

## ICO 187 ANÁLISIS DE DATOS

### CLASE 08: MANIPULACIÓN E INTEGRACIÓN DE DATOS EN EXCEL

Año: 2021  
Profesor: Sebastián Egaña

#### 1. Repaso de la clase pasada: Unión de bases de datos (Joins).

Recordar que vimos algunas formas de generar uniones entre bases de datos. Esto puede cambiar en relación a nuestra necesidad, y la finalidad de nuestro trabajo.

Ahora intentaremos realizar un trabajo aplicado sobre el tema.

#### 2. Set de datos

Considere que se tiene un set de datos clase\_08.xlsx, que contiene 4 tablas:

- Flights: Corresponde a distintos identificadores de vuelos por ruta.
- Routes: Rutas de vuelos.
- Aircraft: Tipos de aviones.
- Airports: Set de datos descriptor de aeropuertos en USA.

#### 3. Import - Tidy - Transform

Se piensa que importar, solo refiere a poder obtener los datos. Desde mi punto de vista, este corresponde a un paso para poder mirar los datos y poder hacernos algunas ideas de los datos. Veamos en detalle cada set de datos.

- ¿Cuáles son las variables de cada set de datos?

Por ejemplo, ¿qué es FuelCostperSeatMile (Cents)?

[Acá](#)

Pero una pregunta más general es, ¿por qué es relevante aprender estos conceptos?

- ¿Se observa algún error en los datos?

Fijémonos en el formato de la fecha.

- ¿Posee la estructura correcta el set de datos?

No se ven errores; en las filas están las observaciones y en las columnas las variables. No existen formatos, ni celdas combinadas.

- ¿Existen identificadores únicos en los set de datos?

Existen múltiples identificadores únicos.

Se debe intentar llegar al análisis relevante del set de datos. En el caso de este set de datos, claramente se puede llegar a un set de datos más completo uniéndolo con los otros. Sobre lo mismo:

- ¿Cuál es el set de datos que debe mantenerse? Veamos alguna manera de llegar a dicha conclusión.

Por último, se puede pensar en la inclusión de datos a la base de datos, que podrían ser de algún interés. En este caso intentaremos agregar los siguientes datos:

[Acá](#)

Hasta este punto, aún no hemos terminado con dicha procedimiento de **batallar con los datos**.

Antes de aplicar la unión de datos, sería bueno poder diagramar la relación que existe entre cada base datos. Acá debemos abordar un contenido relacionado.

## 4. Modelo de datos

Corresponde a una serie de conceptos que pueden utilizarse para describir un conjunto de datos y las operaciones relacionadas para su manipulación.

En la actualidad el modelo de bases de datos más utilizado corresponde al modelo relacional, que se define como el intento de obtener datos de distintas fuentes a través de relaciones o consultas para la obtención de datos. Esto quiere decir, que no necesariamente se tiene una **base gigante**, sino que se tiene la opción de acceder a bases pequeñas que pueden ir complementando los análisis.

### 4.1. Modelo Relacional

Las entidades y relaciones se representan en formas de tablas.

- Las tablas son las relaciones.
- Las filas (tuplas) contienen datos sobre cada entidad.
- Las columnas corresponden a atributos de las entidades.

Se pretende determinar operaciones a realizar: Unión, intersección, diferencia, producto cartesiano, selección, proyección, reunión, etc.

Por otra parte, existen restricciones de integridad de entidad como también de integridad referencial (relacionado con el tema de llaves que vimos anteriormente).

Por ejemplo:

Entidades	Proveedor	Pieza
Atributos	Código	Código
	Nombre	Nombre
	Ciudad	Dimensiones Peso

En donde:

Relación	Suministra
Entidades Participantes	Proveedor - Pieza
Cardinalidad	Muchos a muchos
Atributos	Cantidad

De manera más específica, el modelo entidad/relación corresponde a la técnica basada en la identificación de las entidades y de las relaciones que se dan entre ellas, según la realidad que se intenta modelar.

Se debe recordar acá, el concepto de llave del que hablamos la clase pasada. Una llave (o también llamada clave), corresponde a un conjunto de atributos que permite identificar unívocamente a una entidad dentro de un conjunto de entidades.

Por ejemplo para una Facultad el modelo podría ser:

Entidades Atributos	Asignatura ID Nombre Créditos Curso	Alumno RUT Nombre Dirección Email	Profesor RUT Nombre Categoría Email	Departamento ID Nombre	Aula ID Capacidad	Grupo ID Tipo
------------------------	---	---	---	------------------------------	-------------------------	---------------------

En donde existen las siguientes relaciones:

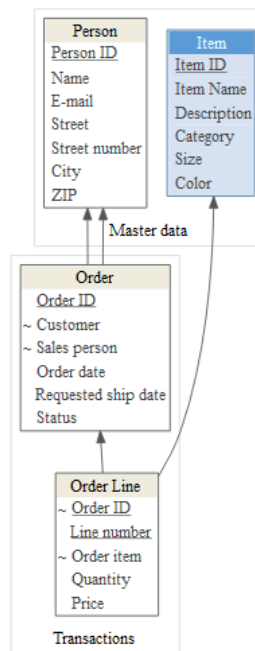
Relación	Entidades participantes	Cardinalidad	Atributos
se matricula en	Alumno - Grupo	N:M	Calificación
enseña	Profesor - Grupo	N:M	
impartida en	Asignatura - Grupo	1:N	
asignada a	Aula - Grupo	N:M	Día, hora
pertenece a	Profesor - Departamento	N:1	
dirige	Profesor - Departamento	1:1	

Esto también puede ser representado de manera grafica. Veamos el siguiente ejemplo:

```
library(datamodelr)
file_path <- system.file("samples/example.yml", package = "datamodelr")
dm <- dm_read_yaml(file_path)

graph <- dm_create_graph(dm, rankdir = "BT")

dm_render_graph(graph)
```



Veamos esto en relación a los datos que estuvimos revisando; primero debemos ingresar las tablas sobre las que estuvimos trabajando.

```
library(readxl)

flights_01 <- read_excel("G:/My Drive/Docencia 2020 - 2021/UST - 2021 01 Análisis de Datos/excel/clase_08.xlsx",
  sheet = "Flights")

routes_01 <- read_excel("G:/My Drive/Docencia 2020 - 2021/UST - 2021 01 Análisis de Datos/excel/clase_08.xlsx",
  sheet = "Routes")

airports_01 <- read_excel("G:/My Drive/Docencia 2020 - 2021/UST - 2021 01 Análisis de Datos/excel/clase_08.xlsx",
  sheet = "Airports")
```

```
aircraft_01 <- read_excel("G:/My Drive/Docencia 2020 - 2021/UST - 2021 01 Análisis de Datos/excel/clase_08.
                        sheet = "Aircraft")
```

```
dm_f <- dm_from_data_frames(flights_01, routes_01, aircraft_01, airports_01)
graph <- dm_create_graph(dm_f, rankdir = "BT", col_attr = c("column", "type"))
dm_render_graph(graph)
```

aircraft_01	
AircraftID	numeric
AircraftType	character
SeatCapacity	numeric
FuelCostperSeatMile (Cents)	numeric

airports_01	
AirportID	character
AirportName	character
Gates (number of available)	numeric

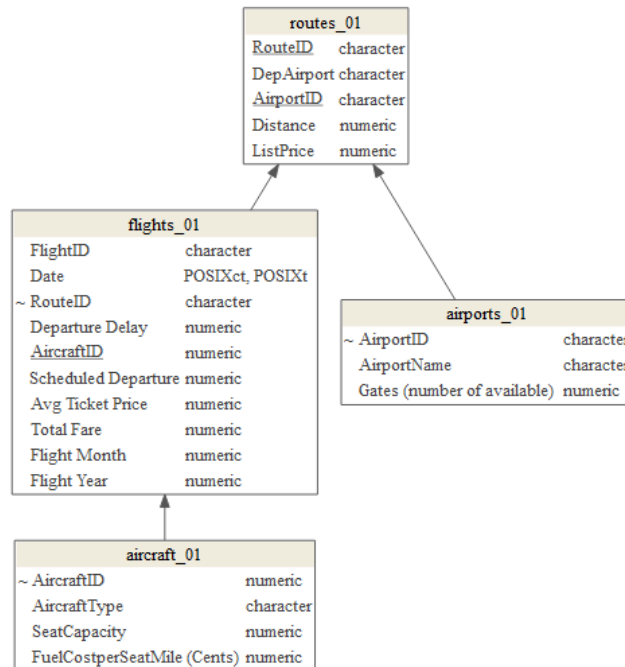
flights_01	
FlightID	character
Date	POSIXct, POSIXt
RouteID	character
Departure Delay	numeric
AircraftID	numeric
Scheduled Departure	numeric
Avg Ticket Price	numeric
Total Fare	numeric
Flight Month	numeric
Flight Year	numeric

routes_01	
RouteID	character
DepAirport	character
AirportID	character
Distance	numeric
ListPrice	numeric

Después debemos generar las relaciones entre cada una de las tablas:

```
dm_f <- dm_add_references(
  dm_f,

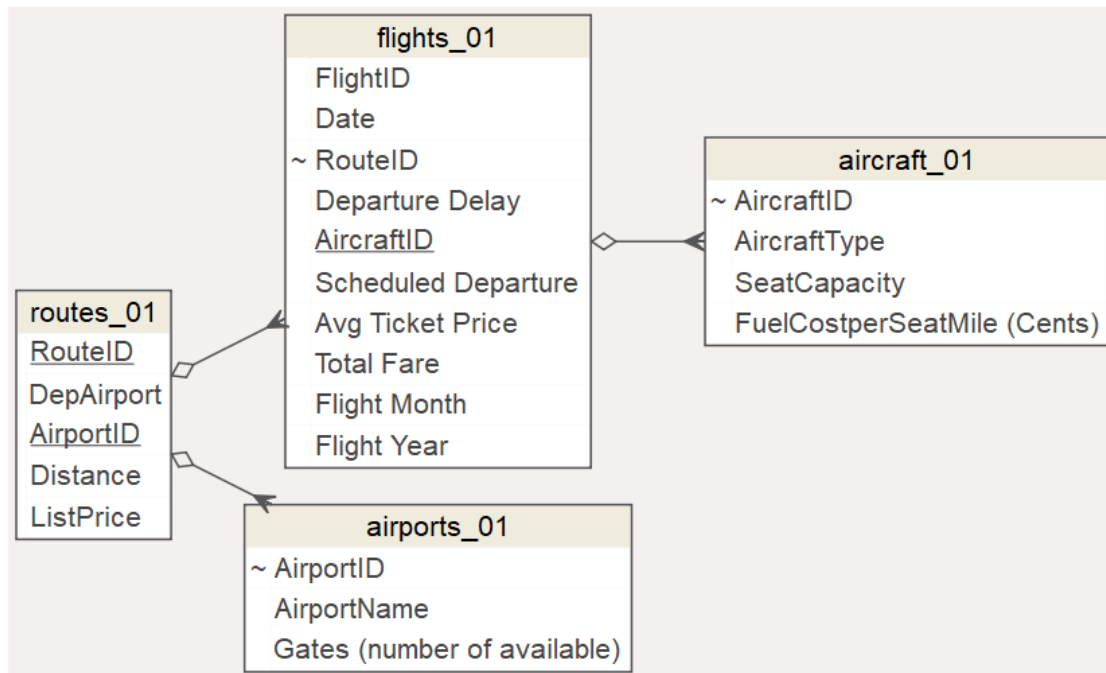
  aircraft_01$AircraftID == flights_01$AircraftID,
  airports_01$AirportID == routes_01$AirportID,
  flights_01$RouteID == routes_01$RouteID
)
graph <- dm_create_graph(dm_f, rankdir = "BT", col_attr = c("column", "type"))
dm_render_graph(graph)
```



Podemos intentar mejorar un poco el formato de la visualización:

```
graph <- dm_create_graph(
  dm_f,
  graph_attrs = "rankdir = RL, bgcolor = '#F4F0EF' ",
  edge_attrs = "dir = both, arrowtail = crow, arrowhead = odiamond",
  node_attrs = "fontname = 'Arial'")

dm_render_graph(graph)
```



## 5. Avisos

- Fue subida la Guía en Grupo N01. Recordar que deben enviar el set de datos para su aprobación y los integrantes del grupo hasta el final de la semana (3 o 4 integrantes, sin excepciones)
- Se actualizo el punto 5 de las instrucciones.

## 6. Fechas Relevantes

Unidad	Evaluación	Ponderación	Fecha
Unidad I	Evaluación diagnóstica		25/03/2021
	Evaluación Individual Participación	(5 %)	05/04/2021
	Evaluación Grupal	(15 %)	27/04/2021 - 04/05/2021
	Evaluación Individual - Sumativa I	(30 %)	11/05/2021
Unidad II	Evaluación Formativa		13/05/2021
	Evaluación Individual Participación	(5 %)	27/05/2021
	Evaluación Grupal	(10 %)	08/06/2021 - 15/06/2021
	Evaluación Individual - Sumativa II	(15 %)	17/06/2021
Unidad III	Evaluación Formativa		22/06/2021
	Evaluación Individual Participación	(5 %)	24/06/2021
	Evaluación Individual Sesión I- Sumativa III	(15 %)	08/07/2021
	Evaluación Individual Sesión II- Sumativa III	(15 %)	13/07/2021