Syeda Marium Faheem

Contact

Email: mariums82@gmail.com
Github: mariumfaheem
Medium: MariumFaheem
LinkedIn : syeda-marium-faheem
Phone No : +971501565042

Skill Highlighted

- Management consulting
- C-Level advisory
- Strong decision maker
- Complex problem solver
- Product development
- Project Management

Cloud Skills

- AWS
- Azure
- GCP

DevOps Skills

- Terraform
- Helm
- Kubernetes, Docker
- CICD Pipelines i.e. Jenkins, Bamboo and CodeCommit

Programming Languages

- Python
- Java

Summary

I am an experienced Data Engineer with a strong focus on building efficient and scalable data platforms for leading IT companies. With over five years of experience in management consulting and e-commerce startups, I have developed big data solutions, MLOps frameworks, and Generative AI applications that help both technical teams and executives make better decisions. I specialize in optimizing cloud-based data pipelines and solving complex challenges in big data processing, cloud engineering, and real-time analytics. My ability to simplify technical concepts into clear, actionable insights makes me an effective bridge between engineering and business teams.

Work Experience

Mckinsey & Company

Senior Data Engineer **December 2024 - present**

- Developed data ingestion pipelines integrating diverse sources into Azure storage.
- Led backend engineering, including Kubernetes/Docker orchestration, API development, and scalable infrastructure deployment.
- Optimized AWS infrastructure using Terraform, cutting provisioning time by 60%.
- Built high-performance data pipelines on Google Cloud using Dataproc Spark for batch and real-time analytics.
- Deployed data-fetching APIs from vector databases into EKS using Helm, AWS ELB, and Ingress.
- Designed distributed data pipelines on Databricks with Spark SQL and PySpark for large-scale ETL workflows.
- Built an end-to-end data warehouse using Informatica ETL.
- Developed GenAI data pipelines using LangChain and RAG for biggest oil company in Abu dhabi
- Engaged with C-level stakeholders to align AI strategies with business goals.
- Managed AI and big data projects for McKinsey and client vendors, overseeing implementation.
- Provided actionable insights from machine learning models to executive management.

Data Engineer **May 2023 – December 2024**

- Designed, developed, and maintained scalable data pipelines for large datasets on on-premise and cloud platforms (AWS, GCP).
- Led AWS big data platform and data operations, architecting data warehouse and data lake infrastructure for efficiency.
- Designed and maintained data models and data products for optimized storage and retrieval in data lakes and warehouses.
- Implemented data warehouses using Kimball and Data Vault methodologies, with optimization strategies like partitioning and indexing.
- Developed complex ETL solutions for large datasets in the Middle East's largest public sector organization.
- Created and managed AWS EKS clusters, deploying big data tools like Apache Airflow, Airbyte, and Jenkins.
- Optimized SQL and Python performance across multiple clients' data warehouses.

Big Data Skills

- BigQuery
- Redshift
- Snowflake
- Elastic Map Reduce – Hadoop
- Apache Hudi
- Apache Arvo
- Apache Kafka
- Apache Trino
- Apache Airbyte
- Apache Airflow
- Apache Superset
- Apache Spark (PySpark)
- MySQL
- Postgres
- Hive
- Hbase
- Cassandra
- MongoDB
- Spark Streaming
- Informatica
- Dataiku
- Kedro
- Dasgter
- FiveTran
- DBT
- Airbyte,
- Power BI
- Google Looker Studio

Machine Learning and MLOps

1. Hugging Face, Llama, OpenAI
2. RAG ,Vector DBs, Langchain
3. Kubeflow , MLflow
4. Vertex AI, SageMaker ,Azure ML

- Built an end-to-end GenAI pipeline for data extraction, conversion (.docx, .ppt, .txt, .pdf), and content generation using LLMs like GPT and Hugging Face.
- Developed GenAI data solutions for multiple clients, integrating structured and unstructured data (text, images, audio) from various sources.
- Built real-time and batch processing data pipelines from databases, APIs, flat files, and streaming data.
- Built a real-time fraud detection pipeline using Kafka (or AWS Kinesis), AWS Lambda, DynamoDB, and SageMaker.
- Orchestrated 500+ data pipelines using Airflow and dbt, diagnosing and resolving performance issues.
- Maintained and troubleshot Airflow environments, including worker nodes, schedulers, and databases.
- Monitored system health and performance metrics, addressing execution and dependency failures in Airflow DAGs.
- Built and maintained self-service datasets to enhance data democratization efforts

Bazaar Technologies

Senior Data Engineer I

January 2023 – April 2023

- As the lead Data Engineer, I designed, built, and scaled a high-performance data platform, ensuring efficient data processing, storage, and retrieval for critical business operations.
- Developed and deployed end-to-end data solutions, enabling seamless data integration, transformation, and analytics across multiple business domains.
- Implemented inference services for a Machine Learning platform, optimizing real-time model predictions and enabling scalable AI-driven inference.
- Established and managed a Model Registry for a Machine Learning platform, ensuring version control, model reproducibility, and streamlined deployment of AI models.
- Architected and managed a complete Machine Learning platform, including MLOps pipelines, automation frameworks, and monitoring systems, enhancing model lifecycle management and scalability.

Data Engineer II

January 2022 – January 2023

- Designed, deployed, and managed EMR on EKS clusters with Apache Hudi, optimizing scalable and efficient data storage solutions.
- Developed Apache Airflow DAGs using Python, scheduling and orchestrating data workflows while submitting jobs to Apache Spark via Apache Livy for data ingestion and transformations.
- Built and maintained highly available, distributed systems for large-scale data extraction, ingestion, and processing, ensuring system reliability and performance.
- machines, leveraging Docker containers for streamlined deployment and scalability.
- Maintained and optimized containerized environments, ensuring stable execution of Apache Airflow and Spark on AWS infrastructure.
- Implemented CI/CD pipelines using AWS CodePipeline, enabling continuous integration and automated deployment of data workflows and infrastructure updates.
- Collaborated with stakeholders to address data challenges and ensure compliance, security, and infrastructure needs across AWS regions.

Education

BE in Computer and Information System **Oct 2020**

NED University of
Engineering and Technology,
Karachi Department of
Computer Engineering

CGPA: 3.747

Professional Certificates

- Databricks Certified Associate Developer for Apache Spark
- AWS Certified Cloud Practitioner
- Data Engineering Foundations Specialization – ***Coursera***
- IBM Data Engineering Professional – ***Coursera***
- AWS Fundamentals Specialization – ***Coursera***
- Data Engineer Career Track –***DataCamp***
- IBM Data Science Professional– ***Coursera***
- Big Data Specialization – ***Coursera***

- Designed and maintained big data pipelines, integrating reporting tools, metadata modeling, and dashboards for business intelligence.

Data Engineer I **August 2020 – December 2021**

- Implemented Kafka for event streaming, enabling real-time data pipelines for high-throughput processing.
- Designed and developed end-to-end Data Lakes and Mesh architectures for scalable data availability and integration.
- Built and deployed AWS-based data warehouses, implementing ETL pipelines with AWS Glue, Redshift, and MySQL RDS for efficient data storage and retrieval.
- Designed and optimized Lakehouse architectures, improving data consistency and accessibility across cloud environments.
- Deployed and managed Kubernetes clusters, using Helm charts for scalable application deployments.
- Developed and maintained data transformation workflows using Python and PySpark, enhancing data preprocessing and feature engineering.
- Built Bazaar's Analytical Platform using Kubernetes, Hadoop, and open-source technologies, enabling large-scale data analytics and reporting.