Mehroos Ali

mehroosali@gmail.com | 2149407050 | Richardson, TX, 75080

Github | LinkedIn | Portfolio

Available: Aug 2023 for Full-time roles

EDUCATION

University of Texas at Dallas, Richardson, TX *Aug 2021 - May 2023 (Expected)*

Masters 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

Motital Nehru National Institute of Technology, Prayagraj, India *July 2014 - May 2018*

Bachelors in Mechanical Engineering

SKILLS

- **Languages -** SQL, Java, Python
- **Frameworks/Tools -** Hadoop, (Spark, PySpark), Hive, Sqoop, Kafka, Nifi, Oozie, Airflow
- **CI/CD -** Docker, Maven, GIT, Jenkins, GitLab
- **Database -** SQL (Oracle, MySQL), NoSQL (HBase)
- **Cloud -** Databricks, GCP (GCS, BigQuery, Dataproc), AWS (S3, Lambda, SNS/SQS, EMR, Cloudformation)
- **Platforms -** Linux, Windows

EXPERIENCE

Data Engineer - Onward Technologies Limited (Chennai, India) *Jan 2021 - Aug 2021*

- Migrated 250 spark jobs from on-premise HDP to Google Cloud Platform which reduced the processing time and increased the computational limit by more than 60%.
- Designed and implemented a real-time scalable data pipeline to process structured and semi-structured data by integrating 550 million raw records from different data sources using Nifi and PySpark and store processed data in BigQuery.
- Authored Airfow DAGs for daily data ingestion and processing from google cloud storage to BigQuery.

Data Engineer - Cognizant Technology Solutions India Pvt Ltd (Chennai, India) *Nov 2018 - Jan 2021*

- Handled sqoop parallelism, incremental data load from Oracle to HDFS, Hive tables for daily data growth.
- Converted some existing sqoop, hive jobs to SparkSQL applications to read data from Oracle using JDBC and write it to hive tables.
- Written hive queries to parse the raw data and store the refined data in partitioned and bucketed tables.
- Designed Nifi workflows for data ingestion from various sources such as RDMS, REST API, Kafka topic, etc.
- Developed shell scripts for dynamic partitions adding to hive stage table, verifying JSON schema change of source files, and verifying duplicate files in the source location.
- Worked with different file formats like JSON, AVRO and parquet along with different compression techniques.
- Developed Lambda function to process SNS/SQS notifications and automated its deployment to AWS via Jenkins and Cloudformation.
- Improved runtime of slow-running spark jobs by 60% by optimizing Spark SQL joins.

PROJECTS

Databricks Formula 1 Racing Analysis [github] *Dec 2021 - Jan 2022*

- Created databricks notebooks to ingest, transform, analyze and create reports on Formula 1 racing data.
- Written Spark SQL queries to find the dominant drivers and teams for visualization.
- Scheduled the pipeline using Azure Data Factory (ADF) for monitoring and alerts.

AWS Batch ETL Pipeline [github] *Jun 2021 - Jul 2021*

- Built functional python script to load songs and logs data from S3 bucket.
- Transformed them to create and store as fact and dimensions tables in redshift.
- Orchestrated the data pipeline using Airflow DAGs and enforced data quality checks.

Twitter Streaming Analysis [github] *Sep 2020 - Dec 2020*

- Designed and implemented a real time streaming and classification system for sentiment analysis on Twitter data.
- Pulled live tweets using Nifi (Twitter API) into Kafka topic for cleaning, parsing and filtering using Spark.
- Applied stanford NLP to get sentiment score for each tweet and visualized using ES-Kibana.

CERTIFICATIONS

- Microsoft Certified: Azure, Data and AI Fundamentals (AZ-900, DP-900, AI-900) [link] [link] [link]
- Astronomer Certified: Airflow Fundamentals [link]