Mehroos Ali

mehroosali@gmail.com | +1(214) 940-7050 | Richardson, TX, 75080
Github | LinkedIn | Portfolio
Work Authorization: F1 OPT

EDUCATION

University of Texas at Dallas, Richardson, TX *Aug 2021 - Dec 2023*
Masters of Science in Computer Science
Relevant Courses—Big Data Management and Analytics, Database Design, Machine Learning, Artificial Intelligence, Natural Language Processing, Design and Analysis of Computer Algorithms, Information Retrieval

Motital Nehru National Institute of Technology, Prayagraj, India *July 2014 - May 2018*
Bachelors of Technology in Mechanical Engineering

SKILLS

- Data Modeling (Relational, Dimensional, ER, EER), Data Pipeline Designing (ETL, ELT, Batch, Streaming, CDC), Data/Delta Lakes, Data Warehousing (Facts, Dimensions), Dashboarding, Cloud Computing
- **Languages** - SQL, Java, Python, Scala
- **Frameworks/Tools** - Hadoop, (Spark, PySpark), Spark Streaming, Hive, Pandas, Boto3, Kafka, Airflow, Tableau
- **CI/CD** - Docker, Maven, Git, GitLab
- **Database** - SQL (MS SQL Server), NoSQL (MongoDB)
- **Cloud** - GCP (GCS, BigQuery, Dataproc), AWS (S3, Lambda, SNS, SQS, EMR, Kinesis, Glue, Athena, Cloud Watch), Azure (Synapse Analytics, ADF, ADLS Gen 2), Databricks

CERTIFICATIONS

- Microsoft Certified: Azure Data Engineer [link]
- Databricks Certified: Data Engineer Associate [link]
- Astronomer Certified: Airflow Fundamentals [link]

EXPERIENCE

Data Engineer Intern - Trinity Industries (Dallas, Texas) *May 2023 - Present*
- Working on a building change data capture solution for migrating data from MS SQL server to AWS using Databricks, Spark Streaming, Debezium, and Kafka.
- Designed and implemented streaming data pipelines using Delta Lake to perform upsert (merge) operations on incoming data streams, ensuring real-time data updates and maintaining data consistency.
- Leveraged Azure Data Factory to seamlessly integrate data across various Azure services, such as Azure Databricks, Azure SQL Database, Azure Blob Storage, and Azure Synapse Analytics.
- Connected Tableau to AWS Athena and created reports and dashboards for railroad mileage comparison.

Data Engineer Intern - Amazon (Boulder, Colorado) *May 2022 - Aug 2022*
- Created data producers and consumers for Kinesis Streams, ensuring uninterrupted end-to-end data flow.
- Authored ETL scripts using PySpark and AWS Glue APIs, implementing complex transformations and business logic to cleanse and enrich data during the ETL process.
- Created Jupyter notebooks on AWS EMR clusters to enable exploratory data analysis of large-scale datasets.

Data Engineer - Onward Technologies Limited (Chennai, India) *Jan 2021 - Aug 2021*
- Migrated 250 spark jobs from on-premise Hadoop to Google Cloud Platform, reducing the processing time and increasing the computational limit by more than 60%.
- Designed and implemented a scalable data pipeline to process structured and semi-structured data by integrating 550 million raw records from different data sources using Kafka and PySpark and storing processed data in MongoDB.
- Authored Airfow DAGs for daily data ingestion and processing from Google Cloud Storage to BigQuery.
- Written hive queries to parse the raw data and store the refined data in partitioned and bucketed tables.

Data Engineer - Cognizant (Chennai, India) *Nov 2018 - Jan 2021*
- Handled sqoop parallelism, incremental data load from Oracle to HDFS, and Hive tables for daily data growth.
- Designed Nifi workflows for data ingestion from various sources such as RDMS, REST API, Kafka topic, etc.
- Improved runtime of slow-running spark jobs by 60% by optimizing Spark SQL joins.
- Developed a notification-based system using SNS, SQS, lambda, and DynamoDB to automate its deployment to AWS via GitLab.
- Stored data from spark as wide tables in Elastic Search for real-time aggregation and visualization in Kibana.
- Involved in integrating back-end systems in NodeJS with the dashboards created using React.