INDRAJIT GHOSH

Chicago, IL | (773) 823-8914 |ighosh2@hawk.iit.edu |LinkedIn| GitHub|Portfolio

EDUCATION

Illinois Institute of Technology, Chicago, IL, USA — Dec 2024
Master of Computer Science
Concentration**:** Artificial Intelligence

Maulana Azad Kalam University, Kolkata, West Bengal, INDIA — Jun 2015
Bachelor of Technology, Electronics & Communication Engineering

SKILLS

- **Programming Languages:** Python, SQL
- **Technologies/Skills:** PySpark, HIVE, Machine Learning, Data Warehouse, ElasticSearch, NoSQL, NumPy, Pandas, Problem-Solving, Data Structure, Algorithms, Data Pipelines, Data Visualization, Data Architecture, Data Management, Data models
- **Tools:** Git, SourceTree, GitHub, Jira, Microsoft Office, AWS, Jasper Studio, Airflow
- **Web Development:** Angular8, Apache Spring4, RESTful API, HTML5, CSS3, Object Oriented Design, JSON/XML, SonarQube, PostgreSQL
- **Software Development Lifecycle:** Agile, Waterfall
- **Languages:** English, Spanish, Bengali, Hindi

CERTIFICATIONS

- **Data Engineer** – Dataquest.io (Validate) — July 2024

PROFESSIONAL EXPERIENCE

Data Engineer
Tata Consultancy Services, Mexico City — Apr 2020 - Nov 2022

- Spearheaded the ETL (Extract Transfer Load) pipeline for Citibanamex bank, and **processed large-scale datasets** using PySpark, promoting data accuracy and **efficiency by 40%** and having a user base >300.
- Engineered an automated **big-data workflow**, incorporating data ingestion, cleansing, and PySpark scripting, streamlined client-side data visualization, and cut data **processing time** by over 30%.
- Managed deployment workflows by orchestrating a **CI/CD** pipeline, conducting in-depth code reviews, and executing comprehensive load testing, resulting **in a 30% boost** in **deployment efficiency** and a **20% reduction** in **system downtime**.
- Implemented **ElasticSearch**, improving search performance and data insights by 35%, reducing data retrieval times by 40%.
- Designed and optimized **data models** and **architecture**, enhancing data management and storage efficiency by 25%.
- Utilized **Airflow to orchestrate workflows**, improving the automation and scheduling of data pipeline tasks by 20%.
- Collaborated with project managers and business analysts on use case development across multiple regions, advising on **SQL report feasibility** and execution timelines, and ensuring alignment with **client data requirements**, resulting in a **20% increase** in **KPI efficiency.**
- Orchestrated tailored workshops training for new engineers, **boosting project execution** efficiency **by 25%** through effective skill transfer.

Software Engineer
Tata Consultancy Services, Mexico City/Kolkata — Sept 2015 - Mar 2020

- Led a **team of 8 engineers** as **Tech Lead**, **automated** manual Excel processes, and web applications by **removing human intervention** using RPA, improving **client satisfaction by 40%** and cutting down process time by over 70%.
- Devised a **scalable Know Your Customer (KYC) process** via OCR (Optical character reader) using Automation Anywhere,downsizing Full-time employees (FTE) by deploying bots on virtual machines, **reducing process time by 70%.**
- Engineered **high-performance front-end** with Angular, HTML, and CSS, integrating Spring framework backend connected to Postgres DB in a three-tier architecture, resulting in **an impressive 40% surge in user engagement KPIs**.
- Designed RESTful APIs to meet service level agreements by **optimizing speed**, security, and code quality, achieving **a 35% improvement in response times** and a **50% reduction in security vulnerabilities.**
- **Integrated NoSQL databases**, enhancing system performance and reliability by 30%, and improving data storage scalability by 40%.
- Employed **RDBMS for structured** data storage and retrieval, increasing data query efficiency by 25% and ensuring robust data management.

Data Engineering Projects

Real-time Socket Streaming and Sentiment Analysis Using Yelp Customer Data and LLM

- Built an end-to-end data pipeline that streams, performs sentiment analysis, and handles large volumes of Yelp customer reviews in real-time using TCP/IP Socket, Apache Spark, GPT-3.5 Turbo API, Kafka, and Elasticsearch & docker. Github Here

Reddit Data Engineering Pipeline

- Built an end-to-end data pipeline that extracts, transforms, and loads (ETL) Reddit API data into an Amazon Redshift data warehouse using Apache Airflow for orchestration, Celery, PostgreSQL, Amazon S3, AWS Glue, Amazon Athena, and Amazon Redshift & docker. Github here

Real-Time Change Data Capture (CDC) for Postgres Using Debezium and Kafka

- Implemented a Change Data Capture (CDC) architecture to track and stream real-time changes in a Postgres database using Docker, Debezium, Kafka, and Apache Zookeeper. Github Here