EDDIE (BO JIAN) HO

edwin.ho.bj@gmail.com • +1 (626) 662-5830 • github.com/noodleslove • linkedin.com/in/eddie-bojian-ho • eddieho.xyz

EDUCATION

University of California - San Diego — **San Diego, CA**

B.S. in Data Science • Minor in Cognitive Science • GPA 3.8 • Provost Honors — **Sep 2021 - Mar 2024**

SKILLS

Programming Languages: Python, R, Scala, SQL, Shell

Machine Learning Stack: Pytorch, Pytorch Lightning, Tensorflow, Keras, XGBoost, scikit-learn, Scipy, numpy, pandas

Big Data Stack: Apache (Hadoop, Spark, Hive, Kafka, Airflow), dbt, BigQuery, ETL/ELT/ETLT Pipeline, snowflake, Data (Warehouse, Lake)

Cloud/Database Stack: AWS (Athena, S3, Glue, EC2, EBS, Kinesis, Lambda), Azure (Blob Storage, DataFactory, Devops), Amazon Redshift, Snowflake, MS SQL Server, PostgreSQL, MySQL, MongoDB, Cassandra

Data Visualization: Tableau, Amazon Quicksight, Looker, plotly, seaborn, matplotlib, Power BI

PROFESSIONAL EXPERIENCE

LESSO America Inc., Supply Chain Management (SCM) — **Corona, CA**

Data Scientist — **May 2024 - Present**

Description: LESSO America is a leading player in the field of plumbing and irrigation. The SCM project enhances the efficiency of the supply chain by developing data-driven models that predict demand, optimize inventory levels, and reduce lead times. As a Data Scientist my responsibilities lie in developing and productionalize data science models as products to optimize business decisions, establishing claims forecasting models to reduce costs.

- Applied **machine learning** algorithms with **scikit-learn** and **deep learning** with **Pytorch** for claims forecasting models, such as **K-means Clustering**, **XGBoost**, and **LSTM** algorithms to predict demand and inventory management resulting in a 30% improvement in inventory turnover and a reduction in excess stock by 25% within 3 months
- Establishing model automation pipeline including data preprocessing, predictive model scoring, validation, and data population, reducing 50% processing time and 60% resource usage
- Utilized **Apache Airflow** for batch processing, loading transformed data into **AWS S3** as a centralized data repository, and subsequently utilized cloud-based data warehouses with **dbt**, **Amazon Redshift**, and **Snowflake** to handle large volumes of structured and unstructured supply chain data for storage and analytics
- Architected optimized **data architecture strategies,** such as **Snowflake schema** and **virtual warehouse** within **Snowflake**, improving data query performance by 40% and reducing storage costs by 25%
- Developed a comprehensive **data governance framework** that included secure **role-based access controls** (**RBAC**) and **data masking** to ensure that sensitive data in Snowflake was protected
- Implemented interactive dashboards with **SQL**, **Power BI** and **Tableau** provide real-time insights into supply chain performance, including demand forecasts and inventory levels to identify **key performance indicators** (**KPI**)

UC San Diego, Analytics Portal — **San Diego, CA**

Student Engineer — **Dec 2023 - Mar 2024**

Description: Analytics Portal is a centralized analytics platform which empowers university departments by providing real-time insights and data-driven decision-making capabilities. My role as a student engineer is responsible for data integration using GraphQL, data pipeline development, and dashboards development.

- Developed custom resolvers in the **GraphQL** schema to handle complex data queries and relationships across different departments; ensured the **GraphQL API** is secure and performant, implementing access controls and optimizing query execution to handle large datasets
- Utilized **AWS S3** as the central storage repository for raw and processed data, and integrated **AWS Glue** to orchestrate and automate data transformation, creating a catalog of metadata to facilitate easy data discovery
- Designed and developed visually compelling dashboards in **Tableau**, incorporating advanced features such as drill-downs, filters, and calculated fields to enhance user interactivity
- Utilized **Agile** methodologies and tools to effectively prioritize and manage quantitative analytics tasks, ensuring efficient utilization of resources and meeting project deadlines

Avanade, Enterprise Data Warehouse (EDW) — **Los Angeles, CA**

Azure Data Engineer Intern — **Dec 2022 - Mar 2023**

Description: Avanade is the leading provider of innovative digital, cloud and advisory services, with over 59,000 employees in over 26 countries. The EDW consolidates data from disparate sources, centralizing data as a single source of truth. As a Data Engineer Intern, I worked on the data storage solution like Azure Data Lake integration, tiered storage strategy, and data partitioning.

- Leveraged **Azure Data Lake** for scalable, cost-effective storage of raw and processed data, enabling future data exploration and advanced analytics
- Implemented a tiered storage approach with **Azure Blob Storage**, categorizing data into **Hot**, **Cool**, **Archive Tier** based on frequency of access and business importance, optimizing storage costs while ensuring quick access to critical data
- Applied partitioning techniques with **Horizontal**, **Vertical**, and **Range Partitioning** to manage large datasets efficiently, improving query performance and reducing resource consumption
- Developed **ETL pipelines** using **Azure Data Factory** and **SSIS**, automating the data ingestion process from on-premises and cloud-based sources, including CRM systems, ERP systems, and external APIs

CERTIFICATES

Azure Data Engineer — **Aug 2024**

IBM Data Engineering — **Jun 2023**