2 de febrero de 2022

NIXON JAVIER VUELE IRENE

UNIVERSIDAD TECNICA PARTICULAR DE LOJA

LOJA

Proyecto DE BASE DE DATOS

MANEJO DE CSV EN MY SQL

**Introducción**

En la actualidad el manejo de bases de datos, datos e información en general es fundamental para el funcionamiento de una entidad, empresa, el mundo entero, debido a que lo que nos rodea son datos he información que pueden ser recopilados y manejados o controlados, por ende en este proyecto tiene como objetivo tener una experimentación practica de la importancia que tienen los datos en el día a día; Para esta práctica se hará uso de un archivo CSV que contiene la información respecto a 4000 películas, también se usara los conocimientos obtenidos en el tercer ciclo de la asignatura fundamentos de base de datos, la misma que es dada por el ingeniero Nelson Piedra.

**Teoría:**

**¿Que es un CSV?**

Son las siglas de "Comma Separated Value" y son las siglas de Comma Separated Value. En otras palabras, un archivo CSV es cualquier archivo de texto en el que los caracteres están separados por comas para formar una tabla organizada en filas y columnas. La columna está definida por cada punto y coma (;), y cada línea está definida por una línea adicional en el texto. De esta forma, los archivos CSV se pueden crear de forma muy sencilla (lo explicaremos más adelante). Es por eso que el archivo .csv está directamente relacionado con la creación del directorio.

**¿Qué es un modelo Conceptual?**

El modelo conceptual se utiliza para que un diseñador pueda transmitir a la empresa lo que ha entendido sobre la información que maneja y así poder crear una base de datos. El modelo conceptual se va construyendo utilizando la información que se encuentra en los requisitos del usuario, es independiente de los aspectos de implementación como por ejemplo los lenguajes de programación o programas de aplicación entre otros que se utilicen para manipularla. Existen varias alternativas una de ellas es recolectar los requisitos para tratar de examinar los diagramas de flujo de los datos u otra es realizar una entrevista a los usuarios que ayuden a examinar los procedimientos, informes, formularios y funcionamiento que tiene la empresa.

**¿Qué es un modelo Logico?**

El modelo lógico tiene como objetivo obtener tablas de bases de datos en su tercera forma normal, una vez obtenidas las tablas se podrá modificar el esquema para lograr una mejor eficiencia. Para cada tabla se debe detallar

**Desarrollo:**

**Modelo Conceptual**

El modelo conceptual se obtuvo a base del análisis de CSV proporcionado por la catedra Integradora, en el cual se saco las entidades, los atributos, sus cardinalidades y participación:

Diagrama

Descripción generada automáticamente

Después de obtener el modelado del CSV se procede a identificar que tipos de datos son de cada atributo de las distintas entidades, cuáles son sus calves primarias, foráneas, si son multivaluados entre otras características que tienen los campos, En la tabla 1 hasta la tabla 12 por cada entidad definida en el modelo Conceptual se muestra los atributos que se encuentran dentro de la entidad, su dominio, el tipo de dato, si son atributos multivaluados y un comentario que lo describen:

**Entidad MOVIES:**

*Table 1. Entidad Movies*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **index** | Char(25) |  | Not Null | False | Índice |
| **id\_movie** | Char(25) | Primary Key | Not Null | False | Id de la película |
| **budget** | Int(20) |  | Not Null | False | Presupuesto disponible |
| **homepage** | Varchar(300) |  | Not Null | False | Página Web de la película |
| **keywords** | Varchar(1000) |  | Not Null | True | Palabras Claves |
| **original\_title** | Varchar(89) |  | Not Null | False | Titulo Original de la película |
| **overview** | Varchar(1100) |  | Not Null | False | Descripción de la película |
| **popularity** | Double(5,9) |  | Not Null | False | Popularidad de la película |
| **release\_date** | Date |  | Not Null | False | Fecha de Estreno |
| **revenue** | Int(15) |  | Not Null | False | Ingresos generados |
| **runtime** | Int(7) |  | Not Null | False | Duración de la película |
| **status** | Varchar(34) |  | Not Null | False | Lema |
| **tagline** | Varchar(350) |  | Not Null | False | Frase final |
| **title** | Varchar(245) |  | Not Null | False | Titulo |
| **vote\_average** | Varchar(30) |  | Not Null | False | Promedio de votaciones |
| **vote\_count** | Int(5) |  | Not Null | False | Número de votos |

**Entidad DIRECTORS:**

*Table 2. Entidad DIRECTORS*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **id\_directors** | Char(25) | Primary Key | Not Null | False | Identificador de los directores |
| **name** | Varchar(350) |  | Not Null | False | Nombre del director |

**Entidad GENRES:**

*Table 3. Entidad GENRES*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **Id\_genres** | Char(25) | Primary Key | Not Null | False | Identificador del género de la película |
| **name** | Varchar(350) |  | Not Null | False | Nombre del género |

**Entidad CAST:**

*Table 4. Entidad CAST*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **id\_cast** | Char(25) | Primary Key | Not Null | False | Identificador del elenco |
| **name** | Varchar(350) |  | Not Null | False | Nombre del elenco |

**Entidad SPOKEN\_LANGUAGES:**

*Table 5. Entidad SPOLEN\_LANGUAGES*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **id\_spoken\_language** | Char(25) | Primary Key | Not Null | False | Identificador del habla del idioma |
| **name** | Varchar(350) |  | Not Null | False | Nombre del idioma hablado |

**Entidad PRODUCTION\_COUNTRIES:**

*Table 6. Entidad PRODUCTION\_COUNTRIES*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **id\_production\_countries** | Char(25) | Primary Key | Not Null | False | Identificador del país de la producción |
| **name** | Varchar(350) |  | Not Null | False | Nombre de la producción en el país |

**Entidad PRODUCTION\_COMPANY:**

*Table 7. Entidad PRODUCTION\_COMPANY*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **Id\_production\_companies** | Char(25) | Primary Key | Not Null | False | Identificador de la compañía que produce |
| **name** | Varchar(350) |  | Not Null | False | Nombre de la compañía |

**Entidad ORIGINAL\_LANGUAGES:**

*Table 8. Entidad ORIGINAL\_LANGUAGES*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ORIGINAL\_LANGUAGES** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **id\_original\_laguage** | Char(25) | Primary Key | Not Null | False | Identificador del idioma original |
| **name** | Varchar(350) |  | Not Null | False | Nombre del idioma |

**Entidad CREW:**

*Table 9. Entidad CREW*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi Valuated** | **Comment** |
| **id\_crew** | Char(25) | Primary Key | Not Null | False | Identificador del equipo |
| **id\_movie** | Char(25) | Foreign Key | Not Null | False | Identificador de la película |

**Entidad WORKERS:**

*Table 10. Entidad WORKERS*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi Valuated** | **Comment** |
| **id\_worker** | Char(25) | Primary Key | Not Null | False | Identificador de los trabajadores |
| **gender** | Varchar(350) |  | Not Null | False | Genero del trabajador |
| **id\_job** | Char(25) | Foreign Key | Not Null | False | Identificación de trabajo |
| **credit\_id** | Char(25) |  | Not Null | False | Identificador de crédito |
| **id\_crew** | Char(25) | Foreign Key | Not Null | False | Identificador del personal |

**Entidad JOB:**

*Table 11. Entidad JOB*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **id\_job** | Char(25) | Primary Key | Not Null | False | Identificador del trabajo |
| **name** | Varchar(50) |  | Not Null | False | Nombre del trabajo |
| **id\_departament** | Char(25) | Foreign Key | Not Null | False | Identificador del departamento |

**Entidad DEPARTMENTS:**

*Table 12. Entidad DEPARTMENTS*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Domain** | **Type** | **Optional** | **Multi**  **Valuated** | **Comment** |
| **Id\_departments** | Char(25) | Primary Key | Not Null | False | Identificador del departamento |
| **name** | Varchar(50) |  | Not Null | False | Nombre del departamento |

Una vez que tenemos los tipos de datos que pertenecen a cada atributo, los valores multivaluados, las primary keys , las relaciones se procede a aplicar la normalización y el modelo conceptual que se realiza después de aplicar las 3F(Formas Normales) presentadas a continuación:

En base de datos existes las 3F, las cuales son las 3 normas de normalización:

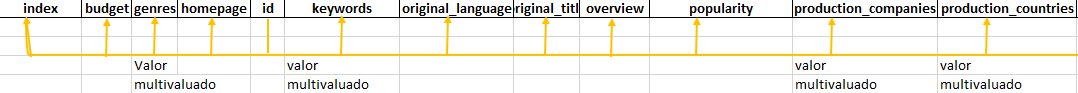
**Primera forma Normal (1F):**

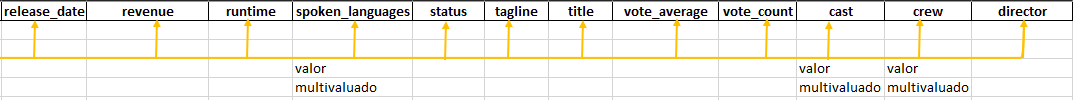
-- Cada celda de tabla debe contener un solo valor.

-- Cada registro debe ser único.

-- Los valores tienen que ser atómicos, no multivaluados, ni compuestos.

**Tabla del CSV:**





Como se puede observas del CSV otorgado por la catedra integradora, existen campos que son multivaluados, según la primera forma de normalización no debe haber campos multivaluados, por ende, se procede a extraer todos los campos multivaluados y se los convierten en tablas:

**Table movie:**

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

**Table genres:**

Tabla

Descripción generada automáticamente

**Table keywords: Table production\_companies:**

Imagen que contiene biombo

Descripción generada automáticamente

**Table production\_countries: Table Spoken\_languages:**

Gráfico

Descripción generada automáticamente

**Table cast: Table crew:**

Tabla

Descripción generada automáticamente

**Segunda forma Normal (2F):**

-- Está en 1FN

-- Todos los atributos que no son clave primaria tienen dependencia funcional completa con respecto a todas las claves existentes en el esquema.

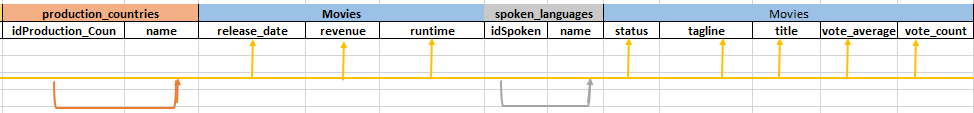
-- Para recuperar un atributo no clave, se necesita acceder por la clave completa, no por una subclave

-- Las 2FN aplican a las relaciones con claves primarias compuestas por dos o más atributos

**Tablas en segunda forma normal:**

Tabla

Descripción generada automáticamente



Gráfico

Descripción generada automáticamente con confianza baja

**Tablas género, kewords, cast :**

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

**Dependencias funcionales completa de movie\_production\_companies:**

Tabla

Descripción generada automáticamente

**Dependencias funcionales completa de movie\_production\_countries:**

Aplicación, Tabla, Excel

Descripción generada automáticamente

**Dependencias funcionales completa de movie\_spoken\_languages:**

Tabla

Descripción generada automáticamente

**Dependencias funcionales completa de movie\_crew:**

Tabla

Descripción generada automáticamente

**Tercera forma Normal (3F):**

-- Está en 2FN

-- Todos los atributos que no son clave primaria no dependen transitivamente de ésta

En este caso la mayoría de las tablas que fueron dadas a conocer en la segunda forma normal, cumplen también la tercera forma normal, sin embargo, la tabla Crew es la única tabla que presenta transitividad, por ende se procede a separa esta transitividad en 2 tablas diferentes:

**Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente**

**Aplicando la normalización las tablas quedarían de la siguiente forma:**

**La tabla movie:**

**Tabla

Descripción generada automáticamente**

**Gráfico

Descripción generada automáticamente**

**Las tablas genres, keywords y cast:**

**Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente**

**La tabla production\_companies con su dependencia funcional completa:**

**Tabla, Excel

Descripción generada automáticamente**

**La tabla production\_countries con su dependencia funcional completa:**

**Tabla

Descripción generada automáticamente**

**La tabla spoken\_Languages con su dependencia funcional completa:**

Tabla

Descripción generada automáticamente

**La dependencia funcional completa de la tabla movie\_crew:**

Tabla

Descripción generada automáticamente

**La tabla crew sin transitividad y separada en 2 tablas:**

Imagen que contiene biombo, reloj, dibujo

Descripción generada automáticamente

**Modelo lógico:**

Una vez que se puso en práctica las 3 formas de normalización, se prosigue con la creación del modelo lógico, el cual es la unión del modelo conceptual y la aplicación de las 3 formas normales:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

El modelado presentado sigue las formas de normalización, una vez que ya tenemos el modelado lógico podemos continuar con el DDL en mi SQL, el cual se ha considerado conveniente dividirlo por 3 partes:

* la primera parte será los Create table
* La segunda parte se dará a conocer los diferentes procedimientos usados para la extracción de datos de lo Json:
* La tercera parte serán la inserción de datos:

**My SQL:**

Para la elaboración de esta practica se usó la herramienta del data Grips, el cual nos permite generar los inserta del movie\_dataset directamente desde este entorno:

**Creación de la tabla movie\_dataset\_cleaned:**

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

**Creación de la tabla movie:**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Creación de la tabla production\_companies:**

**Texto

Descripción generada automáticamente**

**Creación de la tabla movies\_companies:**

**Texto

Descripción generada automáticamente**

**Creación de la tabla production\_countries:**

**Texto

Descripción generada automáticamente**

**Creación de la tabla movie\_countries:**

**Texto

Descripción generada automáticamente**

***Creación de la tabla spoken\_languages:***

**Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media**

**Creación de la tabla movies\_languages:**

**Texto

Descripción generada automáticamente**

**Creación de la tabla género:  
  
Texto

Descripción generada automáticamente con confianza media**

**Creación de la tabla genero\_movie:**

Texto

Descripción generada automáticamente

**Creación de la tabla crew:**

Texto

Descripción generada automáticamente

**Creación de la tabla credits:**

Texto

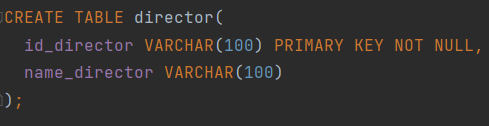
Descripción generada automáticamente

**Creación de la tabla movie\_credits:**

Texto

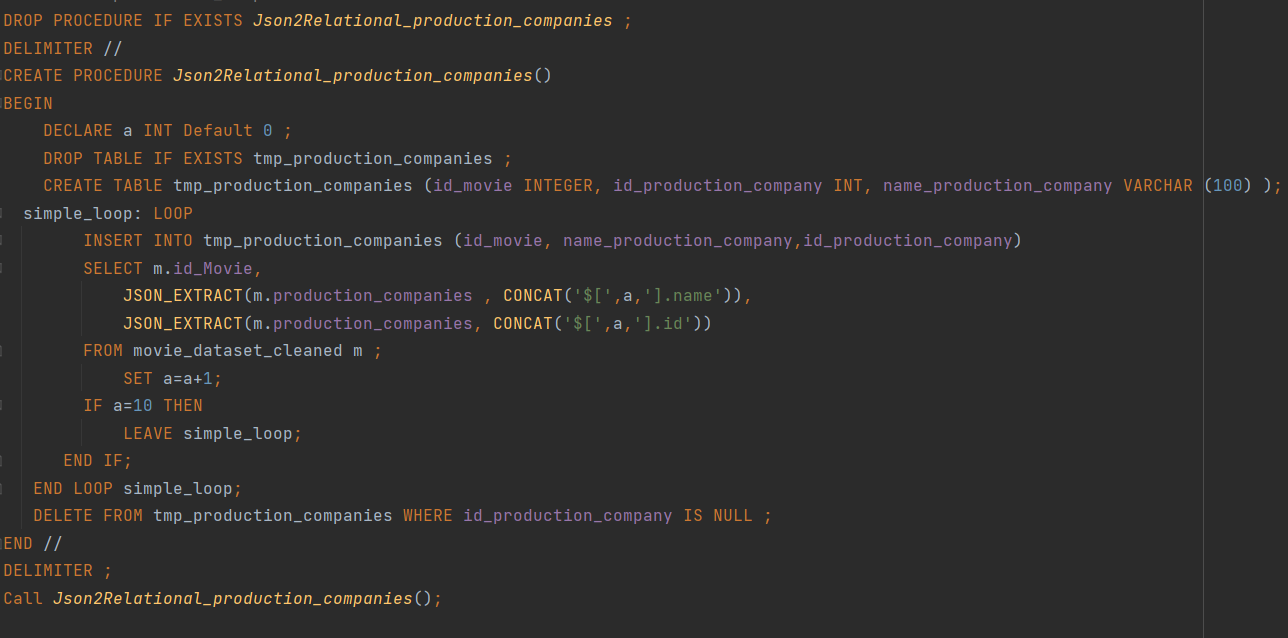
Descripción generada automáticamente

**Creación de la tabla director:**

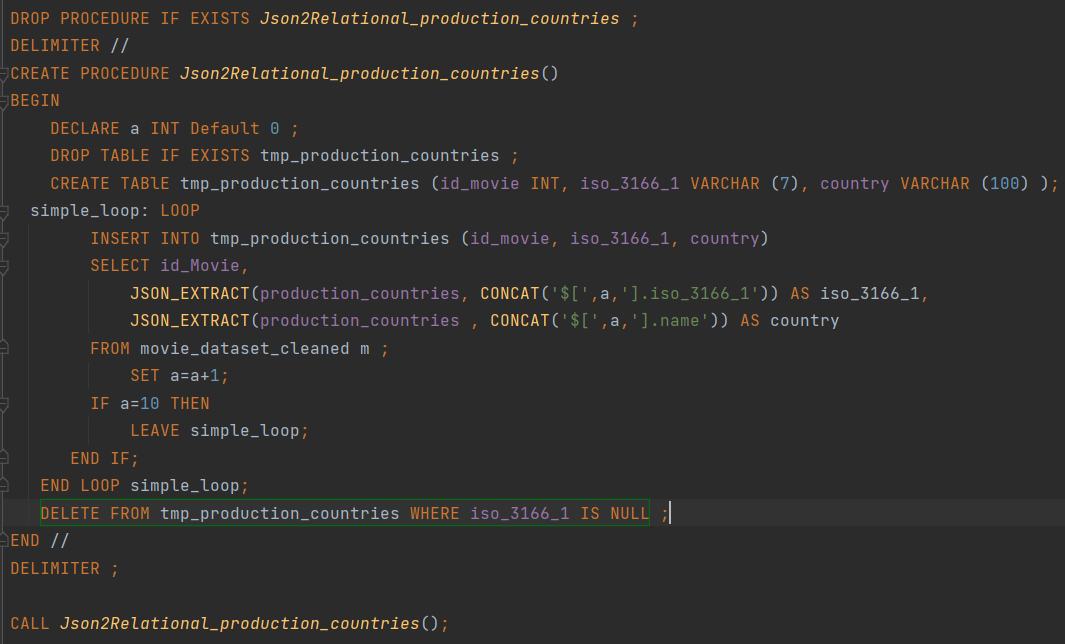


**2. Procedimientos que permiten extraer los JSON:**

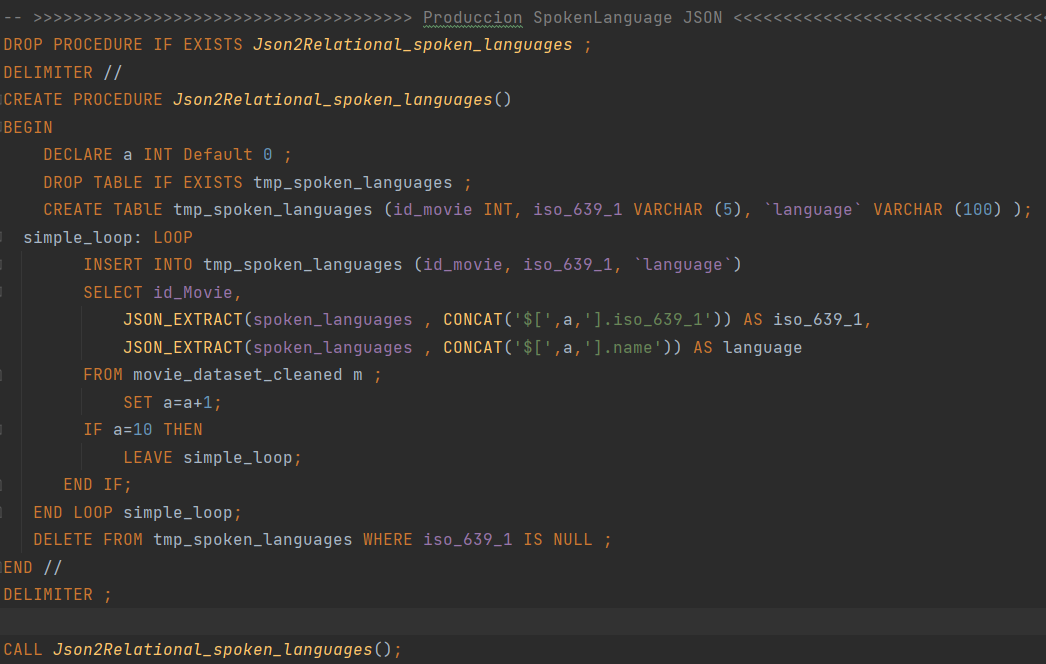
**Procedimiento Json2Relational\_production\_companies:**



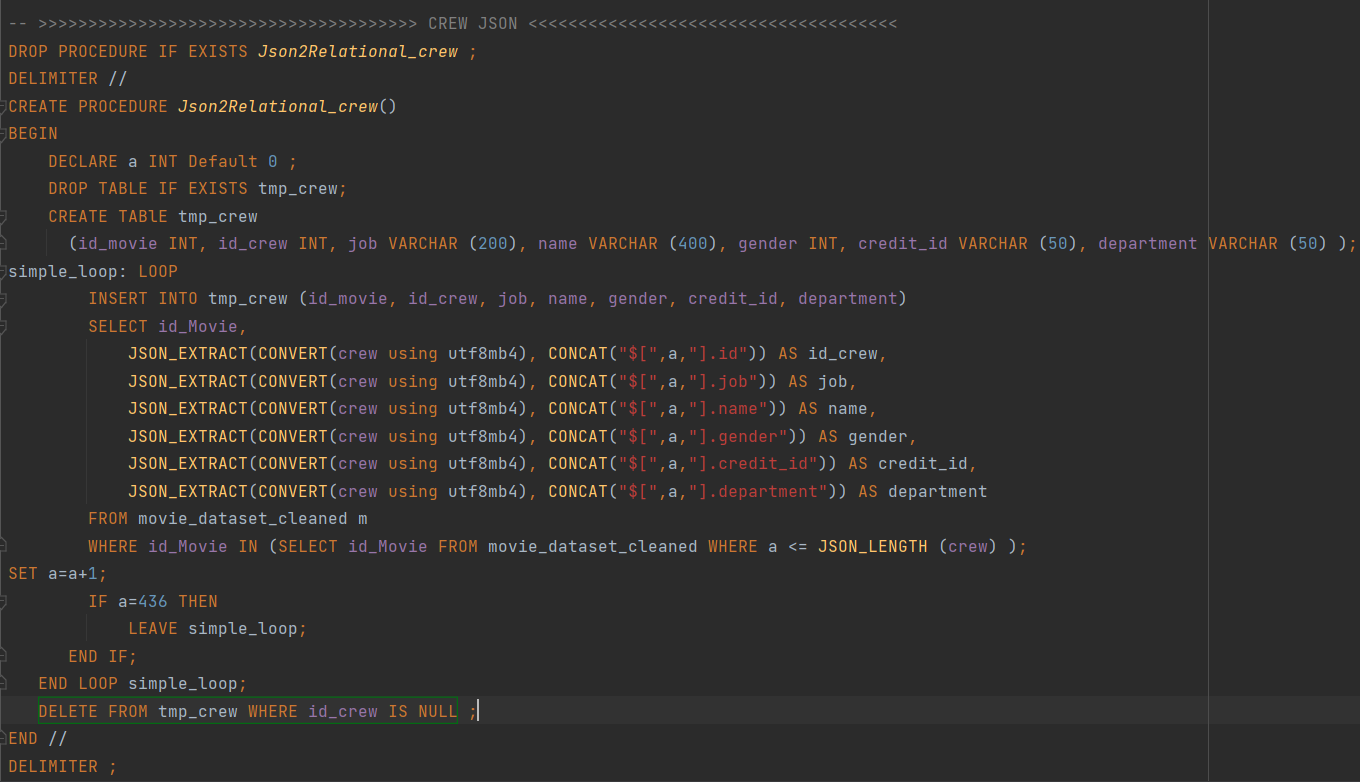
**Procedimiento Json2Relational\_production\_countries:**

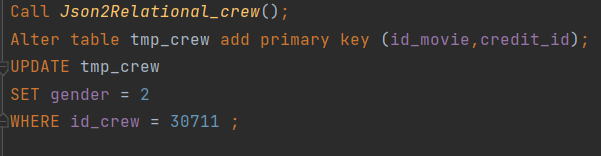


**Procedimiento Json2Relational\_spoken\_languages:**

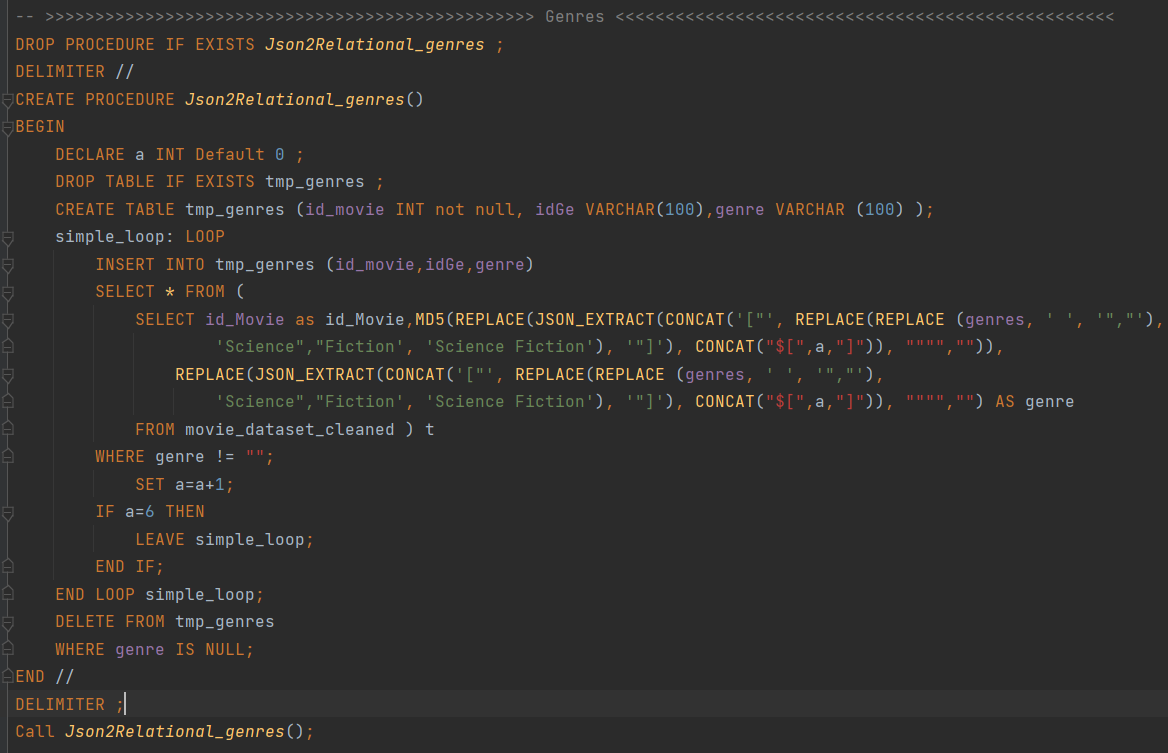


**Procedimiento Json2Relational\_crew:**



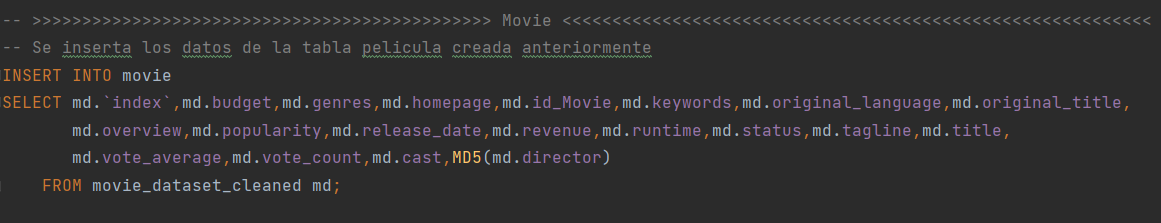


**Procedimiento Json2Relational\_genres:**

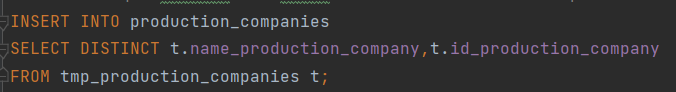


**3. Inserts de los datos en las respectivas tablas:**

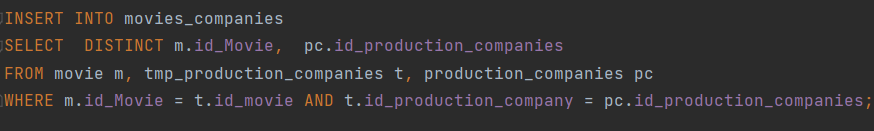
**Inserts de la tabla movie:**

****

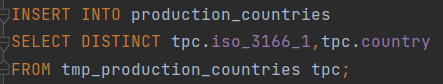
**Inserts de la tabla production\_companies:**

****

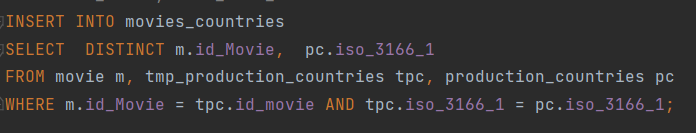
**Inserts de la tabla movies\_companies:**



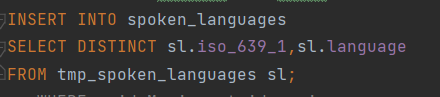
**Inserts de la tabla production\_countries:**



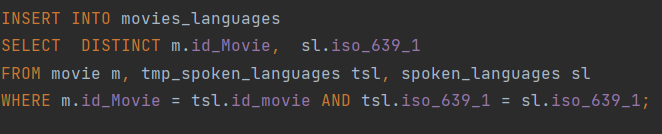
**Inserts de la tabla movies\_countries:**

****

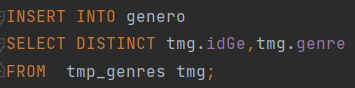
**Inserts de la tabla Spoken\_languages:**

****

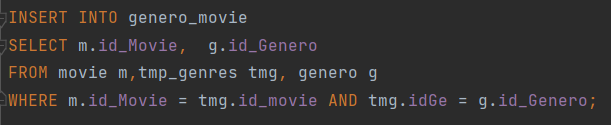
**Inserts de la tabla movies\_languages:**

****

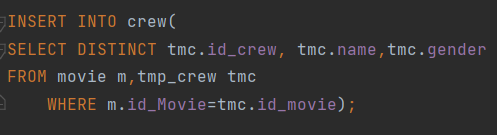
**Inserts de la tabla género:**



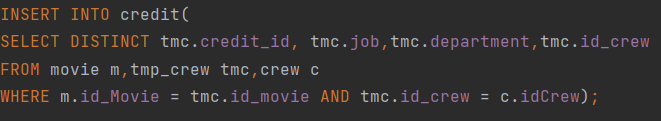
**Inserts de la tabla género\_movie:**



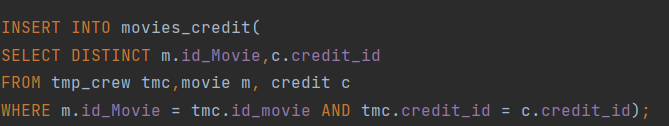
**Inserts de la tabla crew:**



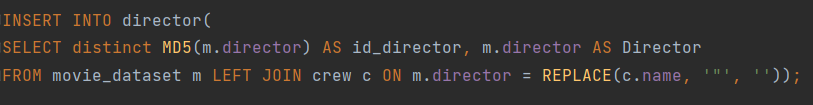
**Inserts de la tabla credit:**



**Inserts de la tabla movies\_credit:**

****

**Inserts de la tabla director**



Para concluir quisiera recalcar que el uso de las 3 formas normales fueron imprescindibles en este esquema, debido a que el CSV estaba con algunos JSON y valores multivaluados, generando un gran conflicto al trabajar con estos campos, sin embargo gracias a la implementación de los procedimientos y el uso del JSON\_EXTRACT facilito en gran medida la extracción de la data, además que esta práctica nos permitió lidiar con un problema que es algo bastante usual en el campo laboral, como lo es el manejo de los JSON, debido a que un gran porcentaje de la data he información no esta del todo proporcionada, como eran algunos de estos campos