

SHANTANU FUKU

☎ 361-466-7316 ✉ shantanufuke1997@gmail.com [LinkedIn](#) [Github](#) [Portfolio](#)

Summary

Data Engineer with 4+ years of experience in designing and automating ETL pipelines, integrating legacy systems with cloud platforms (AWS, Azure, GCP), and delivering scalable solutions using tools like Apache Spark, Airflow, and dbt. Proficient in data lakehouse architecture, cloud-native orchestration (ADF, Databricks), and performance tuning using PL/SQL.

Education

Texas A&M University Kingsville

Master of Science in Computer Science

August 2023 – Aug 2025

Kingsville, Texas

Experience

Texas A&M University Kingsville

Graduate Research Assistant

Jan 2024 – Dec 2024

Kingsville, TX, USA

- Built modular **Python pipelines** to convert large FASTA files into model-ready format, reducing processing time by **60%**.
- Applied **CNN**, **RNN**, and **LSTM** models for DNA sequence classification, improving prediction accuracy by 35% over baseline.
- Structured the pipeline for reusability with clear separation of data loading, transformation, and model training stages.
- Developed and optimized **C++** code to generate mutated **DNA sequences**, accelerating synthetic dataset creation by **70%**.
- Used **Amazon S3** to centralize raw and processed genomic data, streamlining access for preprocessing and training.
- Built a **Databricks workflow** triggered by DNA file uploads to **AWS S3**, using **PySpark** to read sequences and append predictions from a trained **Keras model**, saving results as a CSV file and reducing manual workload by **80%**.

Deloitte Consulting Private Limited

Data Engineer | Oracle PL/SQL Developer

Oct 2021 – Jul 2023

Pune, India

- Streamlined **GCP** to **Oracle RMS** data transfer using **PL/SQL** procedures, reducing stock ledger upload time by **40%**.
- Built **PL/SQL** integration to load **retail invoices** into **SAP ERP**, improving workflow efficiency and **data accuracy** by 60%.
- Led development of an **Oracle Retail Cloud** ETL program to automate location list creation, reducing manual effort by **70%**.
- Developed **BI Reports** for invoices, stock ledger, and deals to enhance insights in the **Oracle Retail Suite**.
- Replicated Oracle transformations in **PySpark/Spark SQL** for benchmarking and cloud-readiness testing.
- Simulated PL/SQL batch jobs in **Apache Airflow** to improve orchestration visibility and recovery.
- Rebuilt legacy Oracle BI logic in **Snowflake** using modular **dbt** models as part of a cloud migration initiative.

Tata Consultancy Services Ltd.

Data Engineer | Oracle PL/SQL Developer

Apr 2019 – Oct 2021

Pune, India

- Built **PL/SQL ETL** to migrate **80% of Coupa PO data** to **Oracle EBS**, saving **20+ hrs/week** via batch automation.
- Automated **user creation** in **Oracle EBS** using PL/SQL and email integration, improving onboarding efficiency by **60%**.
- Enhanced **Oracle Forms/Reports** and applied **performance tuning** to support high-volume transactions.
- Scheduled and monitored **data load jobs** (invoices, POs, users) using **Control-M** for timely execution.
- Simulated Control-M workflows in **Apache Airflow** to assess orchestration, dependencies, and alerting for cloud readiness.
- Prototyped Oracle flat-file ingestion using **Azure Data Factory** to simulate cloud-native orchestration.

Projects

Facebook Clone – Full-Stack Social App | Python, Django, JavaScript, HTML, CSS

Jun 2024 - Dec 2024

- Developed a social networking platform with Django, featuring **user registration and authentication**, and **real-time updates** for posts, comments, and likes via AJAX, along with a friend request system, notifications, **real-time chat using WebSockets**.

Olympic Insights Cloud Data Pipeline | Azure Data Factory(ADF), Databricks, Synapse, Power BI, PySpark

Apr 2025

- Built an end-to-end, cloud-native data pipeline using **ADF** and **Databricks (PySpark)** to ingest, clean, and transform Olympics datasets. Loaded transformed data into **Azure Synapse Analytics** and created interactive dashboards in **Power BI**.

E-Commerce Data Lakehouse Pipeline | Azure Data Lake, ADF, Databricks, Delta Lake, PySpark

May 2025

- Built a medallion architecture pipeline using **Azure Data Factory** and **Databricks (PySpark)** to ingest, transform, and store data in **Delta Lake** and queried Gold layer via **Databricks SQL** for sales insights.

Skills

- Programming Languages:** Python, Java, C++, SQL, PL/SQL, Shell Scripting
- SQL Databases:** Oracle, MySQL, PostgreSQL, SQL Server, SQLite
- NoSQL Databases:** MongoDB, Cassandra, Redis, HBase, DynamoDB, Bigtable
- Cloud Data Warehouses:** Amazon Redshift, Google BigQuery, Azure Synapse Analytics, Snowflake
- Cloud Data Lakes:** Amazon S3, Azure Data Lake Gen2, Google Cloud Storage
- Big Data & Streaming:** Apache Spark, PySpark, Hive, Pig, Presto, Apache Flink, Apache Kafka, Apache Storm
- Cloud Platforms & Services:** AWS (EC2, Lambda, Glue, EMR), Azure (ADF, Databricks, Blob Storage), GCP (Dataflow, Pub/Sub, Cloud Composer)
- Workflow Orchestration:** Apache Airflow, AWS Step Functions, Azure Data Factory, Control-M, dbt
- DevOps & Infrastructure:** Git, GitHub, GitLab, Docker, Kubernetes, Jenkins, GitHub Actions, Terraform, CI/CD Pipelines
- Machine Learning:** Scikit-learn, TensorFlow, Keras
- Data Visualization:** Power BI, Tableau, Looker, Google Data Studio, Oracle BI