

KRUTIK BAJARIYA

| krutikbajariya123@gmail.com | +1 2074094704 | [LinkedIn](#) | [Portfolio](#) | [GitHub](#) |

PROFESSIONAL EXPERIENCE

COASTAL MEASURES

May 2024 - Present

Data Engineer (Aug 2025 – Present)

- Automated XBand data lifecycle management on AWS EC2 with Glue, **reducing storage overhead by 47%**
- Built a scalable AWS Data Lake pipeline to standardize **100K+ wave data records**, enabling selective retrieval and powering Athena dashboards with sub-second insights

Data Scientist and Software Developer Intern (May 2024 – Dec 2024)

- Supporting the U.S. Army Corps of Engineers' mission in water resource management, delivered real-time access to more than **50K daily sensor readings in 37% less time** using S3 and RDS, enabling faster, data-driven decision making
- Partnered with non-technical stakeholders to design **REST APIs** for large file uploads (**10MB+**) with AWS API Gateway and Lambda, enabling seamless data ingestion into Redshift data warehouse and supporting scalable data architecture
- Processed and standardized **11M+ hourly climate records across 264 stations** using Python, creating unified and scalable data pipelines within CUMULUS, Coastal Measures' central data orchestration platform
- Improved tide prediction accuracy using a Logistic Regression model, outperforming NOAA benchmarks by **12%** and strengthening climate resilience planning through high-quality data transformation and data modeling.

TCS (AVIVA Insurance), Data Engineer - Assistant Systems Engineer

Jul 2021 – Jul 2023

- Contributed to Aviva Insurance's enterprise data infrastructure by maintaining and optimizing **ETL** processes in IBM DataStage & PL/SQL, reducing job failures by **23%** and ensuring **full compliance** with data quality assurance protocols
- Led automation of **16 quarterly** business reports using CRON, eliminating manual processes and ensuring timely insights for operational and regulatory teams by integrating Python data pipeline development with reporting workflows
- Worked with a cross-functional team to flag duplicate and suspicious insurance claims by designing a data matching and deduplication logic using Python, leading to a **7% reduction in fraudulent payouts** and reinforcing data integrity
- Enabled faster decision-making for underwriting, fraud detection, actuarial, compliance, and BI teams (**~80 users**) by integrating CI/CD workflows, optimizing end-to-end ETL pipelines, and streamlining cross-team collaboration

PROJECTS

Real-Time Stock Market Data Pipeline ([Vizzes](#)) | PySpark & Technical Analysis

May 2025

- Designed an end-to-end Pyspark pipeline with Airflow to process **55K+** stock records & dashboard for real-time analytics.
- Created financial analytics module calculating technical indicators (RSI, moving averages, volatility metrics) and **identifying top-performing stocks** like NFLX (+23.3%), enabling data-driven investment insights

ReMo Book Recommendation ([Post](#)) | Natural Language Processing, Recommendation System

Jun 2024

- Led a team of 6 to develop a recommendation module using “**facebook/bart-large-mnli**” & “**all-MiniLM-L6-v2**” models, thereby increasing educator & student engagement by **42%**
- Preprocessed the 72K book synopsis data using SpaCy & NLTK and created a **51k x 51k topic modelled cosine similarity matrix**, which generated 25 recommendations based on previous reading history & genre

EDUCATION

Northeastern University, Master of Science in Computer Science (GPA: 4.0 / 4.0)

Jun 2025

University of Mumbai, Bachelor of Engineering in Information Technology

Jun 2021

SKILLS

Data Engineering & ETL: Apache Spark (PySpark), Oracle DataStage, Databricks (basic), dbt, CRON, Apache Airflow, R

Cloud Technologies: AWS (S3, EC2, Lambda, Batch, EMR, Redshift, RDS, API Gateway, IAM, VPC, CloudWatch)

Programming & Database Languages: Python, SQL (PL/SQL, MySQL, PostgreSQL), NoSQL, Shell Scripting

Data Processing & Data Analytics: Pandas, NumPy, Dask, PySpark, Boto3, Matplotlib, Seaborn, MATLAB

Agile Development Tools: Git (GitHub, BitBucket Stash), Jira, Confluence, Docker (basic), Linux/Unix, Jenkins

Business Intelligence (BI Tools): Tableau, Cognos, Qlik Sense, Looker