Data Development and Governance Platform

Product Introduction

Product Documentation

Contents

Product Introduction
Product Overview

Last updated：2024-07-15 17:16:52

Data Development and Governance Platform WeData (hereinafter referred to as WeData) is a cloud-based one-stop data development and governance platform, integrating a full-link DataOps data development capability that includes data integration, data development, and task operations, as well as a series of data governance and operation capabilities such as data maps, data quality, and data security. It aims to help enterprises reduce costs and increase efficiency in data construction and application, maximizing the data value.

Positioning

Target Industries and Users

Applicable to many industries including government, finance, pan-internet, industry, energy, transportation, education, cultural tourism, real estate, retail, health care, and media. The audience includes, but is not limited to:
Technical personnel engaged in data development, algorithm development, and data operations.
Business personnel engaged in data analysis and product operation.
Management personnel responsible for data security compliance.
Management personnel who control the company's core data assets.

Business Challenges and Pain Points

Since the explosion of the information technology revolution, and with the recent rapid development of the mobile internet, along with the continuous deepening evolution and implementation of the Internet+ concept, enterprises across various industries have accumulated more and more data, thereby deriving an urgent need for data processing and application. However, this process also faces many problems and challenges:
Complex infrastructure construction: A dazzling array of big data technologies such as Hadoop, Spark, etc., make construction complex.
Weak technical risk resistance: Disjointed development and testing, high probability of data errors, numerous data tasks, complex dependencies, and lack of effective change control.
Complex data links: Open source projects often solve specific scenarios, requiring combining multiple open source projects to build complete data links.
Difficult data management: Involves cross-departmental and cross-team collaboration, complex team roles, and high communication costs.
Difficult data governance implementation: Data quality, data security, etc., cannot be guaranteed, and upper-level applications dare not use data confidently.

Long business construction period: Data warehouse construction cycle is too long, taking half a year or even a year; slow response to data needs, with two to three days of delayed response.

Core Capabilities

WeData offers comprehensive product services for data production and consumption, with key service capabilities as follows:

Collaboration

Based on the collaborative space around the data value chain, different roles in the data team can collaborate better, breaking down the silos between teams and shortening the path from raw data to data value.
DataOps Philosophy
In large-scale task development scenarios, it enables high-concurrency online execution of data development and testing.
Developers focus on task development and unit testing, avoiding the learning cost of business logic.
Orchestration personnel focus on task orchestration and scheduling configuration, with dedicated personnel for specific tasks to shorten the implementation cycle.
In agile development scenarios, integration of development and orchestration can improve efficiency.
During the process of implementing orchestration business logic, data task development is completed.
Allows simultaneous testing of data logic and business logic.
Implementation Process
Develop first, orchestrate later: Workflow design does not block development work, developers do not need to understand orchestration logic.
After completing the development space, import into the orchestration space, with dedicated personnel for task orchestration.
Suitable for centralized teams with large-scale and high-concurrency development tasks.
Orchestrate first, develop later: Developers understand the business logic, design workflows first, then develop.
Directly orchestrate and develop test tasks in the orchestration space, making it more agile.
Suitable for small-scale or incremental tasks agile development mode for subteams.

Efficiency

Based on DataOps agile iteration, automated processes, and tools to enhance data reliability, it can accelerate the efficiency of data production and link analysis.
Agile and Easy to Use: Supports incremental code development and release; supports automatic code completion; offers visual drag and drop for process design; supports online code debugging and log viewing.
Flexible Development: The development mode is adaptable to multiple scenarios, supporting both development before orchestration and orchestration before development.

High Performance and Scalability: A high-performance scheduling engine supports ten million tasks scheduling per day, can interface with various engines and support engine extensions, and supports more than 20 JDBC interface engines, including EMR, DLC, TBDS, RDS, and others by default.

DataOps Philosophy

Supports version management capabilities such as submission, comparison, and rollback to support the gray release of tasks.

Supports the incremental release of tasks, events, parameters, functions, rather than the traditional periodic release.

Agile development, rapid iteration, to shorten the overall data assetization cycle.

Implementation Process

After data task development is complete, version submission is required to reflect in the workflow.

Different version tasks can be quickly debugged in the same workflow.

Different projects with the same workflow based on different task versions achieve gray release.

Incremental releases are done by date in release management, allowing for rapid iterations.

Integration

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

End-to-end Production Governance: Provides strong quality and security assurance for data production and consumption through pre-planning, in-process exception blocking, post-event quality and cost analysis, and secure control over data circulation.

One-stop Operational Governance: Based on the concepts of data self-service and democratization, it makes searching, understanding, analyzing, and sharing data easier on a stable and secure basis, through data mapping, insights, and sharing.

Quality

Data quality control throughout the pre-, during, and post-phases, integrated into the DataOps pipeline development process to ensure comprehensive data quality improvement.

DataOps Philosophy

Transitioning from post-event quality scoring to in-process quality monitoring, integrated testing comprises both code testing and data testing to ensure high-quality data analysis.

Transitioning from post-event standard benchmarking to pre-event standard implementation to ensure data quality and consistency in statistical caliber during data analysis.

Implementation Process

Data tasks/workflows require online debugging before submission, automatically initiating quality monitoring tasks for corresponding data tables.

Agile data warehouse modeling tools support direct referencing of pre-defined data standards during modeling to ensure early compliance and prevent failure at the source.

Tables following data standards support setting a zero-tolerance threshold for dirty data during data integration tasks to ensure compliance.

Product advantages

Last updated：2024-07-15 17:17:43

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

Based on Open Source

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

Ease of Use

By abstracting core concepts such as workspaces, data sources, and workflows, and organically integrating modules like data maps and data quality, WeData enables users to quickly understand and smoothly use the platform for data development and governance.

Cost Reduction and Efficiency Improvement

WeData provides numerous features to help users reduce costs and improve efficiency. For instance, the data temperature feature in the data map helps identify infrequently used data that incur high costs for cleanup or migration; the canvas feature in workflow development allows easy organization of workflow task dependencies through drag-and-drop controls.

High Security and Stability

Data security modules offer data access control capabilities, enabling pre-approval, interception during the process, and post-event auditing of data access permissions; data content control capabilities allow for business data desensitization, establishing the last line of defense for data security.
With rich and powerful high availability, load balancing, as well as timely and multi-channel monitoring and alarms, WeData ensures the full stability of service states and task operations.

Accelerate 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 research and development as well as delivery.

Meet Self-service Needs for Business

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.

Reduce Enterprise Management Costs

Data development requires cross-team and multi-role collaboration, but the architecture of traditional data tools is relatively fragmented and difficult to coordinate. The product facilitates role-specific duties and effective collaboration by dividing spaces.

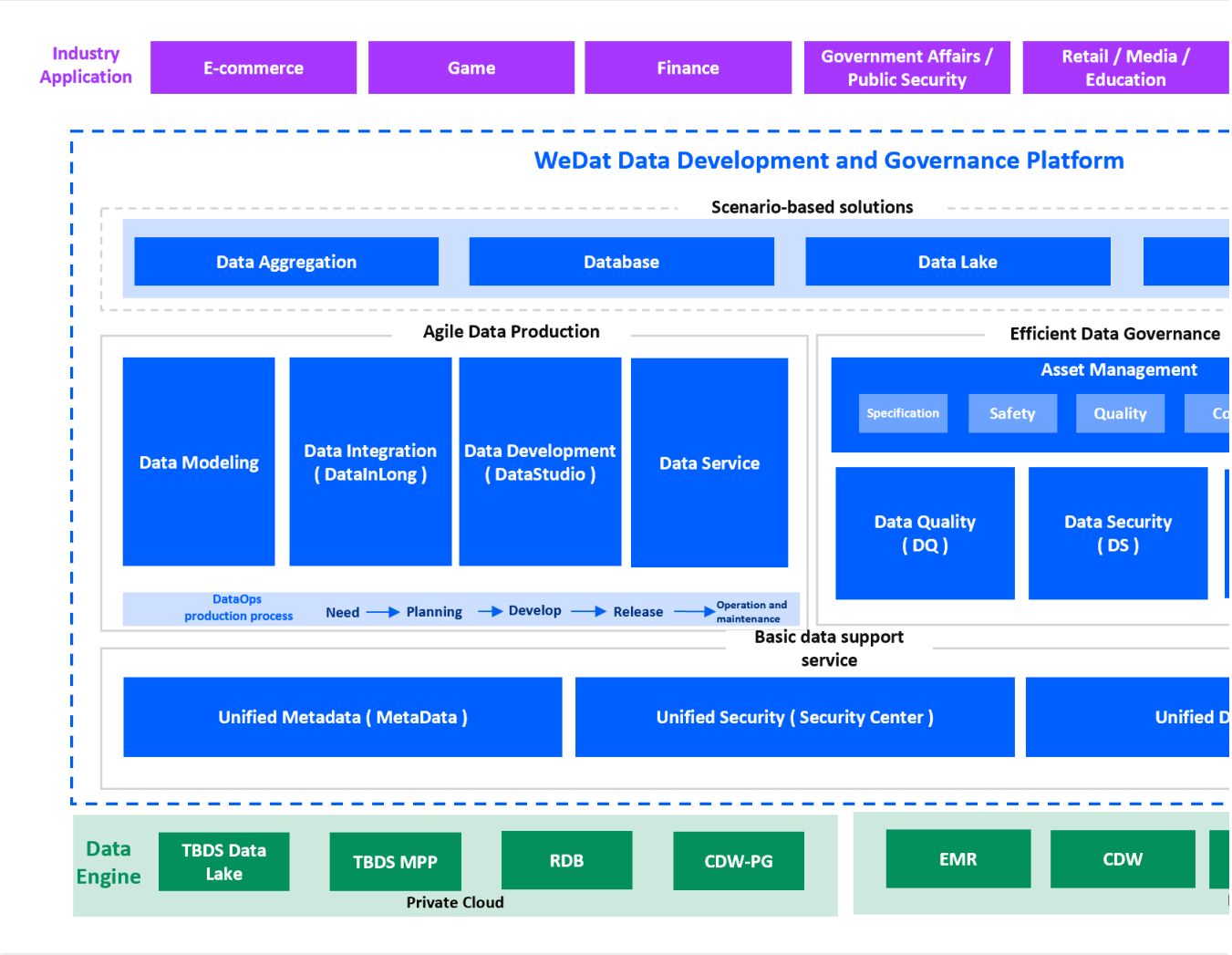
Enhance Enterprise Data Quality Trustworthiness

The product solves the problem of data discrepancy through a pipeline operation across development, testing, and production environments, ensuring data compliance, standardization, quality monitoring, and improvement throughout, to guarantee data quality.

Product Architecture

Last updated：2024-07-15 17:19:56

The Tencent Cloud Data Development and Governance Platform WeData, presents its product solutions and architecture as follows:



The overall architecture of WeData consists of the operational management system and development and operations tools:

The operational management system mainly includes multi-tenant management, multi-environment management, and the open platform, etc.

The operational management system provides foundational support for data environment isolation and secure data circulation. Multi-environment management enables WeData to interface with different data engines, including the private cloud TBDS and most public cloud-based products, such as EMR, Cloud Data Warehouse, DLC, etc. The open platform allows for the exposure of WeData's capabilities through open APIs, open messages, and plugins, facilitating third-party integration and the realization of customized capabilities.

Development and operations tools include DataOps agile data production, end-to-end 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 iteration. End-to-end governance ensures data production and consumption throughout the process—before, during, and after—through data quality, data security, and cost optimization, improving data quality and availability. Integrated operations aim at unleashing data value, enabling users to rapidly complete data discovery, data understanding, data insight, and data application through data map, data insight, and data sharing, lowering the barriers to data use, and shortening the path to data value realization.

Product Features

Last updated：2024-07-15 17:21:13

The core features of WeData include the following:

Project Management

Implement project isolation from the system/tenant perspective, allowing managers to control permissions for users (members) using WeData, configure the underlying computation engine, and manage execution resources.

Data Planning

Provides capabilities for overall data planning and design, including data warehouse layered categorization, logical model design, metric dimension definition, and data standards, helping enterprises unify data warehouse standard design and standard definitions, and automating the transition from the design phase to the development phase.
Data Warehouse Standards: Data warehouse planning is based on a global approach to business object unified planning and standard definitions, layering model management, classifying and managing different domains according to specific business themes, and forming hierarchical business tags.
Model Design: Includes definition and entity relationship design for logical models, encompassing definition, copy, modification, deletion, import/export, and version management capabilities while establishing associations with physical models, metric dimension association mapping, achieving automatic synchronization of the model from the design phase to the development phase.
Standard Management: Includes standard content management and benchmarking task management. Through the design of standard rules and task configuration, standardization of data values, libraries, table structures, table names, and metric dimension tags is achieved.
Business Definition: Metric/Dimension dictionary manages base/derived metrics and dimension conditions (normal dimensions, business limited, time cycles, degenerate dimensions) throughout their lifecycle, and establishes associations with models, achieving automatic generation of metric production code.

Data Integration

Lightweight operations, visualized processes, and open capabilities of data integration support high-speed and stable massive data synchronization between rich heterogeneous data sources in complex network environments.
All-Scenario Synchronization: Includes real-time and offline synchronization.

Multiple Types of Heterogeneous Data Sources: Supports 30+ data sources, providing star structure support for read-write combinations.
T-Transformation
Data Level: Content transformations during synchronization, such as data filtering, Join, etc.
Field Level: Offers single field transformation processing, including custom data fields, format conversion, time format conversion, etc.
Task and Data Monitoring
Read and Write Metrics: Supports real-time statistics on task read and write metrics, including total read and write volume, speed, throughput, dirty data, etc.
Monitoring and Alarms: Supports task and resource monitoring, covering SMS, email, HTTP, and other multi-channel alarms.

Data Development

Through stringent CI/CD process standards and the enhancement of automated test, release, and operation and maintenance capabilities, it shortens the path from raw data processing operations to business application data, improving efficiency while ensuring data quality.
Online Code Development: Supports code development, facilitating easy drag-and-drop orchestration of task workflows, as well as supporting the visual orchestration presentation of large-scale tasks.
Code Development: Supports online code development, debugging, and version management for tasks including HiveSQL, SparkSQL, JDBCSQL, Spark, Shell, MapReduce, PySpark, Python, TBase, DLC SQL, DLCSpark, CDW PostgreSQL, Impala, etc.
Task Testing: Supports testing and version management for tasks and workflows.
Development Assistance: Offers parameter configuration at three granularity levels: project, workflow, and task; supports time parameter calculation and function parameters.
Version Management: Supports version management for events, functions, tasks, and parameters.
Code Management: Provides unified management, import and export for code.
Orchestration Scheduling: Manages task process orchestration and submission scheduling.
Scheduling Method: Supports cyclic, one-time, and event-triggered scheduling, offering crontab configuration for cyclic scheduling.
Dependency Policy: Supports task self-dependency and workflow self-dependency.
Cross-cycle Dependency Configuration: Provides cross-cycle dependency configuration and custom dependency configuration, with the range of upstream and downstream dependency instances being selectable as needed.
Batch Orchestration: Provides the capability to create tasks and dependencies in bulk via Excel, speeding up the efficiency of task dependency orchestration.
Release and Operations: Allows for the deployment of developed tasks to the production environment as needed, and provides unified monitoring and operations of tasks.

Task Release: Supports the release and deployment of development outcomes.
Monitoring and Operations: Manages task process orchestration and submission scheduling.
Analytical Exploration: An intelligent and user-friendly data development approach enhances the efficiency of collaborative task development, helps users view the task processing procedure, and effectively improves the efficiency of ad-hoc data exploration.
Online Editing: Provides a visual interactive analysis IDE.
Run: Offers visualization of execution information.
Development Assistance: Provides efficiency tools for development assistance.

Data Quality

Through flexible rule configuration, comprehensive task management, and multidimensional quality assessment, it provides comprehensive data quality audit capabilities at every stage of the full lifecycle from data access, integration, and processing to consumption.
Multi-source Data Monitoring: Supports monitored data sources and engine types, including EMR Hive, Spark, DLC (public cloud), CDW-PG, TBDS, Gbase (private cloud), etc., offering full data validation capabilities for multi-source data.
Rich Rule Templates: Currently offers 56 common industry-standard table-level and field-level built-in rule templates across six dimensions, truly enabling it out-of-the-box, significantly improving quality control workflow efficiency, and helping users perceive data changes and problematic data generated during the ETL process from various dimensions.
Flexible Quality Control Configuration: Supports system quality rule templates, custom templates, and custom SQL among three rule creation modes, allowing for adjustment of parameters and configuration of task execution policies to easily achieve end-to-end quality control validation.
Global Link Guarantee: Supports related production scheduling and offline periodic detection as execution methods, providing pre-, during, and post-full-link data guarantee operational capabilities, with timely alarms, and interdiction blocking, preventing the downstream spread of dirty data.
Governance Multi-dimensional Visibility: Quality overview and quality report modules provide users with a global perspective, allowing users to be familiar with the operation status of quality tasks, alarm-blocking trends, and quality scores across various dimensions, quickly identify and locate problems, and understand the quality improvement effect.

Data Security

Provides centralized data security management and control and collaboration mechanisms, ensuring the secure and effective circulation of data.

Unified Data Security Management and Control: Integrates security policies deeply with the bound storage and computation engines, unifying data access and simplifying the data usage process.
Permission Approval: Integrates with the ranger permission policy system to realize responsibility assigned to individuals and permission control capability down to the table data granularity. Offers permission application and approval channels to safely open data access control.

Data Operations

Based on powerful underlying metadata capabilities, it offers data asset services such as data catalog, lineage analysis, popularity analysis, asset rating, business classification, and tag management, effectively enhancing the users' understanding, control, and collaboration ability with enterprise-level mass data.
Data Discovery: Unified metadata collection and management.
Data Overview: Provides statistics of data assets, including items, tables, storage volume, data type coverage, as well as data panorama and popular ranking features.
Data Catalog: Supports global database table level, field-level quick search and localization; table details provide full data technology, business information, and features like data lineage, temperature, quality, output and changes, preview, etc.
Database Table Management: Supports management of global database tables.
Business Classification: Supports creating and managing thematic categories, data warehouse layering, and business tags based on business needs, and conducting batch categorization and hierarchical operations on database tables.

Data Services

Provides capabilities covering the full lifecycle of APIs, including API production, API management, and API marketplace, helping enterprises to unify the management of internal and external API services and build a unified data service bus.
Quick API production.
API management and operations.
API security call.

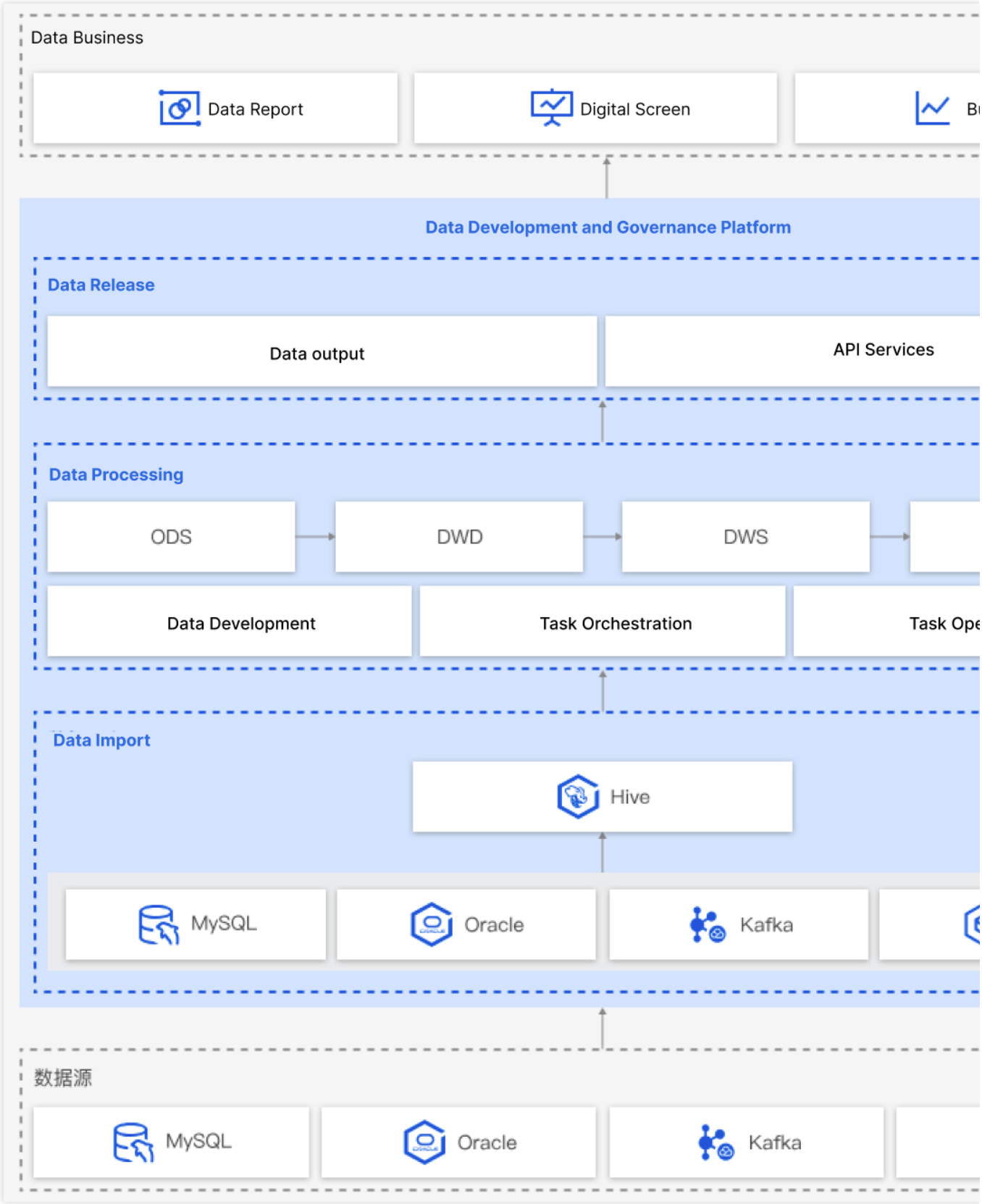
Application scenarios

Last updated：2024-07-15 17:26:35

As a universal data tool product, WeData has a very broad range of scenarios. The following explains some typical scenarios.

Enterprise-level Data Warehouse Construction

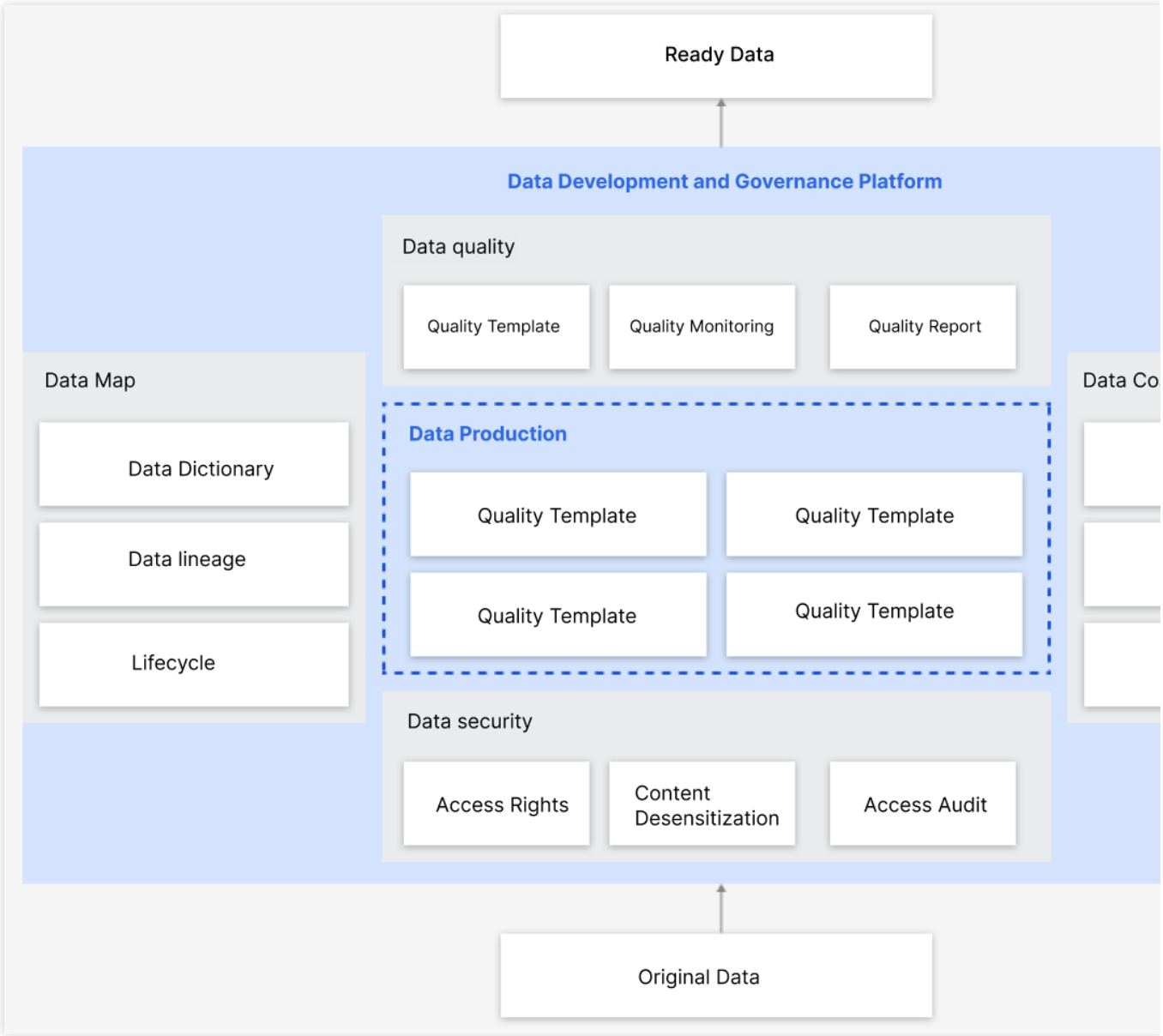
WeData provides complete data processing features, covering the full link of data warehouse construction. From importing data from heterogeneous data sources to supporting the entire process of data development, task orchestration, and task operations through the support of a variety of big data components, it standardizes the production of each layer and data domain of the data warehouse to ensure data standardization, integrity, and timeliness. Finally, by exporting data or via API services, the standardized data produced by the data warehouse is applied to various types of data businesses of the enterprise, empowering the business with data.



Data Asset Governance

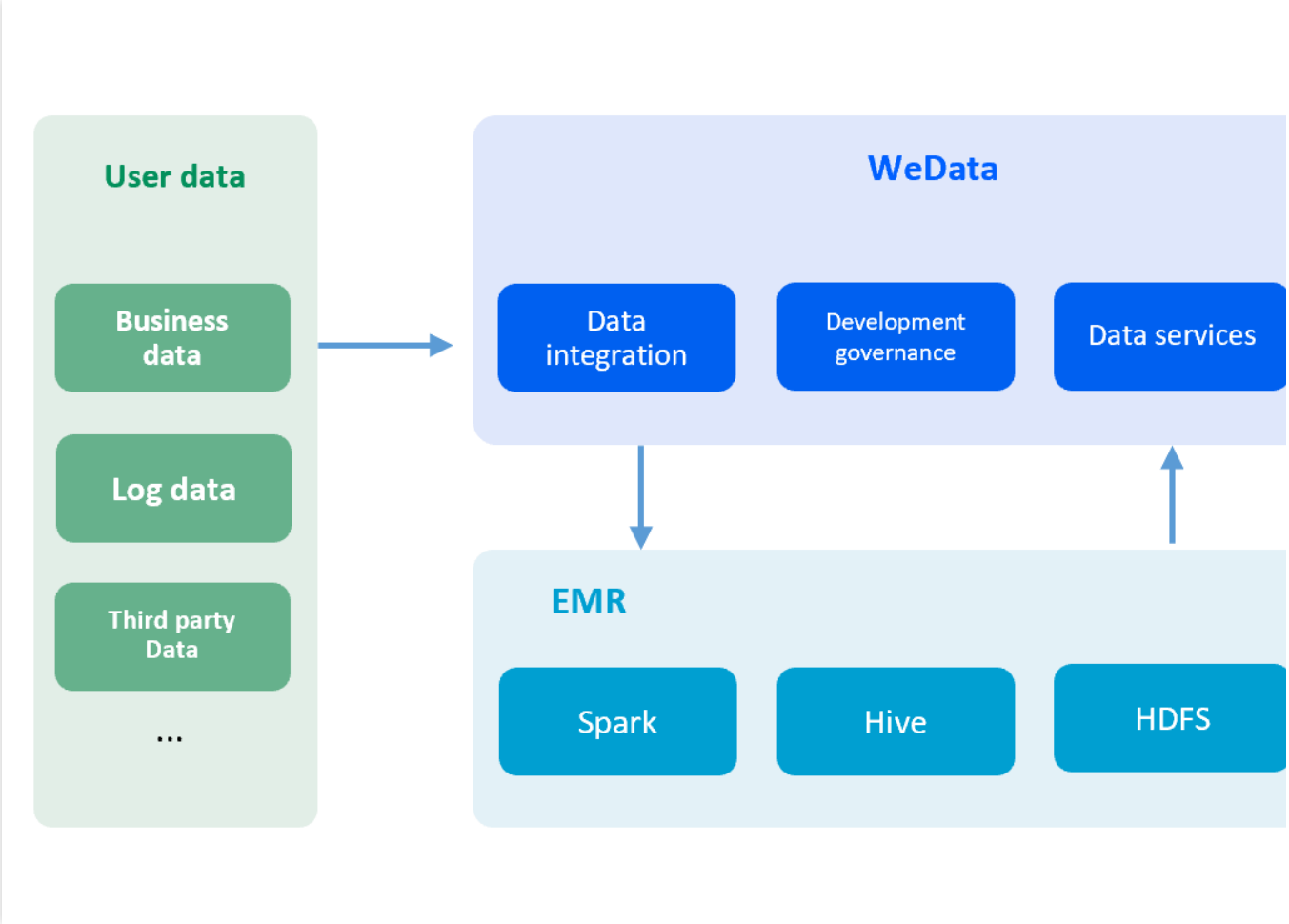
Data assets are an important part of an enterprise's core assets and are the most core assets for data-driven companies. However, the inherent heterogeneity, complexity, and high costs of big data, along with the diversity and complexity of enterprise business, make the ready data processed from raw data exhibit problems such as poor quality, high costs, difficulty in understanding, and insecurity, preventing it from being directly used by the business as expected.

Based on core data processing capabilities, WeData provides a series of data governance capabilities to help enterprises manage all aspects of data, enabling data to be confidently and boldly applied to business, efficiently creating value for business.



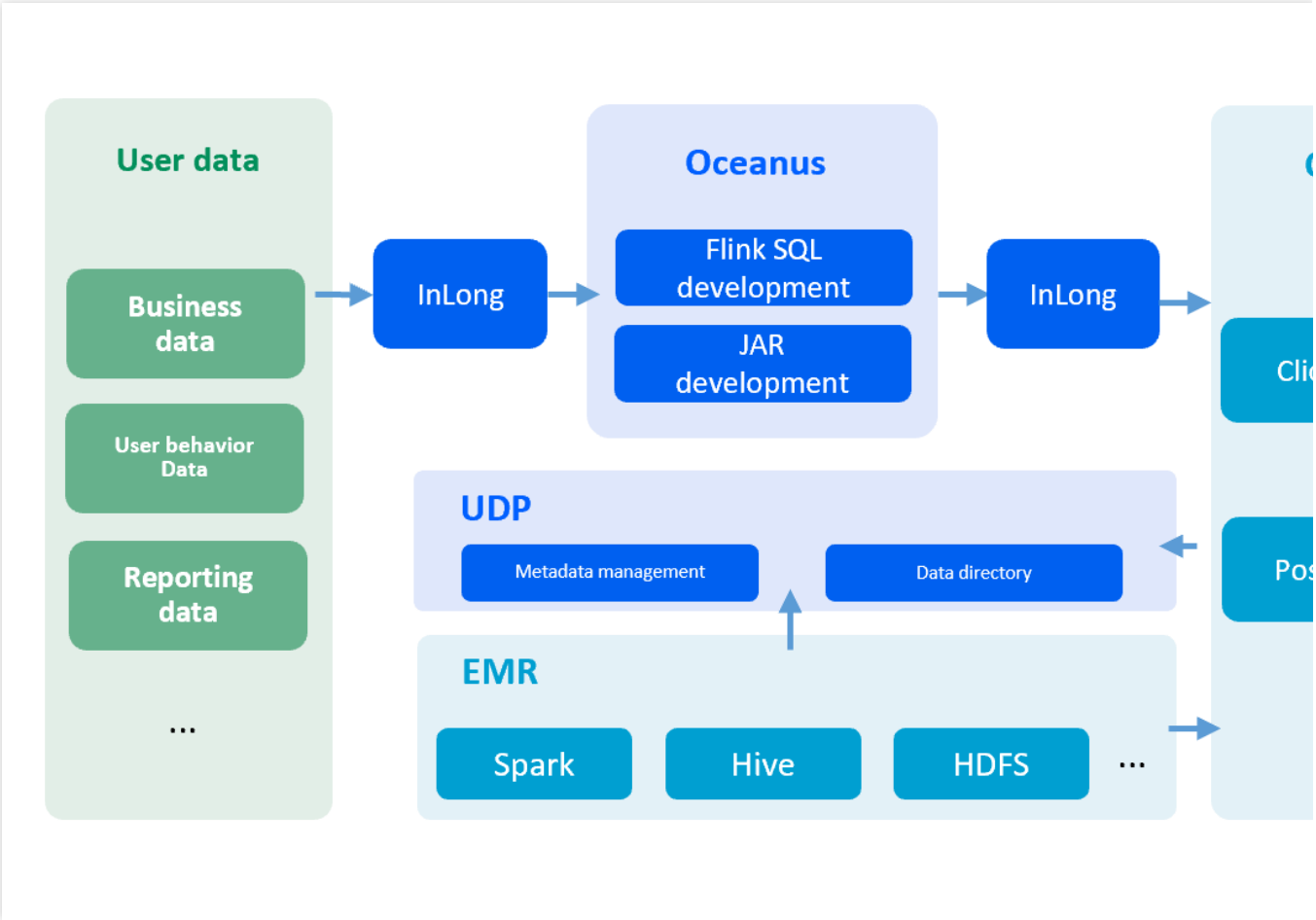
Offline Data Warehouse/Data Platform Solution—WeData+EMR

Based on data integration and development capabilities provided by WeData, it aggregates massive data from the business side. Then, leveraging EMR's powerful PB-level data computing and storage capabilities, it provides a high-performance enterprise-level offline data warehouse solution and can unify security management and data quality governance of the data warehouse through WeData, providing a unified, standardized data warehouse system to support users in further mining data value.



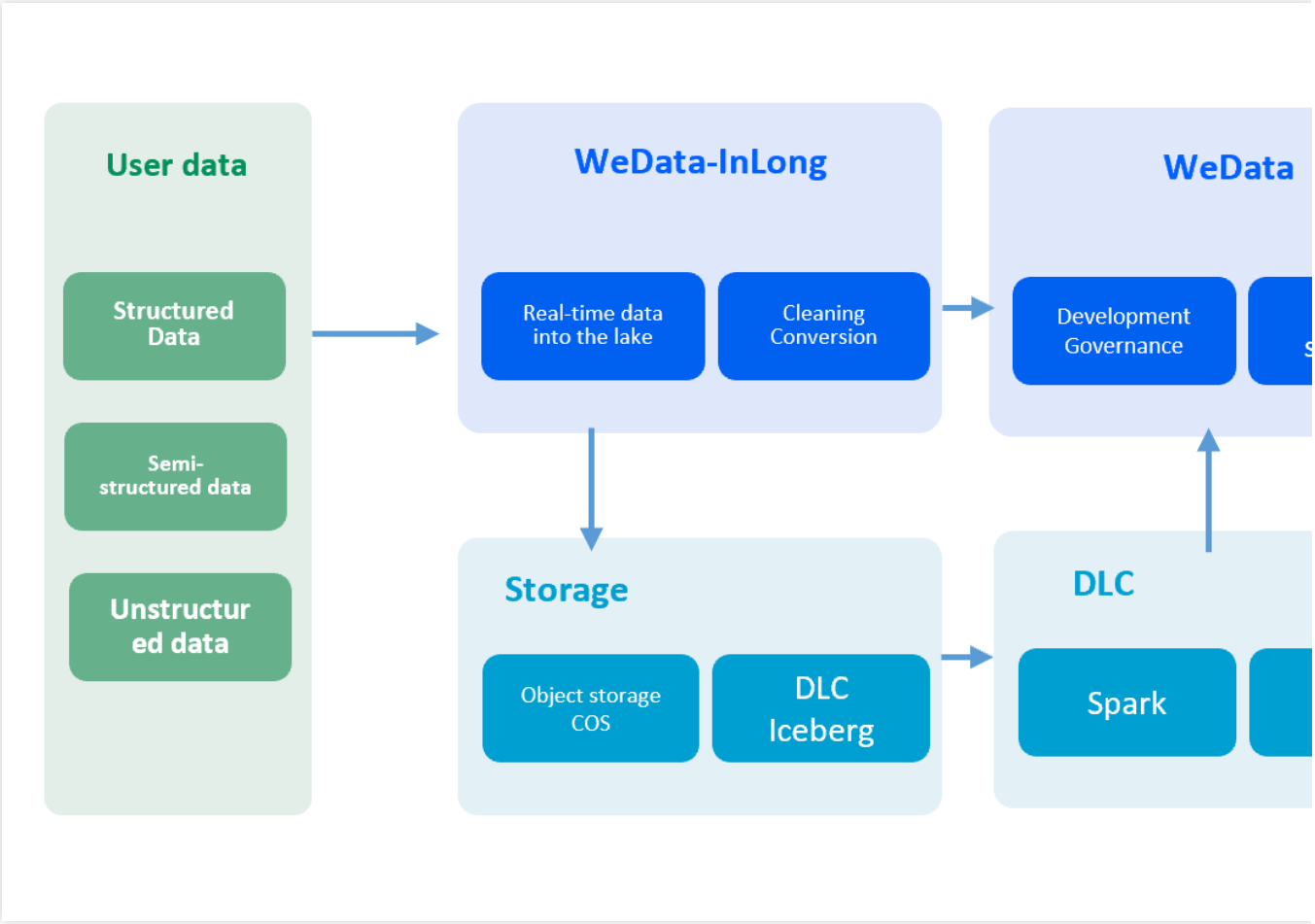
Real-time Data Warehouse Solution—DataInLong/Oceanus+CDW

On top of the massive parallel processing and OLAP analytical capabilities provided by cloud data warehouse CDW-PG/CK, real-time data processing is performed by Oceanus, helping users to quickly aggregate and collect data, and build a lightweight, low-cost, scalable real-time data warehouse. It can be paired with UDP for data catalog management. When further data governance capability is needed, Oceanus can be smoothly upgraded to WeData, providing further data processing, quality governance, and exploratory analysis capabilities.



Fully Managed Data Lake Solution—DataInLong/WeData+DLC

On top of the computation-storage separated massive big data analysis architecture provided by DLC, through DataInLong's real-time data ingestion and cleansing capabilities, it achieves unified aggregation of diverse heterogeneous data from the business side, building a low-cost, highly elastic, fully managed data lake solution. DataInLong can be smoothly upgraded to WeData, providing further data processing, quality governance, and exploratory analysis capabilities.



Data Visualization Solution—WeData+BI

Based on the data warehouse metric management and data service capabilities provided by WeData, connecting data to the "last mile" of business and coupling with BI's self-service report building and visual analysis capabilities, WeData provides a more convenient, simple, and intuitive way for users to work with data. At the same time, leveraging WeData's powerful data asset governance capability, WeData ensures the standardization, accuracy, and consistency of data, making data-assisted decision-making more reliable.