NAVEEN REDDY TUMU

***Sr. Data Engineer***

Phone: +1 (513)-318-8915 | [naveenreddy121510@gmail.com |](mailto:naveenreddy121510@gmail.com%20|) [GitHub](https://github.com/Naveen1203)| [LinkedIn](https://www.linkedin.com/in/naveen-tumu-a4327b1b2/)

# Professional Summary

# Senior Data Engineer with over 5+ years of experience in designing, optimizing, and implementing scalable data pipelines using Python, SQL, and PySpark.

# Proven track record in managing Agile cross-functional teams, improving project delivery speed by 30%, and reducing defects.

# Expertise in developing ETL workflows, utilizing Apache Airflow for automation, and optimizing big data processing in Hadoop, handling structured and unstructured datasets.

# Skilled in implementing data warehousing solutions with Snowflake, enabling fast querying on multi-terabyte datasets and ensuring cost efficiency.

# Experienced in cloud platforms like AWS, managing scalable data storage solutions, and delivering real-time insights through interactive dashboards using Power BI.

# Demonstrated proficiency in optimizing data processing workflows, achieving up to 40% reductions in data processing time through partitioning and advanced optimization techniques.

# Developed machine learning models for fraud detection, achieving a 90% accuracy rate and deploying scalable solutions on AWS for high-volume transactional datasets.

# Proficient in BI tools and dimensional modeling, including Star Schema, to facilitate effective data visualization and reporting.

# Strong background in big data technologies, including Databricks, Snowflake, Hadoop, and Apache Spark, with a focus on performance tuning and workflow orchestration.

# Effective communicator and collaborative team player, skilled in mentoring peers, managing inquiries, and facilitating knowledge-sharing sessions to drive student and team success.

# Technical Skills

# Tech Stack: Informatica, Kafka, Spark, Databricks, Docker, Shell Scripting, Unix, Control-M, Apache Airflow, Tableau, Power BI, Amazon Web Services (EC2, S3, Glue, RDS, DMS, Athena), Azure, GCP, ETL, Data Modeling, Data Profiling, Data Cleaning, Snowflake, Databricks, Hadoop, Hive, SQOOP, Machine Learning, JIRA.

# Programming: MySQL, Oracle, Python, R, Scala, SQL Server, PostgreSQL, MongoDB, BTEQ scripting.

# Professional Experience

## Verizon December 2022 – Present

## *Senior Data Engineer* *Texas, USA*

* Directed Agile, cross-functional teams in designing & implementing data pipelines, increasing project delivery **speed by 30%** by reducing defects.
* Engineered and optimized ETL pipelines using **Python, SQL**, boosting data processing efficiency.
* Optimized big data processing workflows in **Hadoop**, ensuring efficient storage & processing of structured & unstructured data sets.
* Designed and implemented scalable data pipelines using **Apache Spark, PySpark**, processing over 10TB of data daily, reducing data processing time by 40%.
* Implemented data warehousing solutions in **Snowflake**, supporting fast querying on multi-terabyte datasets by setting up **Snowflake Virtual Warehouses** to support multiple workloads simultaneously, ensuring resource efficiency and cost control.
* Industrialized **ETL workflows** using **Apache Airflow**, improving job reliability and scheduling by automating over 200 tasks per week.
* Managed data warehouses using **Snowflake**, enhancing data accessibility and enabling seamless BI reporting; Skilled in using **Jira** for project tracking, ensuring tasks are completed on time and aligned with project goals.
* Collaborated with data analysts to ensure accurate data representation, utilizing **Power BI** for effective data visualization; Applied Dimensional Modeling and Star Schema principles to support analytical queries and facilitate BI processes.
* Designed and deployed scalable data storage and processing solutions on **AWS**, reduced infrastructure costs and created interactive, real-time dashboards using **Power BI**, delivering actionable insights to business stakeholders.
* Spearheaded the creation of a data warehouse using **Amazon Redshift**, resulting in a 70% improvement in data query speed for data analysts and business users.
* Managed data pipelines across multiple cloud platforms, including **AWS** (S3, RDS) and Snowflake, ensuring 99.9% data availability and scalability for diverse business needs.

## Anthem Inc August 2020 – July 2022

## *Data Engineer* *Hyderabad, India*

* Improved data processing **performance by 20%** through data partitioning, optimization techniques, and efficient use of Spark in Databricks, leveraging Python libraries like **Pandas and NumPy** to optimize transformations and prevent data spillage.
* Led the migration of data from **Teradata** to **Snowflake** by designing and developing new ETL pipelines, using Apache Airflow to orchestrate data extraction, transformation, and loading processes across various sources.
* Applied advanced optimization techniques, including **repartitioning**, **coalescing**, and broadcast joins in Spark, reducing data processing time **by 25%** and improving overall system performance, while utilizing SQL and **NoSQL** databases for optimized query execution.
* Tested and implemented new optimization changes in production, resulting in a 40% increase in load performance, and used **Apache Airflow** to automate and monitor data pipelines, ensuring efficient workflow execution.
* Conducted performance tuning and optimization of ETL processes, achieving a **25% reduction** in data processing time through efficient **data transformations** and leveraging data warehousing solutions such as Snowflake for better data storage and retrieval.
* Demonstrated strong **problem-solving skills** and effective **communication** by collaborating with cross-functional teams to ensure smooth execution and optimization of data pipelines in a dynamic environment.

## American Express May 2019 – July 2020

## *Data Scientist / Analyst Intern Hyderabad, India*

* Implemented a **fraud detection system** for credit card transactions using machine learning techniques, resulting in a significant reduction in fraudulent transaction rates.
* Performed data preprocessing tasks such as feature encoding and feature scaling using **pandas** and **scikit-learn** to optimize model performance.
* Trained our model on historical transactional data, using **supervised learning**, achieving **80% improvement** in fraud detection accuracy.
* **Optimized KNN model performance** by tuning number of neighbors (k) and distance metrics to achieve a balanced trade-off between recall and precision.
* Applied confusion matrix and classificationreport using **scikit-learn** to evaluate model performance, achieving an **accuracy of 90%** and an F1 score of **0.85** in detecting fraudulent transactions.
* Leveraged AWS for scalable model deployment and distributed data processing, ensuring high availability and performance for large-scale transaction datasets.
* Worked with big data technologies like **Apache Spark** and **Databricks** to process and analyze large transaction datasets efficiently.
* Used version control systems like **Git** for collaborative code development and managing model updates across environments.
* Utilized **Power BI** for presenting fraud trends and monitoring KPIs to stakeholders.

# Technical Projects

## Stock Market | Python, Kafka, Spark, AWS [[GitHub](https://github.com/Naveen1203/Stock_market_Realtime_data_processing)]

* Developed Data Pipeline which automates the extraction of data from front-end API’s and loads it into AWS S3 by applying transformations which **reduces the processing time by 60%.**

**Uber Data Pipeline** |Erwin, Python, SQL server, Spark **[**[**GitHub**](https://github.com/Naveen1203/Uber_data_pipeline/tree/main)**]**

* Architected and developed a highly scalable ETL (Extract, Transform, Load) data pipeline to ingest, transform, and load data from the Uber front-end systems into SQL Server to **generate real-time reports**.

## Information Retrieval System | Python, Scikit learn, ML, LLM

* Designed and implemented an information retrieval system by using classification and clustering models to help users to get the relevant information within **Top 5 retrieved documents** by calculating the precision, Recall, F1-scores.

# Certifications

* **Power BI** Virtual Experience Certified from PwC.
* **Python** Programming language Certified from Kaggle & Hacker Rank.
* **SQL** Certified from Hacker Rank.

# Education

**Wright State University, Dayton, OH.**

*Master of Science (M.S) in Computer Science*

# Malla Reddy College of Engineering, Hyderabad, India.

*Bachelor of Technology in Computer Science*