数据开发治理平台 WeData

产品简介

产品文档

腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【版权声明】

©2013-2024 腾讯云版权所有

本文档著作权归腾讯云单独所有，未经腾讯云事先书面许可，任何主体不得以任何形式复制、修改、抄袭、传播全部或部分本文档内容。

【商标声明】

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

文档目录

版权所有：腾讯云计算（北京）有限责任公司

产品简介
产品概述

最近更新时间：2024-07-15 17:26:14

数据开发治理平台 WeData（以下简称 WeData）是位于云端的一站式数据开发治理平台，融合了包含数据集成、数据开发、任务运维的全链路 DataOps 数据开发能力，以及数据地图、数据质量、数据安全等一系列数据治理和运营能力，帮助企业在数据构建和应用的过程中实现降本增效，数据价值最大化。

产品定位

目标行业与用户

适用于政府、金融、泛互、工业、能源、交通、教育、文旅、地产、零售、医疗、传媒等众多行业。受众包括但不限于：
从事数据开发、算法开发、数据运维等岗位的技术人员。
从事数据分析、产品运营等岗位的业务人员。
负责数据安全合规工作的管理人员。
把控公司核心数据资产的管理人员。

业务挑战与痛点

从信息技术革命爆发开始，到近年来移动互联网的蓬勃发展，同时伴随着互联网+概念的不断深入演进和落地，各行各业的企业积累了越来越多的数据，由此衍生出迫切的数据处理与数据应用需求。但这个过程中，也面临着诸多的问题和挑战：
基础设施构建复杂：Hadoop、Spark 等大数据技术琳琅满目，构建复杂。
技术抗风险能力弱：开发和测试脱节，数据出错概率大，数据任务多，依赖复杂，缺少有效变更控制。
数据链路复杂：开源项目往往只解决特定场景问题，需要组合多个开源项目才能搭建完整数据链路。
数据管理难度大：涉及跨部门跨团队合作，团队角色复杂，沟通成本高。
数据治理落地难：数据质量、数据安全等无法得到保障，上层应用不敢放心使用数据。
业务搭建周期长：数仓建设周期过长，半年甚至一年起；数据需求响应慢，两至三天的延迟响应。

核心能力

WeData 提供了数据生产和消费全方位的产品服务，核心服务能力如下：

协同

围绕数据价值链基于协作空间使数据团队不同的角色更好的协作，打破团队间孤岛，缩短从原始数据到数据价值的路径。

DataOps 理念

在大规模任务开发场景下，可以高并发的在线执行数据开发与测试。

开发人员专注任务开发与单元测试，避免业务逻辑学习成本。

编排人员专注任务编排与调度配置，专人专项缩短落地周期。

在敏捷开发场景下，开发与编排的一体化以提高效率。

在编排业务逻辑实现的过程中完成数据任务开发。

可以同时测试数据逻辑与业务逻辑。

实现过程

先开发，后编排：工作流设计不阻塞开发工作，开发无需理解编排逻辑。

开发空间完成后导入编排空间，有专人进行任务编排。

适合中心团队大规模高并发的开发任务。

先编排，后开发：开发人员理解业务逻辑，先设计工作流后开发。

直接在编排空间进行任务编排与开发测试，更敏捷。

适合局点团队小规模或增量任务的敏捷开发模式。

效率

基于 DataOps 敏捷迭代、自动化流程和工具提升数据可靠性，加快数据生产和分析链路效率。

敏捷易用：支持增量式代码开发和发布；支持代码自动补全；可视化拖拉拽方式进行流程设计；支持在线代码调试和日志查看。

开发灵活：开发模式适应多场景，支持先开发后编排以及先编排后开发。

高性能可扩展：高性能调度引擎，支持日千万级任务调度，可对接多种引擎并支持引擎扩展，默认支持大多数 JDBC 接口的引擎，包括 EMR、DLC、TBDS、RDS 等20+引擎。

DataOps理念

支持提交、对比、回溯等版本管理能力，以支持任务的灰度发布。

支持任务、事件、参数、函数的增量发布，而非传统的周期性发布。

敏捷开发、快速迭代，以整体上缩短数据资产化的周期。

实现过程

数据任务开发完后需进行版本提交，以反映在工作流中。

不同版本任务可以快速在同一工作流中调试。

不同项目相同工作流基于不同任务版本实现灰度发布。

在发布管理中按照日期进行增量发布，快速迭代。

一体

服务企业数据管理、数据生产、数据应用、数据运营多个角色，给予不同视角一体化的产品体验。

全链路生产治理：通过事前规划、事中异常阻断、事后质量和成本分析以及数据流通安全管控为数据的生产和消费提供有力的质量和安全保障。

版权所有：腾讯云计算（北京）有限责任公司

一站式运营治理：基于数据自服务和民主化理念，在安全稳定的基础上，通过数据地图、数据洞察和共享，让数据的查找、理解、分析和共享更容易。

质量

贯穿事前中后的数据质量控制，融入 DataOps 管道式开发流程，全面保障数据质量提升。

DataOps 理念

从事后的质量评分转为事中的质量监控，一体化测试由代码测试与数据测试两方面组成，以保证数据分析的高质量。

从事后的标准对标转为事前的标准落标，以保证数据分析时的数据质量、统计口径的一致性。

实现过程

数据任务/工作流提交版本前要求通过在线调试，在线调试会自动拉起数据表对应的质量监控任务。

敏捷数仓建模工具在数据建模时支持直接引用事前定义好的数据标准，在源头上做到落标。

遵从数据标准的表在进行数据集成任务时，支持对脏数据设置零容忍阈值来做到贯标。

版权所有：腾讯云计算（北京）有限责任公司

产品优势

最近更新时间：2024-07-15 17:17:25

作为一款大数据开发治理平台，WeData 具有以下优势：

基于开源

WeData 支持开源接入，广泛支持常见大数据开源技术，如 Hadoop、Hive、Spark 等。具有开源软件使用经验的用户，可以很容易地迁移使用经验。

简单易用

通过工作空间、数据源、工作流等核心概念的抽象，并在数据地图、数据质量等模块中有机组合，使得用户能快速理解和流畅使用 WeData 来开发和治理数据。

降本增效

WeData 提供了许多帮助用户降低成本和提高效率的功能，如数据地图中的数据温度功能，可以协助识别使用频率低但占用成本高的数据，以便清理或转移；工作流开发中的画布功能，可以通过拖拽控件的方式，轻松组织工作流任务间的依赖关系。

安全稳定

数据安全模块提供的数据访问管控能力，能对数据访问权限做到事前审批、事中拦截、事后审计；数据内容管控能力，能对数据做业务脱敏等，建立起数据安全的最后防线。
丰富强大的高可用、负载均衡，以及及时、多渠道触达的监控告警，也使得服务状态和任务运行能得到充分的稳定性保证。

快速实现大数据变现

产品通过一体化运营帮助用户快速发现和理解数据，通过DataOps解决复杂的数据流水线开发，解放数据开发生产力，实现快速数据研发和交付。

版权所有：腾讯云计算（北京）有限责任公司

满足业务自助式服务

数据分析师/业务人员能够更加关注业务逻辑本身，结合产品提供的自助的数据发现、探索、分析能力，满足不同角色更加流畅的数据使用需求。

降低企业管理成本

数据开发需要跨团队和多角色协作，但传统数据工具架构较为割裂，较难协同，产品通过空间划分，为不同角色各司其职和有效协同提供工具基础。

提高企业数据质量可信

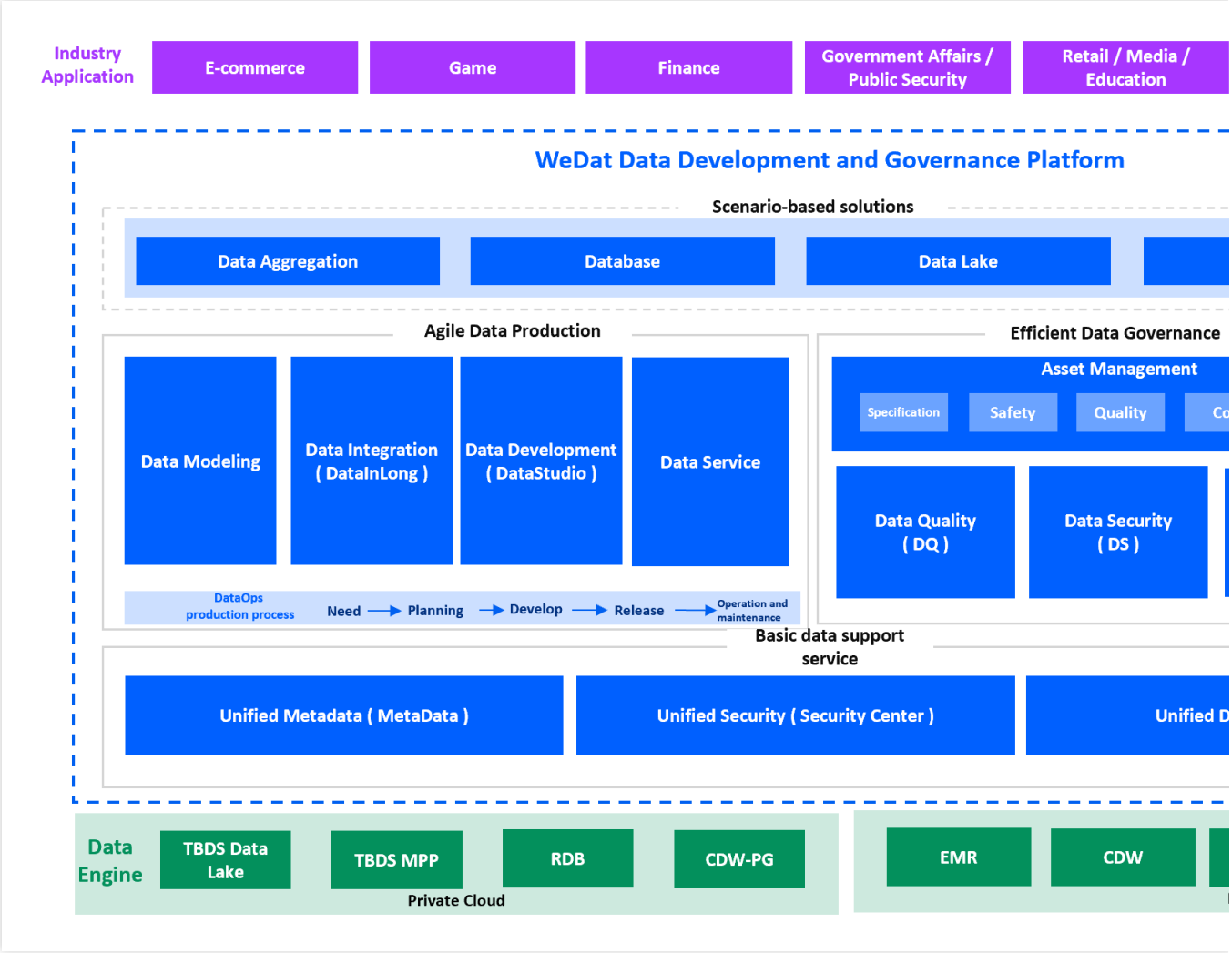
产品通过开发空间、测试空间、生产空间的流水线作业，解决数据“两层皮”的问题，将数据合规、标准化、质量监控和提升始终贯穿其中，确保数据质量。

版权所有：腾讯云计算（北京）有限责任公司

产品架构

最近更新时间：2024-07-15 17:19:20

腾讯云数据开发治理平台 Wedata 产品解决方案和产品架构如图所示：



WeData 整体架构由运营管理体系和开发运营工具构成：

运营管理体系主要包括：多租户管理、多环境管理以及平台开放等。

运营管理体系为数据环境的隔离以及数据的安全流通提供基础保障，多环境管理使得 WeData 支持对接不同的数据引擎，包括私有云 TBDS 以及大部分公有云基础产品，例如弹性 MapReduce、云数据仓库和数据湖计算 DLC 等，平台开放支持将 WeData 的能力以开放 API、开放消息和插件的形式对用户开放，方便与第三方集成和实现定制化能力。

开发运营工具包括：DataOps 敏捷数据生产、全链路数据治理和一体化数据运营。

DataOps 敏捷数据生产遵循 DataOps 理念，规范数据生产过程，通过敏捷迭代的方式提升数据生产效率，全链路治理通过数据质量、数据安全和成本优化在事前、事中和事后保障数据生产和消费全过程，提升数据质量和可用性，

一体化运营以数据价值释放为目标，通过数据地图、数据洞察和数据共享让用户快速的完成数据发现、数据理解、数据洞察和数据应用，降低数据使用门槛，缩短数据价值释放路径。

版权所有：腾讯云计算（北京）有限责任公司

产品功能

最近更新时间：2024-07-15 17:20:49

WeData 核心功能包括如下：

项目管理

从系统/租户层面实现项目隔离，为管理者提供对使用 WeData 的用户（成员）权限、底层计算引擎配置、执行资源的管理能力。

数据规划

提供包含数仓分层分类、逻辑模型设计、指标维度定义、数据标准等数据整体规划设计能力，帮助企业统一数仓规范设计和标准定义，实现设计态到开发态的自动流转。
数仓规范：数仓规划基于全局进行业务对象的统一规划和规范定义，对模型进行分层设计管理，按照特定的业务主题进行分类分域管理，形成具有层级结构的业务标签。
模型设计：对逻辑模型进行定义和实体关系设计，包括定义、复制、修改、删除、导入导出、版本管理能力，同时建立与物理模型、指标维度关联映射，实现模型从设计态到开发态的自动同步。
标准管理：包含标准内容管理和对标任务管理，通过对标准规则的设计和任务配置，实现对数据值、库、表结构、表名、指标维度标签等层面的标准化。
业务定义：指标/维度字典，对基础/衍生指标、维度条件（普通维度、业务限定、时间周期、退化维度）进行全生命周期定义管理，并建立和模型关联关系，实现指标生产代码自动生成。

数据集成

操作轻量化、过程可视化、能力开放化数据集成能力，支持复杂网络环境下、丰富的异构数据源之间高速稳定的海量数据同步。
全场景同步：包括实时同步与离线同步。
多类型异构数据源：支持30+数据源提供星型结构支持读写随机搭配。
T转换
数据级：对同步中的数据进行内容转换，如数据过滤、Join 等。
字段级：提供单个字段转换处理，包括自定义数据字段、格式转换、时间格式转换等。
任务及数据监控
读写指标：支持任务读写实时指标统计，包括读写总量、速度、吞吐、以及脏数据等。
监控告警：支持任务及资源监控，覆盖短信、邮件、HTTP 等多渠道告警。

版权所有：腾讯云计算（北京）有限责任公司

数据开发

通过严谨的 CI/CD 流程规范和自动化的测试发布运维加持能力，缩短从原始数据加工运维到业务应用数据的路径，提升效率的同时保障数据质量。
在线代码开发：支持代码开发，对任务工作流进行易用地拖拽式编排，同时支持大规模任务的可视化编排呈现。
代码开发：支持对 HiveSQL、SparkSQL、JDBCSQL、Spark、Shell、MapReduce、PySpark、Python、TBase、DLC SQL、DLCSpark、CDW PostgreSQL、Impala 等任务进行在线代码开发、调试，以及版本管理。
任务测试：支持任务和工作流测试及版本管理。
开发辅助：提供项目、工作流和任务三种粒度的参数配置，支持时间参数运算以及函数参数。
版本管理：支持事件、函数、任务和参数的版本管理。
代码管理：提供代码统一的管理、导入和导出。
编排调度：对任务进行流程编排及提交调度。
调度方式：支持周期、一次性和事件触发调度，周期调度提供 crontab 方式配置。
依赖策略：支持任务自依赖和工作流自依赖。
跨周期依赖配置：提供跨周期依赖配置及自定义依赖配置，上下游依赖实例范围支持按需自定义选择。
批量编排：提供 Excel 批量创建任务及依赖的能力，加快任务依赖编排效率。
发布运维：对开发完成的任务按需发布到生产环境，并对任务进行统一监控和运维。
任务发布：支持将开发成果发布上线。
监控运维：对任务进行流程编排及提交调度。
分析探索：智能易用的数据开发方式提升任务协同开发效率，帮助用户清晰查看任务处理过程，有效提升数据即席探索效能。
在线编辑：提供可视化的交互式分析 IDE。
运行：提供执行信息可视化。
开发辅助：提供开发辅助效率工具。

数据质量

通过灵活的规则配置、全方位的任务管理、多维度的质量评估，为数据接入、整合、加工到消费的全生命周期各阶段提供全面的数据质量稽核能力。
多源数据监控：支持监控的数据源、引擎类型包括 EMR Hive、Spark、DLC（公有云）、CDW-PG、TBDS、Gbase（私有云）等，提供多源数据全量校验能力。
丰富规则模版：目前提供6大维度、56种业界通用的表级、字段级内置规则模版，真正实现开箱即用，质控工作流得以大幅提效，帮助用户从各个维度感知数据变动及ETL过程中产生的问题数据。
质控灵活配置：支持系统质量规则模版、自定义模版、自定义SQL三种规则创建模式，可按业务需求调整参数，配置任务执行策略，轻松实现全链路质控校验。
全局链路保障：支持关联生产调度以及离线周期检测两种执行方式，提供事前、事中和事后的全链路数据保障运维能力，及时进行告警、阻断拦截，防止脏数据向下游蔓延。

版权所有：腾讯云计算（北京）有限责任公司

治理多维可视：质量概览和质量报告模块为用户提供全局视角，让用户对质量任务运行情况、告警阻塞趋势、各维度质量评分了如指掌，快速发现定位问题，了解质量提升效果。

数据安全

提供集中化的数据安全管控和协作机制，保障数据在安全的条件下进行有效流通。
统一数据安全管控：针对绑定的存算引擎进行安全策略的深度集成，统一数据访问，简化数据使用流程。
权限审批：打通 ranger 权限策略体系，实现责任到人，数据粒度到表的权限管控能力。提供权限申请和审批通道，安全开放数据访问控制能力。

数据运营

基于强大的底层元数据能力，提供数据目录、血缘解析、热度分析、资产评分、业务分类、标签管理等数据资产服务，有效提升用户对企业级海量数据的理解、管控、协作能力。
数据发现：统一的元数据采集和管理。
数据总览：提供数据资产的概览统计，包括项目、表、存储量、数据类型覆盖等基础信息，以及数据全景和热门排行功能。
数据目录：支持全域数据表级、字段级快速检索与定位；表详情提供数据全量技术、业务信息及数据血缘、温度、质量、产出与变更、预览等功能。
库表管理：支持对全域库表进行管理。
业务分类：支持根据业务需求创建、管理主题类目、数仓分层和业务标签，并对数据表进行批量分类分层操作。

数据服务

提供包含 API 生产、API 管理和 API 市场等覆盖 API 全生命周期的能力，帮助企业统一管理对内对外的 API 服务，构建统一的数据服务总线。
快捷 API 生产。
API 管理和运营。
API 安全调用。

版权所有：腾讯云计算（北京）有限责任公司

应用场景

最近更新时间：2024-07-15 17:25:38

数据开发治理平台 WeData 作为一款通用型的数据工具产品，有着非常丰富的应用场景。以下从几个典型的场景加以说明。

企业级数据仓库构建

腾讯云数据开发治理平台 WeData 提供完善的数据处理功能，覆盖数据仓库构建全链路。从异构数据源数据导入，到通过丰富的大数据组件支持对数据开发、任务编排、任务运维等全过程的支持，对数仓各层次各数据域进行规范化生产，确保数据规范性、完整性、及时性等，最后通过数据导出或 API 服务的方式，将数据仓库生产的规范数据应用于企业各类型的数据业务，助力企业用数据赋能经营。

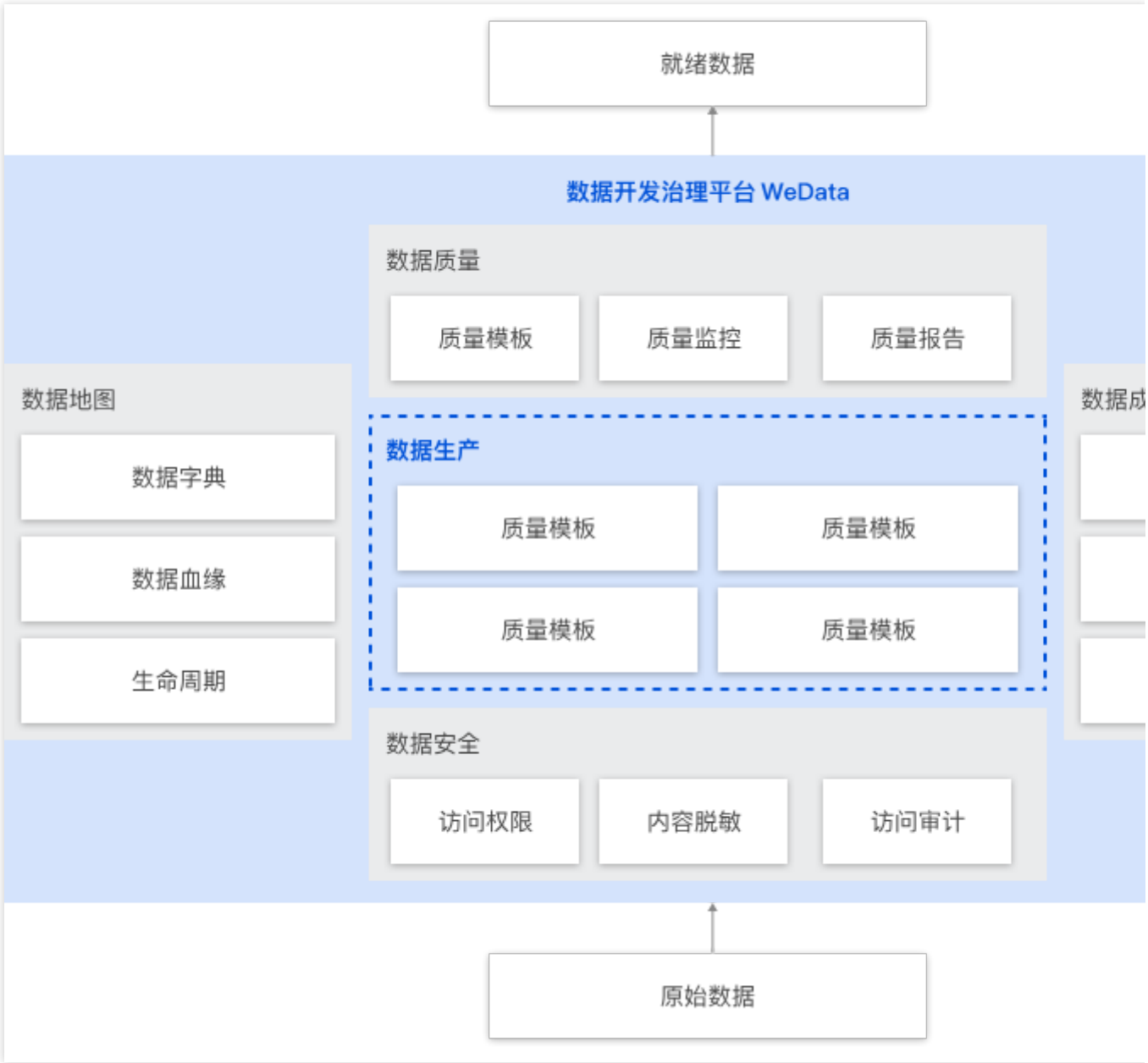
版权所有：腾讯云计算（北京）有限责任公司


数据业务
BI 报表
数字大屏
数据开发治理平台 WeData
数据发布
数据导出
API 服务
数据处理
ODS
DWD
DWS
数据开发
任务编排
数据导入
Hive
MySQL
Oracle
Kafka
数据源
MySQL
Oracle
Kafka

数据资产治理

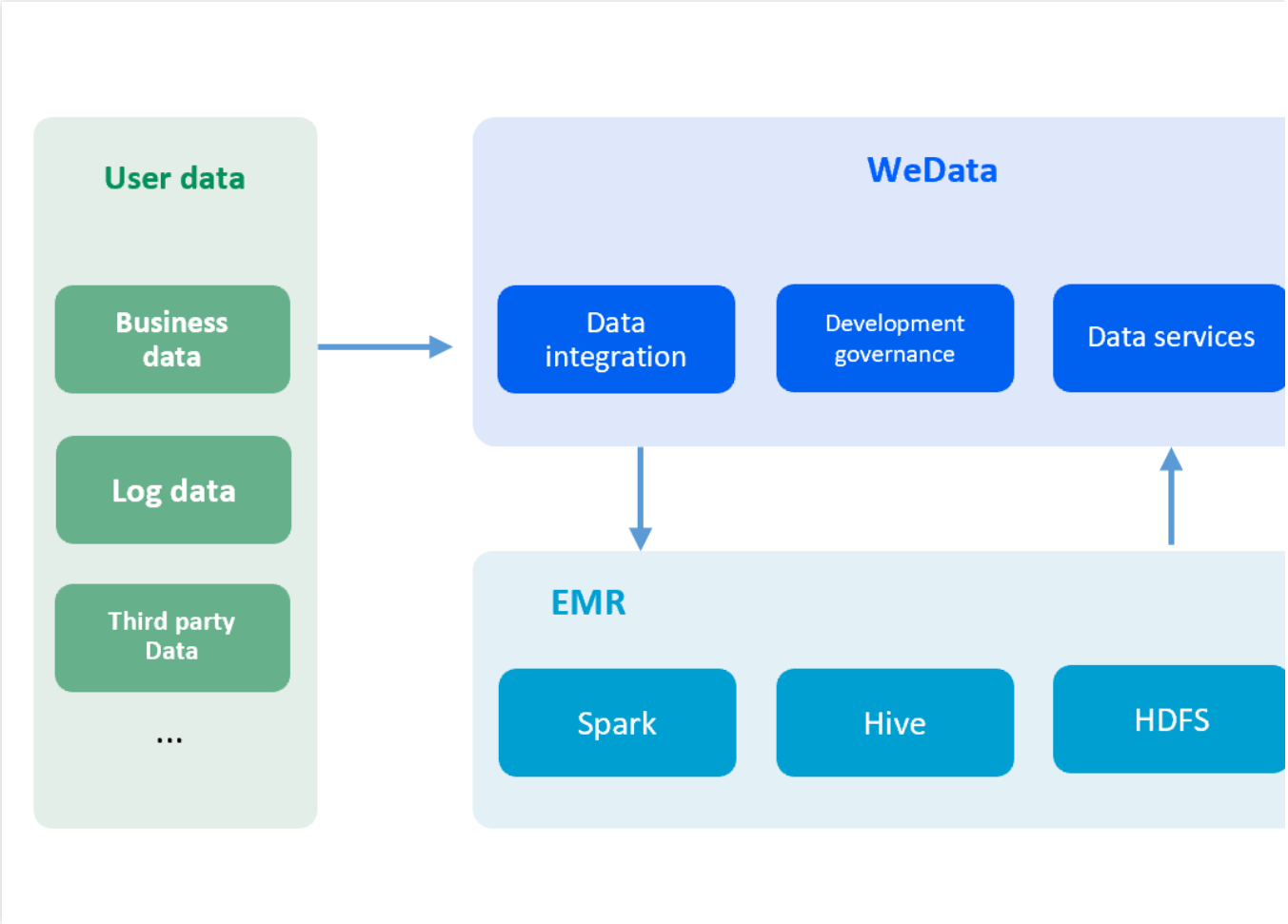
数据资产是企业核心资产的重要组成部分，对于数据驱动型公司，更是最核心的资产。然而，大数据与生俱来的异构性、复杂性、高成本等特性，伴随着企业业务的多样性和复杂性的发展，使得从原始数据经过处理后得到的就绪数据，体现出质量低下、成本高企、理解困难、安全无保障等问题，并未能如预期直接被业务使用。

WeData 在核心数据处理能力的基础上，提供了一系列数据治理能力，帮助企业将数据的方方面面治理起来，使得数据能够被放心大胆地应用于业务，高效为业务创造价值



离线数仓/数据中台解决方案——WeData+EMR

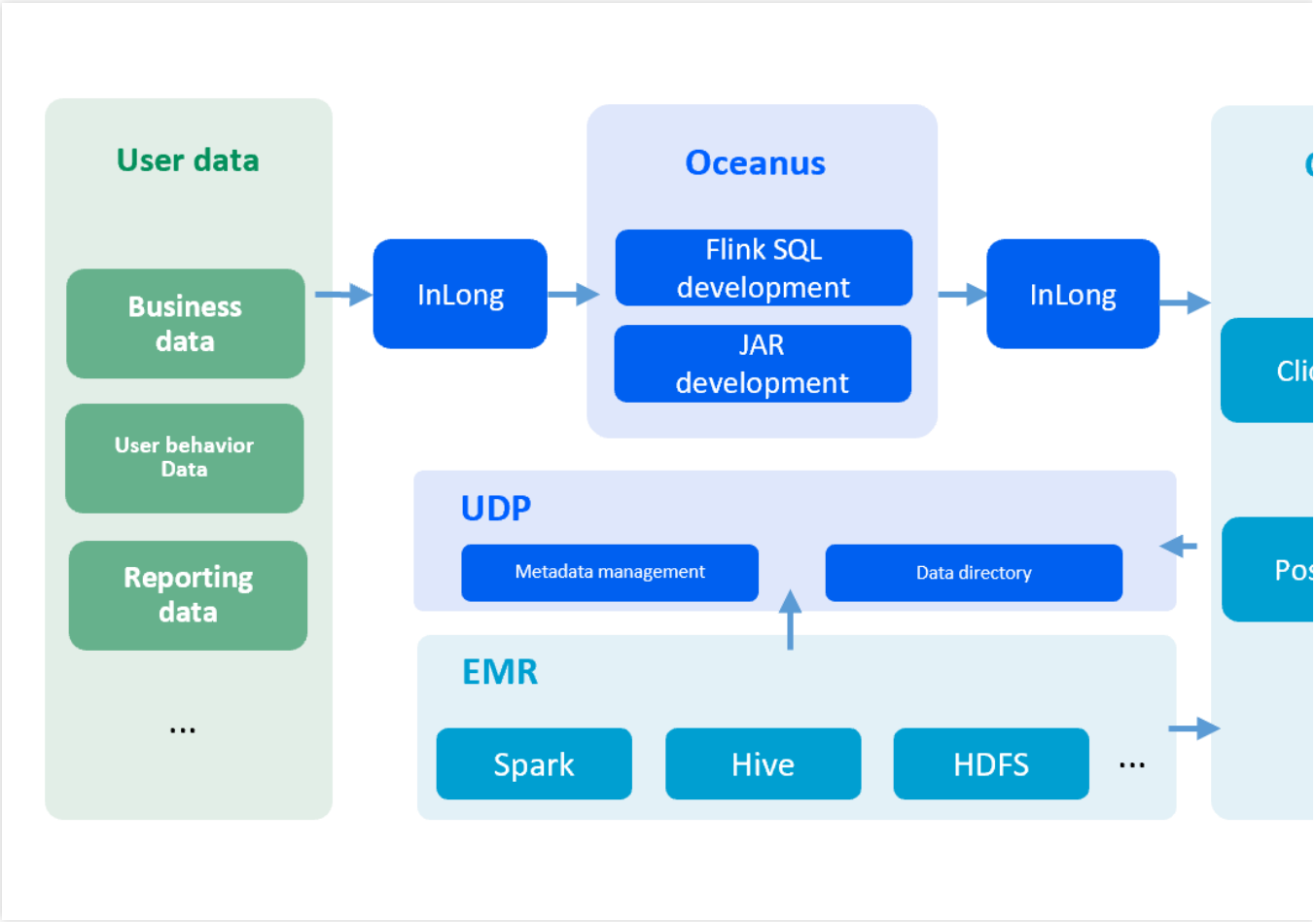
基于 WeData 提供的数据集成和开发能力，汇聚来自业务侧的海量数据，然后借助 EMR 强大的 PB 级数据计算与存储能力，提供高性能企业级离线数仓方案，并可通过 WeData 对数仓进行统一安全管理和数据质量治理，提供统一的、规范化的数仓体系，为用户进一步挖掘数据价值提供支持。



实时数仓解决方案——DataInLong/Oceanus+CDW

在云数仓 CDW-PG/CK 提供的大规模并行处理和 OLAP 分析能力之上，由 Oceanus 进行实时数据处理，帮助用户快速汇聚和收集数据，构建轻量化、低成本、可弹性伸缩的实时数仓，并可配套 UDP 进行数据目录管理，当用户需要进一步的数据治理能力时，Oceanus 可平滑升级至 WeData，为用户提供更进一步的数据加工、质量治理、探索分析能力。

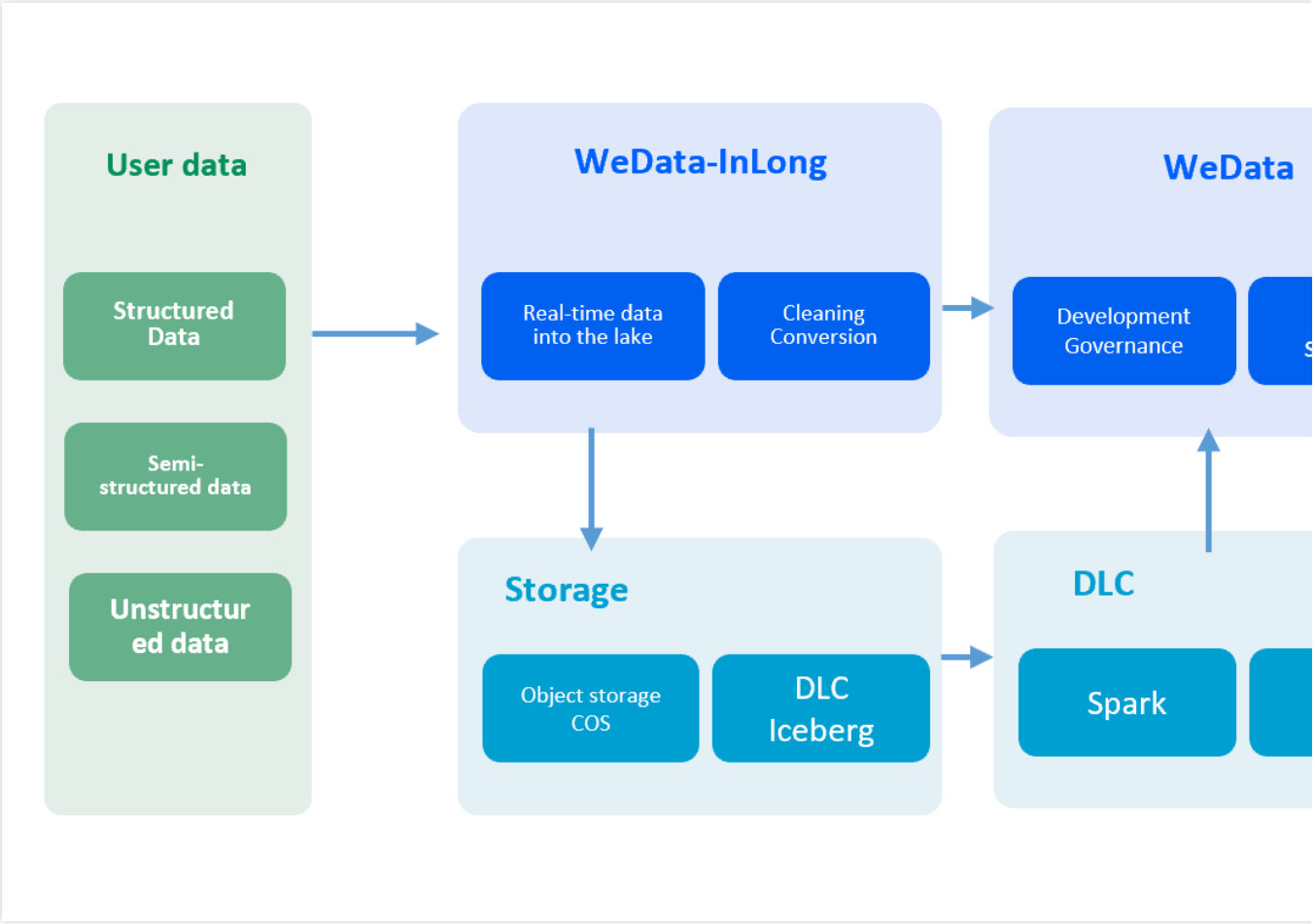
版权所有：腾讯云计算（北京）有限责任公司



全托管数据湖解决方案——DataInLong/WeData+DLC

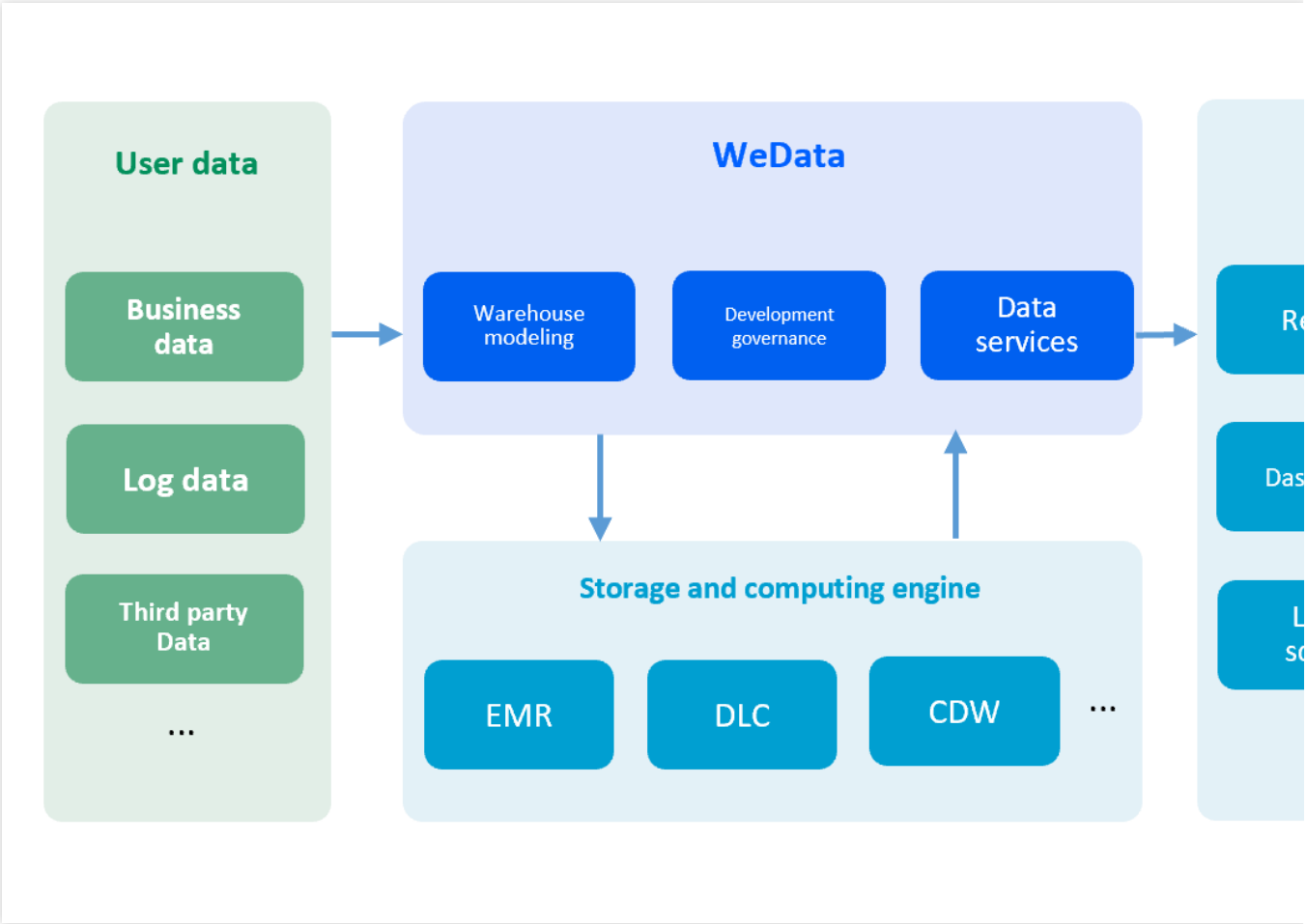
在 DLC 提供的存算分离海量大数据分析架构之上，通过 DataInLong 的实时数据入湖和清洗能力，实现来自业务侧的多元异构数据的统一汇聚，构建低成本、高弹性、全托管数据湖方案，DataInLong 可以平滑升级至 WeData，为用户提供更进一步的数据加工、质量治理、探索分析能力。

版权所有：腾讯云计算（北京）有限责任公司



数据可视化解决方案——WeData+BI

基于 WeData 提供的数仓指标管理和数据服务能力，连通数据到业务的“最后一公里”，配合 BI 提供的自助报表搭建和可视化分析能力，为用户使用数据提供更加便捷、简易、直观的呈现方式，同时，借助 WeData 强大的数据资产治理能力，为数据的规范性、准确性、一致性提供保障，使数据辅助决策业务更加可靠。


User data
Business data
Log data
Third party Data
...
WeData
Warehouse modeling
Development governance
Data services
Storage and computing engine
EMR
DLC
CDW
...