BILAL MOMIN

AI & LLM | Cloud Data Specialist | Data Scientist
Bachelor of Engineering, Mumbai University
Mumbai, India
www.linkedin.com/in/bilal-momin | bm.works00@gmail.com

About Me

LLM-focused Data Scientist with 4+ years of experience building scalable data pipelines and deploying real-world AI solutions using Large Language Models. I've led the design and implementation of NLP systems—including fine-tuned BERT models—for semantic search, topic classification, and sentiment analysis, integrated into APIs and real-time services. Skilled in Python, SQL, Airflow, Kafka, and Docker, with hands-on cloud experience in Google Cloud and AWS. My work spans full ML lifecycles—from data crawling and preprocessing to model training, evaluation, and deployment. As a Google-certified Data Engineer, I specialize in transforming unstructured data into actionable insights, and I'm currently focused on building RAG pipelines, fine-tuning LLMs, and creating production-grade GenAI tools across domains.

What I Bring to the Table?

- **LLM & AI Focus**: Experience building and deploying LLM-based systems for real-world use cases like document QA, semantic search, and text classification using BERT and RAG pipelines.
- **Full-Stack Data Expertise**: From crawling and ingestion to storage, modelling, and API deployment—delivering robust, end-to-end data solutions.
- **Cloud & API Scalability**: Designed and managed scalable APIs and infrastructure on Google Cloud and AWS, optimized for performance, availability, and cost.
- **Real-Time & Big Data Handling**: Built pipelines to process millions of records monthly with tools like Kafka, Airflow, BigQuery, and PySpark.
- **Collaboration & Product Thinking**: Work closely with cross-functional teams to align AI/data efforts with real user problems, focusing on both technical depth and business impact.
- **Certifications & Learning**: Certified **Google Professional Data Engineer**, continuously exploring the latest in LLMs, GenAI, and cloud-native data tooling.

Professional Experience

Data Scientist
Admazes Limited, Hong Kong
December 2021 – Present

- Led the backend development of a Google Trend Prediction platform, processing over **100 million** data points monthly.
- Designed and implemented **ETL** pipelines, developed **machine learning** models, and deployed **APIs** to deliver real-time trend forecasts and demographic insights.
- Built a **Topic Categorization** Engine that automates the classification of search queries using machine learning models.

- Developed and maintained **scalable web crawlers** for platforms like Quora, Reddit, Twitter, LinkedIn Sales Navigator, and TOR, collecting millions of records that fuel **business analytics**.
- Managed **database hosting** and **API deployments on Google Cloud**, ensuring scalability and high availability for business-critical applications.

Python Developer
Codemarket, California
December 2020 – February 2021

- Created web-based services using Python, MongoDB, and ReactJS, meeting complex business requirements. Developed GraphQL APIs to query multiple databases, improving backend efficiency and enabling dynamic front-end queries.
- Deployed scripts on AWS ECR using AWS Fargate and AWS Lambda, automating key business processes.
- Managed multiple projects and GitHub repositories simultaneously, coordinating across teams for smooth operations.

Technical Skills

Programming Languages: Python, SQL
Data Engineering Tools: PySpark, Pandas, Airflow, Kafka
Cloud Platforms: Google Cloud (BigQuery, Dataflow), AWS (Lambda, Fargate, Kafka, ECS)
ETL & Data Processing: Data Pipeline Design, Real-time & Batch Processing
Databases: MongoDB, Postgres, MySQL, ChromaDB (Vector Store)
API Development: FastAPI, Flask, Django, RESTful Services
Machine Learning & LLMs: RAG, BERT, OpenAI/GPT, RAG, Prompt Engineering)
Containerization & Orchestration: Docker, Kubernetes
Version Control: Git (GitHub, GitLab)

Projects

1. Scalable Keyword Trend Forecasting Platform with Google SERP Data

Led the development of a sophisticated platform processing over 100 million monthly data points from Google SERP to forecast keyword trends. Designed and implemented **scalable ETL pipelines**, developed clustering algorithms and **machine learning** models, and deployed APIs for real-time trend forecasting. Managed the **Google Cloud infrastructure**, optimizing it for **scalability** to handle massive datasets and deliver timely insights for stakeholders.

2. LLM-Powered Document Chat API

Developed and deployed a production-grade API that enables users to query their own documents (PDF, TXT, CSV, XLSX) using **natural language**. Handled end-to-end file management in the **cloud**, including secure upload, parsing, and format-specific preprocessing. Engineered a chunking and **metadata-enrichment** strategy to support **context-aware** embedding generation, enabling **high-accuracy** retrieval across multiple document types and sources. Implemented a **Retrieval-Augmented Generation** (RAG) pipeline using LLMs and stored vector embeddings in ChromaDB. Designed the system to handle large-scale API calls with low latency, delivering accurate, contextually relevant responses even across multiple files.

3. Crawlers for Quora, Reddit, Twitter, LinkedIn, and TOR Data

Developed robust web crawlers to scrape high volumes of data from platforms like Quora, Reddit, LinkedIn Sales Navigator, and the TOR network. These crawlers have operated consistently for over two years, collecting **million+** records. The system incorporates real-time error handling, data deduplication, and multi-threading, significantly enhancing efficiency and reliability. Integrated with **MongoDB** for real-time data storage and processing, ensuring seamless access to large datasets.

4. Topic Categorization Engine

Designed and implemented a machine learning-based topic categorization engine using large language models (**LLMs**), specifically fine-tuned **BERT models**, to classify search queries into relevant topics. Built a comprehensive **ETL pipeline** for data ingestion and LLM training, utilizing BERT’s deep contextual understanding to improve the accuracy of text classification. Deployed the system on **Google Cloud** for scalability and high availability, and developed APIs to deliver realtime predictions based on the **trained ML/LLM**, streamlining the categorization process for faster and more precise results.

5. Google Cloud API and Database Hosting

Deployed and managed **multiple APIs and databases on Google Cloud**, utilizing Docker for container orchestration. Ensured secure, scalable database hosting on Google **Cloud Compute Engine and Cloud Storage** and facilitate seamless updates.

6. Large Scale Data Deduplication

Spearheaded the deduplication of **large-scale customer CRM** data on a weekly basis. Managed **data ingestion from multiple sources**, performing checks for duplicates among new data and existing CRM records, with output directed to **BigQuery**. This process enhances **data integrity and ensures reliable analytics**.

7. Portfolio Visualizer

Developed a service for investors to analyze their stock portfolios in a **consolidated view**. This tool allows users to **compare various stock metrics** in one place, enhancing investment decisionmaking. The backend is built in **Python**, while the frontend utilizes **Flask**, providing a seamless user experience.