DOI：10.3969/j.issn.1672-7703.2020.05.011

梦想云数据连环湖建设研究

杨 勇[1,2] 黄文俊[1,2] 王铁成[1,2] 王 华[1,2] 孟令培[1,2] 谭雷军[1,2] 梁 肖[1,2]

（1 中国石油集团东方地球物理勘探有限责任公司；2 北京中油瑞飞信息技术有限责任公司）

摘 要：中国石油信息化在 30 多年的持续发展中，历经分散建设、集中建设、集成应用 3 个阶段，正在迈向协同共享新阶段，努力建成“共享中国石油”。中国石油上游业务板块顺应数字化转型发展需要，针对中国石油上游平台多、数据库多、孤立应用多的“三多”问题，结合实际业务需求和信息技术发展趋势，对上游业务数据资源体系进行了系统的研究，对数据的汇聚、存储和应用进行了深入的分析，组织研发了勘探开发梦想云及数据湖技术，并在梦想云 1.0 统一数据湖的基础上，提出了逻辑统一、互联互通的集团级主湖和油气田公司级区域湖的数据连环湖设计与建设方案，通过试点验证，取得了预期效果，为梦想云 2.0 数据生态建设提供了有力支撑。

关键词：勘探开发梦想云；数据连环湖；数据生态；数据存储；共享应用

中图分类号：TE19 **文献标识码**：A

Research on construction of Data Interlinked Lakes of E&P Dream Cloud

Yang Yong[1,2], Huang Wenjun[1,2], Wang Tiecheng[1,2], Wang Hua[1,2], Meng Lingpei[1,2], Tan Leijun[1,2], Liang Xiao[1,2]

(1 BGP Inc., CNPC; 2 Richfit Information Technology Co., Ltd.)

Abstract: The informatization construction of CNPC has been continuously developed for more than 30 years, and has gone through three stages: decentralized construction, centralized construction, and integrated application. Now it is moving forward to a new stage of collaboration and sharing, so as to establish “sharing CNPC”. In order to meet the requirements of digital transformation and development, in response to the “three-multiple” issues of multiple platforms, multiple databases and multiple isolated applications of CNPC’s upstream business, and combined with the actual business needs and the development trend of information technologies, CNPC has systematically studied the data resource system of the upstream business, deeply analyzed the data aggregation, storage and application, and researched to develop the E&P Dream Cloud and Data Lake technologies. Based on the unified Data Lake of Dream Cloud v1.0, CNPC has put forward the design and construction scheme of Data Interlinked Lakes, with group-level main lake and the oil and gas field company-level regional lakes, which are logically unified and interconnected. Through the pilot verification, the expected results have been achieved, which provides strong support for the construction of data eco-environment of Dream Cloud v2.0.

Key words: E&P Dream Cloud, Data Interlinked Lakes, data eco-environment, data storage, shared application

0 引言

随着信息技术的深入发展，数据量爆炸式增长，为满足日益变化的数据类型和大数据量存储，数据湖技术已经成为各大公司的数据存储与共享应用的首选。中国石油上游业务信息化建设，在“六统一”的原则下，经历了由分散到集中再到集成的建设过程。各项业务基本实现了数字化管理，支撑了专业应用并保护了数据资产；但是存在数据重复采集、综合利用率低、数据孤立、流通困难等问题，难以挖掘数据价值。为了充分利用这些数据，让数据流动起来，充分吸纳国外的先进案例与技术[1]，按照中国石油上游业

基金项目：中国石油天然气股份有限公司投资信息化重点项目“勘探开发一体化协同研究及应用平台（一期）建设”（PetroChina-IT-2017-N104）。

第一作者简介：杨勇（1978-），男，天津人，硕士，2016 年毕业于北京航空航天大学，高级工程师，主要从事石油上游业务信息化及数字油田、智慧油田研究与建设工作。地址：北京市石景山区京原路 7 号东方地球物理公司信息技术中心，邮政编码：100043。E-mail：yangyongbgp@cnpc.com.cn

收稿日期：2020-07-21；修改日期：2020-07-30

务信息化总体蓝图，建成了勘探开发梦想云（E&P Cloud）平台。梦想云统一数据湖（简称数据湖）作为梦想云的核心支撑，统一管理了上游业务油气勘探开发领域，涉及物探、钻井、油气生产等 15 个专业、8 种类型 1.6PB 的数据资产。

数据湖基于中国石油勘探开发数据模型 EPDM 2.0，开发了主数据管理、元数据管理、数据质量管理、数据集成监控、数据安全管理、数据服务等主要功能，搭建了统一数据服务（DaaS，即 Data as a service）体系，初步实现上游全业务链数据的统一治理、管理与共享应用。在应用过程中，也存在以下需要解决的问题：

（1）随着业务应用对数据的深度和广度提出了更高的要求，对大数据分析、认知计算等人工智能技术的依赖不断提高，需要提供灵活快速的数据服务能力。

（2）梦想云数据湖对大块数据，如物探数据体等的存储方式没有得到合理解决，同时大块数据异地访问速度瓶颈问题亟待解决。

（3）经过一年应用实践，上游业务数据共享和应用升级需求更加迫切，勘探开发数据湖需要能够支持更加广泛的应用。

因此本文提出连环湖方案，旨在进一步提升数据质量、实现数据多元化共享，打造新型数据生态系统，达到“连环湖互联、金数据共享、数据价值持续提升、大步迈向智能化”的目标。

1 技术方案

在汲取国内外优秀厂商数据湖建设思路的基础上，结合中国石油勘探开发上游业务特点和先进的数据湖技术，研究确定了梦想云连环湖方案。

1.1 连环湖方案

连环湖方案由主湖与区域湖构成[2]。主湖负责上游业务数据标准、主数据的统一管控，实现上游数据的集中管理与共享应用，形成企业级数据资产；区域湖实现大块数据分布式存储与就近应用访问，承载本单位数据治理工作，并支撑扩展业务数据管理与共享应用。通过连环湖架构实现数据逻辑统一、分布存储、互联互通。梦想云连环湖总体方案如图 1 所示。

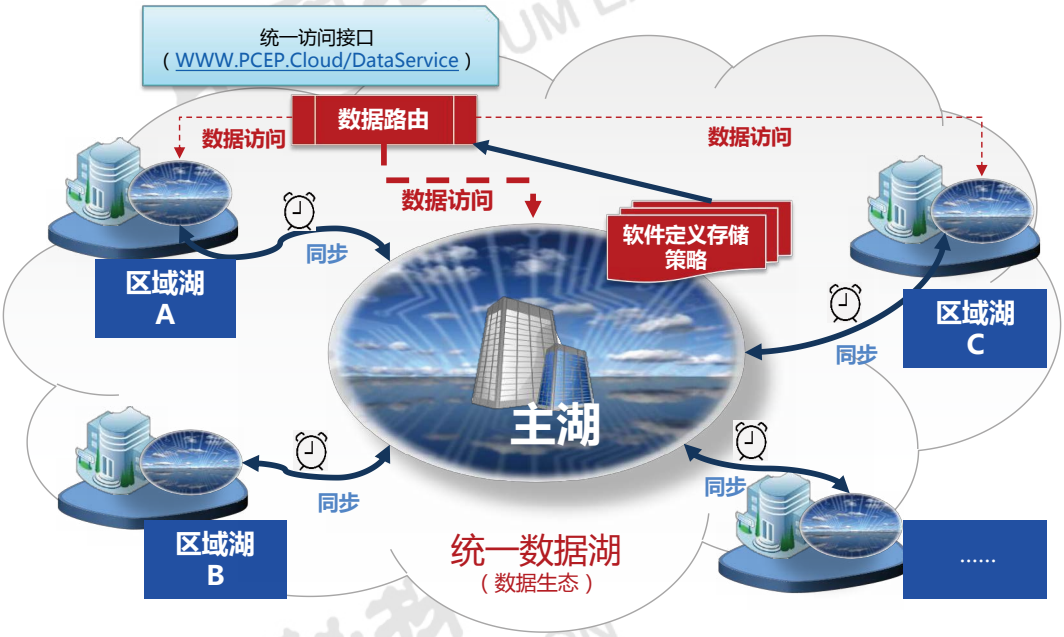

图 1　梦想云连环湖总体方案示意图

Fig.1 Overall scheme of Data Interlinked Lakes of Dream Cloud

1.2 连环湖逻辑架构

基于敏捷迭代、逐步升级的思想，按照“3+3”的架构[3-4]，进一步增强全局数据治理、数据共享和智能分析能力，建立上游业务开放数据生态，实现全业务链的打通，为上层业务提供敏捷服务。连环湖各组成部分的逻辑架构包括 3 个层次，分别是专业数据库层、共享存储层和数据分析层，并通过建立保障体系将 3 层有机融合，通过统一数据服务对外提供应用服务。梦想云连环湖逻辑架构如图 2 所示。

图 2 梦想云连环湖逻辑架构图

Fig.2 Logical architecture of Data Interlinked Lakes of Dream Cloud

3 层逻辑架构中，底层建立专业数据库层，实现采集和业务应用专业库的统一管理，保证数据一致性与溯源能力。第二层为企业核心数据共享存储层，强化数据标准化治理与全局共享。基于统一的模型标准（EPDM 2.0 模型等），针对主数据、结构化数据、非结构化数据、时序数据进行数据治理，建立标准统一、逻辑统一、互联互通的共享存储层。基于共享存储层数据抽取形成满足各种业务应用需要的数据分析层，包括高速索引、领域知识库、分析库、模型库等，基于统一的数据服务地图对外支撑应用。

3 层保障体系包括组织机构与管理制度、业务标准与数据模型、数据湖管理工具。通过管理工具将 3 层架构进行连接，整体对外服务，建立上游业务开放的数据生态。

2 实现方案

敏捷迭代的微服务架构，其核心思想是通过领域划分来实现解耦，各个领域应用独立发展，整体稳定运行，核心标志是数据库环境的相互独立，因此按照连环湖的方式，在逻辑统一、互联互通的前提下，通过应用一系列技术方法实现 3 个维度的解耦，满足原生云应用快速建设的需求，推动企业数字化转型发展。

2.1 专业数据库层

专业数据库层通过提供数据库即服务（DBaaS，即 Database as a service）的能力，帮助梦想云的生态伙伴进行专业库采集和专业库应用的建立和管理，最终达到入湖数据源全面受控管理的目标。

通过元数据的专业库管理模块，实现采集应用与系统标准的分领域统一管理；通过 EPDM 模型管理模块，实现共享存储层标准规范的全领域统一管理；通过数据集管理模块，实现应用规范分领域统一管理；通过元数据关系管理模块，为各层数据建立血缘关系管理，达到解耦后的逻辑统一，并指导数据集成与治理工作。

通过上述能力建设，实现数据采集型应用、数据资产管理型应用和数据服务与统计分析型应用三者之间两个维度的解耦与隔离，满足应用并行建设、持续敏捷迭代的需要（图 3）。

2.2 共享存储层

共享存储层按照连环湖架构建设，建立数据治理环境，存储管理中国石油企业级数据资产及应用共享数据。通过连环湖架构，在统一管控的前提下，支撑集团总部和地区公司之间的应用建设，满足上游业务快速发展需求，推进上游业务全面进入“数字＋智能＋油气田业务”的新业态。

对于结构化数据管理（图 4），在集团总部统一建立上游业务通用应用（统建）的数据贴源层，数据结构基于源头系统。按照上游业务共享数据需要，确定数据管理范围，在共享存储层治理环境的中间库里通过数据治理及 ETL 工具对数据进行标准化后（该类共享数据标准简称 S0）存储到基于 EPDM 2.0 扩展的共享存储层，进而根据业务应用需要，建立面向业务主题的高速服务，支撑分析应用。

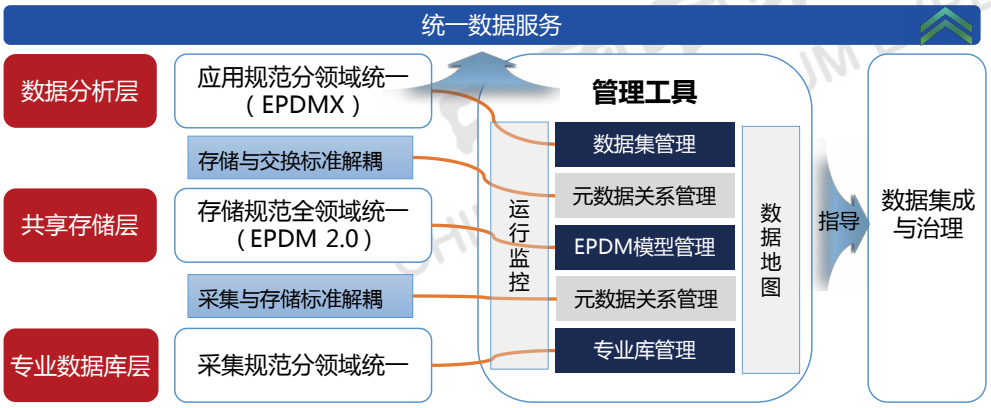

图 3　数据湖管理工具逻辑架构图

Fig.3 Logical architecture of Data Lake management tools

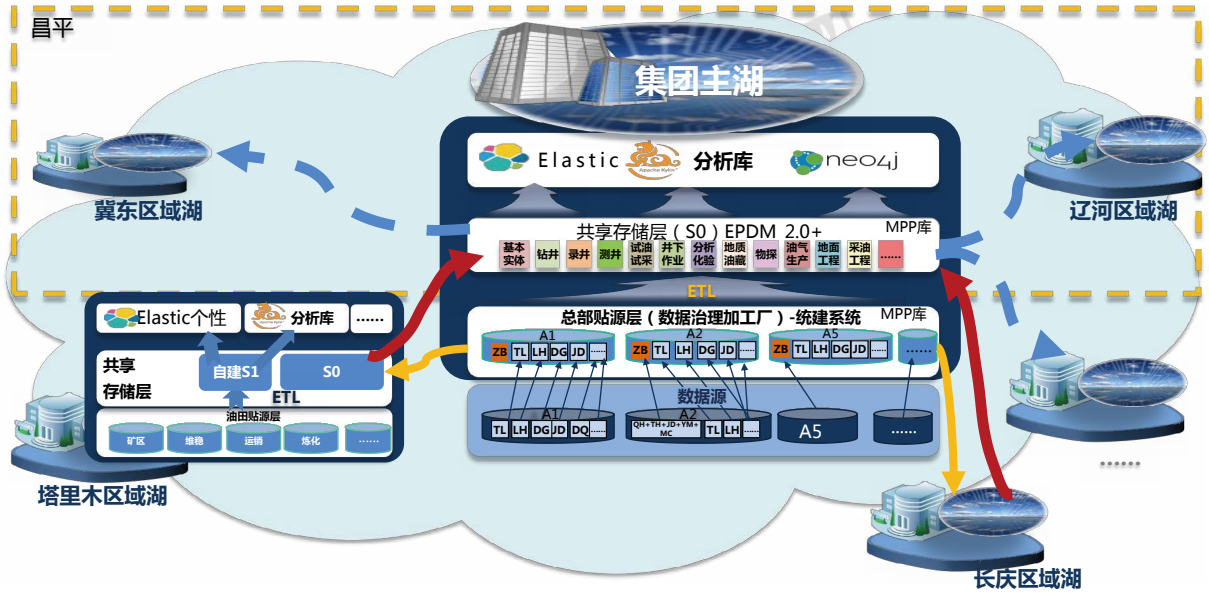

图 4　结构化数据存储逻辑架构图

Fig.4 Logical architecture of structured data storage

各油气田及研究单位部署区域湖，其中结构化数据由两部分组成。第一部分是基于 S0 标准的数据，由各油气田采集并治理后入区域湖，同时统一同步到主湖，以满足通用应用需要。第二部分是油田自建部分，由各油气田自行治理后存储到区域湖共享存储层（区域湖自建共享数据标准，简称 S1），支撑油田本地化应用。

结构化数据共享存储采用 MPP（大规模并行处理器 Massively Parallel Processing ）数据库技术，将任务均衡分解到多个节点同时进行运算，有效解决大规模数据作业计算、缓存和输入输出（IO）带来的性能问题[5]（图 5）。

连环湖中非结构化数据存储，采用基于简单存储服务（Simple Storage Service，简称 S3）标准协议的软件定义分布式文件存储架构，实现地区和集团总部之间分布式存储，主湖主控保证逻辑统一，用户基于统一的 RESTful 服务访问文件内容，支持软件定义数据多镜像与就近访问，满足地震等大块数据存储与高效应用[6]。

根据业务需求，非结构化数据批量上传至存储缓

冲区，经业务审核后，推送至就近的 S3 存储区，存储软件按照预先设置好的策略，自动实现数据多镜像复制，在解决存储管理的同时，满足用户应用体验要求（图 6）。

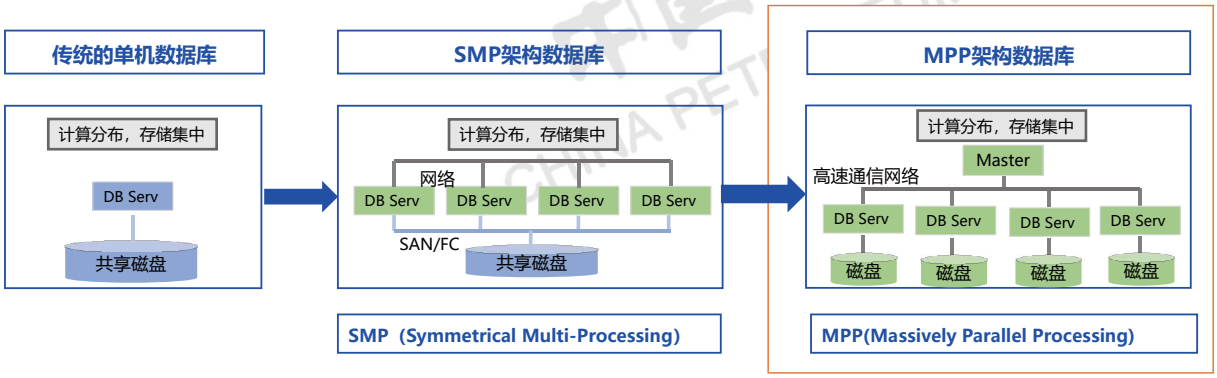

图 5 结构化数据核心存储技术选型示意图

Fig.5 Schematic diagram of technology selection for core storage of structured data

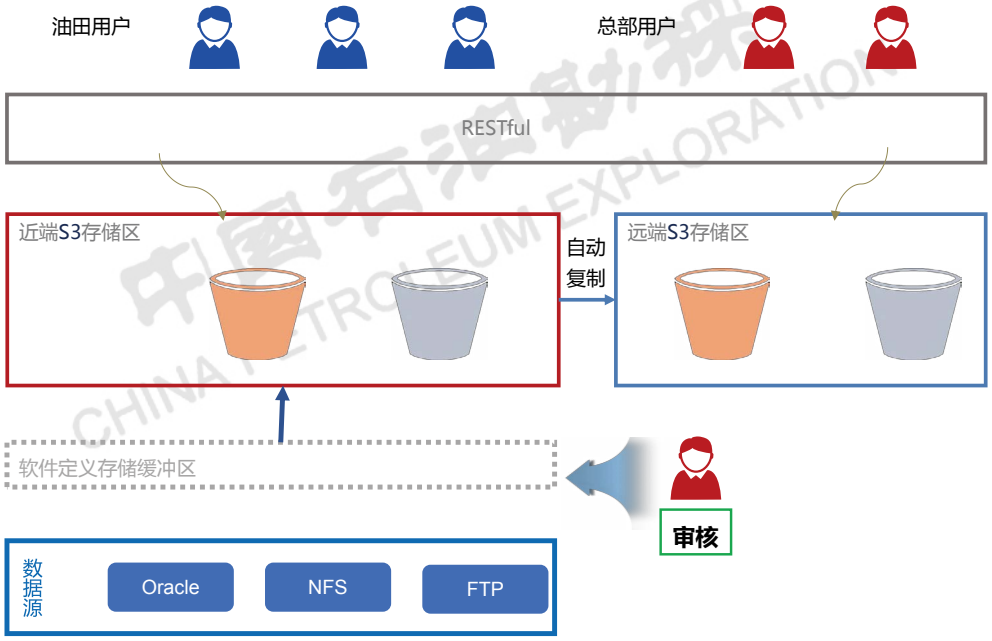

图 6 非结构化数据存储及流转示意图

Fig.6 Schematic diagram of storage and flow of unstructured data

连环湖时序数据存储，采用主流时序数据库技术，分为生产现场、油气田公司、集团总部三级管理。生产现场时序数据在生产网中直接组态应用，满足现场生产指挥需要；按照业务需要，经处理后的时序数据按照统一的模型结构存储在区域湖共享存储层时序数据库中，满足油气田生产监控与指挥应用需求，有条件的可开展大数据和智能化分析应用；地区公司和集团总部之间按照自动复制的方式，按需上传至主湖，满足集团总部大数据及智能化分析应用需求，同时满足备份存储要求（图 7）。

2.3 数据分析层

数据分析层包含高速索引、领域知识库、分析库及模型库，同时构建基于大数据分析和认知计算等技术的分析能力。

2.3.1 高速索引

ElasticSearch 是一个分布式的实时文档存储系统，能够实时分析搜索引擎。采用 ElasticSearch 创建高速索引[7]，能够快速获取检索结果，支持多数据源和文档对象（文档对象指的是非结构化数据或半结

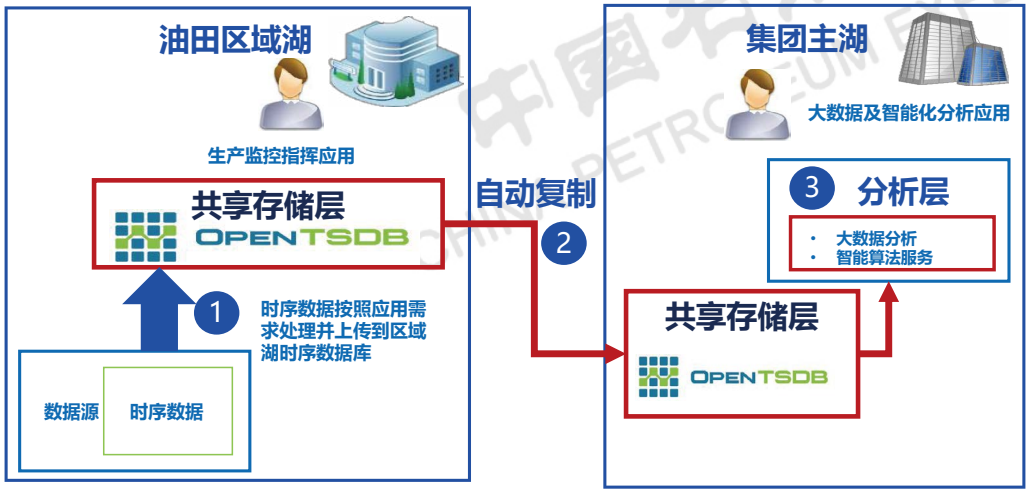

图 7　时序数据存储及流转示意图

Fig.7 Schematic diagram of storage and flow of time series data

构化数据），同时具备高可用、易扩展能力，并支持集群、索引分片和复制。高速索引是实现面向业务主题的高速数据服务，屏蔽物理表访问的复杂性，满足业务按内容搜索应用的需求。

2.3.2 知识库

知识库由知识建模、领域知识库和知识检索三部分构成。

知识建模：基于数据湖中的数据，通过知识获取、图谱化、数据挖掘和知识理解 4 个步骤，实现知识建模。知识图谱描述领域内的实体、概念及其关系，构成全局的语义网络图，实现相关信息连通，支撑智能化应用的建设。

领域知识库：将勘探开发数据进行知识化，将知识进行结构化描述，形成各业务领域的知识图谱，直观、高效展现勘探开发业务对象之间的关系，基于知识本体将知识图谱进行关联，建立领域知识库，采用主流的图数据库进行存储与应用。

知识检索：通过自然语言识别，实现智能语音检索。利用知识图谱对数据湖内数据进行建模，描述领域内的实体、概念及其关系，构建全局语义网络图，实现相关信息连通，提炼复杂业务模型，提供业务驱动、决策支撑的能力，采用丰富的展现方式，适应用户的认知需求。

2.3.3 分析库

通过 ETL 工具将共享存储层的数据进行抽取和清洗，构建应用数据集并加载转换到分析库中[8]。分析库采用星型模型，包含事实表、维表，ETL 工具按照预定周期同步到分析库。分析库提供复杂的通用数据分析业务支撑能力，同时提供海量数据实时分析响应能力。原始数据或业务数据集抽取到分析库，之后采用数据预处理方式，对构造的多维模型数据生成数据立方体，存储在 HDFS 文件系统中，供大数据分析使用。

分析库的主要价值是改变原有的数据库存储过程中进行分析计算的模式，支撑大数据环境下的新型分析计算应用。当与共享存储层的标准化海量数据相结合后，能够迸发出大量创新性分析应用，大幅提升数据价值的挖掘能力。

2.3.4 模型库

目前中国石油的做法是将梦想云数据生态与中国石油认知计算平台（E8）相结合，加强建模工作流对接（数据加载、服务纳管）和模型库、模型管理功能的建设，为梦想云生态用户提供数据科学家工作环境，推动大众创新。

3 应用验证及效果分析

按照梦想云连环湖方案，在塔里木油田进行了试点验证，取得的主要认识为：通过连环湖架构的引入，实现了 3 个维度的解耦，在最大程度继承已有建设成果的基础上，能够改变油气田与集团总部数据治理的格局，建立共享的环境与氛围，提高平台化、微服务化的业务应用建设支撑能力，为智能化应用的快速建设奠定坚实的基础，有助于推动“共享中国石油”战略落地。

3.1 对原有格局产生的影响

数据资产管理流程升级，将过去传统的油气田中心主库提升为统一数据湖作为核心资产的存储环境；企业统一管理的数据类型增加，除结构化数据由集团总部统一标准管理外，增加其他类型数据存储标准要求，数据管理的边界扩大；数据治理工作力度将持续加大，随着云化应用功能的增加，对于高质量数据的需求更加具体，数据治理的工作量将持续加大。

3.2 新生的数据湖生态格局体现出巨大价值

（1）数据共享将更加便捷[9]。统一数据湖相关技术打通了数据共享技术壁垒，将带来更加便捷、高效的数据共享应用体验。

（2）数据质量将明显提升。由于引入了数据治理的多个层次，数据治理的责任主体将更加明确，同时通过业务中台、数据中台建设，进一步强化中国石油上游主数据管理，各专业数据一致性、完整性和准确性将明显提升。

（3）智能应用将更为易建。基于良好的数据基础、更加便捷的应用通道、更高效的数据应用支撑能力，引入大数据平台技术，将为人工智能应用打下坚实基础。

4 结语

梦想云在多年来的勘探开发信息化建设成果积累的基础上，建成了“两统一、一通用”的平台，突破了以往存在的“数据难以共享、业务难以协同”的瓶颈。在连环湖的集团主湖、油气田公司区域湖标准统一的基础上，解决了跨地域的数据入湖、大块数据调用的效率问题，同时有效地调动了油田主动性。

数字化转型之路任重而道远，勘探开发梦想云连环湖方案仅是围绕目前中国石油上游信息化发展现状和对未来发展的愿景进行了有益的探索与实践，最终对应用的支撑效果还需接受实践的深度检验。随着信息技术的不断进步，数据采集与汇聚、存储与管理、治理与应用的方式和方法将会不断地创新，梦想云统一数据湖建设将更快地适应时代的要求，更好地服务石油工业的高效、高质量发展。

参考文献

[1] Amazon. Angling for insights in today’s enterprise data lake [EB/OL]. [2020-05-19].https://s3-ap-southeast-1.amazonaws.com/mktg-apac/Big+Data+Refresh+Q4+Campaign/Aberdeen+Research+-+Angling+for+Insights+in+Today’s+Data+Lake.pdf.

[2] Thought Works. The enterprise data lake[EB/OL]. [2020-05-19].https://www.martinfowler.com/bliki/DataLake.html.

[3] Gorelik A. The enterprise big data lake[M]. Cambridge: O’Reilly Media, 2017.

[4] Carnell J. Spring microservice in action[M]. Beijing: People’s post and Telecommunications Press, 2018.

[5] Greenplum Database. MPP database architecture[EB/OL]. [2020-05-19].https://greenplum.org/gpdb-sandbox-tutorials/introduction-greenplum-database-architecture.html.

[6] Amazon. Amazon simple storage service: Getting started guide [EB/OL]. [2020-05-19].http://docs.aws.amazon.com/AmazonS3/latest/gsg/s3-gsg.pdf.

[7] Gormley C, Tong Z. Elasticsearch: the definitive guide [M]. Cambridge: O’Reilly Media, 2015.

[8] Apache Kylin. Mechanical industry authority guide [EB/OL]. [2020-05-19]. https://github.com/apache/kylin.html.

[9] 杜金虎，时付更，张仲宏，等 . 中国石油勘探开发梦想云研究与实践 [J]. 中国石油勘探，2020,25(1):58-66.
Du Jinhu, Shi Fugeng, Zhang Zhonghong, *et al*. Research and practice of Dream Cloud for exploration and development of PetroChina[J]. China Petroleum Exploration, 2020,25(1):58-66.