阿里云全球技术服务部
ALIBABA CLOUD GLOBAL TECHNOLOGY SERVICES
阿里云数据库产品事业部
ALIBABA CLOUD DATABASE DEPARTMENT

升舱 – 数据仓库升级

交付标准化白皮书

Technology Service Standard for Upgrade - Data Warehouse upgrade

主编单位

阿里云智能全球技术服务部

阿里云智能全球技术服务部 (Global Technology Services，简称 GTS) 是一支面向云智能客户，提供完整生命周期服务的技术履约团队。旨在通过专业化的交付实施、系统化的中台支撑及标准化的服务产品能力，运用云和数据智能的力量，联合生态伙伴一起帮助客户实现业务价值。

阿里云智能数据库产品事业部

作为中国数据库市场的领导者，阿里云数据库产品事业部可以提供一站式全链路数据管理与服务，拥有云原生数据库 PolarDB、云原生数据仓库 AnalyticDB、云原生多模数据库 Lindorm 等丰富的数据库产品和工具组合，通过云上和云下灵活的部署方式，为全球客户提供多样性的选择。

目前，阿里云数据库市场份额稳居全球前三，亚太第一，在国内遥遥领先，并在 2020 年进入 Gartner 全球数据库领导者象限，成为在基础软件领域首次进入领导者象限的中国企业，代表中国数据库真正走进世界一流。

阿里云智能新金融事业部

阿里云新金融事业部致力于为金融行业客户量身定制云服务，运用领先技术全面助推金融行业数字化转型进程，并承载着拉通和联动阿里巴巴所有面向金融行业服务力量的使命，深耕「云 +Fintech」的技术和业务能力整合。

截至目前，借助阿里云的技术方案，已服务 6 家国有大型商业银行、12 家股份制商业银行和 90% 的已上市商业银行，同时覆盖 60% 以上保险企业、证券公司，及三方和四方支付客户。

编委会

总　编　辑：李津、刘湘雯、李飞飞、刘伟光

顾　　问：张瑞、王馨、黄欢欢、王德会、罗龙九、张翅、王伟民、林亮、魏闯先、李圣陶、冯程、薛菲、王德山

主　　编：戈林锋、王真、肖齐治、李靓雯、罗成对

编　　辑：王腾、汪国辉、王宗瑞、卞中明、金川、赵可可、余从佳、杜雄、周晓林、谷磊

视 觉 设 计：马龙泽、王甸甸

版权声明

建议和意见反馈，请通过邮箱：linfeng.glf@alibaba-inc.com 与我们联系。

前言

在全球疫情的深刻影响下，世界经济一片萧条，数字经济展现出独特的魅力，数字化转型也成为各行各业的共识。疫情在某种程度上加速了全球产业数字化转型的步伐，数据成为了主要生产要素，也是数字时代的核心基础资源。基于数据资源，传统产业的生产方式、商业模式、管理模式以及创新模式都在重新定义。

习近平总书记指出：“数字技术正以新理念、新业态、新模式全面融入人类经济、政治、文化、社会、生态文明建设各领域和全过程，给人类生产生活带来广泛而深刻的影响。”2021 年 12 月 27 日，中央网络安全和信息化委员会印发《“十四五”国家信息化规划》，该规划中明确提出了要建立健全规范有序的数字化发展治理体系。这也将推动营造开放、健康、安全的数字生态，加快数字中国的建设进程。“十四五”时期，我国数字经济将继续深化，根据国务院印发的《“十四五”数字经济发展规划》，到 2025 年，数字经济迈向全面扩展期，数字经济核心产业增加值占 GDP 比重达到 10%；其中金融、电信、电力等行业作为数字经济基础行业，随着数字化转型不断升级，自主可控、安全可控被重点提及，要求应用于关键核心业务，赋能产业转型升级，实现应用推广与业务创新。2022 年 5 月，全国政协在京召开专题协商会，围绕“推动数字经济持续健康发展”进行协商议政，国家发改委下一步的工作将围绕规划落实，并研究起草推动数字化转型的指导意见，并组织专项工程和布局数字化转型促进中心，从而提升行业整体的数字化水平和提供相应的数字化转型公共服务。数字技术迭代创新速度加快，在追求较高的社会生产力和资源优化配置的同时，数字化的发展、转型、治理可以有效的营造良好的数字生态。

根据国际数据公司 (IDC) 预测 , 到 2023 年 , 中国的数据量将达到 40zb, 其中超过 80% 是非结构化数据。随着 5G 技术的普及 , 企业数据将成为中国数据的主流，数据市场将成倍增长。而数据仓库作为数据密集、计算密集的数据集中处理平台，是数据承载、数字化转型不可或缺的一环。长期以来，国外数据仓库平台厂商包括 Teradata、Exadata、Netezza 等在过去几十年一直是金融、运营商等重点行业的重要选择，直到今天这些国外的基础技术平台在很多单项技术能力方面仍然具有非常强的竞争力。但随着各行各业数字化转型的深入推进以及 IT 系统自主可控的政策性鼓励，传统数据仓库平台迎来了转型升级的重要窗口期。

出于服务国家“十四五”数字化转型和基础软件国产化战略，阿里云重磅推出“升舱计划”，将阿里标准化产品、多年实践、成熟方法论结合，旨在解决传统数据仓库升级转型面临的诸多新变化，为行业客户提供全新的数仓应用体验。

《 数据仓库升级交付标准白皮书 》白皮书，既是阿里云结合金融、运营商等关系国计民生重点行业大量项目实践经验的分析总结，也是深入研究传统数据仓库如某传统数据仓库厂商等交付流程和价值创造方式的阶段性成果，更是针对传统数据库仓库升级交付的方法、工具、平台等体系化建设的实践指引。

CONTENT

数据仓库交付标准化白皮书

缩略语

- **GTS**，Global Technology Services，全球技术服务部
- **CBM**，Cloud Business Manager，云商务经理
- **SA**， Solution Architect，解决方案架构师
- **CSM**，Customer Success Manager，客户成功经理
- **BA**，Business Architect，业务架构师
- **TM**，Technical Manager，技术经理
- **PMO**，Project Management Office，特指项目可交付性管理
- **PD**，Product Designer，产品经理
- **QA**，Quality Assurance，质量保障
- **TAM**，Technical Account Manager，客户技术经理
- **PDSA**，Product Designer & Solution Architect，产品架构师
- **ISV**，Independent Software Vendors，独立软件开发商

行业趋势与困惑

数据库已经在信息技术领域有了广泛的应用，我们日常的社会生活也宛如一个巨大的数据存储仓库，每个个体产生的活动都是各种各样的数据。作为数据库的一个分支，数据仓库概念的提出晚于数据库概念。

关于何为数据仓库，数据仓库之父比尔 · 恩门（Bill Inmon）在 1991 年出版的“Building the Data Warehouse”（《建立数据仓库》）一书中所提出的定义被广泛接受：数据仓库是一个面向主题的 (Subject Oriented)、集成的 (Integrated)、相对稳定的 (Non-Volatile)、反映历史变化 (Time Variant) 的数据集合，用于支持管理决策（Decision Makeing Support）；我们从该经典定义中也可以看出数据仓库是对有数据模型的、对海量结构化数据进行抽取、转换和加载（ETL）的大规模数据处理平台。

对比数据库的应用设计理念和应用目的，数据库可以被理解为是面向事务的设计，而数据仓库是面向主题的设计；数据库存储在线交易数据，数据仓库侧重存储的多为历史数据；针对数据库的设计秉持避免冗余的原则，而数据仓库的设计则是有意引入冗余；数据库是为捕获数据而设计，而数据仓库是为分析数据而设计。

长期以来，国外厂商的数据仓库平台和最佳实践在过去几十年很好地支撑了金融、运营商等重点行业的业务发展。随着各行各业数字化转型的深入推进以及监管趋势的变革，传统数据仓库平台迎来了转型升级的重大战略阶段。

从传统的 BI（商业智能）到大数据分析，再到当今的 AI（人工智能），传统数据仓库平台一直在演进，并且在 Gartner 年度报告中，某头部传统品牌 T 作为专注于数据分

析的厂商一直出现在领导者象限，并在单一领域保持十余年的领先地位。因此，我们以某传统数据仓库厂商翘楚 T 公司在国内某金融行业用户的发展为例，该数据仓库厂家的传统数仓在过去数十年为该用户的数据价值挖掘方面发挥了重要贡献，主要基于其以下优势：

1）大数据量处理性能较强

具体体现在可线性拓展并有效的实现投资保护、数据库管理和维护相对容易、第三方软件产品丰富；

2）拥有完整的 ETL 方案且配置灵活

3）采集异构数据是，数据以 txt 文件格式，支持一段天数的保留为后续查缺补漏提供时间基础

4）Perl 采集数据时，进行输入插入时速度快

但是随着业务高速发展积累而来的业务数据已经达到了某传统数据仓库厂商该数仓资源容量上限、数据仓库平台多年的高水位运行带来硬件老化的高风险硬件故障，同时由于该传统数据仓库平台本身商业策略调整导致难以继续升级的风险等多方面因素交织，使得传统数据仓库国产化转型升级迫在眉睫。传统数据仓库平台在升级中的不足具体可归纳为下图 1 中描述的五个维度：

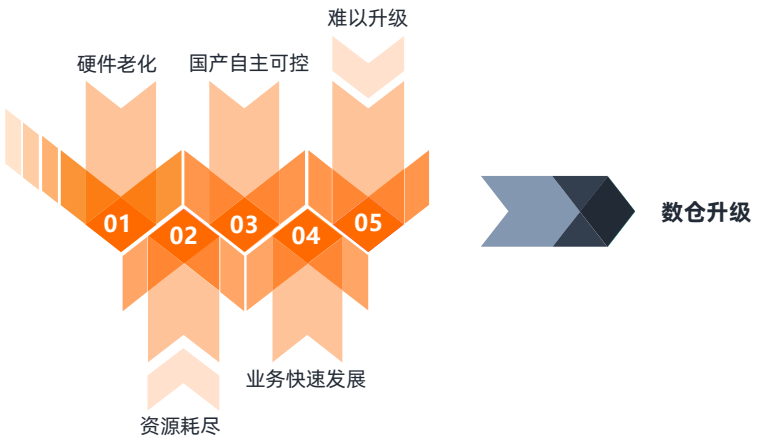


图 1 传统数据仓库升级局限性分析

具体的不足详情体现有以下方面：

1）环境搭建复杂需要专用设备，扩展时，停机时间较长

2）混合负载控制存在问题

3）各系列产品互联困难

4）价格较高导致性价比不佳

5）ETL 过程和 TD 的工具绑定紧密，对数据库的性能要求和依赖度强

6）ETL 环境搭建复杂、非图形化

7）自定义函数过程复杂，需用 C 语言编写并导入

8）数据库在线性扩展时，数据需要重分布，消耗资源量较大，耗时长。

根据我们在部分重点行业用户对于传统数据仓库转型升级方面疑虑和困惑的了解，我们统计归类了如下四类问题：

（一）技术类问题

1. 数仓升级转型的技术难点或者挑战主要有哪些?

2. 现有数据仓库关联上下游系统众多，如何在升级过程中快速集成由不同技术体系架构的应用系统？

（二）实施类问题

1. 数仓升级转型采用怎样的实施路径比较好？

2. 现有数仓上运行的数量众多的作业和模型，是否能尽可能减少改动平迁到新数仓上正常运行?

（三）管理类问题

数仓升级转型需要在项目管理、组织层面如何实现配套?

（四）发展类问题

1. 目前同业在数仓升级转型实践上有哪些成功经验可借鉴?

2. 应用开发商懂业务应用但是不懂升级后新数仓技术，如何推进?

基于阿里云多年实践，本白皮书将从工程角度对以上数仓问题进行深入探索和讨论，并予以解答和参考。

数仓升级成功的关键点

通过多个数仓升级项目实践总结及复盘，要一次性把事情做对，完成一次成功的传统数仓平台升级实施工作难度不亚于飞机在空中更换发动机。总体来看，造成数仓升级实施难度系数较大的原因主要体现在三个层面：技术方案、迁移方法、实施资源。

在技术方案上，需要对每个技术关键点都能考虑周全，深入探索每个技术细节并进行充分的论证和测试；在迁移方法上，需要科学完善的实施方法论，充分考虑迁移项目的工程特点和平滑过渡目标，把迁移风险做到可识别、可分析、可预测、可防范；在实施资源上，不仅需要团队对于新技术具有前瞻性认知和把控能力，更需要对原有数仓体系的盘根错节有深入了解，能够在风险发生时从技术、方案、业务等不同层面提出应对方案，及时化解风险。

管理学大师德鲁克认为，并不是有了工作才有目标，而是相反，有了目标才能确定每个人的工作。为了成功实施数据仓库升级的既定目标，提升升级交付效率与质量，需要秉持“三个目标、四个原则，五个关键动作”标准，下文将基于此标准并结合实践案例详细阐述，图谱如下图所示：

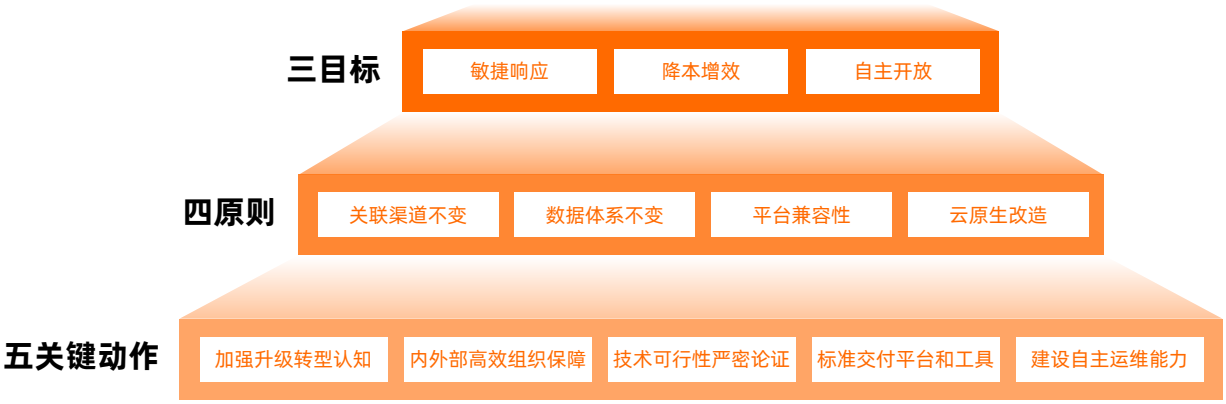


图 2 数据仓库升级实施标准

2.1 数仓升级的三个目标

从我们的调研和实践来看，数据仓库升级不仅仅是一个技术问题，也是一个工程问题，对企业和实施厂商的协调组织、项目管理、技术掌控、交付能力等提出了全方位考验。基于实践，我们总结出“数仓升级稳定模型”（图 3），实现数仓升级需要从三方面进行要素搭建：业务敏捷响应、降本提效、自主开放。

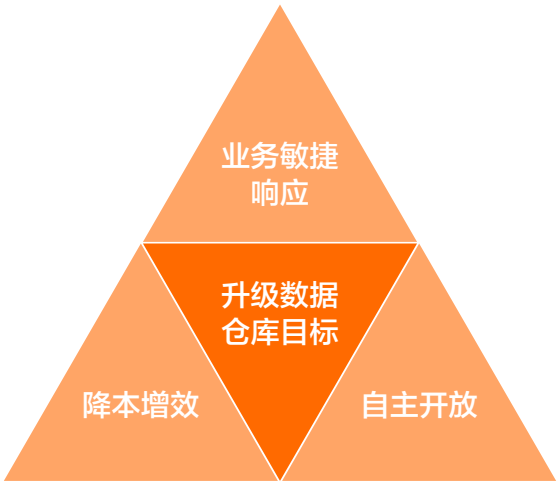


图 3 数仓升级稳定模型

首先，数据仓库具备业务敏捷响应支撑能力，该能力可以有效应对当今时代业务发展不确定性和互联网固有特征。相对于传统数据仓库平台升级，则需要考虑具备兼容传统数仓平台的所有业务处理能力，而兼顾所有业务的同时也带来了问题百出和处理问题低效的现象。所以，业务敏捷响应可以有效的节约时间实现数据的迁移、抓取等操作，数仓升级让业务处理更高效。

其次，通过数仓升级可以有效的降低数据仓库的整体拥有成本。长期以来由于国外数据仓库厂商在数仓建设上存在某种程度的不可替代性，数据仓库市场一直是个卖方市场。据全球权威 IT 研究与顾问咨询公司 Gartner 公司评测，在数据库领导者象限的某数据仓库厂商，对应项目基本都是千万级起步，这还不涵盖每年固定支出的高昂服务支持费用在内；因此，我们希望通过使用成熟的产品、工具和方法论指导降低数据仓库升级和后期维护整体拥有成本（TCO）。

此外，数仓升级可以打破国外厂商的数据仓库平台的技术封闭性问题，实现技术的自主可控。传统国外厂商的数据仓库平台均为软硬一体机的输出模式，且一体机中的网络、存储等对于用户来说是个黑盒；而且随着大数据技术的发展，无法有效与通用大数据技术体系进行有效融合。因此升级传统数据仓库平台需要保持开放性、以自主可控为重要目标进行升级，赋能企业的数字化转型。

2.2 数仓升级的四个原则

要达成传统数据仓库升级成功的三个目标，我们需要的不仅仅是一套技术交付方案，更需要一套能够指引行动的“原则”。正如雷 – 达里奥在《原则》一书中提到：原则犹如指引行动的“灯塔”，它连接着我们的目标与行动。解决不确定性靠敏捷、解决复杂性靠原则，越是复杂的系统越需要一套原则来保证。

过去以国外厂商为主导的数据仓库平台很好地解决了企业从零到一的取数、用数问题。经过多年的长期迭代建设和大量投入，在现有数据仓库平台体系内积累了许多符合企业

自身业务需求的经验和实践，比如在数据架构上采用分层设计思想初步实现了将数据分类处理、建设了大量主题数据模型、建设了统一的数据交换平台，这些都是企业多年经营的心血和结晶，是最重要的数据资产之一，数据仓库平台的升级要尽可能考虑复用企业现有投入，把企业花了大力气积累的数据资产平滑过渡到新的数据仓库平台，避免推倒重来的重复建设开销。

如下图所示，我们将高效实现数仓升级所需要的原则总结为“四边形”原则：

图 4 数仓升级的四边形原则

该模型可细化解读为如下四个方面：

1）原关联上下游渠道不受影响：升级数据仓库涉及到对现网上下游多种类型系统的数据集成工作，要确保升级数仓前后业务渠道和系统功能不变、数据交换平台功能不变、批量调度平台功能不变。

2）原数仓平台数据体系不变：原数仓平台通过分层设计的数据治理模型不变，比如ODS、数据缓冲层、基础数据层、公共汇总层、数据集市等能在新数仓平台复用。

3）原数仓平台兼容性：具有对原有数据仓库平台一定功能上的的兼容能力，确保对原数据仓库平台有一定的使用习惯和功能上的兼容能力。

4）采用云原生技术进行优化改造：考虑技术前瞻性、时代必然性，采用云原生组件对传统数仓平台进行优化改造，并且改造后的业务功能不受影响。

基于以上四个原则，采用云原生组件进行传统数据仓库优化升级可以让企业在云计算时代充分享受云计算的红利。云计算的本质是随着虚拟化技术的突破而带来了用虚拟化技术高效使用来池化底层的硬件资源，最重要的一点是利用好云原生数据仓库的存储和计算资源服务化特性，让用户根据自身的需求灵活选择资源规格、并根据数据量增长需求随时变更资源配置，还能做到对上层应用无感知，达到投入资源收益最大化；此外，分布式并行计算、透明可扩展、高可用、同城容灾等都是传统企业长期以来功能刚需，是云上数据仓库在云上与生俱来的能力，可以让企业实现开箱即用的良好体验；因此，企业在升级传统数仓时选型一套经过大规模验证的、用于支撑新数据仓库平台生产运行的 IaaS 和 PaaS 平台，可以让升级工作事半功倍。

2.3 升级实施的五大关键点

数据仓库平台日常数据处理任务多、对接业务系统广，系统升级切换工作牵一发而动全身，从项目组织管理角度看升级数据仓库是一个大型数据库工程系统的再造。为保证工程质量、工期和风险可控，我们建议企业和实施厂商通过建立相应的实施组织架构，梳理完善的实施工作流程，通过系统工具和完备的方法论指导支撑升级工作的有序进行。实施的五大关键点如下图所示：

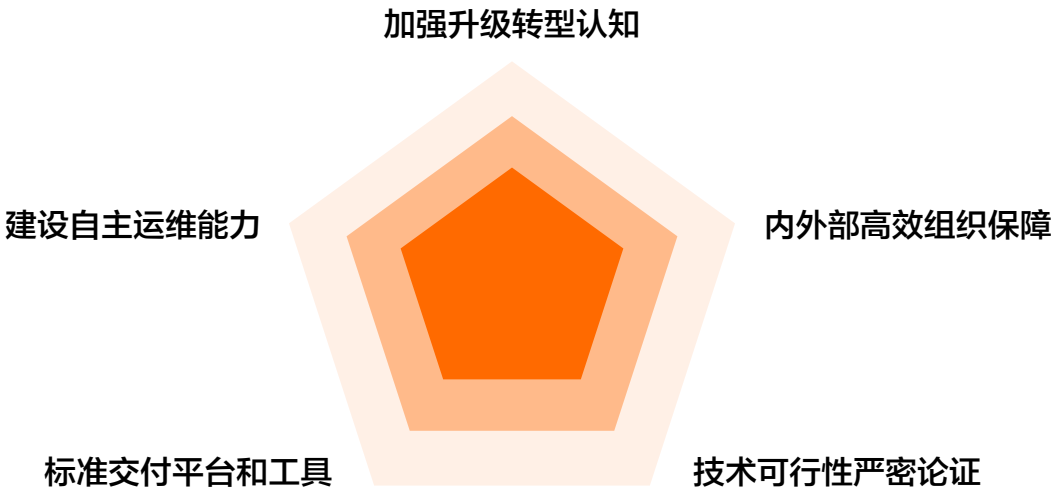


图 5 数据仓库升级实施五大关键点

（1）升级转型认知深入人心

全球公认的领导与变革领域的权威约翰·科特 在《变革之心》一书指出：“想要改变人们的行为，仅仅通过帮助人们进行分析来影响他们的思考是不够的。更重要的是，让他们看到事实，从而影响他们的感受！思考和感受都是需要的，而且在一个成功的组织中两者都要具备。但是让人们改变的核心在于触动感受，所以‘看到－感受－改变’的方式比‘分析－思考－改变’的方式更有效。”在我们过往成功落地的数仓升级案例中，我们深刻意识到在实施升级传统数据仓库平台的过程中同样也有必要引入有效运用变革管理的方法，通过一系列举措降低数仓升级过程中的阻力，保障升级规划的落实。

1）采取各种措施，确保数仓升级规划完成后， 业务和技术部门能够及时深入地学习和了解数仓升级规划内容，并根据规划内容进行下一步工作安排；尽可能将升级转型项目与企业年终绩效机制挂钩。

2）贯彻总体升级规划，扩大升级项目的影响力，降低项目风险，获取项目相关方各部门的广泛理解，确保项目顺利进行。

3）业务部门要广泛深入的参与到升级实施中，增进对项目过程和成果的认同度和支持度。

4）实现知识共享，可以提升应用人员和维护管理人员的新数据仓库平台认识，同时提升对新数仓技术体系的掌控能力。

因此，有必要在实施数据仓库升级全周期过程中，强化企业内各部门宣传实践工作来转变观念、提高认识，推动升级实施的顺利进行。

（2）内外部高效组织保障

升级传统数据仓库平台可以使企业的业务能力、组织能力和技术能力都得到明显提升，但这需要企业能够为升级规划的实施提供充分的资源保障，建议从三个方面进行重点保障：

人员：对内要确保企业科技成员对新数仓技术体系的培养，能够在升级实施过程中全程参与；对外要整合各层次的外部资源与企业自有技术人员形成互补，形成“以外带内”的合作局面，充分把新仓技术体系在企业内部推广落地。

建议结合企业的现状，引入有满足资质的厂商及外部专家对数仓升级转型综合管理和协调，成立专项组；以下组织结构图 6 可作为参考：

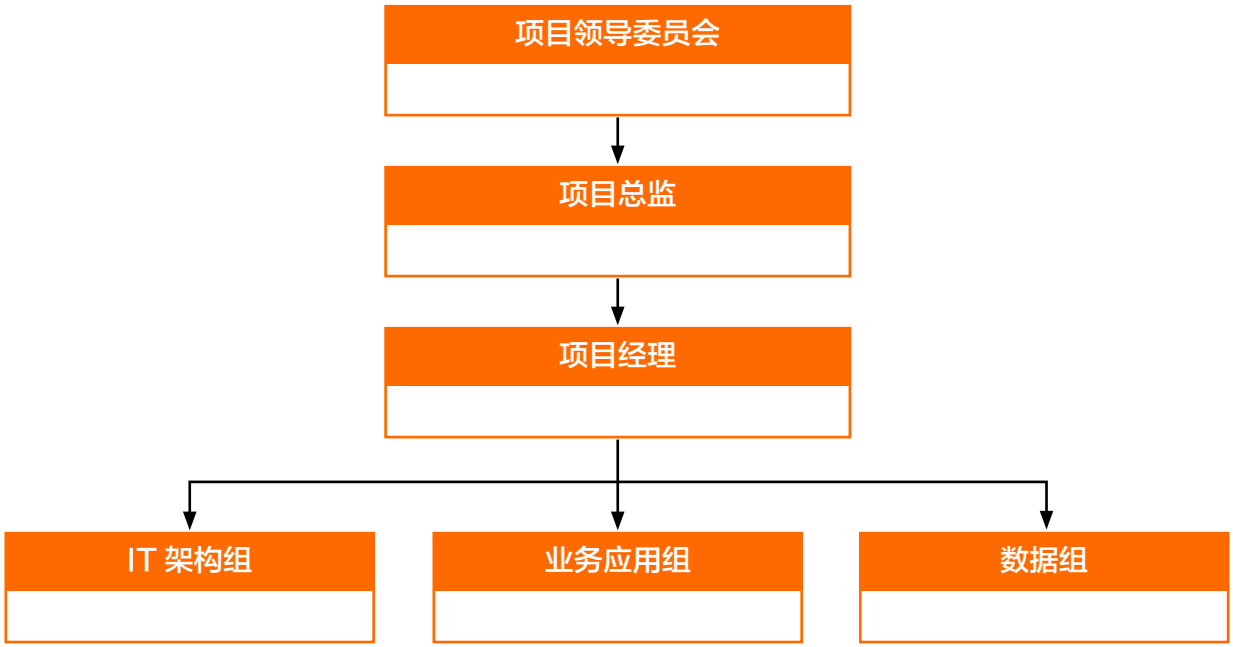


图 6 数据仓库升级组织架构图

对应的组织人员职责说明如下表：

组织单元	职责内容
项目领导委员会	1. 监督、指导项目的方向 2. 确保项目资源 3. 参加项目进展报告会 4. 审核批准项目最终成果
项目总监	1. 把握总体项目策略方向 2. 监督项目整个进程 3. 项目变更进行审核、决策

组织单元	职责内容
项目经理	1. 制定项目计划并监督执行 2. 项目总体的协调、沟通、监控、管理 3. 汇报工作内容与进度
IT 架构组 业务应用组 数据库组	1. 评估现有甲方系统基础架构与现有系统业务、系统概况 2. 制定迁云策略与云资源需求 3. 制定迁移方案和迁移割接方案 4. 提供上云工具与使用技术以协助迁移方案验证、数据迁移与应用部署 5. 协助系统割接与上云后的系统优化 6. 云平台运行问题相应和排查

表 1 组织人员职责说明表

1）**效率：**考虑到数仓升级建设的专业性和时效性，建议在实施过程中在企业的科技管理框架下尽可能“特事特办”，增加灵活性，提升遇到内外遇到卡点问题时的协作效率。

2）**资金：**除了对软硬件的投入以外，最重要的是增大对技术专家、服务顾问等引入外脑方面的投入，外部行业专家的意见和建议可以开拓对数仓升级建设的认识思维边界，同时可以紧贴市场，实现成果对市场第一时间的响应；此外，考虑团队成员工作任务分配的不同，以及偶有一人多职、多项目兼顾的情况，投入专业的技术人员和顾问不仅可以增强项目实施和决策的专业性，同时推进准时结项。

（3）技术可行性严密论证

清晰且细致的实施路径和技术细节实地验证是决定技术可行性的关键因素，也决定了对数仓升级全流程的掌控能力，成功的项目经验和平台工具能力是最重要的体现。无数个细节和核心技术点体现在从调研选型到业务割接上线全流程中，只有成功的项目足够多，可复制的经验才能体现在实施路径的方法论中。而平台和工具能力又决定了数仓升级流程的效率和质量，体现的是标准化的能力。技术细节的实地验证决定实际升级过程是否能够按照预期推进，每个技术细节的理论和实际偏差都需要验证，偏差过大可能导致整体方案不可行，比如源端和目的端数仓的数据模型，任务兼容度如果

偏差 20% 可能只涉及产品或者应用的少量改造，但是如果偏差 50%，则整体的选型设计方案需推倒重来。与此同时，准确识别核心技术点及其验证方法、前后依赖条件是技术可行性论证的核心，比如兼容度评估、数据及任务迁移、数据校验、业务割接等。

（4）标准的交付平台和工具

完备的实施工具集

完备的升级实施工具集是效率大幅提升的重要保障。正如“工欲善其事，必先利其器”，工匠在做工前打磨好工具，操作起来就能得心应手，达到事半功倍的效果。实施工具本身融合了数仓升级设计规范和升级交付产出模板，其产出内容严格按照设计要求，避免人工操作的差异性和不规范性，是交付质量的重要保障手段。同时，工具使用简单、便捷，不需要强依赖经验丰富的专业人员，有效降低了数据仓库升级的实施门槛。

常见的交付实施工具有数据仓库调研评估工具、海量数据迁移同步工具、数据模型迁移工具、数仓作业任务迁移工具、数据校验稽核工具、项目实施流程管控工具等等。

持续的知识库建设

作为工程的一个重要组成部分，数据仓库过程升级中存在项目管理、方案、技术问题、最佳实践、赋能培训等方面的材料沉淀，有必要将实施过程积累下来的技术、经验记录等记录在标准文件中。以免因技术人员的流动而使方案或技术流失，从而方便操作人员经过短期培训，快速掌握云上数据仓库的操作技术；同时，在升级实施的各个阶段，需要有可视化工作台将完成的阶段性任务进行统一组织管理。

（5）自主运维能力建设

数仓升级是客户、ISV、数仓产品及服务提供商通力合作的过程，数仓建设完成后客户的自主运维能力是客户关注的核心问题之一。首先自主运维能力不是一蹴而就的，建议设置短中长期规划并配备相应的组织人员支撑规划。短期以产品及服务提供商为主

客户为辅，中期以客户为主产品及服务提供商为辅，长期目标是客户自主运维、产品及服务提供商承担应急处理的方式；其次，自主运维能力需要理论加实践的结合，如果只是进行一些理论或者技术宣讲，最后大概率只会沦为一些归档文件，自主运维能力必须建立在运维实践的基础上，反复沉淀优化运维和应急资产，迭代组织结构和效率；最后，从数仓升级建设开始，客户人员端到端深度参与也非常重要，这不仅是了解从选型、设计、迁移、优化至运维的契机，也是一个业务和技术双向赋能的过程，每个技术细节的可选方案、权衡取舍都是个人运维能力和团队自主运维能力提升的一小步。

实施标准与实施路径

在具体如何落地数据仓库平台升级的问题上，我们本着求真务实的态度，坚持从实践中来到实践中去。毛泽东曾说“只有人们的社会实践，才是人们对于外界认识的真理性的标准。真理的标准只能是社会的实践。”通过多个数仓升级项目实践总结及复盘，需要通过数据仓库平台升级动作的标准化，把交付过程中的各个要素和环节有机组织起来，使数据仓库平台的升级工程规范化、科学化、程序化，提升数据仓库平台升级交付效率与质量，达成前述升级数仓的三个既定目标。基于项目实践，如图 7 所示，将数仓升级分为三个主流阶段：升级挑战阶段，明确困难点，后期一一进行突破；标准升级流程阶段，遵守四个标准，从而促进标准化的成熟和项目落地；升级标准化成熟度阶段，该阶段同时也反哺第二阶段。

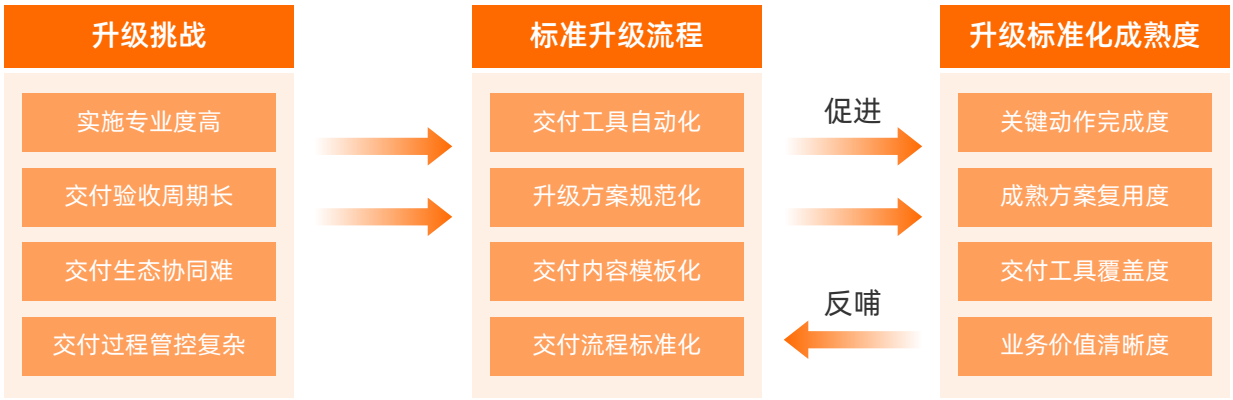


图 7 数据仓库升级的三个主流阶段

整体来看，数据仓库平台升级过程的主要挑战可以归纳有如下方面：

实施专业度高

数据仓库平台升级工作一般都要涉及咨询、数据、技术、运营等，交付技术覆盖范围广、资源需求大，对于交付供应商能力需多方面的考验；而交付供应商常常因行业经验、专业人员、方案成熟度等方面的不足，导致升级实施存在诸多质量问题和不确定性。

交付验收周期长

数据仓库平台升级都会涉及数据迁移、模型迁移、调度任务迁移以及上下游关联系统协同等，其交付周期一般以数月为最小单位，期间人员流动风险、需求变动风险等都对项目工期带来挑战，导致交付的业务价值及质量等级很难做到在线化、可视化评估。

交付生态协同难

数据仓库平台升级工程庞大、环节众多，没有一家公司能够全方面覆盖，往往是多家实施供应商集成工作模式整合彼此优势资源，但是实施供应商之间在能力匹配上有很大不确定性。造成此种不确定性的原因往往集中在实施伙伴间能力成长差异性、伙伴内部对员工的不同组织架构带来的不稳定性以及员工本人对职业路径规划所产生的波动性、伙伴对行业领域知识的缺乏等，这些知识和能力上的参差不齐使得数据仓库升级实施生态协同难。

交付过程管控复杂

在实施交付过程中，从需求调研、方案设计、迁移实施到新 / 旧双系统并行试跑，基本都是在线下由不同角色、甚至不同实施供应商完成的，缺乏项目交付全流程、全生命周期的数字化工作台承载，很难实现对项目全局掌控，各个环节都容易出现不同类型的问题与挑战。

数据仓库平台升级实施面临的这些常见挑战，是高标准、高质量、高效率的去实施数仓升级工程的拦路虎。结合软件工程理论和数据仓库平台升级实施经验，构建标准升级流程和质量评估标准可以有效降低这些挑战带来的风险，即通过交付标准方案、交付标准流程、交付标准工具及数字化工作台，提升交付效率与实施质量，降低项目实施门槛与过程管控复杂度，为“多、快、好、省”数仓升级交付奠定基础。

3.1 实施标准路径

数仓的升级实施流程按照“统一规划，阶段实施”的原则进行建设。从项目生命周期角度，我们将数仓升级实施工作分为两个阶段：轻咨询期和升级建设期。每个不同时期都有服务交付进行兜底管控，同时每一个阶段的产出又是下一个阶段的输入，从而形成一个体系化的升级数据仓库平台的顶层工作任务和路径规划。详情见下图 8:

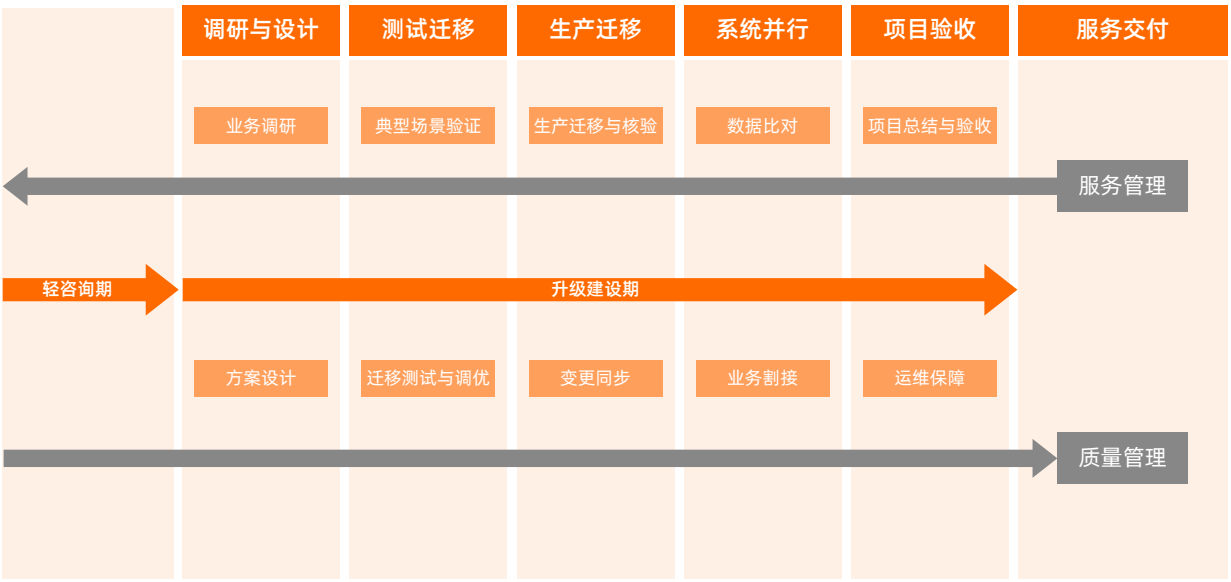


图 8 数据仓库升级路径规划

为了保证升级工程质量、工期和风险可控，最大限度规避和降低升级过程的诸多挑战，升级过程要求以“交付工具自动化、升级方案规范化、交付内容模板化、交付流程标准化”为宗旨进行实施。

（1）轻咨询期

轻咨询期是数仓升级实现规范化的重要阶段，为确保后续升级工作更具有可落地性、可实施性，该阶段主要工作内容如下：

1）**专家咨询与方案输出：**需要企业借助一批有传统数据仓库平台升级落地经验的外部

专家，结合企业自身业务特点规划升级数据仓库顶层设计，并进行严密的技术可行性论证，产出提供数据仓库迁移工作所需的现状调研与需求分析、方案设计、方案可行性验证、交付实施规划、升级所需软硬件资源规划和容量设计、高可用与容灾设计等；

2）**项目组织实施机制确定：**结合同业数仓升级实施经验，组织构建升级交付阵型，明确组织分工和协调机制；

3）**前期人员能力培训：**进行升级实施前的产品、方案的赋能培训工作，确保企业业务和技术部门以及实施供应商能够及时深入地学习和了解数仓升级规划相关产品、技术和工具；

4）**环境准备：**数据仓库平台升级所需环境资源，比如容量规划等；

5）**风险预判与风险对策：**升级实施过程潜在风险点的梳理及应对策略。

通常经过轻咨询阶段后，可以输出数据仓库升级实施整体规划，后续的升级建设基于该规划整体推进将实现从宏观到微观的全面落地，下图为数据仓库升级实施规划流程示意图。

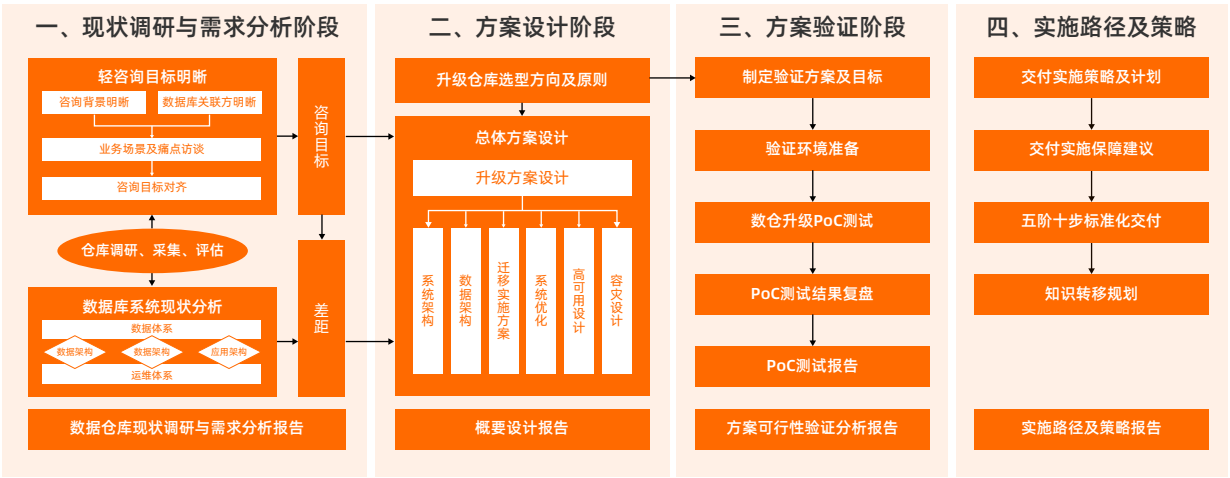


图 9 数据仓库升级实施规划流程图

（2）升级建设期

升级实施建设期，在“不改架构，不动流程，尽力兼容，云化升级”的总体原则下，围绕着“不停机迁移与双核心数仓平台并行”目标，我们进一步将升级实施拆解分阶段总结归纳为“五阶十步”，如下图示：

图 10 “五阶十步”实施图

在具体实施建设期间，高标准、高质量的实施借助标准交付平台和工具的有效利用，通过将交付流程流水线化、工具自动化降低实施厂商人员素质差异带来的交付质量降低风险。实施期间“五阶十步”主要工作内容具体如下，对应产品、工具和平台在后文案例进行介绍。

阶段一：调研与设计

该阶段需要实施厂商对原系统上下游做详实调研，需现有数据仓库平台和业务系统进行充分的信息收集，最后迭代输出调研分析报告，并与企业业务方做深入讨论与修正。调研内容覆盖如下方面：

（1）原数仓系统架构

（2）原数仓数据交互流程

（3）原系统资源盘点

（4）原数仓库表统计

由于数据仓库数据体量一般较大、数据特征复杂，在调研初期需要合理运用技术手段，采集分析现有数据仓库平台数据对象特征、采集分析业务 SQL、关键特性等，并设计合理而科学的迁移实施方案。

阶段二：测试迁移

该阶段主要围绕典型场景验证展开，对现有数据仓库平台涉及的 ETL 作业任务、表 / 视图结构、模型、数据、用户权限等进行迁移的论证工作，通过典型场景验证期望暴露更多技术性问题（如 SQL 兼容性、SQL 复杂度、作业复杂度等），用于正式实施阶段更为准确评估迁移工作量；测试迁移阶段涵盖的主要验证范围如下：

（1）数据源迁移

（2）数仓模型迁移

（3）作业调度迁移

（4）数据迁移

（5）新老数仓系统数据比对

（6）下游系统对接

阶段三：生产迁移

通过测试迁移阶段对潜在风险点、卡点性问题逐一排除后，生产迁移过程压力就会随之降低很多，该阶段依托构建的知识库记录详细的操作流程进行评审和实施，确保正

式迁移中操作质量和效率；生产迁移阶段主要包括数据迁移、模型迁移和调度迁移三大块。

（1）**数据迁移阶段**，涉及表众多且数据量大，迁移步骤涉及建表、数据导出、数据文件传输、数据导入等多个步骤，此过程会占用大量的时间和人员，而且每次操作失误可能都需要重来。因此有必要通过标准数据迁移工具实施迁移，工具是实现上要求基于配置的元数据信息，自动生成建表脚本、数据导出脚本、入库脚本，从而大大减轻人力投入，缩短时间成本。本阶段工作的人力投入主要侧重于脚本的执行调度、执行结果的监控和问题的排查处理。

（2）**模型迁移阶段**，代码量大，全部由人工改写工作量非常大，且不能保证改写的逻辑一致性；针对传统数仓的模型迁移，需要通过模型迁移工具辅助完成绝大部分程序的改写工作，对于特定语法无法改写的还需要输出专门的错误报告供开发人员针对性重构。

（3）**调度迁移阶段**，数据仓库平台上的一个重要工作是每天定时执行调度任务，各个任务之间还相互依赖；如果基于新数仓平台的调度工具无法和原有调度工具有效兼容，将会涉及到原有调度任务元数据重新配置问题，将会是项目如期完成的较大变数，这也是前期轻咨询阶段就需要充分考虑验证的问题，再次说明前期轻咨询阶段的重要性。

阶段四：系统并行

一个企业级的数据仓库平台往往需要对接数十上百个上下游系统。出于业务连续性的考虑，升级过程中想要通过“一刀切”的方式将原有上下游系统切换到新的数据仓库平台上几乎是不可能做到的，因此数据仓库平台升级建设进入到投产期将会存在双核心数仓平台并行。为了降低双系统并行带来的额外人力成本，需要关注以下关键点：

（1）变更同步

在新老系统并行阶段业务触发的变更需要新、老系统同步执行，或者有时间计划的异步执行。在有条件的情况下，尽可能的减少业务变更，避免由于作业变更导致的数据

不一致，进而导致数据稽核（数据完整性和一致性检查）失败而产生大量的排查工作量，延误项目工期。

（2）数据采集

并行阶段数据采集差异易导致后续数据稽核失败，如两套系统配置各自的数据采集服务，但并行阶段为避免对上游系统产生双倍的压力，将会采用错峰采集的方式，若采集时间发生时间差，都有可能导致采集数据的差异，进而导致数据稽核失败，此类问题难以排查，因此尽可能保证数据采集的一致性十分重要。

阶段五：项目验收

不同于业务系统数据库升级，数据仓库平台升级在项目投产末期无法以数据完全一致作为验收标准的。由于双核心系统并行阶段两边数据库特性或者数据采集等方面的差异，数据不一致的情况无法完全避免，因此需要尽早的与业务方明确切实可行的验收标准，并积极推动验证。

此外、在项目验收阶段还需要逐步建立投产后转入运维的保障机制，确保数据仓库平台升级平稳过渡。

3.2 实施质量标准

定义并明晰数据仓库平台升级工作的过程后，还需要必要的过程管控实施纪律确保升级工作实施到位。结合数个数仓升级实施项目经验和 CMMI 能力成熟度模型，我们将传统数据仓库平台升级交付质量进行了等级划分，分别为 L0- 初始级、L1- 标准级、L2- 可重复级、L3- 量化管理级。

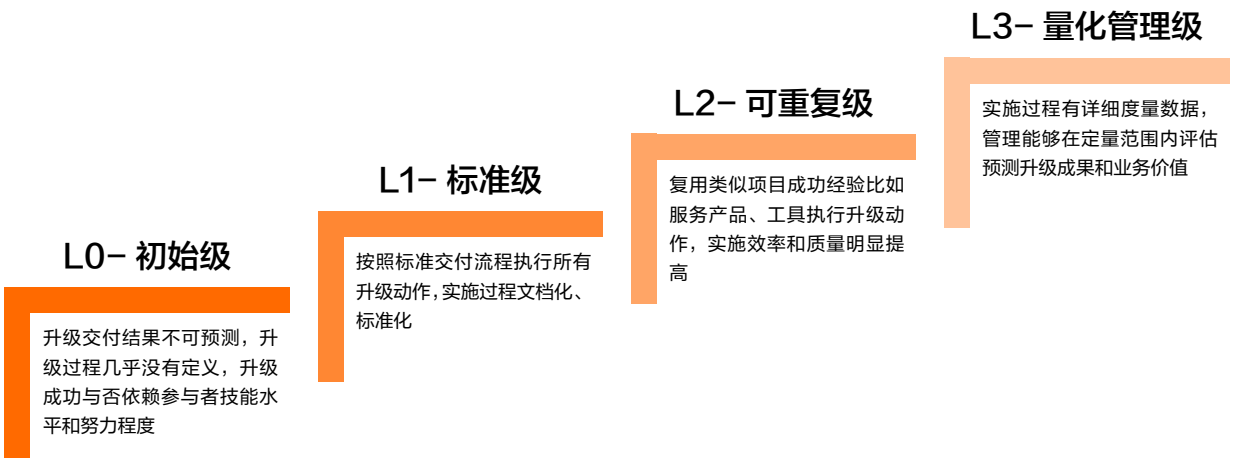


图 11 数据仓库升级交付质量级别图

以下升级数据仓库平台评估级别介绍：

1）**L0– 初始级：**该级别的数据仓库平台升级过程是无序的，有时甚至是混乱的，对过程无法明确定义，成功取决于个人努力；管理是反应式的；升级失败风险比较高。

2）**L1– 标准级：**该标准级别覆盖数仓升级实施完整生命周期，通过约束实施过程中各参与角色的所有动作，给予方法指导，确保每个交付步骤是标准专业的。项目交付团队基于标准的升级流程、动作规范和管理制度，各司其职，沉淀输出项目交付物，完成项目交付。该标准的重要评价维度如下：

a. 有完整的数仓升级交付方案文档和具体动作指导文档

b. 数仓升级过程中关键动作全部完成，没有缺漏

c. 已完成的关键动作按照模版上传交付物文档，如架构图、解决方案、操作手册等

3）**L2– 可重复级：**能重复早先类似升级工程取得的成功经验提高交付效率和客户满意度，主要通过借鉴或复用既有数仓升级项目的交付方案、自动化交付工具、交付服务产品等。重要的评价维度如下：

a. 有可直接复用的数仓升级技术方案，或方案中关键动作可直接复用

b. 有可直接复用的数仓升级自动化工具，或可作为项目中新工具的组件被复用

c. 数仓升级实施动作中，自动化工具覆盖的动作比例，至少达到全部实施动作的 40%

4）**L3- 量化管理级：**对数仓升级全流程有定量的理解和控制，交付过程管控数字化，所有实施动作尽可能在统一工作平台上完成，在线化白屏化完成项目交付，通过自动化工具复用和精细化管控保障项目成功交付。

a. 升级工具库托管在统一工作平台，其中工具可被交付动作自动化调用

b. 升级知识库托管在统一工作平台，可被交付动作直接引用为操作指导文档

c. 升级实施动作在线化完成度达标（且项目、协同在统一工作台在线化交付，在线化实施动作占比 70% 以上）

d. 量化升级工作中的投入产出

上述四个升级质量评估级别的对比如下表 2 所示，供企业考察厂商或者生态伙伴资质时参考。

质量标准等级	判断标准	专业技能要求	过程管控强度	生态协同度	潜在结果展望
L0- 初始级	未达到 L1 的均为 L0	强	弱	弱	不完全交付 正常交付
L1- 标准级	有引入标准升级流程方案 关键动作有完成 有产出标准交付物	高	偏弱	偏弱	不完全交付 正常交付 客户满意
L2- 可重复级	L1 交付物 引用成功经验方案 引用成功经验的交付工具， 全流程工具覆盖率 40%	一般	一般	一般	正常交付 客户满意
L3- 量化管理级	L2 交付物 使用统一交付工作台流水线工具 全流程自动化工具覆盖率 70% 投入产出量化指标	一般	强	强	客户满意

表 2 数据仓库升级质量评估级别表

3.3 升级实施交付阵型

本节主要介绍阿里云在传统数据仓库升级方面的交付阵型介绍，具备专业的升级交付团队是确保成功实施数仓升级的前提；整合内外部专家资源，建立完善的实施阵型是保证数仓升级成功的关键点之一。

角色	参与方	职责
项目领导组	甲方	负责项目整体方向把控和重大事项决策，负责确认项目总体目标，对项目实施中的重大问题做出决策，协调各单位之间的关系
	乙方	
项目管理组	甲方	负责日常项目沟通、风险、时间管理等工作
	乙方	
	ISV	
技术经理	乙方	负责项目产品技术需求管理和技术支持
数据库专家	乙方	负责数据迁移工具、模型迁移工具、调度迁移工具的调优、产品使用过程中问题处理、技术培训等工作
数据架构师	乙方	负责迁移方案的制定，包括数据和模型迁移、调度工具对接、数据校验、容灾架构设计、上下游应用切换方案等工作
业务架构师	甲方	负责业务输入，协助数据架构师制定迁移方案
研发工程师	ISV	负责具体迁移工作的实施
质量工程师	ISV	负责测试，数据质量检验工作

表 3 数仓升级角色及职责对应表

以阿里云交付服务体系实施数仓升级建设全生命周期视角，数仓升级建设需要多个团队和角色协作完成；除了客户和生态伙伴之外，数仓升级企业服务角色构成主要包括 SA、CBM、PMO、CSM、BA、TM、PD、QA 和 TAM，项目不同阶段由不同的团队来负责，每个角色在各个不同阶段及流程中承担的职责如图 12 所示，RACI 责任分配矩阵如下表所示。

	需求方案	风险识别	交付评审	启动规划	业务调研	方案设计	典型场景	测试迁移	测试验证	方案优化	生产迁移	生产验证	变更同步	生产并行	数据比对	兼容适配	知识转移	业务割接	正式上线	售后保障	项目验收
SA	RA	C	C	C																	
PDSA	RA	C	C	C																	
CBM	C	RA	C	C																	I
PMO			RA																		
CSM	C	C	C	RA	C	C	I	I	I	I	I	I	I	I	I	I	C	C	C	C	RA
TM	C	C	C	R	RA	RA	RA	RA	RA	RA	RA	RA	RA	RA	RA	RA	RA	RA	RA	RA	I
PD	C	C	C	I	C	C	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
TAM																		C	C	RA	I
客户	C	I	I	C	C	C	C	C	C	C	C	C	R	C	C	R	C	C	C	C	R
伙伴				I	R	R	R	R	R	R	R	R	RA	R	RA	R	R	R	R	R	I

R：执行人　A：责任人　C：征求意见　I：被告知

图 12 数仓升级企业服务角色与职责矩阵

角色简写	角色英文	角色说明
R	Responsible	谁负责，负责执行任务的角色，具体负责操控项目、解决问题
A	Accountable	谁批准，对任务负全责的角色，只有经同意后项目才能进行
C	Consulted	咨询谁，在任务实施前或中提供指定性意见的人员
I	Informed	告知谁，及时被通知结果的人员，不必向其咨询、征求意见

表 4　RACI 责任分配矩阵说明

通过项目管理工具 RACI 矩阵，项目经理可以有效的实现对项目横向和纵向的检查。横向检查，即可以了解并检查每一个工作包每个人的分工，其中，SA 解决方案架构师负责售前客户需求调研与方案设计，CBM 客户经理负责对售前商务风险等进行识别判断，PMO 负责对项目可交付性进行评审，CSM 对项目启动规划及最终项目验收负责，BA 对业务需求及业务架构设计负责，TM 对数据和技术架构及其落地方案负责，PD 对数据应用及数据产品设计负责，QA 对测试方案及落地负责，TAM 对售后运维负责，各角色在 CSM、BA、TM 的统筹下进行协作，共同对项目交付负责。

成功迁移典型案例：申万宏源数据仓库升级

4.1 行业变化与系统痛点

数字经济已被公认为经济发展的重要力量，也在不断的深刻改变和重塑着当下的经济发展和活动方式。证券行业作为服务实体经济的先锋力量，对于加快数字化转型的发展有着重大意义。近几年监管部门先后出台了一系列政策文件，强调推动证券行业数字化转型并以科技赋能业务发展，实现金融科技与证券业务发展互相促进、良性循环。这为证券行业的数字化转型和高质量发展奠定了坚实的基础。证券行业的数字化转型可总结为三大利好，具体如下：

1）有利于全面提升该行业乃至金融服务实体经济的整体效率与水平，为经济的整体高质量发展提供强大引擎；

2）有利于实现对资本市场发展的强大支撑力量；

3）有利于激发数字化经营新动力。

证券行业对数据的高质量要求正逐步推进证券行业的数字化转型步伐。目前，证券业务的信息系统已积累和储备大量业务数据，这些数据体量大、种类多、更新快、价值高，更为重要的是其数据价值可以有效的为证券业务创新、业务产品升级优化、企业决策、风险管理等层面提供重要支持，而这些价值数据也是证券行业的核心资产。

申万宏源证券作为国内领先券商，一直重视挖掘数据资产的价值，此前采用了十几年的国外传统数据仓库系统为业务发展带来良好的业务支撑。但随着大数据时代的到来，原有的数据仓库架构硬件老化、资源耗尽、扩容和数据迁移成本高、难以升级，无法满足申万宏源日益增长的业务数据需求。在业务快速变化的今天，尤其是在核心业务系统要求满足安全可控的大背景下，数据仓库升级成为券商的普遍选择。

4.2 数据仓库升级主要内容

基于以上的原因，申万宏源决定选择全新的云原生数据仓库进行升级，在综合对比多家厂商性能后，选择携手阿里云通过采用 AnalyticDB for PostgreSQL 等云上组件成功将服役了十几年的某厂家传统数据仓库平台云化升级，在云平台底座上，成功搭建了具备数据接入、数据存储、数据计算、横向扩展、数据共享的先进数据仓库平台。阿里云结合多年行业沉淀的申万宏源的 SDOM 模型经验，高效实现了数据模型从原数据仓库向新平台的平滑迁移。申万宏源的数据仓库系统整体性能提升 40% 以上，支撑了数万张核心表，为运营及业务快速发展提供了可靠、高效的云上基础设施环境。该项目的数据仓库平台升级实施涉及一百多套上游业务源系统、三十余套下游系统、两万多个任务、近 1000 多个服务接口、500 多 TB 数据（日新增 500 多 GB 数据）的业务规模，完成了几乎不可能完成的任务，整体的迁移过程如下图 13 所示。

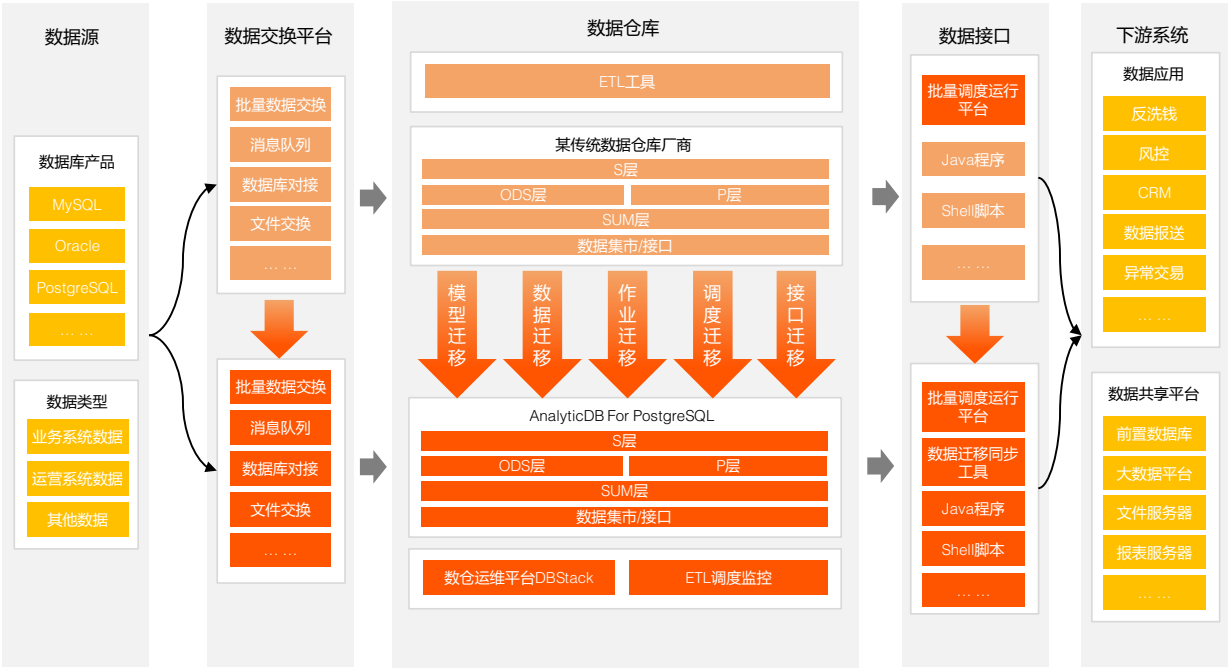


图 13 数据仓库迁移过程图

由于数据协同的上下游系统较多，原厂家的传统数据仓库平台升级替换耗时约 12 个月，在各方的共同努力下，最终完成所有预计目标的平稳迁移，同时保障现有业务平稳有序运转，帮助该证券公司实现自主可控、成本与风险双降，数据快速赋能业务。

那么，数据仓库升级是如何一步一步展开的？如何通过交付标准化手段，围绕交付标准流程、交付工具及数字化工作台，确保数据仓库迁移高质量高稳定履约，帮助客户数据业务转型升级？以下主要将结合该项目中总结沉淀出的数仓升级经验进行介绍。

4.3 升级交付流程标准化

项目启动后，首先进入轻咨询阶段，由申万宏源数据中心、阿里产研、GTS 以及 ISV 合作伙伴一起进行项目前期的交付咨询梳理。包括项目组织实施机制以及前期环境准备等。

此外快速构建起由客户、阿里云和 ISV 伙伴组成的项目交付阵型，明确职责矩阵、分工界面及衔接方式，结合一站式交付工作台 KOC 编排的工作需求说明书及标准流程动作，形成项目交付的里程碑节点计划。

在 CSM、BA、TM 等角色的统筹下，基于项目交付里程碑节点要求、项目执行计划及项目管理章程，高质量高效率推进项目交付标准流程动作。下文以需求调研和方案设计两个关键动作示例说明。

（1）业务需求调研

由于证券业务系统较多，数据仓库升级需要对接的上下游系统不但要保证平稳迁移，更需要根据不同的系统需求反馈，逐步按阶段按计划进行排期，业务调研阶段需对原数据仓库系统上下游每一个系统做详实调研，调研内容包括原数据仓库系统架构及环境、数据调研与分析、模型调研与分析、数仓 ETL 任务调研与分析等，输出调研分析报告，并与业务方做深入讨论与修正。

业务需求调研	交付任务	交付工具
原系统资源盘点清单	业务需求调研会议纪要 业务需求分析说明 技术需求调研会议纪要 数据需求调研会议纪要	SQL 兼容改造评估工具
原数仓系统架构反馈		
原数仓数据交互流程反馈		
原数仓技术性能要求清单		
原数仓库表统计清单		
原数仓 ETL 任务 / 脚本统计清单		
原数仓用户 / 权限统计清单		

表 5 业务需求调研表

交付分工上，整体由解决方案架构师负责，交付技术架构师、客户和 ISV 伙伴协助配合；项目经理负责组织对产出物进行评审；交付技术架构师结合交付工作台产出调研清单进行数据盘点，梳理原有数仓系统架构和数据流向，并参与客户的全局业务架构调研及核心业务流程调研；客户负责整理系统调研和数据盘点所需资料，组织业务调研和技术调

研，并对业务、数据和技术调研的问题进行解答，参与产出物评审及确认；伙伴配合交付技术架构师进行需求调研和数据盘点实施。

（2）迁移方案设计

方案设计阶段是根据前期的业务需求调研以及申万宏源系统实际情况进行量身定做，这个阶段是多方一起参与共创的阶段，最终设计并编制可落地的执行方案。内容包括：系统架构设计、规划设计、迁移方案、并行方案以及实施计划等。

图 15 方案设计

方案设计	交付任务	交付工具
新系统架构说明	数仓升级技术架构设计方案	SQL 兼容改造评估工具
规划设计说明	ADB PG 集群容量规划方案	
迁移方案说明	ADB PG 应用开发方案	
容灾方案说明	上下游系统交付设计	
并行方案说明	某传统数据仓库厂家迁移 ADB PG 方案	
实施计划说明	云上数仓容灾方案	
验收标准说明	升级交付路径规划	

表 6 方案设计表

交付分工上，整体由交付技术架构师负责，解决方案架构师、客户和伙伴协助配合；交付技术架构师负责 ADB PG 集群容量规划、ADB PG 集群应用开发规范、迁移方案、容灾方案等的落地实施方案的设计编写；

解决方案架构师协助确认整体方案的内外部沟通协调，参与方案设计评审；客户协助确认实施的软硬件资源环境，参与评审并确认方案设计，并对业务与技术问题解析解答；伙伴配合技术架构师和业务架构师进行方案设计实施。

4.4 交付内容工具化

（1）测试迁移

迁移测试阶段主要是对阿里云新数据仓库平台的功能性验证和典型场景验证，测试新系统在业务对接上是否存在问题。

同时该阶段还会论证迁移步骤的合理性，评估迁移工具的功能完备性；通过本阶段的典型场景验证，期望找到更多迁移交付中的技术性问题（如 SQL 兼容性、SQL 复杂度、作业复杂度等），用于更为准确评估迁移工作量，以便后续进行合理的工作安排。

测试迁移	交付任务	交付工具
迁移工具验证	数据迁移验证报告 ADAM for TD 评估分析验证报告 调度迁移工具验证报告 测试迁移复盘报告	数据迁移 DTS ADAM for TD
典型场景验证		

表 7 测试迁移表

生产迁移主要工作包括数据迁移、模型迁移和调度迁移三大块，而每一块的迁移都会涉及数据核对。该阶段迁移的主要难度在于三块工作完成迁移后，新数据仓库平台的数据要和原有数据仓库平台的数据一致，包括历史数据一致和新增数据处理的一致，并能够经过一个稳定周期的验证，任何一个环节的错误都将导致整体返工重做。

（2）生产迁移

生产迁移主要工作包括数据迁移、模型迁移和调度迁移三大块，而每一块的迁移都会涉及数据核对。该阶段迁移的主要难度在于三块工作完成迁移后，新数据仓库平台的数据要和原有数据仓库平台的数据一致，包括历史数据一致和新增数据处理的一致，并能够经过一个稳定周期的验证，任何一个环节的错误都将导致整体返工重做。

KOC 阶段：生产迁移	交付任务	交付工具
数据源迁移	数据源迁移及完整性报告	数据迁移 DTS ADAM for TD
模型迁移	模型迁移及完整性报告	
作业迁移	作业迁移及完整性报告	
数据迁移	数据迁移及完整性报告	
数据校验	数据校验验收报告	

表 8 生产迁移表

在其中会用到一系列的交付工具，例如数据迁移工具 DTS，确保铺底数据的原始拷贝和数据自动导出导入，保证数据一致性，简化数据迁移流程；而数据模型迁移工具 ADAM（亚当），则确保不同系统的语法转换，以促进迁移的平滑过度。

由于项目涉及的上下游系统较多，该阶段实施中具体的工作内容可以分为以下几个部分，这里同时总结了本项目中对应的工作量占比。

工作内容	工作量占比
数据源迁移	5%
模型迁移	25%
作业迁移	10%

工作内容	工作量占比
数据迁移	10%
新老系统数据比对	25%
下游系统对接	25%

表 9 项目工作量统计表

交付分工上，本阶段需要由交付技术架构师负责，解决方案架构师、客户和伙伴协助配合；交付技术架构师负责制定迁移方案设计和编写，并参与迁移及数据校验工作；解决方案架构师协助参与方案设计及评审工作；客户协助并确认迁移和校验结果；伙伴配合具体实施迁移工作。

（3）系统并行

本阶段为迁移投产前的数据稽核校验阶段，由于数据仓库平台涉及核心业务系统，稳定性要求极高，有必要在迁移期间保证数据仓库新旧双系统并行，待数据迁移完整性校验通过后，再正式进入生产割接前的最后准备阶段。

KOC 阶段：系统并行	交付任务	交付工具
系统并行	新老数据仓库数据比对分析报告 数据迁移成功确认函	青天鉴
业务割接	业务割接报告	

表 10 系统并行表

由于涉及的上下游系统较多（例如反洗钱，风控，CRM 等均与数据仓库做大量连接），因此在系统并行阶段，交付主要面临数据不一致、新旧平台特性差异以及系统配置不同这三方面的挑战。

数据不一致方面，统计数据迁移不仅涉及数据表及数据量。数据核对的工作更为繁琐，数据核对可分为对象级、记录级、指标级、全表级、每日批量几个级别的核对。每个级别都涉及大量的脚本编写和繁重的数据比对。基于此开发自动核对脚本生成工具，可根据元数据配置自动生成核对脚本，并可以对脚本进行批量调度，快速得出核对结果。这样，核对的主要工作就是发现不一致的数据并解决对应的问题。平台特性差异以及系统配置方面的差异挑战，则主要通过系统调整、推送端适配修改、以及人工校对修改等方式进行完善。

这个阶段也会用到一些标准化交付工具，例如 GTS 研发的数据校验工具青天鉴等，这些工具主要完成数据比对等功能，便于更好的实现双系统并行。

核对大步骤：

1、全量数据核对

2、跑批后增量数据核对

利用虚拟化工具生成核对脚本，自动对比，人工确认

每次核对步骤：

1、基础核对：表数量核对、记录数核对、指标汇总核对

2、完整核对：选定数据范围进行全字段内容核对

图 16 系统并行数据核对

在交付分工上，并行阶段需要由交付技术架构师负责，解决方案架构师、客户和伙伴协助配合；交付技术架构师负责数据比对策略指定及数据差异分析；解决方案架构师协助参与并行方案评审工作；客户协助确认数据比对结果的准确性、一致性；伙伴负责具体数据比对工作。

4.5 小结

该项目的升级交付经过标准化流程梳理后，升级顺利，从最终实施效果来看，云上数据仓库新平台比老平台运行效率高出 40%，这些数据每天都会被业务系统频繁调用，数据基础设施性能提升，增强了响应业务需求的时效性。正如申万宏源证券数据中心部门负责人石宏飞所言，通过数仓升级，申万宏源证券也一步步的向“四个化”进发，即实现数据治理全面化、数据资产标准化、推行数据应用智能化、实现数据服务制度化。

申万宏源证券与阿里云合作的数据仓库升级项目，可以作为金融证券行业数字基础设施进化转型路径的典型样本。数据承载的核心载体，是云原生、高弹性、大并发、快回档、强安全的数智化基础设施和系统级解决方案。申万宏源后续还将基于数据能力驱动，逐步实现全面数据运营。

总结与展望

随着数据仓库升级在金融、电信行业头部及领先企业逐渐落地，服务商和生态伙伴经历了各类业务场景能力沉淀过程，产品技术和实施方法日趋成熟，需求端对新型云原生数据仓库的理解和信任逐步加深，行业认可度和市场规模都在迅速提升。但是，近年来传统数仓掣肘明显，存在软硬绑定，难以升级与维护，成本高昂；架构老化，难以赋能业务创新；体系封闭，受制于人，难以突破等问题。对企业而言，不论是业务创新也好，降本增效也好，立足点是要消除落后的基础体系，构建起面向未来可持续发展的底层平台架构，重塑技术底座，而基于数据仓库的数据平台建设方案在其中发挥了地基作用。

阿里云国产新一代云原生数仓体系化解决以上难题，该类产品日渐成熟，并且在关键行业的关键应用中被成功实施，成果显著。通过帮助金融、电信行业客户将传统数仓全面升级至云原生数仓，或者构建于云原生数仓基础上的数据平台全新架构，有效满足客户对于数据平台实时化、弹性扩展、高性价比及安全可控的诉求，突破传统数仓技术瓶颈，赋能企业数智化创新。

本白皮书对传统数仓升级，称之为“升舱”，未来在金融、电信行业“升舱”后，基于新一代云原生数仓可以构筑起更多的“一体化”技术架构和数据应用的创新，具体体现为以下两点：

1）在离线一体化：基于云原生数据仓库的底层算力基础，建设在线离线混合负载能力，打造跨平台、跨业务融合数据服务能力，数据高效互融互通，创新孵化一批业务运营应用，高效支撑企业复杂实时化的业务场景需求，实现业务体验一体化、高效化；

2）一站式数据平台：集企业级数仓、数据集市、数据开发、AI 融合、数据管理、数据共享、数据服务为一体的一站式解决方案，形成数据计算实时化，数据资产化，数据价值在线化。

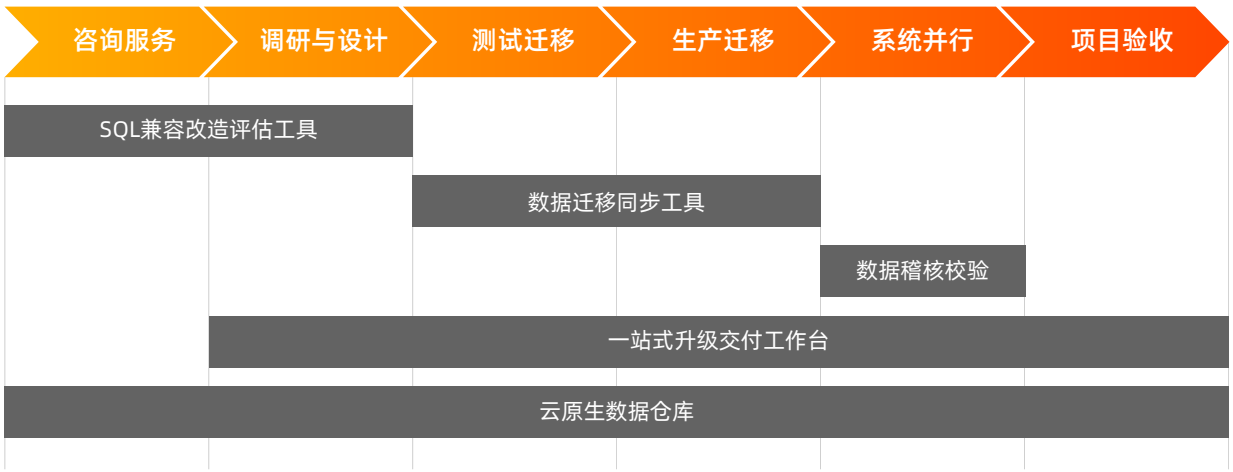
同时，本白皮书提出升舱体系化的指导方法论，并提供了配套升舱最佳实践路线，例如升舱交付阵型与协同矩阵；在“力求不变的云化升级”的实施路径“五阶十步”；围绕升舱的全栈、全链路实用工具组件等。此外，白皮书明确定义升舱交付能力成熟度模型，分别定义 4 个级别，在具体升舱实践中，对提升客户满意度具有重要的牵引指导作用。因此，提炼和总结传统数据仓库升级的服务内容，沉淀行业通用能力，形成标准化的整体解决方案，助力企业客户数字化转型，提升数据仓库服务商和生态伙伴的规模化交付能力，其重要意义不言而喻。

与之升舱交付标准相关的“升舱”计划之一是以阿里云新一代云数仓 AnalyticDB 为平台，欢迎更多生态合作伙伴加入到升舱行动中，包括但不仅限于产品、解决方案、服务等，一同助力赋能金融、电信及其他国家关键重要领域深化数字化转型，再到数智化协同创新。

当然，数据仓库的演进和迭代一直随着发展而改变，但无论怎么变化，新型云原生数据仓库升级交付标准化建设始终以时代发展的趋势和业务需求为出发点，继续围绕客户价值、业务价值为主线持续向前，不断将其推向可持续发展的良性轨道中。

附：数据仓库升级实施云上组件

数据仓库升级是一个大型数据工程再造。除了需要有合适的云原生数据仓库承载某传统数据仓库厂商的业务外，还需要有周边丰富的生态工具来确保升级工作高效完成。该传统数仓厂商往往包含上万个作业，每个作业都涉及到大量的 SQL 语句，几百到上千行不等，同时新老数仓往往很难做到百分百的语法兼容，不能单纯靠人力解决。使用工具可以有效避免大量的重复动作，以下我们重点介绍围绕云原生数据仓库 ADB PG 进行某传统数据仓库厂商升级替换中需要用到的云上实用工具组件。



（1）云原生数据仓库

首先需要一款对标某传统数据仓库厂商的云原生数据仓库——阿里云 AnalyticDB PostgreSQL 版（以下简称 ADB PG），是一种大规模并行处理（MPP）数据仓库服务，可提供海量数据在线分析服务。ADB PG 主要功能特征如下：

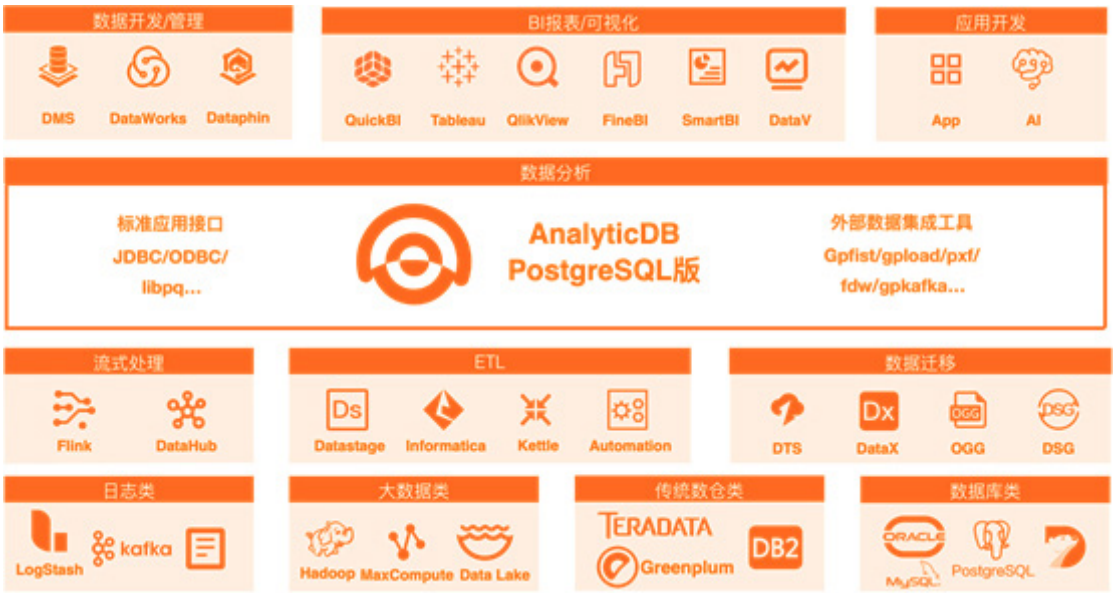
1. **性能优异。**ADB PG 搭载自研向量执行引擎，针对火山模型的缺点和现代硬件的特点，结合向量化计算、即时编译 (Just In TIme, JIT) 等技术，专注于执行性能、特别是 OLAP 场景的执行性能提升。

2. **管理和使用成本低。**ADB PG 基于开源项目 Greenplum 构建，由阿里云深度扩展，兼容 ANSI SQL 2003，兼容 PostgreSQL/Oracle 数据库生态，支持行存储和列存储模式。既提供高性能离线数据处理，也支持高并发在线分析查询，是各行业有竞争力的 PB 级实时数据仓库方案。

3. **生态丰富。**ADB PG 具备完善和开放的生态系统，支持无缝对接业界以及阿里云生态的数据管理工具，BI 报表可视化工具，ETL 工具，数据迁移工具，同时可以与业界流行的流式处理系统，日志类系统，大数据类系统，传统数仓及数据库类系统互通。

4. **低成本。**ADB PG 采用计算节点本地存储的模式，支持行式存储和列式存储，支持非易失存储、固态硬盘、机械硬盘等多种存储介质，支持单节点最大 10TB 存储空间。在此基础之上，ADB PG 进一步提供了存储压缩能力、OSS 外表存储能力和共享存储格式、存储分层等能力，满足用户在不同场景的需求。

5. **优良的可扩展性。**ADB PG 拥有良好的弹性能力，同时支持水平扩展和垂直升降配，以满足不同场景的扩展和弹性需求。



（2）一站式升级交付工作台

阿里云提供一站式的现场交付工作台 KOC，将传统数仓升级这样的大工程由人力驱动升级为流程驱动，将手工操作升级为自动化操作，通过在交付中引入 KOC 实现以下五大能力：

1. **交付流程化：**每一个交付动作串联成 SOP，建立依赖约束，交付各角色围绕 SOP 严格执行到位

2. **交付职责清晰明确：**每个交付动作责任落实到人，对动作交付结果和完成时间是否符合预期负责

3. **操作质量保证：**每个交付动作有核查人，工具执行有日志记录，手工做操作有结果截图证明

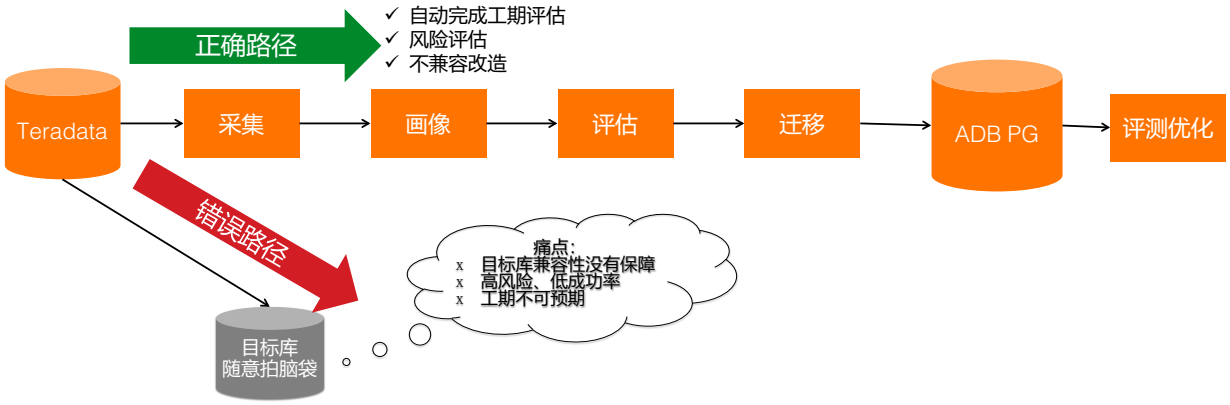
4. **进度里程碑管理：**为项目设定合理的排期、里程碑，据此考察项目进度是否落后，聚焦于项目风险和卡点

5. **关键交付物归集：**将项目过程中持续产生的操作文档、技术方案、工具沉淀收集整理，作为后续复盘和新项目启动时的宝贵资产

KOC 工作台可以有效的确保数仓升级实施高标准、高质量完成，达到前述 L2 自动级及以上的实施标准，弥补统一升级实施过程中技术人员的水平差异、保证操作质量、提高项目管理质量。

（3）SQL 兼容改造工具

针对新旧数仓数据库的 SQL 语法进行兼容性改造，主要用于 ETL 作业脚本和数据库对象 DDL 脚本的改造工作，主要依赖阿里云数据库和应用迁移（Advcanced Database and Appliction Migration，简称 ADAM）工具实现。该工具是阿里巴巴多年去 IOE 的方法论和经验的工具化成果，可以借助该工具自动化评估某传统数据仓库厂商到 ADB PG 在 SQL 语法、DDL 转换方面的兼容点、不兼容点以及改造优化建议。



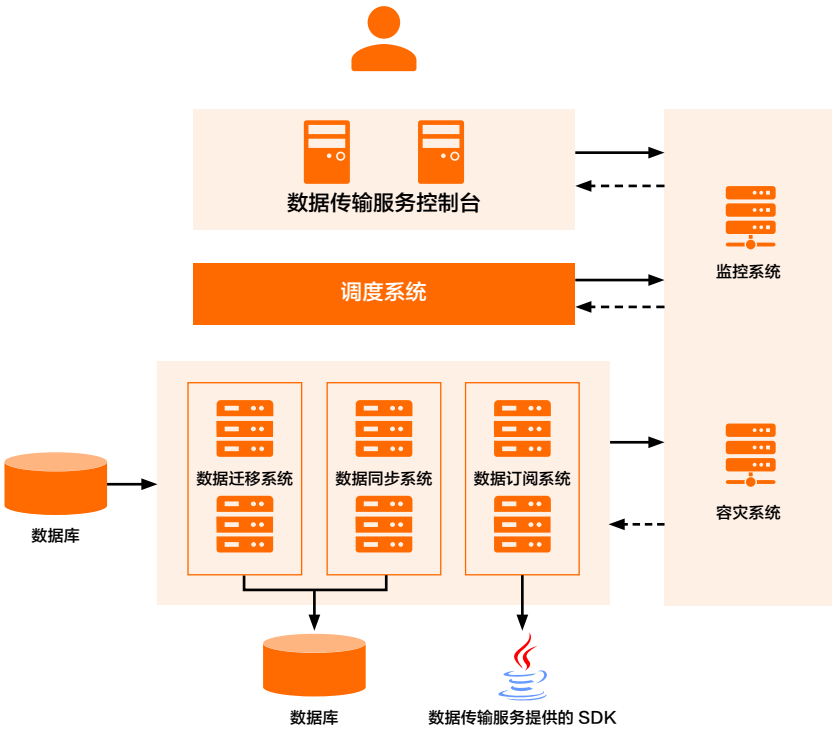
（4）数据迁移同步工具

数据迁移同步可采用阿里云迁移工具 DTS，支持传统数据仓库到阿里云数据库端到端的数据迁移工作。

DTS 可以在两个数据源之间同步正在进行的数据变更。 数据同步通常用于 OLTP 到 OLAP 的数据传输。 数据同步包括以下两个阶段：

1）**同步初始化：**DTS 将源数据库的存量数据加载到目标数据库。

2）**数据实时同步：**DTS 同步正在进行的数据变更，并保持源数据库和目标数据库的同步



数据迁移过程包括三个阶段，即结构迁移、全量数据迁移和增量数据迁移。 如果需要在迁移期间保持源数据库的正常运行，当您在配置迁移任务时，必须将结构迁移、全量数据迁移和增量数据迁移都选为所需的迁移类型。

1）**结构迁移：**在迁移数据之前，DTS 需要在目标数据库中重新创建数据结构。 对于异构数据库之间的迁移，DTS 会解析源数据库的 DDL 代码，将代码转换成目标数据库的语法，然后在目标数据库中重新创建结构对象。

2）**全量数据迁移：**在全量数据迁移阶段，DTS 会将源数据库的存量数据全部迁移到目标数据库。 源数据库保持运行状态，在迁移过程中仍不断进行数据更新。 DTS 使用增量数据读取模块来获取全量数据迁移过程中发生的数据更新。 当全量数据迁移开始时，增量数据读取模块将被激活。 在全量数据迁移阶段，增量数据会被解析、重新格式化并存储在本地 DTS 服务器上。

3）**增量数据迁移：**当全量数据迁移完成后，DTS 会检索本地存储的增量数据，重新格式化，并将数据更新应用到目标数据库中。 此过程将持续下去，直到所有正在进行的数据变更都复制到目标数据库，并且源数据库和目标数据库完全同步。

（5）数据稽核校验工具

数据仓库动则数万张表，升级实施中经常要做数据迁移、ETL 测试，在迁移结束或测试数据时，需要频繁的做数据稽核比对，阿里云青天鉴就是一个支持多源异构数据源的数据校验工具，辅助我们做数据稽核时实现工具化、自动化，青天鉴功能示意图如下

该数据稽核校验工具支持全量校验和指标校验两方面：

1. 全量校验，即源端 A 表到目标端 B 表，所有记录完全一致，主要用于同构数据迁移的场景。

2. 指标校验，即表数据量或内容计算结果一致性校验，即源端和目标端 count()、sum() 等聚合结果对比。尤其是大规模数据仓库迁移，字段比较非常耗时，需要在绝对准确性和效率上追求合理的时间成本和相对准确性。因此，指标校验主要适用于数据仓库迁移后源和目标数据库的数据量校验。数据仓库中常常统计业务含义的指标，青天鉴支持带 group by 的指标校验，即业务方指定的源端指标（一个或多个字段，group by 条件作为 pk）和目标端指标（一个或多个字段，group by 条件作为 pk）对比。

（6）批量调度运行平台

DataWorks（大数据开发治理平台）是阿里云重要的 PaaS（Platform-as-a-Service）平台产品，提供数据集成、数据开发、数据地图、数据质量和数据服务等全方位的产品服务，一站式开发管理的界面，帮助企业专注于数据价值的挖掘和探索。DataWorks 支持多种计算和存储引擎服务，包括离线计算 MaxCompute、开源大数据引擎 E-MapReduce、实时计算（基于 Flink）、机器学习 PAI、云原生数据仓库 AnalyticDB for PostgreSQL，云原生数据仓库 AnalyticDB for MySQL，并且支持用户自定义接入计算和存储服务。DataWorks 为您提供全链路智能大数据及 AI 开发和治理服务。

在过程中，可以使用 DataWorks 进行数据源（如 MaxCompute）到 ADB PG 的数据传输、转换和集成。