

AYUSH KUMAR SAHU

Data Engineer | 2+ Years of Experience | MSc Big Data Science (Distinction)

United Kingdom (Remote) | +44 7405813318 | ayushkusahuk@gmail.com | LinkedIn | GitHub | Portfolio
Graduate visa valid until November 2027 · No sponsorship required

PROFILE

Data engineer with 2+ years across a healthcare diagnostics marketplace and an SQL Server to AWS warehouse migration, plus an **MSc in Big Data Science (Distinction)**. Builds CDC and ETL/ELT pipelines with AWS DMS, S3, Airflow, dbt and Python, owning data modelling, validation and clear documentation end to end.

EDUCATION

MSc Big Data Science (Distinction), Queen Mary University of London

Sep 2024 to Sep 2025

TECHNICAL SKILLS

Languages & Processing: Python (pandas, PySpark, pyodbc, pyarrow), SQL, T-SQL, Spark SQL, ETL and ELT
Change Data Capture: AWS DMS, LSN based CDC, incremental extraction, watermarking, idempotent loads
Data Modelling: dimensional and star schemas, medallion architecture, SCD, surrogate keys, dbt, data quality
Databases & Warehouses: Microsoft SQL Server, PostgreSQL, MySQL, Amazon RDS, Redshift, Snowflake
Cloud & Orchestration: AWS (S3, DMS, EC2, IAM), Azure, Apache Airflow, Databricks, Docker, LocalStack, Git
Streaming, AI & BI: Kafka, Spark Structured Streaming, Delta Lake, LangGraph, Azure OpenAI, Power BI

EXPERIENCE

Data Engineer, BytelQ Analytics, Remote

Jan 2026 to Present

- Designed and built the ingestion layer for a global US relocation services company's SQL Server to AWS warehouse migration (operational and reporting sources), helping shape the target architecture and proposing Airflow for orchestration.
- Engineered a CDC pipeline on SQL Server with AWS DMS, running full load plus ongoing replication into an S3 bronze layer as partitioned Parquet, delivered with a least-privilege reader, runbook and security report.
- Developed a Python CDC extractor with pyodbc and pyarrow that reads SQL Server change tables by LSN range via `fn_cdc_get_all_changes`, writing Parquet in the AWS DMS S3 folder layout.
- Created a config-driven extraction framework where a YAML manifest maps each table to one of four strategies (rowversion, datetime, identity and snapshot hash), with per-table control state for resumable loads.
- Automated loading with Apache Airflow across two DAGs and seven tasks, using an idempotent file-level control table and validation that fails the run on any row-count mismatch between staging and the control log.
- Merged two disjoint source systems (53,860 rows across 24 tables) into a unified SQL Server warehouse with 18 dbt models, recommending hash-based surrogate keys after analysis showed no shared keys.
- Reverse-engineered the client's legacy warehouse, mapping its hub-centred star schema (30 dimensions, 27 fact tables, 95 stored procedures), and authored a 20 section platform reference plus onboarding guide.

Data Engineer, EVE Healthcare, Gurugram, India

Jan 2023 to Aug 2024

- Owned the data pipelines and analytics layer behind a Delhi NCR diagnostics marketplace spanning 3 cities and 25+ diagnostic tests, building Python and SQL ETL/ELT that standardises test catalogues, pricing and slot availability from partner centres.
- Modelled the marketplace schema covering centres, services, pricing, panels and bookings across 7 service categories (MRI, CT, X-ray, ultrasound, blood tests, cardiology and neurology) and 3 empanelment types (CGHS, ECHS and corporate).
- Orchestrated ingestion with Apache Airflow, designing DAGs with scheduling, monitoring and automated retries so pricing and slot availability stayed current across partner centres.
- Built an AWS platform with S3 as the data lake and Amazon Redshift as the warehouse, giving product and business teams visibility into bookings, conversions, top tests and centre performance.
- Implemented data-quality and validation checks behind a price-comparison experience advertising savings of up to 50%, supporting search, near-me discovery and same-day report delivery.

PROJECTS

End to End Aviation Data Pipeline Databricks, PySpark, Delta Lake, Auto Loader, dbt

GitHub

- Built a Databricks lakehouse for flight, passenger and airport data with incremental Auto Loader ingestion, schema drift handling and Bronze, Silver and Gold layers.
- Modelled a Gold star schema with fact_bookings and supporting dimensions, applying deduplication and SCD upserts for queries like airline revenue by departure city.

Real Time Stock Market Pipeline Kafka, Airflow, Snowflake, dbt, Docker, Power BI

GitHub

- Streamed Finnhub API events into Kafka, with Airflow DAGs landing raw JSON in MinIO and loading Snowflake through internal stages and COPY INTO.
- Modelled dbt Silver and Gold layers for parsing, deduplication and KPI and volatility aggregation, feeding live Power BI DirectQuery dashboards from a Dockerised stack.

BrandGuardian, AI Compliance Auditor LangGraph, Azure OpenAI, Azure Video Indexer, FastAPI

GitHub

- Built a multi-agent LangGraph workflow that ingests YouTube videos, extracts transcripts and on-screen text with Azure Video Indexer, and flags compliance issues with GPT-4o RAG behind a FastAPI endpoint.

CERTIFICATIONS

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