Plataforma Interna de Scraping

Análisis Competitivo de Precios — MVP

Versión:	0.1 (para socializar con equipo y usuario final)
Fecha:	15/oct/2025
Patrocinador:	Luis Acosta / Dirección Estrategia
Equipo responsable:	Ruíz / Lozas

Índice

1 Resumen

- **Objetivo:** Plataforma interna que extrae información pública de competidores (**precio**, **promoción**, **disponibilidad**, **vendedor**), la normaliza y la disponibiliza en interfaz simple y mediante descargas/API.
- **Alcance MVP: 3–4 meses**, **3–4 competidores**, **2–3 categorías**, frecuencia **diaria** (sub-diaria cuando sea viable y responsable). Comparación por SKU con *matching* básico (**exacto/variante**). Sitios de alta fricción entran en **Fase 2**.
- **Beneficios esperados:**
  - Habilitar *playbooks* de precio y reacción táctica ante movimientos de mercado
  - Ganar trazabilidad histórica para negociar con marcas
  - Reducir dependencia de proveedor actual que no cumple capacidades/SLAs

2 Antecedentes y problema

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

3 Objetivo del MVP (3–4 meses)

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

4 Alcance del MVP (incluye / no incluye)

INCLUYE

- 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)

NO INCLUYE

- Similaridad avanzada (texto/imagen), *share of search*, *ratings/reviews*
- Cobertura masiva de todos los sitios y categorías
- Sitios con alta fricción (ej.: *marketplaces* globales) ← podría entrar en **Fase II**

5 Casos de uso habilitados

- ***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

6 Requerimientos y supuestos

- **Competidores iniciales:** 3–4 (definir con negocio)
- **Categorías iniciales:** 2–3 (definir con negocio)
- **SKUs canónicos priorizados:** 5,000–15,000 (con *golden set* inicial de **500–1,000** para QA)
- **Frecuencia:** Diaria 01:00/06:00; menor a diario solo si el sitio lo permite y es responsable
- **Cumplimiento:** Solo información pública; revisión de términos y robots.txt por dominio; límites de ritmo por sitio
- **Anti-bot:** Cadencia conservadora, variación de agente de navegador y rotación de IP dentro de límites prudentes
- **Seguridad:** Sin PII; acceso por roles; registro de fuente (URL) y *timestamp* por dato

7 Entrega de datos y disponibilidad

- **Interfaz web:** Filtros (competidor, categoría, marca, SKU, fecha), tabla comparativa y tendencias
- **Descarga:** CSV/Parquet del resultado filtrado
- **API interna:** Endpoints para consulta programática
- **Alertas:** Reglas simples (condición, umbral, destinatarios) con historial
- **Acuerdos operativos:** Ventana de actualización nocturna; tiempos de reintento ante fallas

8 Arquitectura

- **Extracción:** Automatización de navegador (capaz de cargar páginas dinámicas), con plantillas por sitio y reintentos
- **Orquestación:** Planificador de tareas y cola de trabajos con límites por dominio
- **Procesamiento:** Limpieza, normalización y validaciones (precios, moneda, disponibilidad)
- **Almacenamiento:** Archivos columnados (Parquet/Delta) por fecha/sitio y repositorio analítico (tipo *warehouse*) para consultas
- **Publicación:** API interna, descargas y UI
- **Observabilidad:** Bitácoras, métricas, mini-capturas como evidencia y tableros de salud

9 Modelo de datos

Dimensiones (catálogos):

- Producto canónico (marca, familia, clase, departamento, pack/talla/color)
- Sitio/competidor
- Vendedor (1P/3P)
- Categoría estándar
- Tiempo (fecha, semana, mes, año)

Hechos (mediciones):

- **Punto de precio:** precio, precio lista, %descuento, costo de envío (si aplica), disponibilidad, vendedor, moneda, URL, fecha

- **Meta del proceso:** estado de extracción, código de respuesta, latencia, presencia de *captcha* (para soporte)

Vínculo de *matching* (MVP):

- Relación entre producto canónico y producto en sitio con tipo de coincidencia: **exacta** o **variante**

10 KPIs y criterios de éxito (SLOs)

Métricas operacionales:

- **Cobertura:** $\geq 85\,\%$ de SKUs prioritarios con ≥ 1 coincidencia
- **Frescura:** $\geq 95\,\%$ de SKUs con datos de las últimas **24 h**
- **Precisión (precio/promo):** $\geq 97\,\%$ en muestreo estratificado
- **Disponibilidad de procesos:** $\geq 97\,\%$
- **Tiempo de recuperación ante cambio de página:** < 24 h

Definición de éxito del MVP:

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

11 Plan de trabajo (12 semanas)

Fase	Hitos y tareas clave
Semanas 0–2 **Descubrimiento**	**Descubrimiento y base** • Validar competidores y categorías; levantar *golden set* (**500–1,000 SKUs**) • Diseñar modelo de datos y *mockups* de UI • Revisar términos y robots.txt por dominio • Definir KPIs y "Definition of Done"
Semanas 3–4 **Base técnica**	**Infraestructura y primer sitio** • Configurar repos, planificador, bitácoras y almacenamiento • Implementar **primer sitio** (búsquedas + producto) y normalización • UI v0 (tabla + filtros) y descarga básica
Semanas 5–6 **Expansión**	**Más sitios y *matching*** • Implementar sitio 2 y sitio 3; control de calidad por muestreo • *Matching* v1 (exacto/variante) • API y UI v1 (serie de tiempo)
Semanas 7–8 **Consolidación**	**Robustez y alertas** • Detector de cambios de página; reintentos inteligentes • Alertas por umbrales • Demostración con usuarios comerciales y ajustes
Semanas 9–12 **Cierre MVP**	**Ampliación y cierre del MVP** • Sumar categoría #2 (y #3 si aplica); pruebas de carga • **Monitoreo de KPIs durante 4 semanas** • **Decisión Go/No-Go** y backlog de Fase 2

12 Equipo mínimo y roles

- **Líder técnico / Datos Sr (1.0 FTE):** Arquitectura, orquestación, observabilidad, robustez
- **Ingeniero/a de datos (1.0 FTE):** Extracción, normalización, procesos
- **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 estimado: 3.75 FTE para el MVP. (Fase 2 podría requerir **0.5–1.0 FTE** analista/ML para similaridad avanzada).

13 Riesgos y mitigaciones

- **Defensas anti-extracción en algunos sitios**
  - *Mitigación:* Límites por dominio, horarios valle, pausas automáticas ante bloqueos, ventanas reducidas en sitios complejos
- **Cambios en estructura de páginas**
  - *Mitigación:* Detector de cambios, plantillas por sitio y guías de respuesta; **MTTR** < 24 **h**
- **Calidad de comparación por SKU**
  - *Mitigación: Golden set* y muestreo estratificado por categoría/competidor; reglas claras de variante
- **Cumplimiento legal/operativo**
  - *Mitigación:* Matriz por dominio (términos y `robots.txt`), solo info pública y ritmos responsables
- **Costo de conectividad/servicios (IP rotatoria)**
  - *Mitigación:* Medición por dominio (GB/mes), priorización de categorías, cacheo selectivo

14 Costos (*drivers* y escenarios)

Componentes principales:

- **Cómputo/almacenamiento:** Procesos nocturnos y archivos columnados; escala según páginas/día
- **Conectividad/IPs rotatorias:** Principal *driver*; depende del peso de las páginas y frecuencia
- **Monitoreo/observabilidad:** Bitácoras, métricas y alertas

Escenarios de alcance:

- **Conservador: 2 competidores** $\times$ **2 categorías**; frecuencia diaria
- **Base: 3–4 competidores** $\times$ **2–3 categorías**; diaria (menor a diaria selectiva)
- **Ambicioso: 4 competidores** $\times$ **3 categorías**; diaria/menor a diaria, mayor costo de conectividad

15 Gobierno, auditoría y seguridad

- **Comité quincenal:** Avance, riesgos, decisiones
- **Tablero de salud:** Cobertura, frescura, precisión, fallos por sitio, consumo de conectividad

- **Auditoría:** Cada registro con URL y *timestamp*; evidencias (mini-capturas) en casos críticos
- **Seguridad:** Acceso por roles (SSO), cifrado en tránsito y en reposo, registros de acceso

16 *Checklist* para arrancar

☐ Lista de competidores (**3–4**) y categorías (**2–3**), con prioridad

☐ *Golden set* (**500–1,000 SKUs canónicos**) con EAN/UPC/MPN cuando exista

☐ Palabras de búsqueda por categoría (cómo busca el cliente)

☐ Posición legal interna: límites de frecuencia/uso por dominio

☐ Usuarios clave para validar la interfaz y los *playbooks* (nombres y correos)

17 Decisiones solicitadas al *sponsor*

1. **Aprobación del alcance del MVP** (secciones 3, 4 y 8/11)
2. **Definición de competidores y categorías esta semana**
3. **Inicio del Sprint 0 (2 semanas):** Descubrimiento, legal y base técnica
4. **Primer hito visible (Semana 4):** Tabla comparativa funcional con un sitio

18 Glosario breve

- **Extracción (*scraping*):** Obtención automatizada de información pública mostrada en páginas
- ***Snapshot*:** Captura de estado (precio/stock) en una fecha/hora
- **SKU canónico:** Identificador interno para comparar "manzanas con manzanas"
- ***Matching* exacto/variante:** Igual producto; variante = misma referencia con diferencias (talla, color, pack)
- **Sitio de alta fricción:** Página con defensas técnicas que requieren ritmos más bajos o ventanas acotadas
- **MVP (*Minimum Viable Product*):** Producto mínimo viable con funcionalidad básica para validar valor
- **SLO (*Service Level Objective*):** Objetivo de nivel de servicio; métrica cuantificable de calidad operacional
- **MTTR (*Mean Time To Repair*):** Tiempo promedio de recuperación ante fallas o cambios

- **1P/3P:** Primera parte (vendedor directo/retailer) vs. tercera parte (marketplace/vendedor externo)
- **EAN/UPC/MPN:** Códigos de producto estándar (European/Universal Article Number, Manufacturer Part Number)
- ***Golden set*:** Conjunto de referencia de SKUs prioritarios para calibración y validación
- **Normalización:** Proceso de estandarizar formatos (moneda, unidades, categorías) para comparación
- ***Warehouse* analítico:** Repositorio centralizado de datos estructurados para consultas y análisis
- **Cobertura:** Porcentaje de SKUs objetivo con datos disponibles en competidores
- **Frescura:** Antigüedad de los datos; qué tan reciente es la última actualización
- ***Gap* de precio:** Diferencia de precio entre producto propio y competencia
- **Elasticidad:** Sensibilidad de la demanda ante cambios de precio
- ***Captcha*:** Mecanismo de seguridad para distinguir humanos de bots; puede bloquear extracción
- ***Rate limiting*:** Límite de frecuencia de solicitudes por tiempo para evitar sobrecarga o bloqueo
- **Parquet/Delta:** Formatos de archivo columnar optimizados para consultas analíticas a gran escala

19 Anexo A – Variables para estimar ROI (llenar con negocio)

- **Ventas de las categorías objetivo:** % del total
- **Unidades/SKUs críticos:** Ticket promedio
- **Elasticidad de precio:** Aproximado
- **% de tiempo con *gap* detectable:** vs. competidor
- **% de mejora de margen/venta al reaccionar:** *bps*