DWH-DM
Flores González Luis Brandon
Tarea No. “2”
28 - 02 - 2018

1. Extracción, transformación y carga

Análisis de la fuentes de datos:

1. Fuente1.txt

El tipo de dato en este análisis siempre es cadena, ya que proviene de un archivo de texto. El tipo de dato se agregara en la unificación.

No hay reglas de integridad ya que es un archivo de texto. Solo se puede inferir que hay una llave primaria.

Atributo	Tipo de dato	Dominio(Valores válidos)	Reglas de integridad
Orden	Cadena	Número entero de 5 digitos.	Llave primaria
Nombre cliente	Cadena	Nombre y apellido con capitalización.	-
Ciudad cliente	Cadena	Nombre de la ciudad con primera letra mayuscula.	-
Estado/provincia cliente	Cadena	Representación del estado/provincia del cliente con dos caracteres en mayúsculas.	-
Estatus del cliente	Cadena	Estatus del cliente con primera letra en mayuscula. Ejemplos “Gold”, “Silver”, “Platinum”.	-
Fecha de orden	Cadena	Fecha en el formato mmm-dd donde mmm son los primeros caracteres del mes y dd es el número del día.	-
ID producto	Cadena	Identificador del producto con un número entero.	Llave primaria.
Producto	Cadena	Nombre del producto capitalizado.	-
Precio x unidad	Cadena	Precio con signo de pesos y en double(decimal).	-

Cantidad	Cadena	Valor entero mayor o igual que 0.	-
Descuento	Cadena	Valor double mayor o igual que 0.	-
Precio completo	Cadena	Valor double mayor o igual que 0, con signo de pesos.	-
Precio extendido	Cadena	Valor double mayor o igual que 0, con signo de pesos.	-
Descuento total		Valor double mayor o igual que 0, con signo de pesos.	
A	Cadena	Vacio	Vacio
B	Cadena	Vacio	Vacio
C	Cadena	Vacio	Vacio
D	Cadena	Vacio	Vacio

2. **Fuente2.csv**

El tipo de dato en este análisis siempre es cadena, ya que proviene de un archivo de texto. El tipo de dato se agregara en la unificación.

No hay reglas de integridad ya que es un archivo de texto. Solo se puede inferir que hay llave primaria.

Atributo	**Tipo de dato**	**Dominio(Valores válidos)**	**Reglas de integridad**
IDOrden	Cadena	Cadena ANNNNN, donde A es un caracter en mayuscula y N es un digito del 0 al 9.	Llave primaria.
Nombre cliente	Cadena	Nombre(cadena) capitalizado.	
Apellidos cliente	Cadena	Apellido(cadena) capitalizado.	
Ciudad	Cadena	Cadena(ciudad) capitalizada.	
Estado/provincia	Cadena	Cadena(estado/provincia) capitalizada.	
Tipo cliente	Cadena	Número de un dígito.	
fecha	Cadena	Fecha en formato mmm-dd donde mmm es	

		un mes(primeros digitos) y dd es el número del día.	
idprod	Cadena	Número entero.	Llave primaria.
Nombre producto	Cadena	Nombre capitalizado.	
Precio unitario	Cadena	Valor double, con signo de pesos al comienzo.	
Piezas	Cadena	Valor entero mayor o igual que 0.	
Descuento	Cadena	Valor decimal mayor o igual que 0, el primer dígito es un signo de pesos.	
Completo	Cadena	Valor decimal mayor o igual que 0, el primer dígito es un signo de pesos.	
Extendido	Cadena	Valor decimal mayor o igual que 0, el primer dígito es un signo de pesos.	
Descuento total	Cadena	Vacio.	

3. **Fuente3.xls**

El tipo de dato en este análisis siempre es cadena, ya que proviene de un archivo de texto. El tipo de dato se agregara en la unificación.

Atributo	**Tipo de dato**	**Dominio(Valores válidos)**	**Representación de valores nulos.**	**Reglas de integridad**
Customer Name	Cadena	Cadena(nombre o nombres) capitalizada.		-
Previous Credit Line	Cadena	Valor entero mayor o igual que 0.		-
New Credit Line	Cadena	Valor entero mayor o igual que 0.		-
Missed Payments	Cadena	Valor entero mayor o igual que 0 ó cadena “NONE”.	NONE	-
Country	Cadena	Cadena del país capitalizada o siglas este.		-

Reconciliación de datos:

Se modificó el nombre de los atributos, se agregó de donde provienen estos y la transformaciones que se usaron.

Atributo	Fuente	Transformación empleada.
id_orden	- Fuente1.txt: Orden - Fuente2.csv: IDOrden	- No hay duplicados ya que varía en el atributo producto. - A la fuente 2 se le quita una “A” a cada dato para que coincida con el formato de la fuente 1. - Se cambió a número. - Nuevo valor null, 0.
nombre_cliente	- Fuente1.txt: Nombre cliente - Fuente2.csv: Nombre cliente - Fuente3.xls: Customer Name	- Volvemos atómico el campo, para que aparezca solo el nombre del cliente sin el apellido.
apellido_cliente	- Fuente1.txt: Nombre cliente - Fuente2.csv: Apellidos cliente - Fuente3.xls: Customer Name	- Volvemos atómico el campo, para que aparezca solo el apellido del cliente sin el nombre.
ciudad_cliente	- Fuente1.txt: Ciudad cliente - Fuente2.csv: Ciudad	- Sin transformación. - Nuevo valor null, cadena “N”
estado/provincia_cliente	- Fuente1.txt: Estado/provincia cliente - Fuente2.csv: Estado/provincia	- Transformamos los datos de la fuente 1 a la fuente 2(de dos caracteres a nombre completo), para que cada uno tenga solo la representación y así mantener solo una representación. - Nuevo valor null, cadena “N”

estatus_cliente	- Fuente1.txt: Estatus del cliente
	- Fuente2.csv: Tipo cliente
		- Nuevo valor null, cadena “N”
		- Pasamos las celdas de Tipo cliente a la representación de Estatus del cliente, donde:
		  - 1 es
		  - 2 es
		  - 3 es

fecha_orden	- Fuente1.txt: Fecha de orden - Fuente2.csv: fecha	- Nuevo valor null, cadena “N”
id_producto	- Fuente1.txt: ID producto - Fuente2.csv: idprod	- Se cambió a número. - Nuevo valor null, número “0”
producto	- Fuente1.txt: Producto - Fuente2.csv: Nombre producto	- Nuevo valor null, cadena “N”

precio_unitario	- Fuente1.txt: Precio x unidad - Fuente2.csv: Precio unitario	- Se le quita el signo de pesos. - Se cambió a número. - Nuevo valor null, número “0”
cantidad	- Fuente1.txt: Cantidad - Fuente2.csv: Piezas	- Se cambió a número. - Nuevo valor null, número “0”
descuento	- Fuente1.txt: Descuento - Fuente2.csv: Descuento	- Se le quita el signo de pesos. - Se cambió a número. - Nuevo valor null, número “0”
precio_completo	- Fuente1.txt: Precio completo - Fuente2.csv: Completo	- Se le quita el signo de pesos. - Se cambió a número. - Nuevo valor null, número “0”
precio_extendido	- Fuente1.txt: Precio extendido - Fuente2.csv: Extendido	- Se le quita el signo de pesos. - Se cambió a número. - Nuevo valor null, número “0”
descuento_total	- Fuente1.txt: Descuento total - Fuente2.csv: Descuento total	- Se le quita el signo de pesos. - Se cambió a número. - Nuevo valor null, número “0”
credito_previo	- Fuente3.xls: Previous Credit Line	- Sin transformación. - Nuevo valor nulo, 0.
nueva_linea_credito	- Fuente3.xls: New Credit Line	- Sin transformación. - Nuevo valor nulo, 0.
pagos_perdidos	- Fuente3.xls: Missed Payments	- Se cambió a número. - Nuevo valor de vació es 0.
pais	- Fuente3.xls: Country	- Se estandarizó a nombre completo de país. - Nuevo valor null, cadena “N”

2. Limpieza de datos

Análisis de la fuente de datos:

Se analiza el dataset proporcionado y se especifican los atributos y el dominio que estos tienen, al ser un .csv solo podemos dar los identificadores como reglas de integridad.

Atributo	Descripción	Dominio(Valores válidos)	Reglas de integridad
Title	Título de la obra.	Varias cadenas separadas por una coma. Estas mismas forman el título, así lo maneja la página.	-
Artist	Artista de la obra.	Nombres y apellidos del artista. Es decir, cadenas.	-

ConstituentID	Identificador del constituyente.	Número entero identificador.	Llave primaria.
ArtistBio	Nacionalidad, año de nacimiento y año de muerte.	Fecha de nacimiento y de muerte así como la nacionalidad. En formato de cadena.	-
Nationality	Nacionalidad.	Nacionalidad del autor. En cadena.	-
BeginDate	Año de nacimiento.	Año de nacimiento del autor. En formato numérico.	-
EndDate	Año de muerte.	Año de muerte del autor. En formato numérico.	-
Gender	Género.	Género del autor delimitado por paréntesis.	-
Date	Fecha de la obra.	Formato fecha.	-
Medium	Medios que se usarón para crear la obra.	Cadenas separadas por coma.	-
Dimensions	Dimensiones de la obra.	Dimensiones y colores de la obra en formato cadena.	-
CreditLine	Leyenda que describen los créditos(Autor) de la obra.	Formato cadena.	-
AccessionNumber	Número de acceso.	Entero o decimal.	-
Classification	Clasificación de la obra.	Cadena.	-
Department	Departamento al que pertenece la obra.	Cadena.	-
DateAcquired	Fecha en la que se adquirio.	Formato de fecha.	-
Cataloged	Si fue catalogada o no.	Carácter Y ó N.	-
ObjectID	Identificador del objeto.	Número entero.	Identificador.
URL	Enlace de la obra.	Enlace web.	-
ThumbnailURL	Thumbnail de la obra.	Enlace web.	-
Circumference (cm)	Medidas de la obra con respecto a la circunferencia.	Entero o decimal.	-
Depth (cm)	Medidas de la obra con respecto a la profundidad.	Entero o decimal.	-

Diameter (cm)	Medidas de la obra con respecto al diámetro.	Entero o decimal.	-
Height (cm)	Medidas de la obra con respecto a la altura.	Entero o decimal.	-
Length (cm)	Medidas de la obra con respecto a la longitud.	Entero o decimal.	-
Weight (kg)	Medidas de la obra con respecto al peso.	Entero o decimal.	-
Width (cm)	Medidas de la obra con respecto a la anchura.	Entero o decimal.	-
Seat Height (cm)	Medidas de la obra con respecto a la altura del asiento.	Entero o decimal.	-
Duration (sec.)	Medidas de la obra con respecto a la duración.	Entero o decimal.	-

Transformación:

Aquí se muestra la limpieza de datos que se hizo, es decir, transformaciones al dataset.

Atributo	**Inconsistencia**	**Decisión tomada**	**Transformación**
ArtistBio	Valores innecesarios ya que están presentes en otros atributos.	Se elimina la columna ya que es innecesaria.	Removemos la columna ArtistBio.
ConstituentID	- Valores no numéricos. - Valor null.	- Convertir las cadenas en números. - Cambiar representación	- Dividir los identificadores en 2, ya que hay casos en los que hay dos artistas. - Cambio null por “0”.
Artist	Celdas con múltiples artistas. Valor null.	Dividir celdas multivaluadas. Cambiar representación.	Dividir por coma como separador. Cambio null por “N”.
BeginDate	- Signo - al comienzo del año. - Campos multivaluados. - Valor null.	- Eliminar signo. - Dividir celdas multivaluadas. - Cambiar representación.	- Reemplazar - por “”(blanco) para que solo sea numérico. - Dividir por espacio. - Cambio null por “0”.
EndDate	- Signo - al comienzo del año. - Campos	- Eliminar signo. - Dividir celdas multivaluadas. - Cambiar	- Reemplazar - por “”(blanco) para que solo sea numérico.

	multivaluados. - Valor null.	representación.	- Dividir por espacio. - Cambio null por “0”.
Gender	- Campos multivaluados. - Valor null.	- Dividir celdas multivaluadas. - Cambiar representación.	- Dividir por espacio. - Cambio null por ()”.
DateAcquired	- Sin formato - Valor null.	- Cambiar al formato correcto - Cambiar representación.	- Nuevo formato fecha. - Cambio null por “0”.
Date	- Sin formato - Valor null.	- Cambiar al formato correcto - Cambiar representación.	- Nuevo formato fecha. - Cambio null por ()”.
Circumference (cm)	Columna innecesaria.	Eliminar columna ya que solo contiene un dato.	Eliminar columna.
Seat Height (cm)	Columna innecesaria.	Eliminar columna ya que solo contiene un dato.	Eliminar columna.
Duration (sec.)	Columna innecesaria.	Eliminar columna ya que solo contiene un dato.	Eliminar columna.
Weight (kg)	Columna innecesaria.	Eliminar columna ya que solo contiene un dato.	Eliminar columna.
Depth (cm)	Valor null.	Cambiar representación.	Cambio null por “0”.
Diameter (cm)	Valor null.	Cambiar representación.	Cambio null por “0”.
Height (cm)	Valor null.	Cambiar representación.	Cambio null por “0”.
Length (cm)	Valor null.	Cambiar representación.	Cambio null por “0”.
Length (cm)	Valor null.	Cambiar representación.	Cambio null por “0”.
Title	Valor null.	Cambiar representación.	Cambio null por “N”.
Nationality	- Campos multivaluados. - Valor null.	- Dividir celdas multivaluadas. - Cambiar representación.	- Dividir por espacio. - Cambio null por “N”.
Medium	Valor null.	Cambiar representación.	Cambio null por “N”.
Dimensions	Valor null.	Cambiar representación.	Cambio null por “N”.
CreditLine	Valor null.	Cambiar representación.	Cambio null por “N”.
Dimensions	Valor null.	Cambiar representación.	Cambio null por “N”.
AccessionNumber	Valor null.	Cambiar representación.	Cambio null por “N”.
Classification	Valor null.	Cambiar representación.	Cambio null por “N”.

Department	Valor null.	Cambiar representación.	Cambio null por “N”.
Cataloged	Valor null.	Cambiar representación.	Cambio null por “Desconocido”.
ObjectID	Valor null.	Cambiar representación.	Cambio null por “0”.
URL	Valor null.	Cambiar representación.	Cambio null por “N”.
ThumbnailURL	Valor null.	Cambiar representación.	Cambio null por “N”.

Significado de los archivos:

- Fuentes transformadas/
    - Fuente-1.csv: resultado de la fuente 1 transformada.
    - Fuente-2.csv: resultado de la fuente 2 transformada.
    - Fuente3.csv: resultado de la fuente 3 transformada.
    - Fuente-1.openrefine.tar.gz: proyecto correspondiente a la fuente 1.
    - Fuente-2.openrefine.tar.gz: proyecto correspondiente a la fuente 2.
    - Fuente3.openrefine.tar.gz: proyecto correspondiente a la fuente 3.
    - Reconciliacion_fuente1_fuente2.csv: primera unificación de la fuente 1 y 2.
    - Reconciliacion_Fuente1_Fuente2_Fuente3.csv: unificación final de todas las fuentes.

- dataset/
    - MoMA.csv: dataset sin errores.
    - MoMA.openrefine.tar.gz: proyecto correspondiente al dataset sin errores.

- Fuentes/
    - Dentro de esta carpeta se encuentran las diferentes fuentes originales que se usarón para hacer los diferentes análisis.