Sarthak Hatwar

Arlington, TX |+1-682-375-6985 | sarthakhatwar1606@gmail.com | sarthakhatwar1606/LinkedIn

SUMMARY

Data Engineer with 2+ years of experience building scalable data pipelines and real-time analytics solutions using AWS, Azure, Spark, and Kafka. Skilled in anomaly detection, ETL frameworks like Airflow, and creating actionable visualizations with Tableau and Power BI to support data-driven decisions.

EDUCATION

University of Texas at Arlington **Aug 2023 -May 2025**

Master of Science, Data Science

Coursework: Machine Learning, Artificial Intelligence, Probability & Statistic, Data Mining, Project Management, Neural Network

MIT ADT University, India **Aug 2019 - May 2023**

Bachelor of Technology, Computer Science

Coursework: Information Retrieval, Big Data, Algorithms and Data Structure, Database Systems, Operating Systems

EXPERIENCE

Cisco | *Data Engineer Intern* **June 2024 - May 2025**

- Built scalable ETL pipelines using Spark and Glue to process 5TB+ telemetry data/day, optimizing system behavior and network performance by 40% at Cisco.
- Consolidated data from multiple platforms (e.g., Meraki, Webex) into a unified S3-based data lake, managing 3M+ records and reducing duplication by 30%.
- Developed PyDeequ-based anomaly detection models, using CloudWatch and Splunk, improving data reliability to 99% and reducing incident response by 25%.
- Created Tableau dashboards to track device performance, improving regional traffic and capacity planning, with 40% faster report generation.

Codon Technologies | *Data Analyst* **Aug 2022 - Jul 2023**

- Designed and deployed Kafka-Kinesis pipelines for real-time data processing, enabling early bottleneck detectionand improving system throughput predictability.
- Built Power BI dashboards and visualizations for key performance indicators, reducing manual reporting by 40%and empowering self-service analytics.
- Implemented ETL workflows using Python and Airflow to move clinical datasets from GCP to AWS Redshift, ensuring over 90% schema consistency.
- Created Delta Lake pipelines for 2TB+ data processing with 92% accuracy in churn predictions and Flask APIsdeployed on AWS Lambda and EC2.

PROJECTS

Real-time Sports Analytics Platform

- Migrated SQL Server databases to Azure using Azure Data Factory (ADF), achieving 100% data transfer success and real-time synchronization across 50+ databases.
- Enabled centralized data access in Azure Data Lake Gen2, improving data retrieval speed by 30% for downstream analytics.
- Optimized 5TB+ datasets with Azure Databricks and Synapse Analytics, boosting query performance by 40% and enabling near real-time processing.
- Built Tableau dashboards on event metrics like delays and failures, improving operational visibility and speeding up executive responses by 25%.

Customer Churn Prediction

- Built Delta Lake pipelines using Databricks DLT and dbt for modular SQL transformations, processing 2TB+ of data with 92% accuracy.
- Created and deployed Flask-based APIs on AWS Lambda and EC2, streamlining access to prediction models, improving data access speed, and reducing latency by 40%
- Integrated Airflow DAGs for automated scheduling, Dockerized the application, and achieved 50% reduction in model retraining cycle.
- Applied churn modeling to simulate carrier delays and forecast delivery risk, enabling proactive logistics mitigation planning.

Parkinson's Disease Detection using Handwriting

- Devised a handwriting classification pipeline leveraging the VGG16 deep learning model, achieving 78.5% accuracy on test data from patients.
- Executed Hive-based feature extraction processes and optimized data preprocessing steps, reducing pipeline latency by 30%.
- Delivered a diagnostic dashboard for clinicians, significantly improving patient evaluation timelines and clinical decision support.

SKILLS

- **Programming**: Python, SQL, Scala, Java, C++, R
- **Cloud Platforms**: AWS (EC2, Lambda, S3, Redshift, Glue), GCP (BigQuery, GCS), Azure (AZ-900 Certified)
- **Big Data & Streaming**: Apache Spark, Kafka, Hive, Hadoop, Delta Lake, DLT
- **ETL & Orchestration**: Airflow, dbt, Medallion Architecture, Flask APIs
- **Machine Learning:** Scikit-learn, Anomaly Detection, PyDeequ, TensorFlow, Scikit-learn
- **Visualization**: Tableau, Power BI, Looker, Matplot,
- **Databases**: PostgreSQL, MySQL, Redshift, Cosmos DB

CERTIFICATIONS

- AWS Certified Data Engineer – Associate
- Microsoft Azure AZ-900