



**UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO**  
**FACULTAD DE CIENCIAS DE LA INGENIERÍA**  
**INGENIERÍA EN SISTEMAS**



**ASIGNATURA:**

**ANÁLISIS INTELIGENTE DE DATOS**

**CURSO: VII “A”**

**TEMA:**

**SELECCIÓN Y LIMPIEZA DE LOS DATOS**

**AUTOR:**

**ABAD ALAY MARTÍN CARLOS**

**DOCENTE:**

**ING. JARAMILLO CHUQUI IVAN FREDY**

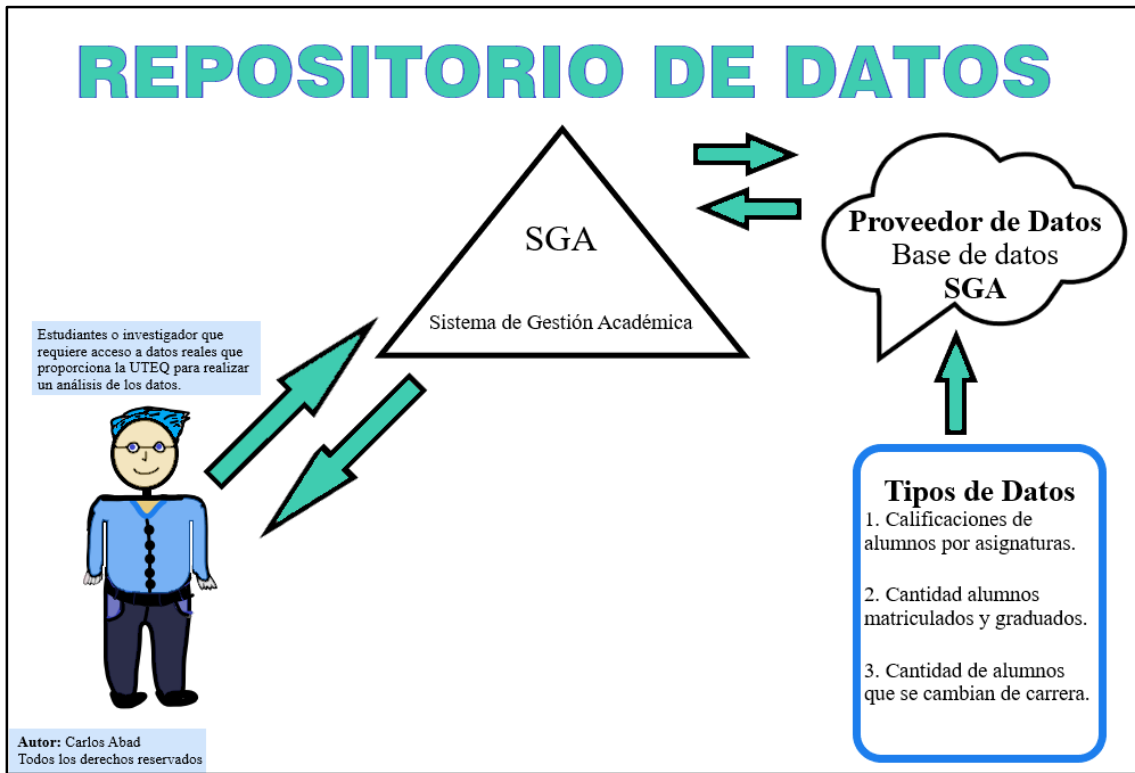
**PRIMER PERIODO ACADÉMICO:**

**2021 – 2022**

# STOP 1

## Los repositorios de acceso libre

¡Importante!, piensa como construir un repositorio de datos para la institución donde estudias. ¿Cuáles serían los proveedores de datos?, ¿Qué tipos de datos se guardarían?, ¿Dibuja un bosquejo de una posible arquitectura?



El estudiante, docente o investigador universitario debería de acceder a los datos registrado en la UTEQ a través del SGA, como, por ejemplo, las calificaciones de los alumnos en función de las asignaturas, los alumnos que se matriculan y se gradúan, los alumnos que se cambian de carrera y demás, se le puede hacer un análisis inteligente de estos datos.

## STOP 2

### Bases de datos transaccionales

¡Importante!, piensa como recolectaría datos desde una Base de Datos local.

Recolectaría datos a través de una aplicación de escritorio que tenga como objetivos principales:

#### Las ventas de productos y generación de facturas

**Facturación**

**Detalle de la orden del cliente**

Buscar Cliente Cliente: 1 ANGÉLICA LALALEO Celular: 0960533704

RUC/C.I.: 2100809637 Dirección: LAGO AGRIO - NUEVA LOJA

**Factura N°: 44012** **Detalle de la orden del cliente**

Factura N°: 44012 Código de Producto: 5 Cantidad: 1 Precio Unitario: 6.00

Agregar al carrito de compra Limpiar tabla Eliminar fila

Factura N°	Código de Producto	Precio Unitario	Cantidad	Descuento	TOTAL
44012	1	13.50	1	0.00	13.5
44012	13	12.00	1	0.00	12.0
44012	7	10.00	3	0.00	30.0
44012	21	10.50	2	0.00	21.0
44012	5	6.00	1	0.00	6.0

ENTREGA→ 100 CAMBIO→ 7,60

FACTURAR Imprimir

SubTotal: \$2,50  
IVA 12% 9,90  
TOTAL: \$2,40

#### Generación de reportes.

**REPORTE** **REGRESAR A REPORTES..!**

**FACTURA N°: 44012**

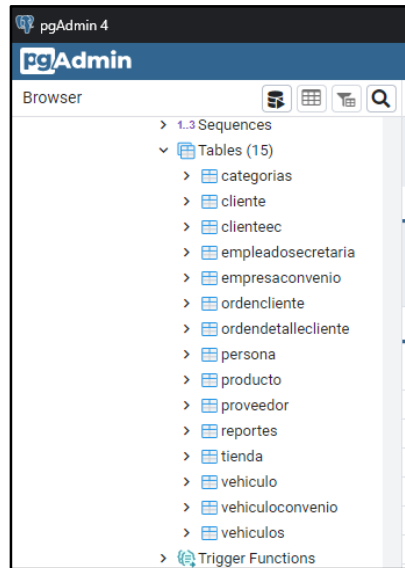
Cliente: LALALEO ANGÉLICA Atendido por: ABAD CARLOS

Dirección: LAGO AGRIO - NUEVA LOJA Fecha compra: 3/01/21 0:00

Producto	Cantidad	PrecioUnitario	SubTotal
Grasa azul 5 Libras	3	10.00	30.00
Filtro de motor	1	13.50	13.50
Agua Destilada o	1	12.00	12.00
Filtro de combustible	1	6.00	6.00
AMSOIL	2	10.50	21.00
SubTotal	null	null	82.50
IVA 12%	null	null	9.90
TOTAL A PAGAR	null	null	92.40

## ¿Cuáles podrían ser los facilitadores de datos?

Haciendo uso de PostgreSQL el cual es un gestor de base de datos que me facilitaría mucho en el almacenamiento de estos datos.



## ¿Qué tipos de datos contienen?

Los tipos de datos que contienen son los registros de clientes, productos y empleados, también de las ventas realizadas. Donde existen datos numéricos y textos.

## ¿Prepara un “Query” para extraer esos datos que te fascinan?

`select * from ordendetalcliente`

	ordenclienteid bigint	productoid integer	preciounitario numeric	cantidad integer	descuento numeric
1	1	1	10.30	10	0.0
2	2	3	15.00	10	0.0
3	8	2	12.30	1	0.0
4	8	1	10.30	1	0.0
5	9	2	12.30	1	0.0
6	1	1	10.30	1	0.0
7	2	2	12.30	1	0.0
8	3	3	15.00	1	0.0
9	4	4	10.00	1	0.0
10	5	5	20.00	1	0.0

## STOP 3

### Mecanismo de captura de datos

**¡Importante!, piensa como recolectaría datos con instrumentos, personas o dispositivos inteligentes**

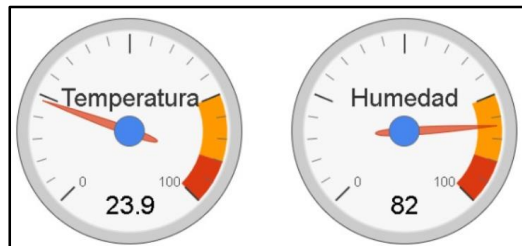
Recolectaría datos de dispositivos inteligentes

**¿Cuáles podrían ser los mecanismos?**

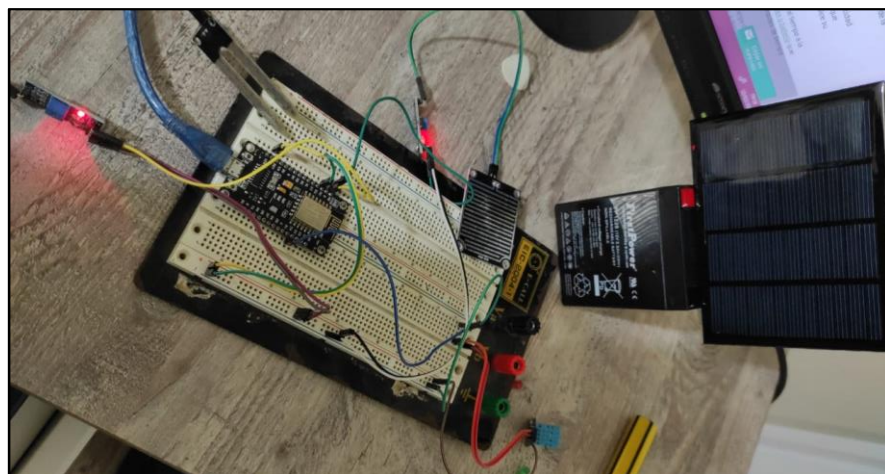
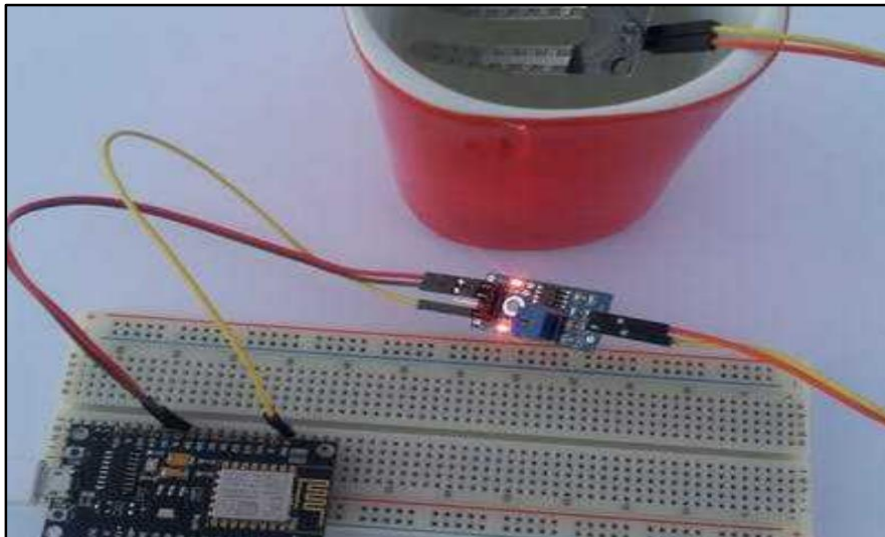
Con la ayuda de red de sensores inalámbricos.

**¿Qué tipos de datos capturarían?**

Recolectaría los datos de temperatura, humedad del suelo y medio ambiente a través de sensores inalámbricos.



**¿Prepara un dibujo sobre tu idea?**



## Aspectos de evaluación práctica

- Localiza el conjunto de datos “iris”.

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa

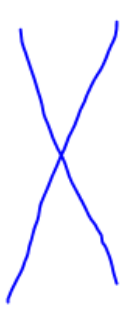
- Inserta aleatoriamente datos faltantes en dos de las tres columnas de tal forma que la tasa de valores perdidos en filas sea del 12%.

```
> #Creamos una variable irisTemp que obtendra el conjunto de datos iris.
> irisTemp=iris
>
> #A las variables n1 y n2 se le asigna números aleatorios no repetidos (pero al 12% de
> n1=sample(1:150,nrow(irisTemp)*0.12,replace=TRUE)
> n2=sample(2:150,nrow(irisTemp)*0.12,replace=TRUE)
>
> #A los números no repetidos se asigna NA
> irisTemp[n1,1]=NA
> irisTemp[n2,2]=NA
>
> #Eliminar la columna de texto
> irisTemp$Species<-NULL
> irisTemp
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	NA	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	NA	3.1	1.5	0.2
5	NA	NA	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	NA	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	NA	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3.0	1.4	0.1
14	4.3	3.0	1.1	0.1
15	5.8	4.0	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	NA	3.8	1.5	0.3

- Eliminamos la columna que contienen datos de tipo texto para evitar conflictos al momento de aplicar KNN

```
> #Eliminar la columna de texto
> irisTemp$Species<-NULL
>
> irisTemp
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1           NA          3.5           1.4          0.2
2          4.9          3.0           1.4          0.2
3          4.7          3.2           1.3          0.2
4           NA          3.1           1.5          0.2
5           NA          NA           1.4          0.2
6          5.4          3.9           1.7          0.4
7          4.6          NA           1.4          0.3
8          5.0          3.4           1.5          0.2
9          4.4          NA           1.4          0.2
```



- Luego aplica las técnicas de completamiento KNN

```
> #Aplica las técnicas de completamiento KNN
> dat.knn=ec.knnimp(irisTemp,k=10)
> dat.knn
  Sepal.Length Sepal.Width Petal.Length Petal.Width
[1,]          5.14          3.50          1.4          0.2
[2,]          4.90          3.00          1.4          0.2
[3,]          4.70          3.20          1.3          0.2
[4,]          4.84          3.10          1.5          0.2
[5,]          4.93          3.35          1.4          0.2
[6,]          5.40          3.90          1.7          0.4
[7,]          4.60          3.06          1.4          0.3
[8,]          5.00          3.40          1.5          0.2
[9,]          4.40          3.07          1.4          0.2
[10,]          4.90          3.10          1.5          0.1
[11,]          5.40          3.70          1.5          0.2
[12,]          4.80          3.40          1.6          0.2
[13,]          4.80          3.00          1.4          0.1
[14,]          4.30          3.00          1.1          0.1
[15,]          5.80          4.00          1.2          0.2
[16,]          5.70          4.40          1.5          0.4
[17,]          5.40          3.90          1.3          0.4
[18,]          5.10          3.50          1.4          0.3
[19,]          5.70          3.80          1.7          0.3
[20,]          5.26          3.80          1.5          0.3
[21,]          5.40          3.40          1.7          0.2
[22,]          5.10          3.70          1.5          0.4
[23,]          4.60          3.60          1.0          0.2
[24,]          5.10          3.30          1.7          0.5
[25,]          4.80          3.40          1.9          0.2
[26,]          5.00          3.00          1.6          0.2
[27,]          5.00          3.40          1.6          0.4
[28,]          5.20          3.50          1.5          0.2
[29,]          5.20          3.40          1.4          0.2
[30,]          4.70          3.20          1.6          0.2
```



➤ **Compara con los valores reales. Escribe tu conclusión.**

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

**Ilustración 1:** Valores reales del conjunto de datos iris

```
> #Aplica las técnicas de completamiento KNN
> dat.knn=ec.knnimp(irisTemp,k=10)
> dat.knn
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
[1,]	5.14	3.50	1.4	0.2
[2,]	4.90	3.00	1.4	0.2
[3,]	4.70	3.20	1.3	0.2
[4,]	4.84	3.10	1.5	0.2
[5,]	4.93	3.35	1.4	0.2
[6,]	5.40	3.90	1.7	0.4
[7,]	4.60	3.06	1.4	0.3
[8,]	5.00	3.40	1.5	0.2
[9,]	4.40	3.07	1.4	0.2
[10,]	4.90	3.10	1.5	0.1

**Ilustración 2:** Valores aplicando completamiento KNN, donde k=10.

### Conclusiones

- En esta práctica se completó los valores faltantes aplicando técnicas de completamiento K Vecinos Cercanos (K-nearest neighbors). En la fila 1 de la ilustración tiene el valor de **5.1**, mientras que con aplicando el completamiento KNN el valor obtenido fue de **5.14**.
- Para los atributos continuos, se reemplaza el valor faltante por la media del atributo en la vecindad de los k vecinos más cercanos (knearest neighborhood).
- Encontrar el valor faltante depende también del valor que tome “**K**”, en función de ese valor se obtiene el valor faltante. Y si el valor de **k** es muy grande, la instrucción tardaría más de lo normal.