

Content Aggregation System

WhatsApp Group Link Aggregator & Semantic Clustering Platform

Transforming Social Media Noise into Structured Intelligence

Business Proposal v1.0 | February 2026

Executive Summary

Content Aggregation System adalah platform canggih yang mengubah noise dari grup WhatsApp yang penuh dengan link media sosial menjadi **insights terstruktur yang actionable**. Sistem ini secara otomatis mengumpulkan, menganalisis, dan mengelompokkan konten dari berbagai platform (Instagram, Twitter/X, TikTok, Facebook) berdasarkan topik semantik, kemudian menghasilkan ringkasan cerdas menggunakan AI.

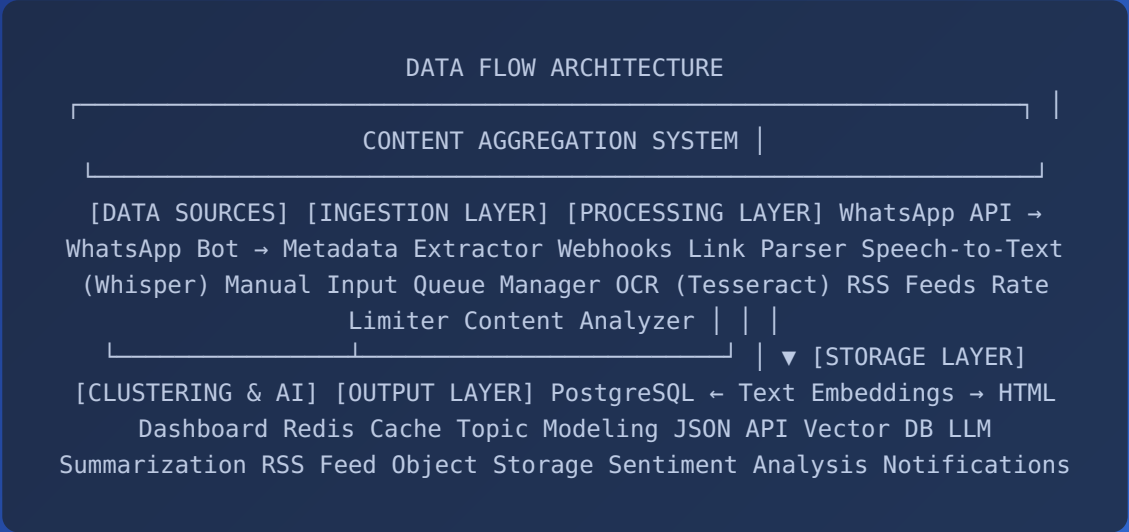
Core Value Proposition: Mengubah ratusan link per hari yang membutuhkan waktu berjam-jam untuk dibaca, menjadi ringkasan 5 menit yang mudah dikonsumsi dan actionable.

Key Features

- Real-time Auto-Ingestion:** Monitoring dan capture link dari WhatsApp secara otomatis

- **Intelligent Metadata Extraction:** Ekstraksi caption, deskripsi, engagement stats, dan thumbnail
- **Speech-to-Text Processing:** Transkripsi otomatis untuk konten video/audio
- **Semantic Clustering:** Pengelompokan konten lintas platform berdasarkan topik menggunakan AI embeddings
- **AI-Powered Summarization:** Ringkasan otomatis per topik dengan key insights
- **Structured Dashboard:** Visualisasi data dalam timeline view, topic clusters, dan searchable feed

System Architecture



Architecture Components

Layer	Technology	Function
Data Sources	WhatsApp API, Webhooks	Capture links dari grup WhatsApp secara realtime
Ingestion	Redis Queue, Bull Queue	Rate limiting dan job distribution
Processing	Python + Node.js Workers	Metadata extraction, S2T, OCR
AI/ML	OpenAI, Local LLMs	Embeddings, clustering, summarization
Storage	PostgreSQL, Redis, Pinecone	Relational, cache, vector database
Frontend	Next.js, Tailwind CSS	Dashboard, timeline, search interface

Technology Stack

Frontend Layer

- **Next.js 14:** React framework dengan App Router untuk optimal performance
- **Tailwind CSS + shadcn/ui:** Modern styling dengan component library
- **React Query:** Data fetching dan caching
- **Recharts/D3.js:** Data visualization dan analytics charts
- **PWA Support:** Progressive Web App untuk mobile experience

Backend Infrastructure

- **Node.js + Fastify:** High performance API server
- **Python + FastAPI:** ML service endpoints untuk AI processing
- **Bull Queue:** Redis-based job queue untuk background processing
- **WebSocket:** Real-time updates ke dashboard
- **Bull Board:** Queue monitoring interface

Database & Storage

- **PostgreSQL:** Primary relational database untuk links, contents, clusters
- **Redis:** Caching layer dan message queue
- **Pinecone/Chroma:** Vector database untuk embeddings
- **MinIO/S3:** Object storage untuk media files
- **Elasticsearch:** Full-text search engine

AI / Machine Learning

- [OpenAI API](#): Text embeddings dan summarization
- [Whisper API](#): Speech-to-text untuk video content
- [Owen \(Local/Cheap\)](#): Alternative untuk summarization dengan cost efektif
- [Hugging Face](#): Topic modeling dan classification models
- [yt-dlp, gallery-dl](#): Media scraping tools

Platform Integration

Platform	Extraction Library	Data Extracted	Speech-to-Text
Instagram	instaloader, gallery-dl	Caption, hashtags, likes, comments, metadata	Whisper (video content)
Twitter/X	ntscraper, snsrape	Tweet text, media attachments, engagement metrics	Whisper (video content)
TikTok	yt-dlp (tiktok-extract)	Caption, hashtags, music info, view/like/share stats	Whisper (required - video primary)
Facebook	facebook-scraper	Post text, reactions, shares, comments	Whisper (video content)
YouTube	yt-dlp, youtube-transcript-api	Title, description, captions, channel info	Caption API / Whisper fallback

Semantic Clustering Algorithm

Sistem menggunakan pendekatan **Hierarchical Clustering dengan Vector Embeddings** untuk mengelompokkan konten secara semantik:

1. **Text Preprocessing:** Menggabungkan transcript, caption, dan description; melakukan cleaning dan normalization.
2. **Embedding Generation:** Menggunakan OpenAI text-embedding-3-small (cost-effective) atau local embedding models untuk menghasilkan vector representation.
3. **Similarity Calculation:** Cosine similarity antara vector konten untuk mengukur kemiripan semantik.
4. **Hierarchical Clustering:** Agglomerative clustering dengan dynamic threshold untuk grouping otomatis.
5. **Topic Labeling:** LLM membaca cluster dan generate nama topik yang deskriptif secara otomatis.
6. **Incremental Update:** Konten baru ditambahkan ke cluster existing atau membuat cluster baru tanpa recompute semua data.

```
PSEUDO-CODE: Cluster Assignment async function
assignToCluster(newContent) { // 1. Generate text embedding const
embedding = await openai.embeddings.create({ model: "text-embedding-3-
small", input: newContent.combinedText }); // 2. Find similar existing
clusters const similarClusters = await vectorDB.query({ vector:
embedding, topK: 5, minSimilarity: 0.85 }); // 3. Assign or create new
cluster if (similarClusters[0]?.similarity > 0.90) { await
addToCluster(newContent, similarClusters[0].clusterId); } else { await
createNewCluster(newContent, embedding); }
```

Dashboard & User Interface

1. Timeline View

Stream kronologis semua link yang masuk, dengan filter by platform, date range, dan keyword search. User dapat melihat evolution of topics over time.

2. Topic/Cluster View

Konten dikelompokkan per topik dengan visualisasi:

- Nama topik (AI-generated dari konten cluster)
- Ringkasan singkat (1-2 paragraf)
- Jumlah konten dan platform breakdown
- Timeline khusus untuk topik tersebut
- Sentiment analysis per topik

3. Summary Feed

Ringkasan harian/mingguan yang mencakup:

- Top 3 Topik Hari Ini
- Trending hashtags dan keywords
- Key insights dengan actionable recommendations
- Highlight konten dengan engagement tertinggi

4. Advanced Search

- **Full-text search:** Elasticsearch untuk pencarian cepat
- **Semantic search:** Vector similarity untuk "related content"
- **Filters:** Platform, date range, sentiment, engagement metrics

Implementation Roadmap

Phase	Duration	Deliverables	Priority
Phase 1: MVP	2-3 weeks	WhatsApp bot integration, basic metadata extraction, simple dashboard	HIGH
Phase 2: Intelligence	2 weeks	Embedding generation, semantic clustering, topic labeling	HIGH
Phase 3: AI Summary	1-2 weeks	Auto-summary generation, sentiment analysis, key insights	MEDIUM
Phase 4: Advanced	2 weeks	Advanced search, notifications, PWA, analytics dashboard	LOW

Total Timeline: 7-9 minggu untuk full-featured platform.

Financial Projections

Infrastructure Costs (Monthly)

Component	Specification	Cost/Month
VPS Server	4 CPU, 8GB RAM, 100GB SSD	\$20-40
Database	PostgreSQL + Redis (self-hosted)	Included
Storage	MinIO local / S3-compatible	\$5-10
Total Infrastructure		\$25-50

API Usage Costs (per 1,000 links processed)

Service	Rate	Est. Cost/1K links
OpenAI Embeddings	\$0.02/1M tokens	~\$0.10
Whisper S2T	\$0.006/minute	~\$0.36 (avg 1 min/video)
LLM Summary (GPT-4o)	FREE (1M tokens)	~\$0.00-2.00
Total API Cost		\$2-5 per 1K links

Cost Optimization Strategy:

- **Qwen for Summarization:** 1 juta token gratis = summary tanpa biaya
- **Local Whisper:** Deploy local model untuk video transcription (higher upfront, zero ongoing cost)
- **Embedding Cache:** Jangan regenerate embeddings untuk konten yang sama
- **Lazy Loading:** Hanya proses media yang benar-benar diperlukan

Risk Assessment & Mitigation

Risk	Impact	Mitigation Strategy
Rate limiting (IG, TikTok)	High	Rotating proxies, exponential backoff, respect robots.txt, implement caching
Private/deleted content	Medium	Graceful degradation, cache metadata immediately upon ingestion
Multi-language content	Medium	Language detection, translate to English until embeddings
Video processing latency	Medium	Async queue processing, parallel workers, transcript caching
Spam/low quality content	Low	Quality scoring, minimum engagement threshold, user reporting

Next Steps

Immediate Actions (Upon Approval)

1. Setup development environment (Docker, repositories)
2. Configure VPS dan deploy initial infrastructure
3. Develop WhatsApp bot integration (Phase 1 MVP)
4. Build basic dashboard untuk demo

Start Simple, Iterate Fast. MVP dapat di-deploy dalam 2-3 minggu dengan fitur dasar: WhatsApp ingestion, metadata extraction, dan simple table view. Advanced features (clustering, AI summary) dapat ditambahkan iteratively berdasarkan user feedback.

Content Aggregation System - Business Proposal v1.0

Prepared for Investment Review | February 2026

For questions or technical clarifications, please contact the development team.