Plataforma Interna de Scraping y Análisis Competitivo de Precios (MVP)

Propuesta Técnica y de Negocio

Patrocinador: Luis Acosta / Dirección Estrategia
Equipo Responsable: Ruíz / Lozas

15 de octubre de 2025

✓ RECOMENDADO - GO

Índice

1. Resumen Ejecutivo

Construiremos una plataforma interna que extrae de forma programada información pública de competidores (precio, promoción, disponibilidad y vendedor), la normaliza y la disponibiliza en una interfaz simple y mediante descargas/API.

El MVP (3/4 meses) se acota a **3/4 competidores** y **2/3 categorías** con frecuencia diaria (o menor solo si es responsable y viable), y comparación por SKU con matching básico (exacto/variante). Sitios de alta fricción entran en Fase 2 con ventanas y frecuencia reducidas.

Beneficio Esperado

Habilitar playbooks de precio y reacción táctica ante movimientos de mercado, ganar trazabilidad histórica para negociar con marcas, reducir dependencia de un proveedor que hoy no cumple capacidades/SLAs.

Evaluación de Viabilidad Técnica

Status: ALTAMENTE VIABLE - Se recomienda proceder

Análisis del Proveedor Actual

- Cobertura deficiente: 54 % de productos sin competidor identificado
- Matching limitado: Solo coincidencias exactas, sin variantes
- Extracción rica: ∼50 atributos por producto (ventaja temporal)
- Sin trazabilidad: No hay evidencia de históricos robustos

Ventajas Competitivas del MVP

- **Matching mejorado:** Target 70-85 % vs 46 % actual
- **Frescura garantizada:** Datos ¡24h en ¿90 % SKUs
- **Control total:** Pipeline propio, sin dependencias
- **ROI positivo:** Break-even en 2-4 meses

2. Antecedentes y Problema

2.1. Situación Actual

- Dependencia de un tercero con resultados irregulares en frescura, cobertura y precisión
- Decisiones comerciales con latencia y poca evidencia histórica
- Necesitamos un pipeline propio con alcance realista, riesgos controlados y KPIs claros

2.2. Análisis de Datos del Proveedor Actual

Archivo 1 - analyse_item_list (107 productos)

- Extracción detallada con ∼50 atributos específicos por categoría
- Incluye: precio, descuento, marca, seller, disponibilidad, envío, planes EMI
- Atributos técnicos: capacidad, modelo, tecnología, certificaciones

Archivo 2 - exact_match_data (200 productos de electrónicos)

- 46 % productos "Out Of Stock"
- 54 % productos "No Competitor" ← Principal problema
- Solo matching básico (campo "Difference" = 0 en todos los casos)
- Categoría: Refrigeradores (Samsung 41, LG 31, Mabe 25, Whirlpool 24)
- Rango de precios: $4,599 - $91,999 MXN

Conclusión: El proveedor actual NO está cumpliendo con las expectativas de cobertura y matching.

3. Objetivo del MVP (3/4 meses)

1. Disponibilizar snapshots programados (diarios; sub-diarios cuando sea viable)
2. Comparar por SKU: precio, precio lista, % descuento, envío (si es visible), disponibilidad y tipo de vendedor (1P/3P)
3. UI simple (filtros, tabla comparativa, series de tiempo) + descargar (CSV/Parquet) + API interna + alertas por umbrales
4. Matching v1 (exacto/variante) usando claves duras (EAN/UPC/MPN/SKU) y reglas simples (pack/talla/color)

3.1. KPIs Objetivo vs Proveedor Actual

Métrica	Proveedor Actual	Target MVP	Mejora
Cobertura efectiva	∼46 %	≥70 %	+52 %
Frescura (¡24h)	¿Semanal?	≥90 %	✓
Precisión precio	¿?	≥97 %	✓
Disponibilidad sistema	¿?	≥97 %	✓

Cuadro 1: Comparativa de métricas: Proveedor Actual vs MVP

4. Alcance del MVP

4.1. Incluye

Funcionalidad Core

- Extracción de listados de búsqueda y páginas de producto en 3/4 competidores y 2/3 categorías
- Normalización de moneda, marca, pack/talla, categoría estándar
- Matching v1: exacto/variante
- UI con filtros, tabla comparativa, serie de tiempo y detalle con evidencias (URL y mini-captura)
- API/exports y alertas (p. ej., caída/subida de precio ¿x % o cambio de disponibilidad)

Atributos a Extraer (MVP)

1. Precio actual
2. Precio lista / precio tachado
3. % Descuento
4. Disponibilidad (In Stock / Out of Stock)
5. Vendedor (1P / 3P + nombre si aplica)
6. URL del producto
7. Marca
8. Categoría
9. SKU del competidor

4.2. No Incluye (Fase 2)

Fuera de Alcance MVP

- Similaridad avanzada (texto/imagen con ML)
- Share of search
- Ratings/reviews
- Cobertura masiva de todos los sitios y categorías
- Sitios con alta fricción (ej.: algunos marketplaces globales)
- **Atributos técnicos detallados** (capacidad, color, tecnología, etc.) ← El proveedor actual los tiene

4.3. Recomendación: Plan para Fase 2

Atributos técnicos (Semanas 13-20):

- Implementar extracción de especificaciones por categoría
- Usar selectores CSS específicos + plantillas configurables
- Considerar LLM API (GPT-4o/Claude) para extracción de atributos no estructurados
- **Prioridad:** Solo si el negocio lo requiere para decisiones comerciales

5. Casos de Uso Habilitados

5.1. Playbooks Operativos

- **Playbooks de precio:** detectar gaps y definir respuesta (mantener/igualar/contraatacar)
- **Oportunidad por OOS competidor:** cuando el competidor queda sin stock
- **Negociación con marcas:** evidencia histórica de movimientos de precio/promoción
- **Alertas operativas:** eventos relevantes para equipos comerciales

5.2. Casos de Uso Adicionales Identificados

Basados en el análisis de datos:

- **Detección de cambios de seller:** Competidor cambia de 1P a 3P
- **Alertas de reposición:** Competidor recupera stock después de OOS
- **Análisis de planes de financiamiento:** Si competidor ofrece MSI o planes EMI más agresivos
- **Monitoreo de envío:** Cambios en costos/tiempos de envío

6. Arquitectura del Sistema

6.1. Componentes Principales

La arquitectura propuesta se organiza en cinco capas principales:

1. Capa de Extracción

- Playwright/Puppeteer para renderizado de JavaScript
- Pool de workers con límites por dominio
- Proxy rotatorio (Bright Data/Smartproxy)
- Detección de cambios (hash estructura HTML)

2. Orquestación

- Prefect/Dagster/Airflow para planificación de tareas
- DAGs por competidor/categoría
- Retry logic inteligente
- Rate limiting por dominio

3. Procesamiento y Matching

- Normalización (precios, moneda, unidades)
- Matching v1: exacto (EAN/UPC) + variante
- Fuzzy matching (fuzzywuzzy/rapidfuzz)
- Validaciones (precios, outliers)

4. Almacenamiento

- Raw: S3/GCS + Parquet (fecha/competidor)
- Procesado: DuckDB/ClickHouse
- Histórico: retención 12-24 meses

5. Capa de Publicación

- API: FastAPI (endpoints REST)
- UI: Streamlit/Retool
- Alertas: Email/Slack

6.2. Recomendaciones Tecnológicas

Componente	**Recomendado**	**Justificación**
Scraping	Playwright	Mejor manejo de SPA, APIs más limpias
Orquestación	Prefect	Más moderno, mejor DX, Python-native
Storage (raw)	S3 + Parquet	Costo, integración
Storage (queries)	DuckDB	OLAP sobre Parquet, zero-config
API	FastAPI	Performance, docs automáticas, async
UI	Streamlit	Prototipado rápido, Python-only
Proxies	Bright Data	Mejor uptime, más IPs residenciales
Matching	rapidfuzz	Más rápido que fuzzywuzzy
Monitoring	Grafana + Loki	Open-source, menor costo

Cuadro 2: Stack tecnológico recomendado

7. Modelo de Datos

7.1. Dimensiones (Catálogos)

Producto Canónico

```
dim_producto:
  - producto_id (PK)
  - ean / upc / mpn
  - marca
  - familia
  - clase
  - departamento
  - pack / talla / color
  - categoria_estandar
```

Competidor

```
dim_competidor:
  - competidor_id (PK)
  - nombre
  - dominio
  - tipo (retailer/marketplace)
  - limite_rpm (rate limit)
```

Vendedor

```
dim_vendedor:
  - vendedor_id (PK)
  - nombre
  - tipo (1P/3P)
  - competidor_id (FK)
```

7.2. Hechos (Mediciones)

Precio Histórico

```
fact_precio:
  - precio_id (PK)
  - producto_id (FK)
  - competidor_id (FK)
  - vendedor_id (FK)
  - fecha_id (FK)
  - precio_actual
  - precio_lista
  - descuento_monto
  - descuento_porcentaje
  - disponibilidad (boolean)
  - moneda
```

```
  - url
  - timestamp_extraccion
  - hash_evidencia (mini-captura)
```

7.3. Tabla de Matching

```
rel_matching:
  - matching_id (PK)
  - producto_canonico_id (FK)
  - competidor_id (FK)
  - sku_competidor
  - tipo_match (exacto/variante/fuzzy)
  - score_similitud (0-100)
  - fecha_validacion
  - validado_manualmente (boolean)
```

Tipos de match:

- **Exacto:** EAN/UPC/MPN coincide 100 %
- **Variante:** Mismo producto, diferente talla/color/pack
- **Fuzzy:** Similitud ¿85 % en nombre normalizado (Fase MVP tardía)

8. KPIs y Criterios de Éxito (SLOs)

8.1. Métricas Core del MVP

KPI	Target MVP	Método de medición
Cobertura	≥70 % de SKUs prioritarios con ≥1 coincidencia	`COUNT(DISTINCT matched_skus) / COUNT(golden_set)`
Frescura	≥90 % de SKUs con datos de las últimas 24h	`COUNT(WHERE timestamp >NOW()-24h) / total_skus`
Precisión (precio/promo)	≥97 % en muestreo estratificado	Validación manual semanal (50 SKUs)
Disponibilidad de procesos	≥97 %	Uptime de procesos de extracción
Tiempo de recuperación ante cambio de página	¡24h	MTTR desde detección hasta fix

Cuadro 3: Métricas y objetivos del MVP

8.2. Definición de Éxito del MVP

Cumplir estos indicadores en ≥**2 competidores** y ≥**2 categorías** durante **4 semanas continuas**.

8.3. Comparativa con Proveedor Actual

Aspecto	**Proveedor Actual**	**Target MVP**	**Mejora**
Matching efectivo	∼46 %	≥70 %	**+52 %**
Frescura	Semanal (?)	Diaria (¿90 %)	✓
Precisión	Desconocida	≥97 %	✓
MTTR cambios	Desconocido	¡24h	✓
Trazabilidad	Limitada	Completa	✓

Cuadro 4: Comparativa de mejoras respecto al proveedor actual

9. Plan de Trabajo (12 semanas)

9.1. Semanas 0-2: Descubrimiento y Base

Objetivos:

- Validar competidores y categorías; levantar golden set (500/1,000 SKUs)
- Diseñar modelo de datos y mockups de UI; revisar términos y robots por dominio
- Definir KPIs y "Definition of Done"

Entregables:

☐ Lista de 3-4 competidores aprobada

☐ Lista de 2-3 categorías con palabras de búsqueda

☐ Golden set con EAN/UPC/MPN cuando exista

☐ Mockups de UI

☐ Matriz legal (ToS y robots.txt por dominio)

Quick win: Presentación de mockups a usuarios finales

9.2. Semanas 3-4: Infraestructura y Primer Sitio

Objetivos:

- Configurar repos, planificador, bitácoras y almacenamiento
- Implementar primer sitio (búsquedas + producto) y normalización
- UI v0 (tabla + filtros) y descarga básica

Entregables:

☐ Repo configurado + CI/CD básico

☐ Primer scraper funcional (1 competidor, 1 categoría)

☐ Storage en Parquet funcionando

☐ UI v0 con tabla comparativa básica

Hito crítico (Semana 4): Demo funcional con datos reales de 1 competidor

9.3. Semanas 5-6: Más Sitios y Matching

Objetivos:

- Implementar sitio 2 y sitio 3; control de calidad por muestreo
- Matching v1 (exacto/variante); API y UI v1 (serie de tiempo)

Entregables:

☐ 3 competidores scraped diariamente

☐ Matching exacto por EAN/UPC

☐ Matching de variantes (talla/color/pack)

☐ API endpoints básicos

☐ UI con serie de tiempo

KPI checkpoint: Cobertura ¿50 % en golden set

9.4. Semanas 7-8: Robustez y Alertas

Objetivos:

- Detector de cambios de página; reintentos inteligentes; alertas por umbrales
- Demostración con usuarios comerciales y ajustes

Entregables:

☐ Detector de cambios de estructura

☐ Sistema de alertas configurables

☐ Retry logic inteligente

☐ Demo con usuarios comerciales

☐ Feedback documentado

Validación de usuarios: 5 usuarios clave validan la plataforma

9.5. Semanas 9-12: Ampliación y Cierre del MVP

Objetivos:

- Sumar categoría #2 (y #3 si aplica); pruebas de carga
- Monitoreo de KPIs 4 semanas; decisión Go/No-Go y backlog de Fase 2

Entregables:

☐ 2-3 categorías completamente operativas

☐ 4 semanas continuas cumpliendo KPIs

☐ Documentación completa

☐ Plan de Fase 2

☐ Decisión Go/No-Go para escalamiento

Decisión final: ¿Proceder con Fase 2 o ajustar?

10. Equipo Mínimo y Roles

Rol	Dedicación	Responsabilidades
Líder técnico / Datos Sr	1.0 FTE	Arquitectura, orquestación, observabilidad, robustez
Ingeniero/a de datos	1.0 FTE	Extracción, normalización, procesos, matching
Ingeniero/a back/-frontend	1.0 FTE	API, autenticación, interfaz, descargas, alertas
PM/PO	0.75 FTE	Roadmap, riesgos, relación con usuarios y sponsors
TOTAL	**3.75 FTE**	**Equipo MVP**

Cuadro 5: Equipo y dedicación requerida

Fase 2 (opcional): +0.5–1.0 analista/ML para similaridad avanzada

11. Riesgos y Mitigaciones

#	Riesgo	Prob.	Impacto	Mitigación
1	Defensas anti-extracción en sitios	Alta	Alto	Proxies residenciales + cadencia conservadora + headers realistas
2	Cambios en estructura de páginas	Media	Alto	Detector de cambios automático. MTTR ¡24h
3	Calidad de matching baja	Media	Medio	Golden set robusto (500-1K SKUs). Fuzzy matching

#	Riesgo	Prob.	Impacto	Mitigación
4	Cumplimiento legal/operativo	Baja	Alto	Matriz por dominio (ToS y robots.txt). Solo info pública
5	Costo de conectividad/IPs	Media	Medio	Medición por dominio. Budget cap mensual
6	Scope creep (features adicionales)	Alta	Medio	**Stick to MVP.** Decir NO a features
7	Cambios frecuentes en marketplaces	Alta	Alto	Empezar con retailers directos
8	Equipo no completo a tiempo	Media	Alto	Pre-asignar equipo antes de Sprint 0

Cuadro 6: Matriz de riesgos y estrategias de mitigación

12. Costos y ROI

12.1. Inversión Inicial (CAPEX)

- Desarrollo del MVP: 3.75 FTE × 3 meses × $10,000 USD promedio
- **Total CAPEX: ~$112,500 USD**

12.2. Costos Operativos (OPEX Mensual)

Concepto	Conservador	Base	Ambicioso
Cómputo/workers	$150	$300	$500
Almacenamiento (S3/GCS)	$50	$100	$200
Proxies/IPs rotatorias	$300	$800	$1,500
Monitoreo (logs/métricas)	$50	$100	$150
Orquestación (Prefect Cloud)	$0	$0	$200
Contingencia (10 %)	$55	$130	$255
TOTAL MENSUAL	**$605**	**$1,430**	**$2,805**
TOTAL ANUAL	**$7,260**	**$17,160**	**$33,660**

Cuadro 7: Costos operativos en diferentes escenarios

12.3. Comparativa Financiera

Concepto	Proveedor Actual	MVP Año 1	MVP Año 2+
Costo total	$36,000-96,000	$129,660	$17,160
Control	Nulo	Total	Total
Cobertura	~46 %	70-85 %	70-85 %+

Cuadro 8: Análisis financiero comparativo

Break-even: Si el proveedor cobra ¿$5,000/mes → **ROI positivo en 2-4 meses**

Ahorro anual (desde Año 2): $36,000-96,000 - $17,160 = **$18,840-78,840 USD/año**

13. Gobierno, Auditoría y Seguridad

13.1. Gobierno del Proyecto

Comité quincenal:

- Sponsor (Luis Acosta)
- PM/PO
- Líder técnico
- Representante usuarios comerciales
- Representante legal (Q&A sobre compliance)

Agenda estándar:

- Avance vs plan (semáforo)
- KPIs actuales
- Riesgos top 3
- Decisiones requeridas
- Budget burn rate

13.2. Seguridad

Acceso

- SSO corporativo (Azure AD / Okta)
- RBAC (roles: admin, comercial, analista, auditor)
- MFA obligatorio para admins

Datos

- Cifrado en tránsito (TLS 1.3)
- Cifrado en reposo (S3 server-side encryption)
- Sin PII de clientes
- Logs de acceso a datos sensibles

Compliance

- Revisión legal de ToS por dominio (cada 6 meses)
- Respeto de robots.txt
- Rate limiting documentado
- Proceso de opt-out si un competidor lo solicita

14. Decisiones Solicitadas al Sponsor

14.1. Decisiones Críticas (Esta Semana)

1. ✓ **Aprobación del alcance del MVP**
2. ○ **Definición de competidores y categorías** (deadline: esta semana)
   - Propuesta: Liverpool, Elektra, Palacio de Hierro
   - Categorías: Línea Blanca, Electrónicos
3. ○ **Inicio del Sprint 0** (2 semanas): descubrimiento, legal y base técnica
4. ○ **Primer hito visible** (Semana 4): tabla comparativa funcional con 1 sitio

14.2. Decisiones Secundarias (2 Semanas)

5. Plan de comunicación a equipos comerciales
6. Proceso de feedback de usuarios durante MVP
7. Criterios de éxito para aprobar Fase 2

15. Próximos Pasos Inmediatos

15.1. Esta Semana

1. **Sponsor aprueba documento y presupuesto**
2. **Definir competidores finales** (Liverpool, Elektra, Palacio + 1?)
3. **Definir categorías finales** (Línea Blanca, Electrónicos + 1?)
4. **Pre-asignar equipo** (4 personas con nombres)

15.2. Semana Próxima (Semana 0 - Inicio)

5. **Kickoff del proyecto** con equipo completo
6. **Construir golden set** (500 SKUs prioritarios)
7. **Revisar ToS y robots.txt** de competidores
8. **Setup técnico inicial** (repo, cloud, proxies trial)

15.3. Semana 4 (Primer Hito)

9. **Demo funcional** con datos reales de 1 competidor
10. **Validación con usuarios** (5 personas comerciales)

16. Glosario

Extracción (scraping)	Obtención automatizada de información pública mostrada en páginas web
Snapshot	Captura de estado (precio/stock) en una fecha/hora específica
SKU canónico	Identificador interno para comparar "manzanas con manzanas" entre competidores
Matching exacto	Mismo producto identificado por EAN/UPC/MPN
Matching variante	Misma referencia con diferencias (talla, color, pack)
Matching fuzzy	Similitud aproximada basada en texto (¿85 % similitud)
1P (First Party)	Producto vendido directamente por el retailer
3P (Third Party)	Producto vendido por un seller externo en marketplace
Golden set	Conjunto de SKUs prioritarios para QA y validación
MTTR	Mean Time To Recovery - tiempo promedio de recuperación ante fallas
SLO	Service Level Objective - objetivo de nivel de servicio

Recomendación Final

✓ RECOMENDACIÓN: GO

Este MVP es técnicamente viable, financieramente justificable, y estratégicamente necesario dado el desempeño deficiente del proveedor actual (54 % sin matching).

Confianza en éxito: Alta (80 %)
Riesgo principal: Scope creep y anti-bot en sitios complejos
Mitigación clave: Disciplina en MVP, empezar con retailers simples

¡Es momento de construir!

Firma de Aprobación

Rol	**Nombre**	**Aprobación**	**Fecha**
Sponsor	Luis Acosta	☐ Apruebo	__/__/__
Tech Lead	Lozas	☐ Apruebo	__/__/__
PM	Ruíz	☐ Apruebo	__/__/__
Legal	[Nombre]	☐ Apruebo	__/__/__

Versión 0.2 - 15 de octubre de 2025
Próxima revisión: Semana 4 (hito de demo funcional)