# MD SAIF UR RAHMAN

## Senior Data Engineer

Github LinkedIn +91 81520 81790 saifurrahman6236@gmail.com Hyderabad, IN

## SUMMARY

More than 6.5 years experienced Software Engineer, proficient in data engineering, building data solution platforms to enable business in utilizing data efficiently in identifying patterns and extracting valuable insights for making data driven decisions

## TECHNICAL SKILLS

**Tools/Frameworks**: **Python, SQL, Snowflake, Databricks, Delta Lake, AWS, Azure, SqlDBM, Git, Bitbucket, Streamsets**

**Data Engineering: Data Pipelines, Big Data, Spark, Data Modelling, ETL, ELT, Data Mesh Architecture, Airflow**

## CERTIFICATIONS

**Snowflake: SNOWPRO CORE**

**Databricks: Data Engineer Associate**

**DP-900: Microsoft Azure Data Fundamentals**

## PROFESSIONAL EXPERIENCE

### Data Engineer (IBM)

Jan '23 - Present

### AT&T (Client)

**DATA PRODUCTS (Data Mesh)**

**Responsibilities**

- Modeled and implemented data **products** for various **departments** including Field Services, Accounts, Workforce, Content Analytics
- Enhanced code functionalities in the Spark **Framework** to extract data from external APIs
- Utilized Databricks PySpark Framework to **ingest** data from sources such as **Teradata Vantage, AWS S3, Salesforce Lightening, APIs, Snowflake shares** into Snowflake stage schema
- Authored **SQL** transformation pipelines on stage tables to create **Dimension and Fact** tables in TGT schema based on Physical Data Model
- Designed **SQL Optimizations** to run the summary data pipeline and significantly reduce the time and resources by 70 percent
- Integrated Data Quality process in pipelines using open source **Great Expectations** package
- Documented Source to Target Mapping (**STTM**) and **Runbooks** for production support and deployment
- Created **Data Models** using SqlDBM and reviewed by **architects and business stakeholders**
- Developed Power BI **Audit dashboards** for each data product for monitoring and alerts
- Performed **jobs orchestration** and workflows using an in-house web application tool

**Key Achievements**

- Implemented **Data Mesh architecture** to create scalable, flexible, and maintainable data products
- Ensured **consistent view** of commonly used dimensions such as Account and Products
- Provided a **strong data foundation for Analytics Pods** and other business **consumers** by developing Data Products
- Collaborated with business consumers to understand their data needs for analytical and **reporting requirements**

- Implemented **domain-driven** approach to ensure inter-connectivity among source hubs

### Data Engineer (Accenture) Dec '21 - Jan '23

### Nationwide (Client)

**Property & Casualty Data Solution(Databricks)**

**Responsibilities**

- Ingestion **of** data from **S3 sources** into **Delta Lake** raw tables using Databricks **Autoloader**
- Engineered **Databricks generic autoloader** to process **6** different types of files such as csv, json etc., into delta database
- Devised **Streamsets pipelines** to ingest data(incrementally) on daily basis from 8 different sources such as salesforce, oracle, Microsoft sql etc., into Delta lake raw tables
- Instituted Databricks Notebook to ingest large database(Oracle, Mssql) tables using **ThreadPoolExecutor asynchronously**
- **Harmonize and Curate** the raw tables as per business requirements and sync into **Snowflake Data warehouse**

**Key Achievements**

- Centralized Data Solution saved almost **2 million USD/year in infrastructure cost**
- Different business users are given access to import data into BI tools for analysis purposes and make **data driven decisions**
- Enabled ML Engineers to collect all the required data from single source and **build models**

### Software Engineer (TCS) Dec '17 - Nov '21

### Qualcomm (Client)

**Modem Software Functionality Prediction**

**Responsibilities**

- Compiled Jira and CR's data from 2 different databases(MySQL and SQL Server)
- Composed feature selection resulted in 2 features, issue summary and issue description and remaining fields dropped
- Built a **LSTM base model with Embedding** by considering **top 30** Software functionality out of a 730

**Key Achievements**

- Achieved a **40% accuracy** based on just issue summary
- Expedited Issue description feature is further used to improve model performance to a of target 90%

Finally the base model is expanded to all 730 different categories

### Software Engineer (TCS)

### Ericsson (Client)

**Spark ML Project (Ericsson)**

**Responsibilities**

- Capturing data from **Cassandra** database into Spark Dataframes by deploying Pyspark for performing in memory computations
- Wrote **Python scripts** to perform different operations based on client requirements
- Dump the Dataframes into **Cassandra** or store in cache or Persist format based on complexity of operations performed

**Key Achievements**

- As the data is in Petabytes, the in memory computations are **100 times faster** than database operations
- Computed data is used by different web applications for **business analysis** purpose

## EDUCATION

### Bachelor of Engineering (ME) Aug '13 - Jun '17

### B.M.S. College of Engineering Bengaluru, IN