YASHI SHUKLA

Data Engineer | AI Engineer | Business Intelligence & Analytics

(+91)-9810824347 | yashishkl1@gmail.com | yashi-shukla | yashi-shukla | yashi-shukla.github.io

ABOUT ME

5 years of experience in data science, engineering, and analytics across government, nonprofit, and private sector projects. Skilled in designing and deploying cloud-native architectures (AWS, GCP, Azure), ETL/ELT pipelines, and AI/ML workflows at scale. Experienced in LLM-based information extraction and MLOps for production deployment. Built GenAI and business intelligence solutions handling millions of records to drive data-driven decisions. Adept at stakeholder management and cross-functional collaboration, delivering solutions that maximize impact.

TECHNICAL SKILLS

Cloud Platforms: AWS (Lambda, Step Functions, Redshift, Athena, Glue, Fargate, ECS, SAM), GCP (BigQuery, Cloud Functions, Gemini, AI Studio, Vertex AI, Looker Studio), Azure (VM Clusters, Azure Kubernetes Service, Data Factory, Databricks, Automation Runbooks)
Languages: Python (Pandas, NumPy, FastAPI, Flask, Google Genai, LangChain, OpenCV, TensorFlow, scikit-learn, Django), SQL, Java, JavaScript
Data, Analytics & AI Tools: Airflow, dbt, Airbyte, Apache Spark, Hevo, Directus, Tableau, Power BI, Superset, LLM fine-tuning (Gemini, GPT), NLP
DevOps/MLOps, Webscraping & Testing: Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, Cypress, Selenium, BeautifulSoup, Postman
Databases/ Data Standards: PostgreSQL, MySQL, Redshift, Athena, BigQuery, Statistical Data and Metadata Exchange (SDMX), FMR

PROFESSIONAL EXPERIENCE

Data Engineer Consultant, World Bank Group, New Delhi

Nov'25 – Present

Working with the Data Science and Engineering teams within the Development Data Group, contributing to Data360 initiatives focused on building scalable data pipelines, analytics, and data products to support global development insights.

- **Data 360 Platform, WBG Product**

- Migrating multiple data pipelines from Kedro to **Databricks** on Azure to improve scalability and maintainability, while standardizing datasets using **SDMX** and **FMR** frameworks under the Data360 platform.
- Collaborated on data engineering efforts for the GAFS Dashboard by developing a scalable **Python**-based system that integrates diverse data sources and APIs, enabling near real-time insights for global food security monitoring.

Data Engineer, IDinsight, New Delhi

Jan'22 – Aug'25

Worked on 10+ workstreams involving large-scale data ingestion, information extraction, MLOps workflows, and dashboarding solutions.

- **Inclusive Financial System, Gates Foundation, India**

- Designed and deployed a **GCP-based pipeline integrating Google's Gemini multimodal LLM** to automate OCR and extract structured data from **10M+ PDFs**, generating high-quality datasets that accelerated a national survey sampling effort.
- Built **autonomous agents** using Gemini's vision-language capabilities for PDF classification, table extraction, and metadata generation, **cutting manual preprocessing time** and expediting survey frame readiness.
- Developed a **Python-based transliteration module** with Gemini APIs to convert addresses from **8 Indian languages into English**, improving surveyor navigation in multilingual regions.
- Leveraged **Gemini 2.5-Flash** for structured PDF outputs, performing **prompt engineering** to refine accuracy and **adding guardrails** for quality control, ensuring reliability of extracted data at scale.

- **National Data Analytics Platform (NDAP) Monitoring Dashboard, Niti Aayog, India**

- Led the design and implementation of a **Redshift**-based architecture to ingest and monitor terabytes of clickstream data, enabling real-time analytics on platform engagement and SLA compliance.
- Defined key indicators and built synchronization workflows, launching a **Superset** dashboard that improved visibility into platform usage for senior government stakeholders.
- Acted as a technical consultant for NDAP, advising external teams on architecture and analytics strategy to ensure scalability and impact.

- **Social Cash Transfer Programme (SCTP) Monitoring Dashboard, Govt. of Malawi, along with UNICEF Malawi**

- Automated raw data transformation workflows using **Python** and **Google Data Studio**, reducing manual update time by 25%.
- Transitioned data loading from manual entry to API-based integration, enhancing data accuracy and laying the groundwork for future phases.
- Co-developed dashboard solutions used by bureaucrats to track cash transfer distribution, strengthening government oversight of the schemes.

- **Indus Action, Non-Profit, India**

- Facilitated an ideation workshop with the vision of expanding government benefit access to 3M+ workers, resulting in three strategic workstreams to guide future implementation.
- Built scalable data pipelines with **AWS Redshift** and **Glue** to process millions of beneficiary records, and developed an **Athena**-based backend for real-time outreach and claim tracking dashboards.

- **Praekelt, Non-Profit, South Africa**

- Developed **SQL data models** and a **FastAPI**-based backend for an Ask-Me-Anything chatbot, reducing frontline support workload by 70%.
- Built a secure data API to enable controlled sharing of sensitive data across different organizational levels.

- **Aspirational Districts Programme (ADP), Niti Aayog, India**

- Created a tool that generated mismatch reports across data collected from different levels, improving visibility into data issues for administrators.
- Designed and developed a low-cost serverless architecture using **AWS Lambda** and **Step Functions** with a monthly cost of less than $20.

- **Kidogo, Non-Profit, Kenya**

- Contributed to a monitoring system for analyzing daycare quality in Kenya by developing a solution for a digital data collection setup using Kobo.

- Created an easy-to-use low-code monitoring solution using **Hevo** and GCP **BigQuery** to enable non-tech users to maintain the system.

- **Uttar Pradesh Data and Evidence Support, U.P. State Govt. India**

- Performed web scraping across multiple websites to support data requests for M&E activities, reducing client-facing working hours by 40%.
- Automated web scraping using **AWS Fargate** and **ECS** to update a dashboard created on **Looker** to analyze budget and expenditure data.

- **Social Justice Movement (SJM), Non-Profit, Kenya**

- Developed a digital data collection system using **Kobo Toolbox**, enabling the organization to transition from paper-based surveys to a centralized, digitized data repository.
- Designed and deployed a low-code architecture using **Directus** on GCP, allowing SJM members to collect key metrics, manage user access, and maintain historical data with minimal technical overhead.

- **Surveystream, IDinsight Product**

- Developed an end-to-end regression testing suite using **Cypress**, reducing manual QA efforts by 30% for the Surveystream web app.
- Designed and developed a pipeline to automate surveyor email notifications, enabling cluster coordinators to efficiently share assignment details, data quality, productivity, and financial reports.
- Refactored legacy email generation workflows in **Airflow** on **Docker** by eliminating redundant DAGs and introducing a modular, config-driven architecture to support scalable email automation.

- **IDinsight Projects (Internal)**

- *Data Development Platform:* Built an automated reporting system using **dbt** and **Airbyte** on GCP's Dalgo framework, paired with a Looker dashboard, reducing monthly finance reporting time by 30%.
- *Satellite Imagery (MOSAIKS):* Optimized MOSAIKS grid sampling by providing **MLOps**-driven infrastructure with **Dask** on **Azure Kubernetes** Service, achieving 10x faster runtime performance.

Analyst, Lentra.ai, Bengaluru *Oct '20 – Dec'21*

- **Bharat Petroleum Corporation Limited, India**

- Designed the architecture to implement observability in the pipeline and conducted a POC on the data coming from the dashboard.
- Conducted website functional testing and API testing using Postman, ensuring reliable system performance across critical applications.
- Validated **Power BI** metrics with Python scripts, strengthening report accuracy and reliability.
- Automated duplicate data analysis using **Azure Data Factory** and **Automation Runbooks**, reducing manual validation time.
- Built unit tests for **ETL Spark Streaming** jobs and developed an observability platform, improving data quality checks at scale.

Associate QA Analyst, Xceedance, Gurugram - *Internship* *Jan '20 – Jun '20*

- **Hollard and Blue Zebra – Australian Insurance groups**

- Worked on two projects and maintained website quality, following **Agile** methodology, and ensured data integrity through regular back-end testing using **SQL**.
- Replicated and logged bugs as well as performed regression testing using **Azure DevOps**, reducing production defects.
- Automated end-to-end sanity flow using **Selenium** in Python, saving ~64 hours of manual QA work per month.

Data Analyst, Alpha Nodus, Bengaluru - *Internship* *May'19 – Jul'19*

- **Focused on Healthcare groups in the US**

- Completed **HIPAA** training and implemented encryption practices to safeguard PII, ensuring compliance with healthcare data regulations.
- Automated appointment scheduling and patient notifications to increase utilization of available time slots and to improve efficiency.
- Developed six statistical metrics in **Tableau** for trend analysis, creating impactful dashboards that supported clinical decision-making.
- Analyzed queue formation and patient prioritization trends using **PostgreSQL** and Python (**Pandas**), reducing manual workload by 50% and increasing patient throughput by 10%.

EDUCATION

Bennett University, Greater Noida | B. Tech, Computer Science Engineering | CGPA: 8.35 /10, GPA: 3.6/4 *Aug '16 - Jun '20*

- **Projects**

- **Sound Recognition System** *(Capstone Project)* - Built a CNN-based audio classification system that transformed input sounds into spectrogram features and classified them into 50 categories, achieving 89.6% accuracy. Designed scalable preprocessing and feature extraction pipelines for large audio datasets, and validated performance using standard evaluation metrics, demonstrating an end-to-end ML workflow.
- **Video Summarisation** - Implemented a YOLO-based video summarisation model that reduced video size by ~83% while preserving key content, demonstrating efficient computer vision techniques for data compression without loss of critical information.
- **Engineered Image Synthesizer** - Built an artificial image generation pipeline using OpenCV to expand a dataset from 60 real images to 50,000 augmented samples, improving CNN model performance by providing diverse, real-world training scenarios.

- **Relevant Coursework:** Big Data Analytics and Business Intelligence, Artificial Intelligence and Machine Learning, Probability and Statistics, Information Management Systems, Deep Learning, Engineering Calculus, Linear Algebra, and Ordinary Differential Equations.

CERTIFICATIONS

- **Microsoft Certified:** Azure Fundamentals
- **Quantium Virtual Experience Program:** Data Analytics
- **Databricks:** i) Developer Essentials, ii) Developer Foundations
- **Coursera:** i) Data Processing and Collection with Python, ii) Intro to TensorFlow (GCP), iii) Building a Web Application in Django