# 概述

PolarDB for AI功能通过一系列MLOps和内置的模型解决了数据、特征和模型的割裂状态，实现了基于数据库的数据智能的一站式服务。本文介绍了PolarDB for AI功能的相关特性。

# 背景信息

随着数据的累积，数据驱动的智能应用（例如：搜索、推荐和问答）已经越来越广泛地被用于各种场景中。在过去几年，数据驱动的智能应用已经逐渐从基于简单的数据分析和统计转型为基于机器学习和深度学习的特征和模型应用。转型的速度远远超出预期，部分原因是机器学习和人工智能方面的技术突破。深度神经网络在图像分析和自然语言处理等学科中取得了前所未有的进步。强化学习成为补充传统监督学习的强大范式。

然而，在目前的数据驱动的智能应用中，数据、特征和模型仍处于割裂状态。首先，数据工程师通过手工编写流程进行数据清洗和数据集成；然后，算法工程师通过自定义的特征工程流程、模型训练脚本以及定时任务脚本进行周期性的生产特征和模型；最后，开发工程师负责模型的上线、稳定性保证和监控运维。这就导致了不同系统之间会进行数据迁移，同一份数据可能在不同源之间冗余，进而出现数据不一致的问题，以及特征难管理、模型难升级等困境。除此之外，数据工程师、算法工程师和开发工程师的人力成本也是当前数据驱动的智能决策应用难以大规模落地的一个阻碍。

事实上，鉴于数据库研究在数据发现、数据管理、版本控制、数据清理和数据集成方面积累的专业知识，PolarDB for AI可以为数据驱动的智能应用提供一站式的数据（包括：数据、特征和模型）服务来解决这种割裂状态，大大减少数据驱动的智能决策开发过程中的人力成本，从而走出当前的困境。

# 简介

PolarDB for AI是基于PolarDB MySQL版的一个数据库内的分布式机器学习组件。其基于云原生的体系架构，通过SQL语句的方式提供了支持机器学习的一系列MLOps，包括：创建模型、训练模型、查看模型状态、查看模型列表、模型评估和模型推理等能力；同时提供了一系列内置的机器学习和人工智能算法，包括：分类算法、回归算法和聚类算法等。基于MLOps和内置的模型，PolarDB for AI为数据驱动的智能应用提供了高效、可靠、方便的数据智能能力，打破了数据库和业务应用之间的系统墙，提供了基于数据库数据智能的一站式服务。

# 版本要求

若要使用PolarDB for AI功能，PolarDB MySQL版集群需满足如下要求：

* 产品版本为**企业版**，系列为**集群版**。
* 内核引擎版本需为8.0.1及以上。
* 数据库代理版本（Proxy）需为2.7.5及以上。

**说明**

关于如何查看或升级内核引擎版本和数据库代理版本，请参见[查看并升级版本](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/upgrade-the-cluster-version%22%20%5Co%20%22)。

# 费用

PolarDB for AI只收取AI节点的费用。普通AI节点按照普通的计算节点收费，其规格详情与价格请参见[计算节点规格](https://help.aliyun.com/zh/polardb/polardb-for-mysql/specifications-of-compute-nodes%22%20%5Co%20%22)和[计算节点价格](https://help.aliyun.com/zh/polardb/polardb-for-mysql/billing-rules-of-compute-nodes/%22%20%5Co%20%22)。

除了普通的计算节点规格之外，AI节点还支持两个GPU规格，其规格详情与价格见下表。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **节点规格码** | **CPU和内存** | **地域** | **包年包月价格** | **按量付费价格** |
| polar.mysql.g8.2xlarge.gpu | 8核30 GB + 一张GU30**说明**GU30的算力与NVIDIA A10性能相当。 | 中国内地 | 6200元/月 | 13元/小时 |
| 美国（硅谷） | 9610元/月 | 20.15元/小时 |
| 印度尼西亚（雅加达） | 10850元/月 | 22.75元/小时 |
| 德国（法兰克福） | 10540元/月 | 22.10元/小时 |
| 印度（孟买）关停中 | 9300元/月 | 19.50元/小时 |
| 美国（弗吉尼亚） | 8060元/月 | 16.90元/小时 |
| 日本（东京） | 9920元/月 | 20.80元/小时 |
| 新加坡 | 11780元/月 | 24.70元/小时 |
| 马来西亚（吉隆坡） | 11780元/月 | 24.70元/小时 |
| 中国（香港） | 11780元/月 | 24.70元/小时 |
| 英国（伦敦） | 11780元/月 | 24.70元/小时 |
| 澳大利亚（悉尼） | 11780元/月 | 24.70元/小时 |
| polar.mysql.x8.2xlarge.gpu | 16核125 GB + 一张GU100 | 中国内地 | 21700元/月 | 45.2元/小时 |
| 美国（硅谷） | 33635元/月 | 70.06元/小时 |
| 印度尼西亚（雅加达） | 37975元/月 | 79.10元/小时 |
| 德国（法兰克福） | 36890元/月 | 76.84元/小时 |
| 印度（孟买）关停中 | 32550元/月 | 67.80元/小时 |
| 美国（弗吉尼亚） | 28210元/月 | 58.76元/小时 |
| 日本（东京） | 34720元/月 | 72.32元/小时 |
| 新加坡 | 41230元/月 | 85.88元/小时 |
| 马来西亚（吉隆坡） | 41230元/月 | 85.88元/小时 |
| 中国（香港） | 41230元/月 | 85.88元/小时 |
| 英国（伦敦） | 41230元/月 | 85.88元/小时 |
| 澳大利亚（悉尼） | 41230元/月 | 85.88元/小时 |

**说明**

polar.mysql.g8.2xlarge.gpu与polar.mysql.x8.2xlarge.gpu这两个规格主要应用于AI模型的训练和推理。

# 技术架构

PolarDB for AI通过数据库代理进行SQL路由，含有/\*polar4ai\*/的SQL语句会被分发到具有AI能力的集群上进行计算，计算的返回结果和普通的SQL返回结果遵循同样的协议。所以，您可以[连接数据库集群](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/connect-to-a-cluster%22%20%5Cl%20%22task-1580301%22%20%5Co%20%22)并执行带有/\*polar4ai\*/的SQL语句来使用该功能。其架构图如下所示：

**说明**

PolarDB数据库代理版本（Proxy）需为2.7.5及以上，如何查看和升级当前数据库代理版本，请参见[小版本升级](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/upgrade-the-cluster-version%22%20%5Cl%20%22task-2449714%22%20%5Co%20%22)。

除普通集群具有的存储功能和计算功能之外，PolarDB for AI还具有三层结构，分别是接入层、特征层及模型层。其中：

* 接入层：负责处理SQL语句，包括SQL解析、SQL校验、代价估算和生成执行计划等，并优化为可执行的SQL语句节点树。
* 特征层：负责处理从数据转换为特征的工作，包括数据接入、特征生成、数据同步和特征更新。
* 模型层：负责处理模型相关的MLOps能力，包括模型创建、模型训练、模型评估、模型推理和模型管理。

# 适用场景

如下图所示，PolarDB for AI将传统的DB+AI的数据智能应用转变为一站式数据智能应用，从数据库服务于传统的数据工程师和算法工程师转变为DB+AI服务于业务工程师，直接与业务系统对接。



PolarDB for AI可用于如下行业场景：

* **ID-Mapping**

ID-Mapping通常适用于平台性的客户，如游戏和电商等。以游戏平台为例，某些游戏平台中的同一个用户可能会有多个游戏账号，这些账号散落在各种游戏或平台本身的数据库中，并没有对账号进行融合。基础数据建设不完善会带来各种问题，例如做精准营销时不能精准定位同一个用户，导致流量在跨渠道使用时效率较低。采用机器学习模型则会将账号统一，不但提高了基础数据质量，而且还为上游业务提供了优质的数据。

* **问答机器人**

问答机器人基于数据库中的内容。根据用户的业务场景，结合AI能力（对话控制、机器学习、自然语言理解等），打造适合企业的对话服务。问答机器人可以实现7×24小时在线服务，能帮助企业接待更多客户、提升客户满意度、提高工作效率和降低运营成本。是企业进行在线咨询、在线营销和在线服务的好帮手。

* **搜索推荐**

在传统的数据库中，用户的搜索能力通常基于数据库固有的全文检索能力，不支持自然语言类的检索需求（如语义检索、同义词匹配等）。采用PolarDB for AI中成熟的搜索解决方案，可以大幅度提升搜索的精确性。

基于PolarDB for AI的智能推荐算法和知识图谱技术，再结合阿里巴巴电商策略，为企业提供贯穿推荐能力的一站式服务，助力企业快速过渡冷启动过程。面向不同的业务场景定制个性化解决方案，持续提升核心业务能力，以实现业务营收增长。

# 优势

PolarDB for AI支持PolarDB内置的模型训练和模型推理等MLOps，无缝兼容MySQL语句，提供达摩院加持的行业智能算法服务，避免数据在两个甚至多个系统间转换和迁移，从而降低数据智能应用的开发成本，加速数据价值业务变现。

# 使用说明

* 快速上手PolarDB for AI
	+ [开启PolarDB for AI功能](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/enable-the-polardb-for-ai-feature)
	+ [使用集群地址连接集群并执行AI SQL](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/connection-mode)
	+ [模型使用流程及说明](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/model-management)
* [自然语言到SQL语言转义（基于大语言模型的NL2SQL）](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/llm-based-nl2sql)
* [分词和向量检索](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/vector-retrieval/)
* 进阶使用PolarDB for AI
	+ [特征管理](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/feature-management)
	+ [算法介绍](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/algorithms/)
	+ [将AI模型推理结果写回数据库](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/bring-ai-generated-content-to-a-database-table-through-aat)
	+ [Row-oriented AI](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/row-oriented-ai/)
* 典型解决方案
	+ [ID-Mapping在游戏领域的解决方案](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/best-practices-for-id-mapping-in-gaming)
	+ [通义千问大模型数据推理和交互](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/qwen-big-model-data-reasoning-and-interaction)

# 核心算法

* [LightGBM算法](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/lightgbm-algorithm)
* [DeepFM算法](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/deepfm-algorithm)
* [K均值聚类算法（K-Means）](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/k-means-clustering-algorithm)
* [随机森林回归算法（Random Forest Regression）](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/random-forest-regression-algorithm)
* [梯度提升回归树算法（GBRT）](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/gbrt-algorithm)
* [梯度提升决策树算法（GBDT）](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/decision-tree-algorithm-gbdt)
* [线性回归算法（Linear Regression）](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/lr-algorithm)
* [支持向量回归算法（SVR）](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/svr-algorithm)
* [IV算法](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/information-value)
* [Pearson相关系数](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/pearson-correlation-coefficient)
* [群体稳定性指标PSI](https://help.aliyun.com/zh/polardb/polardb-for-mysql/user-guide/population-stability-index)