

SHAIKH MOHAMMAD OZAIR

Data Engineer | [Linkedin](#) [GitHub](#)

TECHNICAL SKILLS

- **Programming Languages:** Python, Java, C++, C
- **Python Libraries:** Pandas, Numpy
- **Web Development:** HTML, CSS, Bootstrap
- **Backend Framework:** Android Studio, FastAPI
- **Databases & Connectors:** MySQL, Kafka Connect, MongoDB
- **Data Warehouses:** Hive, BigQuery
- **Message Queue:** Confluent Kafka
- **Distributed Computation Frameworks:** Apache Hadoop, Apache Spark
- **Workflow Management:** Airflow
- **Cloud Platforms:** Microsoft Azure, Google Cloud Platform
- **Azure Services:** Azure Data Factory(ADF), Azure Synapse, Event Hub, Azure SQL Database, Cosmos DB, Key Vault, Function App

PROJECTS

FastAPI Note-Taking App:

- Developed a web-based Notes App using **FastAPI** and **MongoDB**
- Implemented CRUD operations with a dynamic web interface
- Used **Jinja2** for templating and **Bootstrap** for UI
- Planned future enhancements like authentication and deployment

Airbnb Data Processing Pipeline (Batch Processing Every 30 Minutes):

Tech Stack: Azure Data Lake Storage (ADLS), Azure Synapse Analytics, Cosmos DB, Azure Data Factory (ADF), SQL

- **Designed a scalable batch pipeline** to automate the ETL process for Airbnb customer and booking data from ADLS and Cosmos DB into Synapse Analytics, streamlining data processing and improving scalability.
- **Validated and transformed data** through ADF Dataflows, incorporating **conditional splits** and **derived columns**, enhancing data quality by **40%** and ensuring consistency.
- Loaded both customer and booking data into Synapse, leveraging **upsert operations** to maintain accurate and up-to-date records, improving data integrity across systems.
- **Archived processed data** in ADLS as **CSV** files, optimizing storage efficiency by **30%** and ensuring easy access for future reference.
- Automated data aggregation tasks using **SQL stored procedures**, maintaining accurate and real-time reporting capabilities for stakeholders.
- Developed a **Python script** to generate and insert mock booking data into Cosmos DB, simulating real-time data ingestion for testing the pipeline's performance and scalability.
- Orchestrated pipeline execution and ensured seamless data flow using **ADF**, reducing manual intervention and increasing operational efficiency.

Fintech Data Migration Pipeline:

Tech Stack: Azure Synapse, Azure SQL Database, ADLS, PySpark, Delta Lake, Azure Logic Apps

- **Designed and orchestrated a scalable data pipeline** in Azure Synapse, automating the migration of **5 financial tables** from Azure SQL Database to **Azure Data Lake Storage (ADLS)**, reducing end-to-end processing time by **40%**.
- Optimized data extraction and storage by dynamically querying table metadata and writing raw data to the **Bronze layer** in **Parquet format**, improving storage efficiency and query performance by **25%**.
- Developed **PySpark** notebooks to transform data in the **Silver layer**, performing operations such as enrichment, categorization, and schema management, and created **5 fact and dimensional tables** in the Gold layer, enhancing analytics and reporting capabilities by **35%**.
- Integrated **Azure Logic Apps** for real-time success and failure notifications, ensuring **100% communication reliability** to stakeholders and reducing incident response times by **30%**.
- Leveraged Delta Lake for ACID compliance and data versioning in the **Gold layer**, ensuring data consistency and reliability for analytical workloads.
- **Automated pipeline** execution using **Synapse Pipelines**, reducing manual intervention and improving operational efficiency by **20%**.
- **Enhanced data quality** by implementing **data validation checks** and **deduplication** in the Silver layer, ensuring accurate and consistent data for downstream analytics.

EDUCATION

Bachelor of Computer Applications (BCA)

Abeda Inamdar Senior College, Pune

Jul 2022 – May 2025 (Expected), CGPA 8.4/10 (as of 5 semesters)

RECENT ACHIEVEMENTS

- Data Engineering Certificate - Grow Data Skills, 2024:** Earned Data Engineering Certificate from Grow Data Skills, covering key concepts in data processing, storage, and analysis