KHUSHI TRIPATHI

New York | +1(716)3989776 | tripathikhushi214@gmail.com | LinkedIn | Github

Experienced Data Engineer with 2+ years of expertise in building scalable data pipelines, optimizing SQL queries, and architecting AWS-based solutions. Skilled in Python, PySpark, SQL, and Apache Spark for data processing, analysis, and visualization. Proficient in ETL development, data modeling, and data warehousing, leveraging tools like Apache Airflow, AWS Glue, and Databricks to architect high-performance data solutions. Expertise in cloud storage (AWS S3, Snowflake) and real-time big data processing using Kafka, Hive, and Hadoop ecosystem tools (HDFS, MapReduce, Yarn, Sqoop).
Adept at SQL performance tuning, developing Stored Procedures, Triggers, Functions, and Packages for seamless data integration.
A result-driven professional with a track record of improving query performance, automating workflows, and optimizing BI reporting. Passionate about leveraging machine learning (NLP, clustering, predictive analytics) to drive actionable insights. Strong collaborator with cross-functional teams.

Technical Skills

- **Programming & Visualization:** Python, R Programming, Scala, Tableau, Power BI, matplotlib, seaborn, ggplot2
- **Machine Learning & AI:** pandas, numpy, scikit-learn, TensorFlow, PyTorch, NLP (spaCy, NLTK), RNNs, XGBoost
- **Databases & Big Data Technologies:** Oracle SQL, MySQL, PL/SQL PostgreSQL, AWS Redshift, Snowflake, DynamoDB, S3, Hadoop Spark, Kafka, Hive
- **ETL & Data Engineering:** AWS Glue, AWS Lambda, Databricks, Apache Airflow, Data Lake Architecture, SQL Performance Tuning, Event-Driven Pipelines, Docker
- **Analytical Skills:** Data Modeling, Data Warehouse, Predictive Analysis, Anomaly Detection, PCA, Clustering Algorithms, A/B Testing, User Growth Analytics, Customer Segmentation, Marketing Attribution Models, KPI Tracking, Business Intelligence, Feasibility Analysis, Big Data Analytics, Data Analysis, Data Visualization
- **Soft Skills:** Strong Communication, Problem Solving, Decision-making, Leadership, Cross-functional Collaboration, Presentation

Experience

Data Engineer, Match4Action-CrowdDoing **Mar 2024 - Present**

- Designed and implemented scalable ETL pipelines (AWS Glue, Redshift, Airflow) for processing large-scale unstructured text data**,** reducing data ingestion latency by 30%**.**
- Built optimized data models to improve accessibility, enabling self-service analytics via AWS QuickSight for researchers analyzing medicinal herb efficacy.
- Developed NLP-driven classification models (spaCy, NLTK) to categorize herbal remedies based on symptoms, increasing classification accuracy by 20%**.**
- Automated SQL-based reporting workflows**,** reducing manual effort and streamlining real-time data accessibility for key stakeholders.
- Partnered with cross-functional teams**,** including data scientists, engineers, and business strategists**,** to refine research metrics and enhance analytical accuracy.
- Spearheaded process improvements, introducing data validation and governance strategies to ensure data integrity.
- Collaborated with cross-functional teams to refine business metrics and data models, ensuring alignment with research and analytical goals.

Data Engineer, BlueSinga Innovative Solutions Pvt. Ltd. **Jul 2021 – Apr 2022**

- Optimized data pipelines and machine learning solutions for marketing analytics and customer engagement tracking.
- Led the design and deployment of real-time data pipelines (AWS Glue, Redshift, Spark, Airflow) to support marketing analytics and campaign optimization, reducing data latency by 40%.
- Architected scalable data models that enhanced customer segmentation analytics, improving engagement strategies and campaign effectiveness.
- Designed A/B testing and cohort analysis frameworks to evaluate marketing performance, leading to a 22% increase in conversion rates.
- Developed automated data quality checks, root-cause analysis, and real-time alerting pipelines, identifying inconsistencies in customer behavior data and improving data accuracy.
- Fostered collaboration with product managers, software engineers, and marketing teams, aligning analytics insights with business goals.
- Led ETL process optimizations, automating workflows to reduce data processing costs by 30% and improving system efficiency.
- Partnered with business teams to define key KPIs and analytics requirements**,** ensuring reporting tools aligned with decision-making needs.

Data Analyst, C-mos Systems: **Jan 2021 – Jun 2021**

- Engineered automated SQL workflows using Apache Airflow**,** improving report generation speed by 20% and optimizing data integration with AWS Redshift**.**
- Developed predictive customer segmentation models (XGBoost, K-Means, DBSCAN)**,** enhancing marketing personalization and increasing customer retention by 18%**.**
- Built real-time recommendation systems using ML-based clustering and analytics, improving customer engagement by 12%**.**
- Designed interactive dashboards (Power BI, Tableau, AWS QuickSight) for cross-functional teams to access real-time performance insights**.**
- Integrated CI/CD automation for data pipelines**,** reducing deployment failures and ensuring seamless data workflow execution**.**
- Created BI dashboards (Power BI, Tableau) to provide cross-functional teams with automated performance insights.

Project

Scalable Marketing Analytics Data Pipeline

- Developed high-performance ETL pipelines (AWS Glue, Redshift, Spark) to process millions of marketing data points, reducing query latency by 35%. Built automated anomaly detection scripts (Python, SQL) to monitor data quality and trigger alerts for inconsistencies.

Credit Card Holder's Risk:

- Classified high-risk customers and recommended optimal spending limits using PCA (90% variance preserved) and clustering techniques (K-means, DBSCAN).
- Identified six distinct customer segments based on spending habits and purchase history.

Uber Data Analytics:

- Designed an end-to-end data pipeline with Python & Apache Spark for trip analysis.
- Built a scalable ETL system, optimized real-time processing, and queried datasets with SQL to enhance operational insights.

Education

Master of Science: Data Science, University at Buffalo **Feb 2024**
Bachelor of Science: Computer Science, University of Mumbai. **May 2020**

Certifications

Google Data Analytics Professional (Coursera)
SQL Masterclass: SQL for Data Analytics (Udemy)
AWS Fundamentals (Coursera)