Rubab Zahra Sarfraz

rubabzsarfraz@gmail.com · rubabzsarfraz.com · linkedin.com/in/rubabzsarfraz

Summary

I'm a data strategist with 6+ years of deep expertise in data infrastructure development and software engineering for enabling AI/ML products and have championed data-driven transformations in diverse sectors. I have a knack for leading teams and optimizing data strategies and am adept at harnessing data to fuel innovation and growth in both startups and established organizations.

Education

Lahore University of Management Sciences (LUMS) Aug. 2016 – Jun. 2018

M.S. in Computer Science *Lahore, Pakistan*

- Thesis: Measuring the Impact of Fake News in Developing Regions
- Advisor: Dr. Ihsan Ayyub Qazi

University of Engineering & Technology (UET) Oct. 2012 – Jun. 2016

B.Sc. in Computer Engineering *Lahore, Pakistan*

- Thesis: Monitoring Traffic on Virtual Routers of OpenStack
- Advisor: Dr. Irfan Ullah Chaudhary
- GPA: 3.74/4.00

Publications

Rubab Zahra Sarfraz, Samar Haider (2024). "Vizard: Improving Visual Data Literacy with Large Language Models". In *International Workshop on Big Data Visual Exploration and Analytics (BigVis) at VLDB*.

Rubab Zahra Sarfraz (2024). "Towards Semi-Supervised Data Quality Detection in Graphs". In *International Workshop on Quality in Databases (QDB) at VLDB*.

Nida Munawar, **Rubab Zahra Sarfraz**, Maria Costello, David Robinson, Colm Bergin, Elaine Greene (2023). "Risk Factors and Outcomes of Delirium in Hospitalized Older Adults with COVID-19: A Systematic Review and Meta-Analysis". In *Aging and Health Research (Elsevier)*.

Work Experience

BridgeLinx Oct. 2022 – Present

Data Lead *Lahore, Pakistan*

- Designed and implemented a data governance framework across the company with **200+ data points** and **9 teams**, improving data quality from **45% to 98%**. *Python, GX, Prefect, Snowflake*
- Established data infrastructure from the ground up, utilizing Snowflake for data lakes, AWS Lambda for data pipelines, Streamlit for front-end adopted by **40%** of the company force.
- Engineered a real-time bidding system, employing advanced analytics to optimize Return on Capital (ROC) through predictive client settlement behavior analysis upon order booking.
- Developed a real-time financial reporting tool for working capital management, enabling efficient data-driven insights and alerting on time or monetary discrepancies, crucial in maintaining the cash conversion cycle **within 30 days**.

Finja Jan. 2020 – Oct. 2022

Data Lead *Lahore, Pakistan*

- Researched the SME market, leveraging alternate data to drive an AI/ML credit engine serving $3M issuances/month with 0.5% NPLs, optimizing ROI and reducing lending risks. *Machine Learning, AWS*
- Implemented real-time invoice validation system for fraud detection in a low-resourced undocumented SME market. *Tesseract OCR, Python Parser*
- Led a multidisciplinary team to launch Pakistan's first SECP-approved P2P lending, investment and payment engine. *200M+ lending, 3000+ retail stores, Python Flask, GCP*
- Collaborated with SECP to draft P2P lending regulations for Pakistan, enabling NBFCs to obtain P2P licenses.

- Directed a top-performing **10-member team** to transition Finja into a data-centric entity.
- Developed a CRM for loan recovery, enhancing collection efficiency by **30%**. *Python Flask, React, Microservices Architecture*
- Managed regulatory reporting with SBP and SECP, integrated data analysis into product lifecycle, and trained analysts to deliver continuous insights. *Tableau, Google Data Studio*
- Instrumental in securing a **$9M Series A** funding by showcasing data-driven tech and insights to investors.

Finja Oct. 2018 – Dec. 2019
Data Engineer *Lahore, Pakistan*

- Architected the company's data infrastructure, now driving **50+ dashboards** across 8 departments and 3 products. *Google BigQuery, Python*
- Collaborated cross-departmentally to develop ELT data pipelines, enabling data-driven product decisions. Products: Lending, Payments, Cards, Investment, Payroll. *Python, Google Dataflow, Holistics.io*
- Introduced a dual-tier data quality system to auto-flag discrepancies for stakeholders. *Airflow, SQL*
- Developed a rule-driven transaction monitor for compliance alerts. *Python, Stream Processing*
- Designed a sanction screening tool, matching customer KYC with online data **saving $350K**. *Python, REST API*
- Elevated the B2C app's search engine relevancy from **20%** to **80%**. *ELK Stack, Python Flask*

CERN Jul. 2018 – Sept. 2018
Intern, Software Engineer *Geneva, Switzerland*

- Spearheaded the deployment of Ceph clusters with Rook on Kubernetes, leveraging both OpenStack VMs and Ironic hosts.
- Defined and implemented robust evaluation metrics, outperforming the legacy puppet-based deployment methods.
- Advocated for and validated the superior latency and user-friendliness of the new Ceph cluster deployment approach.
- Enhanced the orchestrator CLI with RGW support, enabling rapid S3 service provisioning in just **1-2 minutes**. *Python*
- Published a blog post with the findings in collaboration with CERN and Ceph (RedHat).

Outreachy (RedHat) Dec. 2017 – Mar. 2018
Intern, Software Engineer *Remote*

- Improved distributed cluster management by adding performance dashboards. Pull requests: [1] [2] [3]
- Converted decentralized architecture of Ceph Manager to a centralized one for improved coherence with the codebase.

Meta Feb. 2017 – Mar. 2017
Mentee, Software Engineer *Remote*

- Extended Meta's open source project Osquery by implementing a virtual table in C++ for listing Python packages installed on a server (Linux, Windows & OS X). It was actively used for **securing Meta's data center servers** against vulnerabilities introduced by PyPy in 2017.

Teaching Experience

Instructor, Introduction to Data Science — Summer 2023
Lahore School of Economics

Teaching Assistant, Advanced Operating Systems — Spring 2017
Lahore University of Management Sciences
Instructor: Dr. Muhammad Hamad Alizai

Teaching Assistant, Programming Fundamentals — Spring 2016
University of Engineering & Technology
Instructor: Dr. Irfan Ullah Chaudhary

Awards & Honors

Invited Participant, International Visitor Leadership Program (IVLP), U.S. Department of State — 2024

Winner, USAID FDI Grant of $100K (on behalf of BridgeLinx) — 2024

Finalist, U.K. Climate Finance Accelerator (on behalf of BridgeLinx) — 2023

Runner Up, CERN Openlab Lightning Talks — 2018

Diversity Scholar, KubeCon by the Linux Foundation — 2018

Dean's Honor List, University of Engineering & Technology — 2012 – 2016

Community Involvement

Reviewer, IEEE VIS Workshop on Visualization for AI Explainability (VISxAI)	2024
Speaker, PyCon Pakistan 2024, "Elevating Trust in Your Data with Python"	2024
Panelist, Reshaping the Financial Sector with Artificial Intelligence, Information Technology University	2022
Panelist, Debugging the Startup Space: Tech Careers, LUMS Women in Computing	2022
Panelist, AWS Startup CTO Forum	2022
Panelist, Breaking the Bias, Systems Limited	2021
Panelist, Facebook (Meta) Developers Circle #TechByHer Meeting	2020
Contributor, Ceph Blog: Evaluating Ceph Deployments with Rook	2018
Contributor, Docker Docs Hackathon	2017

Technical Skills

Expertise: Data Engineering, Data Architecture, Data Modeling, Data Product Management

Languages & Scripts: Python, C++, SQL, HTML, LaTeX

Machine Learning: Scikit-learn, PyTorch, Transformers

Data Lake & Quality: BigQuery, Snowflake, S3, Great Expectations, Pandera

Data Visualization: Tableau, Looker Studio, Streamlit

Data Infrastructure: GCP (Pub/Sub, Cloud Functions, Storage, Dataflow, SQL), AWS (SageMaker, Lambda), Prefect, FastAPI, Flask