Anexo 2.3 — Diccionario de Datos y Normalización (ComparaYa)

*Versión 0.2 · Fecha: 2025-09-17 · Autores: Equipo ComparaYa*

# 1. Objetivo

Definir campos maestros, reglas de significado y política de normalización para integrar datasets de Consumidor y Mayorista. Se aborda la variabilidad entre “precio promedio” vs “precio promedio ponderado” y la heterogeneidad en nombres de productos con separadores como '|' o ' | '.

# 2. Campos Maestros (Esquema Normalizado)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Campo | Tipo | Descripción | Valores/Ejemplos | Origen/Regla |
| dataset | enum | Fuente de datos | consumidor | mayorista | Asignado por ETL |
| fecha | date | Fecha del dato | YYYY-MM-DD | Mapeo desde columna fecha/periodo |
| producto\_raw | string | Nombre del producto tal como viene | 'Tomate | primera | granel' | Columna producto original |
| producto\_base | string | Nombre base normalizado | 'Tomate' | Token 1 tras dividir por '|' y limpiar |
| producto\_calidad | string | Calidad/variedad si existe | 'primera' | Token 2 si existe |
| producto\_presentacion | string | Formato/presentación | 'granel' | Token 3 si existe |
| region | string | Región administrativa | 'Región Metropolitana' | Directo si existe; si no, 'Todas' |
| mercado | string | Mercado mayorista (si aplica) | 'Lo Valledor' | Solo para Mayorista |
| unidad\_raw | string | Unidad declarada en fuente | 'kilo','caja 18 kg' | Columna de unidad |
| unidad\_base | enum | Unidad normalizada | kilo|litro|unidad | Tabla de equivalencias |
| factor\_conversion | float | Multiplicador a unidad\_base | 18.0 (para 'caja 18 kg') | Tabla de equivalencias |
| precio\_valor | decimal(12,4) | Precio numérico | 1000.25 | Columna seleccionada |
| precio\_tipo | enum | Tipo de precio | promedio | promedio\_ponderado | mínimo | máximo | Detectado por nombre de columna |
| precio\_fuente | string | Nombre de columna original | 'precio promedio ponderado' | Columna origen |
| moneda | enum | Moneda del valor | CLP | Constante |

# 3. Normalización de Productos

• Separador canónico: '|' (sin espacios). Durante ETL: reemplazar todas las variantes '\s\*\|\s\*' → '|'.  
• Tokenización: producto\_raw.split('|') → [base, calidad, presentación, ...]; trim; lowercase salvo nombres propios.  
• producto\_base: catálogo controlado con alias → canónico (ej.: 'tomate ensalada'→'tomate').  
• Si no hay separador, usar todo el texto como producto\_base.

# 4. Reglas de Precio (Promedio y Promedio Ponderado)

• Detección: si existe columna que contenga 'ponderado', usarla y etiquetar precio\_tipo='promedio\_ponderado'. Si no, buscar 'promedio'; en ausencia, usar 'precio'/'valor' y etiquetar como 'promedio'.

• Definiciones: promedio = media aritmética; promedio\_ponderado = sum(w\_i·x\_i)/sum(w\_i) con pesos w (p. ej., volumen).

• Comparabilidad: marcar 'flag\_ponderado' cuando aplique; mostrar tooltip cuando un período mezcla metodologías.

# 5. Equivalencias de Unidades (ejemplos)

|  |  |  |  |
| --- | --- | --- | --- |
| Unidad\_fuente | Equivalencia\_valor | Unidad\_base | Notas |
| kilo | 1 | kilo | — |
| litro | 1 | litro | — |
| unidad | 1 | unidad | pieza unitaria |
| caja 18 kg | 18 | kilo | factor\_conversion=18 |
| caja 10 kg | 10 | kilo | factor\_conversion=10 |

# 6. Reglas de Transformación (ETL)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regla\_ID | Descripción | Fórmula/Algoritmo | Prioridad | Ejemplo |
| R1 | Normalizar separador en producto | re.sub(r'\s\*\|\s\*','|',producto) | Alta | 'Tomate | primera'→'Tomate|primera' |
| R2 | Tokenizar producto | split('|') → base,calidad,presentación | Alta | 'Tomate|primera|granel' |
| R3 | Mapear unidad a base | tabla equivalencias → (unidad\_base,factor) | Alta | 'caja 18 kg'→(kilo,18) |
| R4 | Seleccionar columna de precio | ponderado>promedio>precio/valor | Alta | si hay ponderado, usarlo |
| R5 | CLP numérico | parsear string a decimal; remover separadores | Alta | '$ 1.234,56'→1234.56 |
| R6 | Completar campos faltantes | mercado solo mayorista; región en ambos | Media | si falta, usar 'Todas' |

# 7. Controles de Calidad de Datos (DQC)

• fecha no nula y dentro de rango (Alta)

• precio\_valor > 0 (Alta)

• unidad\_base presente y factor\_conversion > 0 cuando aplique (Alta)

• producto\_base no vacío (Alta)

• sin duplicados exactos (dataset,fecha,producto\_base,region/mercado,unidad\_base) (Alta)

• flag\_ponderado consistente con precio\_fuente (Media)

# 8. Catálogos y Alias de Producto (plantilla)

|  |  |  |
| --- | --- | --- |
| Alias detectado | Canónico (producto\_base) | Notas |
| tomate ensalada | tomate | Unificar puntuación y separadores |
| tomate|primera | tomate | Unificar puntuación y separadores |
| tomate primera | tomate | Unificar puntuación y separadores |
| tomate (primera) | tomate | Unificar puntuación y separadores |

# 9. Trazabilidad a ERS

Respaldos: RF04–RF10 (ETL, API, manejo de datos) y RNF03–RNF10 (portabilidad, mantenibilidad, calidad).