ICO 187 ANÁLISIS DE DATOS

Clase 09: 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 estamos trabajando con un set de datos relacionado a vuelos. Sobre lo mismo, es necesario analizar de manera más estricta las relaciones entre cada set de datos por lo que debemos hablar ahora de los **modelos de datos**.

2. 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. 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.

2.1. 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 especifica, 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	Asignatura	Alumno	Profesor	Departamento	Aula	Grupo
Atributos	ID	RUT	RUT	ID	ID	ID
	Nombre	Nombre	Nombre	Nombre	Capacidad	Tipo
	Créditos	Dirección	Categoría			
	Curso	Email	Email			

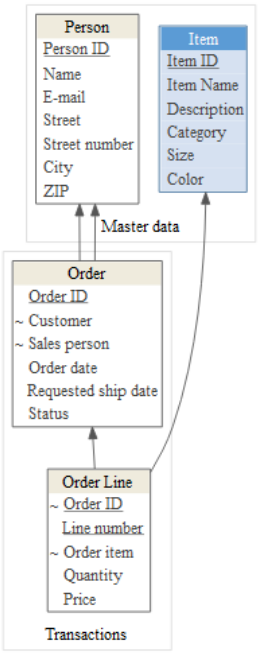
En donde existen las siguientes relaciones:

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:

Código en R:

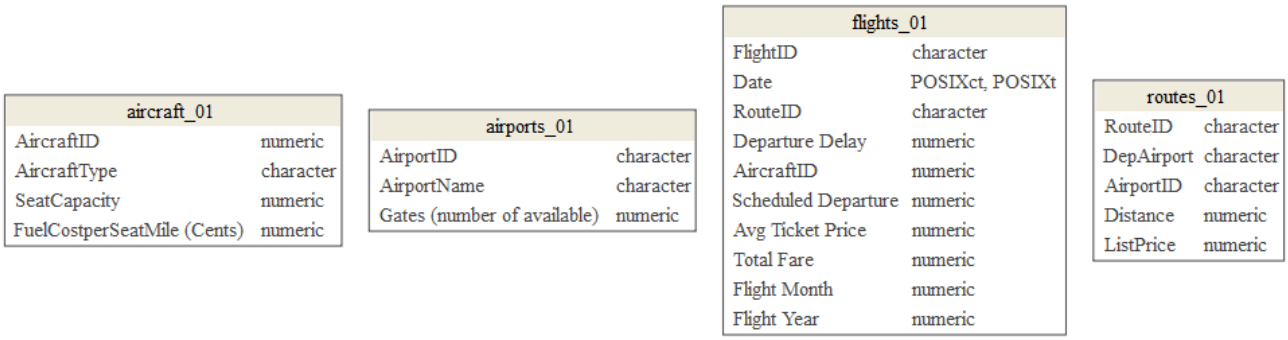
```
library(datamodelr)
file_path <- system.file("samples/example.yml", package = "datamodelr")
dm <- dm_read_yaml(file_path)

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.

Código en R:

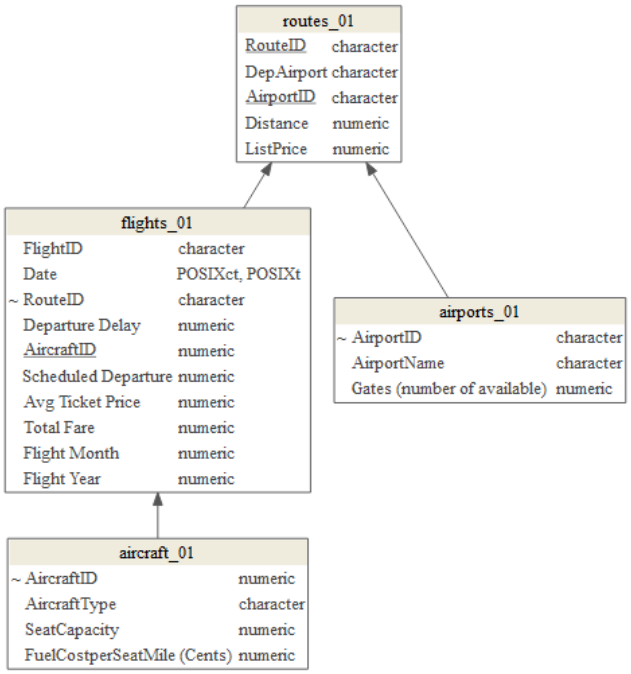
```
library(readxl)

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

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

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



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

Código en R:

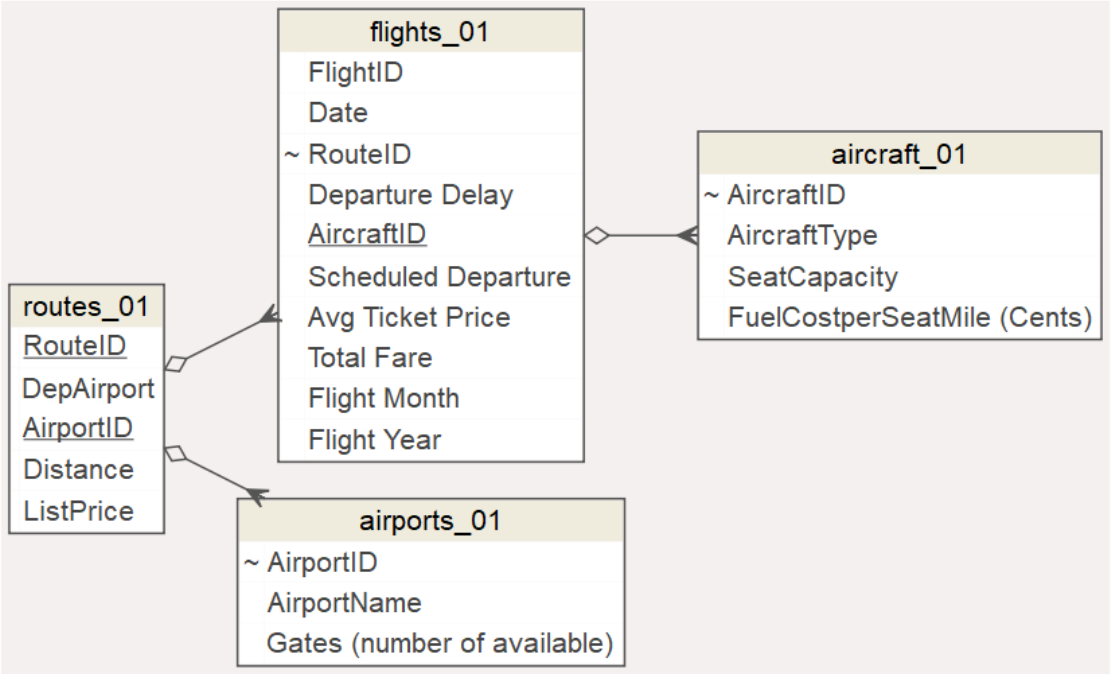
```
dm_f <- dm_add_references(
  dm_f,

  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:

Código en R:

```
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)
```



- Ahora corresponde aplicar esto en el set de datos.

3. Avisos

- Solo tengo claridad en los integrantes de dos gurpos. Nadie me ha enviado la base de datos.

4. 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