YASH KHARE

Boston, MA | (857) 506-4305 | khare.y@northeastern.edu | linkedin.com/in/yash-khare20 | github.com/YashKhare20
yashkhare20.github.io/portfolio

Education

Northeastern University, GPA: 4.0 / 4.0 — Jan. 2024 – Present
Khoury College of Computer Sciences, Candidate for Master of Science in Data Science — *Boston, MA*
Relevant Coursework: Supervised/Unsupervised Machine Learning, Natural Language Processing, Algorithms

MIT World Peace University, GPA: 9.54 / 10.0 — July 2017 – May 2021
B.Tech in Computer Science & Engineering — *Pune, India*
Relevant Coursework: Data Structures & Algorithms, Database Management, Object Oriented Design

Experience

L.L. Bean — Jan. 2025 – Present
Data Architect Intern — *Freeport, Maine*

- Architected a **GCS-to-BigQuery** ingestion pipeline using **Dataflow** and **Cloud Composer**, processing **10K+** daily files and enhancing raw data availability by **90%**.
- Developed a **YAML-driven configuration framework** and implemented **partial record ingestion** with BigQuery audit tables, reducing pipeline setup time by **80%** and reruns by **45%**.
- Designed a **reusable multi-agent framework** using Google's **Vertex AI Agent Development Kit (ADK)** for automated failure analysis across GCP services, achieving **99.9%** uptime with plug-and-play extensibility.
- Implemented **source-controlled CI/CD pipelines** with **GitHub** and **Harness** for agent deployment, integrating **Splunk API** for real-time log extraction and automated **Microsoft Teams/email alerting** with comprehensive test suites.

Quantiphi Analytics Solution Pvt. Ltd. — July 2021 - Nov. 2023
Senior Data Engineer — *Mumbai, India*

- Integrated **SonarQube** & **GitHub Actions** to automate quality checks in **Composer DAGs** and streamlined migration of **10K+** SKUs from **Snowflake** to BigQuery, boosting code security and data workflow productivity.
- Directed the development of a **Kafka-PubSub** real-time **ETL** pipeline for a financial institute, integrating Cloud Composer, BigQuery, & Dataflow to handle **2TB** daily, which resulted in a **40%** increase in operational efficiency.
- Automated molecular enumeration of **1B** molecules using **Dataflow** & **Workflows** for a healthcare innovator and subsequently implemented **CI/CD** pipelines for model training.
- Orchestrated **MySQL** & **MongoDB** to BigQuery migration for a drone racing league, utilizing **Datastream** and **Cloud Run**, streamlined with **Cloud Scheduler** and secured via **Secret Manager**, impacting **5K** drone pilots.

Maker's Lab, Tech Mahindra — Nov. 2020 - May 2021
Data Engineering Research Intern — *Pune, India*

- Developed an automatic image annotation system using **YOLOv4-Tiny**, processing data with **Apache Spark** on **HDFS**, achieving **40 mAP @.5** on the MS COCO dataset and reducing annotation time by **66%**.
- Deployed the system using **Docker** and **Kubernetes**, optimizing CI/CD in **GitLab**, and cutting annotation costs by **89%**.

Projects

HIMAS: Healthcare Intelligence Multi-Agent System for Federated Learning — Sept. 2025 - Dec. 2025

- Architected a **federated learning** pipeline using **Apache Airflow** to process **60K+** ICU records from **MIMIC-IV** across 3 hospitals, implementing **differential privacy** (ε=0.1) and **k-anonymity** for HIPAA-compliant mortality prediction.
- Trained a **TensorFlow** mortality prediction model using **Flower** framework with **FedAvg strategy** and **Bayesian optimization**, achieving **0.92 ROC-AUC**, and **74% recall** across **13K+** test samples with **11.6%** class imbalance.
- Designed a **multi-agent system** using **Vertex AI ADK** with **A2A protocol** for cross-hospital coordination, and implemented **CI/CD** with **GitHub Actions**, **DVC** on **GCS**, achieving **zero data leakage** across data splits.

Technical Skills

Languages: Python, SQL, Shell Script, R, Javascript, C#, HTML/CSS
Frameworks & Tools: Apache Hadoop, Spark, Kafka, Kubernetes, Airflow, Docker, Git, Flask, DVC, Flower, Pytest
ML/AI Platforms: TensorFlow, Keras, Pytorch, Hugging Face, Google ADK, Scikit-learn, Pandas, Numpy
Data Visualization: Tableau, Power BI, Matplotlib, Looker, Plotly, Seaborn
Databases/ETL: BigQuery, Snowflake, MySQL, SQL Server, Hive, HBase, Informatica, Apache NiFi, Google Firebase
Certifications: **Google Cloud** Certified Associate Cloud Engineer, **Professional Data Engineer**