Naga Shrikanth Ammanabrolu

http://www.naaga.me

Email : naga@gatech.edu

Mobile : +1-312-483-7547

Experience

- **Memorial Sloan Kettering Cancer Center** New York, NY
  Data Engineer II *Sept 2020 - Present*
  - **Data Lakehouse**: Architected, implemented, and maintained a data lakehouse platform on AWS leveraging PySpark, Glue, Athena and Apache Hudi to ingest batch and streaming data, migrating away from IBM Cloud
  - **Ingest Automation**: Reduced 4000+ man hours/year by automation of data pipelines, leveraging Glue Blueprints and PySpark to design standard reusable patterns and utility modules, moving away from IBM DataStage
  - **Data Quality**: Implemented data quality framework leveraging PyTest to automate data quality checks for ingestion data pipelines. Built Grafana dashboards for monitoring performance metrics of Hive-based tables
  - **Data Enrichment**: Automated Data Profiling processes leveraging IBM Watson knowledge catalog API and prototyped supervised Machine Learning models for optimizing authoritative data ingestion into the data platform
  - **Communication**: Communicated data architectural design plans, held knowledge sharing, developers forum, and peer programming sessions; collaborated with data engineers and data stewards
- **Icahn School of Medicine at Mount Sinai** New York, NY
  Data Engineer II, Scientific Computing *Apr 2019 - Sept 2020*
  - **Clinical NLP**: Architected proof-of-concept and production-scale Clinical NLP solutions using Apache cTakes and Clinithink by extracting SNOMED terms and PHI de-identification of clinical notes
  - **Streaming Data**: In charge of building and maintaining real-time streaming HL7 (Health Level 7) data processing systems through the Iguana engine and prototyped these pipelines on the open-source Mirth engine in JavaScript
  - **Data Lake**: Developed proof-of-concept solutions for data ingestion pipelines leveraging HIPAA-compliant Microsoft Azure services like Databricks, HDInsight, and Data Factory to replace on-prem ETL framework
  - **Common Data Model**: Assisted in building, maintaining and tuning open-source Healthcare Common Data Model systems like I2B2 and OMOP. Improved performance of the I2B2 PostgreSQL instance by 60x
- **Future plc (formerly Purch Group Inc)** New York, NY
  Associate Data Scientist *Mar 2017 - Apr 2019*
  - **Machine Learning**: Deployed forecasting algorithms using Generalized Additive Models to forecast key KPIs, to aid publisher services. Deployed ML solutions for user segmentation, anomaly detection and yield optimization
  - **Data Lake**: In charge of building and maintaining the BI AWS ETL infrastructure post-acquisition. Reduced 7000+ man hours/year by automation of data ingestion leveraging partner network APIs and web scrapers
  - **Cloud Cost Optimization**: Cut down Amazon Redshift footprint by 75% by migrating legacy data pipelines on AWS Lambda and Glue to use S3 and Amazon Athena, saving the business over $120,000/year
  - **Data Analysis**: Performed ad-hoc reporting, statistical analyses, and automation of common data requests
- **Future plc (formerly Purch Group Inc)** New York, NY
  Data Analyst Intern *Jun 2016 - Aug 2016*
  - **Page Categorization**: Led the ‘Categorization of Purch Websites’ project. Developed web scrapers. Built a database and developed ML algorithms to categorize and analyzing the digital publications under the Purch Group
  - **ETL Performance**: Assisted on the ‘ETL performance’ project which benchmarked the ETL Processes on AWS

Education

- **Georgia Institute of Technology** Remote (part-time), USA
  Master of Science in Computer Science; GPA: 3.6 *Jan. 2021 – Aug. 2023*
- **University of Illinois at Chicago** Chicago, Illinois
  Master of Science in Industrial Engineering; GPA: 3.45 *Aug. 2015 – Dec. 2016*
- **University of Mumbai** Mumbai, India
  Bachelor of Engineering in Electronics Engineering; GPA: First class *Aug. 2011 – May. 2015*

Technical Skills

- **Languages**: Python, SQL, R, Java, C **Cloud Technologies**: AWS, Microsoft Azure, IBM Cloud
- **Machine Learning**: Time series analysis & forecasting, Natural Language Processing (NLP), Deep Learning
- **Data**: AWS (Glue, Lambda, Athena, SNS, SQS, Kinesis), PostgreSQL, Oracle, SQLServer, SSIS, DataStage