

Automatización de Series de Datos de Empleo (Sábanas)

Automatización de Series de Datos de Empleo IMSS

Objetivo del Proyecto

Este proyecto tiene como objetivo automatizar la generación y actualización de las series de datos conocidas como “Sábanas” del IMSS. El sistema permite actualizar las series de tiempo de empleo de manera eficiente y automatizada.

Estado Actual del Proyecto



Completado

- Automatización de desagregaciones de datos
- Flujo de trabajo con targets funcional
- Proceso de actualización de series mediante `tar_make()`



Pendiente

- Generación de resúmenes estadísticos (como los de las sábanas actuales)
- Deployment de la aplicación Shiny en servidor de producción
- Automatización completa del pipeline mensual
- Implementación de pruebas de validación automática

Definiciones y Metodología

Conceptos Principales

Series de Tiempo Generadas

1. Asegurados Trabajadores

- **Por Modalidad:** 11 modalidades del régimen obligatorio (10, 11, 12, 13, 14, 17, 32, 33, 34, 38, 97)
- **Por Sector Económico:** 21 divisiones de actividad económica según clasificación IMSS
- **Universo:** Trabajadores con relación laboral vigente al cierre del mes

2. Patrones (Empleadores)

- **Por Modalidad:** 6 modalidades principales con al menos un trabajador activo
- **Por Rango de Trabajadores:** 6 categorías por tamaño de empresa
 - 1 trabajador
 - 2-5 trabajadores
 - 6-10 trabajadores
 - 11-50 trabajadores
 - 51-500 trabajadores

- Más de 500 trabajadores

3. Salario Promedio

- **Por Modalidad:** Salario base de cotización promedio ponderado por número de asegurados
- **Cálculo:** Media ponderada del salario de cotización mensual
- **Moneda:** Pesos mexicanos corrientes

Desagregación Geográfica

- **Nacional (código 99):** Agregado de todas las delegaciones
- **Delegacional (códigos 01-35):** 35 delegaciones estatales del IMSS
- **Subdelegacional (códigos 001-999):** ~300 subdelegaciones operativas

Modalidades de Aseguramiento

- **Modalidad 10:** Trabajadores permanentes urbanos
- **Modalidad 11:** Trabajadores eventuales urbanos
- **Modalidad 12:** Trabajadores permanentes del campo
- **Modalidad 13:** Trabajadores eventuales del campo
- **Modalidad 14:** Trabajadores eventuales de la industria azucarera
- **Modalidad 17:** Reversión de cuotas
- **Modalidad 32:** Trabajadores al servicio de gobiernos estatales
- **Modalidad 33:** Trabajadores al servicio de gobiernos municipales
- **Modalidad 34:** Trabajadores de universidades públicas
- **Modalidad 38:** Trabajadores domésticos
- **Modalidad 97:** Trabajadores no asalariados

Metodología de Procesamiento

Fuente de Datos Primaria

Archivo Base: pafinalmes* (Padrón de Afiliación Final Mensual) - **Frecuencia:** Mensual, disponible ~5 días hábiles después del cierre - **Cobertura:** Universo completo de trabajadores registrados ante el IMSS - **Registros:** ~20 millones de trabajadores activos por mes - **Variables clave:** NSS, delegación, subdelegación, modalidad, sector, salario, empresa

Proceso de Agregación

```
-- Pseudocódigo del proceso de agregación en SAS
SELECT
    ta,                                -- Tipo de alta (1=vigente)
    div_final,                         -- División económica
    size_cierre,                      -- Rango de tamaño de empresa
    cve_del_final,                    -- Clave delegación
    cve_subdel_final,                 -- Clave subdelegación
    mod,                              -- Modalidad
    tipotrc,                          -- Tipo trabajador
    COUNT(nss) as asegurados,         -- Conteo de trabajadores únicos
    MEAN(sal_cierre) as salario,      -- Salario promedio
    SUM(empresas3) as patrones        -- Conteo de patrones únicos
```

```
FROM pafinalmes_YYYYMM
WHERE aseg_cierre = 1          -- Solo asegurados vigentes
      AND aseg = 1            -- Con alta definitiva
GROUP BY ta, cve_del_final, cve_subdel_final, mod, tipotrc, div_final,
size_cierre;
```

Validaciones Implementadas

1. **Consistencia temporal:** Verificación de continuidad en series
2. **Totales de control:** Validación contra agregados oficiales
3. **Rangos esperados:** Detección de valores atípicos automática
4. **Integridad referencial:** Validación de códigos contra diccionarios

Flujo de Trabajo Principal

Para actualizar las series de datos, siga estos pasos:

1. **Ejecutar el código SAS:** `fetch_employment.sas`
2. **Ejecutar el proceso de targets:** `targets::tar_make()` en R

Este flujo de trabajo actualiza automáticamente todas las series de tiempo exportadas a través de la aplicación Shiny.

Estructura de Archivos

Entradas (Inputs)

Datos Principales

- **fetch_employment.SAS:** Script SAS que extrae datos del archivo más reciente `pafinalmes*` desde la base de datos corporativa del IMSS
- **Sábanas tradicionales (Excel):** Archivos históricos ubicados en `data/raw/SABANAS/` organizados por:
 - **DELEGACIONES/:** Datos agregados a nivel delegacional
 - **SUBDELEGACIONES/:** Datos desagregados a nivel subdelegacional
 - Categorías: ASEGURADOS Y PATRONES/, METAS, RECAUDACIÓN Y COBRANZA/, POBLACIONES/

Diccionarios y Catálogos

- **dictionaries/delegaciones.xlsx:** Mapeo de claves y nombres de delegaciones y subdelegaciones
- **data/dictionary.xlsx:** Diccionario de códigos de modalidad, sectores económicos y rangos
- **data/catalogue.xlsx:** Catálogo automático de archivos Excel disponibles
- **_keys/imss_pwd.txt:** Credenciales para acceso a datos (archivo seguro)

Salidas (Outputs)

Datos Procesados

- **data/sheets/sheets_data.parquet/:** Base de datos principal en formato Parquet particionada por:

- `book_name`: Tipo de serie (ej. “Asegurados trabajadores por modalidad”)
- `sheet_name`: Desagregación específica (ej. “Modalidad 10”, “Sector Industrial”)
- **`temp1.duckdb`**: Base de datos temporal para consultas rápidas
- **`data/sheets.feather` y `data/sheets.RData`**: Formatos alternativos de almacenamiento

Aplicación Web

- **`shiny/app.R`**: Aplicación Shiny interactiva para:
 - Visualización de series de tiempo con gráficos interactivos
 - Filtrado por nivel geográfico (Nacional, Delegacional, Subdelegacional)
 - Selección de entidades múltiples con búsqueda
 - Exportación a CSV y Excel en formato ancho o largo
 - Agregación automática de selecciones múltiples

Reportes

- **`README.html`**: Documentación renderizada del proyecto
- **`notebooks/review.qmd`**: Cuaderno Quarto para análisis y revisión de datos

Código Principal

Pipeline de Datos (`targets`)

- **`_targets.R`**: Orquestador principal que define el flujo de trabajo automatizado:
 - `raw_data_path`: Ubicación de datos históricos
 - `historical_sheets_data`: Lectura de sábanas Excel históricas
 - `fresh_data`: Extracción de datos frescos desde SAS
 - `current_month`: Cálculo automático del mes a procesar
 - `dictionary`: Carga de diccionarios de códigos
 - `fresh_data_cleaned`: Limpieza y transformación de datos frescos
 - `new_sheets`: Generación de nuevas series de tiempo
 - `updated_sheets`: Combinación con datos históricos
 - `saved_sheets_data`: Almacenamiento final en formato Parquet

Funciones Principales (`R/`)

- **`functions.R`**: Funciones centrales del procesamiento:
 - `pull_fresh_data()`: Conecta vía SFTP y extrae datos SAS más recientes
 - `clean_fresh_data()`: Normaliza códigos de modalidad y limpia variables
 - `make_new_sheets()`: Genera 5 tipos de series de tiempo:
 - Asegurados por modalidad y sector económico
 - Patrones por modalidad y rango de trabajadores
 - Salario promedio por modalidad
 - `update_sheets()`: Valida y combina datos nuevos con históricos
 - `save_updated_sheets()`: Guarda en formato Parquet particionado
- **`read_sheets.R`**: Procesamiento de sábanas Excel históricas:
 - `read_content()`: Extrae índices de contenido de archivos Excel
 - `read_sheet()`: Lee hojas individuales y pivotea fechas

- `read_sheets()`: Función principal que procesa todos los archivos Excel
- `read_delegaciones_dictionary()`: Carga diccionario geográfico

Módulos Auxiliares

- **R-other/reset_sheets.R**: Reinicialización completa desde sábanas históricas
- **SAS/fetch_employment.SAS**: Extracción automatizada de datos corporativos:
 - Identifica archivo más reciente `pafinalmes*`
 - Agrega datos por delegación, modalidad, sector y tamaño de empresa
 - Calcula conteos de asegurados, patrones y salarios promedio
 - Exporta a `employment_counts.sas7bdat`

Aplicación Web (shiny/)

- **app.R**: Aplicación modular con:
 - Módulos reutilizables por tipo de serie
 - Interfaz reactiva con filtros geográficos y conceptuales
 - Visualización con Plotly interactivo
 - Exportación flexible en múltiples formatos

Uso del Sistema

Actualización Regular (Mensual)

Paso 1: Extracción de Datos SAS

```
/* Ejecutar en SAS Enterprise Guide o SAS Studio */
%include "SAS/fetch_employment.SAS";
```

Este script: - Identifica automáticamente el archivo más reciente `pafinalmes*` - Procesa aproximadamente 20 millones de registros de trabajadores - Genera agregaciones por delegación, modalidad, sector económico y tamaño de empresa - Guarda resultados en `employment_counts.sas7bdat`

Paso 2: Procesamiento en R

```
# Cargar el pipeline de targets
library(targets)

# Ejecutar todo el pipeline de actualización
tar_make()

# Verificar el estado del pipeline
tar_visnetwork()

# Ver progreso en tiempo real
tar_progress()
```

El pipeline procesará automáticamente: 1. **Datos frescos**: Conexión SFTP para obtener employment_counts.sas7bdat 2. **Limpieza**: Normalización de códigos y fechas 3. **Generación**: Creación de 5 tipos de series de tiempo diferentes 4. **Validación**: Verificación de consistencia con datos históricos 5. **Almacenamiento**: Guardado en formato Parquet optimizado

Visualización y Exportación

Aplicación Shiny Local

```
# Ejecutar la aplicación web
shiny::runApp("shiny/app.R")
```

La aplicación permite: - **Filtrado interactivo**: Por nivel geográfico, entidades y conceptos - **Visualización**: Gráficos de líneas interactivos con Plotly - **Exportación**: Descarga en CSV o Excel, formato ancho o largo - **Agregación**: Suma automática de múltiples entidades seleccionadas


Acceso Programático

```
# Cargar datos directamente
library(arrow)
series <- open_dataset("data/sheets/sheets_data.parquet")

# Ejemplo: Asegurados por modalidad a nivel nacional
nacional <- series |>
  filter(
    series == "Asegurados",
    disaggregation == "Modalidad",
    clave_delegacion == 99,
    is.na(clave_subdelegacion)
  ) |>
  collect()

# Ejemplo: Exportar serie específica
series |>
  filter(
    series == "Patrones",
    level == "Delegación",
    date >= as.Date("2020-01-01")
  ) |>
  collect() |>
  write_csv("patrones_delegacional.csv")
```

Reinicialización Completa

 **Solo en casos excepcionales**: Cuando sea necesario reestablecer completamente las series desde las sábanas tradicionales Excel:

```
# PRECAUCIÓN: Esto sobrescribirá todos los datos automatizados
source("R-other/reset_sheets.R")
```

Cuándo usar reinicialización: - Cambios en la estructura de datos históricos - Correcciones masivas en sábanas Excel originales
- Migración o restauración del sistema - Cambios en diccionarios de códigos

Monitoreo y Mantenimiento

Verificación de Datos

```
# Verificar última actualización
tar_meta(fields = timestamp) |>
  arrange(desc(timestamp))

# Revisar errores en el pipeline
tar_meta(fields = c(name, error)) |>
  filter(!is.na(error))

# Estadísticas de la base de datos
series |> count() |> collect()
series |> summarise(
  min_date = min(date),
  max_date = max(date),
  n_series = n_distinct(paste(series, disaggregation))
) |> collect()
```

Respaldo de Datos

```
# Crear respaldo manual
file.copy(
  "data/sheets/sheets_data.parquet",
  paste0("backup_", Sys.Date(), ".parquet"),
  recursive = TRUE
)
```

Nota: El sistema está diseñado para actualizaciones incrementales mensuales. El flujo normal (tar_make()) solo procesa datos nuevos, manteniendo la integridad histórica y optimizando el rendimiento.

Arquitectura Técnica

Flujo de Datos

```
graph TD
  A[Base Corporativa IMSS] -->|SAS fetch_employment| B[employment_counts.sas7bdat]
```

```

C[Sábanas Excel Históricas] -->|R read_sheets| D[historical_sheets_data]
B -->|SFTP pull_fresh_data| E[fresh_data]
E -->|clean_fresh_data| F[fresh_data_cleaned]
F -->|make_new_sheets| G[new_sheets]
D -->|update_sheets| H[updated_sheets]
G --> H
H -->|save_updated_sheets| I[sheets_data.parquet]
I --> J[Aplicación Shiny]
I --> K[Exportaciones CSV/Excel]

```

Tecnologías Utilizadas

Almacenamiento y Procesamiento

- **Apache Parquet:** Formato columnar optimizado para análisis
- **Apache Arrow:** Interfaz de acceso a datos de alto rendimiento
- **DuckDB:** Motor de consultas SQL embebido para análisis
- **R Targets:** Orquestación de pipelines de datos reproducibles
- **renv:** Gestión de entornos y dependencias de R

Conectividad y Extracción

- **SAS/CONNECT:** Acceso a bases de datos corporativas SAS
- **SFTP:** Transferencia segura de archivos entre sistemas
- **haven:** Lectura de archivos SAS desde R
- **readxl:** Procesamiento de archivos Excel complejos

Visualización y Aplicaciones

- **Shiny:** Framework web reactivo para R
- **Plotly:** Gráficos interactivos y exportables
- **DT:** Tablas interactivas con búsqueda y filtrado
- **Quarto:** Documentación científica reproducible

Especificaciones de Datos

Dimensiones Principales

- **Temporal:** Series mensuales desde enero 2019 hasta presente
- **Geográfica:** 3 niveles (Nacional, 35 Delegaciones, ~300 Subdelegaciones)
- **Conceptual:** 5 tipos de series principales:
 1. **Asegurados:** Por modalidad (11 categorías) y sector económico (21 divisiones)
 2. **Patrones:** Por modalidad (6 categorías) y rango de trabajadores (6 rangos)
 3. **Salario:** Promedio ponderado por modalidad

Estructura de la Base de Datos

```

-- Esquema principal sheets_data.parquet
CREATE TABLE sheets_data (

```



```

    book_id INTEGER,           -- ID del libro/categoría
    sheet_id INTEGER,          -- ID de la hoja/serie específica
    book_name STRING,          -- Nombre del libro (ej. "Asegurados
trabajadores por modalidad")
    sheet_name STRING,         -- Nombre de la serie (ej. "Modalidad 10")
    level STRING,              -- Nivel geográfico: "Nacional", "Delegación",
"Subdelegación"
    date DATE,                 -- Fecha mensual (primer día del mes)
    clave_delegacion INTEGER,  -- Código de delegación (01-35, 99=Nacional)
    clave_subdelegacion INTEGER, -- Código de subdelegación (01-999, 99=Total
delegacional)
    delegacion STRING,         -- Nombre de la delegación
    subdelegacion STRING,      -- Nombre de la subdelegación
    series STRING,             -- Tipo de serie: "Asegurados", "Patrones",
"Salario"
    disaggregation STRING,     -- Tipo de desagregación: "Modalidad", "Sector
económico", etc.
    value DOUBLE               -- Valor numérico de la serie
);

```

Particionamiento

Los datos se almacenan particionados por book_name y sheet_name para optimizar consultas:

```

data/sheets/sheets_data.parquet/
├── book_name=Asegurados trabajadores por modalidad/
│   ├── sheet_name=Modalidad 10/
│   ├── sheet_name=Modalidad 11/
│   └── ...
├── book_name=Patrones por modalidad/
└── ...

```

Rendimiento y Escalabilidad

Optimizaciones Implementadas

- **Lectura incremental:** Solo procesa meses nuevos, no datos históricos
- **Almacenamiento columnar:** Parquet reduce tamaño ~80% vs CSV
- **Particionamiento inteligente:** Consultas filtran solo particiones relevantes
- **Conexión lazy:** Arrow abre datasets sin cargar en memoria
- **Compresión SNAPPY:** Balance óptimo entre compresión y velocidad

Métricas Típicas

- **Datos históricos:** ~50GB en Excel → ~2GB en Parquet
- **Actualización mensual:** 5-10 minutos completos
- **Consultas interactivas:** <1 segundo para agregaciones típicas
- **Capacidad:** Escala hasta millones de series sin degradación

Próximos Pasos

Desarrollo Prioritario

1. **Generación automática de resúmenes estadísticos**
 - Cálculos de variaciones mensuales y anuales
 - Detección automática de anomalías en series
 - Reportes ejecutivos automatizados
2. **Deployment en producción**
 - Configuración de servidor Shiny Pro/Connect
 - Automatización de actualizaciones mensuales vía cron
 - Monitoreo y alertas de fallos en pipeline
3. **Validación y calidad de datos**
 - Pruebas unitarias para funciones críticas
 - Validación cruzada con sábanas oficiales
 - Alertas automáticas por inconsistencias

Mejoras Futuras

4. **Interfaz de usuario mejorada**
 - Dashboard ejecutivo con KPIs principales
 - Comparaciones automáticas período anterior
 - Exportación a formatos estadísticos (SPSS, Stata)
5. **Integración con otros sistemas**
 - API REST para consultas externas
 - Conexión directa con Business Intelligence
 - Sincronización con repositorios oficiales
6. **Análisis avanzado**
 - Modelos de forecasting automático
 - Detección de patrones estacionales
 - Análisis de correlaciones inter-series

Contacto y Soporte

Equipo de Desarrollo: Dirección de Incorporación y Recaudación (DIR), IMSS

Responsable Técnico: Esteban de Getau

Email: [esteban.degetau@imss.gob.mx]

Soporte: - Reportar errores: Issues en repositorio GitHub - Solicitudes de mejora: Contacto directo con equipo DIR - Documentación técnica: Consultar cuadernos en notebooks/

Horarios de mantenimiento: Primer martes de cada mes, 2:00-4:00 AM para actualizaciones automáticas.