FROST & SULLIVAN
沙利文

头豹
LeadLeo

2022年中国AI开发平台市场报告

AI开发平台/AI模型/自动化机器学习/低代码AI开发

2022年11月

弗若斯特沙利文咨询（中国）
头豹研究院

报告说明

沙利文联合头豹发布《 2022年中国AI开发平台市场报告》，该市场报告以中国AI开发平台市场为核心研究对象，研究周期为2022年（数据截至2022年11月25日）。本报告旨在分析中国AI开发平台的概念定义、应用前景、技术动向及发展趋势，并识别AI开发平台市场竞争态势，反映该细分市场领袖梯队品牌的差异化竞争优势。

沙利文联合头豹研究院对AI开发平台市场进行了下游用户体验调查。受访者来自互联网、政务、金融等多个领域，所在组织规模不一，细分领域有别。

本市场报告提供的AI开发平台趋势分析亦反映出行业整体的动向。报告最终对领袖梯队的判断仅适用于本年度中国发展周期。

本报告所有图、表、文字中的数据均源自弗若斯特沙利文咨询（中国）及头豹研究院调查，数据均采用四舍五入，小数计一位。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

研究框架

章节一　AI平台架构

沙利文在本章从基础设施、框架工具以及训练平台三大维度着手，围绕AI开发平台的架构开展分析。

1.1 AI 基础设施

- 人工智能开发平台是一个整合了AI算法、算力和开发工具的平台，为机器学习、深度学习、训练模型等开放了开发架构。它还提供开发所需的算力支持，使开发者能够有效地利用平台中的人工智能能力，通过接口调用进行人工智能产品开发或AI赋能。
- 人工智能开放平台为开发者提供了许多开发工具和框架，有助于降低开发成本，如人工智能数据集、AI模型和算力成本。开发者可以使用该平台的数据集来训练他们自己的模型，或者使用该平台的算法框架来进行功能定制。
- 人工智能开发平台的架构从下到上可以分为基础设施、框架、训练平台和技术服务四层。

1. 基础设施。自主研发的人工智能芯片是企业的核心竞争力，其可以呈现出人工智能架构创新、形态演变、软硬件融合三大趋势。

1.1 底层硬件

主流的人工智能处理器本质上是一个片上系统（SoC），可用于图像、视频、语音和文字处理等相关场景。人工智能处理器的主要架构组件包括一个专门设计的计算单元、一个大容量的存储单元和一个相应的控制单元。通过自主研发的人工智能芯片，企业可以根据自己的算法要求调整芯片线，从而最大限度地提高计算效率，自研的人工智能芯片将逐渐成为AI开发平台厂商的核心竞争力之一。

1.1.1 AI芯片架构创新

云端AI芯片主要用于AI训练场景，计算能力是其核心指标之一。为了适应AI训练中需要用到的应用和算法，供应商需要开发特定领域架构（DSA）芯片来进行架构创新，从而实现性能优化。计算单元作为DSA芯片的计算、存储以及控制三大部件之一，可以进行标量、矢量和矩阵运算。华为对达芬奇架构中的矩阵运算进行了深度优化，并定制了相应的矩阵计算单元，以支持高吞吐量的处理，这样就可以用一条指令完成两个16*16矩阵的乘法运算。

为了解决现有内存访问速度严重滞后于处理器计算速度的问题，新型完全可编程、可重构架构（CGRA）芯片、内存计算芯片以及具有高内存带宽的新型处理器架构IPU可能会引入AI芯片底层生态。

此外，芯片编程方法和软件架构设计也将成为AI芯片创新的重要组成部分。例如，英伟达凭借其CUDA框架大大降低了GPU的编程难度，使得GPU在AI加速领域得到了广泛的应用。未来，更多的人工智能处理器将提供多层软件栈和开发工具链，帮助开发者更有效地利用底层硬件资源，不断提高开发效率，并通过多种软件提高专用芯片的灵活性。

AI开发平台架构

AI技术服务	基于语音识别、图像识别、文字识别、自然语言处理的服务；AI生态的构建
AI训练平台	机器学习平台
AI框架工具	以 TensorFlow 和 PyTorch 为代表的机器学习框架
AI基础设施	包括基础硬件（CPU、GPU、FPGA、NPU、ASIC以及其组合成的SoC）；计算引擎、云容器

1.1.2 AI芯片形态演进

人工智能芯片创新的目标之一是保持芯片的高能效比，同时适应人工智能算法的发展。未来，通用加专用芯片的片上系统形式将成为主流（CPU+NPU、CPU+ASIC等），应用范围更广。

传统的处理器指令集（包括x86和ARM等）是为通用计算而演变的，其基本操作是算术操作（加减乘除）和逻辑操作（有无），在深度学习中完成一个神经元的处理往往需要上百条指令，深度学习的处理效率并不高。为了解决这个次痛点，芯片形式需要打破传统的冯-诺依曼结构。神经网络处理器NPU使用电路来模拟人类的神经元和突触结构。在NPU中，存储和处理被集成在神经网络中，由突触权重反映。例如，寒武纪提出的全球首个深度学习处理器指令集DianNaoYu可以直接面对大规模神经元和突触的处理，通过一条指令就可以完成一组神经元的处理，并为神经元和突触数据在芯片上的传输提供一系列专门支持。在AI训练加速应用方面，寒武纪还推出了最新的MLU370-X8训练加速卡，搭载双芯片四核粒子思源370，在YOLOv3、Transformer、BERT和ResNet101任务中，8张卡并行的平均性能达到350W RTX GPU的155%。

来源：弗若斯特沙利文，头豹研究院

1.1.3 AI芯片软硬一体化

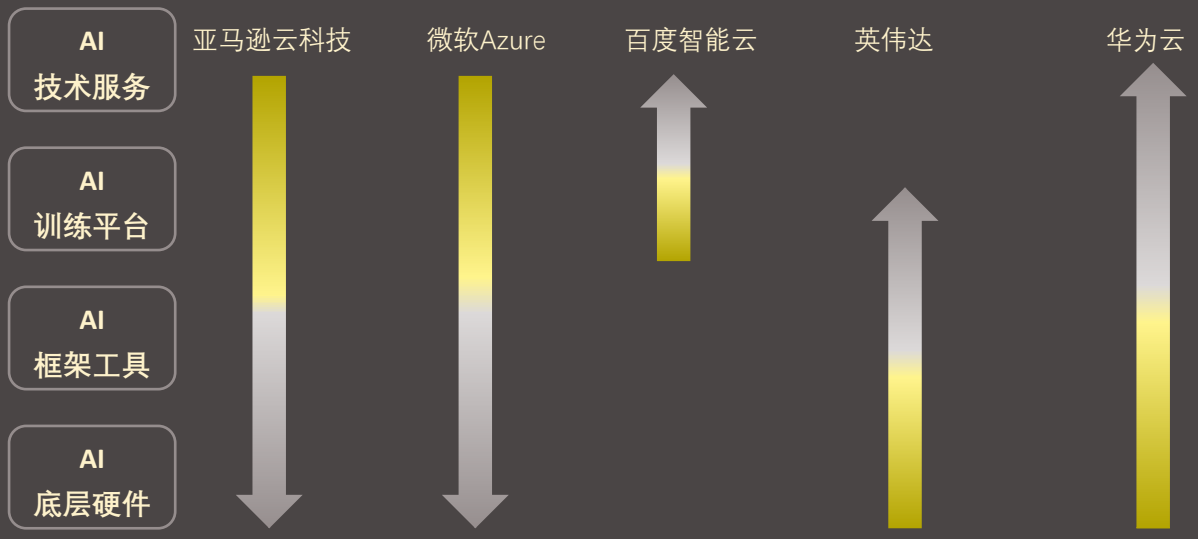
- 围绕AI芯片的软件工具开始从基础计算向场景计算转变。过去，以NVIDIA为代表的芯片企业不断构建以CUDA编程模型为核心，由例如高性能算子库、通信算法、推理加速引擎等多层次基础软件工具组成的生态。
- 在这个阶段，人工智能芯片企业开始构建差异化场景的软硬件一体化平台。商业模式从提供硬件支持服务扩展到提供技术生产工具和技术服务。实现了底层芯片、编程框架、行业算法库、细分场景研发平台的全栈式高效整合，从而培育出多元化行业场景的计算生态圈，并抢占细分市场。
- 同时，企业也可根据客户需求提供模块化服务，进一步提升服务的客制化程度。

1.1.4 AI芯片向移动端拓展

1. 目前传统的AI芯片为了满足计算性能不断扩张的要求，在采用新结构、新工艺的同时，也使得芯片的基础功耗节节攀升。过高的芯片功耗对于芯片使用者和使用场景的供电能力、散热能力等都带来了较高的挑战。为了满足大规模AI芯片在传统电力资源和散热资源上的需求，芯片使用者们不得不增加相关硬件的基础投入，这极大地增加了AI芯片的使用成本，同时也在无形之中抬高了整个AI行业的入门门槛。AI芯片高功耗的情况对于目前主流的固定端应用相对来看挑战较低，使用者们可以通过简单的增加配套硬件设备数量来解决能源功耗问题；但对于移动端来说，AI芯片的功耗是一个不易解决的问题。手机、笔记本电脑、可穿戴设备、汽车自动驾驶等移动端应用受限于产品可用空间的大小、电池等储能设备能力的限制等客观因素，无法上马超高功耗的高性能AI芯片。因此，针对移动端的AI芯片设计将是未来芯片研发的一个主流方向。

2. 在移动端AI芯片上，以三星电子和SK Hynix为代表的韩国厂商走在前列。其中三星电子在2022年ISSCC大会上发表的关于移动端NPU芯片的研究成果显示其最新移动端NPU在优化芯片数据流提升计算单元利用率、优化计算单元以覆盖不同的计算精度、提供不同的工作模式以满足不同功耗和性能的需求上均实现了较大突破。在计算精度方面，三星电子的移动端NPU可以满足INT4、INT8和FP16精度要求，基本覆盖移动端人工智能算法所有需求；而其在模式切换上的突破也极大的解决了手机芯片在日常使用中的痛点。目前三星电子移动端NPU已经在其4nm Exynos SoC中得到了应用。

AI开发平台企业全栈整合



来源：弗若斯特沙利文，头豹研究院

1.1.5 云端原生启用AI基础设施

云原生技术使企业能够在新的动态环境（如公共云、私有云和混合云）中构建和运行弹性和可扩展的应用程序。具有代表性的云原生技术包括容器、服务网格、微服务、不可变的基础设施和声明式的API，这些技术能够构建松散耦合的系统，具有容错性，易于管理，易于观察。与可靠的自动化手段相结合，云原生技术使工程师可以很容易地对系统进行频繁和可预测的重大改变。

1. 在基础设施层面，容器在云基础设施和应用之间解耦应用和技术架构资源。
2. 在应用层面，用户可以根据不同场景选择微服务或无服务器架构。
3. 在复杂的架构场景中，服务组件的通信是通过服务网格控制的。
4. 最后，通过DevOps对系统进行持续迭代和更新。

- 基础设施的完善

云原生容器架基于云原生的深度学习训练平台可以完全容器化部署，基于Kubernetes（K8s），为不同任务提供弹性灵活的资源扩展、资源调度和分配，并向后兼容多种CPU和GPU处理器。因此，基于云原生的人工智能开发平台可以快速适应适当的云原生资源，既可以进行大规模稀疏数据的训练，又可以进行基于感知的场景训练。例如，阿里云PAI可以提供支持近线性加速的内核，让训练任务在多个引擎上实现性能提升和性能加速。

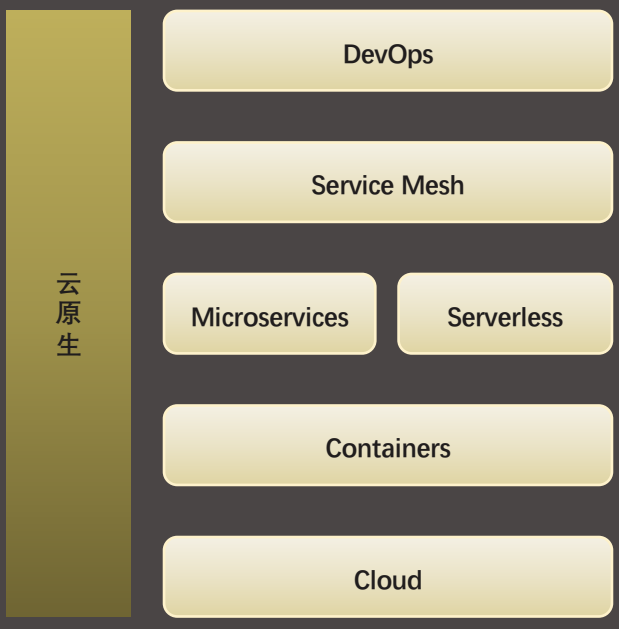
- 训练环节的提升

构可以灵活部署ML训练的计算资源，通过灵活的训练为AI开发降低成本，提高效率。人工智能开发平台的云端可以实时监控资源池的计算能力，在训练中出现闲置计算资源时，将闲置资源分配给任务，提高任务的计算能力，使训练作业能够快速收敛。任务提交后，弹性训练方案还可以根据资源池中空闲资源的使用情况和弹性工作情况，将回收的资源分配给新的ML训练任务，从而保证新的ML训练的计算能力。

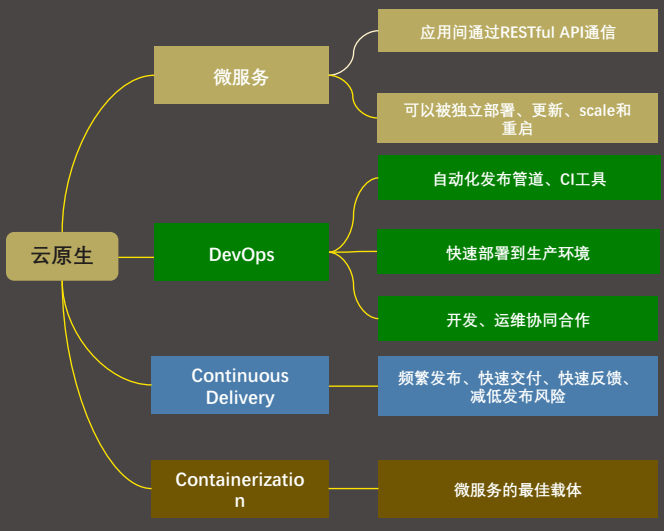
- 用户体验的提升：

云原生应用可以为人工智能开发平台用户（开发者）提供更敏捷、更高质量的应用交付，以及更简单、更高效的应用管理，并提供更快的业务需求响应和更好的用户体验。基于云原生的人工智能平台完全适应团队的人工智能在线协作开发、在线人工智能教学和本地人工智能研发向云端迁移。

云原生概念闭环



云原生：DevOps+持续交付+微服务+容器



来源：弗若斯特沙利文，头豹研究院

1.2.2 深度学习框架市场情况

1.2.2.1主要市场份额占领

- ◆ 全球深度学习框架超90%的份额由谷歌开发的TensorFlow与Meta开发的Pytorch占领：
  - TensorFlow是现阶段最热门的深度学习框架，拥有可视化、性能强悍、多用途等特点。TensorFlow自带tensorboard可视化工具，能够让用户实时监控观察训练过程，同时支持多GPU、分布式训练，跨平台运行能力强。TensorFlow具备不局限于深度学习的多种用途，还拥有支持强化学习和其他算法的工具.
  - PyTorch由脸书开源，具备简洁易用、细节化等特征。PyTorch具备更少的抽象，更直观的设计，建模过程简单透明，所思即所得，代码易于理解，同时可为使用者提供更多关于深度学习实现的细节，如反向传播和其他训练过程等。PyTorch拥有更为活跃的社区，可为开发者提供完整的文档和指南，供用户交流和求教问题，但与Tensorflow的社区相比规模更小。
  - 其他的典型框架还包括Keras（由Google工程师开源）、mxnet（由亚马逊开源）、PP飞桨（由百度开源）、theano（有蒙特利尔大学开源）、CNTK（由微软开源）。
- ◆ 中国的典型框架还包括Keras（由Google工程师开源）、mxnet（由亚马逊开源）、PP飞桨（由百度开源）、theano（有蒙特利尔大学开源）、CNTK（由微软开源）。

深度学习软件框架概况

分类	部署位置		承担任务		基本流程			
具体类型	云端框架	终端框架	训练框架	推理框架	底层运算框架	模型搭建框架	迭代训练框架	跨界框架
定义	即在数据中心完成深度学习的框架	即可以在手机、安防摄像头、汽车、智能家居设备、各种IoT设备等执行边缘计算的终端设备上运行的框架	主要通过输入数据或采取增强学习等非监督学习方法完成深度学习模型训练	主要完成训练模型的优化、部署和推断计算	专注于深度学习基本开发流程中的底层基本运算环节	提供基本模块支持深度学习模型创建，但本身不能接触底层运算模块	提供基本模块，支持深度学习模型训练中的流程优化，但本身不能接触底层运算模块	既能让用户接触底层数据模块，又可以提供现成的基本模块以实现快速建模
关键技术要求	算力、安全以及稳定性	安全、稳定性	生态建设、易用性、性能、支持架构	易用性、性能、底层优化、安全稳定性	运算效率、数据精度、算法设计	模型处理、问题解决		除满足对其他基本流程技术要求，还应满足兼容性、安全性、易用性
案例	TensorFlow PyTorch 飞桨	Tencent TensorFlow Caffe2go 飞桨	TensorFlow PyTorch 飞桨	TensorFlow OpenCV 飞桨	theano TensorFlow PyTorch 飞桨	Caffe mxnet PyTorch 飞桨	Keras TensorFlow PyTorch 飞桨	TensorFlow PyTorch 飞桨

来源：弗若斯特沙利文，头豹研究院

1.2.2.2竞争格局

目前主流深度学习软件框架格局逐步清晰，已从百花齐放向几家逐鹿转变。

- **早期热点退出历史舞台：**微软CNTK、日本初创企业首选网络（preferred networks）Chainer、加拿大蒙特利尔大学主导的Theano等早期热点框架已通过与主流框架合并或直接停止更新的方式退出历史舞台。
- **当下主流格局：**谷歌开发的TensorFlow依托工业界的部署优势持续位于第一，市场关注度第二名PyTorch 3倍以上。Meta的PyTorch（合并Caffe2）凭借其易用性迅速突起，应用数量大幅提升，在各大顶级学术会议论文中占比超过50%，百度推出中国首个开源框架飞桨PaddlePaddle兼具效率与灵活性，有赶超势
- **“开源+巨头支持”成为深度学习软件框架的主流模式：**各平台在稳定性、调试难度、执行速度、内存占用等方面各有所长，主流框架普遍由行业头部企业介入并主导内部应用和搭建。

1.2.2.3竞争焦点

深度学习软件框架竞争焦点已从模型库转移至易用性和硬件适配优化。高级语言接口与硬件适配优化成为开源框架构筑壁垒的关键。

- 一方面，高级语言接口封装后端框架中关键的模型构建、训练等功能，降低研发门槛。目前，三大主流框架加速绑定或构建高级语言接口，已出现合作圈地现象。TensorFlow与Keras形成排他性合作，提升框架易用竞争力，与近期以易用性为优势快速提升地位的PyTorch抗衡；MXNet与Gluon联合，由亚马逊与微软共同维护；PyTorch以Torch和Caffe2作为后端框架，内部先天构筑高级语言接口；百度飞桨PaddlePaddle则具备“动静统一”的编程模式，兼顾灵活性和效率。
- 另一方面，硬件适配优化试图解决多样硬件编译工具导致的适配复杂和性能参差不齐问题，统一编译工具与编译语言成为主流开源开发框架的布局重点。目前，谷歌、脸书加速构建统一的编译语言（IR），试图引导硬件厂商主动适配，获取框架适配的话语权。

深度学习软件框架竞争焦点

来源：弗若斯特沙利文，头豹研究院

1.3 AI训练平台

1.3.1 资源配置

- 根据对实际数据的拟合，AI计算量每年至少增长10倍，速度远超远超摩尔定律的18个月两倍，因此深度学习训练中调整任务资源的能力变得尤为重要。
- 现阶段，随着集群规模的扩大，集群中给定时刻出现机器故障的概率在增加。且随着训练模型复杂度的提升，训练资源与训练时间均显著增长，任务的容错性在下降。此外集群规模的提升让空闲资源的浪费变得不可忽视，集群资源配置的灵活性需求不断。

1.3.2平台功能

- 管理多台训练服务器，尤其是带有GPU的高性能计算服务器，可把训练任务分到到分布式的计算节点上执行计算；
- 集成多种训练框架，抽象训练过程，提供Web界面，上传和指定相关数据和参数，即可启动训练任务并监控和分析训练过程；
- 池化计算资源，尤其是GPU资源，做成“GPU云”。启动训练任务时，平台会自动把训练任务分配到合适的GPU上；
- 打通数据中心，可以直接把数据存储平台中的数据导入到训练节点上中；
- 隔离计算节点中的资源和环境，兼容不同型号的GPU、不同版本的CUDA/CuDNN和不同的深度学习框架。

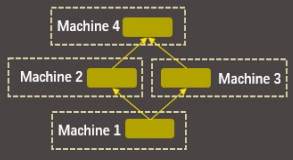
1.3.3 分布式训练

1.3.3.1分布式训练原理

- 分布式训练可提供底层资源的弹性配置，提升系统的资源利用率。例如，百度飞桨通用异构参数服务器可以对任务进行切分，让用户可以在硬件异构集群中部署分布式训练任务，实现对不同算力的芯片高效利用，为用户提供更高吞吐，更低资源消耗的训练能力。但分布式训练的应用也存在较大阻碍。在各个框架上实现弹性控制的模块，以及进行对应调度系统的适配来实现弹性训练需要极大的工作量。此外，如果不同的框架都拥有各自的弹性训练方案，在AI开发平台层面整合不同的框架方案也需要投入很高的维护成本。
- 弹性分布式训练是AI开发平台服务的趋势，可以为用户实现降本增效的体验：当用户需要大量运算资源时扩容，提升算力和稳定性，降低模型训练时间；当用户需求量小时，降低底层资源配置，为客户降低因资源占用而产生的服务费用。

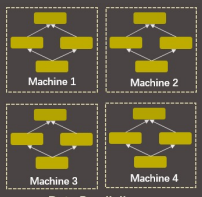
1.3.3.2 分布式训练框架模式

- 模型并行：
  - 将一个模型分拆成多个小模型，分别放在不同的设备上，每个设备跑模型的一部分，由于模型各个部分关系很大，这种方式效率很低，需要不同设备模型之间的频繁通信，一般不使用。



Model Parallelism

- 数据并行：
  - 完整的模型在每个机器上都有，但是把数据分成多份给每个模型，每个模型输入不同的数据进行训练，数据并行是目前最常用的分布式实现方法。



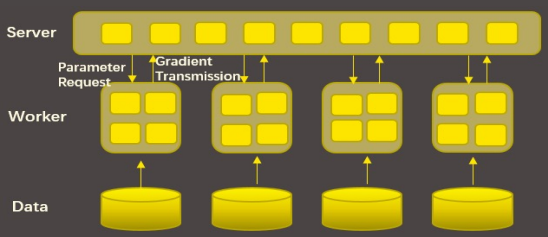
Data Parallelism

来源：弗若斯特沙利文，头豹研究院

1.3.3.3 分布式训练架构

□ 参数服务器架构：

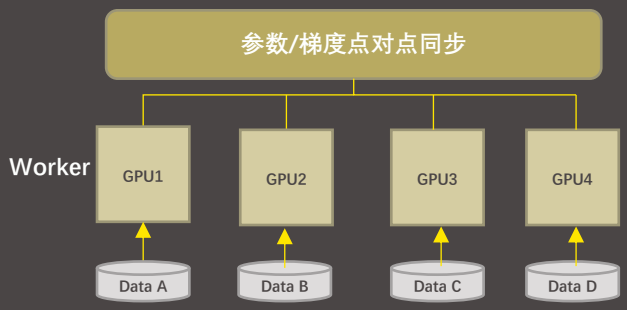
- 多被用于搜索推荐场景中大规模稀疏特征模型的训练任务。中心化架构，
- 该架构采用将模型参数进行中心化管理的方式实现模型参数的更新和分发。参数服务器架构有两个角色Server与Worker，Server与Worker不一定对应实际的硬件，Server负责参数的分片存储与更新，Worker则会保存有完整的模型网络结构，用于执行模型的前向与反向计算。常规的参数服务器的Worker节点，需要使用统一型号的CPU或GPU机器完成模型训练。



Parameter Server Structure

□ Collective架构：

- 多被用于视觉、自然语言处理等需要复杂网络计算的模型训练任务场景；
- 该架构为去中心化，也是近年来非常流行的分布式训练架构。该架构没有所谓管理模型参数的中心节点，每个设备都是Worker，这个Worker同样是进程的概念。每个Worker负责模型的训练同时还需要掌握当前最新的全局信息。



Collective训练

1.3.3.4 常见分布式训练框架

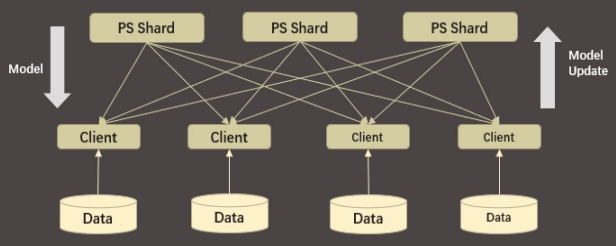
□ Tensorflow-Parameter Server架构

- Parameter Sever架构主要包括1到多个server节点和多个worker节点。其中server节点保存模型参数，worker负责使用sever上的参数以及本worker上的数据计算梯度。

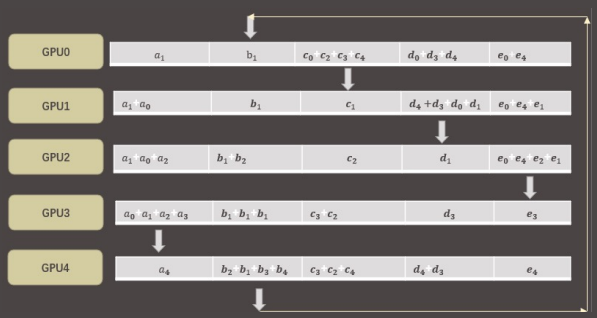
□ Pytorch- Ring AllReduce架构

- 这种架构的运行效率随着工人数量的增加而线性增加。
- 这种架构采用环形结构，在这种架构中没有中央节点服务器，只存在工作节点。
- 在这个架构中，每个工作者都有一个完整的模型参数副本，并进行梯度计算和更新。

Parameter Serve架构运作原理图



Ring AllReduce架构运作原理图



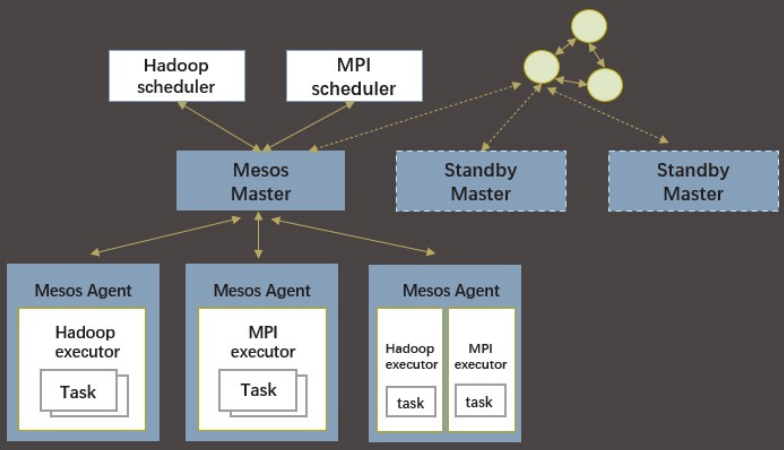
来源：弗若斯特沙利文，头豹研究院

1.3.4关键技术及框架

□ Mesos+Marathon

Mesos是Apache下的开源分布式资源管理框架，它被称为是分布式系统的内核，又被称作是数据中心的操作系统。简单的讲，Mesos实现了一个资源管理的框架，它在数据中心层次上管理集群资源（CPU、GPU、RAM等），提供资源分配和任务调度的能力。为了进一步把资源和任务隔离，Mesos把具体的任务调度能力抽象出来交给具体的Framework来实现，比如Hadoop，Spark，MPI和Marathon等。

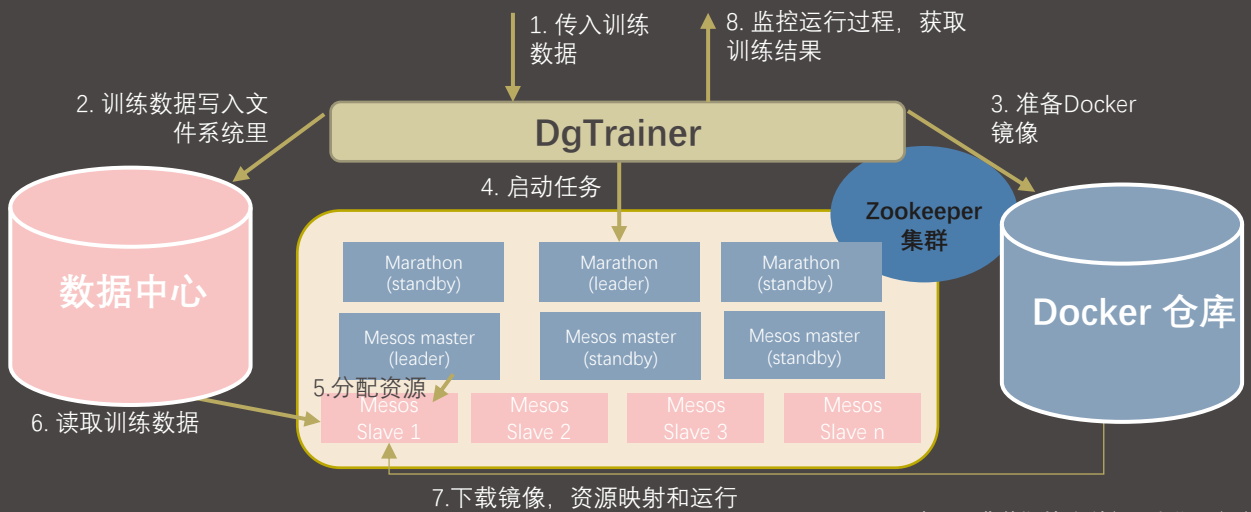
Mesos运作原理图



□ Docker

- 环境隔离：Docker隔离了系统环境和执行环境，即隔离同一台服务器上不同训练任务的环境，实现把同一个任务分发到带有不同型号GPU卡的服务器上，也可以在同一台服务器上同时运行不同CUDA版本、不同深度学习框架的多个任务。
- 资源隔离：Dockers能隔离硬件资源，在满足任务要求的同时，避免多任务对资源的恶性和无序竞争。但Docker对GPU资源的管理没有像CPU那么完善和成熟。
- 代码共享：通过Github + CI + Docker，实现把不同repo和branch的代码打包成完成不同任务的Docker镜像从而实现了更灵活和细粒度的共享。

训练框架流程图

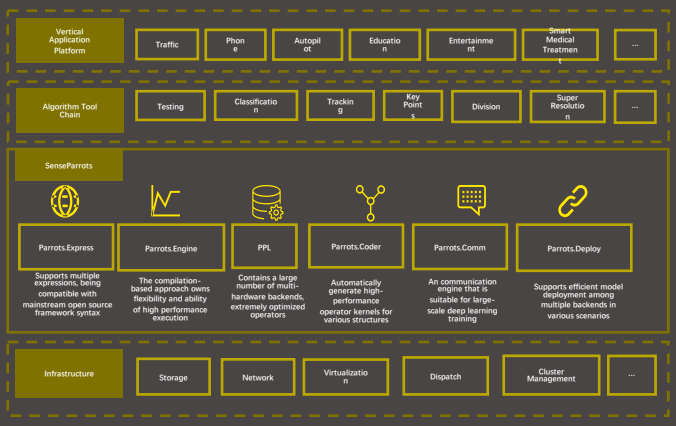


来源：弗若斯特沙利文，头豹研究院

◘ 商汤SenseParrots

- senseparrot不同于Facebook和谷歌的开源培训平台。具有超级计算网络训练、超大规模数据集训练、超大规模端到端复杂应用能力训练。
- senseparrot对计算能力的贡献是在现有的GPU上提供相应的软件系统，并构建相应的系统，使GPU的有效性得到充分发挥，整体研发效率得到极大提高。同样规模的培训，几年前需要几个小时才能完成，而在商汤平台上，90秒就能完成。

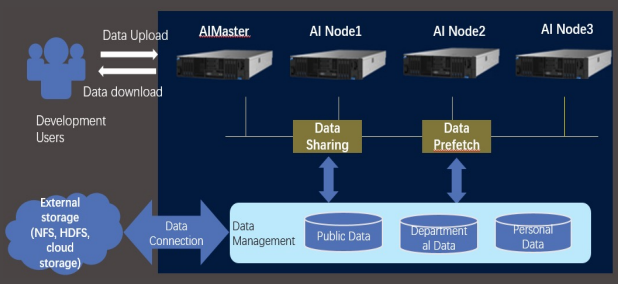
SenseParrots产品介绍图



◘ 浪潮AIStation 训练平台

- AIStation是浪潮面向人工智能企业训练场景的人工智能开发资源平台。
- 实现集体化部署、可视化开发、集中管理等，为用户提供极其高性能的AI计算资源，实现高效的计算支持、精准的资源管理和调度、敏捷的数据集成和加速、面向流程的AI场景和业务集成，有效开放开发环境、计算资源和数据资源，提高开发效率。

AIStation产品介绍图



1.3.5 算法升级

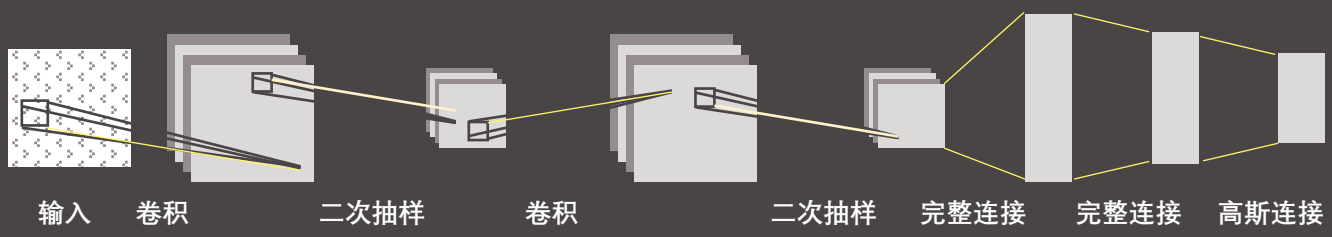
算法是AI与大数据的关联节点。社交媒介、定位技术、搜索引擎等互联网应用实时生成和储存着大量数据。在海量数据的基础上，AI持续对用户的兴趣偏好和需求进行推断，生成不同的用户画像，实现数字文化从生产、传播到接受的全程个性化、精准化定制。

现阶段，AI训练平台已或将集成多种人工智能技术，如计算机视觉、自然语言处理、跨媒体分析推理、智适应学习、群体智能、自主无人系统以及脑机接口等：

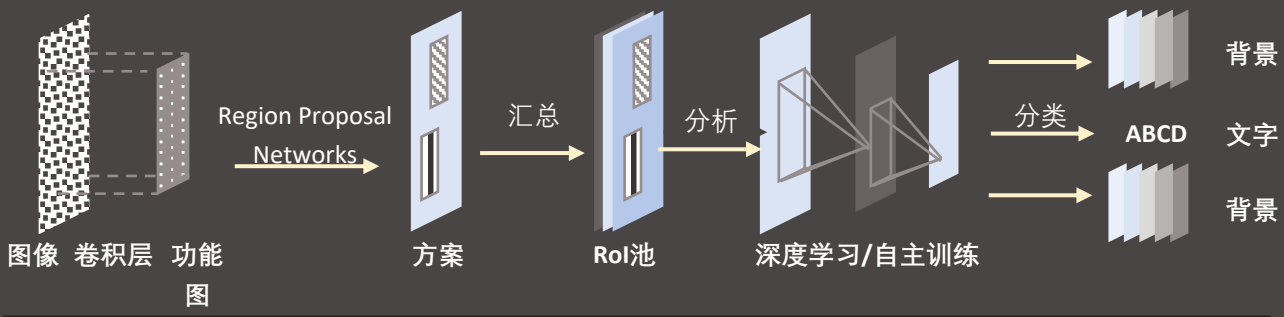
- **计算机视觉技术：**通过摄影机和电脑代替人眼对目标进行识别、跟踪和测量，对环境进行三维感知
- **自然语言处理技术：**通过建立形式化计算模型来分析、理解和处理自然语言
- **跨媒体分析推理技术：**协同综合处理多种形式，如文本、音频、视频、图像等混合并存的复合媒体对象
- **智适应学习技术：**模拟教师学生一对一教学过程，赋予学习系统个性化教学的能力
- **群体智能技术：**集结多个意见转化为决策的过程，降低单一个体做出随机性决策的风险
- **自主无人系统技术：**通过先进技术进行操作或管理而不需要人工干预的系统
- **脑机接口技术：**在人或动物脑与外部设备间建立的直接连接通路，以完成信息交换

来源：弗若斯特沙利文，头豹研究院

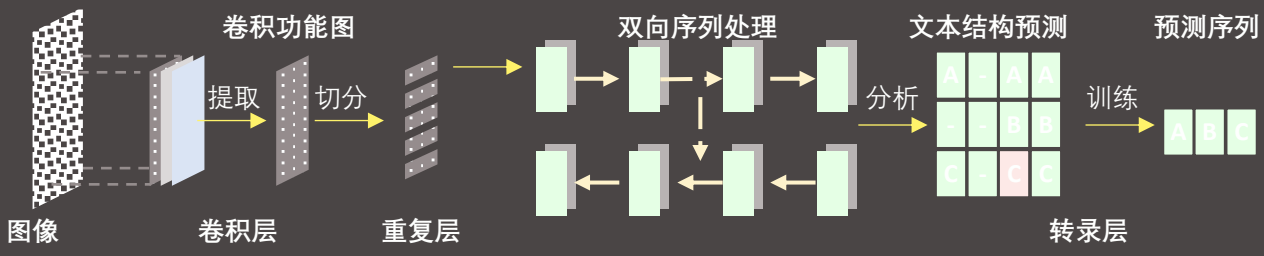
图像预处理技术原理



文字检测技术原理



文本识别技术原理



- 随著AI学习方法在金融、医疗、社交等场景实用化落地，大量数据的哺育将不断完善AI训练算法。例如，CVPR 2021的一篇论文中提出了名为Skip-Convolutions（跳跃卷积）的新型卷积层，它可将前后两帧图像相减，并只对变化部分进行卷积；
- 在图像预处理技术中，基于CNN的神经网络作为特征提取手段，CNN强大的学习能力也可增强AI模型中特征提取的鲁棒性；由多个级联分类器组成的FrameExit的网络可以随着视频帧的复杂度，来改变模型所用的神经元数量，即在视频前后帧差异大的时候，AI会用整个模型计算，而在前后帧差异小的时候，则只用模型的一部分计算。

来源：弗若斯特沙利文，头豹研究院

1.3.6 技术服务：MLOps提升团队协作效率

- 伴随着产业智能化的发展趋势，AI正成为诸多行业转型升级的通用技术。目前，AI最为成熟和广泛的应用领域包括公安、交通、金融、教育等。AI在其他行业的应用需求分散程度高、场景亦具有多样性特征，但AI的应用需求仍广泛存在。针对不同的应用场景，AI开发平台均可提供云端的自然语言理解 、自动语音识别 、视觉搜索、图像识别、文本语音转换、机器学习托管等服务内容。AI开发平台可为开发者或企业用户提供构建高级文本和语音聊天机器人、智能机器学习应用程序等的便捷操作。
- 对于个人或企业开发者而言，开发时间与开发成本是搭建AI应用程序的主要考虑指标。借助云原生及弹性分布式运算的架构可为用户在AI模型的训练与推断层面降本增效，而借助MLOps，团队的开发与部署效率会得到显著提升。
- MLOps是ML的DevOps。数据科学家构建的机器学习（ML）模型需要与其他团队（业务团队、工程团队、运营团队等）紧密合作。团队工作为沟通、协作和协调方面提出了挑战，而MLOps的目标正是通过完善的实践来简化此类挑战。**MLOps为系统带来灵活性与速度：**MLOps通过可靠且有效的ML生命周期管理，减少开发时间并得到高质量的结果；MLOps从DevOps中延续的持续开发（CD）、持续集成（CI）、持续训练（CT）等方法和工具保障AI工作流程和模型的可重复性，开发者可随时随地轻松部署高精度机器学习模型并 集成管理系统可连续监测机器学习资源。
- MLOPs也对平台的数据和超参数版本控制、迭代开发和试验、测试、安全性、生产监控、基础设施等环节提出了更高要求。MLOps平台数据在定义输出时起着与书面代码同等重要的作用，因此数据复杂性较DevOps平台相比有所提升。针对MLOps平台面临的挑战，MLOps的实现流程包括用例发现、数据工程、机器学习管道、生产部署、生产监控等五个阶段，其工作流程主要通过敏捷方式实现。

MLOps概念：MLOps=ML+DevOps

模型设计	模型开发	模型操作
需求工程	数据工程	机器学习模型部署
ML用例优先级	机器学习模型工程	CI/CD管道
数据可用性检查	模型测试和验证	监控和触发

来源：弗若斯特沙利文，头豹研究院

结合人工智能的发展历程和AI开发平台的技术特性来看，AI基础层资源的整体效能水平在不断进化，AI开发平台的发展与人工智能基础层的发展基本一致，大致可以分为三个发展阶段：雏形期、发展期和成熟期。

- 雏形期：

- AI基础层出现粗放式单点工具，产业链逐步清晰。
- 发展支撑点：GPU支撑模型训练对AI算力需求，AI加速落地催化了数据标注等行业星期，通过API输出AI基础算法能力，大部分依赖于人工设计与开发。

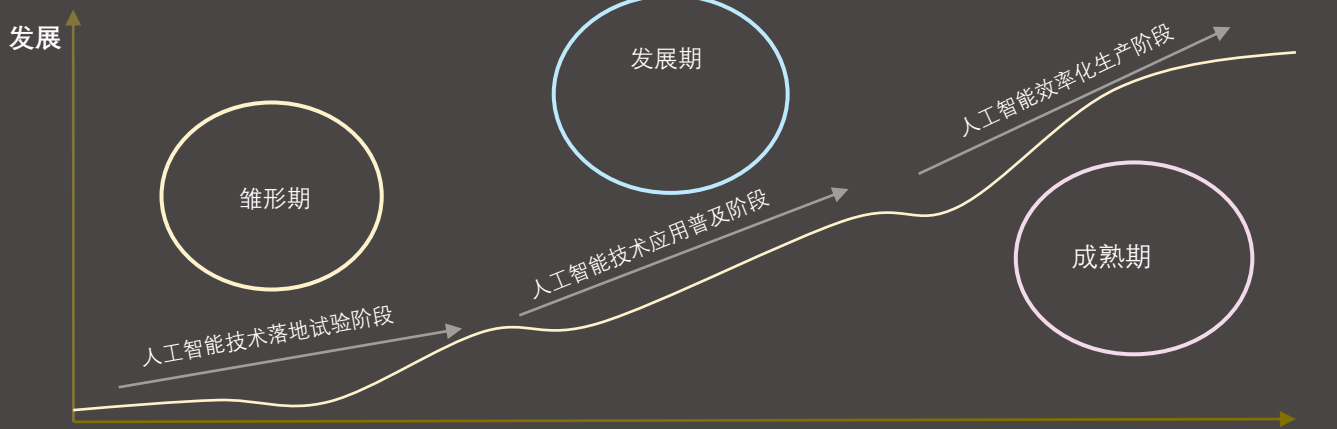
- 发展期：

- AI服务体系逐步完善，市场开始探索产品形态与商业模式。
- 发展支撑点：AI算力提供商将IT基础资源集合形成资源池，需求推动产品优化，算法玩家持续深耕垂直开发领域及AI工程化能力，帮助下游客户降低开发成本，提高模型生产效率。

- 成熟期：

- 工具智能化水平提升，市场进入玩家竞争阶段。
- 发展支撑点：算法、算力、数据各赛道基础层企业提供精细化解决方案，积极参与价值链延伸，塑造核心竞争力；AI供应商布局覆盖数据治理、模型开发、算力资源管理全流程的一站式AI模型开发平台。

AI开发平台发展历程



应用开发主要依靠开源模型或者开发者自发分享的代码进行

随着AI技术发展和需求激增，开源框架陆续出现，助力开发人员快速生成并训练模型，开原框架也成为了众多AI开发平台的基础

专业化的商用AI开发平台的出现，致力于为甲方提供一站式的AI模型生产服务，满足智能化转型需求方对多层次资源的多样性需求

来源：弗若斯特沙利文，头豹研究院

章节二　AI开发平台商业模式

随着市场规模逐步扩大，AI开发平台单客户使用成本将大幅降低，平台利润率将逐步提高。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

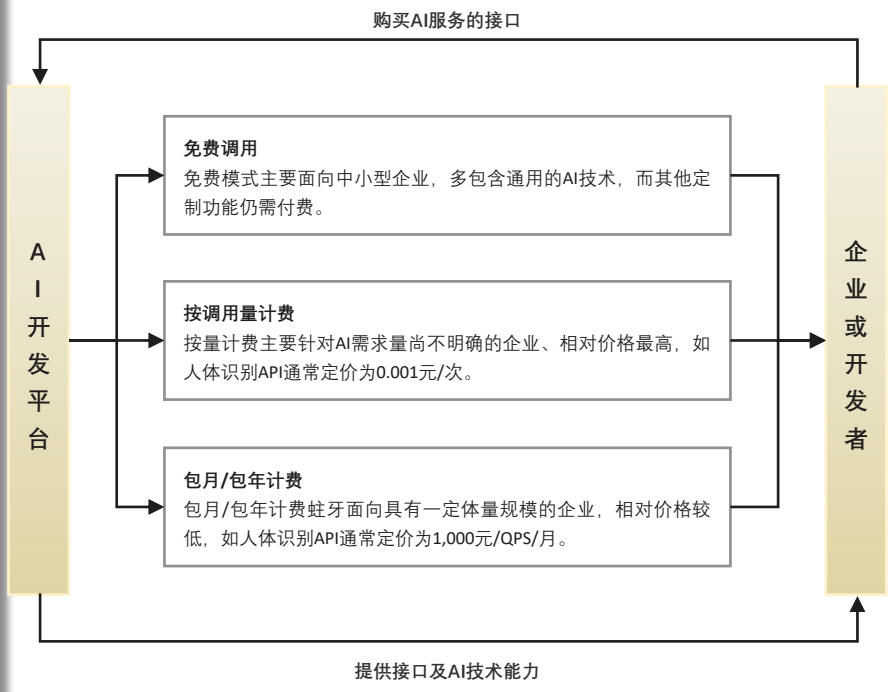
2 商业模式

随着平台规模逐步扩大，AI开发平台单客户平均使用成本将大幅下降，平台利润率将逐步提高。因此，实现规模化经营是AI开发平台的重要发展战略，可帮助平台在降低成本的同时赋予平台更大议价空间。同时该现象也解释了大型厂商在“部分免费”模式下仍能实现盈利的底层商业逻辑，同时也体现了大型厂商相对于中小厂商的竞争优势。

“AI开发平台的商业模式相对简单，开发者数量和平台规模成为其营收的决定性因素。”

- AI开发平台的商业模式是通过为企业或开发者提供AI技术接口或AI开发工具而获利，计费方式主要有免费、按调用量计费、包年或包月等三种。
  - 免费模式为企业或开发者提供如文字识别、人脸识别等常见服务与通用AI技术接口，设有使用限制，通常为1-5QPS/天，主要面向使用量较小的中小企业。免费模式通过数据积累、构建AI生态和提供附加服务从而实现盈利。
  - 与包年或包月的计费方式相比，按量计费的价格较高，适合尚未明确需求量的企业。
- 在产品营销方面，平台运营商可以通过免费试用、补贴、在线教学等方式提高流量转化率。且大型平台可以通过永久免费的通用产品，进一步提高流量至用户的转化。同时平台运营商还可以在客户服务中探索用户的其他增值需求，如云服务、定制AI开发解决方案等。

AI开发平台商业模式



来源：弗若斯特沙利文、头豹研究院

章节三　AI开发平台市场规模

从2016年到2020年，中国AI开发平台的市场规模迅速扩大。2021年，中国AI开发平台的市场规模为234.8亿元。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

3 市场规模

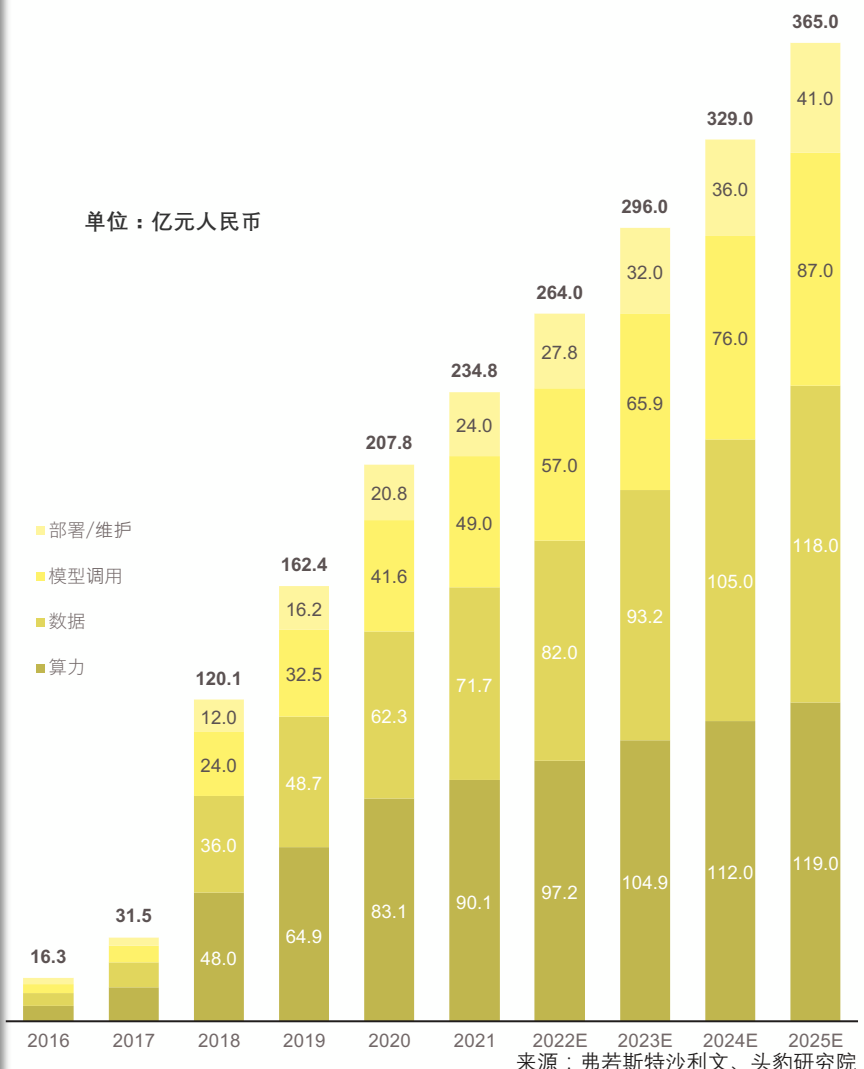
2016-2021年，中国AI开发平台营收规模快速扩张，2021年中国AI开发平台营收为234.8亿元。

在政策红利、行业渗透率以及芯片性能稳步提升的背景下，预计2025年中国AI开发平台市场规模将达365.0亿元。

“2021年中国AI开发平台营收规模为234.8亿元，2025年中国AI开发平台市场规模将达365.0亿元。”

中国AI开发平台市场规模（按营收计），2016-2025年预测

CAGR	2016-2021	2022E-2025E
总计	**56.0%**	**8.4%**
算力	58.0%	5.2%
数据	57.5%	9.5%
模型调用	56.8%	11.2%
部署/维护	45.8%	10.2%



来源：弗若斯特沙利文、头豹研究院

章节四　竞争要素

在本章中，沙利文在本章围绕AI开发平台的市场核心竞争力进行剖析，分为 “提高数据处理能力 ” 的硬实力和 “增强平台易用性 ” 、 “提升生态开放性 "的软实力。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

4 竞争要素

AI开发平台的用户主要为AI产业中个人或企业的开发者，因此如何为开发者提供更高效便捷的开发平台及其他衍生服务将是AI开发平台的核心竞争所在。沙利文将AI开发平台的核心竞争力归纳为“提高数据处理能力”的硬实力和“增强平台易用性”、“提升生态开放性”的软实力。

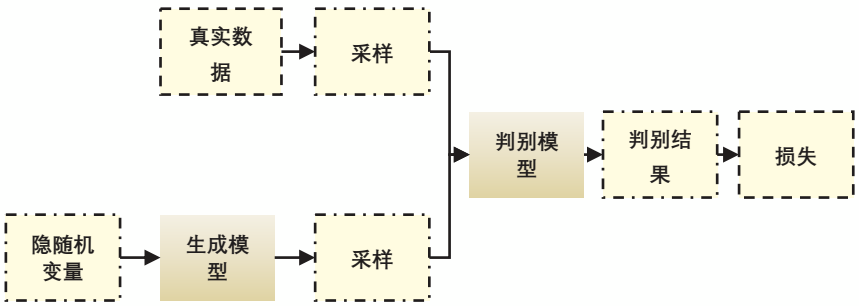
> **AI开发平台的核心竞争将围绕“提高数据处理能力”、“增强平台易用性”、“提升生态开放性”三方面展开。**

AI开发平台厂商以平台底层硬件、算法模型等能力为支撑，为开发者提供更有效的AI开发平台服务。

硬实力之一：智能标注——从“人工”向“智能”的艰难突破

- 整体来看数据标注的智能化替代难度极高，而现阶段标注工具在算法的帮助下已经可以完成基本的标注工作，如自动识别标框、自动识别语音等，未来标注工具的算法将得到进一步优化。
- 对于AI开发平台而言，智能标注功能在优化算法效率、提升用户体验等方面均起重要作用。而AI开发平台上线的智能标注功能包括引入GANs以优化标注效果、采用半监督学习机制以固化标注、引入难例筛选机制以优化标注结果、提供基于难例的数据标注改进建议等，但在实际应用过程中， AI开发平台厂商仍需解决上述方式的局限性。
  - **GANs**：判别器和生成器均需要较高的同步性，而在实际训练过程中易出现“判别器收敛、生成器发散” 的现象，因此判别器与生成器的优化需要极高的设计标准；GANs在训练过程中会出现模型缺失问题（即生成器功能退化），由于在过程中会不断产生相同的样本点，因此导致学习过程无法继续。
  - **半监督式学习**：模型难以纠正自身错误；可能会出现过度平滑的问题，进而导致节点特征无法区分。
  - **硬例选择机制**：只可在模型训练过程中生成难例，无法实现离线的难例挖掘，且用户必须自适应代码，才能修改代码，进而使用在线难例挖掘；难例筛选机制的核心思想是通过自举（bootstrapping）的方式生成难例集，且生成方式仅通过训练样本在训练时的损失值来判断，因此导致评判维度过于单一，无法保证模型精度的提升效果；算法思想不够成熟，无法形成系统性方案。

生产对抗网络GANs算法流程图



来源：极术社区、easyAI、华为云、弗若斯特沙利文、头豹研究院

硬实力之二：机器学习框架——改善框架缺陷，提升用户体验，构建AI生态

- TensorFlow和PyTorch皆为机器学习主流框架，且具有大规模的开发者社区以及大量成熟的可用代码。全球深度学习框架超90%的份额由这两个框架所占据，但TensorFlow和PyTorch两者彼此间具有不同的特征：

TensorFlow：

- **优点：**适用于工业生产环境，模型训练与模型部署皆具备完备的解决方案。
- **缺点：**API多种且风格多样，对新手较不友好；分布式训练的迭代思路较不清晰；对云原生支持度较低。

PyTorch：

- **优点：**编程API风格简约，直观易懂；基于动态计算图构建的深度学习模型，可根据堆栈信息快速debug。
- **缺点：**部署生态尚处成长阶段，因此并不不支持部分操作。

- 开发者数量有限是中国厂商开源机器学习框架的统一缺陷，使用人数较TensorFlow和PyTorch有显著差距，且在语种支持能力略逊一筹，仅支持中文和英文。相比之下TensorFlow和PyTorch可支持部分小语种，因此开发者生态更为完善。
- 目前全球机器学习框架生态基本已稳定，通用框架TensorFlow和PyTorch开源时间较早，因此具备生态优势，而针对机器学习的技术迭代以及TensorFlow和PyTorch的缺陷，中国厂商的自研框架优化了其框架架构，以为开发者提供更好的使用体验。长期来看，中国自研框架的开发者生态多集中于国内，未来也将会有更多企业使用中国的自研生态进行机器学习开发，但全球的机器学习框架格局预计仍将保持TensorFlow和PyTorch为主导的局面。

- 百度、华为等厂商推出机器学习自研框架PaddlePaddle、MindSpore等：

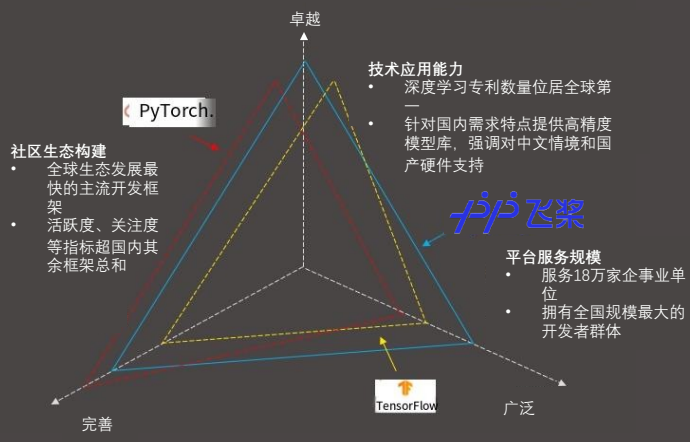
PaddlePaddle：

- **优点：**社区活跃；生态链完整；应用对用户友好；提供全流程能力支持；迭代节奏快；支持大规模异步分布式训练。
- 缺点：个人开发者居多，尚未有大面积厂商部署该框架。

MindSpore：

- 优点：加强版支持可视化、差分隐私、二阶优化算法、图神经网络、量化训练、混合异构、MindSpore Serving、PS分布式训练、MindIR、调试器等；支持多平台运作；倡导软硬件协同设计 ‘’ 支持多种模式分布式训练等。
- 缺点：社区人数较少，部分功能仍待完善。

- 伴随技术、产业、政策等各方环境成熟，AI已跨过技术理论积累和工具平台构建的发力储备期，开始步入以规模应用与价值释放为目标的产业赋能黄金十年。随着AI规模化落地，基于深度学习框架上下延伸、构建智能生态平台成为国内外科技巨头的共同选择。

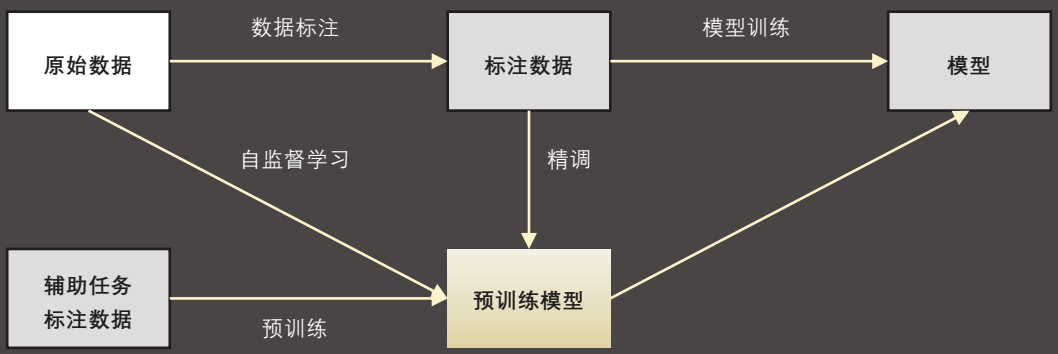


来源：弗若斯特沙利文、头豹研究院

硬实力之三：预训练模型——“大”与“小”兼备

- 未来，AI开发平台厂商会发布支持计算机视觉、语音识别、语音合成、自然语言处理、机器翻译、智能推荐、商业分析及预测、科学计算、多模态数据任务及复合任务等功能的多种预训练模型，而预训练模型也将沿着多技术路径发展。
  - **预训练模型规模将提升（“大”）**：大规模的预训练模型将包括超1,000亿个参数，单次训练成本预计将超1,000万美元，将搭载包括混合精度训练、数据并行、模型并行、Lamb优化器、三维并行训练、稀疏注意力加速等训练优化技术，但该预训练模型过程繁琐，只能布局于云端应用中。
  - **预训练模型将通过压缩与加速提升其灵活性（“小”）**：首先，可基于知识蒸馏、剪枝等现有技术，压缩预训练语言模型；其次，可通过矩阵参数分解、参数共享以及模型结构设计与搜索，实现预训练模型的压缩，旨在去除掉参数矩阵的冗余部分，将模型变“小”；同时，还可以基于量化方法，通过减少数值表示所需要的比特值，压缩预训练语言模型，如把32位浮点数减少至8位浮点数或4位浮点数以简化运算过程。而压缩后的预训练模型可以应用于设备端，应用价值极为广泛。
- 未来，AI开发平台厂商需针对Resnet50-v1.5、SSD-ResNet34、3D UNET、RNNT、Openpose、YOLO、BERT、DLRM等训练模型不断优化原有训练方法，加快训练速度，或提出新的训练方法，提升预训练模型的成熟度。

预训练模型概念图



来源：AI平台建设、弗若斯特沙利文、头豹研究院

AI开发平台厂商通过提供灵活的服务和简化开发者的操作来改善开发者的体验，增强平台易用性，提升生态开放性，进而提高厂商技术水平。

软实力之一：AutoML——降低AI开发门槛，提升AI开发效率

- AutoML是AI领域的重要趋势之一。AutoML能够将迭代过程集成到传统机器学习中，以构建自动化过程，大幅降低了机器学习的门槛：
  - AutoML是一种机器学习过程，旨在通过一系列算法和启发式方法实现从数据选择到建立模型的自动化。研究人员仅需输入元知识（卷积运算过程/问题描述等），该算法即可自动选择合适的数据、自动优化模型结构和配置、自动训练模型并适应它可以部署到不同的设备。

- AutoML可帮助AI开发平台自动完成神经结构搜索、模型选择、特征工程、超参调优、模型压缩等任务。同时依赖于结构化或半结构化数据的分类或回归问题，也可通过AutoML实现自动化，进而大幅提升AI训练的效率。

- 然而，AutoML发展路径上仍存在部分难点需要解决。首先，AutoML仍需要大量算力，因此企业仍需要在研发过程中尝试更多的解决方案；其次，AutoML在提升处理复杂度的同时，仍需保持一定的透明度，以允许模型开发者确认模型质量。最后，AutoML作为自动化工具，在提升工作效率的同，时也具有资源优化和迭代不足、复杂模型处理有限、特征工程低效、特征工程可移植性较低等方面的局限性。

软实力之二：以开发者为中心——增强平台服务能力以构建生态

- 由于AI开发平台是面向开发者的服务，因此平台满足开发者需求、提升平台兼容性、为开发者提供更好的开发体验的能力也应该成为重要的评价标准。

- 在数据准备功能方面，AI开发平台可提供包括本地数据集载入、第三方开源数据集载入、云端数据集调用在内的多种数据接入方式；除此之外，平台还可提供多类型数据标注服务模式，并在操作面板进行数据可视化呈现。

- 在模型训练功能方面，AI开发平台可提高机器学习框架、编程语言、云端IDE工具的兼容性，并提供自定义、模块化算法修改方式。现阶段主流的AI开发平台均可支持弹性训练、计算资源实时监控、硬件设备异构训练、多种并行训练模式与预训练模型迁移等模型管理服务，旨在为开发者提供便捷化的AI开发服务。

- 在模型管理与部署功能方面， AI开发平台研发方向涵盖提供包括提升AI开发平台兼容性（如支持更多编程语言、支持CI/CD、支持第三方AIOps工具等），且支持用户自行构建工作流在内的机器学习工作流构建服务，同时支持模型漂移监测、资源负载监测、自动告警、监控指标可视化呈现在内的模型部署监控服务等。

- 在账户管理功能服务能力方面，部分主流的AI开发平台选择开放部分免费资源（如计算资源、存储资源、数据集资源、模型资源等），为开发者提供平台体验服务。而大部分AI开发平台提供包括按需付费、预付费、订阅付费（如年/月费）在内的多种收费模式，提升平台收费的灵活性，满足不同类型开发者的需求。

来源：华为云、弗若斯特沙利文、头豹研究院

章节五　综合表现

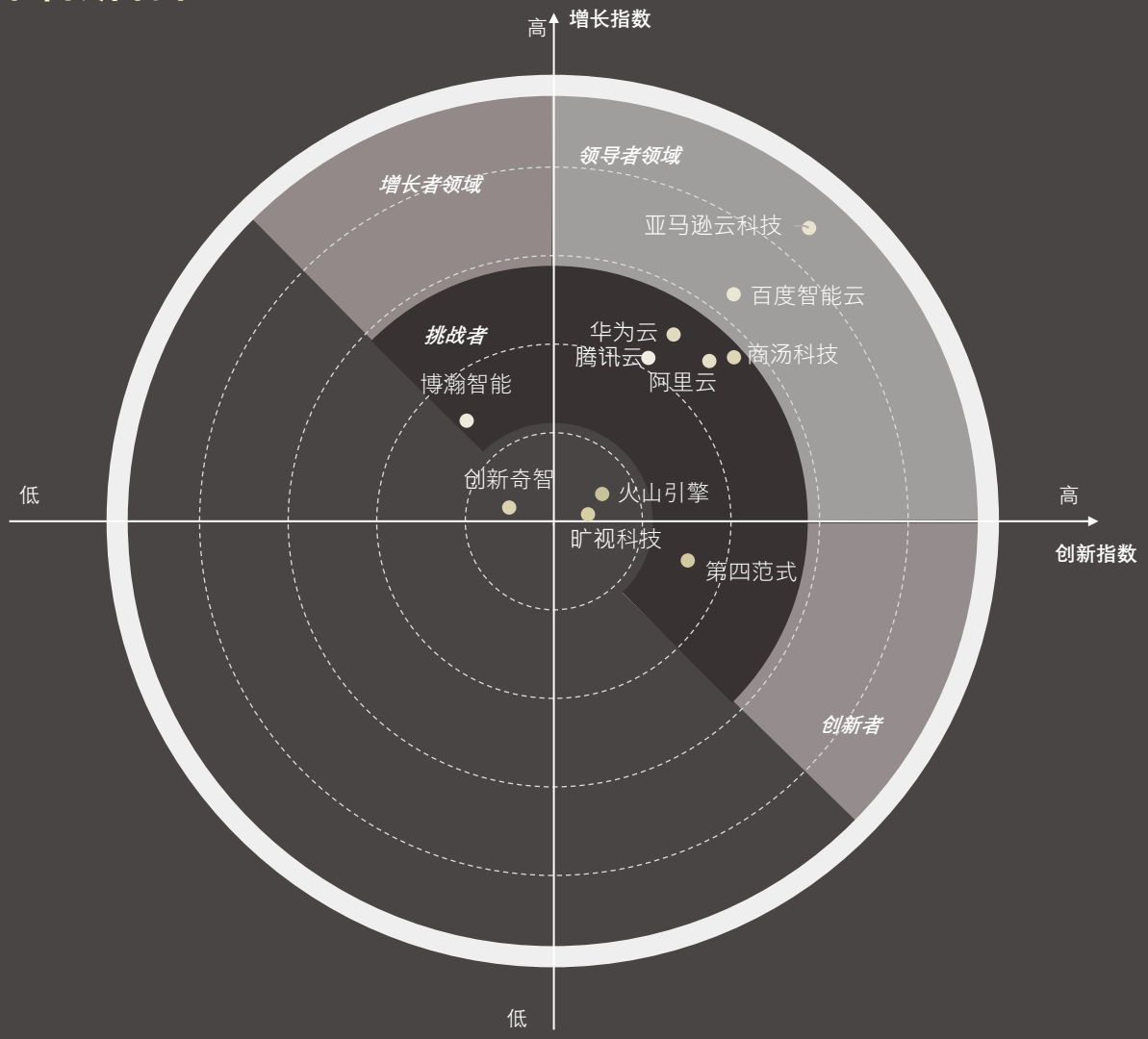
本报告对竞争主体AI开发平台应用产品和服务综合竞争力的分析结论，仅适用于该阶段AI开发平台应用市场发展情况。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

2021年中国AI开发平台市场综合表现分析

弗若斯特雷达 TM



注：圆环按由内向外递增的逻辑对应由低至高的综合评分，竞争力由“创新指数”以及“增长指数”综合得出
中国AI开发平台应用市场发展处于平稳增长期，本报告对竞争主体AI开发平台应用产品和服务综合竞争力的分析结论仅适用于该阶段AI开发平台应用市场发展情况。

- **横坐标代表“创新指数”：**

衡量竞争主体在AI开发平台应用的创新能力，位置越靠右侧，AI开发平台应用服务职能丰富度和产品调优能力越强。

- **纵坐标代表“增长指数”：**

衡量竞争主体在AI开发平台应用产品架构、产品功能、性能增长维度的竞争力，位置越靠上方AI开发平台应用产品增长能力越强。

> **中国AI开发平台市场正处于技术成熟和平台完善的阶段，竞争厂商们在创新和增长能力方面皆具备一定优势。**

章节六　评分维度

沙利文在本章根据从创新指数及增长指数两个维度对AI开发平台进行评分。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

6 评分维度

> 本报告设立创新指数评估体系对AI开发平台进行评价及分析，下设技术创新能力及业务创新能力两大指标。

中国AI开发平台评价维度——创新指数

创新指数		
一级指标	二级指标	要点
技术创新能力	基础硬件	评估厂商在AI芯片及AI服务器研发创新等多个维度的表现
	数据采集与标注	评估厂商在智能数据采集、标注、分析、筛选、增强等多个维度的表现
	深度学习框架	评估厂商在AI深度学习框架研发、框架创新等多个维度的表现
	算法模型	评估厂商在AI算法模型在算法准确率与运行效率、场景通用性等多个维度的表现
业务创新能力	云原生架构	评估厂商在应用云原生架构和优化云开发环境方面的能力
	云安全治理	评估厂商在AI开发平台应用云安全技术、平台综合性安全性能的表现
	大数据管理	评估厂商在AI开发平台应用大数据管理技术、提升平台数据库价值的能力
	可视化开发	评估厂商AI开发平台可视化、零门槛开发功能套件，简化应用开发流程、降低应用开发门槛的能力

6 评分维度

> 本报告设立增长指数评估体系对AI开发平台进行评价及分析，下设服务能力及生态能力两大指标。

中国AI开发平台评价维度——增长指数

增长指数		
一级指标	二级指标	要点
服务能力	算力服务	评估厂商AI开发平台在算力部署、管理能力等多个维度的表现
	数据服务	评估厂商AI开发平台在数据样本体量、数据产品兼容性及数据产品多样性的等多个维度的表现
	算法服务	评估厂商AI开发平台在深度学习框架兼容性、算法模型产品多样性、适配性及可移植性等多个维度的表现
	平台服务	评估厂商AI开发平台在数据管理及安全、AI开发管理、AI应用部署等多个维度的表现
	定价策略	评估厂商AI开发平台在服务定价机制的灵活度及价格弹性等多个维度的表现
生态能力	生态繁荣度	评估厂商AI开发平台在开发社区繁荣度、市场影响力、应用广度等多个维度的表现
	生态发展	评估厂商AI开发平台在生态可持续性发展能力、外部威胁对抗能力等多个维度的表现

章节七　头部厂商案例

在本章中，沙利文介绍了中国AI开发平台厂商。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

7 亚马逊云科技

超过15年以来，亚马逊云科技（Amazon Web Services）一直以技术创新、服务丰富、应用广泛而享誉业界。亚马逊云科技一直不断扩展其服务组合以支持几乎云上任意工作负载，目前提供超过200项全功能的服务，涵盖计算、存储、数据库、网络、数据分析、机器学习与人工智能、物联网、移动、安全、混合云、虚拟现实与增强现实、媒体，以及应用开发、部署与管理等方面；基础设施遍及30个地理区域的96个可用区，并已公布计划在澳大利亚、加拿大、以色列、新西兰和泰国新建5个区域、15个可用区。全球数百万客户，包括发展迅速的初创公司、大型企业和领先的政府机构，都信赖亚马逊云科技，通过亚马逊云科技的服务支撑其基础设施，提高敏捷性，降低成本。

亚马逊云科技全球五大独特优势：

- 广泛而深入的云服务
- 成熟丰富的客户实践
- 覆盖全球的基础设施
- 引领行业的安全合规
- 值得信赖的合作伙伴

“亚马逊云科技旨在让机器学习掌握在每位开发人员的手中，提供高兼容性、高功能模块化的AI开发平台服务。”

- 亚马逊云科技具备完备的AI开发软硬全栈供应水平，从专用基础设施、AI平台到各类场景开箱即用的AI服务解决方案，结合亚马逊云科技的系列云上服务，满足各类型客户的不同需求：

1. **AI 基础设施：**自研的机器学习推理芯片Amazon Inferentia和机器学习训练芯片Amazon Trainium，通过专用芯片实现从推理到训练的端到端机器学习硬件加速。结合自研服务器芯片Amazon Graviton3，亚马逊云科技提供节能高效的机器学习基础设施。
2. **AI 平台：**机器学习平台Amazon SageMaker提供完整、丰富的功能供开发人员、数据科学家和ML工程师使用。Amazon SageMaker Studio Lab提供免费资源，Amazon SageMaker Jumpstart和Amazon SageMaker Canvas提供低代码/无代码的快速上手功能、Amazon SageMaker Pipelines构建全自动ML流程、Amazon SageMaker Ground Truth Plus提供智能标注服务、Amazon SageMaker Data Wrangler内置300多种数据转换、Amazon SageMaker Autopilot自动执行AutoML。
3. **AI 服务：**自然语言理解（NLU）、自动语音识别（ASR）、视觉搜索和图像识别、文本转语音（TTS）及机器学习（ML）托管服务。

- “智能湖仓架构”融合机器学习和数据管理平台，提供数智融合、统一治理的体验。Amazon Redshift ML和Amazon Athena ML均支持以SQL语句的方式发起模型训练请求，Amazon SageMaker Canvas AutoML能力提供模型训练，以SQL形式返回。
- 亚马逊云科技凭借合作伙伴关系和开发人才教育体系打造的完善网络吸引了超过十万客户持续选择。亚马逊云科技拥有80多个ML/AL能力合作伙伴，为客户提供丰富成熟的行业解决方案。在亚马逊云科技 Marketplace 中有来自 300 多个ISV的 1，000 多款机器学习产品，拥有众多标杆案例，覆盖如医疗、零售、金融服务、社交文娱、制造、能源等行业。

亚马逊云科技AI开发平台案例

OPPO在月活过亿的对话式AI产品小布助手业务上，为了达到行业领先的对话语义理解效果，在Amazon EC2 Inf1上创新地研发可支持预训练大模型高效推理服务模块，在部分业务场景上预计可降低35%以上的模型推理服务成本，并期望逐步拓展到越来越多的新场景中。借助Amazon EC2 Inf1，OPPO的机器学习团队不断利用更复杂的算法模型进行创新，并加速改善客户的整体体验。

施耐德电气借助Amazon SageMaker及数据库和计算服务，成功构建智能工业视觉质量检测解决方案“云-边协同AI工业视觉检测平台”。借助Amazon SageMaker，施耐德电气能够成功且准确地构建适应实际制造场景的机器学习模型，通过将生产线的产品图像与合格产品的标准样品进行对照，通过自动化的工业视觉检测来识别产品中的复杂缺陷。该解决方案率先在施耐德电气武汉工厂上线，显著提高了生产线的检测效率，将误检率降低0.5%以内，并实现了零漏检率。

有道乐读基于Amazon Personalize的个性化推荐以及大数据服务，为最终读者提供精准图书推荐。借助Amazon Personalize，有道乐读可以通过简单的API调用来设计个性化图书推荐，无需具备机器学习经验。Amazon Personalize服务开箱即用，在一个月内就有效帮助有道乐读实现图书的精准推荐和预测，有道乐读由此确保优质的用户体验，使得月活跃用户提升20%。此外，相比之前以月为单位的迭代周期，现在基本实现按天交付，甚至实现当天更新当天交付。

来源：亚马逊云科技、弗若斯特沙利文、头豹研究院

7
百度智能云

百度的AI开发平台由全功能AI开发平台(BML)，零门槛AI开发平台(EasyDL)，AI开发实训平台(AIStudio)组成，由百度自主研发的飞桨平台统一进行赋能。目前已经在中国地区累计了477万开发者，18万家企事业单位用户。
百度AI开发平台覆盖不同需求的开发者，涵盖数据处理，算法开发， 模型训练，预测服务部署，以及资源管理等贯穿模型全生命周期的能力。

2022年产品更新

- **数据分析引擎（Data Analytics Engine）**：百度深度优化的高效数据分析引擎，可执行跨数据库的联邦查询，支持数据自动分析及可视化。
- **特征库（Feature Store）**：可确保模型训练使用的特征数据和预测服务使用的特征数据一致性，解耦特征生产环节和消费环节，实现不同团队间的特征共享和复用。
- **XAI**：提供6种模型可解释算法/图表，为大多数常规机器学习模型和深度学习模型提供可解释性，降低使用风险。
- **MLOps**：在模型开发的全生命周期内，提供完备的自动化能力和二次开发SDK。可接入企业CI/CD系统实现模型生产运营一体化，提高开发效率和规范性。
- **大模型（Foundational Models）**：集成百度自主研发的大模型，包括NLP、CV及跨模态的大模型，进一步降低训练成本，提高模型泛化性。
- **模型风险管理**：对模型全生命周期流程进行审核和记录，并对风险进行持续监控和识别，最终降低模型使用中的风险。

百度的AI开发平台由全功能AI开发平台（BML），零门槛AI开发平台（EasyDL），AI开发实训平台（AIStudio）组成，由百度自主研发的飞桨平台统一进行赋能。目前已经在中国地区累计了477万开发者，18万家企事业单位用户。
百度AI开发平台以功能全面性为基础，其能力涵盖了AI模型从立项到部署预测服务的全生命周期流程，并对客户提供在线平台和私有化部署（on-premise）等不同的服务形式。
在前沿技术上， 百度AI中台开始具备了完备的MLOps能力，覆盖数据处理、特征工程、模型开发、训练任务、漂移监控（drift monitoring）、自动重训等， 并可以工作流（workflow）来自动执行，以实现高效的模型生产运营一体化；在XAI领域，百度AI中台通过模型风险管理模块，可实现模型全生命周期的风险管理，满足特定行业和机构的监管需求；同时，在飞桨平台的赋能下，实现了深度学习模型可解释性能力，以及模型鲁棒性与安全性领域的先进能力。在智能标注和AutoML领域相比上一次报告也有了更多新的特性。
得益于百度自研飞桨平台的赋能，百度AI开发平台内置了主流领域的开发套件（kit）和预训练模型，乃至当前热门领域的大模型。其中文心Ernie大模型是国内NLP领域内规模最大的单体模型。
在本土化层面，百度AI开发平台已经支持或适配了多种中国本土芯片，从而形成了一套全层次的本土化解决方案。

百度AI开发平台案例

国家山东省电力公司-输电线路安全巡检：

- 地域分布广、环境复杂多变等给输电线路的安全运行提出了严峻的挑战。通过对输电通道的可视化改造以及智能分析，百度AI开发平台极大地帮助国网山东电力提升了线路安全巡检的效率，为输电线路安全稳定运行提供了可靠保障。

中国邮政储蓄银行-邮储大脑：

- 百度AI开发平台帮助中国邮储银行实现了对AI模型从训练、测试、部署、运行、迭代的全生命周期的研发管理，引入多种机器学习、深度学习先进算法和模型，加速AI应用在全行业务场景的落地。

长沙地铁-智能维修头盔：

- 长沙地铁自主研发的可拆装式结构“智能维修头盔”，通过结合EasyDL物体检测训练工具分类识别模型，累积训练图片500张，耗时一周共迭代了17个版本，其模型准确率达到88.9%。能够实现自动拍照并识别常用工具名称和数量，为工具的查漏盘点提供了及时有效的保障，在智能安全帽的创新应用上取得了突破性进展。

来源：百度智能云、弗若斯特沙利文、头豹研究院

名词解释

- **QPS：**Queries-per-second，每秒查询次数，每秒查询率。QPS是对特定查询服务器在指定时间内处理的流量的衡量。它可以被解释为每秒钟并发的请求数。1QPS调用约86400次。
- API: 应用编程接口。API是一种预定义的功能，其目的是为应用程序和开发人员提供访问基于某种软件或硬件的一套例程的能力，而无需访问源代码或了解内部工作机制的细节。
- 卷积：一个数学概念，通过两个函数f和g生成第三个函数，表示函数f和g的重叠部分函数值的乘积在翻转和平移后的积分长度。
- CGRA：粗粒可重构架构。CGRA是空域中的一种并行计算模式，它将不同粒度、不同功能的计算资源组织在空域硬件结构中。在运行时，根据数据流的特点，将配置好的硬件资源相互连接，形成相对固定的计算路径，接近于计算的 "专用电路"；当算法和应用发生转变时，重新配置成不同的计算路径，执行不同的任务。
- CUDA。计算统一设备架构是英伟达公司基于其GPU（图形处理单元，通常可以理解为显卡）创建的一个并行计算平台和编程模型。
- DevOps：开发和运营的组合，是一组流程、方法和系统的总称，用于促进开发（应用/软件工程）、技术运营和质量保证（QA）部门之间的沟通、协作和整合。
- 数据标注：对文本、视频、图像等元数据进行标注的过程。标注的数据将被用于训练ML模型。
- 云原生：基于容器、微服务、DevOps等技术的一套云技术产品体系，是一种基于分布式部署、统一运营管理的分布式云。

方法论

头豹研究院布局中国市场，深入研究10大行业，54个垂直行业的市场变化，已经积累了近50万行业研究样本，完成近10，000多个独立的研究咨询项目。

研究院依托中国活跃的经济环境，从AI开发平台、深度学习、AI芯片等领域着手，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。

研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。

研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。

研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。

本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。

本报告所涉及的观点或信息仅供参考，不构成任何证券或基金投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告或证券研究报告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。

本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本报告所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本报告所载资料、意见及推测不一致的报告或文章。头豹均不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。