Suraj Sharma

https://github.com/SoeRatch | surajs787@gmail.com | 91-7002524460 | https://linkedin.com/in/surajs787

Primary Skills

Python, FastAPI, Django, PostgreSQL, SQL, Data Modeling, System Design, Redis, GCP, AWS

Secondary Skills

Docker, Apache Kafka, Celery, Apache Airflow, Pub/Sub, Cloud Functions, JavaScript, React.js, Node.js

Experience

Turing
LLM Code Evaluator – Python and JavaScript *Oct 2024 - Mar 2025*

- Validated and debugged 600+ code snippets generated by LLMs, directly contributing to a 25% improvement in feedback quality.
- Identified and fixed logic flaws and test inconsistencies, reducing false negatives by 20% during model validation.
- Wrote 800+ ideal code responses and prompts to support model fine-tuning and improve algorithmic accuracy.

Impact Analytics
Lead Software Engineer *Dec 2021 - Oct 2023*

- Designed and launched a scalable pricing engine to automate product price conversions across 12+ countries, factoring foreign exchange (FX) rates, tax structures, and local costs.
- Engineered an event-driven pipeline with GCP Pub/Sub and Cloud Functions to offload 80% of pricing-related logs and calculations from the main application, doubling throughput and reducing DB contention.
- Achieved sub-2s latency for bulk pricing updates across 10K+ SKUs and 12 countries using parallel queues.
- Mentored a team of 5+ engineers, implemented CI/CD practices, and cut deployment rollback incidents by 40%.
- Delivered complex backend services through cross-functional collaboration, contributing to a 5% revenue increase and a 90% reduction in pricing errors.

Impact Analytics
Senior Software Engineer *Nov 2019 - Nov 2021*

- Spearheaded the development of an automated markdown pricing engine using Django and PostgreSQL, increasing retail sales by 12% through strategic discounting.
- Built an image tagging tool with Flask and React, enhancing catalogue accuracy and search relevance by 35%.
- Refactored real-time markdown updates using FastAPI, server-sent events (SSE), and Google Cloud Pub/Sub, reducing latency from 5 minutes to under 2 seconds.
- Created materialized views and high-performance SQL analytics on 10 million inventory and transaction records, improving REST API and query speed, and reducing dashboard latency by 70%.

Tailorman
Full Stack Engineer *Mar 2019 - Jul 2019*

- Developed and maintained backend services using Express.js and PostgreSQL for order and user management, supporting 5K+ monthly transactions.
- Optimized image delivery using AWS S3 and Lambda, reducing average page load time by 35%.
- Enhanced frontend user experience with React + Redux features, decreasing customer-reported issues by 20%.

Kubric (now Mason)
Member of Technical Staff *May 2018 - Dec 2018*

- Created reusable UI components and REST APIs that shortened development cycles and improved performance.
- Built a custom code editor using Ace and React, improving editor load time by 40% and reducing crash rates.

Projects

Disqueue - Minimal Distributed Job Queue System — May 2025 – Present

- Tech Stack: Python, FastAPI, Redis Streams, Docker, Strategy Pattern, System Design
- Architected a lightweight, production-ready job queue leveraging Redis Streams for log-based message processing and FastAPI for job ingestion and tracking endpoints.
- Engineered strict priority scheduling via stream-based segregation and a custom worker polling mechanism to enforce high-to-low job execution order.
- Devised a pluggable retry strategy supporting both fixed and exponential backoff, and integrated a Dead-Letter Queue (DLQ) for permanently failed jobs.
- Designed idempotency and deduplication controls to guarantee exactly-once execution in distributed job workflows.
- Enabled safe job cancellation, graceful shutdowns, and Redis stream offset tracking to support resilient recovery and reprocessing.
- GitHub: https://github.com/SoeRatch/disqueue

Articles:

Tired of Celery? Here's a Minimal Distributed Queue System You Can Actually Understand
Designing a Pluggable Retry Mechanism in Python Job Queues
Why Is My Job Running Twice? Understanding Idempotency and Deduplication

ETL Pipeline for SaaS Logs — Apr 2025 - May 2025

- Tech Stack: Apache Airflow, Docker, PostgreSQL, GitHub Actions, Slack Integration
- Developed a modular ETL pipeline to simulate and process SaaS logs using Apache Airflow.
- Dockerized the setup with Compose and stored processed data in PostgreSQL for durability.
- Implemented CI with GitHub Actions and real-time Slack alerting for DAG/task failures.
- GitHub: https://github.com/SoeRatch/saas-log-etl

Much Story - Scalable Storytelling Platform — Jul 2019 - Oct 2019

- Tech Stack: MongoDB, Express.js, React.js, Node.js, Slate.js
- Led MERN stack development of a collaborative writing tool with server-side rendering, access control, and token-based authentication.
- Integrated Slate.js for rich-text editing with autosave, story versioning, and performance tuning.
- GitHub: https://github.com/SoeRatch/Much-story-ssr-mern

Education

Bangalore Institute of Technology — Bengaluru, India
Bachelor of Engineering (B.E.), Computer Science and Engineering — *Aug 2013 - Jul 2017*

Recognition

Employee of the Quarter – Impact Analytics — May 2022

Awarded for leading the launch of a pricing engine that reduced errors by 90% and improved revenue by 5%.