**SISTEMA DE BUSINESS INTELLIGENCE:**

PROCESO DE EXTRACCIÓN DE DATOS

NOMBRE: Iker Zabalegui Murugarren

[**1ª FASE: Planificación de la extracción de los datos**](#_i5u8o2dbgduk) **3**

[**2ª FASE: Descarga de los datos**](#_svgt0spnfzvj) **4**

[**3ª FASE: Lectura de los csv**](#_4jji1mebx44x) **5**

[**4ª FASE: Comprobación de los datos**](#_cxa1rvyww3w) **5**

[**5ª FASE: Inserción a la BBDD**](#_22lr4cxvjbfr) **6**

[**Clausura**](#_5h5pnchemvi7) **7**

[**Bibliografía**](#_z7kmlvkgjemx) **7**

## 

## 1ª FASE: Planificación de la extracción de los datos

La primera fase para realizar el script que nos permitirá la extracción de los datos será la fase de planificación. En esta fase se responderá la pregunta de qué es lo que tiene que hacer nuestro programa. Si leemos los requisitos de la tarea tenemos que descargar unos datos de una página web que nos vienen en formato ‘csv’, seguidamente habrá que leer esos csv y tratar los datos y, por último, habrá que conectarse a una Base de Datos que previamente hemos diseñado y cargar los datos conforme a la estructura de la misma.

En mi caso he comenzado primero tomando decisiones en el nivel más bajo (el código) para luego subir hasta el nivel más alto y descender hasta completar el script al completo.

En concreto la primera decisión que he tomado es la del idioma de programación en el que voy a programar: Java con la herramienta NetBeans. Esto se debe a que ya he realizado previos proyectos que requerían la lectura de archivos csv con este idioma de programación y que, con la otra opción que había barajado (Python), no me sentía cómodo después analizar la dificultad del programa a realizar.

Por tanto, ahora toca subir hasta la parte más básica del trabajo y preguntarme qué es lo que realmente tengo que hacer. De nuevo volvemos a los requisitos y nos centramos en el contexto del trabajo. “Se desea disponer de un informe (pueden ser varios gráficos) de la variación de la temperatura media por semana, mes y año; y por comarcas de Navarra (Tierra Estella, Zona Noroeste, Zona Pirineo, Pamplona, Navarra media oriental, Ribera Alta y Tudela - https://es.wikipedia.org/wiki/Comarcas\_de\_Navarra). Haced uso de al menos las estaciones meteorológicas automáticas.”; esto nos dicen los requisitos. De aquí sacamos las siguientes conclusiones: los archivos csv que nos descarguemos tendrán un campo con la temperatura media, nos interesan las estaciones automáticas y estas estaciones habrá que agruparlas por las comarcas indicadas.

## 2ª FASE: Descarga de los datos

Para la descarga de los datos la tarea nos provee de la URL de la página donde se encuentran los datos. Para descargar archivos en Java he modificado el siguiente código que obtuve de internet:

*try (BufferedInputStream inputStream = new BufferedInputStream(new* ***URL****("https://example.com/my-file-path.csv").openStream());*

*FileOutputStream fileOS = new FileOutputStream("****/Users/username/Documents/file\_name.csv****")) {*

*byte data[] = new byte[1024];*

*int byteContent;*

*while ((byteContent = inputStream.read(data, 0, 1024)) != -1) {*

*fileOS.write(data, 0, byteContent);*

*}*

*} catch (IOException e) {*

*// handles IO exceptions*

*}*

Las partes en negrita son las que he cambiado. La primera es la URL de la que descargamos los datos y la segunda es el nombre que se le da al archivo al guardarlo.

Para descargar los archivos la URL debe tener una forma concreta. Para construir la URL necesitamos 3 datos que son los que irán variando:

* El nombre de la estación donde se han recogido los datos.
* El número de esta estación.
* El año del que se desean descargar los datos.

Por tanto es obvio que habrá que iterar por las distintas estaciones y, en cada estación, iterar entre los últimos 15 años que se nos piden. Para esto he creado un archivo csv que contiene todos los nombres de las estaciones con su respectivo número y la comarca a la que pertenecen, ya que luego habrá que agrupar los datos. En el csv ya está realizada la sustitución del carácter espacio por “%20”, necesaria para la conversión de ese nombre a URL. Para la descarga también ha sido necesario cambiar las tildes por otros caracteres, quitar las mayúsculas y demás cambios explicados en el código.

## 3ª FASE: Lectura de los csv

Cada vez que se descarga un archivo csv, se llama a una función que lee los archivos csv y filtra la información que nos interesa para guardarla en una estructura ArrayList. En este caso solo nos interesa la fecha de la medición y la temperatura media de cada día. La única particularidad a la hora de leer los datos es que las celdas de los csv descargados se separan mediante el carácter ‘;’. En las mediciones que corresponden al año 2020, la temperatura media no aparece, por ello, no voy a tener en cuenta este año.

## 

## 4ª FASE: Comprobación de los datos

En este punto ejecutamos el programa para ver si los datos que estamos recogiendo son los correctos.

En algunas estaciones hay varios años que no figuran en la base de datos del gobierno. En otras, la URL cambia. En concreto, las de tipo MAPA hasta el 2018 se llaman MAPAMA. Por último, los archivos de la estación de Goizueta no se descargan y no he llegado a comprender por qué.

## 5ª FASE: Inserción a la BBDD

Tras diseñar la base de datos, cuya estructura se explica en el documento adjunto, es el momento de hacer la inserción de los datos que hemos leído.

Para ello, he utilizado y modificado un trozo de código que encontré en internet:

**Inserción en la BBDD**

*String sDriver = "com.mysql.jdbc.Driver";*

*String sURL = "jdbc:mysql://localhost:3306/lineadecodigo";*

*Connection con = null;*

*Class.forName(sDriver).newInstance();*

*con = DriverManager.getConnection(sURL,"root","password");*

*stmt = con.prepareStatement("INSERT INTO libros VALUES (****?,?,?,?,?****)");*

*String sISBN = "84-9815-212-7";*

*String sTitulo = "Yo, Claudio";*

*String sDescripcion= "Supuesta autobiografía de Claudio...";*

*String sCategoria = "novela histórica";*

*int idAutor = 3;*

*stmt.setString(1,sISBN);*

*stmt.setInt(2,idAutor);*

*stmt.setString(3,sTitulo);*

*stmt.setString(4,sDescripcion);*

*stmt.setString(5,sCategoria);*

*stmt.executeUpdate();*

Cambiando la URL de la BBDD, el “root” y la contraseña, nos conectamos a la BBDD y luego las inserciones se hacen con el smt.

En este caso, sabiendo la estructura, habrá que llenar la tabla de fechas con todas las fechas que hay desde 2005 a 2019, la tabla de Lugar con todas las estaciones y comarcas y la tabla de los hechos con los datos de la temperatura, la clave del lugar, la fecha y la clave del hecho.

La inserción de las fechas y de la tabla Lugar se hará después de haber leído todas las estaciones del csv. Sin embargo, las inserciones a la tabla de hechos se harán cada vez que se lean los datos de cada año en cada estación.

## Clausura

Con este documento espero haber explicado con detalle el proceso que me ha llevado a extraer los datos y las decisiones que he tenido que tomar en el camino.

## Bibliografía

* Descarga de archivos mediante Java: <https://www.it-swarm-es.com/es/java/crear-y-descargar-archivos-csv-java-servlet/1043526617/> (no es concretamente este)
* Inserciones SQL mediante Java: <https://es.stackoverflow.com/questions/359355/inserci%C3%B3n-de-datos-sqlserver-y-java>