SAI PRANEETH KOMMU

2014486713 saipraneethusa7@gmail.com Linkedin GitHub

PROFESSIONAL SUMMARY

Results-driven **Data Engineer** and **SQL DBA** with 3+ years of experience in designing and building ETL pipelines, managing SQL Server databases, and optimizing query performance. Proficient in Python, PySpark, Apache Airflow, Azure Data Factory, and AWS Glue. Experienced with cloud platforms including AWS and Azure, and data warehousing technologies like Databricks, Snowflake, and Synapse. Skilled in database backup, recovery, security and high availability.

EXPERIENCE

Data Engineer | **Database Administrator** Aug 2020 – July 2023

Nichebit Softech Pvt Ltd *Hyderabad, India*

- Developed end-to-end ETL pipelines with Azure Data Factory and Airflow, reducing processing time by 70%.
- Optimized Azure Synapse and PostgreSQL queries, decreasing Tableau dashboard latency by 60%.
- Enhanced data reliability by implementing validation rules in Spark & Airflow, increasing pipeline rate to 99.8%.
- Improved data accuracy by 45% using Apache Spark and Python for cleansing and anomaly detection.
- Built a scalable SQL Server 2022 solution for 900+ clients, improving query speed by 30%.
- Engineered high-throughput data transformation workflows using Databricks & PySpark, reducing runtime by 50%.
- Integrated data pipelines across AWS S3, Glue, RDS, Lambda, and Redshift, enabling a hybrid-cloud data flow and cutting infrastructure cost by 25%.
- Designed and deployed solutions using Azure Data Factory, Data Lake, Synapse, and SQL Database, streamlining enterprise data management and compliance.
- Managed SQL Server 2022 installation, configuration, patching, backups, ensured database uptime & fast recovery.
- Improved database performance and availability by optimizing SQL Server maintenance tasks and implementing robust disaster recovery plans.

PROJECTS

Olympic Data Analytics - Azure Data Engineering Project | *Azure Databricks - PySpark - Azure Data Factory*

- **Process:** Built an Azure data pipeline, achieving full automation, measured by integration of 4 services, by using ADF for ingestion, Databricks for transformation, and ADLS Gen2 for storage.
- **Result:** Improved processing efficiency, achieving 75% reduction in manual effort, measured by pipeline execution time, by automating data workflows with ADF scheduling and PySpark scripts.

Twitter Data Pipeline using Airflow project | *Python - AWS EC2 - Apache Airflow - Pandas*

- **Process:** Built a Twitter data pipeline, measured by automated extraction and transformation of 5K+ tweets, by using Tweepy and Pandas to process data and orchestrating tasks with Airflow.
- **Result:** Deployed an Airflow-managed ETL system, measured by scheduled delivery of processed data to AWS S3, by configuring Airflow DAGs on EC2 to run extraction, transformation, and S3 upload workflows.

Azure NYC Taxi Data Engineering Project | *PySpark - Azure Databricks - Spark SQL - Microsoft Power BI*

- **Process:** Built an Azure-based ETL pipeline, measured by automated ingestion of 12+ months of NYC taxi data, using ADF, ADLS Gen2, and Databricks in a Medallion Architecture.
- **Result:** Delivered BI-ready Delta tables, measured by Gold-layer creation with time travel and versioning, by transforming data with PySpark and integrating with Power BI.

Netflix Azure Data Engineering Project | *Azure Data Factory - Data Lake - Synapse - Databricks - PySpark*

- **Process:** Built a scalable Azure data pipeline, measured by processing 1M+ records across Bronze, Silver, and Gold layers, by integrating ADF, Databricks Autoloader, and Delta Live Tables.
- **Result:** Optimized data transformation workflows, measured by 30% faster execution using parameterized notebooks, by applying PySpark logic with Databricks Workflows and job clusters.

IPL Data Pipeline with PySpark and Databricks | *Apache Spark - SQL - Amazon S3 - Pandas - Seaborn*

- **Process:** Built scalable Apache Spark pipelines on Databricks, processing 5 IPL datasets (600+ matches), using custom schemas, PySpark transformations, DataFrames, and S3 data ingestion.
- **Result:** Analyzed IPL 2017 data with Spark SQL and PySpark, delivering insights on top batsmen, bowlers, and match trends, visualized via Pandas, Matplotlib, and Seaborn.

Stock market real-time data analysis project using Kafka | *Apache Kafka - Amazon S3 - EC2 - AWS Glue Crawler*

- **Process:** Engineered a Kafka-based pipeline, measured by ingesting 10K+ stock events to S3, by simulating real-time data with Python and deploying producer-consumer on EC2.
- **Result:** Automated real-time querying, measured by Athena access to 1K+ S3 JSON files, by using AWS Glue Crawlers to catalog data for SQL analysis.

TECHNICAL SKILLS

Programming Languages: Python, SQL, PySpark, Shell Scripting, Git, Docker

Data Warehousing & ETL: Airflow, Azure Data Factory, AWS Glue, Snowflake, OLAP/OLTP, SSIS, SSAS, Azure Synapse,

Data Visualization & Big Data: Excel Dashboards, Tableau, Power BI, Apache Hadoop, Apache Spark, Apache Kafka

Cloud Platforms: AWS (S3, Glue, RDS, Lambda, Redshift, EC2), Azure (Data Factory, Data Lake, Synapse, SQL Database)

Machine Learning & AI: TensorFlow, Scikit-learn, Keras, PyTorch, Predictive Modeling, DL, NLP, LLMs

Database Administration (DBA): SQL Server, MySQL, PostgreSQL, MongoDB. Installation & Configuration, Partitioning, Performance Tuning, Backup & Recovery, High Availability, Replication, Database Management

EDUCATION

New Jersey Institute of Technology — Sep 2023 – Dec 2024
Masters in Computer Science — *Newark, NJ*

Vardhaman College of Engineering — Mar 2018 – Feb 2022
Bachelors in Computer Science — *Hyderabad, India*