

Diseño de esquema y pipelines

Table of contents

1 Ubicación	1
2 Esquema actual	1
3 Cómo agregar un nuevo dominio	2
4 Diagrama conceptual	3
4.1 Detalle: Presupuestos DIPRES (silver)	3
4.2 Consideraciones de modelado	3

1 Ubicación

- **SQL base:** alembic/versions/ (migraciones versionadas).
- **Plantillas:** data/meta/schema/template.sql (base para nuevos dominios).
- **Pipelines Python:** etl/pipelines/<tema>/<dominio>/.
- **Seeds:** etl/seeds/ (e.g. minimum.py).

2 Esquema actual

Tabla	Descripción	Fuente
dim_region	Dimensión regiones de Chile	Seed manual
dim_fecha	Dimensión fecha (2020-2030)	Seed manual
etl_log	Bitácora de ejecuciones de pipelines	Auto-mantenida
presupuestos	Catálogo de archivos DIPRES normalizados	DIPRES (raw → silver)
presupuesto_partidas	Partidas presupuestarias por presupuesto	presupuestos
presupuesto_capitulos	Capítulos asociados a cada partida	presupuesto_partidas
presupuesto_programas	Programas presupuestarios	presupuesto_capitulos

Tabla	Descripción	Fuente
presupuesto_subtitulos	Subtítulos con clasificación por naturaleza	presupuesto_programas
presupuesto_items	Items presupuestarios	presupuesto_subtitulos
presupuesto_asignaciones	Asignaciones específicas (nivel hoja)	presupuesto_items
presupuesto_ejecuciones	Totales agregados por moneda	presupuestos
presupuesto_clasificacion	Subtítulos y subclases (ingresos/gastos)	data/meta/diccionarios/dipres_subtitulos.
meta_datasets	Catálogo maestro de datasets publicados	data/meta/catalog/*.yaml
meta_files	Inventario de archivos raw y hashes	data/meta/dipres_files_metadata.csv
meta_transforms	Transformaciones registradas (pipelines)	ETL (Alembic 202510090003)
meta_tables	Tablas publicadas (schema, layer, tamaño)	ETL (bronze/silver/gold)
meta_columns	Columnas documentadas con metadatos	ETL (meta_tables)
meta_lineage	Linaje parent→child con referencia a runs	ETL (meta_transforms, jobs_runs)
meta_data_quality	Resultados de checks automáticos por tabla	QA (dbt/SQL, jobs_runs)
jobs_runs	Ejecuciones de pipelines / tareas	Orquestación/ETL
jobs_integrations	Estado de integraciones externas	Orquestación/ETL
monitor_api_metrics	Historial de métricas HTTP (p50/p95/p99)	Middleware API
monitor_loader_metrics	Métricas de ingestión (bytes, filas, fallos)	Pipelines / monitoreo

3 Cómo agregar un nuevo dominio

1. Duplicar data/meta/schema/template.sql y ajustar nombres.
2. Crear migración Alembic:

```
.venv/bin/alembic revision -m "add <dominio>" --autogenerate
.venv/bin/alembic upgrade head
```

3. Registrar dataset en data/meta/catalog/<slug>.yaml.
4. Implementar pipeline en etl/pipelines/<tema>/<dominio>/pipeline.py usando CSVToTable o heredando Pipeline.
5. Añadir pruebas en tests/data/ y tests/api/.

6. Documentar en este archivo y en `api_catalog.md`.

4 Diagrama conceptual

```
raw → bronze → silver → gold
                        API / dashboards
                    ref/meta  seeds (dimensiones)
```

Cada transición debe registrarse en `data/meta/lineage/*.yaml` para asegurar trazabilidad.

4.1 Detalle: Presupuestos DIPRES (silver)

- Raw: `data/raw/dipres/ejecucion_total`, `data/raw/dipres/proyecto_ley` (ignorado en Git).
- Manifiesto: `data/meta/dipres_files_metadata.csv`.
- Tasas de cambio: `data/ref/dipres_exchange_rates.csv`.
- Diccionario de subtítulos: `data/meta/diccionarios/dipres_subtitulos.yaml`.
- Código ETL: `etl/pipelines/dipres_budget.py`.
- DDL de referencia: `data/meta/schema/dipres_presupuestos.sql`.
- Catálogo: `data/meta/catalog/dipres_presupuestos.yaml`.
- API: `/api/{version}/presupuestos/...` (ver docs/api_catalog.md).
- Raw API: `/api/{version}/datasets/{slug}/raw-assets` para `dipres_ejecucion_total`, `dipres_proyecto_ley`, `dipres_ley_presupuesto` y el agregado `dipres_presupuestos`.

Jerarquía resumen:

```
Presupuesto (slug)
  Partidas (ministerios / órganos superiores)
    Capítulos (servicios dependientes)
      Programas (objetivos específicos)
        Subtítulos (naturaleza del gasto / ingreso)
          Ítems
            Asignaciones (nivel hoja)
  Totales por moneda, variante y período
```

4.2 Consideraciones de modelado

- Se excluye **Partida 50 (Tesoro Público)** para evitar doble contabilización de “Ingresos Generales de la Nación” y “Servicio de la Deuda Pública”.
- Los subtítulos se clasifican en **ingreso**, **gasto**, **mixto** o **desconocido** siguiendo el glosario DIPRES (01–15 ingresos, 21–35 gastos).
- Montos procesados son **nominales** (moneda corriente). Aún no se calculan montos reales/deflactados.
- Para ejecuciones históricas sólo se publican acumulados en marzo/junio/septiembre/diciembre; el pipeline normaliza estos trimestres a los meses equivalentes.

- Se mantienen las jerarquías para análisis institucional (partida/capítulo/programa) y la clasificación económica (subtítulo/ítem/asignación).