INDRAJIT GHOSH

Chicago, IL | (773) 823-8914 | ighosh2@hawk.iit.edu | www.linkedin.com/in/indrajit-ghosh |
https://ighosh2.github.io/Indrajit_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, Data Warehousing, ElasticSearch, NoSQL, Data Pipelines, Data Architecture, Data Management, Data Modeling, RDMS, PostgreSQL
- **Data Analysis/Processing Libraries:** NumPy, Pandas
- **Tools:** Git, GitHub, Jira, AWS, Airflow, SourceTree
- **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 — Jan 2019 - Nov 2022

- Spearheaded development of **ETL pipelines with 70 million raw records** using PySpark, Apache Spark, and Cloudera, improving **data accuracy** by 40% and processing efficiency by 30%.
- Implemented and optimized **automated data pipelines** for large datasets using Apache NiFi and Kafka, reducing processing time by 30%.
- Led the management **of CI/CD pipelines** using Jenkins, Git, and SourceTree, enhancing deployment efficiency by 30%.
- Led the end-to-end process of analytic tooling feature development, from requirements evaluation to execution**, QA testing,** and stakeholder communication.
- Ensured **99.8% uptime** while managing data ingestion from multiple sources using Apache Spark, Amazon S3, and Python.
- **Automated & scheduled** ETL processes with AutoSys and Apache Airflow, reducing manual workload by 29% and significantly improving operational efficiency.
- Orchestrated the deployment of ElasticSearch clusters, **enhancing data retrieval speed** by 40% and **improving data insights and ad hoc analysis** by 35%.
- **Translated client business requirements** into actionable use cases using SQL and Python, increasing KPI efficiency by 20%.
- Conducted **workshops on data pipeline best practices** for engineers, including Data Warehousing **and ETL optimization techniques** increasing team productivity by 30%.
- Conducted **comprehensive code reviews** and authored **detailed test case documentation**, ensuring seamless and **error-free code migration** to production environments.
- Built and maintained **reporting dashboards** and **visualizations** to design, create, and track campaign/program KPIs.
- Led **cross-functional** collaboration efforts with product managers, data scientists, and business analysts to ensure data solutions aligned with strategic business goals, resulting in a 15% increase in project delivery speed and a 10% improvement in data accuracy.

Software Engineer

Tata Consultancy Services, Kolkata — Sept 2015 - Dec 2018

- **Automated manual Excel processes and web applications** using RPA & VBA, eliminating human intervention, **reducing process time by 70%**, and improving client satisfaction by 40%.
- **Developed a scalable Know Your Customer (KYC) process** leveraging OCR and Automation Anywhere, **cutting process time by 70%** and reducing the need for full-time employees through bot deployment.
- **Designed** and optimized PostgreSQL databases, focusing on **performance tuning** and **schema design**, which improved query performance by 40% and reduced data retrieval times by 35%.

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