

# Siddharth Patondikar

Rochester, NY (open to relocate) | +1 (585) 202-8026 | [sidpatondikar@gmail.com](mailto:sidpatondikar@gmail.com) | [Portfolio](#) | [Linkedin](#)

## SUMMARY

RIT Data Science MS student (**Dec. 2026**) with **3+ years** building ETL pipelines and ML solutions that solve real business problems. Seeking **full-time** Data Science, Data Engineering, or Analytics roles starting **Jan. 2027**.

## EDUCATION

**Rochester Institute of Technology | Masters of Science in Data Science** Aug. 2024 – Dec. 2026

- **Coursework:** NoSQL, DBMS (SQL), Applied Statistics, Software Engineering for Data Science (Big Data), Applied Data Science (Deep Learning, LLMs), Foundations of Data Science (ML), Predictive Analytics (R) | **GPA: 3.96/4**

**Vellore Institute of Technology | Bachelors of Science in Electronics & Communication** Aug. 2017 – May 2021

- **Coursework:** Statistics, Linear Algebra, Probability Theory, Object Oriented Programming (Java, Python)

## WORK EXPERIENCE

**Information Systems & Technology Co-op | Analytics and Data Engineering** May. 2026 – Dec. 2026

*Rochester Regional Health* Rochester, NY

- Architected a **Jira** analytics platform across **Snowflake**, **Fabric**, and **Power BI** for cross-functional engineering and clinical teams.
- Modeled a **Bronze-to-Gold medallion architecture** in Snowflake using streams and stored procedures for business-ready tables.
- Orchestrated a **Fabric** pipeline with timestamp-based incremental ingestion, cutting Snowflake load runtime by **67%**.
- Built a **30+ table star schema** Power BI semantic model on **OneLake**, with **DAX** measures as the source of truth for reports.
- Delivered **5** multi-page Power BI reports with **20+ KPIs** as a Fabric Workspace App, improving decision-making efficiency.
- Automated a monthly **Fabric/Python** reporting workflow, turning SQL Server data into Excel and distributing via **Power Automate**.

**Data Analyst | Analytics and Data Engineering** Nov. 2023 – Jul. 2024

*Albanero (Company Acquired by Infor in 2024)* Hyderabad, India

- Extracted **10M+ ERP records** from **AWS S3**, **data lakes**, and data warehouses for cross-domain migrations.
- Developed **Python** and **SQL** scripts on Albanero's data mesh platform to clean and transform data, enhancing data quality.
- Collaborated with cross-functional teams and engineered **validation** checks to ensure schema compliance and catch anomalies.
- Built cloud compliant **ETL** workflows with **DAG** orchestration accelerating per-domain migration cycles by **95%** (3 days to 3 hours).

**Data Analyst | Project Engineer** Jul. 2021 – Feb. 2023

*Wipro* Remote, Bengaluru, India

- Reduced data prep time by **60%** for **5M+ records** and built **Power BI** dashboards that cut the reporting cycle by **50%**.

## PROJECT EXPERIENCE

**Customer Segmentation & Retention Analysis | Python, Scikit-learn, XGBoost, SHAP, Streamlit**

- Segmented **5,900+ customers** using **RFM analysis** and **K-Means clustering** on 1M+ transactions, validated with **ANOVA**, **Welch's t-test**, and **Chi-square** confirming significant behavioral differences.
- Built churn prediction models with **GridSearchCV** (Random Forest, XGBoost), achieving **AUC of 0.804** after fixing a data leakage bug that produced misleading perfect scores.
- Estimated **Customer Lifetime Value** with Random Forest to build a **2x2 business action matrix**, identifying **1,164 high-value customers worth \$1.4M** requiring proactive retention.
- Deployed an interactive **Streamlit** dashboard with customer lookup, dynamic churn/CLV sliders, and segment exploration.

**NYC Taxi Data Pipeline (Personal Project) | PySpark, Delta Lake, Kafka, Airflow, AWS, Terraform, Docker**

- Built an end-to-end **medallion architecture** (Bronze-Silver-Gold) processing **9.5M+ NYC taxi records** with **PySpark** and **Delta Lake**.
- Developed a real-time **Kafka/Spark Structured Streaming** pipeline with 5-minute tumbling windows, orchestrating the batch layer with **Airflow DAG** and automatic retries.
- Provisioned **AWS** infrastructure with **Terraform** (S3, Glue, Athena) and deployed Gold layer tables for serverless SQL analytics.

## SKILLS

**Languages:** Python, SQL, R, Java | **Data Engineering:** Apache Spark, Kafka, Airflow, Microsoft Fabric, Power Automate, ETL/ELT Pipelines, Data Modeling | **Databases & Tools:** SQL Server, MySQL, PostgreSQL, MongoDB, Neo4j, Snowflake, GridFS, Docker, Git, CI/CD, Streamlit, Jira, REST APIs | **Cloud:** AWS (S3, Glue, Athena), Terraform, Databricks, Azure | **Machine Learning:** NumPy, pandas, scikit-learn, TensorFlow, PyTorch | **Analytics:** R, Power BI, Tableau, Excel, matplotlib, plotly, A/B Testing, KPI Tracking