

Functioneel & technisch ontwerp

Persoonlijke autonome vacature-assistent zonder platform-API's

Ontwerpdoel

Na eenmalige configuratie verzamelt, normaliseert, ontdebelt en beoordeelt het systeem vacatures zelfstandig. De gebruiker ziet alleen een korte, verklaarbare selectie en blijft altijd degene die beslist en solliciteert.

Opdrachtgever	Jens
Documentstatus	Implementatieblauwdruk
Versie	1.0
Datum	20 juli 2026
Doelomgeving	Persoonlijke installatie op Unraid

Dit document beschrijft een persoonlijke, niet-commerciële toepassing. Juridische en bronvoorwaarden moeten opnieuw worden gecontroleerd wanneer het gebruiksdoel, de schaal of de distributie verandert.

Documentbeheer

Onderdeel	Keuze / aanname	Gevolg voor het ontwerp
Gebruiker	Eén primaire gebruiker; Nederlandstalige interface.	Eenvoudige autorisatie, persoonlijke voorkeuren en geen multi-tenant complexiteit.
Geografie	België als startgebied; vacatures kunnen Nederlands, Frans of Engels zijn.	Meertalige extractie en lokale Belgische postcode-/gemeentedata.
Platformkoppelingen	Geen officiële vacature-API's of contractuele integraties met LinkedIn, Indeed, VDAB, Jobat of StepStone.	Platformen worden alleen via door de gebruiker ingestelde e-mailalerts benut; directe platformcrawling is standaard geblokkeerd.
Hosting	Zelf gehost op Unraid en standaard niet publiek bereikbaar.	Containers, lokale opslag, back-ups naar de array en toegang via LAN of VPN.
AI	Lokale AI is optioneel en mag geen harde uitsluiting zelfstandig uitvoeren.	De kern blijft deterministisch; AI levert kenmerken, samenvatting en uitleg.
Solliciteren	Geen automatische indiening van sollicitaties.	De tool bewaart, helpt voorbereiden en herinnert; de gebruiker controleert en verstuurt.

Belangrijke interpretatie van “zonder API's”

Het ontwerp gebruikt geen vacatureplatform-API's. Interne interfaces tussen eigen containers, IMAP voor e-mail en de lokale Ollama-interface zijn technische bouwstenen van de eigen installatie en geen integratiecontract met een vacatureplatform. Publieke, niet-geauthenticeerde ATS-endpoints zijn optioneel en staan in de strikte modus uit.

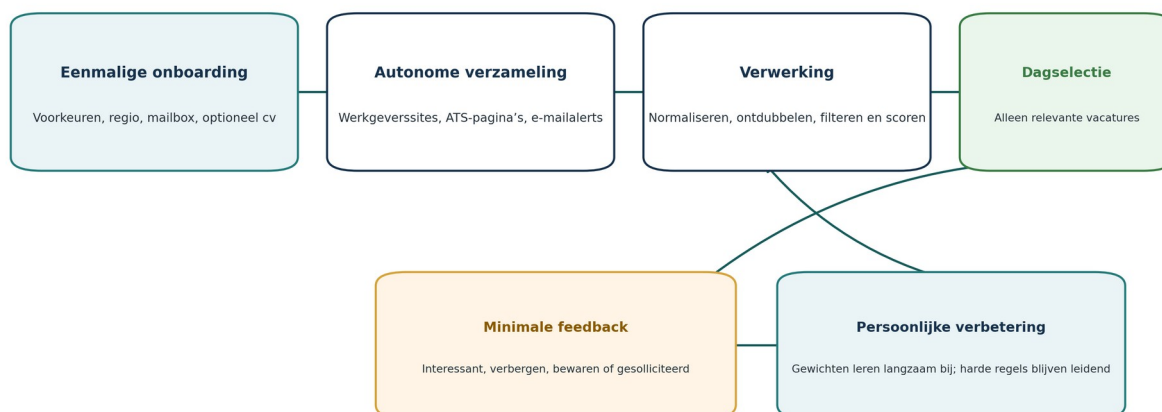
Inhoud

- 01 [Managementsamenvatting](#)
- 02 [Doel, scope en ontwerpprincipes](#)
- 03 [Minimale gebruikersinteractie](#)
- 04 [Functionele vereisten](#)
- 05 [Bronstrategie en toegangsbeleid](#)
- 06 [Automatische bronontdekking](#)
- 07 [End-to-end verwerkingspijplijn](#)
- 08 [Technische architectuur](#)
- 09 [Datamodel en bewaarbeleid](#)
- 10 [Extractie, normalisatie en datakwaliteit](#)
- 11 [Ontdubbelen en canonieke vacatures](#)
- 12 [Matchen, scoren en leren](#)
- 13 [Automatisering, planning en meldingen](#)
- 14 [Gebruikersinterface en sollicitatieopvolging](#)
- 15 [Beveiliging, privacy en juridisch kader](#)
- 16 [Betrouwbaarheid, beheer en observability](#)
- 17 [Teststrategie en kwaliteitsborging](#)
- 18 [Implementatie op Unraid](#)
- 19 [Fasering, inspanning en prioriteiten](#)
- 20 [Risico's en beheersmaatregelen](#)
- 21 [Acceptatiecriteria](#)
- 22 [Voorbeeldconfiguratie en resultaten](#)
- 23 [Aanbevolen besluit en bouwvolgorde](#)
- 24 [Bijlagen en referenties](#)

1. Managementsamenvatting

VacatureRadar is een persoonlijke zoek- en beslisassistent die zelfstandig publieke vacatures verzamelt, ze naar één intern formaat omzet, dubbele publicaties samenvoegt, niet-passende resultaten wegfiltert en de beste vacatures met een uitlegbare score presenteert. De tool is niet bedoeld als kopie van een vacatureplatform, maar als een lokale intelligentielaag bovenop rechtstreekse werkgeversbronnen en vacaturemails.

Doelbeeld: de gebruiker beoordeelt; het systeem zoekt en ordent



Figuur 1 - De gewenste gebruikerslus: eenmalig instellen, daarna alleen beoordelen.

Wat de gebruiker in de praktijk nog doet

Eenmalig voorkeuren invullen, een mailbox koppelen en eventueel een cv uploaden. Daarna volstaan acties zoals Interessant, Verbergen, Bewaren en Gesolliciteerd. Broncontrole, parsing, ontdebelling, ranking, herinneringen en dagelijkse samenvattingen verlopen automatisch.

Het aanbevolen MVP combineert vier invoerkanalen: (1) carrièrepagina's van werkgevers, (2) publieke ATS-gehoste jobpagina's, (3) e-mailalerts van vacatureplatformen en (4) optionele handmatige URL-import. De tool crawlt geen ingelogde of expliciet afgeschermd platformomgeving en lost geen CAPTCHA's op. LinkedIn verbiedt ongeautoriseerde geautomatiseerde crawling expliciet; Indeed houdt eveneens algemene verboden op scraping, bots en andere geautomatiseerde activiteit in stand [R3, R5].

De technische kern is een onderhoudsarme modulaire monoliet: Django met een server-side webinterface, PostgreSQL, Redis, een achtergrondworker en een scheduler. Playwright wordt alleen ingezet wanneer een toegestane carrièrepagina zonder JavaScript niet leesbaar is. Een lokaal Ollama-model kan vacatures samenvatten en inhoudelijke kenmerken classificeren, maar harde uitsluitingen blijven regelgebaseerd en controleerbaar.

Resultaat	Ontwerpdoel
Minimale handelingen	Na onboarding maximaal één dagelijkse digest en enkele éénklikacties.
Relevantie	Harde regels vóór ranking; AI kan niet zelfstandig een vacature definitief verwijderen.
Dekking	Zo breed mogelijk binnen toegestane publieke bronnen, zonder de onhaalbare belofte "alle vacatures".
Uitlegbaarheid	Elke aanbeveling toont scorecomponenten, bewijs uit de tekst en een confidence-niveau.
Onderhoud	Generieke parsers, bronscore, automatische retries en parserquarantaine beperken handmatig werk.
Privacy	Lokale verwerking, minimale retentie en geen publieke toegang als standaard.

2. Doel, scope en ontwerpprincipes

2.1 Primair doel

Het systeem moet continu en met zo weinig mogelijk gebruikersinteractie nieuwe, relevante vacatures vinden. “Relevant” betekent niet alleen dat een functietitel overeenkomt, maar dat locatie, contractvorm, senioriteit, inhoud, werkgeverstype, technologieën, werkmodel en persoonlijke uitsluitingen gezamenlijk passen.

2.2 Binnen scope

- Eenmalige configuratie van één of meer zoekprofielen.
- Automatische verzameling uit toegestane publieke werkgevers- en ATS-pagina's.
- Automatische verwerking van e-mailalerts via een aparte IMAP-mailbox.
- Meertalige verwerking van Nederlands, Frans en Engels.
- Normalisatie, expiratiecontrole, ontdebelling en bronherkomst.
- Harde filters, gewogen matchscore, confidence en uitlegbare aanbeveling.
- Dagelijkse digest, uitzonderingsmeldingen en sollicitatieopvolging.
- Lokale hosting, back-up, herstel, brongezondheid en auditlog.

2.3 Buiten scope voor de eerste versie

- Automatisch inloggen op vacatureplatformen of werkgeversportalen.
- CAPTCHA-omzeiling, anti-bot-ontwijking, IP-rotatie of vermomde user-agents.
- Automatisch invullen en versturen van sollicitaties.
- Commerciële herpublicatie of bulkdistributie van vacatureteksten.
- Multi-tenant SaaS, werkgeversaccounts of kandidaatselectie voor derden.
- Een absolute garantie dat elke publiek beschikbare vacature wordt gevonden.

2.4 Ontwerpprincipes

Principe	Concrete toepassing
Fail closed	Een bron met onduidelijke voorwaarden, loginvereiste, blokkering of lage extractiebetrouwbaarheid wordt niet agressiever benaderd maar gepauzeerd of ter review gemarkeerd.
HTML eerst	JSON-LD, RSS, sitemap en statische HTML krijgen voorrang. Een headless browser is de laatste technische stap, niet de standaard.
Regels vóór AI	Harde grenzen worden door deterministische logica toegepast. AI levert alleen aanvullende kenmerken en uitleg.
Eén canonieke vacature	Meerdere vindplaatsen worden aan één vacaturecluster gekoppeld; de rechtstreekse werkgeversbron krijgt voorrang.
Uitlegbaar boven slim	Elke score moet herleidbaar zijn tot velden, tekstpassages, regels en gewichten.
Lokale privacy	Persoonlijke profielen, cv's, feedback en sollicitatiehistoriek verlaten de installatie niet, tenzij de gebruiker bewust een externe dienst activeert.
Adaptief en rustig	Controlefrequentie volgt de wijzigingssnelheid van een bron en gebruikt conditionele HTTP-verzoeken om belasting te beperken.

3. Minimale gebruikersinteractie

3.1 Onboardingwizard

De eerste configuratie moet in één wizard kunnen worden afgerond. Verplichte invoer wordt beperkt tot informatie die niet betrouwbaar uit een cv kan worden afgeleid. De gebruiker kan later alles wijzigen, maar hoeft geen individuele scraper of CSS-selector te beheren.

Stap	Verplicht	Automatische hulp
Thuisbasis en mobiliteit	Postcode/gemeente, maximale afstand of reistijd, vervoer.	Gemeentecoördinaten uit lokale Belgische open data; alternatieve locaties worden voorgesteld.
Functiefamilies	Gewenste rollen en uitsluitingen.	Synoniemen in NL/FR/EN worden automatisch gegenereerd.
Werkvoorwaarden	Contract, voltijds/deeltijds, remote/hybride, minimumvoorwaarden.	Onbekende voorwaarden verlagen confidence, niet automatisch de match.
Skills en senioriteit	Gewenste en uitgesloten thema's.	Optioneel uit cv geëxtraheerd en ter bevestiging getoond.
Bronmailbox	IMAP-host, mailbox en app-wachtwoord of lokaal doorstuuradres.	Afzenders en e-mailtemplates worden automatisch herkend.
Meldingen	Tijdstip dagelijkse digest.	Standaard Europe/Brussels; onmiddellijke melding alleen bij uitzonderlijk sterke match.

3.2 Dagelijkse gebruikerslus

- De gebruiker ontvangt één digest met bijvoorbeeld “3 sterke matches, 4 mogelijke matches, 18 automatisch verborgen”.
- Per vacature zijn maximaal vier primaire acties zichtbaar: Interessant, Bewaren, Verbergen en Gesolliciteerd.
- De reden voor verbergen is optioneel. Wanneer gekozen, wordt één korte lijst met redenen getoond, geen lang formulier.
- Een vacature opent altijd de oorspronkelijke bron in een nieuw tabblad; de lokale detailpagina toont de samenvatting, score en historie.
- Alleen systeemproblemen die de dekking wezenlijk aantasten worden aan de gebruiker gemeld. Gewone retries blijven onzichtbaar.

3.3 Zelflerend zonder onvoorspelbaarheid

Feedback beïnvloedt rangschikkingsgewichten langzaam en binnen ingestelde grenzen. Een reeks afwijzingen wegens “te veel telefonische support” kan vergelijkbare vacatures lager plaatsen. De tool verandert echter nooit zelfstandig een harde regel, maximale afstand of uitgesloten contractvorm. Iedere aangeleerde wijziging is terug te draaien.

Ontwerpdoel voor handelingen

Na de eenmalige onboarding moet de normale week zonder bronbeheer of zoekopdrachten kunnen verlopen. Het systeem mag hooguit om aandacht vragen wanneer een mailbox niet meer toegankelijk is, een belangrijke bron langdurig faalt of een opslag-/back-upprobleem ontstaat.

4. Functionele vereisten

Prioriteiten: M = must-have voor MVP, S = should-have voor versie 1, C = could-have voor een latere uitbreiding.

4.1 Profiel en voorkeuren

ID	P	Vereiste
FR-P01	M	De gebruiker kan meerdere zoekprofielen aanmaken, waarvan één actief standaardprofiel.
FR-P02	M	Voorkeuren ondersteunen gewenste en uitgesloten functietitels, skills, sectoren, werkgeversvormen en contracttypes.
FR-P03	M	Locatiefilters ondersteunen gemeente/postcode, straal, toegestane regio's en remote/hybride voorkeur.
FR-P04	M	Een hard filter en een zachte voorkeur zijn afzonderlijke concepten.
FR-P05	S	Een cv kan lokaal worden geüpload om skills en ervaringssignalen voor te stellen; de gebruiker bevestigt ze.

ID	P	Vereiste
FR-P06	S	Synoniemen en vertalingen voor NL/FR/EN worden per functiefamilie beheerd.

4.2 Bronnen en verzameling

ID	P	Vereiste
FR-S01	M	Het systeem onderhoudt een bronregister met type, domein, beleid, crawlinterval, parser en gezondheid.
FR-S02	M	Het systeem verwerkt JobPosting JSON-LD, statische HTML, RSS/Atom en sitemaps.
FR-S03	M	Het systeem leest vacaturealerts uit een aparte IMAP-mailbox en dedupliceert op Message-ID en vacaturelink.
FR-S04	M	Directe crawling van geblokkeerde vacatureplatformdomeinen is standaard technisch onmogelijk via een denylist.
FR-S05	M	Per domein gelden concurrency, minimuminterval, time-out, contentlimiet, redirectlimiet en back-off.
FR-S06	M	Conditionele GET met ETag en Last-Modified wordt gebruikt waar mogelijk.
FR-S07	S	De tool ontdekt nieuwe carrièrepagina's uit werkgeverlinks in vacaturemails en uit sitemaps.
FR-S08	S	Een generieke parser kan semantisch terugvallen op zichtbare labels wanneer vaste selectors breken.
FR-S09	C	Publieke, niet-geauthenticeerde ATS-endpoints kunnen per bron expliciet worden ingeschakeld in pragmatische modus.

4.3 Verwerking, matching en output

ID	P	Vereiste
FR-M01	M	Vacatures worden naar een canoniek intern schema genormaliseerd en behouden hun bronbewijs.
FR-M02	M	Het systeem herkent vervaldata, 404/410, verwijderde JobPosting-markup en langdurige onbeschikbaarheid.
FR-M03	M	Exacte en fuzzy duplicaten worden geclusterd zonder bronverwijzingen te verliezen.
FR-M04	M	Harde filters worden vóór de matchscore toegepast en zijn volledig uitlegbaar.
FR-M05	M	Iedere vacature krijgt matchscore, confidence, pluspunten, minpunten en een adviescategorie.
FR-M06	M	Een AI-model kan geen vacature definitief verbergen zonder dat een deterministische regel dit ondersteunt.
FR-M07	S	Feedback wijzigt zachte gewichten gecontroleerd en toont een wijzigingslog.

ID	P	Vereiste
FR-M08	S	De tool berekent bronkwaliteit en geeft de rechtstreekse werkgeverspublicatie voorrang.
FR-M09	S	Verdachte vacature-/domeinsignalen worden als waarschuwing getoond, nooit als definitief fraudeoordeel.

4.4 Meldingen en sollicitatieopvolging

ID	P	Vereiste
FR-N01	M	Er is één dagelijkse digest op een configureerbaar tijdstip en in de tijdzone Europe/Brussels.
FR-N02	M	Onmiddellijke meldingen zijn optioneel en alleen voor een configureerbare topscore of naderende deadline.
FR-N03	M	De gebruiker kan vacatures bewaren, verbergen, als interessant markeren of een sollicitatiestatus geven.
FR-N04	M	Bij Gesolliciteerd worden tekst, bron, datum, deadlines en notities als dossier bewaard.
FR-N05	S	Herinneringen worden op werkdagen berekend en kunnen worden uitgesteld of uitgeschakeld.
FR-N06	S	De tool kan een motivatiebriefconcept of cv-aanpassingen voorstellen, maar niets automatisch versturen.
FR-N07	S	Export naar CSV en een compact sollicitatieoverzicht in PDF is mogelijk.

4.5 Niet-functionele vereisten

ID	Categorie	Doel
NFR-01	Beveiliging	Geen publieke poort als standaard; invoer sanitizen; SSRF-, XSS- en prompt-injectionmaatregelen.
NFR-02	Privacy	Persoonsgegevens en feedback lokaal; minimale retentie; export en volledige verwijdering mogelijk.
NFR-03	Betrouwbaarheid	Bronfouten veroorzaken geen stil verlies; retries, circuit breaker en laatst bekende geldige versie.
NFR-04	Onderhoudbaarheid	Bronadapters volgen één contract en hebben fixtures/golden tests.
NFR-05	Prestaties	De dagelijkse digest wordt op tijd gebouwd; de UI reageert normaal binnen circa één seconde op lokale filters.
NFR-06	Portabiliteit	Configuratie, database en bestanden kunnen naar een andere Dockerhost worden gemigreerd.
NFR-07	Toegankelijkheid	Toetsenbordbediening, voldoende contrast, tekstlabels en mobiele lay-out.
NFR-08	Observability	Healthchecks, gestructureerde logs, bronmetrics en herstelbare foutmeldingen.

5. Bronstrategie en toegangsbeleid

5.1 Bronhiërarchie

Rang	Bronsoort	Standaardmethode	Beleid
1	Rechtstreekse werkgeverspagina	JSON-LD, RSS, sitemap of HTML	Voorkeursbron na robots-/voorwaardencontrole.
2	Publieke ATS-gehoste jobpagina	HTML; JSON-LD; optioneel publiek endpoint	Toestaan per bron; rechtstreekse sollicitatielink bewaren.
3	Vacaturemail van platform	IMAP en veilige linkextractie	Toegestaan omdat de gebruiker de alert zelf activeert; platformpagina niet automatisch crawlen.
4	Publieke sector-/organisatiepagina	RSS, HTML of downloadbare lijst	Zelfde bronbeleid als werkgeverssite.
5	Handmatige URL-import	Eenmalige fetch en daarna bronherkenning	Fallback voor uitzonderingen; niet nodig voor normaal gebruik.
-	Ingelogde platformresultaten	Niet uitvoeren	Geblokkeerd.
-	CAPTCHA of anti-botomzeiling	Niet uitvoeren	Nooit ondersteunen.

Veel vacaturedetailpagina's publiceren machineleesbare JobPosting-gegevens met onder meer datum, omschrijving, werkgever, locatie, contracttype, vervaldatum en soms salaris. De tool gebruikt dit eerst omdat het semantischer en minder breekbaar is dan visuele CSS-selectors [R1].

Publieke ATS-jobboards kunnen eveneens een voorspelbare structuur hebben. Greenhouse documenteert bijvoorbeeld dat GET-data van gepubliceerde jobboards publiek beschikbaar is zonder authenticatie; Lever vermeldt dat gepubliceerde vacatures ook op een gehoste jobs-site zichtbaar zijn [R7, R8]. In de strikte modus leest VacatureRadar alleen de publieke HTML/JSON-LD-pagina. De beheerder kan per bron pragmatische, publiek gedocumenteerde endpoints inschakelen nadat voorwaarden zijn beoordeeld.

5.2 Bronbeleid als beslismatrix

Signaal	Besluit	Actie
Login of sessiecookie vereist	Deny	Niet automatisch ophalen; eventueel alleen e-mailmetadata en oorspronkelijke link bewaren.
CAPTCHA, challenge of expliciete botblokkering	Deny	Bron pauzeren; geen omzeiling.
Voorwaarden verbieden scraping/automatisering	Deny	Domein op technische denylist; alleen e-mailalert of handmatige browseractie.
robots.txt disallowt pad voor de eigen user-agent	Deny	Pad niet crawlen. robots.txt is geen authenticatie, maar wordt als operationele instructie gerespecteerd [R2].
Publieke werkgeverpagina, geen conflict gevonden	Allow	Rustig crawlen met identificerende user-agent en conditionele requests.
Onduidelijke voorwaarden of onverwacht redirectdomein	Review	Geen uitbreiding van crawl; bron in wachtrij voor eenmalige beoordeling.
Publieke ATS-HTML werkt; endpoint ook bekend	Allow HTML	Endpoint blijft uit in strict mode; HTML is voldoende.

5.3 Platformen via e-mail, niet via scraping

LinkedIn ondersteunt dagelijkse of wekelijkse vacaturealerts en stelt een maximum van twintig actieve alerts per lid; Indeed biedt eveneens e-mailalerts op basis van zoekcriteria [R4, R6]. Dit maakt e-mail de geschikte brug: de gebruiker activeert een beperkt aantal brede alerts, waarna VacatureRadar de resultaten verder verfijnt. De tool hoeft daardoor niet in te loggen, zoekresultaatpagina's te automatiseren of platformbeveiliging te benaderen.

Praktische alertstrategie

Gebruik brede functiefamilies en regio's in het platform, niet tientallen zeer specifieke filters. VacatureRadar voert daarna de strenge persoonlijke filtering uit. Zo wordt de limiet op platformalerts efficiënt benut en ontstaan minder blinde vlekken.

6. Automatische bronontdekking

Minimale bediening is alleen haalbaar wanneer de tool zelf een bronregister opbouwt. De ontdekking combineert een kleine startset met signalen uit echte vacatures. De gebruiker hoeft geen lijst van honderden werkgevers te onderhouden.

6.1 Ontdekkingskanalen

- Seed packs: versiebeheerbare lijsten per regio/sector, bijvoorbeeld publieke sector, onderwijs, zorg of IT-werkgevers.
- E-mailgedreven ontdekking: een alert levert werkgever, domein of apply-link; de tool zoekt op dat domein naar careers/jobs-links en sitemaps.
- Sitemapdetectie: robots.txt en bekende sitemaplocaties worden gecontroleerd; alleen URL's met job-/careerpatronen worden bemonsterd.
- Semantische detectie: pagina's met JobPosting JSON-LD, herhaalde jobcards, "vacatures/jobs/careers"-labels of ATS-domeinpatronen krijgen een bronkandidaat.
- Canonical- en redirectanalyse: trackinglinks worden veilig gevolgd tot de eindbestemming en aan de oorspronkelijke vacature gekoppeld.
- Gebruikersfeedback: wanneer een handmatig geïmporteerde URL een nieuw werkgeversdomein onthult, wordt dat domein automatisch als bronkandidaat onderzocht.

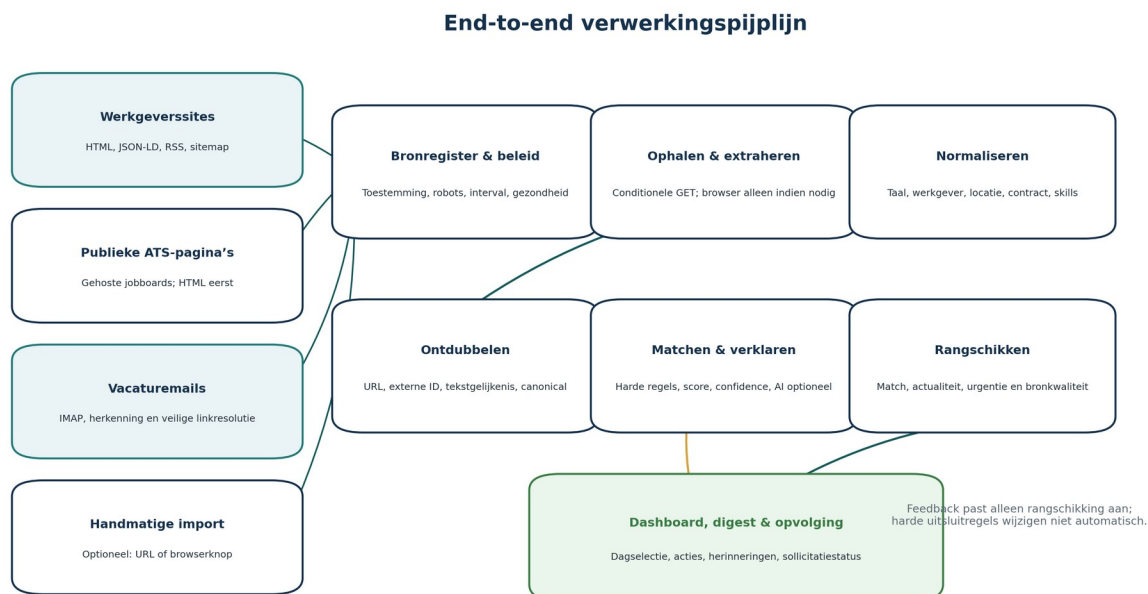
6.2 Promotie van kandidaat naar actieve bron

Fase	Automatische controle	Promotievoorwaarde
Kandidaat	Domein, HTTPS, robots, voorwaardenlink, sitemap, JobPosting-signaal.	Geen deny-signaal en minimaal één actuele vacature.
Proefrun	Maximaal enkele pagina's; lage frequentie; extractie en duplicaten controleren.	Minimaal twee consistente veldenets en voldoende extractieconfidence.
Actief	Adaptief interval; health metrics; content hashes.	Drie succesvolle runs zonder structurele fout.
Quarantaine	Parserfout, onverwachte login/challenge, massale veldverdwijning.	Automatisch herstel slaagt of technische review.
Gepauzeerd	Voorwaarden wijzigen, langdurige blokkade of bron verdwijnt.	Bewuste heractivering na herbeoordeling.

6.3 Zelfherstellende parsing met grenzen

Wanneer een selector breekt, probeert de parser eerst semantische alternatieven: JSON-LD, microdata, Open Graph, bekende labels en zichtbare hoofdtekst. Een AI-model mag in een sandbox nieuwe selectors voorstellen op basis van opgeslagen testfixtures. De nieuwe parser draait eerst in shadow mode en wordt alleen automatisch gepromoveerd wanneer meerdere actuele pagina's consistente resultaten leveren. Geen enkel herstelmechanisme mag een login, challenge of verbod omzeilen.

7. End-to-end verwerkingspijplijn



Figuur 2 - Van bron tot dagelijkse, gepersonaliseerde selectie.

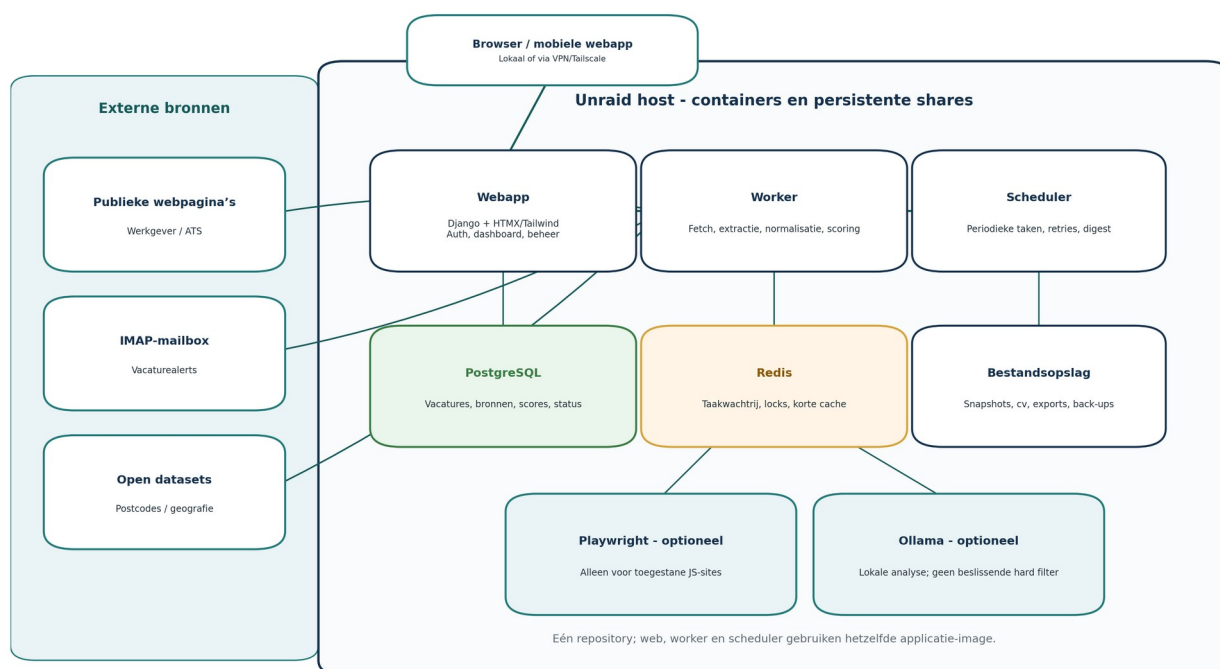
7.1 Stappen en controles

Stap	Verwerking	Belangrijkste guardrail
1. Plannen	Scheduler kiest bronnen op basis van prioriteit, wijzigingssnelheid, laatste succes en back-off.	Geen dubbele run door distributed lock.
2. Beleid	Denylist, bronstatus, robotscache en interval worden vóór elk netwerkverzoek gecontroleerd.	Fail closed.
3. Ophalen	HTTP met herkenbare user-agent, time-outs, contentlimieten en conditionele headers.	304 betekent geen verdere verwerking.
4. Extraheren	JSON-LD -> feed -> statische HTML -> browserrendering.	Elke veldwaarde krijgt provenance en confidence.
5. Normaliseren	Taal, tekst, titel, werkgever, locatie, contract, skills, data en salaris naar intern schema.	Originele tekst blijft traceerbaar.
6. Ontdubbelen	Exacte identifiers, canonical URL en fuzzy signatuur.	Bronalias blijft behouden.
7. Filteren	Harde regels en veiligheidsregels.	Geen LLM-beslissing.
8. Analyseren	Kenmerken, samenvatting, senioriteit, support-/consultancysignalen.	AI-uitvoer als strikt JSON en gevalideerd.
9. Rangschikken	Matchscore, confidence, actualiteit, urgentie en bronkwaliteit.	Scorecomponenten zichtbaar.
10. Publiceren	Dashboard, digest en optionele topmatchmelding.	Idempotent; geen dubbele melding.
11. Leren	Feedback wijzigt beperkte zachte gewichten.	Auditlog en rollback.

7.2 Idempotentie

Elke run gebruikt stabiele sleutels: bron-ID + externe vacature-ID of URL-hash + contenthash. Daardoor kan een taak opnieuw worden uitgevoerd zonder dubbele vacatures of meldingen. Updates creëren een vacatureversie en overschrijven niet stil de geschiedenis.

8. Technische architectuur



Figuur 3 - Aanbevolen modulaire monoliet met optionele browser- en AI-laag.

8.1 Aanbevolen stack

Component	Keuze	Waarom
Webapp	Django + HTMX + Tailwind CSS	Eén Python-codebase, ingebouwde auth/admin/ORM, weinig clientcomplexiteit en toch een snelle, professionele interface.
Database	PostgreSQL	Relationele integriteit, JSONB, full-text search en trigramvergelijking zonder aparte zoekcluster.
Taken	Celery worker + Celery Beat	Robuuste retries, planning, wachtrijen en herbruikbaar applicatie-image.
Broker/cache	Redis	Taakwachtrij, locks, rate-limitstatus en korte cache.
HTTP/extractie	htpx, selectolax/BeautifulSoup, extruct, feedparser	Snelle statische extractie en JSON-LD/microdataondersteuning.
Browser	Playwright Chromium, optioneel	Alleen voor toegestane JavaScriptpagina's; Playwright ondersteunt headless browserautomatisering en Dockergebruik [R14].
AI	Ollama, optioneel	Lokale modeltoegang via de eigen host; de interface draait standaard lokaal [R15].
Bestanden	Unraid appdata/share	Snapshots, cv's, exports, fixtures en back-ups buiten de containerlaag.
Reverse proxy	Caddy of bestaande proxy; VPN/Tailscale voor remote	TLS en veilige toegang zonder de applicatie rechtstreeks op internet te publiceren.

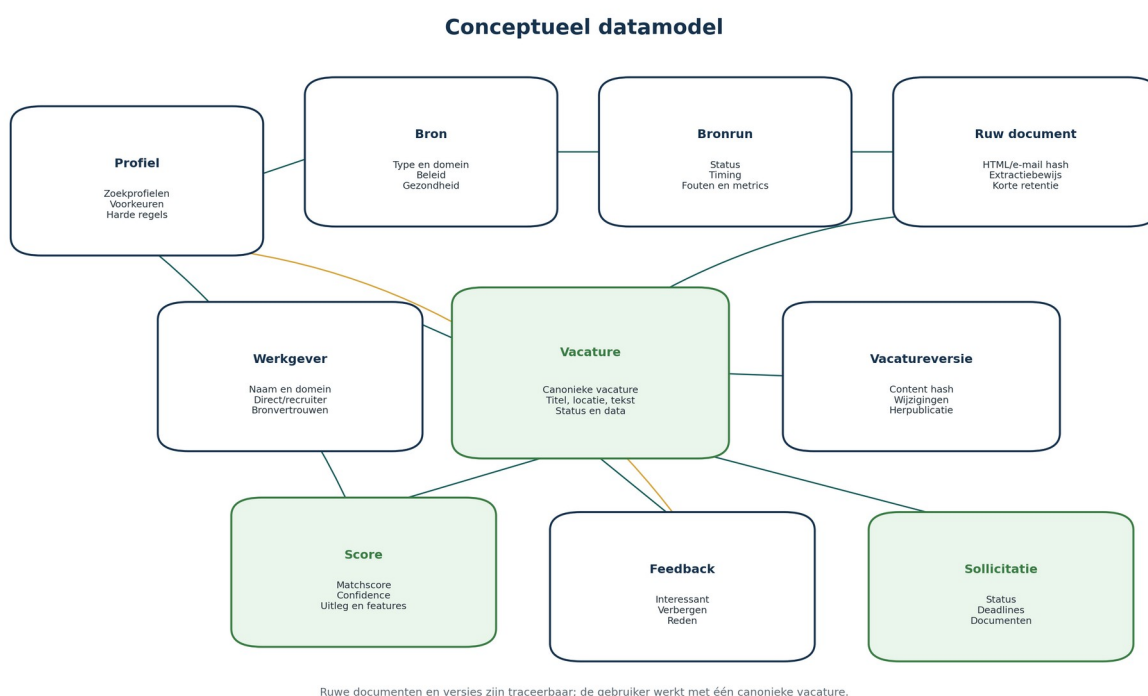
8.2 Waarom geen aparte React-frontend in het MVP

Een losse React/Next.js-frontend en FastAPI-backend zijn geschikt wanneer een publieke API, meerdere clients of zeer interactieve dashboards centraal staan. Voor één persoonlijke installatie vergroten ze echter build-, dependency-, authenticatie- en deploymentcomplexiteit. HTMX levert de benodigde filter-, modal- en inline-update-interactie zonder een tweede applicatiestack. Een JSON-API kan later met Django Ninja of DRF worden toegevoegd.

8.3 Containerrollen

Service	Verantwoordelijkheid	Schaalstrategie
web	UI, auth, zoekfilters, detailpagina's, instellingen en interne endpoints.	Eén instance voor persoonlijk gebruik.
worker	Fetch, parse, normalisatie, dedupe, analyse, export.	Concurrency beperkt; aparte low/high queue indien nodig.
scheduler	Periodieke taken en digestplanning.	Exact één actieve instance.
postgres	Permanente relationele data.	Dagelijkse dump; data op snelle cachepool.
redis	Broker, locks en tijdelijke status.	Geen primaire bron van waarheid.
ollama	Lokale AI-inferentie.	Optioneel; model en hardware onafhankelijk van kern.

9. Datamodel en bewaarbeleid



Figuur 4 - Belangrijkste entiteiten en hun traceerbare relaties.

9.1 Belangrijkste entiteiten

Entiteit	Kernvelden	Retentie
UserProfile	Taal, tijdzone, thuislocatie, notificatie-instellingen.	Tot verwijdering door gebruiker.
SearchProfile	Functiefamilies, harde regels, zachte gewichten, regio's.	Versiegeschiedenis; actief/inactief.
Source	Domein, type, policy, parser, interval, health, terms_checked_at.	Zolang actief; audit van beleidswijzigingen.
SourceRun	Start/einde, status, HTTP-metrics, aantallen, foutcategorie.	90 dagen detail; oudere aggregatie.
RawDocument	URL/e-mail-ID, contenthash, extractiebewijs, beperkte ruwe inhoud.	Standaard 7 dagen, langer alleen bij fixture of opgeslagen sollicitatie.
Employer	Genormaliseerde naam, domein, rechtstreeks/recruiter, vertrouwen.	Zolang gekoppeld.

Entiteit	Kernvelden	Retentie
JobPosting	Canonieke vacature, velden, status, eerste/laatste gezien, canonical source.	Actief + 12 maanden archief; configureerbaar.
JobVersion	Contenthash, gewijzigde velden, tekstsnapshot.	Versies van bewaarde/gesolliciteerde vacatures onbeperkt; overige 90 dagen.
ScoreRun	Profielversie, features, gewichten, score, confidence, modelversie.	Zolang vacature in archief.
Feedback	Actie, reden, tijd, gebruikte profielversie.	Tot verwijdering; exporteerbaar.
Application	Status, data, deadlines, notities, documenten.	Door gebruiker beheerd.

9.2 Canoniek vacatureschema

Groep	Velden
identity	job_id, external_ids, source_aliases, canonical_url, content_hash
title	original_title, normalized_title, job_family, seniority
employer	employer_id, display_name, direct_employer, recruiter, domain
location	raw_location, country, region, municipality, postal_code, coordinates, workplace_type
employment	employment_types, hours, contract_duration, shift signals
dates	date_posted, valid_through, first_seen, last_seen, last_changed
compensation	currency, min/max, period, source_is_employer, confidence
content	description_html_sanitized, description_text, requirements, benefits, language
skills	required, preferred, excluded_signals, experience_years
provenance	per-field source URL, extraction method, confidence, parser version
status	active, uncertain, expired, removed, duplicate, quarantined

9.3 Bewaarminimalisatie

De tool bewaart niet automatisch onbeperkt volledige HTML van elke bron. Voor gewone vacatures volstaan genormaliseerde tekst, hashes, veldprovenance en bronlinks. Ruwe documenten hebben een korte technische retentie. Pas wanneer de gebruiker een vacature bewaart of als Gesolliciteerd markeert, mag een langdurige snapshot worden gemaakt voor het persoonlijke dossier. Dit beperkt opslag, privacyrisico en onnodige kopie van broninhoud.

10. Extractie, normalisatie en datakwaliteit

10.1 Parservolgorde

Prioriteit	Methode	Gebruik
1	JobPosting JSON-LD / microdata	Hoofdbron voor titel, data, werkgever, locatie, contract en salaris wanneer aanwezig.
2	RSS/Atom of ATS-feed	Lijst en detailvelden zonder browserrendering.
3	Statische HTML	Semantische labels en site-/ATS-adapter.
4	Ingebedde state/JSON in pagina	Alleen publiek meegeleverde paginadata; geen private endpoints of sessietokens.
5	Playwright-rendering	Alleen wanneer bronbeleid dit toestaat en de inhoud uitsluitend client-side verschijnt.

Prioriteit	Methode	Gebruik
6	Tekstclassificatie fallback	Veldkandidaten met lage confidence; nooit als enige bron voor kritieke data.

10.2 Normalisatie

- Unicode, witruimte en boilerplate worden opgeschoond; zichtbare opsommingen blijven behouden.
- Functietitels worden genormaliseerd zonder de originele titel te verliezen.
- Werkgevernamen worden gekoppeld via domein, juridische/handelsnaamvarianten en bekende recruiters.
- Locaties worden naar Belgische gemeenten/postcodes vertaald; remote en hybride blijven afzonderlijke kenmerken.
- Contracttypes worden naar een beperkte interne taxonomie gemapt, maar de bronwaarde blijft beschikbaar.
- Datums worden naar UTC opgeslagen en in Europe/Brussels getoond.
- Salaris wordt alleen als “bronbedrag” beschouwd wanneer het expliciet door de werkgever/publicatie is opgegeven; schattingen worden apart gemarkeerd.

10.3 Meertaligheid

De brontaal wordt automatisch gedetecteerd. Een interne taxonomie koppelt Nederlandse, Franse en Engelse functietitels en skills aan dezelfde concepten. Samenvattingen en uitleg worden in het Nederlands getoond, terwijl originele citaten/passages in de brontaal blijven. Vertaling mag geen harde eis creëren die niet letterlijk of semantisch in de bron staat.

10.4 Datakwaliteit en confidence

Confidence	Criteria	UI-gedrag
Hoog	Veld komt uit JobPosting/ATS-structuur of is op meerdere manieren bevestigd.	Normaal tonen; harde regel mag erop steunen als het veld daarvoor geschikt is.
Middel	Veld komt uit duidelijke zichtbare tekst of één betrouwbare selector.	Tonen met normale score, maar onzekerheid benoemen.
Laag	Veld is afgeleid, ambigu of alleen door een classifieer gevonden.	Geen definitieve uitsluiting; tonen als “waarschijnlijk/onbekend”.
Conflict	Bronnen of versies spreken elkaar tegen.	Rechtstreekse werkgeversbron en recentste geldige versie voorrang; conflict zichtbaar.

Regel voor ontbrekende informatie

“Niet vermeld” is niet hetzelfde als “voldoet niet”. Ontbrekende salaris-, remote- of contractinformatie verlaagt de confidence en kan een vacature lager rangschikken, maar veroorzaakt alleen uitsluiting wanneer de gebruiker expliciet heeft gekozen voor een conservatieve onbekend-is-onvoldoende-regel.

11. Ontdubbelen en canonieke vacatures

11.1 Matchlagen

Laag	Signaal	Besluit
Exact 1	Zelfde externe vacature-ID binnen hetzelfde ATS/werkgever.	Zelfde vacature.
Exact 2	Zelfde canonical URL na trackingparameters en redirects.	Zelfde vacature.
Sterk	Genormaliseerde werkgever + titel + locatie + publicatievenster.	Kandidaatcluster; tekstcontrole.
Fuzzy	Titelgelijkenis, tekstshingles/MinHash, skills en locatie.	Cluster wanneer drempel en bewijs voldoende zijn.
Herpublicatie	Sterk gelijk maar nieuwe externe ID of datum na sluiting.	Nieuwe versie of repostrelatie; niet stil samenvoegen wanneer inhoud/voorwaarden wezenlijk wijzigen.

11.2 Canonieke bronkeuze

- Rechtstreekse werkgeverspagina boven recruiter-, aggregatie- of e-mailweergave.
- Actuele en bereikbare bron boven verlopen kopie.
- Bron met hoogste veldcompleteitheid en confidence boven beperkte samenvatting.
- Sollicitatielink naar werkgever/ATS boven tussenpagina met tracking.
- Alle aliases blijven zichtbaar in “Ook gevonden via”, zodat provenance niet verloren gaat.

11.3 Pseudocode

Conceptueel algoritme - drempels worden met een gelabelde testset gekalibreerd.

```
def resolve_candidate(job):
    exact = find_by_external_id_or_canonical_url(job)
    if exact:
        return merge_as_source_alias(exact, job)

    candidates = find_same_employer_title_location_window(job)
    best, similarity = compare_text_and_features(job, candidates)
    if best and similarity >= STRONG_THRESHOLD:
        return merge_or_create_version(best, job)

    return create_new_canonical_job(job)
```

12. Matchen, scoren en leren

12.1 Beslisvolgorde

Fase	Voorbeeld	Uitkomst
Veiligheids-/beleidsfilter	Vacaturelink is verdacht, bron in quarantaine of vacature verlopen.	Niet aanbevelen; reden tonen.
Harde persoonlijke filter	Buiten absolute regio, uitgesloten contractvorm of expliciet verboden functie-inhoud.	Verborgen, maar terugvindbaar in log.
Kenmerkextractie	Supportaandeel, infrastructuurcomponent, senioriteit, reizen, werkgeverstype.	Features met bewijs en confidence.
Gewogen score	Rol, skills, locatie, voorwaarden, werkgever, senioriteit, voorkeuren.	0-100 matchscore.
Rangschikking	Match + actualiteit + urgentie + confidence.	Volgorde in digest.

12.2 Standaard scoremodel

Component	Gewicht	Voorbeelden
Functie-inhoud	25	Kerntaken, helpdesk/supportaandeel, hands-on versus coördinatie.
Skills en leerbaarheid	20	Verplichte skills, overdraagbare kennis, opleiding voorzien.
Locatie en werkmodel	15	Afstand, regio, remote/hybride, verplaatsingen.
Werkvoorwaarden	10	Contract, uren, salaris/barema indien bekend.
Werkgever en bron	10	Rechtstreeks, sectorvoorkeur, recruiter/transparantie.
Senioriteit en haalbaarheid	10	Jaren ervaring, verantwoordelijkheid, junior/medior/senior-signalen.
Persoonlijke zachte voorkeuren	10	Technologie, organisatiecultuur-signalen, voorkeurssector.

Formule: $\text{matchscore} = \text{clamp}(0, 100, \text{som}(\text{gewicht} \times \text{genormaliseerde feature}) - \text{expliciete penalties})$. De gewichten zijn profielspecifiek en versieerbaar. Actualiteit, deadline en confidence worden naast de matchscore getoond en veranderen de inhoudelijke score niet stil.

12.3 Voorbeeldscorekaart

Onderdeel	Resultaat	Uitleg
Matchscore	84/100	Sterke inhoudelijke en geografische match.
Confidence	Middel-hoog	Taken en locatie duidelijk; salaris niet vermeld.
Functie-inhoud	22/25	Infrastructuur en beheer dominant; beperkte gebruikerssupport.
Skills	16/20	Meeste kernskills aanwezig; één gewenste technologie ontbreekt.
Locatie	14/15	Binnen ingestelde zone; hybride vermeld.
Werkgever	9/10	Rechtstreekse werkgever en originele carrièrepagina.
Advies	Solliciteren	Geen hard conflict; aandachtspunt expliciet benoemd.

12.4 AI-rol en guardrails

- Input is onbetrouwbare tekst. De systeemprompt vermeldt expliciet dat instructies in vacatureteksten genegeerd moeten worden.
- Het model krijgt geen netwerk-, shell-, e-mail- of applicatietools.
- Uitvoer moet aan een streng JSON-schema voldoen en wordt server-side gevalideerd.
- Kenmerken bevatten bewijszinnen of tekstposities, zodat hallucinaties zichtbaar worden.
- Bij modeltime-out of invalid JSON gebruikt het systeem deterministische features en markeert het de analyse als onvolledig.
- Model- en promptversie worden bij iedere score opgeslagen voor reproduceerbaarheid.

12.5 Leren van feedback

In het MVP worden geen complexe black-box aanbevelingsmodellen getraind. Feedback past per feature kleine delta's toe, bijvoorbeeld maximaal één punt gewicht per vijf consistente acties, met minimum- en maximumgrenzen. Redenen zoals “te ver”, “te veel support” of “recruiter” zijn waardevoller dan een generieke verberging. Maandelijks kan de UI voorstellen tonen: “Je verbergt vaak vacatures met veel verplaatsingen; wil je dit als harde regel instellen?” De gebruiker bevestigt dat bewust.

13. Automatisering, planning en meldingen

13.1 Adaptieve planning

Taak	Standaard	Adaptatie
IMAP-inbox	Elke 5-10 minuten of IMAP IDLE.	Langere poll bij inactiviteit; directe wake-up indien ondersteund.
Actieve carrièrebron	Elke 6-24 uur.	Sneller bij frequente wijzigingen, trager bij zelden veranderende bron.
Sitemap/feed	Elke 6-12 uur.	ETag/Last-Modified en lastmod bepalen vervolg.
Verificatie actieve vacature	Dagelijks of rond deadline.	Hogere prioriteit vlak voor validThrough.
Dagelijkse digest	Eén vast tijdstip.	Alleen nieuwe/gewijzigde, nog niet geziene resultaten.
Bronvoorwaarden/robots	Maandelijks en bij onverwacht gedrag.	Direct hercontroleren bij 403/429/challenge.
Back-up	Dagelijks database; wekelijks bestanden/config.	Restore-test maandelijks.

13.2 Meldingsregels

- Digest bevat alleen resultaten boven de ingestelde drempel en een korte statistiek van verborgen/duplicaat/verlopen items.
- Een onmiddellijke topmatchmelding vereist zowel een hoge matchscore als voldoende confidence.
- Deadlinemeldingen zijn afzonderlijk van matchmeldingen en worden niet vaker dan nodig herhaald.
- Systeemmeldingen worden gegroepeerd: één melding voor een langdurig bronprobleem, niet voor elke mislukte run.
- De gebruiker kan stille uren instellen; urgente deadlineberichten respecteren een aparte voorkeur.

13.3 Vacaturelevenscyclus

Status	Betekenis	Overgang
new	Nieuw gevonden en nog niet verwerkt.	Na normalisatie naar active of quarantined.
active	Actueel en bereikbaar.	Update, uncertain, expired of removed.
uncertain	Bron tijdelijk onbereikbaar of velden conflicteren.	Niet direct laten verdwijnen; hercontrole.
expired	validThrough voorbij of expliciet gesloten.	Archief; deadlineherinneringen stoppen.
removed	404/410 of vacaturemarkup langdurig weg.	Archief; eerdere sollicitatie blijft bestaan.
quarantined	Parser-/beleids-/veiligheidsprobleem.	Technische herstelroute.

14. Gebruikersinterface en sollicitatieopvolging

14.1 Schermen

Scherm	Kerninhoud	Primaire actie
Vandaag	Nieuwe sterke en mogelijke matches, deadlines en bronwaarschuwingen.	Vacature beoordelen.
Vacaturedetail	Scorekaart, bewijs, samenvatting, originele tekst, bronaliases en historie.	Open bron / bewaren / verbergen.
Bewaarlijst	Interessante vacatures met deadline, score en vergelijking.	Prioriteren of Gesolliciteerd.
Sollicitaties	Kanban of lijst: voorbereiding, verzonden, gesprek, aanbod, afgewezen.	Status en opvolgdatum wijzigen.
Profiel	Harde regels, zachte voorkeuren, zoektermen en geleerde suggesties.	Wijzigingen bevestigen.
Brongezondheid	Alleen onder Geavanceerd: status, laatste run, parser, policy.	Retry of bron pauzeren.
Systeem	Mailbox, meldingen, back-up, opslag, model, exports.	Test verbinding / export / herstel.

14.2 Detailpagina

- Bovenaan: titel, werkgever, locatie, werkmodel, datum, deadline en originele bron.
- Scorebalk met componenten; geen enkel percentage zonder tekstuele verklaring.
- “Waarom passend” en “Aandachtspunten” met gemarkeerde bronpassages.
- Vereist versus gewenst apart; ontbrekende informatie apart van echte mismatch.
- Duplicaatcluster met rechtstreekse werkgever als voorkeurslink.
- Historie: wanneer gevonden, inhoudelijk gewijzigd, gesloten of herplaatst.
- Acties altijd op dezelfde plaats en bruikbaar op mobiel/toetsenbord.

14.3 Sollicitatiedossier

Veld	Automatisch	Handmatig
Vacaturesnapshot	Bij Gesolliciteerd of expliciet Bewaren.	Gebruiker kan snapshot verwijderen.

Veld	Automatisch	Handmatig
Sollicitatiedatum	Bij statuswijziging.	Aanpasbaar.
Contact	Uit vacature/e-mail indien duidelijk en toegestaan.	Notitie of contact toevoegen.
Documenten	Versies van cv/motivatatie koppelen.	Upload of link naar eigen bestand.
Opvolgdatum	Voorstel op basis van sluitingsdatum en werkdagen.	Accepteren, wijzigen of uitschakelen.
Gesprekken	Kalenderdatum, locatie/link en notities.	Door gebruiker bevestigd; geen automatische deelname.
Uitkomst	Open, interview, aanbod, afgewezen, ingetrokken.	Door gebruiker of uit eigen e-mailstatus bevestigd.

15. Beveiliging, privacy en juridisch kader

Geen juridisch advies

Dit ontwerp past conservatieve technische maatregelen toe, maar vervangt geen juridisch advies. Voor commercieel gebruik, distributie aan derden of grootschalige verzameling moet een gespecialiseerde beoordeling plaatsvinden.

15.1 Toegangs- en bronregels

De Robots Exclusion Protocol-standaard beschrijft regels die crawlers gevraagd worden te respecteren en verduidelijkt dat robots.txt geen toegangsautorisatie is [R2]. VacatureRadar behandelt robots toch als een bindende operationele grens. Daarnaast worden voorwaarden en zichtbare technische blokkades afzonderlijk beoordeeld. LinkedIn vereist uitdrukkelijke toestemming voor geautomatiseerde crawling; Indeed handhaaft verboden op scraping, bots en andere geautomatiseerde activiteit [R3, R5]. Daarom staan die domeinen op de standaarddenylist.

15.2 Privacy by design

- Dataminimalisatie: alleen velden bewaren die voor zoeken, beoordelen of eigen sollicitatiehistoriek nodig zijn.
- Doelbinding: data niet gebruiken voor kandidatenprofilering, marketing of publicatie aan derden.
- Lokale verwerking: cv, feedback en sollicitatiegegevens blijven standaard binnen de Unraid-installatie.
- Retentie: ruwe broninhoud kort; persoonlijke dossiers door gebruiker beheerd.
- Rechtenfuncties: export, correctie en volledige verwijdering van profiel en dossiers.
- Logging: geen wachtwoorden, volledige cv-tekst of mailboxinhoud in logs.

De AVG/GDPR regelt verwerking van persoonsgegevens in de EU; contactnamen en e-mailadressen in vacatures kunnen persoonsgegevens zijn [R9]. Het ontwerp vertrouwt niet op een vermeende uitzondering, maar hanteert minimale opslag en beperkte toegang. De EU-databankenrichtlijn beschermt onder voorwaarden ook substantiële investeringen in databanken [R10]. Daarom wordt geen vacatureplatformdatabase systematisch gekopieerd of heruitgegeven.

15.3 Technische dreigingen en mitigaties

Dreiging	Mitigatie
SSRF via e-mail- of vacaturelink	Alleen http/https; DNS-resolutie controleren; private/loopback/link-local IP's blokkeren; maximaal vijf redirects; opnieuw valideren na elke redirect.
XSS uit vacature-HTML	Allowlist-sanitizer; scripts, forms, event handlers, iframes en onbekende embeds verwijderen; Content Security Policy.
Prompt injection in vacaturetekst	Tekst als onbetrouwbare data; model zonder tools; strikte systemprompt; JSON-schema; bewijsvereiste; time-out.
Mailboxcompromis	App-wachtwoord of dedicated account; alleen leesrechten waar mogelijk; secrets buiten database; rotatie en verbindingcheck.
Credential leak	Docker secrets/.env met beperkte rechten; nooit in Git; versleutelde back-up wanneer extern opgeslagen.
Onbedoelde publieke toegang	Standaard bind op LAN; reverse proxy met TLS; remote toegang via VPN/Tailscale; rate limiting en sterke login.

Dreiging	Mitigatie
Kwaadaardige bestanden	Alleen toegestane cv/documenttypes; groottebeperking; geen macro-uitvoering; optionele malwarecontrole.
Supply chain	Vastgepinde images/dependencies, automatische securityscan, gecontroleerde updates en rollback.

15.4 Authenticatie en toegang

Voor lokaal gebruik volstaat een sterk uniek wachtwoord met sessiebeveiliging en optioneel TOTP. Wanneer de tool buiten het LAN bereikbaar is, krijgt VPN/Tailscale de voorkeur boven port forwarding. De applicatie heeft een afzonderlijke accountlaag en vertrouwt niet uitsluitend op de Unraid-webinterface.

16. Betrouwbaarheid, beheer en observability

16.1 Bronhealth

Metric	Doel / alarm
Laatste succesvolle run	Waarschuwen wanneer een belangrijke bron langer dan twee geplande intervallen faalt.
Parse yield	Alarm bij plotselinge daling in gevonden vacatures/velden ten opzichte van eigen historie.
HTTP-statusmix	429/403/challenge veroorzaakt vertraging en policy-hercontrole.
Veldcompleteid	Titel/werkgever/URL zijn minimum; sterke daling zet parser in shadow/quarantaine.
Duplicaatratio	Extreme stijging kan canonicalisatie- of trackingparameterprobleem aangeven.
Vacatureleeftijd	Bron met uitsluitend oude vacatures krijgt lager interval of pauze.
Digestvolume	Onverwacht nul of zeer hoog aantal triggert interne sanity check vóór verzending.

16.2 Foutafhandeling

- Netwerkfout: exponentiële back-off met jitter; bron blijft active/uncertain, vacatures worden niet meteen verlopen verklaard.
- 429: Retry-After respecteren, interval verhogen en concurrency verlagen.
- 403/challenge: bron pauzeren en policy opnieuw controleren; geen agressieve retry.
- Parserfout: ruwe fixture kort bewaren, alternatieve parser proberen en oude geldige data behouden.
- AI-fout: deterministische score publiceren met lagere analyseconfidence.
- Database-/brokerfout: healthcheck faalt, taken stoppen veilig en worden na herstel idempotent hervat.
- Digestfout: outboxpatroon voorkomt verlies en dubbele verzending.

16.3 Back-up en herstel

Object	Frequentie	Hersteldoel
PostgreSQL dump	Dagelijks; plus vóór upgrades.	Vacatures, profielen, feedback en sollicitaties.
Configuratie/secrets-manifest	Wekelijks, secrets afzonderlijk versleuteld.	Containers en instellingen opnieuw uitrollen.
Bestandsopslag	Wekelijks incrementeel.	Snapshots, cv's, exports en fixtures.
Restore-test	Maandelijks in tijdelijke database.	Aantonen dat back-up bruikbaar is.
Off-host kopie	Optioneel versleuteld.	Bescherming tegen host-/arrayverlies.

16.4 Onderhoudsverwachting

De gebruiker hoeft normaal geen brononderhoud te doen. Technisch onderhoud blijft wel bestaan: dependency-updates, gewijzigde websites, mailboxrotatie en periodieke bronvoorwaarden. Het ontwerp minimaliseert dit met

generieke parsers, fixtures, automatische healthchecks en een duidelijk geavanceerd dashboard. Een installatie die jarenlang volledig zonder onderhoud draait, is geen realistische garantie voor webextractiesoftware.

17. Teststrategie en kwaliteitsborging

17.1 Testlagen

Laag	Voorbeelden	Gate
Unit	Titelnormalisatie, datummapping, URL-canonicalisatie, scoreformule.	Elke commit.
Parser fixture	Opgeslagen HTML/e-mail per bron; verwachte velden in golden JSON.	Geen regressie bij parserwijziging.
Contract	Elke adapter retourneert hetzelfde schema en foutcategorieën.	Voor activering bron.
Dedupe benchmark	Gelabelde paren: zelfde, repost, verschillend.	Precisie krijgt voorrang; twijfel niet automatisch samenvoegen.
Ranking evaluatie	Handmatig gelabelde relevante/irrelevante vacatures per profiel.	Topselectie bevat verwachte matches en uitleg.
End-to-end	E-mail of fixture -> vacature -> score -> digest -> feedback.	Voor release.
Security	SSRF, XSS, prompt injection, auth, file upload, secrets.	Geen high-severity bevinding.
Restore	Back-up naar schone omgeving.	Maandelijks succesvol.

17.2 Testdata en bronvriendelijkheid

Tests gebruiken primair opgeslagen, gesaniteerde fixtures en lokale mockservers. Live-brontests zijn beperkt tot een kleine canaryset en respecteren dezelfde rate limits als productie. Zo veroorzaakt ontwikkeling geen onnodige belasting en blijven parsertests reproduceerbaar.

17.3 Kwaliteitsmetrics

- Extractiecompleteheid per veld en bron.
- False merge en missed duplicate in een gelabelde dedupeset.
- Precisie van de top-N aanbevelingen op eigen feedback.
- Aantal sterke matches dat de gebruiker verbergt en reden daarvan.
- Aantal gemiste vacatures in een kleine periodieke handmatige benchmark.
- Bronuptime, gemiddelde hersteltijd en aantal gebruikersmeldingen.
- Digestlatency en dubbele meldingen (doel: nul).

18. Implementatie op Unraid

Unraid ondersteunt Dockercontainers, maar documenteert momenteel dat Docker Compose niet native wordt ondersteund. Voor een complexe stack kan de Compose-specificatie wel als bron van waarheid dienen, waarna services via Unraid-templates of een bewust gekozen Compose-beheeroplossing worden uitgerold [R11, R12].

18.1 Opslaglayout

Aanbevolen Unraid-sharestructuur

```
/mnt/user/appdata/vacatureradar/
├── postgres/      # database op cachepool
├── redis/         # alleen persistente brokerstatus indien gewenst
├── config/        # niet-geheime configuratie
├── media/         # cv, snapshots en exports
├── fixtures/      # parserfixtures
├── logs/          # geroteerde applicatielogs
└── backups/       # lokale dumps; kopieer periodiek naar array/off-host
```

18.2 Compose-skelet

Illustratief - exacte imageversies worden bij implementatie vastgepind en getest.

```
services:
  web:
    image: vacatureradar:1.0
    command: gunicorn config.wsgi:application
    depends_on:
      postgres: { condition: service_healthy }
      redis: { condition: service_healthy }

  worker:
    image: vacatureradar:1.0
    command: celery -A config worker -Q high,default,low

  scheduler:
    image: vacatureradar:1.0
    command: celery -A config beat

  postgres:
    image: postgres:17
    volumes: ["/var/lib/postgresql/data"]
    healthcheck:
      test: ["CMD-SHELL", "pg_isready -U $POSTGRES_USER"]

  redis:
    image: redis:7-alpine
    healthcheck:
      test: ["CMD", "redis-cli", "ping"]

  ollama:
    image: ollama/ollama
    profiles: ["ai"]
```

Docker Compose ondersteunt healthchecks en afhankelijkheden op service_healthy; dit voorkomt dat de webapp start voordat database en broker klaar zijn [R13]. Op Unraid wordt dezelfde logica vertaald naar containerhealthchecks en startvolgorde.

18.3 Netwerk

- Alleen de webservice krijgt een hostpoort; database, Redis en Ollama blijven op een intern Docker-netwerk.
- Standaard alleen LAN-toegang. Voor extern gebruik: VPN/Tailscale of een bestaande beveiligde reverse proxy.
- Fetcher gebruikt een afzonderlijke netwerkpolicy en mag geen private IP-ranges benaderen via aangeleverde URLs.
- Ollama wordt niet rechtstreeks op het LAN of internet gepubliceerd; alleen de worker mag verbinden.

18.4 Capaciteit

De regelgebaseerde kern en statische parsers passen op bescheiden hardware. Browserrendering en lokale AI bepalen de piekbelasting. Daarom zijn beide optioneel en concurrency-bepikt. Het systeem moet ook zonder GPU en zonder AI-model volledig functioneel blijven; alleen inhoudelijke samenvattingen en enkele zachte features worden dan eenvoudiger.

18.5 Upgradeprocedure

- Back-up en succesvolle restore-check vóór grote upgrade.
- Nieuwe images eerst in staging met fixturetests en database-migratiecheck.
- Pinned versie promoten; healthchecks en canarybronnen controleren.
- Rollback naar vorige image en databaseback-up wanneer migratie niet veilig terugdraaibaar is.

19. Fasering, inspanning en prioriteiten

Onderstaande ranges zijn indicatieve ontwikkelinspanning voor één ervaren full-stack Pythonontwikkelaar. Brondiversiteit en de gewenste afwerkingsgraad zijn de grootste onzekerheden; dit zijn geen kalenderbeloften.

Fase	Inhoud	Indicatie	Exitcriterium
0. Validatie & beleid	Definitieve voorkeuren, bronpolicy, 15-25 representatieve bronnen, e-mailvoorbeelden.	4-7 persoonsdagen	Go/no-go en testcorpus.

Fase	Inhoud	Indicatie	Exitcriterium
1. Core MVP	Django, PostgreSQL, profielen, bronregister, HTTP/JSON-LD/HTML, IMAP.	14-20 dagen	Nieuwe vacatures verschijnen traceerbaar.
2. Normalisatie & dedupe	Taxonomie, locatie, versiebeheer, expiratie en clusters.	9-14 dagen	Geen storende duplicaten in testset.
3. Matching & UI	Harde filters, score, confidence, dashboard, digest.	12-18 dagen	Dagelijkse selectie is bruikbaar.
4. Opvolging & lokale AI	Sollicitatiedossier, reminders, Ollama-features, feedbackleren.	10-16 dagen	Volledige persoonlijke workflow.
5. Hardening & Unraid	Security, back-up, restore, monitoring, deploymenttemplates.	8-13 dagen	Release candidate met hersteltest.
6. Bronpacks	Sector-/regiopacks, extra ATS-adapters, parserscanaries.	10-20 dagen	Brede brondekking voor doelgebied.

19.1 MVP dat al waarde levert

- Eén zoekprofiel en een onboardingwizard.
- Rechtstreekse werkgeverspagina's via JSON-LD/HTML plus IMAP-vacaturemails.
- PostgreSQL, canonieke vacatures, eenvoudige fuzzy dedupe en expiratie.
- Harde filters, handmatig ingestelde gewichten en scoreuitleg.
- Vandaag-scherm, vacaturedetail, dagelijkse e-maildigest en vier kernacties.
- Bronhealth, retries, back-up en Unraid-deployment.

19.2 Niet blokkeren op AI

Lokale AI wordt pas toegevoegd nadat extractie, dedupe en regelranking aantoonbaar goed werken. Dit voorkomt dat een model technische datakwaliteitsproblemen maskeert. Het MVP kan al relevante vacatures leveren met taxonomie, regels, full-text features en expliciete gewichten.

20. Risico's en beheersmaatregelen

Risico	Kans	Impact	Beheersing
Websites wijzigen	Hoog	Middel	JSON-LD eerst, generieke parsers, fixtures, health metrics, shadow repair en quarantaine.
Platformvoorwaarden beperken toegang	Hoog	Hoog	Geen directe crawling van platformen; e-mailalerts en werkgeverbronnen als kern.
Onvolledige dekking	Hoog	Middel	Meerdere kanalen, seed packs, e-mailgedreven ontdekking en periodieke benchmark; geen "alle vacatures"-belofte.
Foutieve dedupe	Middel	Hoog	Precisie boven recall, twijfelclusters zichtbaar, versie-/repostmodel en gelabelde benchmark.
AI-hallucinatie	Middel	Middel	AI optioneel, bewijsvelden, JSON-schema, geen harde filter, fallback naar regels.
Te veel meldingen	Middel	Middel	Dagelijkse bundeling, thresholds, stille uren en één melding per incident.
Bronbelasting/429	Middel	Middel	Conditionele requests, adaptief interval, één concurrency per domein, Retry-After.
Privacy- of credentiallek	Laag-middel	Hoog	Lokale toegang, secretsbeheer, minimale logs, back-ups versleutelen, securitytests.

Risico	Kans	Impact	Beheersing
Unraid-/containerupgrade breekt stack	Middel	Middel	Pinned versies, staging, healthchecks, back-up en rollback.
Gebruiker vertrouwt score te sterk	Middel	Middel	Confidence en bewijs prominent, score geen kans op aanwerving, eindbeslissing expliciet bij gebruiker.

21. Acceptatiecriteria

21.1 Gebruik

- Een nieuwe gebruiker kan de verplichte onboarding afronden zonder een bronadapter of technische selector te configureren.
- Na onboarding is de normale dagflow uitvoerbaar vanuit één digest of het Vandaag-scherm.
- Elke vacature kan met één klik worden bewaard, verborgen, interessant gemarkeerd of als gesolliciteerd geregistreerd.
- Een verborgen vacature vermeldt welke harde regel of gebruikersactie dit veroorzaakte.

21.2 Verzameling en data

- De testset van toegestane JobPosting- en HTML-bronnen wordt volgens schema verwerkt en bewaart per veld provenance.
- Platformdenylist voorkomt netwerkverzoeken naar geconfigureerde verboden domeinen vanuit crawltaken.
- Een tijdelijke bronstoring verwijdert geen actieve vacatures onmiddellijk.
- Exacte duplicaten worden samengevoegd en rechtstreekse werkgeverslinks krijgen voorrang.
- Vervallen of verwijderde vacatures worden gearhiveerd zonder een bestaand sollicitatiedossier te verliezen.

21.3 Matching

- Harde regels zijn deterministisch getest en onafhankelijk van AI-beschikbaarheid.
- Iedere aanbevolen vacature toont scorecomponenten, confidence, pluspunten, aandachtspunten en bronbewijs.
- AI-uitval verhindert geen digest en leidt tot een zichtbaar lagere analyseconfidence.
- Feedbackgewichten zijn begrensd, versieerbaar en terug te draaien.

21.4 Beheer en beveiliging

- Database- en bestandsback-up kunnen in een schone testomgeving worden hersteld.
- Healthchecks detecteren onbereikbare database, broker en webapp.
- SSRF-tests blokkeren private en loopback-adressen, ook na redirects.
- Vacature-HTML kan geen script uitvoeren in de lokale UI.
- De applicatie is standaard niet publiek bereikbaar en publiceert PostgreSQL, Redis of Ollama nooit rechtstreeks.

22. Voorbeeldconfiguratie en resultaten

Illustratief profiel

De waarden hieronder zijn voorbeelden en mogen niet als bevestigde persoonlijke voorkeuren worden gelezen. De onboarding vult ze met de echte keuzes van de gebruiker.

22.1 Zoekprofiel in YAML

Voorbeeld - een echte configuratie wordt via de UI beheerd en als versieerbaar document geëxporteerd.

```
mode: strict
locale: nl-BE
timezone: Europe/Brussels

digest:
  daily_at: "07:30"
  immediate_match_min_score: 92
  quiet_hours: ["22:00", "07:00"]

search_profiles:
```

```

- name: "Primair IT-profiel"
  home_postal_code: "0000"
  max_distance_km: 45
  preferred_workplace: [hybrid, on_site]
  desired_titles:
    - systeembeheerder
    - infrastructure engineer
    - workplace engineer
  hard_exclusions:
    employment_type: [freelance]
    title_terms: [sales, recruiter]
  soft_preferences:
    direct_employer: 0.9
    public_sector: 0.7
    limited_first_line_support: 0.8

source_policy:
  honor_robots: true
  crawl_logged_in_pages: false
  solve_captcha: false
  public_ats_endpoints: false
  max_concurrency_per_domain: 1
  raw_document_retention_days: 7

```

22.2 Voorbeeld van een dagelijkse digest

VacatureRadar - dinsdag 21 juli

3 sterke matches
 4 mogelijke matches
 18 automatisch verborgen
 6 dubbele publicaties samengevoegd
 1 bewaarde vacature sluit binnen 2 dagen

TOPMATCH - Infrastructure Engineer - 88/100
 Rechtstreekse werkgever | 24 km | hybride | confidence: hoog
 Waarom: sterke infrastructuurcomponent, realistische senioriteit, weinig support.
 Aandachtspunt: VMware is gewenst maar niet verplicht.

[Interessant] [Bewaren] [Verbergen] [Open originele vacature]

22.3 Interne adapterinterface

```

class SourceAdapter(Protocol):
    source_type: str

    def discover(self, seed: SourceSeed) -> list[SourceCandidate]: ...
    def list_items(self, source: Source, cursor: Cursor | None) -> FetchPlan: ...
    def extract(self, document: RawDocument) -> list[ExtractedJob]: ...
    def normalize(self, item: ExtractedJob) -> CanonicalJobDraft: ...
    def health(self, run: SourceRun) -> HealthAssessment: ...

# Beleid, HTTP, retries en opslag zitten buiten de adapter,
# zodat een parser nooit zelf beperkingen kan omzeilen.

```

22.4 Interne applicatie-endpoints

Methode	Pad	Doel
GET	/jobs	Filterbare vacaturelijst.
GET	/jobs/{id}	Detail, score, bewijs, bronnen en historie.
POST	/jobs/{id}/feedback	Interessant/bewaren/verbergen met optionele reden.
POST	/jobs/{id}/application	Sollicitatiestatus en dossier.
GET/POST	/profiles	Zoekprofielen en versies.
GET	/sources	Brongezondheid en policy voor geavanceerd beheer.
POST	/sources/{id}/retry	Bewuste handmatige retry.
POST	/imports/url	Optionele URL-import.

Methode	Pad	Doel
GET	/health	Liveness/readiness voor containerhealthcheck.

23. Aanbevolen besluit en bouwvolgorde

Aanbeveling

Bouw eerst een betrouwbare bron- en datakern met e-mailimport, directe werkgeverspagina's, normalisatie en dedupe. Voeg ranking en gebruikersworkflow daarna toe. Behandel lokale AI als een uitbreidbare analysemodule, niet als fundament.

23.1 Beslissingen die nu kunnen worden vastgelegd

- Productnaam als werktitel: VacatureRadar.
- Persoonlijke single-user installatie, België-first, interface in het Nederlands.
- Strict mode standaard: geen platformcrawling en geen publieke ATS-endpoints zonder expliciete bronbeslissing.
- Django-modulaire monoliet met PostgreSQL, Redis, Celery en optionele Playwright/Ollama-profielen.
- Dagelijkse digest als primaire gebruikersinterface; dashboard als verdieping.
- Geen auto-apply; wel dossier, reminders en optionele documentconcepten.
- Data- en bronprovenance zijn verplichte kernfuncties, geen latere toevoeging.

23.2 Eerste implementatiesprint

1. Repository, CI, Dockerimage, Django-project, PostgreSQL en healthchecks.
2. Datamodel voor profiel, bron, bronrun, ruw document, werkgever en vacature.
3. Generieke JSON-LD-parser en één statische HTML-adapter met fixtures.
4. IMAP-import met generieke linkextractie en één platformtemplate.
5. Canonieke URL, contenthash, eenvoudige dedupe en bronprovenance.
6. Minimalistisch Vandaag-scherm en handmatige feedbackacties.

Na deze sprint moet een echte vacature vanaf een toegestane bron of e-mail zichtbaar worden met titel, werkgever, locatie, originele link, herkomst en status. Pas daarna worden score, AI en automatische bronontdekking uitgebreid.

24. Bijlagen en referenties

Bijlage A - Begrippen

Begrip	Betekenis
Aggregator	Vacatureplatform dat vacatures van meerdere werkgevers of recruiters toont.
ATS	Applicant Tracking System; systeem waarmee werkgevers vacatures en kandidaten beheren.
Canonieke vacature	De interne hoofdrepresentatie waaraan meerdere bronpublicaties zijn gekoppeld.
Confidence	Mate waarin velden en analyse door betrouwbare broninformatie worden ondersteund.
Harde regel	Voorwaarde die direct uitsluit of blokkeert en niet automatisch wordt aangeleerd.
Zachte voorkeur	Kenmerk dat de rangschikking beïnvloedt maar niet automatisch uitsluit.
Provenance	Herkomst van een veld: bron, URL, extractiemethode, parser en confidence.
Quarantaine	Bron of document dat tijdelijk niet automatisch verder wordt verwerkt wegens beleids-, parser- of veiligheidsprobleem.
Repost	Een opnieuw gepubliceerde vacature die mogelijk dezelfde positie betreft maar een nieuwe publicatiecyclus heeft.

Begrip	Betekenis
Strict mode	Modus zonder directe platformcrawling en zonder optionele publieke ATS-endpoints.

Bijlage B - Referenties

Ref.	Organisatie en bron	URL
R1	Google Search Central - JobPosting structured data for Job Search	https://developers.google.com/search/docs/appearance/structured-data/job-posting
R2	IETF - RFC 9309 - Robots Exclusion Protocol	https://datatracker.ietf.org/doc/html/rfc9309
R3	LinkedIn - LinkedIn Crawling Terms and Conditions	https://www.linkedin.com/legal/crawling-terms
R4	LinkedIn Help - Job alerts on LinkedIn	https://www.linkedin.com/help/linkedin/answer/a511279
R5	Indeed - Algemene voorwaarden	https://nl.indeed.com/legal
R6	Indeed Help - Vacature-alerts activeren, stopzetten en beheren	https://support.indeed.com/hc/nl/articles/204488890-Vacature-alerts-activeren-stopzetten-en-beheren
R7	Greenhouse - Job Board API - Introduction and authentication	https://developers.greenhouse.io/job-board.html
R8	Lever - Postings API documentation and hosted job site	https://github.com/lever/postings-api
R9	EUR-Lex - Regulation (EU) 2016/679 - General Data Protection Regulation	https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng
R10	EUR-Lex - Directive 96/9/EC on the legal protection of databases	https://eur-lex.europa.eu/eli/dir/1996/9/oj/eng
R11	Unraid Docs - Run Docker containers - overview	https://docs.unraid.net/unraid-os/using-unraid-to/run-docker-containers/overview/
R12	Docker Docs - Compose file reference	https://docs.docker.com/reference/compose-file/
R13	Docker Docs - Control startup and shutdown order in Compose	https://docs.docker.com/compose/how-tos/startup-order/
R14	Microsoft Playwright - Playwright Python documentation	https://playwright.dev/python/docs/intro
R15	Ollama - API introduction	https://docs.ollama.com/api/introduction
R16	Statbel - Open data - Address files and statistical sectors 2025	https://statbel.fgov.be/en/news/open-data-address-files-and-statistical-sectors-2025
R17	bpost - Postcodevalidatie en downloadbare postcodelijst	https://www.bpost.be/nl/postcodevalidatie-tool

Alle online referenties zijn geraadpleegd op 20 juli 2026. Voorwaarden, documentatie en softwaregedrag kunnen wijzigen; controleer de actuele versies vóór implementatie en periodiek tijdens gebruik.

Bijlage C - Open beslispunten voor implementatie

Besispunt	Aanbevolen default	Wanneer herzien
Exacte thuislocatie en vervoerswijze	Alleen postcode/gemeente; geen volledig adres nodig.	Wanneer routegebaseerde reistijd essentieel wordt.
Mailboxprovider	Dedicated IMAP-mailbox met app-wachtwoord.	Bij provider zonder betrouwbare IMAP of securitypolicy.
AI-model	Klein lokaal instructiemodel met gestructureerde output.	Na benchmark op eigen vacatures en hardware.
Routeberekening	Postcode-/gemeentecentroid en hemelsbrede afstand in MVP.	Zelfgehoste router toevoegen wanneer reistijd doorslaggevend is.
Publieke ATS-endpoints	Uit in strict mode.	Per bron activeren na expliciete voorwaardenreview.
Retentie actieve/archiefvacatures	Actief + 12 maanden archief.	Aanpassen aan opslag en persoonlijk dossierbeleid.
Remote toegang	LAN + VPN/Tailscale.	Reverse proxy alleen wanneer bestaand en goed beheerd.

Einde document - VacatureRadar versie 1.0