GAURAV GANESH ACHARYA

Chicago, Illinois | +1 773-681-2485 | gauravacharya511@gmail.com | LinkedIn | GitHub

SUMMARY

Data Engineer specializing in building scalable data pipelines and distributed systems using Python, SQL, Kafka, and Spark. Experienced in real-time data processing, ETL pipeline development, and cloud-based deployment on GCP. Built end-to-end data systems handling high-throughput workloads and enabling analytics at scale.

EDUCATION

Illinois Institute of Technology, Chicago, IL September 2024 - May 2026
Master of Science, Computer Science

Atharva College of Engineering, Mumbai University, Mumbai, India April 2018 - May 2022
Bachelor of Technology, Information Technology

SKILLS

- **Languages:** Python, SQL, Scala
- **Data Engineering:** Apache Spark, Kafka, Airflow, dbt, ETL/ELT Pipelines
- **Streaming & Real-Time:** Spark Structured Streaming, Kafka Streams, Event-Driven Architecture
- **Data Warehousing:** Snowflake, BigQuery, Redshift, PostgreSQL, MongoDB
- **Lakehouse & Storage:** Delta Lake, Apache Iceberg, Parquet, Medallion Architecture
- **Cloud Platforms:** AWS (S3, Glue, EMR, Lambda), GCP (BigQuery, Pub/Sub, Dataflow)
- **Data Quality & Governance:** Great Expectations, Data Validation, Data Lineage, HIPAA Compliance
- **Concepts:** Data Modeling, Distributed Systems, Partitioning, Schema Evolution, Query Optimization
- **DevOps & Infra:** Docker, Kubernetes, CI/CD (GitHub Actions), Terraform
- **Tools:** Git, Tableau, Grafana

WORK EXPERIENCE

Data Analyst
MERKLE Inc, Mumbai June 2022 - July 2024

- Designed and optimized data ingestion pipelines processing 10,000+ survey responses across 50+ projects using Python and SQL
- Built scalable ETL workflows to clean, transform, and validate datasets, improving data reliability and usability for analytics
- Developed structured data models and reporting pipelines, enabling a 15% improvement in client decision-making efficiency
- Standardized data capture systems using Decipher, Qualtrics, Confirmit, increasing engagement by 50% and ensuring consistent data collection
- Implemented data validation and quality checks to maintain accuracy and consistency across multiple data sources
- Partnered with analysts and stakeholders to deliver high-quality, analytics-ready datasets under tight deadlines.

PROJECTS

Real-Time Ad Analytics Pipeline March 2025 - April 2025

- Architected a real-time data pipeline using Apache Kafka to ingest and process high-volume user and ad event streams.
- Developed distributed data processing jobs with Apache Spark (PySpark) for scalable streaming and batch transformations.
- Deployed a containerized, multi-service pipeline using Docker Compose, ensuring environment consistency and rapid setup.
- Designed and implemented data storage in BigQuery and PostgreSQL, enabling low-latency analytical queries.
- Improved query performance through partitioning and schema optimization, reducing processing overhead and execution time.
- Delivered near real-time analytics capabilities for campaign performance monitoring and data-driven decision-making.

GitHub Issues Forecasting System March 2025 - April 2025

- Architected an end-to-end data pipeline leveraging GitHub APIs to ingest and process data across multiple repositories, enabling real-time analytics and reporting.
- Built high-throughput Flask microservices handling 1K+ daily requests with ~95% uptime, optimizing API performance and fault tolerance.
- Deployed containerized pipelines using Docker on GCP, improving scalability and reducing environment inconsistencies.
- Automated CI/CD workflows with GitHub Actions, cutting deployment downtime by 60% and accelerating iteration cycles.
- Designed modular ETL pipelines powering analytics, forecasting models, and visualization systems.

Dog Breed Prediction System September 2021 - February 2022

- Designed and implanted a CNN model with 97% accuracy, classifying 120+ dog breeds.
- Evaluated a dataset of over 20,000 dog images from Stanford University to enhance model performance.
- Developed a mobile application allowing user to process dog images in real-time, delivering breed predictions and detailed information on average height, weight, lifespan and other attributes in less than 3-sec display time.
- Optimized a new integrated web application and experienced a hike in user interaction by 25%.

Online Voting System September 2020 - December 2020

- Created a web-based voting platform assisting to host elections for over 1,500 students and college events.
- Implemented an encapsulated voting platform utilizing C and Java programming, HTML and CSS for frontend design, and MySQL for a database with over 10,000 records, resulting in a 30% increase in voter participation.

CERTIFICATIONS AND ACHIVEMENTS

- **WORK**: MVP award at Merkle Inc.
- IBM Data Engineering, Google Data Analytics, Meta Data Analytics, IBM Data Analytics, IBM z/OS Mainframe Practitioner Professional Certificate, Python for Everybody, C Programming, Demystifying IOT Security for Digital Word