ABINASH ACHARYA

Software (Data) Engineer

Kathmandu, Nepal · acharyabinash@gmail.com · (+977) 9846856324

Summary

Software engineer specializing in distributed systems, high-performance architectures, and large-scale data extraction. I build scalable Go backends and production-grade Python data pipelines, and stay current with high-productivity AI workflows that compress iteration cycles for backend and data work. I'm motivated by problems where engineering, data, and life sciences converge, work that is technically challenging and creates real-world impact. I gravitate toward roles where I can translate executive vision into concrete solutions and resolve technical bottlenecks. I prefer composing proven primitives over rebuilding what already works.

Experience

Data Engineer, Kavaya Research | 108 Capital Pte. Ltd *Jan/2025–Now*

Go · Python · PySpark · Kubernetes · Seleniumbase · Camoufox · Clickhouse · PostgreSQL · Iceberg · Polaris

Fast-paced research-driven quant startup; small high-ownership team with weekly to bi-weekly iteration cycles.
End goal: turn applied research into revenue, requiring rapid prototyping and production hardening in tight loops.

- **Scalable Go Backend Systems:** Built self-healing, stateless Go services on NATS JetStream with gRPC/REST/NATS interfaces, delivering high-throughput ingestion backed by distributed tracing, metrics, and logs via OpenTelemetry and OpenObserve.
- **Blockchain Data at Scale:** Engineered ingestion and transformation pipelines for Base chain on-chain data, processing high-volume blockchain events for research-driven trading and market analytics.
- **Bot Orchestration at Scale:** Built a distributed, multi-platform automation system coordinating thousands of headless browser instances (Seleniumbase, Camoufox, Patchright) with adaptive fingerprinting, proxy rotation, behavioral modeling, and automated bypass of Turnstile and reCAPTCHA v3 anti-bot systems.
- **ClickHouse-Centric ETL at Billion-Scale:** Developed production-grade ETL workflows in Python (Prefect + dbt, PySpark) on ClickHouse/PostgreSQL supporting billion-scale daily ingestion on VPS infrastructure.
- **Kubernetes-Based ETL Platform:** Designed an ETL flow where job artifacts are rsynced to S3-compatible object storage and Spark jobs are submitted to the cluster via Airflow's SparkKubernetesOperator or Prefect workers, with self-hosted GitLab CI/CD runners on Vultr handling build/deploy.

Software Engineer | Customer Solutions Engineer, Grepsr *Dec/2023–Dec/2024*

PHP · Python · Delta Lake · SQL · AWS

Dedicated web-scraping company serving enterprise clients; client-facing role bridging pre-sales engineering and delivery.
Hybrid culture combining production crawler operations with applied research into lakehouse storage and ML for product intelligence.

- **Large-Scale Web Data Extraction:** Worked at a dedicated web-scraping company, leading technical feasibility analysis for enterprise extraction projects and translating client requirements into crawler architecture, throughput estimates, risk assessments, and delivery strategies.
- **Massively Scalable Crawling Systems:** Engineered and operated 1000+ concurrent scrapers and

crawlers across highly dynamic targets (social platforms, e-commerce, JS-heavy sites), integrating fingerprinting, proxy rotation, and fault recovery for sustained high-volume data acquisition.

- **Delta Lake & Lakehouse Research:** Researched and prototyped Delta Lake-based storage layers for crawler output, evaluating ACID-compliant ingestion, schema evolution, and time-travel queries.
- **OCR-Driven Text ETL Pipelines:** Developed production-grade OCR-to-analytics pipelines using Tesseract, PySpark, MongoDB, and S3 to extract, normalize, and store unstructured document data at scale.

Software Intern, SurTaal *Jun/2023–Sept/2023*

Kotlin · Git

Early-stage audio-tech startup focused on real-time pitch and vocal analysis on mobile.
Lean intern role with direct backend and mobile app ownership and exposure to applied research on audio signals.

- **Backend Latency Optimization:** Designed event-driven microservices with optimized DB access patterns and Redis-backed caching layers, achieving a 30% latency reduction and stable performance under high-concurrency production traffic.
- **Audio Signal Processing & Deep Learning:** Built low-level real-time audio control pipelines using Oboe and investigated Transformer-based models for vocal note prediction, enabling high-resolution pitch tracking and advanced audio feature extraction.

Robotics Volunteer, Robotics Club | Pulchowk Campus, Team Nepal *Oct 2018–Oct 2022*

STM32 · C/C++ · ROS · KiCad

University robotics team representing Nepal at ABU Robocon 2020 (Fiji); cross-disciplinary ownership across firmware, electronics, and navigation.

- **Localization & Embedded Control:** Implemented MCL and EKF for real-time pose estimation and SLAM-based navigation on STM32, alongside motor control and sensor fusion for teleoperated and autonomous robots.
- **Electronics & PCB Design:** Designed custom THT PCBs for power distribution, motor drivers, and sensor interfaces for competition-grade robots.

Projects

Nepali Paraphrase Generation in Devanagari Script *2023*

Undergraduate Major Project · Python · m-BERT · Google Compute · Regex

- **Model Fine-tuning:** Fine-tuned multilingual BERT (m-BERT) transformer for Nepali paraphrase generation, achieving over 80% accuracy on a custom test dataset.
- **Dataset Creation:** Curated and annotated a novel Nepali paraphrase corpus from diverse sources, addressing the low-resource challenge for Devanagari NLP tasks. Corpus creation was done using regex-based cleaning pipelines with pandas to handle Devanagari-specific UTF-8 encoding issues and normalize text for model training.

Real-Time Stream Processing Pipeline *2023*

Python · Kafka · PySpark · dbt · Docker

- **Event Streaming:** Built a fully Dockerized, end-to-end Kafka-based data pipeline for real-time event processing and transformation.
- **Stream Processing:** Implemented PySpark streaming jobs for data transformation and aggregation with dbt for data modeling and quality assurance.

Distributed Batch Processing System *2023*
Python · PySpark · Google Cloud Storage · BigQuery

- **Distributed Computing:** Designed a scalable batch processing system using PySpark on Google Cloud Platform for large-scale data analytics.
- **Data Warehouse Integration:** Orchestrated data pipelines from GCS to BigQuery, enabling efficient querying and business intelligence workflows.

Education

Bachelor of Engineering in Electronics, Communication & Information Engineering *2018–2023*
Tribhuvan University, Pulchowk Campus, Nepal - top-ranked engineering institute in Nepal

Relevant Coursework: Data Science, Big Data Technologies, Numerical Methods, C/C++, Object-Oriented Programming, Data Structure & Algorithm, Artificial Intelligence, Embedded Systems, Digital Signal Analysis & Processing, Biomedical Instrumentation

Skills

- **Languages:** Go, Python, SQL, PHP, Kotlin
- **Data Engineering:** PySpark, Kafka, Prefect, dbt, Airflow, Delta Lake, Iceberg, Polaris, Trino
- **Endpoints:** REST, gRPC, NATS
- **Databases:** PostgreSQL, ClickHouse, MongoDB, Redis, BigQuery
- **Cloud & Infra:** Kubernetes, Docker, GitLab CI/CD, AWS, GCP, Hetzner, Vultr, Cherry Servers
- **Scraping & Automation:** Seleniumbase, Camoufox, Patchright
- **Observability:** OpenTelemetry, OpenObserve, Grafana
- **Embedded & IoT:** MQTT(AWS IoT Core), ROS, Sensor Fusion
- **ML & AI:** PyTorch, ONNX, Transformers

References

Sanjivan Satyal
Assistant Professor, Department of Electronics and Computer Engineering
Institute of Engineering, Pulchowk Campus, Tribhuvan University, Kathmandu, Nepal
Email: sanziwans@gmail.com
Courses Taught: Wireless Communication, Communication Systems, Telecommunications

Pramod Phuyal
Software Engineer, Grepsr, Kathmandu, Nepal
Email: pramod.phuyal@outlook.com
Relevance: Work Colleague, Same Team

Kabish Bhattarai
Data Engineer, Kavaya Research, Remote
Email: kabishbrt@gmail.com
Relevance: Work Colleague, Same Team