KEVAL SAKHIYA

Web Scraping Specialist - Distributed Crawling at Scale

+91-9687591750 | kevalsakhiya@gmail.com | linkedin.com/in/kevalsakhiya | github.com/kevalsakhiya | Anand, Gujarat, India

SUMMARY

Python developer and web-scraping specialist with 7+ years building and operating **large-scale distributed crawlers and data-extraction systems** for US companies. Deep Scrapy expertise - spiders, item pipelines, middlewares, and Redis-backed distributed crawling deployed via Scrapyd / Scrapy Cloud / Gerapy - with async and multithreaded architectures tuned for high-throughput, long-running jobs. Proven anti-bot depth (rotating residential/datacenter proxies, TLS/JA3 & browser fingerprinting, CAPTCHA solving, stealth headless automation with Playwright and Selenium) and rigorous data-quality validation, dedup, and normalisation into MongoDB / PostgreSQL. Owns the full path from web-scale crawl to ETL to delivery - large-scale processing with PySpark, orchestration with Apache Airflow, and production monitoring/alerting with Spidermon. Open-source contributor to Scrapy.

CORE COMPETENCIES

Distributed Crawling at Scale • Async & Multithreaded Python • Anti-Bot & Stealth • Data Quality & Validation • Pipeline Monitoring & Alerting • NoSQL (MongoDB) & Redis • Cloud at Scale (AWS) • ETL & PySpark

PROFESSIONAL EXPERIENCE

Turing

Jul 2023 - Feb 2026

Web Scraping Specialist & Data Engineer | Remote (USA) | Full-time contract

— Built and operated **large-scale distributed web crawlers** - production Scrapy with Redis-backed request queues and shared state, tuning concurrency, AutoThrottle, and memory profiles for reliable long-running jobs.

— Engineered **anti-bot strategies at scale** - rotating residential/datacenter proxies, TLS/JA3 and browser-fingerprint handling, UA/header rotation, and CAPTCHA solving (2Captcha, Anti-Captcha).

— Built Scrapy item pipelines for **data-quality validation, deduplication, and normalisation**, with monitoring and alerting via Spidermon - crawl-health checks and alerts on validation failures and coverage drops.

— Processed large-scale datasets with **PySpark and distributed computing** for big-data transformations on scraped sources, alongside Pandas/AWS Glue ETL flows.

— Orchestrated multi-stage pipelines with **Apache Airflow** (retry, logging, error handling) and exposed cleaned datasets through FastAPI / Django REST services.

— Delivered an e-commerce ETL pipeline that cut manual processing time by **60%**.

Heirlift Estate

Mar 2021 - May 2023

Web Scraping Specialist | Remote (USA) | Full-time contract

— Designed and maintained Scrapy spiders to extract **real-estate listings at scale** - property attributes, agent details, price history, and images.

— Handled **paginated, JavaScript-rendered, and AJAX/session-protected pages** using scrapy-playwright and Selenium with stealth configurations; managed login flows and cookie persistence.

— Engineered anti-bot bypass with rotating residential proxies, randomized request fingerprints, and CAPTCHA-solving middleware to sustain reliable extraction.

— Built Scrapy item pipelines for validation, deduplication, and normalisation into **PostgreSQL and MongoDB**.

— Loaded cleaned property data through an **AWS Glue + S3 ETL** flow feeding downstream analytics.

Upwork (Top-Rated)

2019 - 2021

Freelance Python Developer - Web Scraping | Remote

— Built production Scrapy spiders for international clients across e-commerce, B2B directories, and aggregator sites, with structured item pipelines and rotating proxy pools.

— Achieved **Top-Rated** status on Upwork through consistent delivery and client satisfaction across long-running engagements.

— Exposed crawled datasets via FastAPI services and collaborated directly with clients in English to scope requirements and define data schemas.

KEY PROJECTS

County Property & Mortgage Data Pipeline *Distributed crawl | 500k+ records/day | 40+ scrapers*

Harvested public mortgage and property records scattered across 40+ US government county portals into one cleaned, monitored ETL pipeline, with instant alerts the moment a run fails.

Self-Healing Broker-Intelligence Pipeline *Data quality | 15,500+ brokers/week | 95.6% coverage*

Weekly Scrapy system scraping two registries that validates, dedupes, and repairs bad rows mid-run - delivering 15,500+ broker profiles with zero manual cleanup between runs.

Reddit Sentiment Pipeline for ML Signals *NLP | 1M+ posts/day | 2,000+ subreddits*

Apache Airflow ETL pulling 1M+ Reddit posts/day across 2,000+ subreddits, scored with NLP and embedding models and delivered ready for downstream prediction models.

ML-Ready Sports Data Feature Store *AI training data | 3M+ data points | model-ready*

Scrapers across many sports sites feeding a pipeline that cleans, normalises, and vectorises match and player data into a model-ready feature store, auto-refreshed as new matches complete.

EDUCATION

Bachelor of Science (B.Sc.), Chemistry | Gujarat, India

OPEN SOURCE

Scrapy - documentation contribution reviewed & merged into the official `scrapy/scrapy` repository — Open Source

Public scraping projects & utilities — github.com/kevalsakhiya

TECHNICAL SKILLS

Languages: Python (7+ yrs), SQL, JavaScript, Bash

Web Scraping & Crawling: Scrapy, scrapy-playwright, Spidermon, BeautifulSoup, lxml / parsel, Requests, Selenium, Playwright

Distributed & Async: Redis-backed distributed crawling, asyncio, multithreading, concurrency / AutoThrottle tuning, scaling long-running jobs

Anti-Bot & Stealth: Rotating residential/datacenter proxies, CAPTCHA solving (2Captcha, Anti-Captcha), TLS/JA3 & browser-fingerprint handling, headless stealth automation, session/cookie management, UA/header rotation

Web & APIs: HTTP/HTTPS, HTML, CSS selectors, XPath, DOM, AJAX / dynamic content, REST APIs, GraphQL, API reverse engineering, WebSockets

Databases: MongoDB, Redis, PostgreSQL, MySQL

Data Processing: ETL pipeline design, data-quality validation, Pandas, NumPy, PySpark, distributed computing, JSON / XML / CSV

Orchestration & Deployment: Apache Airflow, Docker, Kubernetes, Scrapyd, Scrapy Cloud, Gerapy, cron, Git

Cloud & OS: AWS (EC2, S3, Glue, Lambda), Azure Synapse, Linux/Unix

ML & MLOps: Scikit-learn, TensorFlow, MLflow, model deployment via FastAPI