FROST & SULLIVAN
沙利文

头豹
LeadLeo

2022 China AI Development Platform Market Report

AI development platform/AI Model/AutoML/Low-Code AI Development

November 2022

Frost & Sullivan (China)
LeadLeo

Instruction

Frost & Sullivan and LeadLeo publish *2022 China AI Development Platform Market Report*. This report aims to analyze the definition, application, technology trends and development trends of AI development products in the AI Development market in China and identify the competitive landscape of AI development platform market, reflecting the differentiated competitive advantages of the leading brands in this market segment.

Frost & Sullivan and LeadLeo conducted a downstream user experience survey of the AI development platform market. Respondents came from a variety of enterprises of different sizes and in different industries.

The analysis of AI development platform market trends provided in this market report also reflects the overall movement of the industry. The final judgment of the market ranking and leadership is only applicable to this year's China AI development platform development cycle.

All figures, tables and text in this report are derived from Frost & Sullivan Consulting (China) and LeadLeo Research Institute surveys, and data are rounded off to one decimal place.

头豹 LeadLeo

FROST & SULLIVAN 沙利文

Framework

Chapter I AI Platform Structure

Frost & Sullivan focuses on the structure of AI development platform in this chapter, starting from following dimensions: infrastructure, framework and training platform.

1.1 AI Infrastructure

- The AI development platform is a platform that integrates AI algorithms, computing power and development tools, and opens up the development architecture for machine learning, deep learning, training models, etc. It also provides the computing power support required for development, and enables developers to efficiently use the AI capabilities in the platform for AI product development or AI empowerment through interface calls.
- The AI Open Platform provides developers with many development tools and frameworks that help reduce development costs, such as AI datasets, AI models and computing power. Developers can use the platform's datasets to train their own models, or use the platform's algorithmic framework to customize their own functions.
- The AI development platform architecture can be divided into four layers: infrastructure, framework, training platform, and technical services from bottom to top.

1. AI Infrastructure: Self-developed AI chips are the core competitiveness of enterprises, and self-developed chips show the trend of architectural innovation, morphological evolution, software and hardware integration.

1.1 Underlying hardware

The mainstream AI processor is essentially a system-on-chip (SoC), which can be used in scenarios related to image, video, voice, and word processing. the main architectural components of the AI processor include a specially designed computing unit, a large-capacity storage unit, and a corresponding control unit. By self-researching AI chips, companies can adapt the chip line architecture to their own algorithms to maximize computing efficiency, and self-researching AI chips will gradually become one of the core competencies of AI development platform companies.

1.1.1 AI chip architecture innovation

The cloud AI chip is mainly used for AI training scenarios, and computing power is one of its core metrics. In order to adapt to the applications and algorithms that need to be used in AI training, suppliers need to develop domain specific architecture (DSA) chips to carry out architecture innovation to realize performance optimization. As one of its three major components (calculation, storage and control), the computing unit can perform scalar, vector and matrix operations. Huawei has deeply optimized the matrix operations in the da Vinci architecture anetnd customized the corresponding matrix computing units to support high throughput matrix processing, so that it can use one instruction to complete the multiplication of two 16 * 16 matrices.

In order to solve the problem that the existing memory access speed is seriously lagging behind the computing speed of the processor, the new fully programmable, reconfigurable architecture (CGRA) chips, memory computing chips, and the new processor architecture IPU with high memory bandwidth may introduce the AI chip bottom ecology.

In addition, chip programming methods and software architecture design will also become an important part of AI chip innovation. For example, NVIDIA has greatly reduced the programming difficulty of its GPU by virtue of its CUDA framework, making GPU widely used in AI acceleration. In the future, more AI processors will provide multi-layer software stacks and development tool chains to help developers use underlying hardware resources more effectively, improve development efficiency, and reduce the low flexibility of special chips through software diversity.

AI Development Platform Structure

AI Technology Services	Services based on speech recognition, image recognition, character recognition and natural language processing; Construction of AI ecology
AI Training Platform	ML platform
AI Framework	ML framework represented by TensorFlow and PyTorch
AI Infrastructure	Including basic hardware (CPU, GPU, FPGA, NPU, ASIC and its combined SoC); Computing engine and cloud container

1.1.2 Evolution path of AI chip

One of the goals of AI chip innovation is to maintain a high energy efficiency ratio of the chip while adapting to the evolution of AI algorithms. In the future, the system-on-a-chip form of general-purpose plus dedicated chips will become mainstream (CPU+NPU, CPU+ASIC, etc.) and have a broader scope of application.

Traditional processor instruction sets (including x86 and ARM, etc.) evolve for general-purpose computing, and their basic operations are arithmetic operations (addition, subtraction, multiplication and division) and logical operations (with or without), which often require hundreds of instructions to complete the processing of a neuron in deep learning, and the processing efficiency of deep learning is not high. To solve the sub-pain point, the chip form needs to break the traditional von Neumann structure. The neural network processor NPU uses circuitry to simulate human neuron and synapse structures. In NPU, storage and processing are integrated in neural networks, which are reflected by synaptic weights. For example, the world's first deep learning processor instruction set DianNaoYu proposed by Cambrian can directly face the processing of large-scale neurons and synapses, which can complete the processing of a set of neurons through a single instruction, and provides a series of specialized support for the transmission of neuron and synaptic data on the chip. In AI training acceleration applications, Cambrian also launched the latest MLU370-X8 training acceleration card equipped with dual-chip quad-core particle SiYuan 370, in YOLOv3, Transformer, BERT and ResNet101 tasks, the average performance of 8 cards in parallel up to 155% of 350W RTX GPU.

Source: Frost Sullivan, LeadLeo

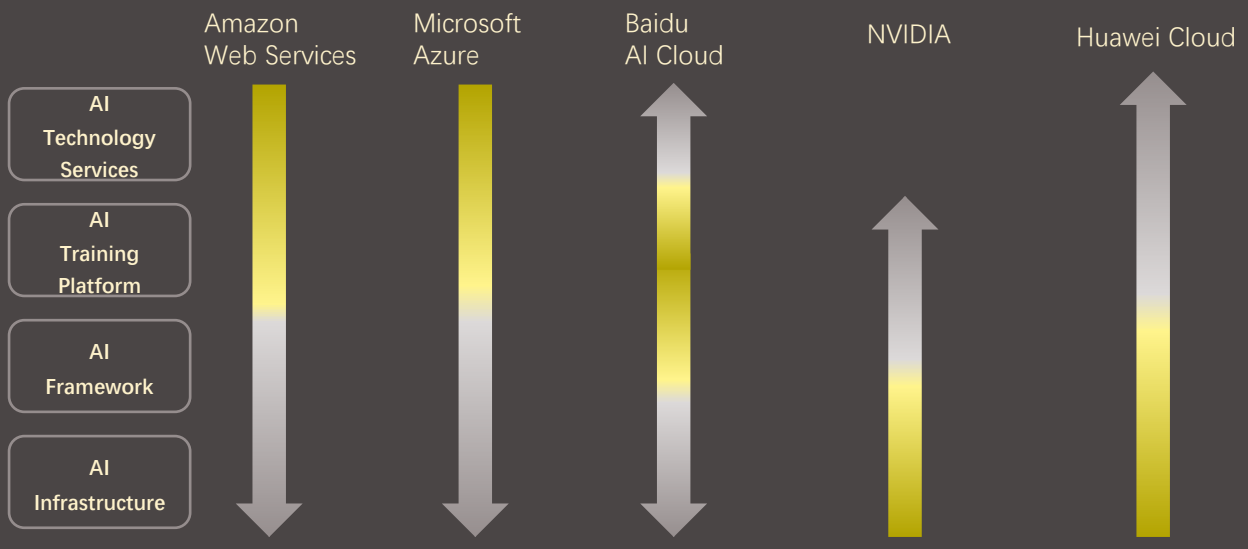
1.1.3 Software-Hardware Integration

- Software tools around AI chips began to transform from basic computing to scenario computing. In the past, as a representative of chip suppliers, NVIDIA have been building multi-level basic software tool ecology, such as high-performance operator library, communication algorithm, reasoning acceleration engine, etc., with CUDA programming model as the core.
- At this stage, the AI chip enterprises began to build a software and hardware integrated platform for differentiated scenarios. The business model has expanded from providing hardware support services to providing technical production tools and technical services. It has achieved efficient integration of the whole stack of underlying chips, programming frameworks, industry algorithm libraries, segmentation scenario research and development platforms, so as to cultivate a computing ecosystem of diversified industry scenarios and seize the segmentation market.
- At the same time, enterprises can also provide modular services according to customer needs, provide customers with services with weak capabilities, and improve the degree of customization of services.

1.1.4 Emgerging AI chips for mobile

1. The current traditional AI chip in order to meet the requirements of the continuous expansion of computing performance, while using new structures, new processes, but also make the chip's basic power consumption is climbing. The excessive chip power consumption poses a high challenge to the power supply capability and heat dissipation capability of the chip users and usage scenarios. The high power consumption of AI chips is a relatively low challenge for the mainstream fixed-end applications, which can be solved by simply increasing the number of supporting hardware devices. However, for the mobile side, the power consumption of AI chips is a problem that cannot be easily solved. Mobile applications such as cell phones, laptops, wearable devices, and autonomous vehicles are limited by the size of the available space and the capacity of batteries and other energy storage devices, and other objective factors that prevent the launch of ultra-high-power, high-performance AI chips. Therefore, AI chip design for the mobile side will be a mainstream direction for future chip development
2. In mobile AI chips, Korean manufacturers represented by Samsung Electronics and SK Hynix are in the forefront. Among them, Samsung Electronics' research results on mobile NPU chips presented at ISSCC 2022 show that its latest mobile NPUs have achieved major breakthroughs in optimizing chip data flow to enhance computing unit utilization, optimizing computing units to cover different computing accuracies, and providing different operating modes to meet different power consumption and performance requirements. In terms of computational accuracy, Samsung Electronics' mobile NPU can meet the accuracy requirements of INT4, INT8 and FP16, basically covering all the needs of mobile AI algorithms; and its breakthrough in mode switching also greatly solves the pain points of mobile chips in daily use. At present, Samsung Electronics mobile NPU has been applied in its 4nm Exynos SoC.

The pathways to full-stack integration of the AI development platform vendors



Source: Frost Sullivan, LeadLeo

1.1.5 Cloud native Enabling AI Infrastructure

Cloud-native technologies enable organizations to build and run elastic and scalable applications in new dynamic environments such as public, private and hybrid clouds. Representative cloud-native technologies include containers, service grids, microservices, immutable infrastructure, and declarative APIs, which enable the construction of loosely coupled systems that are fault-tolerant, easy to manage, and easy to observe. Combined with reliable means of automation, cloud-native technologies make it easy for engineers to make frequent and predictable major changes to systems.

1. At the infrastructure level, containers decouple application and technical architecture resources between cloud infrastructure and applications;
2. At the application level, users can choose between microservices or serverless architecture depending on the scenario;
3. In complex architectural scenarios, service component communication is controlled through a service grid;
4. Finally, the system is continuously iterated and updated through DevOps.

- **Infrastructure improvement**

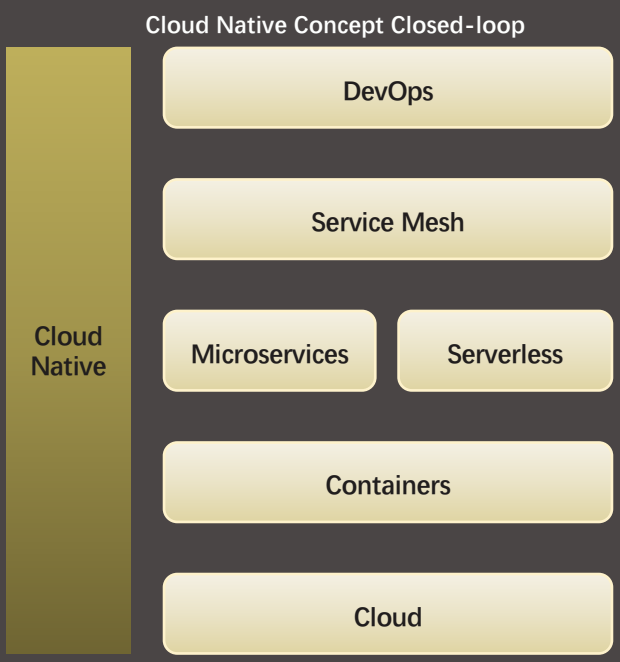
  The cloud-native-based deep learning training platform can be fully containerized and deployed, and based on Kubernetes (K8s), it provides elastic and flexible resource scaling, resource scheduling and allocation for different tasks, and is backward compatible with multiple CPU and GPU processors. Therefore, cloud-native-based AI development platform can be quickly adapted to appropriate cloud-native resources for both training of large-scale sparse data and training of perception-based scenarios. For example, AliCloud PAI can provide kernels that support near-linear acceleration, allowing training tasks to achieve performance enhancement and performance acceleration on multiple engines.

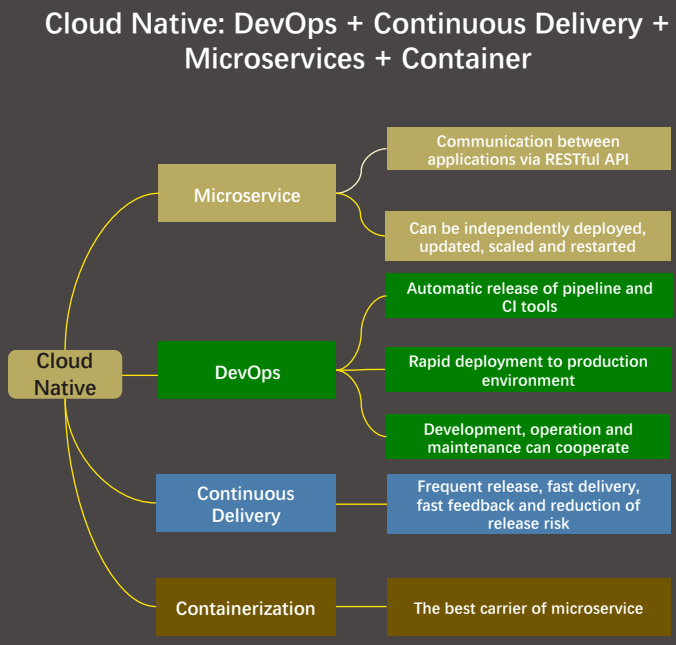
- **Training session enhancements**

  The cloud native container architecture can flexibly deploy computing resources for ML training, and reduce costs and increase efficiency for AI development through flexible training. The cloud of the AI development platform can monitor the computing power of the resource pool in real time, allocate the idle resources to the task in training when there are idle computing resources, improve the computing power of the task, so that the training job can converge quickly. After the task is submitted, the elastic training scheme can also allocate the recovered resources to the new ML training task according to the use of the free resources in the resource pool and the elastic work, so as to ensure the computing power of the new ML training.

- **User experience improvement :**

  Cloud-native applications can provide AI development platform users (developers) with more agile and high-quality application delivery as well as simpler and more efficient application management, and provide faster business demand response and better user experience. Cloud-native based AI platform is perfectly adapted to the team's AI online collaborative development, online AI teaching and local AI R&D migration to the cloud.





Source: Frost Sullivan, LeadLeo

1.2.2 Market situation of deep learning framework

1.2.2.1 Market share occupation

Over 90% of the share of the global deep learning frameworks market are occupied by TensorFlow and Pytorch, which are developed by Google and Meta respectively.

- TensorFlow is the most popular deep learning framework at this stage, with visualization, strong performance, versatility, etc. TensorFlow comes with a tensorboard visualization tool that allows users to monitor and observe the training process in real time, while supporting multi-GPU, distributed training, cross-platform running capabilities. TensorFlow has a variety of uses that are not limited to deep learning, but also has tools to support reinforcement learning and other algorithms.
- PyTorch is open-sourced by Facebook and features simplicity, ease of use, and detail. PyTorch has a more active community, providing complete documentation and guides for users to communicate and ask questions, but is smaller than Tensorflow's community.
- Other typical frameworks differ in terms of stability, debugging difficulty, execution speed and memory usage, such as Mxnet, PaddlePaddle, and CNTK.

  - Currently, "open source + top company support" is the mainstream model of deep learning frameworks on the market, which means the main frameworks are generally intervened by the industry's top companies and lead the internal application and construction.

Overview of Deep Learning Framework

Category	Deployment Location		Task		Basic process			
Type	Cloud Framework	Terminal Framework	Training Framework	Reasoning Framework	Underlying Computing Framework	Model Building Framework	Iterative Training Framework	Cross Border Framework
Definition	Mainly completed through data input or unsupervised learning methods such as reinforcement learning	Can run on mobile phones, security cameras, cars, smart home devices, IoT devices and other terminal devices that perform edge computing	Framework for completing deep learning in the data center	Mainly complete the optimization, deployment and inference calculation of the training model	Focus on in-depth learning of the underlying basic computing links in the basic development process	Provided Basic modules to support the creation of deep learning models, but they cannot contact the underlying computing modules	Provided basic modules to support process optimization, but the underlying computing module cannot be contacted by itself	Allows users to access the underlying data modules, and also provides ready-made basic modules to achieve rapid modeling
Key Technical Requirements	Computing power Security and stability	Security and stability	Ecological construction, ease of use, performance, support architecture	Usability, performance, underlying optimization, security and stability	Operation efficiency, data precision, algorithm design	Model processing and problem solving		Meet the technical requirements for other basic processes, and satisfied for compatibility, security and utility
Case	TensorFlow, PyTorch, 飞桨	Caffe2go, 飞桨	TensorFlow, PyTorch, 飞桨	OpenCL, TensorFlow, 飞桨	theano, TensorFlow, 飞桨, PyTorch	Caffe, mxnet, 飞桨, PyTorch	Keras, TensorFlow, 飞桨, PyTorch	TensorFlow, 飞桨, PyTorch

Source: Frost Sullivan, LeadLeo, CAICT

1.2.2.2 The competitive landscape

The current landscape of deep learning frameworks is gradually becoming clearer and shifting to oligopoly competition

- **Early exit**

Microsoft CNTK, Japanese startup preferred networks Chainer, and Theano led by the University of Montreal, Canada, and other early preliminary frameworks have exited the stage of history by merging with mainstream frameworks or simply stopping updates.

- **The current mainstream landscape**

Google-developed TensorFlow continues to sit at number one, relying on its industrial deployment advantage, with more than three times the market attention of second-place PyTorch. Meta's PyTorch (merged with Caffe2) is rapidly surging with its ease of use, with a dramatic increase in the number of applications and more than 50% of papers at major top academic conferences, and Baidu's launch of China's first The open source framework Flying PaddlePaddle has both efficiency and flexibility, and has the momentum to catch up.

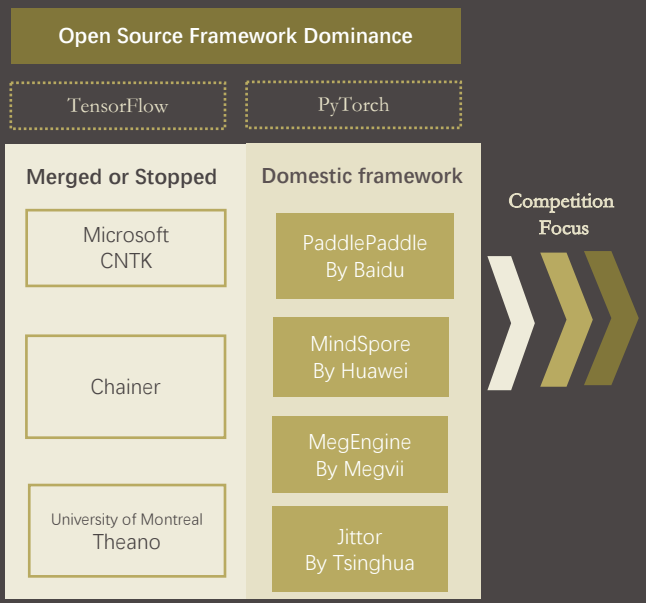
- **Future pattern**

Open source + support form industry giant has become the mainstream model of deep learning software framework: each platform has its own strengths in terms of stability, debugging difficulty, execution speed, memory consumption, etc. The mainstream framework is generally intervened by the industry's head enterprises and lead the internal application and build.

1.2.2.3 The Competiton Focus

1. The high-level language interface encapsulates the key model building, training and other functions in the back-end framework, reducing the R&D threshold. At present, the three mainstream frameworks accelerate binding or building high-level language interfaces, and have emerged to cooperate to encircle the land; TensorFlow and Keras form an exclusive cooperation to enhance the framework's ease of use and competitiveness, and rival PyTorch, which has recently been rapidly rising in status with the advantage of ease of use; MXNet is united with Gluon, jointly maintained by Amazon and Microsoft; PyTorch uses PyTorch uses Torch and Caffe2 as its back-end framework, with an inherently high-level language interface; Baidu's PaddlePaddle has a "dynamic and unified" programming model that balances flexibility and efficiency.
2. Hardware adaptation optimization tries to solve the problem of complex adaptation and uneven performance caused by diverse hardware compilation tools, and unified compilation tools and compilation languages have become the focus of the layout of mainstream open-source development frameworks. At present, Google and Facebook accelerate the construction of a unified compilation language (IR), trying to guide hardware manufacturers to take the initiative to adapt and gain the right to speak about framework adaptation.

Competition Focus on Deep Learning Frameworks



High-level language interface ease of use

At present, the three mainstream frameworks to accelerate the binding or build high-level language interface, has appeared to cooperate the phenomenon of encirclement. TensorFlow and Keras to form an exclusive cooperation, MXNet and Gluon joint, PyTorch to Torch and Caffe2 internal innate construction of high-level language interface, Baidu fly paddle PaddlePaddle is equipped with The "dynamic and unified" programming model.

Hardware adaptation optimization

Hardware adaptation optimization tries to solve the problem of complex adaptation and uneven performance caused by diverse hardware compilation tools, and unified compilation tools and compilation languages have become the focus of the layout of mainstream open source development frameworks. At present, Google and Facebook accelerate the construction of a unified compilation language (IR), trying to guide hardware manufacturers to take the initiative to adapt, to obtain the framework adaptation of the right to speak.

Source: Frost Sullivan, LeadLeo

1.3 AI Training platform

1.3.1 Resource allocation

Based on fits to actual data, AI computation is growing at least 10 times per year, at a rate well beyond the 18-month doubling of Moore's Law, so the ability to adjust task resources in deep learning training becomes particularly important.

At this stage, as the cluster size increases, the probability of machine failure at a given moment in the cluster is increasing. And as the complexity of the training model increases, the training resources and training time both increase significantly, and the fault tolerance of the task decreases. In addition, the increase in cluster size makes the waste of idle resources non-negligible, and the flexibility of cluster resource allocation is in constant demand.

1.3.2 Platform Functions

- Managing multiple training servers, especially high-performance computing servers with GPUs, which can distribute training tasks to distributed computing nodes to perform computations.
- Integrating multiple training frameworks, abstracting the training process, providing a web interface, uploading and specifying relevant data and parameters, and starting training tasks and monitoring and analyzing the training process.
- Pooling computing resources, especially GPU resources, to make a "GPU cloud". When a training task is started, the platform will automatically assign the training task to the appropriate GPU.
- Connecting to the data center so that data from the data storage platform can be directly imported to the training nodes.
- Isolating the resources and environment in the computing nodes, compatible with different models of GPUs, different versions of CUDA/CuDNN and different deep learning frameworks.

1.3.3 Distributed Training

1.3.3.1 Principle of distributed training

Distributed training can provide elastic allocation of underlying resources and improve the resource utilization of the system. For example, the Baidu Flying Paddle Universal Heterogeneous Parameter Server can slice and dice tasks, allowing users to deploy distributed training tasks in hardware heterogeneous clusters to achieve efficient utilization of chips with different computing power, providing users with higher throughput and lower resource consumption training capabilities.

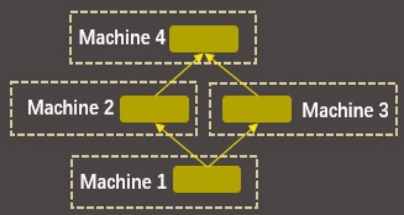
However, there are major obstacles to the application of distributed training. It takes a lot of work to implement the module of elastic control on each framework and to adapt the corresponding scheduling system to achieve elastic training. In addition, if different frameworks have their own resilient training solutions, integrating different framework solutions at the AI development platform level requires a high maintenance cost.

Elastic distributed training is the trend of AI development platform service, which can realize the experience of reducing cost and increasing efficiency for users: when users need a large amount of computing resources to expand capacity, improve arithmetic power and stability, and reduce model training time; when user demand is small, reduce the underlying resource allocation, and reduce the service cost for customers due to resource occupation.

1.3.3.2 Distributed training framework mode

❑ Model parallelism:

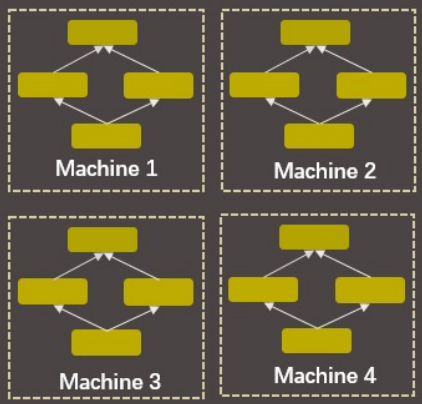
Splitting a model into multiple smaller models on different devices, with each device running a part of the model. Also due to the frequent communication required between models of different devices, this model is generally not used because of low efficiency.



Model Parallelism

❑ **Data parallelism:**

Each GPU has the complete model, while dividing the data into multiple copies to each model, and then each model is fed with different data for training. This model is currently the most commonly used distributed training method.



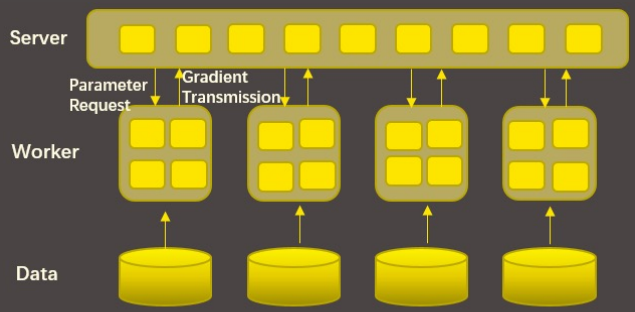
Data Parallelism

Source: Frost Sullivan, LeadLeo

1.3.3.3 Distributed training architecture

❑ *Parameter server architecture:*

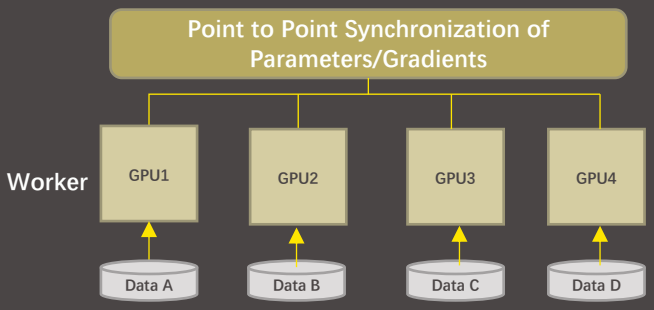
- Mostly used for the training task of large-scale sparse feature models in search recommendation scenarios.
- This architecture is a centralized architecture, which adopts the centralized management of model parameters to achieve the update and distribution of model parameters.
- The parameter server architecture has two roles, Server and Worker. Server and Worker do not necessarily correspond to the actual hardware, while Server is responsible for the slice storage and update of parameters, and Worker will keep the complete model network structure for performing forward and backward computation of the model. The Worker node of the regular parameter server needs to use a uniform model CPU or GPU machine to complete the model training.



Parameter Server Structure

❑ *Collect architecture:*

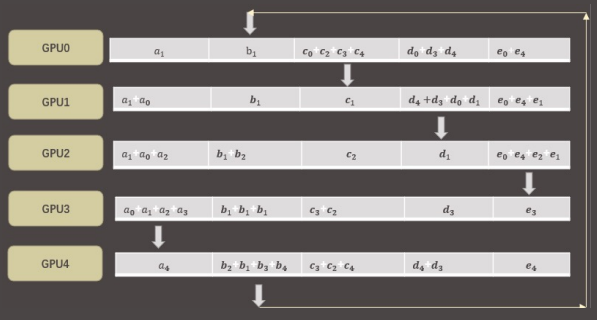
- It is often used for model training task scenarios requiring complex network computing such as vision and natural language processing;
- It is a decentralized and recently popular distributed training architecture. Each device is a worker, which also represents a training process. Each worker is responsible for the training of the model and also needs to have the latest global information..



Collective Structure

❑ *PyTorch- Ring AllReduce Architecture*

- This architecture's operational efficiency increases linearly as the number of workers increases.
- This architecture uses a ring structure, and there is no central node server in this architecture, only worker nodes exist.
- In this architecture, each worker has a complete copy of the model parameters, and performs gradient calculation and update.

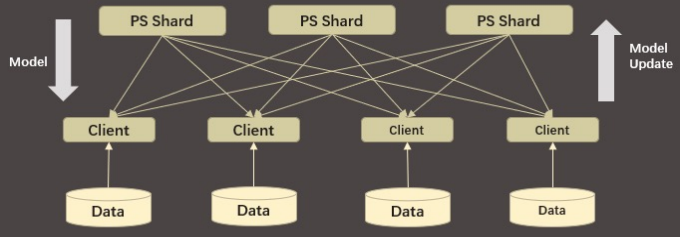


Ring AllReduce Architecture Operation Schematic Diagram

1.3.3.4 Common distributed training frameworks

❑ *TensorFlow-Parameter Server Framework*

The parameter server architecture mainly includes 1 to multiple server nodes and multiple worker nodes. The server node saves the model parameters, and the worker is responsible for using the parameters on sever and the data on this worker to calculate the gradient.



Schematic Diagram of Parameter Serve Architecture

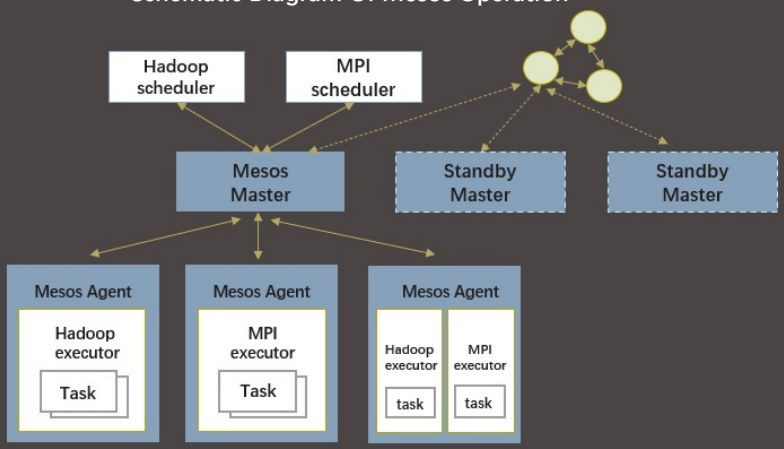
Source: Frost Sullivan, LeadLeo

1.3.4 Key technologies and framework

❑ **Mesos+Marathon**

Mesos is an open-source distributed resource management framework under Apache. It is called the kernel of distributed system and also called the operating system of data center. Mesos implements a resource management framework. It manages cluster resources (CPU, GPU, RAM, etc.) at the data center level, and enables resources allocation and tasks schedule. To further isolate resources and tasks, Mesos abstracts specific task scheduling capabilities to specific frameworks for implementation, such as Hadoop, Spark, MPI and Marathon.

Schematic Diagram Of Mesos Operation



❑ **Docker**

- **Environment isolation**

Docker isolates the system environment and execution environment, that is, it isolates the environment of different training tasks on the same server, so as to distribute the same task to servers with different GPU cards, or it can run multiple tasks with different CUDA versions and different depth learning frameworks on the same server at the same time.

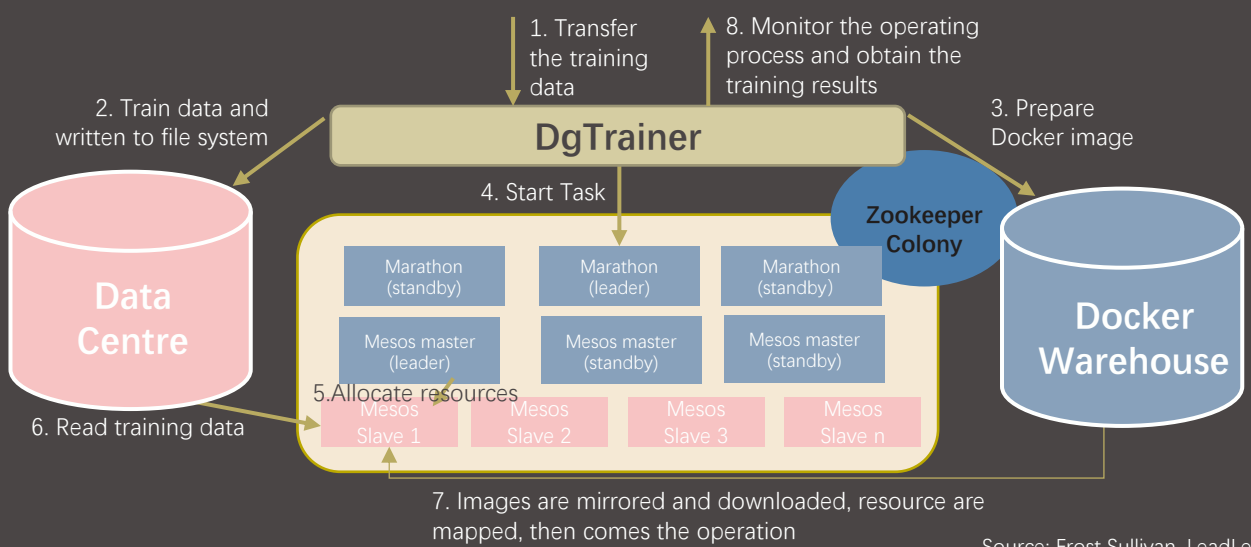
- **Resource isolation**

Dockers can isolate hardware resources to meet task requirements and avoid vicious and unordered competition between multiple tasks and resources. However, Docker's management of GPU resources is not as complete and mature as CPU.

- **Code sharing**

Through Github+CI+Docker, code from different repos and branches can be packaged into Docker images to complete different tasks, thus achieving more flexible and fine-grained sharing.

Training Framework Flow Chart

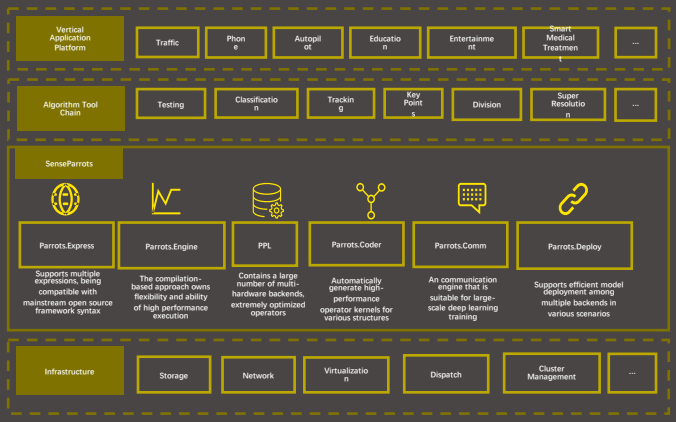


Source: Frost Sullivan, LeadLeo

SenseTime SenseParrots

- SenseParrots is different from the open source training platform of Facebook and Google. It has super computing network training, super large-scale data set training, and super large end-to-end complex application ability training.
- SenseParrots' contribution to computing power is to provide the corresponding software system on the existing GPU and build the corresponding system, so that the effectiveness of the GPU can be fully exerted and the overall R&D efficiency can be greatly improved. The training of the same scale took a few hours to complete a few years ago, and could be completed in 90 seconds on the platform of SenseTime.

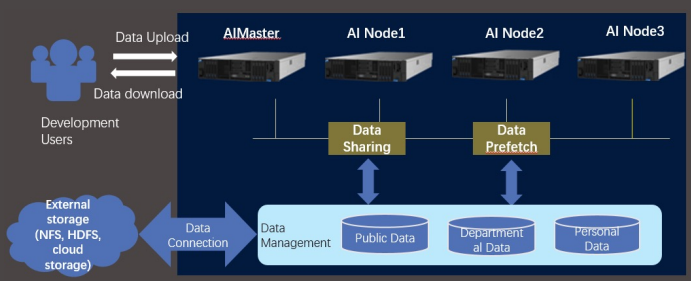
SenseParrots Product Introduction



Inspur AIStation Training Platform

- AIStation is an AI development resource platform of Inspur for AI enterprise training scenarios.
- Realize containerized deployment, visual development, centralized management, etc., provide users with extremely high-performance AI computing resources, achieve efficient computing support, accurate resource management and scheduling, agile data integration and acceleration, and process oriented AI scenarios and business integration, effectively open up the development environment, computing resources and data resources, and improve development efficiency.

AIStation Product Introduction



1.3.5 Algorithm upgrading

The algorithm is an association node between AI and big data. Social media, location technology, search engines and other Internet applications generate and store large amounts of data in real time. On the basis of massive data, AI continues to infer users' interests, preferences and needs, generate different user portraits, and achieve personalized and precise customization of digital culture from production, dissemination to reception.

At this stage, AI training platform has or will integrate a variety of AI technologies, such as computer vision, natural language processing, cross media analysis and reasoning, intelligent adaptive learning, swarm intelligence, autonomous unmanned system, brain computer interface, etc:

- **Computer vision technology**

Identify, track and measure the target through cameras and computers instead of human eyes, and perceive the environment in three dimensions.

- **Natural language processing technology**

Aanalyze, understand and process natural language by establishing a formal computing model.

- **Cross-media analysis and reasoning technology**

Collaborative and comprehensive processing of multiple forms, such as text, audio, video, image and other mixed and coexisting composite media objects.

- **Intelligent adaptive learning technology**

Simulate the one-to-one teaching process of teachers and students, and endow the learning system with the ability of personalized teaching.

- **Group intelligence technology**

The process of gathering multiple opinions and transforming them into decisions to reduce the risk of a single individual making random decisions.

- **Autonomous unmanned system technology**

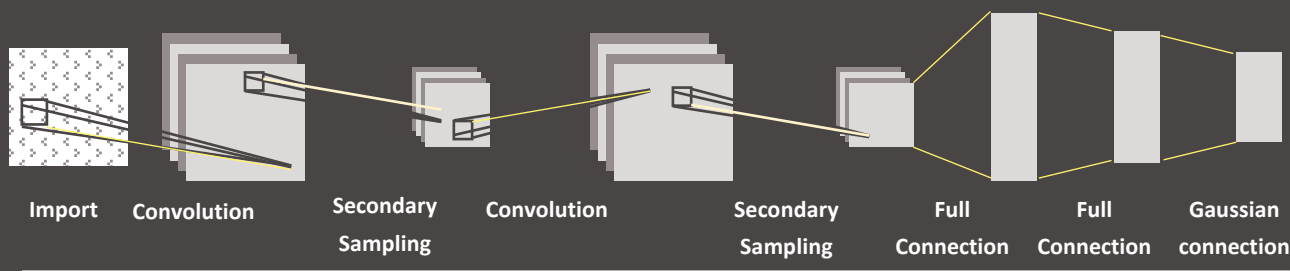
A system operated or managed by advanced technology without manual intervention.

- **Brain computer interface technology**

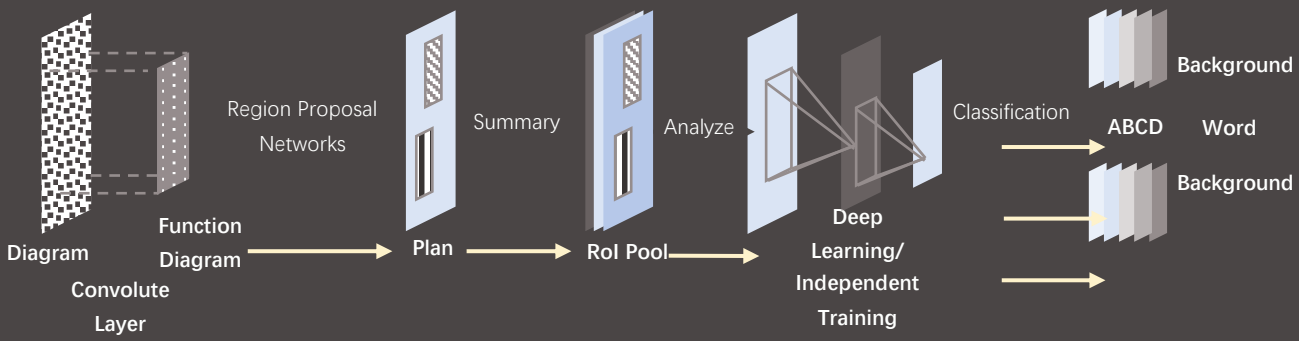
A direct connection between human or animal brain and external devices to complete information exchange.

Source: Frost Sullivan, LeadLeo

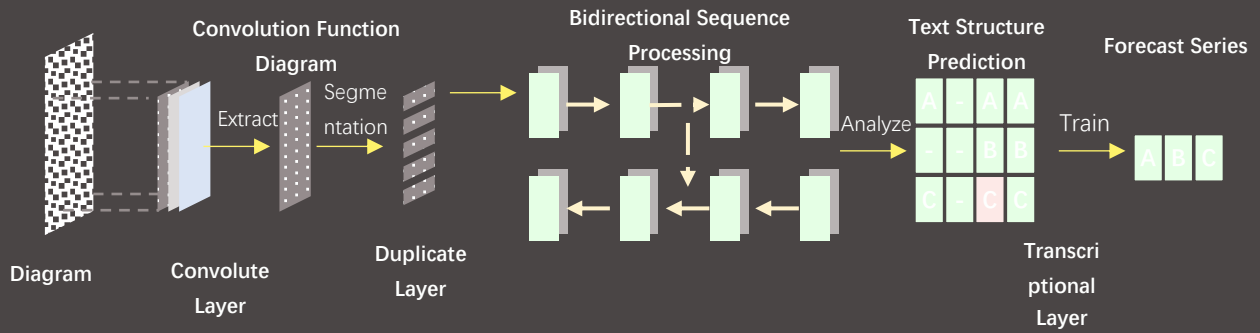
Principle of Image Preprocessing Technology



Principle Of Text Detection Technology



Principles of Text Recognition Technology



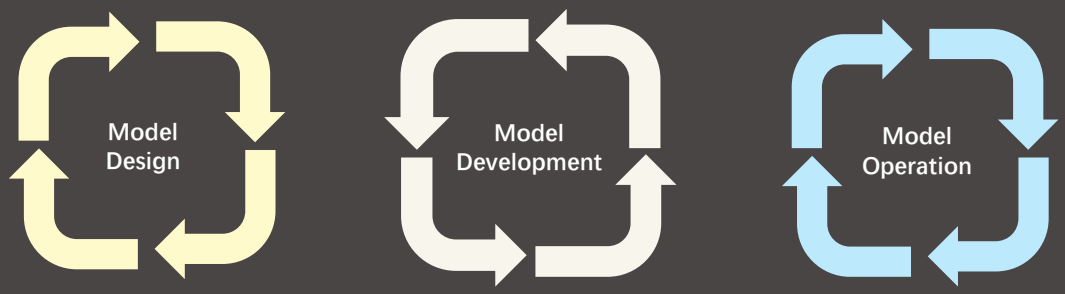
- With the practical implementation of AI learning methods in financial, medical, social and other scenarios, the nurturing of large amounts of data will continue to improve AI training algorithms. For example, in a paper of CVPR 2021, a new convolution layer named Skip Convolutions is proposed, which can subtract the two frames before and after, and only convolve the changed parts;
- In the image preprocessing technology, the neural network based on CNN is used as a feature extraction method, and the strong learning ability of CNN can also enhance the robustness of feature extraction in AI model; The FrameExit network composed of multiple cascaded classifiers can change the number of neurons used in the model according to the complexity of the video frame, that is, when the difference between the front and back frames of the video is large, AI will use the whole model to calculate, while when the difference between the front and back frames is small, AI will only use a part of the model to calculate.

Source: Frost Sullivan, LeadLeo

1.3.5. Technical service: MLOps improves team collaboration efficiency

- Along with the development trend of industrial intelligence, AI is becoming a common technology for transformation and upgrading in many industries. Currently, the most mature and widespread application areas of AI include public security, transportation, finance, education, etc. The demand for AI applications in other industries is highly fragmented and the scenarios are diverse, but the demand for AI applications still exists widely. The AI development platform provides cloud-based natural language understanding, automatic speech recognition, visual search, image recognition, text-to-speech conversion, and machine learning hosting services for different application scenarios, and provides developers or enterprise users with convenient operations for building advanced text and voice chatbots and intelligent machine learning applications.
- For individual or enterprise developers, development time and development cost are the main metrics to consider when building AI applications. With cloud-native and elastic distributed computing architecture, users can reduce cost and increase efficiency at the training and inference level of AI models, and with MLOps, the development and deployment efficiency of teams will be significantly improved.
- MLOps is DevOps for ML. machine learning (ML) models built by data scientists need to work closely with other teams (business teams, engineering teams, operations teams, etc.). MLOps brings flexibility and speed to the system: MLOps reduces development time and delivers high-quality results through reliable and effective ML lifecycle management; MLOps carries over from DevOps to continuous development (CD), continuous Continuous Development (CD), Continuous Integration (CI), Continuous Training (CT), and other methods and tools carried over from DevOps guarantee the repeatability of AI workflows and models, allowing developers to easily deploy high-precision machine learning models anytime, anywhere and integrate management systems to continuously monitor machine learning resources.
- MLOPs also place higher demands on the platform in terms of data and hyperparameter version control, iterative development and experimentation, testing, security, production monitoring, infrastructure, etc. MLOps platform data plays an equally important role in defining output as written code, thus increasing data complexity compared to DevOps platforms. In response to the challenges faced by the MLOps platform, the MLOps implementation process includes five phases: use case discovery, data engineering, machine learning pipeline, production deployment, and production monitoring, and its workflow is mainly implemented through an agile approach.

MLOps Definition：MLOps=ML+DevOps

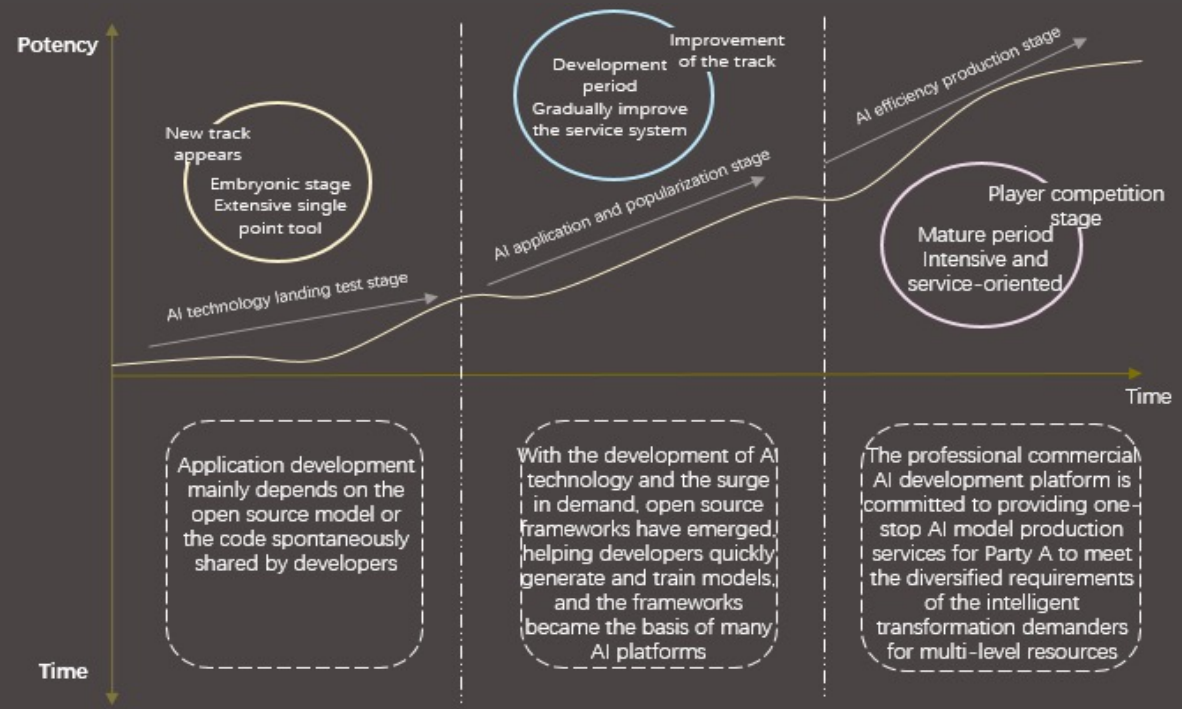


Source: Frost Sullivan, LeadLeo

According to the development history of AI and the technical characteristics of AI development platform, the overall efficiency level of AI foundation layer resources is constantly evolving. The development of AI development platform is basically consistent with the development of AI foundation layer, which can be roughly divided into three development stages: Embryonic stage, Development stage and Maturity stage：

- **The embryonic stage**:
  - Extensive single point tools appear in the AI foundation layer, and the industrial chain is gradually clear.
  - **Development support**: GPU support model training requires AI computing power. AI accelerated landing catalyzes data annotation and other industry weeks. AI basic algorithm capability is output through API, which mostly depends on manual design and development.
- Development period:
  - The AI service system is gradually improved, and the market starts to explore product forms and business models.
  - **Development support:** AI computing power providers gather IT basic resources to form a resource pool, and demand drives product optimization. Algorithm players continue to deepen vertical development and AI engineering capabilities, helping downstream customers reduce development costs and improve model production efficiency.
- Maturity period:
  - the level of tool intelligence has been improved, and the market has entered the stage of player competition.
  - Supporting points for development: the enterprises at the basic level of the algorithm, computing power and data tracks provide refined solutions, actively participate in the extension of the value chain, and shape the core competitiveness; The AI supplier layout is a one-stop AI model development platform covering the whole process of data governance, model development and computing resource management.

Development History Of AI Development Platform



Source: Frost Sullivan, LeadLeo

Chapter 2 AI Platform Business Model

With the gradual expansion of the scale, the average cost of a single customer of the AI development platform will decrease significantly, and the service profit margin will gradually increase

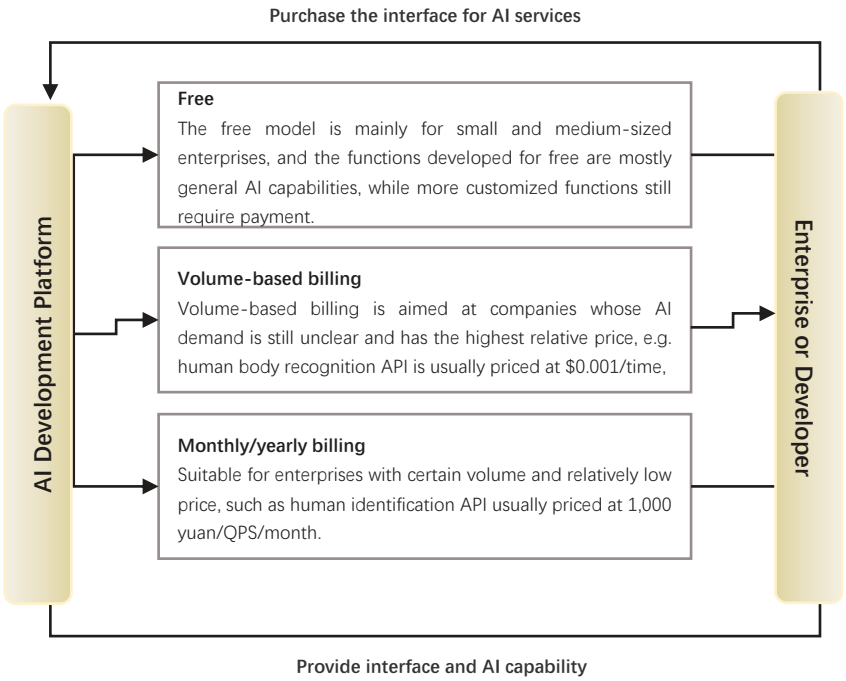
2 Business Model

With the gradual expansion of the scale, the average cost of a single customer of the AI development platform will decrease significantly, and the service profit margin will gradually increase. Therefore, realizing large-scale operation is an important development strategy of the AI development platform, which can help the platform reduce costs while giving the platform greater bargaining space. This phenomenon also explains the bottom business logic that large manufacturers can still make profits under the "partially free" mode, and also reflects the market competitive advantage of large manufacturers compared with small and medium-sized manufacturers.

"AI development platform business model is relatively simple, developer volume and platform scale become its revenue decisive elements."

- The AI development platform business model is to profit from providing AI technology interfaces or AI development tools for enterprises or developers, and the billing methods mainly include free, per-call, annual or monthly.
    1. The free model provides enterprises or developers with common and general AI technology interfaces such as text recognition and face recognition, with a usage limit, usually 1-5QPS/day, mainly for small and medium-sized enterprises with low usage. The free model achieves profitability through data accumulation, building AI ecology and providing additional services.
    2. Compared with annual and monthly billing, the volume-based billing is higher, which is suitable for enterprises with unclear demand.
- In terms of product marketing, platform operators can improve the conversion rate of traffic through free trial, subsidies, online teaching, etc. Large platforms can further improve the conversion of traffic to users through permanent free universal products. Platform operators can also explore other value-added needs of users in customer service, such as cloud services, customized AI development solutions, etc.

AI Development Platform Business Model



Source: Frost Sullivan, LeadLeo

Chapter 3 AI Platform Market Size

From 2016 to 2020, market size of China's AI development platform expanded rapidly. In 2021, market size of China's AI development platform exceeded 23.5 billion yuan.

3 Market Size

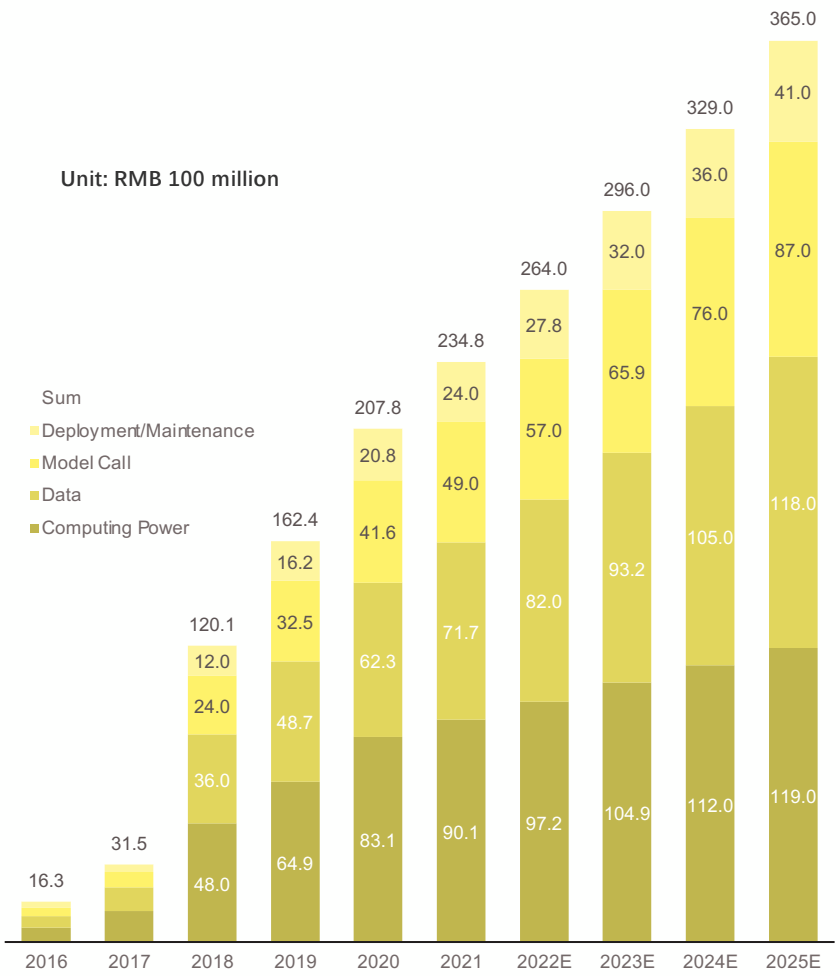
- From 2016 to 2021, market size of China's AI development platform expanded rapidly. In 2021, market size of China's AI development platform exceeded 23.5 billion yuan.
- With the background of policy support, industry penetration and chip performance, the market size of China AI development platform is expected to reach 36.5 billion yuan in 2025.

> In 2021, market size of China ‘s AI development platform is 23.5 billion yuan, and China ‘s AI development platform’ s market size in 2025 is 36.5 billion yuan.

China AI Development Platform Market Size (by Revenue)

Forecast to 2025

CAGR	2016-2021	2022E-2025E
Sum	**56.0%**	**8.4%**
Computing Power	58.0%	5.2%
Data	57.5%	9.5%
Model Call	56.8%	11.2%
Deployment/Maintenance	45.8%	10.2%



Source: Frost Sullivan, LeadLeo

Chapter 4
Competitive Elements

In this chapter, Frost & Sullivan analyzes the core competitiveness of AI development platform in the market, which is divided into hard power of "improving data processing capability" and soft power of "enhancing platform usability" and “increasing ecological openness ".

4 Competitive Elements

The users of AI development platforms are individual or developers of enterprises in the AI industry, and the core competition of AI development platforms will focus on how to provide developers with a more efficient and convenient development platform and other derivative services. Sullivan summarizes the core competencies of AI development platforms as the hard power of "improving the capability of data processing" and these two soft powers of "enhancing the ease of use of the platform" and "improving the openness of the ecology".

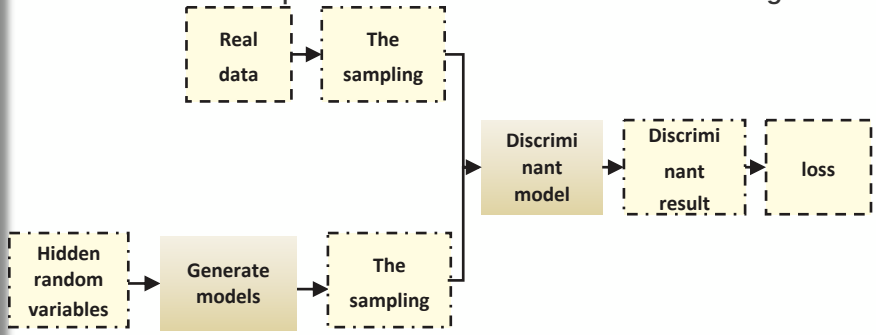
> **"The core competition of AI development platforms will focus on "improving capability of data processing", "enhancing ease of use of platform", and "improving openness of ecology"."**

AI development platform providers offer more effective AI development platform services for developers with platform hardware, algorithm models and other capabilities as the underlying support.

The 1st hard power: Intelligent annotation -- a difficult breakthrough from "artificial" to "intelligent"

- The intelligent replacement of data annotation is extremely difficult. At this stage, with the help of algorithms, annotation tools can already complete the basic annotation work, such as automatic recognition of labeled frames and recognition of speech, while the algorithms of annotation tools are being continuously developed and optimized.
- For AI development platforms, intelligent annotation is of high importance in optimizing the efficiency of self-developed algorithms and user experience. The available intelligent annotation for AI development platforms include introducing GANs to optimize the annotation effect, using semi-supervised learning mechanisms to solidify annotation, introducing difficult case screening mechanisms to optimize annotation results, and providing data annotation based on difficult cases to improve recommendations. However, in the actual application process, providers still need to address the limitations of the above approaches.
  - **GANs :** During the training process, the generator and discriminator will need high synchronization, but in the actual training, it is easy to generate scenarios where the discriminator converges and the generator diverges, which also requires extremely high standards for the optimization of the discriminator and generator; GANs will have the problem of model missing, during the training process, which means that the generator function degrades to generate the same sample points, resulting in the inability to continue the deep learning process.
  - **Semi-supervised learning:** Difficulty for models to correct itself; problems of over-smoothing, resulting in indistinguishable features of nodes.
  - **Hard example selection mechanism:** The hard cases can only be generated in the process of model training, which means that offline hard case mining is not possible, and users must modify themselves to use the online hard case mining function. The core of the difficult case screening mechanism is to generate the difficult case set by bootstrapping, and the generation method is only judged by the loss value of the training samples during training, which is a single dimension for judging and cannot guarantee the improvement effect of the model accuracy; the algorithm idea is not mature enough to form a systematic scheme.

Flow chart of production adversarial network GANs algorithm



Source: Polar Community, easyAI, Huawei Cloud, Frost & Sullivan , LeadLeo

The 2nd hard power : Machine Learning framework – improve framework defects, enhance user experience, and build AI ecology

- TensorFlow and PyTorch are the dominant machine learning frameworks, with large developers, and many mature and available code. These two have over 90% of global market share of deep learning frameworks. However, TensorFlow and PyTorch have different features from each other.

TensorFlow:

- Pros: Suitable for industrial production environment; complete solution for both model training and deployment.
- Cons: Has too much different styles of APIs; newbie unfriendly; unclear updating ideas for distributed training; low support for cloud-native.

PyTorch:

- Pros: Simple, intuitive and understandable API style of programming; based on the deep learning model built by dynamic computational graphs, developers can debug quickly based on the stack information.
- Cons: The deployment of the ecology is not yet completed, and some services are not supported.

- The limited number of developers is a uniform shortcoming of open-source machine learning frameworks developed by Chinese providers, with a significant gap in developers compared to TensorFlow and PyTorch, and only support Chinese and English. In contrast, TensorFlow and PyTorch support some minority languages, so the developer ecosystem is more complete.
- The global ecology of machine learning frameworks has basically stabilized. The general-purpose frameworks, such as TensorFlow and PyTorch, were open-sourced earlier and thus have an ecological advantage. The framework developed by Chinese providers themselves optimizes the framework architecture in terms of technical iterations of machine learning, and flaws of TensorFlow and PyTorch. Meanwhile, Chinese providers' self-developed frameworks can provide a better experience for developers. In the long term, the developer ecosystem of Chinese providers' self-developed frameworks is mostly concentrated in China, and more companies will use Chinese providers self-developed frameworks for machine learning in the future, but the global machine learning framework market is expected to remain dominated by TensorFlow and PyTorch.
- Baidu, Huawei and other Chinese providers launch machine learning self-developed framework, such as PaddlePaddle and MindSpore.

PaddlePaddle:

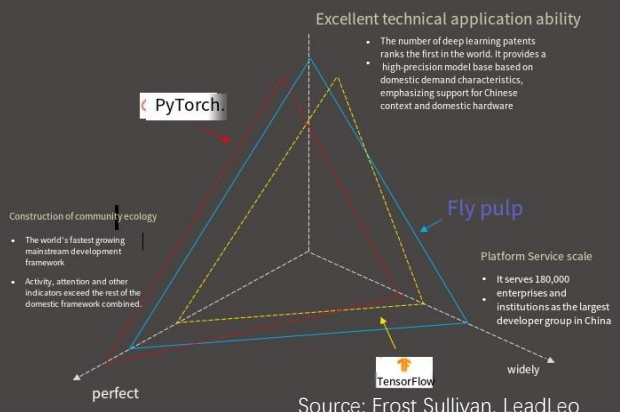
- Pros: Active community, complete ecosystem, user-friendly application, full-process support, fast-updating, supports large-scale asynchronous distributed training
- Cons: Individual developers dominate, not deployed by large-scale companies.

MindSpore:

- Pros: Support visual boosting, differential privacy, second-order optimization, graph neural networks, quantization training, hybrid heterogeneity, MindSpore Serving, PS distributed training, MindIR, debugger; support multi-platform; advocate software and hardware collaborative design; support multiple modes of distributed training.
- Cons: Small number of users in the community; some functions need to be improved.

- Along with the maturity of technology, industry and policy, AI has crossed the period of accumulation of technology theory and construction of tool platform, and started to enter the golden decade of industry empowerment with the goal of large-scale application and high-value release. With the implementation of AI, the up-and-down extension and construction of intelligent ecological platform, based on deep learning framework, will become the common choice of domestic and foreign technology top companies.

ML frameworks compete for elements

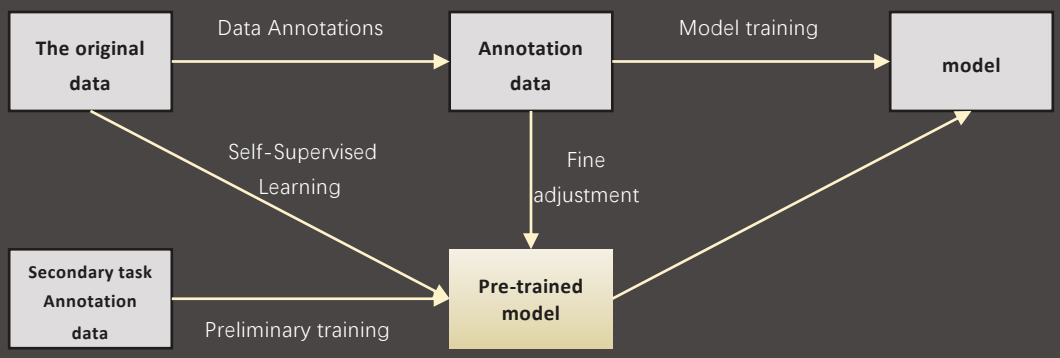


Source: Frost Sullivan, LeadLeo

The 3rd hard power: Pre-trained models -- both "large" and "small"

- In the future, providers of AI development platforms will release a variety of pre-trained models, including computer vision, speech recognition, speech synthesis, natural language processing, machine translation, artificial intelligence recommendation, business analytics & prediction, scientific computing, multimodal data tasks, and composite tasks. Pre-training models will also evolve along multi-technology paths.
  - **The scale of pre-training models will be increased (" large"):** large-scale pre-training models will include over 100 billion parameters, and the cost of a single training is expected to be over $10 million. Therefore, the pre-trained model will be equipped with optimized training techniques, including mixed precision training, data parallelism, model parallelism, Lamb optimizer, 3D parallel training, and sparse attention acceleration. However, this pre-training model process is tedious and can only be laid out with cloud-based applications.
  - **The pre-trained model will improve its flexibility by compression and acceleration ("small"):** the model can be compressed by pre-training language based on knowledge distillation and knowledge pruning, or it can be compressed by matrix parameter decomposition, parameter sharing and model structure design & search, which means removing the redundant parts of the parameter matrix and making the model "smaller". In other words, with quantization-based model compression for pre-trained languages, the value reduced by compression represents the required bit , such as reducing the computation of a 32 bits to 4 bits, in order to simplify the computing process. The compressed pre-trained models can be applied on the device with great value.
- In the future, AI development platform providers need to continuously optimize the original training methods for Resnet50-v1.5, SSD-ResNet34, 3D UNET, RNNT, Openpose, YOLO, BERT, DLRM and other training models, in order to accelerate the training speed and propose new training methods, for the purpose of improving the maturity of pre-trained models.

Concept map of the pre-trained model



Source: Frost Sullivan, LeadLeo

AI development platform vendors improve developers' experience by providing flexible services and simplifying operation of developers, which enhances platform usability and increases ecological prosperity, increase vendors 'technology level.

Soft power 1: AutoML -- lower the threshold of AI development and increase the efficiency of AI development

- AutoML is one of the most important trends in artificial intelligence. AutoML will be able to integrate iterative processes into traditional ML to build an automated process that dramatically lowers the bar for ML: AutoML is a ML process that automates everything from data selection to modeling through a series of algorithms and heuristics. When researchers input meta-knowledge (such as convolutional process and problem description), this algorithm can automatically help them select appropriate data, optimize the model structure and configuration, train the model, and deploy it to different devices.
- AutoML can help AI development platforms automatically complete tasks such as neural architecture search, model selection, feature engineering, hyperparameter optimization, and model compression. Classification or regression problems that rely on structured or semi-structured data can be solved automatically by AutoML, significantly improving the efficiency of AI training.
- However, there are still some difficulties to be solved in the development path of AutoML. First, AutoML still requires a large amount of computing power, and companies still need to try more solutions in the R&D process. Secondly, AutoML needs to increase processing complexity while maintaining some transparency to allow users to confirm the model's quality. Morever, AutoML, as an automation tools, has limitations in resource optimization and updating, complex model processing and feature engineering, while improving work efficiency.

Soft power 2: Developer-centric -- enhance platform service capabilities to build a technology ecosystem

- AI development platform is a developer-oriented service. Therefore, the degree of satisfying developers' needs, compatibility, usability, and development experience are also important criteria for evaluating AI development platform.
  - For data preparation, AI development platform can provide local datasets, third-party open-source datasets, cloud datasets and other data access methods. The platform also provides multi-type data annotation services, and allows data visualization in the operation panel.
  - For model training, AI development platforms can improve the compatibility of machine learning frameworks, programming languages, and cloud-based IDE , and provide customized and modular algorithm modification methods. At this stage, mainstream AI development platforms can support model management services, such as elastic training, real-time monitoring of resources, and heterogeneous training of hardware devices, providing developers with convenient AI development services.
  - For model management and deployment, R&D direction of AI development platform covers platform compatibility, such as support for more programming languages, CI/CD support, third-party AIOps tools, and machine learning workflow building services.AI development platform supports model deployment and monitoring services, such as model drift monitoring, resource load monitoring, automatic alarms, and monitoring metrics visualization.
  - For account management and services, some mainstream AI development platforms choose to open some free resources, such as computing resources, storage resources, dataset resources, and model resources. In addition, most AI development platforms provide hours billing, prepaid, subscription and other payment methods, thus enhancing the flexibility of charging to meet the needs of different types of developers.

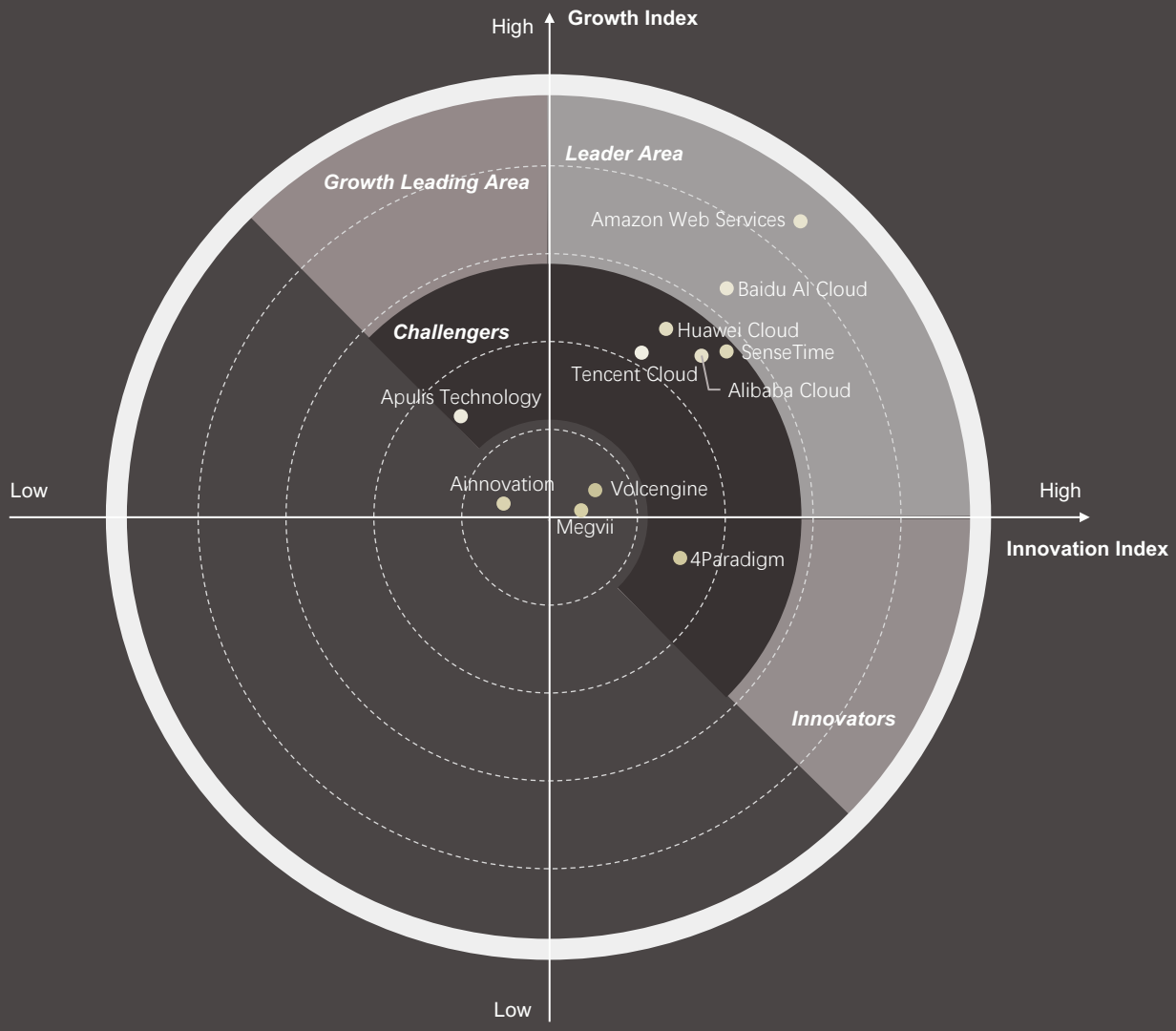
Source: Frost Sullivan, LeadLeo

Chapter 5
Comprehensive Performance

The conclusion of this report on the comprehensive competitiveness performance of AI development platform application products and services of providers is only applicable to the development of its application market at this stage.

Competitiveness Analysis of China's AI Development Platform Market in 2021

Frost Radar ™



Note: The circle diagram above measures the competitiveness of firms, in which more competitive firms are in the outer ring and less competitive firms, inner ring. The competitiveness is measured by "innovation index" and "growth index".

China's AI development platform application market is in a steady growth phase. The analysis and conclusions of this report are only applicable at current stage.

- **Vertical Axis: "Growth Index"**

Growth Index measures the product structure, functionality, and growth potential. Higher position represents stronger growth capability in AI development platform application.

- **Horizontal Axis: "Innovation Index"**

Innovation Index measures the innovation capacity of firms. The more to the right, the richer the service function and stronger product optimization ability.

> **"China's AI development platform market is in the stage of technological maturity and platform improvement, where competitors have advantages in both innovation and growth capabilities."**

Source: Frost Sullivan, LeadLeo

Chapter 6
Dimensitions of Scoring

In this chapter, Frost & Sullivan Consulting (China) scores AI development platforms based on two dimensions: innovation index and growth index.

6 Dimensitions of Scoring

> "In this report, the growth index evaluation system is set up to evaluate and analyze the artificial intelligence development platform, which consists of two indexes: technological innovation ability and business innovation ability"

Evaluation Dimension of China's AI Development Platform -- Innovation Index

Innovation Index		
Level Indicators	**The Secondary Indicators**	**Main Points**
Technological innovation capability	Hardware	Evaluate the performance of vendors in multiple dimensions such as AI chip and AI server
	Data Collection and Annotation	Evaluate vendor's performance in multiple dimensions such as intelligent data collection, annotation, analysis, screening, and enhancement
	Deep Learning Framework	Evaluate vendor's performance in multiple dimensions such as R&D and innovation of deep learning framework
	Algorithm Model	Evaluate vendor's performance of algorithm model in multiple dimensions, such as algorithm accuracy and operation efficiency, and scene universality
Business innovation capability	Cloud Native Architecture	Evaluate vendor's performance in applying cloud native architecture, and optimizing cloud development environment
	Cloud Security Governance	Evaluate vendor's performance in applying cloud security technology and improve the comprehensive security performance of the platform
	Big Data Management	Evaluate vendor's performance in applying big data management technology and enhance the value of the platform database
	Visual Development	Evaluate vendor's performance in visualizing, developing functional suites with zero barriers, simplifying the application development process, and lowering the application development barriers

Source: Frost & Sullivan , LeadLeo

6 Dimensitions of Scoring

> "In this report, the growth index evaluation system is set up to evaluate and analyze the AI development platform, which consists of two indexes: service capability and ecological capability"

Evaluation Dimension of China's AI Development Platform -- Growth Index

Growth Index		
Level Indicators	The Secondary Indicators	Main Points
Service capability	Computing Power Service	Evaluate vendor's computing power performance in multiple dimensions, such as computing power deployment and management ability
	Data Service	Evaluate vendor' s performance in data sample size, data product compatibility and data product diversity
	Algorithm Service	Evaluate vendor' s performance in multiple dimensions, such as deep learning framework compatibility, algorithm model product diversity, adaptation and portability
	Platform Service	Evaluate vendor' s performance in platform service such as data management and security, AI development management and AI application deployment
	Pricing Strategy	Evaluate vendor' s performance in pricing strategy such as flexibility and elasticity of service pricing mechanism
Ecological capacity	Ecological Prosperity	Evaluate vendor' s performance in community prosperity, market influence, application breadth and other dimensions
	Ecological Development	Evaluate vendor' s performance in ecological development such as ecological sustainability and external threat resistance

Source: Frost & Sullivan , LeadLeo

Chapter 7
Leader Vendors Cases

In this chapter, Frost & Sullivan introduces China AI Development Platform providers.

Amazon Web Services

For more than 15 years, Amazon Web Services has been known for its technological innovation, abundant services and broad adoption.
Amazon Web Services continues to expand its portfolio of services to support virtually any workload on the cloud, and now offers more than 200 full-featured services across computing, storage, database, network, data analytics, Machine Learning and AI, IoT, mobile, security, hybrid cloud, virtual and augmented reality, media, and application development, deployment and management; with infrastructure in 96 availability zones across 30 geographic regions and has announced plans for 5 new regions and 15 availability zones in Australia, Canada, Israel, New Zealand and Thailand.
Millions of customers around the world, including fast-growing startups, large enterprises and leading government agencies, trust Amazon Web Services to support their infrastructure, improve agility and reduce costs with Amazon Web Services' services.

Unique Advantages：

- **Broad and abundant cloud service**
- **Various customer practice**
- **Global infrastrcture**
- **Leading security compliance**
- **Trustworthy partner**

> **“AWS aims to bestow ML on developers by providing a highly compatible, highly functional and modular AI development platform service”**

- Amazon Web Services has a full-stack supply of hardware and software for AI development, including infrastructure, AI platforms, out-of-the-box solutions for various scenarios. Integrated with the series of cloud services, Amazon Web Services has the ability of satisfying the diversified demands of various types of customers.:

1. **AI Infrastructure：** Self-developed ML inference chip, Amazon Infrentia, together with the ML training chip, Amazon Trainium, enable end-to-end machine learning hardware acceleration from inference to training. Combined with server chip Amazon Graviton3, Amazon Web Services provides energy efficient and effective machine learning infrastructure.
2. **AI Platform：** Amazon SageMaker offers a set of feature-rich capabilities for developers, data scientists and ML engineers. Amazon SageMaker Studio Lab offers free resources. Amazon SageMaker Jumpstart and Amazon SageMaker Canvas provide low-code/no-code quick-start features. Amazon SageMaker Pipelines provides automated ML processes. Amazon SageMaker Ground Truth Plus provides intelligent annotation services. Amazon SageMaker Data Wrangler has more than 300 built-in data transformations. Amazon SageMaker Autopilot automates AutoML execution.
3. **AI Services：** Natural Language Understanding (NLU), Automatic Speech Recognition (ASR), Visual Search and Image Recognition, Text-to-Speech (TTS) and Machine Learning (ML) hosting services.

- The "Lake House Architecture" combines machine learning and data management platforms to provide an integrated data intelligence unified data governance experience. Amazon Redshift ML and Amazon Athena ML both support model training requests in the form of SQL statements, and Amazon SageMaker Canvas AutoML capabilities provide model training and return in the form of SQL.
- Amazon Web Services has attracted more than 100,000 customers with the partnership and talent ecology. Amazon Cloud Technologies has more than 80 ML/AL capability partners to provide customers with abundant and well-established industry solutions. There are more than 1,000 machine learning products from more than 300 ISVs in the Marketplace, covering a range of customers from healthcare, retail, financial services, social entertainment, manufacturing, energy, etc.

Amazon Web Services AI Platform user cases

OPPO's conversational AI product "Xiaobu" with over 100 million monthly activities, in order to achieve industry-leading conversational semantic understanding, innovates on Amazon EC2 Inf1 to develop efficient inference service modules that can support pre-training large models, which is expected to reduce the cost of model inference services by more than 35% on some business scenarios, and aims to gradually expand to more and more new scenarios. With Amazon EC2 Inf1, OPPO's machine learning team continues to innovate with more sophisticated algorithmic models and accelerate improvements in the overall customer experience.

With Amazon SageMaker and its database and computing services, Schneider Electric has successfully built an intelligent industrial vision quality inspection solution, the "Cloud-Side Collaborative AI Industrial Vision Inspection Platform". With Amazon SageMaker, Schneider Electric was able to successfully and accurately build machine learning models adapted to real-world manufacturing scenarios to identify complex defects in products through automated industrial vision inspection by comparing product images from production lines with standard samples of qualified products. The solution was first launched in Schneider Electric's Wuhan factory, significantly improving the inspection efficiency of the production line, reducing the false detection rate to within 0.5% and achieving a zero-miss detection rate.

Youdao LeRead is based on Amazon Personalize's personalized recommendations and big data services to provide accurate book recommendations for end-users. With Amazon Personalize, you can design personalized book recommendations through simple API calls without the need to have machine learning experience. Amazon Personalize service works out of the box and effectively helps you achieve accurate book recommendations and predictions within a month, thus ensuring a quality user experience and increasing monthly active users by 20%. In addition, compared with the previous monthly iteration cycle, the delivery is now basically on a daily basis, even on the same day of the update.

Source:AWS, Frost & Sullivan, LeadLeo

Source: Amazon Web Services, Frost & Sullivan

Baidu AI Cloud

Baidu's AI development platform consists of a full-featured AI development platform (BML), a zero-code AI development platform (EasyDL), and an AI development training platform (AI Studio). It has accumulated 4.77 million developers and 180,000 enterprises in China. Baidu AI development platform covers the life cycle of AI model application, including data processing, algorithm development, model training, etc.

2022 Product Updates:

- **Data Analytics Engine**: a data analytics engine, which can perform cross-database federal queries and support automatic data analysis and visualization.
- **Feature Store**: be able to ensure consistency of feature data used for model training and prediction services, decouples feature production and consumption, and enables feature sharing and reuse among different teams.
- **XAI**: provides 6 model interpretation algorithms and graphs, providing interpretability for most conventional machine learning models and deep learning models. Also, it can reduce risk of use.
- **MLOps**: Provide complete automation capability and customized development SDK in lifecycle of model development. It can be connected to enterprise CI/CD system to realize model production and operation integration and improve development efficiency and standardization.

- Baidu's AI development platform consists of a full-featured AI development platform (BML), a zero-code AI development platform (EasyDL), and an AI development training platform (AI Studio), which are unified and enabled by Baidu's self-developed PaddlePaddle platform. It has accumulated 4.77 million developers and 180,000 users from enterprises and institutions in China.
- Baidu's AI development platform provides rich and comprehensive functions, which cover the entire lifecycle process of AI models from project creation to deployment of inference services. It also provides different forms of services to customers, including online platform and on-premise deployment.
- In terms of cutting-edge technology, Baidu's AI development platform has started to have complete MLOps capabilities, covering data processing, feature engineering, model development, training tasks, drift monitoring, automatic retraining, and workflow to automate execution for efficient integration of model production and operation; In the XAI domain, Baidu AI development platform can realize the risk management of models' lifecycle through the MRM(model risk management) module to meet the regulatory needs of specific industries and organizations; at the same time, more advanced features, such as the deep learning model interpretability and model robustness/security capability, are realized with the empowerment of the PaddlePaddle. There are also more new features in the areas of intelligent annotation and AutoML than in the previous report.
- With the support of Baidu's own PaddlePaddle platform, Baidu AI development platform has built-in development kits and pre-trained models for mainstream domains, as well as foundational models for currently popular domains. Among them, the Wenxin ERNIE is the largest single NLP model in China. .
- At the localization, Baidu AI development platform has supported or adapted to a variety of Chinese local chips, thus forming a full-level localization solution.

Shandong Electric Power: Transmission line safety inspection

The wide geographical distribution and complex and changing environment pose a serious challenge to the safe operation of transmission lines. The visualization of transmission channels and intelligent analysis greatly enhance the efficiency of transmission line safety inspection and provide a reliable guarantee for the safe and stable operation of transmission lines.

Postal Saving Bank Intelligent System

Realizing the R&D management of the whole life cycle of AI models from training, testing, deployment, operation and iteration, introducing various machine learning and deep learning advanced algorithms and models, and accelerating the implementation of AI applications in business scenarios of the whole bank.

Changsha Metro's intelligent maintenance helmet

By combining EasyDL object detection training tool with classification recognition model, accumulating 500 training pictures and iterating 17 versions, Changsha Metro developed a detachable structure "intelligent maintenance helmet", whose model accuracy rate reached 88.9%. It can automatically take pictures and identify the name and quantity of common tools, which provides timely and effective guarantee for the inventory of tools and makes a breakthrough in the innovative application of smart helmet.

Source: Baidu AI Cloud, Frost & Sullivan, LeadLeo

Terms

- **QPS:** Queries per second, query rate per second. QPS is a measure of the traffic handled by a specific query server within a specified time. It can be interpreted as the number of concurrent requests per second. 1QPS calls about 86400 times.
- **API:** Application Programming Interface. API is a pre-defined function, which aims to provide the ability for applications and developers to access a set of routines based on a certain software or hardware without accessing the source code or understanding the details of the internal working mechanism.
- Convolution: a mathematical concept that generates the third function through two functions f and g, representing the integral of the overlapping length of the product of the overlapping part function values of function f and g after flip and translation.
- **CGRA:** Coarse grained Reconfigurable Architecture. CGRA is a parallel computing mode in the airspace, which organizes computing resources with different granularity and different functions in the airspace hardware structure. In the runtime, according to the characteristics of the data flow, the configured hardware resources are interconnected to form a relatively fixed computing path, which is close to the "dedicated circuit" for computing; When the algorithm and application are transformed, they are re configured into different computing paths to perform different tasks.
- **CUDA:** Compute Unified Device Architecture is a parallel computing platform and programming model created by NVIDIA based on their GPUs (Graphics Processing Units, which can be commonly understood as graphics cards).
- **DevOps:** a combination of Development and Operations, which is a general term for a group of processes, methods and systems, and is used to promote communication, collaboration and integration among development (application/software engineering), technical operation and quality assurance (QA) departments.
- **Data annotation:** the process of annotation of metadata such as text, video, image, etc. The marked data will be used to train ML models.
- **Cloud native**: a set of cloud technology product system based on container, micro service, DevOps and other technologies, which is a distributed cloud based on distributed deployment and unified operation management.

Methodology

- Frost & Sullivan has conducted in-depth research on the market changes of 10 major industries and 54 vertical industries in China with more than 500,000 industry research samples accumulated and more than 10,000 independent research and consulting projects completed.

- Rooted on the active economic environment in China, the research institute, starting from data management and big data fields, covers the development of the industry cycle, follows from the enterprises' establishment, development, expansion, IPO and maturation. Research analysts of the institute continuously explore and evaluate the vagaries of the industrial development model, enterprise business and operation model, Interpret the evolution of the industry from a professional perspective.

- Research institute integrates the traditional and new research methods, adopts the use of self-developed algorithms, excavates the logic behind the quantitative data with the big data across industries and diversified research methods, analyses the views behind the qualitative content, describes the present situation of the industry objectively and authentically, predicts the trend of the development of industry prospectively. Every research report includes a complete presentation of the past, present and future of the industry.

- Research institute pays close attention to the latest trends of industry development. The report content and data will be updated and optimized continuously with the development of the industry, technological innovation, changes in the competitive landscape, promulgations of policies and regulations, and in-depth market research.

- Adhering to the purpose of research with originality and tenacity, the research institute analyses the industry from the perspective of strategy and reads the industry from the perspective of execution, so as to provide worthy research reports for the report readers of each industry.

Legal Disclaimer

- The copyright of this report belongs to LeadLeo. Without written permission, no organization or individual may reproduce, reproduce, publish or quote this report in any form. If the report is to be quoted or published with the permission of LeadLeo, it should be used within the permitted scope, and the source should be given as "LeadLeo Research Institute", also the report should not be quoted, deleted or modified in any way contrary to the original intention.

- The analysts in this report are of professional research capabilities and ensure that the data in the report are from legal and compliance channels. The opinions and data analysis are based on the analysts' objective understanding of the industry. This report is not subject to any third party's instruction or influence.

- The views or information contained in this report are for reference only and do not constitute any investment recommendations. This report is issued only as permitted by the relevant laws and is issued only for information purposes and does not constitute any advertisement. If permitted by law, LeadLeo may provide or seek to provide relevant services such as investment, financing or consulting for the enterprises mentioned in the report. The value, price and investment income of the company or investment subject referred to in this report will vary from time to time.

- Some of the information in this report is derived from publicly available sources, and LeadLeo makes no warranties as to the accuracy, completeness or reliability of such information. The information, opinions and speculations contained herein only reflect the judgment of the analysts of leopard at the first date of publication of this report. The descriptions in previous reports should not be taken as the basis for future performance. At different times, the LeadLeo may issue reports and articles that are inconsistent with the information, opinions and conjectures contained herein. LeadLeo does not guarantee that the information contained in this report is kept up to date. At the same time, the information contained in this report may be modified by LeadLeo without notice, and readers should pay their own attention to the corresponding updates or modifications. Any organization or individual shall be responsible for all activities carried out by it using the data, analysis, research, part or all of the contents of this report and shall be liable for any loss or injury caused by such activities.