

Sachin Ramesh Shet

✉ shet.sa@northeastern.edu 📞 +1 617-505-8176 🌐 in/SachinShet73 🌐 sachinshet73.vercel.app

Education

Master of Science in Computer Software Engineering

Northeastern University • 3.667/4.0

Boston, MA
September 2024 – Present

- **Advanced Coursework:** Data Science Engineering Methods and Tools, Program Structures and Algorithms, Database Systems.

Experience

Infosys

Senior Systems Engineer

Bangalore, India
November 2021 – February 2024

- Architected end-to-end Extract Transform Load (ETL) pipeline using Azure Data Factory to manage 2GB+ daily data flow between server, Linux environment, and Azure Blob Storage, implementing automated validation checks.
- Built Python automation scripts handling **55,000+ rows daily**, cutting processing time by **60%** through automated table creation and data insertion.
- Created reliable data migration workflows between **Azure Blob Storage** and **Snowflake**, maintaining **99.9% data accuracy** across **180+ tables** with automated validation.
- Implemented ETL workflows using **Apache Airflow**, decreasing manual tasks by **90%** and maintaining complete data integrity during 6-month deployment.
- Received the Blue Badge award (top 5% of engineers) and earned two promotions within **2.5 years** for technical contributions.
- Led technical discussions with international clients to convert business requirements into specific technical solutions, completing 3 major deliverables ahead of schedule.

RoveLabs

Machine Learning Intern

Davangere, India
April 2020 – September 2020

- Authored a real-time candidate verification system using **OpenCV** and **dlib**, achieving **95% accuracy** in face detection and maintaining **30 FPS** processing speed.
- Implemented **concurrent processing** for multimodal data analysis, integrating video streams and audio pattern detection with **TensorFlow** and **librosa**, resulting in less than **2-second latency**.
- Engineered an audio anomaly detection system achieving **90% accuracy** while reducing false positives in suspicious activity detection by **40%**.

Projects

Financial Research Intelligence Platform | Airflow, FastAPI, Snowflake, Docker

- Engineered an automated ETL pipeline using **Apache Airflow DAGs** to scrape and process **CFA** research publications, implementing **custom operators** for **PDF extraction** and **metadata validation**, resulting in successful processing of **100+ documents** with **5MB+ file sizes**.
- Architected a scalable document storage solution leveraging **AWS S3** for raw PDFs and **Snowflake** for structured data, designing optimized **table schemas**, and implementing automated **validation checks** to maintain data integrity across **2GB+ of research documents**.
- Built document analysis system using OpenAI API with custom prompt engineering and caching, achieving consistent 5-second response times for queries.
- Built **RESTful API** architecture using **FastAPI**, implementing **10+ endpoints** with **async support** for document operations, **JWT authentication**, and **rate limiting**, ensuring secure and efficient access to research data.
- Created an interactive analytics dashboard using **Streamlit** with custom visualization components, implementing efficient **search algorithms** and **caching mechanisms** to reduce document retrieval time from **minutes to under 10 seconds**.
- Implemented a comprehensive **CI/CD pipeline** using **Docker** and **GitHub Actions**, containerizing multiple services, including **FastAPI**, **Streamlit**, and **Airflow**, reducing deployment cycles by **70%**, and ensuring consistent environment setup.

Real-time Social Media Sentiment Analysis Pipeline | Kafka, Spark, Docker, PostgreSQL, Elasticsearch

- Engineered a real-time data pipeline using **Twitter API** and **Apache Kafka**, processing **1,000+ posts/hour**, reducing brand monitoring time from days to **minutes**, and enabling immediate responses to customer sentiment.
- Developed real-time visualization dashboards using **Kibana**, allowing teams to track sentiment trends and launch targeted marketing initiatives, driving **data-driven** decision-making.
- Implemented distributed data processing using **Apache Spark**, achieving **70%** sentiment analysis accuracy with **2-second** processing latency.
- Designed a hybrid storage solution combining **PostgreSQL** for structured data and **Elasticsearch** for efficient text search, reducing query response times to under **3 seconds**.
- Containerized the entire application stack with **Docker**, ensuring consistent deployment across development environments and achieving a **90%** deployment success rate.

Skills

Data Engineering: Apache Airflow, Apache Kafka, Apache Spark, Azure Data Factory, Python (NumPy, Pandas)

Cloud & Infrastructure: AWS S3, Azure Blob Storage, Docker, Linux, GitHub Actions, CI/CD, Git

Data Storage & Analytics: Snowflake, PostgreSQL, Elasticsearch, MongoDB

Data Visualization: Kibana

AI & Machine Learning: OpenAI API, TensorFlow, OpenCV, dlib, librosa

Web & API Development: FastAPI, RESTful APIs, Streamlit

Programming Languages & Core: Python, Java, SQL