Sergio Hervás Aragón

Sergiohervas9@gmail.com

Brief description

This project is a case study of a Business Intelligence (BI) solution for a telecommunications multinational. It involves designing the data analysis system architecture, modeling the database, loading extracted data, and performing data exploitation to support decision-making.

Final practice

The job of the data scientist

Index

Content

[1. Analysis of initial data and proposed data model. 2](#_Toc188957924)

[1.1. Understanding initial data 2](#_Toc188957925)

[1.2. Data processing 3](#_Toc188957926)

[1.3. Design of the proposed data model 4](#_Toc188957927)

[1.4. Why this model? 5](#_Toc188957928)

[2. Model development in MySQL 6](#_Toc188957929)

[2.1. Creating tables and defining keys 6](#_Toc188957930)

[2.2. Loading initial data 8](#_Toc188957931)

[3. Developing loads with Pentaho Data Integration 9](#_Toc188957932)

[3.1. Staging 9](#_Toc188957933)

# Analysis of initial data and proposed data model.

## Understanding initial data

Dataset data:

* **Fecha de creación**: Specifies the time when the ticket is created. It will be used to perform the temporal analysis.
* **Numero de incidente**: Unique ticket identifier.
* **Descripción**: Brief description of the request or incident.
* **Servicio**: The different services that the company provides to its users and clients.
* **Tipo de servicio**: Specify whether it is a request or an incident:
  + **Incidencias**: Restoration of infrastructure or restoration of service to user.
  + **Peticiones**: Service request by user
* **Prioridad**: Specifies the priority with which the ticket should be treated.
  + There are four priorities, ordered from most important to least: Critical, High, Medium, and Low.
  + Only incidents can have critical priority and should only be considered in production environments.
* **Estado**: Indicates the status of the ticket.
  + A ticket is considered finished if it is in Closed or Resolved status.
  + A ticket is considered open if it is pending or assigned.
  + Cancelled tickets do not count.
* **Torre**: It is a functional grouping of services. Each service belongs to a management tower.
* **Entorno**: Indicates the environment that the ticket affects. Production environments are all those that contain a PRO.
* **Estado cumplimiento**: Indicates whether the ticket has been resolved within the time required by the client or not.
  + Within the service objective: ticket resolved appropriately
  + Service objectives not met: ticket resolved late

**Duración días**: Number of days it took to resolve the ticket.

## Data processing

An analysis of the data is performed to verify nulls, and/or data that may be erroneous, in a first instance of .isnull():

Imagen que contiene texto, grande

Descripción generada automáticamente

With the empty data, proceed by columns to:

* Descripción: Nulls are filled with the value “No description”.
* Servicio: Nulls are filled with the value “Unspecified”.
* Torre: Nulls are filled with the value “Unspecified”.
* Entorno: Nulls are filled with the value “Unspecified”.
* Estado cumplimiento SLA: Nulls are filled with the value “Unspecified”.

Imagen que contiene texto, tabla, calle, grande

Descripción generada automáticamente

## Design of the proposed data model

After analyzing the data, the star model was chosen, which will consist of a central fact table that is related to several dimension tables:

* Fact table (fincidencias): Table which will contain numerical information about the incidents.
* Dimension tables:
  + dsevicio: Contains details about the services.
    - id
    - servicio
  + dtipo\_servicio: Contains details about service types.
    - id
    - tipo\_servicio
  + dprioridad: Contains details about priority.
    - id
    - prioridad
  + destado: Contains details about the status.
    - id
    - estado
  + dtorre: Contains details about management.
    - id
    - torre
  + dentorno: Contains details about the environment.
    - id
    - entorno
  + dsla: Contains details of the compliance status.
    - id
    - sla

## Why this model?

The use of this model is based on the use of the term ‘DISTINCT’ in MySQL queries by columns of our Dataset, and based on the data obtained, if it is a large amount, the column is discarded as an identifier:

* fecha\_creacion: 58376 different records, so it is not in our model.
* num\_incidente: 58564 distinct records, so it is not in our model.
* descripción: 51925 distinct records, so it is not in our model.
* servicio: 114 different records, so it is in our model.
* tipo\_servicio: 4 different records, so it is in our model.
* prioridad: 4 different records, so it is in our model.
* estado: 5 different records, so it is in our model.
* torre: 19 different records, so it is in our model.
* entorno: 881 different records, so it is in our model.
* sla: 3 different records, so it is in our model.
* duración\_dia: 55467 distinct records, so it is not in our model.

# Model development in MySQL

## Creating tables and defining keys

Based on the chosen model:

* Dimension tables:
  + dentorno:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| identorno | INT | X | X |  |  |  |  |  |  |
| entorno | VARCHAR |  |  |  |  |  |  |  |  |

* + destado:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| idestado | INT | X | X |  |  |  |  |  |  |
| estado | VARCHAR |  |  |  |  |  |  |  |  |

* + dprioridad:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| idprioridad | INT | X | X |  |  |  |  |  |  |
| prioridad | VARCHAR |  |  |  |  |  |  |  |  |

* + servicio:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| idservicio | INT | X | X |  |  |  |  |  |  |
| servicio | VARCHAR |  |  |  |  |  |  |  |  |

* + sla:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| idsla | INT | X | X |  |  |  |  |  |  |
| sla | VARCHAR |  |  |  |  |  |  |  |  |

* + tipo\_servicio:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| idtipo\_servicio | INT | X | X |  |  |  |  |  |  |
| tipo\_servicio | VARCHAR |  |  |  |  |  |  |  |  |

* + torre:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| idtorre | INT | X | X |  |  |  |  |  |  |
| torre | VARCHAR |  |  |  |  |  |  |  |  |

* Fact table:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| id | INT | X | X |  |  |  |  |  |  |
| identorno | INT | X | X |  |  |  |  |  |  |
| idestado | INT | X | X |  |  |  |  |  |  |
| idprioridad | INT | X | X |  |  |  |  |  |  |
| idservicio | INT | X | X |  |  |  |  |  |  |
| idsla | INT | X | X |  |  |  |  |  |  |
| idtipo\_servicio | INT | X | X |  |  |  |  |  |  |
| idtorre | INT | X | X |  |  |  |  |  |  |

## Loading initial data

When loading the data, it involved creating a table that reflected the structure of the data contained in the XLS file, so a table was designed with column names that matched the column headers in the XLS file, and the data types of each column were adjusted accordingly.

Initial data loading table (staging):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column Name | Datatype | PK | NN | UQ | B | UN | ZF | AI | G |
| fecha\_creacion | DATETIME |  |  |  |  |  |  |  |  |
| num\_incidente | VARCHAR |  |  |  |  |  |  |  |  |
| descripción | VARCHAR |  |  |  |  |  |  |  |  |
| servicio | VARCHAR |  |  |  |  |  |  |  |  |
| tipo\_servicio | VARCHAR |  |  |  |  |  |  |  |  |
| prioridad | VARCHAR |  |  |  |  |  |  |  |  |
| estado | VARCHAR |  |  |  |  |  |  |  |  |
| torre | VARCHAR |  |  |  |  |  |  |  |  |
| entorno | VARCHAR |  |  |  |  |  |  |  |  |
| sla | VARCHAR |  |  |  |  |  |  |  |  |
| duración\_dia | DOUBLE |  |  |  |  |  |  |  |  |

# Developing loads with Pentaho Data Integration

## Staging