M4 R1 C1C Fuentes y Tipos de Datos

Clase 2

stucom

Objetivos de la Clase

- Entender la importancia de combinar diferentes fuentes y tipos de datos en proyectos de análisis de datos.
- **Conceptos Clave:** integración de datos, limpieza de datos, y uniones de datos.

stucom

“Los datos son el nuevo petróleo”

- Martin Hilbert

Imagen: The Economist Magazine - 2017

stucom

DATOS VS. INFORMACIÓN

Característica o atributo **sin procesamiento**, el cual no informa nada por sí solo.

Unión de **datos procesados**, que se complementan para informar un hecho.

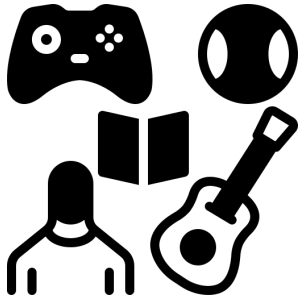
stucom

TIPOS DE DATOS

stucom

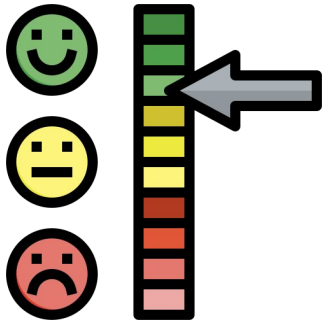
Cualitativos o Categóricos

Nominales



Género, nacionalidad, color Habilidades, preferencias, comida o pasiones.

Ordinales



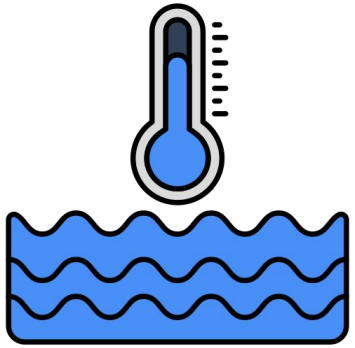
Ratings (bueno, regular, malo) o Niveles de educación (primaria, secundaria, universidad).

Cuantitativos

Discretos

Pueden tomar ciertos valores y no pueden ser más precisos: estudiantes en clase BI & IA en Stucom

Continuos



Pueden tomar cualquier valor dentro de un rango determinado: temperatura, peso, altura.

https://www.guvi.in/blog/types-of-data-in-data-science/

stucom

TIPOS DE DATOS	Cualitativos o Categóricos	Cuantitativos
Nominales	Etiquetas sin orden	
Ordinales	Etiquetas con orden	
Discretos		Valores fijos
Continuos		Rangos de valores

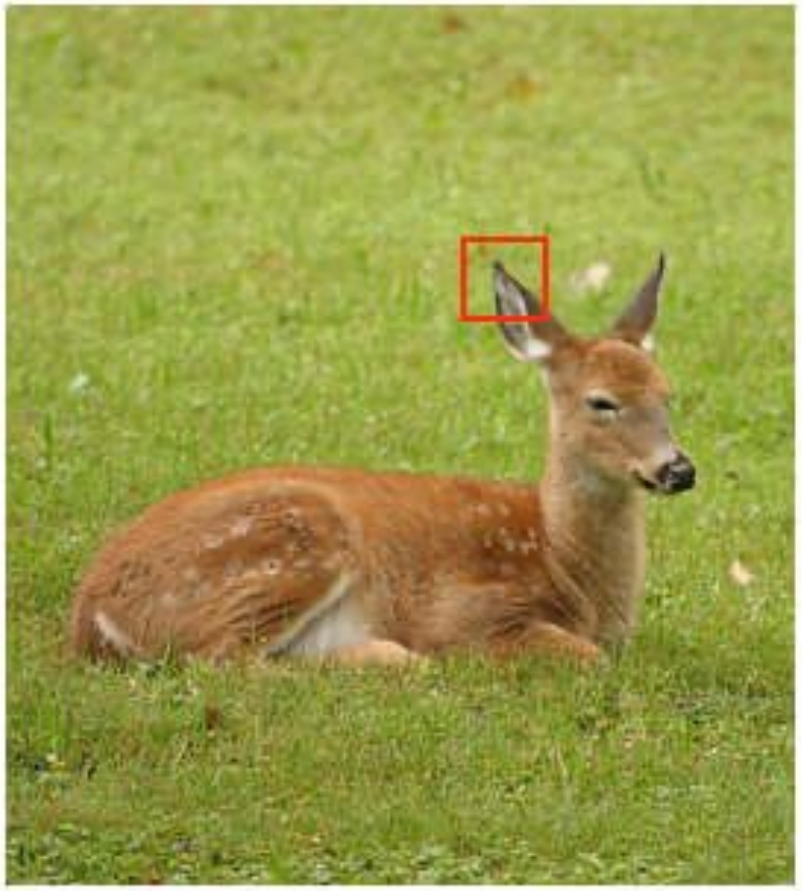
Cualitativos o Categóricos	Cuantitativos
Mi mejor amigo pesa 75 kilos	Mi mejor amigo _____
Mi madre ha conducido _____	Mi madre conduce en un coche rojo
El equipo de natación entrenó ____	El abuelo es divertido y sabe escuchar
El terminó la maratón en 3 hr y 45 mins	El tiene unos ojos _____ y un cabello ___

stucom

OTROS TIPOS DE DATOS

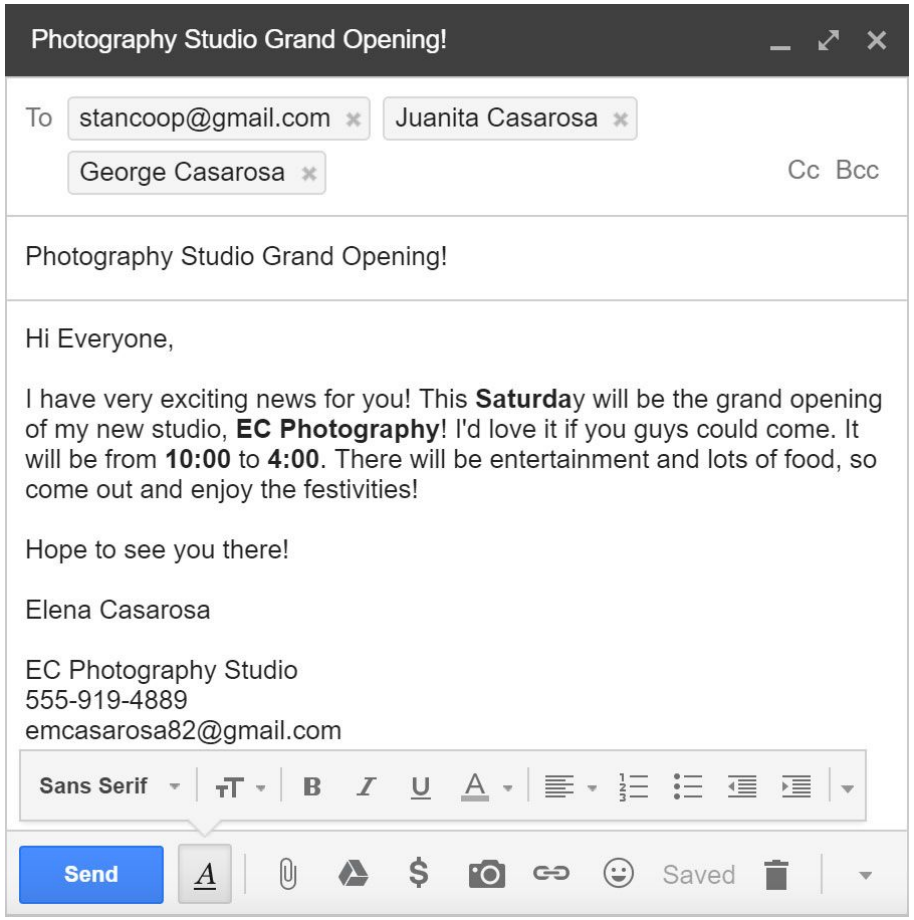
stucom

DATOS DE IMAGEN



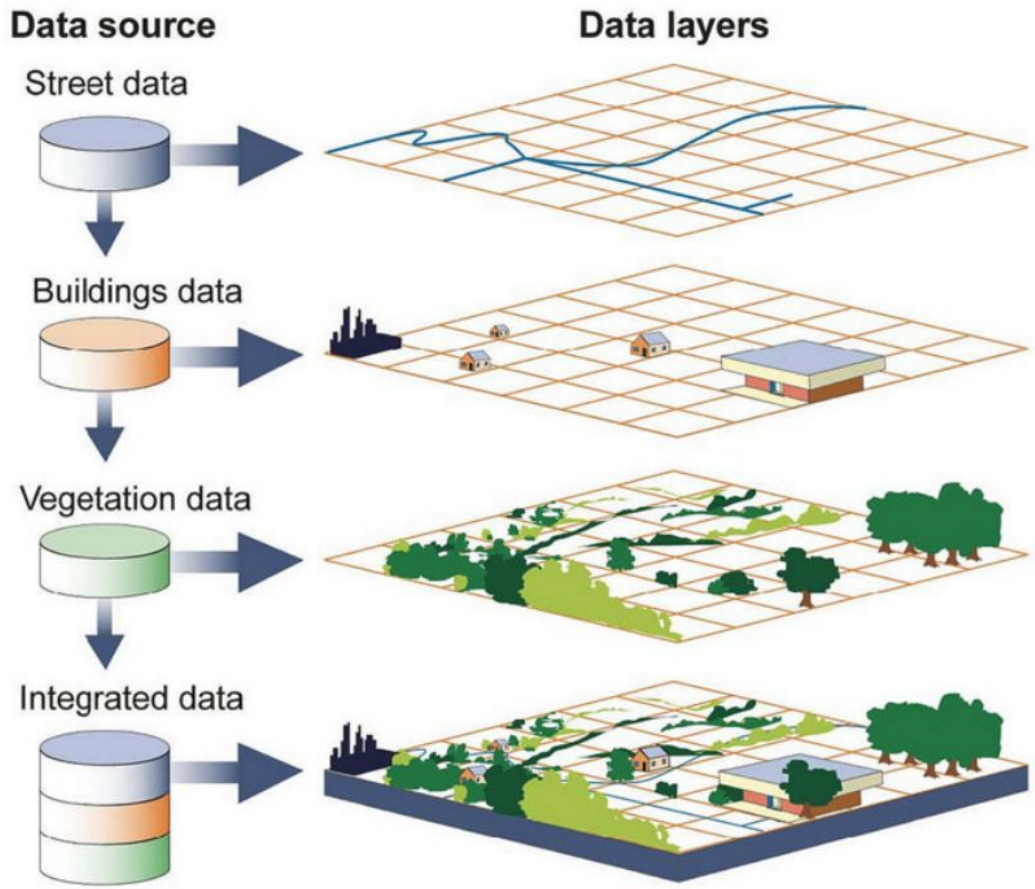
stucom

DATOS DE TEXTO



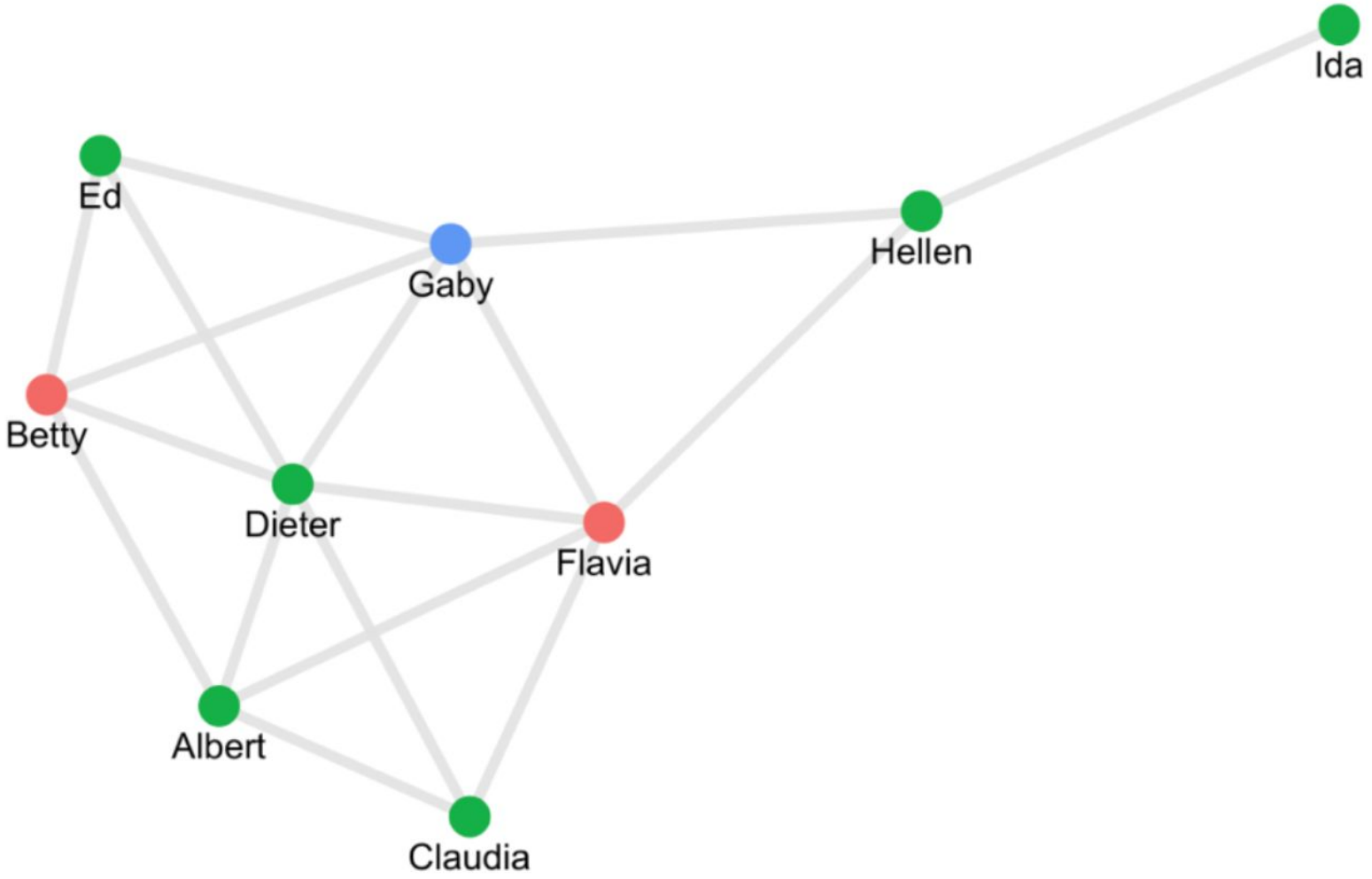
stucom

DATOS GEOESPACIALES



stucom

NETWORK DATA

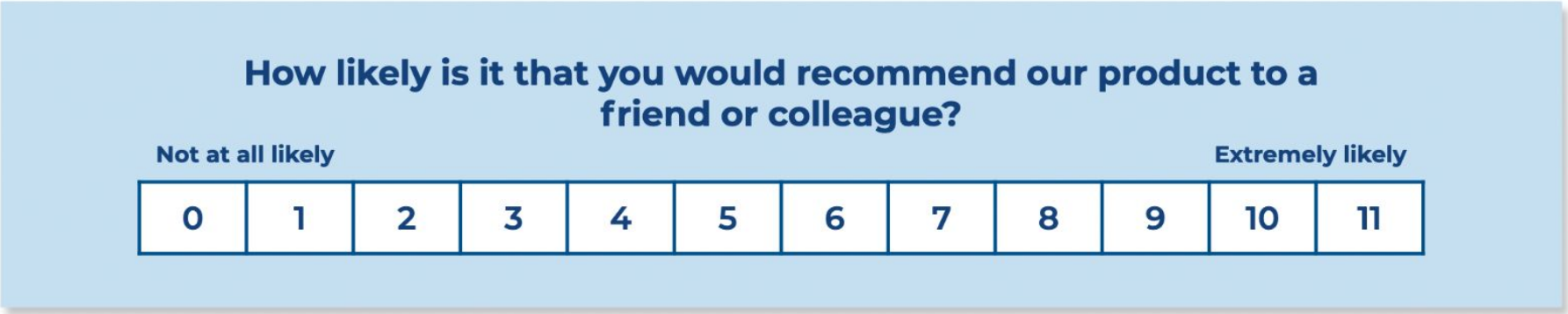


stucom

Net Promoter Score (o NPS) es una métrica común que las empresas utilizan para realizar un seguimiento del éxito de un producto o sitio web. Se mide haciendo una simple pregunta:

¿Qué tan probable es que recomiendes [insertar marca/sitio web/servicio/producto] a un amigo o colega?

Los usuarios responden en una escala de 0 a 10, siendo 0 que no es en absoluto probable que se recomiende y 10 es extremadamente probable que se recomienda.



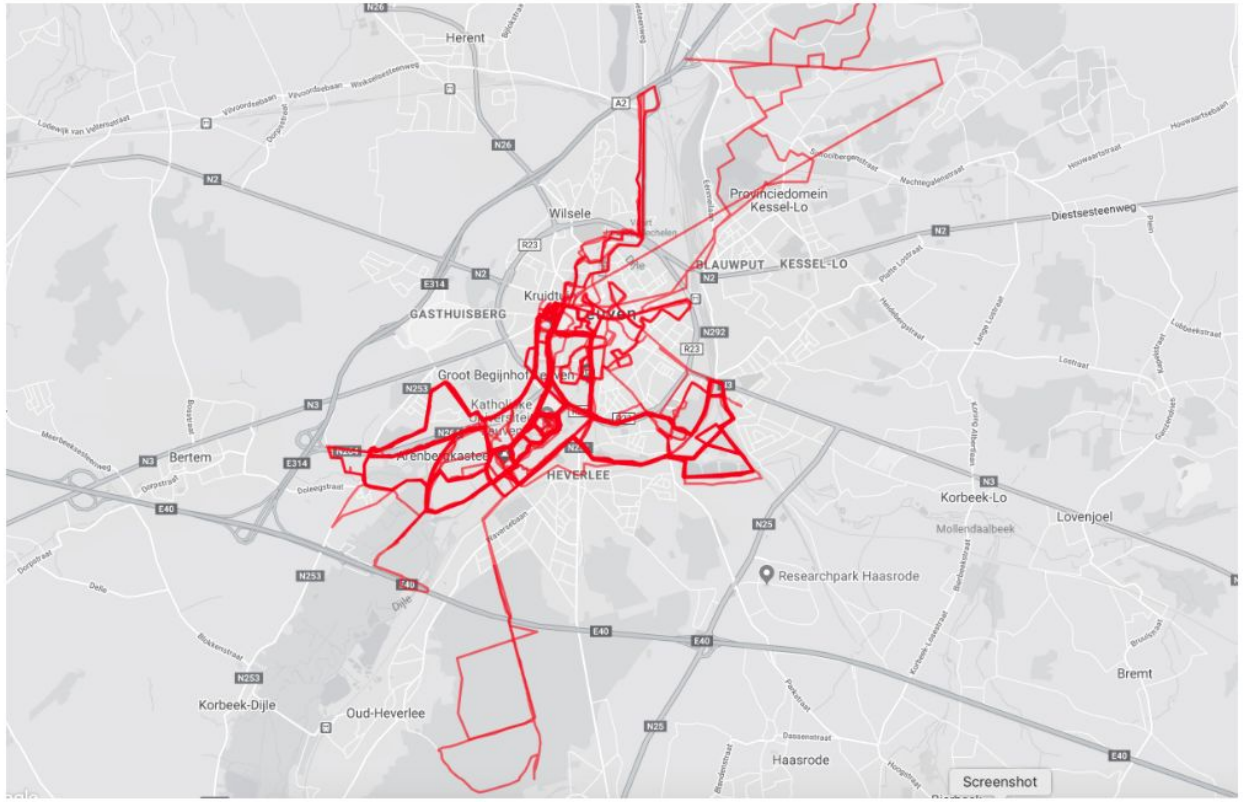
Cualitativo

Cuantitativo

stucom

La resolución de Año Nuevo de Jane este año era ponerse en la mejor forma de su vida. Para ayudarla a alcanzar este objetivo, decidió invertir en un rastreador de actividad. Después de algunos meses de rastrear su actividad, hay bastantes datos disponibles.

La empresa que fabricó el rastreador de actividad tiene una API pública que permite el acceso a sus datos personales. Jane está específicamente interesada en los datos GPS de sus carreras porque quiere hacer un mapa de calor que muestre sus rutas de carrera más comunes.



Imagen

Texto

Geoespacial

Network

#StucomÉsDemà

stucom

¡IMPORTANTE!

stucom

UBICACIÓN (ALMACENAMIENTO)

Solución de Almacenamiento Paralelo

- ➔ Alto rendimiento local
- ➔ Acceso simultáneo y rápido

CLOUD (La Nube)

- ➔ Accesibilidad
- ➔ Flexibilidad

stucom

TIPOS DE ALMACENAMIENTO DE DATOS

- NO ESTRUCTURADA
- ESTRUCTURADA

stucom

NO ESTRUCTURADA

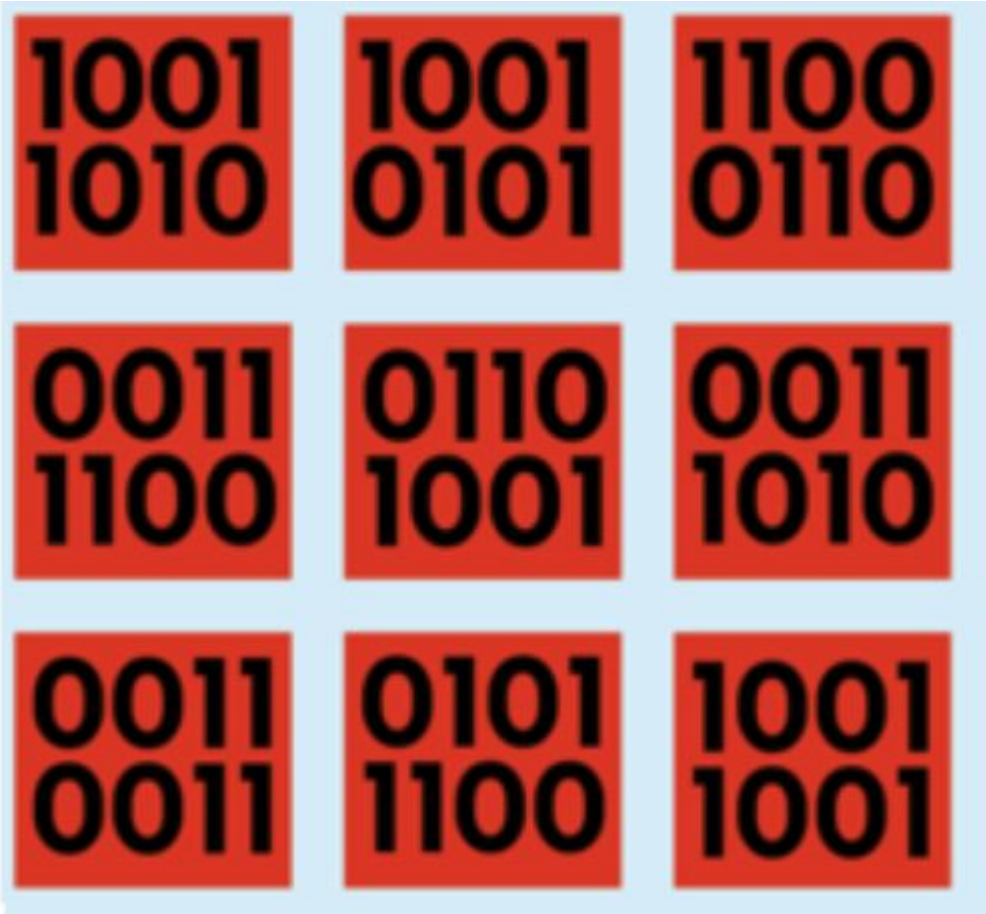
Tipos de Datos
Email’s
Texto
Páginas Web
Redes Sociales

Almacenamiento
Base de Datos Documental

Recuperación - Lenguaje Query
NoSQL

stucom

ESTRUCTURADA

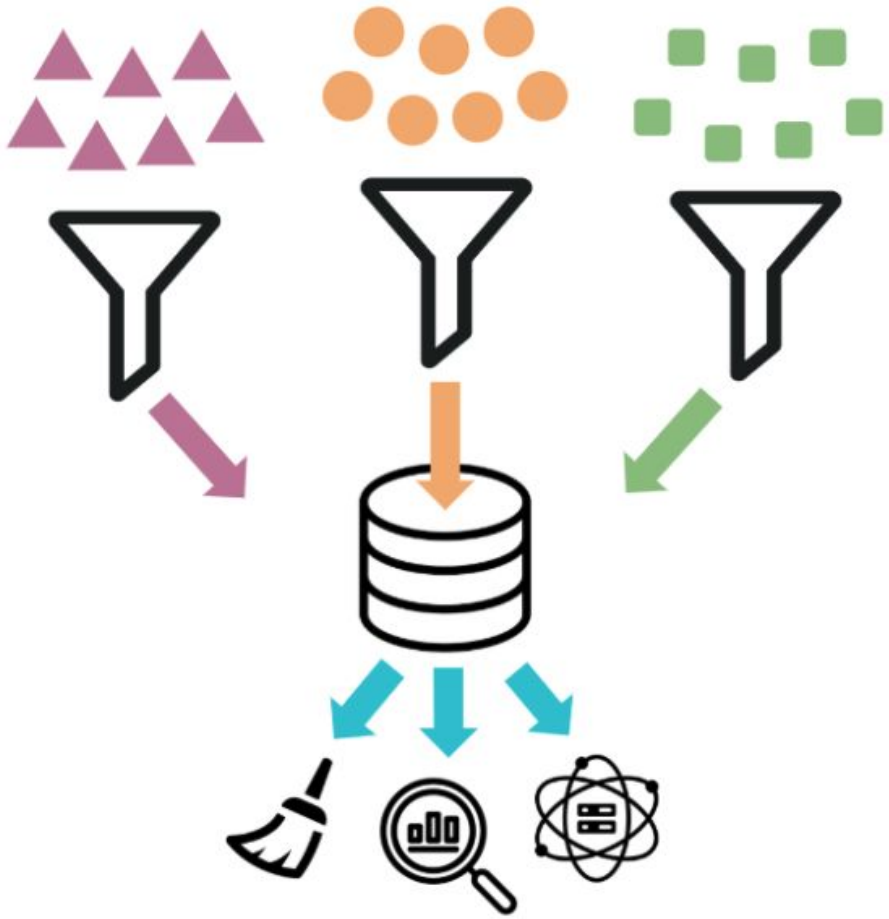


Tipos de Datos
Tabulares
Bases de Datos
Google Sheets
Archivos Excel

Almacenamiento
Base de datos relacionales

Recuperación - Lenguaje Query
SQL (Lenguaje de Consulta Estructurada)

stucom



¿Cómo escalamos?

Más de una fuente de datos:

- Bases de Datos
- APIs
- Registros Públicos

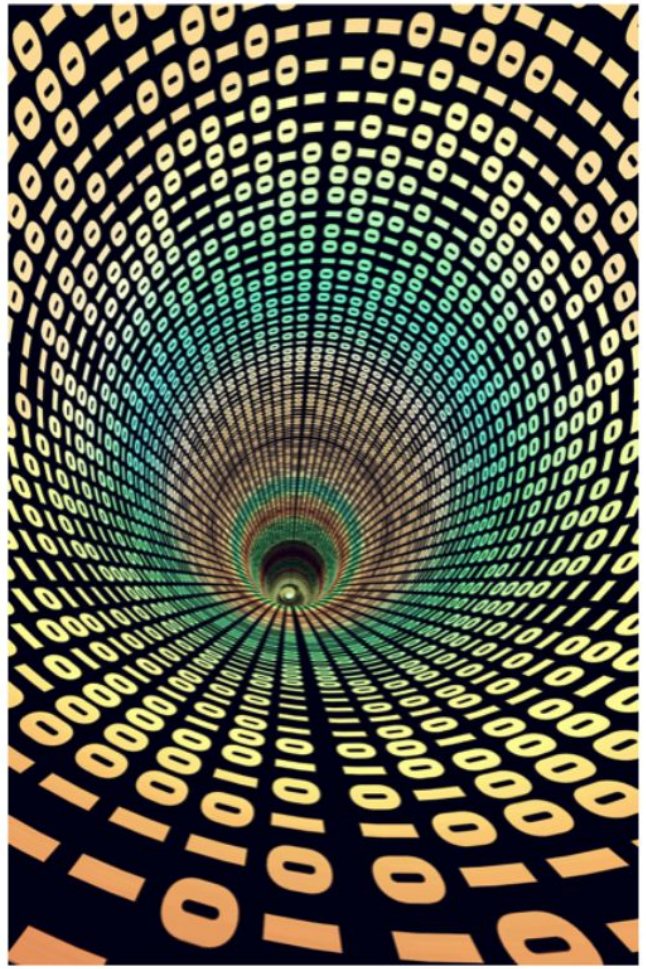
Diferentes tipos de datos:

- Datos estructurados
- Datos Tabulares
- Datos transmisión en tiempo real. Ejemplo: Tweets

stucom

Qué es un PIPELINE DE DATOS?

- Movilidad de datos en etapas definidas
- Recolección y almacenamiento de datos automatizado
  - Programados por hora, diariamente, semanalmente, mensualmente, etc
  - Desencadenado por eventos
- Registros Públicos
- Necesarios para grandes proyectos de datos
- Extract Transform Load (ETL)



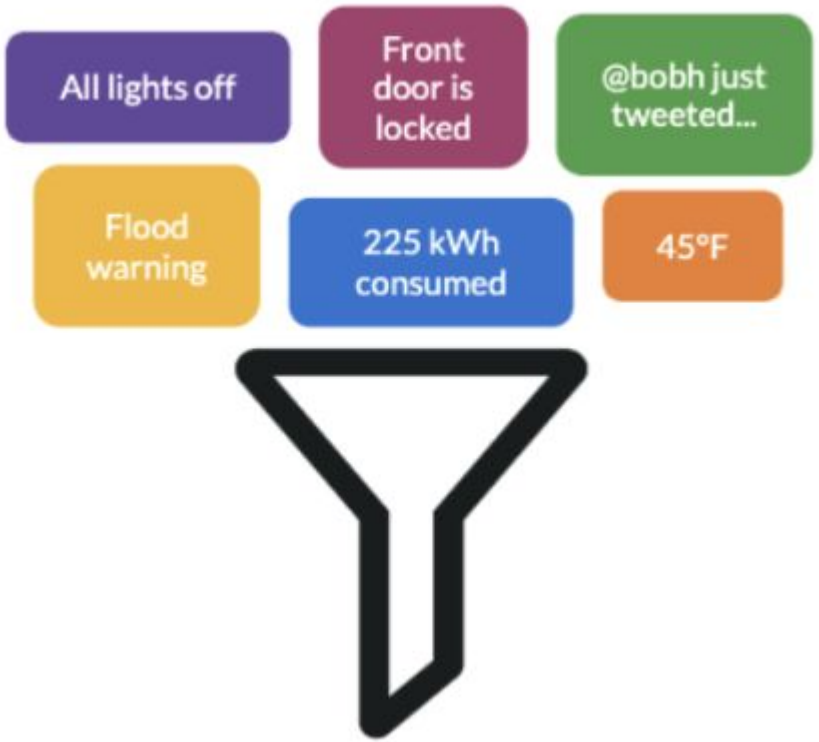
stucom

CASO DE ESTUDIO

Dato	Fuente	Frecuencia
Condiciones Ambientales	National Weather Service API	Cada 30 mins
Tweets en tu área	Twitter API	Tiempo-real
Temperatura (Interna)	Smart home thermostat	Cada 5 mins
Estado de luces	Smart light bulbs	Cada 1 min
Estado de bloqueos	Smart door locks	Cada 15 segundos
Consumo de energía	Smart meter	Semanalmente

stucom

EXTRACCIÓN



Dato	Fuente	Frecuencia
Condiciones Ambientales	National Weather Service API	Cada 30 mins
Tweets en tu área	Twitter API	Tiempo-real
Temperatura (Interna)	Smart home thermostat	Cada 5 mins
Estado de luces	Smart light bulbs	Cada 1 min
Estado de bloqueos	Smart door locks	Cada 15 segundos
Consumo de energía	Smart meter	Semanalmente

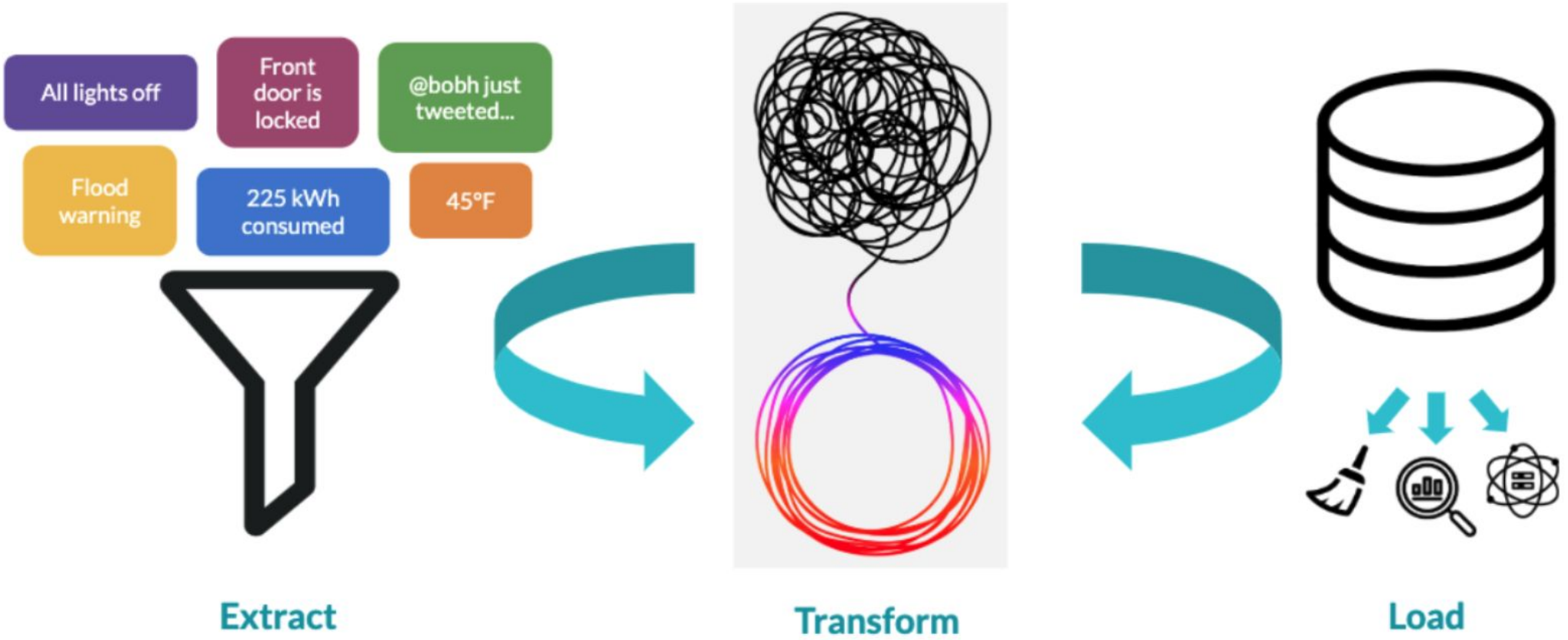
stucom

TRANSFORMACIÓN

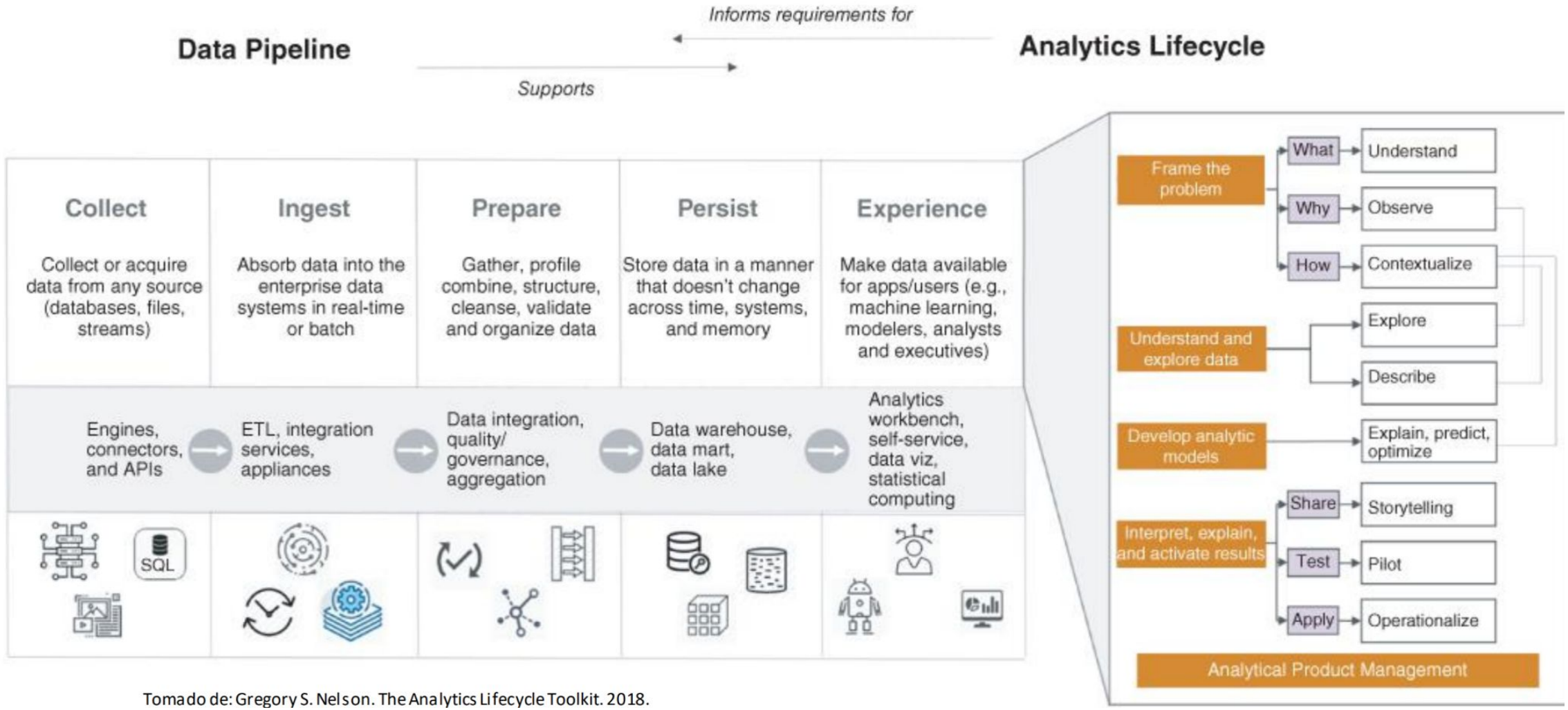
- Unir diferentes fuentes de datos en un solo conjunto de datos
- Convertir la estructura de datos para que se ajuste al esquema de las bases de datos
  - Ejemplo id_cliente es char y queremos que sea numérico
- Removiendo datos irrelevantes
- La preparación de datos y la exploración de los mismos no ocurren hasta esta fase.

stucom

CARGA Y AUTOMATIZACIÓN

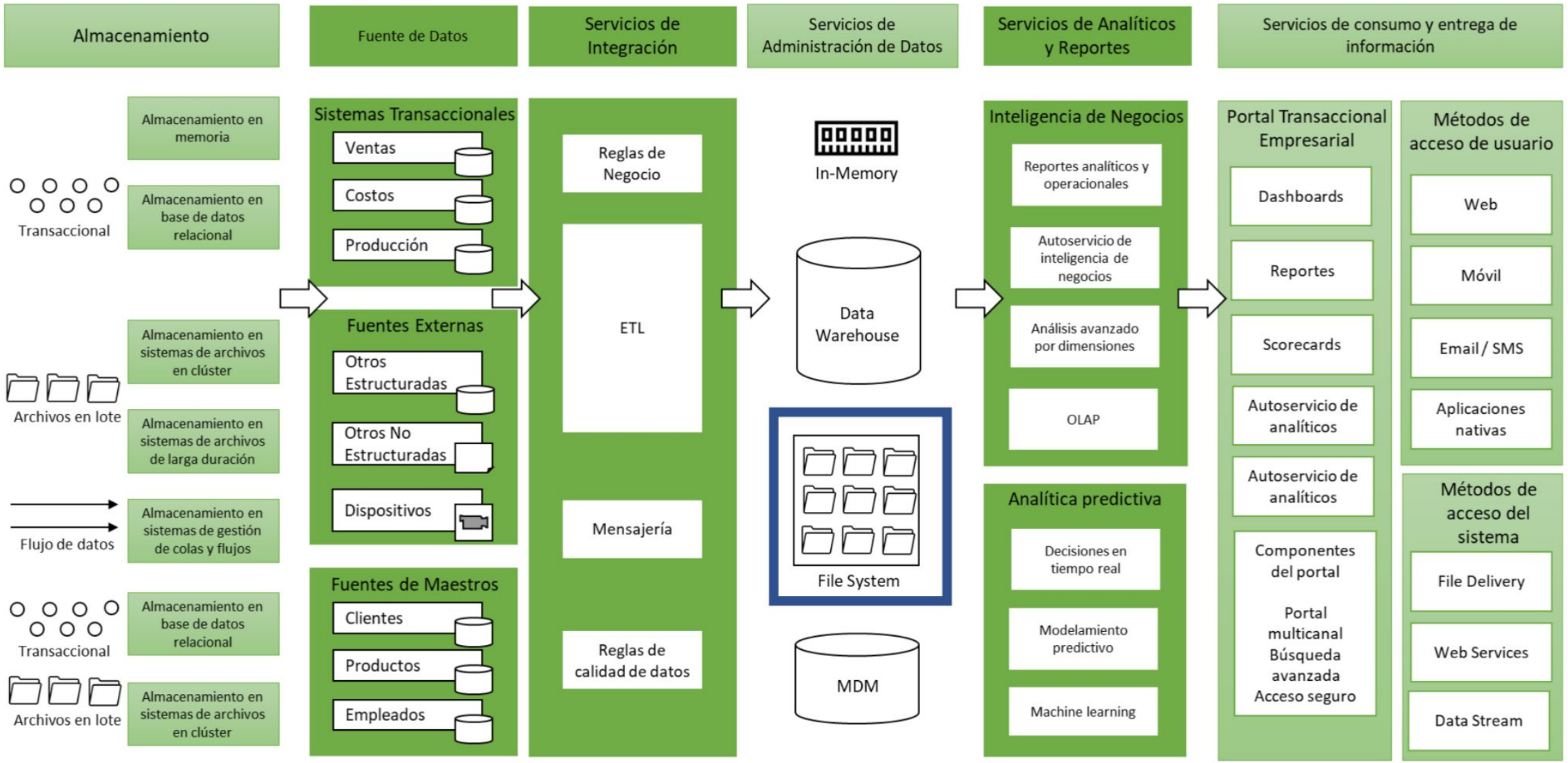


stucom



Tomado de: Gregory S. Nelson. The Analytics Lifecycle Toolkit. 2018.

stucom


Almacenamiento
Fuente de Datos
Servicios de Integración
Servicios de Administración de Datos
Servicios de Analíticos y Reportes
Servicios de consumo y entrega de información
Almacenamiento en memoria
Transaccional
Almacenamiento en base de datos relacional
Archivos en lote
Almacenamiento en sistemas de archivos en clúster
Almacenamiento en sistemas de archivos de larga duración
Flujo de datos
Almacenamiento en sistemas de gestión de colas y flujos
Transaccional
Almacenamiento en base de datos relacional
Archivos en lote
Almacenamiento en sistemas de archivos en clúster
Sistemas Transaccionales
Ventas
Costos
Producción
Fuentes Externas
Otros Estructuradas
Otros No Estructuradas
Dispositivos
Fuentes de Maestros
Clientes
Productos
Empleados
Reglas de Negocio
ETL
Mensajería
Reglas de calidad de datos
In-Memory
Data Warehouse
File System
MDM
Inteligencia de Negocios
Reportes analíticos y operacionales
Autoservicio de inteligencia de negocios
Análisis avanzado por dimensiones
OLAP
Analítica predictiva
Decisiones en tiempo real
Modelamiento predictivo
Machine learning
Portal Transaccional Empresarial
Dashboards
Reportes
Scorecards
Autoservicio de analíticos
Autoservicio de analíticos
Componentes del portal
Portal multicanal
Búsqueda avanzada
Acceso seguro
Métodos de acceso de usuario
Web
Móvil
Email / SMS
Aplicaciones nativas
Métodos de acceso del sistema
File Delivery
Web Services
Data Stream

stucom

IDE’s para el desarrollo de Python

Permite trabajar en un entorno no local y la creación de Notebooks 🚀
Las herramientas que mostramos anteriormente no son las únicas en donde compilar código de Python…

stucom

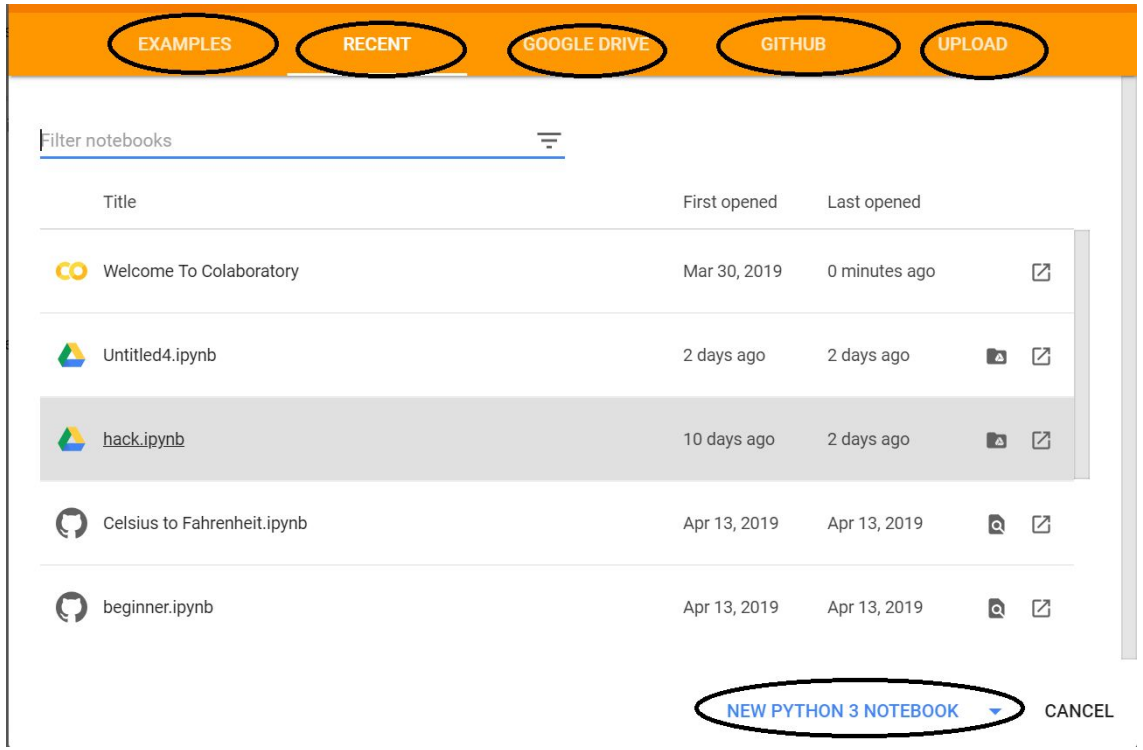
Google Colab

Permite trabajar en un entorno no local y la creación de Notebooks 🚀

- ✔ Es un producto de Google Research. Está especialmente adecuado para tareas de aprendizaje automático, análisis de datos y educación.
- ✔ Jupyter es el proyecto de código abierto en el que se basa Colab.
- ✔ Nos permite compartir notebooks sin la necesidad de descargar ningún software extra.
- ✔ El código se ejecuta en una máquina virtual dedicada a tu cuenta y pueden eliminarse luego de cierto tiempo.

stucom

Cómo usar Google Colab



EXAMPLES: Contiene ejemplos de Jupyter notebooks con diversos ejemplos.

RECENT: Jupyter notebooks que has trabajado recientemente.

GOOGLE DRIVE: Jupyter notebooks en tu google drive.

GITHUB: Puedes añadir Jupyter notebooks desde Github pero es necesario conectar Colab con GitHub.

UPLOAD: Si deseas subir un Jupyter notebook desde tu equipo local.

Ir al siguiente enlace: https://colab.research.google.com

stucom

VAMOS A APLICAR !!!

https://colab.research.google.com/drive/1FC7wngzYTGcO_n_Lv4RB3rHU0r2BVPKm?usp=share_link

https://www.kaggle.com/datasets/manovirat/groceries-sales-data


Gracias