Nitin Gopala Krishna Sontineni

700 Health Sciences Drive, Stony Brook, New York, 11790 | Open to Relocation

(631)633-0763 | nitingopalakr.sontineni@stonybrook.edu | linkedin.com/in/nitinsontineni/ | github.com/nitin-sontineni

Education

Stony Brook University **Aug 2023 - May 2025(Expected)**

Masters of Science in Computer Science New York, USA

Coursework: Data Science, Advanced Machine Learning, Computer Vision, Data Visualization, Operating Systems, Generative 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, Databases, Data Mining, Deep Learning

Technical Skills

Programming Languages: Python, Go, C, C++, Java, JavaScript

Tools and Databases: Excel, Git, Airflow, Docker, Kubernetes, Tableau, Kafka, Hive, SQL, PostgreSQL, MongoDB

Frameworks, Libraries: React, Django, CUDA, Spark, Hadoop, Tensorflow, PyTorch, NumPy, Pandas, D3.js, GPT, LLaMA

Cloud Services: AWS(EC2, EMR, S3, Lambda, Cloud Watch), GCP(Data Proc, BigQuery, Vertex AI, Looker, Cloud Storage)

Experience

American Express **Jul 2022 - Jul 2023**

Data Engineer *Bengaluru, India*

- Developed a scalable JupyterLab environment hosted on AWS and built RESTful APIs in Python to provision AWS EC2, EMR clusters, enhancing the modeling journey for 500+ ML modelers and enabling efficient distributed AI/ML workloads.
- Implemented resources allocation controls for each notebook sessions leveraging Unix shell scripting, improved shared environment experience by managing system resources, ensuring equitable compute access for over 3400 users.
- Orchestrated microservices with Airflow for automation of Data Purging in AWS S3 buckets reducing unwanted costs, sending reminders to users based on AWS usage patterns.
- Designed and created Postgres database with optimized schema for efficient storage of AWS log data and created an interactive dashboard using Tableau for visualizing usage KPIs, enabled real-time cost monitoring and management.
- Built a recommendation system by creating multi-class classification XGBoost model that suggests the optimal AWS clusters based on user inputs, resulting in saving $25,000 quarterly through optimized cluster usage.
- Spearheaded the migration of over 30 AI/ML models to CUDA programming for distributed GPU computing, improving the model training time by ∼70x, which led to the faster development and experiments.

American Express **Jul 2021 - Dec 2021**

Data Engineer Intern *Gurgaon, India*

- Created CI, CD pipelines for machine learning models deployment in Kubernetes using docker and helm, reduced the effort spent onboarding a new model deployments by 60%.
- Developed a PoC for Model Retraining and integrated into deployment pipeline for auto-retraining of deployed models by creating a ETL pipeline using Kafka, reduced the manual intervention and increased operational efficiency.
- Revamped Hive queries and Optimized the data pipelines for usecases like customer segmentation, loyalty programs success, fraud detection by using indexed views, table partitioning, optimized joins and subqueries, achieved 20% improvement in retrieval time.

Strand Life Sciences **May 2021 - Jul 2021**

Software Development Intern *Bengaluru, India*

- Developed a SaaS application using React, Django, PostgreSQL, AWS SDKs 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.

TIBIL Solutions **May 2020 - Jul 2020**

Data Science Intern *Bengaluru, India*

- Engineered a Hybrid ARIMA-BiLSTM wind power forecasting model with PyTorch utilizing 40GB of wind turbine timeseries data, achieved 7.2% MAPE. Transformed wind turbine data using SQL, segmenting data by temporal and environmental factors.

Projects

AI for Health | *Generative AI, Deep Learning, LLMs, Spark, PyTorch | Research Assistant under Prof.Fusheng Wang* **May 2024**

- Streamlined 6TB of medical data pipelines with PySpark for distributed processing, decreased processing time by 24%.
- 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.

Dashboard for Visualizing US Tech Layoffs | *Data Visualization, D3.js, React, Django, Python* **Apr 2024**

- Developed a interactive dashboard for visualizing layoff trends across US using D3.js, React, Django and Python.
- Implemented dynamic user interactions like time-range selection, sector filtering, geographical location selection.

Chat application for APIs | *LLM, Python, GPT, SQL, Kafka, Streamlit, SQL, Airflow, Spark* **Dec 2023**

- Built a GPT-powered application to interact with and extract information from any API, streamed using Kafka. Web app built using the streamlit framework. LangChain's React SQL Agent used to interact with the SQL database.

Flight Delay Prediction | *Predictive Modeling, Feature Engineering, LSTM, XGBoost, Selenium* **Nov 2023**

- Combined diverse data sources like weather patterns, air traffic, and historical flight data using Python and performed exploratory data analysis to evaluate causes of flight delays across the US.
- Built predictive models for flight delays utilising LSTM and XGBoost and achieved a R2-squared coefficient of 0.98.