解决方案实践

华为云通用数据使能解决方案实践

文档版本 1.1

发布日期 2024-04-23

华为技术有限公司

版权所有 © 华为技术有限公司 2024。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

版权所有 © 华为技术有限公司

目 录

版权所有 © 华为技术有限公司

1 方案概述

应用场景

疫情时代大幅加速了企业数字化建设，导致全球数字化进程整体提前7年，亚太地区提前10年（麦肯锡）。数字化不再被认为困难重重，企业做事速度比原先预想的快20~25倍，部分先进企业已将混合办公模式作为新常态。

2020年4月，中央将数据列为新的生产要素，作为数字经济发展和科技创新的关键驱动力，在企业数字化转型中发挥着难以替代的作用。如何释放数据生产要素价值，是每个企业值得思考的问题。

数据使能解决方案，旨在通过数据使能企业从容应对风险及不确定性，保障业务平稳运营，同时从变化中找到机会。

图 1-1 图示

深耕数字化是企业实现跨越式发展的必然选择。企业实现数字化转型，要基于战略将流程、组织、IT治理进行重构。整个解决方案涵盖了管理、业务、技术三个视角，也对解决方案和专业服务提出了巨大挑战。

华为云数据使能解决方案为客户提供坚实稳定可扩展可演进数据平台底座，也为客户提供性各种可选余地的高性价比数据设计、治理、实施专业服务，更为客户提供丰富的各行业数据应用。

方案架构

华为云数据使能解决方案，承载华为数字化转型能力外溢使命，具备“**方法论+技术平台+数据应用+行业场景**”所需的全套能力体系，从**设计到落地**，为大型政企客户量身定制“**跨越孤立系统、承载业务数字孪生、感知业务**”的数据管理解决方案；帮助企业从多角度、多层次、多粒度挖掘数据价值，实现数据驱动运营，完成数字化转型。

版权所有 © 华为技术有限公司

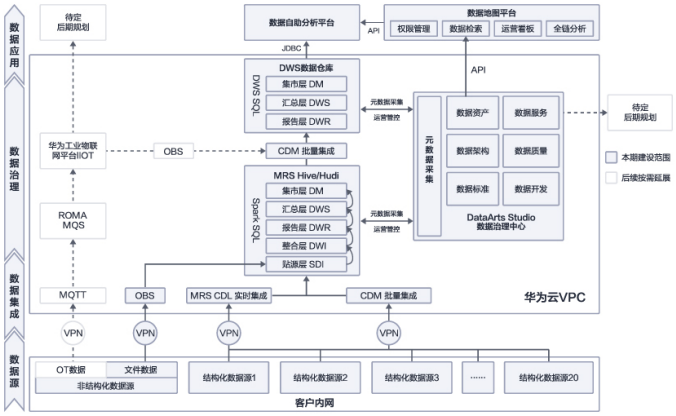
华为云通用数据使能解决方案，以数据治理为基础，数据智能为动力，驱动企业加速发展，主要由三部分组成：

1. 丰富的企业应用转型方案：以华为云云商店为平台，面向企业研发、生产、营销、经营等领域，联合生态合作伙伴，推出一系列企业应用提升方案。
2. 专业的数据治理实施服务：结合华为“数据之道”方法论 + 业界的数据治理实践，为企业落地立而不破的数据治理解决方案。
3. 先进的数据基础设施平台：面向数据湖、数据治理、数据资产管理、数据应用构建、数据可信流通等数据场景，提供先进、高效、可靠、经济的数据基础实施平台。

图 1-2 业务架构

方案采用云服务的方式部署交付，整体部署架构如下：

图 1-3 部署架构



架构描述：

本架构基于某零售行业客户数据使能项目实践输出，作为最佳实践案例，不代表数据使能的完整集成架构。客户作为国内零售行业龙头企业，随着规模不断扩张，对数字化转型提出了明确的需求。数据使能作为支撑数字化转型的核心解决方案，旨在梳理并打通客户数据的获取、存储、治理、流转、服务、应用等通路，构建数据管理体系，支撑业务人员自助式开发可视化报表、看板等应用，实现数据价值的深度挖掘与

版权所有 © 华为技术有限公司

快速变现。项目分两期进行，一期聚焦IT数据，二期聚焦OT数据，本实践以项目一期为主进行展开，其中数据治理部分通过CDM云数据迁移服务和MRS服务的CDL组件实现定时/实时集成客户结构化数据源，置入MRS服务+DWS数据仓库服务所组成的湖仓中进行存储与计算，通过DataArts Studio数据治理中心完成数据治理与作业编排；数据消费部分通过数据地图平台完成数据检索、授权与审批，通过数据自助分析平台完成数据的分析与报表展示。

方案优势

华为云数据使能解决方案依托华为云，联合众多生态伙伴能力，为企业数字化转型奠定数据基础。该方案具备以下几个优势：

1. **端到端的数据全生命周期管理解决方案**

   为客户提供坚实稳定、可扩展、可演进数据平台底座，提供性各种高性价比的数据设计、治理、实施专业服务，以及丰富的各行业数据应用。

2. **生态伙伴取长补短、优势融合**

   基于华为“数据之道”方法论和华为云全面稳定的数据平台服务，结合各行各业的生态伙伴优势能力，强强联合，为各企业数字化转型提供一体化解决方案。

3. **适合各行各业的通用数据使能解决方案**

   组合了华为云服务、数据治理实施专业服务、伙伴联营应用方案，支持公有云云服务、私有云部署、边缘部署形态，针对政府、制造、企业、矿业亦有专门场景化方案，面向各行各业，提供多元化服务。

版权所有 © 华为技术有限公司

2 资源和成本规划

资源规划说明

以某行业客户为例，客户的需求为构建全集团统一的数据平台，在数据平台中对数据进行治理，并支撑上层的数据应用。

假设客户的数据量规模在10T左右，考虑到增量的数据集成方式、数据入湖入仓的膨胀系数、以及客户对平台性能的要求，设计了以下的资源与成本清单。实际收费应以账单为准：

资源与成本清单

1. 云服务清单

表 2-1 云服务清单

资源类型	服务	规格		数量
数据治理平台	数据湖治理中心 (**DataArts**: Data Governance Center)	企业版：5,000次/天的数据开发调度，附带8作业并发， 最大1.5Gbps带宽的数据集成能力，并且支持管理5,000个技术资产、100个数据模型。		1
	数据治理中心（DataArts Studio）技术资产数量增量包	package.da.10k，增加1万张管理数据表规模。		1
数据迁移	DataArts批量数据迁移增量包 (**CDM**: Cloud Data Migration)	cdm.xlarge： 16核/32GB | 10/4 Gbit/s | 100 并发任务		2
数据计算	MapReduce服务 (**MRS**: MapReduce Service)	离线集群-MRS Master节点	MRS服务管理费用 *3; 规格:X86计算 | 通用计算增强型 | c6.8xlarge.4 | 32核 | 128GB *3; 系统盘:通用型SSD | 480GB *3; 数据盘:通用型SSD | 200GB *3;	1

版权所有 © 华为技术有限公司

资源类型	服务	规格		数量
		离线集群-MRS分析Core节点	MRS服务管理费用 *16; 规格:X86计算 | 通用计算增强型 | c6.8xlarge.4 | 32核 | 128GB *16; 系统盘:超高IO | 500GB *16; 数据盘:超高IO | 600GB | 2个 *16;	1
		实时集群-MRS Master节点	MRS服务管理费用 *3; 规格:X86计算 | 通用计算增强型 | c6.8xlarge.4 | 32核 | 128GB *3; 系统盘:通用型SSD | 480GB *3; 数据盘:通用型SSD | 200GB *3;	1
		实时集群-MRS分析Core节点	MRS服务管理费用 *4; 规格:X86计算 | 通用计算增强型 | c6.8xlarge.4 | 32核 | 128GB *4; 系统盘:超高IO | 500GB *4; 数据盘:超高IO | 600GB | 2个 *4;	1
数据存储	对象存储服务 (**OBS**: Object Storage Service)	对象存储 | 标准存储单AZ存储包 | 100TB;		1
数据仓库	数据仓库服务 (**DWS**: Data Warehouse Service)	dwsx2.16xlarge.m7 | 云数仓 | X86 | 64 vCPUs | 512 GB | 3T * 5 节点;		1
数据应用（取决于应用自身需求）	弹性云服务器 (**ECS**：Elastic Cloud Server)	32GiB|c6s.4xlarge.2 通用计算增强型 16 vCPUs | 32 GB | Ubuntu 20.04 server 64bit 系统盘 高IO 500 GB x 1 数据盘		3
	弹性公网IP (**EIP**：Elastic IP)	弹性公网10 Mbit/s (可选，如需外网访问系统)		2
	云数据库RDS (**RDS**：Relational Database Service)	实例类型：主备 | 数据库引擎版本：MYSQL 8.0 | 性能规格：4 vCPUs | 16 GB | 存储空间：500G | 网络：内网		1
	分布式缓存服务Redis版	实例类型：主备 | 版本：Redis 5.0 | 性能规格：4GB | 副本数:2 | 网络：内网		1

2. 专业服务清单

   本案例所涉及的数据管理实施专业服务报价项如下，不同报价项的价格仅供参考，实际以收费账单为准：

版权所有 © 华为技术有限公司

表 2-2 专业服务清单

类别	报价项	量纲
数据管理成熟度诊断	数据管理成熟度诊断	套
数据使能信息架构设计	数据使能信息架构设计-基础包	套
	数据使能信息架构设计-增量包	套
数据使能技术架构设计	数据使能技术架构设计	套
数据使能技术平台集成实施	数据使能技术平台集成实施	套
	IT数采集成实施	套
	OT数采集成实施	套
数据使能方案设计	数据湖治理方案设计-基础包	套
	数据湖治理方案设计-增量包	套
	ITOT融合方案设计	套
	数智融合方案设计	套
数据使能方案实施	数据湖治理方案实施-基础包	套
	数据湖治理方案实施-增量包	套
	ITOT融合方案实施	套
	数智融合方案实施	套
数据应用集成设计与实施	数据应用集成设计与实施	套

版权所有 © 华为技术有限公司

3 实施步骤

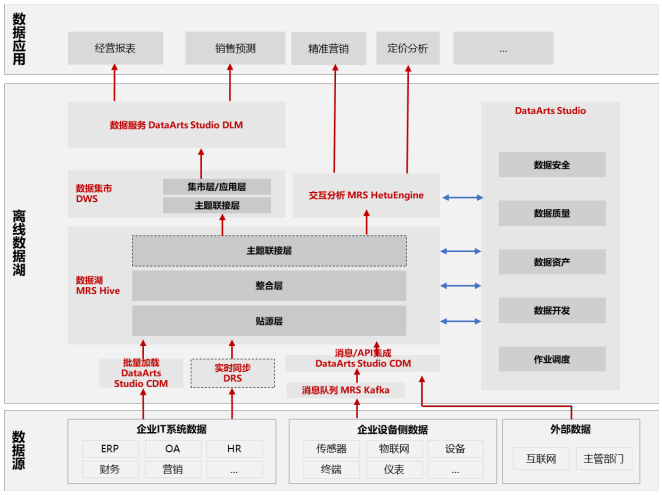
3.1 数据治理平台

数据平台总体架构

本项目一期以离线数据分析为主，按照华为云数据使能方案的离线数据湖子方案，以华为公有云为载体，为客户建设离线数据湖平台，作为本次项目数据分析应用的数据底座。

离线数据湖平台整体架构如下，核心由三个云服务组成，大数据平台MRS、数据仓库平台DWS、数据集成治理平台DataArts Studio。

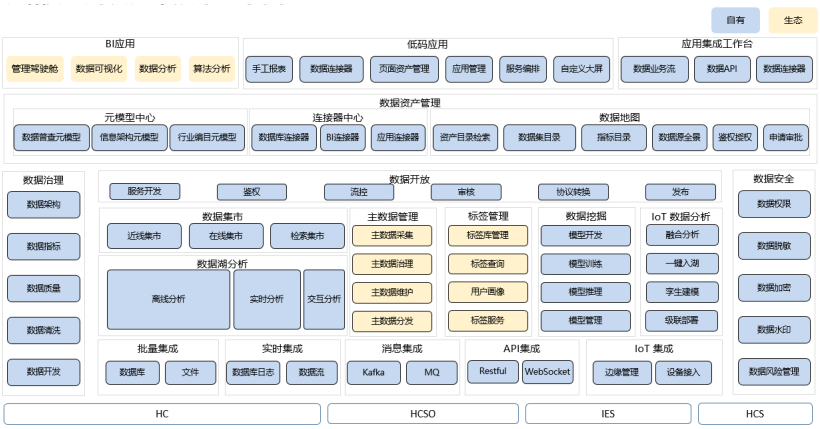
图 3-1 离线数据湖整体架构



离线数据湖平台可以向实时数据湖、ITOT融合数据湖、数据资产平台、数据可信流通等其他子方案演进，整体演进方案如下：

版权所有 © 华为技术有限公司

图 3-2 华为云数据使能方案总体演进



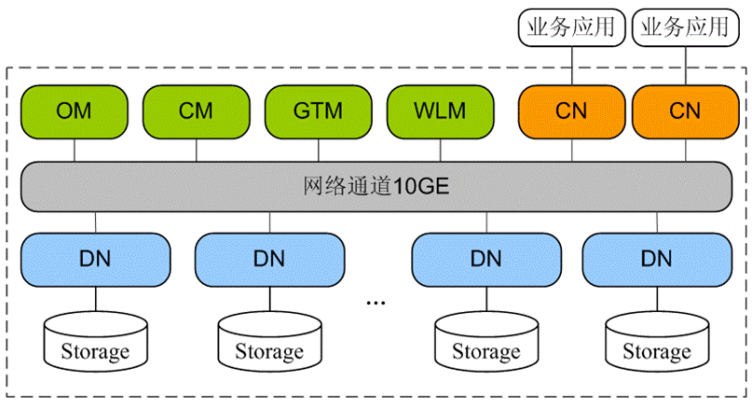
华为云数据使能方案为客户提供全栈大数据管理方案，覆盖“采存算管用”的全数据生命周期处理环节，支持公有云、混合云、边缘云等多种基础设施形态，支持向未来平滑演进。

数据仓库服务 DWS

GaussDB(DWS)是企业级的大规模并行处理关系型数据库。其采用MPP（Massive Parallel Processing）架构，支持行存储与列存储，提供PB（Petabyte，2的50次方字节）级别数据量的处理能力。数据仓库服务（GaussDB(DWS)，简称DWS）是一种在线数据处理数据库，提供即开即用、可扩展且完全托管的分析型数据库服务。DWS是基于华为融合数据仓库GaussDB产品的云原生服务，兼容标准ANSI SQL 99和SQL 2003，同时兼容PostgreSQL/Oracle数据库生态，为各行业PB级海量大数据分析提供有竞争力的解决方案。

GaussDB(DWS)在核心技术上跟传统数据库相比有巨大优势，可以解决很多行业用户的数据处理性能问题，可以为超大规模数据管理提供高性价比的通用计算平台，并可用于支撑各类数据仓库系统、BI（Business Intelligence）系统和决策支持系统，统一为上层应用的决策分析等提供服务。DWS可广泛应用于金融、车联网、政企、电商、能源、电信等多个领域，已连续两年入选Gartner发布的数据管理解决方案魔力象限，相比传统数据仓库，性价比提升数倍，具备大规模扩展能力和企业级可靠性。

图 3-3 数据仓库产品架构

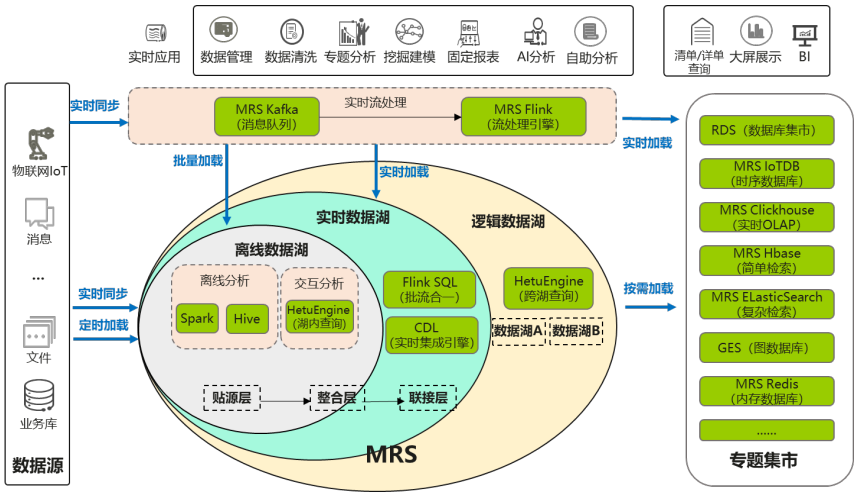


版权所有 © 华为技术有限公司

华为并行数据库基于Shared-nothing/MPP架构，面向开放x86平台，数据跨所有节点均匀分布，所有节点以并行方式工作，提供标准SQL接口，支持SQL92,99,2003标准，支持JDBC/ODBC标准接口，提供多达256个物理节点PB级数据存储分析的扩展能力。

大数据服务 MRS

图 3-4 云原生数据湖全景



基于MRS，建设企业级云原生数据湖，云原生数据湖主要包括数据湖，数据集市：

- **数据湖**：企业内多种格式数据源汇聚的大数据平台，通过严格的数据权限和资源管控，将数据和算力开放给各种使用者，为数据湖。一份数据支持多种分析，是数据湖最大的特点。数据湖又分为三个阶段：
  a. **离线数据湖**：将企业内多种格式数据源汇聚的大数据平台，通过严格的数据权限和资源管控，将数据和算力开放给各种使用者。其中数据从数据源产生后到进入到数据湖存储，无法做到实时，通常超过15分钟。离线数据湖主要用来支撑企业内部T+1小时级别的离线分析和处理。
  b. **（离线数据湖是客户大数据平台的必选，一般的客户做大数据处理都要使用离线数据湖，但是离线数据湖的时效性很低，只能做到小时级处理，已经开始逐渐无法满足各行业需求，因此除非客户坚持目前和未来如果干年对时效性都没有要求，不建议选择离线数据湖）**
  c. **实时数据湖**：将企业内多种格式数据源汇聚的大数据平台，通过严格的数据权限和资源管控，将数据和算力开放给各种使用者。其中数据从数据源产生后，可以实时进入到数据湖存储，通常在1到15分钟之间。实时数据湖既可以用来支撑企业内部T+1小时级别的离线分析和处理，也可以支撑企业内部实时分析和处理。
  d. **逻辑数据湖**：将企业内多种格式数据源汇聚的大数据平台，通过严格的数据权限和资源管控，将数据和算力开放给各种使用者。其中数据并不是在物理上汇聚到了一个数据平台上，而是如果干个物理分开的数据平台形成虚拟数据湖。
- **数据集市**：企业内存储特定格式数据，提供给特定类型查询分析，满足特定的业务场景，针对特定用户的，特定的数据平台。各个集市之间，数据会有重复。

版权所有 © 华为技术有限公司

数据治理中心 DataArtsStudio

数据治理中心DataArts Studio是针对企业数字化运营诉求提供的数据全生命周期管理、具有智能数据管理能力的一站式治理运营平台，包含数据集成、数据架构、数据开发、数据质量监控、数据目录管理、数据服务等功能，支持行业知识库智能化建设，支持大数据存储、大数据计算分析引擎等数据底座，帮助企业快速构建从数据接入到数据分析的端到端智能数据系统，消除数据孤岛，统一数据标准，加快数据变现，实现数字化转型。

图 3-5 数据治理方案图

- **数据集成**

  支持批量数据迁移、实时数据集成和数据库实时同步，支持20+异构数据源，全向导式配置和管理，支持单表、整库、增量、周期性数据集成。

- **数据架构**

  作为数据治理的一个核心模块，承担数据治理过程中的数据加工并业务化的功能，提供智能数据规划、自定义主题数据模型、统一数据标准、可视化数据建模、标注数据标签等功能，有利于改善数据质量，有效支撑经营决策。

- **数据开发**

  大数据开发环境，降低用户使用大数据的门槛，帮助用户快速构建大数据处理中心。支持数据建模、数据集成、脚本开发、工作流编排等操作，轻松完成整个数据的处理分析流程。

- **数据质量**

  数据全生命周期管控，数据处理全流程质量监控，异常事件实时通知。

- **数据目录**

  提供企业级的元数据管理，厘清信息资产。通过数据地图，实现数据目录的数据血缘和数据全景可视，提供数据智能搜索和运营监控。

- **数据服务**

  标准化的数据服务平台，提供一站式数据服务开发、测试部署能力，实现数据服务敏捷响应，降低数据获取难度，提升数据消费体验和效率，最终实现数据目录的变现。

- **数据安全**

  数据安全为数据治理中心提供数据生命周期内统一的数据使用保护能力。通过敏感数据识别、分级分类、隐私保护、资源权限控制、数据加密传输、加密存储、数据风险识别以及合规审计等措施，帮助用户建立安全预警机制，增强整体安全防护能力，让数据可用不可得和安全合规。

- **智能数据湖**

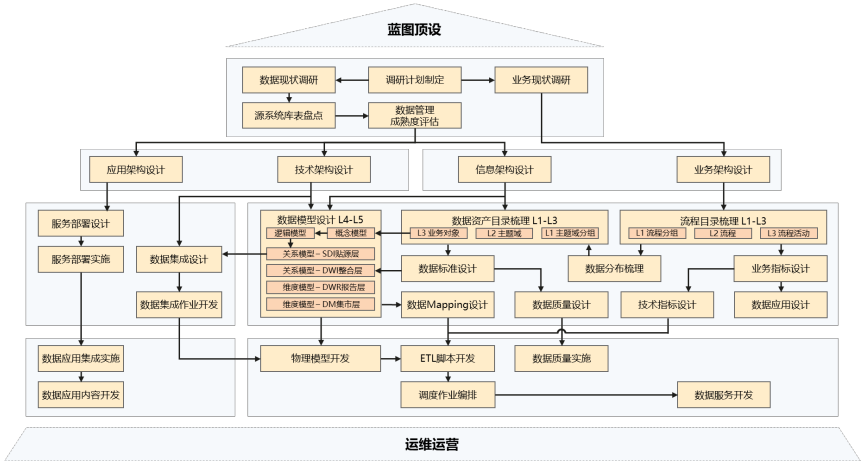
  DataArts Studio集成了丰富的数据引擎，支持对接所有华为云的数据湖与数据库云服务，例如MapReduce服务MRS、数据仓库服务DWS等，也支持对接企业传统数据仓库，例如Oracle。

版权所有 © 华为技术有限公司

3.2 数据治理实施专业服务

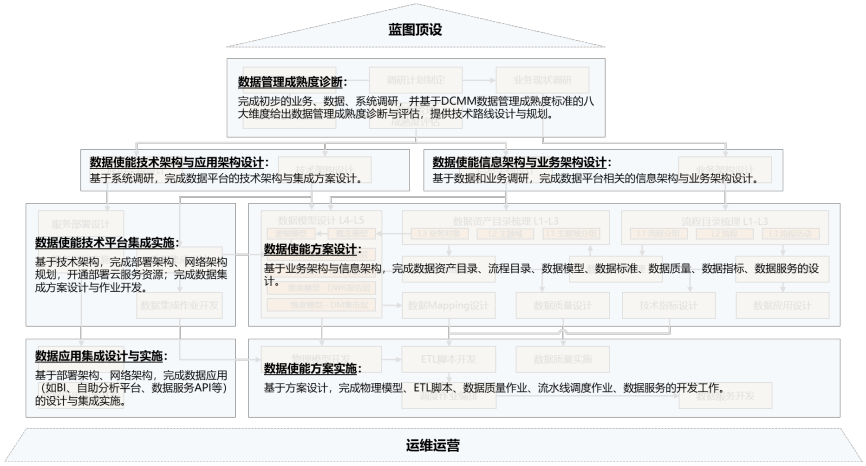
数据使能设计与实施流程如下图所示：

图 3-6 流程图



上图中的各个流程，可以归类为数据使能专业服务的各个报价项，如下图所示：

图 3-7 报价项



- 数据管理成熟度诊断：完成初步的业务、数据、系统调研，并基于DCMM数据管理成熟度标准的八大维度给出数据管理成熟度诊断与评估，提供技术路线设计与规划。
- 数据使能技术架构与应用架构设计：基于系统调研，完成数据平台的技术架构与集成方案设计。
- 数据使能信息架构与业务架构设计：基于数据和业务调研，完成数据平台相关的信息架构与业务架构设计。
- 数据使能技术平台集成实施：基于技术架构，完成部署架构、网络架构规划，开通部署云服务资源；完成数据集成方案设计与作业开发。
- 数据使能方案设计：基于业务架构与信息架构，完成数据资产目录、流程目录、数据模型、数据标准、数据质量、数据指标、数据服务的设计。

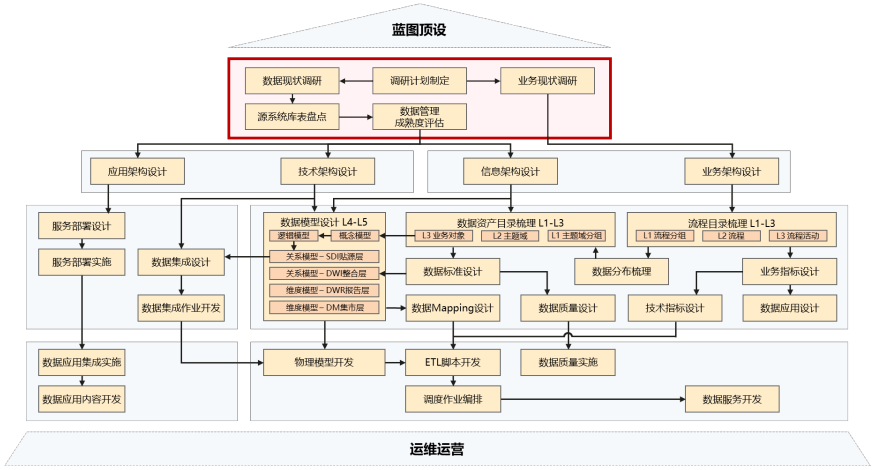
版权所有 © 华为技术有限公司

- 数据应用集成设计与实施：基于部署架构、网络架构，完成数据应用（如BI、自助分析平台、数据服务API等）的设计与集成实施。
- 数据使能方案实施：基于方案设计，完成物理模型、ETL脚本、数据质量作业、流水线调度作业、数据服务的开发工作。

本实践的整体实施流程将围绕上述7个模块展开，以某零售客户场景为例，聚焦水平流程。各流程步骤将在下一章节实施步骤中进行详细讲解。

3.2.1 数据管理成熟度诊断

图 3-8 数据管理成熟度诊断



数据管理成熟度评估是一种基于能力成熟度模型框架的能力提升方案，描述了数据管理能力初始状态发展到最优化的过程。在数据管理实施专业服务的整体工作流程中，数据管理成熟度评估作为第一个步骤，串联了数据调研与业务调研，作为一个结果型输出件呈现；一方面在项目早期可以快速与客户建立关系，另一方面在项目交付时更容易讲清数据治理的价值。在交付团队没有参与蓝图顶设的项目中，数据管理成熟度评估是引出数据管理体系设计（咨询）、数据使能方案设计（实施）的关键步骤。

业务现状调研

- **调研目的**

  一般情况下，数据管理实施专业服务开展工作的第一步都是从调研开始的。考虑到数据使能方案往往需要拉通业务与IT，因此调研工作需要并行向业务和数据开展。本章节先讲解业务调研。

  业务调研的目的在于深入了解客户的核心业务流程、需求、挑战和目标，以某零售行业客户为例，具体的调研目标可能包括以下几个方面：

  a. 分析客户的核心业务流程和组织结构，理解客户的业务模式。

  b. 识别业务中的关键挑战和机会，为业务优化提供方向。

  c. 明确业务目标和战略方向，为业务增长提供支持。

- **调研方式**

  业务调研的方式涵盖访谈交流、现场观察、市场分析和竞争对手研究等多种方法，用以深入了解客户的业务现状。以某零售行业客户为例，具体内容可能包括：

版权所有 © 华为技术有限公司

a. 访谈交流： 通过有针对性的沟通了解业务现状，调研对象分为：

管理层人员：深入了解客户企业战略和发展方向，理解客户在当前市场中的定位和竞争态势，了解客户的组织结构与管理流程等。

业务侧人员：调研客户日常业务的流程和效率，了解客户业务当前的需求，以及周边对客户产品或服务的满意度。

b. 实地考察：实地考察参观客户的零售门店或工厂，观察实际运营流程。与不同岗位的人员深入交流，了解各方面的核心诉求。

c. 市场分析：通过市场报告、行业分析等方式，了解客户相关的市场趋势、竞争态势；分析主要竞争对手的业务模式和战略，评估客户在市场中的地位。

d. 集体研讨：包括调研会议、赋能会议、联合周例会等，确保多方面的信息汇聚达成一致。

数据现状调研

- **调研目的**

除了对客户进行业务调研外，数据管理实施专业服务在前期还需要进行数据调研，目的在于了解客户当前数据的质量、一致性、可信度和可用性。以某零售行业客户为例，具体的调研目标可能包括以下几个方面：

a. 了解客户整体数据资产及其关系：通过调研，可以深入探究客户不同业务系统间的数据资产，并揭示数据之间的相关性和相互影响。

b. 识别并改进数据质量问题：调研过程能够识别客户的数据质量问题，包括数据准确性、完整性、时效性等，并针对这些问题提出具体的改进措施。

c. 评估数据治理成熟度：通过对组织的数据治理能力进行深入评估，确保公司的数据管理能力符合不断复杂化和快速变化的市场需求，以及相关的法律法规要求。

d. 明确数据治理的目标与策略：通过数据调研，帮助确定客户的数据治理目标和策略，涵盖数据采集、处理、存储、分析等各个方面的具体需求和目标。

- **调研方式**

数据调研的方式涵盖访谈交流、数据探查、实地考察、集体研讨等多种方法，用以深入了解客户的数据现状。以某零售行业客户为例，数据调研的具体内容包括：

a. 访谈交流： 通过有针对性的沟通了解数据现状，调研对象分为：

研发侧人员：深入了解整体数据体系、主数据、数据性能、数据质量和数据标准现状，收集相关数据问题。

业务侧人员：与运营和门店店主沟通，以理解业务对数据的需求，并提出针对性建议。

b. 数据探查：包括数据血缘和数据质量探查，涉及系统间调用、作业依赖、存储过程依赖、数据表关系、代码字段、数据一致性、数据完整性和数据准确性等方面的分析。

c. 实地考察：实地考察客户总部办公大楼和门店，与不同岗位的人员深入交流，了解各方面的核心诉求。

d. 集体研讨：包括调研会议、赋能会议、联合周例会等，确保项目的协同和效率。

整体调研不仅需要借鉴华为的数据治理实践经验，同时也要参考国内外广泛应用的管理体系，旨在为深入理解客户的数据现状和解决存在的问题提供坚实的支持。这样全面而系统的调研方法确保了项目的专业性和实用性，对客户的持续发展具有积极推动作用。

版权所有 © 华为技术有限公司

源系统数据库表盘点

在数据现状调研过程中，需要对源系统数据库表进行盘点，来实现以下几个目的：

1. 了解客户的数据资产。通过盘点清楚地知道客户拥有哪些数据库、表、字段等数据资产，以便为客户提供相关的数据服务。
2. 分析数据资产的状态。通过分析源端系统库表的结构、属性、关系等，可以了解数据资产的完整性、一致性等状态，有利于后续的数据整合和应用。
3. 识别数据重合和冗余。通过比较不同源端系统库表之间的结构和内容，可以识别出数据重合和冗余的情况，为后续的数据整合提供依据。
4. 评估数据质量。通过对源端系统库表的内容和结构进行审查，可以初步评估数据的质量，识别缺失和错误的数据。
5. 确定数据整合和迁移方案。了解源端系统库表的情况，可以帮助确定最佳的数据库整合和数据迁移方案，以实现数据的同源共享。
6. 为数据应用提供依据。了解源端系统库表的详细情况，可以为后续的数据分析和应用提供依据，帮助构建有效的数据模型和架构。

在盘点动作执行前，需要客户提供所要盘点的数据库IP地址、数据库端口号、数据库名称以及至少有源库表和视图只读权限的账号。

对源系统盘点的方法有很多，例如Haydn平台交付中心的探源功能，可以通过配置相应的探源规则，对待集成的源端系统数据进行探源，方便用户快速获取到要进行数据实施的源系统数据结构，且支持基于探源结果生成入湖清单。

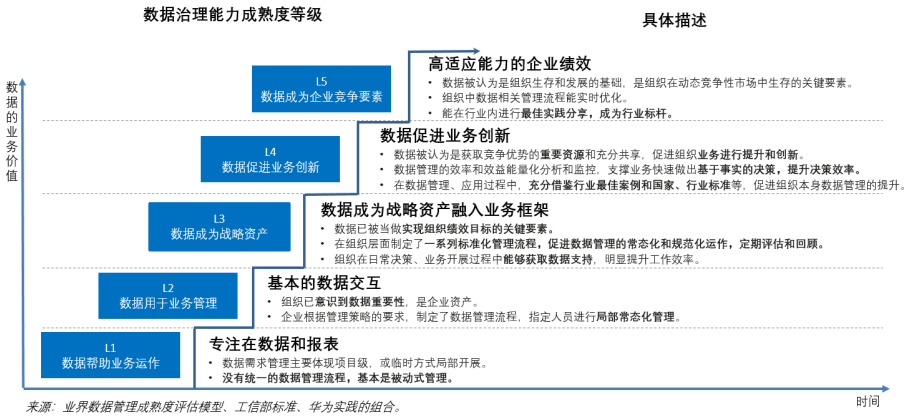
数据管理成熟度评估

数据管理成熟度评估是一种基于能力成熟度模型框架的能力提升方案，描述了数据管理能力初始状态发展到最优化的过程。成数据管理成熟度模型通过描述各阶段能力特点来定义成熟度的级别。当一个组织满足某阶段能力特征时，就可以评估其成熟度等级，并制订一个提高能力的计划。它还可以帮助组织在等级评估的指导下进行改进，与竞争对手或合作伙伴进行比较。在每一个新等级，能力评估会变得更加一致、可预测和可靠。

数据管理成熟度评估将组织内部数据能力划分为数据政策与流程、数据组织、数据标准、数据架构、数据应用、数据质量、主数据、元数据管理、数据安全等9大能力域。每个能力域下面又分为不同的子能力域，共计28个子能力域；每个评估细项按照5个等级模型进行评估，综合平均后计算出整体数据管理成熟度。数据管理成熟度评估成熟度评估等级如下：

版权所有 © 华为技术有限公司

图 3-9 评估等级



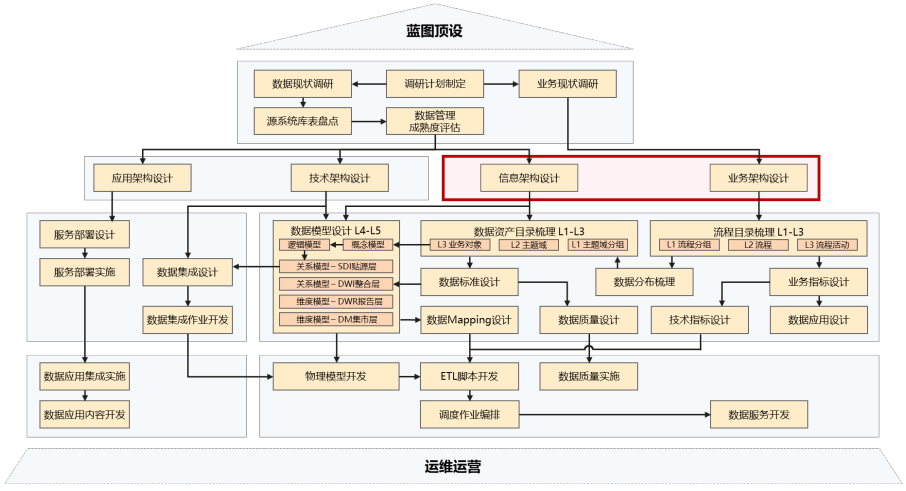
来源：业界数据管理成熟度评估模型、工信部标准、华为实践的组合。

数据管理成熟度评估对企业有巨大的意义：

- 准确把握现状：通过数据管理成熟度评估对组织数据管理的现状进行全面分析，总结当前数据管理工作的优势和劣势
- 明确建设方向：通过数据管理成熟度评估明确数据治理的薄弱项以及和业界标杆的差距，结合企业数字化发展需求，识别亟待完善的数据管理能力，明确数据管理工作的建设方向
- 持续提升管理水平：将评估结果纳入企业数字化转型考核评价体系，定期对组织及其下属单位进行评价考核，实现数据管理能力持续提升

3.2.2 数据使能信息架构与业务架构设计

图 3-10 架构



在完成数据调研与业务调研之后，即可着手进行客户的信息架构与业务架构设计。信息架构和业务架构都是企业架构的重要组成部分。信息架构确保了信息的有效组织和流通，而业务架构确保了企业的流程和结构与其战略目标保持一致。这两者共同为企业提供了高效运作、持续创新、风险管理和客户满意度提升的基础。

版权所有 © 华为技术有限公司

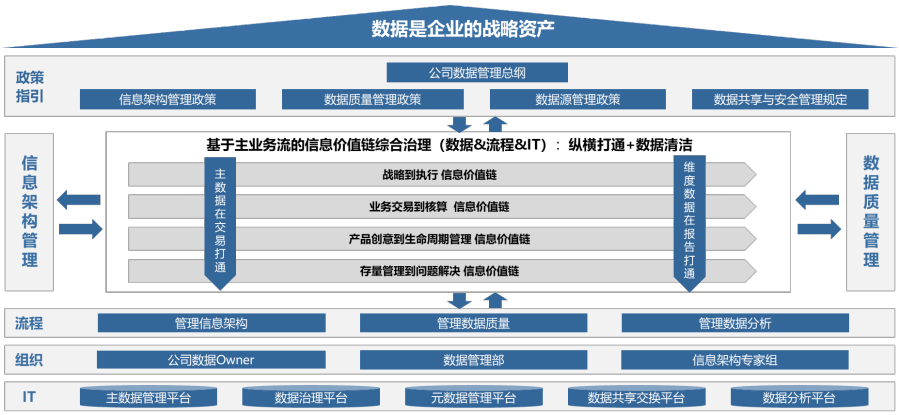
数据使能信息架构设计

信息架构是以结构化的方式描述在业务运作和管理决策中所需要的各类信息及其关系的一套整体组件规范。作为企业架构的数据层面，信息架构强调数据和信息的结构、标准化、整合和治理等方面。它有助于企业有效利用和管理其庞大的数据资产，支持业务流程、决策制定和合规要求，是企业成功运营的关键支柱。

企业在构建和维护有效的信息架构时会面临一些挑战，如数据的复杂性和多样性、技术的快速变化、组织文化和沟通障碍等。正确理解并解决这些挑战是成功实施信息架构的关键。有效的信息架构，可以更好的支撑业务“做正确的事”，支撑技术“正确的做事”。帮助客户构建、优化信息架构，也是数据管理实施专业服务的重中之重。

信息架构承载了企业如何管理数据资产的方法，需要从整个企业层面制订统一的原则，同时也要求一个组织承接这个工作，即数据管理部：

图 3-11 数据管理部 1



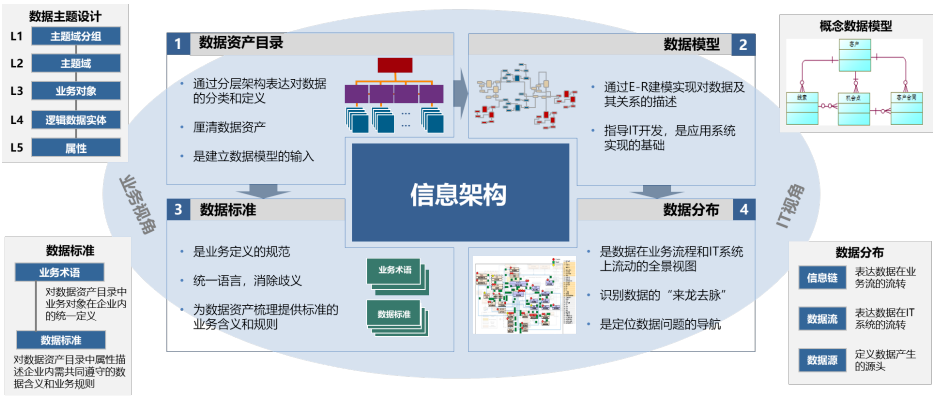
数据管理部负责保证数据治理和管理的效果，保证公司所有业务部门都遵守信息架构原则。通常来说，数据使能解决方案会推荐客户数据管理部制定并遵守以下五条架构原则：

- 原则一：数据按对象管理，明确数据Owner
- 原则二：从企业视角定义信息架构
- 原则三：遵从公司的数据分类管理框架
- 原则四：业务对象结构化、数字化
- 原则五：数据服务化，同源共享

根据华为方法论，数据管理部在保障信息架构原则落地的同时，还要维护好信息架构的四个要素：数据资产目录、数据模型、数据标准、数据分布。

版权所有 © 华为技术有限公司

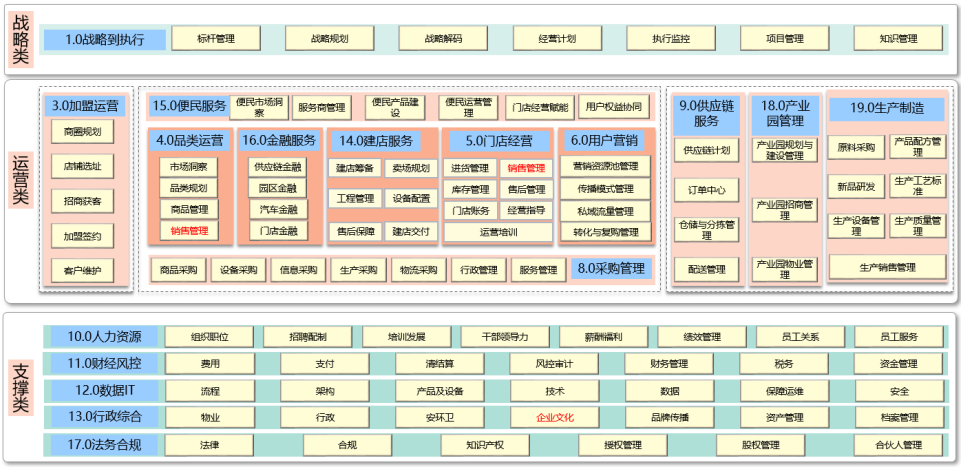
图 3-12 数据管理部 2



数据使能业务架构设计

业务架构（Business Architecture，简称BA）是业务的结构化表达，描述组织如何运用业务的关键要素来实现其战略意图和目标。常见业务架构如下

图 3-13 常见业务架构



业务架构的价值如下：

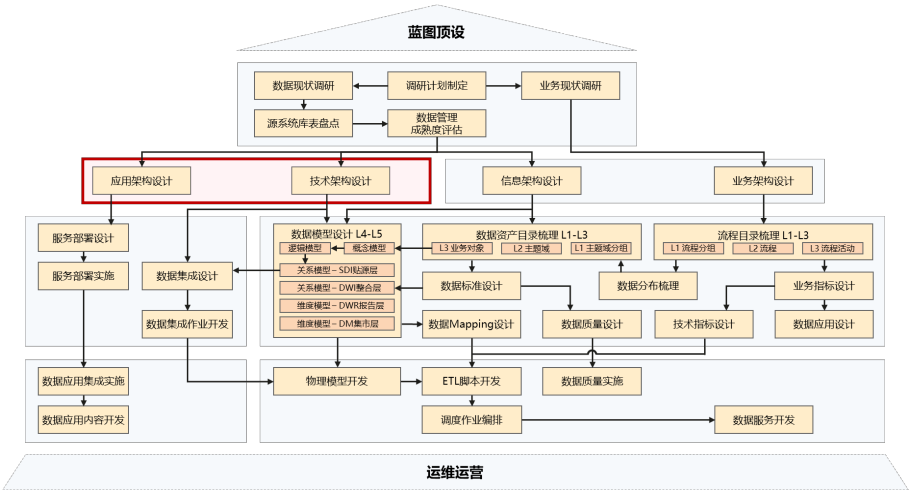
- 业务架构描述了人员，流程，技术，数据等内容如何有效的一起协同工作，为客户提供产品和服务
- 业务架构是企业的业务蓝图，从不同的视角、全面地描述业务，从而确保企业所有人员对业务有一个共同理解
- 业务架构设计是从战略到执行的第一步，将宏观的企业战略进行分解，从战略范畴落实到战术范畴，通过运营支撑业务目标达成
- 业务架构支撑变革：用于组织，流程，IT的建设，支撑数字化转型

业务架构的设计方法分为四个步骤，即分析、设计客户业务的价值流、业务场景、业务流程，并将价值流与业务流程联系。

版权所有 © 华为技术有限公司

3.2.3 数据使能技术架构

图 3-14 数据使能技术架构



应用架构和技术架构是企业架构的技术层面，它们分别关注软件应用和硬件基础设施，以支持企业的业务流程和信息管理。良好的应用架构能够确保软件应用满足业务需求，提供灵活性和可维护性。而合适的技术架构则确保了整体系统的性能、可靠性、扩展性和安全性。这两者共同为企业提供了强有力的技术支持，助力企业实现战略目标，降低成本，增强竞争力。

数据使能技术架构设计：

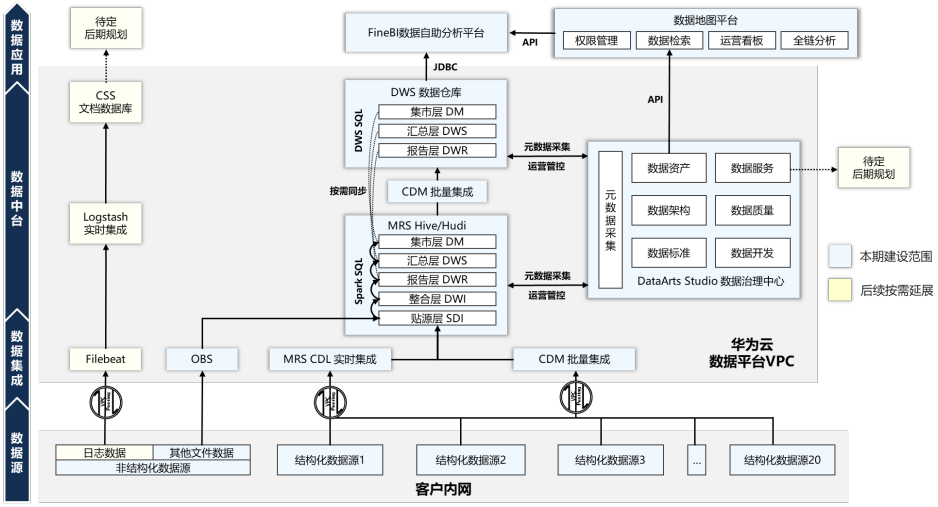
针对企业具体需求，华为云数据管理实施专业服务提供涵盖基础设施、治理、资产管理、可信交换、应用服务的端到端最佳实践和技术方案，解决客户数据实时性差、数据劣质、架构割裂、信息孤岛、缺乏运营的五大问题，从数据的采存算管用出发，华为云提供数据平台所需的全部技术。

以某零售客户为例，客户希望分期完成其企业级数据中台的建设。客户源系统既有结构化数据、又有非结构化数据，同时客户需求部分结构化数据实时进入数据中台，准实时消费。在数据中台的基础上，客户诉求数据使能解决方案帮助其构建企业级的数据地图平台，完成全企业的元数据采集、展示、和审批流构建，同时通过BI实现一键审批+自助分析，打破其找数、取数、用数功能割裂的现状。

针对该客户的需求，设计如下图所示的技术架构，并分两期规划完成数据中台的建设工作：

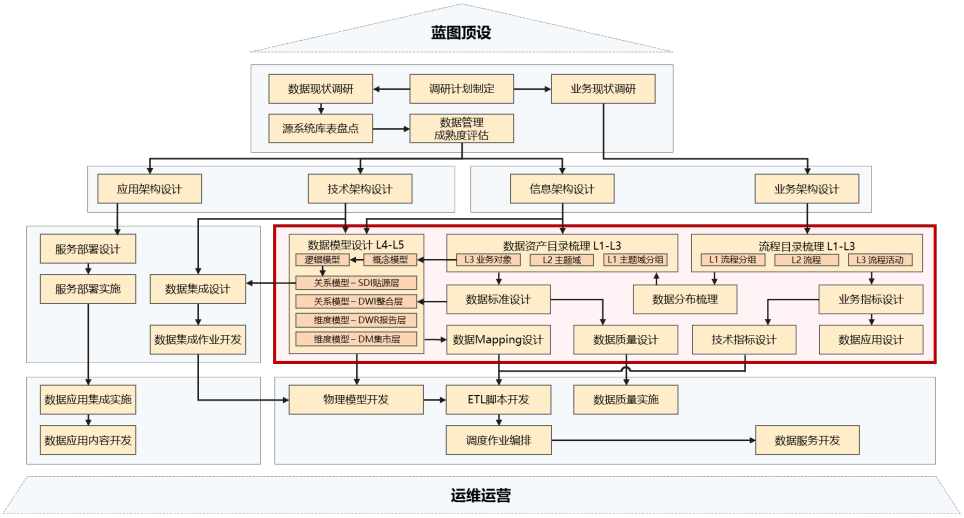
版权所有 © 华为技术有限公司

图 3-15 数据使能技术架构设计



3.2.4 数据使能方案设计

图 3-16 数据使能方案设计



在完成数据使能的4A架构设计后，即可进行数据使能方案设计。数据使能方案设计是数据管理实施专业服务的核心工作，在这个过程中，交付团队会完成流程目录梳理、数据资产目录设计、数据分布梳理、数据标准设计、指标数据梳理、数据模型设计、分层Mapping设计、数据质量设计、业务指标设计、技术指标设计、数据应用设计等工作。最终支撑数据使能解决方案的落地。

流程目录梳理

流程目录（Process Catalog）是一个用于记录和组织企业各类业务流程的集合，它包含了组织内外各个层级的业务流程、主题域分组、主题域、业务对象等信息。流程目录的设计和维护有助于企业在业务管理、流程优化和信息共享方面取得更好的效果。

输出的内容主要有以下两种，业务架构图：

版权所有 © 华为技术有限公司

图 3-17 业务架构图

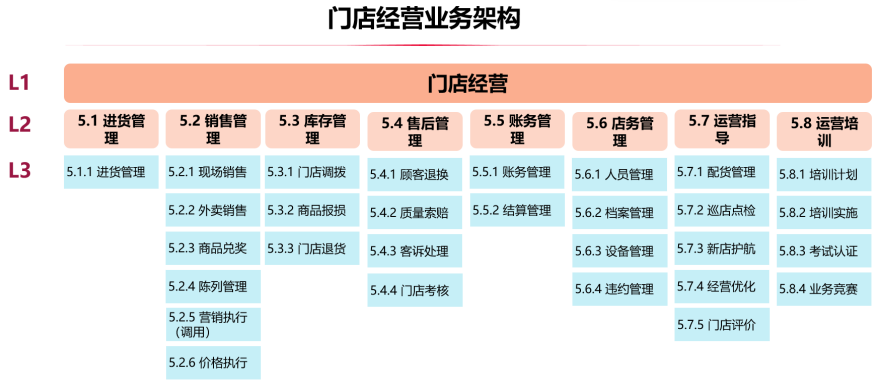
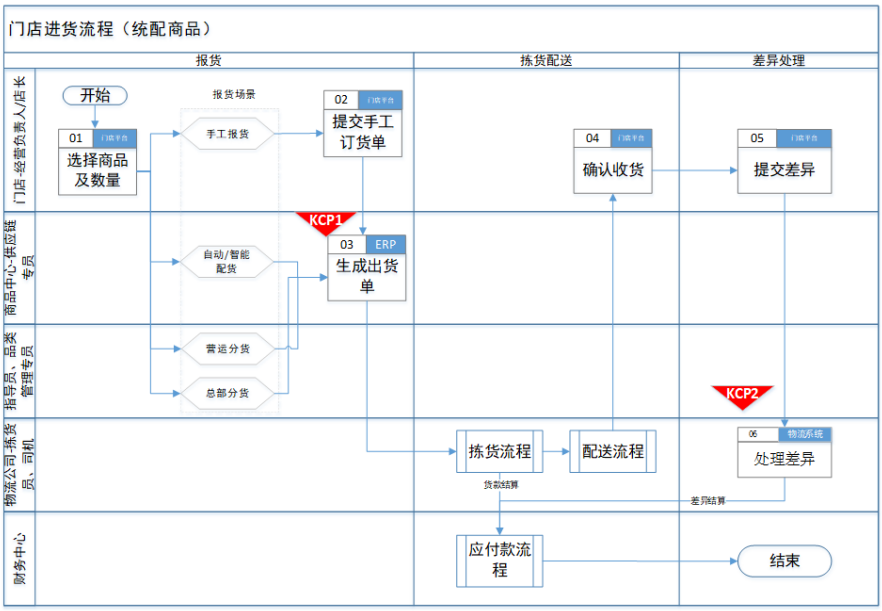


图 3-18 业务流程图



数据资产目录设计

随着数字化转型的推进，企业面临着越来越大的复杂的数据资源，在信息爆炸的背景下，企业内部的数据往往分散在各种系统的部门中，导致数据孤岛现象，造成数据的冗余和低效利用，同时数据管理和治理面临着越来越严峻的挑战。为有效应对这些问题，通过资产的目录的梳理，建立一个全面的、统一和可视的数据资产清单，涵盖企业内部所有数据资源。资产目录将为数据管理和治理提供基础，帮助企业更好地优化数据使用、共享和流转，降低数据管理的复杂性。

信息架构（Information Architecture）：企业级信息架构是以结构化的方式描述在业务运作和管理决策中所需要的各类信息及其关系的一套整体组件规范。信息架构包括数据资产目录、数据标准、企业级数据模型和数据分布四个组件。

数据资产目录是一个组织或企业中用于管理和组织数据资产的结构化文档或系统。它记录了组织内部存在的各种数据资产，包括但不限于数据库、数据集、文件、报告、元数据等信息。数据资产目录的主要目的是为了帮助组织更有效地管理、控制和利用其数据资源。

版权所有 © 华为技术有限公司

数据资产目录有如下作用：

- 数据资产清单：资产目录提供了企业内部所有数据资产的全面清单，包括数据库、表、文件等。它帮助组织了解所有数据资产的属性、用途、所属部门和数据血缘关系，为数据资产提供统一的视图和描述
- 数据管理和治理：资产目录为数据管理和治理提供了基础。通过明确数据资产的归属、负责人和使用规则，资产目录帮助企业更好地管理数据资源，减少数据冗余和重复存储，提高数据质量和安全性
- 数据流程优化：资产目录揭示了数据资产之间的关联和流转路径，帮助企业优化数据流程和数据使用。它使数据流程更加高效，减少数据的滞留和延误，提升数据使用的效率和价值
- 数据决策支持：通过资产目录，企业能够更准确地了解数据资源，从而做出更准确的数据驱动决策。它为业务洞察和智能决策提供依据，推动业务增长和竞争优势
- 数字化转型基础：资产目录是数字化转型的基础设施之一。它帮助企业在数字化转型过程中优化业务流程，提升数字化能力，实现业务模式的创新和提高竞争力
- 数据合规和隐私保护：通过资产目录，企业能够更好地管理和控制敏感数据，确保数据的合规性和隐私保护，降低数据泄露和安全风险

数据分布梳理

数据分布指的是数据在不同的存储系统、节点或位置之间的分布情况。了解数据在哪里存储，如何分布，以及分布情况的变化对数据处理、查询性能和数据安全都非常重要。

在数据资产目录中记录数据的分布信息可以帮助数据使用者更好地了解数据的物理存储位置。这对于查询性能优化很有帮助，使用者可以根据数据分布情况选择更合适的查询方式。此外，了解数据存储位置也有助于数据的隐私和安全管理

在数据流程目录中了解数据的分布情况非常重要。如果数据在不同的节点上分布，数据流程需要考虑如何处理数据移动和传输。避免不必要的数据复制和传输可以提高流程的效率，并减少资源开销。同时，了解数据分布还可以影响数据转换和处理步骤的设计，尽量减少性能问题。

综上所述，数据分布在数据管理中具有重要的影响，涉及到性能、安全性和一致性等多个方面。了解数据分布情况，能够更好地优化数据的使用、处理和流程，并确保数据的质量和安全性。

数据标准设计

数据标准（Data Standards）是进行数据标准化的主要依据，构建一套完整的数据标准体系是开展数据标准管理工作的良好基础，有利于打通数据底层的互通性，提升数据的可用性。

数据标准是指保障数据的内外部使用和交换的一致性和准确性的规范性约束，是对数据的名称、含义、结构、取值等信息的统一定义和规范，以达成对数据的业务理解、技术实现的一致。

数据标准管理是指数据标准的制定和实施的一系列活动，包括明确组织职责和制度规范、构建工具、制标和落标等。通过统一的数据标准制定和发布，结合制度约束、系统控制等手段，实现数据的完整性、有效性、一致性、规范性、开放性和共享性管理，为数据资产管理提供管理依据。

版权所有 © 华为技术有限公司

数据标准是进行数据标准化的主要依据，通过数据标准化，有利于拉通数据，有效提升业务效率和数据质量、促进数据共享。

提升业务效率：数据标准统一了业务语言，明确了业务规则，规范了业务处理过程，从而提升组织整体业务效率，满足管理决策对信息及时性的要求。

提升数据质量：数据标准明确了数据填写及处理要求，规范了数据源的格式，同时提供了管控方面的保障，因此数据标准将直接提高数据质量。

促进数据共享：数据标准统一了各类系统的数据定义，降低了系统间集成的复杂度，提高了系统间交换效率，并为管理分析系统提供了一致的分析指标和分析维度定义。

指标数据梳理

指标数据是指按照确定的计算逻辑，基于交易数据或主数据的一个或多个数据项值加工得到的新数据项，一般由指标名 、指标值、统计口径、指标阈值等组成，又称衍生数据。通过指标数据的标准化，可以统一组织各部门对于指标的理 解，有利于提升统计分析的数据质量。

根据调研报告、IT 系统数据调研表和数据探查结果，明确指标数据的业务用途和目标，确定与目标相关的关键业务指标，如销售额、来客数、经营天数等，进行详细的需求分析，形成需求分析文档：需求指标确认清单（原子指标、衍生指标、复合指标）、指标口径确认清单（业务过程、度量、维度），如下图：

图 3-19 指标数据梳理

售罄次数
1.原始buy2表增加字段：是否会员、是否异常、支付渠道（微信/支付宝等）、业务类型（云销售/智能售卖机）
2.通过清洗整合生成：buy_yc(异常流水表)、buy_stsale_hour(门店每日时段销售表)、buy_gdsale_hour(门店商品每日时段销售表)、buy_sortsale_dl_hour(门店大类每日时段销售表)、buy_sort
3.字段见：附录1

界面模块1	指标类目	指标	指标别名	计算维度	统计维度	*设置目的	*指标定义
经营分析	经营天数	经营天数		本值	按月统计：日、本月、月度	判断门店是否有营业	门店有销售产生的自然日，为有经营的一天
经营分析	销售额	销售总额		本值、同比、环比	按时段统计：昨日、7日、当月、月度 按日统计：昨日、近7天、近30天、当月、月度、年度		销售订单总额
经营分析	销售额	香烟销售		本值、同比、环比	按日统计：昨日、近7天、近30天、当月、月度、年度	通过不同维度了解门店的销售额指标表现	香烟商品的订单总额
经营分析	销售额	线下含烟销售		本值、同比、环比	按日统计：昨日、近7天、当月、月度		线下销售的订单总额，flag为mpos销售和冲单
经营分析	销售额	线下非烟销售		本值、同比、环比	按日统计：昨日、近7天、当月、月度		线下销售的非烟商品订单总额，flag为mpos销售和冲单
经营分析	销售额	会员销售		本值、同比、环比	按时段统计：昨日、7日、当月、月度 按日统计：昨日、近7天、近30天、当月、月度、年度		会员销售的订单总额
经营分析	销售额	会员非烟销售		本值	按日统计：近30天、当月、月度、年度		会员销售的非烟订单总额
经营分析	销售额	会员香烟销售		本值	按日统计：近30天、当月、月度、年度		会员销售的香烟总额
				本值、同比	按时段统计：昨日、7日、当月、月度 按场景统计：不限、工作日、周末		

经营分析业务指标 | 附录1 | 附录2

数据模型设计

DataArts Studio数据架构以关系建模、维度建模理论支撑，实现规范化、可视化、标准化数据模型开发，定位于数据治理流程设计落地阶段，输出成果用于指导开发人员实践落地数据治理方法论。

DataArts Studio数据架构建议的数据分层如下

SDI (Source Data Integration)，又称贴源数据层。SDI是源系统数据的简单落地。

DWI (Data Warehouse Integration)，又称数据整合层。DWI整合多个源系统数据，对源系统进来的数据进行整合、清洗，并基于三范式进行关系建模。

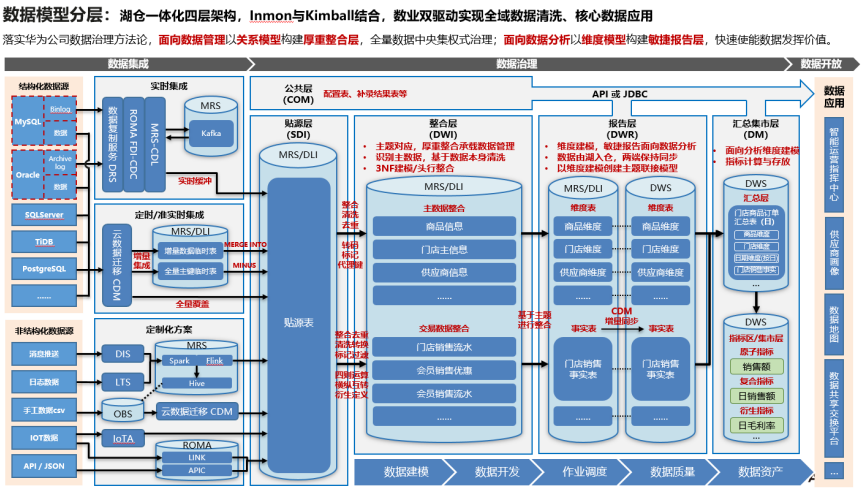
版权所有 © 华为技术有限公司

DWR (Data Warehouse Report)，又称数据报告层。DWR基于多维模型，和DWI层数据粒度保持一致。

DM (Data Mart)，又称数据集市。DM面向展现层，数据有多级汇总。

华为方法论示意图，如下：

图 3-20 华为方法论示意图



分层 Mapping 设计

在数据仓库和ETL（抽取、转换、加载）领域中，"mapping" 指的是一种规则和逻辑的集合，用于描述如何从源数据抽取、转换和加载到目标数据仓库中的过程。映射定义了源系统中的数据如何映射到目标系统中，以满足数据仓库的数据需求和分析目标。

做Mapping的目的如下：

- **数据转换：** 数据从源系统到数据仓库的过程中，往往需要进行各种数据转换，包括单位转换、日期格式标准化、数据清洗、计算等。编写映射可以明确这些转换规则，确保数据在转换过程中的准确性和一致性
- **数据整合：** 数据仓库通常集成来自多个不同源系统的数据，这些数据可能具有不同的格式和结构。编写映射可以将这些不同的数据整合到一个统一的数据模型中，以便进行分析和报告
- **数据质量：** 数据质量是数据仓库的关键因素之一。编写映射时，可以实施数据清洗、去重、标准化等步骤，从而提高数据的质量，减少错误和不一致性
- **业务逻辑应用：** 在数据仓库中，可能需要应用特定的业务逻辑，例如计算指标、创建层级等。通过编写映射，可以确保这些业务逻辑在数据加载过程中得到正确的应用
- **性能优化：** 编写映射时，可以考虑性能问题，使用合适的索引、分区等方法，以提高数据加载和查询性能
- **文档和可维护性：** 编写映射规则和逻辑可以帮助团队成员理解数据转换和加载的过程。这些文档可以作为日后维护和调整的参考
- **可复用性：** 编写映射可以将数据转换规则和逻辑进行抽象和封装，从而实现可复用性，减少重复劳动

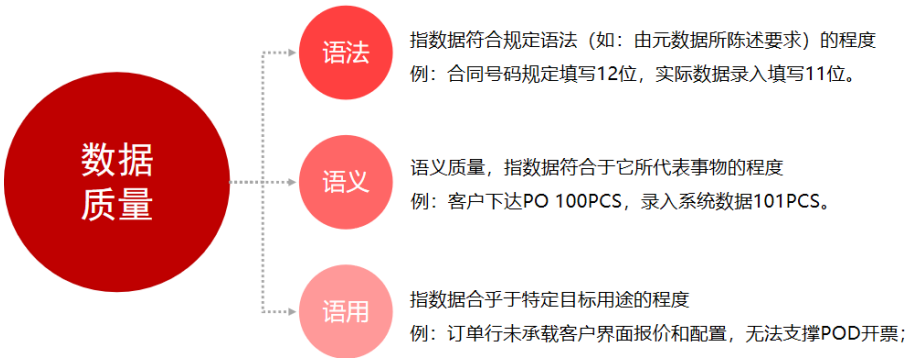
版权所有 © 华为技术有限公司

数据质量设计

随着数据类型、数据来源的不断丰富以及数据量的飞速增长，企业面临数据质量问题的概率显著增加。数据质量是一个复杂问题，往往是多种因素综合作用的结果，解决数据质量问题要从机制、制度、流程、工具、管理等多个方面发力。

ISO8000定义：从语法、语义、语用三个方面去定义和衡量数据质量

图 3-21 数据质量设计



企业数据来源于多个不同的业务系统，数据流转、处理环节多，用“Garbage in Garbage out”原则保证数据质量已成为数字化转型企业的共识。企业数据质量管理是一个系统性的工程，华为数据质量从数据质量领导力、数据质量持续改进、数据质量能力保障三方面展开，有机结合形成联动。华为数据质量指“数据满足应用的可信程度”，从以下六个维度对数据质量进行描述。

- **完整性：**指数据在创建、传递过程中无缺失和遗漏，包括实体完整、属性完整、记录完整和字段值完整四个方面。完整性是数据质量最基础的一项，例如员工工号不可为空。
- **及时性：**指及时记录和传递相关数据，满足业务对信息获取的时间要求。数据交付要及时，抽取要及时，展现要及时。数据交付时间过长可能导致分析结论失去参考意义。
- **准确性：**指真实、准确地记录原始数据，无虚假数据集信息。数据要准确反映其所建模的“真实世界”实体。例如员工的身份信息必须与身份证件上的信息一致。
- **一致性：**指遵循同一的数据标准记录和传递数据和信息，主要体现在数据记录是否规范、数据是否符合逻辑。例如同一工号对应的不同系统中的员工姓名需一致。
- **唯一性：**指同一数据智能有位移的标识符。体现在一个数据集中，一个实体只出现一次，并且每个唯一实体有一个键值且该键值只指向该实体。例如员工有且仅有一个有效工号。
- **有效性：**指数据的值、格式和展现形式符合数据定义和业务定义的要求。例如员工的国籍必须是国家基础数据中定义的允许值。

业务指标设计

业务指标是用于度量和评估组织或业务活动绩效的衡量标准。它们是量化的、可衡量的数据点，用于衡量业务的成功、进展和表现。业务指标通常用来帮助组织了解其绩效状况，监控趋势，做出决策和制定战略。

版权所有 © 华为技术有限公司

设计有效的业务指标是一个关键的过程，它需要深入了解业务需求、关键绩效指标以及如何从数据中衡量这些指标。以下是设计业务指标的一般步骤

- **理解业务目标：** 首先，深入了解组织的业务目标、战略和重要驱动因素。与业务领导和相关团队交流，确保理解业务的核心需求和关注点
- **确定关键绩效指标（KPIs）：** 从业务目标中识别出关键的绩效指标，这些指标能够最直接地反映业务的成功。关键绩效指标应该能够定量地衡量业务的核心结果
- **SMART目标设置：** 为每个绩效指标设置SMART目标，确保它们具有明确的特定性、可衡量性、可实现性、相关性和时限性。这有助于确保指标是具体且有意义的
- **选择适当的度量单位：** 为每个指标选择适当的度量单位，如货币、百分比、数量等，以便进行比较和分析
- **建立度量标准：** 为每个指标定义不同层次的表现标准，例如“优秀”、“良好”、“一般”等。这有助于评估业务绩效
- **数据源和计算逻辑：** 确定每个指标的数据来源，以及如何从底层数据计算或聚合出指标。清楚指标的计算逻辑是确保其准确性的关键
- **数据质量和一致性：** 确保指标所使用的数据源具有高质量和一致性。数据的准确性对于有效的指标分析至关重要
- **时效性和更新频率：** 考虑指标的时效性和更新频率。有些指标可能需要实时更新，而其他指标可以更适合定期更新
- **与业务团队合作：** 与业务团队保持紧密合作，确保指标设计与业务需求保持一致，并及时进行反馈和调整
- **持续改进：** 定期审查和更新指标设计，以确保其仍然适用于不断变化的业务环境

技术指标设计

业务指标用于指导技术指标，用于定义指标的设置目的、计算公式等，并不进行实际运算，可与技术指标进行关联。而技术指标是对业务指标的具体实现，定义了指标如何计算。在华为的数据治理方法论中，技术指标直接关联到业务目标，通过将业务需求翻译为可操作的技术指标，确保数据质量和系统性能达到支持业务决策和运营的水平。这种转化过程将抽象的业务需求转变为具体的度量标准，如数据准确性、数据完整性、数据可用性等，以此来量化业务的影响。这种紧密的关联确保了技术指标的有效性，从而为数据质量的实际提升提供了清晰的路径。通过业务指标与技术指标之间的相互转化，华为能够更加有针对性地设计和执行数据治理策略，实现数据对业务的支持和驱动。

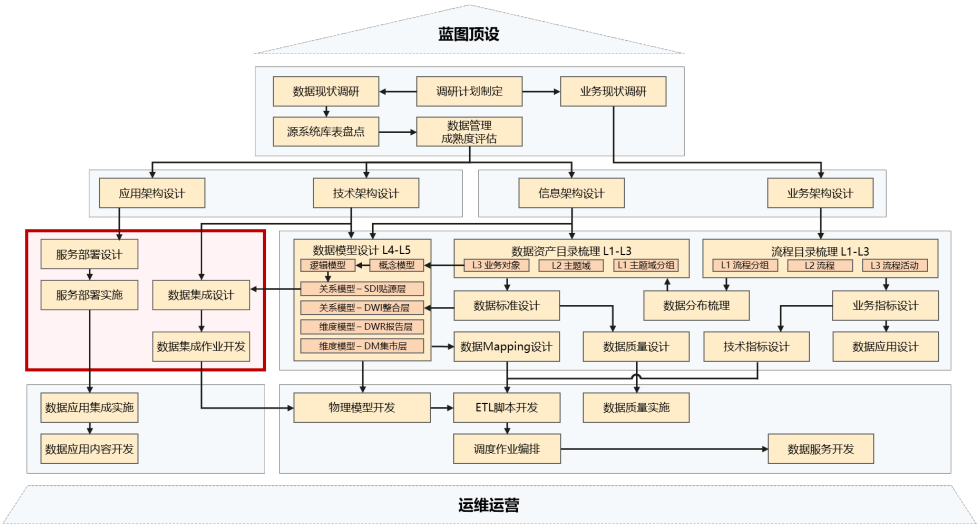
根据华为数据治理方法论，技术指标包含：原子指标，衍生指标，复合指标：

- **原子指标**=业务 + 业务过程 + 度量
- **衍生指标**=修饰词 + 业务规则 + 原子指标
- **复合指标**=计算规则 + 衍生指标 / 原子指标

版权所有 © 华为技术有限公司

3.2.5 数据使能技术平台集成实施

图 3-22 数据使能技术平台集成实施



服务部署设计

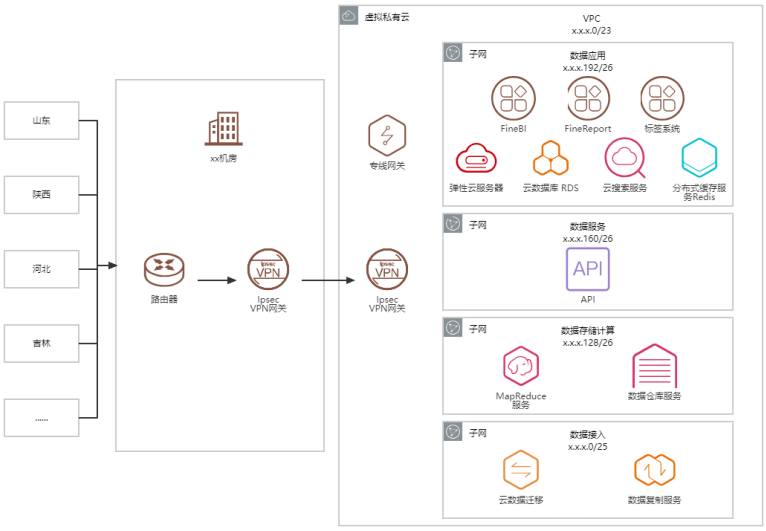
在为客户进行服务部署设计时，网络架构规划是一个非常重要的环节。一个好的网络架构规划可以提高系统的可用性、性能和安全性，同时还能帮助节省成本。在进行网络架构规划时，除了考虑传统的网络因素外，还需要了解相关的技术架构和云服务。技术架构涉及到系统的整体设计和组件之间的交互方式，而云服务则提供了各种功能和工具来支持应用程序的部署和运行。因此，在进行网络架构规划之前，了解所使用的云服务和技术架构是至关重要的。这样可以确保网络架构与技术架构的协调，从而实现系统的高效运行和优化资源利用。以下是进行网络架构规划时需要考虑的一些关键因素：

1. 需求理解和收集：首先要充分理解和收集客户的业务需求和技术需求，包括应用类型、数据流量大小、性能要求、安全性需求、业务发展预期等。这将为后续的网络架构设计提供基础。
2. 网络拓扑设计：根据需求，设计网络的物理布局和逻辑布局。包括决定网络的层次结构，选择合适的网络设备和技术，规划网络地址和路由，等等。
3. 性能考虑：网络架构需要能够满足客户的性能需求，包括带宽、延迟、吞吐量等。可能需要采用负载均衡、冗余链接、多路径路由等技术来提高性能。
4. 安全规划：网络安全是极为重要的，需要考虑如何防止各种安全威胁，如DDoS攻击、数据泄露等。可能需要使用防火墙、入侵检测系统、VPN、数据加密等安全措施。
5. 可扩展性和灵活性：网络架构需要考虑未来的业务增长和技术发展。设计时要考虑到网络的可扩展性和灵活性，以便在不影响现有业务的情况下进行升级和扩展。
6. 成本和ROI：在满足业务需求的同时，还要考虑成本和投资回报率。这包括硬件、软件、维护、升级等各方面的成本。

结合上述关键因素及实际情况，可以给客户提供一个合适的网络架构。下图以本项目为示例，与客户沟通了解到集团正在规划骨干网，因此先期选用VPN接入方式后期转为专线网络，在云上划分一个VPC，该VPC下划分为4个子网分别供给数据接入、数据存储计算、数据服务、数据应用使用。

版权所有 © 华为技术有限公司

图 3-23 服务部署设计



完成网络规划后，可根据网络架构展开细化部署架构。

数据集成设计

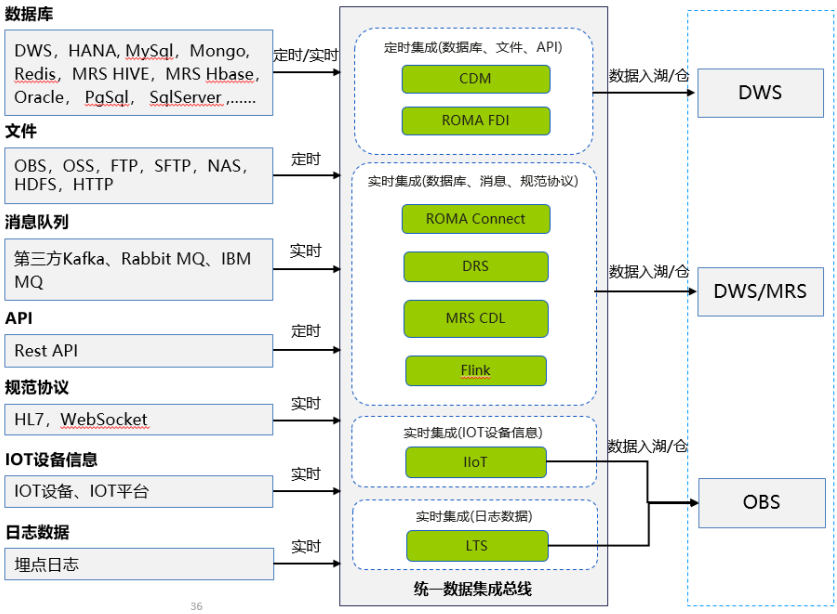
在数据集成中，针对不同场景可以选择不同的数据集成技术栈。以下是一些常见的数据集成技术栈及其适用场景：

- 批量数据集成：使用CDM，适用于批量数据迁移和同步，支持多种数据源和目标数据库。
- 实时数据集成：使用DRS、CDL、Kafka、Flink、IIoT，适用于实时数据流处理和传输以及异步消息传递和解耦，支持多种消息协议和数据格式，低延迟、高吞吐量的数据处理。
- 非结构化数据集成：使用OBS、FTS，适用于大规模文件的传输、同步、存储、访问，支持多种文件格式和存储策略。
- API集成：使用Roma、数据服务，适用于API数据集成、管理、发布，提供安全、高性能的API访问控制和管理功能。

华为云提供多种数据集成服务，针对数据库、文件、消息、API、协议、IOT类数据集成支持可视化快速配置，配套专业服务赋能，大幅降低实施成本。

版权所有 © 华为技术有限公司

图 3-24 数据集成设计



数据集成作业开发

数据集成作业开发是数据使能解决方案中的重要环节，它涉及将不同数据源的数据整合、转换和传输到目标系统的过程。在华为云数据使能解决方案中，提供了多种数据集成技术栈，以满足不同场景的需求。

对于批量数据集成，推荐使用华为云的CDM（Cloud Data Migration）服务。CDM支持多种数据源和目标数据库，可以实现批量数据的迁移和同步。通过可视化界面和配置，用户可以快速设置数据源和目标数据库的连接，并进行数据映射和转换，实现高效的批量数据集成。

对于实时数据集成，提供了多种技术和工具。DRS和CDL可以实现实时数据流处理和传输，支持多种消息协议和数据格式。Kafka和Flink是流式处理框架，可以实现异步消息传递和解耦，具有低延迟和高吞吐量的数据处理能力。IIoT则专注于工业物联网领域的实时数据集成和处理。

对于非结构化数据集成，提供了OBS和FTS。OBS可以用于大规模文件的传输、同步、存储和访问，支持多种文件格式和存储策略。FTS则提供了高速的文件传输服务，可以快速传输大文件和大量文件。

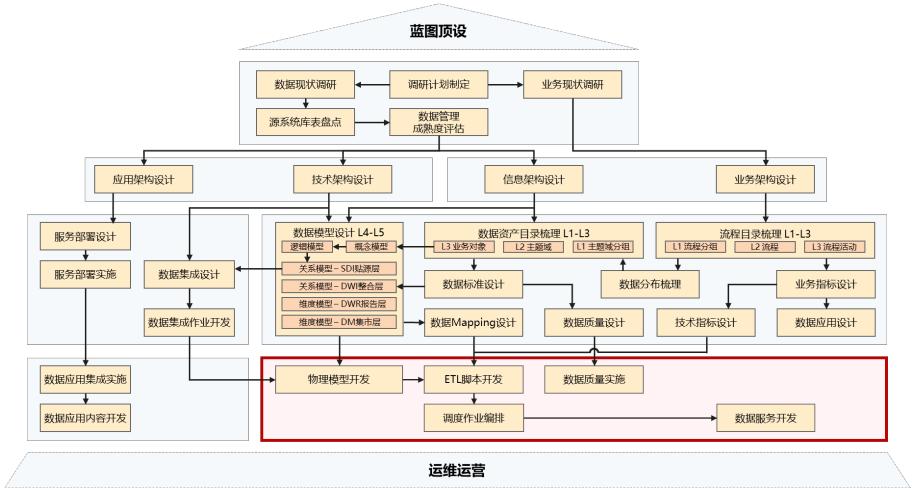
此外，还提供了API集成的解决方案。Roma和数据服务可以帮助用户进行API数据集成、管理和发布。它们提供了安全、高性能的API访问控制和管理功能，使用户能够轻松构建和管理API接口。

在数据集成作业开发过程中，华为云的数据集成服务提供了可视化的快速配置和专业的支持服务，大幅降低了学习成本。用户可以根据具体的需求选择适合的数据集成技术栈，并通过配置和定制来实现高效、可靠的数据集成作业。这将帮助用户实现数据的整合和流动，为业务计算和应用提供可靠的数据基础。

版权所有 © 华为技术有限公司

3.2.6 数据使能方案实施

图 3-25 数据使能方案实施



物理模型开发

数据采集、数据存储和数据处理等关键组件相互协作，为企业提供了高效、可靠的数据处理能力。在模型设计部分，详细介绍了如何设计数据模型，包括数据表的结构、字段定义和关系等。强调了良好的模型设计对于数据处理的重要性，并提供了一些最佳实践和建议。

在物理模型开发中，除了设计数据模型的结构和字段定义，还需要进行以下步骤：

1. 建立模型目录名称：为了组织和管理数据模型，建议在物理模型开发之前先建立一个模型目录。模型目录可以根据业务需求和数据分类进行命名，例如按照功能模块、数据主题或业务流程等进行分类。这样可以方便团队成员查找和维护模型，提高工作效率。
2. 创建逻辑模型：在物理模型开发之前，通常需要先创建逻辑模型。逻辑模型是基于业务需求和数据分析结果设计的模型，它描述了数据之间的关系和业务规则。逻辑模型可以使用实体关系图（ER图）或其他建模工具进行设计和表示。在创建逻辑模型时，需要考虑数据的实体、属性、关系和约束等。
3. 转化物理模型：一旦逻辑模型设计完成，就可以开始转化为物理模型。物理模型是逻辑模型的具体实现，它定义了数据表的结构、字段定义、索引、分区等细节。在转化物理模型时，需要考虑数据库的特性和限制，选择合适的数据类型、约束和索引等。可以使用建模工具或数据库管理工具来创建和管理物理模型。

在建立模型目录名称、创建逻辑模型和转化物理模型的过程中，需要与业务团队和数据开发团队紧密合作，确保模型的准确性和一致性。同时，建议遵循一些最佳实践和建议，如命名规范、数据类型选择、索引优化等，以提高模型的性能和可维护性。

通过良好的物理模型开发，可以确保数据在存储和处理过程中的准确性和一致性，为后续的数据处理和分析提供可靠的基础。

ETL 脚本开发

在开发过程中，开发人员需要仔细阅读并参考开发规范文档，遵循其中的命名规范，并根据mapping表和逻辑文档进行开发，以确保代码的一致性和可读性。

版权所有 © 华为技术有限公司

本示例项目以某零售行业客户为例，采用MRS Hudi+DWS湖仓一体化架构。因此ETL开发主要使用两种数据库：MRS HUDI数据库（使用Spark SQL）和DWS数据库（使用DWS SQL）。

ETL是数据处理中的重要环节，它是一个缩写，代表了数据处理的三个主要阶段：

1. 提取（Extract）：在这个阶段，数据从源系统中提取出来。这可能涉及到连接到数据库、读取文件、调用API等操作，以获取源数据。提取的过程需要考虑数据的完整性、准确性和安全性。
2. 转换（Transform）：在这个阶段，提取的数据经过一系列的转换操作，以满足目标系统的需求。转换操作可以包括数据清洗、数据格式转换、数据合并、数据计算等。转换的目的是将数据转化为目标系统所需的结构和格式，并进行必要的数据处理和修正。
3. 加载（Load）：在这个阶段，经过转换后的数据被加载到目标系统中，通常是一个数据仓库或数据湖。加载的过程需要考虑数据的完整性、一致性和可用性。这可能涉及到数据验证、数据校验、数据分区等操作，以确保数据的质量和可靠性。

在现代企业中，数据量庞大且来源多样化，来自不同的数据源和系统。这些数据可能存在于关系型数据库、日志文件、API接口、云存储等各种形式。ETL的目标是将数据从源系统提取出来，并经过转换后加载到目标系统中，以实现数据的集成、一致性和可用性。通过ETL过程，企业可以将分散的数据整合起来，为数据分析、报告和决策提供可靠的基础。ETL还可以帮助清洗和修复数据，提高数据质量，并支持数据的历史追溯和审计。

华为云的DataArts Studio数据治理中心是一个强大的ETL工具和技术，它可以帮助开发人员设计、编写和管理ETL脚本。以下是DataArts Studio在这些方面的主要功能和优势：

- 可视化的ETL设计：DataArts Studio提供了一个直观的可视化界面，使开发人员能够以图形化方式设计和配置ETL流程。通过拖放组件和连接线，开发人员可以轻松定义数据提取、转换和加载的步骤，而无需编写复杂的代码。
- 内置的数据转换和处理功能：DataArts Studio提供了丰富的内置转换和处理组件，如数据清洗、数据格式转换、数据合并、数据计算等。开发人员可以直接使用这些组件，而无需自行编写转换逻辑，从而加快开发速度并减少错误。
- 强大的数据连接和集成能力：DataArts Studio支持与各种数据源的连接和集成，包括关系型数据库、文件系统、云存储、API接口等。开发人员可以轻松地配置数据源连接，并直接从这些数据源中提取数据。
- 可扩展的脚本编写和管理：虽然DataArts Studio提供了可视化的ETL设计界面，但它也支持自定义脚本编写。开发人员可以使用内置的脚本编辑器编写自定义的ETL脚本，以满足特定的需求。此外，DataArts Studio还提供了ETL脚本的版本控制和管理功能，方便团队协作和脚本的维护。
- 实时监控和调试：DataArts Studio提供了实时监控和调试功能，开发人员可以实时查看ETL流程的执行状态、数据处理的结果和错误信息。这有助于快速发现和解决问题，提高ETL脚本的质量和可靠性。

通过使用华为云的DataArts Studio数据治理中心，开发人员可以更高效地设计、编写和管理ETL脚本。它提供了可视化的ETL设计界面、内置的数据转换和处理功能、强大的数据连接和集成能力、可扩展的脚本编写和管理功能，以及实时监控和调试功能。这些功能使开发人员能够更快速、更准确地开发和维护高质量的ETL脚本。

版权所有 © 华为技术有限公司

数据质量实施

本章节基于数据质量设计，在DataArts上配置质量作业并运行。整体流程可分为以下步骤：

- 质量作业：质量作业将创建的规则应用到建好的表中进行质量监控。
- 新建质量作业：在“质量作业”页面单击“新建”，配置相关参数。
- 管理质量作业：支持对单个质量作业的操作如：运行、启动调度、编辑等。也支持批量运行质量作业，一次最多可批量运行200个。
- 导出质量作业：支持批量导出质量作业，最多可导出200个。选择要导出的质量作业并单击“导出”。
- 导入质量作业：支持批量导入质量作业，最大可导入4M数据的文件。
- 导入质量作业：选择“导入配置”页签，选择模板名称重名策略。如果质量作业名称有重复，则全部导入失败。
- 作业调度：数据质量作业调度采用流水线方式，配置在整合层作业后面，详细配置方法见相关文档。
- 运维管理操作：监控质量作业运行状态，包括成功、失败、运行中和告警。成功表示实例正常结束，结果符合预期。
- 数据质量评分模型：基于DataArts数据质量监控-质量报告，质量评分的满分可设置为5分、10分或100分。默认为5分制，基于表关联的规则评分。评分基于规则评分的加权平均值。
- 数据质量呈现：查看DataArts Studio数据质量监控->质量报告。

作业调度编排

作业调度编排是将之前开发的集成作业、ETL脚本和数据质量作业有机地组合在一起，以实现数据流的自动化处理和监控。通过作业调度编排，可以根据业务需求和时间要求，合理安排和管理数据处理流程，确保数据的准确性和及时性。

在进行作业调度编排时，可以使用DataArts Studio数据治理中心提供的作业调度功能。该功能允许创建调度任务，并将之前开发的集成作业、ETL脚本和数据质量作业作为任务的组成部分。通过定义任务的触发条件、依赖关系和执行顺序，可以实现复杂的数据处理流程，并确保每个作业在正确的时间和顺序下执行。

此外，作业还可以调用自定义脚本，以实现更高级的调度和编排功能。通过这样的集成，可以进一步提升作业调度的灵活性和可扩展性，满足不同业务场景下的需求。

综上所述，作业调度编排是将集成作业、ETL脚本和数据质量作业结合起来，根据业务需要进行自动化调度和编排的重要环节。通过合理规划和管理数据处理流程，可以确保数据的质量和及时性，为业务决策提供可靠的数据支持。

数据服务开发

数据服务模块提供快速将数据表生成数据API的能力，同时支持将现有的API快速注册到数据服务平台以统一管理和发布。通过数据服务模块，可以将数据仓库集市层中的业务数据转化为易于访问和使用的API。这样，数据应用可以直接从API中获取所需的数据，无需直接访问底层数据表。这种方式不仅提供了更高的数据安全性，还能够简化数据应用的开发和维护过程。

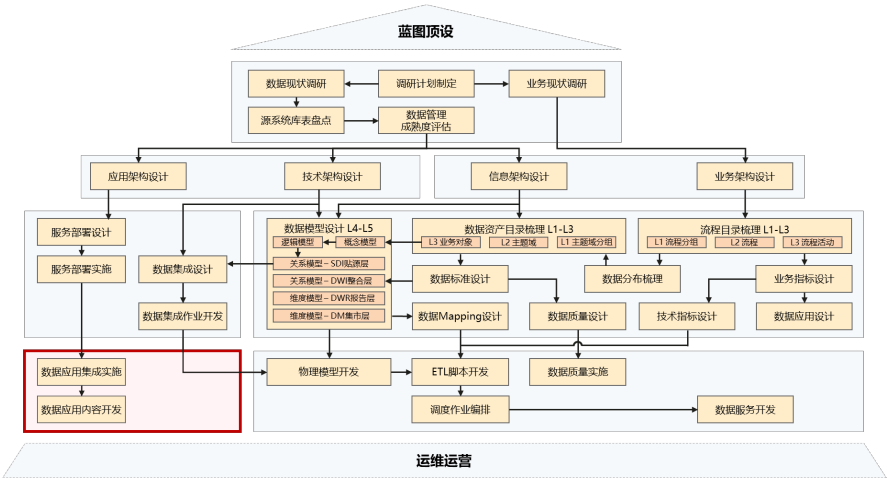
版权所有 © 华为技术有限公司

通过将数据转化为API，可以实现数据的实时更新和动态查询。数据应用可以根据业务需求，灵活地调用API获取最新的数据，并进行实时分析和决策。同时，API还可以提供数据的标准化和格式化，确保数据的一致性和可靠性。

综上所述，通过数据服务模块，可以实现数据应用的高效消费和灵活使用。这一步骤将为企业提供更多的数据驱动能力，促进业务的创新和发展。

3.2.7 数据应用集成设计与实施

图 3-26 数据应用集成设计与实施



数据应用集成设计与实施作为数据管理实施专业服务中的一个非必选项，主要面向客户数据消费场景多，IT能力弱的场景提供服务。工作内容包括数据应用集成实施和数据应用内容开发，将多个数据应用进行打通组合，或针对特定业务场景进行报表设计与开发。这部分工作偏向于IT侧和业务应用侧，不属于数据平台上线的必要条件。

数据应用集成实施：

数据应用集成实施的目标是将多个数据应用进行打通组合，以实现数据的无缝流动和共享。这涉及到将不同的数据应用系统进行整合，确保它们能够相互协作，实现数据的互通。通过数据应用集成实施，企业可以消除数据孤岛，提高数据的可访问性和可用性，从而更好地支持业务决策和运营。

数据应用集成实施还包括针对特定业务场景的报表设计与开发。通过深入了解客户的业务需求和数据消费场景，数据应用集成团队可以设计和开发定制化的报表，以满足客户对数据分析和可视化的需求。这些报表可以帮助客户更好地理解和利用数据，支持业务决策和业务优化。

需要注意的是，数据应用集成实施更偏向于IT侧和业务应用侧，而不是数据平台上线的必要条件。它主要关注数据的整合和应用，以提供更好的数据支持和业务价值。因此，在数据应用集成实施过程中，需要密切协作的是IT团队、业务应用团队和数据管理团队，以确保数据的有效集成和应用。

综上所述，数据应用集成实施是数据管理实施中的一个重要环节，它通过整合数据应用系统和开发定制化报表，为客户提供数据的无缝流动和共享，以支持业务决策和优化。在实施过程中，需要充分协作和沟通，确保数据的有效整合和应用。

版权所有 © 华为技术有限公司

3.3 数据应用

3.3.1 自助分析平台

自助分析平台简介

基于华为云商店产品，帆软FineReport和FineBI，华为和帆软共同为本项目开发了自助分析平台。

FineReport，企业级web报表工具，支持通过简单拖拽操作便可制作中国式复杂报表，轻松实现报表的多样展示、交互分析、数据录入等需求。借助于FineReport的无码理念，可以轻松地构建出灵活的数据分析，网络直报等应用系统，大大缩短项目周期，减少实施成本。FineReport支持普通报表、聚合报表、决策报表三种模式。

图 3-27 报表类型

FineBI，新一代自助大数据分析的BI工具，旨在帮助企业的业务部门用户充分了解和利用数据。FineBI支持在dashboard面板中简单拖拽操作，便能制作出丰富多样的数据可视化信息，并可以进行数据钻取，联动和过滤操作，自由地对业务经营过程中产生的数据进行分析和探索，及时地做出经营决策调整，让大数据释放出更多未知潜能。

FineReport与FineBI结合，为企业实现多重视角的报表加分析的自助分析平台。

图 3-28 报表使用场景

版权所有 © 华为技术有限公司

自助分析平台具体实现

面向客户零售连锁自助分析场景，以本期项目数据治理为基础，基于销售、订货、生产、配发、库存、会员、店员等各个业务主题的数据，自助分析平台实现多维度自助分析，打造业务赛马场，激发门店执行力，数字化会员门店战报、红黑榜，刺激门店拉新动力。同时嵌入各业务系统，提高工作效率，指导一线员工执行业务动作。生鲜品类毛利率同比提升15%-30%，门店整体毛利率同比提升5%-10%。整个自助分析平台具体情况如下：

图 3-29 自助分析平台系统状况

3.3.2 用户推荐平台

用户推荐平台简介

本项目建设的用户推荐平台基于华为云客户数据平台CDP。用户推荐平台以治理后的数据为基础，使用高效、准确的用户推荐模型，建设消费者数据中心并建立人群画像，挖掘用户行为数据，驱动业务决策实现智能化全场景营销体验，助力企业实现精细化运营和营销。

用户推荐平台具体实现

以华为云客户数据平台为基础建设本项目用户推荐平台，以本期项目数据治理为数据基础，主要基于用户域和销售域数据，使用CDP客户数据平台和UGA增长分析平台，进行用户偏好分析、用户行为洞察分析、基于成交的投放效果分析等内容，达到邀约进店率提升20%+，养客成交率提升15%+。

本期建设以企业内部私域数据为主进行用户推荐，具体技术架构如下：

版权所有 © 华为技术有限公司

图 3-30 技术架构 1

未来将纳入企业外部数据，建设公域推荐平台：

图 3-31 技术架构 2

版权所有 © 华为技术有限公司

4 修订记录

表 4-1 修订记录

发布日期	修订记录
2024-04-23	规范词、敏感词专项处理，章节优化
2022-08-28	第一次正式发布。

版权所有 © 华为技术有限公司