

Definiciones Clave

- Data anonymization : proceso de protección de datos privados / sensibles mediante la eliminación de identificadores
- PII (personally identifiable information) :
 - información que puede restar la identidad de una persona
 - función por si sola o combinada con otros datos (atopos de entorno)

2) Técnicas de anonimización

métodos estandar para transformar columnas en datos

- Blanting : Borrado total del campo
- Masking : otorgar valores ligeramente más altos precision identificable

3) Rol del analista

- responsabilidad general : establecer qué datos requieren anonimizaciones

Ejecución

Normalmente : No realizan anonimización / recibe datos empíricos

• Ejemplo : entornos de desarrollo / testing

4) industria de alto riesgo

Requiere desinformación (de identificación) rigurosa :

- Salud : Historiales médicos, diagnósticos
- Finanzas : Números de cuenta, transacciones, apuntes de crédito

5

Checklist : datos a anónimizar

- Nombres
- # teléfono
- Correo
- IP
- ID
- Placas

- numero de cuenta
- registros medios
- Fotos

Definición de datos abiertos

Subconjunto de la élite de los datos. Se refiere al libre acceso, uso e intercambio.

- Los 3 requerimientos técnicos
 - Disponibilidad
 - Reutilización y redistribución
 - Permeabilidad universal

Defecto: apertura vs profundidad

Beneficios de open data

Apolazamientos de base de datos
Credible

colaboración científica y práctica
en investigación

mejora en la toma de decisiones
Bases de datos externas

El nexo es la representación de
individuos y pueblos

③ Tipos de datos críticos en el debate

A) Datos de 3eros

Datos recopilados por encuestadores sin relación directa con el sujeto

Creación de perfiles audiencia y publicidad dirigida

disponibilidad masiva y consentimiento del usuario

Información identificativa personal

los datos pueden identificar a una persona o revelar su identidad

dirección / email

id / Tarjetas

Base de datos relacionales

- Conexión de tablas interconectadas mediante relaciones lógicas
- Relacional VS no relacional
 - Relación estructurada VS agrupación masiva
 - Facilita búsqueda compleja VS análisis crudo
 - Reduce errores de duplicidad

2) Normalización

Proceso de organizar los datos para que:

- Eliminen la redundancia
- Aumentar integridad de datos
- Reducir complejidad

tablos pequeños y espesos
en lugar de una "sabana" de
datos gigante

3) Sistemas de llaves

El mecanismo que crea las relaciones
entre tablas

[• Llaves primarias]

- identificando inicio de cada registro
- valores únicos (unicos), nunca se repiten y nunca nulos
- orden_id en una tabla

• Have foreign

- Compo que apunta a la PK
de otra tabla
- Crea el vínculo / las relaciones

- Una tabla puede tener multiples FKs
- clave compuesta
- PK construida usando multiples columnas
- Necesaria cuando una sola columna no garantiza unicidad

(4)

SQL en este contexto

- SQL : Lenguaje interfoz para Crear y Comunicarse con RPDMS
- Funciones analíticas : permite reconstruir la visión completa de los datos usando (joins) definidos por los usuarios

① Objetivo de la inspección

- Validar viabilidad técnica de las preguntas de negocio
- Identificar necesidad de fuentes externas
- Entender la remontada externa

② Círculo de estudio dentro de hechos

Análisis entre los bruchos y los datos disponibles:

A) Sobre qué mercados?

Dato disponible: Units - Sold

- Dato faltante: ingreso por robos

Decisiones: Definir "popularidad por volumen (unidades) no por rentabilidad o buscar límite de precio externo

B) Impacto de la temperatura

- dato disponible : Temperatura, Salvo
- Antiguedad : En Salvo el total del dia o una fracción

366 filas / año bisiesto \rightarrow Sugiere

granularidad
diaria

Confirmar dato con el autor si
el orden es cronológico

C) Efecto o finales de semana

- Dato disponible \rightarrow Dato <
- Términos : features engineering
- extra dia de remora por el fin de semana
- Crear búsquedas is - weekend

D) Rebalizado nuevo VS recurrentes

- Datos disponibles : Solo ventas generales
- No hay customers_id ni históricos de clientes
- Se requiere Data Blending con un CRM o tablas de clientes

③ Tipos de problemas

- The Data is not there : alcance imposible. Con el dataset actual
- the Data is insufficient : faltan datos de tiempo o granularidad
- the Data is incorrect : errores de rebalizado

Propósito y difusión [metadatos]

Es el contexto necesario para interpretar, gestionar y confiar en los datos.

[Metadata is important as the data itself]

Funciones $W + H$: Responde a quién, qué, cuándo, dónde y cómo sobre el dato

set

② Elementos extendidos sobre los metadatos

identidad : tipo de archivo Nombre

Temporalidad : created, modified, last accessed

Autoría : created-by

Ubicación :

Acceso : permisos

③ Géneros terminos por fuentes

A) Foto

- Punto oculto en imágenes digitales
- modelo de Camara, iso, apertura
- GPS

B) Email

Vital para ciberseguridad y rastreo

IP del servidor remitente, ruta
de servidores

C) Bases de datos

- Los metadatos viven en el information_Scheme
- Definen tipos de datos, relaciones
llaves primarias

D) Web

para SEO y navegadores

4) Herramientas prácticas

Bash

ls -l

File nombre - archivo

SQL : Muestra los metadatos de las columnas

Beneficios de los Metadatos

optimizan los toma de decisiones
el garantizar

- Fidabilidad : Valida la salud del dato antes de usarlo
- Consistencia : - uniformidad en formato
 - permite comprender "datos con peros" entre diferentes fuentes

2. Repositorios de metadatos

- Base de datos especializada que almacena y gestiona metadatos
- Centraliza la estructura, ubicación y orígenes de los datos
- Evita la búsqueda manual archive nos archivos

3) Clasificación de datos extremos

Guloso para adquisiciones y alcance de datos

- 1- Party data : datos propios
- 2- party data : datos de adquirir mas, comprados directamente de la fuente
- 3- Datos : datos de agregadores extremos sin relación directa con el origen

4)

Checklist de validación

- Accesibilidad
- Frecuencia
- Propiedad

Definición Técnica:

CSV

• Nomenclatura : Archivo de texto plano

• Esquema : Matriz tabular

Filas = líneas de texto

Columnas = Separadas por delimitadores

Ventaja : interoperabilidad total
(Excel, SQL, Python, R)

Desventaja : eficiencia en lectura / escritura
(l./o) vs Formatos Unarios

② Análisis de módulo

• Parámetros / criterios de lectura

• Delimitar / separador ; ; ; , \t

Header : La primera fila
contiene nombre de
columnas ?

Encoding : UTF-8 vs ISO-8859-1

Quoting : Manejo de delimitadores
dentro del texto

(3)

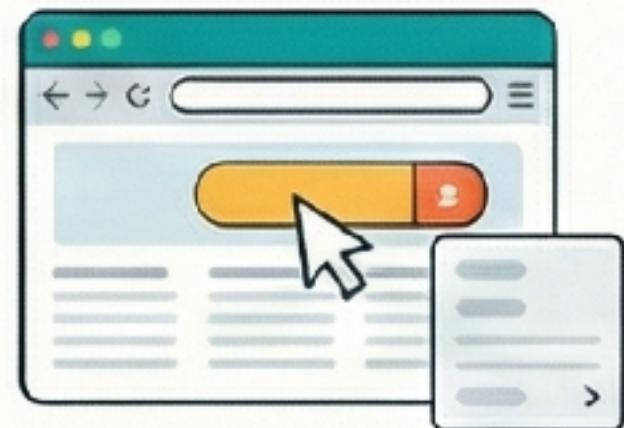
Métodos de obtención

Manual (ad-hoc)

Descarga directa

Mastering CSVs: The Bridge to Data Analysis

Phase 1: Acquiring CSV Data



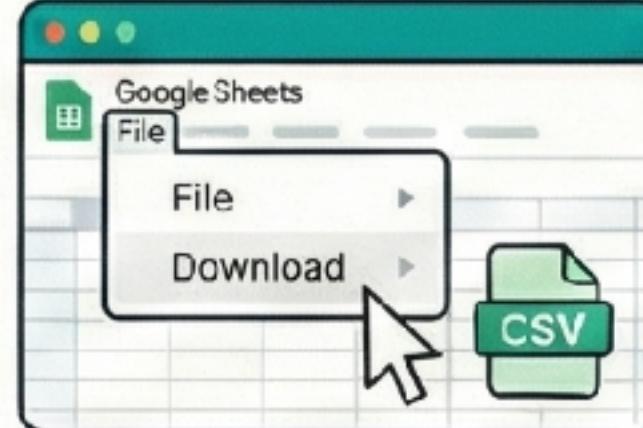
Standard Download Methods

Click the direct link or right-click the data element to "Save As" a .csv file.



Force a Direct Download

Hold the Alt key while clicking a link to bypass browser previews and save directly.



The Chrome Browser Workaround

Save as a Google Sheet first, then select Download > Comma Separated Values (.csv).

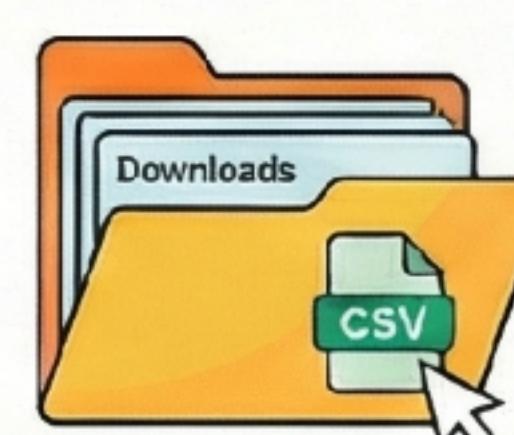


Phase 2: Importing for Analysis



Locate the Import Interface

Find the designated "Upload," "Import," or drag-and-drop zone within your analysis platform.



Retrieve from Local Storage

Select your file, typically found in your device's Downloads folder after acquisition.



Verify Platform Requirements

Check for specific file size or formatting restrictions before initiating the final upload.

Mastering CSV Data Management

CSV Definition

Plain text files

Table structure

Comma-separated values

Rows and columns

Wide compatibility

Easy to edit and manipulate

Ideal for data transfer

Facilitates visualization

Benefits for Data Analysts

Downloading CSV Files

Direct click on links

Right-click and Save As

Force download with Alt key

Google Chrome handling

Uploading CSV Files

Locate Upload or Import option

Select file from local device

Monitor progress bars

Check size and format restrictions

Applications

Data cleaning

Statistical extraction

Analysis programs

Data visualization

Mastering Dynamic Tables

Mastering Dynamic Data: Automating Google Sheets

Explaining the benefits of dynamic data import over manual methods and providing a quick guide to the primary Google Sheets functions used for automation.

STATIC VS. DYNAMIC DATA

MANUAL UPDATES



Manual vs. Automatic Updates

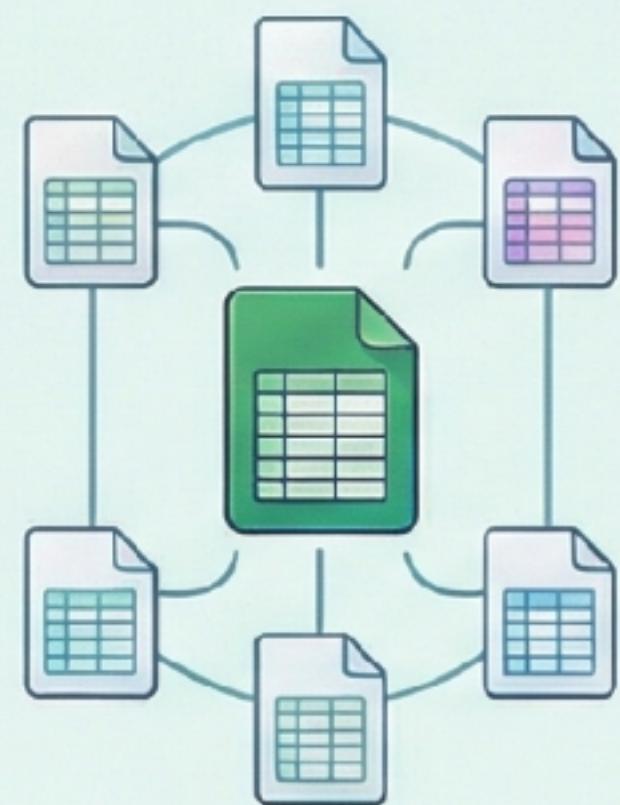
Static imports require constant manual re-uploading, while dynamic imports update automatically as source data changes.



Reduced Human Error

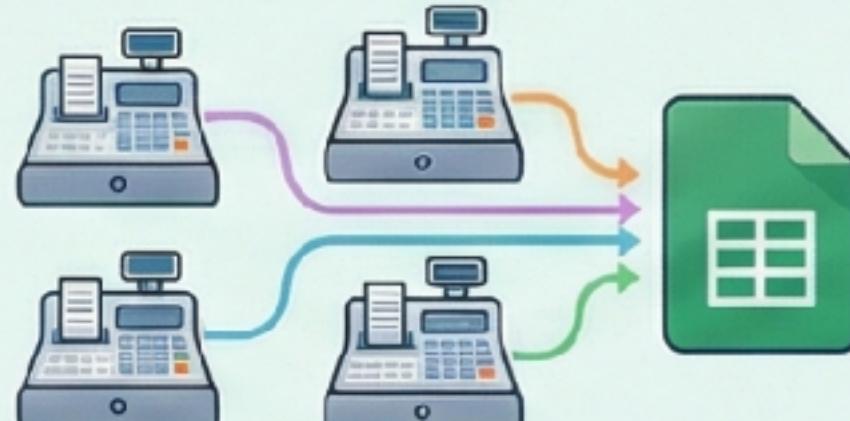
Automating data flow between spreadsheets prevents the discrepancies that occur when maintaining multiple manual records.

AUTOMATIC UPDATES



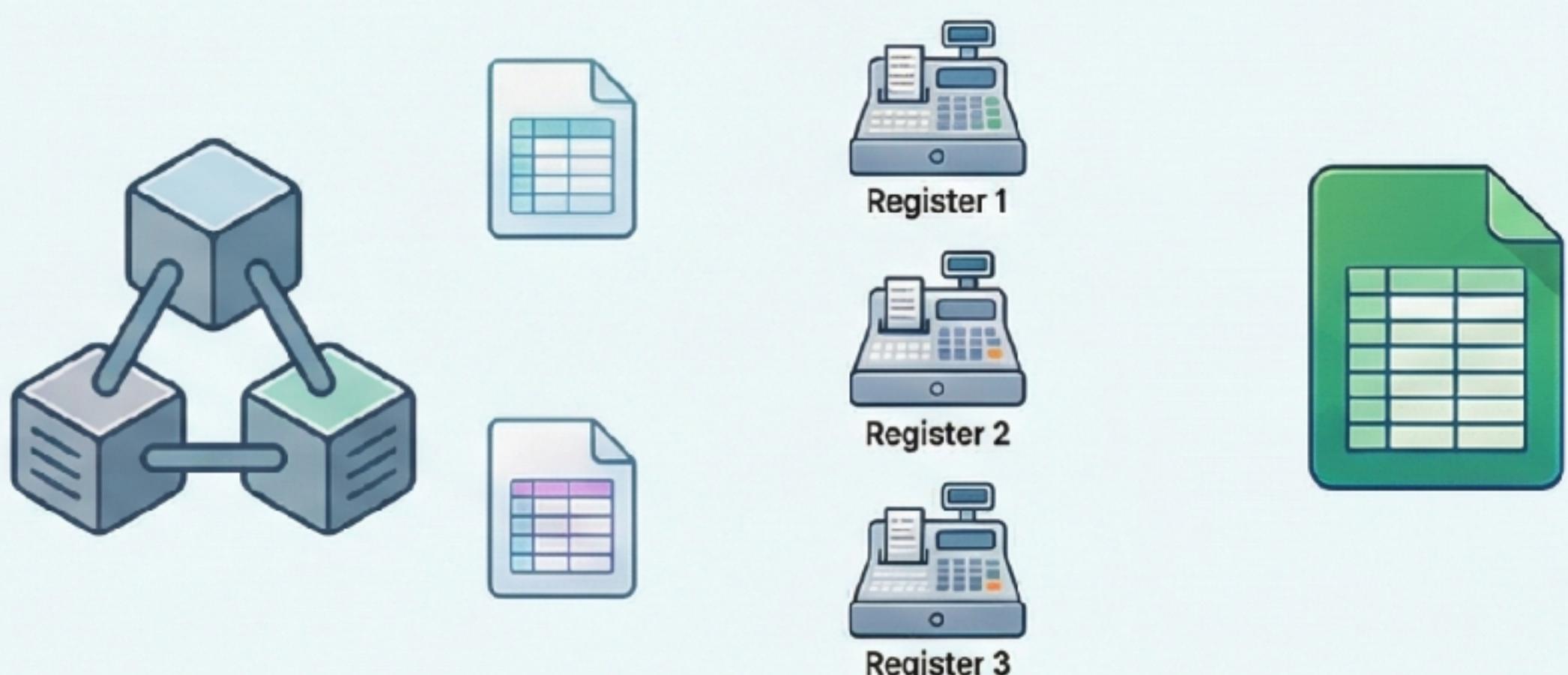
The Multi-Register Retail Model

A shop owner can automatically consolidate individual sales from three registers into one master spreadsheet.



The Multi-Register Retail Model

A shop owner can automatically consolidate individual sales from three registers into one master spreadsheet.



ESSENTIAL IMPORT FUNCTIONS

IMPORTRANGE for Sheet-to-Sheet Sync



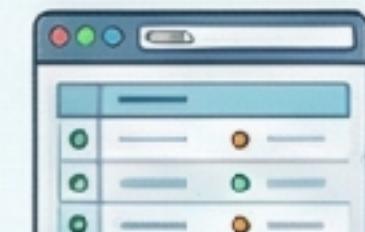
=IMPORTRANGE("URL", "sheet_namchange")

Source Sheet
(Spreadsheet URL)



Use =IMPORTRANGE("URL", "sheet_nametrangle") to pull data from an entirely different Google Sheets file.

Function Requirement	Description
Spreadsheet URL	The full web address of the source Google Sheet.
Range String	The specific sheet name and cell range (e.g., "Sheet1!A1:D10").
Access Permission	A one-time "Allow Access" click required when first connecting sheets.



IMPORTHTML for Web Scraping

This function extracts structured tables or lists directly from public websites into your spreadsheet.



IMPORTDATA for External Files

Automatically fetches data from online .csv or .tsv files using a simple URL reference.

