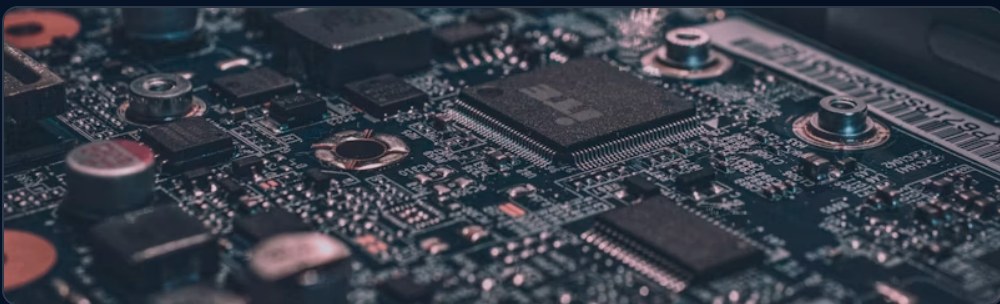




AI Lead Extraction Architecture

Enterprise Technical Specification & Blueprint Document



Architectural Reference Design • Version 2.4

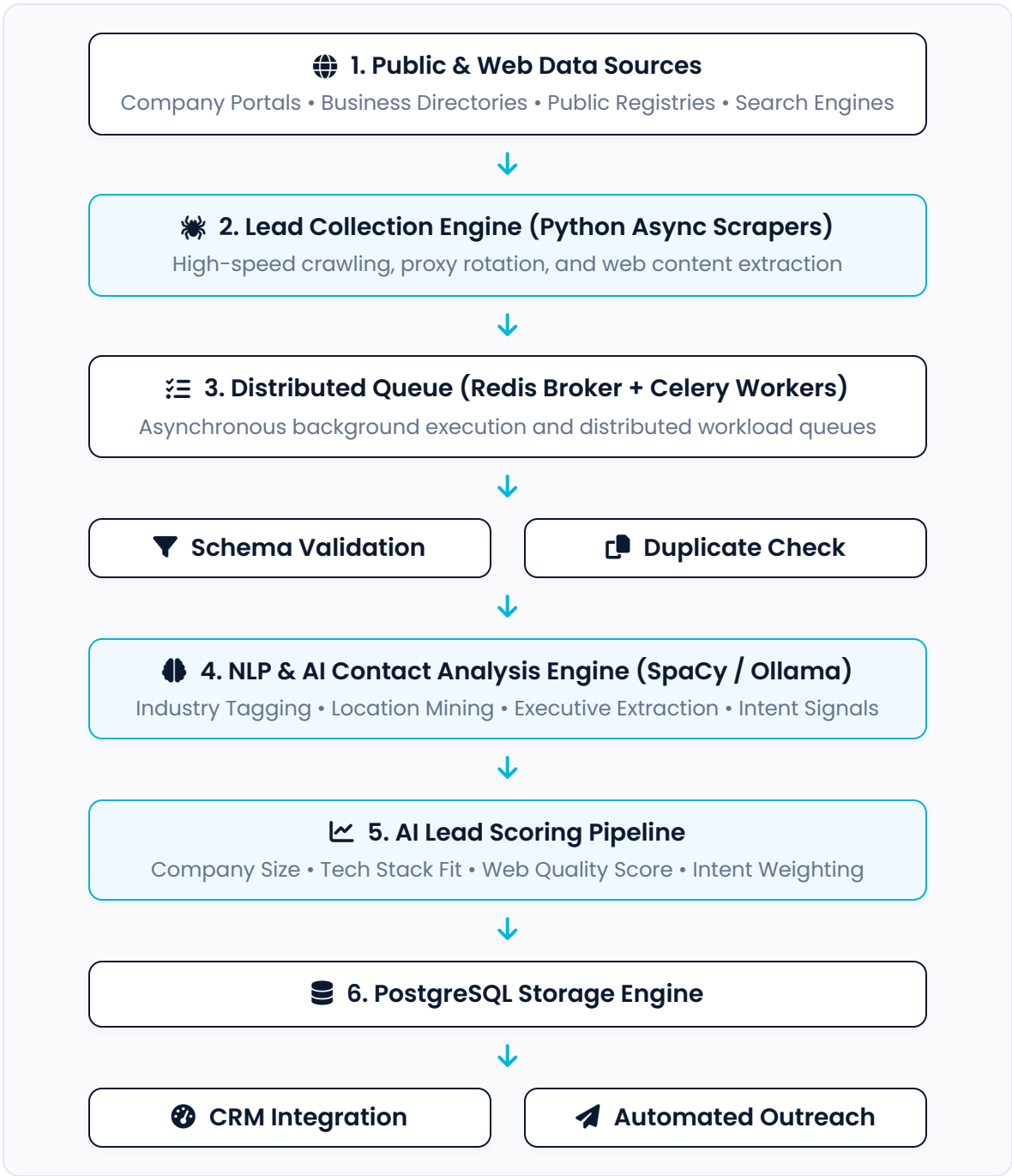
Engineered by Voxora Solutions (voxorasolutions.in)



System Purpose & Pipeline Flow

Executive Summary: The AI Lead Extraction Architecture is an enterprise-grade platform designed to continuously collect B2B data, parse contact metadata using offline LLMs (Ollama) and SpaCy NLP pipelines, score prospect intent, and route high-value leads directly into CRM systems.

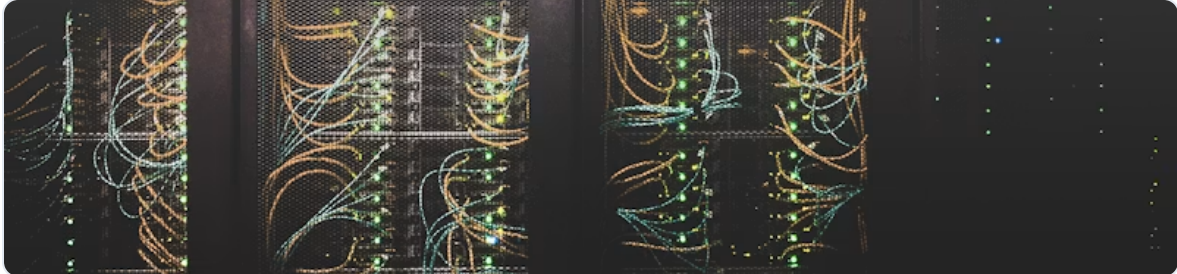
High-Level System Flow





1. Lead Collection Engine & Task Queue

The ingestion layer utilizes Python `Playwright` and `BeautifulSoup` wrapped in asynchronous `asyncio` events to scrape targets while adhering to robots.txt guidelines and anti-bot mitigation strategies.



Proxy & Header Rotation

Dynamic residential proxy networks with automated user-agent rotation to maintain seamless scraping pipelines with 99.4% uptime.

Redis Task Queue

In-memory Redis instance serving as a message broker for distributed Celery tasks, allowing scaling up to millions of crawl jobs daily.

2. Validation & Deduplication Engine

Intelligent Deduplication Algorithm

To prevent stale or duplicated company records, incoming leads pass through a strict hash matching and fuzzy lookup pipeline before entering the NLP engine:

- **Exact Hash Matching:** Normalized domain names matched using SHA-256 string keys.
- **Fuzzy Name Matching:** Levenshtein distance calculations on corporate entity names.
- **Email Verification:** Real-time MX records lookup and dynamic SMTP handshake validation.



3. NLP Analysis (SpaCy + Local Ollama LLM)

Raw HTML payloads are converted into plain text and analyzed locally without sending sensitive client data to third-party APIs. This ensures complete privacy and zero per-token cost.

A SpaCy Named Entity Recognition

Parses executive designations (e.g., CEO, CTO, VP of Engineering), phone numbers, physical addresses, and corporate registration codes.

Local Ollama (Mistral / Llama 3)

Infers business verticals, key selling propositions, employee counts, and modern tech stacks directly from raw website text.

4. Algorithmic Lead Scoring Engine

Leads are scored dynamically using a customizable weighting matrix to highlight high-value prospects for sales teams:

Scoring Factor	Evaluation Method	Max Score Impact
Company Size	Estimated employee count parsed by local LLM	+25 Points
Tech Stack Alignment	Matches target technologies (e.g., React, Python, AWS)	+30 Points
Decision Maker Contact	Direct C-level email and phone availability	+25 Points
Website Quality	Page speed, SSL certificate, modern web layout	+20 Points

</> Structured JSON Output Payload

```
{ "company_domain": "exampletech.com", "industry": "Enterprise Software", "lead_score": 92, "decision_makers": [ { "name": "Alex Mercer", "role": "CTO", "email": "a.mercer@exampletech.com" } ], "tech_stack": ["Python", "FastAPI", "PostgreSQL", "AWS"], "priority": "HIGH" }
```



 Technology Stack

Component	Technology
Language	Python 3.11+
Task Queue	Redis + Celery
NLP / Local LLM	SpaCy, Ollama
Database	PostgreSQL
API Engine	FastAPI / Flask
Containerization	Docker, Nginx

 Core Capabilities

- ✓ Multi-source web extraction
- ✓ AI contact & firmographic parsing
- ✓ Deduplication algorithm
- ✓ Lead scoring & prioritization
- ✓ Native CRM & email sync

VOXORA SOLUTIONS

Engineering Enterprise-Grade AI Platforms

 Contact Information

-  **Website:** <https://voxorasolutions.in>
-  **Email:** contact@voxorasolutions.in
-  **Phone:** +91 70577 66762
-  **HQ:** Vasai, Maharashtra, India



Scan to open <https://voxorasolutions.in>