

Được. Đây là **Tài liệu Đặc tả Kỹ thuật** (Technical Specification Document) đầy đủ cho dự án của bạn, được tổng hợp từ toàn bộ các phân tích của chúng ta.



## Tài liệu Đặc tả Kỹ thuật: Project Auto-SEO

Tên dự án: Project Auto-SEO

Phiên bản: 1.0

Ngày: 9/11/2025

Kiến trúc sư: Gemini AI

### 1. Tổng quan & Mục tiêu

**Project Auto-SEO** là một Nền tảng Tự động hóa Content và SEO (Content & SEO Automation Platform) với khả năng tự học, tự tối ưu và tự vận hành.

**Mục tiêu:** Xây dựng một hệ thống đa-người dùng (multi-tenant) cho phép quản lý nhiều trang blog, tự động phân tích từ khóa, lập chiến lược nội dung, tự động viết bài chuẩn SEO, đăng bài, và tự học (self-learning) dựa trên hiệu suất thực tế (Google Analytics, GSC) để cải tiến chiến lược và tối ưu chi phí.

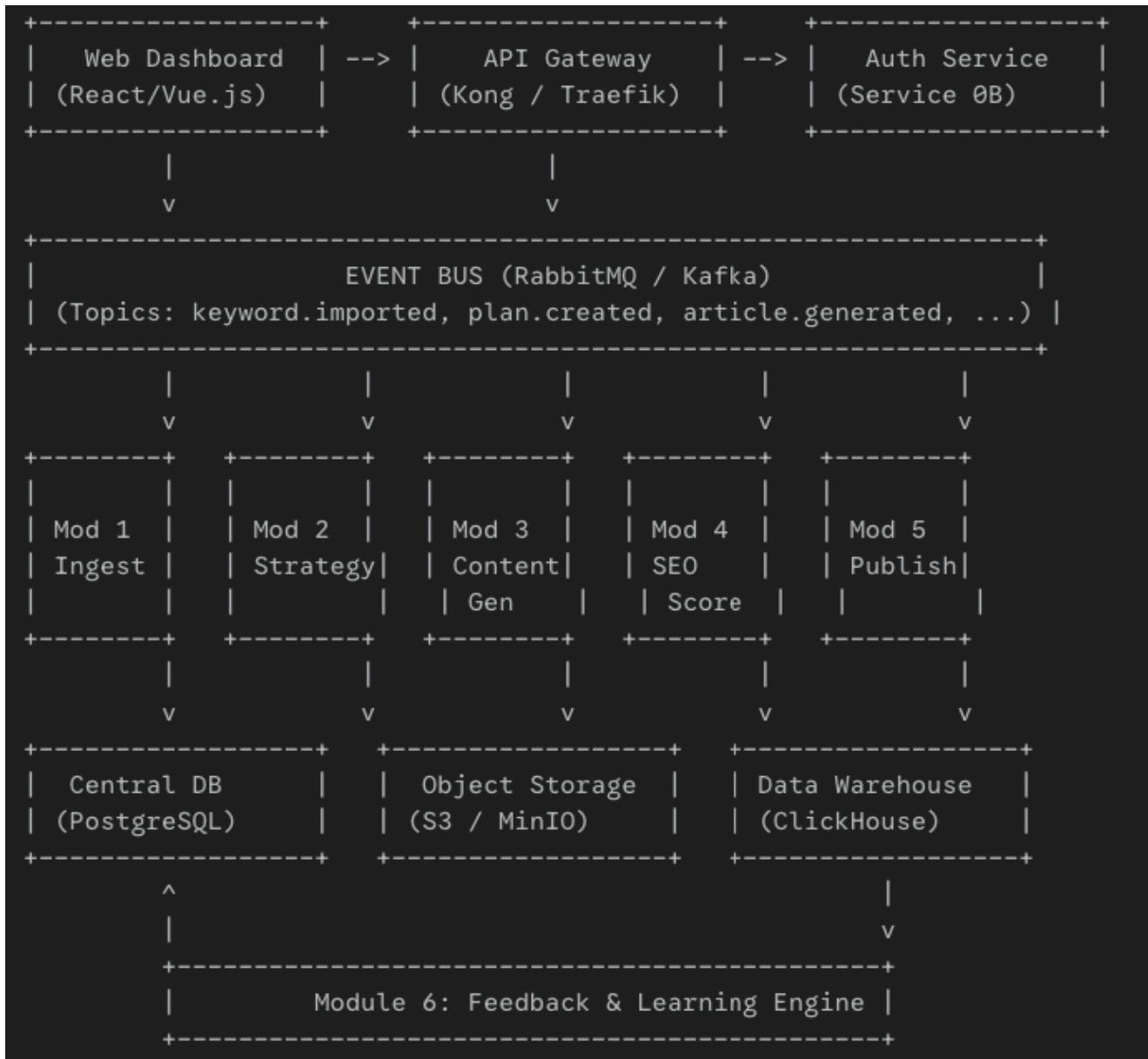
#### Triết lý Kiến trúc:

- Microservices:** Hệ thống được chia nhỏ thành các "engines" độc lập, dễ dàng phát triển, nâng cấp và mở rộng quy mô.
- Event-Driven (EDA):** Các module giao tiếp với nhau bất đồng bộ qua một Bus Sự kiện (Event Bus). Điều này đảm bảo tính ổn định và khả năng phục hồi.
- Polyglot:** Chọn ngôn ngữ lập trình tốt nhất cho từng công việc (ví dụ: Python cho AI, Node.js cho I/O).
- Cloud-Native & Kubernetes-Native:** Thiết kế để chạy trên Kubernetes, tận dụng khả năng tự động co giãn (auto-scaling) và tự phục hồi (self-healing).

### 2. Sơ đồ Kiến trúc Tổng thể (High-Level Architecture)

Luồng vận hành cốt lõi: Người dùng tương tác qua Dashboard. Yêu cầu đi qua API Gateway, được xác thực và sau đó được chuyển thành các "sự kiện" (events). Các Microservices

(Engines) lắng nghe và xử lý các sự kiện này một cách độc lập.



### 3. Module 0: Nền tảng Kỹ thuật & Dịch vụ Lõi (Technical Foundation)

Đây là "xương sống" và "hệ điều hành" của toàn bộ dự án.

### 3.1. Ngăn xếp Công nghệ (Technology Stack)

Hạng mục	Lựa chọn Đề xuất (Recommended)	Lý do & Tùy chọn (Rationale & Options)
Ngôn ngữ (Polyglot)	Python (Chủ đạo), Node.js, Go	<p><b>Python:</b> Bắt buộc cho AI/ML, NLP, Data (Module 2, 3, 4, 6).</p> <p><b>Node.js:</b> Hiệu suất I/O tốt cho API Gateway, BFF, gọi API bên thứ 3 (Module 1, 5).</p> <p><b>Go (Tùy chọn):</b> Dùng cho các worker cần concurrency cao (Scraping, Internal Linker).</p>
Bus Sự kiện	RabbitMQ (Giai đoạn 1-2)	<p>Hoàn hảo cho Task Queue &amp; Pub/Sub. Dễ vận hành.</p> <p><b>Nâng cấp lên Kafka (Giai đoạn 3)</b> khi cần xử lý streaming dữ liệu hiệu suất cực cao.</p>
CSDL (Primary)	PostgreSQL (v15+)	Ổn định, hỗ trợ JSONB, có pgvector cho semantic search cơ bản.
CSDL (Cache)	Redis	Dùng cho cache session, cấu hình, và hàng đợi job nhỏ.
CSDL (Analytics)	ClickHouse (hoặc BigQuery)	Bắt buộc cho Module 6 (OLAP). <b>ClickHouse</b> có thể tự host, tối ưu chi phí.
Frontend	React (hoặc Vue.js)	Xây dựng Dashboard quản

		lý.
--	--	-----

### 3.2. Dịch vụ Lõi (Core Services)

- **Service OA (API Gateway):** "Cổng" duy nhất cho mọi request. Chịu trách nhiệm điều tuyến (routing), rate limiting.
- **Service OB (AuthN/AuthZ):** Dịch vụ Xác thực & Phân quyền. Quản lý JWT, Users, Workspaces, và mã hóa (trong Vault) API keys (WordPress, Google, AI...).
- **Service OC (Workspace & Site Mgmt):** Service CRUD để quản lý Sites (lưu URL WP, API endpoint) và Workspaces (khách hàng).
- **Service OD (Notification):** Dịch vụ gửi thông báo (Email, Slack, UI) khi có sự kiện (ví dụ: article.published, job.failed).

### 3.3. Hạ tầng & Vận hành (Infrastructure & DevOps)

- **Containerization: Docker** (Toàn bộ microservices phải được đóng gói).
- **Orchestration: Kubernetes (K8s)** (Bắt buộc. Dùng để Tự động Scale, Tự phục hồi).
- **Monitoring (Observability): Prometheus** (Metrics), **Grafana** (Dashboards), **Loki** (Logs).
- **CI/CD: GitHub Actions hoặc GitLab CI** (Tự động build, test, deploy lên K8s).
- **Storage: MinIO** (Tự host, tương thích S3) hoặc **AWS S3** (Lưu trữ file, nội dung bài viết).

---

## 4. Chi tiết các Module Nghiệp vụ (Engines)

Chi tiết lộ trình phát triển 3 giai đoạn cho từng engine.



### Module 1: Keyword & Data Ingestion Engine

- **Mục đích:** Nhận và chuẩn hóa dữ liệu từ khóa đầu vào.
- **Lộ trình Phát triển:**

Tính năng	Giai đoạn 1 (MVP)	Giai đoạn 2 (Automation)	Giai đoạn 3 (Self-Learning)
<b>Nhập dữ liệu</b>	Upload thủ công file <b>CSV/TXT</b> (Google Keyword Planner).	<ul style="list-style-type: none"> <li>- <b>API Connectors:</b> Tự động kéo từ khóa từ Ahrefs/SEMrush.</li> <li>- Cho phép <b>paste</b> danh sách vào UI.</li> </ul>	<ul style="list-style-type: none"> <li>- <b>Real-time:</b> Tích hợp Google Trends API để tự động phát hiện từ khóa "nóng".</li> </ul>
<b>Xử lý dữ liệu</b>	Đọc file, chuẩn hóa (lowercase, xóa trùng).	<ul style="list-style-type: none"> <li>- Tự động gọi API (miễn phí) để lấy "Search Intent" cơ bản.</li> </ul>	<ul style="list-style-type: none"> <li>- Sử dụng <b>NLP model nội bộ</b> để tự động phân loại Intent (Informational, Commercial...).</li> </ul>

- **Mô hình Dữ liệu (PostgreSQL):** Keywords(id, list\_id, text, status: 'pending', intent: 'unknown')
- **Output (Sự kiện):** keyword.list.imported

## Module 2: SEO Strategy & Planning Engine

- **Mục đích:** Phân tích, xây dựng ma trận từ khóa (Topic Cluster) và lên kế hoạch (Content Plan).
- **Lộ trình Phát triển:**

Tính năng	Giai đoạn 1 (MVP)	Giai đoạn 2 (Automation)	Giai đoạn 3 (Self-Learning)
<b>Phân cụm (Clustering)</b>	Thủ công. UI "Kéo-Thả" (Drag-Drop) để user tự nhóm từ khóa thành <b>Topic Clusters</b> .	<ul style="list-style-type: none"> <li>- <b>Tự động cơ bản:</b> Dùng <b>TF-IDF/K-Means</b> để gợi ý cụm. User duyệt và "Accept".</li> </ul>	<ul style="list-style-type: none"> <li>- <b>Tự động ngữ nghĩa:</b> Dùng <b>Embeddings Model (SBERT)</b> để nhóm theo ý nghĩa (semantic).</li> </ul>
<b>Phân tích đối thủ</b>	Thủ công. User tự	- Tự động gọi API	- <b>SERP Analysis</b>

	nhập URL đối thủ.	(Ahrefs/SEMrush) lấy Top 10 đối thủ.  - Phân tích cơ bản (đếm chữ, H1, H2).	<b>Engine:</b> Service chuyên dụng tự "cào" (scrape) Top 10 Google, phân tích cấu trúc, entities.
<b>Tạo Kế hoạch</b>	User tự tạo Content Plan và set Priority thủ công.	- <b>Rule-Based Engine:</b> Tự động gán Priority theo Volume/KD.	- <b>Predictive Model:</b> Dựa trên data Module 6, ML dự đoán "từ khóa nào sẽ rank" và tự động tạo Content Plan.

- Mô hình Dữ liệu (PostgreSQL):** TopicClusters(id, name, type: 'pillar'), ContentPlans(id, cluster\_id, title, priority, competitors\_data\_json)
- Output (Sự kiện):** content.plan.created

### **Module 3: Smart Content Generation Engine**

- Mục đích:** "Nhà máy" sản xuất nội dung, với bộ định tuyến chi phí thông minh.
- Lộ trình Phát triển:**

Tính năng	Giai đoạn 1 (MVP)	Giai đoạn 2 (Automation)	Giai đoạn 3 (Self-Learning)
<b>Lên lịch (Scheduler)</b>	<b>Thủ công.</b> User bấm nút "Viết bài ngay".	- <b>Lên lịch tự động:</b> Cron job tự động quét ContentPlans và tạo ArticleJob theo priority.	- <b>Lịch thông minh:</b> Lịch đăng dựa trên dữ liệu traffic (từ Module 6).
<b>Viết bài (Generation)</b>	- <b>Một model duy nhất</b> (ví dụ: GPT-3.5-Turbo) với prompt cứng.	- <b>LLM Gateway Service:</b> Lớp trùm tượng để gọi <b>nhiều API</b> (Gemini, GPT, Grok...).	- <b>RAG (Retrieval-Augmented Generation):</b> Tự động tìm thông tin mới (dùng SERP data từ Module 2)

		<ul style="list-style-type: none"> <li>- <b>Cost-Optimization Router:</b> Logic IF-THEN để chọn model theo priority (High -&gt; GPT-4o; Low -&gt; DeepSeek).</li> </ul>	và "nhồi" vào prompt cho AI.
<b>Nguồn ảnh</b>	<b>Thủ công.</b> User tự upload.	<ul style="list-style-type: none"> <li>- <b>Tự động (Miễn phí):</b> Tích hợp API Pexels, Unsplash.</li> </ul>	<ul style="list-style-type: none"> <li>- <b>AI Image Generation:</b> Tích hợp DALL-E 3 / Stable Diffusion để tạo ảnh độc quyền.</li> </ul>

- **Mô hình Dữ liệu (PostgreSQL):** Articles(id, plan\_id, status: 'draft', ai\_model\_used, cost)
- **Mô hình Dữ liệu (Object Storage):** bucket/articles/<article\_id>.html (lưu nội dung).
- **Output (Sự kiện):** article.generated



## Module 4: SEO Scoring & Self-Correction Engine

- **Mục đích:** Tự chấm điểm SEO và thực thi "Vòng lặp tự học chiến thuật" (Tactical Loop).
- **Lộ trình Phát triển:**

Tính năng	Giai đoạn 1 (MVP)	Giai đoạn 2 (Automation)	Giai đoạn 3 (Self-Learning)
<b>Chấm điểm (Scoring)</b>	<b>Thủ công.</b> UI hiển thị checklist (Tiêu đề, H2, ảnh...) cho user tự tick.	<ul style="list-style-type: none"> <li>- <b>Internal Scorer (Tự động):</b> Service (Python/BeautifulSoup) tự phân tích HTML và chấm điểm (ví dụ: 82/100).</li> </ul>	<ul style="list-style-type: none"> <li>- <b>Scorer tự học:</b> Dựa trên data Module 6, hệ thống tự điều chỉnh trọng số (weight) của checklist.</li> <li>- (Tùy chọn) Tích hợp API SurferSEO.</li> </ul>
<b>Tự sửa (Correction)</b>	<b>Không có.</b> User tự sửa tay.	<ul style="list-style-type: none"> <li>- <b>Tactical Correction Loop (Tự động):</b></li> </ul>	<ul style="list-style-type: none"> <li>- <b>Sửa lỗi ngữ nghĩa:</b> Dùng AI (Gemini/GPT-4) để đọc bài viết và đọc</li> </ul>

		<p>IF score &lt; 80 AND reason == 'missing_keyword_in_h2'</p> <p>-&gt; Auto-update prompt</p> <p>-&gt; Bắn lại sự kiện 'article.generate.request' (quay lại Module 3).</p>	<p>lý do fail -&gt; AI tự viết <i>lại</i> prompt hoàn toàn mới để sửa lỗi.</p>
--	--	--	--

- **Mô hình Dữ liệu (PostgreSQL):** SeoScore(article\_id, auto\_score: 82, checklist\_json: '[...]', status: 'passed')
- **Output (Sự kiện):** article.approved\_for\_publishing (nếu score > 80) HOẶC article.generate.request (nếu score < 80, lặp lại).



## Module 5: Publishing & Linking Engine

- **Mục đích:** Đăng bài lên WP và quản lý, xây dựng mạng lưới internal link thông minh.
- **Lộ trình Phát triển:**

Tính năng	Giai đoạn 1 (MVP)	Giai đoạn 2 (Automation)	Giai đoạn 3 (Self-Learning)
<b>Đăng bài (Publishing)</b>	<b>Copy-Paste.</b> User tự copy HTML từ Dashboard và paste vào WP.	<ul style="list-style-type: none"> <li>- <b>WP Publisher Service:</b> Tự động đăng bài lên WP qua <b>WP API (Application Passwords)</b>.</li> </ul>	<ul style="list-style-type: none"> <li>- <b>Multi-Platform:</b> Mở rộng để đăng lên <b>Ghost, Medium, Webflow...</b></li> <li>- Tự động ping Google Indexing API.</li> </ul>
<b>Internal Linking</b>	<b>Không có.</b>	<ul style="list-style-type: none"> <li>- <b>Cơ bản (Cron Job):</b> Hàng đêm, Job chạy, tìm các bài cũ có chứa</li> </ul>	<ul style="list-style-type: none"> <li>- <b>Internal Link Injector (Thông minh):</b></li> </ul>

		<p>chính xác từ khóa (string matching) của bài mới và chèn link.</p>	<p>1. Dùng <b>Embeddings (SBERT)</b> để tìm các bài cũ <i>liên quan về mặt ngữ nghĩa</i>.</p> <p>2. Dùng AI (Gemini) để viết <i>lại</i> một câu trong bài cũ để chèn link mới vào một cách <i>tự nhiên nhất</i>.</p>
--	--	--	--

- **Mô hình Dữ liệu (PostgreSQL):** PublishedPosts(article\_id, site\_id, wp\_post\_id: 123, auto\_url: '...'), InternalLinkMap(from\_post\_id, to\_post\_id, anchor\_text)
- **Output (Sự kiện):** article.published

## Module 6: Feedback & Strategic Learning Engine

- **Mục đích:** "Bộ não chiến lược" (Strategic Loop). Hoàn thiện vòng lặp tự học dựa trên hiệu suất thực tế.
- **Lộ trình Phát triển:**

Tính năng	Giai đoạn 1 (MVP)	Giai đoạn 2 (Automation)	Giai đoạn 3 (Self-Learning)
Thu thập Dữ liệu	Thủ công. User tự xem GA/GSC và nhập rank vào Dashboard.	- <b>Data Ingestor (Tự động):</b> Worker hàng ngày tự gọi API của GA & GSC, kéo dữ liệu (Impressions, Clicks, Position) cho các URL đã đăng.	- <b>Thu thập dữ liệu chi phí</b> (từ Module 3) và so sánh Cost vs Rank.
Báo cáo (Reporting)	Báo cáo cơ bản (dựa trên data user nhập).	- <b>Dashboard Phân tích:</b> Hiển thị biểu đồ cho user thấy bài nào đang rank,	- <b>Correlational Analysis:</b> Báo cáo "sâu": "Các bài viết dùng GPT-4o có

		bài nào không.	<i>avg_position</i> cao hơn 30% so với <i>GPT-3.5</i> ."
<b>Tự học (Learning)</b>	<b>Không có.</b> User tự rút kinh nghiệm.	<p>- <b>Alerting:</b> Hệ thống tự động báo (Alert) cho user: "Cụm chủ đề 'XYZ' không có hiệu suất tốt."</p>	<p>- <b>Strategy Re-evaluation Engine (Tự động):</b></p> <ol style="list-style-type: none"> <li><b>Phát hiện:</b> "Cluster 'XYZ' (dùng Prompt_A, Model_Gemini1.0) đã 30 ngày không rank."</li> <li><b>Hành động:</b> Tự động <i>quay lại</i> Module 2 và tạo một ContentPlan mới để "Viết lại (Rewrite)" cluster 'XYZ' với chiến lược mới (ví dụ: Prompt_B, Model_Gemini1.5).</li> </ol>

- Mô hình Dữ liệu (ClickHouse/BigQuery):** FactPerformance(date, url\_hash, impressions, clicks, position, ai\_model\_used, prompt\_id, cost\_usd)
  - Output (Sự kiện):** content.plan.regenerate (gửi cho Module 2).
- 

## 5. Kiến trúc Dữ liệu Tổng thể (Data Architecture Summary)

- Central DB (PostgreSQL):** Cơ sở dữ liệu nghiệp vụ (OLTP). Lưu trữ "trạng thái" của hệ thống (users, sites, keywords, content\_plans, articles\_metadata, jobs).
- Object Storage (S3 / MinIO):** Lưu trữ các "vật thể" (artifacts) lớn. (File CSV đầu vào, nội dung HTML/Markdown của bài viết, ảnh, logs chi tiết).
- Data Warehouse (ClickHouse):** Cơ sở dữ liệu phân tích (OLAP). Lưu trữ dữ liệu hiệu suất (từ Module 6) để thực hiện các truy vấn phân tích phức tạp, phục vụ cho "bộ não" tự học.

---

## 6. 🚀 Khả năng Mở rộng & Tương lai (Scalability & Future-Proofing)

Kiến trúc này được thiết kế để "mở" từ đầu:

- **Mở rộng AI:** Thêm một model AI mới (ví dụ: Claude 5) chỉ đơn giản là thêm một "driver" mới vào **LLM Gateway Service (Module 3)**.
- **Mở rộng Nền tảng:** Đăng bài lên Webflow/Ghost? Chỉ cần thêm một "Publisher Service" mới (trong **Module 5**) lắng nghe cùng sự kiện article.approved\_for\_publishing.
- **Mở rộng Tính năng:** Đăng bài lên Mạng xã hội? Thêm một "Social Media Service" lắng nghe sự kiện article.published để tóm tắt (dùng AI) và đăng.
- **Mở rộng Quy mô:** Nếu Module 3 (Viết bài) bị quá tải, **Kubernetes** sẽ tự động scale (tăng số lượng pods) chỉ cho service đó mà không ảnh hưởng đến phần còn lại của hệ thống.

Tài liệu đặc tả này cung cấp một lộ trình rõ ràng để xây dựng một hệ thống từ Giai đoạn 1 (MVP) đến Giai đoạn 3 (Hoàn toàn tự học), đảm bảo nền tảng kỹ thuật vững chắc cho sự phát triển lâu dài.