Nitin Gopala Krishna Sontineni

700 Health Sciences Drive, Stony Brook, New York, 11790 | Open to Relocation | Portfolio: nitinsontineni.netlify.app/
(631)633-0763 | nitingopalakr.sontineni@stonybrook.edu | linkedin.com/in/nitinsontineni/ | github.com/nitin-sontineni

Education

Stony Brook University **Aug 2023 - Dec 2024(Expected)**
Masters of Science in Computer Science New York, USA
Coursework: Distributed Systems, Operating Systems, Data Science, Advanced Machine Learning, Data Visualization, Gen AI

Birla Institute of Technology and Science Pilani **Aug 2018 - May 2022**
Bachelor of Engineering in Computer Science Hyderabad, India
Relevant Courses: Data Structures and Algorithms, Object Oriented Programming, Database Management, Software Engineering

Technical Skills

Programming Languages/OS: Python, C, C++, Java, Go, Bash, Linux
Web Technologies: React, Redux, JavaScript (ES6+), Node, Express, Django, HTML5, CSS
Tools/Databases/OS: Git, Airflow, Docker, Kubernetes, Kafka, Jenkins, Postman, Hive, SQL, PostgreSQL, MongoDB, Linux/Unix
Frameworks/Protocols/Methodologies: Jest, Selenium, MVC, gRPC, HTTP, REST, GraphQL, Agile, TDD
Cloud Services: AWS(EC2, EMR, S3, Lambda, RDS, SDKs), GCP(Data Proc, Terraform, Vertex AI, BigQuery, Spanner)

Experience

Research Foundation for SUNY **Jun 2024 - Present**
Software Engineer *Stony Brook, New York*

- Created a Docker-based workbench, streamlining the development process by creating a consistent and reproducible environment, which reduced setup time and improved team productivity.
- Automated the process of combining the healthcare data from multiple relational databases and staged the data to standard OHDSI format using Spark to facilitate the faster modeling experiments.
- Maintained and optimized database through sharding and indexing techniques, ensuring efficient data storage, retrieval, and performance tuning, which improved query response times by 12%.

American Express **Jul 2022 - Jul 2023**
Software Engineer *Bengaluru, India*

- Developed backend for a scalable JupyterLab-based workbench hosted on AWS, enhancing modeling journey for 1000+ users.
- Created RESTful APIs in Python for provisioning AWS resources, catering to the infrastructure needs of intensive workloads.
- Optimized cloud resources utilization by deploying Nvidia RAPIDS on AWS, enabling complex models to run entirely on GPUs, leveraged Dask clusters on Amazon EKS and ECS, recuding model training time by ∼70x
- Implemented dynamic resource allocation controls for Jupyter notebook sessions using Unix shell scripting with Bash, optimizing compute resource distribution across a multi-tenant environment and ensuring equitable access for over 3,400 users.
- Designed and implemented PostgreSQL databases with optimized schemas, improving data retrieval times and efficient storage of infrastructure usage data and created an interactive dashboard using React and Node for visualizing KPIs.
- Integrated automated testing with Jest and Selenium to minimize production bugs and achieved 86% code coverage.

American Express **Jul 2021 - Dec 2021**
Software Engineer Intern *Gurgaon, India*

- Leveraged React, Node, MySQL to develop a model deployment platform, maintained security measures and compliance protocols, including HTTP and OAuth for secure API access, significantly enhancing the user experience.
- Developed backend CI/CD pipelines for machine learning models deployment in Kubernetes using docker and helm, reduced the effort spent on onboarding a new model deployments by 70%.
- Revamped HiveQL queries to optimize the data pipelines for usecases like customer segmentation, marketing, fraud detection by using indexed views, table partitioning, optimized joins and subqueries, improved the execution time by 20%.

Strand Life Sciences **May 2021 - Jul 2021**
Software Development Intern *Bengaluru, India*

- Developed a SaaS application using React, Django, MongoDB hosted on AWS enabling processing of large-scale NGS data.
- Automated exporting results to designated AWS S3 buckets and termination of EC2 instances post job completion, optimizing resource utilization and cost efficiency.

Projects

AI for Health | *Generative AI, Deep Learning, LLMs, Spark, PyTorch | Research under Prof.Fusheng Wang*

- Created dashboard utilizing Tableau, integrated with Postgres database of 3TB medical records, facilitating data-driven decisions.
- Implemented a Generative AI model using LLaMA 2(Large Language Model Meta AI) and Bi-Attention GNN to predict patient care outcomes, boosted AUROC by 4.2% for readmission, F1-score by 8.4% for drug recommendation.

Video Streaming Server | *Cloud Computing, Media Streaming, Golang, AWS*

- Created a Go-based video streaming server, integrated TLS Protocol, AWS S3 for storage, and adaptive bitrate streaming to deliver high-quality video content dynamically based on user connectivity and device capabilities.

EditRox | *Java, Spring Boot, React, WebRTC*

- Developed a application using Java, Spring Boot and React, a pair programming platform with end-to-end encrypted HD video calls via WebRTC, real-time code collaboration, and integrated compiling, supporting multiple languages and themes.

CourseHub | *Full Stack Web Development, React, Node, Express, MongoDB*

- Designed and deployed an online learning management system (LMS) using MERN Stack, providing an intuitive platform akin to Canvas for professors, and students. Successfully scaled to support 11 instructors and over 400 active student.