

**E**scola **S**uperior de **E**nxeñaría **I**nformática

|  |
| --- |
| PRÁCTICA INTELIGENCIA DE NEGOCIO Y ALMACENES DE DATOS  (curso 2020-2021)  **Código Grupo: G2\_1**  **Título: HOTEL** |

Alumna/o: Rubén Gómez Dopazo

Alumna/o: Antonio Cebreiro Bernárdez

Alumna/o: Manuel Certal Puga

Alumna/o: César Gabriel Márquez Rodríguez

**ÍNDICE**

# Descripción del proceso

* 1. **Descripción textual de la organización**

Elite Hotels es una compañía hotelera que cuenta con numerosos hoteles de lujo. Cada uno cuenta con uno o varios edificios, distribuidos en diferentes plantas. Los hoteles de la cadena cuentan con todo tipo de instalaciones para el disfrute de los huéspedes, desde restaurantes y gimnasios hasta campos de golf o tenis.

En sus hoteles, la compañía ofrece dos tipos de hospedaje: apartamentos y habitaciones. Además, existen espacios de trabajo únicamente accesibles para los trabajadores del hotel.

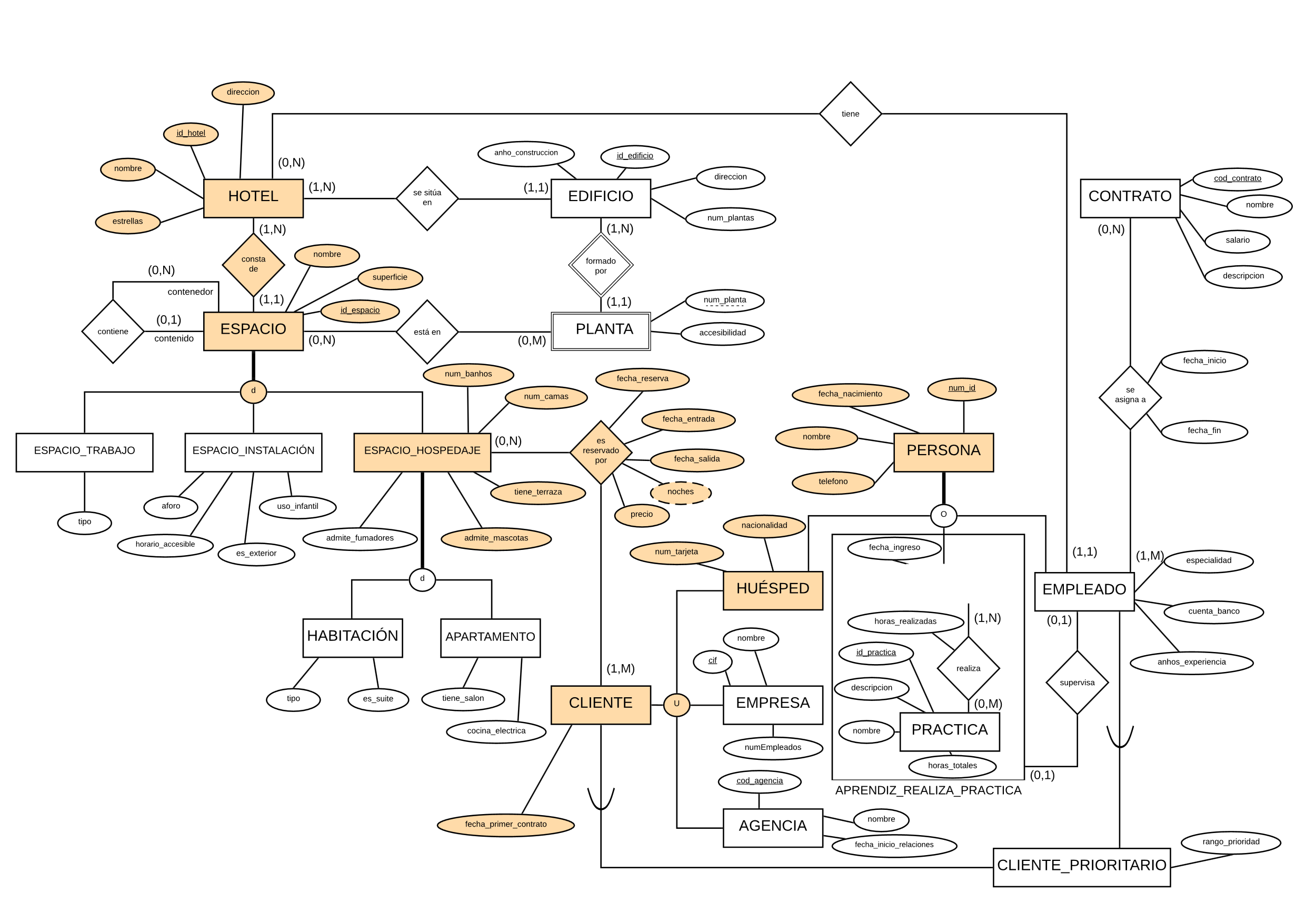
La cadena hotelera maneja información de distintos tipos de personas: huéspedes, empleados con contrato y aprendices. Estos últimos realizan prácticas que son supervisadas por empleados de la cadena. Además, en el caso de que alguien sea empleado y a la vez cliente, se le trata como cliente prioritario, por lo que tiene ciertas ventajas al utilizar los servicios de la cadena.

Los huéspedes no son el único tipo de clientes con los que la compañía mantiene relaciones, sino que esta también trabaja con agencias de viajes y otras empresas de todo tipo. De esta manera, cuando una empresa, agencia o huésped quiere contratar los servicios de la cadena, ésta los trata a todos por igual como clientes. Estos realizan reservas de los espacios de hospedaje que ofrece la cadena dentro de sus hoteles para una fecha de entrada y salida concretas.

* 1. **Modelo E/R original del sistema OLTP preexistente**

En un primer momento, la compañía hotelera diseñó e implementó un sistema OLTP para registrar sus operaciones diarias. Este sistema queda descrito por el siguiente modelo E/R, en el que se indican las entidades, atributos y relaciones que modelan la actividad de la organización.

Sobre el diagrama mostrado se ha coloreado el área que será de interés para la creación del nuevo sistema.



***Figura 1.*** *Diagrama del modelo E/R correspondiente al sistema OLTP preexistente.*

* 1. **Modelo E/R con las entidades que almacenan información de interés para el mercado de datos**

***1.3.1. Explicar la incorporación de nuevos atributos a los preexistentes en la organización, en su caso.***

A continuación, se muestran los nuevos atributos que han sido añadidos a las relaciones señaladas como interesantes en la Figura 1. Además, se da una justificación de por qué se considera necesario para el nuevo sistema de mercado de datos.

Para la relación **HOTEL**:

|  |  |  |
| --- | --- | --- |
| **Atributo(s)** | **Tipo** | **Justificación** |
| PAIS y CIUDAD | varchar(32) | Interesa conocer cómo se desarrolla la actividad de los hoteles de diferentes ciudades y países. |
| PARKING | enum('ABIERTO',  'CERRADO') | Se podrá ver cómo influye el tipo de aparcamiento sobre el rendimiento de los hoteles. |
| TRASLADO\_  AEROPUERTO,  TIENE\_ASCENSOR,  TIENE\_SH | enum('Sí', 'No') | Permitirá comprobar si contar con servicios como traslado al aeropuerto, presencia de ascensores o servicio de habitaciones influye sobre la actividad de los hoteles. |
| VALORACION | decimal(2,1) | Podrá mostrar la relación entre las características del hotel y las opiniones de los clientes. |
| TIPO | enum('HISTORICO',  'CONCEPTUAL',  'MODERNO',  'TRADICIONAL') | Se podrá ver cómo influye el tipo de hotel sobre el rendimiento de los hoteles con relación a factores como qué tipos de cliente los visitan, en qué épocas... |
| PRESUPUESTO, PRES\_MARKETING e INGRESOS | decimal(11,2) | Se podrá ver cómo influyen el presupuesto total, el presupuesto dedicado a marketing y los ingresos generados de los hoteles dadas sus características, las características de los clientes que los visitan, etc. |
| DISTANCIA\_CENTRO, DISTANCIA\_PLAYA | enum('< 500m', '< 1Km', '< 2Km', '< 3Km', '< 5Km') | Permitirá comprobar cómo influyen las distancias a lugares de interés como playas o el centro de la ciudad sobre la actividad de los hoteles. |

Para la relación **ESPACIO**:

|  |  |  |
| --- | --- | --- |
| **Atributo(s)** | **Tipo** | **Justificación** |
| TIENE\_INTERNET, TIENE\_VISTAS, ESTA\_ACLIMATIZADO | enum('Sí', 'No') | Interesa saber cómo influyen características del espacio sobre su rendimiento de cara a reservas, uso por parte de los clientes… |
| TIPO\_ESPACIO | enum('ESPACIO\_  HOSPEDAJE', 'ESPACIO  \_INSTALACION', 'ESPACIO  \_TRABAJO’) | Permitirá facilitar el trabajo para la carga de información en el datamart cargando ´la información sólo del tipo que nos interese (ESPACIO\_HOSPEDAJE). |

Para la relación **ESPACIO\_HOSPEDAJE**:

|  |  |  |
| --- | --- | --- |
| **Atributo(s)** | **Tipo** | **Justificación** |
| ADMITE\_FUMADORES,  CANCELACION\_GRATIS,  MOVILIDAD\_ADAPTADA | enum('Sí', 'No') | Interesa saber cómo influyen características del espacio sobre su rendimiento de cara a reservas, uso por parte de los clientes… |
| TIPO\_HOSPEDAJE | enum('APARTAMENTO', 'HABITACION') | Se podrá ver la relación entre el tipo de hospedaje y los perfiles de clientes que los reservan. |
| PRECIO\_NOCHE | decimal(5,2) | Permitirá relacionar categorizar a los alojamientos por precio y relacionarlos con los perfiles de cliente que los reservan. |
| LUMINOSIDAD | enum('BAJA', 'MEDIA', 'ALTA') | Se podrá ver cómo afecta el tipo de luminosidad al a las reservas de los espacios de hospedaje. |

Para la relación **PERSONA**:

|  |  |  |
| --- | --- | --- |
| **Atributo(s)** | **Tipo** | **Justificación** |
| GENERO | enum('Masculino', 'Femenino', 'Otro') | Se podrán realizar análisis clasificando a la gente por género. |

Para la relación **HUESPED**:

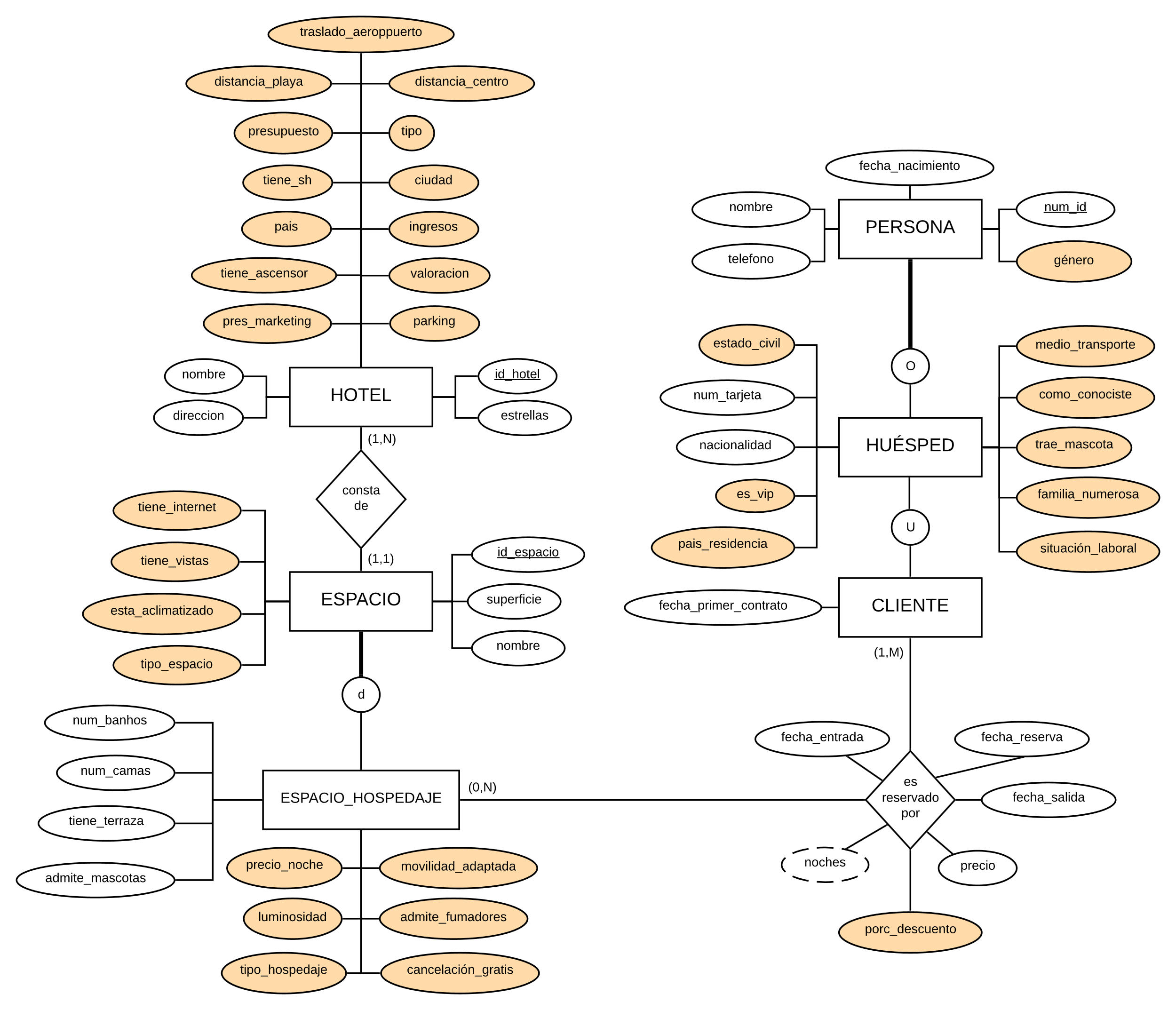
|  |  |  |
| --- | --- | --- |
| **Atributo(s)** | **Tipo** | **Justificación** |
| FAMILIA\_NUMEROSA | varchar(20) | Se podrán realizar análisis clasificando a la gente por género. |
| ESTADO\_CIVIL | enum('APARTAMENTO', 'HABITACION') | Se podrá ver la relación entre el tipo de hospedaje y los perfiles de clientes que los reservan. |
| SITUACION\_LABORAL | enum('Empresario','Autónomo',  'Asalariado', 'Cooperativista',  'En paro', 'Jubilado') | Información como la situación laboral o el medio de transporte en el que viajaron nos permitirán categorizar a los clientes, |
| MEDIO\_TRANSPORTE | enum('Coche','Avión','Tren',  'Moto','Bus','Camión') |
| COMO\_CONOCISTE | enum('Internet','Publicidad',  'Conocidos','Otros') | Permitirá ver qué medios de difusión son más efectivos en función de los tipos de cliente. |
| TRAE\_MASCOTA,  ES\_VIP | enum('Sí', 'No') | Se podrá ver cómo características como el traer mascotas o ser vip influyen sobre las reservas de los clientes. |

Para la relación **CLIENTE\_RESERVA\_ESP\_HOSPEDAJE**:

|  |  |  |
| --- | --- | --- |
| **Atributo(s)** | **Tipo** | **Justificación** |
| PORC\_DESCUENTO | decimal(2,0) | Permitirá saber cómo afectan los porcentajes de descuento a las reservas y relacionar esto con los clientes que las hacen |

Además de todos los atributos descritos anteriormente, se ha incorporado a cada tabla de entidad un nuevo campo ‘ULTIMA\_ACTUALIZACION’. Este es una marca de tiempo que nos permitirá llevar la información nueva (o actualizada) al nuevo sistema.

***1.3.2. Diagrama del modelo E/R con las entidades de interés para el mercado de datos y los nuevos atributos***

En el siguiente diagrama E/R se muestran solamente las entidades y relaciones de interés señaladas en la Figura 1, incluyendo los nuevos atributos descritos en el apartado anterior. Los nuevos atributos han sido coloreados para distinguirlos de los que ya formaban parte del E/R original.

***Figura 2.*** *Diagrama del modelo E/R ampliado.*

* 1. **Descripción de las fuentes externas**

En las siguientes tablas, se presentan las fuentes externas que le serán incorporadas al nuevo sistema de mercado de datos. De cada una, se hace una breve descripción y se señalan los atributos específicos que se manejan.

|  |  |  |
| --- | --- | --- |
| **FUENTE** | **World Bank: Total Population** | |
| **Descripción** | World Bank ofrece un dataset que contiene la evolución poblacional de los diferentes países del mundo desde el año 1960 hasta la actualidad. El formato de los datos es XLS. | |
| **ATRIBUTO** | **TIPO** | **DESCRIPCIÓN** |
| **País** | String | El país al que corresponden los datos. |
| **Año** | Integer | Año al que corresponden los datos del país. |
| **Población** | Integer | La población del país para el año y país |

|  |  |  |
| --- | --- | --- |
| **FUENTE** | **World Bank: Countries classified by income group** | |
| **Descripción** | World Bank ofrece un dataset en el que agrupa a los diferentes países del mundo en función de su nivel de ingresos. Además, los agrupa a su vez según la región en la que se localicen. El formato de los datos es CSV. | |
| **ATRIBUTO** | **TIPO** | **DESCRIPCIÓN** |
| **País** | String | El país al que corresponden los datos. |
| **Nivel de ingresos** | Enum: bajo, medio-bajo, medio-alto, alto | Muestra el grupo de nivel de ingresos del país. |

|  |  |  |
| --- | --- | --- |
| **FUENTE** | **Kaggle: Average Temperature per country per year** | |
| **Descripción** | Kaggle ofrece un dataset en el que se muestran las temperaturas medias de los países del mundo en los últimos años. El formato de los datos es CSV. | |
| **ATRIBUTO** | **TIPO** | **DESCRIPCIÓN** |
| **País** | String | El país al que corresponden los datos. |
| **Año** | Integer | Año al que corresponde la temperatura media del país. |
| **Temperatura** | Float | Temperatura media del año para el país. |

* 1. **Descripción textual de la actividad a modelar**

La actividad de la organización a modelar en el nuevo sistema de mercado de datos es la reserva de espacios de hospedaje en los diferentes hoteles realizadas por los huéspedes de la cadena.

# Selección de la granularidad

* 1. **Descripción del gránulo de la actividad a modelar y de cada una de las dimensiones**

Se almacenan las **reservas diarias** que un **huésped** realiza para un **espacio de hospedaje** de un hotel. En la siguiente tabla se muestran los gránulos para cada una de las dimensiones del nuevo sistema de mercado de datos.

|  |  |  |
| --- | --- | --- |
| **Dimensión** | **Gránulo Menor** | **Gránulo Mayor** |
| **Quién** | Huésped | Huésped |
| **Qué** | Espacio hospedaje | Espacio hospedaje |
| **Cuándo** | Día | Año |
| **Dónde** | Hotel | País |

* 1. **Razonar por qué no se ha optado por una granularidad mayor o menor**

**¿Quién? Huésped**

Se quiere recopilar los datos personales (edad, nacionalidad, país de residencia, etc.) de cada **huésped** para un análisis de comportamiento a la hora de realizar una reserva, en función de su perfil.

**¿Qué? Espacio hospedaje**

Nos interesa manejar la información del objeto de reserva **a nivel del espacio de hospedaje** elegido por un cliente. Así, se agrupará la información sobre las características comunes de **habitaciones y apartamentos**. De esta manera, se podrá ver cómo influyen estas características sobre las reservas.

**¿Cuándo? Día y Año**

Como gránulo menor se ha escogido **‘día’** porque es la unidad mínima de tiempo en la que se pueden hacer reservas. Como gránulo mayor se ha optado por **‘año’**, porque el período de vida de la empresa es de 10 años.

**¿Dónde? Hotel y País**

Cómo gránulo mínimo interesa que el lugar de reserva sea **‘hotel’**, de manera que podamos utilizar información como su número de estrellas o distancia a la playa o al centro, para el futuro análisis. Como gránulo mayor se quiere utilizar **‘país’**, para poder analizar tendencias a nivel internacional.

# Identificación de las dimensiones

* 1. **Descripción textual de las dimensiones (a nivel general, sin detalle de los atributos)**

**Dimensión "Dónde"**

Esta dimensión intenta responder a consultas en las que la pregunta es "Dónde es la reserva realizada". En ella se puede encontrar diferentes atributos para poder conocer las características geográficas de los hoteles en los que se hacen las reservas.

**Dimensión "Qué"**

En esta dimensión las consultas que se intentan resolver son las relacionadas con "Qué ha sido reservado". De esta manera, se podrá realizar un análisis centrado en los rasgos de los espacios de hospedaje reservados.

**Dimensión "Cuándo"**

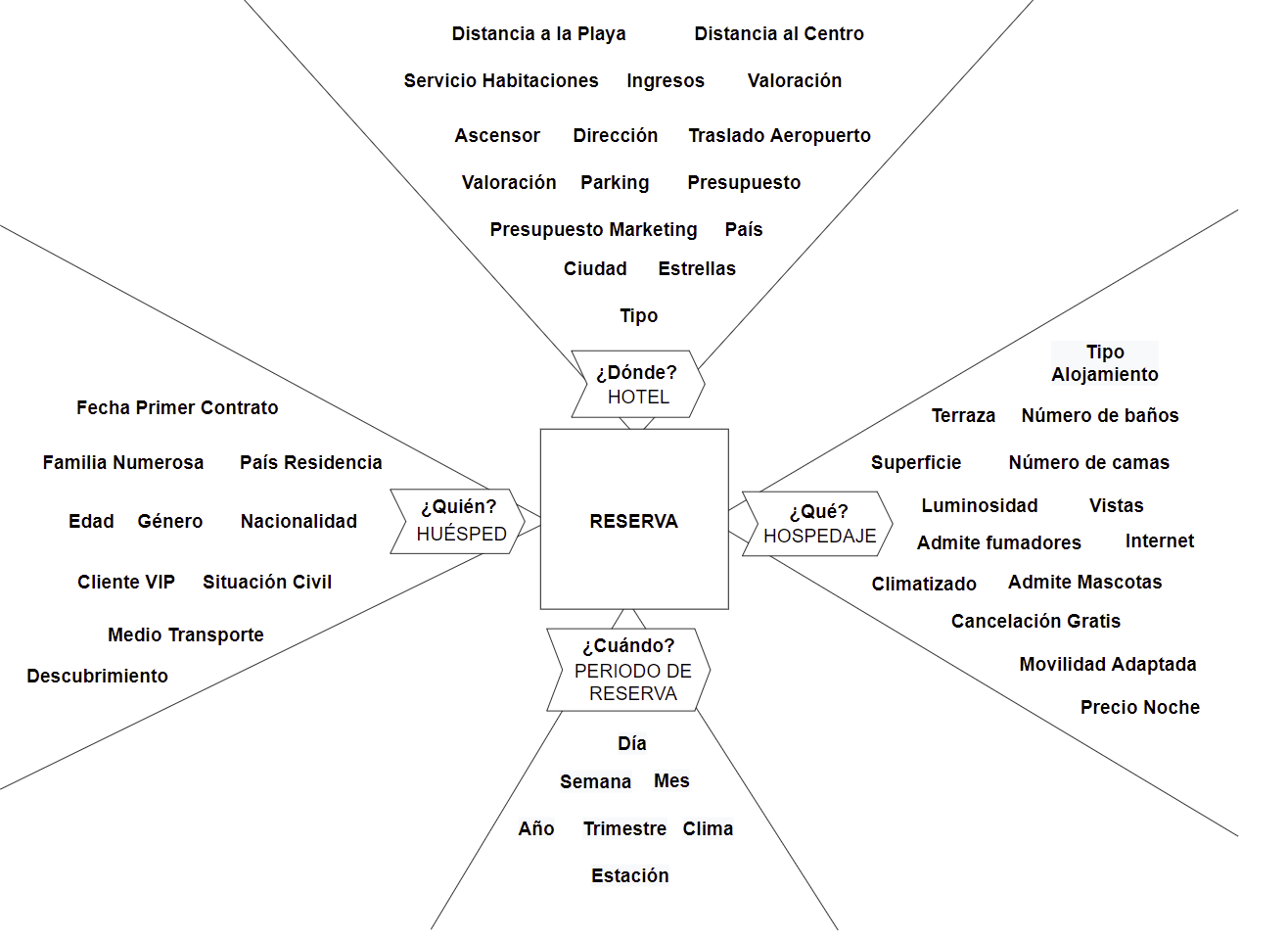
Esta servirá para contestar a consultas del tipo "Cuándo se ha realizado la reserva". De este modo, se podrá consultar en qué medida aspectos como la época del año o de eventos puntuales (internos o externos a la organización) afectan a las reservas.

**Dimensión "Quién"**

Esta dimensión se utilizará en consultas en las que se desea saber "Quién realiza las reservas". De esta manera, se puede identificar cuál es el tipo de cliente más frecuente de la organización.

* 1. **Diagrama de las dimensiones**

A continuación, se muestra un diagrama en el que se muestran las dimensiones del datamart, tal y como fueron descritas en las secciones anteriores.

Además, se indican los atributos que han sido identificados para cada una de las dimensiones. Estos serán tomados tanto de los atributos del sistema OLTP preexistente (viejos y nuevos), como de las fuentes externas descritas anteriormente.

***Figura 3.*** *Diagrama de las dimensiones del datamart sin las métricas.*

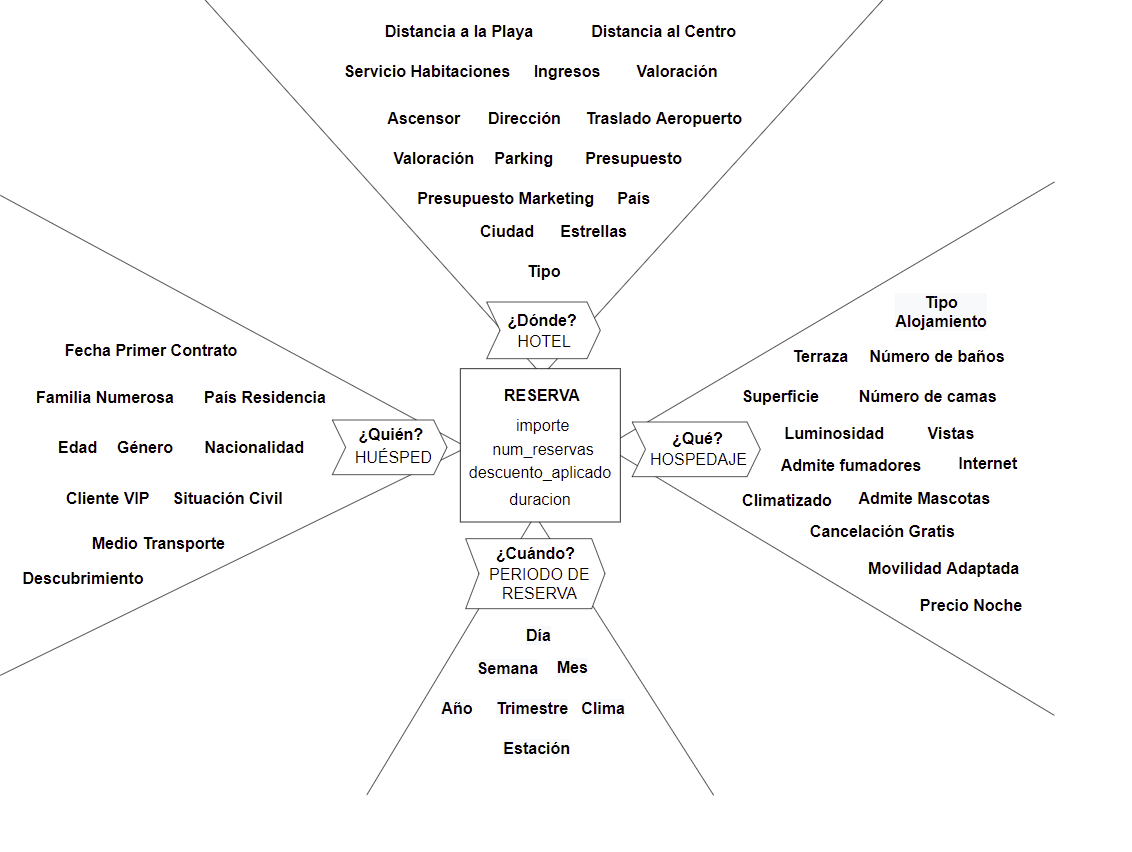
# Selección de las medidas asociadas al hecho

* 1. **Descripción textual de las medidas y de dónde se obtienen (a qué atributo(s) se corresponde en la BD OLTP).**

En la siguiente tabla se presentan las medidas que se emplearán para evaluar el rendimiento de la actividad modelada en el datamart.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MEDIDA** | **TIPO** | **DESCRIPCIÓN** | **TABLA/FUENTE ORIGEN** | **ATRIBUTO/CAMPO ORIGEN** |
| **importe** | *float* | Sumatorio de los importes de las reservas | CLIENTE\_  RESERVA\_  ESPACIO\_  HOSPEDAJE | PRECIO |
| **num\_reservas** | *int* | Número de reservas realizadas | CLIENTE\_  RESERVA\_  ESPACIO\_  HOSPEDAJE | ID\_RESERVA |
| **descuento\_aplicado** | *float* | Sumatorio de los descuentos aplicados | CLIENTE\_  RESERVA\_  ESPACIO\_  HOSPEDAJE | PORC\_DESCUENTO  \* PRECIO \* 0.01 |
| **duración** | *float* | Número medio de noches de las reservas | CLIENTE\_  RESERVA\_  ESPACIO\_  HOSPEDAJE | NOCHES = FECHA\_ENTRADA – FECHA\_SALIDA |

* 1. **Diagrama con la incorporación de las medidas**

En el siguiente diagrama se incluyen las métricas descritas en el apartado anterior al diagrama mostrado en la Figura 3.

***Figura 3.*** *Diagrama de las dimensiones del datamart con las métricas.*

# Almacenamiento de valores precalculados en la tabla de hechos

* 1. Definición de valores precalculados y descripción acerca de cómo se obtienen
  2. Diagrama con incorporación de los valores precalculados

# Terminación de las tablas de dimensión

* 1. Descripción detallada de las dimensiones. Para cada uno de sus atributos indicar:
     1. Cómo se obtiene (atributo(s) específico(s) en OLTP, campo calculado, fuente externa X, etc.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DIMENSIÓN | (nombre dimensión) | | | |
| Descripción de la dimensión |  | | | |
| ATRIBUTO | TIPO | DESCRIPCIÓN | TABLA/FUENTE ORIGEN | ATRIBUTO/CAMPO ORIGEN |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

* + 1. Establecimiento de la jerarquía entre los atributos de cada dimensión (p.ej: día, mes, año)

|  |  |
| --- | --- |
| DIMENSIÓN | (nombre dimensión) |
| Jerarquía |  |

* + 1. Conversión de datos: numerización/etiquetado, discretización (hacer al menos 2)

|  |  |  |
| --- | --- | --- |
| DIMENSIÓN |  | |
| ATRIBUTO | TIPO DE  CONVERSIÓN | DESCRIPCIÓN |
|  |  |  |
|  |  |  |

* + 1. Cómo se debe tratar en caso de datos erróneos

|  |  |
| --- | --- |
| DIMENSIÓN | (nombre dimensión) |
| ATRIBUTO | DESCRIPCIÓN DEL TIPO DE ERROR Y CÓMO RESOLVERLO |
|  |  |
|  |  |

* + 1. Cómo se debe tratar en caso de datos faltantes

|  |  |
| --- | --- |
| DIMENSIÓN | (nombre dimensión) |
| ATRIBUTO | DESCRIPCIÓN DEL TIPO DE PROBLEMA Y CÓMO RESOLVERLO |
|  |  |
|  |  |

* 1. Diagrama final de diseño de la BD Datawarehouse versión ROLAP (incorporando claves)

# Selección de la duración de la base de datos

* 1. Indicación de la duración elegida y justificación

# Control de las dimensiones lentamente cambiantes

* 1. Determinación de las dimensiones lentamente cambiantes y cuáles son los atributos afectados.
  2. Para cada uno de los atributos cambiantes indicar cómo será tratado (Tipo 1, 2 o 3). *Es necesario incluir al menos un atributo cambiante Tipo 2.*

# Modos de consulta y prioridades de ejecución

* 1. Descripción textual de las consultas OLAP