吴富林♂

On job, open for new job · 38 · Bachelor · 14 years 4 month experience

13026748216 664737299@qq.com 13026748216

Self-description

R&D Capability Model

1、Proficient in multiple programming languages, including C++, Python, Golang, Vue/React, Java, and Rust. Meanwhile, familiar with middleware such as Kafka, Redis, and MQTT, demonstrating strong technical adaptability.

2、Skilled in cloud-native microservice architecture based on Docker and Kubernetes, as well as DevOps CI/CD processes, capable of effectively improving the deployment efficiency and stability of systems.

3、Master Numpy, Pandas, PyTorch, and other deep learning frameworks, with basic knowledge of deep learning. Familiar with CUDA, TensorRT, Triton, and vLLM, and possess in-depth research and practical application experience in MaaS (Model-as-a-Service) and AI cloud-edge collaboration.

4、Capable of designing, building, and optimizing large-scale GPU computing power platforms, establishing efficient MLOps systems, and holding in-depth professional knowledge in high-performance inference optimization for large models.

5、With over 6 years of technical R&D management experience, proficient in team building, guiding, and motivating teams to ensure the achievement of company goals.

Work Experience

HKSTP **2025/04-To Present**

HPC IT Manager

1. Responsible for the daily management and maintenance of the HPC AI cluster composed of Nvidia DGX (H100/H800) nodes.

2. Build an AIOps/MLOps platform based on the AI cluster, and use this platform to conduct model SFT training across multiple GPUs and multiple nodes.
3. Responsible for the architecture design and performance optimization of the GPU computing power platform. This architecture is mainly built on Kubernetes to develop the MaaS platform, while performing performance optimization for AI Inference.

Qianhai Aixun Technology (Shenzhen) Co., Ltd., Guangzhou Branch (Hong Kong-funded) **2023/11-2025/03**

Head of AI Edge Computing Technology

1、Led a 3-person development team to complete the commercialization and market launch of an AI video analysis system and AI cloud platform from scratch within less than 6 months. The AI video analysis system is compatible with Nvidia Jetson chips, Huawei Ascend chips, and Rockchip RK3588 chips.
2、Developed a cloud-edge collaboration platform based on Kubernetes+KubeEdge, supporting unified Kubernetes management and scheduling of AI edge devices with GPU/NPU, which serves as a key infrastructure for the AI cloud platform.
3、Designed and developed an AI algorithm inference engine pipeline (implemented in C++) for edge/end devices, enabling containerized algorithm deployment and seamless integration and calling of algorithm modules through standardized interfaces.
4、Optimized the performance of AI algorithm inference to achieve high-performance video analysis inference for 8-channel, 16-channel, and 32-channel streams on edge hardware boxes with limited resources.
5、Developed a personalized recommendation system based on LLM large models and RAG, providing users with more accurate and personalized product recommendations by analyzing user interests, shopping behavior data, and understanding large amounts of product text content.
6、Developed an intelligent customer service system based on LLM large models and RAG, which can quickly understand user inquiries, retrieve accurate answers from a large number of knowledge base documents and frequently asked questions, and provide timely and accurate services to improve customer satisfaction.

Rootcloud **2020/07-2023/09**

Manager of Platform R&D Department

1、Responsible for R&D management of the data middle platform of the company's Rootcloud platform, leading a team of up to 22 people in technical research, coordinating product, R&D, testing, and operation teams for iterations, and promoting platform reconstruction.
2、Responsible for the architectural design of Asset Management (EAM), Data Warehouse (IDF), Data Permissions (IAM), and Indicator System (ABI), guiding R&D personnel in solving technical difficulties and providing optimization/solutions.

Achievements:
1、Won the 2021 Outstanding Employee Award (top 1%).
2、Obtained 1 authorized invention patent.
3、Led the team to successfully launch the IDF data warehouse and ABI functional modules of the indicator system.

Guangdong Jiatai Intelligent Technology Co., Ltd. **2017/12-2020/06**

R&D Director

1、Built the R&D team from scratch, with a team structure including an embedded hardware group (6 people), a platform application software group (5 people), a software and hardware testing group (2 people), and a product group (4 people). Responsible for the entire R&D system construction, recruitment, work management, and technical development.
2、Designed and built an IoT PaaS cloud platform from scratch, realizing functions such as device access, protocol conversion, device management, rule engine, and OTA upgrades.
3、Developed a Linux-based data acquisition gateway with edge computing capabilities, achieving initial cloud-edge collaboration management.
4、Built a K8S DevOps CI/CD R&D process system based on Gitlab, Rancher, and Jenkins, significantly improving R&D and operation efficiency.

Achievements:
1、Helped the company obtain tens of millions of Pre-A round financing.
2、Applied for multiple invention patents, with 1 authorized and most others under substantive examination.

Project Experience

MLOps Platform and Large Model Inference Optimization Based on NVIDIA DGX Cluster 2025/04-To Present

Manager

Description

This project aims to build an ultra-large-scale HPC/AI cluster consisting of dozens of Nvidia DGX (H100/H800) nodes, and on this basis, integrate a complete MLOps platform to provide unified, high-performance computing power and service support for the company's large model training, fine-tuning, and inference.

Responsibilities

As a core development team member, I participated in the entire process, from infrastructure management and operation & maintenance to the architecture design and optimization of the upper-layer platform.

Achievement

1、Platform Architecture: Led the design of a Kubernetes-based MaaS (Model-as-a-Service) platform, which manages large-scale AI clusters composed of Nvidia DGX (H100/H800) units. This platform enables unified pooling of computing resources and elastic scheduling.
2、MLOps Practice: Built an end-to-end MLOps pipeline on this platform, which efficiently supports large-model SFT (Supervised Fine-Tuning) tasks in multi-machine multi-GPU environments. It also realizes experiment tracking, model version management, and automated training.
3、Inference Performance Optimization: Conducted in-depth optimization of large-model inference services. By integrating key technologies such as TensorRT-LLM/vLLM/SGLang, and applying model quantization and continuous batching, the inference throughput was increased by 5x, latency was reduced by 60%, and service costs were significantly lowered.

WeVision - AI Video Cloud Platform 2024/12-To Present

Architect + Core Developer

Description

A multi-tenant video cloud platform with main functions including controlling edge gateway devices, upgrading edge gateway device systems, delivering algorithms, and providing standard data dashboards based on alarms/events reported by edges. It helps customers combine the real-time performance of edge computing with the overall perspective of the cloud, providing efficient and intelligent video data processing and management services, improving device management efficiency, and creating greater business value through algorithm delivery and data analysis.

Responsibilities

As the team architect, led a 3-person development team in technical selection for cloud-edge collaboration, data link design (cloud-edge data integration), overall architectural design (including AI model delivery to edge devices for upgrades, algorithm mall management, alarm and event data storage, multi-tenant design, and algorithm license authorization design), and cloud platform interface design.

Achievement

1、Completed the multi-tenant, cloud-edge collaboration-based AI video cloud platform, including a mini-program version, within 2 months.
2、Secondary development of KubeEdge, developing one-click installation packages for both cloud and edge/end sides, enabling rapid deployment in new private environments and significantly accelerating delivery efficiency.

EdgeTurbo - AI Video Analysis Platform **2024/04-2024/08**

Committer and Architect

Description

EdgeTurbo is a system based on edge computing devices (NVIDIA Jetson boxes, Rockchip RK3588 boxes, Huawei Ascend boxes) that automatically analyzes, identifies, and processes video streams through computer vision (AI deep learning algorithms, e.g. yolov8). Its main functions include target detection and tracking, target classification, intrusion detection, passenger flow analysis, behavior analysis, vehicle recognition, and industrial defect recognition. It helps customers solve monitoring, management, analysis, and optimization

problems, saving labor costs, improving analysis accuracy and real-time performance, and is widely used in security, medical care, transportation, education, and other fields.

Responsibilities

As a committer and architect, led a 3-person development team to complete the commercialization and market launch of the AI video analysis system from scratch within less than 6 months. The system is compatible with Nvidia Jetson chips, Huawei Ascend chips, and Rockchip RK3588 chips.

Achievement

1、Designed and developed an end-side AI algorithm inference engine, implementing an algorithm pipeline in C++ that can flexibly combine various nodes (FFmpeg stream pulling, video hardware decoding, preprocessing-inference-postprocessing, target tracking, OSD, alarm/event pushing, video hardware encoding, FFmpeg stream pushing) to achieve different business logics.
2、Designed a set of rules to enable the AI cloud platform to deliver and update models on the edge/end sides, realize containerized deployment of algorithms, and achieve seamless integration, scheduling, and updating of algorithm modules through standardized interfaces.
3、Optimized the performance of the AI algorithm inference end using methods such as model quantization, hardware-accelerated encoding/decoding, video stream frame extraction without decoding, CUDA/CANN kernel functions for preprocessing and postprocessing logic, OSD implementation on YUV, TensorRT Runtime acceleration, and calling various coprocessor chips for acceleration. This enabled high-performance video analysis inference for 8-channel, 16-channel, and 32-channel streams on edge hardware boxes with limited resources.
4、Adapted the inference engine for different AI chips (Nvidia Jetson, Ascend 310B, RK3588) using the same set of code, reducing maintenance costs.

Data Warehouse **2022/01-2023/09**

Team Leader

Description

Building a data warehouse for the IoT platform to provide IT and OT data integration capabilities, enabling the Rootcloud platform to effectively use data scattered and managed in

various business systems to support the calculation of operational indicators and provide a basis for subsequent production and operation decisions.

Responsibilities

1、As the microservice owner, responsible for service splitting based on external business requirement analysis, data structure design, database table design, API specification design, task arrangement, and organizing solution reviews.
2、As a committer and architect, mainly responsible for internal performance optimization solutions, overall architectural design, database high-reliability design, and solving related technical difficulties.

Achievement

1、Designed the architecture for business data modeling, using custom metadata schemas to dynamically generate tables and their indexes at runtime based on metadata definitions, enabling users to create business tables (corresponding to the data warehouse ODS layer) on the platform with zero coding. Converted business tables (ODS layer) into data warehouse DWD and DIM layers using the Data Vault 2.0 unified modeling methodology and dbt tools.
2、Designed the dynamic generation architecture for wide tables in the data warehouse, defining the structure and views of wide tables through drag-and-drop operations, and using CDC to monitor tables related to the views. When table data changes, the range of changes is calculated in a timely manner and updated to the wide tables, ensuring that updates to business tables are synchronized to the wide tables to meet strong real-time requirements.

Data Permissions (IAM) **2021/11-2022/06**

Team Leader

Description

Enhancing the capabilities of the platform's IAM module, implementing fast data permission queries using the ABAC model + SDK on the basis of the RBAC model to meet permission management and performance requirements.

Responsibilities

1、Proposed a data permission design based on the ABAC model.
2、Initially simplified permission calculation using Postgresql RLS features, and later converted condition parsing into SQL for permission calculation expressions.
3、Designed the IAM SDK class diagram and developed functions such as expression parsing.

Achievement

1、Designed the IAM SDK to simplify the integration of each service with the platform's permission system and improve interface query speed using SDK caching.
2、Reconstructed the permission system, changing the original permission calculation method based on RLS to parsing Condition expressions into SQL Where statements, improving query efficiency by more than 10 times.

IoT PaaS Platform **2019/09-2020/04**

R&D Director

Description

An IoT platform for managing the access of split Bluetooth Mesh gateways, Bluetooth Mesh collectors, and integrated Linux collection gateways (including encrypted certificate management, network activation management, protocol configuration management, etc.), as well as data forwarding, historical data storage, and firmware management.

Responsibilities

1、Responsible for the overall technical architecture design of the platform.
2、Responsible for the design of protocols for inter-cloud platform docking.
3、Responsible for writing core code for key technical difficulties.

Achievement

1、Designed the overall framework of the IoT platform data flow, completing functions such as device access, rule engine, scenario linkage, batch OTA, and device historical data storage.
2、Designed protocols for communication between direct-connected devices, gateways, and the IoT platform, including detailed designs of MQTT TOPIC division principles, MQTT uplink protocols, MQTT command delivery protocols, and transparent transmission protocols.
3、Designed and implemented the northbound output of the platform, enabling IoT data to

support multiple third-party cloud platforms and realize inter-cloud forwarding.
4、Designed an IoT big data storage solution using ClickHouse features to achieve second-level OLAP computing and statistical analysis of big data.

Linux Data Acquisition Gateway **2018/01-2019/06**

R&D Director

Description

The Linux data acquisition gateway is a device that performs unified configuration management with Jiatai Intelligent IoT PaaS platform through TCP/IP, 4G communication, etc. It has functions of remote PLC program debugging and downloading, and industrial data acquisition.

Responsibilities

1、Responsible for the design of gateway hardware schematics and PCBs, and prototype debugging.
2、Responsible for embedded Linux kernel tailoring and Linux application development, including front-end development in Vue and back-end development in Golang.
3、Responsible for upper computer software development and VPN link communication setup.
4、Responsible for the development of communication protocols between the gateway and devices, the gateway and the IoT PaaS platform, as well as dynamic scripts and edge algorithms.

Achievement

1、Created a prototype of the gateway hardware in only 2 months, with successful customer POC demonstrations.
2、Realized remote downloading and debugging of programs and data acquisition for PLCs of multiple mainstream brands through the gateway.

Education

Sun Yat-sen University (985, 211) **2007/09-2011/06**

Geographic Information System

Bachelor

Languages

Mandarin（Mastery）；Cantonese（Mastery）

Additional Info

1、"ESRI" China University Student GIS Software Development Second Prize and Best On-site Performance Award.

2、87 Economic Geography Scholarship of the School of Geographic Sciences and Urban Planning (top 10%).

3、2007 Sun Yat-sen University Excellent Scholarship First Prize (top 5%).