Chetana Thorat

☎ 812-553-1900 ✉ thoratchetana4@gmail.com LinkedIn GitHub

Education

Master of Science - Data Science **August 2024 – Present**
Indiana University Bloomington **GPA: 3.5/4.0**

Bachelor of Engineering - Computer Engineering **July 2019 – May 2023**
Savitribai Phule Pune University **GPA: 9.5/10**

Technical Skills

Programming Languages: Python, SQL, Java
Cloud & AWS: S3, Glue, Athena, Lambda, Redshift, EMR, CloudWatch, SQS, KMS, Secrets Manager, Kinesis
Databases & Visualization: PostgreSQL, Snowflake, Amazon RDS, Tableau, Power BI, AWS QuickSight, Alteryx
DevOps & Monitoring: Docker, Git, CI/CD, Prometheus, Grafana
API & Deployment: OpenAPI/Swagger, RESTful Services, Jest, PyTest, Flask, FastAPI, Streamlit, REST APIs, GraphQL
Data Engineering: Apache Spark (PySpark), Apache Airflow, Apache Kafka, Apache Flink, dbt, Data Modeling, Terraform, ELK, Talend, Fivetran
AI/ML & LLMs: Scikit-learn, XGBoost, LightGBM, CatBoost, TensorFlow, Keras, PyTorch, Hugging Face Transformers, BERT, LLaMA, LangChain, GroqCloud, Retrieval-Augmented Generation (RAG), CNNs, RNNs, LSTMs

Work Experience

Marketing Data Analyst **June 2025 - Present**
Indiana University Bloomington **Indiana, USA**

- Developed and maintained 100+ automated data pipelines using **SQL**, **Python**, **dbt**, and **Airflow** to support scalable email campaigns and ensure seamless integration between **Salesforce CRM** and subscription systems.
- Designed self-service dashboards and executive-level reports in **Looker** and **Tableau**, delivering insights that increased email click-through rates by 32% and open rates by 18%.
- Conducted advanced funnel analysis and **A/B testing** to optimize campaign performance, resulting in a 27% improvement in targeting precision, 22% lower unsubscribe rates, and a 33% increase in application completions.

Data Engineer **Feb 2025 - May 2025**
Indiana University Bloomington **Indiana, USA**

- Developed an **ETL pipeline** using **Azure Data Factory**, ingesting 50+ public CSV datasets into **Data Lake Gen2**.
- Transformed 10K+ records using **PySpark**, handling nulls and resolving 12+ column schema issues.
- Built 5+ external tables in **Azure Synapse Analytics** to query and rank using optimized **T-SQL** scripts.
- Automated the entire workflow using **ADF triggers** and integrated Databricks notebooks for scheduled transformations.
- Designed interactive dashboards in **Tableau Public**, boosting data exploration speed by **60%** and enabling insights for 100+ countries.

Data Analyst Intern **Oct 2024 - Dec 2024**
Indiana University Bloomington **Indiana, USA**

- Analyzed unstructured data from **12M+ trademark case records** stored in multiple CSV files using **Dask**, **Pandas**, and **NumPy** for scalable parallel processing.
- Uncovered patterns in opposition filings and legal outcomes, optimizing queries to reduce execution time by **60%**.
- Designed a reusable data integration framework to merge event logs, ownership history, and case files for analytics.
- Built interactive dashboards using **Power BI**, enabling real-time exploration of case trends and improving stakeholder reporting turnaround by **50%**.

Software Development Engineer **Aug 2023 - July 2024**
SAS R&D **Pune, India**

- Created **Jenkins** pipelines with **Groovy** scripts, automating **CI/CD workflows** and reducing deployment time by **35%**.
- Automated **Docker-based** deployments through **Kubernetes**, improving system scalability by **30%**.
- Enhanced test coverage by **40%** for the **SWARM SECURITY** project, ensuring **Spring Boot 3.2** compatibility.
- Upgraded SAS Java common libraries in BOM from Spring Boot 2.7 to 3.0 for Viya 4, ensuring compatibility and stability.
- Strengthened system security by integrating third-party CVE patches, reducing vulnerabilities by 50%.

Academic Projects

SalesData ETL Pipeline

- Built an end-to-end ETL pipeline using Apache Airflow to extract sales data from Azure Blob Storage, transform it with Pandas, and load aggregated output into a PostgreSQL table.
- Defined DAGs using PythonOperator and PostgresOperator to manage task dependencies and retries via the Airflow UI.
- Containerized the setup using Docker Compose to deploy Airflow, PostgreSQL, and pgAdmin, and built Power BI dashboards for visualizing and analyzing aggregated sales data.

IPL Data Analysis (Databricks & Pyspark)

- Created an Azure Resource Group and configured a Databricks cluster for scalable Spark-based development.
- Built an end-to-end PySpark pipeline to process 1.5M+ IPL records, performing data cleaning, joins, aggregations, and window functions using Spark SQL and the DataFrame API to deliver actionable insights.
- Visualized analytical results by converting Spark DataFrames to Pandas and created plots using Matplotlib and Seaborn.

Vaani – Real-Time Voice-to-Text

- Developed Vaani, a real-time transcription system using **Whisper** for speech-to-text conversion, **pyannote.audio** for speaker diarization, and **VADER** for sentiment analysis, reducing transcription time by **40%.**
- Integrated Groq's **LLaMA3-8B** model for generating structured summaries, action items, and key decisions from transcriptions, improving meeting insight accuracy by **30%** and enhancing decision-making efficiency.

DocVerse – ChatBot [Live Demo]

- Crafted a **RAG-based document processing app** using **LangChain** for text chunking and **Hugging Face embeddings**, reducing processing time by **30%** and manual analysis by **80%**.
- Implemented **FAISS** for fast similarity search and integrated **Gemma2-9b** for context-aware responses, enabling structured summarization and boosting search accuracy by **70%**. Deployed the chatbot using **Streamlit** for real-time document query handling.

Caries Risk Predictor – AI Web App [Live Demo]

- Developed an AI-powered dental caries risk predictor using **XGBoost**, achieving over **91% accuracy** on patient data and handling class imbalance with **SMOTE**.
- Built an interactive **Streamlit** web app to collect user inputs, predict risk level, and display results in real time.
- Integrated a **Groq-powered LLaMA3 chatbot** that provides step-by-step oral health guidance based on risk level and user symptoms.

Mineral Map Visualization Platform [Live Demo]

- Developed an interactive web platform using **React.js** and **ArcGIS**, visualizing **100+ U.S. mineral projects** with real-time geospatial rendering based on mineral type, land ownership, and project status.
- Integrated custom **color-coded markers** and **shape-based legends** to distinguish **6+ minerals** and **5+ project statuses**, improving clarity and user navigation by **60%**.
- Built a scalable frontend with **Node.js**, **HTML/CSS**, and dynamic filters, enabling real-time search and detail views from structured geocoded CSV data, reducing exploration time by **40%**.

Achievements

- **Women in Tech - InnovateHer:** Secured **3rd place** in the Technology Showcase and Competition.
- **Teaching Assistant** – Object-Oriented Software Methods at Indiana University.
- **Luddy Hackathon:** Participated with project *DocuVerse*, a document-based intelligent chatbot.
- **Google Developer Student Club (GDSC):** Served as a Management Team Member in the Computer Department.
- **Hacktoberfest:** Contributed to open-source projects as part of the global Hacktoberfest challenge.
- **Copyrighted Research Work:**
  - *Breast Cancer Detection Using Deep Learning* – Copyright No: **14643/2023-CO/L**
  - *Hobby Recommender System for Kids using Machine Learning* – Copyright No: **L-92820/2020**

Certifications

- **Alteryx Fundamentals** [View Certificate]
- **Introduction to Airflow in Python** [View Certificate]
- **Introduction to Databricks** [View Certificate]
- **Introduction to Power BI** [View Certificate]
- **Introduction to Snowflake** [View Certificate]
- **Introduction to Apache Kafka** [View Certificate]