Mahesh **Arumugam**

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

Seeking a software architect role in distributed systems that focusses on data and learning insights from data

Technical professional with over 17 years of diverse and progressively responsible experience in network analytics, security, network infrastructure, research and development, team leadership, product development, engineering, problem-solving and mentoring. Excellent verbal and written communication skills, attention to detail, customer focus, and a strong commitment to work. Strong research and analytic skills that are backed up by the ability to deliver results within time constraints.

Recent Experience

Present – October 2021 | **Architect, Zscaler, Inc., San Jose, CA**

Network Analytics | SaaS APIs | Microsoft Azure | Azure ADX | ChatGPT | Golang | Java | Python

October 2021 – May 2021 | **Principal Software Engineer, Molecula, Remote, HQ : Austin, TX**

- Responsible for the architecture and design of Molecula as a Service platform that enables customers to (1) store features in an optimized format, (2) retrieve responses for complex analytical queries quickly, (3) serve features to analytical and ML engines. In this role, contributed to development of control plane for Molecula as a Service that handles management of organizations/tenants, users, roles, deployment of feature store instances, queries to tables on a deployment, etc.

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

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

- Responsible for design and development of connector framework for Tetration . The control plane includes provisioning, configuring, monitoring, and troubleshooting connectors. And, designed and developed various connectors that collect network telemetry from routers, switches (NetFlow / IPFIX), middleboxes, and VPN endpoints. In addition, designed and developed annotation framework for LDAP annotations on flow telemetry.
- Designed and developed a full-stack cloud security posture management system for AWS accounts (e.g., CIS AWS benchmark), with periodic cloud posture PDF report generation.

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

Education

Expected : 2023 | **MIDS : Masters in Data Science** *(in Final Capstone project)*, University of California, Berkeley, CA
Capstone Project : Restor-AI-tion : Ancient Handwritten Japanese Text Restoration
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
Ph.D. Dissertation : Rapid Prototyping and Quick Deployment of Sensor Networks

Skills

Programming Languages	**Golang**, **Python**, **R**, **C**, **C++**, **Java** (Beginner), **Javascript** (Beginner)
Data	**Streaming** : Kafka (Beginner), **RDBMS** : Post-gres, MySQL, **NoSQL** : Microsoft ADX, MongoDB, RocksDB, **Graph** : Neo4j, ArangoDB, **Big Data** : Hadoop, Hive, Databricks (Beginner)
Containers and Orchestration	**Docker**, **Kubernetes** (Beginner), **Ansible** (Beginner)
Cloud	**AWS**, **GCP**, **Microsoft Azure** (Beginner)

Selected Projects

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, the verification of the sources is as necessary as the information they provide. To battle misinformation and disinformation through digital objects, it is salient to provide consumers the ability to verify whether or not information (or data) provided by such sources was altered prior to its use (e.g., publication). 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, Signet-ring supports the lifecycle management of source identities (using cryptographic keys) and relationships between objects and sources. This lifecycle management includes the revocation of source identity keys and previously accepted object-source relationships.

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

Lifecycle management of Tetration Connectors — Cisco Tetration

Connectors bring in 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 the data collected 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, including : (i) creation of the connector integration, (ii) configuration management of the connectors, and (iii) troubleshooting infrastructure.

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

StyleBooks : A declarative configuration language for Citrix ADC — Citrix Systems, Inc.

StyleBooks is a declarative language that allow users to consume NetScaler (now, called Citrix ADC) services in a variety of data center configurations and cloud architectures, providing both configuration simplification and smart operational visibility. It captures useful NetScaler configuration and includes operational aspects (health, counters, logs). New StyleBooks can be created by cloning and modifying existing ones, or by composing existing StyleBooks into new ones, thus, allowing for modular and incremental design. In this project, I was responsible for the following : (i) compiler for StyleBooks that generates an equivalent Python package, (ii) design of the runtime engine that instantiates a compiled StyleBook to create an actual configuration, (iii) design of config audit and config diffs for computing the differences when an existing configuration is updated, (iv) design of the REST APIs.

Citrix ADC | Infrastructure as Code | Compiler | Python

Other Information

MIDS Project Portfolio Papers and reports from MIDS projects. MIDS

Patents 15+ patents (pending and approved).

Publications Published in peer-reviewed conferences and journals. Papers

References

Available upon request.

What made the Pacific Northwest move during the July 4th weekend ?

Statistics for Data Science

The effects of the COVID-19 pandemic have been felt and are still being felt around the world. In particular, the tourism industry suffered extreme economic losses between the many travel restrictions, cessation of communal social activities, and widespread fear of contracting the virus. Specifically, in 2020, there was a 35% reduction in domestic travel, resulting significant revenue loss. By contrast, there was an uptick in domestic travel in 2021. In this project, we seek to understand how vaccination rates are impacting vacation mobility in the Pacific Northwest during summer 2021. We explore the relationship between number of trips between 100 and 500 miles and COVID-19 vaccinations controlled for key related attributes, including population density and number of COVID-19 cases. In addition, we also consider several additional factors that influence travel, such as employment status, state restrictions, and political ideology. We also note that this study only focuses on a small time period -July 4th weekend.

Prediction | Forecasting | R | R Studio

Will Your Flights be Delayed ?

Machine Learning @ Scale

Air travel has a significant role in shaping economies of states and this it is important to optimize every aspect of the flight cycle to increase the quality of airline's services, especially, the consistent on-time departure of flights. However, due to multiple factors, airline flight delay is inevitable, and thus have to be accounted for by the airline company executives in order to optimize value to multiple stakeholders –from the shareholders to the customers. In this project, we seek to address the following question : can weather and airline data collected two hours before the scheduled departure time, predict whether the flight will be delayed 15 minutes or more ? The models we developed in this project, predict the delay with an accuracy of 82%.

Logistic Regression | Decision Trees | Gradient Boosted Trees | XGBoost | SVM | Databricks | Python

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

Natural Language Processing

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++, extensions to CodeT5 where we introduce novel pre-training tasks that capture relevant source code features most useful in code summarization tasks. Specifically, we pretrain 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

Urban Land Use Classification from Public Domain Imagery

Computer Vision

In this project, we investigate the urban land use dataset from UC Merced that consists of 21 land use categories from various urban areas across the United States. We explore various features of the images, starting from simple ones such as mean color values of the color channels and histogram of oriented gradients, to complex features such as bag of visual words, contours of an image, and embeddings from pre-trained deep neural network models (e.g., VGG16 and ResNet101). We perform principal component analysis to reduce the dimensionality of our dataset for better generalizability. We identify features suitable for classifying the images into their respective classes through experiments with various classification models. Finally, we compare and contrast these classification models for efficiency (with respect to the time required for training and prediction) and model accuracy.

Vision Transformers | Python

Restor-AI-tion : Ancient Handwritten Japanese Texts Restoration

Final Capstone Project *in-progress*

Much like cursive writing in English, the Japanese language has Kuzushiji, roughly translating to squished writing. This writing style was used for over a thousand years beginning in the 8th century but is not taught in modern day curricula. As a result, most Japanese natives, aside from a select few experts, can read Kuzushiji. Most conversion of these works from Kuzushiji to modern Japanese is done by hand and takes countless hours for just a single document. However, with the relatively recent effort by museums and libraries to digitize ancient works in an effort to safeguard these priceless artifacts against natural disasters, the use of AI for these conversions has become suddenly more realistic. This project looks to build on existing work to restore deteriorated ancient texts through the use of generative AI such as GANs or the more advanced Diffusion Models. Converting these works into modern legible Japanese can help better understand, as well as preserve the history and culture of the Japanese people.

Vision Transformers | OCR | LLM | Python