

# Sheharyar Akhtar

Cell: +92 301 5388946 | Email: [sheharyar7146@gmail.com](mailto:sheharyar7146@gmail.com) | Portfolio: [LinkedIn](#) [Portfolio](#)

## Professional Experience

### Data Consultant, Systems Limited

Aug'23 – Present

- Developed a **Flask**-based data ingestion web application leveraging **Snowpark** for streamlined data integration
- Implemented robust data integrity checks in the Python backend, utilizing **Pandera** for schema validation against sample files stored in **Azure Blob Storage**
- Implemented **CI/CD** pipelines on **Azure DevOps**, automatically refreshing the web application on **Azure App Service** upon changes being pushed
- Analyzed various HR data tables hosted on Snowflake to extract predictive variables for **multivariate forecasting** models
- Employed **AzureML datasets** and Databricks managed **MLflow** to version datasets and models, ensuring reproducibility and traceability
- Utilized the **Snowflake** table within a **Plotly Dash** web application hosted on Azure App Service to present a dashboard, enabling HR business users to derive insights from the predictions
- Developed an end-to-end credit scoring pipeline utilizing **PySpark** and **DataRobot**, enabling predictive analysis of high-risk customers and empowering the client to make informed decisions regarding loan disbursements

### Data Scientist, Afiniti

Jan'21 – Aug'23

- Developed CRM **SCD** and automated **ETL jobs** using **Talend** and **PostgreSQL on Greenplum**, ensuring immediate data processing with triggers for a telecom client.
- Engineered **hierarchical join logic** and **materialized views**, optimizing data retrieval across multiple tables and enhancing data access efficiency.
- Implemented **data integrity alerts and quality controls**, resolving issues in billions of rows, ensuring data reliability across terabytes of information.
- Built and managed **feature engineering pipelines** using **SQL, Talend, and Python**; consolidated predictive variables into MySQL views for **AI model consumption**.
- Developed a **Webex notification bot** to provide **real-time KPI updates**, streamlining data validation and immediate issue resolution.
- Generated **comprehensive data integrity** and **BI reports in RMarkdown**, disseminating critical insights and model performance to stakeholders.
- Led key accounts across Europe, **driving \$4.0 million in monthly revenue** through **strategic data solutions**.
- Designed and maintained **CI/CD pipelines**, achieving data mismatch rates below 1% across billions of records through rigorous testing and deployment processes.
- Utilized Apache Superset and SQL to create dashboards and sensors, enabling **intuitive visualization of data integrity, drift, and join rates** for stakeholders.
- Collaborated with **cross-functional teams** to integrate data pipelines with AI models, ensuring **seamless data flow** and enhancing the **accuracy of predictive analytics**.

## Education

### B.S. Mechanical Engineering, Ghulam Ishaq Khan Institute

Aug'16 – Jun'20

**Relevant Courses:** Engineering statistics, Operations Management, Numerical Analysis, Linear Algebra, Robotics, Control Systems, Microcontrollers, Digital Logic Design

## Skillset

**Programming Languages:** Python | R | SQL | PostgreSQL | noSQL | Stan | Bash

**Data Science:** Tabular ML | Deep learning | Large Language Models | RAG | Vector DB | Multi-Agent Workflows | Causal Modeling

**Databases:** MySQL | GreenPlum | Snowflake | MongoDB | DynamoDB | ChromaDB | Elasticsearch

**Cloud Technologies:** Azure (App Service | Databricks | ML | Blob Storage | DevOps | Pipelines), AWS (AI Services | Lambda | S3 | EC2 | CloudWatch | Glue | Kinesis | Boto3)

**Tools:** RStudio | Anaconda | Jupyter | CI/CD | Airflow | MLOps | Git | Docker | AirFlow | Talend | Apache Spark | Apache Kafka | Power BI

**Soft Skills:** Stakeholder Management | Project Management | Cost Optimization | Solution Architecture | Collaboration | Innovation

## Projects

### Real-Time Data Ingestion on AWS

Developed a comprehensive real-time data pipeline to ingest, process, and visualize data using AWS services. Automated data ingestion from a local generator to Kinesis, transformed and stored data in S3 as Parquet files, cataloged data with Glue Crawler, enabled querying via Athena, and created dashboards with QuickSight.

**Technologies Used:** AWS S3, Kinesis Data Streams, Glue ETL, Glue Crawler, Athena, QuickSight, Python, Boto3

### Data Integration and CRM Enrichment for Call Center Analytics

Engineered a multi-layered data integration pipeline to combine call records, outcomes, CRM snapshots, and run-time CDR data for enhanced customer analytics. Implemented complex hierarchical joins, created CRM Slowly Changing Dimensions (SCDs) for tracking customer products, and designed real-time KPI monitoring to optimize AI model performance.

**Technologies Used:** Talend, PySpark, Python, Cron, MySQL, PostgreSQL, CRM Systems, Data Warehousing