ADITYA DUBEY

✉ adi.dubey552@gmail.com | 📞 (312) 918-8779 | in LinkedIn | ⬤ GitHub

SUMMARY

Creator of MyGol.ai | Data Engineer | Cloud Architect Certified | AI enthusiast
Innovative Data Engineer (**GCP/AWS/Snowflake**) with **5+** years of experience building scalable, cloud-native data platforms, advanced **ETL/ELT**, and **APIs**. Drove **$100K/month** BigQuery savings, cut delivery **SLAs by 87%**, and shipped **GDPR/CCPA**-compliant data services, Recognized for architecting high-performance, cost-optimized data ecosystems that power business growth and innovation.

CERTIFICATIONS

- Google Cloud Certified – **Professional Cloud Architect**.
- Certified ScrumMaster® (CSM)

SKILLS

Data Engineering: Google BigQuery (GBQ), Snowflake, Cleanroom, Airflow, Spark, Kafka, dbt, Redshift, AWS Glue, AWS Lambda, AWS S3, Tableau, Redis rate limiter, GCP-SDK, websockets, Rest API

Cloud: Google BigQuery, Cloud SQL, Cloud Run, Cloud Functions, Dataflow, Pub/Sub, GCS, Google API Gateway, AWS

Databases: PostgreSQL, MySQL, MongoDB, Cassandra, Firebase, Redshift, NoSQL

Programming Languages: Python, Go, Java, Scala, C, C++, TypeScript, JavaScript

Frameworks & APIs: FastAPI, Flask, OpenAI, RAG, React, Next.js, Streamlit

CI/CD & DevOps: Docker, GitHub Actions, Terraform, Vercel, dbt CI/CD, Agile, Git

Monitoring & Observability: Grafana, logz.io, Google Cloud Tracing & alerts

AI & ML: LLM (OpenAI,Gemini,Claude), RAG pipelines (LangChain, LlamaIndex), Vector Databases (Pinecone), Vertex AI, Agentic AI, NLP, Prompt engineering & guardrails (e.g., hallucination mitigation)

EXPERIENCE

Data Engineer   *Feb 2021 - Present*
Viant Technology, Chicago

- Integrated an **AI-powered** LLM-based **chatbot** to automatically generate targeted audience segments from natural language prompts, reducing manual segmentation effort by **90%,** accelerating campaign launch times by **85%,** and driving a **15%** increase in targeted campaign adoption across the platform, did hallucination mitigation.
- **Solely designed and implemented** the **CRM Delivery Platform** for secure PII data handling — including multi-source data collection, encryption + salting, and chunked delivery to **SFTP**, **GCS**, and **S3** destinations. Reduced delivery time from **2 days to 6 hours**, improving client turnaround by **>87%**.
- **Architected and developed** a high-performance **First Party Segment API** on **Google Cloud Run**, integrating **Google API Gateway**, **OpenTelemetry tracing**, and async workflows (Opt-Out, 1st Party) to ensure scalable, cost-efficient, and observable operations.
- **Optimized API latency** by implementing pre-validation logic at the gateway layer, resulting in significantly faster response times.
- **Built E2E automated testing** integrated with **GitHub Actions**, ensuring deployment stability and reducing release issues.
- **Built and deployed GDPR- and CCPA-compliant opt-out, audience creation, and expiration endpoints**, strengthening data privacy safeguards.
- **Implemented Snowflake Cleanroom** for secure, privacy-preserving data collaboration, integrating auto-generated CRM delivery templates into the UI — reducing campaign setup time by **80%.**
- Proposed and built a **centralized email spine & profile graph** by aggregating and linking user identities (emails, IPs) from multiple data marketplaces, increasing data accuracy by **30%** and boosting customer satisfaction.
- **Optimized BigQuery workloads** and developed advanced SQL stored procedures for IMP platform, saving **$100K/month** in query costs.
- Integrated and processed data from **30+ external data partners** (terabytes-scale) with automated cleaning, refinement, and caching (Java serialized HashMap) for faster UI delivery.
- **Built CI/CD ETL pipelines** and automated data refreshes, improving deployment speed and reliability.
- Migrated workloads from cloud instances to **Google Cloud Run** for cost efficiency and scalability.
- **Developed and automated Bash scripts with Cron jobs** to schedule and orchestrate ETL pipelines.

Data Engineer Fellow - Live Demo GitHub   *Sep 2020 - Jan2021*
Insight, Boston

- **Designed and implemented** a distributed **data analytics pipeline** to detect and analyze traffic congestion, delivering actionable business intelligence that informed **city-wide traffic optimization strategies**.
- **Ingested and processed 60GB+ of historical traffic data** using **Apache Spark**, integrating real-time and static datasets (traffic signal timings, incident reports) to generate comprehensive congestion insights.
- **Engineered a data warehouse in Amazon Redshift**, optimizing schema design and query execution; reduced query processing time by **45%** through efficient indexing and multi-level LEFT JOIN strategies.
- **Developed a Tableau dashboard** to visualize traffic trends and analyze the impact of variables such as **weather conditions, crash incidents, and peak hours** on congestion patterns, enabling data-driven policy decisions.
- **Created and deployed a Traffic Density Classification algorithm** that scored and categorized congestion into **3 severity levels**, improving response planning for high-traffic zones and events.

Research Project   *Jun 2020 - Aug 2020*
Illinois Institute of Technology, Chicago

- **Integrated Fitbit and Apple Watch step-tracking data** into a **Fantasy Sports League web application**, enhancing user engagement by gamifying real-world fitness activity for sports enthusiasts.
- **Developed the web application** using **Python** and **Flask**, ensuring a responsive, user-friendly interface and scalable backend architecture.
- **Led a cross-functional team** to design and implement a **cost-effective data pipeline**, optimizing ingestion, transformation, and storage of wearable device data.
- **Conducted detailed data analysis** to evaluate user engagement metrics, uncovering behavioral trends that informed new feature development and boosted platform retention rates.

Graduate Teaching Assistant   *Aug 2020 - Dec 2020*
Illinois Institute of Technology, Chicago

- **Collaborated with faculty** to design curriculum, lab exercises, and assignments for a graduate-level **Database Security** course, enhancing practical, hands-on learning outcomes.
- **Provided technical guidance** to students by resolving complex database security queries and troubleshooting issues during lab sessions.
- **Evaluated and graded** lab assignments, offering constructive feedback to improve student understanding and application of database security concepts.

PROJECTS

- **Earthquake Prediction**(Big Data Pipeline-PySpark, MLlib, Bokeh, Flask,) **-GitHub**
- **Crop Yield Prediction** (Regression, Python)- **GitHub- Report, code**
- **COVID-19 Analysis** (PowerBI, Tableau)- **GitHub- Dashboards**
- **Rossman Sales Prediction** (R, Predictive modelling, MS Excel) - **GitHub- Report, code**

EDUCATION

Master of Science, Illinois Institute of Technology   *Aug 2019 - May 2021*
Information Technology Management – **Data Analytics**   **GPA: 3.9**
Coursework: Big Data, Database Management, Data Warehousing, Data Analytics, Data Mining

---

Bachelor of Engineering, SGSITS   *Aug 2014 - May 2018*
Computer Science   **GPA: 3.9**