Universidad de A Coruña

Facultad de Informática

Grado en Ingeniería Informática

# INFORME DE CALIDAD DE DATOS

**Álvaro Fernández González –** *alvaro.fernandezg*

**Ana Armenteros López –** *ana.armenteros*

**Krizia Paola González García –** *krizia.gonzalez*

**Rubén Portos Rey –** *ruben.prey*

A Coruña

Marzo de 2022

# ÍNDICE DE CONTENIDO

# ÍNDICE DE FIGURAS

# ÍNDICE DE ANEXOS

# INTRODUCCIÓN

En este informe se exponen las carencias encontradas, respecto a la calidad de los datos, de la base de datos de la empresa, así como las recomendaciones para tratar de solucionarlas o evitarlas. Para ello, se ha hecho uso de la herramienta Talend Open Studio for Data Quality y se ha estructurado según las diferentes dimensiones "clave" de calidad de los datos y usando el fichero *tbi_new.sql* como apoyo.

# EXACTITUD

Se ha analizado la forma en la que se representan los datos y las inexactitudes de sus respectivos valores. Para ello, se ha diferenciado entre la **exactitud sintáctica**, que busca que los datos sigan un patrón sintáctico preestablecido, y la **exactitud semántica**, que persigue una correspondencia de los datos con el mundo real.

## 1. Deficiencias

Para la columna "**email**" de la tabla "**customer**" se ha comprobado, aplicando un patrón *Regex* específico para emails, que no se cumple con la exactitud sintáctica (ver Figura 1). Como se puede apreciar, el 42.79% de los valores contienen espacios, símbolos especiales no permitidos, etc. que no respetan con dicho patrón, lo cual puede conllevar a consultas insatisfactorias sobre estos datos.

En la columna "**marital_status**" de la tabla "**employee**" se ha comprobado mediante un indicador de "value_frecuency" la presencia de un valor "3" con una frecuencia de 19.57% (ver Figura 2), el cual no corresponde a ningún valor real para esa columna e incumple con la exactitud semántica. En la vida real, los valores corresponderían con "Married" ("M") o "Single" ("S").

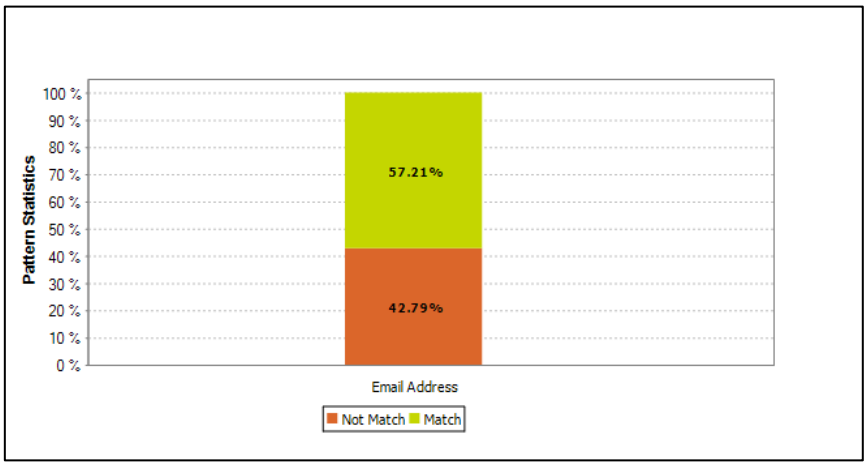


*Figura 1.Porcentaje de exactitud con un patrón para email*

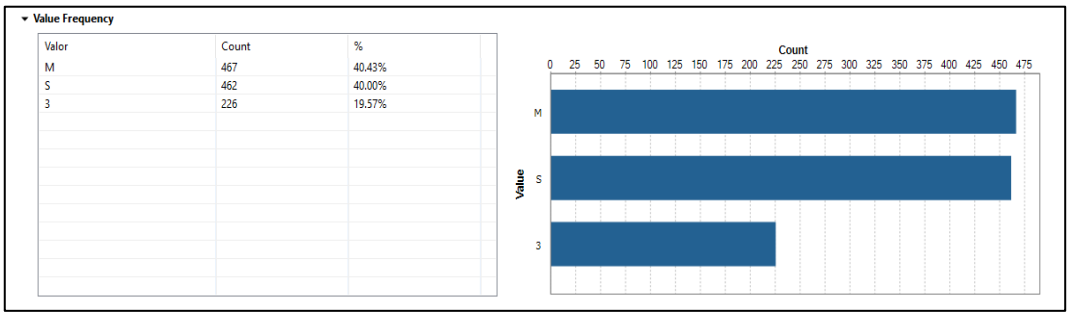


*Figura 2. "value_frecuency" para los valores de la columna "employee.marital_status"*

## 2. Recomendaciones

La mejor forma de solucionar errores con la exactitud sintáctica es documentar qué patrón deben seguir las columnas que pueden verse afectadas (p.e. la columna "**email**", que tiene que seguir el patrón *usuario@dominio*).

La columna **"gender"** de la tabla **"customer"** es un gran ejemplo de cómo respetar la exactitud semántica, ya que sus valores representativos son Female (F) y Male (M) con un nivel de exactitud de un 99.45% (ver Figura 3). Como se puede ver, F y M, que son los valores reales de género, son los predominantes. Existen otros valores en la columna que, al tener tan baja frecuencia, se pueden considerar como errores a la hora de introducir los datos.

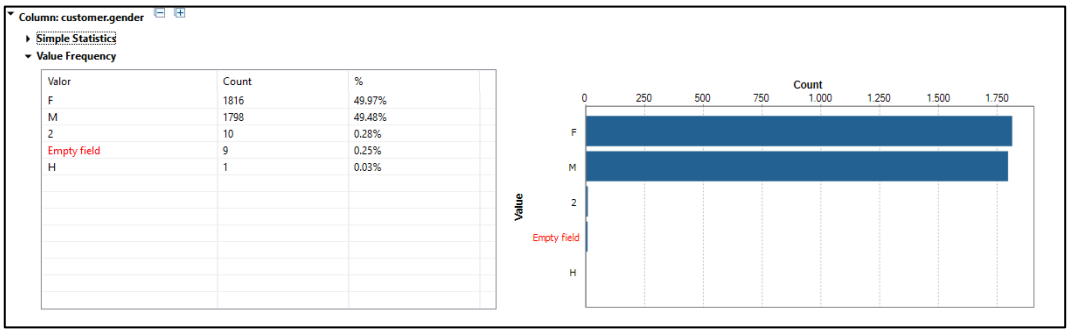


*Figura 3. "value frecuency" para los valores de "customer.gender"*

# COMPLETITUD

Se han analizado la presencia y significado de los valores nulos en las columnas de cada tabla. A menor completitud, mayores son las posibilidades de que se pierdan datos o de que una consulta sea insatisfactoria.

## 1. Deficiencias

La tabla **"company"** no tiene datos almacenados para ninguna de sus columnas, por lo que está incompleta. Este mismo caso ocurre para "**fiscal_period**" en la tabla "**time_by_day**"; y para **"warehouse_owner_name"** en la tabla "**warehouse**".

La tabla **"customer"** tiene una columna **"mi"** con numerosos valores nulos y dos de ellos en blanco (ver Figura 4 y Figura 5). Al no tener una documentación y el nombre ser poco descriptivo, no se ha podido evaluar el significado de los valores nulos correctamente. Este problema también se presenta para "**address1**", "**address2**" y "**address3**" de las tablas "**customer**"; y para "**wh_address2**", "**wh_address3**", "**wh_address4**" en la tabla "**warehouse**".

| Label | Count | % |
|---|---|---|
| Row Count | 3634 | 100.00% |
| Null Count | 1540 | 42.38% |
| Distinct Count | 49 | 1.35% |
| Unique Count | 19 | 0.52% |
| Duplicate Count | 30 | 0.83% |
| Blank Count | 2 | 0.06% |

*Figura 4. Muestra en forma de tabla del análisis de los valores de la tabla "mi"*

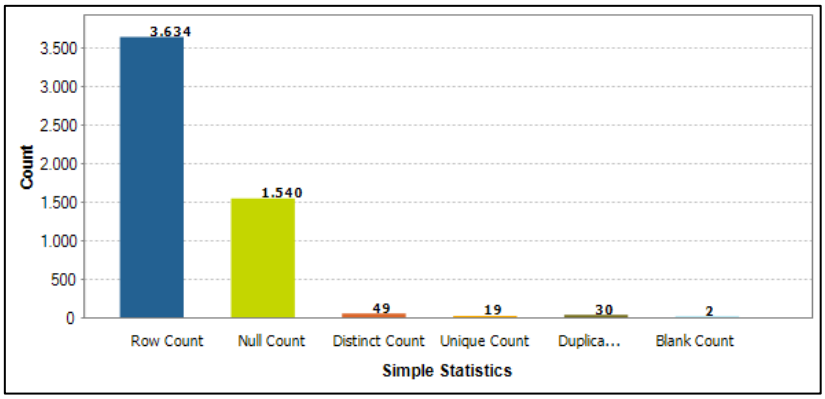


*Figura 5. Muestra en forma de diagrama de barras del análisis de los valores de la tabla "mi".*

En la columna **"country"** de la tabla "**customer**", aunque hay un pequeño porcentaje de valores nulos (3.08%), debería conocerse para todos los casos. El mismo caso aplica para "**phone1**" en la tabla "**customer**"; para "**email**" y "**position_title**" en la tabla "**employee**"; y para "**the_day**" de la tabla "**time_by_day"**.

Si se desea ver en más detalle los porcentajes resultantes de los valores nulos anteriormente mencionados (ver ANEXO A).

## 1. Recomendaciones

Aquellas tablas en las que se conocen los datos antes de su creación, como la de la propia empresa, deben rellenarse en el momento que se crean. Si se considerasen datos que deban ser introducidos a lo largo del tiempo, se debe indicar de forma clara por medio de documentación.

Se debe hacer uso de la cláusula *NOT NULL* para aquellas columnas cuyos valores sean significativos. De esta manera, no se permitiría nunca la inserción de un valor nulo en dicha columna, lo que evita una pérdida de datos debida a un error humano (ver Figura 6). Por otro lado, si se consideran algunos datos como no estrictamente necesarios, también se debe reflejar en la debida documentación.

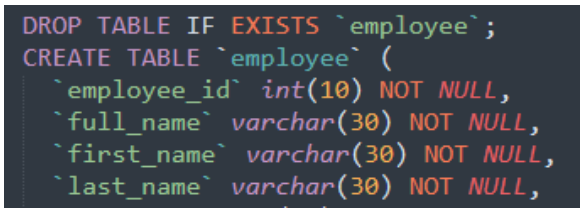


*Figura 6. Ejemplo correcto de restricción de valores no nulo en la tabla "employee".*

# CONSISTENCIA

Se ha analizado el cumplimiento de las reglas semánticas definidas sobre los datos. Para ello, hemos hecho hincapié en las siguientes restricciones de integridad:

1) **Restricciones de dominio**. Definir un rango de valores posibles para una columna disminuye la posibilidad de introducir valores erróneos.
2) **Restricciones de clave principal.** Toda tabla debe tener una clave principal que sea única, de forma que se puedan identificar de manera unívoca y se evite la duplicidad de los datos. Estas claves no deben tener valores nulos.
3) **Restricciones de clave foránea.** Si una tabla hace referencia a otra, ésta debe hacerlo hacia la clave principal de la otra tabla.
4) **Dependencias funcionales.** Ciertas columnas pueden tener un valor que determine el valor de otra columna. Por ejemplo, el valor de la columna "**warehouse_city**" puede determinar el valor de otra columna. "**warehouse_country**".

## 1. Deficiencias

No existe documentación que especifique el rango de valores posibles para cada columna y tampoco se hace uso de la cláusula *CHECK* en la creación de cada tabla para limitar dichos valores, por lo que no se hace uso de restricciones de dominio.

Las tablas "**product**", "**sales_fact**" y "**warehouse**" no tienen una clave principal asignada, por lo cual no cumplen con las restricciones de clave principal (ver Figura 7).

Debido a lo anterior, tampoco se cumple con las restricciones de clave foránea. La columna "**prod_id**" de la tabla "**sales_fact**" debería hacer referencia a "**prod_id**" de la tabla "**product**", pero sería imposible al no ser esta última una clave primaria.

Entre las columnas "**state_province**" y **"country"** de la tabla "**customer**" existe un bajo nivel de dependencia, pero no tenemos los datos suficientes para concluir si se cumple con las dependencias funcionales. Por falta de documentación, desconocemos si pueden existir los mismos nombres para diferentes países o si, en algunos casos, el país puede ser nulo (ver Figura 8).

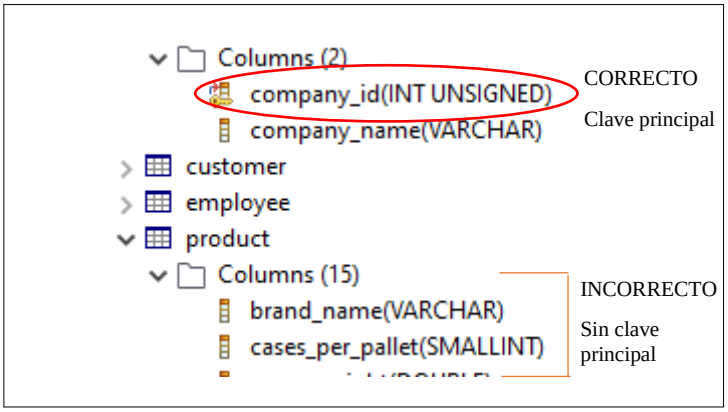


*Figura 7. Visualización de claves principales en Talend.*

| state_province | country |
|---|---|
| BC | <null> |
| BC | Canada |

*Figura 8. Muestra de valores distintos del país para la misma provincia.*

## 2. Recomendaciones

Es aconsejable añadir documentación respecto a los rangos de valores que deben de tomar las columnas, así como hacer uso de la cláusula *CHECK* para imposibilitar la introducción de valores fuera de los rangos definidos. Un ejemplo de uso de esta cláusula sería *ADD CONSTRAINT chkSalaryEmployee CHECK(salary ≥ 5000)*, que restringiría el salario anual de un empleado a un valor mayor o igual de 5000.

Se debe asignar siempre una clave principal, ya sea conformada por una columna o varias, a todas las tablas de las que se dispongan (ver Figura 9).

En todo caso, las claves foráneas deben apuntar a claves primarias.

Se deberían especificar, por medio de documentación, las dependencias funcionales a tener en cuenta para evitar inconsistencias.

```
DROP TABLE IF EXISTS `company`;
CREATE TABLE `company` (
  `company_id` int(10) unsigned NOT NULL AUTO_INCREMENT,
  `company_name` varchar(45) NOT NULL,
  PRIMARY KEY (`company_id`)
) ENGINE=InnoDB AUTO_INCREMENT=505 DEFAULT CHARSET=latin1;
```

*Figura 9. Ejemplo correcto de clave principal*

.

# INTERPRETABILIDAD

Se ha analizado la documentación y metadatos para saber si es suficiente para interpretar correctamente el significado y las propiedades de las fuentes de datos.

## 1. Deficiencias

No se ha proporcionado ningún tipo de documentación. Esto se trata de una deficiencia significativa que se debe tratar con urgencia, puesto que, en cualquier caso, la documentación es imprescindible.

En la tabla "**customer**", existen columnas que tienen un nombre muy inespecífico como "**address1**", "**address2**", "**address3**" y "**address4**" para el caso de la dirección (ver Figura 10). Resulta difícil entender a qué parte de la dirección corresponde cada uno. Ocurre lo mismo para la columna "**mi**" de esta misma tabla; y para las columnas "**wa_address1**", "**wa_address2**", "**wa_address3**" y "**wa_address4**" de la tabla "**warehouse**".

| | address1 | address2 | address3 | address4 |
|---|---|---|---|---|
| 1 | 2433 Bailey Road | road gred | high street | bt 01-v7 |
| 2 | 2433 Bailey Road | road gred | high street | bt 01-v7 |
| 3 | 2219 Dewing Ave... | <null> | <null> | <null> |
| 4 | 7640 First Ave. | <null> | <null> | <null> |
| 5 | 337 Tosca Way | <null> | <null> | <null> |
| 6 | 8668 Via Neruda | <null> | <null> | <null> |
| 7 | 1619 Stillman Co... | <null> | <null> | <null> |
| 8 | 2860 D Mt. Hood... | <null> | <null> | <null> |
| 9 | 6064 Brodia Court | <null> | <null> | <null> |
| 10 | 7560 Trees Drive | <null> | <null> | <null> |

*Figura 10. Muestra de las columnas "addressX" para customer.*

## 3. Recomendaciones

Se debe escribir una documentación adecuada, que explique correctamente las propiedades de los datos con los que se trabajan y sus respectivas restricciones, para poder recurrir a ella en caso de dudas. También debe mejorar nombres de columnas que resultan confusos o incompletos, así como añadir descripciones que los explique correctamente.

# ANEXO A. COMENTARIOS ADICIONALES

No se ha realizado el análisis de las dimensiones temporales y de las dimensiones de sincronización y accesibilidad ya que consideramos que no disponemos de la información necesaria para determinar si son correctas o no.

# ANEXO B. DATOS ADICIONALES

| | *address2* | *address3* | *address4* | *country* | *phone1* |
|---|---|---|---|---|---|
| *Valores nulos* | 95.27% | 99.94% | 99.94% | 3.08% | 3.08% |

*Figura 11. Muestra de valores nulos para la tabla "customer".*

| | *email* | *position_title* |
|---|---|---|
| *Valores nulos* | 8.66% | 2.77% |

*Figura 12. Muestra de valores nulos para la tabla "employee".*

| | *wh_address2* | *wh_address3* | *wh_address4* | *Warehouse_owner_name* |
|---|---|---|---|---|
| *Valores nulos* | 95.27% | 99.94% | 99.94% | 100% |

*Figura 13. Muestra de valores nulos para la tabla "warehouse".*

# ANEXO C. REFERENCIAS

Microsoft Corporation. (26 de Mayo de 2021). *Restricciones UNIQUE y restricciones CHECK*. Obtenido de https://docs.microsoft.com/es-es/sql/relational-databases/tables/unique-constraints-and-check-constraints?view=sql-server-ver15

Microsoft Corporation. (29 de Enero de 2022). *Restricciones entre claves principales y claves externas*. Obtenido de https://docs.microsoft.com/es-es/sql/relational-databases/tables/primary-and-foreign-key-constraints?view=sql-server-ver15

Roldán, R. C. (s.f.). *Restricciones de integridad*. Obtenido de http://gpd.sip.ucm.es/rafa/docencia/bdsi/apuntes/TEMA05.pdf

Talend S.A. (s.f.). *Talend Help Center*. Obtenido de https://help.talend.com/