# Sarthak Sahoo

571-303-9819 | sarthaksahoo49@gmail.com | https://www.linkedin.com/in/sarthaksahoo511/

**SUMMARY**

Highly skilled Data Engineer with 3+ years of experience in designing and implementing ETL pipelines using Python and Spark. Proficient in cloud services (AWS: S3, Glue, EMR, Lambda | Azure: Data Bricks, Data Factory) and Snowflake. Experienced in automating data architectures with CI/CD pipelines using Jenkins for enhanced performance and reliability.

**SKILLS**

| * Programming: Python, SQL, Pyspark * Cloud: AWS (S3, Lambda, EC2), Azure (Data Bricks, Data Factory) * CI/CD: Jenkins, AWS CodePipeline | * Data engineering: Spark, Glue * Big Data: Spark, EMR, Hive * Database: PostgreSQL, MySQL, Oracle * Data Warehouse: Redshift, Snowflake |
| --- | --- |

**PROFESSIONAL EXPERIENCE**

**CVS Health, Richmond VA (remote)** **Jan 2022 - Present**

**Cloud Data Engineer**

* Designed and implemented ETL pipelines with Glue scripts, Python, Pandas, and Spark(including RDD and DataFrame API) within EMR for data processing and analysis. Loaded data into Snowflake and AWS Redshift for efficient batch processing workflows with SQL
* Orchestrated CI/CD pipelines with Jenkins, Git, and AWS CloudWatch, reducing deployment time by 25%, improving error handling, and enhancing overall efficiency by 20%.
* Leveraged Docker and Kubernetes for containerization within ECS and EKS to ensure seamless deployment and scalability of ETL process
* Collaborated with cross-functional teams to align data strategies with business objectives, improving data-driven decision-making accuracy.

**Performance Food Group, Richmond VA Jan 2021 - Dec 2021**

**Data Engineer**

* Engineered and optimized ETL pipelines in Azure Databricks for Snowflake, processing sales, inventory, and consumer data with Python, PySpark, and SQL.
* Conducted statistical and machine learning analysis using Scikit-learn and Scipy to extract actionable insights.
* Utilized Matplotlib and Seaborn, and Power BI for data visualization, creating comprehensive reports to empower stakeholders with sales and inventory insights.

**EDUCATION**

M.S. Applied Computer Science (Data Science Focused) **Aug 2020-Aug 2022**

Frostburg State University, Frostburg MD

B.S. Chemical Engineering

Virginia Commonwealth University, Richmond VA **Aug 2016 - May 2019**

**PROJECTS**

**Twitter HashTag Project** **Aug 2021 - Aug 2022**

* Developed a real-time Twitter data collection pipeline using Python and Twitter API, integrated with Apache Kafka for ingestion and buffering.
* Designed and implemented a Python algorithm to extract, transform, and filter relevant information from streamed tweets, storing processed data in an AWS S3 bucket and providing real-time access via AWS Lambda and API Gateway.

**PORTFOLIO**

https://sahoos511.github.io/