Ah, good pivot — if you want to **replicate what Clay is doing**, but with **open-source / free data sources**, there are a lot of places you can legally tap into for lead generation. These won’t be as polished as paid providers like Clearbit or Apollo, but they’re powerful if you build the right scrapers, pipelines, and enrichment workflows.

Here’s a structured list:

1. **🌐 1. Public Business & Company Directories**

* **Crunchbase (free tier + open datasets)** → basic company info, funding, industry tags.
* **OpenCorporates** → world’s largest open database of company registries.
* **Chamber of Commerce / Gov registries** (e.g., SEC EDGAR in US, Companies House UK).
* **Wikipedia + DBpedia + Wikidata** → structured info about notable organizations.

1. **💼 2. Professional & Social Data**

* **LinkedIn (via open-source scrapers like linkedin-scraper, SerpAPI, or PhantomBuster-like tools)** → job titles, company employees, hiring trends.
* **GitHub APIs** → developers, projects, companies contributing to OSS. Great for technical leads.
* **Twitter/X APIs (or scrapers)** → people tweeting about niches, product launches, job changes.
* **Reddit / Forums** → niche communities discussing products/services.

1. **🛠 3. Job Boards & Hiring Signals (strong intent data)**

* **Indeed, Glassdoor, LinkedIn Jobs, AngelList** → hiring signals = company is growing/spending.
* **Open source scrapers** (scrapy, BeautifulSoup) can pull postings and map to companies.
* **RemoteOK / WeWorkRemotely** → startup-heavy job signals.

1. **🏢 4. Company Websites (scraping with open tools)**

* Crawl for: “About” pages, “Team” pages, contact emails, phone numbers.
* Tools: scrapy, playwright, puppeteer, trafilatura (for extracting clean text).
* Free signals: tech stack (via Wappalyzer OSS or BuiltWith-like parsers).

1. **📊 5. Open Data APIs**

* **Google Maps API / OpenStreetMap** → local businesses (addresses, categories).
* **Yelp API (free tier)** → small/local business info.
* **G2 / Capterra (public profiles)** → software vendors & reviews.
* **Product Hunt API** → new startups & launches.

1. **📚 6. Academic & Research Orgs (if your ICP is education/research)**

* **CORDIS (EU projects)** or **NSF awards (US)** → funded research projects with contact info.
* **ORCID** → researcher profiles.

1. **📨 7. Email Discovery & Validation (open tools)**

* **Hunter.io (limited free tier)** or open-source alternatives like theHarvester for email hunting.
* **Debounce.io API (free tier)** → email validation.
* **CommonCrawl** → raw web crawl data (requires processing, but huge source).

1. **⚙️ 8. Open-Source Tools to Automate Lead Generation**

* **Scrapy** (Python) → framework for building web scrapers.
* **Apify actors (many are open)** → prebuilt scrapers for LinkedIn, Google Maps, Crunchbase.
* **SerpAPI (freemium)** → structured Google search results for queries like “Top SaaS companies 2024”.
* **Prospect tools OSS**:
  + [OpenLead](https://github.com/OpenLeadHQ/openlead) – community OSS for prospect enrichment.
  + [OSINT Framework](https://osintframework.com/) – tons of open tools for gathering info.

1. **🚀 Strategy (Mimicking Clay with Open Sources)**

Clay’s **biggest trick** isn’t just the sources, but the **combination & enrichment pipeline**:

1. **Seed leads** from directories, job boards, product launches.
2. **Enrich** via GitHub/LinkedIn/website scraping → add job titles, tech stack, contact info.
3. **Score** with signals (hiring fast? using your competitor’s tech?).
4. **Export** into outreach workflows (email campaigns, CRMs).

✅ So to answer you:  
The best **open-source lead generation sources** are **OpenCorporates, Crunchbase free, Wikidata, LinkedIn scrapers, GitHub API, job boards (Indeed, AngelList, RemoteOK), Product Hunt API, Google Maps/OpenStreetMap, and CommonCrawl** — all combined via **scrapers + enrichment tools like Scrapy, Apify, OSINT libraries**.

Do you want me to sketch a **step-by-step “Clay-like” open-source pipeline** (with tools + APIs to pull, enrich, and score leads automatically)? That way you can literally replicate Clay’s backend flow without paying for premium data providers.