SAMIKSHA BARASKAR

Boston, MA | +1(857)-390-5787 | baraskar.s@northeastern.edu | LinkedIn | GitHub | Portfolio |

EDUCATION

Northeastern University, Boston, USA **Dec 2024**
Master of Science in Data Analytics Engineering, **GPA : 3.93/4.0**
Relevant Coursework: Database Management, Data Mining, Computation & Visualization, Data Warehousing & Integration

Dr. Vishwanath Karad's MIT World Peace University, Pune, India **July 2021**
Bachelor of Technology in Computer Science and Engineering, **GPA : 10 / 10**
Relevant Coursework: Data Structures & Algorithms, Data Warehousing, Big Data Analytics, [Innovation Patent]

WORK EXPERIENCE

Graduate Teaching Assistant (IE7275 Data Mining and Engineering), *Northeastern University* **Sep 2023 – May 2024**
- Mentored 50+ students on data analysis, classification, clustering, and predictive modeling using Python.
- Assisted in designing coursework, grading assignments, and guiding industry-based projects.

Data Engineer , *TietoEvry Private Ltd, Pune, India* **Aug 2021 – Dec 2022**
- Implemented a **distributed and scalable ETL** system using **SQL** and **Python** scripts to **automate document transfer**, aligning with business requirements and improving deployment efficiency by 20% for Dochotel and Share@Poyry.
- Utilized **SQL Server Management Studio** (**SSMS**) for advanced index optimization and schema redesign, by optimizing **relational** and **non-relational databases** for efficient data storage and **data retrieval** performance.
- Optimized complex **SQL** queries and leveraged **Informatica** and **SSIS** for data quality monitoring, eliminating duplicate records and ensuring data integrity and consistency across the system.
- Collaborated with **data owners and cross-functional teams** to resolve **inconsistencies** and operational reporting.

Data Engineer, *Erai Technologies Private Ltd, Pune, India* **Jan 2021 – May 2021**
- Developed cloud-based Smart Care AI system, improving skill mapping accuracy by 40% using Google Cloud AI.
- Built and automated **real-time ETL workflows** using **Google Cloud Dataflow**, **Pub/Sub**, and **BigQuery**.
- Improved processing efficiency by **27%** through optimized data extraction and transformation techniques.
- Worked with business analysts and engineering teams to resolve data quality issues and improve reporting accuracy.

ML – AI Intern, *Erai Technologies Private Ltd, Pune, India* **June 2020 – Dec 2020**
- Designed and developed the **Smart Care AI system** using **Python**, **TensorFlow**, and **NLP**, significantly improving skill mapping accuracy by 40% through robust data modelling and mining practices.
- Built scalable **ETL pipelines** integrating **Apache Kafka** with **Elasticsearch** to enable real-time data streaming and indexing of candidate resumes, optimizing data retrieval and processing workflows.
- Enhanced data processing efficiency and application responsiveness by 27% by automating **Kafka pipelines** for seamless data collection, transformation, and distribution.
- Collaborated with stakeholders and **cross-functional teams** to design scalable data architectures, leveraging **Elasticsearch** for efficient storage, retrieval, and indexing, ensuring data consistency and accuracy.

PROJECTS

Medi-Mart [AWS, Python, Spark, Airflow, ETL] **Sept 2024 - Dec 2024**
- Built a **scalable ETL pipeline** using **AWS Glue, S3, and Redshift** to process structured and unstructured data.
- Automated data transformation with **PySpark Glue Jobs** and **Airflow DAGs**, improving pipeline efficiency by 30%.
- Enabled real-time insights using **AWS Athena**, reducing decision-making time by 25%.

Redfin Real Estate Data Pipeline [Python, Airflow, Snowflake, Power BI] **May 2024 - June 2024**
- Engineered a robust ELT pipeline to process real estate data into **AWS S3** and **Snowflake**, optimizing data workflows.
- Automated reporting and visualization with **Power BI**, delivering interactive, **real-time dashboards**.

Spotify Personalized Recommendation System [Python, Spark, GCP, ML] **Jan 2024 – Apr 2024**
- Built scalable data pipelines with **Apache Spark** to process large-scale Spotify datasets, improving throughput by 40%.
- Developed ML models with **85% accuracy** for real-time user preference predictions and **predictive analytics**.
- Integrated recommendation outputs with backend services on **GCP** for real-time predictions.

TECHNICAL SKILLS

Certifications:	**AWS Cloud Practitioner, AWS Associate Data Engineer**
Programming Languages:	Python, SQL, T-SQL, PySpark, Scala, Shell Scripting
Big Data & ETL:	Apache Spark, Kafka, Hadoop, Airflow, dbt, Informatica, Elasticsearch
Databases:	MySQL, PostgreSQL, MongoDB, NoSQL, Hive, DynamoDB
Cloud Computing:	AWS (S3, Lambda, Glue, Athena, Redshift, Kinesis), GCP (BigQuery, Dataflow)
DevOps & Deployment:	Docker, Kubernetes, CI/CD, REST APIs, Flask, Fast API, GitHub, Confluence