NILANJAN CHATTERJEE

Phone: 1-704-345-9929|E-mail: nilanjan.9325@gmail.com

LinkedIn: *https://www.linkedin.com/in/nil68657* | GitHub: *https://github.com/nil68657*

AWS | GCP certified polyglot architect with 12+ years of experience in Data and ML Engineering leadership with problem-solving acumen and has proven ability in leading GTM strategy of turn-key solutions and 0-1 products, looking for leadership roles.

EXPERIENCE ~ 12 years

TransUnion – Sr. Engineering Manager – Austin, TX, USA Apr 2023 – Present

- Strategically leading a team of 11 Data Engineers, Data Scientists, SDEs in developing 0-1 turn-key solutions and performing scoping and architectural reviews, streamlining progress on projects for various cutting-edge ML and AI initiatives on fraud, device and credit and non-credit risk, cyber and churn in sprints.
- Spearheading Go-to-Market strategy for 0-1 features on CDP and Data Cleanroom efforts supporting **Lambda** | **Kappa** Data Marts & MLOps for Model execution| registry | serving | monitoring | governance and also enabled GenAI with RAG pipeline and enabled lift of **32% YoY** for fraud and risk identification by efficient identity keying -matching-linking and scoring pipeline using **Feast, Redis, Aerospike**, **dbt**, **MLFlow**, **BigQuery**, **Airflow**, **PubSub**, **Kubectl**, **K8s** and **Trino.**
- Engineered **Aerospike** based Feature Store using **Feast**, **Tecton**, **Kubeflow** to serve offline/online modes for efficient real time and batch inferencing, reducing inference time for Scoring, Fraud, AML checks by 80% from 35ms to 7ms enabling $1.2Mn/y savings**.**
- Designed A/A and A/B testing pipelines for credit and non-credit feature rollout with custom treatment traffic bifurcation.
- Spearheaded ML monitoring framework development to track [data | performance drift | inference] drift using **MLFlow**, **Bento ML**, **Prometheus, Mage** and driving BI layer using **Iceberg**, **BigQuery**, **Superset** & **Grafana** driving $3Mn+ revenue growth YoY
- Spearheaded MLOps serving infra using **Spanner**, **Terraform**, **Kong**, **Swagger** for **GraphQL** endpoints for batch inferencing on GCP, AWS and Azure for **Parquet**, **Protobuf**, **Arrow, Iceberg** files for fraud identification leading to $1Mn+ YoY revenue.
- Operationalized cost efficient, & performant datasets, data lakehouses & orchestration jobs supporting business critical needs using Kafka, Flink, Spark, **Hudi**, **Iceberg**, Airflow S3, Redshift, **Timescale DB**, **Aerospike** and data serialization formats **Avro, Parquet**.
- Operationalized enterprise architecture and data governance, lineage, storage and privacy architectures complying with PCI, HIPPA, GDPR, CCPA, GLBA, NIST, SOC 2, ISO 27001 using **Collibra** and making strategic product decisions based on governance principles.
- Developed distributed and scalable Gen AI solutions including RAG using **GPT-4** and **Llama** and custom vectors indexed on **Milvus** and used **Streamlit** serving, offloading existing NLP engine saving 10+ clients $2.5M+ annually.
- Architected and developed HPC clusters to support Gen AI capabilities with RAG and RLHF pipeline using SOTA models
- Developed platform capabilities using **Java Springboot**, **Vue.js,** Terraform and GraphQL & REST APIs on **Swagger**, **Postman.**
- Engineered low latency data ingestion pipelines for threat and intrusion detection solutions using **CrowdStrike**, **SIEM**, **Cloudflare**, **Splunk**, **Beats** and **ELK** stack.

Amazon – Sr. Data Architect – Arlington, VA, USA July 2020 – Apr 2023

- Directed a team of Data Architects, Data Scientists, ML Engineers including task assignment and monitoring with task assignment, weekly review and growth tracking on multiple analytics customer engagements and all-round development leading to promotions.
- Operationalized ML models for pre-post analysis, cost benefit analysis, causal analysis for different A/B, A/A tests for client metrics via automated pipelines using **Lambda**, **SageMaker**, **DynamoDB**, **API Gateway, CloudFormation** leading to **66%** faster TAT.
- Developed ETL pipelines using **Redshift**, **Lambda**, **Glue** and helped teams in optimizing queries and reduced latency delay by 70% by implementing **Iceberg** with **Nessie** and optimized data lake caching design, saving **400 resource-hours annually**.
- Architected Data Strategy and roadmap by efficient Lambda | Kappa | Delta pipelines via Data Lakehouse | Lake |Meshe | Mart and CP/AP data delivery strategy using **Redshift, Aurora** ,**Kinesis**, **Firehose**, **DynamoDB**, **dbt**, **Fivetran, Athena, Airflow** for multiple FinTech | Education | Retail | FMCG clients .
- Architected enterprise data roadmap complying with p9999 latency SLA for clients, defined QA process, and socialized resolution process to ensure data is flowing accurately through data creation to presentation layers and reduced SEV-2 issues by 50% annually.
- Designed and built the data strategies and roadmaps necessary for ML/AI/BI using **Spark**, **EMR**, **Dynamo**, **Glue**, **Redshift**, **Elastic**, **Kibana** to serve analytics and data science needs for the clients reducing delivery misses due to data blockers by 90%.
- Implemented code review best practices and acted as PoC for client engineering teams .
- Led capacity planning attribution model development for Amazon Seller & Buyer starting from seller onboarding till shipment, boosting operation performance by 200bps and enabled proactive RBAC based data delivery using **Kinesis**, **Glue**, **Dynamo**
- Set up Airflow environment for job orchestration and created DAG's and fork scripts for local executor to schedule python jobs to run weekly and update S3 and AWS Batch. Developed DAGs with multiple executors for real-time processing with efficient caching by integrating **Kafka**, **Hudi**, **Flink** and improved failover consistency with Python scripts and migration from **cron**, **Jenkins** integration.

Hughes Network Systems – Sr. Data Mining Scientist – Germantown, MD, USA Jan 2018 – July 2020

- Worked on anomaly and outlier detection for network usage data and integrating it to Tableau reports for QA team.
- Developed NLP/NLU processing pipelines using AWS **Sagemaker**, **Textract**, **spacy**, **gensm**, **AllenNLP**, **ELMO** and **RoBERTa** for text summarization from survey response data for multiple different LOB operations with custom corpus for training.
- Set up GCP clusters with AutoML jobs for image classification for installer team and wrote Bigquery queries and PubSub jobs to streamline the data ingestion process.
- Created NLP models for NER, Topic modelling, POS tagging, n-gram models, Sentiment Analysis and sarcasm detection m and other feature extraction from Survey feedback data to understand driver metrics using Python, Oracle, Trifacta, Excel, R, and Tableau. Used **vowpal-wabbit**, spacy and H20 for NLP task reporting to Tableau dashboard with periodic refresh.
- Developed hive UDF for WiFi diagnostic data analysis using Python and MapReduce.
- Built data warehouses & lakes using **Oracle 11g, 12c [Data Guard, ASM, RMAN, RAC]** , enabling data pumping, patching, upgrade.
- Set up sqoop jobs for importing data from Oracle tables into HDFS and managing them by Hive querying and wrote customized Hive UDF's and also worked with Impala for data aggregation finally stored the data in S3 buckets. Data from S3 was used for analytical processing using AWS Kinesis, Athena, Glue, Lambda and SAS VA, SAS DI and connecting to Tableau.

The information is correct to the best of my knowledge| Date: 10-Apr-25 | H1B Visa | Open to Relocation

- Implemented ARIMA with exponential smoothing and effective differencing as per ACF/PACF/Dicky Fuller tests and in addition to Kalman filter for better prediction accuracy. Also working on k-NN, mixed model ANOVA, Poisson regression in SPSS modeler, Python for product assortment.
- Minimized supply-chain and logistics bottleneck for retail product assortment across distribution center by 4% enabling 6% lift in profit. Also, built dashboards integrating S3, Tableau with Hive connector and SAS **Viya**.

IBM Corporation - Data Science Intern - Durham, NC, USA **June 2017- December 2017**

- Developed solutions for BlueMix team by feature engineering of usage pattern data/ customer churn & developing algorithms like PCA, SVD, Market Basket Analysis, kNN, GBM, Hierarchical Clustering, Logistic Regression with Bayesian Belief, MCMC.
- Worked with HR Analytics team implemented NLP, IR, ranking algorithms to predict employee attrition rate and identify them.
- Have designed model deployment and optimization strategy using **CPLEX**, **Gurobi**, **LAPack** with Spark **MLib** platform.
- Have created visualizations using Python (Matplotlib, Seaborn), R(ggplot2) , Created roll out and pilot testing strategies along with Sales/Pre-Sales team based on model performance and setup A/B testing for challenger vs champion model on Bluemix.

Cognizant Technology Solutions – Senior Data Engineer- Bangalore, India **July 2013- August 2016**

- Led a team of 3 Data Engineers towards implementing data engineering pipelines using Informatica, Talend, Cognos BI, SAP Business Objects following best practices for Marketing team leading **$500K** in YoY revenue growth.
- Data mining on 10TB+/daily of data using **OpenRefine**, **Trifacta**, **Kafka** to enable XGBoost development.
- Developed predictive models using R, Python, Scala and presented on scrum meetings for consumer analytics. Had setup 50+ node Spark cluster for stream processing of 10TB+ data for pilot phase using Kafka, **Flume**, **Impala**, **HBase**, **Oozie**.
- Qualitative data and sentiment analysis using Topic modelling & build NLP models and visualizations using R.
- Migrated DW's and data lakes to 50+ node Hadoop cluster with Pig, Hive, HBase, Impala, Oozie, Flume, Ambari, Mahout.
- Extensive experience in importing data from various sources in Tableau and creating dashboard and presented in MBR and QBRs.
- Worked on AWS for cloud integration using S3, EBS, EMR, Dynamo, RedShift, Lambda, Lex following CRISP-DM.
- Implemented end to end CI/CD tooling using Git, Maven, Jenkins, Chef, Ansible, Puppet, Artifactory, Nexus, Splunk and Selenium.
- Worked on DW integration and management using Erwin Data Modeler, MSSQL, Power BI, Pentaho, Cognos, UNIX. Have experience managing lifecycle of application deployment by technologies like SQL Server, SSIS, SSMS, Toad, Unix, Putty
- Experience of creating and automating dashboard, reports, scorecards, and visualizations in Tableau, Qlikview, Microstrategy
- Have written efficient advanced SQL queries and cubes using SSIS, SSAS, SSRS, PlSQL & T-SQL, with Microstrategy MDX query

Education

- **BS** in **Electronics & Communication Engg.** – West Bengal University Of Technology, India **2009-2013** **GPA-3.7/4.0**
- **MS** in **Computer Science** - The University of North Carolina at Charlotte, NC, USA **2016-2017** **GPA -3.8/4.0**

Certifications

- AWS Certified Developer Associate | AWS Certified Machine Learning Specialty (ML-S)

Technical Skills

- **Languages:** Python, Scala, C++, Java, Rust || **Cloud:** AWS, Google Cloud, Azure
- **Databases:** Mysql, Oracle, MongoDB, Cassandra, Redis, DynamoDB, Aerospike, Bigquery, Trino, Redshift, Athena, DuckDB
- **Tools:** Git, Jupyter, TFS, Jenkins, SAS Eminer, Orange, WEKA
- **Visualization:** QlikView, Tableau Desktop/ Server, Superset, Quicksight, Power BI, ggplot, bokeh, seaborn,.
- **Machine Learning:** Classification, Regression, Feature Engineering, Clustering, Ensemble Models, Transfer Learning, Reinforcement Learning, Gen AI
- **Statistics**: t/z test, CHAID, Causality, Time Series forecasting, Hypothesis(A/B) testing, ANOVA, Data Mining.
- **Deep Learning:** RNN, CNN, SOM, RBM, GAN, MLP, DBN, Q-Learning, Transformers, Encoders.

Projects:

- **Healthcare claims success data via web scraping** - Developed a web scraper using Python(scrapy) and scraped multiple websites to get healthcare claims data and their success rate. Finally built a dashboard in Tableau with incremental refresh
  - **Technologies:** ***Python(pandas, scrapy, randomforest , xgboost, catboost), Tableau, Oracle, Airflow***
- **Quora Duplicate question using Python -** Developed a character based RNN implementation using neural networks and SVM as Word2Vec, Bag Of Words models with fine tuning via scikit-learn GridSearchCV, Keras, Caffe, Tensorflow, PyTorch.
  - **Technologies:** ***Python, pandas, numpy, scikit-learn, Word2Vec, GridSearchCV, Caffe, Tensor Flow, Glove.***
- **Image segregation algorithm for classifying different types -** Developed an algorithm to classify different types of images from a broad category into classes and label them using PyTorch, WEKA and Orange extensions.
  - **Technologies:** ***NLTK, Stanford-NLP, Tensorflow, Caffe, Apache Solr, HDFS, Hive, Kafka, Druid, HBase.***
- **Page rank Algorithm in Spark-** Developed mappers and reducers allocations for page rank algorithm on Spark.
  Technologies: ***Apache Nutch, Lucene, SolR, Java, Eclipse, Tomcat, Hadoop, Sqoop, Flume, Zookeeper***