Ming Chen

+86-13911397477 | 2727524@qq.com
Beijing
vxjiulidaji

SUMMARY

- 21 years of experience in the IT industry, 12 years in cloud computing development and operations, 6 years in machine learning, and 2 years in large model application incubation and development. I have experienced both successes and failures across various products and have spent the past 8 years dedicated to developing and operating desktop cloud monitoring and data platforms, integrating AI and large model applications into monitoring platforms.
- Fully experienced in the evolution of monitoring services from private deployment to cloud services and then to intelligence-driven transformation. Led the transition to cloud-native and intelligent monitoring.
- 16 years of team management experience, having built and scaled teams from scratch and managed teams as a senior leader. Currently leading a 26-member development team focused on desktop cloud monitoring and data platforms.
- Actively involved in the practical application of AI, machine learning, and large language models, successfully incubating the first large language model application (RAG + LLM) within a data platform.
- Throughout my 20-year career, I've applied many Java, database, and Linux technologies. For the past 8 years, I've focused on cloud-native development based on AWS services and SpringBoot, building data platforms centered around Kinesis/Kafka, Cassandra, Elasticsearch, Presto, and DatawareHouse.
- Consistently summarizes and promotes best practices to improve engineering efficiency through automation, tooling, process optimization, etc.
- Passionate about innovation, served as the Innovation Evangelist for VMware China EUC, actively organizing, promoting, and hosting multiple annual Innovation Week events. Successfully helped the team incubate several award-winning ideas into components of EUC products. Holds five published patents.

PROFESSIONAL EXPERIENCE

VMware China By Broadcom Sep 2013 - Present

Senior Manager I Manager I Staff Engineer EUC, Beijing

Senior Manager(2018/8 - Present); Manager(2015/3 - 2018/7)

- Leads a 26-member development team focused on desktop cloud monitoring services and data platforms, embracing openness, intelligence, real-time capabilities, and security.
- Seized every critical milestone in digital transformation, leading the evolution of products from On-Premise deployment to public and hybrid cloud architectures, and further advancing to intelligent solutions powered by AI/LLM, consistently driving technological innovation and change.
- Current Focus: 1) User-customizable data aggregation, 2) Large model applications in data platforms, 3) Digital employee experience management.

Phase 3 - Intelligent VDI Cloud Monitoring & Digital Employee Experience Management

- Incubated a multi-source data query platform based on LLM, approaching its first version. Managed tasks from POC to MVP, including model training, data labeling, Antlr post-processing, business post-processing, pre-filtering, and E2E deployment coordination with cross-product teams.
- Developed a digital employee experience management solution for Horizon Desktop Cloud based on the monitoring data platform.

- Currently incubating an RAG-powered on-call issue assistant bot that provides suggestions based on Slack chat history, Confluence knowledge base, and Git codebase for on-call issues.

Phase 2 - VDI Cloud Monitoring

- Led two major improvements to the data platform based on business needs: 1) Transitioned from Cassandra + Spark + ECS to ElasticSearch + S3 + AWS Lambda + EKS, 2) Added Presto and AWS Redshift.
- Strengthened auditing of AWS cloud platform costs, reducing operating expenses by 25% through network optimization, storage optimization, and the use of Spot instances.
- Introduced PagerDuty to establish a robust DevOps framework for cloud services, covering Runbook, Oncall, and Escalation processes, and integrated a series of CloudWatch metrics to assess the health of core services.
- Drove automation and CI/CD efforts to ensure infrastructure as code, minimizing manual deployments, and is proficient in GitHub Workflow.
- The service has five production deployments, serving over 350 global customers. It processes approximately 230,000 documents per minute, about 330 million documents per day, with a daily traffic volume of around 500 GB.

Phase 1 - OnPrem Deployment for VDI Monitoring

- Responsible for collecting monitoring data from Horizon Connection Server and agent data reported by team members.
- Led the team to develop an end-to-end monitoring solution based on vRealize Aria, covering data collection, ETL, storage, aggregation, querying, and display.
- Proposed the idea for Horizon Helpdesk, which became a product after collaborating with the PM to build a demo.

Staff Enginner(2013/9 - 2015/2)

- Mirage project team lead and Scrum Master, managing a team of six engineers responsible for developing and testing Mirage Gateway, Mirage Report, Mirage API, and Mirage CEIP.
- Implemented coding standards, submission protocols, and unit testing guidelines.
- Promoted agile development practices (unit testing, stand-up meetings, CI, iteration cycles) within the team and actively advocated for test automation.
- Keyword: Nginx, C#, GCE, Sencha, SQLServer, Jasper, Python, Mongo

IBM CDL Apr 2005 - Sep 2013

Advisory Engineer | Staff Engineer China Development Lab Beijing

Advisory Enginner (2011/11 - 2013/9)

- Served as a technical expert, responsible for the integration of IBM's Sametime RTAV and Polycom AV Server, coordinating with PolyComm to resolve critical issues.
- Led quality assurance for IBM's Sametime product, strongly advocating for test automation and code coverage, and introduced Scrum practices. Developed small tools to track automation and Scrum progress.
- Introduced Agile Practices to the team and Acted 1st Scrum Master in Sametime Beijing Team

Staff Engineer (2005/4 - 2011/10)

- Performance/System Team Lead, established and led a team responsible for the performance analysis of IBM Connections, an enterprise social software.
- Conducted design and critical code reviews, specializing in analyzing complex performance issues, OS optimization, JVM GC optimization, lock analysis, and SQL tuning.
- Customer Engagement: HSBC, Dupont, China Mobile Research Institute, HuaWei, AT&T, BAS

Beijing Rewin Network Technology Co. Ltd. Apr 2003 - Mar 2005

Software Engineer Beijing

- Developed and tested the web portal for New Era Securities, focusing on receiving and importing F10 data, and backend development by utilizing Java + WebLogic + JDBC.
- Developed and tested a real-time market quoto applet by using Java Applet + multithreading + JDBC.

PROJECT EXPERIENCE

Query AI - LLM-Based Text-to-Query Service (LLM/RAG + SpringBoot)

Tech Stack: Springboot/SpringAI/Python/GPT4o-mini/ANTLR4/AWS SQS/Redis/AWS SageMaker

Technical Challenges:

- Bridging semantic gap between natural language and domain-specific query/aggregation
- Effectively, efficiently selecting and injecting the most relevant database schemas into the limited LLM context
- Achieving precise SQL syntax tree conversion and optimization
- Finding the cost effective and accurate way for text to query conversion

Key Implementations:

- Developed querypilot micro service to provide REST API (Spring Boot) to handle natural language-to-query conversion
- Integrated AWS SQS for async query refinement, decoupling user requests from compute-intensive LLM/SQL conversion steps and maybe time consuming query task.
- Implemented schema-aware caching using Redis and similar text to SQL request using vector search to reduce redundant LLM calls
- Built a query rewriter using ANTLR4 to canonicalize LLM-generated SQL into WS1 Data Platform-compatible syntax.
- Implemented the schema filter and few shot selector based on RAG architecture with 2000+ vertical domain SQL templates knowledge base
- Compared SFT vs RAG, Llama vs GPT4o-mini
- Synthesized/Labeled 6k+ cold start questions/query dataset

Queryfox: Unified Query Engine for Workspace ONE Data Platform

Tech Stack: Java/SpringBoot/Elasticsearch/Redshift/Trino/Redis/AWS Glue

Technical Challenge:

- Multiple heterogeneous data stores for different purpose
- Complext query / aggregation, No consistent model
- New aggregation requirements for time series data

Key Implementations:

- Smart Query Routing Engine
  - Dynamically selects optimal data source
    - Elasticsearch - warm data
    - Redshift – OLAP Engine for latest snapshot data,
    - Presto-S3 – Cold/Raw/Aggregated time series data
  - Implements rule based routing algorithm based on the data source and the purpose
    - Multi-Granularity Aggregation Processing
    - Supports time-tiered aggregation (minute/hour/day/year levels)
    - Implements rollup preprocessing for 10x faster annual trend queries

- Model Query/Aggregation
  - Query Definition for both query and aggregation
  - Add new counters like latest, any, count by minute etc. to support more complex aggregation
- Performance Optimization
  - Columnar storage adaptation and merge for Glue/Presto accelerates scan queries

SpringBoot 2 to SpringBoot 3 Update

Tech Stack: SpringBoot, MicroService, CICD, Spring Security, Apache HTTP Client

Technical Challenges:

- **Jakarta EE Migration**: Transitioned from javax to jakarta namespace, requiring updates to dependencies and import statements across 100+ microservices.
- **Apache HttpClient Upgrade**: Migrated from HttpClient 4.x to 5.x, impacting common code and direct usage in services.
- **Spring Security 6.0 Adjustments**: Deprecated methods (e.g., oauth2ResourceServer(), jwt()) replaced with lambda-based configurations; updated SecurityContext implementations.
- **Auto-Configuration Changes**: Updated META-INF/spring.factories to META-INF/spring/org.springframework.boot.autoconfigure.AutoConfiguration.imports format.
- **Testing Framework Updates**: Addressed HttpSecurity.logout() deprecation and NestedServletException replacements.

Key Implementation:

- **Automated Dependency Analysis**: Developed tools to auto-resolve 85% of compatibility issues, including Jakarta EE namespace updates.
- **Zero Downtime Rollout**: Phased migration to minimize downtime, validated via canary deployments. Zero-downtime upgrade achieved for 100+ microservices.
- **CI/CD Pipeline Refactoring**: Improved build efficiency through automated issue resolution and modernized CI/CD pipelines.
- **Documentation & Training**: Created migration guides and upskilled teams on new Spring Security patterns and dependency management.

EDUCATION

Beijing Institute of Technology Sep 2000 - Apr 2003

Automation Master Beijing

Beijing Institute of Technology Sep 1996 - Jul 2000

Automation Bachelor Beijing

PATENTS

- US10409577B2, HYBRID APPLICATION DELIVERY THAT COMBINES DOWNLOAD AND REMOTE ACCESS
- US9769515B2, Method and system for improving application sharing by dynamic partitioning
- US9641452B2, Resolving a convex optimization problem to optimize network traffic in a distributed system
- US10333821B2, Method and system for optimizing network traffic in a distributed system with a point of convergence
- US10608955B2, Reverse breadth-first search method for optimizing network traffic in a distributed system with a point of convergence