Mirza Elaaf Shuja

elaaf.shuja@gmail.com | linkedin.com/in/elaaf | github.com/elaaf

Technical Skills

Languages:	Python, C/C++, SQL, Shell Scripting, JavaScript
Platforms:	Linux (Ubuntu, Alpine, CentOS), Microsoft Azure, Google Cloud Platform
Tools:	Git, Docker, Hadoop, Spark, Apache Kafka, Apache Airflow, Sqoop, HP Vertica, Informatica
Philosophies:	Agile Development (Scrum), Test Driven Development

Work Experience

Data Engineer @ KEYRUS - keyrus.be *Aug 2021 – Present* (Lahore Pakistan)

- Working on on-prem Enterprise Data Warehouse migration to Azure Cloud and internal Data Integration tool based on Apache Airflow and Spark.

Technology Stack: Apache Airflow, Apache Spark, Apache Kafka, Docker, Azure Storage, Azure Data Factory, Azure Synapse, Databricks

Big Data & Machine Learning Engineer @ ADDO AI - addo.ai *Apr 2019 – Aug 2021 (*Lahore Pakistan)

***Project**: Azure Big Data and Machine Learning Platform for Mercy HealthCare*:

- Designed Informatica Data Integration jobs for data loading from secure-agent VM to Azure Data Lake Gen2 and data movement between the Data Lake layers (Raw/Curated/Semantic) to feed different use-cases.
- Designed Informatica Mass Ingestion jobs for ingesting data from external on-prem DBs to Azure Data Lake Gen2.
- Performed EDA of historic data sourced from Oracle DBs, for different ML use-cases to gauge their feasibility.
- ML use case implementation on Azure ML Studi/Azure DataBricks: Forecasting of star-rating, CHF (Congestive Heart Failure) prediction, and Market Growth Analysis to name a few.

Technology Stack: Informatica (Data Integration, Mass Ingestion), Azure Data Lake Gen2, Azure Machine Learning Studio, Azure Synapse, Azure DataBricks.

***Project**: Data Lake/Warehouse Migration from TeraData for Telenor Pakistan:*

- Responsible for the development, unit/system integration/user-acceptance testing of ETL pipelines (Spark jobs) of over 35 distinct business streams and over 12 dimensions of varying load and frequency on Data Lake (HIVE).
- Designed and Implemented the strategy for the PII data masking and data movement of different business streams between the Data Lake layers (Raw/Curated/Serving).
- Optimization of spark jobs and figuring out their most appropriate scheduling/triggers using shell scripts based on business requirements and fact dependencies.
- Analysis of existing Teradata SQL with the Data Modelling team and their conversion to PySpark jobs and Spark SQL.

Technology Stack: Apache Spark, Hive, HDFS, Vertica , Talend, Sqoop, Shell Scripts.

Machine Learning Researcher @ Information Technology University *Sep 2018 - Feb 2019 (*Lahore Pakistan)

- Unsupervised Image to Image Translation using Attention guided Cycle GANs
- Textual MisInformation Analysis in a semi-closed Social Network (WhatsApp)

Education

Masters in Data Science @ Information Technology University (ITU) *2018 – 2021 (*Lahore Pakistan)

- **Courses**: Machine Learning, Deep Learning, Big Data Analytics, Computer Vision, Information Retrieval Systems.
- **Honors:** Awarded Graduate Fellowship based on Academic performance.

Bachelors in Electrical Engineering @ National University of Sciences and Technology 2013 – 2017 (Islamabad Pakistan)

- **Courses**: Object-Oriented Programming, Data Structures, and Algorithms, Embedded System Design.

Publications

A First Look at COVID-19 Messages on WhatsApp in Pakistan (ASONAM 2020) - arxiv

A novel study of prevalent textual and image-based misinformation in semi-closed social networks (WhatsApp) in Pakistan.

Accepted as a Full Paper @ International Conference on Social Networks Analysis and Mining (ASONAM 2020) Dec 2020