AYUSH KUMAR SAHU

+44 7405 813318 | ayushkusahuk@gmail.com | LinkedIn | GitHub | Portfolio

SUMMARY

Data Engineer with an MSc in Big Data Science (Distinction) from Queen Mary University of London and Azure AZ-104 certification. I build data pipelines, real-time streaming systems, and AI applications using Python, PySpark, Databricks, Kafka, and Kubernetes. I like solving messy data problems and turning them into clean, reliable systems that people can actually use.

EXPERIENCE

Data Engineer | ByteIQ Analytics (Data Analytics Startup) remote *Jan 2026 – Present*

- Set up and managed cloud infrastructure on Azure and NIC Cloud for 10+ services. Built CI/CD pipelines with automated tests and rollback so deployments don't break anything. We haven't had a single failed release, and uptime has stayed at 99.9%.
- Wrote Python ETL scripts using Pandas and SQLAlchemy to move and sync over 500K records between MySQL and PostgreSQL. Made sure everything followed proper 3NF structure and handled transactions cleanly across multiple tables. Data issues dropped to zero after this.
- Took over the analytics side on Databricks. Built ELT pipelines using PySpark and dbt so the team could run SQL transformations with proper version control. Added partitioning and Z-order indexing, which brought query times down by about 40% and saved on compute costs.

PROJECTS

BrandGuardian — AI Brand Compliance Agent [GitHub]
LangGraph **Azure OpenAI (GPT-4o)** **Azure AI Search** **FastAPI** **Streamlit**

- Built a tool that checks if marketing content follows brand guidelines. It reads brand PDFs, stores them in Azure AI Search as vectors, and uses GPT-4o to compare submitted content against those guidelines. It flags violations and gives each one a severity score.
- Used LangGraph to set up different stages the system first analyses the content, then decides whether to escalate or approve it, all depending on how serious the issues are. You can plug in different brand rules without touching the code.
- The backend runs on FastAPI and the frontend is a Streamlit app where users upload content, see results on a dashboard, and download compliance reports as PDFs.

End-to-End Flight Analytics Pipeline [GitHub]
Databricks **PySpark** **Delta Live Tables** **dbt** **Medallion Architecture**

- Built a full data pipeline on Databricks following Bronze → Silver → Gold layers. Raw flight, passenger, and airport CSVs (200K+ records) come in through Autoloader, get cleaned and validated in Delta Live Tables, and end up as a proper star-schema in the Gold layer using dbt.
- Set up incremental loads so the pipeline only picks up new data, and used CDC to keep dimension tables up to date in near real-time. The whole thing runs end-to-end in under 8 minutes.
- Everything is scheduled through Databricks Jobs with retry logic and alerts if something fails. All notebooks are synced to GitHub so the team can track changes and roll back if needed.

Real-Time Stock Market Analytics Platform [GitHub]
Apache Kafka **PySpark Streaming** **PostgreSQL** **Docker** **Grafana**

- Wrote a Kafka producer that pulls live stock prices from a financial API and pushes them into partitioned topics at around 5K messages per second. On the other end, a PySpark Streaming job picks them up and calculates indicators like moving averages, RSI, and VWAP in near real-time.
- Results go into PostgreSQL, and I built Grafana dashboards on top that refresh automatically and send alerts when prices hit certain thresholds. Good for spotting trends and anomalies quickly.
- Packaged the whole setup — Kafka, Zookeeper, Spark, Postgres, Grafana — into Docker Compose so anyone can spin it up with one command locally or on a server.

Kubernetes Microservices Deployment Pipeline
Docker **Kubernetes** **Helm** **GitHub Actions** **Prometheus** **Grafana**

- Created a CI/CD setup where each microservice gets its own multi-stage Dockerfile (cut image sizes by about 60%) and Helm chart. Deployments go through rolling updates with health checks and autoscaling, so nothing goes down during releases.
- Hooked it all into GitHub Actions pushing to main automatically builds the image, runs tests, pushes to the registry, and deploys to the cluster. Set up Prometheus and Grafana to track pod health, CPU/memory usage, and response times with alerts if anything looks off.

TECHNICAL SKILLS

Cloud & DevOps: Azure (AZ-104), Databricks, Docker, Kubernetes, Helm, GitHub Actions, CI/CD, Prometheus, Grafana
Data & AI: Python, PySpark, SQL, Kafka, Delta Lake, DLT, dbt, Pandas, SQLAlchemy, LangGraph, LangChain, Azure OpenAI, RAG, FastAPI, Streamlit
Databases: PostgreSQL, MySQL, Delta Lake, Azure AI Search | **Other:** Git, Medallion Architecture, Star Schema, ETL/ELT, Data Quality

EDUCATION

MSc Big Data Science — Distinction — Queen Mary University of London (Russell Group) *Sep 2024 – Sep 2025*
B.Tech Electronics & Communication Engineering — First Class — SRM, Chennai *May 2019 – Dec 2023*

CERTIFICATIONS

Microsoft Azure Administrator (AZ-104) | Google Data Analytics Professional Certificate (Coursera) | SQL Associate (DataCamp)