**KOUSHIK REDDY SAMA** Mobile no: 316-519-0532

LinkedIn: https://www.linkedin.com/in/koushik-sama-0b4048289/ Email: **Koushik.sama9@gmail.com \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# EDUCATION

**Master of Science in Computer Science**  **Jan 2020 – May 2021**

University of Missouri, Kansas City

# PROFESSIONAL SUMMARY

Data Engineer with around 5+ years of expertise in data solutions, specializing in ETL optimization and infrastructure design. Proficient in Python and SQL for data analysis and manipulation. Skilled in managing data warehouses, including Snowflake, and adept at ensuring data integrity and security. Experienced in migration projects and proficient in cloud technologies, with a focus on AWS and Azure. Eager to leverage technical skills to drive data excellence in diverse projects.

# TECHNICAL SKILLS/CERTIFICATION:

* **Technologies**: Python, SQL, ETL, SQL/NoSQL DBs, CICD, GIT, Restful APIs, Control-M, Airflow, AWS Services, Hadoop Map Reduce, Spark, Pyspark, Databricks, Snowflake, Informatica, PL/SQL, Power BI, Tableau, business intelligence.
* **Version Control Tools:** GIT, SVN, TFS, Docker, Jenkins.
* **SCM:** Github, Bitbucket, GitLab, Azure Devops
* **Mics:** Data Structures, Algorithms, problem solving, Design Patterns and Data Modelling.
* **Databases:** MySQL, SQL Server, Oracle, DB2, PostgreSQL, Netezza, **Cloud**: AWS, Azure
* **Operating Systems:** Windows, Linux, Unix
* **AWS Certified Solutions Architect - Associate.**

**PROFESSIONAL EXPERIENCE**

**CAREFIRST BCBS**

**Data Engineer Sep 2023 –Present**

* Worked on Multi-phase data integration projects, delivering risk analysis data for the business LexisNexis Risk Solutions.
* Designed data models for SQL Server tables to implement SCD Type 2, for membership data integration into SQL Server.
* Developed and optimized **ETL** **workflows** on Databricks with **Apache** **Spark** and **IICS**, leveraging data partitioning and parallelism to accelerate the integration of **membership**, **policy**, and **claim** data into **Snowflake** and **SQL** **Server**.
* Enhanced processing efficiency by implementing **Spark** **SQL** and DataFrames for complex transformations and secure file transfers via **SFTP**, reducing data transformation times by over 30%.
* Monitored and fine-tuned **Spark** **jobs** on **Databricks** for peak performance, utilizing adaptive query execution and cluster configuration adjustments, ensuring reliable and timely data delivery.
* Developed complex SQL queries to source data from Snowflake, utilizing advanced techniques Partitioning, window functions, CTEs, subqueries, and joins to optimize data retrieval and ensure accuracy in ETL processes.
* Developed **Python** **RESTful** **APIs** with OAuth for secure data integration and created API documentation using Swagger.
* Engineered DevOps Ingestor tools for creating tables and loading data in Snowflake and for Azure storage integration.
* Managed containerized ETL applications with **Docker** and **Kubernetes**, ensuring consistency, scalability, and reliability across deployments.
* Completed **HIPAA** compliance training, ensuring adherence to healthcare privacy and security standards.

**Technologies Used**: Snowflake, Python, Azure Devops, Apache spark, Configuration tools, SQL, IICS, Microsoft SQL Server Management Studio, Datawarehouse solutions, Data Modelling, SCD, CDC, CI/CD, Azure Synapse analytics, Data Visualization, Docker, Kubernetes.

**PLANET FITNESS**

**AWS Python Developer Dec 2022 –Sep 2023**

* Worked on Data Modernization project, standardizing data transfers from **DynamoDB** to **AWS** **S3** and **Snowflake** Data Warehouse.
* Re-engineered **Snowflake** **stored** **procedures**, optimizing query performance and indexing.
* Built and standardized Data Lakes, focusing on data ingestion, cataloging, analytics, security, and governance.
* Automated data exports using **AWS** **Lambda** and Firehose **Kinesis** for **real**-**time** data transfer.
* Developed AWS **Glue** ETL jobs for data processing and orchestrated jobs with Step Functions.
* Utilized **Databricks** with **Apache** **Spark** for real-time data ingestion and processing, integrating seamlessly with Kinesis and S3 to enable high-speed data streaming and scalable processing.
* All the scripts are developed in **Python**, **PySpark**, and shell scripting, leveraging AWS EMR clusters for large dataset processing and connected Tableau servers to AWS Athena for querying.

**Technologies Used**: Hadoop, HDFS, SQL, Python, Pyspark, Apache Spark, Snowflake, Databricks, AWS Dynamo DB, Lambda, Step Functions, Firehouse Kinesis, Glue jobs, APIs, Athena, Tableau, S3, CI/CD.

**USAA**

**Data Engineer May 2021 – Dec 2022**

* Led end-to-end data migration from **Oracle** and **Netezza** to AWS **Snowflake**.
* Automated ETL processes using **Python**, **SQL**, and **Shell** **scripting**, developing reusable and robust migration scripts.
* Utilized Control-M for developing batch services/user interface for single and bulk asset registrations and Airflow for job scheduling, creating dags and ensuring seamless migration execution.
* Created complex queries and stored procedures in Oracle PL/SQL for efficient data extraction and transformation.
* Designed and developed **Python** **RESTful** **APIs** for data integration between Oracle and AWS Snowflake.
* **Built and maintained complex ELT models in DBT (Data Build Tool)**, creating modular, reusable transformations within Snowflake to streamline data migration and ensure data quality and consistency during the transfer from Oracle and Netezza.
* Employed **Databricks** and **Kafka** for real-time data ingestion, transformation, and streaming services.
* Utilized Snowpipe for real-time data ingestion into AWS Snowflake during the migration from Oracle and Netezza, enhancing data availability and processing efficiency.

**Technologies Used:** Hadoop, HDFS, Hive, SQL, Snowflake, Python, Pyspark, Databricks, Spark, Shell, AWS EC2, S3, IAM, Snowflake, Unix scripts, Kafka, Control-M Job Schedular, DBT ELT Tool, Oracle, PL/SQL, Netezza, Power BI, CSV Files, XML files, CI/CD.

**Techmatrics Solutions**

**Data Engineer Jan 2017 – Dec 2019**

* Built scalable data pipelines using Python (Pandas, SQL Alchemy) and SQL to integrate multiple data sources like PostgreSQL, AWS RDS, and APIs into a centralized data warehouse on AWS Redshift.
* Optimized complex SQL queries across AWS Redshift and PostgreSQL, reducing query execution times by up to 40% using indexing, partitioning, and query tuning strategies.
* Automated ETL workflows with Apache Airflow, orchestrating and scheduling daily data ingestion and transformation processes for a 50+ TB data warehouse, ensuring robust data flow management.
* Integrated data from diverse sources such as AWS S3, relational databases, and external APIs, utilizing AWS Glue to automate transformation and data loading processes.
* Enhanced data quality by implementing advanced data cleansing and normalization routines, achieving 99.9% data accuracy for downstream BI tools like Tableau and Looker.
* Led the migration of on-premises databases to AWS, leveraging S3 for data storage and AWS Redshift for warehousing, ensuring secure and seamless data transfer with zero downtime.
* Developed Python-based automation scripts for data validation, anomaly detection, and error reporting, increasing ETL process efficiency by 60% and reducing manual intervention.

**Technologies Used:** Data Integration, Analysis, Python, Pandas, Hadoop, Spark, Hive, ETL, SQL, SQL Alchemy, Dynamo DB, Mongo DB, Postgres SQL, Redshift, Stored Procedures, Data Structures, Airflow, CI/CD, GIT, AWS, S3, Airflow, Tableau, Looker .