KRUTIK RAJESH PANCHAL

Rochester, NY | 585-957-2641 | krutikpanchal44@gmail.com | LinkedIn/panchalkrutik | Github/Krutik4

Education

Rochester Institute of Technology (RIT), Rochester, NY   **Aug 2023 - May 2025**
Master of Science in Data Science

University of Mumbai, Mumbai   **Jul 2016 - May 2020**
Bachelor of Engineering in Information Technology

Skills

Programming: Python (boto3, PySpark, scikit-learn, pandas, NumPy), SQL, Java
Databases: MySQL, PostgreSQL, MongoDB, Neo4j, Redshift
Cloud & DevOps: AWS, Azure, Docker, Kubernetes, Jenkins CI/CD , Git, GitHub, GitLab, Linux
Big Data: Apache Spark, Apache Airflow, Kafka
Visualization: Power BI(Microsoft Certified Data Analyst), Tableau, Qlik Sense, Excel, matplotlib, seaborn
Machine Learning & AI: Supervised/Unsupervised Learning, Feature Selection, Hyperparameter Tuning, NLP (LLMs, RAG)

Experience

Data Engineer   **Aug 2020 - Jul 2023**
Larsen and Toubro Infotech   *Mumbai, India*

- Architected and managed end-to-end data solutions on **AWS** using services like **EC2, S3, Redshift, Glue**, integrated with ETL tools like **Talend, dbt** resulting in a 40% improvement in data processing speed
- Optimized **ETL** and **data warehouse** processes by embedding automated validation checks, resulting in a 36% reduction in data discrepancies and improved data integrity
- Implemented real-time dashboards with **Power BI** and **Tableau**/ reducing processing time from 2 hours to 1 minute and accelerating data-driven decision-making across multiple projects
- Collaborated with cross-functional teams and stakeholders in **agile** environments, transforming business needs into effective data-driven solutions that supported key project deliverables

Graduate Teaching Assistant   **Jan 2025 - May 2025**
Rochester Institute of Technology   *New York*

- Designed and validated ETL pipelines for batch and stream processing workflows using **Python(PySpark), SQL, MongoDB, and Linux**, reducing processing times by 30% and ensuring data integrity
- Streamlined deployment and automation by integrating **GitLab CI/CD** and **Apache Airflow**, accelerating deployment cycles by 25% while enhancing system scalability and error handling
- Mentored 33 students across 5 hands-on projects in Software Engineering for Data Science (**Data Engineering**), fostering expertise in scalable big data systems and end-to-end ETL pipelines

Projects

LambdaSight : Serverless ETL & Analytics   **Dec 2024**

- Created a serverless ETL pipeline using **AWS Lambda, Glue, Python**, and **SQL** to extract, aggregate, and load data from **Amazon Redshift**, cutting down execution time by 50%
- Leveraged AWS CloudWatch for monitoring, **EventBridge** for event-driven orchestration, and **IAM** for secure access control to automate data workflows, cutting manual tasks by 30%
- Developed interactive **PowerBI dashboards** that provided real-time insights for stakeholders, enhancing decision-making through optimized **SQL** queries and data transformations

StockSage : Predictive Inventory Management   **Jan 2025**

- Constructed a robust **PySpark**-based ETL solution to integrate data from **MySQL, NoSQL**(MongoDB Atlas, and JSON files) into **medallion architecture**, reducing processing time by 30% and ensuring high data accuracy
- Orchestrated efficient batch processing workflows with **Apache Airflow** to transform raw order transactions into structured tables for dynamic inventory management, leveraging **Docker** for containerized deployments
- Integrated a **time series forecasting model** to predict sales and profit, delivering critical performance metrics (MAE, MSE) to support strategic decision-making

DataNexus : Data Modelling & Dashboard   **Mar 2025**

- Performed data transformation tasks via **PySpark**-driven workflows to process and **normalize** over **1M+ order records** into a structured **star-snowflake schema**, improving data query efficiency and accessibility
- Applied robust **data validation**, removing incomplete transactions and generating **10 core dimension** and **fact** tables along with **store-level partitioned datasets** to support modular data analysis and reporting
- Created an interactive **Tableau dashboard** showcasing KPIs like revenue trends, store-wise sales, supplier performance thus enhancing visibility for business stakeholders and reducing manual reporting time by 40%

TickSync : Cloud-Native Stock Stream with Kafka & S3   **Apr 2025**

- Deployed an end-to-end **real-time data pipeline** using **Apache Kafka on AWS EC2** to simulate stock trades from a **1M+ record dataset** using a Python-based producer with **Boto3**
- Streamed data to Kafka **consumer application** which processes and stores incoming trade data to **Amazon S3** in **Parquet format**, optimizing downstream processing and reducing query latency by **45%**
- Automated metadata management with **AWS Glue Crawlers** and queried live trading metrics using **Amazon Athena**, achieving **sub-second insights** into price trends and trading volumes