Abdul Rafay Ahmed Khan

+1 (303) 9615167 rafay.ahmedkhan@colorado.edu rafaykhan11 rafayak1

Education

University of Colorado Boulder **2024 – 2026**
Masters of Science in Computer Science - ***cGPA: 4***
Relevant Courses: Data Center Scale Computing, Data Mining, Database Systems, Big Data Architecture

Lahore University of Management Sciences **2020 – 2024**
Bachelors of Science in Computer Science - ***Dean's Honor List***

Technical Skills

Languages: Python, SQL, R, JavaScript, C++
Data & ML: Pandas, NumPy, PySpark, PyTorch/TensorFlow, Tableau, Power BI, Plotly, Seaborn, ggplot2
Platforms/Cloud: Apache Spark, Airflow, Databricks, BigQuery, dbt, Presto; GCP, AWS; Docker, Kubernetes
Databases/Tools: PostgreSQL, MySQL, Snowflake, Oracle, SQLite, Neo4j, Firestore; Git

Work Experience

Meta **May 2025 – August 2025**
Data Engineering Intern - Menlo Park, California, United States

- Built and deployed scalable batch pipelines using **dbt**, **Apache Spark**, **Apache Airflow**, and **Apache Hive** to **process TBs of records daily**, supporting high-throughput ML model training workflows.
- Collaborated with research scientists to productionize LLM features, improving model recall by **9%** and reducing query latency by **35%** using optimized **Presto** and **Spark** transformations.
- Developed interactive dashboards using **Tableau** and advanced **SQL** (**Presto**, **Hive**, **SparkSQL**) to expose model outputs and metrics, decreasing manual query time by **70%** for **50+ internal users**.
- Built and launched an **AI-powered onboarding assistant** for an internal platform, reducing onboarding time by **60%** across **7+ teams** and eliminating over **40** recurring meetings through automated, self-serve documentation guidance.

Infolyze Solutions **June 2024 – August 2024**
Data Engineer - Woodbridge, Virginia, United States (Remote)

- Engineered scalable data infrastructures for 6 client projects using Python, Google Cloud, and AWS, **processing over 10 GB of data daily**. Optimized ETL processes with **Apache Airflow**, reducing data integration time by **30%**.
- Implemented advanced data processing systems using **Apache Spark on Databricks**, enhancing data retrieval speeds by **20%** and reducing latency for real-time analytics.
- Automated data validation on **BigQuery** with Python-driven checks, reducing manual validation by **50%** and increasing client satisfaction by **40%**.

Projects

DataBuff | *Big Data Architecture* **Jan 2025 – May 2025**

- Built an AI-driven web app powered by the **Mistral 14B LLM** for dataset understanding and transformation via natural language. Engineered dynamic code generation (Pandas) and execution pipeline, enabling zero-code data science through prompt-based AI interaction.
- Integrated **GenAI** reasoning into real-time data cleaning, charting, and ML workflows. Optimized LLM prompts and feedback loops to improve model response quality and safety.
- Leveraged **Google Cloud Storage** for dataset and visualization management; deployed on **GKE** with optimized load balancing and security.

MetaPulse | *Independent Project* **April 2025 – May 2025**

- Built a production-style batch pipeline for **1.2M+** NYC Taxi trips using **Airflow (4-stage DAG) to Pandas to PostgreSQL**; packaged in a **4-container** Docker Compose stack for one-command, reproducible deployments with **idempotent upserts**.
- Modeled a **star schema** (fact_trips + **3** dims: vendor, ratecode, datetime) with **date partitioning** and targeted indexes; enabled interactive **Metabase** dashboards over **30/90-day** windows without pre-aggregation.
- Implemented **data-quality checks** (row-count parity, NULL/type coercion, duplicate detection) and task-level alerts in Airflow to prevent bad loads and make failures observable end-to-end.

AI Community & Leadership

Artificial Intelligence Community of Pakistan **June 2023 – August 2024**
President Academics

- Led a team of **30+** across outreach, curriculum, and platform operations.
- Designed and hosted **3** national coding challenges focused on ML/data challenges (**250+ participants**).