**AKRAM SHAIK**

+1 (302) 514-3362 | akramshaik2326@gmail.com | LinkedIn

## PROFESSIONAL SUMMARY

- Experienced Data Engineer with 3.8 years of expertise in delivering impactful data-driven projects and innovative solutions.
- Proficient in designing scalable data pipeline architectures and optimizing ETL processes using Python, Spark, PySpark, Databricks, and Azure resources.
- Adept at data validation and system automation, with a strong ability to manage diverse forms of data from multiple sources.
- Proven track record in leading teams, collaborating with cross-functional stakeholders, and presenting complex data insights.
- Strong focus on developing frameworks for data lineage, automating validation processes, and achieving significant cost savings through workflow optimization.

## SKILL SET

- **Primary Skills:** Python, PySpark, Spark, Hadoop, SQL, T-SQL, C, C++, JAVA, HTML, CSS
- **Tools & Platforms:** Databricks, Azure Resources (ADF, Blob and Data Lake Storage, Azure DevOps, Azure Repos), Elastic, Git, Hive, GCP, BigQuery, AWS, Snowflake
- **Secondary Skills:** Visual Studio, Visual Studio Code
- **Database Used:** SQL Server, PostgreSQL, MySQL, Oracle
- **Certifications:** ***Azure Data Engineer Associate***

## EXPERIENCE

*Client: Greenbrier [Portland, Oregon]*

*Role: Data Engineer* Nov 2021 - Present

- Delivered a world-class data platform from the ground up, enhancing the data management solutions for **top U.S. railroads** in the railcar industry, with the potential to save billions of dollars by optimizing operations and decision-making.
- Planned and delegated complex projects with broad scope by managing diverse forms of data from over **90** data sources.
- **Mentored** and grew early career developers, providing guidance on onboarding, knowledge transfer, and best practices in data engineering.
- Facilitated technical discussions, guiding the team toward the most effective approaches to problem-solving and ensuring alignment with business needs and technical requirements.
- **Engaged with stakeholders across** the business to ensure their needs were met, **presenting bi-weekly project progress** updates to business stakeholders and cross-functional teams.

- Built, upgraded, and maintained data-related infrastructure and monitoring across multiple clouds and other systems & services.
- Wrote performant, readable, and reusable code and Infrastructure-as-Code.
- Reviewed code for the team to ensure a high standard of technical quality across our codebases.
- Automated validation of multi-format datasets across different stages, maintaining results with lineage tables for enhanced traceability.
- Delivered end-to-end data pipeline solutions, managing ETL/ELT workflows with Databricks and Azure resources, including Blob Storage, Data Lake, and Azure DevOps.
- Developed a framework for data lineage, spanning **10+** phases of complex transformations, enabling traceability across the data lifecycle.
- Delivered dashboards in Databricks to track project KPIs and system health.
- Implemented asynchronous logging using Elastic Logger for enhanced observability and system diagnostics.

**Key Achievements:**

- Successfully processed over **40M** daily records with complex transformations in under **7 minutes**.
- Designed a scalable architecture for asynchronous data validation across multi-phased pipelines.
- Achieved a **30%** cost reduction by identifying redundant resource utilization.
- Enabled **onboarding of new clients** by demonstrating MVP results to business stakeholders, earning accolades.

*Client: Walmart (Arkansas)*

*Role: Data Engineer* June 2021 – Nov 2021

- Led a comprehensive **data migration** project, transferring data from various Azure resources to Google Cloud Platform (GCP). Migrating and restructuring data transformation notebooks from Databricks to local development environments using Visual Studio, ensuring continuity and efficiency in workflows.
- Utilized a proprietary **data flow tool**, **designed to replace Azure Data Factory** (ADF) services, built on top of **Apache Airflow** for enhanced orchestration and automation capabilities.
- Conducted Proof of Concepts **(POCs)** to evaluate the feasibility of the new data flow tool, created extensive documentation, and **identified feature gaps**.
- Collaborated with the team responsible for developing the new service, providing feedback on limitations and requesting **essential feature enhancements**.
- Devised and implemented workaround solutions for missing features, thoroughly tested the migration process to ensure data integrity and reliability.
- Guided and supported team members from international locations, coordinating cross-border efforts for testing and validation.

- Engaged with multiple teams and business leaders to regularly update them on project progress, addressing concerns, and aligning objectives.
- Leveraged Google services such as **GCP** and **BigQuery** extensively, as well as other relevant technologies.
- Planning, designing, and developing code in **Python, PySpark, and SQL** and deploying SQL code based on mapping and business logic across multiple environments.
- Designing and developing ADF pipelines for data extraction from relational sources (Teradata, Hadoop, SQL Server) and non-relational sources (Flat files, SharePoint).
- Regularly updating and refining code to reflect business requirements, involving daily communication with stakeholders.
- Conducting **Unit Testing, System Integration Testing (SIT), and Integration Testing** for all developed code using Databricks Notebooks and BigQuery.

## EDUCATION

*Master of Science in Computer Science*
Specialization: Data Analytics, Northern Illinois University, USA — May 2021

*Bachelor of Technology in Computer Science and Engineering*
Jawaharlal Nehru Technology University, INDIA — June 2017

## ACADEMIC PROJECTS

*Perceiving Health Posts on Instagram:*
Analyzed Instagram datasets to extract topics on COVID-19, health, and lifestyle using clustering techniques on captions and hashtags. Implemented filtering by location and generated meaningful insights through image analysis. *Technologies Used:* Python, Pandas, Numpy, Scikit-learn, COCO Algorithm, TensorFlow, PyTorch, NLTK, Gensim, GeoPy, Matplotlib, Seaborn, Plotly

*Analyzing the Impact of COVID-19 and its Vaccine on Social Media:*
Processed and analyzed 10M tweets to extract sentiment and location data using Python libraries like geopy. Applied topic modeling to identify user concerns and perspectives regarding vaccines. *Technologies Used:* Python, Pandas, Numpy, Scikit-learn, TensorFlow, PyTorch, NLTK, GeoPy, Matplotlib, Plotly

*Cloud Workflow Scheduling with Deadlines and Time Slot Availability:*
Developed a hybrid cloud solution combining public and private resources, optimizing workflow scheduling based on deadlines and time slots. *Technologies Used:* Java, Quartz Scheduler, Java Concurrency API, Spring Boot, MySQL, PostgreSQL, MongoDB

## ADDITIONAL EXPERIENCE

*Graduate Teaching Assistant* (C++, Python), NIU — Jan 2020 – May 2021
*Peer Educator*, Recreation and Wellness Center, NIU — July 2019 – Dec 2019