HZ BOOKS
华章图书

ENTERPRISE DATA AT HUAWEI

华为公司数据管理部 ◎ 著

华为数据之道

华为质量与流程IT、华为云、华为大学**联合出品**

总结华为公司数据治理、数字化转型方面的实践经验

机械工业出版社
China Machine Press

华为数据之道

ENTERPRISE DATA AT HUAWEI

华为公司数据管理部　著

ISBN：978-7-111-66704-9

本书纸版由机械工业出版社于2020年出版，电子版由华章分社（北京华章图文信息有限公司，北京奥维博世图书发行有限公司）全球范围内制作与发行。

版权所有，侵权必究

客服热线：+ 86-10-68995265

客服信箱：service@bbbvip.com

官方网址：www.hzmedia.com.cn

新浪微博 @华章数媒

微信公众号 华章电子书（微信号：hzebook）

目　录

序一

序二

序三

前言

第1章　数据驱动的企业数字化转型

1.1　非数字原生企业的数字化转型挑战

1.1.1　业态特征：产业链条长、多业态并存

1.1.2　运营环境：数据交互和共享风险高

1.1.3　IT建设过程：数据复杂、历史包袱重

1.1.4　数据质量：数据可信和一致化的要求程度高

1.2　华为数字化转型与数据治理

1.2.1　华为数字化转型整体目标

1.2.2　华为数字化转型蓝图及对数据治理的要求

1.3 华为数据治理实践

1.3.1 华为数据治理历程

1.3.2 华为数据工作的愿景与目标

1.3.3 华为数据工作建设的整体思路和框架

1.4 本章小结

第2章 建立企业级数据综合治理体系

2.1 建立公司级的数据治理政策

2.1.1 华为数据管理总纲

2.1.2 信息架构管理政策

2.1.3 数据源管理政策

2.1.4 数据质量管理政策

2.2 融入变革、运营与IT的数据治理

2.2.1 建立管理数据流程

2.2.2 管理数据流程与管理变革项目、管理质量与运营之间的关系

2.2.3 通过变革体系和运营体系进行决策

2.2.4 数据治理融入IT实施

2.2.5 通过内控体系赋能数据治理

2.3 建立业务负责制的数据管理责任体系

2.3.1 任命数据Owner和数据管家

2.3.2 建立公司层面的数据管理组织

2.4 本章小结

第3章 差异化的企业数据分类管理框架

3.1 基于数据特性的分类管理框架

3.2 以统一语言为核心的结构化数据管理

3.2.1 基础数据治理

3.2.2 主数据治理

3.2.3 事务数据治理

3.2.4 报告数据治理

3.2.5 观测数据治理

3.2.6 规则数据治理

3.3 以特征提取为核心的非结构化数据管理

3.4 以确保合规遵从为核心的外部数据管理

3.5 作用于数据价值流的元数据管理

3.5.1 元数据治理面临的挑战

3.5.2 元数据管理架构及策略

3.5.3 元数据管理

3.6 本章小结

第4章 面向“业务交易”的信息架构建设

4.1 信息架构的四个组件

4.1.1 数据资产目录

4.1.2 数据标准

4.1.3 数据模型

4.1.4 数据分布

4.2 信息架构原则：建立企业层面的共同行为准则

4.3 信息架构建设核心要素：基于业务对象进行设计和落地

4.3.1 按业务对象进行架构设计

4.3.2 按业务对象进行架构落地

4.4 传统信息架构向业务数字化扩展：对象、过程、规则

4.5 本章小结

第5章 面向“联接共享”的数据底座建设

5.1 支撑非数字原生企业数字化转型的数据底座建设框架

5.1.1 数据底座的总体架构

5.1.2 数据底座的建设策略

5.2 数据湖：实现企业数据的“逻辑汇聚”

5.2.1 华为数据湖的3个特点

5.2.2 数据入湖的6个标准

5.2.3 数据入湖方式

5.2.4 结构化数据入湖

5.2.5 非结构化数据入湖

5.3 数据主题联接：将数据转换为“信息”

5.3.1 5类数据主题联接的应用场景

5.3.2 多维模型设计

5.3.3 图模型设计

5.3.4 标签设计

5.3.5 指标设计

5.3.6 算法模型设计

5.4 本章小结

第6章 面向“自助消费”的数据服务建设

6.1 数据服务：实现数据自助、高效、复用

6.1.1 什么是数据服务

6.1.2 数据服务生命周期管理

6.1.3 数据服务分类与建设规范

6.1.4 打造数据供应的“三个1”

6.2 构建以用户体验为核心的数据地图

6.2.1 数据地图的核心价值

6.2.2 数据地图的关键能力

6.3 人人都是分析师

6.3.1 从“保姆”模式到“服务+自助”模式

6.3.2 打造业务自助分析的关键能力

6.4 从结果管理到过程管理，从能“看”到能“管”

6.4.1 数据赋能业务运营

6.4.2 数据消费典型场景实践

6.4.3 华为数据驱动数字化运营的历程和经验

6.5 本章小结

第7章 打造“数字孪生”的数据全量感知能力

7.1 “全量、无接触”的数据感知能力框架

7.1.1 数据感知能力的需求起源：数字孪生

7.1.2 数据感知能力架构

7.2 基于物理世界的“硬感知”能力

7.2.1 “硬感知”能力的分类

7.2.2 “硬感知”能力在华为的实践

7.3 基于数字世界的“软感知”能力

7.3.1 “软感知”能力的分类

7.3.2 “软感知”能力在华为的实践

7.4 通过感知能力推进企业业务数字化

7.4.1 感知数据在华为信息架构中的位置

7.4.2 非数字原生企业数据感知能力的建设

7.5 本章小结

第8章 打造“清洁数据”的质量综合管理能力

8.1 基于PDCA的数据质量管理框架

8.1.1 什么是数据质量

8.1.2 数据质量管理范围

8.1.3 数据质量的总体框架

8.2 全面监控企业业务异常数据

8.2.1 数据质量规则

8.2.2 异常数据监控

8.3 通过数据质量综合水平牵引质量提升

8.3.1 数据质量度量运作机制

8.3.2 设计质量度量

8.3.3 执行质量度量

8.3.4 质量改进

8.4 本章小结

第9章 打造“安全合规”的数据可控共享能力

9.1 内外部安全形势，驱动数据安全治理发展

9.1.1 数据安全成为国家竞争的新战场

9.1.2 数字时代数据安全的新变化

9.2 数字化转型下的数据安全共享

9.3 构建以元数据为基础的安全隐私保护框架

9.3.1 以元数据为基础的安全隐私治理

9.3.2 数据安全隐私分层分级管控策略

9.3.3 数据底座安全隐私分级管控方案

9.3.4 分级标识数据安全隐私

9.4 “静”“动”结合的数据保护与授权管理

9.4.1 静态控制：数据保护能力架构

9.4.2 动态控制：数据授权与权限管理

9.5 本章小结

第10章 未来已来：数据成为企业核心竞争力

10.1 数据：新的生产要素

10.1.1 数据被列为生产要素：制度层面的肯定

10.1.2 数据将进入企业的资产负债表

10.1.3 数据资产的价值由市场决定

10.2 大规模数据交互的企业数据生态

10.2.1 数据生态离不开底层技术的支撑

10.2.2 数据主权是数据安全交换的核心

10.2.3 国际数据空间的目标与原则

10.2.4 多方安全计算强化数据主权

10.3 摆脱传统手段的数据管理方式

10.3.1 智能数据管理是数据工作的未来

10.3.2 内容级分析能力提供资产全景图

10.3.3 属性特征启发主外键智能联接

10.3.4 质量缺陷预发现

10.3.5 算法助力数据管理

10.3.6 数字道德抵御算法歧视

10.4 第四个世界：机器认知世界

10.4.1 真实唯一的“物理世界”和五彩缤纷的“人类认知世界”

10.4.2 映射“物理世界”的数字孪生——“数字世界”

10.4.3 “数字世界”中的智能认知——“机器认知世界”

10.5 本章小结

序一

第三次工业革命带来了机器的进步，但却不能解决一个行业或者一家企业的运营问题，运营效率低下带来的运营成本居高不下，已经成为一个时代性的难题。随着第四次工业革命的来临，数字化生产已经成为普遍的商业模式，其本质是以数据为处理对象，以ICT平台为生产工具，以软件为载体，以服务为目的的生产过程。数字技术和数字平台能够从根本上解决这个时代性的问题，使企业有可能在产品、体验和成本三要素上同时做到最优。

全球各行各业都在积极探索和开展数字化建设，期望通过数字化技术来支撑业务的长期、持续增长。华为作为一家拥有30多年历史、全球领先的ICT基础设施和智能终端供应企业，同样有着数字化转型的强烈愿望。

华为为什么要进行数字化转型?

华为是一家业务范围涵盖研发、营销、制造、供应、采购、服务等领域的非数字原生企业，在信息化时代初期建立了很多相对独立的IT系统，典型的特点是形成了“一类业务、一个IT系统、一个数据库”的封闭式IT架构。其带来的直接问题就是“数据孤岛”：IT系统中的数据语言不统一，不同IT系统之间的数据不贯通，同样的数据需要在不同IT系统中重复录

入，甚至不同IT系统中的同一个数据不一致等。这些问题限制了运营效率的提升和效益的改进，华为迫切需要数字化转型来改变这种状况。

华为规划的数字世界是什么样子的?

其内容无外乎就是业务对象、业务过程和业务规则的数字化，华为希望构建一个实现感知、联接和智能的数据平台。感知是物理世界与数字世界之间形成完整且有效的映射，联接是把各种离散的数据相互联系成有机整体，智能是在这个基础上加入一些大数据和高级模型算法。

华为如何进行数字化转型?

首先，要抓住数据治理这个“牛鼻子”。华为的IT系统和数据有太多的历史包袱，要进行数据治理并不容易，到今天为止，我们所做的也只能说“刚刚及格”。我们想要在构建新的数据平台时不对原有的信息系统进行颠覆性改造。因此我们一方面通过感知能力实现业务数据的自动采集，另一方面通过一些技术手段，把现有的各个相对独立的数据库中的数据按一定的标准进行汇聚和联接。这就带来了“数据湖”的全新体验，先初步解决“数据孤岛”的问题，然后再来进行深入的数据治理。

数字化转型是当前各个行业的各个企业最关心的话题，是一次大的机遇，也是一次大的挑战。现在业界的数字化转型过

多地强调了技术的动因，而我认为数字化转型应该首先强调业务价值。根据Paul Romer的《 内生经济理论》，我们在做数字化转型时要反复问自己：

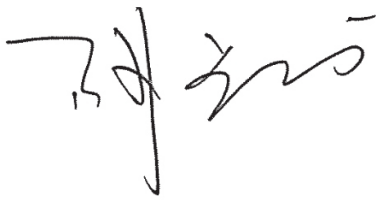
第一，数字化转型到底要解决客户的什么问题？用户到底需要什么？用户和客户关心的问题在哪？

第二，业务战略到底要解决业务的什么问题？

第三，变革是否有一个好的规划和持续的架构？

数字化转型是一个持续优化的过程，只有起点，没有终点。

本书是华为视角的数据治理总结，其中的内容都是华为在数字化转型实践中的经验和教训。数据治理是一件很专业的事情，我们希望本书能给同行提供借鉴、带来启发，也能促成业界与我们的深入探讨和研究，共同推进企业的数字化转型。



陶景文

华为董事、质量与流程IT总裁、CIO

序二

2017年初，数字化大潮方兴未艾，华为轮值董事长郭平在公司“817变革战略规划”中提出，要在内部率先实现数字化转型，并把实现ROADS体验、全面提升运营效率作为公司各业务单元和功能领域的共同变革目标。对于集研发、制造、采购、供应、销服于一体，横跨ToB、ToC业务领域，运营30余年的一家传统企业，如何用数字化的手段来全面改造公司的流程和IT，改变支撑近20万人有效运作的运营模式，成为华为公司变革指导委员会讨论的焦点。

不同于数字世界的“原住民”，非数字原生企业的数字化转型是企业的一次巨大变革。这场变革涉及商业模式、运营模式的变化，需要完成流程、组织、IT、文化等多方面的转变，对于飞速发展的华为来说，相当于在高速路上换轮胎。华为当时面临的局面是，存量的IT“烟囱”遍布各个业务但又支撑着海量的交易和分析，各种短期见效的数据搬家、自动化小工具逐渐从“帮手”变成了“帮凶”，数据被“私有化”为各个业务部门的“资产”，“表哥表姐”为了实现数字化运营加班加点整理Excel，高薪招来的数据科学家却因为没有数据而闲得离职……

变革指导委员会经过充分的讨论达成共识：数字化转型要坚持业务和技术的双轮驱动，而连接双轮的“轴”就是数据。只有建立统一、清洁、智能的数据底座，才能支撑公司不断发

展的新业务，支撑各个区域市场的差异化需求，实现“数据实时可视、海量业务自动、算法支撑决策”，实现“万物互联的智能世界”。

2017年10月，“统一数据底座建设”项目立项。针对数据搬家多、找不到、读不懂、获取难、不敢信等痛点，将“打破数据孤岛，支撑数字化转型，实现数据随需共享、敏捷自助、安全合规”作为项目目标。项目组一手抓数据入湖与联接，一手抓数据消费，经过两年多的努力，终于基本完成了数据底座的建设。今天，数据底座支撑着华为在全球170多个国家的差异化运营，支撑着公司各BG海量的交易与分析，驱动了交付、供应、财经等诸多领域的运营模式（在线、远程、集中）转型，也帮助公司实现了在美国极限施压下的快速分析与应对。数据底座成为华为数字化转型的基石。

数字化变革不仅仅是技术的变革，华为数据底座的建设过程充分说明了这一点。本书是华为数字化转型过程中数据变革实战的阶段总结，希望能给数字化转型道路上的企业一些帮助，同时也欢迎各方朋友交流、指正。



熊康

华为企业架构与变革管理部部长

序三

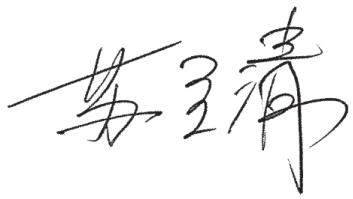
企业数字化转型是大势所趋，通过数据科学治理、数据平台建设、数据分析与建模，把数据变成服务，使数据能在企业内顺畅流动起来，为企业带来巨大的价值。数据是企业数字化转型的重要基础。

为了实现数字化转型，企业需要构建以云为基础、以数据为驱动的新型企业IT架构。但是，多年积累下来的存量IT系统和大量历史数据怎么办？华为数字化转型的核心理念是“双模IT、立而不破”，通过建立新老环境融合的双模（Bi-Model）IT架构，把企业的新老数据和应用与正在及未来将要产生的IoT数据连接在一起，构建统一的数据与应用平台，并与机器学习、人工智能等技术手段相结合，使数据产生更大的价值。

数据的潜在应用场景有很多，但是企业只有将数据与自身业务相结合，从业务实际问题出发，结合数据分析技术找到解决方案，并及时变现，才有真正的意义。数据时代已经到来，华为在深入数字化转型的过程中，在全球供应链、松山湖制造工厂以及170多个国家的工程交付现场，以“对象数字化、过程数字化和规则数字化”为基本原则，通过IT工具引入和人工智能元素的叠加，实现从过去的“人拉肩扛”到现在的线下、线上高效协同，为业务创造了很大的价值。

本书对华为公司多年数据治理和数据消费变革历程进行了系统性总结，从治理体系、架构方法、流程规范、IT工具、数据组织等多方面总结了企业在数据治理中面临的挑战及其解决方案，并介绍了一些华为独有的创新成果，如数据底座、数据湖、主题联接、数据地图、数据生态等。

相信华为在数字化转型历程中所积累的实践、方法和思考，会使奋战在各行各业的企业家和IT同人产生共鸣。让数据治理更简单，使用更简单，能够更方便地挖掘数据价值，是我们共同的期许。企业数字化转型不可能一蹴而就，而是一个长期的过程。本书将华为在数据管理和数字化转型实践中沉淀的能力对外分享，与志同道合的业界同人一起切磋、联合创新，一定可以加速企业数字化转型的进程。



苏立清

华为云副总裁、华为云首席数字化转型官

前言

随着数字化转型的深入开展，数据成为新的生产要素。对于非数字原生企业，数据治理的重要性越来越突出。如何有效地开展数据治理工作、提升数据质量、打破数据孤岛、充分发挥数据的业务价值，成了业界的热门话题。本书基于华为数据治理的历程，介绍了华为数据工作的愿景、整体思路框架，阐述了企业级数据综合治理体系和方法论，回顾了华为数据底座的建设过程，总结了华为数据治理和数字化转型的经验。

华为公司作为典型的非数字原生企业，发展初期基本是以物理世界为中心构建的，缺乏以软件和数字平台为核心的数字世界架构，在数字化转型过程中，面临着巨大挑战。华为从小到大、从弱到强几十年不断发展的历程中，伴随着一次次重大业务变革、信息化建设和数字化转型。本书理论结合实践，通过对华为公司数据治理体系和数据底座建设方法与实践的介绍，讲述了数据工作如何支撑业务变革，如何驱动数字化转型，总结了华为数据工作的发展历程、经验和对未来的思考。书中所述的方法、规范、解决方案都经过华为内部的充分实践，相信对企业数字化转型的领导者、设计者、实施者和参与数据治理的同行，会有一定的启发和借鉴意义。

内容简介

本书共10章，内容从逻辑上可以分为四部分。

第一部分为第1～3章。第1章以非数字原生企业在数字化转型时面临的挑战为引导，阐述了数据驱动的企业数字化转型理念，介绍了华为公司的数据治理框架；第2章从企业政策和架构协同的角度，介绍了企业级的数据综合治理体系，理顺了数据与变革、运营、IT之间的协同关系，明确了数据管理的责任主体在业务；第3章以数据特性的差异为维度，详细阐述了不同类型数据的不同管理方式，明确了结构化数据、非结构化数据、外部数据、元数据的核心管理要点。

第二部分为第4～6章，介绍了数据治理工作中的三项重点建设任务：信息架构、数据底座、数据服务。第4章介绍了信息架构的四个组件，给出了建设原则和核心要素，并引出了业务对象、过程、规则三项数字化的建设方向；第5章提出了数据底座建设的整体框架，分别介绍了数据湖和数据主题联接两个层次的建设实践；第6章以自助、高效、复用为数据服务的目的，提出了对数据进行搜索、加工和分析的消费过程管理方案。

第三部分为第7～9章，介绍了数据治理的三项关键能力：数据的全量感知、综合质量提升、可控共享。第7章以数字孪生的全量、无接触感知为目标，介绍了数据的硬感知和软感知两类能力；第8章基于PDCA框架，介绍了对企业业务数据异常的全面监控，从而助力数据质量综合水平的提升；第9章介绍如何构建以元数据为基础的数据安全隐私保护框架，如何建立动静结合的数据保护与授权管理方案。

第四部分为第10章。基于对“机器认知世界”的理解，我们提出了对数据治理未来的思考，畅想了AI治理、数据主权和数据生态建设。未来已来，让我们共同努力，把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界。

读者对象

- 企业管理者：CEO，CIO，CDO，数字化转型项目的领导者、设计者和实施者。
- 数据从业人员：数据架构师、数据工程师、数据质量工程师、数据产品经理、数据分析师。
- IT从业人员：应用架构师、数据库专家、业务架构师。

致谢

本书主要作者有马运（Yun Ma）、杜浩、王强、陈实和周剑锋。参与写作的还有公司数据管理部的专家韦冬、廖华赟、赵子文、傅琨等。各位作者在数据领域耕耘和坚持多年。

本书的写作目的来自内外两方面：首先，公司内部希望我们对数据工作进行系统的梳理总结，积累企业数字化转型的经验；其次，华为云和华为中国地区部也希望我们能够写本书，以便和客户分享华为数据治理的实践，输出数据驱动企业数字

化转型的理念和方法。在此感谢陶景文、熊康、邓涛、洪方明、韩晓、苏立清等领导的建议、支持和指导。

本书的内容来自华为数据体系多年的工作实践，特别是最近几年在数字化转型方面的探索。公司数据治理的实践经验是在公司变革体系、质量运营体系、流程体系、IT和数据体系等各部门的共同努力下取得的。感谢郝健康、张印臣两位领导对数据体系建设的贡献！感谢数据体系的全体同人！感谢所有支持数据工作的同事！

感谢华为云殷宏、祝向党，华为大学陈小宇、蒋文俏以及胡小敏等人从读者视角对本书进行审视，并在编写过程中给予了我们很大的帮助。

数字化转型波澜壮阔，华为数据治理也历经十多年沉淀，成果丰富，本书呈现出来的内容实属挂一漏万，加上编写时间仓促，书中难免会出现一些失误或者不准确的地方，恳请读者批评指正。

华为公司 数据管理部

第1章

数据驱动的企业数字化转型

随着通信与数字技术的发展，网络化和数字化给人类带来更多的精彩和无限的可能，推动我们进入全联接的信息时代和大数据时代。因此，如何响应这个时代的变化是当前所有企业都需要考虑的问题。

在这样的时代背景下，数字化转型正在改变许多企业和行业的运作模式，无论是数字原生企业，还是非数字原生企业，都在积极探索数字化转型。社会经济大环境的变化、行业趋势的变化、竞争对手的压力、公司的战略优化、自身经营的改善等是企业数字化转型最主要的驱动力。

IDC（国际数据公司）预测，鉴于竞争对手和产业都在进行数字化转型，如果企业不能快速实现数字化转型，到2022年，它们逾三分之二的目标市场会消失。过去几年里，IT厂商和传统企业始终专注于数字化转型，它们利用第三平台技术（云计算、移动、大数据/分析、社交）重组企业架构，而物联网（IoT）、人工智能（AI）和增强与虚拟现实（AR/VR）等创新加速器更进一步推动了这一进程。随着数字覆盖面的扩大、智能技术的广泛普及、应用程序与服务开发的爆发式增长，企业不断释放出“倍增创新”能力，数字化转型已步入第二阶段。

在这个技术与商业日 新月异的环境中，企业竞相加强自己的数字化创新能力，以便在快速数字化的全球经济中提升竞争力，实现繁荣发展。

企业要想在这样的数字时代生存下来，要么是数字原生企业，要么数字化转型成功，成为重生后的数字企业。

1.1　非数字原生企业的数字化转型挑战

数字原生企业在设立之初就以数字世界为中心来构建，生成了以软件和数据平台为核心的数字世界入口，便捷地获取和存储了大量的数据，并开始尝试通过机器学习等人工智能技术分析这些数据，以便更好地理解用户需求，增强数字化创新能力。部分数字原生企业引领着云计算、大数据、人工智能技术的发展，推动了数字化时代的发展。在这些数字原生企业中，整个企业的战略愿景、业务需求、组织架构、人员技能、管理文化、思考方式都是围绕着数字世界展开的。

与数字原生企业不同，非数字原生企业在成立之时，基本都是以物理世界为中心来构建的。绝大部分企业在创建的时候，是围绕生产、流通、服务等具体的经济活动展开的，天然缺乏以软件和数据平台为核心的数字世界入口，这也就造成了非数字原生企业与数字原生企业之间的显著差异。所以在数字化转型过程中，非数字原生企业面临着更大的挑战。

华为公司作为典型的非数字原生企业，在数字化转型过程中面临着与大多数非数字原生企业相似的问题。

1.1.1　业态特征：产业链条长、多业态并存

非数字原生企业，特别是大中型生产企业，往往有较长的业务链路，从研发到销售全产业链覆盖。以传统的钢铁企业为

例（如图1-1所示），完整工艺包括采矿、选矿、烧结、炼铁、炼钢、热轧、冷轧、硅钢等，辅助生产工艺包括焦化、制氧、燃气、自备电 、动力等，在各个工艺流程中沉淀着大量的复杂数据。

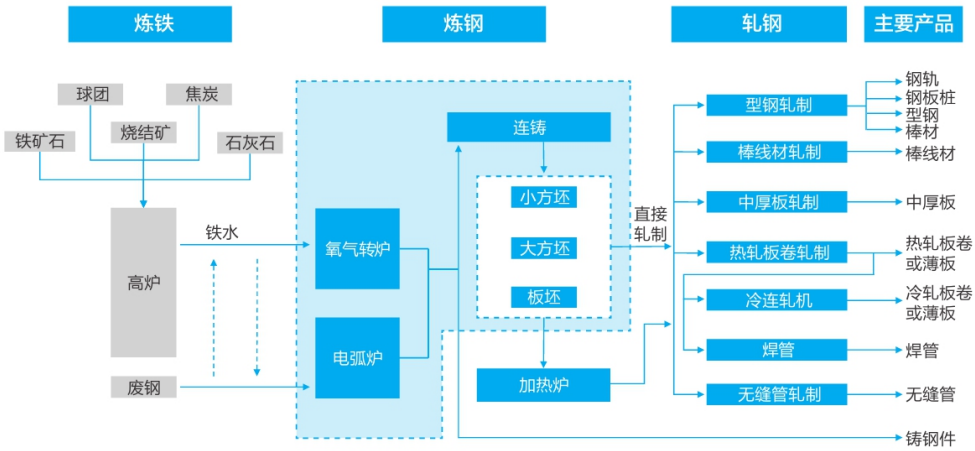


图1-1　钢铁企业工艺流程简 图

华为公司在构建面向客户价值流的过程中，同样形成了从研发到销售、供应、交付、运维的长链条，同时产品类型包括电 信基站、服务器、CPU、电 脑、手机、耳机等，横跨多个产业。这在某种程度上造成了各条块分割、业务组织强势、变革困 难、变革复杂度极高等问题。

1.1.2　运营环境：数据交互和共享风险高

非数字原生企业，特别是注重实物生产、交易的大中型企业，还面临着场景复杂的特点，比如交易复杂、风险周期长、内外部风险多等。生产过程中需要关注原材料供应、人工成本、物流过程；交易过程中涉及进出口的还需要关注外汇汇

率、当地政治环境、海关、法律法规、安全隐私、环境保护等多种信息；对于设备需要异地安装的情况，还需要考虑地理环境、道路环境、施工条件、运输条件、用工政策和安全防护等复杂因素。

华为公司的服务对象从运营商、企业客户到个人消费者，服务范围和雇员遍布全球100多个国家和地区，需要严格遵守各个国家和地区的进出口管制措施、环保条例、安全隐私法规等。这些业务形态上的特点，导致包括华为在内的诸多非数字原生企业对数据共享（特别是生产、销售侧数据的对外共享）有更多顾虑，更容易形成客观上的“数据孤岛”。

1.1.3 IT建设过程：数据复杂、历史包袱重

非数字原生企业普遍有较长的历史，组织架构和人员配置都围绕着线下业务开展，大都经历过信息化过程。很多制造型企业随着不同阶段的发展需求，保留着各个版本的ERP软件和各种不同类型的数据库存储环境，导致数据来源多样，独立封装和存储的数据难以集中共享，也不敢随意改造或替换，IT系统历史包袱沉重。Oracle ERP历年的版本信息如图1-2所示。

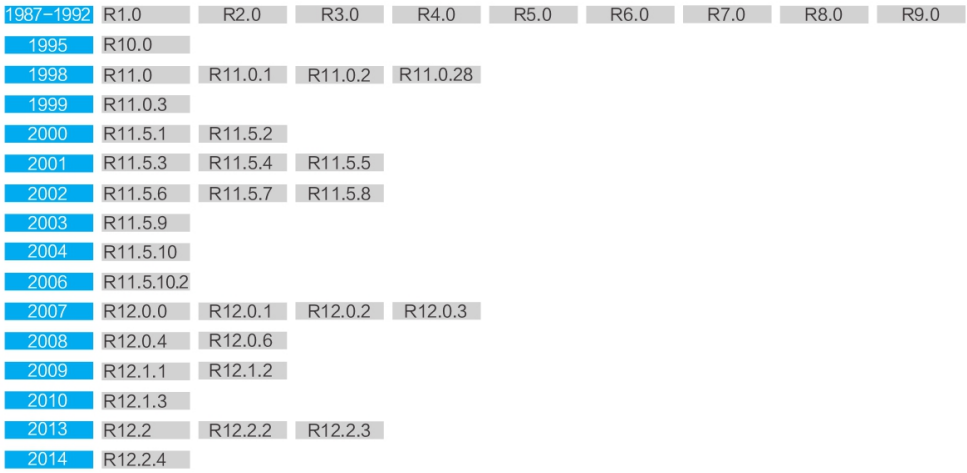


图1-2　Oracle ERP历年的版本信息（资料参考：Oracle）

目前，华为公司的主业务流程中存在几千个系统模块，有多版本的ERP、多种集成方式，系统间存在大量复杂的集成和嵌套。各业务领域开发了上千个应用系统模块，包含上百万张物理表、几千万个字段，这些数据又分别存储在上千个不同数据库中，共享困难；数据链路呈“长网”状，典型链路达12层以上，部分链路甚至高达22层。

1.1.4　数据质量：数据可信和一致化的要求程度高

基于业务特征和运营环境的特点，非数字原生企业对数据生成质量有更高的要求。数据产生时的质量高低不仅直接影响产品质量，而且直接影响整个内部业务的运作效率和成本。例如，华为公司会对合同录入质量进行严格度量和控制，以确保下游各环节能够及时、准确、完整地获得所需数据，并在整个端到端链条中对异常数据进行严格监控。数据质量要求严格，

需要配置多重精确规则，基于客观事实多重校验，确保数据可信、一致。

非数字原生企业在消费数据时对数据质量的要求也更高，一般会更聚焦于与业务流程相关的特定场景，更关注业务流程中问题的根因和偏差，数据挖掘、推理、人工智能都会聚焦于对业务的理解，面向业务去做定制化、精细化的算法管理，因此消费数据时的质量容错空间非常小。

上面所列出的非数字原生企业的特点，是我们基于华为的发展和对行业的认知所总结的，包括对非数字原生企业存在的问题和历史包袱等的表述，只是管中窥豹。联合国 工业体系分类中525门小工业体系的差异，足以说明非数字原生企业数字化转型的复杂性。精益管理技术下的不合格产品的“小数据”，让制造业AI难以基于这样的数据量训练出性能良 好的产品质检模型，同样说明非数字原生企业的数字化转型不可能是对数字原生企业的简单复刻。

1.2　华为数字化转型与数据治理

传统企业通过制造先进的机器来提升生产效率，但是未来，如何结构性地提升服务和运营效率，如何用更低的成本获取更好的产品，成了时代性的问题。数字化转型归根结底就是要解决企业的两大问题：成本和效率，并围绕“多打粮食，增加土地肥力”而开展。

1.2.1　华为数字化转型整体目标

2016年华为变革战略规划，明确要面向用户（企业客户、消费者、员工、合作伙伴、供应商）实现ROADS体验，持续提升效率、效益和客户满意度。明确要用五年时间完成业务数字化转型，数字化转型成为华为唯一的变革。

2017年华为提出了企业的新愿景：“把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界”。同时，华为公司董事、CIO陶景文提出了**“实现全联接的智能华为，成为行业标杆”的数字化转型目标**（如图1-3所示）。



图1-3　数字化转型目 标

对内，各业务领域数字化、服务化，打通跨领域的信息断点，达到领先于行业的运营效率。逐步构建以“面向客户做生意”和“基于市场的创新”两个业务流为核心的“端到端”的数字化管理体系。管理方式从定性走向定量，实现数据驱动的高效运作。

对外，对准5类用户的ROADS体验，实现与客户做生意更简单、更高效、更安全，提升客户满意度。华为首先从用户体验的视角表达了对行业的最新判断，并将其总结为ROADS，即实时（Real-time）、按需（On-demand）、全在线（All-online）、服务自助（DIY）和社交化（Social）。

1.2.2　华为数字化转型蓝图及对数据治理的要求

2017年，华为基于愿景确定了数字化转型的蓝图和框架，统一规划、分层次开展，最终实现客户交互方式的转变，实现内部运营效率和效益的提升。华为数字化转型蓝图包括5项举措（如图1-4所示）。

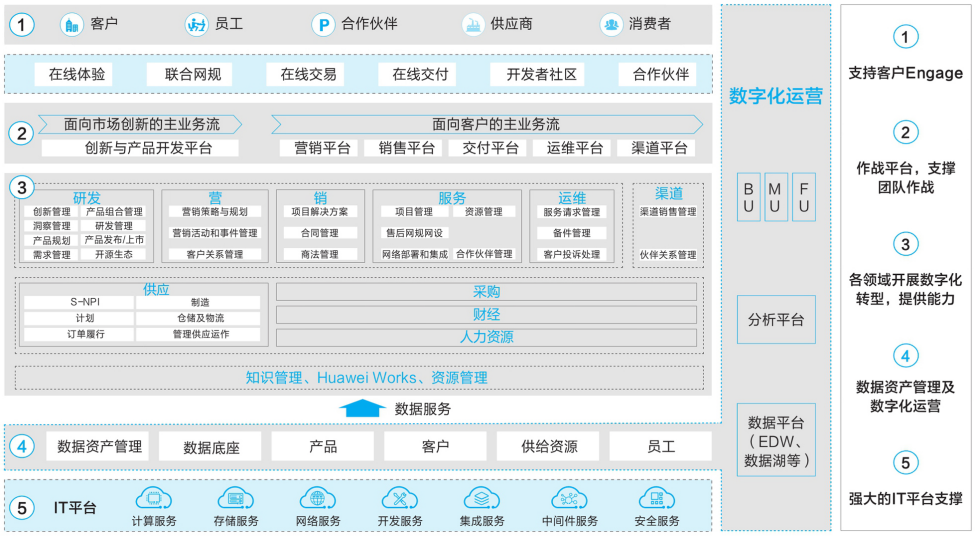


图1-4　华为数字化转型蓝图

举措1：实现“客户交互方式”的转变，用数字化手段做厚、做深客户界面，实现与客户做生意更简单、更高效、更安全，提升客户体验满意度，帮助客户解决问题。

举措2：实现“作战模式”的转变，围绕两大主业务流，以项目为中心，对准一线精兵团队作战，率先实现基于ROADS的体验，达到领先于行业的运营效率。

举措3：实现“平台能力”提供方式的转变，实现关键业务对象的数字化并不断汇聚数据，实现流程数字化和能力服务化，支撑一线作战人员和客户的全联接。

举措4：实现“运营模式”的转变，基于统一数据底座，实现数字化运营与决策，简化管理，加大对一线人员的授权。

举措5：云化、服务化的IT基础设施和IT应用，统一公司IT平台，同时构建智能服务。

其中，举措4涉及数据治理和数字化运营，是华为数字化转型的关键，承接了打破数据孤岛、确保源头数据准确、促进数据共享、保障数据隐私与安全等目标。华为数字化转型对数据治理的要求如下：

1）基于统一的数据管理规则，确保数据源头质量以及数据入湖，形成清洁、完整、一致的数据湖，这是华为数字化转型的基础。

2）业务与数据双驱动，加强数据联接建设，并能够以数据服务方式，灵活满足业务自助式的数据消费诉求。

3）针对汇聚的海量内外部数据，能够确保数据安全合规。

4）不断完善业务对象、过程与规则数字化，提升数据自动采集能力，减少人工录入。

1.3　华为数据治理实践

华为从2007年开始启动数据治理，历经两个阶段的持续变革，系统地建立了华为数据管理体系。第一阶段近十年的持续投入为华为在2017年开始的数字化转型打下了坚实的基础。同时，在数字化转型对数据治理的新要求下，正式进入第二阶段，数据治理工作也迎来了新的挑战和发展。

1.3.1　华为数据治理历程

1. 第一阶段：2007～2016年

在这一阶段，华为设立数据管理专业组织，建立数据管理框架，发布数据管理政策，任命数据Owner，通过统一信息架构与标准、唯一可信的数据源、有效的数据质量度量改进机制，实现了以下目标。

1）**持续提升数据质量，减少纠错成本：**通过数据质量度量与持续改进，确保数据真实反映业务，降低运营风险。

2）**数据全流程贯通，提升业务运作效率：**通过业务数字化、标准化，借助IT技术，实现业务上下游信息快速传递、共享。

2. 第二阶段：2017年至今

在这一阶段，华为建设数据底座，汇聚企业全域数据并对数据进行联接，通过数据服务、数据地图、数据安全防护与隐私保护，实现了数据随需共享、敏捷自助、安全透明的目标，支撑着华为数字化转型，实现了如下的数据价值。

1）**业务可视，能够快速、准确决策：**通过数据汇聚，实现业务状态透明可视，提供基于“事实”的决策支持依据。

2）**人工智能，实现业务自动化：**通过业务规则数字化、算法化，嵌入业务流，逐步替代人工判断。

3）**数据创新，成为差异化竞争优势：**基于数据的用户洞察，发现新的市场机会点。

华为数据治理的发展历程如图1-5所示。

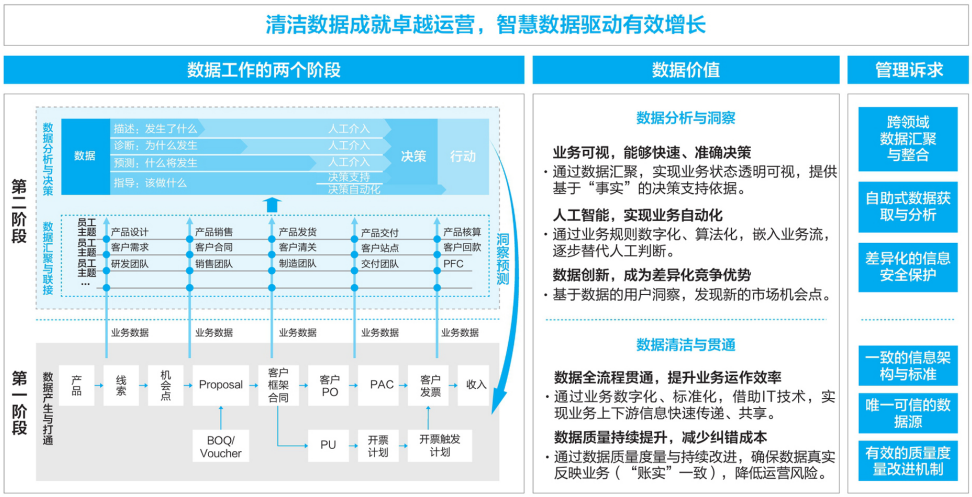


图1-5　华为数据治理的两个阶段

1.3.2　华为数据工作的愿景与目标

华为公司基于多业务、全球化、分布式管理等业务战略规划和数字化转型诉求，明确了华为数据工作的愿景，即“实现业务感知、互联、智能和ROADS体验，支撑华为数字化转型”。华为数据工作的目标为“清洁、透明、智慧数据，使能卓越运营和有效增长”。为确保数据工作的愿景与目标达成，需要实现数据自动采集、对象/规则/过程数字化、数据清洁、安全共享等特性（如图1-6所示）。

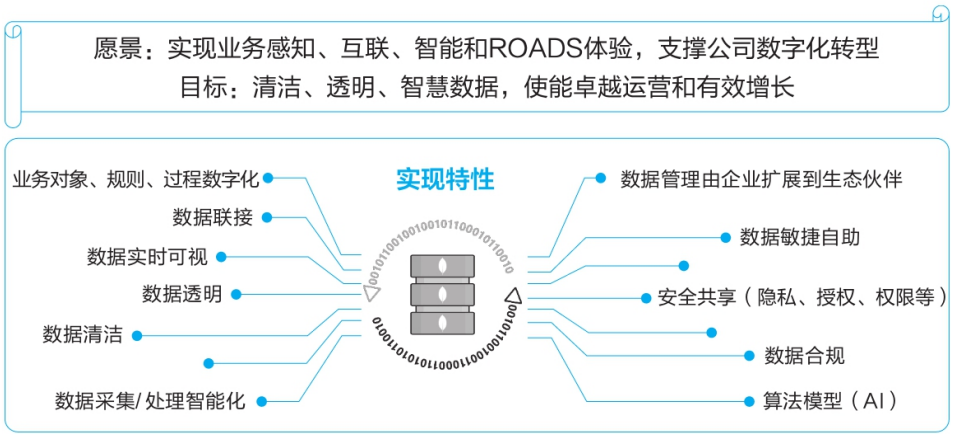


图1-6　华为数据治理的愿景与目标

1.3.3　华为数据工作建设的整体思路和框架

作为非数字原生企业，我们认为数字化转型的关键要素之一是在现实世界的基础上构建一个跨越孤立系统、承载业务的“数字孪生”的数字世界。通过在数字世界汇聚、联接与分析数据，进行描述、诊断和预测，最终指导业务改进。在实现策略上，数字世界一方面要充分利用现有IT系统的存量数据资产，另一方面要构建一条从现实世界直接感知、采集、汇聚数

据到数字世界的通道，不断驱动业务对象、过程与规则的数字化。华为数据工作建设的整体思路如图1-7所示。

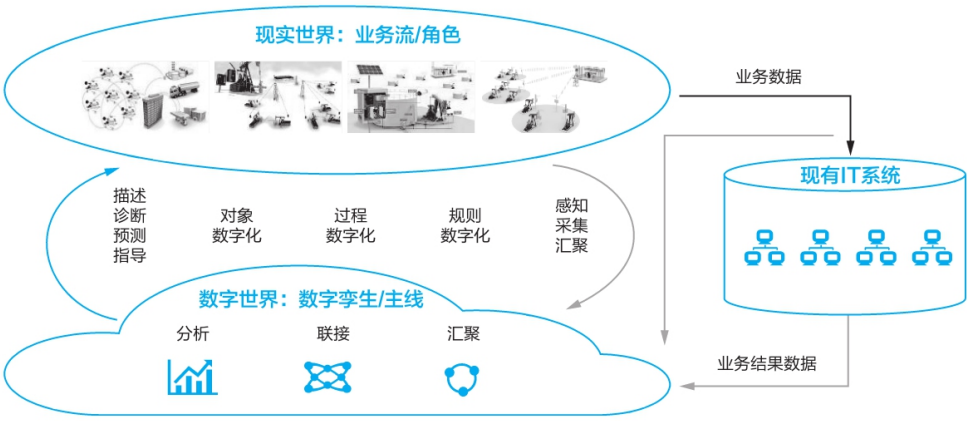


图1-7　华为数据工作建设的整体思路

华为经过多年实践，形成了一套数据工作框架。

1）**数据源：**业务数字化是数据工作的前提，通过业务对象、规则与过程数字化，不断提升数据质量，建立清洁、可靠的数据源。

2）**数据湖：**基于“统筹推动、以用促建”的建设策略，严格按六项标准，通过物理与虚拟两种入湖方式，汇聚华为内部和外部的海量数据，形成清洁、完整、一致的数据湖。

3）**数据主题联接：**通过五种数据联接方式，规划和需求双驱动，建立数据主题联接，并通过服务支撑数据消费。

4）**数据消费：**对准数据消费场景，通过提供统一的数据分析平台，满足自助式数据消费需求。

5）**数据治理：**为保障各业务领域数据工作的有序开展，需建立统一的数据治理能力，如数据体系、数据分类、数据感知、数据质量、安全与隐私等。

数据体系建设的整体框架（如图1-8所示），基于统一的规则与平台，以业务数字化为前提，数据入湖为基础，通过数据主题联接并提供服务，支撑业务数字化运营。



图1-8　华为数据工作建设的整体框架

1.4　本章小结

本书是华为数据治理方法论和数据建设实践经验的总结，本章从宏观上介绍了华为在数字化转型和数据治理方面面临的挑战和整体策略。后续内容将从数据体系建设和数据分类开始，围绕信息架构、数据底座、数据服务建设，结合数据感知、数据质量和安全合规能力打造，详细阐述华为在数据治理和数字化转型方面的经验。

第2章

建立企业级数据综合治理体系

数据作为一种新的生产要素，在企业构筑竞争优势的过程中起着重要作用，企业应将数据作为一种战略资产进行管理。数据从业务中产生，在IT系统中承载，要对数据进行有效治理，需要业务充分参与，IT系统确保遵从，这是一个非常复杂的系统工程。

华为公司经过十多年的实践证明，只有构筑一套企业级的数据综合治理体系，才能确保关键数据资产有清晰的业务管理责任，IT建设有稳定的原则和依据，作业人员有规范的流程和指导；当面临争议时，有裁决机构和升级处理机制；治理过程所需的人才、组织、预算有充足的保障。综合上述因素，最终建立有效的数据治理环境，数据的质量和安全得到保障，数据的价值才能真正发挥出来。

华为的数据治理体系框架如图2-1所示。

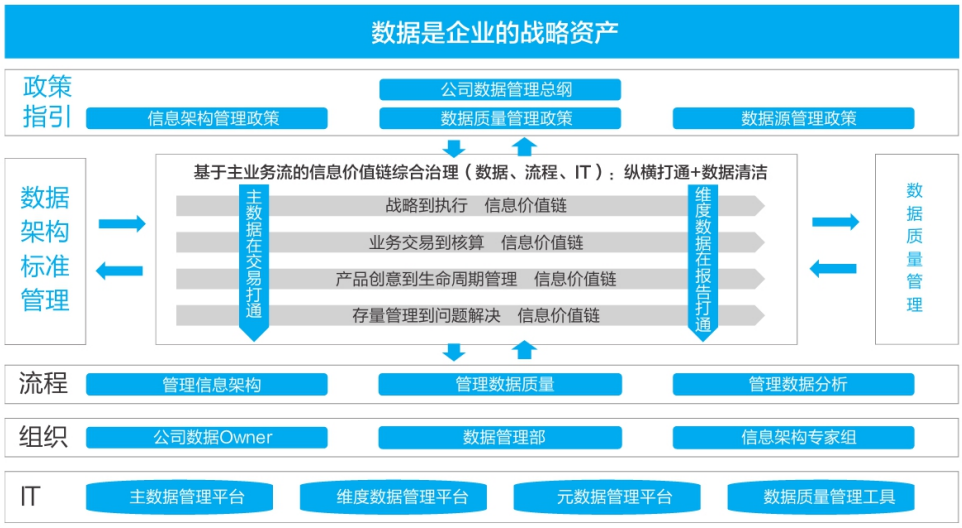


图2-1 华为数据治理体系框架

2.1　建立公司级的数据治理政策

数据治理政策是华为数据治理的顶层设计，该政策在华为公司EMT（经营管理团队）汇报通过后，由总裁签发，该政策明确了数据工作在华为公司治理体系中的地位，体现了公司管理层对数据工作重要性的统一认知。

2.1.1　华为数据管理总纲

华为数据管理总纲明确了数据治理最基本的原则，包括信息架构、数据产生、数据应用及数据质量的职责和分工等，确保数据治理环境的有效构建。

（1）信息架构管理原则

第一条：建立企业级信息架构，统一数据语言。

第二条：所有变革项目须遵从数据管控要求。对于不遵从管控要求的变革项目，数据管控组织拥有一票否决权。

第三条：应用系统设计和开发应遵从企业级信息架构。关键应用系统必须通过应用系统认证。

（2）数据产生管理原则

第一条：数据规划对齐业务战略，业务战略规划必须包含关键数据举措及其路标规划。

第二条：公司数据Owner拥有公司数据管理的最高决策权，依托ESC（变革指导委员会）决策平台议事。各数据Owner承担数据工作路标、信息架构、数据责任机制和数据质量的管理责任。

第三条：关键数据须定义单一数据源，一点录入，多点调用。数据质量问题应在源头解决。

第四条：谁产生数据，谁对数据质量负责。数据Owner负责基于使用要求制定数据质量标准，且须征得关键使用部门的同意。

（3）数据应用管理原则

第一条：数据应在满足信息安全的前提下充分共享，数据产生部门不得拒绝跨领域的、合理的数据共享需求。

第二条：信息披露、数据安全管理、数据保管和个人数据隐私保护等必须遵守法律法规和道德规范的要求。公司保护员工、客户、商业伙伴和其他可识别个体的数据。

（4）数据问责与奖惩管理原则

各数据Owner应建立数据问题回溯和奖惩机制。对不遵从信息架构或存在严重数据质量问题的责任人进行问责。

2.1.2 信息架构管理政策

信息架构是公司统一的数据语言，是业务流打通、消除信息孤岛和提升业务流集成效率的关键要素。华为公司通过明确对信息架构的管理要求，规范信息架构的建设和遵从原则，使公司的信息资产得到有效管理和重用。

（1）管理信息架构的角色与职责

第一条：公司数据Owner负责批准企业级信息架构，裁决重大信息架构问题和争议。

第二条：各数据Owner负责其所辖数据的信息架构建设和维护，承接及落实公司的数据规划要求。

第三条：公司的数据管理专业组织作为公司数据工作的支撑组织，负责组织信息架构的建设、维护、落地及遵从管控，负责协调跨领域的信息架构冲突。各领域各事业群（BG）数据管理专业组织协助完成本领域信息架构建设和维护工作。

第四条：数据管控组织作为信息架构专业评审机构，确保信息架构的质量和集成。

（2）信息架构建设要求

第一条：关键数据应被识别、分类、定义及标准化，数据的定义在公司范围内应唯一，数据标准制定要考虑跨流程要

求。

第二条：数据资产目录必须承接公司各业务环节的使用需求和报告分析最小粒度的要求。

第三条：信息架构驱动应用架构设计，合理规划数据分布。

第四 条：应用系统数据库的设计和开发要遵循信息架构，减少数据冗余，实现接口标准化。

（3）信息架构遵从管控

第一条：变革项目必须遵从已发布的信息架构，变革项目的交付件须包含信息架构内容。对现有架构的遵从是关键评审要素，对于不满足要求的变革项目，数据管控组织拥有一票否决权。

第二条：业务流程设计必须遵从已发布的信息架构，在流程说明文件、操作指导书或模板类文件中体现。对于不满足要求的流程，不予发布。

第三条：应用系统设计必须遵从已发布的信息架构。在应用架构交付件和应用系统设计交付件中体现。对于不满足要求的应用系统，不予上线。

2.1.3　数据源管理政策

数据同源是华为数据治理的核心观点之一。数据源是指业务上首次正式发布某项数据的应用系统，经过数据管理专业组织认证，作为唯一数据源头被周边系统调用。本政策通过明确华为公司在数据源建设和数据源使用方面的总体原则和要求，确保数据源头的统一，以及跨流程、跨系统数据的唯一性和一致性。

（1）数据源管理原则

第一条：所有关键数据必须认证数据源。关键数据是指影响公司经营、运营报告的数据，在公司范围内统一发布。

第二条：数据管理专业组织为关键数据指定源头，数据源必须遵从信息架构和标准，经信息架构专家委员会认证后成为数据源。

第三条：所有关键数据仅能在数据源录入、修改，全流程共享，其他调用系统不能修改。下游环节发现的数据源质量问题，应当在数据源进行修正。

第四 条：所有应用系统必须从数据源或数据源镜像获取关键数据。

第五条：数据Owner确保数据源的数据质量，对不符合数据质量标准的数据源，必须限 期整改。

（2）数据源认证标准

数据的源头通过认证成为数据源，在遵从公司相关政策和规定的前提下，还必须符合以下标准。

第一条：数据源是在信息链上正式发布数据的第一个数据存储系统。

第二条：数据源是某项数据唯一的录入点。

第三条：数据源必须是数据维护最为及时、正确、完整的数据存储系统。

第四条：数据源所在系统的性能和可用性应当满足其他调用系统的数据访问需求。

2.1.4 数据质量管理政策

数据质量的持续提升是华为数据治理的核心目标。通过制定数据质量管理政策，明确数据在创建、维护、应用过程中的规则及质量要求，确保数据真实可靠。

（1）数据质量管理职责及要求

第一条：各数据Owner负责保障所辖数据的质量，承接公司数据Owner设定的数据质量目标，制定数据质量标准及测评指标，持续度量与改进。

第二条：公司全员在业务执行的过程中应确保业务记录满足数据质量要求。

第三条：财经各级CFO组织应遵循职业道德准则，诚实记录和报告财经数据，承担财务监控和及时报告责任。

第四条：公司各级数据管理专业组织为数据Owner提供数据质量管理专业支撑。

第五条：内控组织应将数据质量管控要素的执行情况纳入SACA（Semi-Annual Control Assessment，半年度控制评估）评估范围，推动数据质量问题的闭环管理。

第六条：内审部门作为独立机构，负责重大数据问题的审计和责任回溯。

（2）数据质量管理的业务规则和管理要求

数据创建、维护、应用是数据生命周期管理的关键活动，应遵循以下规则及要求。

第一条：流程建设应考虑数据质量要求，将数据的关键质量控制要素纳入关键控制点。

第二条：数据Owner负责基于使用要求制定数据质量标准，且须征得关键使用部门的同意。

第三条：数据创建应确保录入正确，关键数据应进行复核或审批。录入、复核和审批人员应掌握数据质量要求才能上岗。

第四条：对影响关键经营指标的数据造假行为（如伪造文档、提供与业务实质不符的信息等）采取零容忍态度。

第五条：上游环节应保证数据的真实、完整并及时传递到下游环节。下游环节为核实数据质量问题可调阅所需的上游环节的数据。

第六条：因外部原因频繁变化的基础数据（如汇率、税率等），数据Owner应及时维护并统一发布最新数据，各环节应适时刷新或引用。

第七条：数据质量应持续进行度量。数据Owner应主动解决长期影响业务运营和经营管理的数据问题。

第八条：报告与分析的层级和最小粒度应适度，能与最小业务信息单元相匹配。数据加工规则应相对稳定，报告加工过程可检视，数据可回溯、可解释。

2.2 融入变革、运营与IT的数据治理

华为公司依托变革管理体系，进行流程、数据与应用系统建设，同时持续优化运营体系。数据从业务中产生，在IT系统中落地，决定了数据治理工作必须充分融入业务运营与IT系统建设中。

2.2.1 建立管理数据流程

为了支撑企业数据资产从架构设计、质量管理到数据分析应用的全生命周期管理，需要在企业的流程架构中建立一个管理数据流程，明确数据管理的关键活动、角色，以及与周边组织的协作关系。华为将“管理数据”流程定位为“管理BT&IT”流程下的一个L2流程，下设“管理信息架构”“管理数据质量”“管理数据分析”3个子流程，如图2-2所示。

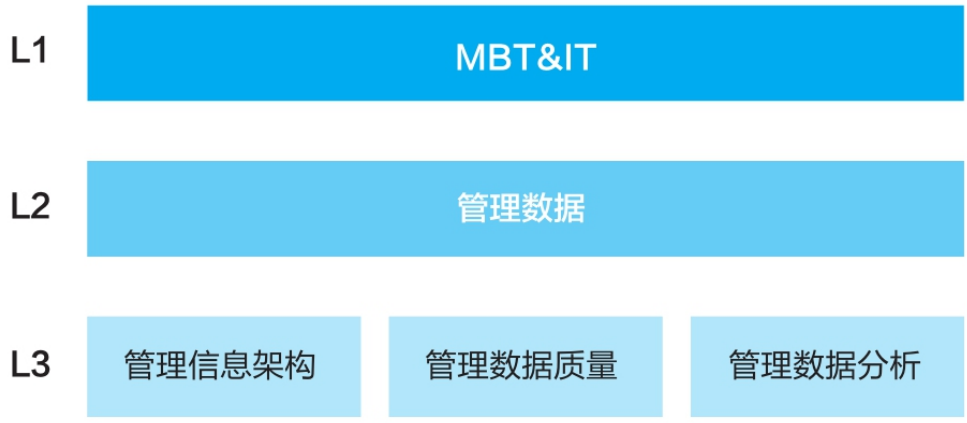


图2-2 华为管理数据流程

管理数据流程关键角色及职责设置如表2-1所示。

表2-1　管理数据流程关键角色及职责设置

序　号	角色名称	职责描述
1	信息架构工程师	数据架构设计和管控 数据分类、定义及标准化 企业级信息架构、业务侧概念模型开发和维护 数据标准开发 数据源认证 数据流、信息链开发
2	数据治理工程师	聚焦数据资产建设和治理 数据治理和数据质量监控 识别和定位数据质量问题，实施根因分析 组织制定数据质量标准和数据质量监控计划 定义和制定数据质量评测指标 实施测评和报告
3	数据平台工程师	数据分析平台规划和运营 数据采集和预处理
4	数据分析师	聚焦价值实现 数据分析和挖掘 业务数据模型开发 数据分析报告拟定 数据可视化设计
5	数据科学家	聚焦技术研究和攻关 基础数据模型和算法的开发 数据产品设计 数据分析问题攻关

2.2.2　管理数据流程与管理变革项目、管理质量与运营之间的关系

企业在运营过程中，能力的提升和架构的调整依托于变革项目和改进项目的实施。变革项目和改进项目需要交付业务解决方案、数据解决方案、IT解决方案，其中数据解决方案包含信息架构设计、数据质量度量、改进方案和数据分析方案。支撑数据解决方案的角色为数据经理，数据经理统筹管理信息架构工程师、数据治理工程师、数据分析师和数据科学家，共同完成项目数据解决方案的交付和验证。具体的管理数据流程与管理变革项目、管理质量与运营之间的关系如图2-3所示。

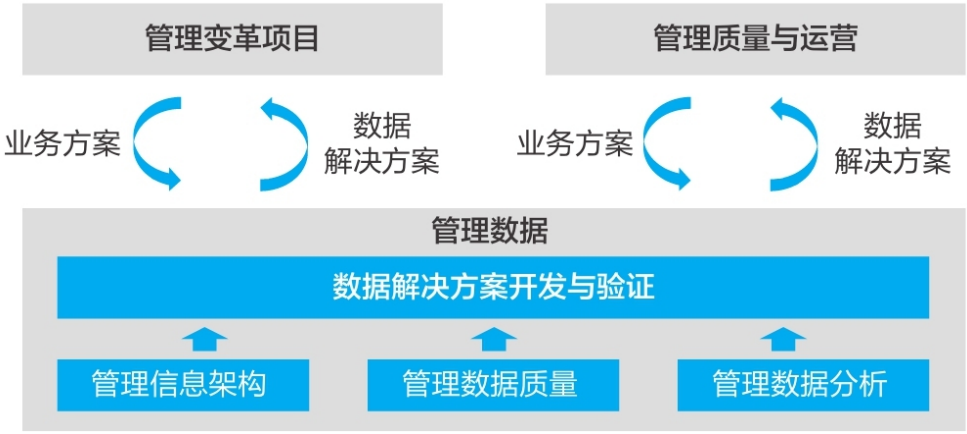


图2-3　华为管理数据流程与管理变革项目、管理质量运营之间的关系

2.2.3　通过变革体系和运营体系进行决策

在华为的数据治理实践中，数据相关的重大决议由企业变革指导委员会决策，通过变革管理体系和流程运营体系实现落地，如图2-4所示。

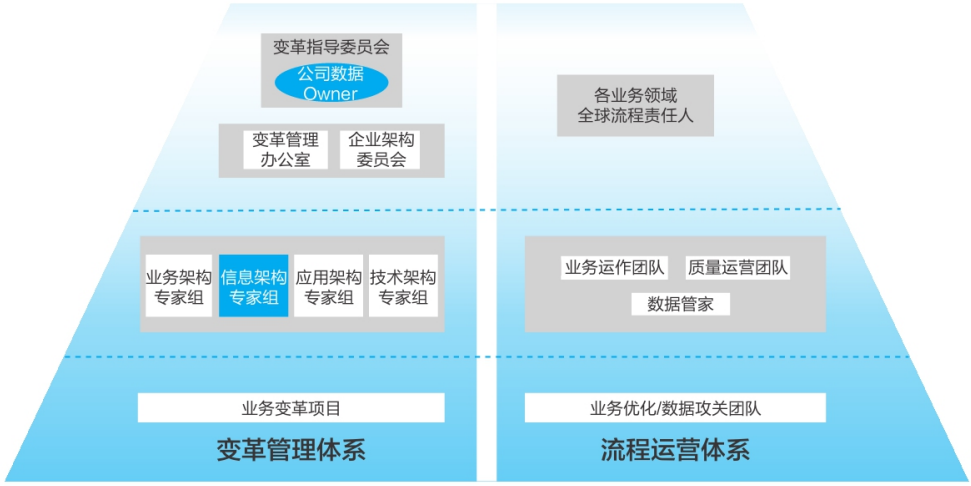


图2-4　华为数据治理决策体系

其中，信息架构的设计和变更分两层评审决策，在信息架构专家组进行专业评审，在企业架构委员会进行流程、数据与IT集成评审和争议裁决。

2.2.4　数据治理融入IT实施

业务人员通过使用IT产品提供的功能和服务提升作业效率，因此，对业务数据的管理要求，必然要落实到IT产品的操作界面和数据库设计中，这样才能落实数据治理的要求。在华为的数据治理实践中，在IT产品团队中设置系统架构师和数据架构师角色，负责界面设计、数据库设计、数据集成方案设计，向上承接信息架构的设计要求。同时，在管理IT流程的设计规范中，明确界面的字段要遵从数据标准的定义，数据库表和字段的设计要承接信息架构的设计要求，从而达到数据治理融入IT实施流程的目标。

2.2.5 通过内控体系赋能数据治理

要对华为这样的大型企业实施数据治理是件非常复杂的事情，涉及上千个业务对象、上百个变革和优化改进项目的协同，仅仅通过数据管理部门对各个项目和部门的培训、指导、人员支持，不足以确保公司的治理目标和要求有效地贯彻到位。因此，华为通过内控体系，每年实施SACA评估和数据专项内部审计，揭示数据治理过程的问题，确定改进目标和责任人，从而保证数据治理机制的有效运作。

2.3　建立业务负责制的数据管理责任体系

业务即行为，行为即记录，记录即数据。华为公司的每一个数据，必须由对应的业务部门承担管理责任，且必须有唯一的数据Owner。**业务负责制的数据管理责任体系，是华为数据治理体系多年实践经验的结晶，是确保体系发挥作用的基石。**

2.3.1　任命数据Owner和数据管家

华为按分层分级原则任命数据Owner，在公司层面设置公司数据Owner，在各业务领域设置领域数据Owner，这样既能确保公司数据工作统筹规划，也能同时兼顾各业务领域灵活多变的特征。

公司数据Owner是公司数据战略的制定者、数据文化的营造者、数据资产的所有者和数据争议的裁决者，拥有公司数据日常管理的最高决策权，职责如下所示。

第一条：制定数据管理体系的愿景和路标。

第二条：传播数据管理理念，营造数据文化氛围。

第三条：建设和优化数据管理体系，包括组织与任命、授权与问责等。

第四条：批准公司数据管理的政策和法规。

第五条：裁决跨领域的数据及管理争议，解决跨领域的重大数据及管理问题。

各级流程Owner就是该流程域的数据Owner，在公司数据Owner的统筹下负责所管理流程域的数据管理体系的建设和优化。各业务部门是执行规则，保证数据质量，进而推动规则优化的关键环节。通过主管机构正式任命各数据主题域和业务对象的数据Owner和数据管家，数据Owner的职责可以归纳为以下五条。

第一条：负责数据管理体系建设。数据Owner要负责所辖领域的数据管理体系建设和优化，传播数据管理理念，营造数据文化氛围。

第二条：负责信息架构建设。数据Owner要负责所辖领域的信息架构建设和维护，确保关键数据被识别、分类、定义及标准化，数据的定义在公司范围内唯一，数据标准制定要考虑跨流程要求。

第三条：负责数据质量管理。数据Owner要负责保障所辖领域的数据质量，承接公司设定的数据质量目标，制定数据质量标准及测评指标，持续度量与改进。

第四条：负责数据底座和数据服务建设。数据Owner要负责所辖领域数据入湖，建设数据服务，满足公司各个部门对本领域数据的需求。

第五条：负责数据争议裁决。数据Owner要建立数据问题回溯和奖惩机制，对所辖领域的数据问题及争议进行裁决，对不遵从信息架构或存在严重数据质量问题的责任人进行问责。

数据管家是数据Owner的助手，是数据Owner在数据管理方面的具体执行者。

2.3.2 建立公司层面的数据管理组织

华为数据管理组织如图2-5所示。为支撑公司实施数据治理，华为在企业范围内建立了一个公司级数据管理部，代表公司制定数据管理相关的政策、流程、方法和支撑系统，制定公司数据管理的战略规划和年度计划并监控落实。建立并维护企业信息架构，监控数据质量，披露重大数据问题，建立专业任职资格管理体系，提升企业数据管理能力，推动企业数据文化的建立和传播。

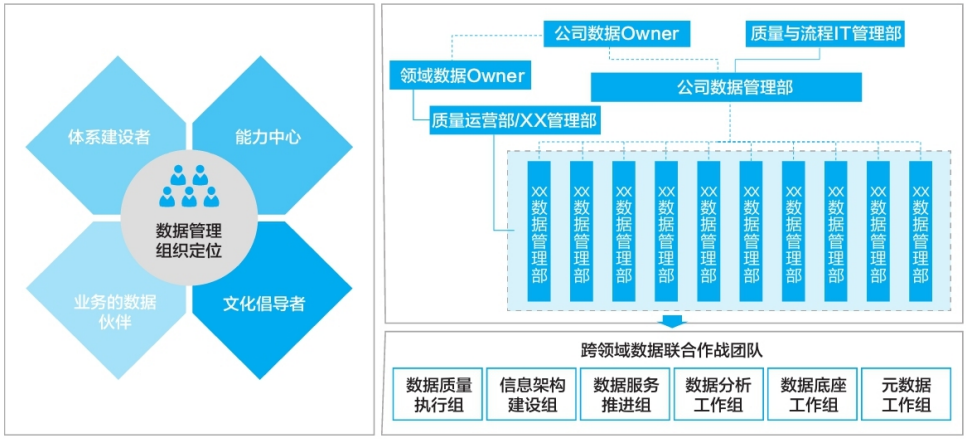


图2-5 华为数据管理组织

为落实公司制定的数据管理目标，在各业务领域要建立实体化的数据管理专业组织，实线向GPO（各业务领域的全球流程Owner，通常是业务领域的最高主管）汇报，承接并落实GPO的数据管理责任；虚线向公司数据管理部汇报，遵从公司统一的数据管理政策、流程和规则要求。

华为虚实结合的数据组织设置，是确保数据工作能充分融入业务，同时能够在应用系统中有效落地的关键。

数据管理组织中各个组织的职责和分工如下所示。

1）体系建设者

第一条：负责数据管理的战略、规划、政策、规则的制定。

第二条：负责数据管理体系建设。

第三条：数据架构及核心数据资产管理。

第四 条：确保公司数据质量水平。

2）能力中心

第一条：构建数据管理的方法、工具、平台。

第二条：负责专业能力的开发和建设，包括数据架构、数据分析、信息管理、数据质量管理。

3）业务的数据伙伴

第一条：面向业务，提供数据解决方案，解决业务数据痛点。

第二条：支撑业务数据需求。

第三条：向业务提供标准化的主数据或基础数据服务。

4）文化倡导者

第一条：在公司范围建设追求卓越、“谁创建（录入）数据，谁对数据质量负责”的文化。

第二条：用数据支撑业务决策的文化。

同时，在数据工作的不同阶段，分场景组建了不同的虚拟数据团队，如信息架构建设工作组、数据质量执行组、元数据工作组等，以保障跨领域数据工作的有序开展。

当面对数字化转型这一时代挑战时，华为建立的一整套数据治理体系，使得华为公司拥有从容面对的底气。2017年华为启动数字化转型后，也极大提升了华为的数据治理能力，在实践中形成了数据全生命周期的治理规范与方案，如图2-6所示。

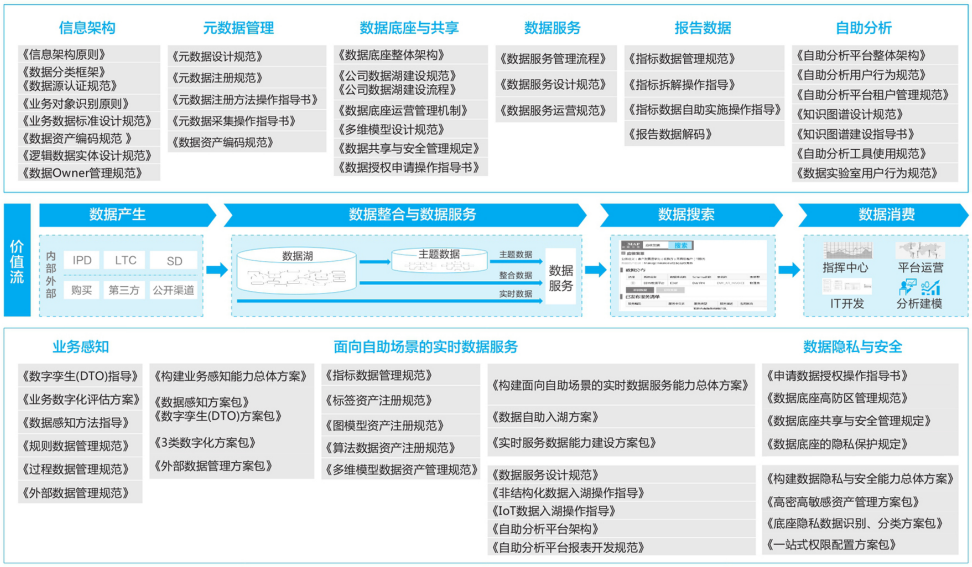


图2-6　华为数据全生命周期治理规范与方案

2.4　本章小结

华为自2007年建立数据实体组织以来，走过了13年的数据治理历程。从最初的数据管理体系搭建，到目前主干业务流全场景覆盖，确保了各个业务流程产生的数据准确。华为数据治理一方面实现了业务运作效率的提升，一方面夯实了企业有效内控的基础，数据文化与价值深入人心。

第3章

差异化的企业数据分类管理框架

不同的企业或组织基于不同的目的，可以从多个角度对数据进行分类，如结构化数据和非结构化数据、内部数据和外部数据、原始数据和衍生数据、明细数据和汇总数据等。华为在业界的数据分类基础上，结合自身多年的实践，已形成完整的数据分类管理框架。华为对数据进行分类的目的，是为了针对不同特性的数据采取不同的管理策略，以期实现最大的投入产出比。

3.1　基于数据特性的分类管理框架

华为根据数据特性及治理方法的不同对数据进行了分类定义：内部数据和外部数据、结构化数据和非结构化数据、元数据。其中，结构化数据又进一步划分为基础数据、主数据、事务数据、报告数据、观测数据和规则数据。华为数据分类管理框架如图3-1所示。

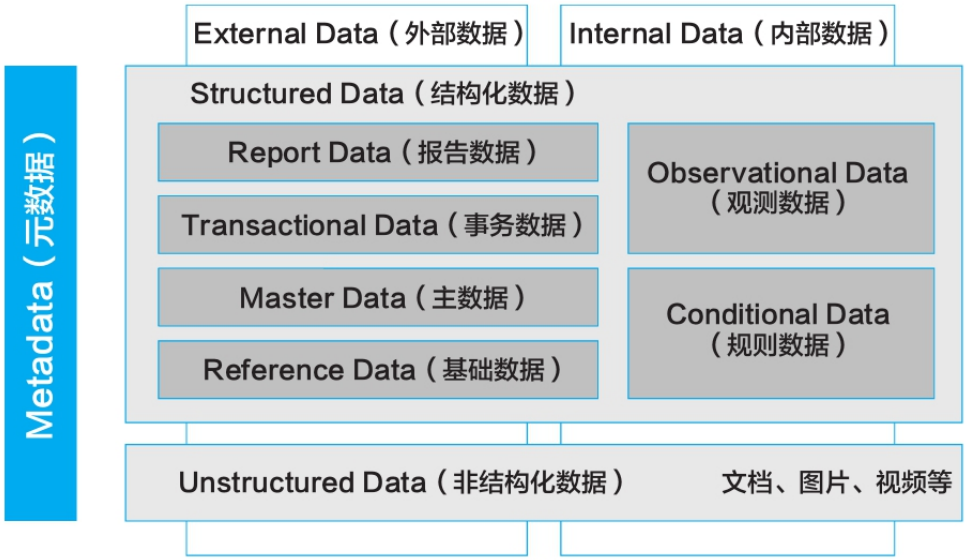


图3-1　华为数据分类管理框架

对上述数据分类的定义及特征描述见表3-1。

表3-1　数据分类定义及特征描述

分类维度	数据分类名称	定　义	特　征	举　例
按数据主权所属华为外部 / 内部分类	External Data（外部数据）	华为通过公共领域获取的数据	客观存在，其产生、修改不受我司的影响	国家、币种、汇率
	Internal Data（内部数据）	企业内经营产生的数据	在企业的业务流程中产生或在业务管理规定中定义，受企业经营影响	合同、项目、组织
从数据存储特性为结构化或者非结构化分类	Structured Data（结构化数据）	可以存储在关系数据库里，用二维表结构来表达实现的数据	1）可以用关系数据库存储 2）先有数据结构，再产生数据	国家、币种、组织、产品、客户
	Unstructured Data（非结构化数据）	形式相对不固定，不方便用数据库二维逻辑表来表现的数据	1）形式多样，无法用关系数据库存储 2）数据量通常较大	网页、图片、视频、音频、XML
	Reference Data（基础数据）	用结构化的语言描述属性，用于分类或目录整编的数据，也称作参考数据	1）通常有一个有限的允许 / 可选值范围 2）静态数据，非常稳定，可以用作业务 / IT 的开关、职责 / 权限的划分或统计报告的维度	合同类型、职位、国家、币种
	Master Data（主数据）	具有高业务价值的、可以在企业内跨流程跨系统被重复使用的数据，具有唯一、准确、权威的数据源	1）通常是业务事件的参与方，可以在企业内跨流程、跨系统重复调用 2）取值不受限于预先定义的数据范围 3）在业务事件发生之前就客观存在，比较稳定 4）主数据的补充描述可归入主数据范畴	实体型组织、客户、人员基础配置

（续）

分类维度	数据分类名称	定　义	特　征	举　例
从数据存储特性为结构化或者非结构化分类	Transactional Data（事务数据）	用于记录企业经营过程中产生的业务事件，其实质是主数据之间活动产生的数据	1）有较强的时效性，通常是一次性的 2）事务数据无法脱离主数据独立存在	BOQ、支付指令、主生产计划
	Observational Data（观测数据）	观测者通过观测工具获取观测对象行为 / 过程的记录数据	1）通常数据量较大 2）数据是过程性的，主要用作监控分析 3）可以由机器自动采集	系统日志、物联网数据、运输过程中产生的 GPS 数据
	Conditional Data（规则数据）	结构化描述业务规则变量（一般为决策表、关联关系表、评分卡等形式）的数据，是实现业务规则的核心数据	1）规则数据不可实例化，只以逻辑实体形式存在 2）规则数据的结构在纵向和横向两个维度上相对稳定，变化形式多为内容刷新 3）规则数据的变更对业务活动的影响是大范围的	员工报销遵从性评分规则、出差补助规则
	Report Data（报告数据）	是指对数据进行处理加工后，用作业务决策依据的数据	1）通常需要对数据进行加工处理 2）通常需要将不同来源的数据进行清洗、转换、整合，以便更好地进行分析 3）维度、指标值都可归入报告数据	收入、成本
从描述数据的手段上分类	Meta-data（元数据）	定义数据的数据，是有关一个企业所使用的物理数据、技术和业务流程、数据规则和约束以及数据的物理与逻辑结构的信息	是描述性标签，描述了数据（如数据库、数据元素、数据模型）、相关概念（如业务流程、应用系统、软件代码、技术架构）以及它们之间的联系（关系）	数据标准、业务术语、指标定义

不同分类的数据，其治理方法有所不同。如基础数据内容的变更通常会对现有流程、IT系统产生影响，因此基础数据的

管理重点在于变更管理和统一标准管控。主数据的错误可能会导致成百上千的事务数据错误，因此主数据的管理重点是确保同源多用、重点进行数据内容的校验等。

3.2　以统一语言为核心的结构化数据管理

结构化数据包括基础数据、主数据、事务数据、报告数据、观测数据、规则数据。结构化数据的共同特点是以信息架构为基础，建立统一的数据资产目录、数据标准与模型。本节将重点介绍六类结构化数据的治理方法。

3.2.1　基础数据治理

基础数据用于对其他数据进行分类，在业界也称作参考数据。基础数据通常是静态的（如国家、币种），一般在业务事件发生之前就已经预先定义。它的可选值数量有限，可以用作业务或IT的开关和判断条件。当基础数据的取值发生变化的时候，通常需要对流程和IT系统进行分析和修改，以满足业务需求。因此，基础数据的管理重点在于变更管理和统一标准管控。

基础数据在支撑场景分流、流程自动化、提升分析质量方面起着关键作用，治理基础数据的价值如图3-2所示。

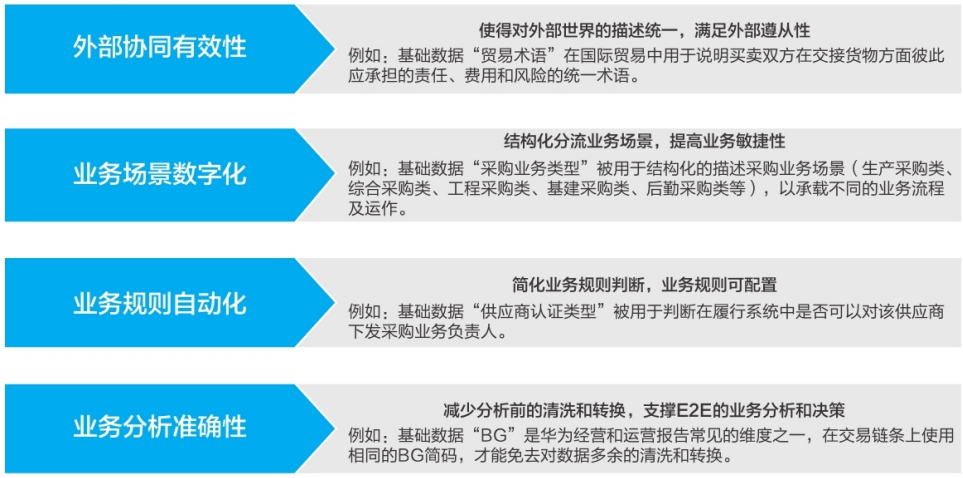


图3-2　基础数据治理的价值

因此，有效地管理基础数据对企业来说可以产生巨大的收益。以“运输方式”为例，基础数据的管理收益如图3-3所示。

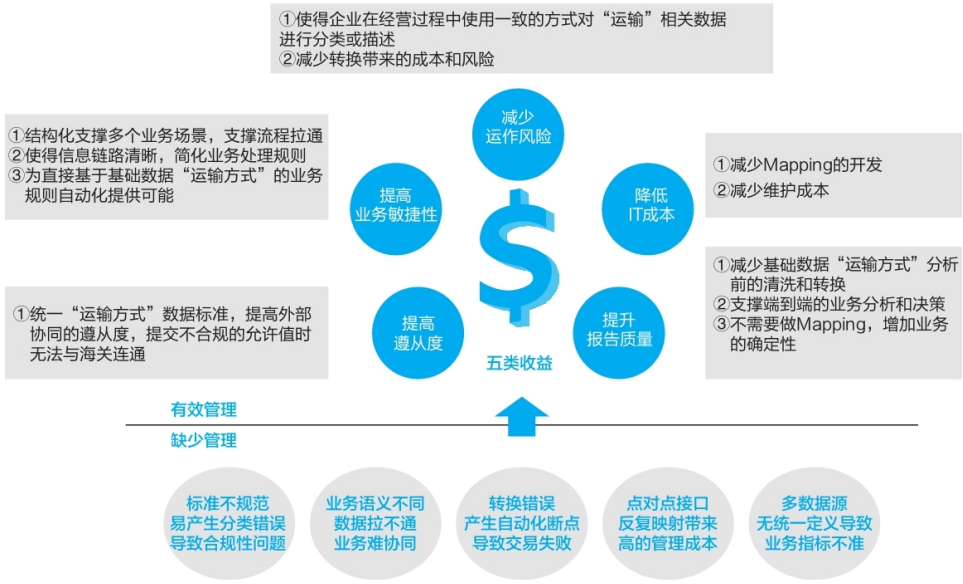


图3-3　基础数据“运输方式”案例

华为建立了一个完整的基础数据管理框架（如图3-4所示），通过明确各方的管理责任、发布相关的流程和规范以及建立基础数据管理平台等来确保基础数据的有效管理。

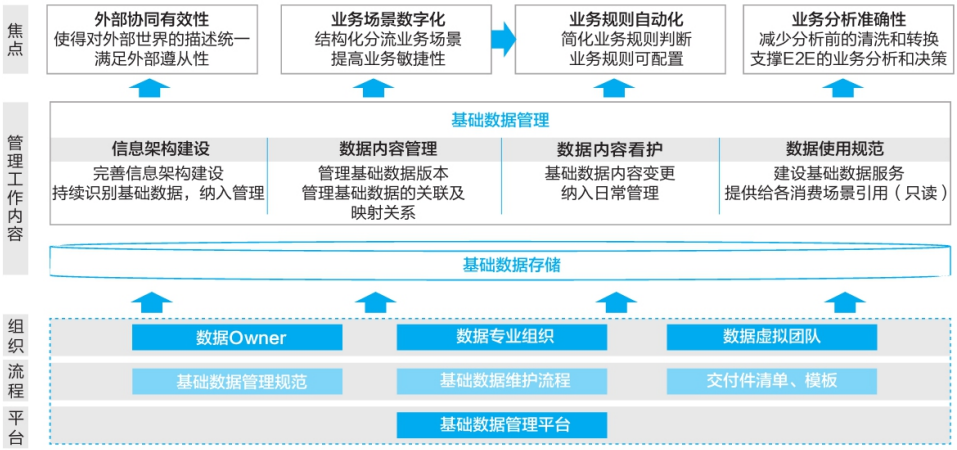


图3-4　基础数据治理框架

3.2.2　主数据治理

主数据是参与业务事件的主体或资源，是具有高业务价值的、跨流程和跨系统重复使用的数据。主数据与基础数据有一定的相似性，都是在业务事件发生之前预先定义；但又与基础数据不同，主数据的取值不受限于预先定义的数据范围，而且主数据的记录的增加和减少一般不会影响流程和IT系统的变化。但是，主数据的错误可能导致成百上千的事务数据错误，因此主数据最重要的管理要求是确保同源多用和重点进行数据内容的校验。华为的主数据管理策略如图3-5所示。



图3-5　主数据治理策略

华为的主数据范围包括客户、产品、供应商、组织、人员主题，每个主数据都有相应的架构、流程及管控组织来负责管理。

鉴于主数据管理的重要性，对于每个重要的主数据，都会发布相应的管理规范，数据管家依据数据质量标准定期进行数据质量的度量与改进。

同时，对于主数据的集成消费按照如下管理框架进行管理，如图3-6所示。

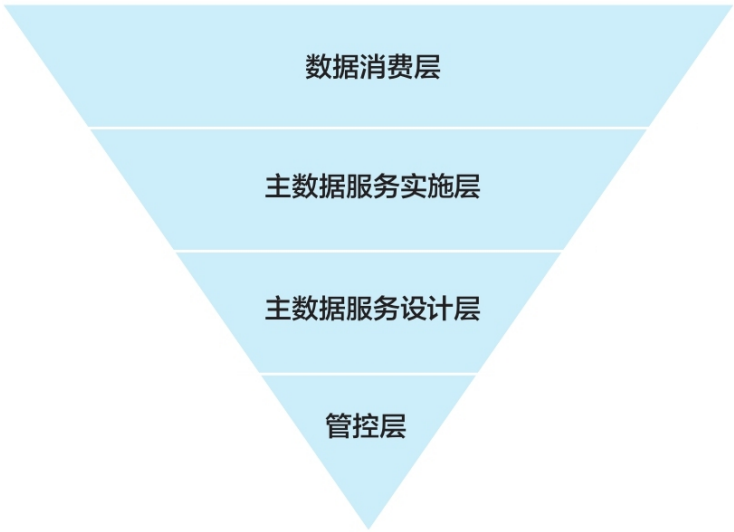


图3-6　主数据治理框架

- **数据消费层：**数据消费层包括所有消费数据的IT产品团队，负责提出数据集成需求和集成接口实施。
- **主数据服务实施层：**负责主数据集成解决方案的落地，包括数据服务的IT实施和数据服务的配置管理。
- **主数据服务设计层：**为需要集成主数据的IT产品团队提供咨询和方案服务，负责受理主数据集成需求，制定主数据集成解决方案，维护主数据的通用数据模型。
- **管控层：**管控层由信息架构专家组担任，负责主数据规则的制定与发布，以及主数据集成争议或例外的决策。

接下来介绍客户主数据治理的实践。客户数据是企业最重要的主数据之一，几乎贯穿所有业务经营活动。客户数据在全流程中的及时性、准确性、完整性、一致性、有效性、唯一性是业务高效运作、经营可控的重要保障。随着业务发展，华为客户数量迅速增长，客户数据种类复杂多样，因此要构建客户

数据管理和服务化能力，以满足经营分析、交易打通、内外部遵从、客户价值挖掘等核心要求，支撑面向多BG的战略转变。

在客户数据治理和服务化改造前，客户历史数据质量较差，一个客户编码存在多个BG属性，导致无法直接基于客户维度生成BG报告，同时无法支持基于不同业务特点对同一客户授信、控制备发货。

下游系统违规录入客户数据会影响财报的准确性，风险等级高。财报内控管理建议书中指出，“风险等级评定高，部分同源的主数据在不同系统中维护，可能导致各系统间不一致，增加维护的工作量”。

经过对3大BG，财经、供应链、变革项目组等24个部门的情况进行收集和分析，客户数据的问题根源在以下几个方面。

- 客户信息不完整，且下游系统未严格遵循数据源头所定义的标准。
- 数据架构不灵活、紧耦合，不能有效支撑多BG的业务管理。
- 下游系统集成管理不严格，存在多源头录入。
- 客户数据源头的数据质量管理控制点无法延伸到下游的各集成IT系统中。

为彻底解决客户数据问题，华为制订了客户数据管理及服务化架构方案，以客户数据质量为核心，严控数据流入与流出两个端口，搭建客户数据管理及服务平台，统一数据架构和标

准，通过服务化架构实现“数出一孔”，提升财报准确性、提升运作效率、降低运营风险（如图3-7所示）。

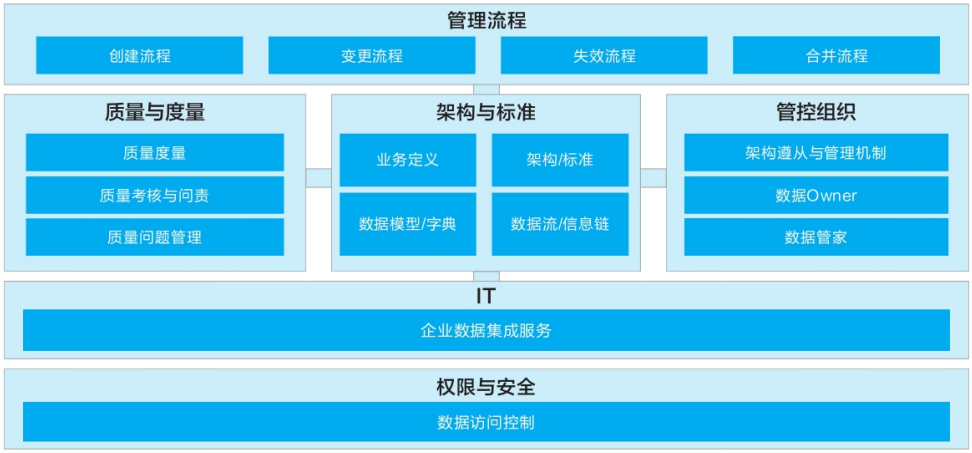


图3-7　客户主数据治理框架

以客户数据架构的重构和管理为基础，制订了Account & Legal Entity两级架构。Account用于华为公司市场拓展、销售管理及数据归集等内部经营管理，是不具备与华为公司签约资格的对象；Legal Entity（法人客户）是依法具有民事权利能力和民事行为能力，依法独立享有民事权利和承担民事义务，具备与华为公司签约资格的对象，包括企业、国家机关、事业单位和社会团体等。Account数据确保客观、稳定，各BG、各流程、各系统一致；而Legal Entity基于BG分层解耦，按内容性质区分“身份证”信息和其他业务信息，满足多BG业务管理。客户主数据架构如图3-8所示。

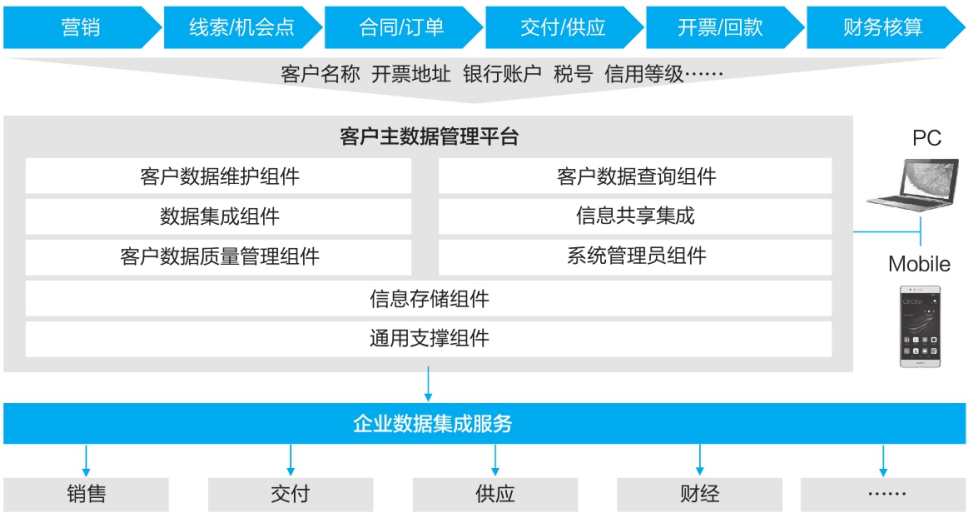


图3-8　客户主数据平台架构

以客户数据架构的优化为基础，重点通过数据服务化方式对整个华为公司原有的集成方式进行改造，包括下游的136个IT系统和应用，3大类近2000个改造点，从根本上消除了原有的不合理的数据集成方式，具体包括如下4点。

1）确保下游IT系统或应用不从非数据源系统集成客户数据。例如：A系统从B系统（非数据源）集成主数据，并且在A系统落地了物理表。

2）确保下游IT系统或应用集成合法数据源且不修改属性。例如，修改了展现业务含义的字段，将编码改为编号。

3）确保下游IT系统或应用中不补录数据。例如，客户数据从合法的数据源集成，集成后对客户数据进行行记录的新增或补录。

4）确保下游IT系统或应用不向后传递数据。例如，某系统未以数据服务方式从数据源获取数据，而是直接调用中间系统（非数据源）的客户数据。

通过服务化改造提升了全流程数据的一致性，同时为各个环节带来了明显的业务价值，包含如下几点。

1）实现“数出一孔”，提高数据质量。提高数据准确性与及时性，减少不同部门之间的对账成本，帮助提高财经等报告的准确性。

2）满足内外部应遵从的要求，降低华为公司风险。实现数据“一点录入，多点调用”，满足财报内控及内外部审计要求，提高客户数据真实性，降低合同造假等业务运营风险。

3）支持交易流打通，提升运营效率。满足各流程对客户数据的要求，降低合同非正常变更及退票风险。

4）支持经营分析和价值评价。支持基于客户视角生成BG管理报告与各业务部门经营管理分析。

5）支持价值挖掘，聚焦优质客户。支持客户360度分析，驱动优质资源瞄准优质客户，提高市场响应效率。

3.2.3　事务数据治理

事务数据在业务和流程中产生，是业务事件的记录，其本身就是业务运作的一部分。事务数据是具有较强时效性的一次性业务事件，通常在事件结束后不再更新。

事务数据会调用主数据和基础数据。以客户框架合同为例，核心属性有32个，其中调用基础数据和主数据24个，占75%；客户框架合同本身特有的属性8个，占 25%。同时，框架合同也引用了机会点的编码和投标项目的编码等事务数据的信息。

因此，事务数据的治理重点就是管理好事务数据对主数据和基础数据的调用，以及事务数据之间的关联关系，确保上下游信息传递顺畅。在事务数据的信息架构中需明确哪些属性是引用其他业务对象的，哪些是其自身特有的。对于引用的基础数据和主数据，要尽可能调用而不是重新创建。

3.2.4 报告数据治理

报告数据是指对数据进行处理加工后，用作业务决策依据的数据。它用于支持报告和报表的生成。

用于报告和报表的数据可以分为如下几种。

- 用于报表项数据生成的事实表、指标数据、维度。
- 用于报表项统计和计算的统计函数、趋势函数及报告规则。
- 用于报表和报告展示的序列关系数据。

- 用于报表项描述的主数据、基础数据、事务数据、观测数据。
- 用于对报告进行补充说明的非结构化数据。

报告数据涵盖的范围较广，如主数据、基础数据等，这些数据类别本身已经有相应的管理机制和规范，这里我们重点对部分新的细分数据类型进行说明。

1）事实表：从业务活动或者事件中提炼出来的性能度量。其特点为：

- 每个事实表由颗粒度属性、维度属性、事务描述属性、度量属性组成；
- 事实表可以分为基于明细构建的事实表和基于明细做过汇聚的事实表。

2）维度：用于观察和分析业务数据的视角，支持对数据进行汇聚、钻取、切片分析。其特点为：

- 维度的数据一般来源于基础数据和主数据；
- 维度的数据一般用于分析视角的分类；
- 维度的数据一般有层级关系，可以向下钻取和向上聚合形成新的维度。

3）统计型函数：与指标高度相关，是对指标数量特征进一步的数学统计，例如均值、中位数、总和、方差等。其特点为：

- 通常反映某一维度下指标的聚合情况、离散情况等特征；
- 其计算数值在报告中通常呈现为图表中的参考线。

4）趋势型函数：反映指标在时间维度上变化情况的统计方式，例如同比、环比、定基比等。其特点为：

- 通常将当期值与历史某时点值进行比较；
- 调用时，需要收集指标的历史表现数据；
- 其计算数值在报告中通常呈现为图表中的趋势线。

5）报告规则数据：一种描述业务决策或过程的陈述，通常是基于某些约束下产生的结论或需要采取的某种措施。其特点为：

- 将业务逻辑通过函数运算体现，通常一个规则包含多个运算和判断条件；
- 规则的计算结果一般不直接输出，需要基于计算结果翻译成业务语言后输出；
- 规则通常与参数表密切相关。

6）序列关系数据：反映报告中指标及其他数据序列关系的数据。

3.2.5 观测数据治理

观测数据是通过观测工具获取的数据，观测对象一般为人、事、物、环境。

相比传统数据，观测数据通常数据量较大且是过程性的，由机器自动采集生成。不同感知方式获取的观测数据，其数据资产管理要素不同。

观测数据的感知方式可分为软感知和硬感知。软感知是使用软件或者各种技术进行数据收集，收集的对象存在于数字世界，通常不依赖于物理设备，一般是自动运行的程序或脚本；硬感知是利用设备或装置进行数据收集，收集的对象为物理世界中的物理实体，或者是以物理实体为载体的信息，其数据的感知过程是数据从物理世界向数字世界的转化过程。

观测数据的特征有如下几点：

1）观测数据通常数据量较大且是过程性的，主要用作监控分析。例如，视频监控器产生的视频数据、操作系统产生的日志记录数据等；

2）观测数据由机器自动采集生成。例如，各种传感器或探针记录观测对象产生的数据；

3）观测数据是观测工具采集回 来的原始数据（Raw Data），仅转换结构和格式，不做任何业务规则解析。

观测数据的管理模型如图3-9所示。

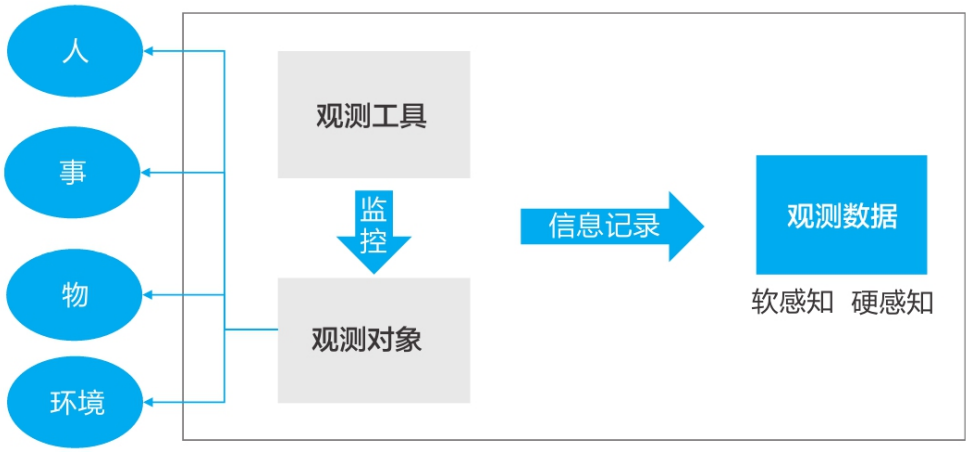


图3-9　观测数据管理模型

观测工具的元数据可以作为数据资产管理：软感知（埋点、日志收集、爬虫）观测工具抽象成业务对象，由IT部门担任数据Owner进行统一管理；硬感知观测工具作为资源类数据，也建议作为业务对象由相应的领域担任数据Owner进行管理。

原则上，观测对象要定义成业务对象进行管理，这是观测数据管理的前提条件。

观测数据需要记录观测工具、观测对象。针对不同感知方式获取的观测数据，其资产管理方案也不尽相同。例如，以用户界面浏览记录为例，如果是对销售机会点的查询访问观测，应当归属到相应业务领域；如果是对页面性能、页面UV、PV的观测，应当归属到IT部门。

3.2.6　规则数据治理

在业务规则管理方面，华为经常面对“各种业务场景业务规则不同，记不住，找不到”“大量规则在政策、流程等文件中承载，难以遵守”“各国 规则均不同，IT能否一国 一策、快速上线”等问题。

规则数据是结构化描述业务规则变量（一般为决策表、关联关系表、评分卡等形式）的数据，是实现业务规则的核心数据，如业务中普遍存在的基线数据。

规则数据主要有以下特征：

1）规则数据不可实例化；

2）规则数据包含判断条件和决策结果两部分信息，区别于描述事物分类信息的基础数据；

3）规则数据的结构在纵向（列）、横向（行）两个维度上相对稳定，变化形式多为内容刷新；

4）规则数据的变更对业务活动的影响是大范围的。

其基本原则为：

1）规则数据的管理是为了支撑业务规则的结构化、信息化、数字化，目标是实现规则的可配置、可视化、可追溯。

2）不同于标准化的信息架构管理，规则数据的管理具有轻量化、分级的特点。重要的、调用量大、变动频繁的业务规则需要通过规则数据管理，使其从代码中解耦，进行资产注册；

使用广泛的、有分析需求的规则数据需要通过注册入湖，实现共享和复用。

3）业务规则在架构层次上与流程中的业务活动相关联，是业务活动的指导和依据，业务活动的结果通过该业务活动的相关业务对象的属性来记录。业务规则通过业务活动对业务事实、业务行为进行限制，业务人员可以根据业务规则判断业务情况，采取具体行动。

4）业务规则包含规则变量和变量之间的关系，规则数据主要描述规则的变量部分，是支撑业务规则的核心数据（如图3-10所示）。

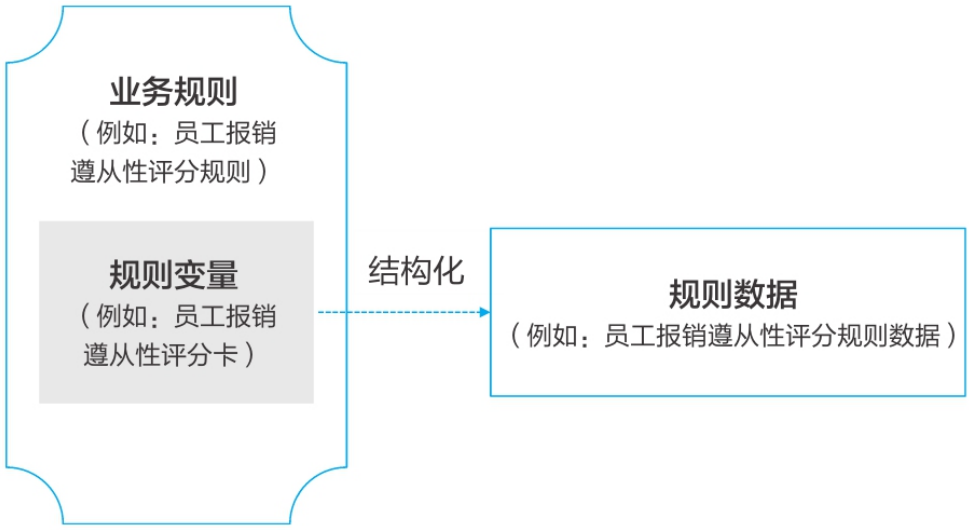


图3-10　业务规则与规则数据之间的关系

此外，运行规则所需要的输入数据、输出数据，包括动态数据库访问对象、内存表缓存、Excel、XML处理类等，主要起支撑作用，不在规则数据的范畴。

规则数据必须有唯一的数据Owner，其负责开展规则数据的信息架构建设与维护、数据质量的监控与保障、数据服务建设、数据安全授权与定密等工作。相应的数据管家支持数据Owner对所管辖的业务中的规则数据进行治理，包括建设和维护信息架构、确保架构落地遵从、例行监控数据质量等。

规则数据的元数据要记录与业务规则的关系（规则数据定义前应先完成业务规则的识别和定义）。一个业务规则可以包含零个、一个或多个规则数据，一个规则数据在信息架构上对应一个逻辑数据实体，在物理实现上一般对应一个物理表。规则数据要遵从信息架构资产管理要求（包括明确规则数据的Owner、制定数据标准、明确数据源等），按照信息安全要求定密，以方便规则数据的管理、共享和分析。

3.3 以特征提取为核心的非结构化数据管理

随着业务对大数据分析的需求日 益增长，非结构化数据的管理逐渐成为数据管理的重要组成部分。非结构化数据包括无格式文本、各类格式文档、图像、音频、视频等多种异构的格式文件，较之结构化数据，其更难标准化和理解，因此在存储、检索以及消费使用时需要智能化的IT技术与之匹配。华为的非结构化数据包括文档（邮 件、Excel、Word、PPT）、图片、音频、视频等。

相较于结构化数据，非结构化元数据管理除了需要管理文件对象的标题、格式、Owner等基本特征和定义外，还需对数据内容的客观理解进行管理，如标签、相似性检索、相似性连接等，以便于用户搜索和消费使用。**因此，非结构化数据的治理核心是对其基本特征与内容进行提取，并通过元数据落地来开展的。**非结构化数据的管理模型如图3-11所示。

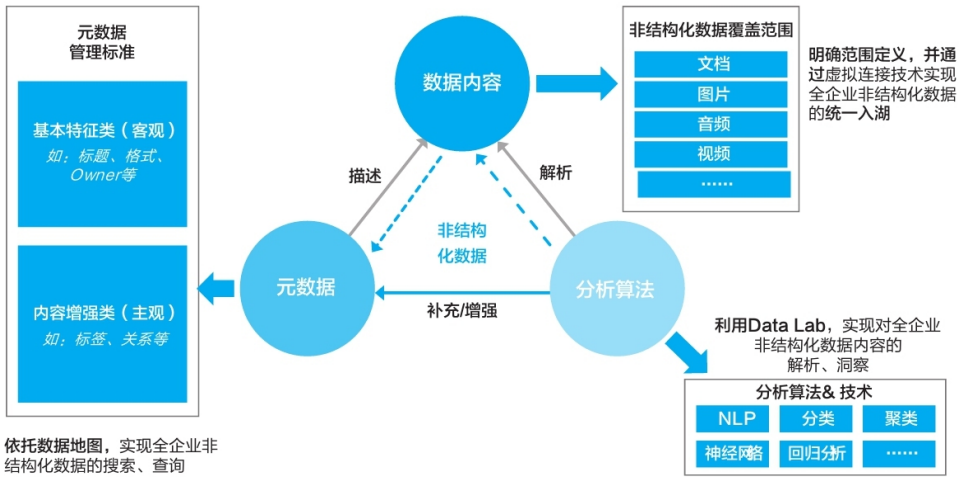


图3-11 非结构化数据管理模型

非结构化数据的元数据可以分为基本特征类（客观）和内容增强类（主观）两类。

1）基本特征类：参考都柏林十五个核心元数据，实现对非结构化数据对象的规范化定义，如标题、格式、来源等。

2）内容增强类：基于非结构化数据内容的上下文语境，解析目标文件对象的数据内容，加深对目标对象的客观理解，如标签、相似性检索、相似性连接等。

非结构化数据的元数据管理采用统分统管的原则，即基本特征类属性由公司进行统一管理，内容增强类属性由相关承担数据分析工作的项目组自行设计，但其分析结果都应由公司元数据管理平台自动采集后进行统一存储。

元数据管理平台通过“基本特征类元数据流”和“内容增强类元数据流”两条线来实现对非结构化数据的元数据管理和消费使用。

1）基本特征类元数据流

元数据管理平台基于收集到的各类非结构化数据源信息，自动完成基础特征类元数据的采集工作，按照管理规范和要求通过标准化、整合后存储在元数据管理平台中，并在完成元数据过滤、排序后将结果在元数据报告中进行可视化展示，以供用户消费使用。

2）内容增强类元数据流

基于元数据管理平台中基本特征类元数据的信息，各数据分析项目组解析目标非结构化对象的数据内容，并将分析结果通过元数据采集、元数据标准化&整合后统一存放在元数据管理平台中，以供用户一并消费使用，增强用户体验。

非结构化数据的处理过程如图3-12所示。

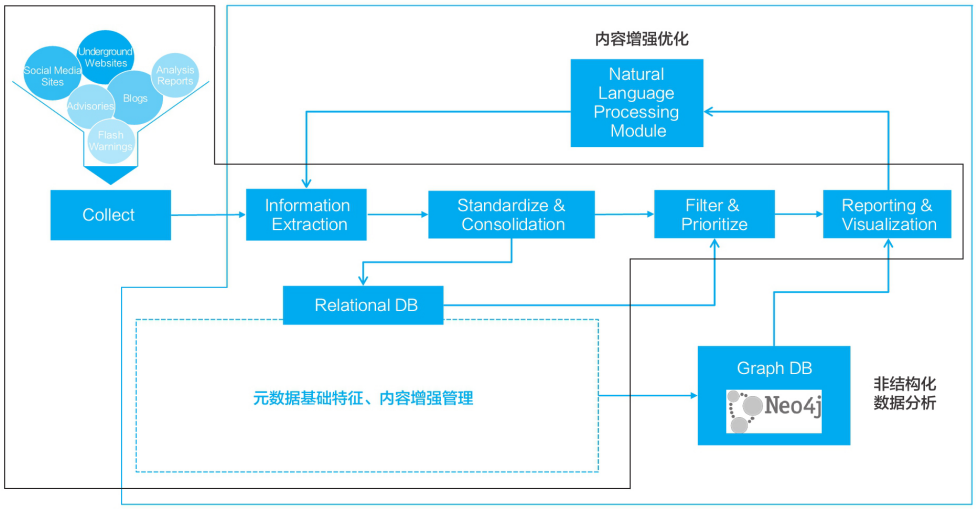


图3-12　非结构化数据处理过程

3.4　以确保合规遵从为核心的外部数据管理

外部数据是指华为公司引入的外部组织或者个人拥有处置权利的数据，如供应商资质证明、消费者洞察报告等。外部数据治理的出发点是合规遵从优先，与内部数据治理的目的不同。

外部数据的治理主要遵循以下原则。

1）**合规优先原则：**遵从法律法规、采购合同、客户授权、公司信息安全与公司隐私保护政策等相关规定。

2）**责任明确原则：**所有引入的外部数据都要有明确的管理责任主体，承担数据引入方式、数据安全要求、数据隐私要求、数据共享范围、数据使用授权、数据质量监管、数据退出销毁等责任。

3）**有效流动原则：**使用方优先使用公司已有数据资产，避免重复采购、重复建设。

4）**可审计、可追溯原则：**控制访问权限，留存访问日志，做到外部数据使用有记录、可审计、可追溯。

5）**受控审批原则：**在授权范围内，外部数据管理责任主体应合理审批使用方的数据获取要求。

在以上原则指导下，我们要求所有采购的外部数据要注册，在合规的前提下鼓励数据共享，避免重复采购。其他方式引入的外部数据，由管理责任主体决定登记方式。根据法律条款和授权范围，外部数据管理责任主体有权决定外部数据是否入数据湖，如果需要入数据湖，必须遵从数据湖建设相应的流程和规范。同时，外部数据管理责任主体有义务告知使用方合规使用外部数据，对于不合规的使用场景，不予授权；数据使用方要遵从外部数据管理责任主体的要求，对不遵从要求所引起的后果承担责任。

3.5 作用于数据价值流的元数据管理

无论结构化数据，还是非结构化数据，或者外部数据，最终都会通过元数据治理落地。华为将元数据治理贯穿整个数据价值流，覆盖从数据产生、汇聚、加工到消费的全生命周期。

3.5.1 元数据治理面临的挑战

华为在进行元数据治理以前，遇到的元数据问题主要表现为数据找不到、读不懂、不可信，数据分析师们往往会陷入数据沼泽中，例如以下常见的场景。

- 某子公司需要从发货数据里对设备保修和维保进行区分，用来不对过保设备进行服务场景分析。为此，数据分析师需面对几十个IT系统，不知道该从哪里拿到合适的数据。
- 因盘点内部要货的研发领料情况，需要从IT系统中获取研发内部的要货数据，面对复杂的数据存储结构（涉及超过40个物理表和超过1000个字段）、物理层和业务层脱离的情况，业务部门的数据分析师无法读懂物理层数据，只能提出需求向IT系统求助。
- 某子公司存货和收入管理需要做繁重的数据收集与获取工作，运行一次计划耗时超过20个小时。同时，由于销售、供应、交付各领域计划的语言不通，还需要数据分析师进行大量人工转换与人工校验。

以上场景频繁出现在公司日 常运营的各个环节，极大地阻碍了公司数字化转型的进行，其根本原因就在于业务元数据与技术元数据未打通，导致业务读不懂IT系统中的数据。并且缺乏面向普通业务人员的准确、高效的数据搜索工具，业务人员无法快速获取可信数据。元数据管理的痛点如图3-13所示。



图3-13　元数据管理痛点

为解决以上痛点，华为建立了公司级的元数据管理机制。制定了统一的元数据管理方法、机制和平台，拉通业务语言和机器语言。**确保数据“入湖有依据，出湖可检索”成为华为元数据管理的使命与目标。**基于高质量的元数据，通过数据地图就能在企业内部实现方便的数据搜索。

元数据是描述数据的数据，用于打破业务和IT之间的语言障碍，帮助业务更好地理解数据。元数据通常分为业务、技术和操作三类。

- **业务元数据：**用户访问数据时了解业务含义的途径，包括资产目录、Owner、数据密级等。
- **技术元数据：**实施人员开发系统时使用的数据，包括物理模型的表与字段、ETL规则、集成关系等。
- **操作元数据：**数据处理日志及运营情况数据，包括调度频度、访问记录等。

在企业的数字化运营中，元数据作用于整个价值流，在从数据源到数据消费的五个环节中都能充分体现元数据管理的价值。

- **数据消费侧：**元数据能支持企业指标、报表的动态构建。
- **数据服务侧：**元数据支持数据服务的统一管理和运营，并实现利用元数据驱动IT敏捷开发。
- **数据主题侧：**元数据统一管理分析模型，敏捷响应井喷式增长的数据分析需求，支持数据增值、数据变现。
- **数据湖侧：**元数据能实现暗数据的透明化，增强数据活性，并能解决数据治理与IT落地脱节的问题。
- **数据源侧：**元数据支撑业务管理规则有效落地，保障数据内容合格、合规。

3.5.2 元数据管理架构及策略

元数据管理架构包括产生元数据、采集元数据、注册元数据和运维元数据。

- **产生元数据：**制定元数据管理相关流程与规范的落地方案，在IT产品开发过程中实现业务元数据与技术元数据的连接。
- **采集元数据：**通过统一的元模型从各类IT系统中自动采集元数据。
- **注册元数据：**基于增量与存量两种场景，制定元数据注册方法，完成底座元数据注册工作。

- **运维元数据：**打造公司元数据中心，管理元数据产生、采集、注册的全过程，实现元数据运维。
- **元数据管理方案：**通过制定元数据标准、规范、平台与管控机制，建立企业级元数据管理体系，并推动其在公司各领域落地，支撑数据底座建设与数字化运营。

华为元数据管理整体方案如图3-14所示。

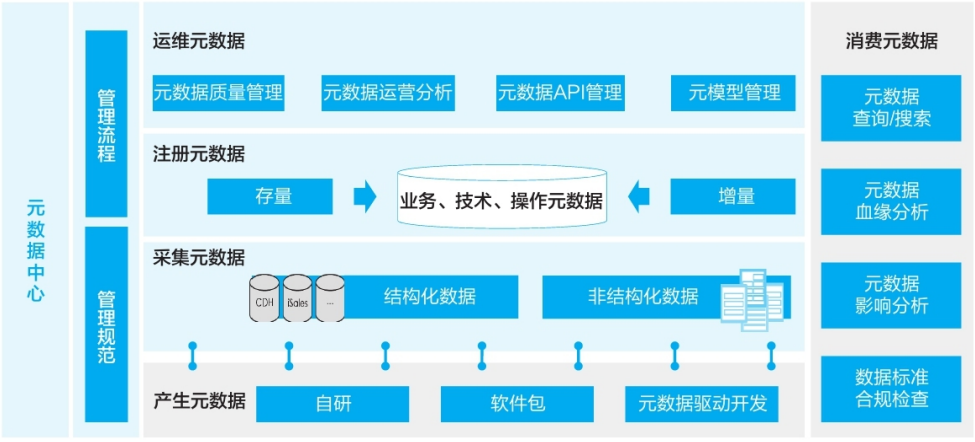


图3-14　华为元数据管理整体方案

3.5.3　元数据管理

1. 产生元数据

（1）明确业务元数据、技术元数据和操作元数据之间的关系，定义华为公司元数据模型，如图3-15所示。

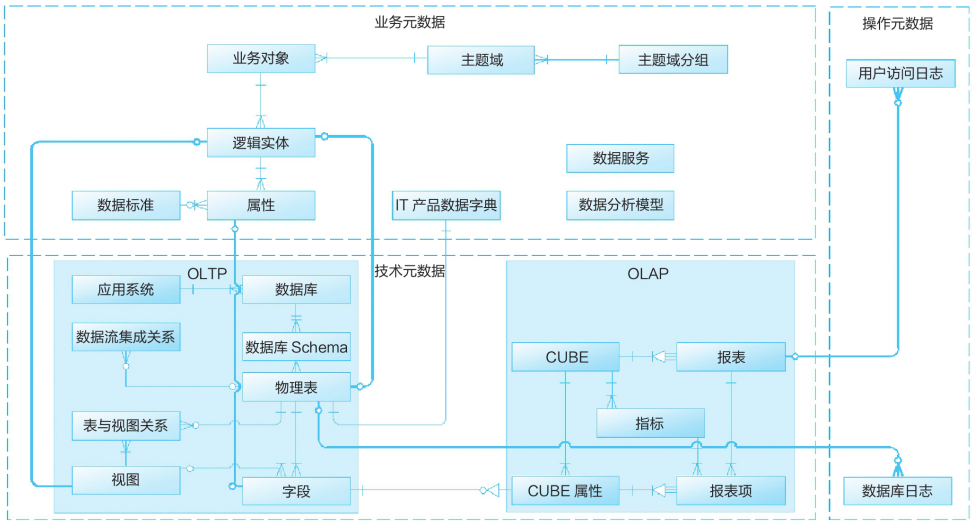


图3-15　华为元数据模型

（2）针对找数据及获取数据难的痛点，明确业务元数据、技术元数据、操作元数据的设计原则。

1）业务元数据设计原则

- 一个主题域分组下有多个主题域，一个主题域下有多个业务对象，一个业务对象下有多个逻辑实体，一个逻辑实体下有多个属性，一个属性有一个数据标准。
- 每个数据标准可被一个或多个属性引用，每个属性归属于一个逻辑实体，每个逻辑实体归属于一个业务对象，每个业务对象归属于一个主题域，每个主题域归属于一个主题域分组。

2）技术元数据设计原则

- 物理表设计须满足三范式，如为了降低系统的总体资源消耗，提高查询效率，可反范式设计。

- 物理表、视图和字段的设计须基于用途进行分类。
- 承载业务用途的物理表、虚拟表、视图必须与逻辑实体一一对应，承载业务用途的字段必须与属性一一对应。
- 系统间的数据传递须优先采用数据服务。

3）操作元数据设计原则

日志目的不同的进行分类设计，日志目的相同的进行相同设计（非自研场景按软件包适配）。

（3）规范数据资产管理，设计数据资产编码规范

1）数据资产编码规范

华为数据资产编码的主要包括业务元数据和技术元数据两大类，其中业务元数据包含主题域分组、主题域、业务对象、逻辑实体、属性、数据标准；技术元数据包含物理数据库、Schema、表、字段。具体的定义与描述如表3-2所示。

表3-2　数据资产编码规范

数据资产大类	数据资产子类	定义 / 描述
业务元数据	主题域分组	公司顶层信息分类，通过数据视角体现公司最高层面关注的业务领域
	主题域	互不重叠数据的高层面的分类，用于管理其下一级的业务对象
	业务对象	业务领域重要的人、事、物，承载了业务运作和管理涉及的重要信息
	逻辑实体	描述业务对象的某种业务特征属性的集合
	属性	描述所属业务对象的性质和特征，反映信息管理最小粒度
	数据标准	用于描述公司层面需共同遵守的属性层数据的含义和业务规则，相关标准一旦确定且发布，全公司范围内需严格遵守
技术元数据	数据库	按照数据结构来组织、存储和管理数据的仓库
	Schema	数据库对象的集合，一个用户一般对应一个Schema
	表	分为物理表和虚拟表，物理表为数据库的核心组件，由行和列组成。行包括若干列信息项，一行数据称为一个或一条记录；列又称为字段，用于描述相关数据的特征 虚拟表基于物理表进行定义，用于提供数据服务，但不实际存储数据，其数据使用方式和物理表一致
	字段	表中的列信息

2）数据资产编码原则

数据资产编码（DAN）是通过一组数字、符号等组成的字符串 去唯一标识华为公司内部每一个数据资产，基于此唯一标识，保证各业务领域对同一数据资产的理解和使用一致，它的设计遵循以下原则。

- **统一性原则：**华为公司内部只能使用一套数据资产编码，以方便不同业务部门之间的沟通和IT应用之间的数据交换。
- **唯一性原则：**每一个数据资产只能用唯一的数据资产编码进行标识，不同数据资产的编码不允许重复，同一个编码也只能对应到一个数据资产上。
- **可读性原则：**数据资产编码作为数据资产分类、检索的关键词和索引，需要具备一定的可读性，让用户通过编码就能初步判断其对应的数据资产类型。
- **扩展性原则：**数据资产的编码要从数据管理角度适当考虑未来几年的业务发展趋势，其编码长度要能适当扩展，同时不影响整个编码体系。

3）业务元数据资产编码规则

业务元数据资产编码规则主要包含三个部分：第一部分为主题域分组的编码规则，主题域分组的编码由公司统一分配；第二部分为主题域、业务对象、逻辑实体、属性的编码规则，这部分主要由数据治理平台按照编码规则自动生成；第三部分主要为业务元数据包含的子类对应的数据资产类型代码。

2. 采集元数据

元数据采集是指从生产系统、IT设计平台等数据源获取元数据，对元数据进行转换，然后写入元数据中心的过程。元数据的来源可分为如表3-3所示的六类。

表3-3　元数据来源

序　号	元数据来源类型	元数据来源
1	关系数据库	Oracle、MS SQLServer、DB2 等
2	建模工具	ERWin、PowerDesigner 等
3	数据集成工具	DataStage、PowerCenter 等
4	BI 报表工具	Cognos、SQL Server Reporting Services 等
5	调度工具	Automation
6	开发语言及脚本	Perl（日志方式）、SP（注释方式）
7	其他	元数据采集虚拟库等

1）选择适配器

适配器是指针对不同的元数据来源，采用相应的采集方式获取元数据的程序，元数据的来源种类繁多，因而须选择相对应的适配器及元模型。

2）配置数据源

配置数据源是采集元数据的关键，在确定数据源所选择的适配器类型、适配器版本、元模型的基础上，配置数据源的名称、连接参数和描述。

3）配置采集任务

采集任务为自动调度的工作单元，为元数据的采集提供自动化的、周期性的、定时的触发机制。

3. 注册元数据

大多数企业的数字化建设都存在增量和存量两种场景，如何同时有效地管理这两种场景下的元数据就成了问题的关键。华为通过标准的元数据注册规范和统一的元数据注册方法，实现了两种场景下业务元数据和技术元数据的高效连接，使业务人员能看懂数据、理解数据，并通过数据底座实现数据的共享与消费。

（1）元数据注册原则

元数据注册的原则包括如下三点：

- 数据Owner负责，是谁的数据就由谁负责业务元数据和技术元数据连接关系的建设和注册发布；
- 按需注册，各领域数据管理部根据数据搜索、共享的需求，推进元数据注册；
- 注册的元数据的信息安全密级为内部公开。

（2）元数据注册规范

通过“元数据注册三步法”完成元数据注册，如图3-16所示。

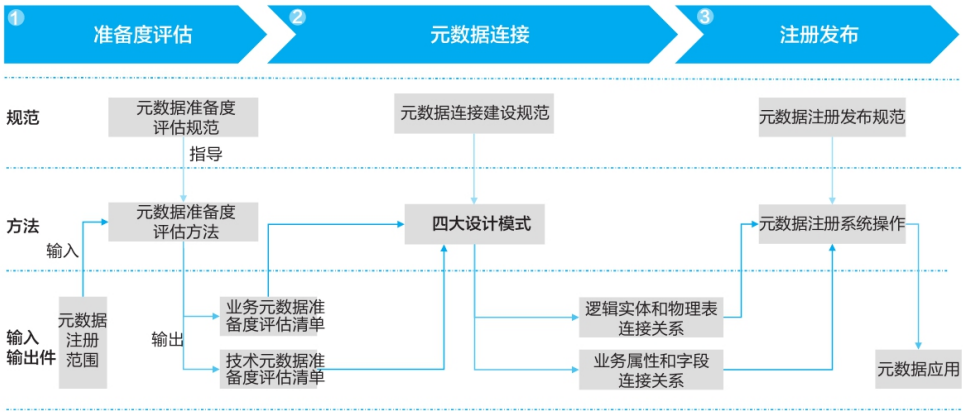


图3-16　元数据注册方法

1）准备度评估项包括如下检查要点：

- IT系统名称必须是公司标准名称；
- 数据资产目录是否经过评审并正式发布；
- 数据Owner是否确定数据密级；
- 物理表/虚拟表/视图名。

2）元数据连接需遵从以下规范。

- 逻辑实体和物理表/虚拟表/视图一对一连接规范：在业务元数据与技术元数据连接的过程中，必须遵从逻辑实体和物理表/虚拟表/视图一对一的连接原则，如果出现一对多、多对一或多对多的情况，各领域需根据实际场景，参照元数据连接的设计模式进行调整。
- 业务属性与字段一对一连接规范：除了逻辑实体与物理表/虚拟表/视图要求一一对应外，属性和非系统字段（具备业务含义）也要求遵从一对一的连接原则，如出现属性与字

段匹配不上的情况，可参考元数据关联的设计模式进行调整。

完成元数据注册后，通过元数据中心自动发布。

（3）元数据注册方法

元数据注册分为增量元数据注册和存量元数据注册两种场景。

增量场景相对容易，在IT系统的设计与开发过程中，落实元数据的相关规范，确保系统上线时即完成业务元数据与技术元数据连接，通过元数据采集器实现元数据自动注册。

针对存量场景，华为设计了元数据注册的四 大模式。在符合元数据设计规范的前提下，进行业务元数据与技术元数据的连接及注册。

模式一：一对一模式

适用场景

适用于数据已发布信息架构和数据标准且物理落地，架构、标准与物理落地能一一对应的场景。

解决方案

- 将逻辑实体和物理表一对一连接。
- 逻辑实体属性和物理表字段一对一连接。

应用实例

具体的应用实例如图3-17所示。

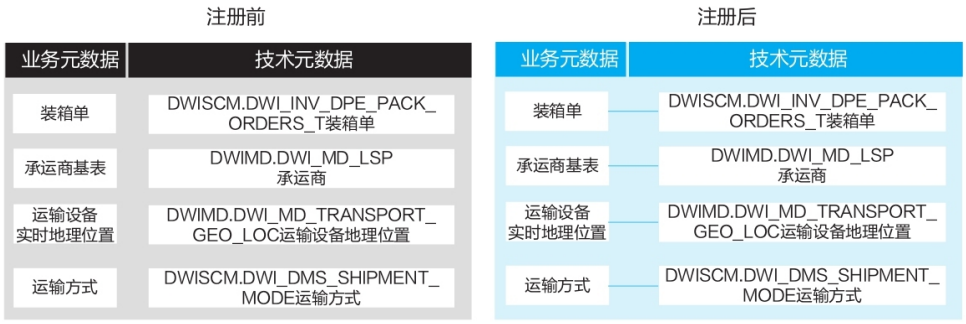


图3-17　元数据注册一对一模式样例

模式二：主从模式

适用场景

适用于主表和从表结构一致，但数据内容基于某种维度分别存储在不同物理表中的场景。例如，按时间或项目归档，或按区域进行分布式存储。

解决方案

- 识别主物理表和从属物理表。
- 以主物理表为核心，纵向UNION所有从属物理表，并固化为视图。
- 将视图、逻辑实体、字段和业务属性一对一连接。

应用实例

具体的应用实例如图3-18所示。

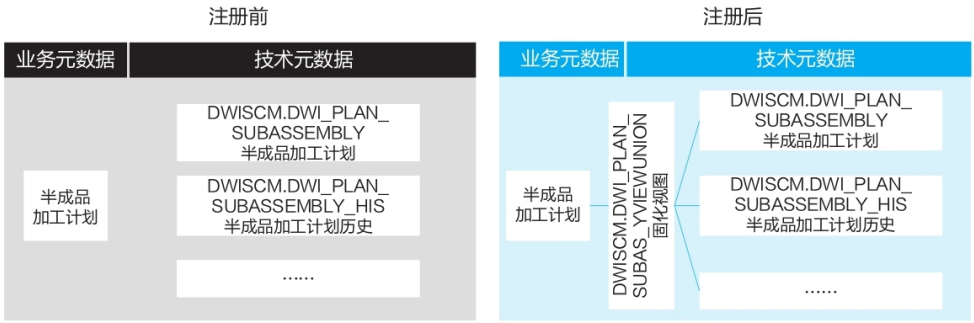


图3-18　元数据注册主从模式样例

模式三：主扩模式

适用场景

适用于逻辑实体的大部分业务属性在主物理表，少数属性在其他物理表中的场景。

解决方案

- 识别主物理表和扩展物理表。
- 以主物理表为核心，横向JOIN所有扩展物理表，完成扩展属性与主表的映射，并固化为视图。
- 将视图、逻辑实体、字段和业务属性一对一连接。

应用实例

具体应用实例如图3-19所示。

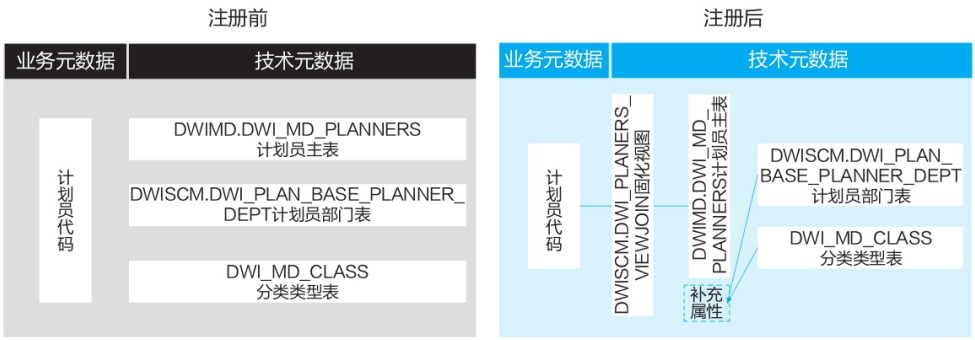


图3-19　元数据注册主扩模式样例

模式四：父子模式

适用场景

适用于多个逻辑实体业务属性完全相同，按不同场景区分逻辑实体名称，但落地在同一张物理表的场景。

解决方案

- 识别一张物理表和对应的多个逻辑实体。
- 将物理表按场景拆分和多个逻辑实体一对一连接。
- 将物理表字段和多个逻辑实体属性一对一连接。

应用实例

具体应用实例如图3-20所示。

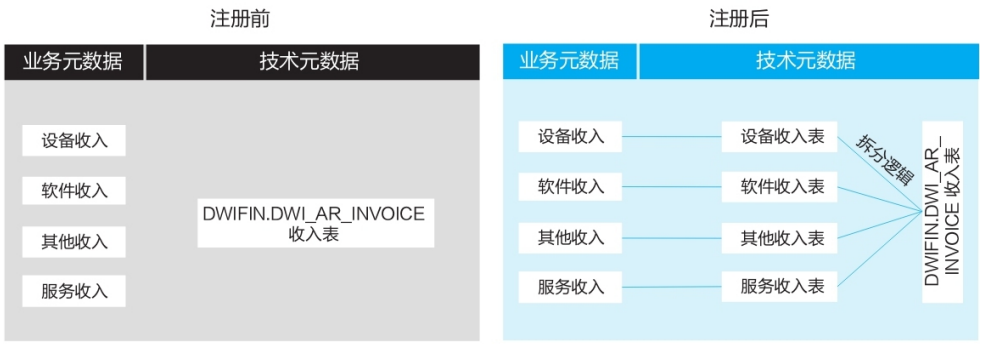


图3-20　元数据注册父子模式样例

4. 运维元数据

运维元数据是为了通过对元数据进行分析，发现数据注册、设计、使用的现状及问题，确保元数据的完整、准确。通过数据资产分析，了解各区域/领域的数据注册情况，进而发现数据在各信息系统使用过程中存在的问题。通过业务元数据与技术元数据的关联分析，反向校验架构设计与落地的实施情况，检查公司数据管理政策的执行情况。

主要分为如下四 个场景。

场景一：基于数据更新发现，数据源上游创建，下游更新；

场景二：通过数据调用次数发现，某数据源上游调用次数<下游调用次数；

场景三：虽制定了架构标准，但不知落地情况，比如某个属性建立了数据标准，但是却找不到对应落地的物理表字段；

场景四：通过物理表的字段分析，发现很多字段缺少数据标准。

3.6　本章小结

华为经过多年实践，已经建立了相对完整的数据分类管理框架，为数据治理奠定了基础。随着数字化转型的深入开展，尤其是面向未来海量的非结构化数据、IoT场景的观测数据、外部合规日 趋严格的外部数据等，华为将不断丰富 每一类数据的治理实践。

第4章

面向“业务交易”的信息架构建设

华为过去的信息架构建设主要是为了实现“信息化”或“业务上ERP”，信息架构往往隐藏在系统中、隐藏在IT功能下。对于大部分业务作业人员和管理者而言，他们的关注点更多聚焦在“功能是否完善”或“业务是在系统中完成还是手工完成”上。此时，对信息架构的要求仅限于支撑好各类IT系统的落地，或在一定范围内对IT建设提供指导。

随着企业数字化转型的推进，华为公司越来越认识到信息架构的价值并不应局限于“支撑IT建设落地”，而是更好地管理企业数据资产，更好地提升整个业务交易链条的效率，甚至基于信息架构重新审视业务边界的划分和整合。

4.1 信息架构的四个组件

企业在运作过程中，首先需要管理好人和物等“资源”，然后管理好各类资源之间的联系，即各类业务交易“事件”，再对各类事件的执行效果进行“整体描述和评估”，最终实现组织目标和价值。

以一个通用的工业企业运营为例（如图4-1所示），企业要管理关键的“员工、组织、产品、客户、供应商”等资源。在企业价值实现的过程中，企业会与客户签订销售合同，与供应商签订采购合同，组建各种交付项目，制定供应计划，财务部门会对成本、费用、收入进行核算，记录客户的应收、供应商的应付，建立合法合规的会计记账体系。然后，通过报告体系按月度、季度、年度发布各种经营、考核报告用于企业决策。

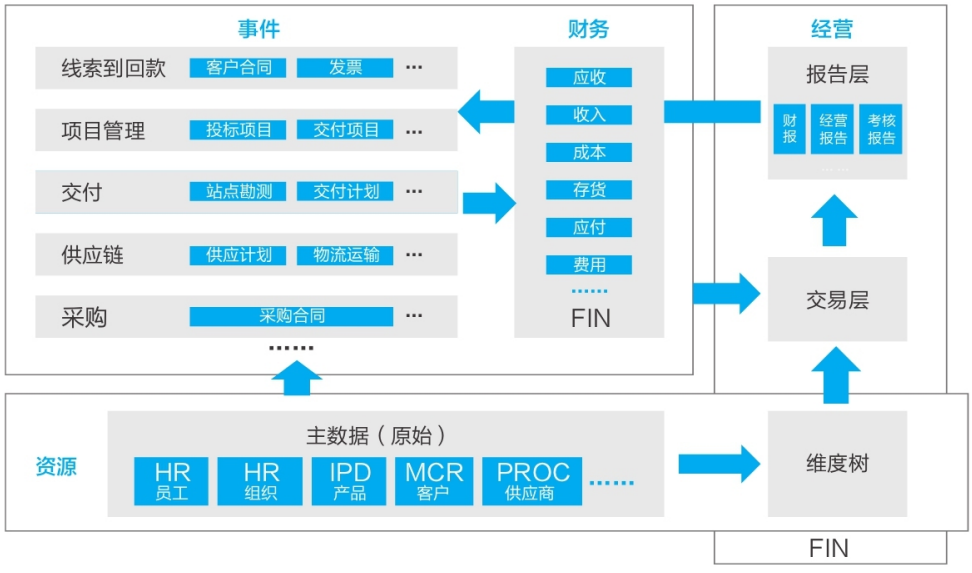


图4-1 信息架构视图示例

信息架构的目的就是定义好整个运作过程中涉及的各种人、事、物资源，并实施有效的治理，从而确保各类数据在企业各业务单元间高效、准确地传递，上下游流程快速地执行和运作。

华为在实践中构建了一套对业务运作数据进行有效管理的信息架构方法论，用于指导企业内部各部门的信息架构建设工作，让管理者、专家和员工之间有共同语言。

华为的企业级信息架构（Information Architecture）是指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，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，如图4-2所示。

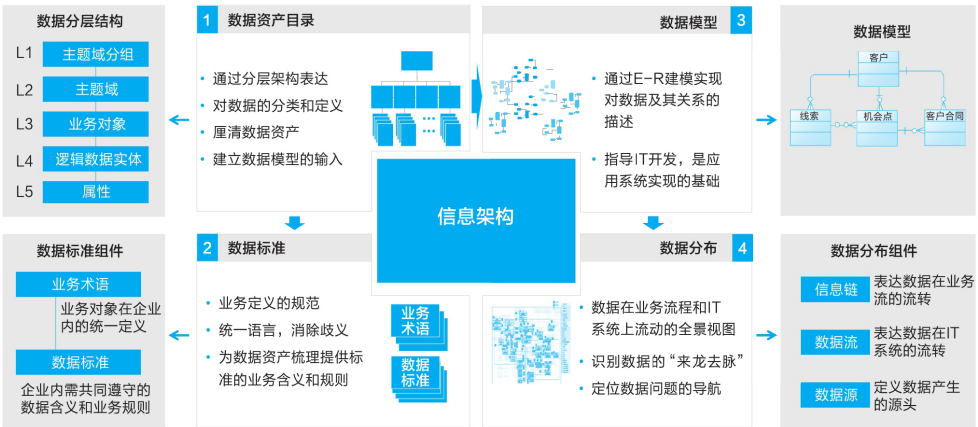


图4-2　华为企业级信息架构的四个组件

4.1.1　数据资产目录

数据资产目录形成完善的企业资产地图，也在一定程度上为企业数据治理、业务变革提供了指引。基于数据资产目录可以识别数据管理责任，解决数据问题争议，帮助企业更好地对业务变革进行规划设计，避免重复建设。

数据资产目录分为5层，涵盖华为公司的所有业务数据资产，如图4-3所示。

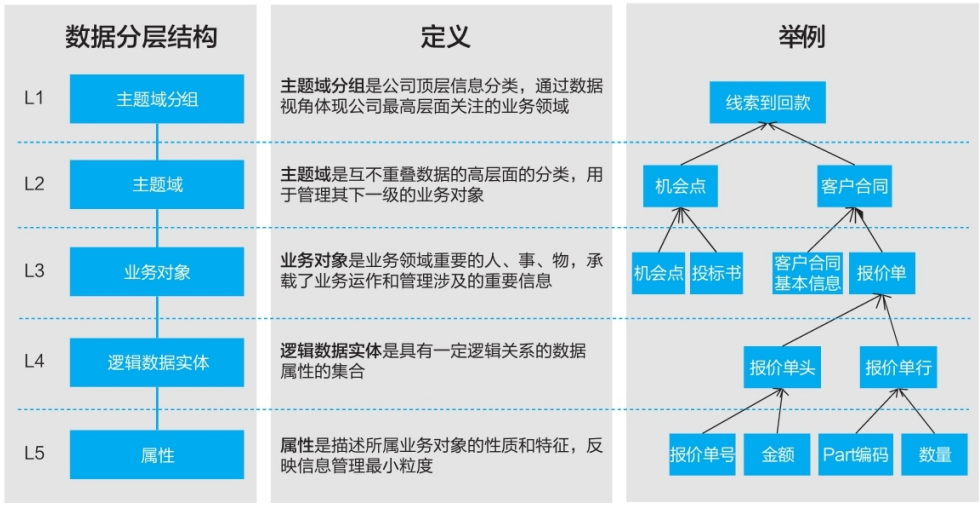


图4-3　数据资产的5层结构

L1为主题域分组，是描述公司数据管理的最高层级分类。业界通常有两种数据资产分类方式：基于数据自身特征边界进行分类和基于业务管理边界进行分类。华为公司为了强化企业内业务部门的数据管理责任，更好地推进数据资产建设、数据治理和数据消费建设，采用业务管理边界划分方式，即将L1主题域分组与流程架构L1相匹配，数据资产和华为业务GPO（全球流程责任人）相匹配，有利于更好地推进各项数据工作。

L2为主题域，是互不重叠的数据分类，管辖一组密切相关的业务对象，通常同一个主题域有相同的数据Owner。

L3为业务对象，是信息架构的核心层，用于定义业务领域重要的人、事、物，架构建设和治理主要围绕业务对象开展。同时，在企业架构（EA）的范畴内，信息架构（IA）也主要通过业务对象实现与业务架构（BA）、应用架构（AA）、技术架构（TA）的架构集成。

L4是逻辑数据实体，是指描述一个业务对象在某方面特征的一组属性集合。

L5为属性，是信息架构的最小颗粒，用于客观描述业务对象在某方面的性质和特征。

4.1.2 数据标准

数据标准是在企业范围内确保数据一致的关键，因此有必要多花一些篇幅来详细介绍。

数据标准定义公司层面需共同遵守的属性层数据含义和业务规则，是公司层面对某个数据的共同理解，这些理解一旦确定下来，就应作为企业层面的标准在企业内被共同遵守。

例如，合同是公司最重要的数据之一，因此有必要对合同编号制订统一的数据标准，包括编号的位数、具体的编码规则等，一旦合同编号数据标准制订下来，那么整个公司所有业务

部门都必须共同遵守，除了数据Owner以外，任何部门都不允许自定义合同编号。如果随着业务发展需要对合同编号进行变更，那么相关需求也应该统一由数据Owner受理，统一制订变更方案。一旦不同业务环节各自定义，那么数据就无法在上下游业务之间快速流转，往往需要额外的人工转换和翻译，这会极大地增加不必要的人工成本、延长业务执行周期、降低业务效率。

表4-1展示了华为在数据治理早期阶段，各个BG的分类定义不统一问题。

表4-1　数据定义不一致问题示例

主要系统	运营商BG	企业BG	消费者BG	其　他
产品设计	carrier network	enterprises	consumers	other
合同处理	operator	enterprise	consumer	other
订单处理	operator	enterprise	consumer	other
开票	CNBG	EBG	CBG	other BG
经营报告	carrier network	enterprises	consumers	other

以基础数据“运营商BG”为例，有的系统叫“carrier network”，有的系统叫“operator”，还有的系统叫“CNBG”，那么在统计“运营商BG”视角的财务报告时，数据处理人员需要将标识为“carrier network”“operator”“CNBG”的数据进行分类清洗和处理，而不能直接卷积得到结果，增加了额外的处理时间和投入。如果数据处理人员不清楚“运营商BG”这个数据在诸多系统中的名称不一致的情况，那

么很可能在统计时丢失部分数据，导致最终报告的失真，最终可能误导决策。

华为公司对业务数据标准有严格的限定，每个数据标准应该覆盖以下三方面。

- **业务视角要求：**用于统一业务侧语言和理解，明确定义每个属性所遵从的业务定义和用途、业务规则、同义词，并对名称进行统一定义，避免重复。
- **技术视角要求：**对IT实施形成必要的指引和约束，包括数据类型、长度，如果存在多个允许值，则应对每个允许值进行明确的限定。
- **管理视角要求：**明确各业务部门在贯彻数据标准管理方面应承担的责任，包括业务规则责任主体、数据维护责任主体、数据监控责任主体，因为很多情况下这些责任并不是由同一个业务部门来负责，所以必须在标准制订时就约定清楚。例如，“客户合同”中某些条款的规则制订者可能是财经部门，负责与客户达成约定并在系统中录入的可能是销售业务部门，而对整个客户合同数据质量进行跟踪、监控的可能是数据专业部门。

但是，企业的每个业务数据标准的定义和维护都需要一定的成本，很多大型企业的IT系统中可能存在上百万、上千万属性，即使去掉冗余、重复的部分，数据量也相当大，因此其实并不需要对IT系统内所有字段都进行定义。为了实现在统一定义的必要性和成本之间取得平衡，华为公司制订了数据标准规范，明确了在不同情况下哪些数据应该制订统一的标准。

- 描述业务对象的特有属性应作为本业务对象的属性进行定义，并明确业务数据标准。
- 引用其他业务对象的属性，如果属性值可随本业务对象确定和更改，就应作为本业务对象的属性进行定义，并明确业务数据标准。
- 引用其他业务对象的属性，如果属性值取自引用业务对象相应时点的数值且后续不变更，就应纳入本业务对象的数据标准范围，并明确相应取值规则。
- 引用其他业务对象的属性，如果属性值与引用业务对象同步，就不需要重新定义数据标准。
- 引用其他业务对象/逻辑数据实体的身份标识属性，应作为本业务对象的属性进行定义，但只能在业务数据标准中定义出处及引用规则，而不允许修改或重新定义该属性本身的业务含义及业务规则。

4.1.3 数据模型

数据模型是从数据视角对现实世界特征的模拟和抽象，根据业务需求抽取信息的主要特征，反映业务信息（对象）之间的关联关系。数据模型不仅能比较真实地模拟业务（场景），同时也是对重要业务模式和规则的固化。例如在某个物流业务数据模型中，“运输申付单”与“运输委托”建立一对一关系，而“运输委托”与“派送任务”建立多对多关系，那么这意味着业务部门可以根据发货效率和成本的考虑将“运输委

托”拆成分多个“派送任务”，但“派送任务”必须在将一个运输委托完整执行后，才能申请向供应商付款。

4.1.4 数据分布

如果说前三个组件主要是从静态角度对数据、数据关系进行定义，那么数据分布则定义了数据产生的源头及在各流程和IT系统间的流动情况。**数据分布组件的核心是数据源，指业务上首次正式发布某项数据的应用系统，并经过数据管理专业组织认证，作为企业范围内唯一数据源头被周边系统调用。**华为公司规定所有业务数据必须认证数据源，并在公司范围内统一发布。为了更好地识别、管理数据在流程和IT系统间的流动，可以通过信息链、数据流来进行描述，体现某一数据在流程或应用系统中是如何被创建（Create）、读取（Read）、更新（Update）、删除（Delete）的。

4.2 信息架构原则：建立企业层面的共同行为准则

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，这些原则不仅是对数据专业人员的要求，也是对业务的要求，因为业务才是真正的数据Owner。所以，公司所有业务部门都应该共同遵从信息架构原则。

华为首先确定了“数据同源一致”的治理目标，围绕目标的实现，制定了五条架构原则。各业务领域和变革项目应按照架构原则设计其信息架构，并由EAC（企业架构委员会）、IA-SAG（信息架构专家组）指导和监督各领域落实企业架构原则，在一套规则的约束下，共同建设一个企业级的信息架构。

原则一：数据按对象管理，明确数据Owner

数据要发挥作用，必然会在多个IT系统和流程中流转，并且越是重要的数据资产，所流经的业务环节就越多。例如，产品、人员、客户的数据几乎在所有流程中都会涉及，客户合同数据也会在整个业务交易链条中流转，因此不应该以IT系统、业务流程边界来管理数据，而应该从数据本身出发，按对象进行数据全生命周期管理。（更d书f享搜索雅 书.YabooK）

几乎所有的企业数据都是由业务产生的，IT人员无法对数据的定义、质量负责，因此需要在公司层面确定数据Owner。华

为公司按照业务对象任命数据Owner，并且每个数据都只能有唯一的数据Owner。数据Owner要负责所辖领域的信息架构建设和维护，负责保障所辖领域的数据质量，承接公司各个部门对本领域数据的需求，并有责任建立数据问题回溯和奖惩机制，对所辖领域的数据问题及争议进行裁决，公司有权对不遵从信息架构或存在严重数据质量问题的责任人进行问责。

原则二：从企业视角定义信息架构

任何一个数据Owner都不只代表自己所辖业务范围的数据管理诉求，而是代表公司对数据进行管理。华为在数据治理实践中，为了拉通各部门所产生的数据结构和流转路径，实现数据在企业内共享和流通的目标，明确要求各业务领域都需站在企业的视角定义信息架构，充分考虑数据的应用场景、范围和用户群体，参考业界实践和主流软件包，平衡和兼顾AS-IS（现状）和TO-BE（未来）诉求，在流程设计和IT实现中得到落实。

以前面的合同编号为例，销售部门作为数据Owner有责任定义合同信息架构，但不应只考虑销售环节对合同编号的管理诉求，而是应该综合考虑供应、交付、财经等各个环节对合同的诉求，合同在整个交易链条中延伸的范围就是相应数据Owner所综合覆盖的范围。在这个链条中，任何业务部门对合同编号的诉求，都可以提交给数据Owner；同时，合同数据Owner对所辖数据在整个企业范围内的架构的合理性和一致性负责，如果某个业务环节私自定义了合同信息架构，那么数据Owner有责任对该架构进行统一和整改。

原则三：遵从公司的数据分类管理框架

为了协同企业内各业务领域的数据治理，华为在实践中总结了各类数据的内在特性，制定了统一的数据分类管理框架，公司所有业务领域按照统一的分类框架进行数据治理。

原则四：业务对象结构化、数字化

华为在长期的数据治理过程中，制定了业务对象结构化、数字化的架构设计原则，实现数据处理效率的提升，构建数据的处理和应用能力，支撑业务管理。

业务对象内容包括业务结果、业务规则、业务过程，并应打造相应的数字化能力。

原则五：数据服务化，同源共享

随着企业业务规模的不断扩大，往往会随之产生大量的IT系统，这样很容易出现数据多源头的情况，导致数据不可信、不可管。为了有效地避免这些问题，华为制定了数据同源共享的架构原则，每一个数据有且只有单一数据源，数据使用方应从数据源获取数据，数据更改应在数据源进行。为了克服企业业务和IT的复杂性这一客观现实，华为公司持续推进数据服务建设，要求各数据Owner通过数据服务向各业务环节提供数据，各业务环节也有责任通过服务来合理获取数据，从而在整个企业层面实现数据的“一点定义、全局共享”。

4.3　信息架构建设核心要素：基于业务对象进行设计和落地

4.3.1　按业务对象进行架构设计

业务对象是指业务领域中重要的人、事、物对象。业务对象承载了业务运作和管理涉及的重要信息，是信息架构中最重要的管理要素。

业务对象同时还是业务和IT的关键连接点，也是实现IA（信息架构）、BA（业务架构）、AA（应用架构）、TA（技术架构）集成的关键要素。

以一个简化的交易场景为例（如图4-4所示），要完成一个交易，实现商业价值的兑现，企业内的某个子公司，需要与法人客户签订客户合同，在客户合同中，要明确交易的产品。在这个场景中，子公司、法人客户、客户合同、产品是企业需要管理和控制的核心对象，要作为业务对象进行管理。

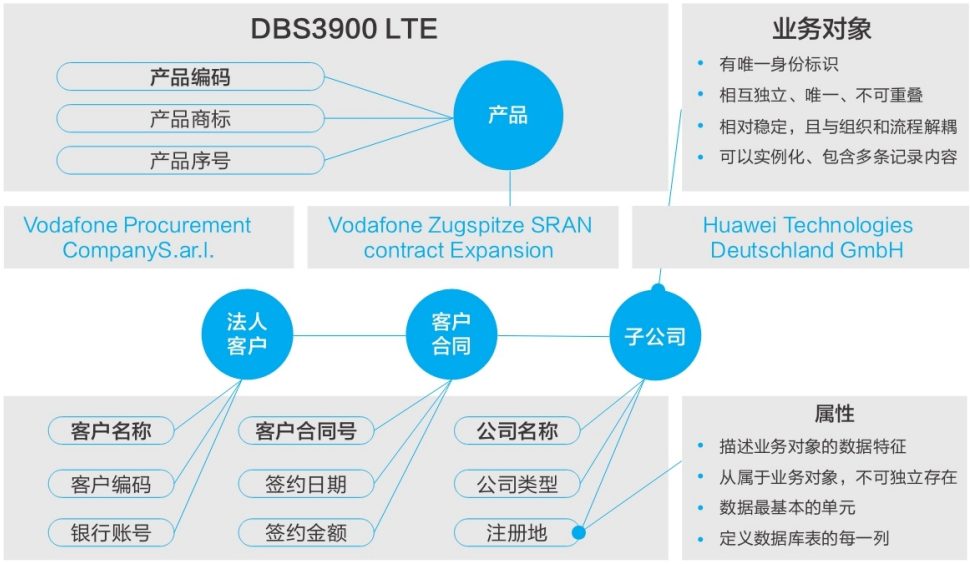


图4-4　业务对象和属性示例

在进行信息架构设计时，架构师、业务代表、数据Owner通常会对业务对象的判定存在理解上的偏差，从而产生争议。数据治理部门需要制定一套确定性的规则，通过确定性的规则促进形成稳定的架构。华为通过以下四条原则来判定业务对象。

原则一：业务对象是指企业运作和管理中不可缺少的重要人、事、物

企业在设计业务对象时，围绕支持企业运作和管理的重要的人、事、物去识别。通常，一个业务对象会有相应的管理流程、管理组织，以及支持运作的IT系统。比如“客户”这个对象，企业通常会建立类似客户管理部这样的组织，会采购或者开发CRM客户管理系统来支撑客户管理，会建立客户信息管理的一系列流程和规范来确保客户信息的准确、合理、合规。为了避免管理上的冲突，业务对象通常在企业内只能有一个唯一的

数据Owner，由数据Owner制定相关的架构、标准和管理规则，用于监控和提升数据质量。

原则二：业务对象有唯一身份标识信息

企业要对业务对象进行管理，需要对所有业务对象的实例进行编码，确保每个对象的实例在企业范围内都有唯一的标识。比如员工，企业需要为每个员工分配一个唯一的工号，如果工号出现重复，则可能引起管理上的混乱，比如工资错发，任务指令接收不到等。又比如产品，企业需要给每一种产品分配精确的分类编号，确保在研发组织内部、制造工厂、物流运输、销售回款各个部门和阶段，相同的产品使用唯一相同的编号，不同的产品绝不出现相同编号。企业的研发、生产、销售、核算各环节均采用产品的唯一编码进行标识和处理。

原则三：业务对象相对独立并有属性描述

业务对象需要通过大量属性来描述其各个方面的性质和特征，因此属性必定依附于某个业务对象而不可独立存在。比如“名称”是个属性，单纯地记录“名称”这个属性，无任何业务含义，因为“客户”有“名称”属性，“供应商”也有“名称”属性，“员工”也有“名称”属性。业务对象可以独立地存储、传输、使用，业务对象之间可以有关联、依赖关系，但不应有包含或从属关系。

以“销售订单”为例（如图4-5所示），“销售订单”通常包含两个方面的信息。一方面是销售订单中所销售产品的公共信息，比如归属的订单编号、订单名称、订单总价等，这类信息集中起来形成一个叫“订单头”的逻辑数据实体。另一方面是销售订单中某个产品的个性化信息，一个销售订单通常会销售多种产品，每种产品的价格和数量可能不一样，这些信息需要用另一个逻辑数据实体来记录，并用一个“订单编码”属性来表示这些明细的销售产品归属于该销售订单里，同时不同产品按不同“订单行号”展示。而“订单行号”是无法独立存在的，企业能够确保所有“订单编码”不会重复，但无法确保所有“订单行号”不会重复，并且这也没有必要，因为任何订单行都是隶属于某个订单的。因此从这个例子可以看出，订单行是无法作为一个独立的业务对象而存在的，必须归属于“销售订单”这个业务对象。

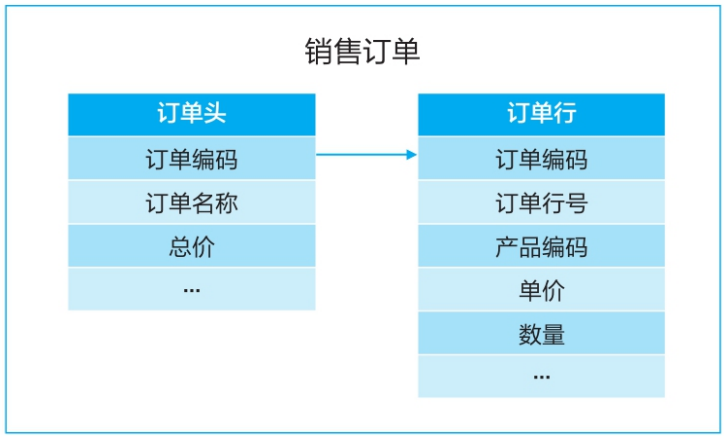


图4-5　业务对象示例：销售订单

原则四：业务对象可实例化

在现实世界中，业务对象有大量的实例存在，并可感知、可获取。以员工为例，就算是规模很小的企业，通常至少会有经理、业务员、会计等不同岗位的人员，每个员工的信息都可以视为一个实例；而“员工入职类型”是对员工入职信息的一种分类，其本身是无法实例化的，因此“员工入职类型”这一基础数据应从属于员工业务对象下，而不能独立存在，员工业务对象Owner也应该同时负责“员工入职类型”数据的生命周期管理。

4.3.2 按业务对象进行架构落地

信息架构向IT侧落地的主要交付件是数据模型。数据模型本身有相对比较成熟的方法体系支撑，不同企业之间可能名称存在差异，但本质差别不大。华为公司将数据模型分为三层：概念数据模型、逻辑数据模型、物理数据模型，如图4-6所示。

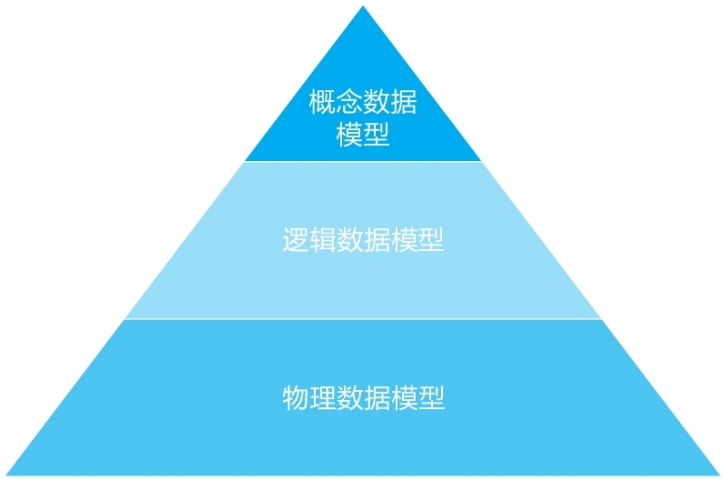


图4-6 数据模型分层框架

概念数据模型是通过业务对象及业务对象之间的关系，从宏观角度分析和设计的企业核心数据结构。逻辑数据模型是利用逻辑数据实体及实体之间的关系，准确描述业务规则的逻辑实体关系。物理数据模型是按照一定规则和方法，将逻辑数据模型中定义的逻辑数据实体、属性、属性约束、关系等内容，如实转换为数据库软件能识别的物理数据实体关系。

为了确保架构在落地过程中“不走形”，要控制好两个关键点：一个是概念模型与逻辑模型的一致性，主要通过逻辑数据实体的设计管理来实现；另一个是逻辑模型与物理模型的一致性，主要通过一体化建模管理来实现。

1. 逻辑数据实体设计

逻辑数据实体本质上是对描述业务对象的众多属性的归类，业务对象无法直接指导IT系统的物理实现，也无法基于业务对象来审视物理设计是否满足业务需求，因此需要通过逻辑数据实体及相应的逻辑数据模型来指导IT系统层面的数据设计。在设计逻辑数据实体时，可参考如下几条主要规则。

1）逻辑数据实体不能脱离业务对象独立存在，因此某个逻辑数据实体一定是用来描述一个特定的业务对象的，业务对象与逻辑数据实体的关系是一对一或一对多，不允许多对一的情况出现。

2）描述业务对象不同业务特征的密切相关的一组属性集合，可以设计为一个逻辑数据实体。

3）逻辑数据实体设计要遵循第三范式。在设计一个业务对象的逻辑数据实体时，每个逻辑数据实体的属性不要重复定义，不应包含其他逻辑数据实体中的非关键字类型的属性。

4）提供数据服务或跨业务领域使用的基础数据，要单独设计逻辑数据实体。描述业务对象的若干属性，如果能够组合起来形成独特价值的数据服务，满足下游的数据消费需求，可以设计成一个逻辑数据实体。

5）两个业务对象间的关系也可以设计成关系型逻辑数据实体，在数据资产目录中，可按业务发生的时间先后顺序，归属于后出现的业务对象。

2. 一体化建模管理

华为公司过去长期存在信息架构与IT开发实施“两张皮”的现象，数据人员和IT开发实施人员缺乏统一和协同，数据架构遵从无法进行实质、有效管理，信息架构资产和产品实现的物理表割裂、不匹配，同时各种数据模型资产缺失。

- 针对应用系统设计应遵从信息架构设计的政策要求，在相关项目、产品的流程中，缺乏显性化的且有实操指导的角色和活动。
- 信息架构设计大多集中在变革项目层的设计输出和领域层的例行刷新，未与系统落地有效拉通。
- IT产品聚焦在版本交付，产品级的数据模型与数据字典缺少有效看护和及时维护。

为了解决这个问题，华为推行了一体化模型设计（如图4-7所示），不仅在工具上实现了一体化设计和开发，而且在机制上形成了信息架构设计与IT开发实施的有效协同。通过一体化设计不仅确保了元数据验证、发布和注册的一致性，而且实现了产品数据模型管理和资产可视。同时，由于促成了产品元数据的持续运营，进而能够持续对物理模型进行规范，如整改、清理各类作废表等。

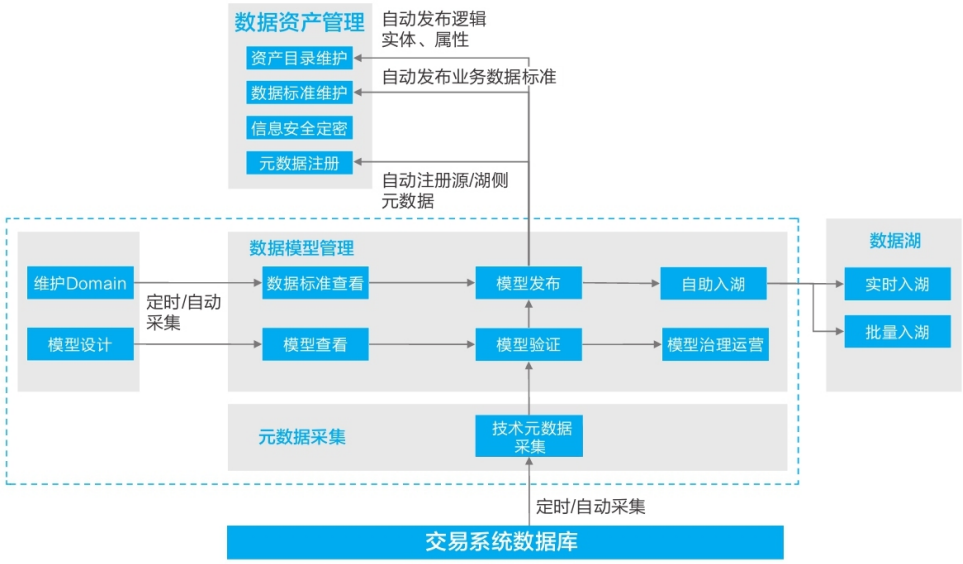


图4-7　一体化建模架构

基于良 好的一体化建模架构，不仅可以打通设计和物理实现，而且能够对设计、发布、管理运营等完整生命周期进行融合管理，包括：

- 产品逻辑模型和物理模型一体化设计，元数据管理和数据模型管理融合；
- 构建数据标准池，实体属性只能从数据标准池选择；
- 产品元数据和数据库自动比对和验证；

- 产品元数据发布认证和信息资产打通；
- 基于交易侧产品元数据自助入湖。

4.4　传统信息架构向业务数字化扩展：对象、过程、规则

本章前面提到，华为公司的信息架构最初是为了满足“信息化”和“业务上ERP”过程中提出的数据管理要求，但随着数字化转型的深入，发现既有信息架构已经无法满足自身业务需要，主要体现在以下几个方面。

1）大量业务和作业所产生的数据并没有完整地被管理

很多情况下，并不是所有业务和作业所产生的数据都在系统中承载，因为大量IT系统中已经承载的数据，往往都是为了满足流程的标准化需要而存在的。例如，每个与客户签订的合同都非常复杂，包括诸多的条款，华为公司签订的合同通常会有上百页内容，但信息架构往往只定义了数百个数据属性，IT系统中也只承载了这部分内容，而大量的数据是以文档的形式存在的。当要对某个合同进行签订前的评审时，如果想基于过去已签订的合同的条款进行基线参考和验证，那么是无法通过自动化手段高效实现的，只能通过人工翻阅历史合同来实现，完整性和准确性都无法保障，而且效率很低。

2）大量业务过程没有形成可视、可管理的数据

业务在执行某个具体活动时是有大量作业过程的，但这部分数据往往并没有得到管理。例如，过去只记录了物流各个节

点的实际到达信息，但缺乏过程信息记录，假如想实时了解具体的物流状态，只能通过电话、邮件一次次询问，增加大量沟通成本，并且信息的及时性也得不到保障。

3）大量业务规则缺乏管理、无法灵活使用

在业务执行中存在大量规则，但绝大部分规则都缺乏有效管理，往往只能通过文件和文档管理，即使有部分规则固化到了IT系统中，也是无法灵活调整的。例如，有业务人员经常抱怨，由于每年都会发布一些文件来制订业务规范，因此自己不知道哪些是最新的，以及多个历史规范之间是否有重叠和矛盾；另外，如果想基于业务变化对规则进行刷新，但这些规则都固定在IT代码中，IT系统动辄需要数个月才能完成修改，而此时业务可能又发生了新的变化。

因此，华为公司在传统信息架构的基础上，提出了面向数字化转型的扩展：对象数字化、过程数字化、规则数字化，并打造与之相应的能力。

1）对象数字化

对象数字化的目标是建立对象本体在数字世界的映射。这种映射不是传统意义上基于流程要求的少量数据的管理，而是管理某个对象的全量数据，如图4-8所示。

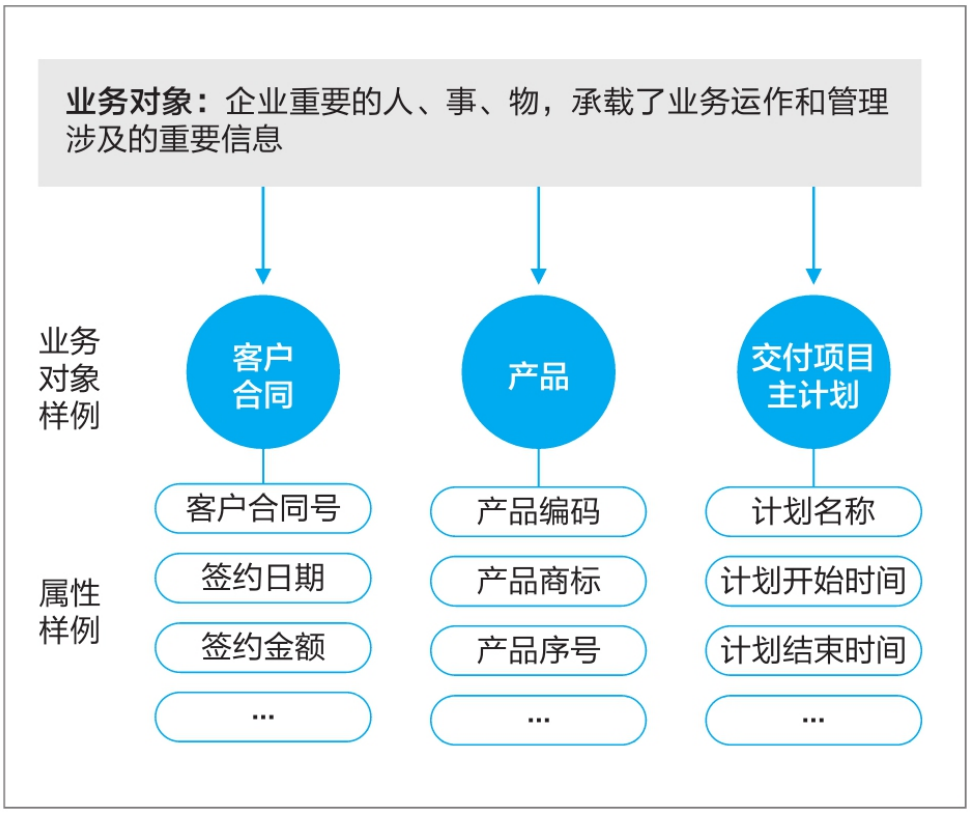


图4-8　业务对象数字化

以产品研发和设计为例，信息架构过去只管理产品数据进入ERP管道所必需的少量内容，如产品编码、描述、BOM清单等，而基于对象数字化则需要建立完整的数字孪生（Digital Twin），也要管理与之相应的完整信息架构。过去，供应部门经常抱怨产品研发部门所提供的“重量”“体积”信息不准确，而研发部门又没有足够的人力在产品进入生产环节前精准测量每个产品、部件、元器件。但是，实际上研发在设计过程中会多次产生并使用这些重量和体积信息，因为这对研发设计也同样重要。在推行对象数字化后，就可以通过数据感知等手段在设计的各个环节记录上述这些数据，并按项目编码进行更新，这样就可以向供应环节提供准确并且全量的数据。

2）过程数字化

仅仅管理好结果还不够，有时我们需要把作业过程记录下来，了解过程进度或者反过来改进结果。这种记录首先是不干预业务活动的，并且能够自动记录（例如，车辆行驶中自动监控是否存在交通违规）。

过程数字化要实现业务活动线上化，并记录业务活动的执行或操作轨迹，一般通过观测数据来实现轨迹记录，如图4-9所示。



图4-9　业务过程数字化

以前面举过的物流场景为例，华为公司通过推进业务过程数字化，实现供应链对各类物流状态的实时感知和可视，大幅缩减了发货后反复人工沟通的成本。

3）规则数字化

规则数字化的目的是把复杂场景下的复杂规则用数字化手段进行管理。良 好的规则数字化管理，应该能实现业务规则与IT应用解耦，所有关键业务规则数据要实现可配置，能够根据业务的变化灵活调整，如图4-10所示。

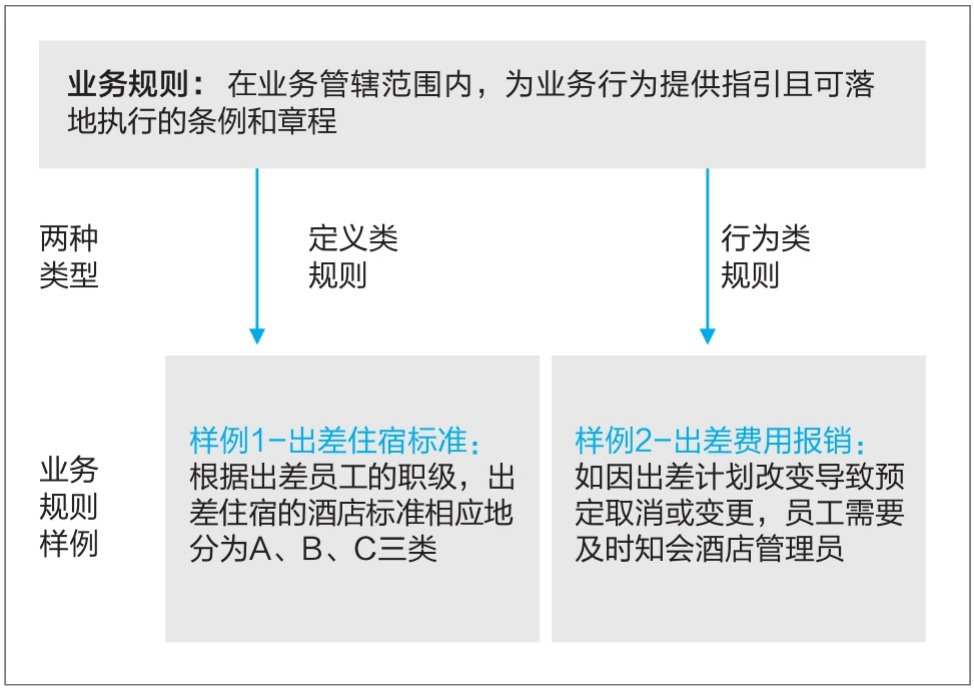


图4-10　业务规则数字化

同样以物流场景为例，通常业务希望基于计划对各个环节的物流任务进行监控和预警，这需要大量的预警规则。例如，某个部件的物流周期是1周，当5天后要交付而对应物流还未发货，则应该预警。但是，不同物料、不同场景、不同国 家的供应能力往往是有差异的，并且随着环境经常动态变化，这就需要将对应的规则数据从IT应用中解耦出来，单独定义这类数据资产的信息架构，从而使之能够灵活调整。这样，不同国 家的

业务人员就可以根据需要随时调整规则，而不用对现有IT系统进行大的改动，最大程度地满足业务灵活性的要求。

4.5 本章小结

在企业数字化转型过程中，企业信息架构的定位、内涵和管理方式都在不断地演进。信息架构的定位发生了根本性的变化，不再是对准IT功能或实现，而是对准整个企业的业务管理目标；信息架构的内涵也进行了极大的扩展，不再只是聚焦于进入类似ERP系统的结构化数据，而是对准整个企业在业务中产生的各种结构化数据、非结构化数据、内外部数据、过程类数据、规则类数据、IoT数据等；信息架构的管理方式也发生了颠覆性的改变，不再是抽象化的、预先定义好的、一次定义覆盖所有场景的标准，而是全量的、实时产生的、满足差异化要求的，甚至是随需定义的标准。

在这样的背景下，需要对原有信息架构框架和方法论不断进行审视和优化，可能两年前刚刚确定的框架已经不能满足要求，甚至一年前发布的架构规则就要重新修订。在企业实现数字化转型的过程中，信息架构管理的结构、技术、组件、标准可能永远不会稳定，永远在进化。

第5章

面向“联接共享”的数据底座建设

在从信息化向数字化转型的过程中，企业积累了海量的数据，并且还在爆发式地增长。数据很多，但真正能产生价值的数据却很少。数据普遍存在分散、不拉通的问题，缺乏统一的定义和架构，找到想要的、能用的数据越来越难。

本章将讲述华为数据底座的总体架构和建设策略，详细说明华为如何通过数据湖和数据主题联接的建设，实现数据的汇聚和联接，打破数据孤岛和垄断，重建数据获取方式和次序。数据底座在华为数字化转型中起着关键作用。

5.1　支撑非数字原生企业数字化转型的数据底座建设框架

华为通过建设数据底座，将公司内外部的数据汇聚在一起，对数据进行重新组织和联接，让数据有清晰的定义和统一的结构，并在尊重数据安全与隐私的前提下，让数据更易获取，最终打破数据孤岛和垄断。通过数据底座，主要可以实现如下目标。

1）统一管理结构化、非结构化数据。将数据视为资产，能够追溯数据的产生者、业务源头以及数据的需求方和消费者等。

2）打通数据供应通道，为数据消费提供丰富的数据原材料、半成品以及成品，满足公司自助分析、数字化运营等不同场景的数据消费需求。

3）确保公司数据完整、一致、共享。监控数据全链路下的各个环节的数据情况，从底层数据存储的角度，诊断数据冗余、重复以及“僵尸”问题，降低数据维护和使用成本。

4）保障数据安全可控。基于数据安全管理策略，利用数据权限控制，通过数据服务封装等技术手段，实现对涉密数据和隐私数据的合法、合规地消费。

5.1.1 数据底座的总体架构

华为数据底座由数据湖、数据主题联接两层组成，将公司内外部的数据汇聚到一起，并对数据进行重新的组织和联接，为业务可视化、分析、决策等提供数据服务，如图5-1所示。



图5-1 华为数据底座总体架构

数据湖是逻辑上各种原始数据的集合，除了“原始”这一特征外，还具有“海量”和“多样”（包含结构化、非结构化数据）的特征。数据湖保留数据的原格式，原则上不对数据进行清洗、加工，但对于数据资产多源异构的场景需要整合处理，并进行数据资产注册。

数据入湖必须要遵循6项标准，共同满足数据联接和用户数据消费需求。

数据主题联接是对数据湖的数据按业务流/事件、对象/主体进行联接和规则计算等处理，形成面向数据消费的主题数据，具有多角度、多层次、多粒度等特征，支撑业务分析、决策与执行。基于不同的数据消费诉求，主要有多维模型、图模型、指标、标签、算法模型5种数据联接方式。

5.1.2 数据底座的建设策略

数据底座建设不能一蹴而就，要从业务出发，因势利导，持续进行。具体来说，**华为数据底座采取“统筹推动、以用促建、急用先行”的建设策略，**根据公司数字化运营的需要，由公司数据管理部统一规划，各领域分别建设，以满足本领域和跨领域的数据需求。其中，数据Owner是各领域数据底座建设的第一责任人，各领域数据部负责执行。数据底座资产建设遵从下面四 项原则。

1）数据安全原则

数据底座数据资产应遵循用户权限、数据密级、隐私级别等管理要求，以确保数据在存储、传输、消费等全过程中的数据安全。技术手段包括但不限于授权管理、权限控制、数据加密、数据脱敏。

2）需求、规划双轮驱动原则

数据底座数据资产基于业务规划和需求触发双驱动的原则进行建设，对核心数据资产优先建设。

3）数据供应多场景原则

数据底座资产供应需根据业务需求提供离线/实时、物理/虚拟等不同的数据供应通道，满足不同的数据消费场景。

4）信息架构遵从原则

数据底座数据资产应遵从公司的信息架构，必须经IA-SAG（信息架构专家组）发布并完成注册。

5.2　数据湖：实现企业数据的“逻辑汇聚”

5.2.1　华为数据湖的3个特点

华为数据湖（如图5-2所示）是逻辑上对内外部的结构化、非结构化的原始数据的逻辑汇聚。数据入湖要遵从6项入湖标准，基于6项标准保证入湖的质量，同时面向不同的消费场景提供两种入湖方式，满足数据消费的要求。经过近两年的数据湖建设，目前已经完成1.2万个逻辑数据实体、28万个业务属性的入湖，同时数据入湖在华为公司也形成了标准的流程规范，每个数据资产都要入湖成为数据工作的重要标准。

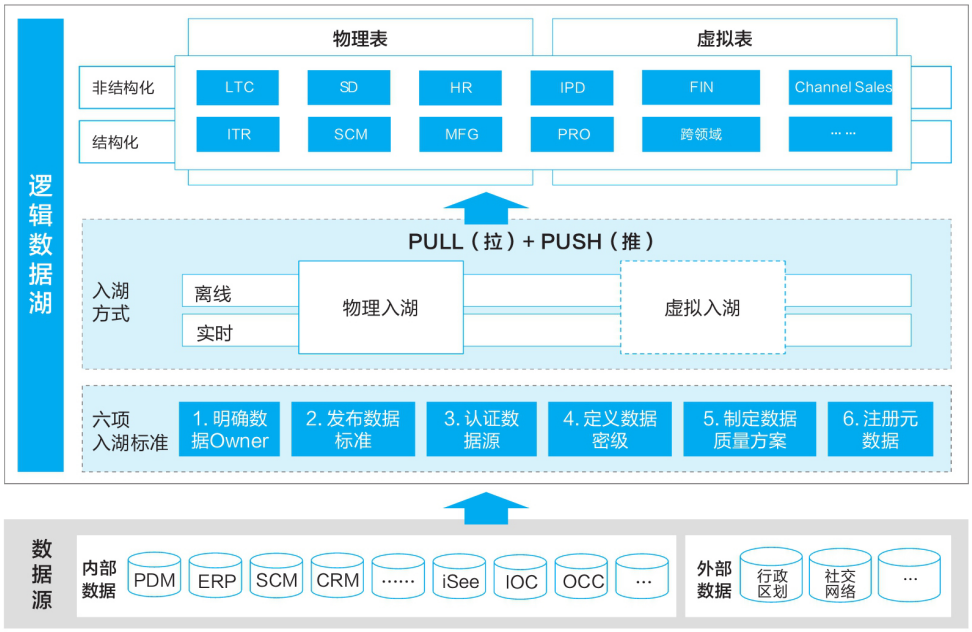


图5-2　数据湖总体视图

华为数据湖主要有以下几个特点。

1）逻辑统一

华为数据湖不是一个单一的物理存储，而是根据数据类型、业务区域等由多个不同的物理存储构成，并通过统一的元数据语义层进行定义、拉通和管理。

2）类型多样

数据湖存放所有不同类型的数据，包括企业内部IT系统产生的结构化数据、业务交易和内部管理的非结构化的文本数据、公司内部园区各种传感器检测到的设备运行数据，以及外部的媒体数据等。

3）原始记录

华为数据湖是对原始数据的汇聚，不对数据做任何的转换、清洗、加工等处理，保留数据最原始特征，为数据的加工和消费提供丰富的可能。

5.2.2　数据入湖的6个标准

数据入湖是数据消费的基础，需要严格满足入湖的6项标准，包括明确数据Owner、发布数据标准、定义数据密级、明确数据源、数据质量评估、元数据注册。通过这6项标准保证入湖的数据都有明确的业务责任人，各项数据都可理解，同时都能在相应的信息安全保障下进行消费。

（1）明确数据Owner

数据Owner由数据产生对应的流程Owner担任，是所辖数据端到端管理的责任人，负责对入湖的数据定义数据标准和密级，承接数据消费中的数据质量问题，并制定数据管理工作路标，持续提升数据质量。

（2）发布数据标准

入湖数据要有相应的业务数据标准。业务数据标准描述公司层面需共同遵守的“属性层”数据的含义和业务规则，是公司层面对某个数据的共同理解，这些理解一旦明确并发布，就需要作为标准在企业内被共同遵守。数据标准的信息如表5-1所示。

表5-1　数据标准说明

数据标准内容		说　　明
数据资产目录	主题域分组	公司顶层数据分类，通过数据视角体现最高层面关注的业务领域
	主题域	互不重叠数据的高层面分类，用于管理下一级的业务对象
	业务对象	业务领域的重要人、事、物，承载了业务运作和管理涉及的重要信息
	逻辑数据实体	具有一定逻辑关系的业务属性集合
	业务属性	描述所属业务对象的性质和特征，反映信息管理的最小粒度
定义及规则	引用的数据标准	说明该业务属性是否引用已定义的数据标准
	业务定义	对业务属性的定义，解释业务属性是什么，对业务的作用
	业务规则	业务属性的业务规则，包括但不限于业务属性在各场景下的变化规则和编码含义等
	数据类型	业务定义的数据类型，例如文本、日期、数字等
	数据长度	业务定义的数据长度
	允许值	业务属性对应的允许值清单
	数据示例	属性实例化的样例，用以帮助其他人员理解此业务属性
定义及规则	同义词	业务对于同一属性可能有不同的称呼，在此列出业务对此属性的其他称呼
	标准应用范围	业务数据标准在全公司范围、领域或区域范围内遵从
责任主体	业务规则责任主体	业务规则制定的责任部门
	数据维护责任主体	数据维护的责任部门
	数据质量监控责任主体	数据质量监控责任部门

（3）认证数据源

通过认证数据源，能够确保数据从正确的数据源头入湖。认证数据源应遵循公司数据源管理的要求，一般数据源是指业务上首次正式发布某项数据的应用系统，并经过数据管理专业组织认证。认证过的数据源作为唯一数据源头被数据湖调用。当承载数据源的应用系统出现合并、分拆、下线情况时，应及时对数据源进行失效处理，并启动新数据源认证。

（4）定义数据密级

定义数据密级是数据入湖的必要条件，为了确保数据湖中的数据能充分地共享，同时又不发生信息安全问题，入湖的数据必须要定密。数据定密的责任主体是数据Owner，数据管家有责任审视入湖数据密级的完整性，并推动、协调数据定密工作。数据定级密度在属性层级，根据资产的重要程度，定义不同等级。不同密级的数据有相应的数据消费要求，为了促进公司数据的消费，数据湖中的数据有相应的降密机制，到降密期或满足降密条件的数据应及时降密，并刷新密级信息。

（5）数据质量评估

数据质量是数据消费结果的保证，数据入湖不需要对数据进行清洗，但需要对数据质量进行评估，让数据的消费人员了解数据的质量情况，并了解消费该数据的质量风险。同时数据Owner和数据管家可以根据数据质量评估的情况，推动源头数据质量的提升，满足数据质量的消费要求。

（6）元数据注册

元数据注册是指将入湖数据的业务元数据和技术元数据进行关联，包括逻辑实体与物理表的对应关系，以及业务属性和表字段的对应关系。通过联接业务元数据和技术元数据的关系，能够支撑数据消费人员通过业务语义快速地搜索到数据湖中的数据，降低数据湖中数据消费的门槛，能让更多的业务分析人员理解和消费数据。

5.2.3 数据入湖方式

数据入湖遵循华为信息架构，以逻辑数据实体为粒度入湖，逻辑数据实体在首次入湖时应该考虑信息的完整性。原则上，一个逻辑数据实体的所有属性应该一次性进湖，避免一个逻辑实体多次入湖，增加入湖工作量。

数据入湖的方式主要有物理入湖和虚拟入湖两种，根据数据消费的场景和需求，一个逻辑实体可以有不同的入湖方式。两种入湖方式相互协同，共同满足数据联接和用户数据消费的需求，数据管家有责任根据消费场景的不同，提供相应方式的入湖数据。

物理入湖是指将原始数据复制到数据湖中，包括批量处理、数据复制同步、消息和流集成等方式。虚拟入湖是指原始数据不在数据湖中进行物理存储，而是通过建立对应虚拟表的集成方式实现入湖，实时性强，一般面向小数据量应用，大批量的数据操作可能会影响源系统。

数据入湖有以下5种主要技术手段。

- 批量集成（Bulk/Batch Data Movement）

对于需要进行复杂数据清理和转换且数据量较大的场景，批量集成是首选。通常，调度作业每小时或每天执行，主要包含ETL、ELT和FTP等工具。批量集成不适合低数据延迟和高灵活性的场景。

- 数据复制同步（Data Replication/Data Synchronization）

适用于需要高可用性和对数据源影响小的场景。使用基于日志的CDC捕获数据变更，实时获取数据。数据复制同步不适合处理各种数据结构以及需要清理和转换复杂数据的场景。

- 消息集成（Message-Oriented Movement of Data）

通常通过API捕获或提取数据，适用于处理不同数据结构以及需要高可靠性和复杂转换的场景。尤其对于许多遗留系统、ERP和SaaS来说，消息集成是唯一的选择。消息集成不适合处理大量数据的场景。

- 流集成（Stream Data Integration）

主要关注流数据的采集和处理，满足数据实时集成需求，处理每秒数万甚至数十万个事件流，有时甚至数以百万计的事件流。流集成不适合需要复杂数据清理和转换的场景。

- 数据虚拟化（Data Virtualization）

对于需要低数据延迟、高灵活性和临时模式（不断变化下的模式）的消费场景，数据虚拟化是一个很好的选择。在数据虚拟化的基础上，通过共享数据访问层，分离数据源和数据湖，减少数据源变更带来的影响，同时支持数据实时消费。数据虚拟化不适合需要处理大量数据的场景。

5种数据入湖方式的对比可以参考表5-2。

表5-2　数据入湖方式对比

说明				数据搬家	实时性	源系统性能要求	批量数据处理	历史数据处理
物理入湖	批量集成	ETL/ELT工具	拉	需要	非实时	低	支持（强）	支持（强）
		FTP 工具	推	需要	非实时	低	通常不支持	通常不支持
	数据复制同步	CDC 工具	拉	需要	实时	中	通常不支持	通常不支持
	消息集成	MQ 工具	推	需要	实时	中	通常不支持	通常不支持
	流集成	Pipeline工具	推	需要	实时	中	通常不支持	通常不支持
虚拟入湖	数据虚拟化	虚拟化工具	拉	不需要	实时	高	支持（弱）	支持（弱）

可以通过数据湖主动从数据源PULL（拉）的方式入湖，也可以通过数据源主动向数据湖PUSH（推）的方式入湖。数据复制同步、数据虚拟化以及传统ETL批量集成都属于数据湖主动拉的方式；流集成、消息集成属于数据源主动推送的方式（如表5-3所示）。在特定的批量集成场景下，数据会以CSV、XML等格式，通过FTP推送给数据湖。

表5-3　PULL（拉）& PUSH（推）方式入湖

入湖方式	数据源	数据湖
PULL（拉）	被动：当被请求时提供数据	主动：决定何时获取数据
PUSH（推）	主动：按自己节奏提供数据	被动：响应接收数据

5.2.4　结构化数据入湖

结构化数据是指由二维表结构来逻辑表达和实现的数据，严格遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理。

触发结构化数据入湖的场景有两种：第一，企业数据管理组织基于业务需求主动规划和统筹；第二，响应数据消费方的需求。

结构化数据入湖过程包括：数据入湖需求分析及管理、检查数据入湖条件和评估入湖标准、实施数据入湖、注册元数据（如图5-3所示）。

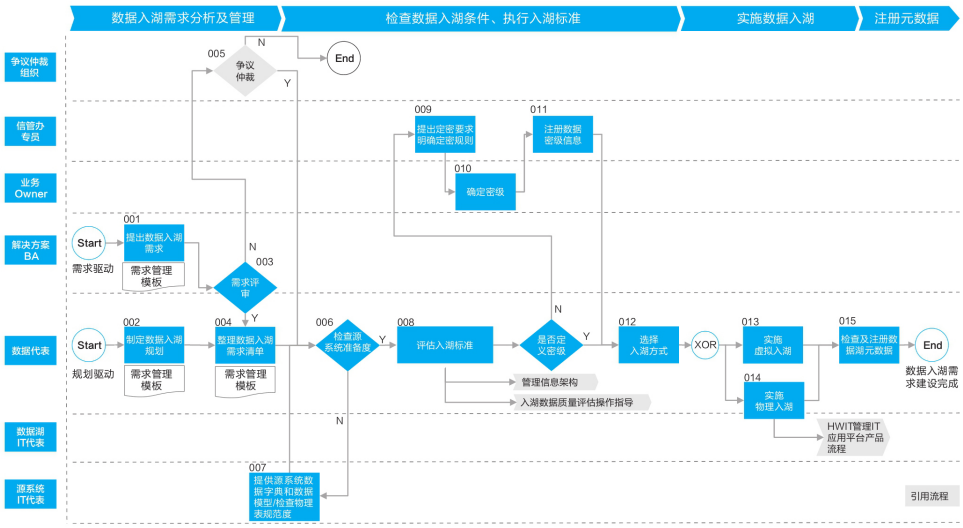


图5-3　结构化数据入湖流程

1. 数据入湖需求分析及管理

对于规划驱动入湖场景而言，由对应的数据代表基于数据湖的建设规划，输出入湖规划清单，清单包含主题域分组、主题域、业务对象、逻辑实体、业务属性、源系统物理表和物理字段等信息。

对于需求驱动入湖场景而言，由数据消费方的业务代表提出入湖需求，并提供数据需求的业务元数据和技术元数据的信息，包括业务对象、逻辑实体、业务属性对应界面的截图。

无论是主动规划还是被动响应需求，入湖需求清单必须通过业务代表和数据代表的联合评审。当业务代表和数据代表就评审结论发生争议时，可到专业评审组织申请仲裁。

2. 检查数据入湖条件和评估入湖标准

在数据入湖前要检查数据源准备度和评估数据入湖标准。

（1）检查数据源准备度

数据有源是数据入湖的基本前提，数据源准备度检查不仅需要源系统的IT团队提供源系统的数据字典和数据模型并检查源系统的物理表规范度，而且需要数据代表评估源系统的数据质量。

（2）评估入湖标准

入湖标准包括以下几点。

- **明确数据Owner：**为保证入湖数据的管理责任清晰，在数据入湖前应明确数据Owner。
- **发布数据标准：**入湖数据应有数据标准，数据标准定义了数据属性的业务含义、业务规则等，是正确理解和使用数据的重要依据，也是业务元数据的重要组成部分。
- **认证数据源：**原则上以初始源进湖，数据源认证是保证数据湖数据一致性和唯一性的重要措施。
- **定义数据密级：**定义完整、明确的数据密级是数据湖数据共享、权限控制等的关键依据。信息安全管理专员向业务Owner提出定密需求，并与业务Owner确定定密规则，确定数据密级、定密时间、降密期/降密条件等，然后由信息安全管理专员在信息架构管理平台注册密级信息。
- **评估入湖数据质量：**对入湖数据做质量评估，给入湖数据打质量标签。

如果不满足上述任意一条入湖标准，就应推动源系统数据代表完成整改，满足要求后方可实施数据入湖。

3. 实施数据入湖

数据代表依据消费场景合理选择入湖方式，在不要求历史数据、小批量数据且实时性要求高的场景，建议虚拟入湖；在要求历史数据、大批量数据且实时性要求不高的场景，可以物理入湖。

虚拟入湖由数据代表实施，数据代表负责设计和部署虚拟表。

物理入湖由对应数据湖的IT代表承接IT实施需求，设计集成方案和数据质量监测方案，实施数据入湖。数据代表组织UAT测试、上线验证。

4. 注册元数据

元数据是公司的重要资产，是数据共享和消费的前提，为数据导航和数据地图建设提供关键输入。对元数据进行有效注册是实现上述目的的前提。

虚拟表部署完成后或IT实施完成后，由数据代表检查并注册元数据，元数据注册应遵循企业元数据注册规范。

5.2.5　非结构化数据入湖

1. 非结构化数据管理的范围

非结构化数据包括无格式的文本、各类格式的文档、图像、音频、视频等多样异构的格式文件。相较于结构化数据，非结构化数据更难以标准化和理解，因而非结构化数据的管理不仅包括文件本身，而且包括对文件的描述属性，也就是非结构化的元数据信息。

这些元数据信息包括文件对象的标题、格式、Owner等基本特征，还包括对数据内容的客观理解信息，如标签、相似性检索、相似性连接等。这些元数据信息便于用户对非结构化数据进行搜索和消费。非结构化数据的元数据实体如图5-4所示。

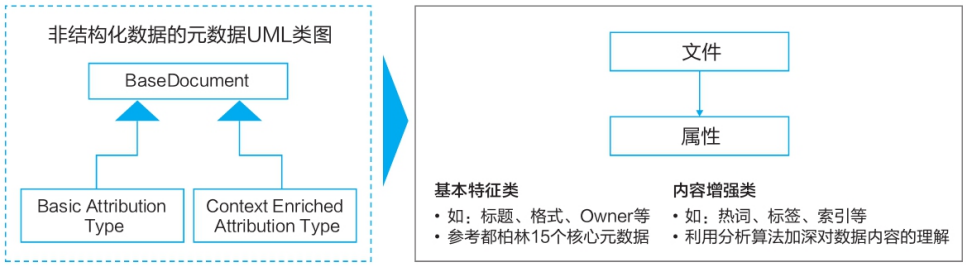


图5-4　非结构化数据的元数据实体

都柏林核心元数据是一个致力于规范Web资源体系结构的国际性元数据解决方案，它定义了一个所有Web资源都应遵循的通用核心标准。

基本特征类属性由公司进行统一管理，内容增强类属性由承担数据分析工作的项目组自行设计，但其分析结果都应由公司元数据管理平台自动采集后进行统一存储。

2. 非结构化数据入湖的4种方式

非结构化数据入湖包括基本特征元数据入湖、文件解析内容入湖、文件关系入湖和原始文件入湖4种方式，其中基本特征元数据入湖是必选内容，后面三项内容可以根据分析诉求选择性入湖和延后入湖，如图5-5所示。

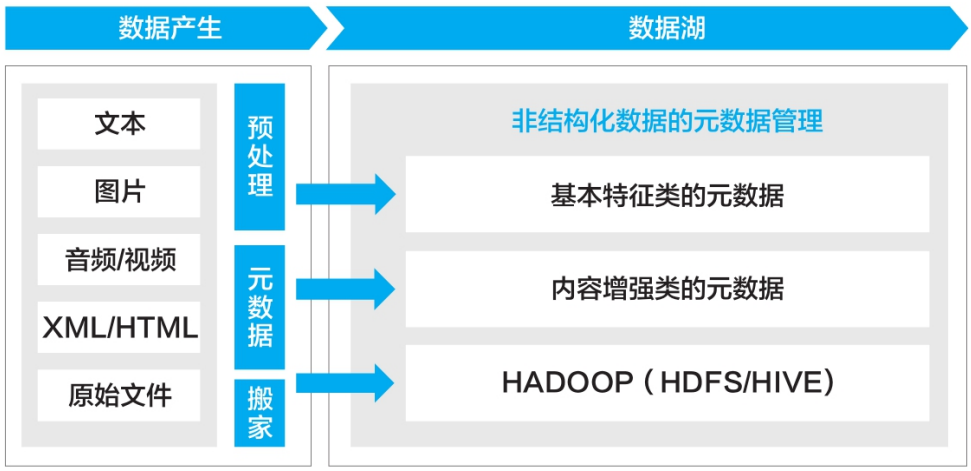


图5-5　非结构化数据入湖

1）基本特征元数据入湖

主要通过从源端集成的文档本身的基本信息入湖。入湖的过程中，数据内容仍存储在源系统，数据湖中仅存储非结构化数据的基本特征元数据。基本特征元数据入湖需同时满足如下条件。

- 已经设计了包含基本特征元数据的索引表。
- 已经设计了信息架构，如业务对象和逻辑实体。
- 已经定义了索引表中每笔记录对应文件的Owner、标准、密级，认证了数据源并满足质量要求。

参考都柏林核心元数据，非结构化数据的基本特征类属性元数据规范如表5-4所示。

表5-4　非结构化数据的基本特征类属性

元数据实体	元数据元素	定义及规则	数据类型	数据长度	是否必填	有允许值
属性（基本特征类）	数据（Code）	文件的唯一标识	文本	32	是	否
	是否必填（Title）	赋予文件资源的名称	文本	256	是	否
	类型（Type）	文件资源所属的类别，包括文档、图片、音频、视频	文本	32	是	有
	格式（Format）	文件的物理格式，包括 doc、xls、ppt、jpg、bmp 等	文本	16	是	否
	创建者（Creator）	创建资源内容的主要责任者	文本	32	是	否
	主题（Subject）	资源内容的主题描述	文本	64	否	否
	描述（Description）	资源内容的解释	文本	256	否	否
	发布者（Publisher）	使资源成为可获得的责任实体	文本	32	否	否
	其他责任者（Contributor）	资源生存期中做出贡献的其他实体，除制作者 / 创作者之外的其他撰稿人和贡献者，如插图绘制者、编辑等	文本	32	否	否
	创建日期（Create Date）	资源创建的时间	日期	/	是	否
	发布日期（Publish Date）	资源发布的时间	日期	/	否	否
	最后修改时间（Last Modify Date）	资源最近被修改的时间	日期	/	是	否
	生效时间（Effective Date）	资源有效的开始时间	日期	/	是	否
	失效时间（Failure Date）	资源有效的结束时间	日期	/	是	否
	版本（Version）	资源的版本信息	文本	8	是	否

（续）

元数据实体	元数据元素	定义及规则	数据类型	数据长度	是否必填	有允许值
属性（基本特征类）	标识符（Identifier）	资源的唯一标识，如ISBN（国际标准书号）、ISSN（国际标准刊号）、URI（统一资源标识符）、DOI（数字对象标识符）等	文本	64	否	否
	语言（Language）	描述资源知识内容的语言、语种。文档、文本类资源的必填项	文本	16	否	否
	来源（Source）	对当前资源来源的参照，包括组织、人、IT系统等	文本	64	否	否
	关联（Relation）	与其他资源的索引关系，用关联ID来标引参考的相关索引、资源	文本	256	否	否
	覆盖范围（Coverage）	资源使用的范围，如适用区域（地理位置）、业务领域、客户群、角色等	文本	256	是	否
	密级（Security Classification/Rights）	文件的访问密级权限信息	文本	256	是	否

2）文件解析内容入湖

对数据源的文件内容进行文本解析、拆分后入湖。入湖的过程中，原始文件仍存储在源系统，数据湖中仅存储解析后的内容增强元数据。内容解析入湖需同时满足如下条件。

- 已经确定解析后的内容对应的Owner、密级和使用的范围。

- 已经获取了解析前对应原始文件的基本特征元数据。
- 已经确定了内容解析后的存储位置，并保证至少一年内不会迁移。

3）文件关系入湖

根据知识图谱等应用案例在源端提取的文件上下文关系入湖。入湖的过程中，原始文件仍存储在源系统，数据湖中仅存储文件的关系等内容增强元数据。文件关系入湖需同时满足如下条件：

- 已经确定文件对应的Owner、密级和使用的范围。
- 已经获取了文件的基本特征元数据。
- 已经确定了关系实体的存储位置，并保证至少一年内不会迁移。

4）原始文件入湖

根据消费应用案例从源端把原始文件搬入湖。数据湖中存储原始文件并进行全生命周期管理。原始文件入湖需同时满足如下条件。

- 已经确定原始文件对应的Owner、密级和使用的范围。
- 已经获取了基本特征元数据。
- 已经确定了存储位置，并保证至少一年内不会迁移。

5.3　数据主题联接：将数据转换为“信息”

5.3.1　5类数据主题联接的应用场景

在数字化转型的背景下，华为的数据消费已经不再局限于传统的报表分析，还要支持用户的自助分析、实时分析，通过数据的关联，支持业务的关联影响分析以及对目标对象做特征识别，进行特定业务范围圈定、差异化管理与决策等。这些分析需求也不再是对单一数据的分析，往往需要对跨领域的数据进行联接后再进行综合分析。

目前，数据湖汇聚了大量的原始数据，用户不再需要到各个源系统调用数据，而是统一从数据湖调用。由于数据湖中的数据零散且数据结构都与源系统一致，严格遵从三范式，即使每个数据都有详细的定义和解释，用户也很难知道数据之间的关联关系。例如，消费者BG做设备收入预测需要的数据有产品、订单、计划等超过150个物理表信息，这些表没有进行联接，没有形成有用信息，是很难支撑用户进行分析的。

华为在数据湖的基础上通过建立数据联接层，基于不同的分析场景，通过5类联接方式将跨域的数据联接起来，将数据由“原材料”加工成“半成品”和“成品”，支撑不同场景的数据消费需求，如图5-6所示。

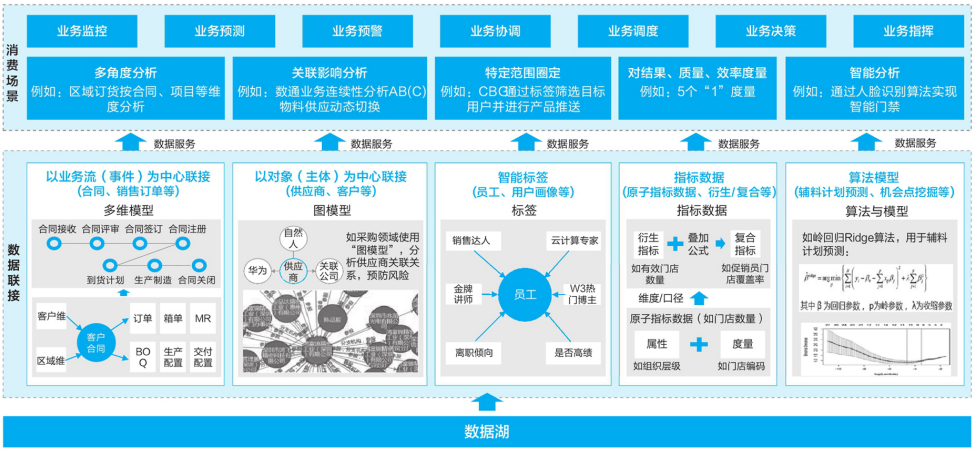


图5-6　5类数据主题联接

多维模型是面向业务的多视角、多维度的分析，通过明确的业务关系，建立基于事实表、维度表以及相互间联接关系，实现多维数据查询和分析。例如，对订货数据从时间、区域、产品、客户等维度进行多视角、不同粒度的查询和分析。

图模型面向数据间的关联影响分析，通过建立数据对象以及数据实例之间的关系，帮助业务快速定位关联影响。例如，查看某国家原产地的项目的数据具体关联到哪个客户以及合同、订单、产品的详细信息时，可以通过图模型快速分析关联影响，支撑业务决策。

标签是对特定业务范围的圈定。在业务场景的上下文背景中，运用抽象、归纳、推理等算法计算并生成目标对象特征的表示符号，是用户主观观察、认识和描述对象的一个角度。例如，对用户进行画像，识别不同的用户群，为产品设计和营销提供策略支持。

指标是对业务结果、效率和质量的度量。依据明确的业务规则，通过数据计算得到衡量目标总体特征的统计数值，能客观表征企业某一业务活动中业务状况。例如，促销员门店覆盖率指标就是衡量一线销售门店促销员的覆盖程度。

算法模型是面向智能分析的场景，通过数学建模对现实世界进行抽象、模拟和仿真，提供支撑业务判断和决策的高级分析方法。例如，预测未来18个月的销售量，需要数据科学家根据数据湖中的历史订单、发货等数据通过决策树和基因算法进行数据建模，支持业务决策。

5.3.2 多维模型设计

多维模型是依据明确的业务关系，建立基于维度、事实表以及相互间连接关系的模型，实现多角度、多层次的数据查询和分析。如何设计出稳定、易扩展、高可用的数据模型来支持用户消费对数据主题联接至关重要。

多维模型设计有4个主要步骤，包括确定业务场景、声明粒度、维度设计和事实表设计。

（1）确定业务场景

分析业务需求，识别需求中所涉及的业务流及其对应的逻辑数据实体和关联关系。如业务负责人（PO）履行全流程可视，首先需要识别监控的具体业务环节（如发货、开票等），

再根据这些业务环节识别其对应的逻辑数据实体及关联关系，如图5-7所示。

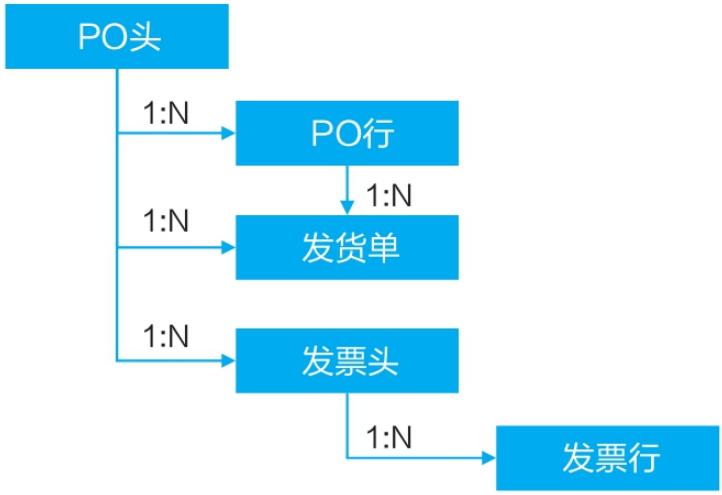


图5-7　PO履行全流程可视的数据范围

（2）声明粒度

粒度表示数据单元的细节程度或综合程度，细节程度越高，粒度越细；细节程度越低，粒度越粗。声明粒度是维度和事实表设计的重要步骤，声明粒度意味着精确定义事实表的每一行表示什么。针对监控PO履行这个场景，在做设计时首先要确认是监控PO的履行，还是具体到每个PO行的履行，不同的粒度会对应不同的事实表。

（3）维度设计

维度是用于观察和分析业务数据的视角，支持对数据进行汇聚、钻取、切片分析，如图5-8所示。维度由层次结构（关系）、层级、成员、属性组成。维度可以分为基础树和组合

树，维度基础树提供统一定义的、完整的层级结构和成员；维度组合树根据业务使用场景进行定制。

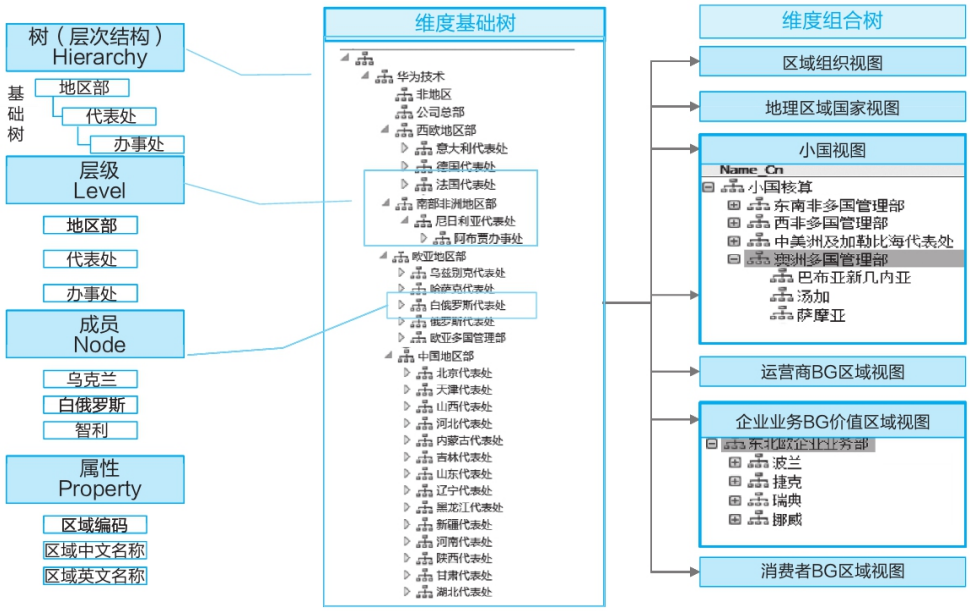


图5-8　维度示例

维度设计需要满足单一性、单向性和正交性。

1）单一性

有且仅有一个视角，在同一个维度中不能穿插其他经营分析的视角，例如，区域维不含客户视角，产品维不含客户视角等。图5-9中区域维度客户视角不满足单一性要求。

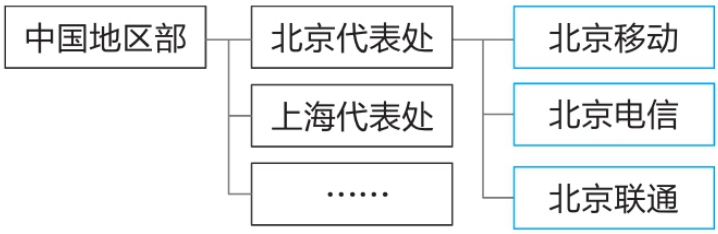


图5-9　不满足单一性示例

2）单向性

“上大下小”，维度只能支撑自上而下的分解和自下而上的收敛，每个成员只能存在向上的收敛路径，不能具备向上和向下两个方向的收敛逻辑。图5-10中代表处维度低于国 家维度，不满足单向性要求。

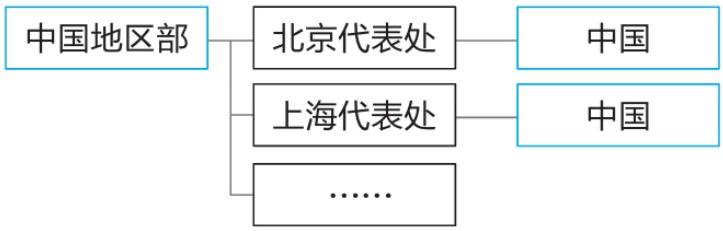


图5-10　不满足单向性示例

3）正交性

成员两两不相交，同一成员不能同时拥有多个上级成员，以产品维为例，华为向客户提供的设备或服务都只能被准确地分配到唯一叶子（最底层）节点，并以此路径进行收敛。图5-11中最小粒度成员“无线专业服务”同时归属不同的上层节点，不满足正交性要求。

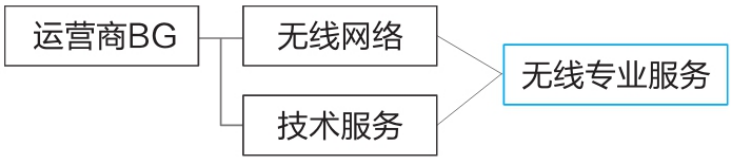


图5-11　不满足正交性示例

（4）事实表设计

事实表存储业务过程事件的性能度量结果，由粒度属性、维度属性、事实属性和其他描述属性组成，如图5-12所示。

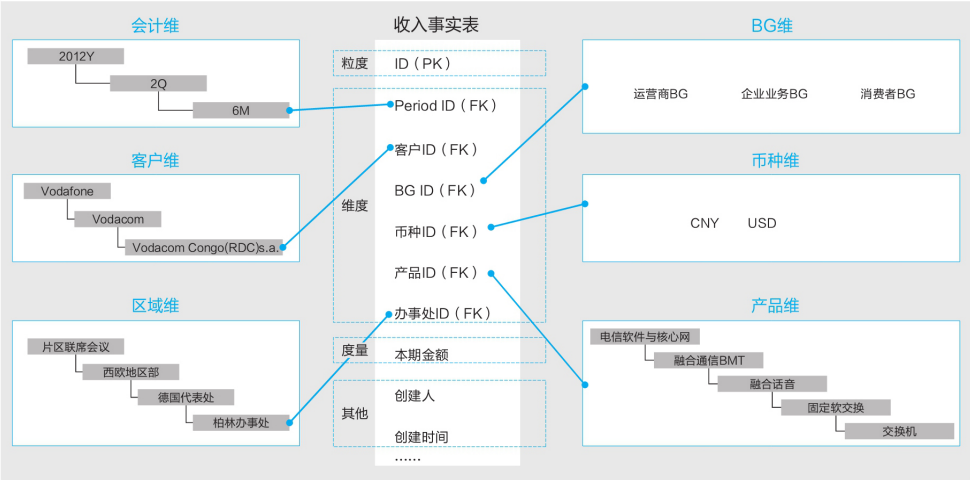


图5-12　事实表示例

粒度属性是事实表的主键，通常由原始数据的主键或一组维度属性生成。

维度属性是从维度中继承的属性，可以只继承主键作为事实表的外键，也可以继承维度中全部或其他部分的属性。在上述例子中，事实表中除了有币种ID，还可以带有币种编码和币种名称等属性。

- 事实属性是可以对该颗粒度的事实进行定量的属性，大多数的事实表包括一个或多个事实字段。
- 同一事实表中不能存在多种不同粒度的事实，比如PO行明细事实表中不应该包含PO总金额，否则PO总金额累加时会出现错误。

- 尽可能包含所有与业务过程相关的事实，不包含与业务过程无关的事实，比如在设计“订单下单”这个业务过程的事实表时，不应该存在“支付金额”这个支付业务过程的事实。
- 对于不可相加的事实，需要分解为可加的事实。比如比率，需要分解为分子和分母。
- 事实的数值单位要保持一致。

其他属性主要包括创建人、创建时间、最后修改人、最后修改时间等审计字段。

5.3.3 图模型设计

图模型作为当前流行的信息处理加工技术，自提出以来，迅速在学术界和工业界得到了普及，在智能推荐、决策分析等方面有着广泛的应用。

图模型由节点和边组成。节点表示实体或概念，边则由属性或关系构成。实体指的是具有可区别性且独立存在的某种事物，如某一个人、某一个城市、某一种植物、某一种商品等，是图模型中的最基本元素；概念是对特征的组合而形成的知识单元，主要指集合、类别、对象类型、事物的种类，例如人物、地理等；属性主要指描述实体或概念的特征或特性，例如人员的国籍、生日 等。我们以“哲学家”为例设计图模型，如图5-13所示。

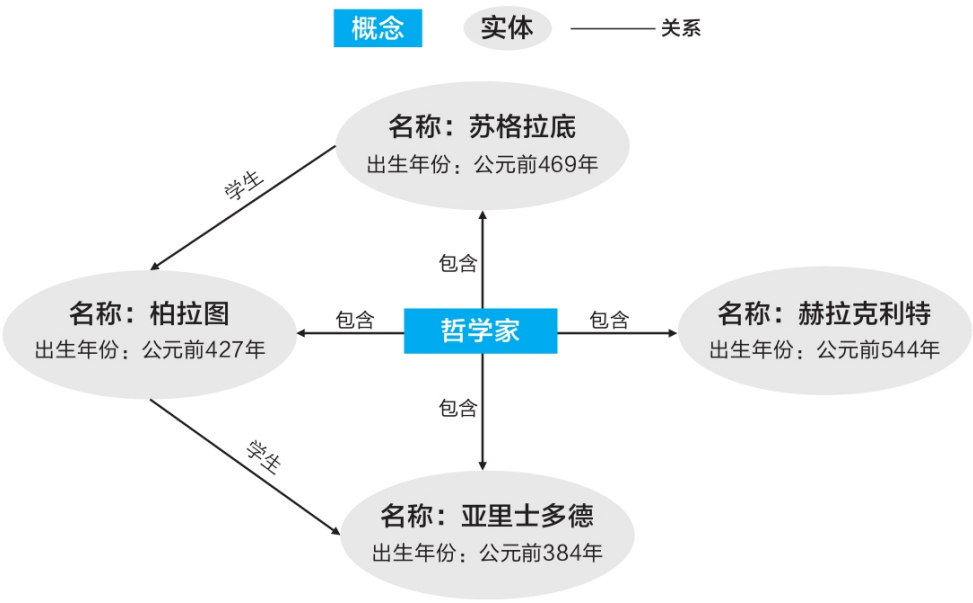


图5-13　图模型示例

图模型构建包含几个关键步骤，如图5-14所示。

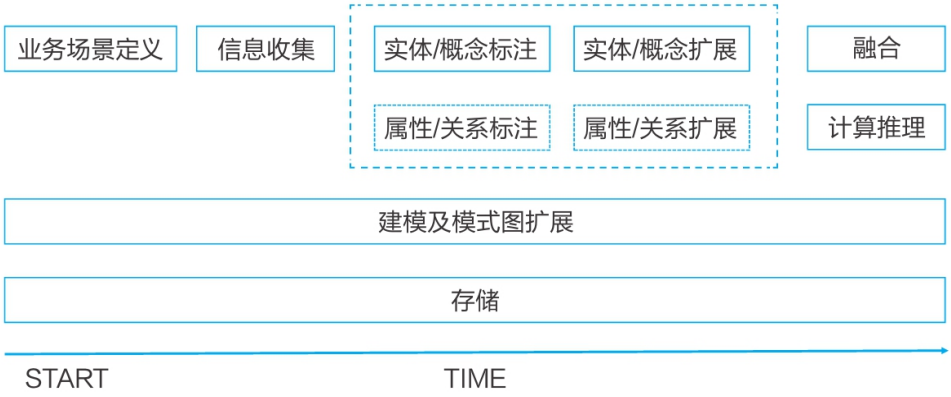


图5-14　企业图模型构建步骤

第一步：业务场景定义。

业务场景决定信息涵盖范围，以及信息颗粒度的表示。以支撑业务连续性为例，因为不可抗力的影响，部分区域的供应商工厂无法正常生产和发货，涉及的信息包括供应商的信息、

产能、元器件及内部物料、合同和客户信息，要求能够根据用户输入的当前物料储备和合同状态，获取影响内部物料、产品、合同交付和客户的清单和范围。这种应用涉及对产品目录和配置的解读，需要对收集的信息进行最小采购器件的抽取。

信息颗粒度在图模型建设中是个不可忽视的问题，根据应用场景决定信息颗粒度以及图模型的精确性与有效性。比如手机，有品牌、型号、批次，直至手机整机。同样的信息范围，颗粒度越细，图模型应用越广泛，关系越丰富，但冗余越多，知识消费越低效。信息颗粒度的原则是“能满足业务应用的最粗颗粒度”。

第二步：信息收集。

信息的选取要考虑两个方面的内容。

1）与应用场景直接相关的信息。例如，判断不可抗力供应中断影响的范围，直接相关的信息有物料信息、产品配置、合同信息等。

2）与应用场景间接相关，但可辅助理解问题的信息。这包括企业信息、专业领域信息、行业信息以及开放域信息。

第三步：图建模。

相同的数据可以有若干种模式的定义，良 好的模式可以减少数据冗余，提高实体识别的准确率，在建模的过程中，要结

合数据特点与应用场景来完成。同样的数据从不同的视角可以得出不同的图模型。

第四步：实体、概念、属性、关系的标注。

企业图模型中涉及的实体和概念可分为三类：公共类，如人名、机构名、地名、公司名、时间等；企业类，如业务术语、企业部门等；行业类，如金融行业、通信行业等。

第五步：实体和概念的识别。

企业图模型中实体、概念的识别可将业务输入与数据资产中已有的信息作为种子，运用命名实体识别（NER）的方法扩展出新实体概念，经业务确认后，列入实体、概念库。

第六步：属性识别与关系识别。

企业图模型中的属性与关系一般是根据业务知识在模式层设计时定义，属性与关系相对稳定，其扩展场景不是很多。

企业图模型的存储技术要综合考虑应用场景、图模型中节点和联接的数量、逻辑的复杂度、属性的复杂度，以及性能要求。一般建议采用混合存储方式，用图数据库存储关系，关系型数据库或键值对存储属性。偏重逻辑推理的应用场景用RDF的存储方式，偏重图计算的应用场景选择属性图的存储方式。发挥两类数据存储和读写的各自优势。

知识计算主要是根据图谱提供的信息得到更多隐含的知识，如通过模式层以及规则推理技术可以获取数据中存在的隐含信息。知识计算涉及三大关键技术：图挖掘计算、基于本体的推理、基于规则的推理。图挖掘计算是基于图论的相关算法，实现对图谱的探索和挖掘。图挖掘计算主要分为如下6类。

- 图遍历：知识图谱构建完之后可以理解为是一张很大的图，可以去查询和遍历这个图，要根据图的特点和应用场景进行遍历。
- 图里面经典的算法，如最短路径。
- 路径的探寻，即根据给定两个实体或多个实体去发现它们之间的关系。
- 权威节点的分析，这在社交网络分析中使用较多。
- 族群分析。
- 相似节点的发现。

图挖掘计算如图5-15所示。

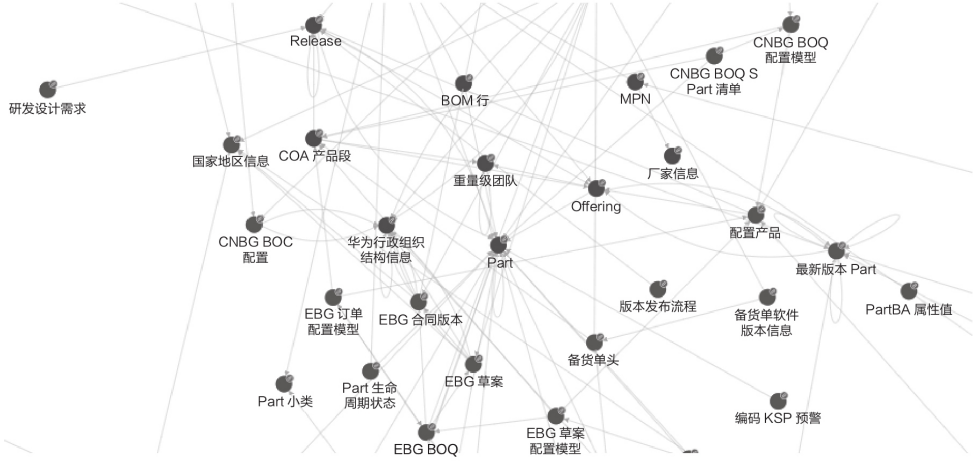


图5-15　图模型示例

图挖掘计算在当前的应用场景中，基于业务连续性，通过查询遍历图模型，识别影响节点和影响范围，基于最短路径，辅助决策物流线路，在企业中的应用较为普遍。

图模型在企业中的价值，很大程度上取决于企业基于对象节点可以构建多完善的关系，这个关系的构建是一个逐步完善的过程，基于业务场景不断补充和完善关系，这就是图模型的优势。当形成一个足够完善的企业级图模型后，领域分段的业务场景应用只需要裁剪部分节点和关系，就可以满足业务的需求，达到快速响应业务需求、降低开发成本的目的。

5.3.4 标签设计

标签是根据业务场景的需求，通过对目标对象（含静态、动态特性）运用抽象、归纳、推理等算法得到的高度精练的特征标识，用于差异化管理与决策。标签由标签和标签值组成，打在目标对象上，如图5-16所示。

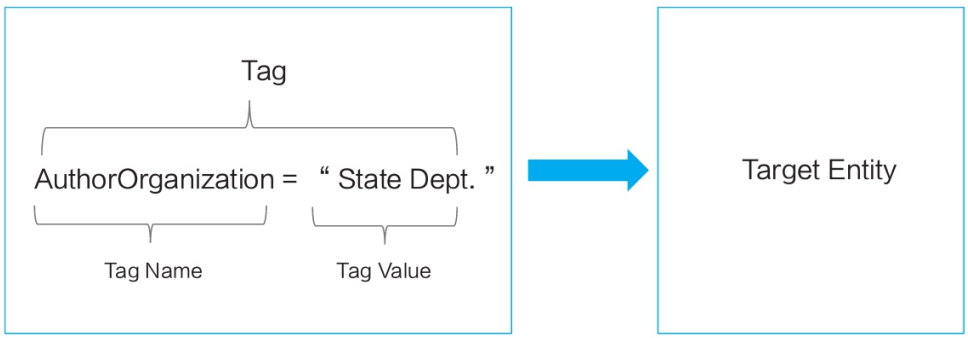


图5-16 打标签示例

标签由互联网领域逐步推广到其他领域，打标签的对象也由用户、产品等扩展到渠道、营销活动等。在互联网领域，标签有助于实现精准营销、定向推送、提升用户差异化体验等；在行业领域，标签更多助力于战略分级、智能搜索、优化运营、精准营销、优化服务、智慧经营等。

标签分为事实标签、规则标签和模型标签，如图5-17所示。

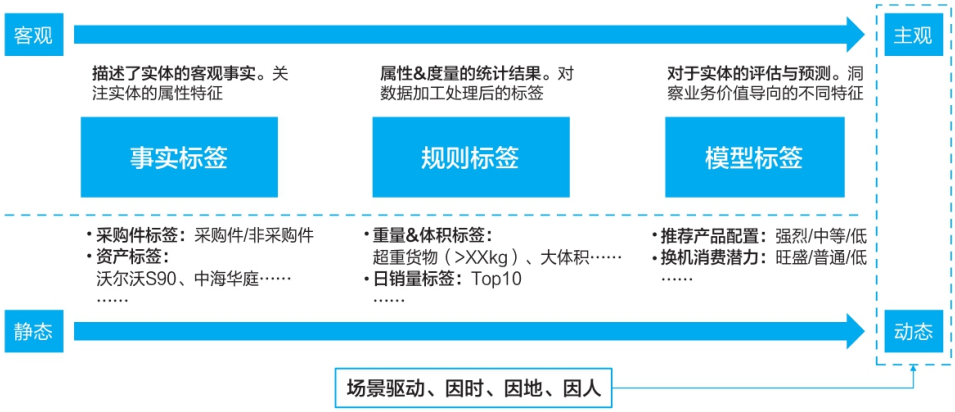


图5-17　三种类型的标签

事实标签是描述实体的客观事实，关注实体的属性特征，如一个部件是采购件还是非采购件，一名员工是男性还是女性等，标签来源于实体的属性，是客观和静态的；规则标签是对数据加工处理后的标签，是属性与度量结合的统计结果，如货物是否是超重货物，产品是否是热销产品等，标签是通过属性结合一些判断规则生成的，是相对客观和静态的；模型标签则是洞察业务价值导向的不同特征，是对于实体的评估和预测，如消费者的换机消费潜力是旺盛、普通还是低等，标签是通过属性结合算法生成的，是主观和动态的。

标签管理分为标签体系建设和打标签。

1. 标签体系建设

（1）选定目标对象，根据业务需求确定标签所打的业务对象，业务对象范围参考公司发布的信息架构中的业务对象。

（2）根据标签的复杂程度进行标签层级设计。

（3）进行详细的标签和标签值设计，包括标签定义、适用范围、标签的生成逻辑等：

- 事实标签应与业务对象中的属性和属性值保持一致，不允许新增和修改；
- 规则标签按照业务部门的规则进行相关设计；
- 模型标签根据算法模型生成。

2. 打标签

（1）打标签数据存储结构

打标签是建立标签值与实例数据的关系，可以对一个业务对象、一个逻辑数据实体、一个物理表或一条记录打标签。

为了方便从“用户”视角查找、关联、消费标签，可增加用户表，将标签归属到该“用户”下，这里的“用户”是泛

指，可以是具体的人，也可以是一个组织、一个部门、一个项目等。

（2）打标签的实现方法

- 事实标签：根据标签值和属性允许值的关系由系统自动打标签。
- 规则标签：设计打标签逻辑由系统自动打标签。
- 模型标签：设计打标签算法模型由系统自动打标签。

5.3.5 指标设计

指标是衡量目标总体特征的统计数值，是能表征企业某一业务活动中业务状况的数值指示器。指标一般由指标名称和指标数值两部分组成，指标名称及其含义体现了指标在质的规定性和量的规定性两个方面的特点；指标数值反映了指标在具体时间、地点、条件下的数量表现。

根据指标计算逻辑是否含有叠加公式，可以把指标分为原子指标和复合指标两种类型。

原子指标是指标数据通过添加口径/修饰词、维度卷积而成，口径/修饰词、维度均来源于指标数据中的属性。

复合指标由一个或多个原子指标叠加计算而成，其中维度、口径/修饰词均继承于原子指标，不能脱离原子指标维度和

口径/修饰词的范围去产生新的维度和口径/修饰词。指标和数据的关系如图5-18所示。

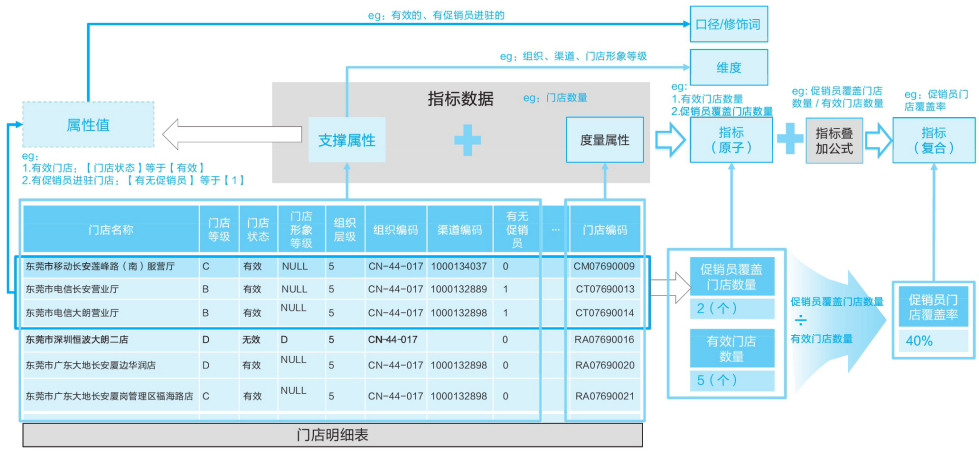


图5-18　指标和数据关系 样例

- 指标数据：承载原子指标的数据表，例如门店明细表，其中度量为门店数量，通过【 门店编码】卷积；属性包括门店等级、门店状态、门店形象等级、组织等级等。
- 维度：从属性中选取组织、渠道、门店形象等级。
- 口径/修饰词：【 门店状态】等于【 有效】，【 有无促销员】等于【 1】 。
- 原子指标：由指标数据通过添加口径/修饰词、维度卷积而成，包括促销员覆盖门店数量、有效门店数量。
- 复合指标：由2个或2个以上指标叠加计算而成，包括【 促销员门店覆盖率】=【 促销员覆盖门店数量】÷【 有效门店数量】 。

如何按需求进行指标拆解，是将指标对应到数据资产并进行结构化管理，支持指标服务化与自助需求的关键。指标的拆

解过程主要包括指标拆解需求澄清、指标拆解设计、指标数据与数据资产匹配3个阶段，如图5-19所示。

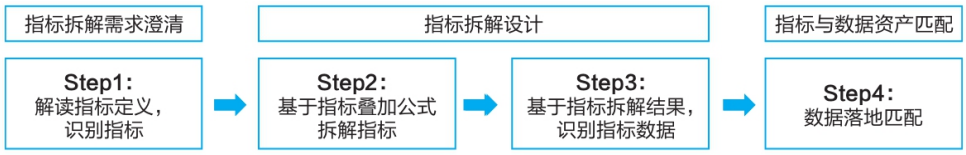


图5-19　指标拆解过程

- **解读指标定义，识别指标：**通过与指标定义的业务管理部门沟通（通常为指标解释部门的业务人员），从业务角度了解指标基本信息、所需统计维度、指标度量场景以及各场景下的计算逻辑和口径（包括剔除规则）、指标发布信息等。
- **基于指标叠加公式拆解指标：**根据指标计算逻辑识别原子指标，明确原子指标中需要的口径/修饰词、维度信息，以及原子指标与复合指标间的支撑关系。
- **基于指标拆解结果，识别指标数据：**识别原子指标的度量属性和支撑属性，并根据原子指标中的维度、口径修饰词匹配已发布业务对象的属性，形成指标数据。
- **数据匹配落地：**补充指标、指标数据中的标准属性名称以及对应的落地物理表，支持用户自助实现指标计算，拉通指标设计和落地。

5.3.6　算法模型设计

算法是指训练、学习模型的具体计算方法，也就是如何求解全局最优解，并使得这个过程高效且准确，其本质上是求数学问题的最优化解，即算法是利用样本数据生成模型的方法。算法模型是根据业务需求，运用数学方法对数据进行建模，得到业务最优解，主要用于业务智能分析。

算法模型在数据分析流程中产生，算法模型管理框架包括建模、模型资产管理和模型消费。公司各领域已相继开发出大量基于算法模型的分析应用，通过对算法模型资产注册逐步打造公司级的算法模型地图。

算法模型的设计步骤主要有需求评估、数据准备、方案设计和建模与验证。

（1）需求评估

1）业务驱动的分析需求识别

- 如果要识别与业务运营优化相关的分析需求，就需要梳理业务需求的背景、现状与目标。
- 若由战略或变革提出可能的分析需求，则应进行战略目标解耦，识别分析需求，了解业务现状与制定目标。
- 初步识别分析结果的应用场景。

2）数据驱动的分析需求识别

- 在集成的数据环境中进行数据挖掘，探索可能的分析应用。

- 识别分析需求和确认应用领域。
- 初步识别分析结果的应用场景。

3）价值与可行性评估

- 确定数据分析主题。
- 分析需求的业务价值评估，包括业务基线、分析主题的业务影响与可增进的效益。
- 分析前提与可行性，包括识别目前业务流程与可能的影响因素，探讨业务现状因素，并制定对应的分析解决方案，呈现出对应解决方案可提升的效益，对方案所需资源和数据的可行性进行评估。
- 根据相关的历史数据，进行假设和分析，并明确业务范围。

（2）数据准备

- 深入探索数据资产目录，识别与分析主题可能相关的数据。
- 提供数据源、数据标准、数据流等信息。
- 收集与整合原始数据，生成分析数据集。
- 根据分析需求进行数据筛选和质量分析。

（3）方案设计

- 明确要分析的业务目标与相关假设。
- 定义数据集中的分析目标、样本与筛选条件。
- 设计所需变量、指标、可能的分析方法和产出。

- 规划分析的应用场景。

（4）建模与验证

1）**决定是否需要分析建模：**根据技术复杂度、业务效益和资源评估该分析需求是否需要分析建模。若需要分析建模且通过项目评审，则应进行高阶分析；若不需要建模分析，则运用BI分析。

2）**建模与验证：**根据数据分析方案创建模型，对模型的参数和变量进行调整，根据应用场景选择适用的模型，并与业务分析师确认模型成效与应用，并进行优化，进行模型相关验证（如准确度和稳定度评估）及效益评估。

3）**试算分析：**对数据分析方案中不需分析建模的场景和应用，根据数据分析方案进行分析结果的计算，并选择合适的展示方式。

4）**编写数据分析线下验证报告：**

- 记录分析结果与发现。
- 根据洞察发现，建议业务应用场景。
- 建议模型监测方式。

5）**决定是否需要IT开发：**根据模型验证成果（分析建模）、预估业务效益、IT开发所需的成本和资源来评估分析结果是否需要IT开发。若需要，则通过评审后转入IT开发流程；若不需要，则进入业务应用并结束流程

6）**模型线上验证：**

- 设定线上验证范围与场景。
- 进行线上验证，制定模型监控机制（含监控频次和监控要素），生成分析模型线上验证报告。
- 进行业务试运行与推广。

7）**转运营：**与数据分析模型所属领域的业务代表确认转正式运营计划，启动业务正式运营。

5.4 本章小结

企业数据治理的最终目的是让数据更有效地服务于业务目标，创造价值。对于数字原生企业而言，原生入口提供的大规模、高质量的数据，可以快速地封装成企业级的API，满足业务侧的应用。华为作为非数字原生企业，在实践探索中发现，数字化转型的关键在于打通数据供应链，通过理解业务、识别数据资产、建设数据架构来推动组织间的共享和协作，标识安全隐私标签，从源头提升数据质量，并通过数据底座建设构建数据湖和数据主题联接两层，形成数据的逻辑集合，为业务可视化、分析、决策等数据消费提供数据服务，让企业数据成为能为业务带来价值的数据资产。

第6章

面向“自助消费”的数据服务建设

数据底座建设的目标是更好地支撑数据消费，在完成数据的汇聚、整合、联接之后，还需要在供应侧确保用户更便捷、更安全地获取数据。一方面业务人员希望尽可能快速地获取各种所需的数据，另一方面要确保数据获取过程中满足安全、合规的要求。同时，业务人员消费数据时，也希望能够有更加灵活的使用数据、分析数据的方式，业务人员希望消费数据的自主性更强，并且不能忍受过去冗长、呆板的报表呈现方式。

在数据供应侧和消费侧的双重推动下，华为公司进行了基于数据服务提供“自助消费”的实践，打造了从数据供应到消费的完整链条。

6.1　数据服务：实现数据自助、高效、复用

过去，数据获取大部分依赖于传统集成方式，即将数据从一个系统复制到另一个系统。随着企业规模的扩大，需要在几十个甚至上百个IT系统中进行数据集成，这样一来，随着系统集成的复杂度的提升，会带来一系列数据质量问题。

我们来看两个不同场景下的相似案例。

图6-1是以交易侧（OLTP）客户合同为例，涉及近100个系统/工具、近200个集成关系，多种集成技术同时并存，复杂的集成关系导致出现各种数据质量问题，给企业的正常运作带来了很大的风险。

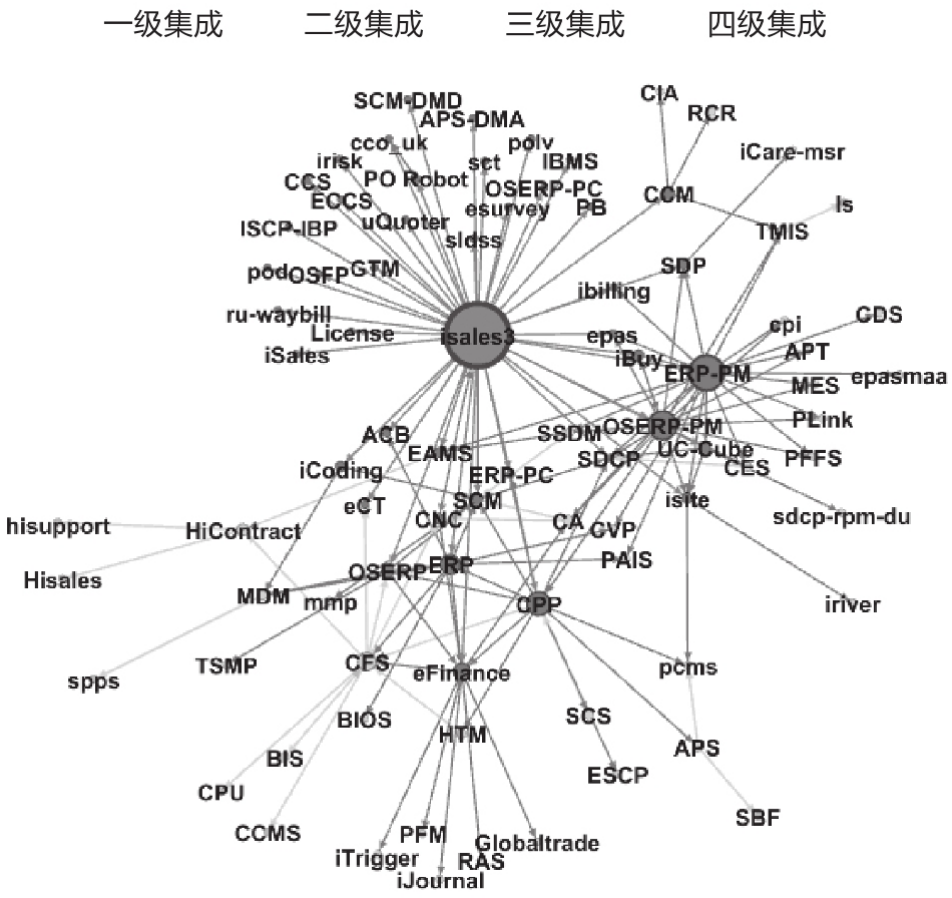


图6-1　合同数据集成视图（示例）

我们换个视角，从业务分析侧（OLAP）看看数据“搬家”造成的问题。以图6-2的经营数据为例，涉及7个领域的30多个系统的集成。在OLAP侧由于类似情况造成的数据“搬家”，可能需要上千万元的IT开发费用。而且，各分析系统对财务数据的使用和再加工，以及数据集成的时间差等，会造成与集团报告不一致，甚至导致安全审计风险。

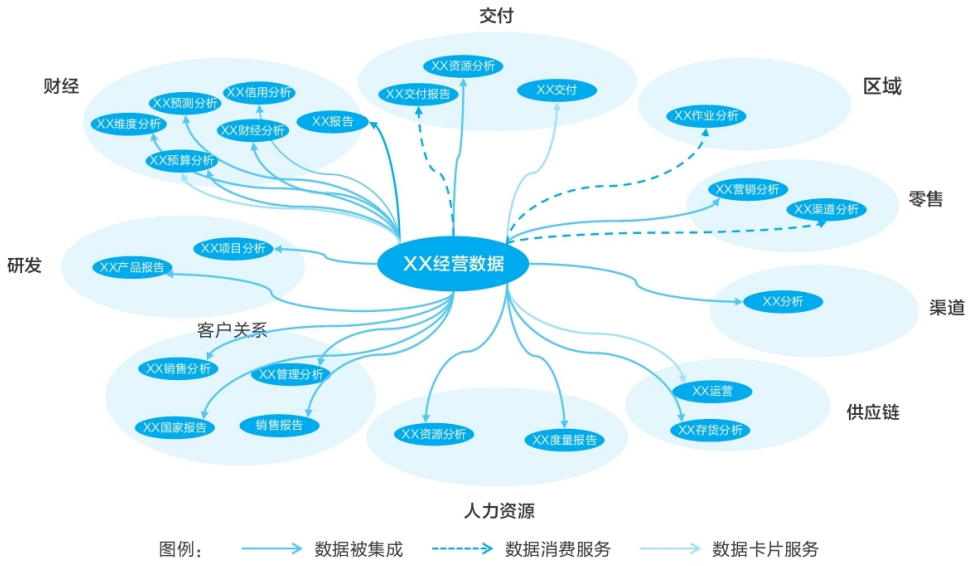


图6-2　业务分析侧数据集成情况示例

从这两个案例可以看出，数据在不同的系统间不断“搬家”，数据的一致性很难得到保障，尤其是经过多次搬家后，源头数据往往和下游各系统之间的数据差异巨大。

同时，较复杂的数据集成还会导致企业管理成本上升，每个系统都存在数据的大量重复构建，这样一来，每当源头数据出现变化时，整个业务流上的相关系统都要执行变更。

这种通过集成获取数据的方式不仅会导致当前的诸多问题，而且会给未来的业务发展带来更大的挑战。以图6-3为例，如果不能满足新的协同方式、安全、数据使用模式等要求，将很难应对未来企业之间的交互协作。

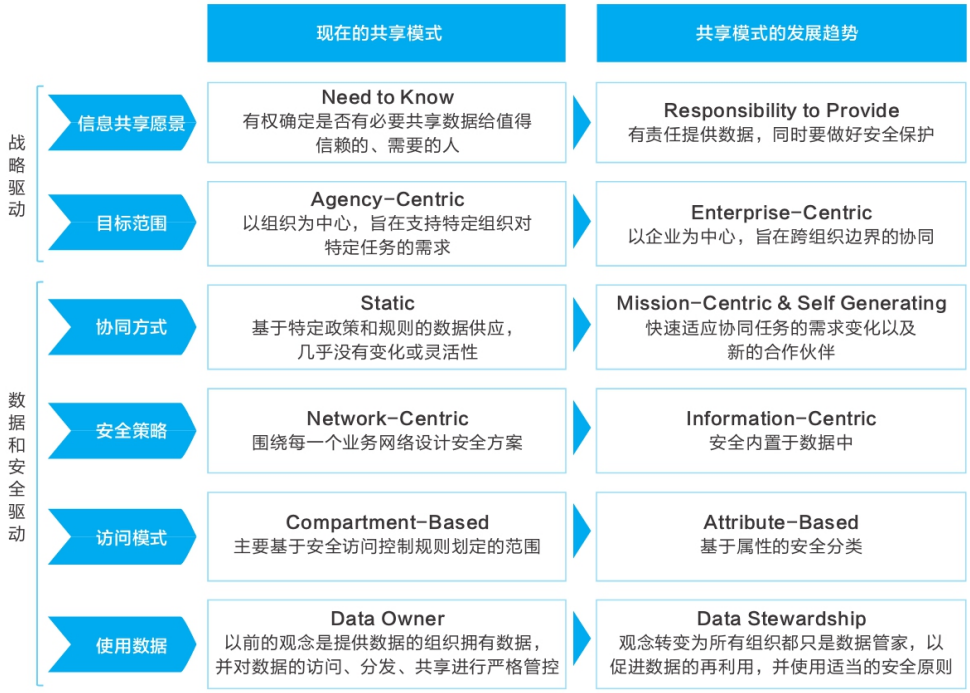


图6-3　数据共享模式的发展趋势（参考：美国智慧社区信息共享战略）

在这样的背景下，华为公司进行了大规模的数据服务建设，通过数据服务替代原有数据集成方式，解决了数据交互过程中的诸多问题，取得了数据获取效率和数据安全之间的平衡。

6.1.1　什么是数据服务

参考IEEE规范，华为公司给出了数据服务的定义。数据服务是基于数据分发、发布的框架，将数据作为一种服务产品来提供，以满足客户的实时数据需求，它能复用并符合企业和工业标准，兼顾数据共享和安全。

以图6-4为例，数据服务和传统集成方式有很大区别，数据的使用方（不仅仅是IT系统人员，也可以是具体业务人员）不再需要点对点地寻找数据来源，再点对点地进行数据集成，从而形成错综复杂的集成关系，而是通过公共数据服务按需获取各类数据。

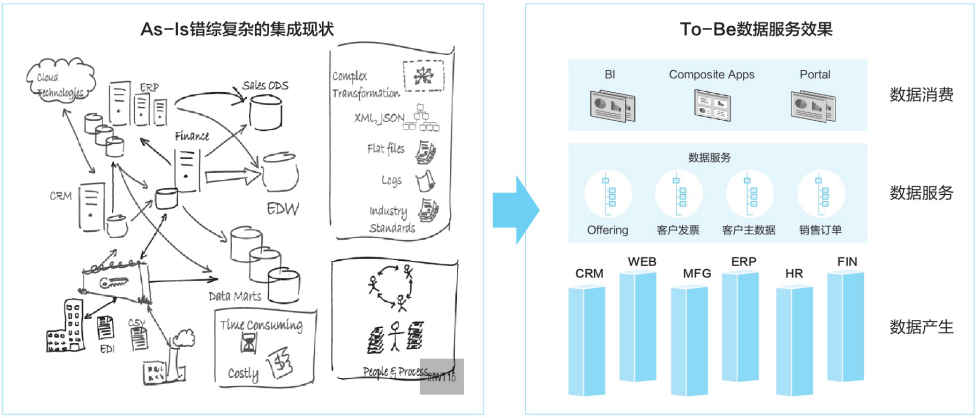


图6-4　数据服务和传统集成方式对比

1. 数据服务给企业带来的价值

数据服务给企业带来了如图6-5所示的价值。



图6-5　数据服务化的价值

1）保障“数出一孔”，提升数据的一致性。通过服务获取数据的方式类似于“阅后即焚”，大部分情况下数据并不会在使用方的系统中落地，因此减少了数据“搬家”，而一旦数据的使用方并不拥有数据，就减少了向下游二次传递所造成的数据不一致问题。

2）数据消费者不用关注技术细节，可以满足不同类型的数据服务需求。对于数据消费者而言，不用再关心“我要的数据在哪里”，例如用户不需要知道这些数据来自哪个系统、哪个数据库、哪个物理表，只需要清楚自身的数据需求，就能找到对应的数据服务，进而获取数据。

3）提升数据敏捷响应能力。数据服务一旦建设完成，并不需要按使用者重复构建集成通道，而是通过“订阅”该数据服务快速获取数据。

4）满足用户灵活多样的消费诉求。数据服务的提供者并不需要关心用户怎么“消费”数据，避免了供应方持续开发却满足不了消费方灵活多变的数据使用诉求的问题。

5）兼顾数据安全。所有数据服务的使用都可管理，数据供应方能够准确、及时地了解“谁”使用了自己的数据，并且可以在数据服务建设中落实各种安全措施，确保数据使用的合规。

2. 数据服务建设策略

数据服务建设过程中，首先应该在企业层面制定统一的数据服务建设策略，如图6-6所示。这个策略不能只关注数据服务的设计，而应该覆盖全生命周期的各个环节。

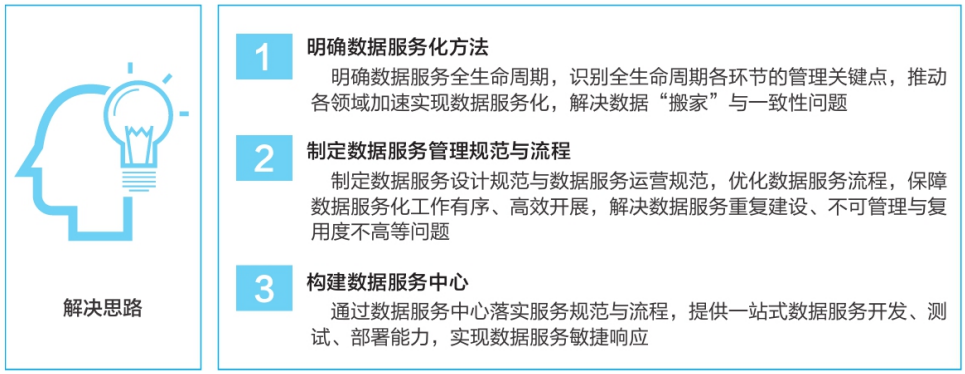


图6-6　数据服务建设策略

1）要制定数据服务建设的方法，确保每个从事数据服务建设的人都明白数据一致性的要求，并且所提供的数据是可信的和清洁的。

2）要建立数据服务流程，以确保各个环节的有效协同，定义整个生命周期中每个角色的责任和有效输出。在企业中，应该有明确的部门负责数据服务流程的建设和看护，一方面要确保所制定的流程能够在实际工作中落地，另一方面随着技术的演进和企业业务环境的变化，持续对流程进行优化和完善。

3）要构建统一的数据服务能力中心，负责数据服务建设方法、规范、流程的落地，数据服务不同于传统集成方式，因此应该有统一的平台提供能力保障。

在数据服务建设中，应该为各个供应方树立统一的标准，并将这些标准以规范的形式进行固化，使所有数据服务建设都遵循同样的标准。

1）数据服务要满足可重用性、减少数据“搬家”。

- 数据服务在实际或者可预见的时间范围内会被多个需求方消费。
- 数据服务面向场景进行消费时，无须重复落地。

2）服务提供方在规划服务时应明确服务的用户是谁，并针对用户的场景和需求进行服务设计，同时定义SLA服务水平承诺。

- 服务要有业务Owner，业务Owner负责组织业务和IT一体化团队，主动进行服务规划和设计。
- 服务规划和设计人员在规划和设计任何服务时，都应考虑到服务可能会被重用。

- 服务规划需考虑价值，并优先对高价值的服务进行投资。
- 服务消费方应对服务提出改进需求，促进服务能力的持续提升。

3）应用只能通过服务接口向其他应用开放其数据和功能，服务接口要稳定，应用间的通信也必须通过这些服务接口进行。

需要说明的是，应用如果需要向其他应用开放其数据和功能，只能通过服务接口，服务接口应该易理解、易使用，达到服务市场准入标准。

4）所有的服务需在统一的服务管控平台中进行注册和发布。

需要说明的是，华为公司的IT服务（HIS）负责提供服务管控平台的注册和发布功能，通过HIS可查询到发布的所有服务。

5）应根据不同场景选择合适的服务化架构粒度。

需要说明的是，服务化要采用合适的架构粒度，不是越“微”越好，也不是越“灵活”越好。

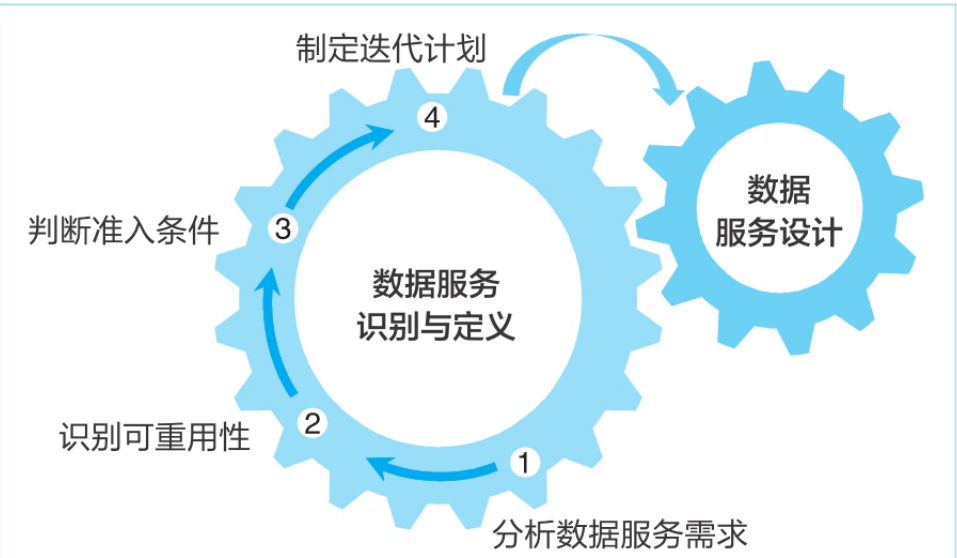
6.1.2　数据服务生命周期管理

完整的数据服务生命周期包括服务识别与定义、服务设计与实现、服务运营三个主要阶段。

- **服务识别与定义：**业务与数据握手，识别服务的业务价值、准入条件与服务类型，减少数据服务的重复建设，提升数据服务的重用度。
- **服务设计与实现：**业务、数据、IT三方协同，使设计、开发、测试与部署快速迭代以实现服务的敏捷交付，缩短数据服务的建设周期。
- **服务运营：**通过统一数据服务中心及服务运营机制，保障服务SLA与持续优化。

1. 数据服务的识别与定义

针对数据需求，规范数据服务识别过程，列出数据服务识别过程需要了解的关键内容，明确数据服务的实现方式和准入条件，提高数据服务的可复用性，减少重复建设，如图6-7所示。



服务识别与定义框架（Service Identification Framework）

图6-7　数据服务识别与定义的关键要素

1）**分析数据服务需求：**通过数据需求调研与需求交接，判断数据服务类型（面向系统或面向消费）、数据内容（指标/维度/范围/报表项）、数据源与时效性要求。

2）**识别可重用性：**结合数据需求分析，通过数据服务中心匹配已有的数据服务，判断以哪种方式（新建服务、直接复用、服务变更）满足业务需求。对于已有数据服务，必须使用服务化方式满足需求，减少数据“搬家”。

3）**判断准入条件**：判断服务设计条件是否已具备，包括数据Owner是否明确、元数据是否定义、业务元数据和技术元数据是否建立联接、数据是否已入湖等。

4）**制定迭代计划：**根据数据服务需求制定敏捷交付计划，快速满足数据服务需求。

在数据服务的识别和定义中，要特别注意数据服务的可重用性。所有的数据服务都是需求驱动产生的，如果没有需求方，那么这个数据服务就没有存在的价值。但是，重复建设也就成了数据服务建设中最大的风险之一。数据供应方很容易基于不同的数据消费方开发出不同的数据服务，而他们的数据需求往往是相似的，但数据供应方可能因为响应周期、客户（数据消费方）体验等压力，并没有花费精力去对来自各数据消费方的需求进行筛选和收敛，这就导致所建设的数据服务往往只是满足某个特定客户（数据消费方）的需求。如果数据服务不

具备可重用性，那么它与传统的数据集成的方式相比，就并不具备优越性，有时甚至会导致更大的重复投资。

因此，以数据服务需求分析为输入，通过服务可重用性判断已有数据服务对需求的满足情况，给出满足服务需求的策略，并结合准入条件的关键问题判断服务需求是否能够被快速满足。

通常可以从数据服务提供形式、数据服务提供内容两方面来判断服务的可重用性，如图6-8所示。

服务提供形式	服务提供内容	需求满足策略
相同	内容完全相同	复用
相同	如果内容不满足，需要依据数据服务封装规范和要求判断是在原有的服务上变更还是新增服务	变更/新增
不相同	-	新增

图6-8　服务可重用性判断矩阵的示例

在数据服务识别和定义阶段还应重点关注的另一个要素是数据资产的可服务性，通常用于数据服务准入条件的判断，即某个数据资产是否已经具备对外提供服务的条件。

在进行数据可服务性（准入条件）判断时，至少应充分考虑以下因素：

- 数据Owner是否明确？
- 数据是否有明确的安全密级定义？

- 元数据是否定义？
- 业务元数据和技术元数据是否建立联接？
- 面向数字化运营分析场景时，数据是否已入湖？

2. 数据服务的设计与实现

在服务设计与实现阶段，要定义服务契约和数据契约，重点明确服务契约所涉及的服务责任主体、处理逻辑，并以数据契约规范服务的数据格式与数据的安全要求。

通过服务契约与数据契约，可以有效地管理在数据交互中可能存在的安全风险，数据供应方可以将一些安全要求通过契约实现。例如，对某些高密级控制属性进行屏蔽。同时，供应方能够通过契约了解哪些消费者获取了数据以及使用的数量和频率等，如图6-9所示。

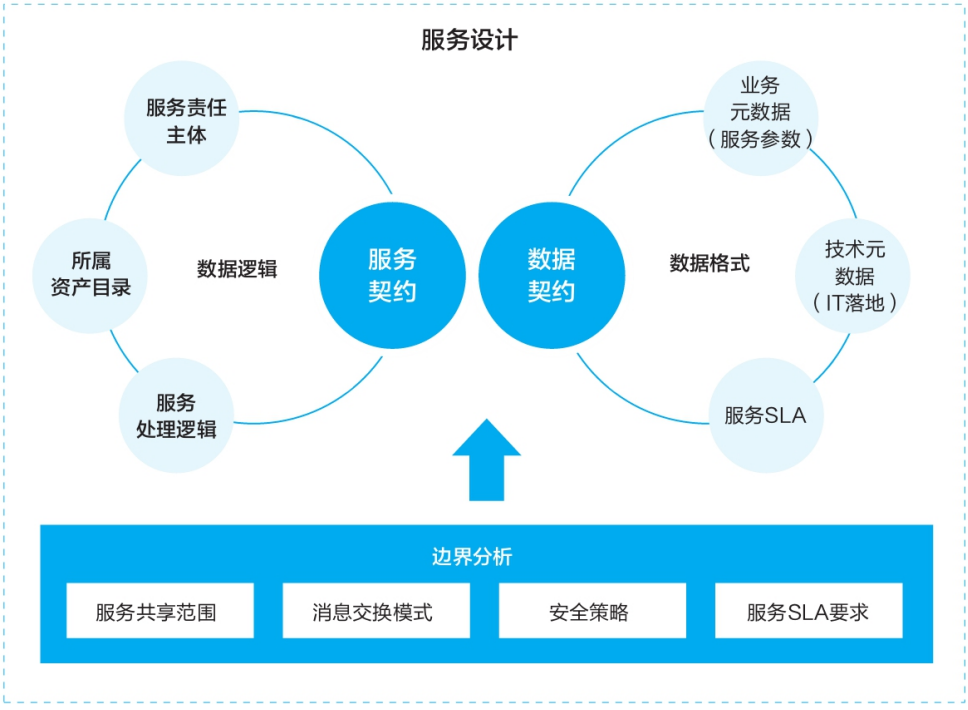


图6-9　数据服务设计的关键要素

- 服务契约：包括服务的基本信息（数据服务提供方、数据服务的类型）、能力要求（服务的时效性、服务的处理逻辑、服务的安全策略、服务的SLA要求）等。
- 数据契约：包括数据契约描述、输入和输出参数、业务数据资产编码、物理落地资产编码等。

数据服务设计中应强调数据服务的颗粒度，数据服务颗粒度的大小直接影响着服务的可重用性，细粒度的服务更容易被重用。但是，如果我们只考虑可重用性，将导致产生大量颗粒度很小的数据服务，这将对系统的整体性能带来严重的影响，因此必须在服务粒度设计上维护一种平衡。

数据服务颗粒度通常应考虑以下原则。

- **业务特性：**将业务相近或相关、数据粒度相同的数据设计为一个数据服务。
- **消费特性：**将高概率同时访问、时效性要求相同的数据设计为一个数据服务。
- **管理特性：**综合考虑企业在数据安全管理策略方面的要求。
- **能力特性：**将单一能力模型设计为一个服务。

基于上述原则，可以形成一些具体的用于指导实际执行的参考规范，如下所示。

- 同一种提供形式下，一个数据只能设计在一个数据服务中。
- 按主题（业务对象）将相同维度的数据设计为一个数据服务。如果同一个主题下的指标数量过多，则需要考虑按“高概率同时使用、业务关联度紧密”的原则再进行划分。
- 将同一个逻辑实体的数据设计为一个数据服务。
- 将单一功能的算法、应用模型设计为一个服务。

为确保服务设计后能快速、有序落地，要建立数据服务的开发、测试、部署流程，通过技术、自动化工具、管理协同机制，确保数据服务敏捷交付，缩短数据服务建设周期，如图6-10所示。

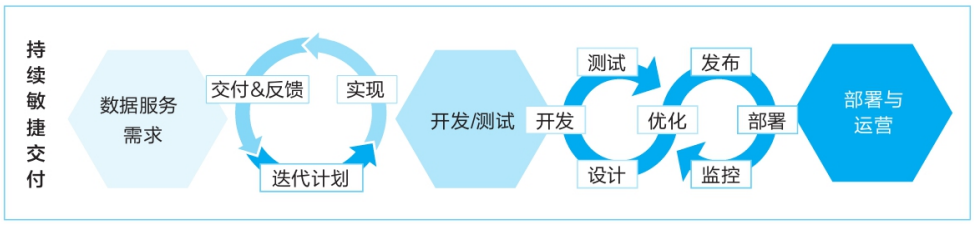


图6-10　数据服务开发及部署过程

服务开发、测试、部署中应重点构建以下能力。

- **服务需求接收与管理：**明确数据管理部门、IT部门、业务代表的具体职责，通过三方共同协作，解决需求理解不透彻导致开发过程反复的问题。
- **构建自助式开发平台：**通过简单配置的方式实现数据服务的开发，降低数据服务的开发门槛，缩短数据服务的开发

周期。

- **代码自动审查：**通过自动化手段校验服务开发代码的性能及不规范等问题，阻断代码提交，直到问题解决，做到事前规避。
- **数据自动验证：**构建数据自动测试能力，实现数据服务的数据量差异、字段差异、数据准确性差异的验证。
- **功能自动测试：**构建功能自动化测试能力，自动对数据服务SLA、查询出入参数进行自动检查，构建容错机制。
- **服务部署：**数据服务不涉及对数据的修改，采用实时部署的方式可缩短数据服务的实现周期，提升数据服务的敏捷响应能力。

3. 数据服务的变更与下架

随着业务需求的不断变化，数据服务需要随之进行调整，因此在建设中也应做好数据服务的变更管理和下架管理。

（1）数据服务变更管理

可参考以下因素：

- **服务变更内容：**包括数据服务的时效性、出入参数、服务处理逻辑、数据安全策略等。
- **服务变更影响：**业务连续性影响、变更成本影响等。

（2）数据服务的下架管理

因为数据服务是基于用户需求产生的，而业务需求是动态变化的，因此需要持续将调用量很少甚至为零的数据服务从市场中下架，确保数据消费者总能拿到“最好的”数据服务。通常，数据服务的下架有两种情况：一种是由服务消费方主动提出的数据服务下架申请，可以称为“主动下架”；另一种是通过运营度量策略判断需要下架的数据服务（例如，三个月内无服务调用、重复的数据服务等），可以称为“被动下架”。

企业应制定针对不同场景的数据服务下架流程，确保数据服务下架前进行充分的影响度评估，并具备面向所有相关方的消息通知能力，确保实际下架前各消费方的知情权。另外，还应构建一定的自动化实施能力，在各方就数据服务下架达成一致后，系统自动执行数据服务下架动作。

6.1.3 数据服务分类与建设规范

数据服务是为了更好地满足用户的数据消费需求而产生的，因此数据消费方的差异是数据服务分类的最关键因素。具体包括两大类：数据集服务和数据API服务。

1. 数据集服务

（1）数据集服务定义

比较常见的数据消费者有两类：一类是真实的“人”，一类是“IT系统”。企业越来越强调各业务部门的自我运营，因

此产生了大量自助分析消费者，这类消费者就是业务人员，甚至可能是管理者，他们通过各种数据分析工具，直接使用、消费数据。这种情况下，消费者是“访问”某个相对完整的“数据集”，这种消费方式称之为“数据集服务”。

数据集服务最主要的特征是由服务提供方提供相对完整的数据集合，消费方“访问”数据集合，并自行决定接下来的处理逻辑，如图6-11所示。

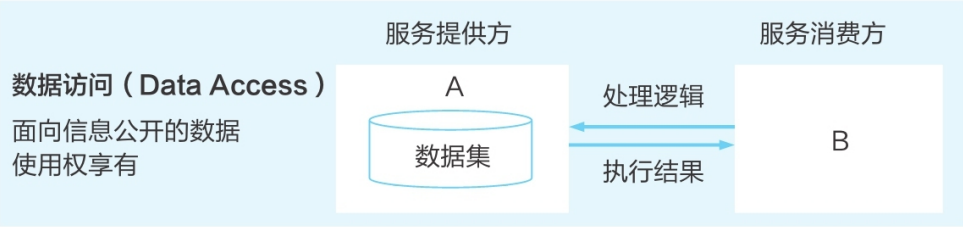


图6-11　数据集服务特征

- 数据服务提供方被动地公开数据以供数据消费方检索。
- 数据服务提供方并不定义数据处理逻辑，但数据和数据处理逻辑仍然由其控制。
- 数据服务的生命周期即数据访问授权的有效期。

举例来说，数据服务供应方提供信息搜索、查询服务，但并不清楚用户的真实意图，用户可以自由地在服务提供方的地盘上“玩”数据。

（2）数据集服务建设规范

数据集服务主要面向自助分析场景提供相对完整的数据集合，因此所提供的数据主要来自数据底座，包括“数据湖”和

“主题联接”。

当所提供的数据来自数据湖时，建设规范如图6-12所示。

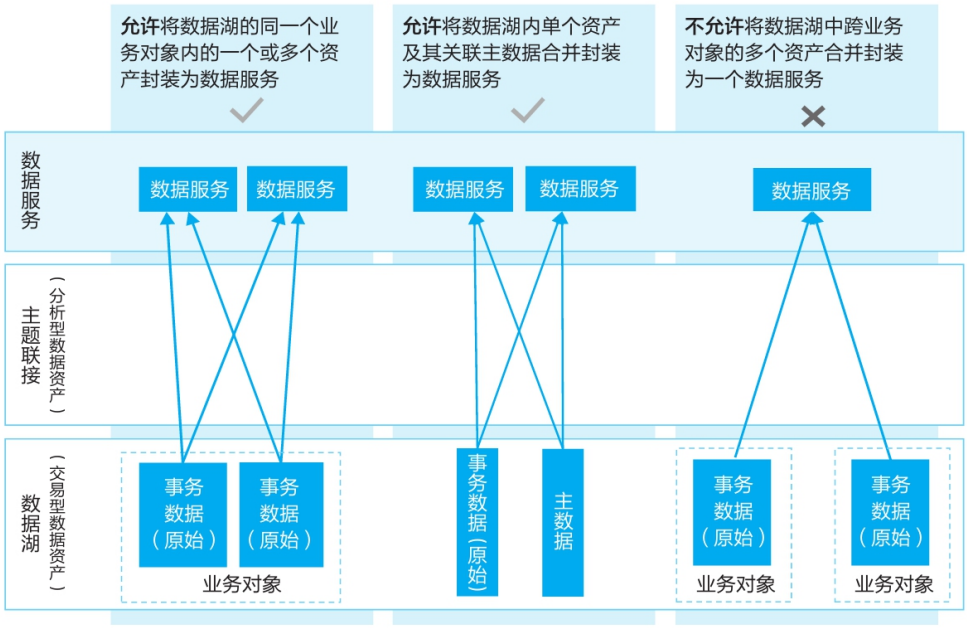


图6-12　数据集服务建设规范（数据湖）

1）允许将数据湖的同一个业务对象内的一个或多个资产封装为数据服务。

在部分实时性要求极高的场景下，例如，对于某个地区所有销售投标项目的实时状态可视化场景，可以将“投标项目（Proposal）”这个业务对象下的多个逻辑数据实体封装在一起，设计成可以支撑投标的实时可视化的数据服务。

2）允许将数据湖内单个资产及其关联主数据合并封装为数据服务。

在部分实时数据服务需求场景下，需要向用户提供相对完整的主数据或基础数据信息，以便于用户自助分析。例如，某个业务部门可能需要交付项目实施计划的数据服务，以便进行实时监控和指挥。当通过IT系统或应用实现该功能时，只需获取数据湖中原始的事务数据（交付项目实施计划明细），但在自助分析场景下，由于数据服务面对的是具体的业务人员，而业务人员不可能读懂任务ID、区域组织ID等物理层主键或外键，并且没有必要让每个自助分析人员都重复进行共性数据联接，因此可以在数据服务封装时，将必要的数据联接在一起，比如将“任务与任务资源关系”或“任务与区域组织关系”与交付项目实施计划明细合并封装为一个数据集服务。

3）不允许将数据湖中跨业务对象的多个资产合并封装为一个数据服务。

要注意数据服务合并封装的边界，数据服务的本质是将已有数据资产以服务的形式提供给消费者，而不是在服务中创建一个新的数据资产，面向OLAP的数据资产创建应该在数据主题联接完成，这在一定程度上也可以避免出现数据服务大量重复建设的情况。

当所提供的数据来自于主题联接时，建设规范如图6-13所示。

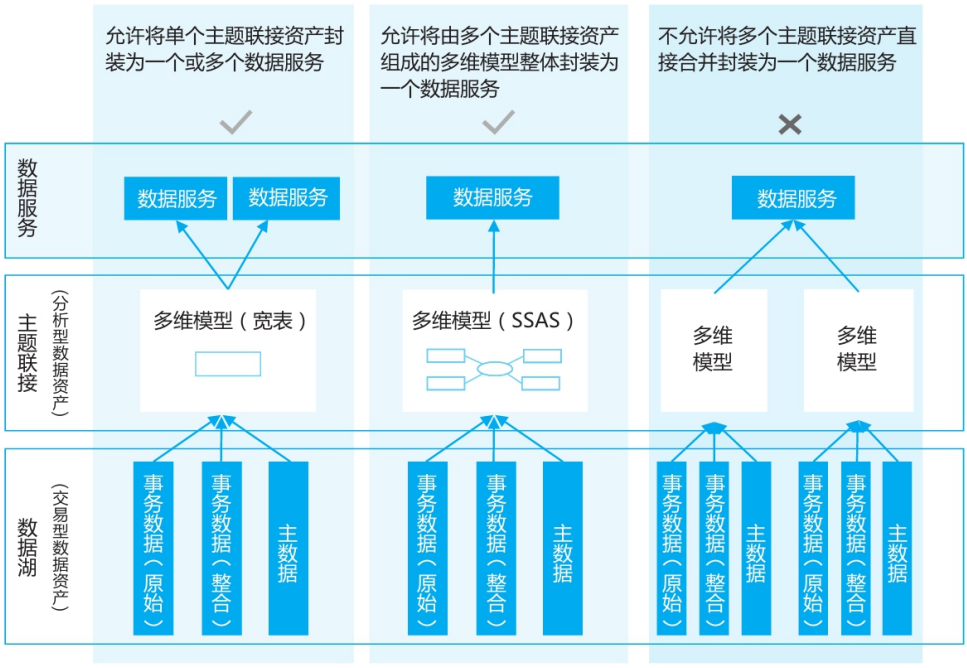


图6-13　数据集服务建设规范（主题联接）

1）允许将单个主题联接的数据资产封装为一个或多个数据服务。

数据服务在面对不同消费者的不同需求时，可以适当地拆分为多个数据服务，以便更好地提供给数据消费者，减少冗余数据，提升用户体验。例如，在封装“区域损益明细实际数据”服务时，集团职能部门和具体业务部门的需求可能是不同的，具体业务部门不需要精细到产品L3以下的明细数据。如果把产品L1～L5的所有明细都提供出来，数据量将会以百倍的规模增加，会极大地影响数据分析性能，这显然是不必要的。比较恰当的方式是将两类需求分别封装为不同的数据服务，并确保这些数据服务的数据来源于同一个主题联接数据资产。

2）允许将由多个主题联接数据资产组成的多维模型整体封装为一个数据服务。

在部分情况下，主题联接数据资产并不是以宽表的形式落地，而是以多维模型的形式存在，此时可以将多维模型整体封装为一个数据集服务。例如，可以将“预测多维分析模型”中的“区域组织维表、产品维表、预算事实表”等封装为一个服务，满足区域组织经营管理的需要。

3）不允许将多个主题联接数据资产直接合并封装为一个数据服务。

数据资产之间的联接属于主题联接范畴，应该首先沉淀为公共数据主题联接资产，再封装为服务。

2. 数据API服务定义

数据服务的另外一类消费者是“IT系统”，即面向某个IT系统提供数据事件驱动的“响应”，这种服务的封装方式与前面所提到的数据集不同，称为“数据API服务”。

（1）数据API服务特征

服务提供方“响应”消费方的服务请求，提供执行结果，如图6-14所示。

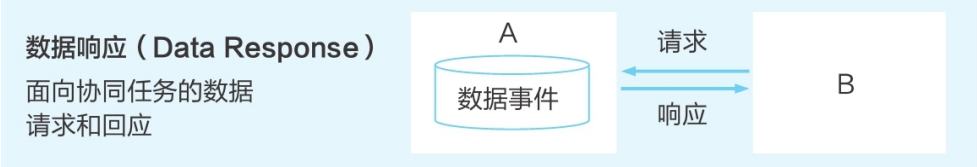


图6-14　数据API服务特征

- 数据服务提供方基于随机的数据事件主动地传送数据。
- 数据服务提供方会基于事件定义数据处理逻辑，由消费方提前订阅并随机触发。
- 服务的生命周期跟着事件走，事件关闭了，服务就终止了。

例如，华为公司给OBS（Object Storage Service，对象存储服务）提供面向客户的服务能力评估和报价复核服务。

数据API服务是对用户随机数据事件的响应，这个需求往往伴随着用户的某个任务产生，随着任务的结束，整个服务也就完成了。通过数据API服务，用户可以及时地获知任务的协同情况，并基于服务方的反馈结果，做出相应的调整。服务供给方和消费方是协同关系（互操作），而非交接棒关系（交换情报），有效提升了面向协同任务的互操作一致性。

（2）数据API服务VS数据集成服务

数据API服务与传统系统集成相比有非常明显的优势，如图6-15所示。

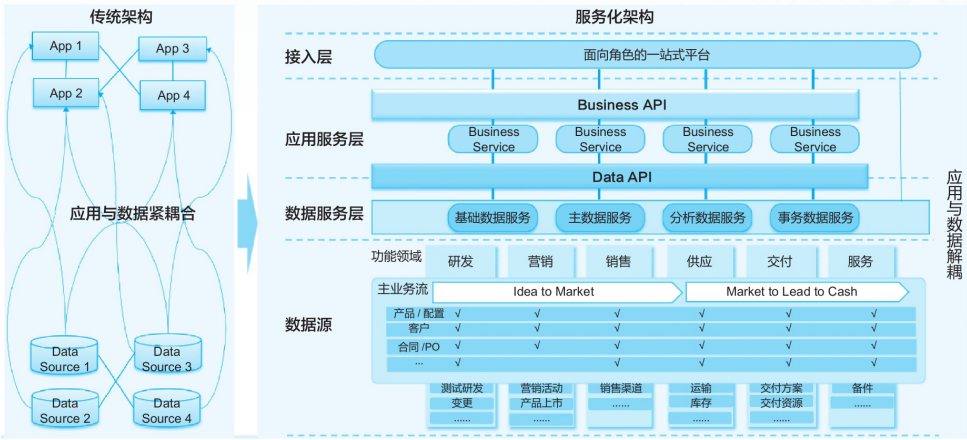


图6-15　数据API服务与传统集成方式的对比

- **供应/消费数据服务：**应用组件间传递的是基于数据服务契约的消息，即传递对数据进行逻辑操作的结果。
- **高聚合：**订单服务使业务逻辑变得更加集中，易于数据同源管控。
- **松耦合：**业务逻辑的变化对服务消费方没有直接影响。

数据API服务的设计规范业界相对统一，不在这里详细说明。

6.1.4　打造数据供应的“三个1”

数据服务改变了传统的数据集成方式，所有数据都通过服务对外提供，用户不再直接集成数据，而是通过服务获取。因此，数据服务应该拉动数据供应链条的各个节点，以方便用户能准确地获取数据为重要目标。

数据供应到消费的完整链条如图6-16所示，当用户所需数据处于链条上的不同节点时，提供服务的周期是有差异的。

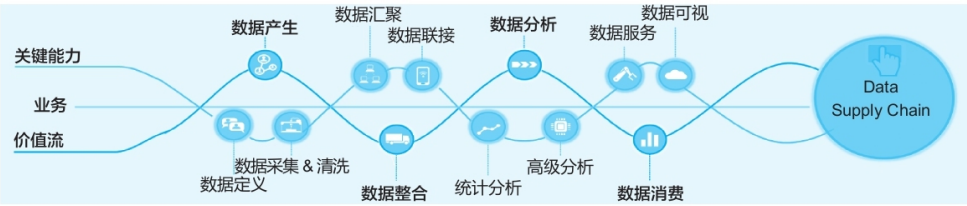


图6-16　数据供应链（资料参考：埃森哲数据供应链方法论）

华为公司为确保整个数据供应链条的高效协同，制订了“三个1”作为拉通各个供应环节的整体目标，确保每个环节能够形成合力并对准最终用户，如图6-17所示。

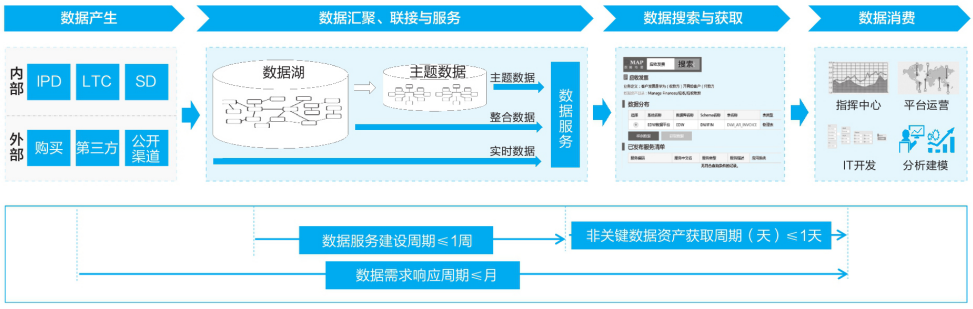


图6-17　数据服务供应SLA

“三个1”是数据供应的整体目标，起点是需求方提出数据需求，终点为需求方拿到数据并可立即进行消费，具体衡量标准包括如下内容。

- 1天：对于已发布数据服务的场景，从需求提出到消费者通过服务获取数据，在1天内完成。
- 1周：对于已进底座但无数据服务的场景，从需求提出到数据服务设计落地、消费者通过服务获取数据，在1周内完

成。

- 1月：对于已结构化但未进底座的场景，从需求提出到汇聚入湖、数据主题联接、数据服务设计落地、消费者通过服务获取数据，在1个月内完成。

数据供应的“三个1”并不是单纯的度量指标，而是一整套瞄准最终数据消费体验的能力体系以及确保数据供应能力的管理机制，还包括组织职责的明确、流程规范的制定与落实、IT平台的建设和管理，如图6-18所示。

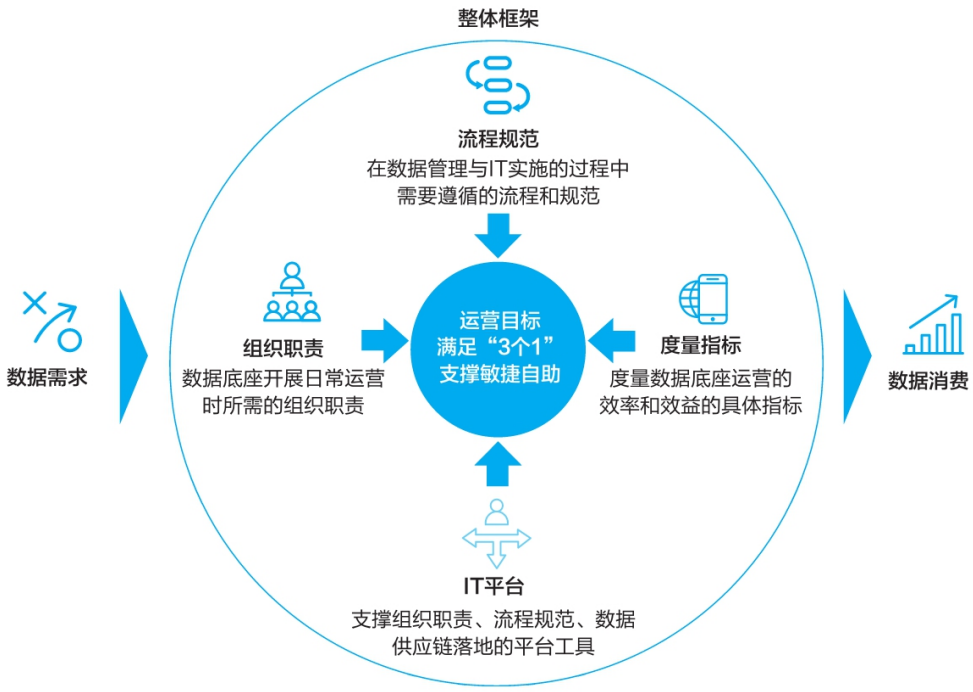


图6-18　实现数据服务供应SLA的关键要素

（1）组织职责的明确

- 构建专业的评审及仲裁组织。
- 承接各细分工作内容的角色职责。

（2）流程规范的制定与落实

- 统一的工作细分流程。
- 配合工作流程制定相应的管理规范，以指导开展工作。
- 配合IT平台制定相应的管理规范，以指导开展工作。

（3）IT平台的建设

- 度量、评价数据底座运营的效率和效益的具体指标。
- 支撑组织职责、流程规范、度量指标落地的IT工具。

（4）面向需求方的效率承诺度量

对所有供应团队形成明确、清晰的工作要求。在华为的内部实践中，会以图6-19为例进行明确的承诺，并在公司层面进行公示，请所有数据需求方共同对供应能力进行监督。

数据供应SLA承诺值			
IT活动	承诺范围	承诺时间	SLA承诺值
数据主题联接资产建设	全球	2018年	7天
数据集成申请&实施	全球	2018年	1天
数据服务封装实施	全球	2018年	3天
数据按租户授权切片	全球	2018年	1天

图6-19　数据供应团队面向消费方的SLA承诺示例

6.2　构建以用户体验为核心的数据地图

在解决数据的“可供应性”之后，企业应该帮助业务更便捷、更准确地找到它们所需要的数据，这就需要打造一个能够满足用户体验的“数据地图”。

6.2.1　数据地图的核心价值

数据供应者与消费者之间往往存在一种矛盾：供应者做了大量的数据治理工作、提供了大量的数据，但数据消费者却仍然不满意，他们始终认为在使用数据之前存在两个重大困难。

1）找数难

企业的数据分散存储在上千个数据库、上百万张物理表中，已纳入架构、经过质量、安全有效管理的数据资产也会超过上万个，并且还在持续增长中。例如，用户需要从发货数据里对设备保修和维保进行区分，以便为判断哪类设备已过保（无法继续服务）提供准确依据，但生成和关联的交易系统有几十个，用户不知道应该从哪里获取这类数据，也不清楚获取的数据是否正确。

2）读不懂

企业往往会面对数据库物理层和业务层脱离的现状，数据的最终消费用户无法直接读懂物理层数据，无法确认数据是否能满足需求，只能寻求IT人员支持，经过大量转换和人工校验，才最终确认可消费的数据，而熟悉物理层结构的IT人员，并不是数据的最终消费者。例如，当需要盘点研发内部要货情况的时候，就需要从供应链系统获取研发内部的要货数据，但业务用户不了解该系统复杂的数据存储结构（涉及超过40个表、1000余个字段），也不清楚每个字段名称下所包含的业务的含义和规则。

企业在经营和运营过程中产生了大量数据，但只有让用户“找得到”“读得懂”，能够准确地搜索、便捷地订阅这些数据，数据才能真正发挥价值。

数据地图（DMAP）是华为公司面向数据的最终消费用户针对数据“找得到”“读得懂”的需求而设计的，基于元数据应用，以数据搜索为核心，通过可视化方式，综合反映有关数据的来源、数量、质量、分布、标准、流向、关联关系，让用户高效率地找到数据，读懂数据，支撑数据消费。

数据地图作为数据治理成果的集散地，需要提供多种数据，满足多类用户、多样场景的数据消费需求，所以华为公司结合实际业务制定了如图6-20所示的数据地图框架。

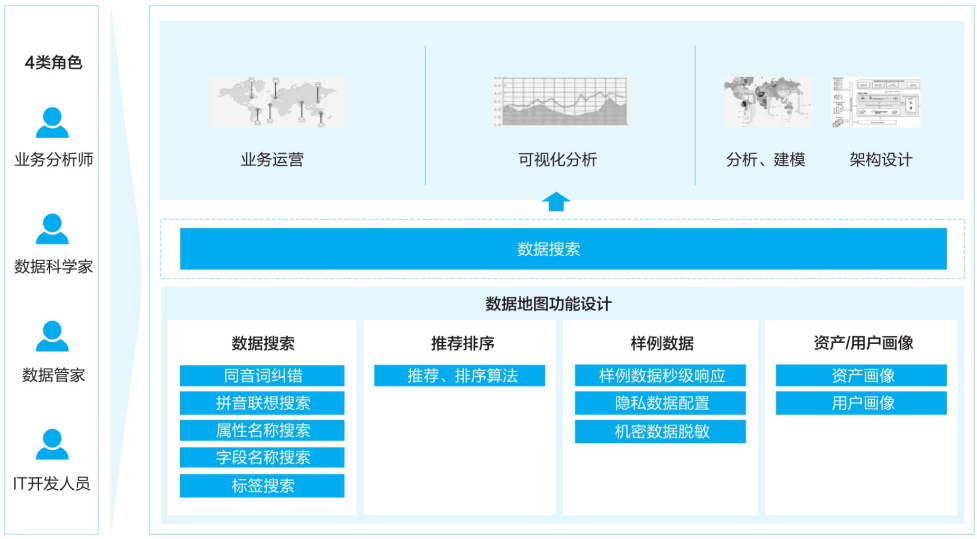


图6-20　数据地图整体框架

数据地图为四类关键用户群体提供服务。

1）业务分析师

业务分析师是企业最大的数据消费群体，具有良好的业务背景，有些数据分析师本身就是业务人员，了解业务需求实质，理解业务含义，与利益相关者有良好的沟通。通过对数据的识别，借助数据分析工具，生成可供阅读的图表或者仪表板，使用分析结果识别问题，支撑决策。对数据可信度、业务含义、数据定位有强烈诉求。

2）数据科学家

数据科学家是指能采用科学方法、运用数据挖掘工具对复杂异构的数字、符号、文字、网址、音频或视频等信息进行数字化重现与认识，并能进行新的数据洞察的工程师或专家。对业务含义、数据关系有强烈诉求。

3）数据管家

公司数据管理体系的专业人员，负责协助数据Owner对数据信息架构进行管理，包括定义信息架构中的责任主体、密级/分类，为数据安全管理提供重要输入。通过信息架构设计，统一业务语言，明确管理责任，设定数据质量标准，拉通跨领域信息流，支撑运营和决策。对数据质量、信息架构、数据关系有强烈诉求。

4）IT开发人员

主要为企业的数据仓库开发人员，通过对物理表进行定位、识别和ETL，创建满足业务分析师或者应用平台所需要的模型或维表。对数据定位、数据关系有强烈诉求。

6.2.2　数据地图的关键能力

数据地图重点提供数据搜索、排序推荐、数据样例、资产/用户画像等关键能力。

（1）数据搜索

数据搜索可以提高用户的搜索准确度，使用户能快速理解搜索出来的数据内容，通过组合搜索、筛选分类，数据标签等持续提升用户体验。

通过界面封装搜索引擎，只向用户暴露单一的搜索栏，通过搜索栏的单一或者组合搜索，发现数据。

以图6-21为例，当用户搜索“数据标准”时，既可以精确匹配名称的资产，通过关联搜索带出完全匹配的资产并进行展示，也可以在输入的关键词无法直接匹配逻辑实体或者物理表名称的情况下，执行模糊逻辑搜索，对所涉及的前分词、后分词、中间分词进行匹配，除了逻辑实体名称，也会涉及属性名称、业务描述等更多内容的匹配。当没有完全匹配的直接资产时（如“人员”），会根据前后分词进行搜索，这样整体的结果记录会比较多，并会涵盖搜索属性名称或者业务定义中的“人员”关键词。



图6-21　数据搜索结果示例

（2）排序推荐

排序推荐能让用户更容易地找到高质量、可消费的数据资产，缩小搜索结果集范围，减少数据识别和判断的时间，最终目标是让用户实现“所搜即所得”的效果。

对应搜索结果的推荐排序，主要在功能侧提供了两类服务，以便用户通过被动式和主动式的办法管理搜索结果。

1）被动响应推荐排序

用户在前端无须操作，通过推荐排序逻辑对结果集进行处理，完全基于数据管理分类、用户行为分析等输入。优点是提升了用户的体验，无须操作即可以大概率定位到自己需要的数据资产；缺点是与用户之间缺乏交互，准确度因人而异，需要积累大量管理分类和用户行为的输入作为参考。

2）主动管理推荐排序

通过数据管理侧的分类和通用性的标签进行分类管理，用户在使用过程中可以通过分类标签对搜索结果集进行再次过滤和定位。优点是与用户有一定的交互，让用户在使用的时候可以主动管理；缺点是基于管理侧和通用性收敛上来的标签难以满足个性化的需求。

接下来我们看两个示例。

示例1：以属性名称组合搜索为例，一组属性名称串 联起来，连用“订单履行经理，BU，CF_EPD，ETRAK标识”组合搜索，结果集中全匹配、部分匹配的结果会按照前后的顺序进行

排列，匹配程度越高的数据资产排序会越靠前，如图6-22所示。



图6-22　数据搜索结果示例1

示例2：以“合同”为关键词，相同关键词匹配程度、相同密级的情况下，合同维表的消费频率高于合同与项目关联信息的消费频率，排序优先，如图6-23所示。



图6-23　数据搜索结果示例2

（3）数据样例

“读懂数据”是用户进行数据消费的基础，用户只有读懂数据，才可以准确判断数据的来源、质量、可信度等关键信息。除了可以通过提供数据资产的各类元数据信息来辅助用户读懂数据外，生产环境的实时数据对用户而言更有参考价值，

因为生产环境直接采集的数据的内容，与用户可消费的数据内容是一致的。

接下来看一个样例数据查看的示例：

用户在搜索结果中点击样例数据，能够自动读取数据库名称，并根据对象编号，查找数据库记录表，实现生产环境数据的样例查看，如图6-24所示。

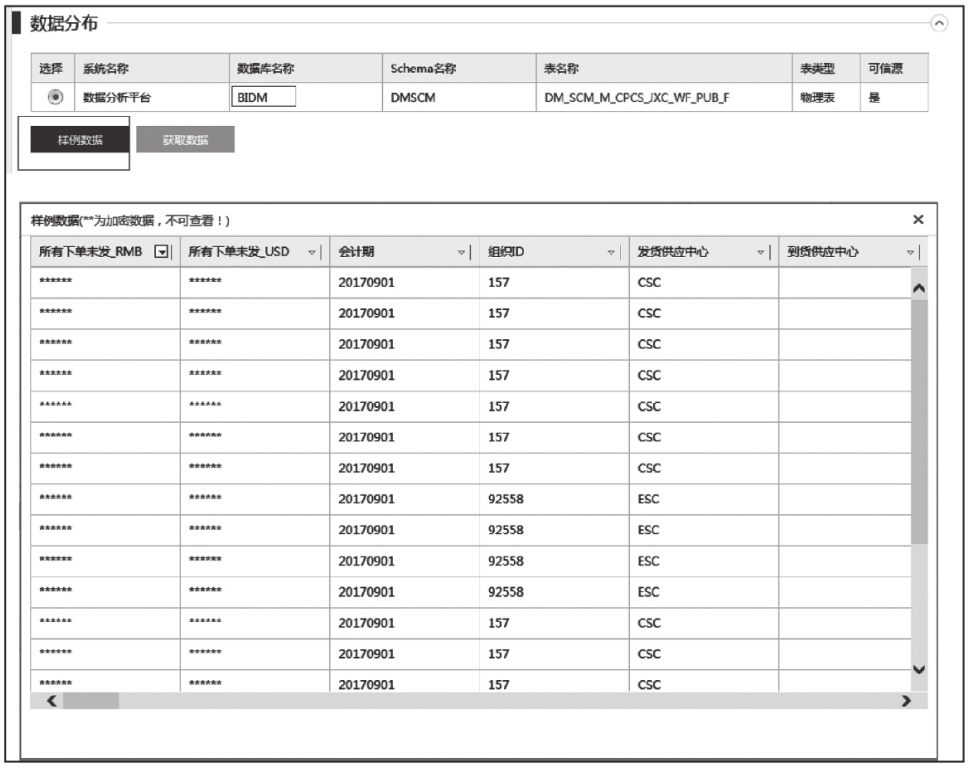


图6-24　样例数据查看示例

（4）资产/用户画像

资产/用户画像通过标签化的手段来对资产和用户进行清晰地描绘，有助于数据搜索和推荐排序的不断优化，贴近用户真

实需求。

基于用户画像、经验模型库和资产画像理解文本语义，可以提高搜索准确性，解决资产查不到、难挑选等问题，并通过业务运营不断优化资产搜索能力，如图6-25所示。

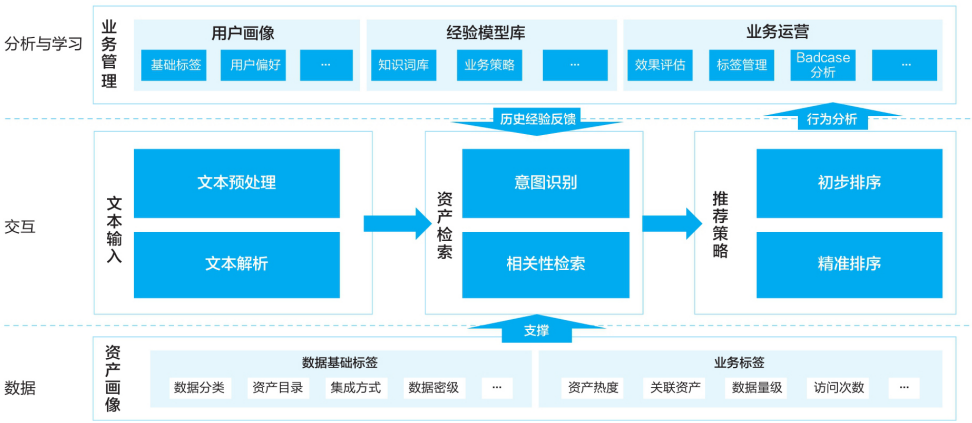


图6-25　资产/用户画像框架

6.3　人人都是分析师

数据服务解决了“可供应性”，数据地图解决了“可搜索/可获取性”，当消费方获取数据后，提供“可分析”能力，帮助数据消费者结合自身需要获取想要的分析结果。

6.3.1　从“保姆”模式到“服务+自助”模式

过去，各业务部门的分析诉求往往通过公司总部“保姆式”开发模式来满足，即业务部门只负责提出需求，所有的方案从设计到开发实现，统一由总部完成。这也是传统意义上的数据仓库的标准报告生成方式，强依赖于IT人员，贯穿整个数据分析过程，从获取数据、建模到设计报告，均需要IT人员的支撑，如图6-26所示。这种模式存在多个问题，如下所示。

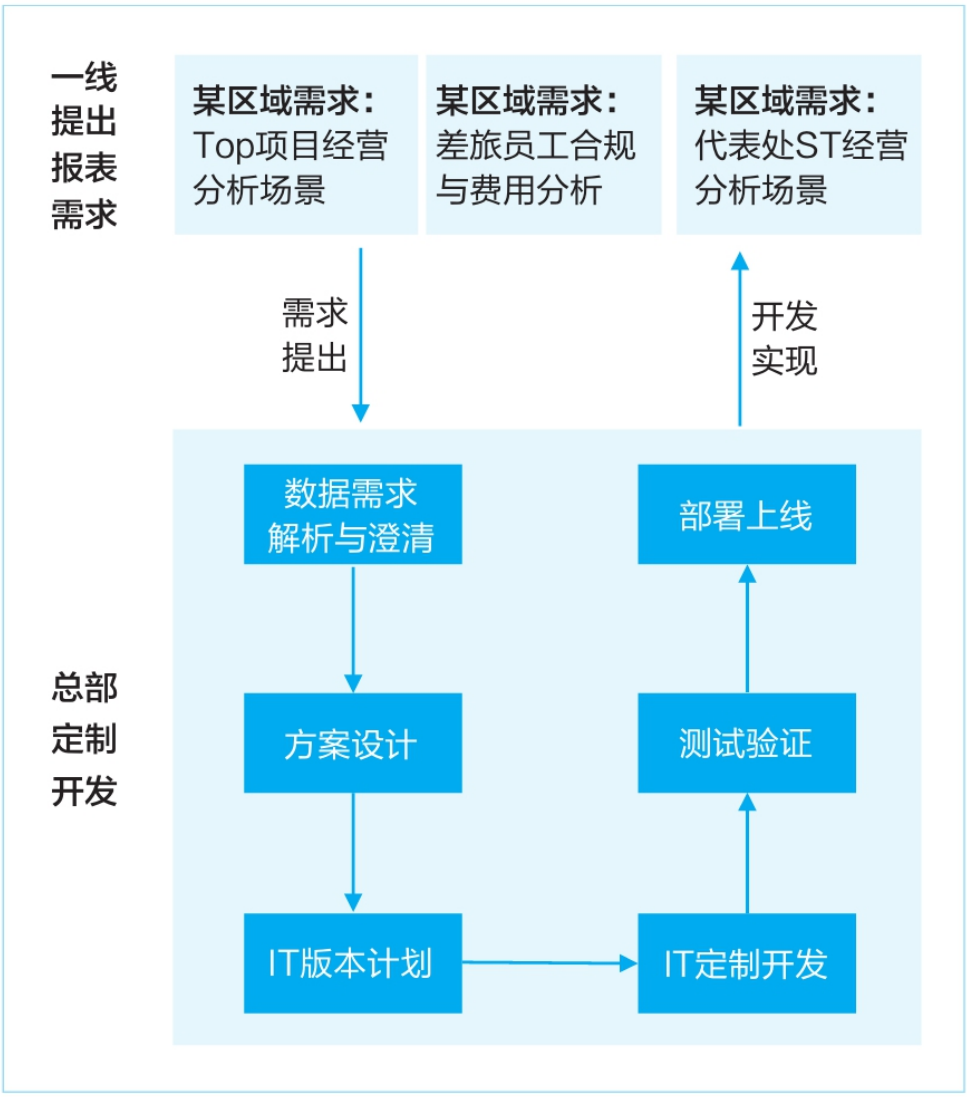


图6-26　传统总部定制开发模式

1）总部开发周期长，通常从需求提出到开发实现，需要多轮次需求解析和澄清。由于总部并不了解业务部门的实际业务，即使在方案设计阶段也可能需要再次对需求进行确认。IT开发完成后还需要进行严格的测试验证和部署，因此整个周期通常最短也需要30天左右。

2）无法满足灵活多变的业务要求。业务运营和业务作业不同，作业模式相对稳定，当大的场景不发生变化时，作业模式是基本稳定的。而业务运营是按需开展的，往往是从问题出发，在业务开展过程中，可能出现的问题、风险是经常变化的，很可能任何一个内外部因素的变化就会带来新的业务运营关注点，而总部开发模式不可能实时满足所有区域的要求。

在这种背景下，提出了“服务+ 自助”模式（如图6-27所示），即公司总部只提供统一的数据服务和分析能力组件服务，各业务部门可以根据业务需要进行灵活的数据分析消费，数据分析的方案和结果由业务自己完成。这一模式有如下价值。

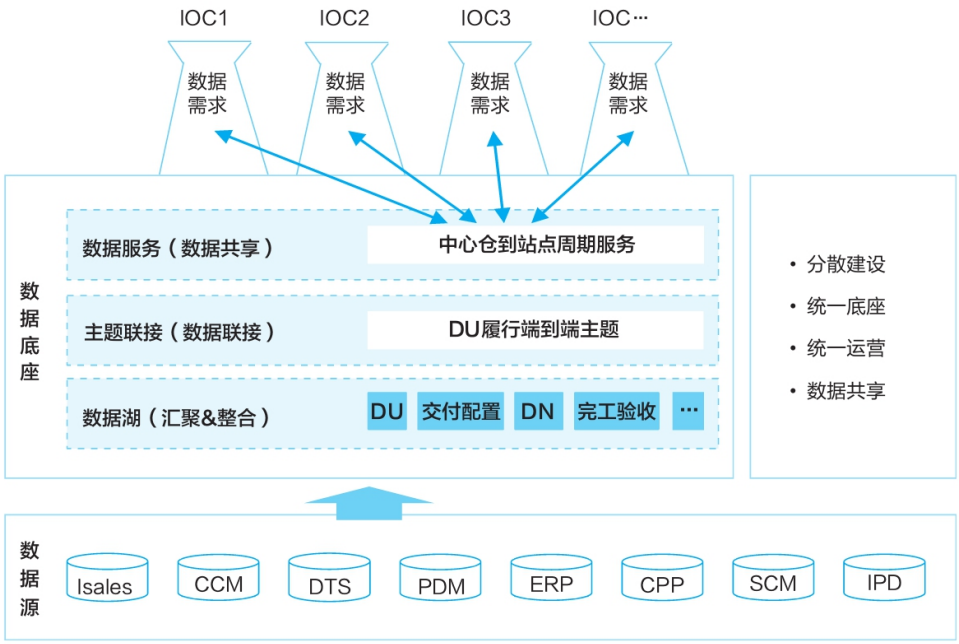


图6-27 “服务+自助”模式

1）数据分析消费周期极大缩短。当各业务部门需要进行数据分析消费时，可以直接调用已建好的数据服务进行自助分

析，整个报表开发周期缩短为1～2天。

2）发挥业务运营主观能动性。俗话说“高手在民间”，各业务部门是业务作业的责任主体，同时也对业务及经营结果负责，因此各业务部门是业务运营的第一责任人，同时也是最了解业务自身现状与问题的。通过自助模式，可以更有效地发挥各业务部门的主观能动性，真正将数据分析消费与业务运营改进相结合。

3）减少“烟囱式系统”的重复建设。各业务部门在保证数据分析消费灵活性的同时，并不需要重复构建支撑消费的数据基础，所有公共的数据汇聚、数据联接都统一建设，在遵从隐私保护和安全防护要求的前提下以数据服务的形式充分共享。

6.3.2 打造业务自助分析的关键能力

华为公司将自助分析作为一种公共能力，在企业层面进行了统一构建。一方面，面向不同的消费用户提供了差异性的能力和工具支撑；另一方面，引入了“租户”概念，不同类型的用户可以在一定范围内分析数据、共享数据结果。

1. 针对三类角色提供的差异性服务

面向三类角色的分析架构能力如图6-28所示。

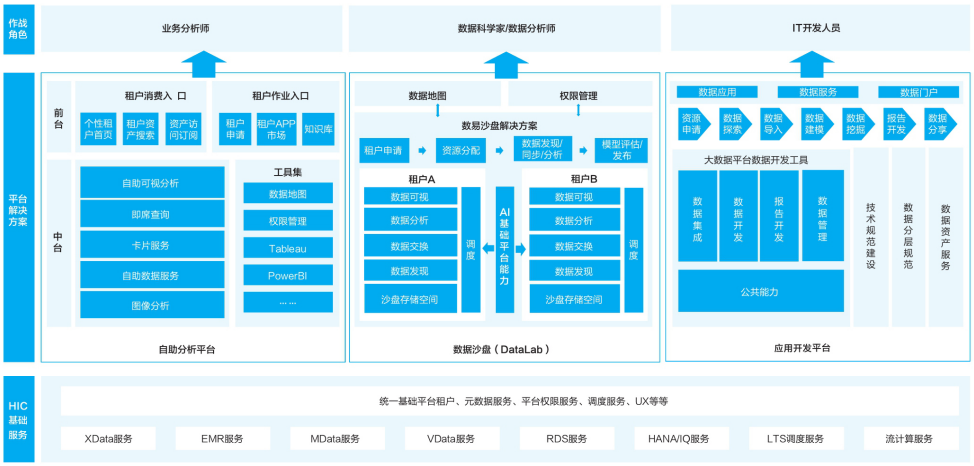


图6-28　面向三类角色的分析能力架构

（1）面向业务分析师，提供自助分析能力，业务人员通过“拖、拉、拽”即可快速产生分析报告

- 基于多租户环境，提供数据资产订阅、报表作品搜索、服务订阅等能力。
- 实现从数据查询到数据拖拽式分析的端到端的一站式自助作业，增强数据即席查询和数据建模等功能。
- 提供数据搜索、数据获取、自助分析、数据消费等一站式自助分析服务，缩短报表开发周期。
- 支持租户管理、工具集管理、日志管理功能，集成数据底座权限模型，提供稳定的分析环境。

（2）面向数据科学家，提供高效的数据接入能力和常用的数据分析组件，快速搭建数据探索和分析环境

- 集成数据可视化、数据建模能力，降低数据分析门槛，提高平台的易用性。

- 识别公共诉求，提供R Studio、Zeppelin等工具集，增强NLP基础服务、人工智能等分析装备对于机会点的支撑能力，支撑各种大数据分析应用场景。
- 提供源系统到分析平台的数据实时同步功能。
- 为数据科学家提供数据目录导航入口。
- 提供数据分析环境，支持权限 申请和计算资源的分配，缩短建模周期。

（3）面向IT开发人员，提供云端数据开发、计算、分析、应用套件，支撑海量数据的分析与可视化，实现组件重用

- 整合数据接入、数据计算、数据挖掘、数据展现等能力，提供高效、安全的数据集成、数据开发、报告开发、数据管理等服务，减少重复建设，实现组件重用。
- 整合第三方资源，依托HIC能力通道，提供自助、按需、在线的基础数据服务，包括分布式处理、实时处理、内存计算等。

2. 以租户为核心的自助分析关键能力

（1）多租户管理能力

租户是指把数据、分析工具、计算资源有机组合的工作环境，用户可以在租户内自助完成数据搜索、数据加工、在线分析、报表共享等工作。

多租户技术也称多重租赁技术，是一种软件架构技术。多租户技术可以实现多个租户之间共享系统实例，同时也可以实现租户的系统实例的个性化定制。通过使用多租户技术可以保证系统共性的部分被共享，个性的部分被单独隔离。例如，按国家设定不同租户，这样在本租户内共享该国的经营分析结果，共同进行异常分析和经营改进；同时，该租户数据对其他国家屏蔽，避免了数据扩散等安全风险。

为了合理分配软硬件资源，满足各领域在线、自助、个性化的数据分析诉求，促进数据的安全共享和价值变现，明确了租户申请、租户命名、数据准备、数据同步、数据加工、数据申请、权限管理、安全与隐私、运维与运营等方面的要求，旨在通过正确的引导，确保数据消费的便捷、高效与安全合规，支持公司的数字化转型。

在多租户建设中，相对于技术层面的解决方案，租户管理的职责需要在企业里建立共识，将共识以标准规范的形式固化下来。租户自助分析能力架构如图6-29所示。

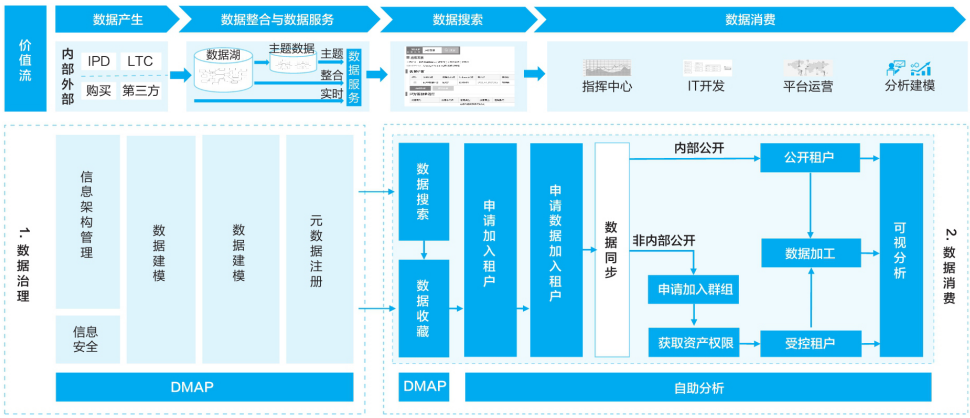


图6-29　租户自助分析能力架构

租户的4个关键角色如下所示。

- **租户Owner：**租户管理的第一责任人，由公司正式任命的管理者或者变革项目经理担任，是租户内数据消费的总责任人。
- **租户管理员：**由租户Owner指定并授权，是对租户内资产、用户、报告的日 常维护、配置、授权承担具体管理职责的人员。
- **查看者：**申请并被允许加入租户，只对租户内的报告有查看权限 的租户用户。
- **分析师：**申请并被允许加入租户，对数据资产可执行申请数据入租户、申请租户授权、通过分析工具分析数据、制作报告、查看报告、分享报告等操作的租户用户。

（2）数据加工能力

在同一个租户空间内，对数据进行关联、过滤等操作，满足最终分析报告的数据需求。

用户可将多个数据进行关联，构建自己的宽表，可对宽表进行数据过滤，选择合适的字段以及增加计算字段，如图6-30所示。

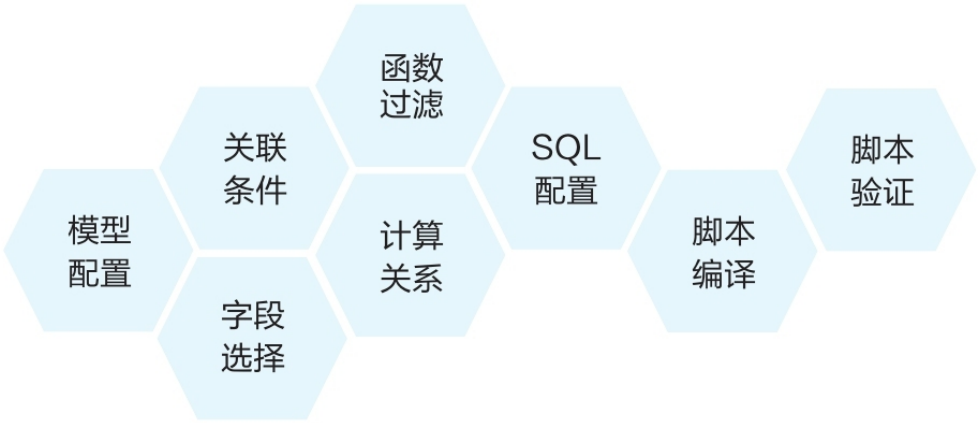


图6-30　数据加工关键能力

（3）数据分析能力

基于消费场景，利用租户内授权的数据资产，通过分析工具对数据进行分析并生成可视化报告。

用户可以选择即席查询自行配置各类条件后的结果数据，再基于这些数据直接链接到不同的分析工具，进行进一步的数据分析。

1）即席查询

提供通过筛选条件展示结果数据的能力，如图6-31所示。

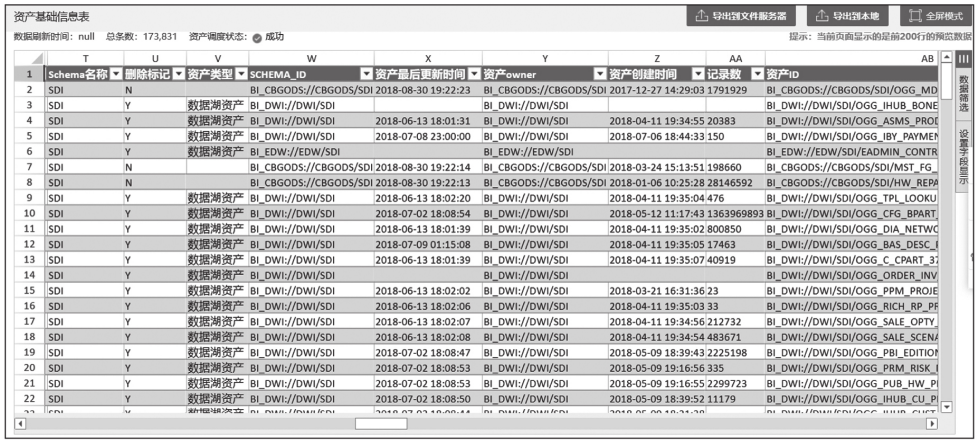


图6-31　即席查询样例（通过对Schema名称过滤，保留名称为SDI的技术资产）

- 提供生产环境的实时数据内容，有助于用户通过筛选后的结果数据判断能否满足最终的分析需求。
- 分析结果支持以文件服务器的方式下载，满足本地化处理的需求，同时避免数据被过度共享。

2）可视分析

查看已授权并加工好的数据的详情，进入可视化分析阶段，充分利用企业现有的分析工具，或打通主流的商业分析工具，减少开发成本，降低学习成本，如图6-32所示。



图6-32　可视分析案例

- 数据打通，已授权加工后的数据可以直接进入分析工具进行分析操作。
- 最大程度利用各种分析工具的已有功能。

（4）自助分享能力

基于自助分享能力，可以对分析报告进行密级设定和权限管理，向租户个人或者群体分享报告，不仅可以分享给本租户内指定的用户，而且可以进行跨租户分享。这样一方面可以扩大报告的使用范围，降低报告重复建设过程中的成本，另一方面也有助于解决分析结果不一致的问题。

- 对报表提供浏览和编辑能力，查找需要浏览的报表，选择查看、编辑、分享、删除功能。
- 提供对生成的报告定义密级的能力，报告生成者作为报告的Owner，定义密级和管控分享范围。

6.4　从结果管理到过程管理，从能“看”到能“管”

数据分析和消费本身只是一种手段而不是目标，数据消费要真正产生价值，必须与业务密切结合。业务数字化运营是华为公司数据消费的最重要场景之一。

6.4.1　数据赋能业务运营

业务运营的本质是围绕业务战略“RUN”流程。运营过程中业务沿着流程周而复始地运转，在作业过程中识别并解决问题，应基于PDCA循环（戴明环）进行有质量的运营。而数字化运营旨 在利用数字化技术获取、管理和分析数据，从而为企业的战略决策与业务运营提供可量化的、科学的支撑。

数字化运营归根结底是运营，旨 在推动运营效率与能力的提升，所以它体现在具体的业务中，而不是一块新的业务，更多是在现有标准流程的基础上改进和完善。数字化运营的核心是数据，以及基于数据的精细化管理和科学决策分析。企业业务数字化运营模式如图6-33所示。



图6-33　企业业务数字化运营模式（资料参考：罗兰贝格）

业务数字化运营的目的不应只有“察（数字化看板）”，还应该进一步实现“打（业务决策与执行）”，即支撑业务运营作战模式转变，提升运营效率。业务数字化运营要发挥对业务的指挥作用，要能够让上下同步感知业务运行态势，通过分工协作解决业务运作中的问题，提升运作效率。

业务数字化运营要同时具备多个能力，包括战略落地、业务可视化、预测预警、作业指挥、跨领域问题解决和联动指挥等。

基于数据底座的数字化运营模式（如图6-34所示）支撑着华为公司大量相对独立的业务作战单元和业务部门，基于自身业务进行持续运营，提升业务效率和业务盈利水平。同时，也解决了过去普遍存在的“线下会议多、手工报告多”等问题，避免让大量作战人员将精力消耗在事务性工作中。

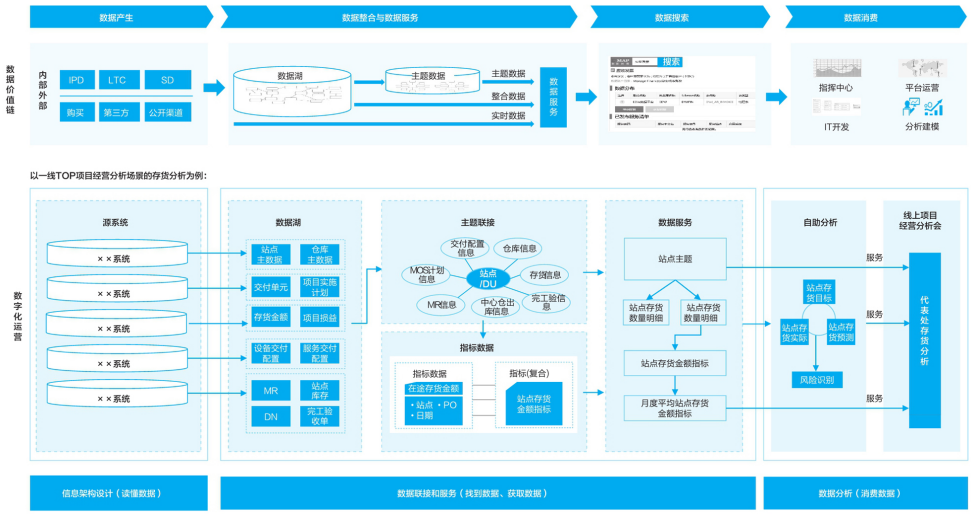


图6-34　基于数据底座的数字化运营过程

（1）满足业务运营中数据实时可视化的需求

过去，数据的展现时间与数据在业务中的真实发生时间往往有相当长的时间间隔。某个具体的作业数据进入系统后，要经过复杂的集成和长时间的等待才能体现在报表中，业务部门无法第一时间掌握真实的工作进度，也就无法在第一时间进行业务管理。例如，分包商进行站点作业时，相关数据往往要第二天或更晚才能体现在监控报告中，等到平台能根据这些信息采取行动时，也无法在作业时发挥作用。假如，站点出现物料问题或安装质量问题，可能会导致分包商重复上站，额外增加了成本。通过实时数据入湖和联接方案，业务可以第一时间获取作业监控信息，从而根据实际情况进行灵活调整，一次性把工作做对。基于实时数据的可视化能力，还能够支持业务在线会议，通过数据底座提供的服务将数据层层下钻，从过去的纸面报告或PPT材料变成可视化的业务场景。

（2）满足业务运营中及时诊断预警的需求

通过数字化手段，在获取业务运营数据的同时，可以根据实际场景差异，灵活配置各种规则类数据，通过分析平台的规则引擎，帮助业务提前感知业务问题、自动预警潜在风险，从而有效支撑业务的快速响应。例如，根据每个国家的供应能力预设不同的预警规则，当供应出现瓶颈或问题时，会自动触发预设的预警规则，从而对下一步的设备交付环节进行风险预警，提醒业务部门及时对交付计划进行调整，避免影响客户界面的交付承诺。

（3）满足业务运营中复杂智能决策的需求

通过数据分析模型对数据底座中的海量数据进行挖掘，智能分析业务问题的本质，洞察趋势并推荐方案，从而支持业务的客观、精准决策。例如，基于数据分析的统计预测构建不同的计划方法模型，支撑供应网络多级库存优化、多场景计划决策，应用数字化技术提高计划决策的效率和质量。

6.4.2　数据消费典型场景实践

为了更好地实现数据消费，华为公司通过5个步骤来管理从需求到自助分析的过程，包括业务需求提出、数据解析、数据搜索和获取、数据服务提供、自助报告设计和展示，如图6-35所示。

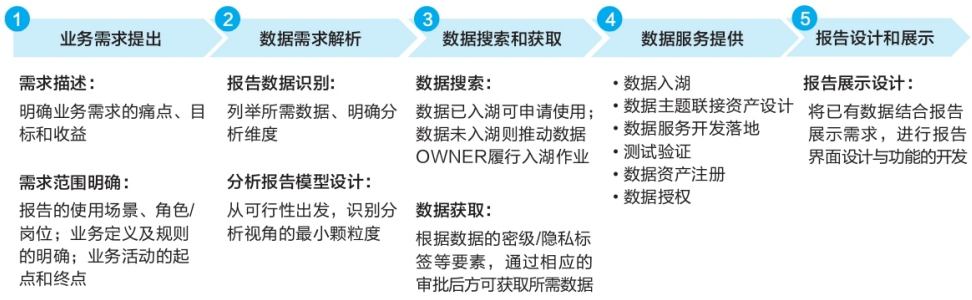


图6-35　从需求提出到消费的过程

当业务部门产生数据消费需求后，可以方便地通过数据地图进行搜索。如果已经有数据服务可以支撑，那么就能够快速地订阅服务，从而直接进行数据分析消费；如果暂时没有数据服务可用，那么数据专业人员也只需要提供相应的数据服务，而不需要为业务部门开发定制分析报告，这样就实现了数据供应和数据消费的高效协同。

可以通过一些具体的数字化运营实例看看数据如何发挥作用。

实例1：经营管理实践

过去，经营管理多为事后管理，通过每月下任务令进行业务改进，整个问题从发现到解决周期长，且存在大量人工动作，耗散大量业务精力。

- 数据滞后，经营过程非实时可视，主要为事后管理。
- 按月做经营分析，从经营数据产生到下发任务令需0.5～1个月。
- 缺少集成管理平台，手工做经营分析，线下管理任务令。

传统经营管理过程如图6-36所示。

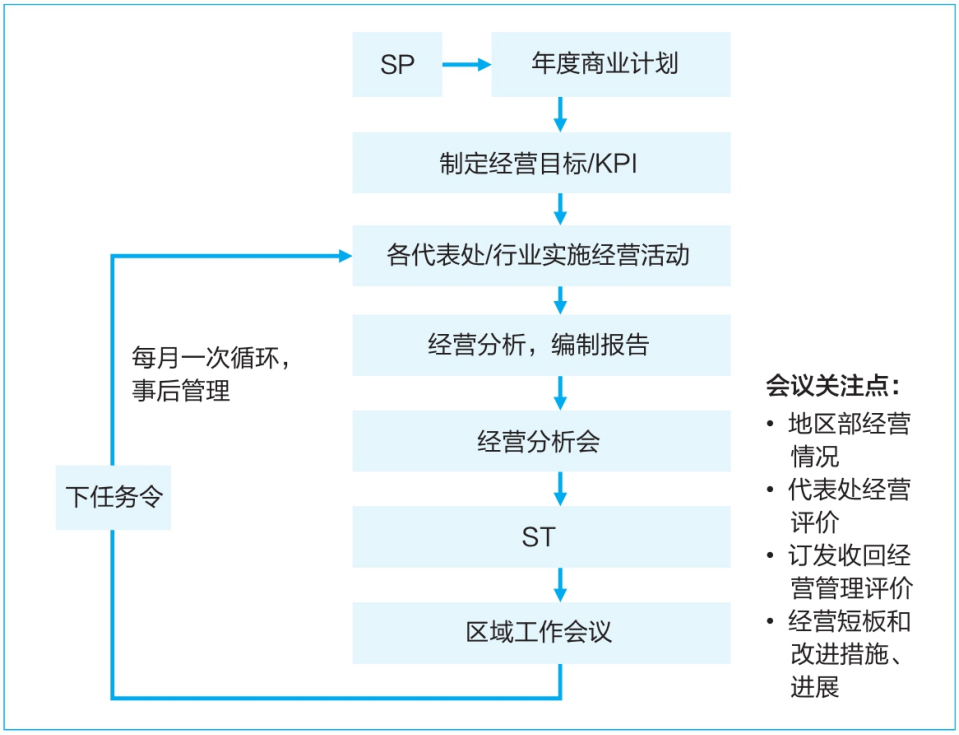


图6-36　传统经营管理过程示意

通过数字化运营实现事中监控和及时改进，达到经营过程可视化、报告在线分析、实时社交讨论、及时经营改进的目的。

- 数据在线转变为事中监控，随时可拉通审视规模、盈利、效率全经营指标。
- 经营管理模式改变，实时分析经营风险，会上聚焦重大风险决策，高效运作。
- 在统一平台进行监控、预警、协调、任务管理，从战略、执行到经营过程和结果实现闭环。

基于数字化运营的经营管理过程如图6-37所示。

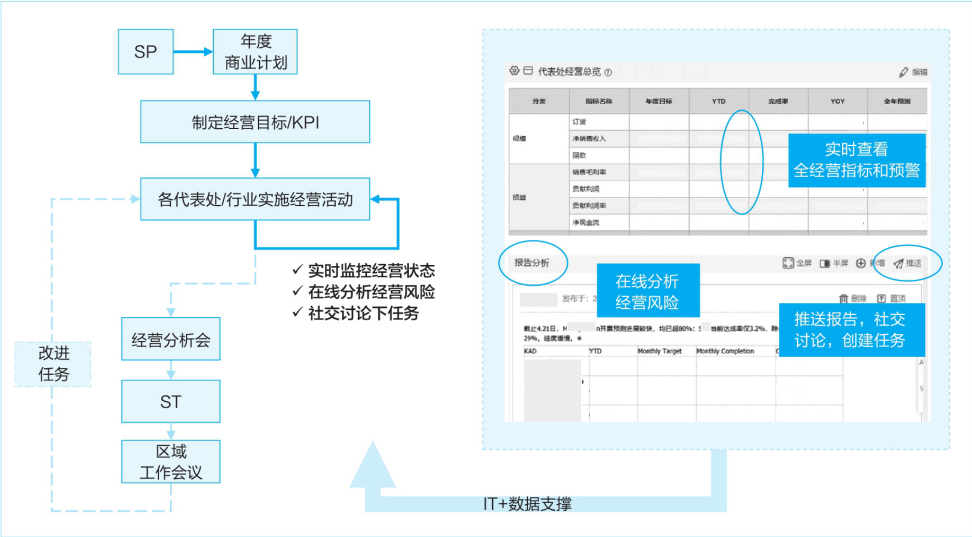


图6-37　基于数字化运营的经营管理流程示意

实例2：风险管理实践

过去，业务风险管理主要依赖事后审查，业务发生一段时间后再通过CT（Compliance Test，流程遵从性测试）、SACA（Semi-Annual Control Assessment，半年度控制评估）、稽核、审计等方式进行事后管理，业务根据审查要求提供相应“证据”。每一次核查都要先定位系统，先由核查人员自己在系统的海量数据中找问题，再约谈相关业务人员确定问题及其原因，这个过程中又存在多轮PK，然后再由责任人负责改进，最终还要由核查人员检查实际落实情况。整个过程不仅消耗大量精力，而且业务往往是被动改进，并不是主动进化。

- 落实主要依靠流程规则，细则多，执行难。
- 海量线下数据整理和分析，定位内控问题原因难。

- 流程控制评估、遵从性评估、主动性审视、稽核、审计多方事后清理。

传统风险管理方式如图6-38所示。

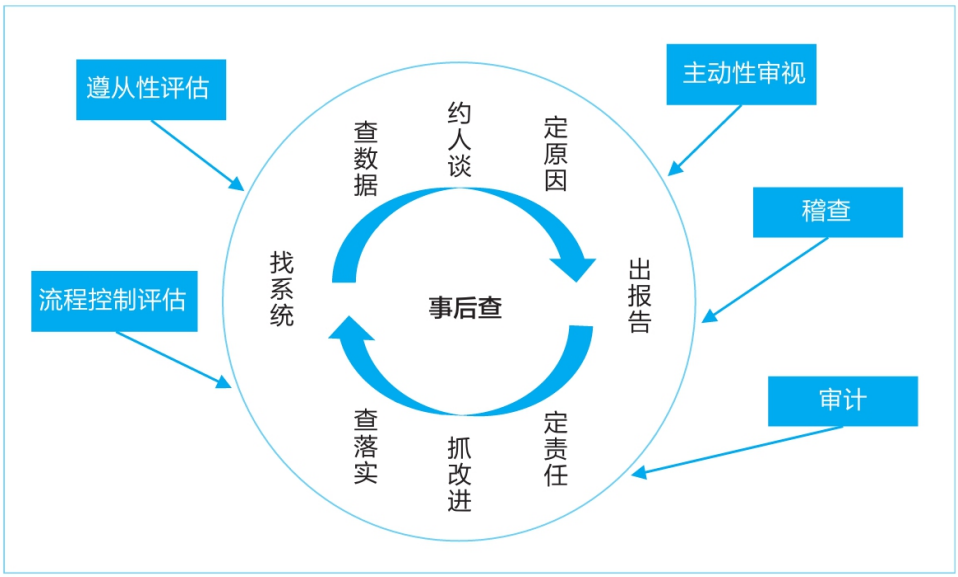


图6-38　传统风险管理方式

通过数字化运营，可以实现业务实时自检，风险实时在线审视和预警，风险任务快速关闭，不需要完全依赖事后核查，而是业务人员主动遵从。用数据规则替代人工的分析、检查和回溯，大大提高了风险管理的实时性，从事后管理变为事中管理，部分业务部门的量化风险降低超过50%。

- 在业务风险控制点打探针，及时发现问题。
- 实现风险量化可视，实时预警。
- 实时在线下达风险点的改进任务令，及时改善。

基于数字化运营的风险管理方式如图6-39所示。

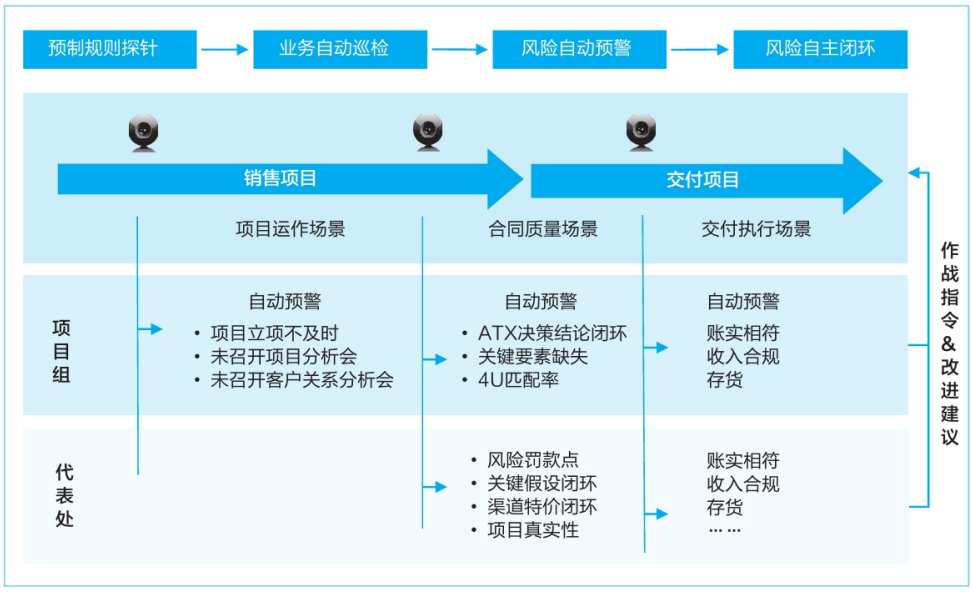


图6-39　基于数字化运营的风险管理

6.4.3　华为数据驱动数字化运营的历程和经验

1. 华为数字化运营的不同阶段

数字化运营本身是一种实践，每个企业数字化运营的道路都不相同，华为公司通过数据赋能业务运营是从2016年开始的，中间经历了不同阶段，也走过一些弯路，如图6-40所示。



图6-40　华为数字化运营历程

（1）“ 从行走到公交”阶段

大部分情况下采取“机关建、业务部门用”的方式。虽然通过数据治理达到了“数据清洁”的目标，也能够实现一定程度的经营和运营数据的可视化，但仍存在机关开发永远满足不了业务部门需求的问题，尤其是不同国 家的业务场景各有差异，机关开发无法满足根据业务场景进行灵活运营的目的，机关开发往往疲于奔命，而业务部门满意度仍然不高。

（2）“ 从公交到自驾”阶段

由于自研和引 进了业界先进的分析工具，各区域开始按需以自助的形式生成各种分析报表。在满足各国 家、各业务部门特定需求和灵活变化方面，收到了一些效果。但在该阶段，数据从供应到消费的整个过程处于“无序”状态，业务自助分析“野蛮生长”，大量数据是以离线手工获取方式为主，数据的完整性和可靠性存在很大问题，也存在大量的数据安全隐患 。

（3）“ 从无序到有序”阶段

通过数据底座建设实现生态共建、平台共享的效果。随着数据服务的大量建设，逐步覆盖了80%的场景，使得各种数据消费的效率更高、安全性更好，减少了大量重复建设。通过前台统一入口，大大提升了运营分析的性能体验。另外，还打造了从“业务部门回 到机关”通道，对各个国 家的优秀典型场景进行归纳，业务部门的优秀实践反向纳入数据底座形成公共数据

服务和报表卡片，各个国家可以通过自助分析平台实现对优秀实践的快速复制。

（4）“人工到智能”阶段

在原有的数据实时可视化的基础上，逐步增加动态及时预警能力、智能分析和方案推荐能力、任务自动执行能力，支撑业务数字化运营达到更高层面。业务数字化运营逐步从单一报表可视化，扩展到业务监控、预测、预警、协调、调度、决策、指挥等7个场景的持续打造，最终实现企业数字化运营“平时值班、战时指挥、察打一体”目标，如图6-41所示。



图6-41　华为数字化运营关键特性

2. 做好数字化运营的“三个要点、两个基础”

业务数字化运营不是某个能力的单独应用，需要多种能力的联合推动。在初期孵化过程中，尤其应该关注企业数字化运

营的“三个要点、两个基础”。

“三个要点”是指数字化运营中的“发育、激励、分享”。

在面向各业务部门的数字化运营能力的“发育”过程中，要做好自助分析能力赋能，识别关键核心人员并通过培训与实战的方式帮助他们掌握自助分析的基本能力，同时机关专家要做好现场支持，帮助各业务分析人员“上马”。在数字化运营实践中要充分激励原创，采取各种方式保护原创，同时鼓励各业务部门充分共享优秀实践。同时，机关要发挥归纳和总结作用，从各地优秀实践中识别真正具有共性的典型场景和典型数据联接模型，不仅可以使自助分析性能更好，还可以推动优秀实践在各个业务部门快速复制，达到“从1到N”的效果。

“两个基础”是指数字化运营中的“数据服务和IT平台”。

数据服务是整个数字化消费的关键，也是业务数字化运营的重要基础。IT平台包括分析平台和数据分析结果呈现前台，其中分析平台承载企业的公共分析能力建设，并重点面向业务分析师提供自助分析能力；数据分析结果呈现前台承载了公共场景的市场能力，支撑典型场景的快速分享。

具体如图6-42所示。

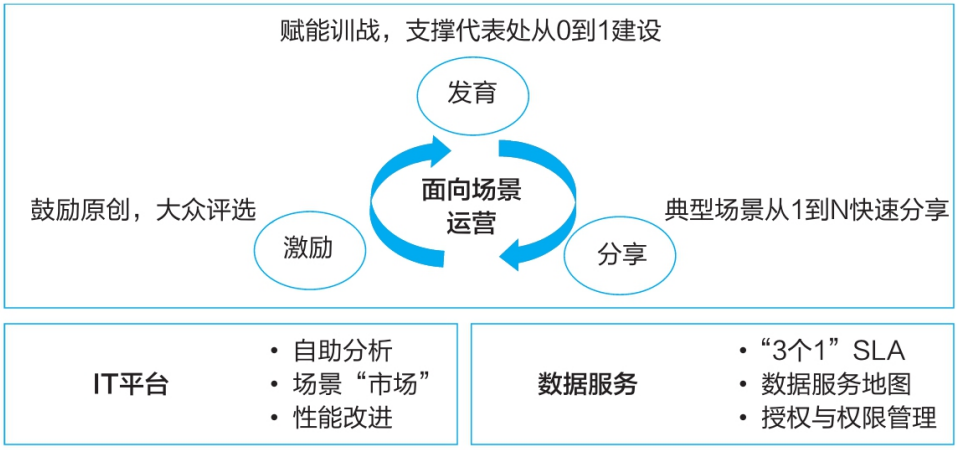


图6-42　企业开展数字化运营的关键要素

6.5 本章小结

对于非数字原生企业来说，打造企业级的全面的数据服务和面向业务的自助消费，将会是一个长期的过程，中间会遇到大量的困难和挑战。一方面，面临着大量遗留资产和平台的改造；另一方面，面临着诸多技术上的不确定因素。数据服务本身的能力还需要不断改进，在稳定性方面与传统方式还存在一定的差距；整个企业的认知还需要不断加强，很多业务部门还没有完全适应这种模式的转变；人员能力还需要不断提升，很多“业务专家”还不具备“业务数据分析专家”的能力。

但是，未来的方向是确定的，企业数字化转型的决心是确定的，尽管有这样或那样的困难，但数据服务是数据成为业务的“可消费产品”的最关键要素之一，我们应该坚持数据服务化的道路，充分学习和利用各种先进实践和先进技术，不断充实和加强自身能力，让数据服务发挥更大的价值。

第7章

打造“数字孪生”的数据全量感知能力

在信息化时代构建的IT系统，基本上是功能化、烟囱化、封闭式的，只能给企业内部经过培训的专业人员使用，所有的决策数据和我们信任的IT系统基本都是靠人来录入数据。但是，人如果犯错呢？

数字化转型是在解决工业革命时代没有解决的效率和成本问题，所以如果转型依赖的数据，还是需要组织大量专业人员去录入、去校验，那么就并没有从源头上解决数字化转型要解决的效率和成本问题。数字化转型要从根本上加强数据的可获得性，围绕我们构建的数据主题和对象丰富数据感知渠道。要追求更加实时、全面、有效、安全的数据获取。

7.1 “全量、无接触”的数据感知能力框架

7.1.1 数据感知能力的需求起源：数字孪生

2003年，Michael Grieves教授首次提出了“与物理产品等价的虚拟数字化表达”的概念，并给出定义：一个或一组特定装置的数字复制品，能够抽象表达真实装置并可以此为基础进行真实条件或模拟条件下的测试。该概念源于对装置的信息和数据进行更清晰的表达的期望，希望能够将所有信息放在一起进行更高层次的分析。数字孪生（Digital Twin，DT）即由此概念衍生而出并沿用至今。

在复杂的企业数字化变革过程中，非数字原生企业往往需要协调众多业务流，极具挑战性，但同时也是成功完成转型的关键。所以基于DT衍生出来的DTO（Digital Twin of an Organization，企业数字孪生）是一种动态的软件模型。模型需要输入组织的运营及其他类型的相关数据，以实现组织运营模型在虚拟世界中的映射，并能更新实时状态、应对外界变化、部署相应资源和产生预期客户价值。DTO虽然概念脱胎于DT，但是两者之间在适用对象、模型数据等方面，有着显著的差异，我们参考Gartner的文章归纳出了表7-1。

表7-1 DT和DTO对比

	起　　源	适用对象	模型数据	运作方式	目的意义
DT	起源于数字化制造中产品全生命周期的管理问题	涵盖物、人、流程、地点、复杂对象等几乎所有物理实体对象，支撑物联网的发展；但往往只关注单个设备、产品或是它们的组合	多来自仿真模型、CAD、BOM清单等物理对象的功能模型，数据多来自传感器、应用系统、维护日志等	利用人工智能、机器学习对数字孪生对象进行仿真分析，通过数字孪生体对物理对象进行控制和仿真	用以实时监控物理对象的运作、使用仿真分析实现对物理对象发展趋势的预测、控制以及优化
DTO	概念脱胎于数字孪生，并将其升华至组织的高度，起源于企业的数字化变革问题	主要集中在企业这一复杂对象内部，关注流程、运营和绩效指标之间的相关关系，是DT概念在组织管理方向的延伸；将“人”这一元素融入数字孪生中	企业内部的组织运作模型，数据来自组织流程、交易流、运营数据、绩效指标以及各类外部数据，例如客户用户体验反馈、市场行情变化等	制定战略目标，设立实现这一目标的价值链，建立多重综合评价指标，树立场景意识，动态把握组织所处的环境，并最终制定决策	帮助管理者实时了解企业运营情况，为企业数字化变革提供建议。并对一些不确定的因素做情景/模拟分析，为决策提供支持

Gartner预测2020年将有超过200亿个联网的传感器和端点，将会有数十亿个物件存在数字孪生。企业领导者开始有意识构建并不断改进企业的数据感知能力，希望提高物理对象的操作意识，并力求优化与这些对象的变化状态相关的决策，提升产品全生命周期数据收集和可视化能力，运用合适的分析工具和规则，高效地达成业务目标。通过操作数据或其他数据可以了解组织如何实施业务模型，连接其当前状态，部署资源，应对变化，以提供预期的客户价值，提升项目投资回报率，提升物理对象的性能并降低运营风险，从而创建更灵活、更动态、更迅捷的流程，自动应对数字化时代不断变化的形势。业务数字化整体方案如图7-1所示。

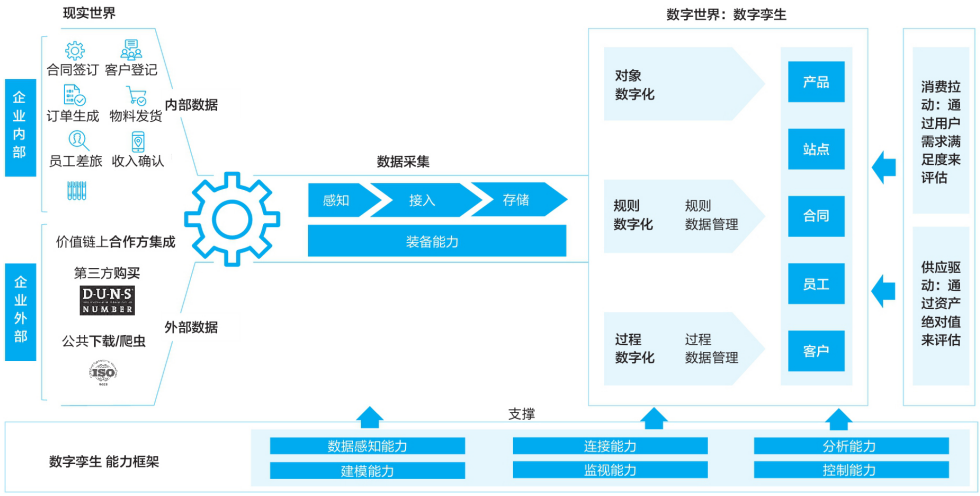


图7-1　业务数字化整体方案

很多非数字原生企业的数据管理能力不足、信息化程度较低，DTO还遥不可及，但这又是数字化转型的趋势，所以可以先着手构建数据采集能力，完成数据感知、接入和存储，先让企业具备DTO应用的基础。

7.1.2　数据感知能力架构

随着企业业务数字化转型的推进，非数字原生企业对数据的感知和获取提出了新的要求和挑战，原有信息化平台的数据输出和人工录入能力已经远远满足不了企业内部组织在数字化下的运作需求。企业需要构建数据感知能力，采用现代化手段采集和获取数据，减少人工录入。数据感知能力架构如图7-2所示。

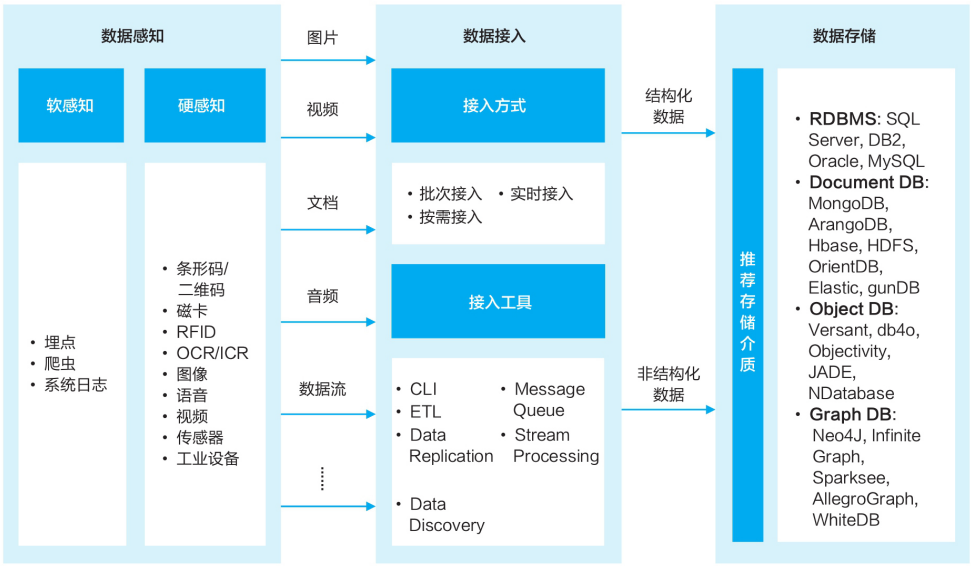


图7-2　数据感知

数据感知可分为“硬感知”和“软感知”，面向不同场景。“硬感知”主要利用设备或装置进行数据的收集，收集对象为物理世界中的物理实体，或者是以物理实体为载体的信息、事件、流程等。而“软感知”使用软件或者各种技术进行数据收集，收集的对象存在于数字世界，通常不依赖物理设备进行收集。如图7-3所示。

	软感知	硬感知
感知方式	使用软件或者各种程序进行数据收集，收集的对象存在于数字世界，通常**不依赖物理设备进行收集**	利用**设备或装置**进行数据的收集，收集对象为物理世界中的物理实体，或者是以物理实体为载体的信息、事件、流程、状态等
感知过程	数据感知的过程发生在**数字世界**，通常是自动运行的**程序或脚本**	数据的感知过程是数据**从物理世界向数字世界的转化过程**，有些数据感知需要人的操作
典型应用	埋点、System Log、网络爬虫	语音、视频、OCR、RFID、条形码/二维码、传感器、工控设备……

图7-3　感知分类

感知产生的数据还是孤立的物理对象的镜像，需要在企业这一复杂对象内部与其他数据资产一起，与流程、运营和指标之间建立关系，纳入企业的信息架构进行管理，才能真正打通从数据感知、生成到消费的链路。

当然，这一切的最终目的是生成企业级的感知数据，形成数字孪生的基础，满足企业利用人工智能、机器学习对数字孪生对象进行仿真分析、控制并优化制定战略目标的需求，帮助企业动态把握组织所处的环境，帮助管理者实时了解企业运营情况，为企业数字化变革提供建议，通过这些数字化的手段持续变革创新、获取业务价值。

7.2　基于物理世界的“硬感知”能力

7.2.1　“硬感知”能力的分类

数据采集方式主要经历了人工采集和自动采集两个阶段。自动采集技术仍在发展中，不同的应用领域所使用的具体技术手段也不同。基于物理世界的“硬感知”依靠的就是数据采集，是将物理对象镜像到数字世界中的主要通道，是构建数据感知的关键，是实现人工智能的基础。

基于当前的技术水平和应用场景，我们将“硬感知”分为9类，每一类感知方式都有自身的特点和应用场景，如图7-4所示。

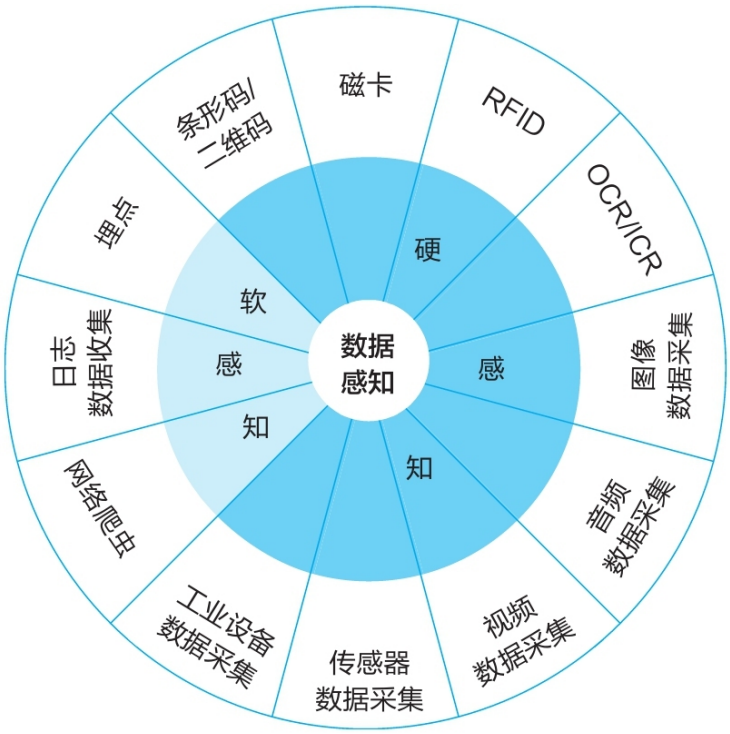


图7-4　9类“硬感知”

1. 条形码与二维码

条形码或者条码是将宽度不等的多个黑条和空白，按一定的编码规则排列，用以表达一组信息的图形标识符，通常一维条形码所能表示的字符集不过10个数字、26个英文字母及一些特殊字符，条码字符集所能表示的字符个数最多为128个ASCII字符，信息量非常有限。

二维码是用某种特定的几何图形按一定规律在平面上分布的黑白相间的图形，用来记录数据符号信息。二维码拥有庞大的信息携带量，能够把使用一维条码时存储于后台数据库中的信息包含在条码中，可以直接阅读条码得到相应的信息，并且二维码还有错误修正及防伪功能，增加了数据的安全性。

2. 磁卡

磁卡是一种卡片状的磁性记录介质，利用磁性载体记录字符与数字信息，用来保存身份信息。视使用基材的不同，可分为PET卡、PVC卡和纸卡三种；视磁层构造的不同，又可分为磁条卡和全涂磁卡两种。

磁卡的优点是成本低，这是它容易推广的原因，但缺点也比较明显，例如卡的保密性和安全性较差，使用磁卡的应用系统需要有可靠的计算机系统和中央数据库的支持。

3. RFID

RFID（Radio Frequency Identification，无线射频识别）是一种非接触式的自动识别技术，通过无线射频方式进行非接触双向数据通信，利用无线射频方式对记录媒体（电子标签或射频卡）进行读写，从而达到识别目标和数据交换的目的。

基于特别业务场景的需求，在RFID的基础上发展出了NFC（Near Field Communication，近场通信）。NFC本质上与RFID没有太大区别，在应用上的区别如下。

- NFC的距离小于10cm，所以具有很高的安全性，而RFID距离从几米到几十米都有。
- NFC仅限于13.56MHz的频段，与现有非接触智能卡技术兼容，所以很多的厂商和相关团体都支持NFC。而RFID标准较

多，难以统一，只能在特殊行业有特殊需求的情况下，采用相应的技术标准。

- RFID更多地被应用在生产、物流、跟踪、资产管理上，而NFC则在门禁、公交、手机支付等领域发挥着巨大的作用。

4. OCR和ICR

OCR（Optical Character Recognition，光学字符识别）是指电子设备（例如扫描仪或者数码相机）检查纸上打印的字符，通过边检测暗、亮的模式确定其形状，将其形状翻译成计算机文字的过程。如何除错或利用辅助信息提高识别正确率，是OCR的重要课题。

ICR（Intelligent Character Recognition，智能字符识别）是一种更先进的OCR。它植入了计算机深度学习的人工智能技术，采用语义推理和语义分析，根据字符上下文语句信息并结合语义知识库，对未识别部分的字符进行信息补全，解决了OCR的技术缺陷。

一个OCR识别系统，从影像到结果输出，须经过影像输入、影像预处理、文字特征抽取、比对识别，最后经人工校正将认错的文字更正，将结果输出。

目前OCR和ICR技术在业界有较为成熟的解决方案供应商，非数字原生企业不需要自行研发就可以完成相关技术的部署和数据的采集。

5. 图像数据采集

图像数据采集是指利用计算机对图像进行采集、处理、分析和理解，以识别不同模式的目标和对象的技术，是深度学习算法的一种实践应用。

图像数据采集的步骤如图7-5所示。



对象	描述
指纹	通过取像设备读取指纹图像，然后用计算机识别软件分析指纹的**全局特征**和指纹的**局部特征**
虹膜	虹膜识别技术是利用虹膜终身不变性和差异性的特点来识别身份的。虹膜是一种在眼睛中瞳孔内的织物状的各色环状物，每个虹膜都包含一个独一无二的基于水晶体、细丝、斑点、凹点、皱纹和**条纹等特征**的结构
视网膜	人体的**血管纹路**也是具有独特性的，人的视网膜上面血管的图样可以利用光学方法透过人眼晶体来测定
面部	面部识别技术通过对**面部特征**和它们之间的**关系**（眼睛、鼻子和嘴的位置以及它们之间的相对位置）来进行识别，用于捕捉面部图像的两项技术为**标准视频**和**热成像技术**，视频摄像头不同，热成像技术并不需要较好的光源，即使在黑暗情况下也可以使用
掌纹	掌纹与指纹一样也具有稳定性和唯一性，利用掌纹的**线特征**、**点特征**、**纹理特征**、**几何特征**等完全可以确定一个人的身份，因此掌纹识别是基于生物特征身份认证技术的重要内容
人耳	一套完整的人耳自动识别系统一般包括以下几个过程：人耳图像采集、图像的预处理、人耳图像的边缘检测与分割、特征提取、人耳图像的识别

图7-5　图像采集步骤

6. 音频数据采集

语音识别技术也被称为自动语音识别（Automatic Speech Recognition，ASR），可将人类的语音中的词汇内容转换为计算机可读的输入，例如二进制编码、字符序列或者文本文件。

目前音频数据采集技术在业界也有较为成熟的解决方案供应商，可以很便捷地通过解决方案供应商的技术，完成技术的部署和数据的采集。

采集来的声音作为音频文件存储。音频文件是指通过声音录入设备录制的原始声音，直接记录了真实声音的二进制采样数据，是互联网多媒体中重要的一种文件。音频获取途径包括下载音频、麦克风录制、MP3录音、录制计算机的声音、从CD中获取音频等。

7. 视频数据采集

视频是动态的数据，内容随时间而变化，声音与运动图像同步。通常视频信息体积较大，集成了影像、声音、文本等多种信息。

视频的获取方式包括网络下载、从VCD或DVD中捕获、从录像带中采集、利用摄像机拍摄等，以及购买视频素材、屏幕录制等。

8. 传感器数据采集

传感器是一种检测装置，能感受到被检测的信息，并能将检测到的信息按一定规律变换成信号或其他所需形式的信息输出，以满足信息的采集、传输、处理、存储、显示、记录等要求。信号类型包括IEPE信号、电流信号、电压信号、脉冲信号、I/O信号、电阻变化信号等。

传感器数据的主要特点是多源、实时、时序化、海量、高噪声、异构、价值密度低等，数据通信和处理难度都较大。

9. 工业设备数据采集

工业设备数据是对工业机器设备产生数据的统称。在机器中有很多特定功能的元器件（阀门、开关、压力计、摄像头等），这些元器件接受工业设备和系统的命令开、关或上报数据。工业设备和系统能够采集、存储、加工、传输数据。工业设备目前应用在很多行业，有联网设备，也有未联网设备。

工业设备数据采集应用广泛，例如可编程逻辑控制器（PLC）现场监控、数控设备故障诊断与检测、专用设备等大型工控设备的远程监控等。

7.2.2 “硬感知”能力在华为的实践

“硬感知”在非数字原生企业有广阔的前景，因为在数字化时代，非数字原生企业大量存在的产线、流程工艺、实体货物、物流设备等，都需要通过“硬感知”来实现数据的感知和采集。华为作为典型的非数字原生企业，9类数据“硬感知”能力在各领域中都得到了一定的应用，并已发挥了实际的业务价值。

1. 门店数字化

如图7-6所示，采用7种数据采集方式，支撑持续提升运营效率与消费者体验。通过光线传感器和温度传感器，自动调节

窗帘、灯光，温度随环境改变，并与店门、窗帘、灯光、空调、屏幕、防盗系统联动，打造智能绿色门店环境。通过实物管理感知，样机自动申报位置与状态，异常告警，自动上报消费者在门店体验过程中的行为，结合消费者体验情况优化陈列、营销设计、产品设计。通过视频感知客流与热区，管理门店各片区人流密度与停留时间，优化陈列与营销，实时调整服务人力与资源配置。



图7-6　门店数字化

2. 站点数字化

如图7-7所示，站点主要在高层或者在野外环境中，勘测和日 常维护难度都比较大，通过360度全景拍照和OCR，构建站点物理对象完整的围栏尺寸、塔高、机房尺寸、设备尺寸、天线挂高、走线距离、天线的方位角、下倾角、扇区等数字镜像，实现在数字化站点勘测规划，现实站点直接施工，避免在现场反复勘测、设计调整。

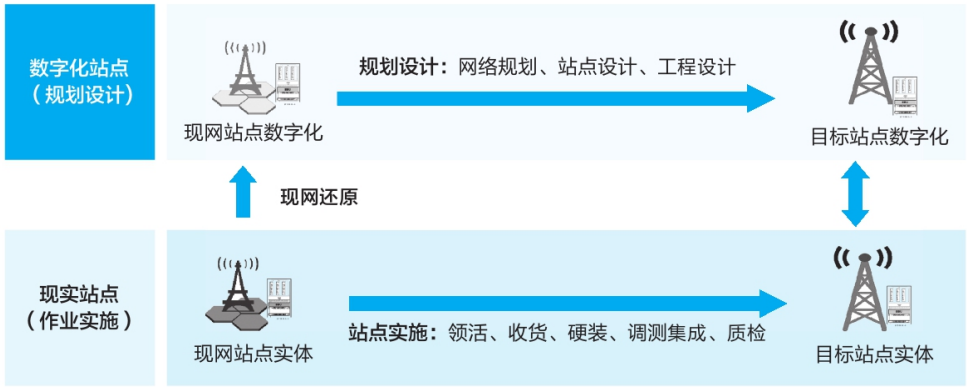

数字化站点
（规划设计）
现网站点数字化
规划设计：网络规划、站点设计、工程设计
目标站点数字化
现网还原
现实站点
（作业实施）
现网站点实体
站点实施：领活、收货、硬装、调测集成、质检
目标站点实体

图7-7　站点数字化

7.3　基于数字世界的“软感知”能力

7.3.1　“软感知”能力的分类

物理世界的“硬感知”是将物理对象构建到数字世界中的主要通道，是构建数据孪生的关键，而已经存在于数字世界中的那些分散、异构信息，可通过“软感知”能力来利用。目前“软感知”比较成熟，并随着数字原生企业的崛起而得到了广泛的应用。我们将“软感知”分为3类，如图7-8所示。

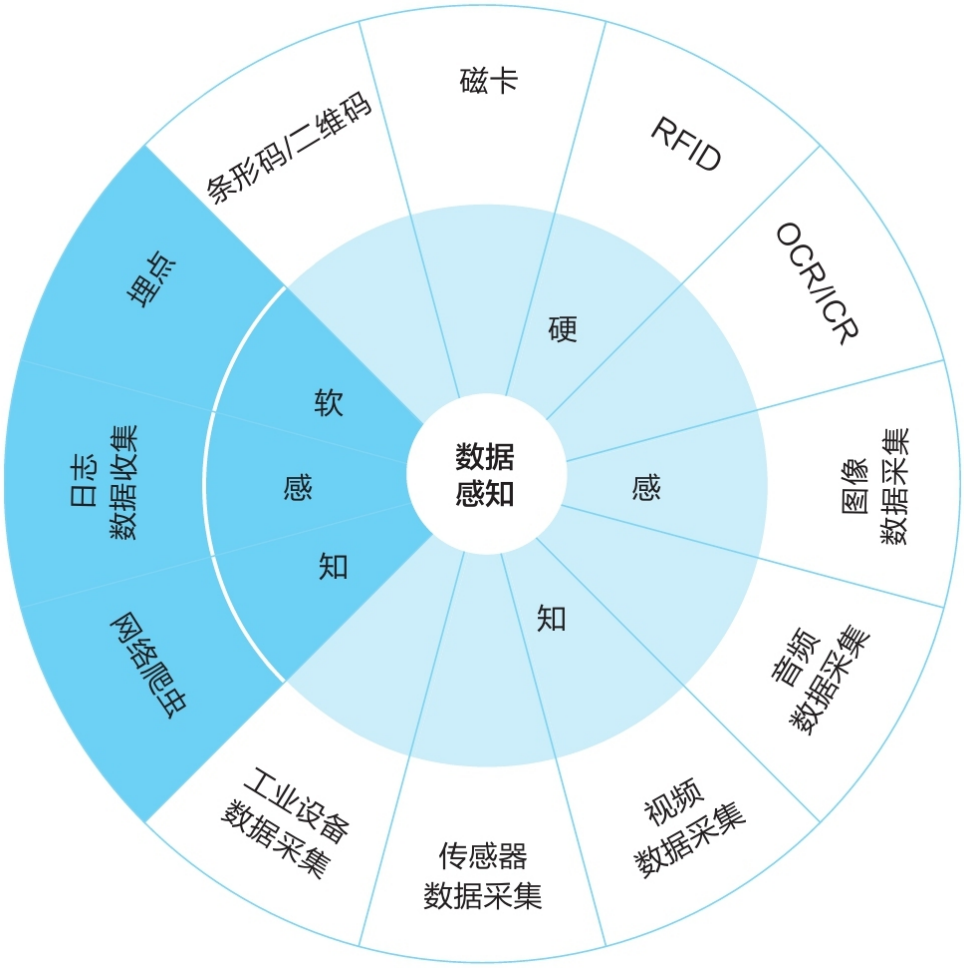


图7-8　3类“软感知”

1. 埋点

埋点是数据采集领域，尤其是用户行为数据采集领域的术语，指的是针对特定用户行为或事件进行捕获的相关技术。埋点的技术实质，是监听软件应用运行过程中的事件，当需要关注的事件发生时进行判断和捕获。

埋点的主要作用是能够帮助业务和数据分析人员打通固有信息墙，为了解用户交互行为、扩宽用户信息和前移运营机会

提供数据支撑。在产品数据分析的初级阶段，业务人员通过自有或第三方的数据统计平台了解App用户访问的数据指标，包括新增用户数、活跃用户数等。这些指标能帮助企业宏观地了解用户访问的整体情况和趋势，从总体上把握产品的运营状况，通过分析埋点获取的数据，制定产品改进策略。

埋点技术在当前主要有以下几类，每一类都有自己独特的优缺点，可以基于业务的需求，匹配使用。

- 代码埋点是目前比较主流的埋点方式，业务人员根据自己的统计需求选择需要埋点的区域及埋点方式，形成详细的埋点方案，由技术人员手工将这些统计代码添加在想要获取数据的统计点上。
- 可视化埋点通过可视化页面设定埋点区域和事件ID，从而在用户操作时记录操作行为。
- 全埋点是在SDK部署时做统一的埋点，将App或应用程序的操作尽量多地采集下来。无论业务人员是否需要埋点数据，全埋点都会将该处的用户行为数据和对应产生的信息全采集下来。

2. 日志数据采集

日志数据收集是实时收集服务器、应用程序、网络设备等生成的日志记录，此过程的目的是识别运行错误、配置错误、入侵尝试、策略违反或安全问题。

在企业业务管理中，基于IT系统建设和运作产生的日 志内容，可以将日 志分为三类。因为系统的多样化和分析维度的差异，日 志管理面临着诸多的数据管理问题。

- 操作日 志，指系统用户使用系统过程中的一系列的操作记录。此日 志有利于备查及提供相关安全审计的资料。
- 运行日 志，用于记录网 元设备或应用程序在运行过程中的状况和信息，包括异常的状态、动作、关键的事件等。
- 安全日 志，用于记录在设备侧发生的安全事件，如登录、权限 等。

3. 网络爬虫

网络爬虫（Web Crawler）又称为网页蜘蛛、网络机器人，是按照一定的规则自动抓取网页信息的程序或者脚本。

搜索和数字化运营需求的兴起，使得爬虫技术得到了长足的发展，爬虫技术作为网络、数据库与机器学习 等领域的交汇点，已经成为满足个性化数据需求的最佳实践。Python、Java、PHP、C#、Go等语 言都可以实现爬虫，特别是Python中配置爬虫的便捷性，使得爬虫技术得以迅速普及，也促成了政府、企业界、个人对信息安全和隐私的关注。

7.3.2 “软感知”能力在华为的实践

“软感知”主要面向产品持续运营提供服务，基于对产品日志、用户行为的感知，改善产品功能。以华为内部数据管理平台为例（如图7-9所示），数据管理平台的数字化运营，需要识别用户行为，进而提升运营效率与用户数据消费的体验。通过对平台埋点，捕捉用户在界面上从数据定位到最终消费的浏览过程和停留时间等信息，并关联用户的部门、职位、所在地等信息，自动生成用户画像和数据画像，确定细分用户范围，界定相同认知背景和业务场景的用户，提供可识别的分类资产用于搜索，界定数据资产分类，面向不同用户界定不同的资产范围，减少匹配差异和搜索引擎复杂度，训练搜索引擎和推荐算法，提供最优数据推荐结果和排序位置。

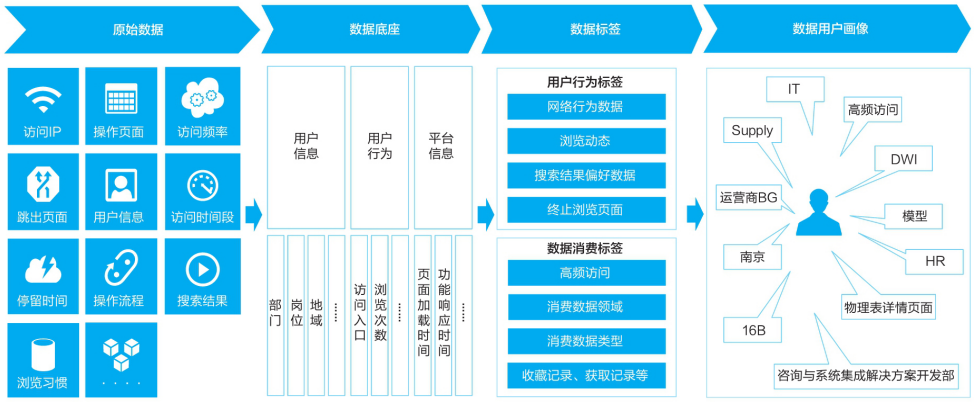


图7-9　数据管理平台 用户标签

12类感知能力在企业中的应用，突破了原有人工维护数据的局限。但是不管是“软感知”还是“硬感知”，产生的数据在没有纳入企业整体的数据管理体系情况下，如果只以独立数据的形式存在，是无法应对复杂的企业数字化变革的。

7.4　通过感知能力推进企业业务数字化

7.4.1　感知数据在华为信息架构中的位置

感知可以应用于广泛的物理世界和数字世界，感知范围可以从人、物、作业、地点扩展到复杂环境。成熟的用例倾向于以物和人为中心。而在企业中，只有将感知数据纳入整体的数据体系中，才能发挥感知数据的价值。

华为数据治理下的感知能力对接了数据供应链（Data Supply Chain），数据从感知采集到最终的分析消费，都纳入公司级的信息架构，作为数据资产来进行管理，如图7-10所示。

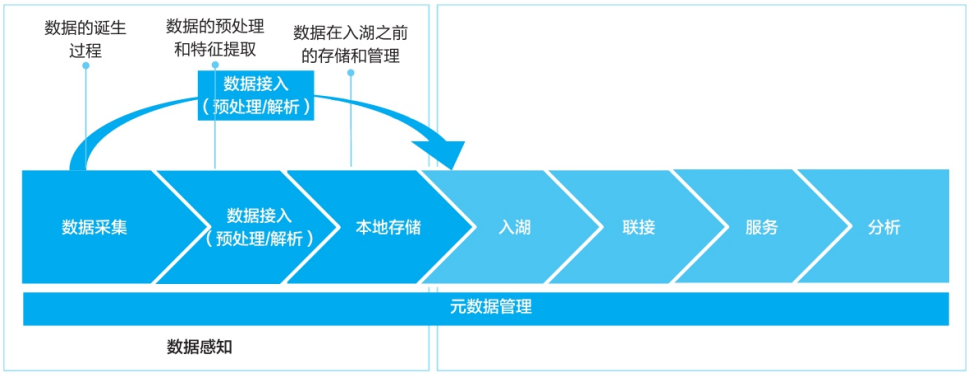


图7-10　数据感知到分析消费

感知数据生成后，需要通过连接进入下一步环境，通过不同的数据类型，选择不同的数据接入方式。在确定数据接入方式之前，需要重点考虑以下几个问题。

- 数据源的可用性分析。
- 接入的数据量大小。
- 数据接入过程是连续的还是按一定的时间间隔进行。
- 数据接入是拉（Pull）的方式还是推（Push）的方式。
- 在数据接入的过程中，是否需要做数据校验或数据标准化。
- 在接入的过程中，是否需要对数据做进一步的处理，如数据聚合、数据分类等。

感知数据的接入方式与工具如图7-11所示。

接入方式	触发方式	频率	数据源	数据量	数据类型	工具类型
批次接入 Batch Data	固定时间间隔触发接入	频率较低，数据接入完成时间通常是小时级	RDBMS, Flat File, ERP、Cloud、DWH等	支持大批量接入和小批量接入	结构化或者非结构化数据	• CLI • ETL
按需接入 Ad-Hoc	按需要触发接入	频率较高，数据接入完成时间通常是分钟或者秒级	通常是数据文件，spread sheet、RDBMS等	通常比较小	结构化或者非结构化数据	• CLI • ETL • Data发现与预处理工具
实时接入 Real Time	数据源产生新数据时触发接入	数据接入完成时间频率是毫秒级	传感器，机器，网络路由器等	数据量比较大，但单条数据记录通常比较小	结构化数据	• Message Queue • Stream处理工具

图7-11　感知数据接入方式及工具

根据不同的数据采集方式、采集内容和接入方式，选择合适的存储介质。在选择存储介质的时候需要考虑如表7-2所示的因素。

表7-2　感知数据推荐存储介质

序号	采集方式	数据类型	接入方式	推荐存储介质	典型数据库
1	埋点	结构化数据	实时	RDBMS	RDBMS：SQL Server、DB2、Oracle、MySQL 基于文件存储的数据库：MongoDB、ArangoDB、Hbase、HDFS、OrientDB、Elastic、gunDB 基于对象存储的数据库：Versant、db4o、Objectivity、JADE、NDatabase 图数据库：Neo4J、Infinite Graph、Sparksee、Allegro-Graph、WhiteDB
		非结构化数据	批次、点对点	Document DB、Flat File	
2	日志收集	结构化数据	实时	RDBMS	
		非结构化数据	批次、点对点	Flat File	
3	爬虫	结构化数据	实时、半实时	RDBMS	
		非结构化数据	批次、点对点	Document DB、Flat File、Graph DB	
4	条形码 / 二维码	结构化数据	实时、半实时	RDBMS	
5	磁卡	结构化数据	实时、半实时	RDBMS	
6	RFID	结构化数据	实时、半实时	RDBMS	
7	OCR/ICR	非结构化数据	批次、点对点	Flat File	
8	图像	非结构化数据	批次、点对点	Graph DB	
9	SRA	非结构化数据	批次、点对点	Flat File	
	音频	非结构化数据	批次、点对点	Object DB	
10	视频	非结构化数据	批次	Object DB	
11	传感器	结构化数据	实时	RDBMS	
12	设备	结构化数据	实时	RDBMS	

作为数据资产管理的核心，感知元数据管理应该包含两个方面的内容，如图7-12所示。

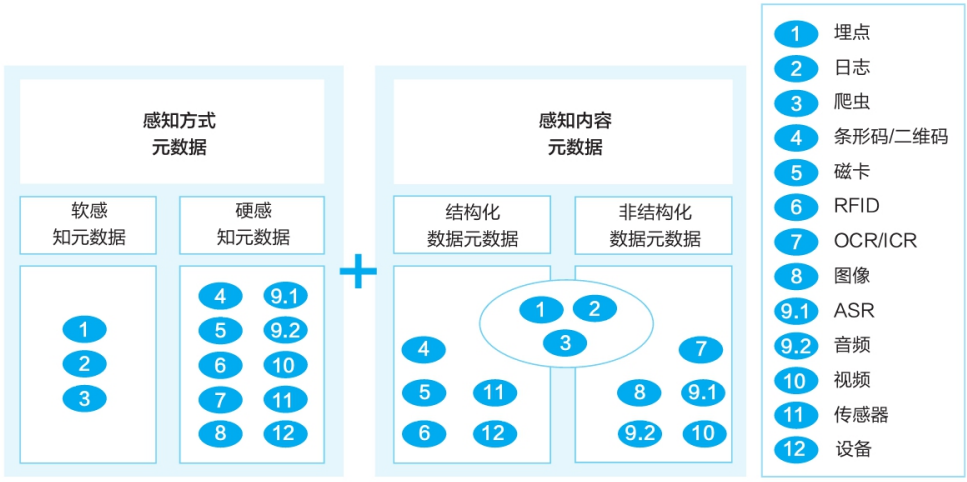


图7-12　感知元数据管理

- 感知方式元数据：对数据感知的方法进行登记注册的过程，在后续的数据消费的过程中可以知道数据来源。

- 感知内容元数据：感知内容包括结构化数据和非结构化数据，所以元数据管理也分为结构化数据元数据和非结构化数据元数据。

感知得到的数据是企业信息架构的一部分，在数据分类中需要基于感知采集方式的差异，制定不同的管理办法。

观测工具和观测对象都要纳入信息架构中，定义业务对象对其进行管理。观测数据在资产管理中识别业务对象时，可以采用以下两个建议：

- 观测对象是一个时，观测数据挂靠在该业务对象下。
- 观测对象是多个时，观测数据按大股东原则判定数据Owner和挂靠业务对象。

7.4.2　非数字原生企业数据感知能力的建设

因为非数字原生企业的业务特征、数字化基础和数据管理阶段都不一样，数据感知和采集工具的成熟度也不一致，考虑技术发展和成本的制约因素，企业一般会逐步构建感知能力，完善企业数字孪生。我们参考埃森哲关于数字孪生的调查总结出了图7-13所示的数字孪生成熟阶段。

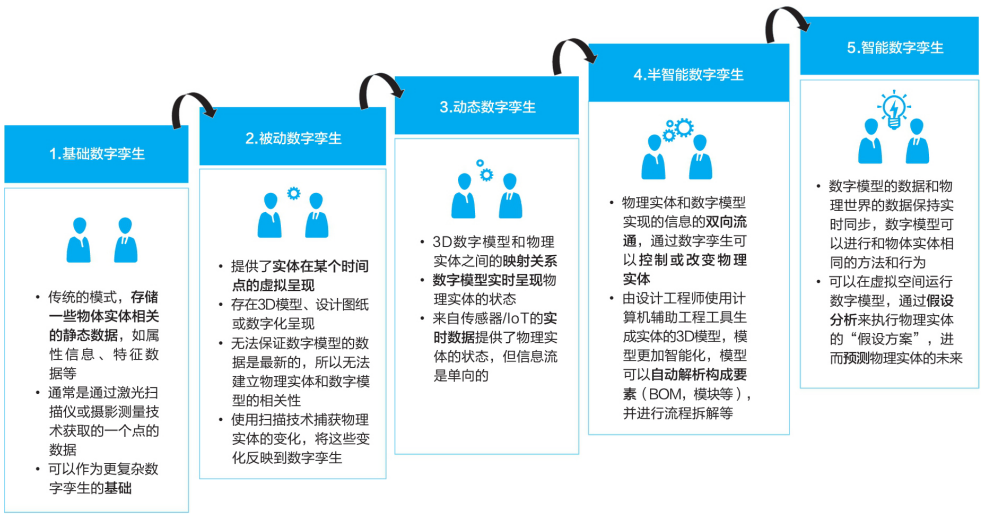


图7-13　数字孪生成熟阶段

如果非数字原生企业需要构建感知能力，可以考虑从以下几个方向来选择，关键是能力的构建始终要贴合业务，尽快促成业务价值的呈现。

- 开发一个独特的物理对象感知能力可以获得收益的方向，包括改善运营、降低运营风险、降低成本、更好地为客户服务的机会，或者通过拥有质量更高、更全面的数据来进行更好的业务决策。
- 在更复杂、更昂贵的环境（例如工业机器和企业资产）中，更有可能抵消感知能力构建的实现成本。
- 组织是否拥有相关感知能力的前身，比如可以利用现有的、详细的元数据和模型（例如BOM、CAD和仿真模型）。
- 需要一个模型来支持极端的操作环境，比如远程或环境恶劣的地方。
- 探索技术或商业模式的创新，比如增强现实的应用，或者实现资产货币化的新方法，或者提供前所未有的、差异化

的服务水平等领域。

7.5　本章小结

随着非数字原生企业数字化转型项目的推进，感知能力构建的最终对象逐渐从单一节点发展到获得完整物理对象的数字孪生。考虑到物理对象的维度和可能的数据量，构建一个全量感知的企业数字孪生的成本可能会相当惊人。所以一个成功的数字化转型项目要构建的感知规模一定要面向应用，由业务价值驱动。非数字原生企业不可能构建物理对象100%的镜像数字孪生，也完全没必要这么做。每个数字孪生实际上只是对象的最有业务价值的一个或几个方面的数字模型，我们只需利用适当的技术满足特定的业务目标，优化回报，分阶段利用感知获取的数据创造价值，同时最大限度地降低成本，逐步完成全量的数据感知能力，打造“孪生”的数字世界。

第8章

打造“清洁数据”的质量综合管理能力

越来越多的企业应用和服务都基于数据而建，数据质量是数据价值得以发挥的前提。例如企业运营效率主要依赖于数据获取的准确性和及时性，企业客户关系管理系统中的错误或不完整数据将导致客户沟通不顺畅，影响客户满意度。

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。

本章讲述数据质量基本概念和管理框架，详细说明数据质量控制、数据质量改进、数据质量度量的基本方法。

8.1 基于PDCA的数据质量管理框架

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（垃圾进，垃圾出）”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。

8.1.1 什么是数据质量

ISO9000标准对质量的定义为“产品固有特性满足要求的程度”，其中“要求”指“明示的、隐含的或必须履行的需求或期望”，强调“以顾客为关注焦点”。

在Won Kim的论文“A Taxonomy of Dirty Data”中，数据质量被定义为“适合使用”，即数据适合使用的程度、满足特定用户期望的程度。

数据质量不是追求100%，而是从数据使用者的角度定义，满足业务、用户需要的数据即为“好”数据。

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。

1）**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。

2）**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。

3）**准确性：**指真实、准确地记录原始数据，无虚假数据及信息。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保持一致。

4）**一致性：**指遵循统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。

5）**唯一性：**指同一数据只能有唯一的标识符。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。

6）**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。

8.1.2　数据质量管理范围

提到数据质量管理，经常有人会问：数据质量和流程质量有什么区别？流程质量是基于流程结果评估业务执行的好坏，数据质量更关注业务对象、业务规则、业务过程、业务结果等数据是否得到了及时记录。以采购验收为例，采购验收及时性属于流程质量，送达到验收所需时间满足3天的SLA即属于流程质量合格；而验收数据录入及时性属于数据质量，验收到录入所需时间满足1天的SLA即属于数据质量合格。

8.1.3 数据质量的总体框架

华为以ISO8000质量标准体系为依据，设计了PDCA（Plan、Do、Check、Action、计划、执行、检查、处理）持续改进的数据质量管理框架，如图8-1所示。

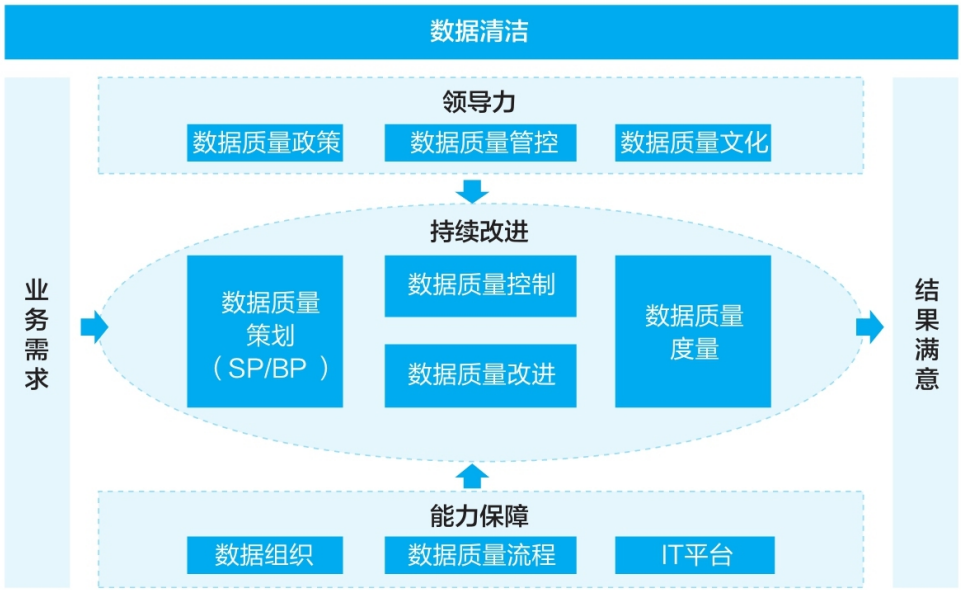


图8-1 数据质量管理框架

数据质量管理以数据清洁为目标，以业务需求为驱动，通过PDCA的循环，提升数据质量，达到数据质量结果满意。领导力模块通过制定政策、规范来构建数据质量管理机制，对数据质量的工作起牵引作用。能力保障模块构建完整的数据组织、流程和工具，起到支撑作用。

（1）自上而下打造数据质量领导力

数据质量政策应该有不同的层次，数据质量的管控要兼顾宏观方面的指导原则以及微观层面的具体操作要求，引导正确的业务行为，提升企业成员的数据质量意识。

（2）全面推进数据质量持续改进机制

提升数据质量是为了满足业务应用，业务战略变化会产生新数据，对数据应用提出更高的要求，使得数据质量管理范围、目标发生变化，因此数据质量管理是动态、持续的循环过程。

（3）不断加强数据质量能力保障

数据质量管理具有专业性，需要专业团队制定数据质量管理策略、流程、规范等，通过技术工具实现自动融入日常业务。通过不断提升数据质量管理组织的管理水平、改善数据质量工具平台，使企业数据质量获得进一步提高。

8.2　全面监控企业业务异常数据

不论做了多少数据质量预防措施，实施多严格的数据质量过程控制，只要涉及人为干预，总会存在数据质量的问题。为了避免或降低数据质量对业务的影响，要能及时发现数据质量问题。问题的发现既可以“正向”主动监控，也可以“逆向”通过下游环节反馈问题来识别。主动发现、制定解决方案、采取行动，比被动采取补救措施效果更好，并且代价更小。数据质量监控环节必不可少，本节重点讲述基于异常数据的数据质量监控。

8.2.1　数据质量规则

异常数据是不满足数据标准、不符合业务实质的客观存在的数据，如某位员工的国籍信息错误、某位客户的客户名称信息错误等。

数据在底层数据库多数是以二维表格的形式存储，每个数据格存储一个数据值。若想从众多数据中识别出异常数据，就需要通过数据质量规则给数据打上标签。

数据质量规则是判断数据是否符合数据质量要求的逻辑约束。在整个数据质量监控的过程中，数据质量规则的好坏直接影响监控的效果，因此如何设计数据质量规则很重要。

依据数据在数据库落地时的质量特性及数据质量规则类型，设计如下四类数据质量分类框架。

1）**单列数据质量规则。**关注数据属性值的有无以及是否符合自身规范的逻辑判断。

2）**跨列数据质量规则。**关注数据属性间关联关系的逻辑判断。

3）**跨行数据质量规则。**关注数据记录之间关联关系的逻辑判断。

4）**跨表数据质量规则。**关注数据集关联关系的逻辑判断。

华为结合ISO8000数据质量标准、数据质量控制与评估原则（国标SY/T 7005—2014），共设计了15类规则，具体如图8-2所示。

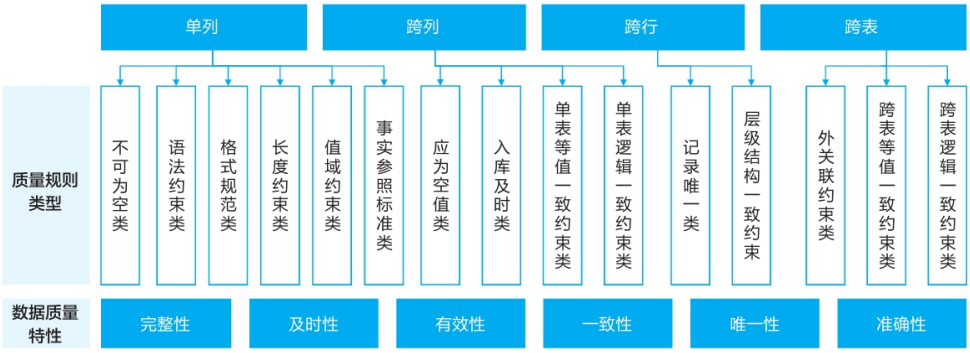


图8-2　数据质量规则

规则类型的详细说明如表8-1所示。

表8-1　规则分类内容及示例

业务对象落地	质量特性	规则类型	类型描述	示　例
单列	完整性	不可为空类	属性不允许或在满足某种条件下不允许出现空值	员工工号不可为空
	有效性	语法约束类	属性值满足数据语法规范取值约束	邮箱地址需满足有效邮箱格式，身份证号满足国家标准
	有效性	格式规范类	属性值必须满足展现格式约束	日期有多重格式，对于同一属性指定同一类格式
	有效性	长度约束类	属性值须满足约定的长度范围	密码的长度至少要 8 位，不超过 16 位
	有效性	值域约束类	属性值必须满足已定义的枚举值列的约束	合同的合同主类型及子类型必须是合同类型基础数据中定义的枚举值
	准确性	事实参照标准类	存在事实数据或者事实参考标准数据，与该事实或事实参照标准对比一致的约束	中国电信通信有限公司的信息必须与国家法人数据库中的信息保持一致
跨列	完整性	应为空值类	属性满足某种条件下不能维护值	敏感站点不允许维护经纬度信息
	一致性	单表等值一致约束类	某一属性值与本实体其他属性计算值相等的约束	合同的 RMB 签约金额必须等于 USD 签约金额与汇率的乘积

（续）

业务对象落地	质量特性	规则类型	类型描述	示　例
跨列	一致性	单表逻辑一致约束类	某一属性值与本实体其他属性满足逻辑关系约束（大于或小于）	合同的关闭日期不能早于注册日期
	及时性	入库及时类	数据进入系统的及时性约束，通常要包括数据原材料获取时间和入库时间才能进行规则设计	通过 HRMS 系统中员工的入职日期和系统创建日期判断员工入职信息维护及时性
跨表	一致性	外关联约束类	引用其他业务对象属性时，所维护的属性值必须在其他业务对象中存在的约束	合同的签约客户必须为客户主数据中定义的法人客户
	一致性	跨表等值一致约束类	某一属性值与其他实体的一个或多个属性值的函数计算结果相等的约束	合同的金额与合同按产品拆分后的金额之和一致
	一致性	跨表逻辑一致约束类	某一属性值满足其他实体的一个或多个属性值的函数关系的约束（大于或小于）	员工的任命日期早于员工的到岗日期
跨行	唯一性	记录唯一类	记录不重复，存在可识别的业务主键进行唯一性判断，是对数据集内部是否存在相似或重复记录的约束规则	法人客户中国移动通信股份有限公司只能存在唯一一笔
	一致性	层级结构一致约束类	存在层级结构的属性，同层级属性结构一致	所有子网类型的客户，满足总部 – 分部 – 子网的三层结构

当我们发现某个数据格的数据异常时，往往会思考这一列其他的数据格是否也存在同样的问题，是否应该对这一列的其他数据格进行检查。因此数据质量规则一般以业务属性（即数据列）为对象，数据质量规则类型为颗粒度进行设计和应用。

这样既方便获取业务属性的整体数据质量状况，又可清晰定位异常数据、识别严重问题、制定解决方案，同时数据质量规则也不会因互相交织而过于庞大，方便后续的运营维护。

我们以员工“邮箱地址”业务属性为例设计数据质量规则进行数据质量检查。根据业务问题反馈、数据源剖析及15类数据质量规则对数据遍历的综合结果，我们设计了“不可为空类”“语法约束类”“格式规范类”三个数据质量规则进行数据质量检查。同时对这三个子规则向上收敛，形成“邮箱地址”业务属性的完整的主规则，这种层级关系我们称之为“规则树”，如图8-3所示。



图8-3　规则树示例

通过规则树，我们既能统计出共有多少员工的“邮箱地址”数据异常，又可分别统计各子规则的异常数量，从而快速识别出当前哪个问题更严重（异常数量越多，问题越严重）。

因此我们在制定相应的解决方案时，可能会优先解决问题严重的子规则。

在如图8-4所示的规则应用结果中，我们可以看到6位员工的“邮箱地址”有异常，其中“不可为空类”的异常有5个，占比最大，且解决此问题的技术手段简单，成本较低。因此我们决定先解决邮箱地址“不可为空”的问题，在数据产生系统中根据数据质量规则增加防呆设计。

数据对象				数据质量规则			
ID	区域	国家	邮箱地址	Rule_ 邮箱地址	不可为 空类	语法 约束类	格式 规范类
E001	中国	中国	email@example.com	**F**	T	T	**F**
E002	东北欧	拉脱维亚		T	T	T	T
E003	西欧	比利时		**F**	**F**	T	T
E004	西非	加纳	email@example.com	T	T	T	T
E005	南美南	智利		**F**	**F**	T	T
E006	北美	美国	email@example.com	**F**	T	**F**	T
E007	中国	中国	email@example.com	T	T	T	T
E008	东南亚	泰国	email@example.com	T	T	T	T
E009	北美	美国		**F**	**F**	T	T
E010	东北欧	丹麦		**F**	**T**	T	T

图8-4　规则应用结果

这里需要强调的是，并不是每一个属性都会涉及上述15类规则，例如“记录唯一类”规则，适用于“员工ID”但不适用于“员工姓名”；“值域约束类”规则，仅适用于有枚举值列表的业务属性。同时，随着解决方案的落地、历史数据的清理、新需求的开发，需要进行监控的数据质量规则也会随之新增、变更、取消。例如上面所提到的“邮箱地址”的“不可为空类”规则，当IT系统实现了防呆功能且完成历史数据清理

后，监控持续一段时间里异常率都为0，则规则可下线。所以，数据质量规则的生命周期是随着数据治理范围的扩大和数据治理程度的深入而更新的。

8.2.2 异常数据监控

质量控制是通过监控质量形成过程，消除全过程中引起不合格或不满意效果的因素，以达到质量要求而采用的各种质量作业技术和活动。要保证最终交付质量，必须对过程进行质量控制，通常是在过程中设置关键质量控制点。例如，可以在数据录入阶段设置规则程序，从源头避免不可接受的数据进入系统。

数据质量控制的目的是致力于满足数据质量要求，消除或减少异常数据。数据质量控制可以在数据的生命周期内的不同时点被应用，来测试数据的质量和其是否适合于其所在的系统。

华为通过数据质量监控平台，以异常数据管理为核心，实施数据质量控制，如图8-5所示。

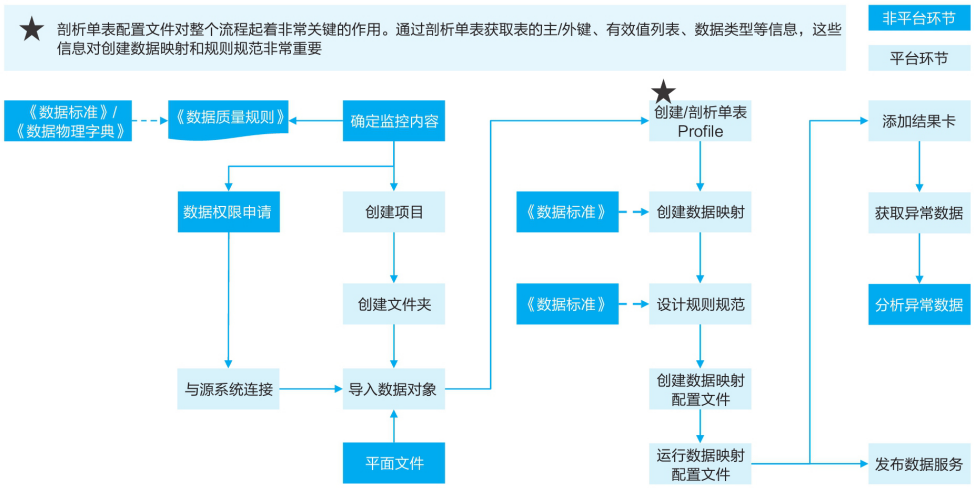


图8-5　数据质量异常监控业务流

1. 识别监控对象范围，确定监控内容

数据质量控制从明确业务需求开始，根据业务规划和数据相关方的需求，阶段性确定数据质量控制范围。

从定性、定量两个维度识别关键数据，定性维度参考以下原则。

（1）重要性原则

- 关键主数据和基础数据：公司级、领域级主数据，如产品、客户、供应商、组织、人员、站点。
- 关键的事务数据：主交易流的核心事务数据，如客户合同、BOQ、工程服务采购PR、S&OP计划、采购PO。
- 痛点问题：领域业务运营痛点问题、公司级变革、攻关项目、业务核心KPI等涉及的对象纳入度量，如产品Item。

（2）成本效益原则

- 运作成熟且质量较高的数据，或度量成本很高但预期的改进很少的数据，可不优先考虑。
- 数据管家也可通过收集业务需求、数据质量问题等其他途径从中筛选当前需监控的数据。

领域数据质量策划样例如图8-6所示，图中领域数据质量监测范围选择，以当期业务运营重点工作“1个全局运营场景、4个核心领域运营场景、2个质量运营场景”作为基础范围，再按照重要性打分排序，筛选出数据质量监控优先级高的业务对象。之后，通过运营管理重点、影响范围、数据质量痛点、下游流程路径选择、与其他属性有逻辑关系等因素，确定业务对象的关键属性集。明确关键数据范围后，通过IT工具配置数据质量规则，识别异常数据。

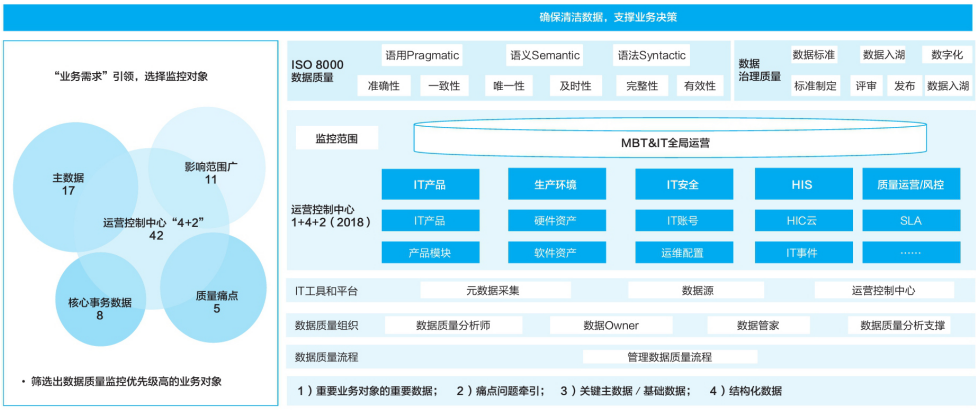


图8-6　领域数据质量策划样例

2. 数据源剖析

在着手设计数据质量规则前，需对数据进行快速数据剖析，目的是分析数据源的内容、质量和结构，同时发现和分析数据源中的所有数据不规范问题和使数据项目处于危险中的隐藏数据问题。

数据剖析摘要视图示例如图8-7所示，其显示了配置文件中所有列和规则的属性。摘要视图包含属性的可视化表示形式。



图8-7　数据剖析摘要视图示例

1）**数据源内容：**如从上述数据源剖析结果的摘要视图中，我们可以了解到此表包含员工工号、姓名等内容，即列信息等。

2）**数据源结构：**包括技术结构和业务结构。技术结构指空值频率、相异值频率、值范围（最大值、最小值）、模式、长度、数据类型。业务结构如组织结构存储是平面结构还是树状结构。

3）**数据源质量：**根据数据标准分析剖析结果的数据质量，例如必填字段是否有空值存储，有允许值列表中的值个数与相异值频率是否一致等。

数据剖析可以更好地识别需要监控数据的质量要素，如图8-8所示。

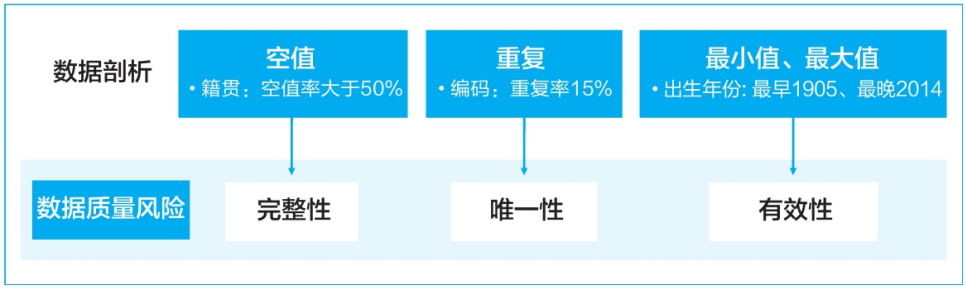


图8-8　数据剖析结果分析

3. 设计和配置监控规则，自动监测异常数据

上一节已详细讲述了如何设计质量规则，部署质量监控规则，对目标数据进行质量监控，并对发现的数据异常情况进行告警。目前华为数据质量监控平台已实现质量规则的可配置、数字化、快速部署、自动监控识别异常数据等能力，并可随时间推移，制定周期性监控计划，监视数据质量的进展情况，并通过虚拟化的方式快速、灵活发布监控结果，如图8-9所示。

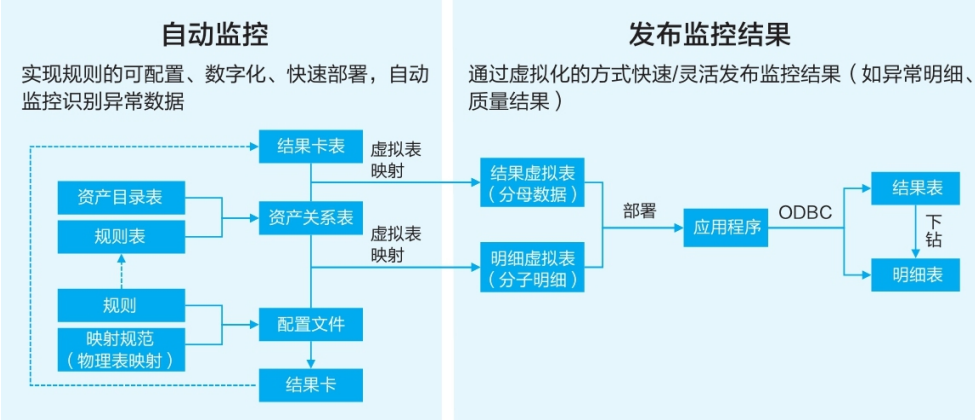


图8-9　数据质量自 动监控逻辑

可利用自助分析工具开发在线数据质量分析报告，通过前端工具不仅能够查看监控结果汇总数据，而且能够通过钻取功能查看异常明细数据，以便业务人员准确定位业务系统的异常数据。

8.3　通过数据质量综合水平牵引质量提升

通过数据质量度量综合评价公司整体数据质量水平，制定数据质量基线，披露数据质量问题与短板，促进问题改进，推动数据Owner承接数据质量改进目标，持续提升数据质量，实现数据清洁。

8.3.1　数据质量度量运作机制

（1）度量模型

过程设计与执行结果并重，设计质量评估信息架构的建设，执行质量评估数据清洁。数据质量度量模型如图8-10所示。

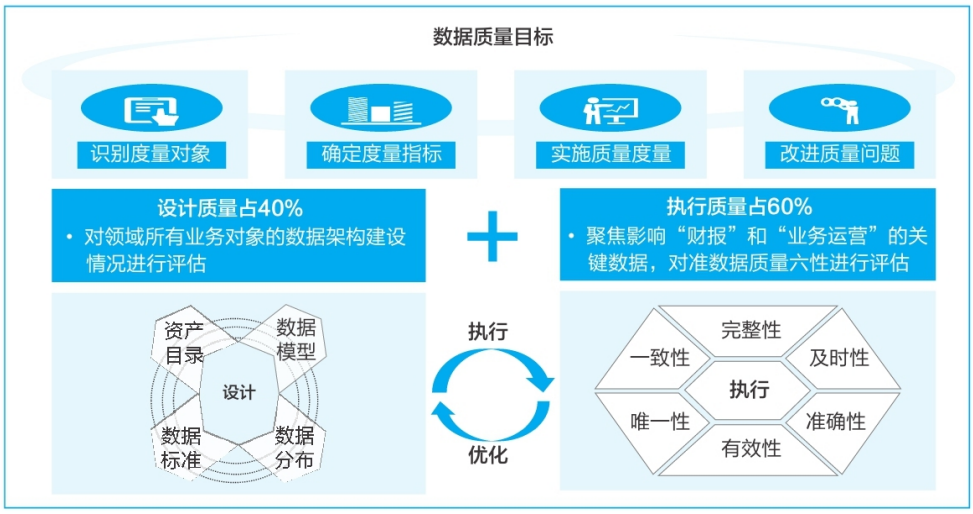


图8-10　数据质量度量模型

（2）数据Owner职责要求

1）公司数据Owner：下达数据质量目标，并签发数据质量度量报告；基于数据质量结果及改进状况，对相应数据Owner进行奖励及问责。

2）各领域数据Owner：承接公司数据Owner设定的数据质量目标；明确数据质量问题改进责任人，并推动问题闭环管理；对数据质量度量结果负责，依据要求向公司数据Owner述职。

（3）专业支撑组织职责要求

1）公司数据管理部：根据公司数据管理工作规划，制定数据质量目标；组织数据质量度量工作开展，发布公司数据质量度量报告；组织评审数据质量标准及指标，并验收数据质量问题闭环状况。

2）各领域数据管理部：基于公司数据质量度量工作要求，拟定数据质量标准并设计指标，执行数据质量度量；组织各领域业务专家，分析数据质量问题根因，制定改进举措及闭环管理。

（4）度量规则

1）度量对象选定原则：聚焦业务运营痛点数据和影响财报的关键数据。

2）**度量频率：**一年度量两次。上半年度量期间为1月～6月，重点监控质量改进状况；全年度量期间为1月～12月，综合评价质量达成水平。

3）**度量方法：**从“设计”及“执行”两个方面开展，通过“设计”明确架构及标准，通过“执行”反映其质量结果。

4）**评价标准：**统一采取百分率的方式评价，并根据度量得分划分如表8-2所示的五档。

表8-2　满意度等级

评分标准（百分率）	定级评价（5等级）
80% ～ 100%	1等（满意）
60% ～ 80%	2等（基本满意）
40% ～ 60%	3等（略不满意）
20% ～ 40%	4等（不满意）
0 ～ 20%	5等（很不满意）

8.3.2　设计质量度量

为确保设计质量标准稳定，从信息架构的四个角度（数据资产目录、数据标准、数据模型、数据分布）进行综合评估，其范围覆盖度量期间内已通过IA-SAG评审发布的所有数据资产。当实际业务有例外场景时，可向IA-SAG专业评审团申请仲裁，若评审通过，则可采用白名单的方式进行管理。

（1）数据资产目录

1）业务对象需有明确、唯一的数据Owner，并对该业务对象全流程端到端质量负责，如是否有定义数据质量目标、是否有数据质量工作规划等。

2）业务对象的元数据质量，如数据分类是否完整、业务定义是否准确、数据管家是否有效等。

3）资产目录完整性。

（2）数据标准

1）数据标准元数据质量，如数据标准是否唯一、业务用途及定义是否准确、各责任主体是否有效等。

2）所有业务对象应准确关联数据标准。

3）数据标准在IT系统及其对应的业务流程中应得到应用和遵从。

（3）数据模型

1）开发概念模型和逻辑模型，并通过IA-SAG评审。

2）物理数据模型设计应遵从逻辑数据模型设计，数据库中物理表的落地应遵循物理模型。

（4）数据分布

1）已认证数据源，并通过IA-SAG评审。

2）交易侧完整的信息链和数据流，并通过IA-SAG评审。

3）交易侧业务资产、数据湖、主题联接、数据服务、自助分析之间完整准确的血缘关系。

（5）设计质量打分模型

质量打分细则如图8-11所示。

得分	数据资产目录	数据标准	数据模型	数据分布
				
1分（差）	• 业务对象没有明确的定义及Owner，无法对该业务对象全流程端到端质量负责	• 未发布完整的数据字典	• 未发布IA-SAG评审后的概念模型	• 未认证数据源
2分（中）	• 业务对象有明确的定义及Owner，承诺对该业务对象全流程端到端质量负责	• 开发完整的数据字典，并通过IA-SAG评审	• 开发和维护概念模型，并通过IA-SAG评审	• 数据源认证通过IA-SAG评审
3分（良）		• 数据标准通过GPO、BPO或GPC签发 • 数据标准在IT系统及其对应的业务流程中得到应用和遵从	• 开发和维护逻辑模型，并通过IA-SAG评审	• 明确数据血缘关系，并制定出合理的整改方案及路标
4分（优）			• 管理逻辑模型和物理模型之间关系，并确保物理模型的开发及维护遵从已发布的逻辑模型	• 完成数据血缘信息链治理，确保数据高效传递，避免不必要的人工干预
5分（满分）	• 管理面向未来3～5年的数据架构蓝图 • 全面实现满足业务主体数据化、业务场景数据化、业务规则数据化、业务决定算法化、业务流程自动化的要求 • 信息资产成为公司战略核心竞争力			

图8-11　质量打分细则

8.3.3　执行质量度量

执行质量度量主要是从数据质量六性（一致性、完整性、及时性、唯一性、有效性、准确性）评估数据内容的清洁度，涉及三个要素：客户关注重要性、法律财务风险性、业务流程战略性。业务领域也可根据阶段性的管理重点和诉求调整评估的要素。

- **客户关注重要性：**给客户运营带来直接影响的数据的客户关注重要性就高，如合同、PO、验收标准、开票数据等。
- **法律财务风险性：**与法律、财务的关联性强，一旦发生质量问题，会触犯法律或带来相关财务损失，那么该数据的法律财务风险性就高，如收入、成本等数据。
- **业务流程战略性：**数据所产生的业务流程如果是公司核心交易流程（如LTC流程）或战略地位高的流程（如IPD流程），那么数据的业务流程战略性普遍会得到较高关注；如果是相关支撑或使能流程（如变革流程、IT开发流程等），那么数据的业务流程战略性相对较弱。

关键数据对象评分表如图8-12所示。

交付项目主题域关键数据对象评分表

数据对象	业务流程战略性	法律财务风险性	客户关注重要性	问题发生的频度和影响程度(反向验证)	得分	重要性等级
交付项目	5	3	3		11	M
交付子项目	5	3	5		13	H
项目WBS	5	3	1		9	L
主计划	5	3	3		11	M
实施计划	3	3	3		9	L
供应需求计划	3	3	1		7	L
开票触发计划	5	5	5	5	20	H
收入触发记录	5	5	5		15	H
---	---	---	---		.....	

备注：从各数据对象实际问题发生的频度、影响程度等要素出发，反向验证和补充所选定的关键数据对象

SAMPLE

图8-12　关键数据对象评分表

1. 确定度量指标

与度量对象一样，数据质量度量指标也往往来源于日 常监控的数据质量规则，将业务属性层主规则通过叠加公式变成业务对象层度量指标。

数据质量规则的设计应让相关业务人员参与，以满足业务的使用场景。但当某些业务场景的规则不够清晰，或当前的技术手段无法较为准确地识别异常数据时，这类数据质量规则往往只能用于警示，不建议纳入度量。例如数据标准的唯一性规则，通过判断数据标准被业务属性的引用次数来定义。当某数据标准被引用次数少于10次时，我们认为这类数据标准可能存在冗余的风险，但不能完全确定为异常数据。此类规则若纳入度量考核，后续需投入大量的人工核对成本。其次，数据质量规则应可支撑持续度量。例如某些完整性的数据质量规则，可设置必填项，一次性解决其数据质量问题，此类数据质量规则不建议纳入数据质量度量。

数据质量指标同时参考5项原则进行设置。

- **重要性原则：**对核心数据、痛点问题较严重的数据，需重点考虑设计度量指标。
- **成本效益原则：**运作成熟且质量较高的数据，或度量成本很高但预期改进很少的数据，可以考虑简化度量指标或不度量。
- **明确性原则：**指标设计清晰、可衡量。
- **分层分级原则：**可根据不同层级的管理诉求，设计分层分级的指标。
- **持续度量原则：**一次性就可解决问题的数据不需要度量。

一个业务对象下有如此多的数据质量规则，如何叠加形成数据质量度量指标呢？对于叠加公式，我们建议使用以下计算规则。

1）逻辑实体数据质量度量指标=Σ 属性数据质量异常数量/Σ 属性数据总量，我们称之为数据格面积算法。

2）业务对象数据质量度量综合指标= Average（逻辑实体数据质量度量指标）。

不直接在业务对象层采用数据格面积算法，是为了避免重要的错误数据被“淹没”。我们以业务对象“采购PO”中的逻辑实体“PO头信息”和“PO行信息”为示例进行阐述。

1）每年“PO头信息”的数据量大概为“PO行信息”的数据量的1/100。

2）“PO头信息”中业务属性“汇率类型”异常率为50%，即100个PO头信息中有50个汇率类型错误。“PO行信息”中业务属性“品类”异常率为10%，即10 000个PO行信息中，有1000个“品类”信息。

3）若我们在业务对象层级采用数据格面积算法作为其度量指标，则业务对象综合数据质量异常率为：（50 + 1000）/（100 + 10 000）≈10.4%。这就基本忽略了“PO头信息”中业务属性“汇率类型”这个重要异常率。

当然企业也可根据公司自身的数据特点，制定相应的叠加公式进行综合计算。例如可以对业务对象下逻辑实体异常率进行加权平均，而权重比例可参考其数据量的差异倍数进行设置。

2. 确定数据质量衡量标准

数据质量衡量标准是指指标测评结果与用户质量诉求的关系。华为主要采用五个等级（差、中、良 、优、满分）来衡量和拉通数据质量满足消费者的应用程度，如表8-3所示。为了让读者能更深入地了解五分制的目的和用途，这里我们列举两个对比示例。

表8-3　数据质量五分衡量标准

得分	用户感知	得分说明	主数据	事务数据
1分	差	指标未度量，或度量结果远不能达到数据质量要求，存在严重的数据质量问题	小于三西格玛（93.32%）	• 事务数据指标按照主业务流、同类指标拉通原则 • 按业务流：客户交易流、产品配置流、交付作业流、集成计划流等 • 按六性：结合六性情况对指标相对拉通
2分	中	度量结果不能达到数据质量要求，存在较多的数据质量问题，影响较大	三西格玛（93.32%）～四西格玛（99.3797%）	
3分	良	度量结果基本达到数据质量要求，存在少量的数据质量问题，影响一般	四西格玛（99.379%）～五西格玛（99.977%）	
4分	优	度量结果完全达到数据质量要求，且基本不存在数据质量问题	五西格玛（99.977%）～六西格玛（100%）	
5分	满分	零缺陷（所有属性都没有任何异常，且长期不存在任何数据质量问题）	零缺陷	

示例一：企业供应商数量为2000家，其供应商账号信息中的“收款账号”信息准确率为90%，即有200位供应商的“付款

账号”数据错误，这意味着有10%的应付可能出现付款错误。对于这个准确率数据，消费者是不能接受的。

示例二：企业度量“员工现居住地址准确率”结果为90%。调研统计结果表明有50%的员工处于租房状态，因此数据消费者认为当前的数据质量可以满足他们的应用要求。

从上述两个示例我们会发现，不同数据的数据消费者对其要求不同，不能单纯以度量结果来衡量数据质量的好坏，因此需要有一个衡量标准。为避免衡量标准的参差不齐，我们有一些原则性的建议。

1）主数据绝对拉通，采用业界通用的六西格玛要求。

2）事务数据可依据各业务流进行相对拉通，但对于完整性和及时性这类较简易的数据质量要求，应相对严格。

3）衡量标准的划分，数据管家应组织数据生产者和数据消费者共同协商讨论，达成一致。数据管家应从数据专业视角给予建议，数据生产者从其当前的数据管理、IT工具、人员技能等方面预估当前的数据质量水平，数据消费者从数据的使用视角提出数据质量要求。

这里还需要说明一点，同一个业务对象下的不同业务属性的消费者不同。那么如何综合所有消费者的诉求，在业务对象层级划分数据质量衡量标准呢？这里我们建议同一业务对象下的可保留 部分独立的数据质量规则，再逐一对业务对象下的所

有度量指标划分质量衡量标准，最后再通过加权平均的方法收敛到业务对象层，即得到业务对象分。

3. 执行度量

数据质量度量已流程化，因此我们可将其作为一次小型变革项目进行管理。根据度量运作机制，由公司数据管理部定期启动公司级数据质量度量。召开启动会议，明确本次数据质量度量细则，如数据质量度量目标、度量期间、度量范围、度量指标、计划进度等相关事宜，以确保数据质量度量工作有序、高效地开展，同时也确认数据质量度量结果的公正、有效。

8.3.4 质量改进

数据质量改进致力于增强满足数据质量要求的能力。数据质量改进消除 系统性的问题，对现有的质量水平在控制的基础上加以提高，使质量达到一个新水平、新高度。

质量改进的步骤本身就是一个PDCA循环。质量改进包括涉及企业跨组织的变革性改进（BTMS）和企业各部门内部人员对现有过程进行渐进的持续改进（GPMS）。华为公司也出现过针对一些问题年年改进但是年年问题再次发生的现象，最为关键的原因就是没有真正按照质量改进的步骤开展工作，没有用质量改进的方法把真正的根因识别出来加以改进并固化到流程体系中。因此，规范改进过程并按照过程规范实施管理改进是非常关键的。

华为定义的改进过程框架是一个大的PDCA循环。通过管理层（ST）的管理评审以及变革与改进的规划，识别变革与改进项目，每个项目按照规范的项目群管理运作流程或者改进过程框架实施改进。改进成果固化到流程及管理体系中并实施推广执行，执行后再通过质量组织进行客户满意度管理、度量、审核与变革进度指标评估等，将再次识别改进作为管理评审的输入，最终形成大的改进循环。

数据质量改进流程如图8-13所示。

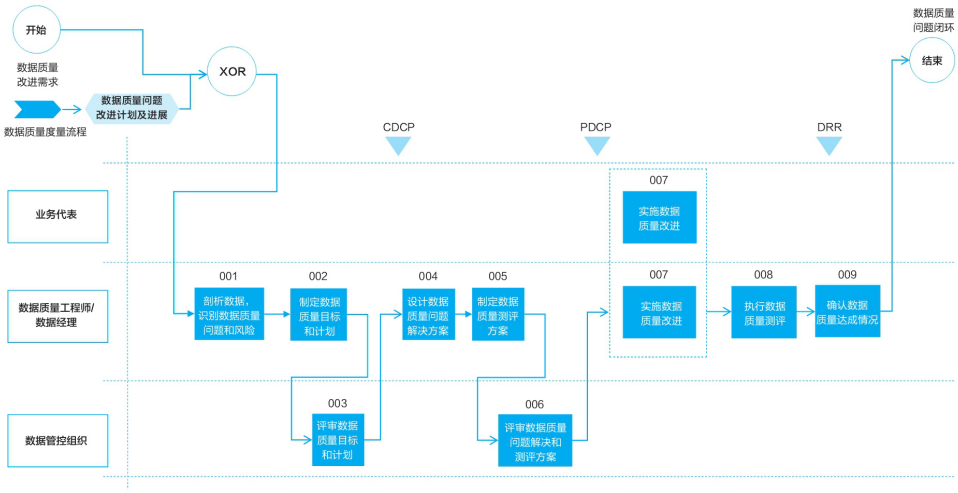


图8-13　数据质量改进流程

在这里我们说明一下数据质量控制和数据质量改进的关系。质量活动通常分为两类：维持与改善。维持是指维持现有的数据质量水平，其方法是数据质量控制；改善是指改进目前的数据质量，其方法是主动采取措施，使数据质量在原有的基础上有突破性的提高，即数据质量改进。

从结果的角度来说，数据质量控制的目的是维持某一特定的质量水平，控制系统的偶发性缺陷；而数据质量改进则是对某一特定的数据质量水平进行“突破性”的提升，使其在更高的目标水平下处于相对平衡的状态。控制是日 常进行的工作，可以纳入流程体系的“操作规程”中加以贯彻执行，最好的手段就是纳入流程体系进行标准化。质量改进则是一项阶段性的工作，达到既定目标之后，该项工作就完成了。质量改进的最终效果比原来维持下的效果好得多，这种工作必然需要精心策划。质量改进要固化在流程体系中进行标准化，通过质量控制使得标准化的流程得以实施，达到新的质量水平。

质量控制是质量改进的前提，控制就意味着维持以前的质量水平，是PDCA改进循环中保证水平不下降的“努力的楔子”，是保证下一次改进的起点，而改进则是在起点基础上的变革和突破。如果不做好质量控制，质量水平就会下降，下次又在低水平重复，因此不能只关注质量改进，改进后关键还是要实施质量控制，二者交替进行，相辅相成。

8.4 本章小结

数据质量管理应成为企业持续、例行的工作，企业数据质量管理水平直接影响数据应用的效果和数字化转型的成效。华为数据质量管理框架由三个部分构成，包括自上而下打造数据质量领导力、全面推进数据质量持续改进机制、不断加强数据质量能力保障。通过制定数据质量政策，并依托公司变革体系和流程运营体系实现质量管控的落地，同时以多种方式在全公司营造质量氛围和文化。其中最重要的是建立了企业数据质量持续改进的机制，即基于质量管理的PDCA循环——数据质量策划、控制、度量和改进。最后通过组织、流程、IT三个方面的能力保障，使数据质量管理“系统化”“持续化”“常态化”。

第9章

打造“安全合规”的数据可控共享能力

在企业进行数据治理，开展数据底座建设工作之前，用户经常面临的一个问题：使用数据做分析洞察的时候找不到数据，数据分散，或数据获取困难。

为了消除数据“孤岛”，我们构建了公司统一的数据底座，汇聚、联接大量的企业数据。但是，大量的数据汇集在一个湖中，如何在内外部合规的基础上，确保业务能够迅速获得所需数据，可控共享。这是企业在数字化转型过程中面临的共同问题，数据资产作为企业的核心战略资产，作为生产要素，锁在独立硬盘中是发挥不了价值的，那么，如何让数据在安全合规的前提下最大程度地发挥价值？这是数字化转型中的关键问题，如果数据的安全问题得不到妥善解决，那么宁愿数字化转型慢一点，或者不转型，也不能在错误的方向上渐行渐远。

9.1　内外部安全形势，驱动数据安全治理发展

9.1.1　数据安全成为国家竞争的新战场

随着近年来大数据、数字化转型的兴起，数据的价值获得了极大的提升，数据已成为企业和国家的“战略资源”和“生产要素”。

在工业时代，政府通过控制货物、人员、资金的流动来形成国家壁垒、实现国际影响力；到了数字时代，货物、人员、资金可以全世界自由流动，而跨区的数据流动反而受限制。数据管控力将成为衡量国家竞争能力的重要指标。

通过分析各国对网络安全、数据保护、隐私保护的立法进展，可以看出各国的立法进度都在加快。隐私保护立法都在向欧盟GDPR看齐，从原来依靠道德约束保护隐私，上升至法律约束。数字时代带来了新的发展机遇，也给数据安全带来了新的挑战。

9.1.2　数字时代数据安全的新变化

伴随着大数据、人工智能、区块链、物联网、5G等新技术的快速迭代和持续创新，各种新兴技术被越来越广泛地应用到各行各业，企业内、不同企业间、不同行业间的数字化迁移速

度不断加快。但是我们一定要认识到，现在的网络是不安全的，包括“云”也不是绝对安全的，2019年以来数据泄露事件频发，从数据上看，勒索软件、网络攻击的次数与2018年相比翻了25倍。“网络不安全”不再是偶然，而是常态，不容忽视。

从泄露的根因分析统计来看，随着数字化技术和能力的普及，泄露的路径越来越多元，已经不再限于“黑客攻击”，更多的是企业内部人员、离职员工、第三方外包的泄露行为。所以说“堡垒再坚固，也容易从内部攻破”，这些泄露都不是由技能高超的黑客造成的，而只是因为企业自身安全管理上的疏忽。图9-1为安华金和发布的《数据安全治理白皮书》中关于数据泄露的相关案例截图。

分类	序号	事件名称	事件时间	泄露人员	泄露数据量或非法所得
政府部门	22	南京公务员泄露居民信息	2018 年 1 月	内部人员，副主任科员刘某	82 万条
	23	明星购房信息泄露	2016 年 3 月	内部人员，青岛市不动产登记中心工作人员	明星购房信息
	24	国家旅游局安全事件	2015 年	外部不法分子	6000 万客户、6 万多个旅行社账户
	25	12306 网站用户信息外泄事件	2014 年 12 月	黑客，撞库黑客	13 万条
	26	车管所违章记录被篡改	2014 年 11 月	合作开发方，公安车管系统软件供应商	1.4 万条
	27	国家宏观经济数据泄露	2010 年～ 2011 年	内部人员，原国家统计局干部孙某、原中国人民银行干部伍某某、4 名证券行业从业人员	多次泄露
教育	28	学信网疑被拖库	2016 年 4 月	黑客，疑似拖库黑客	35GB（蓝点网）
	29	教育考试信息泄露	2016 年 8 月	黑客，直接攻击的黑客	非法所得 5 万
社保	30	篡改退休人员数据非法牟利	2010 年～ 2011 年	内部人员，某市社保局退管中心蔡某、市社保局信息中心陈某	非法所得 280 万
	31	非法获得养老金	2005 年～ 2008 年	内部人员，外部勾结；某区社保事业管理处副主任王某、某银行电脑维护员向某	非法所得 190.5 万
	32	冒领他人社保	2005 年～ 2009 年	内部人员，某市社保局支付股股长胡某	非法所得 90 万
医疗	33	疾控中心信息泄露	2016 年 7 月	黑客	30 个省的 275 例
	34	上海新生儿信息外泄	2016 年 7 月	离职人员，韩某原是上海疾控中心工作人员，张某原是黄浦区疾控中心工作人员	20 万新生儿信息

图9-1　数据泄露类型

大量新技术所带来的数据安全风险也急剧上升，以下三个事实无法回 避：数字时代丰富 的数据必然成为国 家与国 家间、企业与企业间竞争的关键；攻击者的攻击手法更加多样，数字化加速了泄露的便捷性；不管是传统数据库还是云端，基于网络边界的防 护必然会被突破。如何从安全能力建设的源头进行标准化风险预防，在安全可控的前提下最大程度释放数据共享的价值，是所有企业共同面对的课题。

9.2　数字化转型下的数据安全共享

数据安全是从决策到技术、从管理制度到工具支撑，自上而下贯穿 整个组织的完整链条。

非数字原生企业信息化程度差，存在割裂的信息孤岛，阻碍了企业的数字化转型。随着非数字原生企业的逐步转型，企业拥有的数据资产越来越庞大。商品的价值原理告诉我们：“买方的市场需求决定一件商品的价值。”那数据安全的核心价值就是“让数据使用更安全”。换句话说，数据安全与隐私保护的目标就是解决如何在安全前提下充分共享数据，如图9-2所示。



图9-2　数据安全治理价值

如果数据安全共享这件事情没有做好，不单数字化转型的成果无法呈现业务价值，很有可能企业多年积累的经营成果也付之一炬。

华为最近几年推进数字化转型，并梳理全部数据资产，明确了数据分类、数据标准、数据分布、元数据注册、访问方式、使用频率等。数据进底座、生成“数据地图”“数据随需共享”，成了华为数据治理的主要目标，让数据充分共享并为业务带来价值则是数据治理的主题。华为在全球范围内共享的数据服务有几万个，覆盖全球多个国家和不同的业务场景。

数据是不能藏起来的，数据存储起来就是为了消费，为了创造价值，支撑业务的决策、运营、经营、现场作业。大量的共享对安全提出了更高的要求，必须在安全、合规的基础上，才能共享数据。

企业在加速数字化转型，通过挖掘数据的价值抓住巨大的发展机遇。同时，企业面临的风险也在剧增，这些风险源于外部法律要求、网络安全威胁，也有来自内部数据的大量汇聚和充分共享。

充分认识到数据安全、隐私保护的价值后，我们还需要知道具体怎么解决数据的安全隐私问题。数据安全治理绝不是一套IT工具组合的产品级解决方案，而是从决策层到技术层、从管理制度到工具支撑，自上而下贯穿整个组织架构的完整链路。

9.3　构建以元数据为基础的安全隐私保护框架

9.3.1　以元数据为基础的安全隐私治理

有决策权的公司高层已经意识到安全隐私的重要性，在变革指导委员会以及各个高层会议纪要中都明确指明安全隐私是变革优先级非常高的主题，安全是一切业务的保障。

基于这个大前提，我们构建了以元数据为基础的安全隐私保护框架。在实际的管理流程中，如何利用元数据来管理好我们的安全隐私呢？安全隐私保护好比治疗过程，我们需要先做全面的体检（元数据发现），建立病历（信息架构、数据分类等），然后由专业的医生给出治理策略，也就是策略制定与执行数据保护和控制。整个过程都是以元数据为基础的，如图9-3所示。

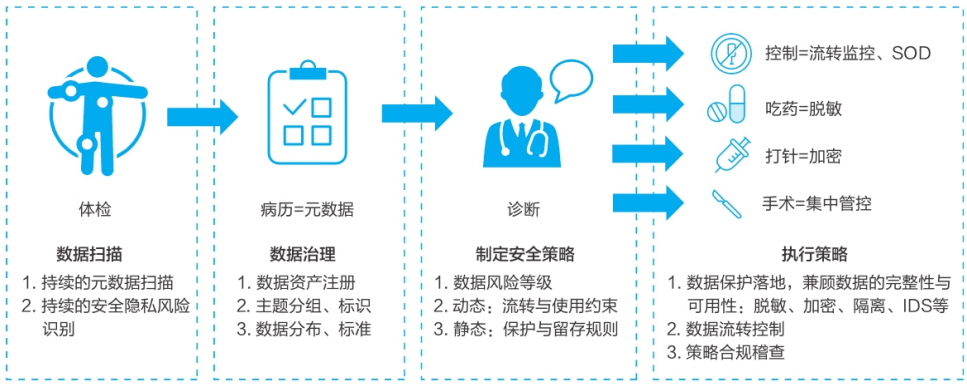


图9-3　元数据对安全隐私保护的作用

元数据就是描述数据的数据，即数据的上下文。而数据的管理要求、信息安全要求、隐私、网络安全要求等，都是数据的管理要素，当然也可以由元数据承载，用元数据来组织、来描述安全隐私管理策略和约束，如图9-4所示。



图9-4　元数据承载管理要素

治理安全隐私方案的思路，就是站在数据治理和元数据管理的基础上，构建对数据共享业务影响低且非介入式的治理框架。安全隐私保护的愿景是“让数据使用更安全”。为了让大家快速理解数据安全隐私保护的核心价值，整个数据安全隐私保护过程都要以元数据为基础，也就是都是以数据治理成果为基础来推进的。

9.3.2　数据安全隐私分层分级管控策略

在数据安全隐私管理政策一致性上，全球网络安全与隐私保护办公室和公司信息安全部发布了整体的管理策略，对整个

信息安全管理、隐私保护治理体系进行了分层映射，共同管理，如图9-5所示。



图9-5　数据安全隐私保护与公司安全隐私治理体系之间的关系

从公司层面，通过对整体内外部安全隐私管理政策的解读，将内部信息密级维度分为五类，要求组织间共享时一致遵从。

1）外部公开：指可以在公司外部公开发布的信息，不属于保密信息。

2）内部公开：指可以在全公司范围内公开，但不应向公司外部扩散的信息。

3）秘密：是公司较为重要或敏感的信息，其泄露会使公司利益遭受损害，且影响范围较大。

4）机密：是公司非常重要或敏感的信息，其泄露会使公司利益遭受较大损害，且影响范围广泛。

5）**绝密：**是公司最重要或敏感的信息，其泄露会使公司利益遭受巨大损害，且影响范围巨大。

基于业务管理的诉求，以内部信息密级维度为基础，从资产的维度增加两类划分，进行针对性管理。

1）**核心资产：**对应绝密信息，特指公司真正具有商业价值的信息资产。

2）**关键资产：**属于机密信息，特指对我司在消费者BG、5G领域领先战略竞争对手，在市场竞争中获胜起决定性作用的信息资产。

基于对GDPR的解读和企业内部的管理需求，将涉及潜在隐私管控需求的数据分为五类进行管理。

1）**个人数据：**与一个身份已被识别或者身份可被识别的自然人（数据主体）相关的任何信息。

2）**敏感个人数据：**指在个人基本权利和自由方面极其敏感，一旦泄露可能会造成人身伤害、财务损失、名誉损害、身份盗窃或欺诈、歧视性待遇等的个人数据。通常情况下，敏感个人数据包括但不限于可以揭示种族或血统、政治观点、宗教或哲学信仰、工会成员资格的数据，用于唯一识别自然人的基因数据、生物数据（如指纹），与自然人的健康、性取向相关的数据。

3）商业联系个人数据：指自然人基于商业联系目的提供的可识别到个人的数据。

4）一般个人数据：除敏感个人数据、商业联系人以外的个人数据，作为一般个人数据。

5）特种个人数据：GDPR法律中明文确定的特殊种类个人数据，严禁物理入湖，严禁共享及分析。

9.3.3 数据底座安全隐私分级管控方案

数据底座是“数据随需共享”的关键。在数据底座建设工作开始之前，业务经常面临的一个问题是，在做数据分析洞察时数据获取难，甚至有时即使知道数据在哪儿也拿不到。

例如某经营单元在构建其自身的经营分析指标沙盘时，需要用到61项指标。但由于数据获取没有明确的规则，需要逐个向各数据Owner索取相应授权，造成了“自己产生的数据自己无权访问”的现象。导致以上问题的原因就在于，公司虽然发布了数据共享的政策，但由于政策未完全落地，数据授权和权限控制机制不完善，导致数据获取需要通过邮件等方式层层审批，尤其是跨业务领域的数据获取需求，往往需要审批一二个月，极大地影响了业务决策的效率。

但是在数据底座建设好以后，我们又面临另一个重大问题，那就是在大量的数据汇聚到数据底座之后，如何才能保证这些数据的安全？当前在数据底座中已有超过数万个逻辑数据

实体，上百万张物理表，如果没有完善的安全管理措施，这些数据一旦泄露将会是一个巨大的灾难。

所以公司数据底座从建设起，就与公司安全、隐私治理体系之间建立了紧密的关系，在遵从公司安全隐私管理策略的同时，数据底座根据需要，发布相关操作流程、规范，指导数据工作。在应用数据安全与隐私保护框架和方法基础上，构建了数据底座的安全隐私五个子方案包。

1）**数据底座安全隐私管理政策：**说明数据底座的责任边界，数据风险标识标准、数据加工、存储、流转规范。

2）**数据风险标识方案：**平台提供的数据标识能力。

3）**数据保护能力架构：**数据底座分级存储架构能力。

4）**数据组织授权管理：**数据在组织内共享的规则。

5）**数据个人权限管理：**个人访问数据的权限管理方案。

数据底座的安全隐私保护方案如图9-6所示。

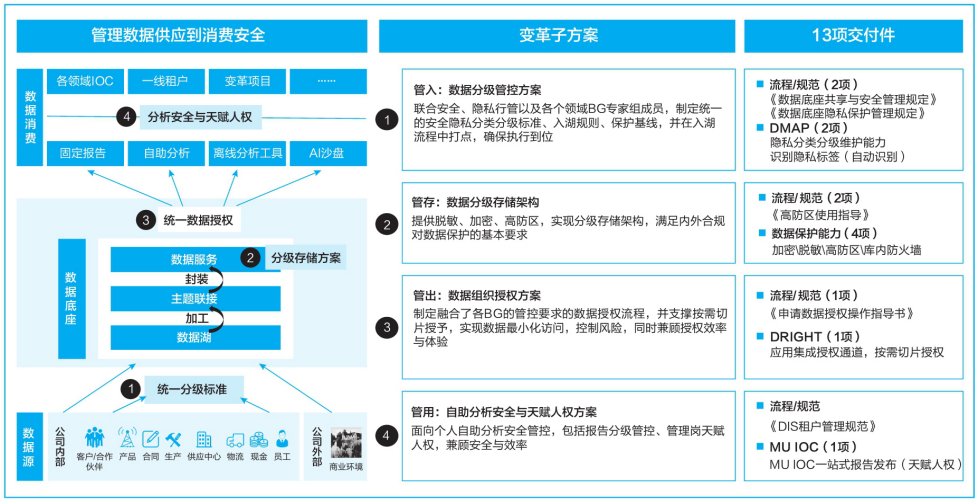


图9-6　数据底座安全隐私保护方案

在数据安全方面，根据公司信息保密规定，数据底座安全管理总体原则与数据管理原则是一致的，即“**核心资产安全优先，非核心资产效率优先**”。数据安全管理的基本原则如图9-7所示。



图9-7　数据底座安全管理

数据安全规范主体包括三部分。

1）数据密级分级标准：数据定密的标准，包括外部公开、内部公开、秘密、机密、绝密五个等级。

2）存储保护的基线：描述每一个级别的数据资产的存储要求以及入湖原则。

3）流转审批层级：描述每一个级别的数据资产在申请数据共享时应该经过哪些控制审批。一般控制审批流程下，内部公开数据不需要审批，在流程中自动存档并知会数据消费方直属主管。秘密数据由消费方直属主管审批即可，机密数据需要数据生成方和消费方双方数据Owner共同审批。

在隐私保护方面，根据公司隐私保护总体纲领文件和数据底座自身的特点，发布了数据底座隐私保护规定，总体原则是“个人数据原则上不入湖，并尽可能脱敏处理”。数据底座隐私保护管理原则如图9-8所示。

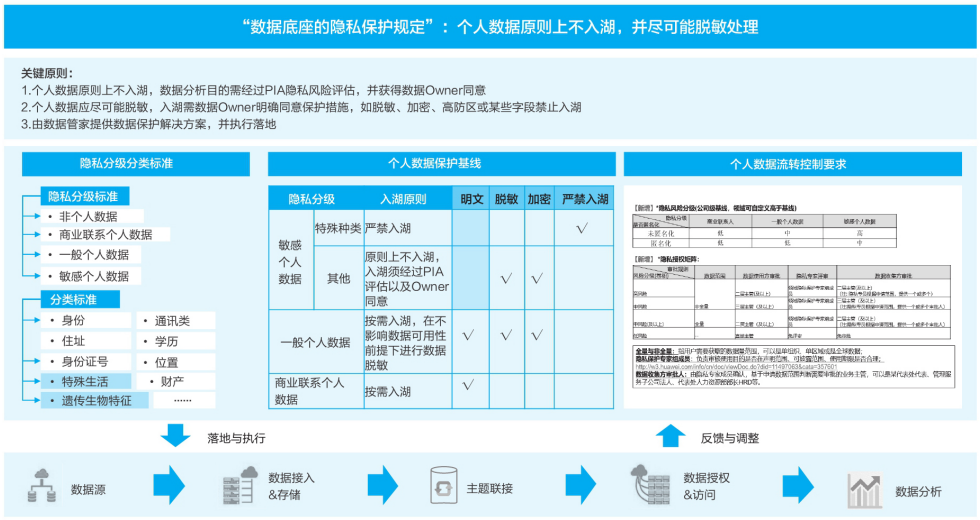


图9-8　数据底座隐私保护

隐私保护规范主体包括三部分。

1）个人数据分类、分级标准：非个人数据、商业联系个人数据、一般个人数据、敏感个人数据，共4个级别。

2）个人数据保护基线：根据个人数据分级，敏感个人数据、一般个人数据、商业联系人分别需要做不同程度的数据保护，其中法律明文规定的特种个人数据严禁入湖。

3）流转审批层级：隐私审批层级基本与安全一致，但新增了隐私专员的介入，以专家评审身份，参与控制数据流转业务，判别数据消费的目的限制以及最小化授权。

9.3.4　分级标识数据安全隐私

在明确数据分类分级标准的基础上，还需要有具体的平台支撑数据风险标识。这就包括传统的元数据人工标识方案以及通过规则、AI自动推荐方案。

1）人工识别数据风险。数据安全隐私分级标识必须基于元数据管理平台，在平台中构建对数据字段级别的风险标识。

2）基于规则与AI的自动识别。在数字时代，随着数据资产的膨胀，数据风险标识工作量非常巨大，字段的数量是数据表数量的100倍有余，依靠人工的方式无法识别全面，需要通过工具，基于规则（正则表达式）以及AI机器学习 的方式，构建自动推荐、识别风险标识的能力。

在数字时代，数据的安全隐私也得到了越来越多企业的重视，这其中也包括非数字原生企业，因为这类企业手中的生产工艺和研发数据中往往包含大量的专利成果和机密配方，而识别数据资产、管理元数据、标识安全隐私，只是“安全合规”共享数据的第一步。

9.4 “静”“动”结合的数据保护与授权管理

9.4.1 静态控制：数据保护能力架构

在充分识别数据风险并标识数据安全隐私后，数据底座产品还需要提供不同程度的数据保护能力。数据保护能力包括存储保护、访问控制、可追溯三种，每种保护能力都面向不同的业务管理需求，如图9-9所示。

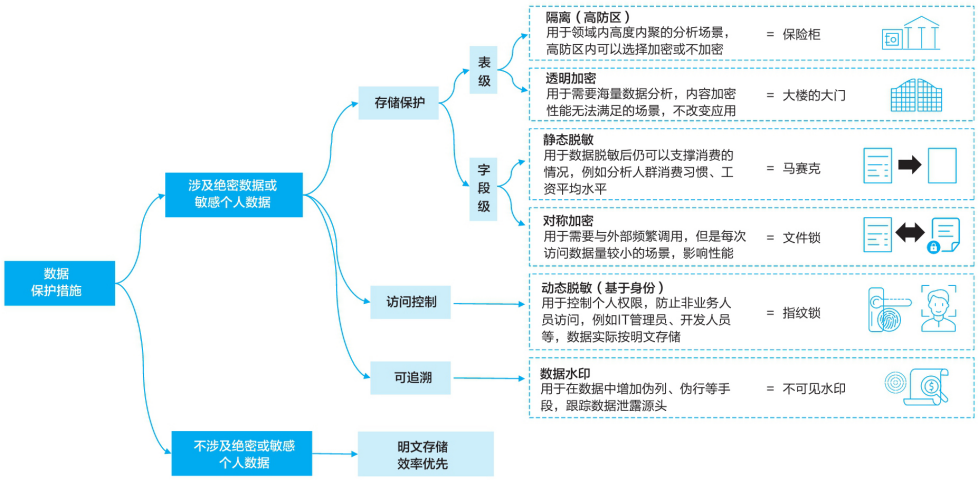


图9-9 数据隔离：数据底座高防区

1. 存储保护

存储保护能力包括面向表级管理的高防区隔离、透明加密和基于字段级的对称加密和静态脱敏。

1）高防区隔离：高防区隔离就是我们通过在数据底座独立部署单独的防火墙以及配合流向控制、堡垒机等措施，对高密资产重点防护。关键要点就是有独立的防火墙，并且内部区分脱敏开发区以及明文业务访问区，让数据开发人员在脱敏区工作。高防区数据经过审核后才能发布到明文区，给业务部门使用。

2）透明加密：透明加密就是对表空间进行加解密，进入表空间的表自动加密，有权限的应用读取表空间的表时就自动解密。主要用于防止黑客把库文件搬走。

3）对称加密：对称加密指应用对数据字段应用对称加密算法进行加密，需要配合统一的密钥管理服务使用。

4）静态脱敏：首先需要从技术角度制定出脱敏标准。脱敏不是单一的技术能力，而是多种脱敏算法的合集，包括加噪、替换、模糊等，每种数据类型应该有不同的脱敏标准。我们在ETL集成工具中增加脱敏API能力，可以对具体的字段进行脱敏，每类数据字段都依据脱敏标准执行。

2. 访问控制

静态脱敏用于存储保护，而动态脱敏则是一项基于身份的访问控制。通常Web应用都是使用自己的菜单和角色权限进行职责分离，对于数据权限，很难做到字段级别的控制。而动态脱敏可以对某些数据表、数据字段根据身份进行脱敏，从而做到更细颗粒度的保护。

3. 可追溯

在可追溯方面，业界有比较成熟的数据水印技术。简单来说，是直接改动数据，在数据行、数据列中增加水印，不影响数据的关联与计算，适用于核心资产或敏感个人数据。一旦发生泄露，可以溯源定责。

9.4.2 动态控制：数据授权与权限管理

对数据的保护，只是采取合理和适当的措施保护信息资源，但是数据在组织内部肯定是要流动的，需要被加工、消费，需要创造价值。而脱离业务流，脱离生产、决策的数据，是死的字节，不能称其为数据资产。

1. 数据授权管理

数据授权和数据权限是两个不同的概念。数据授权主要是面向组织，指数据Owner对组织授予数据访问权的过程，让数据与组织绑定，为组织提供长期的数据订阅权限。数据授权包含两个场景。

1）数据加工授权：由于数据主题联接资产建设中需要跨组织进行数据联接、加工、训练需要转移数据而发生的数据授权场景。

2）**数据消费授权：**由于业务用户数据的分析需要订阅数据服务而发生的数据授权场景。

数据授权管理要基于数据风险标识和数据保护能力，既能在数据流转中落实安全隐私控制策略，让数据安全隐私政策落地，又能作为数据架构治理的抓手，融入架构审核，避免重复建设。

2. 数据权限管理

数据权限管理是基于访问管控规范，对授予的数据访问权限进行管理的过程。面向个人和面向与岗位绑定的综合管理者的管理策略不同。

面向个人，指业务制定数据访问管控规范，授予个人数据访问权限的过程，具有与个人绑定、短期有效的特点。基于消费数据类型的差异，个人数据权限分为两大场景（如图9-10所示）。

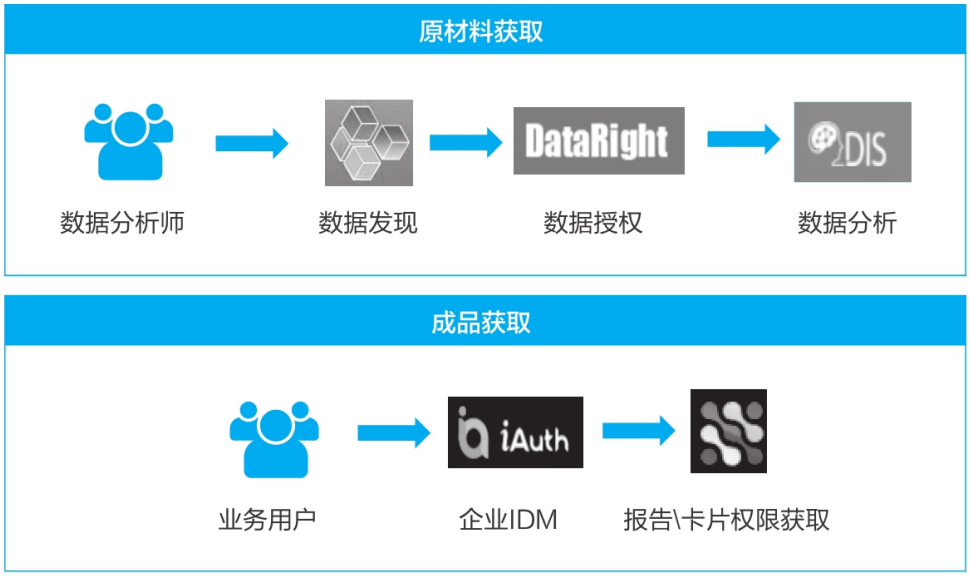


图9-10　区分原材料与成品的访问权限管理

1）业务分析师获取数据资产（原材料场景）。

2）业务用户获取报告访问权限（成品场景）。

基于企业IAM（身份识别与访问管理）和IDM（账号权限管理），结合数据分级管理机制，让数据权限随人员流动而改变，并统一规则、集中管控高风险数据，实现对个人权限授予、销权、调动全生命周期集中管控。

而对于综合管理者，引用人力资源管理岗的信息，当管理者被任命或者调动交接后，会执行相应的授权和销权操作。这个过程是全自动的，无须管理层的操作，在有效权限管理的基础上提升了用户在权限管理下进行数据消费的效率和体验，如图9-11所示。

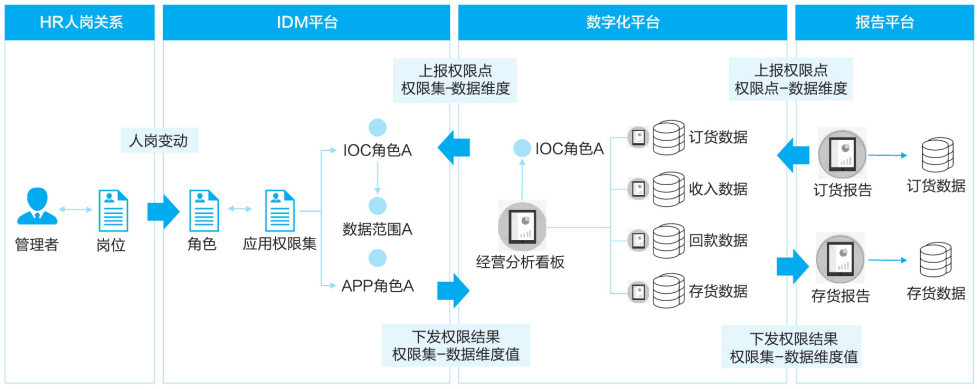


图9-11　管理岗自动赋权逻辑

为打造“安全合规”的数据可控共享能力，我们践行了数据安全隐私管理不仅仅是一套IT工具组合的思路，基于安全隐私的两个公司级治理文件，通过“数据底座共享与安全管理规定”和“数据底座的隐私保护规定”，落实管理要求，分别建设了数据标识、存储保护、授权控制、访问控制的能力。同时平台调用了传统IT安全措施，通过态势感知、堡垒机、日志服务等，结合数据安全治理方法与传统的IT安全手段，做好数据的内外合规，形成完整的数据安全与隐私保护，实现让数据使用更安全这一目标。数据安全与隐私保护能力架构如图9-12所示。



图9-12　数据安全与隐私保护能力架构

9.5　本章小结

数字技术正在构建一个全新世界。在数字时代这个大风暴中，数据的安全隐私管理无异于风暴之眼，纷乱的外部因素与企业自身特定的安全威胁正在共同影响着整体安全隐私态势，既要求企业可以减轻安全威胁，避免内外安全隐私风险带来的信誉损失和经济损失，又要求企业最大化利用数据、共享数据，面向大数据和机器学习，达成业务目标，发挥数据价值。所以数据保护和数据共享作为一对矛盾体，将不断引入新的理念。国际数据空间技术、“链条控制”转向“集中管控”、构建基于元数据管理的影响小、非介入式的公司级数据安全隐私保护平台，都会在数字时代不断演进，不断发展。

第10章

未来已来：数据成为企业核心竞争力

数字化转型不能一蹴而就，数据治理也不是一朝一夕之功。数字化转型带来机遇的同时，也给整个企业的数据治理带来了新的挑战。

基于对华为公司数字化转型的解读，我们建立了数据综合治理体系，发布了信息架构，构建了数据湖、数据底座，打造了数据感知、安全合规能力，提升了数据质量。但是，在数据成为新的生产要素，数据成为企业核心竞争力的情况下，未来已来，面对这样一个新的、复杂的内外部环境，非数字原生企业在数据治理的问题上，做了哪些思考？我们应当如何应对？

10.1　数据：新的生产要素

数字化变革改变了人们看待数据的方式。数据不再仅仅被视为商业活动的副产品，而是战略资源，是发展和提供新型数字产品与服务、建立新型数字商业模式的基础。

10.1.1　数据被列为生产要素：制度层面的肯定

从制度层面来看，2019年10月，中国共产党第十九届中央委员会第四次全体会议审议通过了《中共中央关于坚持和完善中国特色社会主义制度、推进国家治理体系和治理能力现代化若干重大问题的决定》（以下简称《决定》）。《决定》指出：“健全劳动、资本、土地、知识、技术、管理、数据等生产要素由市场评价贡献、按贡献决定报酬的机制。”这是首次将数据作为与劳动、资本、土地、知识、技术、管理并列的生产要素，从制度层面确立了数据作为一种新的生产要素的重要地位。如何促进数据要素有效参与价值创造和分配，成了数据新时代交给我们的重要课题。

生产要素这一概念，自古典经济学时代以来，就是经济学家们关注和讨论的重点，他们在著作中不惜花费大量篇幅来定义、解析、演绎、推论各项生产要素。人类社会进入数字时代后，数据因其在生产过程中的巨大增值作用而被列为生产要素，这是历史的进步。

2020年4月9日，《 中共中央、国 务院 关于构建更加完善的要素市场化配置体制机制的意见》（以下简称《 意见》）正式公布。《 意见》分类提出了土地、劳动力、资本、技术、数据五个要素领域改革的方向，明确了完善要素市场化配置的具体举措。《 意见》提出要从三个方面加快培育数据要素市场：推进政府数据开放共享；提升社会数据资源价值；加强数据资源整合和安全保护。

数据作为生产要素的美好前景在我们眼 前展开：加快数据要素价格市场化改革，健全数据要素市场运行机制，并提供组织保障已经提上了议 事日 程。各地区、各部门间的数据共享交换将得到快速推进，数据共享的责任将进一步明确，公共的、基础的数据资源将得到有效流动。在数据开发利用规范化和数据采集标准化的基础上，数字经济会不断涌 现新产业、新业态和新模式。与此同时，数据管理制度得到进一步统一规范，数据质量和规范性不断提高，数据分类分级安全保护制度不断完善，企业商业秘密和个人隐私数据得到更好的保护。

10.1.2 数据将进入企业的资产负债表

从企业管理的层面来看，“大数据商业应用第一人”维克托• 舍恩伯格在其2012年出版的新作《 大数据时代》一书中意味深长地说道：“虽 然数据还没有被列入企业的资产负债表，但这只是一个时间问题。”当下我们回 顾这句话，更加感觉这个“时间”已临近。

根据企业财务管理对于资产的定义，只有被企业拥有和控制，并且能够用货币 计量，能够为企业带来经济利益的数据，才能成为企业的生产要素和资产。这三个方面，无疑对我们的数据管理工作和企业的数字化转型提出了努力的方向和要求。也就是说，除了要发挥数据的价值之外，还要重视和关注数据的主权与定价。

数据能创造价值，但数据创造价值的功能并不能由数据自身来直接实现，数据要素也不能直接参与价值分配，而是要经过数据创造、加工并传输给数据要素使用者后，才能创造价值，进而参与价值分配。由此可见，在数字时代，能否掌握数据资产并将其有效转化为生产要素，已经成为衡量一个企业核心竞争力的决定性因素。

从本书前面的内容大家也可以看出，我们始终围绕数据这一新的生产要素的两个方面来论述：第一，如何提高数据资产的利用率；第二，如何降低数据的运行维护成本。

衡量一项资产的价值要从资产未来带来的经济利益来看。对于数据资产而言，应该从最终的应用价值出发来衡量数据价值，而我们采集和整理数据的时候，往往很难确切预计到数据资产后续会如何被使用、会产生多少数据价值，所以对数据资产的价值评估是一个持续更新的动态过程。

在经济学中，国民 收入是指物质生产部门劳动者在一定时期所创造的价值，是一个国 家的生产要素所有者在一定时期内提供生产要素所得的报酬，即工资、利息、租金和利润等的总

和。这表示，劳动、土地、资本、管理等传统的生产要素对最终价值的贡献是加成的。不难想象，因为数据能够提升劳动者能力、加速资本周转、加速知识转化、推进技术进步、提高管理水平，所以数据对最终所得收益的贡献将是一个乘数因子，而非简单的加成。数据对前面几项生产要素所得的报酬或多或少都有提高作用。

10.1.3 数据资产的价值由市场决定

数据资产的定价权掌握在市场手中，这就意味着数据资产价值的公允计量取决于市场参与者通过使用资产而创造经济利益的能力，只有努力提高企业自身利用数据资产提升生产效率的能力，才能主导数据资产的市场定价权。我们利用数据资产来提升生产效率的能力，主要体现在提升营收、提高效率、提高质量、提升速度、增加柔性、降低成本、减低风险、品牌增值等方面。而传统做法中，基于数据自身属性（比如数据的数量、质量、及时性、稀缺性、稳定性以及法律限制等）来决定价格的方式，相比之下就缺乏对数据价值的想象力。

作为生产要素的数据资产，其折旧损毁和保值增值是一对永恒的矛盾。数据资产如果不加以有效利用就会逐渐贬值，甚至成为资产负债表上的负资产。实现数据的保值增值，要从扩大数据生态、提高数据活性两方面入手。数据生态会带来数据有效连接的扩大，网络效应获得的数据价值提升往往是指数级的。数据的活性既包括数据自身的时效与质量，又包括使用者对数据的认知和黏性。

当然，数据与传统的生产要素的特点不同，数据的交易、定价、主权保护、收益分配等方面也还存在很多理论空白。可以想见，在不久的未来，在经济学界将掀起对这一生产要素的研究热潮。比如，通过数字化转型促进了企业资产的保值增值，但企业的数据资产自身的变现与保值增值仍然是一个开放的研究课题。

10.2　大规模数据交互的企业数据生态

相对于数字原生企业来说，非数字原生企业更需要通过数字化转型来实现变革，共同构建以数据为中心的商业生态系统。平台化的数据生态系统为商业模式创新带来了更多机遇，在生态系统中，众多企业围绕共同的客户价值主张，通过竞争与合作方式发展各项能力。相比之前，生态伙伴的作用及彼此之间的数据流正在发生根本性的变化。生态系统是数据经济中的一种多边组织形式，它与其中的个体及整个共同体均是互利双赢的关系。

10.2.1　数据生态离不开底层技术的支撑

在万物互联的智能世界，数据工作所依赖的绝不是一个封闭的系统，而是一个开放的生态。尤其是在大数据时代到来之后，多类型、多形态的数据可能会给我们带来超出预期的收益。对数据生态的梳理和发掘，也会降低依赖于单一数据供应来源所导致的单点失效的风险。

未来的业务不断变化和发展，企业对于数据的需求没有边界。如何建立跨越企业边界的数据共享平台，建立安全可信的数据生态，以尽可能低的成本交换数据、共享数据，将是一个长期而具有挑战性的课题。

由于数据的可复制特性，我们在实施企业内外部的数据共享场景时，存在很大的痛点：在企业间大规模数据交互的场景下，双方通过协议合同定义了严格的数据处理流程条款，包括对逾期数据的及时删除等，这就需要企业投入不少的IT及业务人力来保障条款的落实，以满足内外部合规的要求。

因此，我们的数据生态建设目标是：从依赖管理手段到依赖自动化技术，增强数据管理的可信、透明；通过基于密码学和区块链技术的智能合约代码化，支撑商业生态系统的数据安全交换；构建统一标准的数据交换空间，实现与客户、合作伙伴协同的数据生态体验。

10.2.2　数据主权是数据安全交换的核心

在数据生态系统中，大规模交互的数据是一种战略资源。数据主权是数据生态系统中数据安全交换的核心。数据主权是自然人或公司实体对其数据进行排他性自决的权利。

数据主权的提出，旨 在建立一种便于在数据生态圈 内交换数据同时确保数据主权的架构方法，使企业能够在安全可信的数据生态系统中发挥数据的价值。基于数据主权保护的原理，数据所有者在将数据发送给数据消费者之前，需要将访问及使用控制信息附 加到数据中，数据消费者只有完全同意该原则才可以使用该数据。

数据主权与云主权、数据采集组件的主权共同构成了完整的生态主权。

那么，数据主权管理与我们所熟知的数据所有权管理有什么区别呢？数据所有权管理针对的是数据提供，确保数据同源可信，数出一孔；数据主权管理针对的是数据访问与使用，确保数据安全共享，防止数据滥用。数据生态下的安全交换架构如图10-1所示。

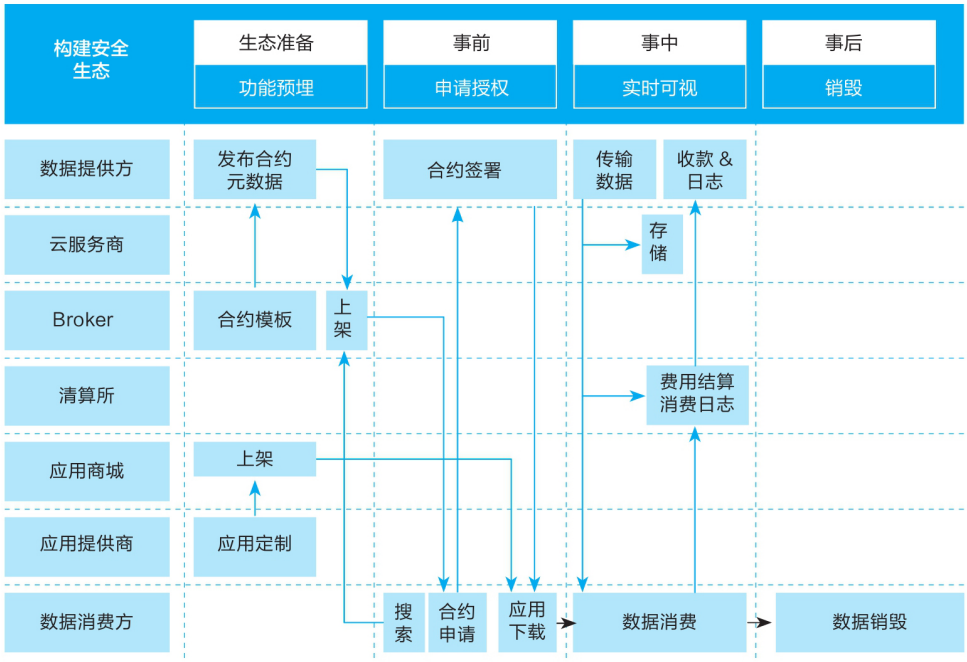


图10-1　数据生态下的安全交换架构

企业内部的数据孤岛可以通过数据底座的建设融合在一起，而企业间的数据孤岛就没有这么容易改变了。近年来，以欧盟GDPR为代表的数据隐私监管力度加强，企业对数据共享的态度更为谨慎了。企业间数据生态的建设呼唤新技术的支持，

传统的电子数据交换模式已经不能适应现代的业务和监管要求。

10.2.3 国际数据空间的目标与原则

国际数据空间（International Data Spaces，IDS）是一个虚拟数据空间。它利用现有数据标准和数据技术以及公认的数据治理模型，推动数据在可信的商业生态系统中进行安全、标准化的交换并促进数据连接，从而为各种智能服务场景及跨公司业务流程提供基础设施，同时确保参与其中的数据所有者的数据主权得到保障。为了履行这项职能，IDS不存储信息，而是通过身份提供者及经认证的连接器组件，在经认定的通信伙伴之间建立安全交换。外部数据交换是企业的业务交流过程中的一个重要方面，来自外部合作伙伴的数据可用于提升运营服务。

IDS协会作为一个非营利组织，推动了一系列研发活动以及标准化工作。来自世界各地不同行业、不同规模的众多公司均已加入该协会，成为不可或缺的一部分，确保整体的架构得到发展。

数字化会促使各组织之间共享更多的数据。IDS很早就认识到，即使在其组织之外，数据所有者也需要掌控其数据资产。因此，IDS倡议将数据主权作为体系架构开发的一个核心层面。数据主权可定义为自然人或法人实体对其数据享有完全自决的

权利，IDS倡议为该项特定权利及在商业生态系统中进行安全、可信的数据交换的需求等相关内容提出一个参考架构模型。

IDS的目标是满足以下战略需求：

- **信任：**信任是IDS的基础。通过全面的身份管理，重点确定参与者的身份，并根据对所有参与者进行的组织评估和认证结果提供有关参与者的信息，来确保他们彼此之间的信任。
- **安全与数据主权：**IDS的各个组成部分依赖于现有的安全措施。除了架构规范之外，还通过对各个组成部分进行评估和认证来确保安全和数据主权。根据确保数据主权的核心要求，IDS中的数据所有者在将其数据传输给数据消费者之前，会给其数据附上使用限制信息。只有完全接受数据所有者的使用策略，数据消费者才能使用这些数据。
- **数据生态系统：**IDS架构无须具备集中的数据存储能力，相反，IDS架构追求分散化数据存储，这实际上意味着，在传递给可信任的另一方之前，数据仍然由各自的数据所有者所有。该方式要求将数据源和数据作为资产进行全面描述，并能够为数据集成特定领域的词汇表。生态系统中的代理，可以实现全面实时数据搜索。
- **标准化的互用性：**IDS连接器作为架构的核心组件，由不同的供应商提供并以不同的形式展现。然而，各连接器都能够与IDS生态系统中的其他连接器或组件进行交流互动。
- **增值应用：**IDS允许将应用嵌入各个连接器，以便在纯数据交换的基础上提供服务。这包括数据处理服务、统一的数

据格式和数据交换协议以及通过远程执行算法进行数据分析。

- **数据市场：**IDS允许创建使用数据应用、基于数据的新型服务。IDS还通过提供结算、计费和创建特定领域的代理与数据市场，为这些服务打造全新的商业模式。此外，使用限制和法律问题也作为模板提供，并提供相关方法供参考。

另外，作为研究项目的核心交付件，IDS参考架构模型构成了各种软件实施的基础，因此也是各种商业软件和服务的基础。一系列研发活动以及标准化工作遵循以下准则开展。

- **复用现有技术：**组织间信息系统、数据互用性和信息安全是确立已久的研发课题，市场上有大量技术可供选择。IDS倡议相关工作遵循的理念不是“白费力气做重复工作”，而是尽可能使用现有技术（如来自开源领域的技术）和标准（如万维网联盟的语义标准）。
- **标准化贡献：**由于本身是为了建立国际标准，IDS倡议支持标准化架构堆栈的想法。

为了构建安全的数据生态，需要在生态准备阶段预置相应的功能模块，在事前处理申请授权，在事中做到数据流动与结算实时可视，在事后保证完成共享使命的数据能够得到及时的、合规的、安全的销毁。整个过程需要生态中的各个组件角色（数据提供方、云服务商、Broker、清算所、App商城、App提供商、数据消费方）的通力协作。

10.2.4　多方安全计算强化数据主权

除了基于公钥基础设施（PKI）的数据共享以外，联邦学习也是构建良 好的数据生态、突破企业数据墙有力的技术武器。联邦学习 技术底层依赖于同态加密、秘密共享、散列、梯度交换等多种多方安全计算机制，在算法层面上可以灵活支持多方安全计算模式下的逻辑回归、Boosting（提升法）、联邦迁移学习 等多种算法模型，可以实现在保护本地数据的前提下让多个数据拥有方联合建立共有的数据挖掘模型，从而实现以隐私保护和数据安全为前提的互利共赢。基于联邦学习 技术构建的新型数据生态，有利于打破企业间数据壁垒，实现大范围、高密度的数据空前的融合协作，辅以密码学、区块链等相关技术，必将实现一幅前所未有的惠及各行各业的数据生态场景。

在联邦学习 的机制框架下，不仅数据不需要传输出企业边界，算法模型本身也不需要传输，因为对有些企业，算法模型也是非常重要的信息资产，是核心竞争力，安全要求很高。

参考描述网络价值的梅特卡夫定律，数据生态圈 作为一个由数据拥有者实体组成的网络，其价值与加入成员数的平方成正比。加入的成员数目越多，那么整个数据生态网络和该网络内的每个成员所能贡献和收益的价值也就越大。随着生态网络规模在多个领域的扩张，生态里会涌现出各个不同垂直领域的数据生态圈，如电 信运营商的数据生态圈 等。当然也会有跨领域的数据生态合作，比如涉及公共事务的数据分析，往往需要跨越多个领域获取数据。数据生态的构建和参与能力，以及在

其所处生态圈中的地位和影响力，将成为未来企业的核心竞争力。

10.3　摆脱传统手段的数据管理方式

10.3.1　智能数据管理是数据工作的未来

在以传统方式对数据实施管理和治理的过程中，数据工作者和业务方都需要投入相当多的人力和资源，才能达成管理目标，其中的艰辛，相信各位业内人士都深有体会。而随着智能大数据时代的到来，各行各业都看到了摆脱传统工作方式的希望。在数据工作方面更是如此，因为我们的工作对象天然具有极高的数字化程度、极具规模的体量、强大的内生关联度，我们更需要应用智能化、数字化的新方法来提升工作效率和效果，借助于数据挖掘、机器学习、数据可视化等方法来更深入地了解海量、复杂、多维、高度互联的数据，让企业的海量数据更加透明、可知、易用。

10.3.2　内容级分析能力提供资产全景图

举个例子，初步完成数据的架构工作并构建了企业级的数据湖之后，我们就可以基于多维数据特征的可视化分析技术，对数据质量进行内容级分析，采用特征工程方法，建立数据内容的多维模型，在高维空间进行多维度聚类，利用可视化投影技术在二维平面进行渲染展示。与传统的表格式数据展示不同，这种基于内容解析的数据资产智能分析会有诸多强大的应

用场景，全景展示所有已经进入企业数据湖的表字段及其关系结构只是其中最为直接和显而易见的应用。

10.3.3 属性特征启发主外键智能联接

数据表之间的主外键关系是ER模型中的重要组成部分，蕴含了对后续数据加工利用有重大价值的信息。然而，出于对性能等因素的考量，很多实现场景并未将这一信息传递到数据供应链的下一阶段，造成重要信息丢失，给数据管理带来了不小的困扰。传统IT系统及其开发造成的这一困境，是否可以利用先进的数据分析技术予以弥补乃至解决呢？我们观察到，在全景图中若干个属性字段投影位置重叠，表明它们的数据指纹几乎一致，很有可能是可以做主题连接的主外键。基于这一启发，辅以对主外键关系存在诸多属性约束的条件的帮助，通过实验证实，我们可以以很高的准确率重建已经丢失的主外键关系，加速主题连接的创建和拓展，让已有数据通过更多、更准确的连接发挥更大的业务价值。

10.3.4 质量缺陷预发现

数据质量话题，在前面已经有专门章节论述，这里不再赘述。我们想补充的是，除了已有的基于规则对质量的方方面面进行有尺度的微观管控和宏观治理之外，我们也可以利用大数据分析方法，进行介观层面的分析管理。之所以称之为介观层面，是因为通过大数据分析与可视化方法，我们可以以极快的

速度在宏观和微观之间切换，以前所未有的人机交互的方式观察数据分布和异常，从而在很大程度上提升管理水平和效率。简单来说，比如我们观察到，相似类型的数据通常呈聚集状态，远离数据群的属性节点则往往需要质量人员的更多关注。

10.3.5 算法助力数据管理

另外，我们可以利用基于密码学的资产指纹技术来更好地管理数据架构。由于大量数据表中含有相同或相似的字段，且判断两张数据表是否同源比较耗时，因此我们对每张数据表的字段名进行快速编码，实现数据表快速比对判重，而不受表中各字段排列顺序影响。我们已经为物理级数据资产建立了数据架构指纹库，支持快速查询、资产去重、篡改发现、资产比对。

随着计算能力的不断提升和智能算法的不断优化，我们越来越能够对数据的实质内容而不仅仅是元数据进行深入分析。相信在不久的将来，我们会看到越来越多的智能数据分析算法应用于企业内部的数据管理和治理任务中，让我们数据工作者从繁重的数据处理分析中解脱出来，有更多的时间思考、设计和解决数据管理的本质问题，既能下沉到数据里触摸到落地的细节，又能上升到整个全景把握好宏观趋势。

10.3.6 数字道德抵御算法歧视

基于数据的算法因其黑盒的特性而在某种程度上诱导人类让出了自己的决策权，我们应该如何重建数据空间里的信任关系呢？数据道德准则的建立迫在眉睫。我们需要对数据流程上的各个环节所受的影响进行分类，谨慎评估潜在的道德和伦理风险，充分测试、模拟和评估数据系统，提高算法模型的透明度，遵循最佳实践进行数据分享。采集数据之前要取得知情同意，对数据匿名化的能力和限度有充分认知，从而有效地保护数字道德不受到我们自己亲手构建的系统的伤害。

10.4　第四个世界：机器认知世界

宇宙自诞生起，一直以恒定的规律在运转，从而构成了我们身处的、已知或未知的物理世界。当代的考古学家和人类学家发现六百万年前已经有了人类的踪影。几百万年来，人类经历了漫长的演化，首先是直立行走、解放双手、制作工具，然后大脑发育出不同的区域，开始了对物理世界的感知，产生了语言与思维，发展了抽象与分析能力，创建了数学、物理学、天文学、生物学、社会学等科学。人类自产生以来就开始了对宇宙这个物理世界及其运行规律的永无止境的认知。

10.4.1　真实唯一的“物理世界”和五彩缤纷的“人类认知世界”

在数字技术出现之前，人类通过文字和数据记录了对现实“物理世界”的探索和抽象过程，形成了“人类认知世界”。人类对世界的认知水平，受限于其自身的知识、智力、经验和当时的科技水平。亚里士多德认为重的物体比轻的物体下落快，后来被伽利略证明是错的，哥白尼通过天体观测提出了日心说，牛顿的微积分和万有引力定律对世界的认知达到了数理演绎的阶段，爱因斯坦的相对论又进一步颠覆了绝对时空观。对“物理世界”中的同一事物，每个人的认知是不同的，带有明显的个体或群体的责任标识，认知不同，世界观不同，流派

层出，形成了百花齐放、百家争鸣、五彩缤纷的人类认知世界，对“物理世界”的持续不断的认知、再认知，产生了以文学、哲学、科学、宗教为标志的人类文明，推动着社会进步。

以本书的主题数据为例，数据是对真实世界中的对象、事件和概念的某一属性的抽象表示，可以说数据创建这一抽象过程，就是人类对“物理世界”的认知过程。例如在大型企业的信息系统中，“产品”是一个非常重要的主数据，对产品的定义主要受两方面因素影响：

第一，不同的功能部门对产品的认知是不同的。销售部门定义的产品应该是可销售单元，其产品结构是由可销售单元组成的；产品研发部门眼中的产品是从功能和系统的角度定义的；供应链关心的是产品的制造单元和交付单元；实施部门需要有清晰的产品安装单元和结构；财务关心的则是产品盈亏核算单元。可见针对同一个销售给客户并安装运行的产品，销售、研发、制造、供应、实施、财务等部门都会对“产品”进行定义，也就是进行不同的特性抽象和认知。

第二，对于大型企业，往往存在多个产品事业部，存在多个信息系统，对“产品”数据定义除了受到“功能部门”视角的影响，还取决于不同架构师的经验和抽象能力。

“人类认知世界”带有明显的个体差异性，“一千个人心中有一千个哈姆雷特”，也就是说不同个人和组织对同一现实对象的认知很大程度上是不同的。常常听到大型集团的CIO（首席信息官）抱怨，同类产品数据在不同系统中的数据标准和结

构不一致，拉通汇总困难。人类个体的认知能力取决于其知识、智慧和经验，其认知能力和认知视角的差异正是形成企业“数据孤岛”的根本原因。本书前面谈到的数据治理，其核心应该是在一个企业内部对如何描述业务的数据语言形成统一的认知，遵守统一的原则，这样可以大幅降低企业的数据处理成本，提升交流沟通效率，促进对未知事物的认知。这就是为什么企业、行业、国家、国际组织都在努力制定一个个“标准”，在某种程度上统一“认知”，形成我们所说的共识。

10.4.2 映射“物理世界”的数字孪生——“数字世界”

二十世纪四十年代，数学家香农提出的采样定理是数字化技术的重要基础，即在一定条件下用离散的序列可以完全代表一个连续函数。基于香农定律的现代数字技术，我们已经可以通过对物理世界的感知，构建出完整映射的数字世界，“数字孪生”“数字世界”等概念应运而生。“数字世界”的形成，使我们可以通过对“数字世界”的认知来达到对现实“物理世界”的认知，消除了认知过程中在时间和空间上的约束，大大提升了人类对物理世界的认知能力。

目前，“物理世界”“数字世界”的表述和概念已被广为接受，当然还有我们前面提到的“人类认知世界”。按照概念出现的先后顺序，我们不妨称“物理世界”为第一世界，“人

类认知世界”为第二世界，“数字世界”为第三世界（如图10-2所示）。

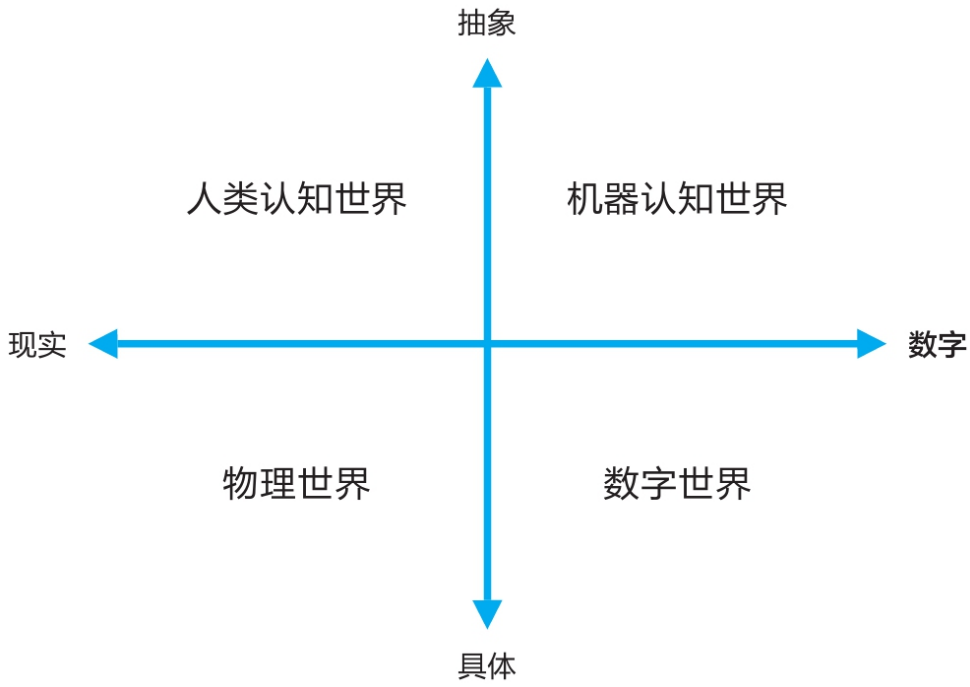


图10-2　四大世界的象限表示

10.4.3　“数字世界”中的智能认知——“机器认知世界”

随着以算法、算力和数据为基础的人工智能的发展和广泛应用，我们可以认为出现了第四个世界——“机器认知世界”，即基于大量数据，各种人工智能“机器”按照各自的算法对映射到数字世界中的事物进行认知，其认知结论会直接影响人类的决策和行动，如流行的购物网站的智能推荐、汽车自动驾驶的智能判断、股票交易员数据处理分析的智能助手等。

我们很容易发现，对于同一个事物，给到不同的算法和数据会得出不同的“机器认知”结果，进而采取不同的行动。这和“人类认知世界”是一样的，对相同的事物，每个人的认知可能是不同的。但是，人类的认知带有明显的个体或群体的责任标识，谁的理论、谁的标准、谁的观点，是非常清晰的。人类在几千年的实践中，已经总结出了很多对物理世界的认知方法和体系，形成了一系列处理不同认知的冲突的体系。但在“机器认知世界”则不同，以机器学习和算法为核心的人工智能，强烈依赖于用于训练的数据集，很多机器学习算法本身是一个黑盒，依赖的数据集也大概率是偏倚的，得出的模型和结论缺乏明显的责任标识，社会的大量数据掌握在政府和少数寡头手中，个人或群体无法参与监督和辩论。个人无法知道你所得到的各种推荐是基于什么认知，无法知道你在各个数据库里被打上了多少个不同的标签，也不知道这些标签会对你的生活、就业、升迁等产生什么影响。

对于企业来说，未来的常态是基于各种人工智能的算法，做出一系列的决策与行动。那么谁对决策的结果负责？是算法？还是数据集？如何规避风险，提升决策质量？企业的运作成功与否将在很大程度上取决于其对“数字世界”和“机器认知世界”的治理和管理水平，用通俗的说法，也就是需要建立对智能世界的治理体系，在这方面我们会面临全新的问题和挑战。

10.5　本章小结

数据成为企业的生产要素，将带来数据确权体系和数据市场基础设施建设的浪潮。大规模数据交互将构成庞大的企业数据生态，数据管理手段也将全面智能化。“物理世界”“人类认知世界”“数字世界”和“机器认知世界” 将构成全新的“智能世界”，数据将成为四个世界相互联接转换的枢纽，成为智能世界的支柱之一。数据治理将面临一系列全新的问题与挑战。

未来已来，让我们共同努力，把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界。