

Mian Ali Shah

+92-3314393249 | mianali1720@gmail.com | [linkedin.com/in/mianalishah](https://www.linkedin.com/in/mianalishah)

Technical Skills

Languages: Python, SQL

Tools & Technologies: Apache Spark, Apache Airflow, Docker, NoSQL Databases, Relational Databases, Git, Jenkins, Terraform, Microsoft Power Automate

Cloud Technologies: Databricks, Azure Data Factory, Azure Data Lake, Azure Synapse, Azure Logic Apps, Azure Function App, Delta Lakehouse, Delta Live Tables, Unity Catalog, AWS S3

Certifications

AWS: [AWS Certified Cloud Practitioner \(CLF-C01\)](#)

Databricks: [Databricks Certified Data Engineer Associate](#)

Astronomer: [Astronomer Certification for Apache Airflow Fundamentals](#)

Experience

Data Engineer, [Digifloat](#) – Lahore, PK

June. 2023 – July 2024

Client: [GAIG, US](#)

- Automated data ingestion through batch processing by scheduling daily API fetches and converting zip-compressed files to parquet using Airflow.
- Accomplished a **35%** improvement in data upload efficiency as measured by chunked uploads to Azure Blob Storage, ensuring reliable data handling.
- Performed data cleaning and transformation using Azure Databricks PySpark notebooks triggered by Airflow, creating aggregated tables.
- Delivered real-time data visualization through a Streamlit application deployed with Jenkins, providing up-to-date insights.
- Streamlined environment setup with Terraform for resource provisioning in Azure, enhancing infrastructure management.

Client: [IMEC, Belgium](#)

- Developed Metadata Driven Ingestion Framework to automate data ingestion from different sources; JDBC, and File Systems, simplifying data movement and orchestration for the end user.
- The Framework consists of 3 modules; Apache Spark in Databricks to Ingest Data, Airflow for Orchestration and Web UI to collect user's data sources' metadata information.
- Developed Dynamic dags and Tasks in Airflow to handle users' custom Ingestion Pipeline configurations and frequencies.
- Developed Dynamic Task dependencies in Airflow to cater to users' custom ETL Pipeline tasks and frequencies.
- Developed different Ingestion Modes in Spark; Incremental, Snapshot, and SCD Type 2.
- Implemented a tiered data storage architecture (Bronze, Silver, Gold layers) with Bronze as the raw, Silver for transformed and cleaned, and Gold for data modeling, including the design and implementation of fact and dimensional tables for efficient data representation and analytics purposes.

Client: [Atlas Copco, Belgium](#)

- Developed a monthly recurring Spark job in Databricks to scan all catalogs, schemas, and tables, continuously updating and aggregating metadata into a single metadata table for dashboard creation. This facilitated insights into user contributions, and table space utilization based on user, catalog, and location.
- Enhanced code efficiency by **34%** by optimizing complex transformations, leveraging Spark native functions over loops typically utilized by other developers.
- Utilized Unity Catalog within Databricks to enforce data governance policies, maintain data lineage, and ensure data quality, enabling streamlined data access and governance across projects.

Education

Ghulam Ishaq Khan Institute for Engineering Sciences and Technology

Bachelor of Science in Computer Science

Aug. 2019 - May 2023

Topi, PK