Máster de Formación Permanente
en BIG DATA
e Inteligencia Artificial

Universitat d'Alacant
Universidad de Alicante

# Module 5
# Analítica de datos y extracción de conocimiento mediante técnicas de IA
# (Data Analytics)

## Task 2: Multidimensional Logical and Physical Design

### Description:

The use case refers to a technological company which gathers data related to its orders and shipments. The company wishes to better understand the evolution of its sales and returns globally. In a more concrete way, the company is absolutely worried with improving its order's preparation time. To this aim, the company is gathering data related to the exact date when (i) the order is received, and (ii) the shipment is ready and placed in the shipment company. Apart from this critical issue, the company would like to better understand its clients, the main ordered products as well as the main returned products.

The collected data is distributed into three main datasets:

1.- Orders.

Basically, an order represents products ordered by clients. The key information gathered about the order of a product is the number of units (quantity), the sales value (sales), discount, profit, and the shipping cost of an order. The type of orders and the priority are features chosen by clients (also provided in the data set).

The main information provided for customers are their names, customer number, type of customer, zip code, city, state, country, region and market to which they belong to. This latter feature has a lot of potential as it will allow us to analyze potential markets among other factors, as well as assess which are our best clients or identify potential clients.

The main information gathered from products are the name of the product and their different categories depending on their types.

Finally, we are also interested in knowing which products have been returned or not for any reason.

2.- Returns

It contains information on which product from which region has been returned.

3.- Markets

This data set provides further information about which country belongs to which Region and to which Market. In this particular case study, this information has been provided in another different data file.

Try to help your company improve its situation by creating a data analytical repository (data warehouse) that contains the information that will allow us to contain the main data that facilitates its ulterior analytical queries. In order to create this data warehouse, this first task is to specify a multidimensional conceptual schema.

**Submit a zip file** that contains a (i) **logical multidimensional** (MD) schema, based on the fact that the data warehouse will be deployed in a relational database server, and (ii) **the script** that allows us to create the database schema in an Oracle database server. Therefore, the logical MD schema should follow the star schema (and/or its variants). The logical schema should provide a relational schema and the description on the main relationships, data attributes, data types and so on (please, describe the logical schema in a PDF document where the schema and its description are easily presented). The script should allow us to run it in SQL developer and create the corresponding database in an Oracle database server.

Remember that the file submitted must **include the full name(s) of the student(s) involved**.

## To do:

Design the main elements of the (i) logical MD Schema (star schema) as defined in the Lesson 1 and also using other tips learnt from other modules (tables, primary keys, foreign keys, and so on). In other to implement/deploy this star schema, we need to the data types of all the attributes/columns/variables. Please, remember to define the same data types for the attributes that will be used as foreign keys and the corresponding primary keys to be pointed out by them.

The data types to be used depends on the data to be stored in them. Some tips for data types maybe CLOB for undefined length text fields, NUMBER (10) for integers, VARCHAR2 (45 CHAR) for fix text fields (for example of 45 char length), or BINARY_DOUBLE for float numbers.

The defined tables should represent a star schema that will be populated with the data from the data files provided in Lessons. This population will be accomplished in the following Task 3.