**Fundamentos de Base de Datos**



**UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA**

***La Universidad Católica de Loja***

**INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN**

*Proyecto Final - Ciclo de vida de bases de datos relacionales normalizada*

**Por:**

**Edgar Martínez (edmartinez8@utpl.edu.ec)**

**Docente:**

**Ing. Nelson Piedra (**[**http://investigacion.utpl.edu.ec/nopiedra**](http://investigacion.utpl.edu.ec/nopiedra)**)**

**Link del proyecto en Github:**

**https://github.com/Edmartinez28/Proyecto-Final-Base-de-Datos.git**

**octubre 2021 – febrero 2022**

*Ciclo de vida de bases de datos relacionales normalizada*

*Antecedentes*

Con el desarrollo de las materias de tercer ciclo de la carrera de computación de la UTPL se toma las carreras de Fundamentos de Base de datos y de Practicum 1.1 las cuales establecen que se debe desarrollar un proyecto integrador para asegurar el conocimiento impartido por los profesores por ello se dio la tarea de extraer datos de un documento CSV e introducirlos en una base de datos con ello debíamos desarrollar el modelo de base de datos relacional por lo cual para hacer más fácil la operación se da el proceso de Normalización de los campos

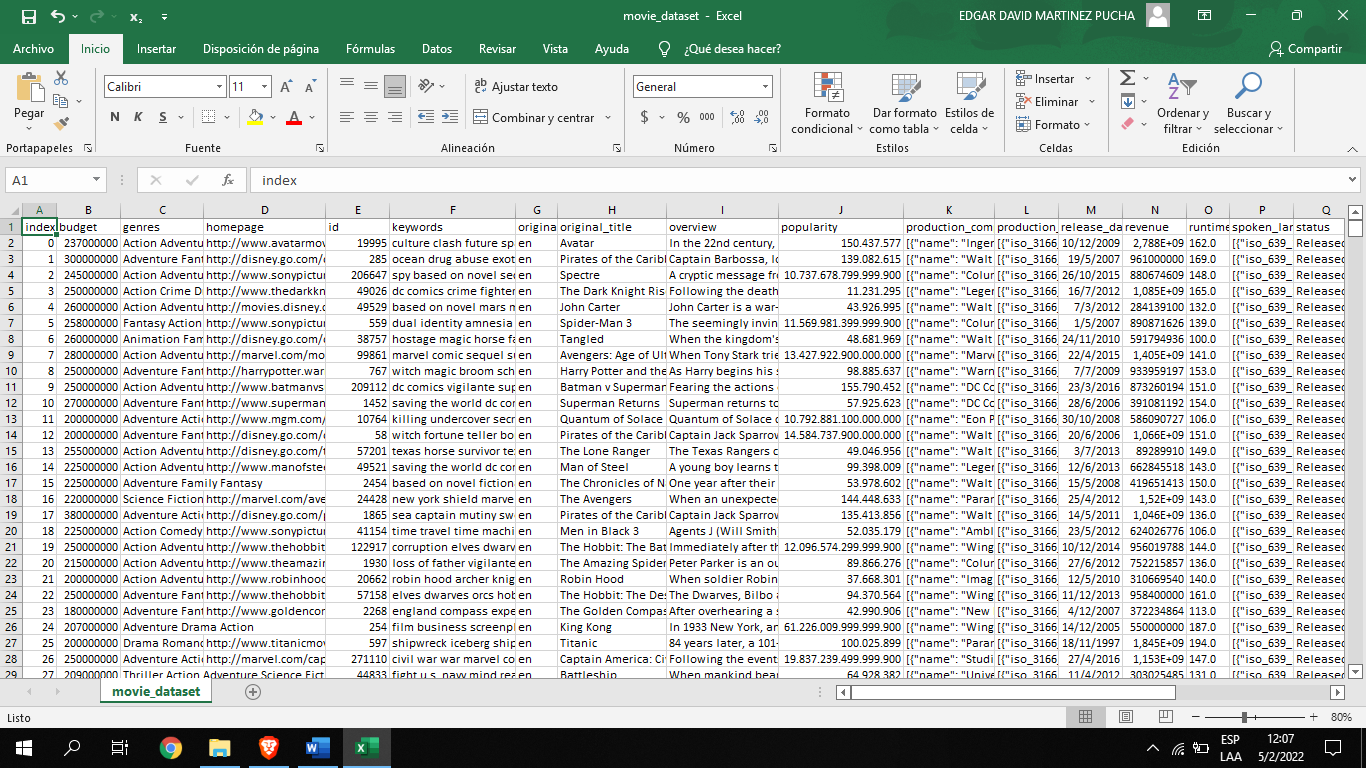
*Introducción*

Los datos son un activo muy valioso en los últimos años por lo cual la carrera se enfoca en estos como uno de los principales puntos que deben desarrollar los estudiantes para sus habilidades profesionales, con ello el proyecto tiene como objetivo aprender a manipular los datos, ingresarlos, consultarlos entre otras acciones que se puede hacer con estos mismos.

*Desarrollo*

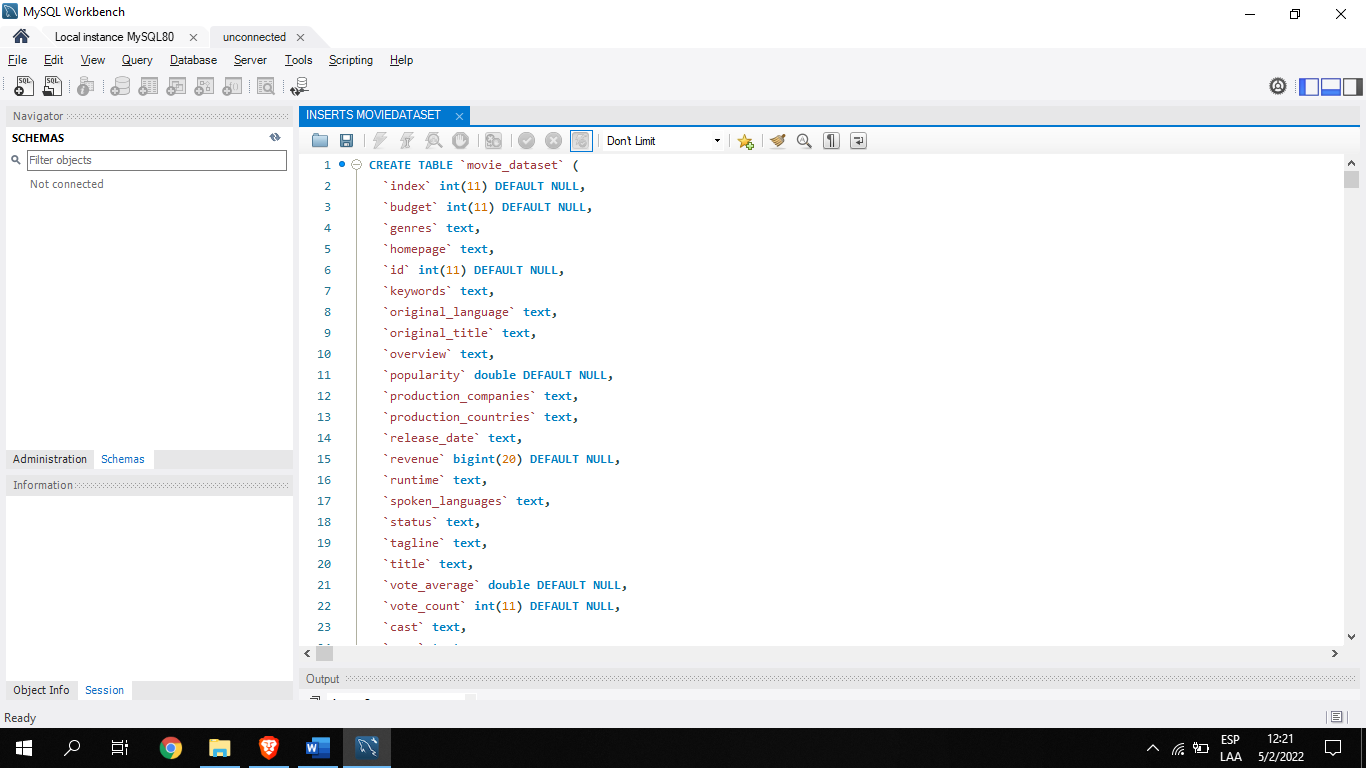
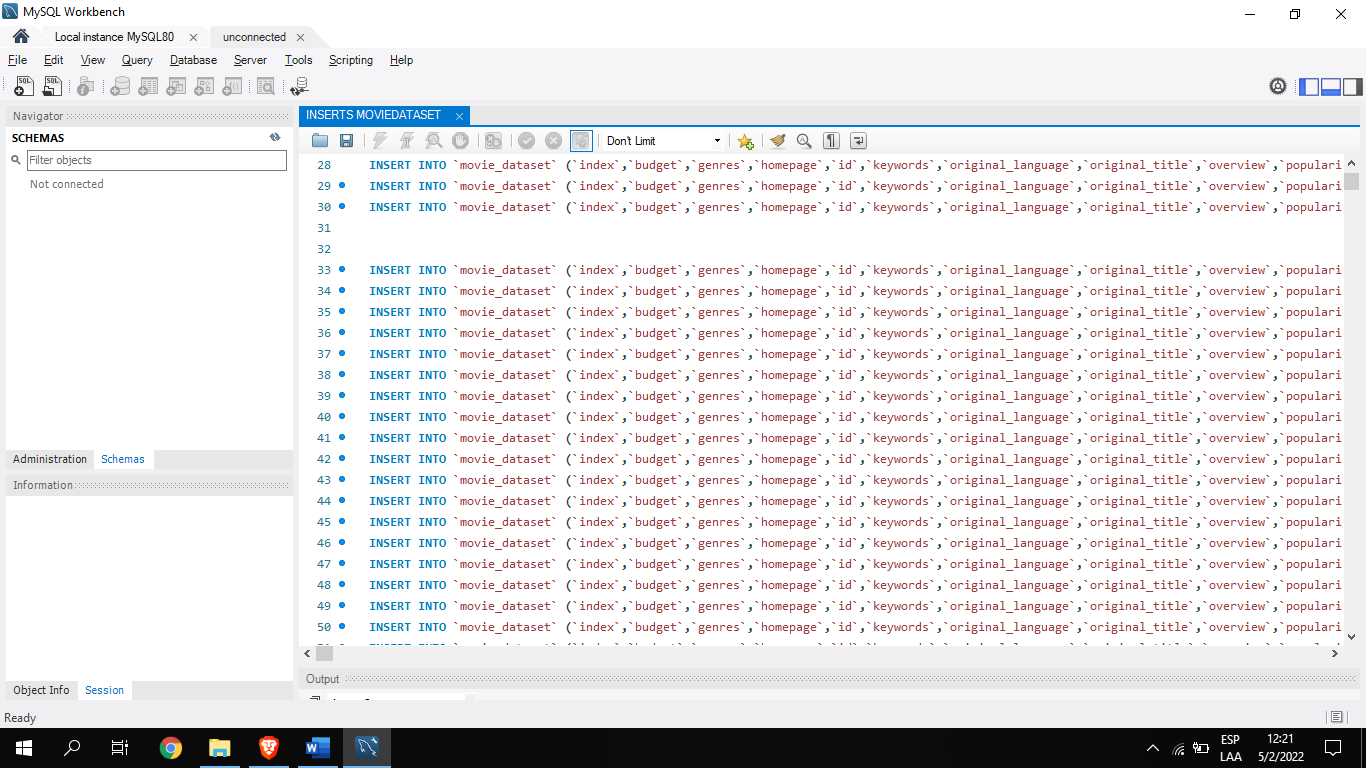
*Descripción de la Data*

El análisis de los datos antes de comenzar un proceso de normalización es fundamental para cualquiera dedicado al procesamiento de los datos, por ello se debe comenzar analizando los datos con los que se cuenta, en nuestro caso es del archivo CSV Movies Data Set, este es un archivo que cuenta con la información de un aproximado de 4800 películas con información fundamental de cada una de ellas



El documento cuenta con 24 campos de los cuales se incluyen campos de tipo multivaluados que son los casos de genres, keywords y cast y también cuenta con campos que además de ser multivaluados son compuestos, estos serían los campos de production\_countries, production\_companies, spoken\_languages y crew. Estos son los campos que se debe analizar más y tratar de colocarlos de una manera lo más optima posible para su manejo.

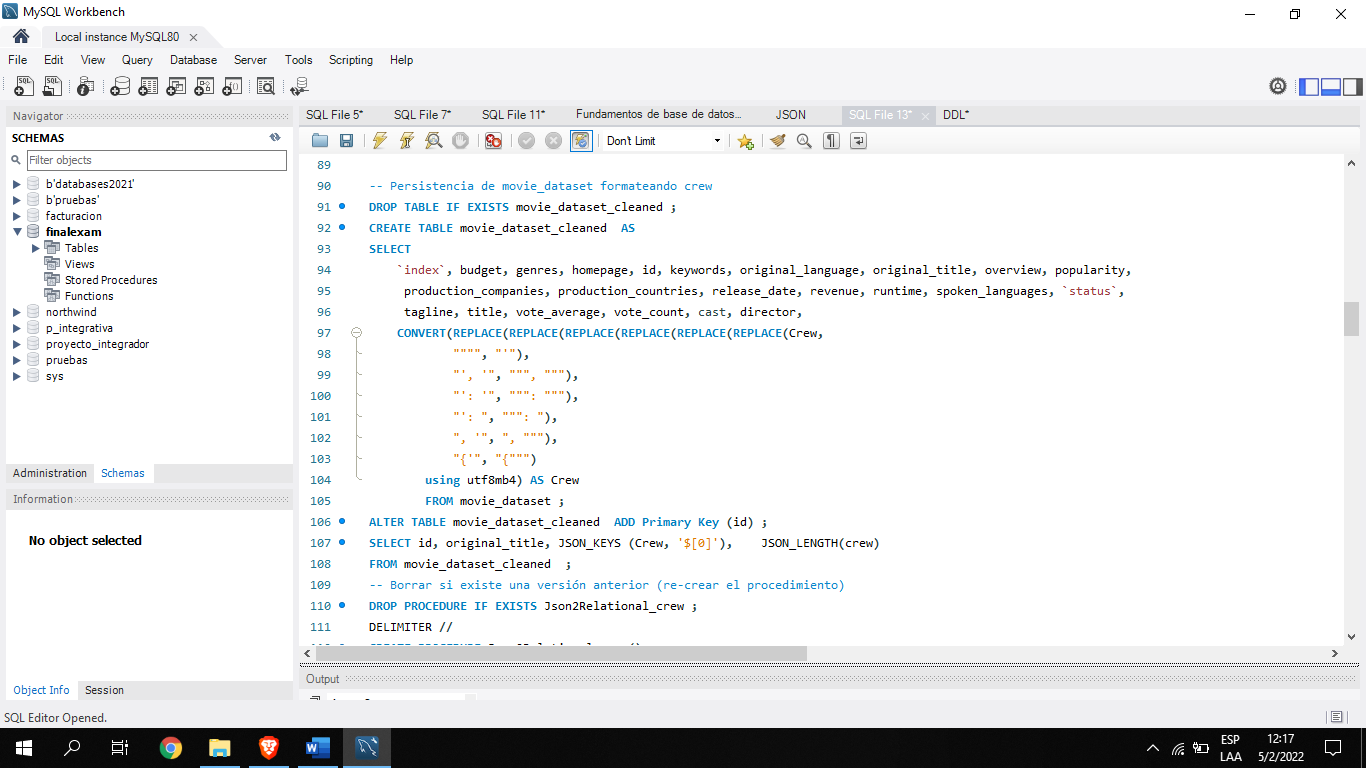
Primero lo que se debe es de insertar la data en MySql esto para poder trabajarla desde herramientas como el MySQL Workbrench y en este programa manejarla, implementarla, modificarla, etc. Todos estos procesos se hacen a través de código que se pude crear para la inserción de datos en MySQL Workbrench.

*Limpieza de Data*

El comenzar haciendo limpieza de los datos es fundamental, esto quiere decir que se debe hacer un remplazo de caracteres que podrían afectar al procesamiento de de los datos de manera autónoma por lo que se debe de limpiarla, en el caso de Movies Data Set los campos o la data que se debe de hacer limpieza son los campos compuestos multivaluados, esto debido a que estos campos se encuentran descritos en un formato de almacenamiento de información conocido como tipo JSON, para el procesamiento de estos datos existen caracteres y estructuras indispensables para la separación de los archivos, por ello se aplica un código de tipo sql para poder optimizar los datos.

En el caso de estos campos los caracteres más importantes son (“ ”), ({ }) y ([ ]) por ello se aplican funciones REPLACE para evitar que el contenido de la información tipo JSON pueda afectar a la lectura de esta misma.



Primero se debe analizar el archivo CSV que se nos proveyó para hacer la extracción de datos y analizar sus campos para poder aplicar las formas normales. Para ello primero hacemos la búsqueda de dependencias funcionales

*Visión del Proyecto*

Con este proyecto que se tienen planteado el objetivo principal es el de poner y a la vez adquirir nuevas técnicas y habilidades en cuanto a la materia de Base de Datos, con ello para el proyecto se pone como objetivo la lectura de datos de un archivo CSV el cual es un formato estandarizado para la importación y exportación de Data a nivel mundial, por ello con este proyecto se busca desarrollar una integración que nos permita hacer la lectura del archivo CSV Movies Data Set y poder ingresarlos en una base de datos en este caso MySQL pero de una manera organizada basándonos en un modelo relacional el cual será creado con el análisis de los datos y con ello tener acceso a una rápida manipulación y acceso a la Data.

*Proceso de Normalización*

*Dependencias Funcionales*

*Tabla Directory*

Index 🡪 Id

*Tabla Movie*

Id 🡪 Budget, Homepage, Homepage, Original\_Language, Original\_Title, Overview, Popularity, Status, Cast, Director, Gender, SpokenLanguages, ProductionCountries, ProductionCompanies, Crew

*Tabla OtherTitle*

Id, Original\_Title 🡪 Title

*Tabla Identifiers*

Id, Overview 🡪 Keywords, Tagline

*Tabla Posproduction*

Id, Status 🡪 Relase\_Date, Revenue, Runtime, VoteAverange, VoteCount

*Tabla SpokenLanguages*

ISO639 🡪 Name

*Tabla ProductionCountries*

ISO3166 🡪 Name

*Tabla ProductionCompanies*

IdCompany 🡪 Name

*Tabla Crew*

IdEmployee 🡪 Name, Gender, Department, Job, CreditId

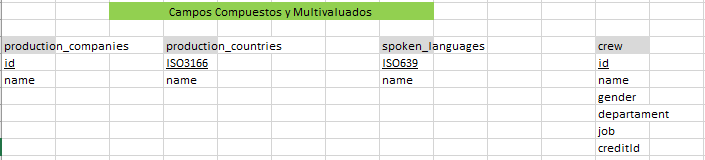
Para llegar a las tablas finales es necesario crear el proceso de normalización paso a paso por las 3 formas normales por ello:

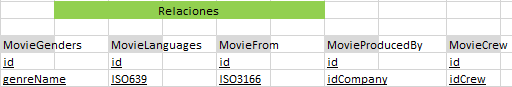
*Primera forma Normal*

Se separan todos los campos que no son atómicos, en este caso se determinó los campos atómicos, campos multivaluados y campos que son compuestos y multivaluados a la vez para los campos que son multivaluados o compuestos se los separa en tablas distintas y se crea tablas que las relacionan con la tabla fuerte movie.





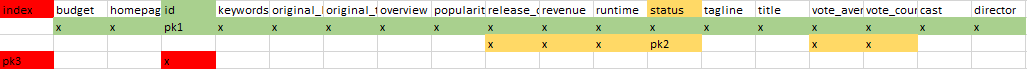


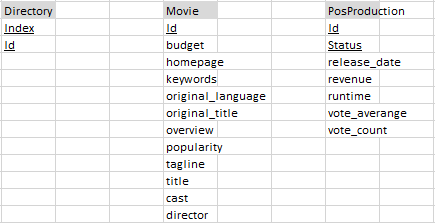


En el caso de los campos simples multivaluados se analizó la complejidad y cuan optimo era de separarlos en tablas distintas o en relaciones y se llegó al punto en que solo el campo de genres era óptimo para poder separarlo en una tabla distinta esto debido a su sencillez y ya que es posible el poder separar los valores de este campo en tokens, caso contrario de keywords o cast en donde los tokens vendrían a ser compuestos de 2 o más y ya que no existen caracteres separadores de estos no se puede sacar la información deseada.

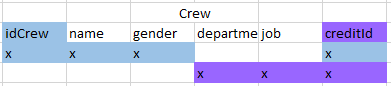
*Segunda Forma Normal*

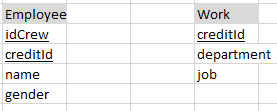
Para la segunda forma normal se tomó como tabla fuerte a la tabla movie y a partir de esta siendo id su clave candidata se sacó analizando 1 a 1 los atributos cuales deberían pertenecer a esta tabla y se determinó una segunda tabla que no solo depende de id sino también de status. Esto ya que ciertos campos dependían parcialmente de id y parcialmente de status por lo cual fue necesario dividir los componentes en tablas separadas. Esto se llegó a concluir después de determinar las dependencias funcionales de los campos.

**

**

También para la segunda forma normal se analizó la tabla crew con la cual se mantenía en la primera forma normal, sin embargo, de mismo modo se analizó las dependencias funcionales y se llegó al siguiente resultado



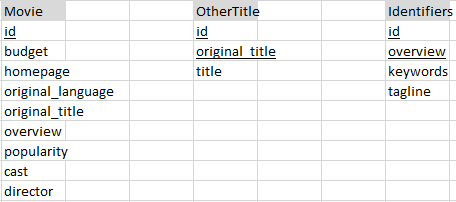


*Tercera Forma Normal*

Con la tercera forma normal determinamos si hay campos transitivos y se determinaron dos tablas adicionales con campos que dependen de otro que sería el caso de keywords, tagline y title, ya que estos son determinados de otros campos.

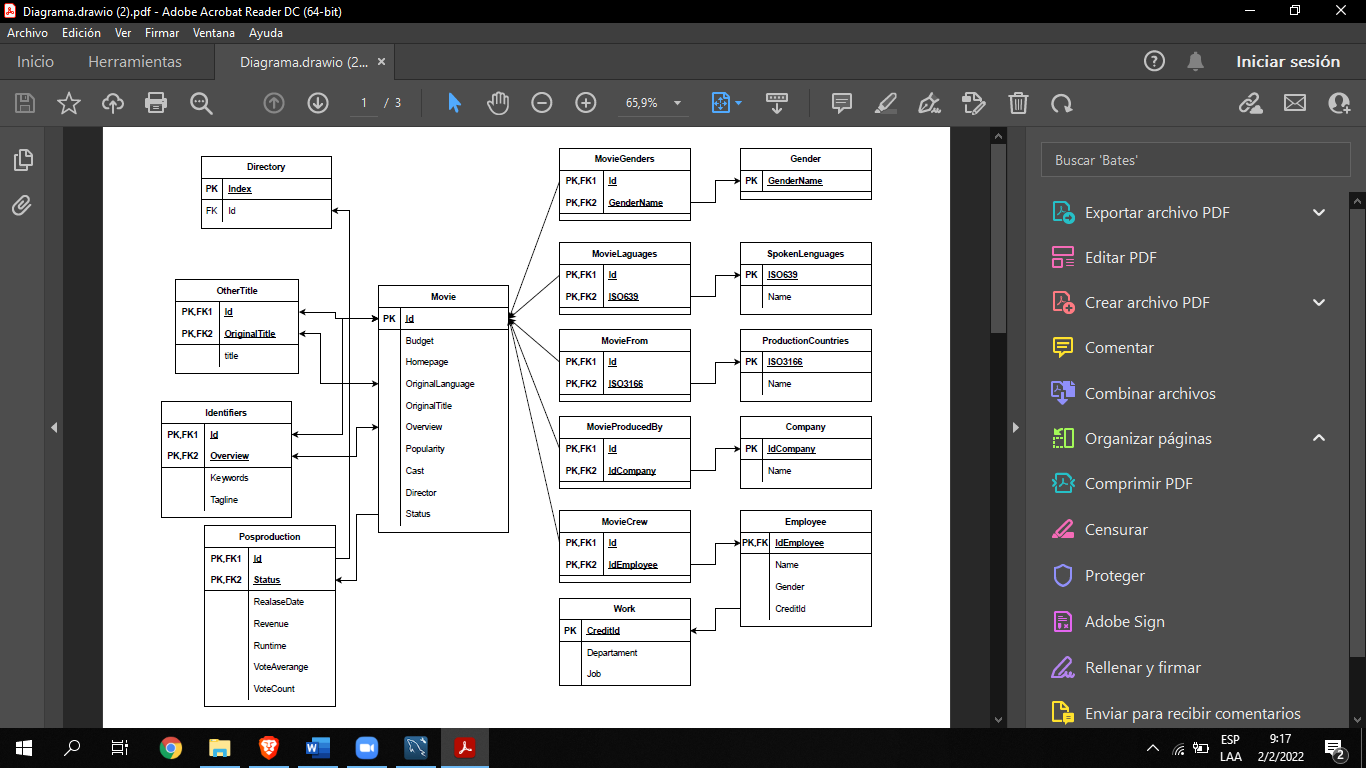
En el caso de title este va determinado por las claves primarias de id y original\_title con esto se da una relación 1 a 1 por lo cual simplemente se creó una nueva tabla sin que exista alguna que de relación a la tabla movie y a la nueva tabla, del mismo modo ocurre con los campos keywords y tagline ya que estos van determinados por la combinación de claves primarias de id y overview.

**

**

*Diagrama de entidades*

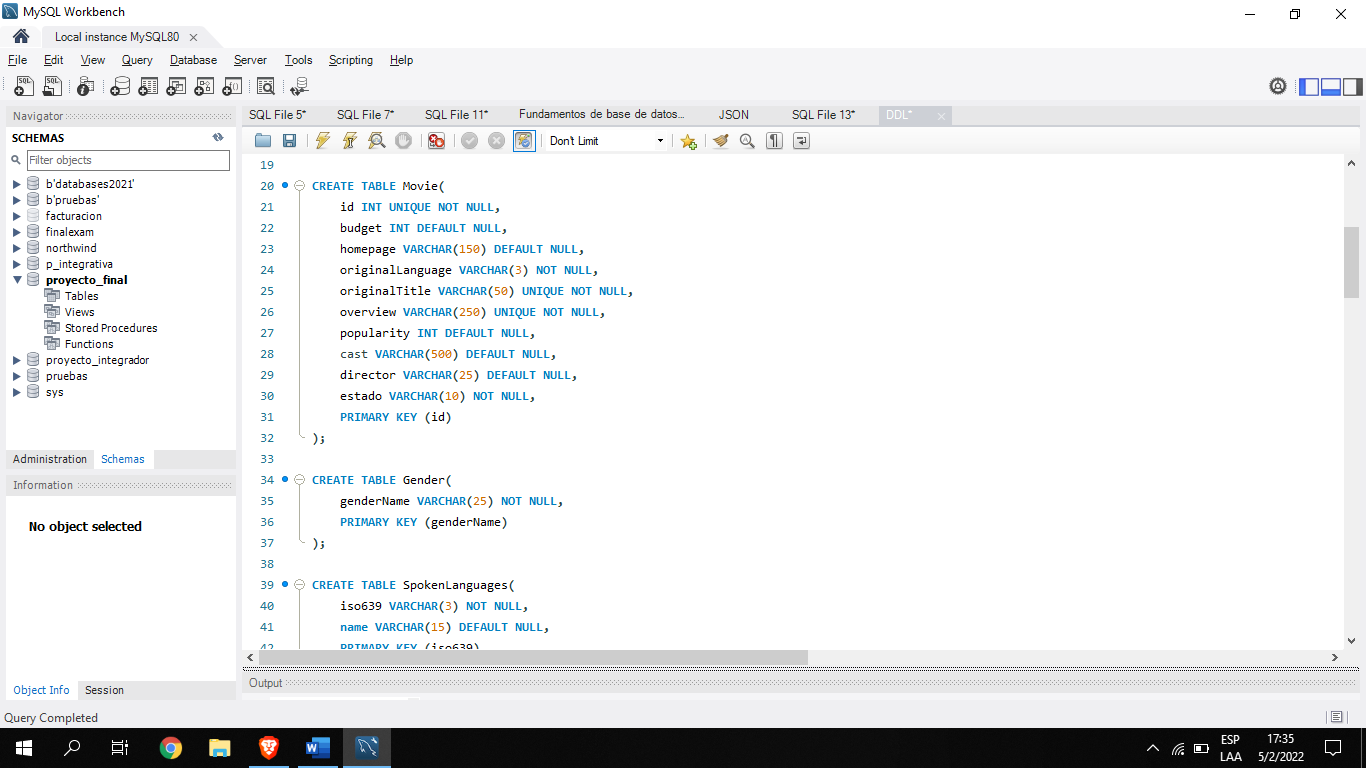
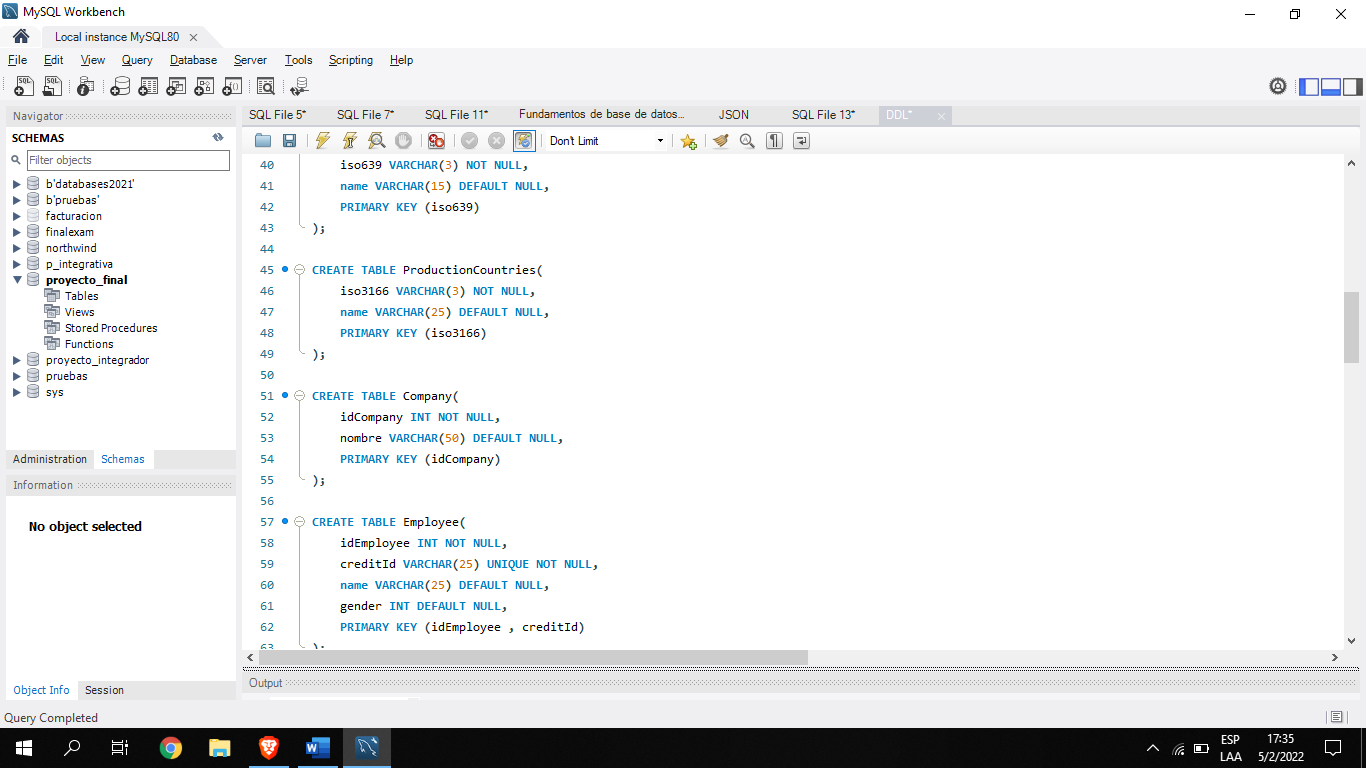
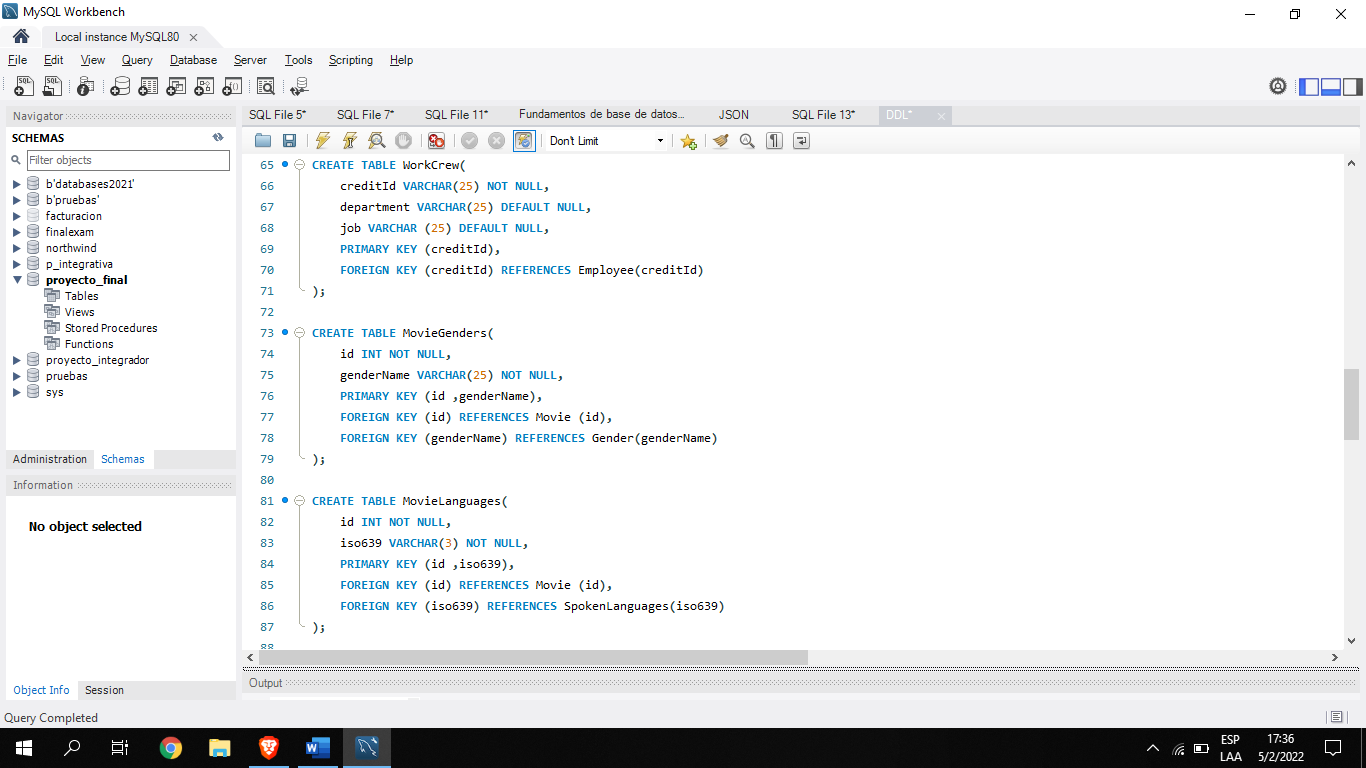
Con las tres formas normales realizadas se llega a establecer las relaciones y con esto determinar el diagrama entidad relación con el cual se partirá para el ordenamiento, almacenamiento y manejo de los datos

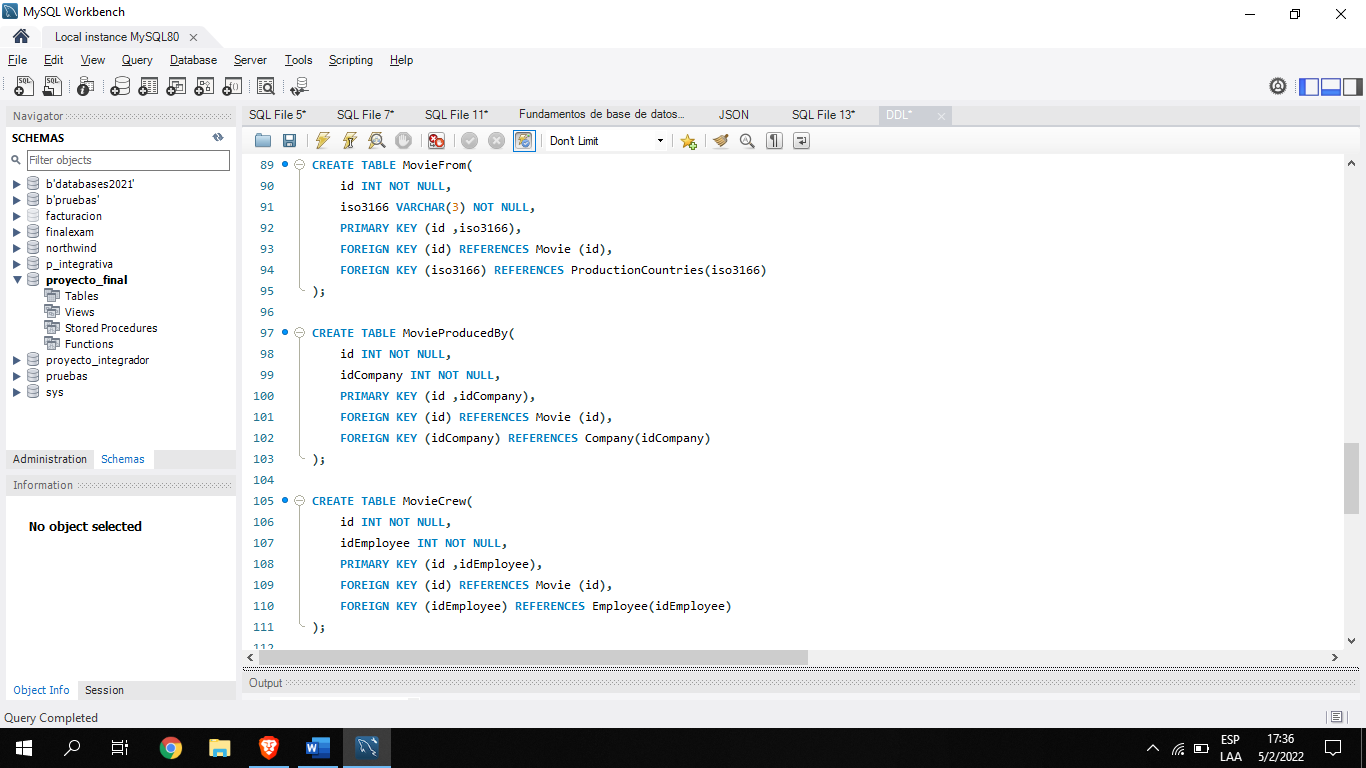
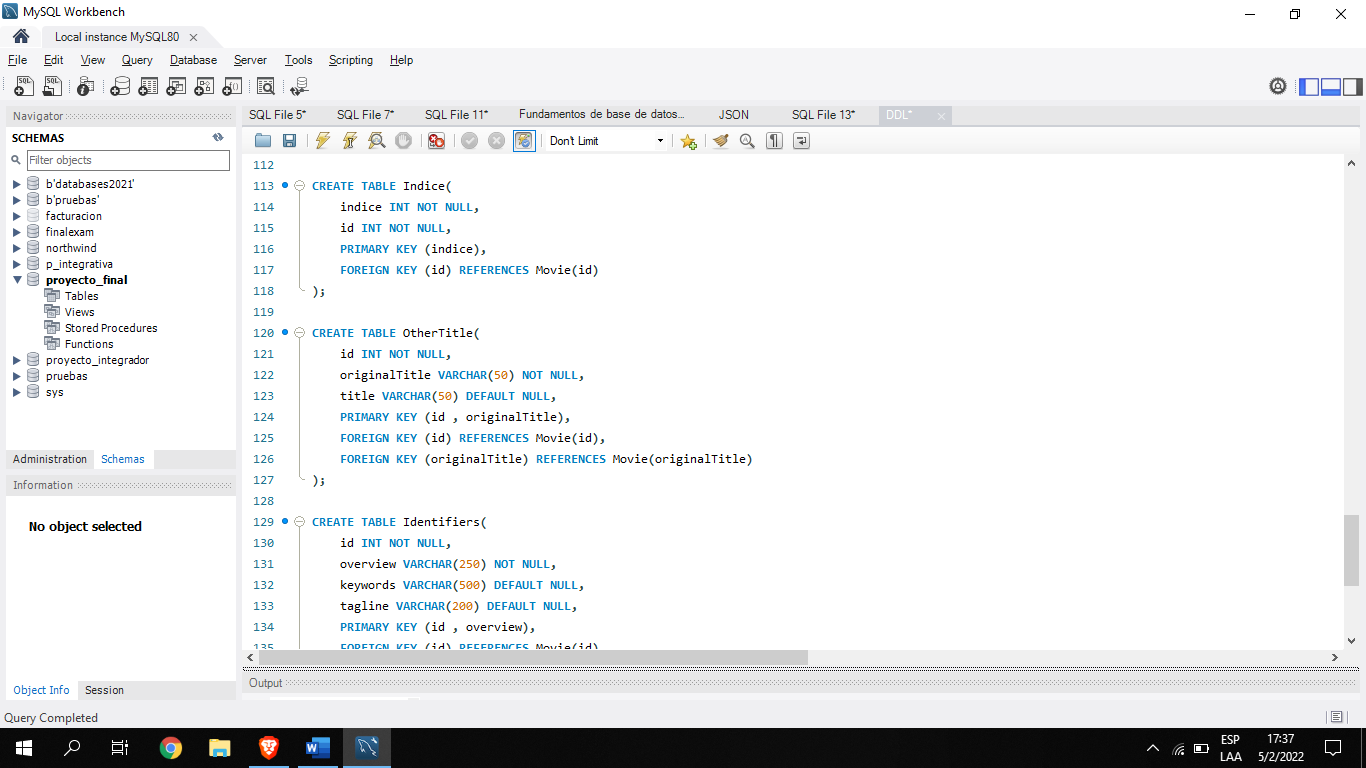
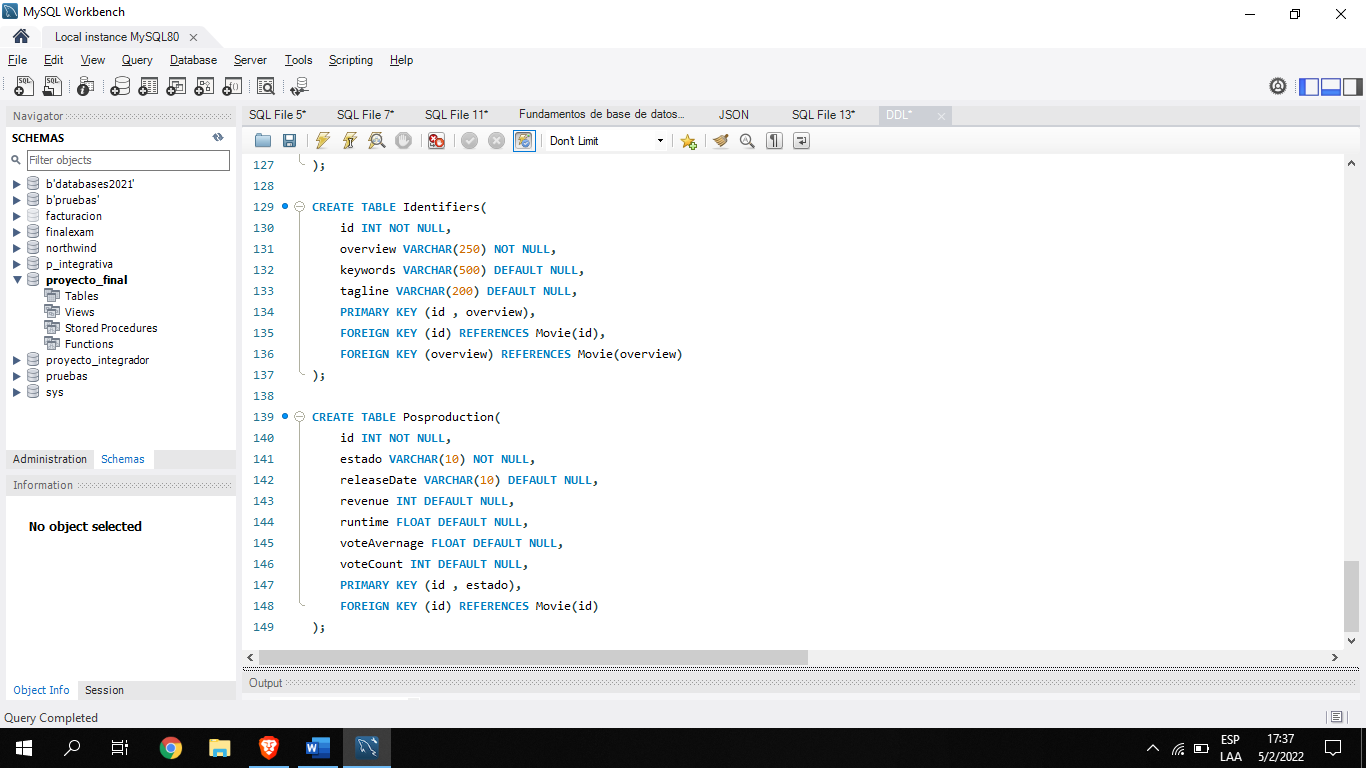


*Creación del DDL*

*DDL*

Con el modelo entidad relacion desarrollado comenzamos a escribir el códigopara crear el DDL (Data Definition Language) del modelo de Base de Datos o de manera más sencilla escribir el código de las tablas del modelo de base de datos al cual llegamos gracias al proceso de normalización.

Una vez creadas las tablas del modelo de base de datos lo siguiente es poblar las tablas pero antes de ello se da una descripción detallada de cada uno de los campos del modelo relacional y de las tablas ya creadadas.

*Detalles de los campos*

Definición de la entidad Indice

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Index | Int | Not Null | No | Primary Key | Numero en Catalogo |
| Id | Int | Not Null | No | Foreign Key | Identificador de la película |

Definición de la entidad Movie

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key | Identificador de la película |
| Budget | Int | Not Null | No |  | Presupuesto |
| Homepage | VarChar(100) |  | No |  | Página Web |
| OriginalLanguage | VarChar(3) | Not Null | No |  | Lenguaje Inicial |
| OriginalTitle | VarChar(50) | Not Null | No |  | Titulo Inicial |
| Overview | VarChar(250) | Not Null | No |  | Descripción |
| Status | VarChar(10) | Not Null | No |  | Estado de producción |
| Popularity | Int |  | No |  | Popularidad |
| Director | VarChar(25) |  | No |  | Director |
| Cast | VarChar(200) |  | No |  | Actores |

Definición de la entidad OtherTitle

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key Foreign Key | Identificador de la película |
| OriginalTitle | VarChar(50) | Not Null | No | Primary Key Foreign Key | Titulo Inicial |
| Title | VarChar(50) | Not Null | No |  | Titulo según región |

Definición de la entidad PosProduction

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key Foreign Key | Identificador de la película |
| Status | VarChar(10) | Not Null | No | Primary Key Foreign Key | Estado de producción |
| RealaseDate | VarChar(10) | Not Null | No |  | Fecha de Estreno |
| Revenue | Int | Not Null | No |  | Ingresos |
| Runtime | Float |  | No |  | Tiempo de duración |
| VoteAverange | Float |  | No |  | Calificación por personas |
| VoteCount | Int |  | No |  | # personas que calificaron |

Definición de la entidad Identifiers

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key Foreign Key | Identificador de la película |
| Overview | VarChar(250) | Not Null | No | Primary Key Foreign Key | Descripción |
| Keywords | VarChar(500) |  | No |  | Palabras de Reconocimiento |
| Tagline | VarChar(200) | Not Null | No |  | Línea descriptiva |

Definición de la entidad Gender

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| GenderName | VarChar(15) | Not Null | No | Primary Key | Nombre del genero |

Definición de la entidad SpokenLenguages

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| ISO639 | VarChar(3) | Not Null | No | Primary Key | Siglas del lenguaje |
| Name | VarChar(15) | Not Null | No |  | Nombre del lenguaje |

Definición de la entidad ProductionCountries

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| ISO3166 | VarChar(3) | Not Null | No | Primary Key | Abreviatura del país |
| Name | VarChar(25) | Not Null | No |  | Nombre del País |

Definición de la entidad Company

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| IdCompany | Int | Not Null | No | Primary Key | Identificador de compañia |
| Name | VarChar(50) | Not Null | No |  | Nombre compañia |

Definición de la entidad Employee

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| IdEmployee | Int | Not Null | No | Primary Key | Identificador de empleado |
| Name | VarChar(25) | Not Null | No |  | Nombre Emp |
| Gender | Int |  | No |  | Genro Emple |
| CreditId | VarChar(25) |  | No |  | Cuenta de credito |

Definición de la entidad Work

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| CreditId | VarChar(25) | Not Null | No | Primary Key | Cuenta de credito |
| Departament | VarChar(15) |  | No |  | Departamento al que pertenece |
| Job | VarChar(25) |  | No |  | Cargo que ocupa |

***RELACIONES***

Relacion entre las entidades Movie y Gender

MovieGenders

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key  Foreign Key | Identificador de la pelicula |
| GenderName | VarChar(15) | Not Null | No | Primary Key | Nombre del genero |

Relacion entre las entidades Movie y SpokenLanguages

MovieLanguages

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key  Foreign Key | Identificador de la pelicula |
| ISO639 | VarChar(3) | Not Null | No | Primary Key  Foreign Key | Siglas del lenguaje |

Relacion entre las entidades Movie y ProductionCountries

MovieFrom

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key  Foreign Key | Identificador de la pelicula |
| ISO3166 | VarChar(3) | Not Null | No | Primary Key  Foreign Key | Abreviatura del país |

Relacion entre las entidades Movie y Company

MovieProducedBy

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key  Foreign Key | Identificador de la pelicula |
| IdCompany | Int | Not Null | No | Primary Key  Foreign Key | Identificador de compañia |

Relacion entre las entidades Movie y Employee

MovieCrew

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Campo | Dominio | Opcionalidad | Multivaluado | Tipo | Comentario |
| Id | Int | Not Null | No | Primary Key  Foreign Key | Identificador de la pelicula |
| IdEmployee | Int | Not Null | No | Primary Key  Foreign Key | Identificador de empleado |

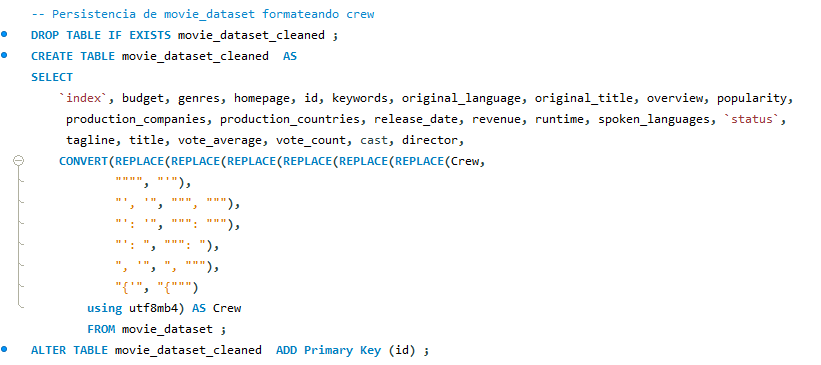
*Estrategia de importación de datos*

Para poder crear las tablas con los datos poblados se sigue la siguiente estrategia

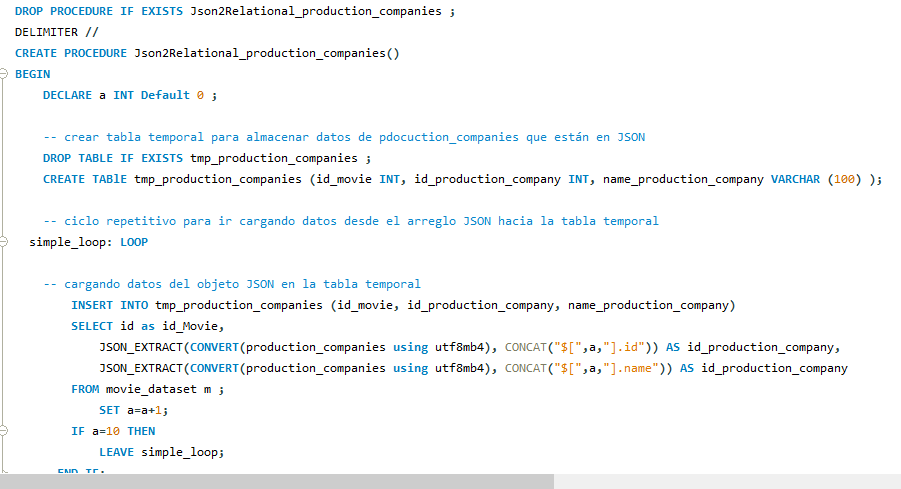
Primero lo que se debe de desarrollar es realizar un ingreso de la data original o del archivo csv donde se almacene en una tabla única y referencial que servirá para ir extrayendo los datos y posteriormente poder colocarlos adecuadamente en las tablas correspondiente.



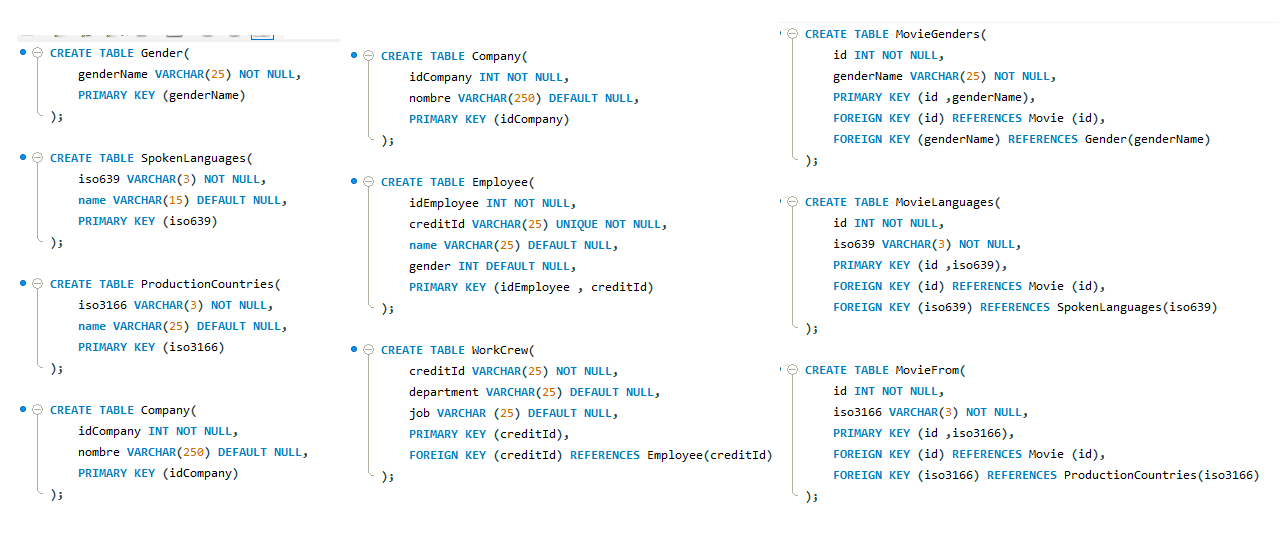
Segundo se tiene que procesar la información para hacerle una limpieza y poder trabajar con esta misma para ello aplicamos una serie de funciones replace para hacer este proceso



Luego será necesario la creación de tablas temporales extrayendo la información de la tabla a la cual se le aplico la limpieza de datos, para ello es necesario la creación de procesos para la creación de las tablas con la data necesaria y específica para estas mismas



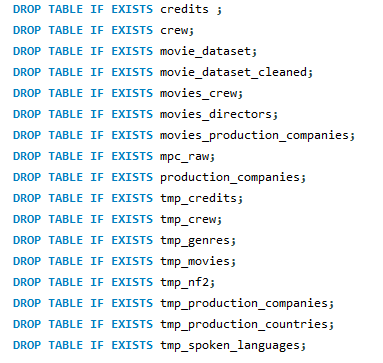
Luego de crear las tablas temporales de da la creación de las tablas finales para ello creamos y corremos el DDL para poder sacar la estructura de las tablas finales



Continuamos realizando un traspaso de la información de las tablas temporales a las tablas finales



Para finalizar hacemos un DROP TABLE de las tablas temporales para quedarnos solo con las tablas finales y con esto nuestro modelo de base de datos se encontraría creado y con la información adecuada



*Conclusiones*

El proyecto nos ayudo a mejorar nuestras habilidades en el campo de las bases de datos ya que este reto planteado fue muy útil y a la vez nos permitió experimentar y desarrollar casos reales que se pueden encontrar en el mundo de los datos, siendo este proyecto muy importante para nuestra futura vida profesional.