

Muhammad Uzair Khan

PHONE: +92 311 2025620

E-MAIL: muhammaduzairkhan329@gmail.com

GITHUB: <https://github.com/MuhammadUzair1/>

LINKEDIN: <https://www.linkedin.com/in/muhammaduzairkhan1/>

Summary

Data Engineer with **4+ years of experience** delivering production-grade pipelines across **Azure, Snowflake, Databricks, and AWS**. Specializes in **medallion architecture, ELT orchestration, and optimization**, with real **deployment experience** in healthcare and FMCG environments. Hands-on with **Snowpipe, SnowSQL, Airflow, and ADF**; strong **Python** and **PySpark** fundamentals. Comfortable **bridging engineering and business stakeholders** to deliver reliable, **auditable data systems**.

Work Experience

Data Engineer at WeCrunch (UAE):

Remote – Full Time

December 2024 – Present

- Developed an **autonomous Sharepoint ingestion framework** using **Logic Apps, Blob Storage, and Delta Lake**, eliminating manual file handling and cutting **pipeline trigger time by 70%**.
- Deployed **idempotent, config-driven pipelines** across Bronze, Silver, and Gold Delta layers, reducing post-release **intervention by 40%** and enforcing **data governance standards** across all ingestion layers.
- Implemented **schema-drift detection** and **automated email alerting** within DQ pipelines, cutting mean fault-resolution time **from hours to under 15 minutes**.
- Optimized **Gold-layer PySpark transformations** in Databricks, accelerating **Power BI executive dashboard** refresh cycles for **FMCG multi-country** market strategy.
- Cut **monthly compute and storage costs by 30%** by optimizing cluster configurations in **Databricks**, **pruning unused tables**, and implementing retention policies on Blob Storage.
- Integrated **Azure Event Grid** with **Databricks streaming jobs** to deliver micro-batch market share data, **enabling near-real-time reporting** for operational teams.



AI Engineer at OctDaily (USA):

Remote – Full Time

May 2024 – December 2024

- Refactored **fragmented EHR & clinical databases** into **normalized schemas**, building **high-availability SQL views** that reduced **data inconsistency incidents** by **35%** across downstream applications.
- Provisioned scalable **ADLS and Blob Storage data lake architecture**, cutting **query latency by 25%** and improving data accessibility for **ML and analytics workloads**.
- Developed **end-to-end ELT pipelines** on **Azure Data Factory and Databricks** integrating multiple **EHR/EMR sources**, **improving data flow throughput by 30%**.
- Delivered a **clinical decision-support module** for **medicine recommendations** and **patient history visualization**, accelerating diagnostic prep time for medical staff.



Data Engineer at Create Impact:

Hybrid - Full Time

August 2021 – February 2024

- Productionized **ETL pipelines on AWS Glue and Lambda** processing large-scale transactional datasets, cutting end-to-end processing time by **35%** and **improving SLA adherence**.
- Designed **Snowflake data models** and managed **staged ingestion workflows using SnowSQL**, enabling reliable data promotion **from raw to analytics-ready layers**.
- Maintained **Git-based version control and CI/CD deployment workflows** for **pipeline releases**, **ensuring reproducible builds** and reducing environment-related deployment failures.
- Enforced **data validation and integrity checks** at ingestion, **reducing downstream data errors** and **increasing stakeholder confidence** in analytical outputs.



Education

BS (Computer Science and IT), NED University:

Specialized in **OOP**, **Data Structures**, **Data Warehouse Management**, **NLP**, and **AI**; applied **C++** and **Python** in projects, and independently learned **R**, **SQL**, **AWS**, and **Excel** to enhance data science skills.



Projects

Real-Time Sales Analytics Pipeline – AWS Data Engineering

- Engineered an end-to-end, event-driven data pipeline using **AWS Glue**, **Lambda**, **S3**, and **Event Bridge**, enabling real-time ingestion and processing of high-volume transactional data from retail sources like APIs.
- Designed a **multi-zone architecture (Raw → Staging → Curated)** using **AWS Lake Formation**, improving data reliability and cutting processing latency by **45%**.
- Deployed a **metadata-driven workflow** using **Step Functions** and **SNS**, automating failure alerts, retries, and dependency handling to maintain pipeline integrity.

Customer Information Data Warehousing– Snowflake & AWS Integration

- Designed a production **Snowflake warehouse** integrating **S3**, **CRM**, and **transactional** sources via **Snowpipe**, cutting ingestion **time by 60%** through **automated JSON/CSV staging**.
- Configured **Snowflake Streams and Tasks for incremental processing**; managed staging-to-production promotion via **SnowSQL scripts** with schema versioning controls.

Snowflake Pipeline with CI/CD & Data Quality Framework – Snowflake, Great Expectations

- Designed and deployed a production-grade **Snowflake ELT pipeline** using **SnowSQL** and **Snowpark Python API**, loading **structured** and **semi-structured JSON data** from S3 into staging and analytics layers.
- Implemented a **data quality framework** using **Great Expectations** to **validate schema conformance**, null thresholds, and referential integrity at each pipeline stage — **blocking non-conforming data from reaching Gold layer**.
- Built a **CI/CD pipeline** using **GitHub Actions** to **automate linting**, pytest unit tests, and deployment of **SnowSQL** migration scripts, **achieving zero-downtime** schema changes through version-controlled migrations.
- Enforced **data governance policies** including **lineage tracking**, **audit logging via Snowflake Streams**, and access control standards — making the pipeline **audit-ready from day one**.

Achievements

Runner-Up UBL Datathon by United Bank Limited of Pakistan:

April 2022

Secured **2nd place** out of **300 teams** by developing an **AI-based Instant Decision Loan Portal** with integrated document verification and income proof.



Runner-Up Data Science Competition by IEEE Student Branch NEDUET:

July 2021

Achieved **2nd place** among **30 teams** by proposing a predictive model for **ATM downtime** using machine learning algorithms, resulting in a **20% reduction** in service interruptions.



Professional Certifications

Data Science Certification by **IBM**

Time Series Forecasting by **DataCamp**

Data Engineering Certification by **IBM**

Hands-On Essentials: Data Warehousing Workshop by **Snowflake**

DataBricks - Build Data Pipelines with Lakeflow Spark Declarative Pipelines

SQL by **365DataScience**

Advance SQL by **365DataScience**

Natural Language Processing Specialization by **Deeplearning.ai**

Hands-On Essentials: Data Engineering Workshop by **Snowflake**

Hands-On Essentials: Data Application Builders Workshop by **Snowflake**