Anirban Kar Chaudhuri

Singapore| 84297220 | anirban.karchaudhuri@gmail.com
www.linkedin.com/in/anirban-kar-chaudhuri-7913737b/
https://github.com/Anirbankc1992/ML-Engineering-Projects

PROFESSIONAL SUMMARY

I, Anirban Kar Chaudhuri, have acquired considerable work experience and educational knowledge as a data scientist and engineer over 6.5 years of approximately. I've created big data and machine learning architectures and subsequently translating them to reliable code pipelines. I am a **certified Azure Data Engineer.** I'm aware of and leveraged various cloud based data storage & processing solutions, machine learning model scoring and pre-trained AI models services. My commitment to **engineering innovation** is demonstrated through my design and implementation of modular, reusable and generalisable custom-made python software frameworks build based on recommended best practices. They're built on a foundation of strong programming ethics. My robust data engineering pipelines ensure efficient data storage, querying and quality assurance. I have also undertaken work in data modelling, following best practices for data management and governance that underpin efficient information organisation.

I hold a **master's degree in Artificial Intelligence** from NUS Institute of System Science. I patiently and systematically explore and drill down into relevant datasets, both structured & unstructured, to provide actionable and accurate analysis to users as well as for **creating impactful features for AI models**. I keep my calm temperament even under heat. I understand the need for striking a critical balance between model accuracy and business relevance, as well as automatic model retraining to maintain and enhance model accuracy over time, adapting to new emerging data trends automatically. I enjoy reading widely and sharing technical knowledge and best practices, leading by example. I also communicate complex data insights to relevant stakeholders, facilitating informed decision-making and strategic planning.

SOFT SKILLS

Avid reader and inquisitive knowledge explorer
Deep thinker and pattern seeker
Problem solving using analytical approaches
Excellent written and verbal communication
Teamwork, networking, coordination

TECHNICAL SKILLS

- [Data Engineering]
- [Data Analysis]
- [ETL, ELT]
- [SQL, NoSQL databases]
- [Data Warehousing]
- [Python, R]
- [Hadoop, Spark, HIVE]
- [Machine & Deep Learning]
- [Natural Language Processing]
- [Pytorch, Tensorflow, Sklearn]
- [Numpy, Pandas]
- [Kafka]
- [Azure]
- [Optimisation Algorithms]
- [Reinforcement Learning]
- [Predictive, Descriptive Analytics]
- [Prescriptive Analytics]
- [Git, Jenkins, Sonarqube]
- [Visualisation]
- [Requirements Gathering]
- [Linux, Bash]
- [Batch Processing]
- [Stream Processing]

WORK EXPERIENCE

Data Engineer **Jun 2022 – Present**
DBS Bank Singapore

Responsibilities:

- Proactively design & document enterprise architecture of data ETL pipeline connecting miscellaneous flatfiles(csv, excel, json), APIs and OLTP databases in Confluence. Specify data mappings between source to target and the metadata as well.

- Displayed advanced python programming skills using functional and OOP programming paradigm while creating modular, reusable and generalisable python ETL framework for extracting data from miscellaneous sources (json, excel, csv, etc). Separating pipeline configuration in yaml files from Python code. Experimented with different data structures.
- Enabled **error-handling capabilities** into python programs and logging of errors using python logging module. Ingest log into elasticsearch index, monitoring pipeline error trends via Kibana visualisation dashboard. Quick error troubleshooting to make pipeline workable, iteratively improving ETL framework design.
- Keeping track of data volume and growth, latency, consistency and integrity, lineage & dependency tracking, pipeline resource utilisation & performance metrics and quality metrics. Independently come up with quality assurance and validation tests.
- Based on business user cases, create analytical queries and automate creation of required datasets/reports and visualisations.
- Star schema dimensional modelling techniques to create fact and table dimensions for datawarehouse for business analytics. Pay attention to hierarchies, granularity, join keys.
- Became adept at querying **elasticsearch** database index using REST API, creating indexes with custom mappings, visualising index data via kibana dashboard.
- **Apache Kafka:** Listening to kafka topics from python for instant messages pushed from DBS techdebt UI, processing event on the fly based on inferred schema, pushing to elasticsearch index. Purpose is to test UI properties and workability.
- **Apache Airflow for automatic-scheduling** of daily ingestion and transformation tasks with DAG defined workflow sequences and dependencies, specifications for frequencies and timings of tasks.

Data Engineer **Dec 2021 – May 2022**

Capgemini Singapore Pte Ltd

Client: Toyota

Responsibilities:

- Develop pyspark scripts for cleansing, transformation of data (structured and unstructured) for end user data ingestion in **Azure CosmosDB** using Azure Spark databricks. Develop data quality and validation test cases both at source (Blob storage) and target (**CosmosDB**). **Azure Data Factory** as data syncing and transformation ETL pipeline orchestrator.
- Work with stakeholders including the Product owner, Data and Design teams to assist with data-related technical issues and discuss understand business problems at hand and to design data architectural needs.
- Insightful Analytics and Visualization: Mining Toyota's data to identify servicing trends, creating a PowerBI dashboard that highlights vehicles likely needing maintenance.
- Detailed investigation and analysis into source duplicates at SAP side. Find ways to stream data using Spark streaming API to provide live insights for vehicle reparation status.
- Optimize query performance by designing appropriate partition keys, composite indexes, and query patterns tailored to the application's access patterns and workload characteristics. Develop efficient queries and analytics pipelines using SQL-like query language (SQL API) supported by Azure Cosmos DB, leveraging built-in functions and operators for filtering, aggregation, and projection. Utilize Cosmos DB Change Feed to capture and process incremental changes to the data in near-real-time, enabling reactive data processing and downstream analytics.
- Implement advanced optimizations such as partitioning, caching, and cluster autoscaling to improve the speed and efficiency of data transformations.
- Leverage Databricks notebooks for collaborative development and experimentation, ensuring best practices in code versioning and documentation.

Big Data Engineer **Dec 2020 – Dec 2021**

Citibank Singapore

Vendor Payroll: Virtusa Singapore Pte Ltd

Project: Anti-Money Laundering Datamart migration from legacy Oracle systems to HIVE & Spark based distributed platform & miscellaneous implementation activities for Datamart maintenance & improvisation

Highlights: Performance rating among top 4-5% Consultants in Virtusa Singapore. Part of dynamic, high performing tea whose output resulted in 60% reduction in ingested data ETL timing after migration.

Responsibilities:

- **Deep Diving into Data Architecture:**
  - Intensely analysing and understanding the data dictionary and metadata, meticulously analyzing various database constructs such as tables, views, functions, and the intricacies of Oracle SQL package body of length above 3000 lines. We are looking at customer transaction, account, self-profile and geographical travel data.
- **Building & Finetuning Pyspark Functions:**
  - Requirements gathering for complicated business transformation rules (e.g addresses, bank account, transaction number format, datetime format), using proper joins, window and groupby aggregations. Translate requirements into accurate, tested code.
  - Monitoring Spark UI page, looking out for unusually long running pyspark functions. Finetuning spark jobs using techniques like cache, repartitioning/coalesce, customising executor memory and their numbers.
- **Autosys Auto-scheduler:**
  - Autosys definition of jobs (scripts or batch files), their execution dependencies, job workflows and their schedules, actions to take on success or failure.
  - Adeptly & routinely running and monitoring batch jobs across various environments (Dev, SIT, & UAT) in distributed big data platform, ensured seamless data flow and processing. Utilizing Autosys CA for automation workflow management. I expertly managed job dependencies, scrutinized logs for success or errors, and executed prompt corrective actions when necessary.
  - Accelerated enhancements and accompanying data defects in SIT, UAT & production big data environment. Helping foster team culture of continuous improvement and operational excellence.

Data Science Freelance **Mar 2020 – Dec 2020**
Organisation: Flownote AI

MACHINE LEARNING ENGINEER (Contract) **Aug 2019 – Feb 2020**
Keppel Data Centres, Singapore

Achievements: Achieved 98% classification of data centre physical asset faults on approximately generating 20,000 device events daily whose number projected to grow. Ease of equipment maintenance and real time condition reporting.

Responsibilities:

- **Azure Blob Storage** for scalable and cost-effective storage of raw and processed sensor data, data quality checks and validation processes to ensure the integrity and accuracy of the sensor data.
- **Integrated Azure Stream Analytics** for real-time processing of streaming sensor data, enabling rapid detection of anomalies and insights into equipment performance. Developed streaming analytics jobs to aggregate, filter, and enrich sensor data in real-time, facilitating proactive maintenance actions. Utilized windowing functions and temporal querying to analyze patterns in streaming data, identifying equipment faults and performance degradation
- Integrated **Azure Stream Analytic**s with **Azure Databricks** and **Azure Data Factory** for end-to-end data processing workflows. Leveraged Azure Databricks for advanced analytics and model training, while Azure Data Factory orchestrated data movement and transformation tasks. This seamless integration ensured efficient data processing from ingestion to insights, empowering proactive maintenance actions based on real-time and historical data analysis.
- Developing LSTM Neural Network by feature engineering critical sensor obtained data points on data centre chiller and coolant voltage, current, temperature, and pressure to construct a nuanced, multi-dimensional feature set. Predicted equipment failures at specific time instances and the remaining useful life of physical assets within the data centre accurately. Consequently, this made proactive maintenance scheduling and resource allocation feasible, significantly reducing downtime and operational risks.
- Leveraging MLflow module in Azure databricks for meticulous tracking of model experiments, parameters, and performance metrics. Rigorous logging of model metrics and versions, enabling transparent audit trails, facilitated seamless model evaluations and comparisons.
- Automating the model retraining process, intelligently triggered by any signs of performance degradation attributable to data drift. This also bolstered our model governance practices, ensuring compliance and consistency across all machine learning initiatives.
- Exposing model as REST APIs, enabling interaction between our machine learning models and operational systems, driving automation and intelligence at the core of our operational processes.
- View prediction trends in a Grafana dashboard in real-time, finding patterns in faults. Compare them with voltage, current and temperatures of different equipments. Empowered senior management with actionable insights.

DATA SCIENTIST (Intern) Dec 2018 – June 2019
Smart Consulting Solutions Pte Ltd, Singapore

Project: Weighted average price prediction and forecasting of real estate

Achievements: Built predictive model successfully and rigorously tested used by client to save cost by liaising with desired housing developers and choosing best micro-markets to build residential estates.

Responsibilities:

- Utilized K-means clustering for effective market segmentation into 15 distinct real estate categories, chosen for its efficiency over hierarchical clustering based on performance metrics.
- Design of Experiment framework to systematically treat different segments for nuanced analysis of real estate segments, identifying key variables affecting pricing and liquidity to unveil causal relationships. Setting independent variables categorised into economic, location-specific, property features factors.
- Performed detailed statistical analysis (mean, median, range, mode, skew) and utilized visualization tools (histograms, scatterplots, heatmap, boxplot, kernel plot) to uncover segment trends, patterns, relationships, facilitating hypothesis generation on market behaviours.
- Applied XGBoost, Random Forest, and Linear Regression for accurate price predictions, complemented by ARIMA for future price forecasting, aiding in strategic client advisories. Assessed construction project scopes beyond standard benchmarks, conducting risk analyses to guide profitable developer partnerships. Feature selection using chi-square and t-test, feature manually generated based on insights derived.

R&D ENGINEER Sep 2017 – Nov 2018
Exicom Telesystems PTE Ltd, Gurgaon, Delhi-NCR region, India

EDUCATION BACKGROUND QUALIFICATIONS

Masters of Technology in Intelligent Systems(Part-Time), **National University of Singapore**, Jan 2020 – Jan 2023
Bachelor of Engineering in Electrical Engineering, **National University of Singapore**, Jul 2017

TECHNICAL CERTIFICATES

Azure Data Engineer Associate, Azure**,** Mar 2023
CCA175 - Spark & Hadoop Developer – Python (Pyspark), Udemy, April 2021
Data Warehouse Concepts, Basic to Advanced Concepts, Udemy, Feb 2021
CNN for Computer Vision with Keras and Tensorflow in Python, Udemy, July 2020
NLP - Natural Language Processing with Python, Udemy, July 2020
REST APIs with Flask & python, Udemy, Feb 2021
Mastering Tableau, May 2019

Publication

Edge Detection Methods Comparison | by Anirban Kar Chaudhuri | Medium, May 2021

Notable Academic Project

Chatbot Creation for diagnosis of depression causes and conversation for recommendation, integration with telegram, July 2021