Mahesh **Arumugam**

linkedin.com/in/aumahesh github.com/aumahesh
aumahesh@gmail.com aumahesh.github.io scholar.google.com +1 (408) 904-9746 San Jose, CA

Software Architect | Distributed Systems | Data Science | AI/ML | LLM

Recent Experience

November 2023 – October 2021 — Architect, Zscaler, Inc., San Jose, CA

- Designed the Zscaler Business Insights platform that enables customers to gain insights into application access patterns and identify optimizations (that includes cost saving opportunities, overlapping applications, and user engagement).
- Ensured the solution provided an excellent day-zero experience by highlighting insights learned from customer traffic logs through Zscaler Zero Trust Network Exchange (ZTNA).

SaaS APIs | API Integration | Microsoft Azure | Azure ADX | ChatGPT | Golang | Java | Python

October 2021 – May 2021 — Principal Software Engineer, FeatureBase (formerly, Molecula), Remote, HQ : Austin, TX

- Designed Molecula as a Service platform to (1) store features in an optimized format, (2) retrieve responses for complex analytical queries quickly, and (3) serve features to analytical and ML engines.

Feature Store | Machine Learning | Analytics | Data Science | MLOps | Cloud | Golang | Python

November 2020 – June 2017 — Principal Engineer, Cisco Tetration, Palo Alto, CA

- Designed and developed a connector framework for data ingestion from the customer data center to Tetration. The framework supports provisioning, configuring, monitoring, and troubleshooting connectors.
- Developed connectors that helped onboard over 100 customers who wanted to understand the applications in their data center without installing yet another agent in their ecosystem.
- Designed and developed a full-stack Cloud Security Posture Management (CSPM) for insights into how workloads and configurations in a public cloud follow standards. CSPM provides a policy framework that flags/notifies configurations and entities not following standards and customer-defined rules.

Zero Trust | Micro-segmentation | NetFlow | IPFIX | Docker | AWS | Kafka | Golang | Python

Education

December 2023 — **Master of Information and Data Science (MIDS)**, **University of California, Berkeley, CA**
Project Portfolio : aumahesh.github.io
Relevant Coursework : Research Design, Statistics for Data Science, Applied Machine Learning, Machine Learning @ Scale, Natural Language Processing, Privacy Engineering, Time Series Analysis, Computer Vision

2006 — **Ph.D. & M.S.** *Computer Science and Engineering*, **Michigan State University, East Lansing, MI**

2001 — **B.E.** *Computer Science and Engineering*, College of Engineering, Guindy, **Anna University, Chennai, India**

Selected Projects

Backend architecture for Zscaler Business Insights — Zscaler

To understand the application and workplace landscape of organizations onboarded on the Zscaler Zero Trust Network Access (ZTNA) platform, we designed and developed the Zscaler Business Insights platform. Zscaler Business Insights platform provides insights into application access patterns and identifies optimizations. As part of this project, I designed the backend architecture for data ingestion, processing, and analytics from multiple sources, including the traffic data from Zscaler Zero Trust Network Access (ZTNA), SaaS applications (APIs), and IdP providers. I developed integrations with various SaaS application APIs to collect user profile and engagement data for richer insights. I contributed to the design and development of integrations for the eight most popular applications among the organizations in the Zscaler ZTNA platform. In addition, I developed ChatGPT integration to build a catalog of applications seen in Zscaler ZTNA to provide context about applications accessed by organizations. The application catalog provides metadata for more than 40000 applications, which includes business categories, available SKUs, and alternatives.

SaaS APIs | API Integration | Microsoft Azure | Azure ADX | ChatGPT | Golang | Java | Python

Signet-ring : A framework for authenticating sources and lineages of digital objects — MIDS

Abstract Presentation at High Confidence Software and Systems Conference 2023, Annapolis, MD

Verifying sources of information is vital in assessing the credibility of facts and data in our increasingly digital world; often, verifying the sources is as necessary as the information they provide. To battle misinformation and disinformation through digital objects, the ability to establish whether or not information (or data) provided by such sources was altered before its use (e.g., publication) is very important. To address these concerns, we designed and implemented Signet-ring. Signet-ring registers and authenticates all participants in the origination and publication process, potentially including the sources, publishers, and applications. It manages the following critical workflows : (1) documentation and verification of the relationships between objects and sources (certification), (2) documentation and verification of the relationships between different related objects (lineage), and (3) authentication of sources to each other (handshake). Furthermore, the Signet ring supports the lifecycle management of source identities (using cryptographic keys) and relationships between objects and sources. This lifecycle management includes revoking source identity keys and previously accepted object-source relationships.

Authenticity | Certificate Authority | Lineage | Provenance | Security | Privacy | Python

CodeT5++ : A Pre-trained Programming Language Model for Code Summarization Task — MIDS

There has been considerable research in building pre-trained models for programming language tasks, such as CodeBERT and CodeT5, that enable several downstream tasks, including code summarization, generation, and translation. In this paper, we focus on the task of automated code summarization that translates Python source code into a natural language docstring. Towards this end, we propose CodeT5++, an extension to CodeT5 where we introduce novel pre-training tasks that capture relevant source code features most useful in code summarization tasks. Specifically, we pre-train the model to (1) predict masked return values of Python functions, (2) detect whether a docstring and source code pair is an accurate representation of the function, and (3) predict masked function names of Python functions. Subsequently, we fine-tune the models for the code summarization task and evaluate the performance using a smoothed BLEU-4 score, a precision-based metric applicable in translation tasks. Finally, we analyze how the pre-training steps help improve the summarization tasks.

Transformers | LLM | CodeT5 | Python

Lifecycle management of Tetration Connectors — Cisco Tetration

Connectors bring telemetry and analytics data from various vantage points in a data center. Typical connectors include network switches and routers, application delivery controllers such as F5 Big-IP and Citrix NetScaler, and firewalls. Tetration uses data from such connectors to baseline the behavior in a network and automatically organize the workloads in the data center. In addition, Tetration also recommends Zero-Trust policies and enforces them. As part of this project, I designed and developed a framework for the lifecycle management of connector integration, specifically, to create/deploy, configure, and troubleshoot a connector integration.

NetFlow | IPFIX | AWS VPC Flow Logs | Cisco AnyConnect | LDAP | Docker | Kafka | Golang

Skills

Programming Languages	**Golang**, **Python**, **R**, **C**, **C++**, **Java**
Data	**RDBMS** : PostgreSQL, MySQL, **NoSQL** : Microsoft ADX, MongoDB, RocksDB, Redis, **Graph** : Neo4j, ArangoDB, **Big Data** : Hadoop, Hive, Databricks, **Streaming** : Kafka
Containers and Orchestration	**Docker**, **Kubernetes**
Cloud	**AWS**, **GCP**, **Microsoft Azure**

Other Information

MIDS Project Portfolio	aumahesh.github.io
Publications and Patents	scholar.google.com