

知识湖存储

产品介绍

文档版本

01

发布日期

2025-06-19



版权所有 © 华为云计算技术有限公司 2025。保留一切权利。

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

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

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

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

目 录

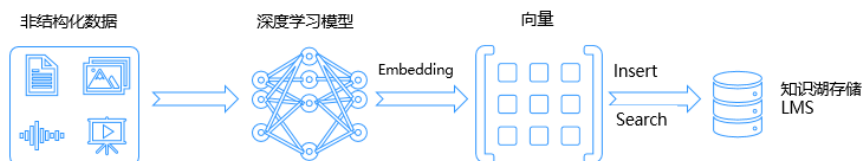
1 什么是知识湖存储.....	1
2 产品优势.....	2
3 应用场景.....	3
4 产品功能.....	4
5 计费说明.....	5

1 什么是知识湖存储

知识湖存储（Long-term Memory Service, LMS），是一款用于存储、检索、分析多维向量数据和标量数据的云原生服务。

知识湖存储服务支持多种索引类型和相似度计算方法，支持千亿向量规模和毫秒级查询时延。

知识湖存储不仅能为大模型提供外部知识库，提高大模型回答的准确性，还可广泛应用于推荐系统，自然语言处理，智能客服等AI领域。



访问方式

提供了Web化的服务管理平台，即管理控制台和基于HTTPS请求的API（Application programming interface）和SDK的管理方式。控制台

- 控制台方式
通过控制台可以创建知识仓实例。
- API方式
创建知识仓实例后，可以通过Restful API来创建或删除collection、插入向量、查询向量等操作。
- SDK方式
创建知识仓实例后，可以通过python SDK来创建或删除collection、插入向量、查询向量等操作。

2 产品优势

知识湖存储作为一种存储和检索向量数据的服务提供给用户，在高性能、高可用、大规模、简单易用、安全可信等方面体现出显著优势。

- **高性能：**知识湖存储支持千亿级向量数据规模及毫秒级查询延迟。
- **高可用：**知识湖存储提供多副本高可用特性，其多可用区和分布式多节点的架构，可用性可达99.99%，显著提高系统的可靠性和容错性，确保知识湖存储服务在面临节点故障和负载变化等挑战时仍能正常运行。
- **大规模：**知识湖存储是云原生的服务，架构支持水平扩展，可支持百亿向量规模，满足 AI 场景下的向量存储与检索需求。
- **简单易用：**支持丰富的向量检索能力。用户通过 HTTP API 或者 SDK 接口即可快速使用知识湖存储服务，开发效率高。
- **安全可信：**充分复用华为云对象存储的安全可信能力，实现安全隔离和全链路数据安全可信保障。

3 应用场景

知识湖存储可进行高性能向量存储和检索，主要适用于以下应用场景：

- 文本/图像检索：向量数据库对输入的图像和文本信息进行向量存储与检索，会找到最匹配输入信息的文本或图像结果。
- 大规模知识库：企业的私域数据存储在向量数据库中可构建外部知识库，帮助企业更好地管理和利用自己的数据资源。
- 问答系统：向量数据库会基于问题信息进行向量存储与检索，并返回最相关的问题与对应的答案。
- 推荐系统：向量数据库会基于用户特征进行向量存储与检索，最终筛选用户可能感兴趣的物品推荐给用户。

4 产品功能

表 4-1 知识湖存储功能

功能名称	功能描述
向量和标量存储	将数据以向量和标量的形式存储在对象存储服务OBS中
向量检索	计算目标向量与OBS中向量之间的相似度，获取TOP K个最相似的向量
标量检索	基于数据的标量字段进行精确匹配的检索
混合检索	稠密向量 Dense vector 和 稀疏向量 Sparse vector 混合检索

5 计费说明

在使用知识湖存储中的知识检索功能时，需要先创建知识仓实例，使用知识仓实例，会收取一定的费用。

计费说明

资费项	计费项	含义	适用的计费模式	计费公式
知识检索计费	知识仓实例费用	根据知识仓实例的CU规格、CU数量以及使用时长（小时）进行按需计费。 使用时长不足一小时的，以实际使用时长为准。	按需计费	CU规格单价 * CU数量 * 使用时长（小时）

计费示例

须知

以下示例中出现的费用价格仅供参考，实际价格请以控制台显示为准。

假设某用户于2025年5月27日14:00:00创建了一个知识仓实例，该知识仓实例的CU规格为2xlarge，CU数量为3个，假设2xlarge的单价为0.3元/个/小时，那么截止2025年5月27日18:36:00，该知识仓实例的计费如下：

2xlarge规格的计费单价 * CU数量 * 使用时长 = 0.3元/个/小时 * 3个 * 4.6小时 = 4.14元

说明

按需计费模式，使用时长不足一小时的根据实际使用时长计费，而36分钟换算为小时，为0.6小时，因此本示例中实际使用时长为4.6小时。