# Babin Joshi

Kathmandu, Nepal babin.joshi97@gmail.com (+977)9843633450 Website babin-joshi
babin411

## Profile Summary

Insightful, ambitious, and solution-oriented Computer Science graduate and Data Engineer by profession with a passion for transforming data into impactful insights. I specialize in optimizing data workflows, designing scalable ETL/ELT pipelines, and leveraging technologies like SQL, Spark, Azure, and AWS to drive efficiency and cost savings. From automating complex processes to simplifying legacy systems, I thrive on building data solutions that empower organizations to make smarter, data-driven decisions. Always eager to learn and grow, I'm dedicated to pushing the boundaries of what data can achieve.

## Education

**Kathmandu University, Dhulikhel, Nepal** *August 2017 – March 2022*
*Bachelor of Science in Computer Science*

- GPA: 3.59/4.0
- **Coursework:** Database Management System,Algorithm and Complexity, Machine Learning, Data Structures and Algorithm

## Experience

**Data Engineer** *Kathmandu, Nepal*
*Fusemachines* *May 2022 – Present*

- **Cloud Migration & Cost Optimization:** Led a critical project to migrate data pipelines from Azure Synapse to Databricks, implementing the Medallion Architecture (Bronze, Silver, Gold). This initiative resulted in a staggering 90% reduction in monthly cloud costs and an 85% reduction in peak costs, significantly improving operational efficiency.
- **Data Pipeline Development ETL Framework:** Architected and implemented end-to-end data pipelines for ingesting data from diverse sources like APIs, MongoDB, HubSpot, and Google Sheets. Developed robust ETL processes using AWS Glue and PySpark and on Azure Databricks, transforming raw JSON and other data formats into structured, business-ready datasets.
- **Automation Data Governance:** Automated pipeline triggers using Terraform scripts and established a robust monitoring system with alerts and notifications integrated with Slack and PagerDuty for pipeline failures. Implemented data validation processes using Spark to ensure data integrity and consistency between Azure SQL databases and ADLS Gen2 files.
- **Fraud Detection System:** Engineered a complete data pipeline for fraud detection, using AWS Boto3 to extract data from Amazon Redshift and store it in Amazon S3. Cleansed and feature-engineered the data with scikit-learn and orchestrated the entire workflow using Apache Airflow.
- **Mentorship Training:** Provided instruction and mentorship to new data engineering trainees on key topics including SQL, Apache Spark, Azure services (ADLS Gen2, Data Factory, Synapse), and Apache Kafka.

**Machine Learning Intern** *Kathmandu, Nepal*
*Fusemachines* *December 2021 – March 2022*

- Performed EDA, outlier handling, and feature selection on various datasets. Implemented ML algorithms and prepared detailed notebook documentation.
- Trained multiple ML algorithms (KNN, Logistic Regression, Random Forest, etc.) for stroke prediction, emotion classification, and credit risk analysis. Conducted hyperparameter tuning and built basic UIs using Streamlit and Gradio.
- Built a chatbot using Rasa for weather information. Created a blog site using Ruby and Jekyll, and deployed

it on GitHub Pages.

## Projects

**Sandstone Microscopic Image Segmentation** *GitHub*

- Uses image filters and random forest model for performing semantic image segmentation to locate the Quartz, Pore, Clay and Heavy materials in the sandstone microscopic image.
- Applied image processing filters like Gaussian, Sobel, Entropy and Gabor filters to prepare the data features for model training.
- Built a chatbot using Rasa for weather information. Created a blog site using Ruby and Jekyll, and deployed it on GitHub Pages.
- Tools Used: python, scikit-learn, semantic-segmentation

**BreakfastScoop** *GitHub*

- Scrapes news articles from Nepali news sites using BeautifulSoup
- Uses Naive Bayes classifier to classify news into 10 categories with an f1-score of 81%.
- Tools Used: python,webscraping

**Nepali License Plate Detection and Recognition** *GitHub*

- Uses YOLOv2 state of art object detection model to detect license plates from a video.
- Applied Otsu threshold technique to segment and extract license plate characters
- Developed Nepali character recognition model using Keras with an accuracy of 96%.
- Tools used: Python, YoLO,Neural Network

## Certifications

- **Microsoft Certified: Fabric Data Engineer Associate:** Microsoft — May 2025
- **Databricks Certified Data Engineer Associate:** Databricks — March 2025
- **Microsoft Certified: Fabric Analytics Engineer Associate:** Microsoft — Jan 2025
- **Astronomer Certification for Apache Airflow Fundamentals:** Astronomer — Nov 2024
- **Databricks Certified Associate Developer for Apache Spark 3.0:** Databricks — Dec 2022
- **Tableau Fundamentals:** DataCamp — Aug 2022

## Achievements

- **Winner of AI Hackathon:** Played a key role in the development of *Fuse Inventory Counter*, an innovative solution for inventory management. Contributed significantly to dataset generation using image augmentation techniques, enabling the system to leverage AI, computer vision, and ML effectively for stock monitoring and demand prediction.
- **Mentor and Instructor:** Acted as an instructor and mentor for Apache Airflow, SQL, Apache Spark, and Azure technologies, conducting training sessions and guiding new data engineering trainees in advanced implementation techniques.

## Skills

**Programming Languages:** SQL(3 +years), Python (3 +years)C++, C, Java,

**Database and Tools:** PostgreSQL, SQLServer, MSSQL, MySQL, Git, Postman, Docker, Github

**Cloud Technologies:** AWS(S3, AWS Glue, Athena, MWAA), Azure ( ADLSGen2, Azure Databricks, Azure Data Factory, Azure Synapse, Azure Logic Apps), GCP(BigQuery), Microsoft Fabric, Databricks

**Orchestration Tools:** Apache Airflow

**Frameworks:** pyspark, pandas, matplotlib, numpy, scikit-learn, OpenCV

**Languages:** Nepali, Hindi and English