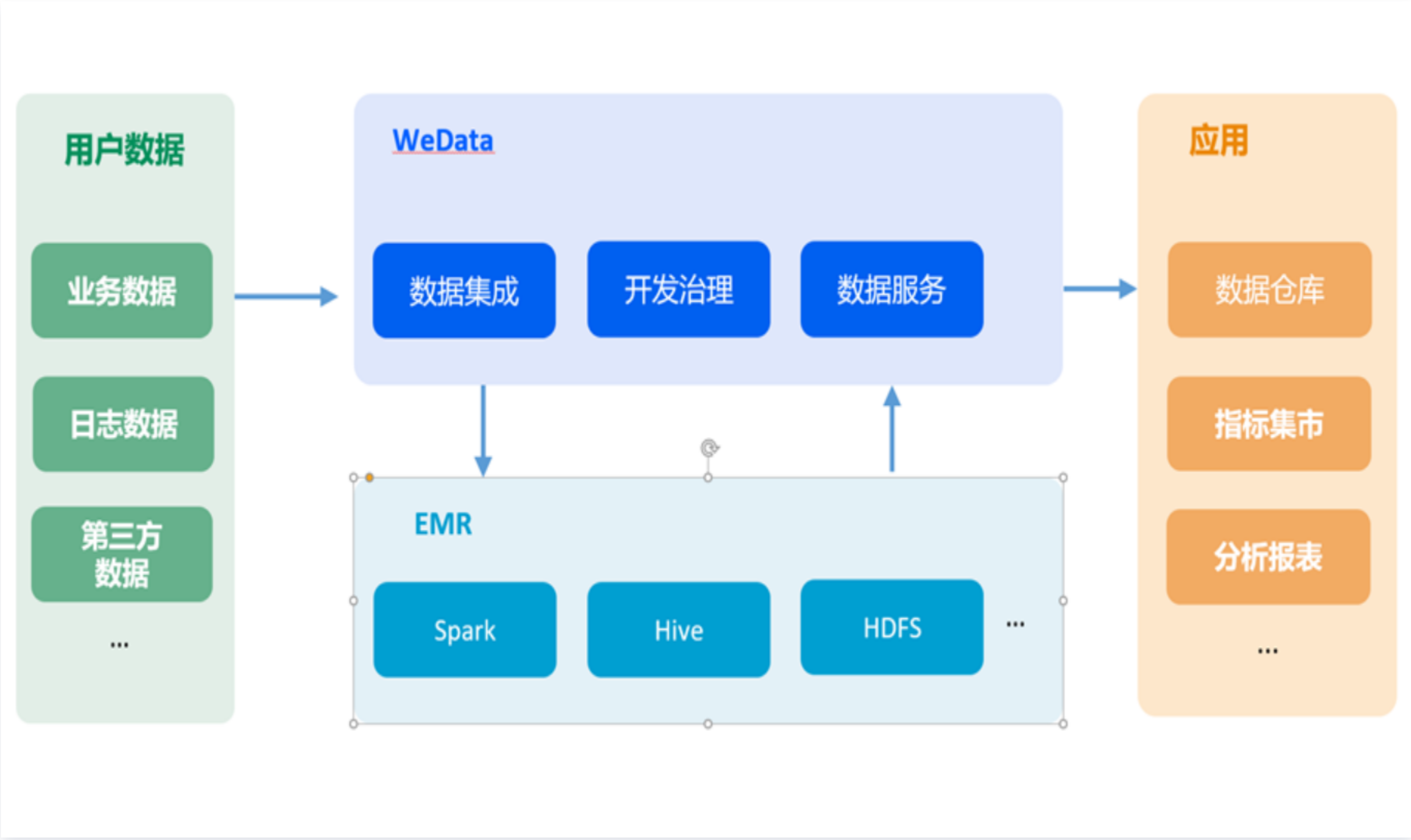
Data Asset Governance

Data assets are a crucial part of an enterprise's core assets, especially for data-driven companies, where they represent the most vital assets. However, the inherent heterogeneity, complexity, and high cost of big data, combined with the diversity and complexity of enterprise business, lead to issues such as poor quality, high costs, difficult comprehension, and lack of security for the ready data processed from raw data, failing to be directly used by the business as expected.

WeData, based on its core data processing capabilities, offers a range of data governance capabilities, helping enterprises manage all aspects of their data, ensuring that data can be confidently applied to business, efficiently creating value for the business.

Offline Data Warehouse/Data Platform Solution—WeData + EMR

Using WeData's DataInLong and development capabilities, massive data from the business side is aggregated. Then, leveraging the powerful PB-level data computing and storage capabilities of EMR, it provides a high-performance enterprise-level offline data warehouse solution. Additionally, WeData can perform unified security management and data quality governance on the data warehouse, delivering a unified and standardized data warehouse system, supporting users in further extracting data value.

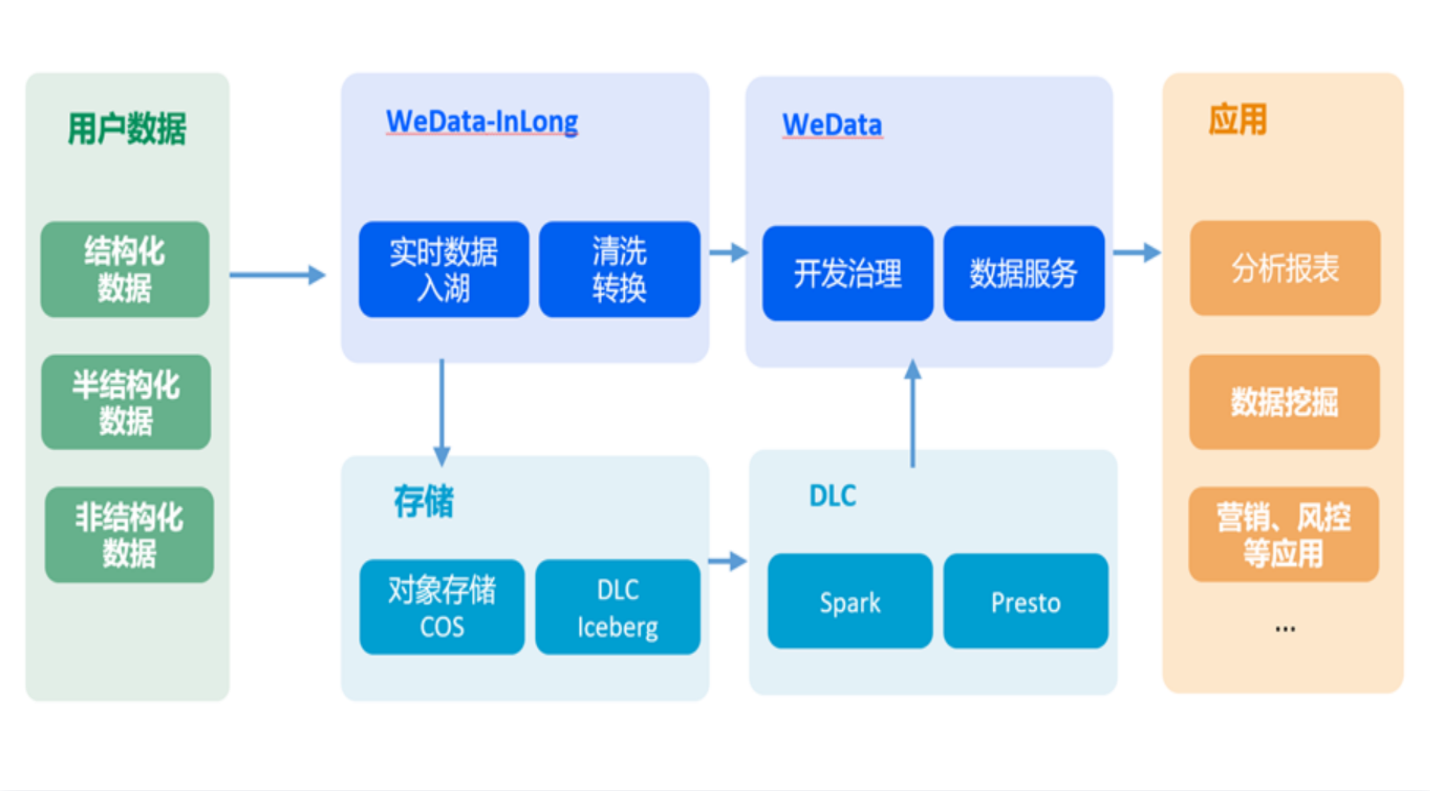


Real-time Data Warehouse Solution—DataInLong/Oceanus + TCHouse-P

Built on TCHouse-P's large-scale parallel processing and OLAP analysis capabilities, Oceanus provides real-time data processing, helping users quickly aggregate and collect data to construct lightweight, low-cost, AS capable real-time data warehouses. It can be accompanied by UDP for data catalog management. When users require advanced data governance capabilities, Oceanus can smoothly upgrade to WeData, offering further data processing, quality governance, and exploratory analysis capabilities.

Fully Managed Data Lake Solution—DataInLong/WeData + DLC

Built on the massive big data analysis architecture of DLC's compute-storage separation, DataInLong's real-time data ingestion and cleaning capabilities achieve unified aggregation of diverse heterogeneous data from the business side, constructing a low-cost, highly elastic, fully managed data lake solution. DataInLong can smoothly upgrade to WeData, offering advanced data processing, quality governance, and exploratory analysis capabilities to the users.



Data Visualization Solution—WeData + BI

Based on the data warehouse metrics management and data services provided by WeData, connecting data to the "last mile" of business, and in coordination with BI's capabilities for DIY report building and visual analysis, a more convenient, simple, and intuitive presentation method is offered to users making use of data. Simultaneously, with the help of WeData's powerful data asset governance capabilities, the standardization, accuracy, and consistency of data are ensured, making data-assisted business decisions more reliable.