Harsh Gupta

930-904-3651 – hagupta@iu.edu – linkedin.com/in/harshg1347 – github.com/Harsh1347 – harsh1347.github.io

Education

Indiana University Bloomington Aug 2024 – May 2026
Master of Science in Data Science (GPA: 4/4) *Bloomington, IN*

Manipal University Jul 2017 – May 2021
B.Tech in Electronics and Communication, Minor in Data Science (GPA: 8.13/10) *Manipal, India*

Technical Skills

Programming: Python, R, PySpark, SQL
Cloud & Data Platforms: AWS, Azure, OracleDB, Postgres, Redshift, Amazon Aurora, Hadoop, PowerBI, Tableau
Data Engineering: AWS Glue, Spark, Redshift, ETL Pipelines, Airflow, Data Ingestion, Feature Engineering, Data Modeling (Fact/Dimensional), Data Warehouse Design, Docker, CI/CD Pipelines
Generative AI: LangChain, LangGraph, MCP, Semantic Kernel, Azure AI Studio, Ollama, Hugging Face, RAG

Experience

Indiana University Bloomington Campus Auxiliaries Jun 2025 - Aug 2025
AI & Data Engineer Intern | *Azure AI Search, Azure ML, Python, FastAPI, LangChain* *Bloomington, IN*

- Developed a **RAG-based internal knowledge assistant** using Azure AI Search & LLMs, **enabling 250+ staff to query 12,000+ Confluence pages** with natural language, **reducing manual search time.**
- Engineered an **end-to-end ingestion pipeline** (extraction → context-aware chunking → embeddings → AI Search Index), enabling **automated updates of 12,000+ existing and new documents.**
- **Enhanced response accuracy from 76% to 93%** by implementing **metadata filtering & small-to-large context expansion**, improving document retrieval accuracy and user confidence.

Indiana University Bloomington Oct 2024 - Present
Part-Time AI & Data Engineer | *Azure, Python, SQL, Tableau, LLM* *Bloomington, IN*

- Developed **time-series prediction pipelines** for **dining operations** using Azure Data Factory and SQL.
- Crafted the **Data Ingestion** and **Feature Engineering pipeline** utilizing **Azure Data Factory** to transform transactional data into an ML dataset enriched with multiple features, resulting in a significant **21% increase** in prediction **accuracy**.
- **Boosted pipeline throughput 6×** using distributed Azure compute, achieving a **residual error of 5%**.

Deloitte - Strategy & Analytics Sep 2021 - Jul 2024
Data Engineer - FSI Domain | *AWS, Spark, OracleDB, Postgres* *Bengaluru, India*

- **Migrated legacy mainframe systems to AWS cloud**, building data ingestion & curation pipelines using AWS Glue & PySpark; developed a reusable PySpark-based framework, **cutting development effort by 25%.**
- Designed and operated **end-to-end data ingestion pipelines** using AWS Glue, Spark, and Redshift, **processing 15M+ records/day** into a unified cloud data warehouse. Modeled fact and dimension tables to support analytics, reducing time-to-insight and development cost.
- Built a scalable **data validation, logging, and metadata framework** to ensure ETL data quality and lineage in production pipelines, **reducing issue resolution time by 33%.**

Data Scientist - Customer Strategy & Pricing | *Python, R, Machine Learning, AWS S3, Tableau, Excel*

- Delivered **pricing strategy** by modeling demand elasticity patterns across 5,000+ stores, **increasing profit by $1M–$3M while maintaining <1% guest loss**. Enabled business stakeholders to make data-driven pricing decisions.
- **Streamlined** the **Pricing-as-a-Service data pipeline, boosting efficiency by 30%** and supporting client acquisition through success story presentations.

Projects

CodETL | *PySpark, Flask, Celery, Postgres, AWS S3, AWS EC2, Apache Airflow*

- Spearheaded the development of a **platform-agnostic ETL engine** with standardized transformation maps and **Airflow-orchestrated**, topologically sorted schedules, **enabling 25+ transformations**, maximizing parallel processing, and **reducing development efforts by 40%**—earning Deloitte's Outstanding Award.

DataLens | *Python, LangChain, LLM*

- Developed an **LLM-powered search engine** and Q&A portal for **structured and unstructured data using RAG** for postgres database, and documents.