Herramienta de Gestión de Calidad de Datos para procesos de ETL.

Informe Final

2 de junio de 2014

Valentina Ramos

Tutor:
Adriana Marotta

Instituto de Computación
Facultad de Ingeniería - Universidad de la República
Montevideo - Uruguay

Resumen

La calidad de los datos ha cobrado cada vez más importancia dentro de las organizaciones, siendo fundamental evaluar la calidad y encontrar las causas de la mala calidad. No siempre se encuentran las causas reales de los problemas de calidad de datos. En los últimos años han aumentado de manera notoria la cantidad de herramientas de calidad de datos que intentan encontrar las posibles causas de una mala calidad, prevenirla, y mejorarla.

Este trabajo realiza una evaluación de las herramientas de calidad de datos existentes buscando carencias en las mismas. Se evalúan tanto las herramientas comerciales (libre y no libres, gratuitas y no gratuitas), así como las herramientas desarrolladas en la academia. En esa evaluación surge la necesidad de clasificar las herramientas teniendo en cuenta distintos aspectos, por ejemplo si son específicas de calidad de datos o si están embebidas en una herramienta más general.

Esa evaluación sustenta y motiva el principal objetivo de este proyecto: desarrollar un prototipo de herramienta de gestión de calidad de datos más general, más completa, y con un enfoque riguroso. La herramienta es integrable en procesos de extracción, transformación y carga de datos, ya que se entiende que es de gran utilidad para la toma de decisiones en el momento de carga de datos. Además, la herramienta puede conectarse a un módulo de servicios que consume servicios de distintas herramientas de calidad de datos ya existentes.

Índice general

Capítulo 1

Introducción

La calidad de los datos se ha tornado de gran interés en los últimos años para cualquier organización de cualquier rama. Las consecuencias de una mala calidad son experimentadas a diario, sin embargo en general no se hace una conexión con las causas reales. Un ejemplo de esto, es el envío de cartas a través del correo postal, si la carta no llega en general se piensa que el correo no funciona adecuadamente. Si se hiciera un análisis más profundo de ese problema, arrojaría causas relacionadas a los datos, como pueden ser errores en la dirección originadas en la base de datos. La mala calidad de datos tiene serias consecuencias de gran trascendencia para la eficiencia y efectividad de las organizaciones. Por otro lado, en la Academia, en los últimos años se ha avanzado mucho en la investigación en el área de Calidad de Datos. [18][31]

Cada vez existen más herramientas de calidad de datos de distintos tipos, y el uso de las mismas dentro de las organizaciones está en aumento. Existen muchas herramientas comerciales, todas ellas atacando diferentes tareas de gestión de calidad de datos. A su vez, estas herramientas son de diferentes tipos, algunas son específicas para limpieza de datos, otras son para dominios específicos, etc.

Este gran interés en el área y la aparición constante de nuevas herramientas motivan la realización de un estudio de las herramientas existentes. De esta manera se puede tener una visión general de lo que existe hoy en día y las carencias que tienen estas herramientas. Dichas carencias motivan el desarrollo de una herramienta de calidad de datos más general, más completa, y con un enfoque riguroso que dé solidez a la gestión de la calidad de los datos.

1.1. Objetivos del proyecto

El objetivo general de este proyecto es desarrollar un prototipo de herramienta de calidad de datos integrable en procesos de ETL.

Para desarrollar este objetivo es necesario cumplir con los siguientes objetivos particulares:

- Realizar un estudio de las herramientas de calidad de datos existentes, para detectar fortalezas y debilidades.
- Proponer una forma de trabajo en la que se gestione la calidad de los datos en forma integrada al proceso ETL.

- Diseñar una herramienta de calidad de datos que pueda ser utilizada desde un entorno de ETL.
- Implementar un prototipo de dicha herramienta.
- Desarrollar un caso de estudio que muestre la aplicabilidad de la propuesta.

El estudio de las herramientas de calidad existentes debe considerar tanto las herramientas comerciales como las herramientas desarrolladas en la academia. Además es de interés clasificar las herramientas existentes según el tipo de tarea de calidad de datos en la que se focalizan, así como clasificarlas según el tipo de herramienta que son (si son específicas de calidad de datos, si están embebidas en herramientas más generales, etc.).

En cuanto al prototipo a desarrollar, es deseable que sea una herramienta para la medición de la calidad de datos que incluya un marco riguroso y estricto para la definición de las medidas a considerar.

1.2. Organización del informe

Este informe consta de cuatro capítulos, a continuación describiremos como está organizado el resto del documento. En el capítulo 2 se presentan los conceptos básicos necesarios para poder entender el trabajo. Se detallan en este los conceptos básicos de calidad de datos y los conceptos básicos en proceso ETL.

En el capítulo 3, se describen los trabajos existentes en evaluaciones de herramientas tanto comerciales como desarrolladas en la academia. Además se mencionan las herramientas desarrolladas en la academia con una breve descripción de la funcionalidad.

El capítulo 4, contiene la evaluación de herramientas realizada. Esta evaluación incluye una clasificación de las herramientas de calidad de datos, y un análisis general de las herramientas evaluadas.

En el capítulo 5, se detalla la solución propuesta, esto es: el análisis, el diseño y la implementación del prototipo desarrollado.

El capítulo 6, incluye un caso de estudio que muestra la aplicabilidad de la herramienta.

En el capítulo 7, se presenta las conclusiones del presente trabajo, y los posibles trabajos a futuro.

Por último se describe la bibliografía consultada para desarrollar este trabajo, y el anexo referente al manual de usuario de la misma.

Capítulo 2

Conceptos Básicos

En este capítulo se presentarán los principales conceptos básicos necesarios para una mejor comprensión del trabajo realizado. Se explicarán los conceptos básicos de calidad de datos y los conceptos básicos de los procesos ETL.

2.1. Conceptos Básicos en calidad de datos

A continuación se hará una descripción del significado y la importancia de la calidad de datos, así como un detalle del enfoque utilizado. Además se explicará que actividades conforman la gestión de la calidad de datos.

2.1.1. Calidad de datos y su importancia

En el contexto de este trabajo los datos serán vistos como una representación de objetos del mundo real que pueden ser almacenados, recuperados y elaborados por algún procedimiento de software. Actualmente los datos son de gran importancia, en el caso de las organizaciones estos pueden ser utilizados para la toma de decisiones. Otro claro ejemplo de uso, es internet, considerando a este como un gran almacenador de datos, que brinda información a sus usuarios.

Asimismo la calidad se verá como un conjunto de propiedades que hacen que algo se adecue al uso, logrando alcanzar o exceder las expectativas del consumidor. La calidad en sí también se ha tornado algo importante en todos los ámbitos, para ello se usan cada vez más las normas y certificaciones de calidad.

Se puede ajustar la calidad de datos a lo que los consumidores desean de los datos: que sean relevantes para su uso, correctos y sin inconsistencias, lo más actualizados posibles, de fácil acceso y que sean visibles de manera adecuada en las aplicaciones. La calidad de datos suele reducirse a exactitud de datos, sin embargo el primero es un concepto multifacético incluyendo distintas dimensiones, siendo exactitud una de ellas.

Una mala calidad en los datos puede traer aparejado severas y variadas consecuencias como ser económicas, sociales, de ineficiencia, etc. Un claro ejemplo es el que se muestra en un reporte de EEUU en el cual problemas de calidad en los datos conllevaron a errores en medicaciones, afectando a 1,5 millones de pacientes al año [5]. Existen muchísimos problemas de calidad de datos que

hacen que sean considerados de mala calidad, algunos ejemplos son: datos incorrectos, datos inconsistentes con la realidad, datos desactualizados y datos poco fiables debido a la fuente a la que pertenecen. A su vez, estos problemas se pueden generar en distintos momentos: en la producción, en el almacenamiento o en la utilización de los datos. En cada una de estas etapas se pueden generar problemas debido a distintas causas, por ejemplo en la etapa de producción se pueden generar datos incorrectos debido a que la información es ingresada manualmente.

2.1.2. Dimensiones de calidad de datos

A la hora de realizar un análisis de calidad de datos éste se debe basar en algún enfoque. El enfoque de calidad usado en este trabajo, caracteriza la calidad según varias dimensiones o atributos que ayudan a calificar los datos, y define una jerarquía de conceptos de calidad.

Una *dimensión* captura una faceta de la calidad y puede verse como un agrupamiento de *factores* de calidad que tienen el mismo propósito. Un *factor* representa un aspecto particular de una dimensión. Un mismo factor de calidad puede medirse con diferentes *métricas*. Una *métrica* es un instrumento que define la forma de medir un factor de calidad. Una misma métrica puede ser medida por diferentes *métodos*. Un *método* es un proceso que implementa una métrica. Como se puede ver en la figura 2.1.1 este enfoque genera una jerarquía entre estos conceptos.

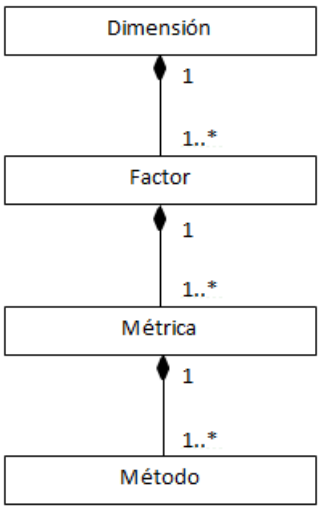

Figura 2.1.1: Jerarquía de conceptos

A continuación se detallan las principales dimensiones de calidad consensuadas en la bibliografía [28][18][33][29][22][17][30], con sus factores correspondientes.

2.1.2.1. Dimensión exactitud

La dimensión *exactitud* indica qué tan precisos, válidos y libres de errores son los datos. Algunos de los factores de esta dimensión son *correctitud semántica*, *correctitud sintáctica* y *precisión*. La *correctitud semántica* mide la correspondencia entre los valores reales y los valores del sistema de información (SI de ahora en más). En general, para realizar esta medición se hace una comparación contra un referencial considerado como válido u otra base de datos debido al alto costo de la comparación del SI con el mundo real. Con la *correctitud sintáctica* interesa medir si los valores del SI corresponden a valores válidos del dominio, sin importar si los mismos son valores reales o no. Para verificar la correctitud sintáctica normalmente se utilizan diccionarios, los cuales contienen una lista de valores para un dominio determinado. En general tiene un gran costo validar la correctitud semántica ya que implica comparar contra el mundo real, mientras que validar la correctitud sintáctica suele ser algo de menor costo. El factor *precisión* mide que tan detallados son los datos del SI.

2.1.2.2. Dimensión completitud

La dimensión *completitud* indica si el SI contiene toda la información de interés. Asociados a esta dimensión tenemos los factores: *densidad* y *cobertura*. La *densidad* mide cuánta información se tiene sobre las entidades del SI. La *cobertura* mide la porción de datos de la realidad contenidos en el SI, este factor requiere de una comparación del SI con el mundo real, lo cual suele ser muy costoso, con lo cual como alternativa se podría estimar el tamaño de un posible referencial.

2.1.2.3. Dimensión frescura

La dimensión *frescura* refiere a qué tan viejos son los datos, ésta involucra la perspectiva temporal de los datos. Factores asociados a esta dimensión son *edad* y *actualidad*. El factor *edad* mide qué tan viejos son los datos, esto es el tiempo que hace que fueron creados o modificados. Mientras tanto, el factor *actualidad* mide que tan vigentes son los datos en el SI, esto es el tiempo desde que se hizo la extracción de una fuente de datos hasta el momento en que se realiza una consulta en nuestro SI.

2.1.2.4. Dimensión consistencia

La dimensión *consistencia* captura la satisfacción de reglas semánticas definidas sobre los datos, tanto reglas de integridad de una base de datos como reglas definidas por los usuarios. Los factores de esta dimensión son *integridad de dominio*, *integridad intra-relación* e *integridad referencial*. En todos los casos de los factores lo que interesa medir es qué tan bien se satisfacen las reglas de integridad.

2.1.2.5. Dimensión unicidad

Por último, la dimensión *unicidad* indica el nivel de duplicación entre los datos. Factores de esta dimensión son *duplicación* y *contradicción*. En el primer

caso se mide la cantidad de repetidos y en el segundo caso se mide la cantidad de repetidos donde hay conflicto entre los datos de una misma entidad.

2.1.3. Gestión de calidad

La gestión de la calidad en un SI abarca varias tareas, como se puede ver en la figura 2.1.2. En primer lugar se deben analizar los procesos de negocio involucrados, lo cual nos permite conocer el dominio del sistema de información. Posteriormente hay una etapa de análisis de datos (Data Profiling), la cual es un primer acercamiento con la calidad de nuestro SI, en esta etapa se hacen estimaciones de la calidad del mismo. Una vez finalizado esto se puede definir el modelo de calidad de datos a utilizar, esto es las dimensiones involucradas, los factores a tomar en cuenta y las métricas para los mismos. Como paso siguiente se pueden realizar mediciones de calidad y tomando en cuenta el análisis hecho de negocio se puede detectar las posibles causas de mala calidad. Luego de las mediciones de calidad se puede aplicar limpieza de datos en base a ésta, así como pensar una restructuración del sistema. Asimismo es importante siempre monitorear la calidad de nuestro sistema tomando en cuenta el modelo de calidad y realizando mediciones del mismo.

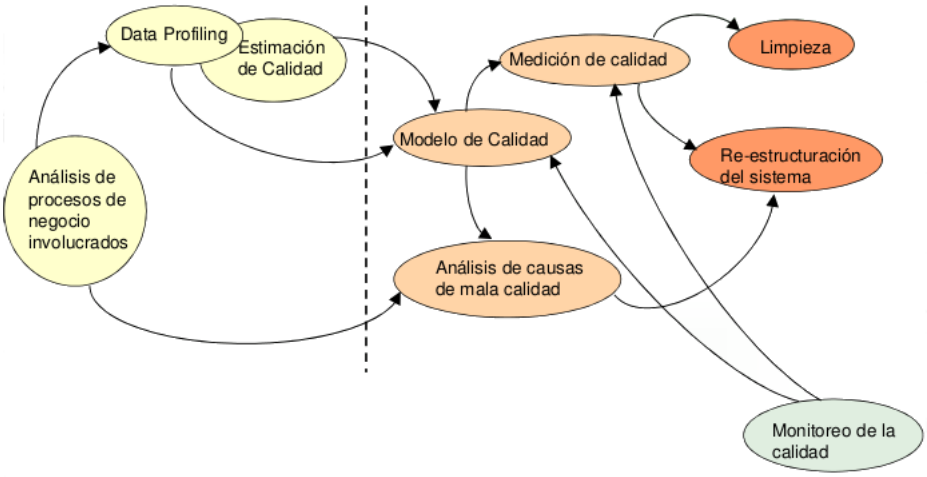

Figura 2.1.2: Gestión de calidad

2.1.3.1. Análisis de datos y estimación de calidad (Data Profiling)

El análisis de los datos se hace con el fin de poder estimar la calidad de un SI, en general basándose en estadísticas. Es muy importante porque da un panorama de los posibles problemas de calidad existentes en el SI, y en base a esta información se pueden evaluar medidas a tomar. Por ejemplo, en esta etapa se puede ver la cantidad de datos con formato incorrecto o sin información (o nulos) en cierto tiempo, si este campo es importante para el SI en cuestión se harán en una etapa posterior mediciones de calidad más específicas sobre el mismo.

2.1.3.2. Modelo de Calidad de Datos

Con el fin de lograr un enfoque más riguroso para la medición se define el Modelo de Calidad. Una vez hechas las estimaciones de calidad que nos interesa para el SI en cuestión, debemos definir el Modelo de Calidad que aplica a este dominio. Se definen entonces, las dimensiones a aplicar, los factores y las métricas (especificadas sobre los datos particulares), con el fin de poder tomar mediciones. El Modelo de Calidad da un marco riguroso, lo cual es importante ya que los números por sí solos no dicen nada. Se debe saber qué se mide, con qué granularidad, cuál es el rango de valores esperado, lo cual se especifica en la definición de las métricas. Este marco riguroso no solamente soporta la medición de calidad, sino que también guía la limpieza de datos y el análisis de causas de una mala calidad en el sistema.

2.1.3.3. Medición de calidad de datos

Una vez que se define el Modelo de Calidad se elige algún método para medir las métricas asociadas a los factores elegidos. Las mediciones hechas son almacenadas en una base de Metadatos de Calidad, con el fin de analizar los resultados obtenidos.

2.1.3.4. Análisis de causas de mala calidad

El análisis de las causas de la mala calidad es muy importante ya que se pueden detectar errores sistemáticos que se estén haciendo al procesar los datos, o al usar el sistema. Para detectar las causas se debe tener en cuenta los procesos de negocio involucrados. Este análisis ayuda a pensar una posible reestructuración del sistema en el cual se mejorarían los pasos que empeoran la calidad del sistema.

2.1.3.5. Limpieza de datos (Data cleansing)

Una vez que se hicieron las mediciones de calidad correspondientes se puede definir qué limpiezas se realizarán en los datos con el fin de mejorar la calidad del sistema.

2.1.3.6. Re-estructuración del sistema

La reestructuración del sistema debe tenerse en cuenta cuando tanto el análisis de las causas de mala calidad y las mediciones de calidad realizadas detectan problemas en el sistema que ocurren sistemáticamente.

2.1.3.7. Monitoreo de la calidad de datos

La calidad de un sistema debe ser monitoreada permanentemente, tomando mediciones de calidad y tal vez redefiniendo nuestro modelo de calidad.

2.2. Conceptos Básicos en procesos ETL

A continuación se detallarán los conceptos básicos en procesos ETL tanto en Data Warehouse como en otros casos.

2.2.1. Procesos de ETL en Data Warehouse

Uno de los usos más comunes de los procesos de ETL es para la carga de los sistemas de Data Warehouse.

Un Data Warehouse es una base para el procesamiento de la información, este contiene datos integrados, históricos y con distintas granularidades. Los Data Warehouse son utilizados con el fin de tener una visión global de la empresa, analizando los datos correspondientes. Con distintas granularidades se pueden obtener distintas visiones de los mismos datos, por ejemplo si se quiere una visión financiera se van a ver los datos de una manera distinta que cuando se quiere una visión de marketing. Los datos son los mismos, lo que cambia es la granularidad con la que se analizan los mismos. Otra gran ventaja que tienen estos sistemas es que almacenan datos históricos, con lo cual se puede analizar la evolución de los datos a lo largo del tiempo. Algunos de los desafíos de la construcción de un Data Warehouse son la integración de datos provenientes de fuentes heterogéneas, el cálculo de indicadores complejos y el manejo de grandes volúmenes de datos.

En los procesos ETL (Extraction Transform and Load), se convierten los datos extraídos de las fuentes en datos corporativos del Data Warehouse. Una manera de ver este proceso es descomponiéndolo en las actividades que este involucra:

- **Extraer (Extraction):** Es el proceso de extraer datos de una o más fuentes, pueden existir diversas fuentes de datos. Por ejemplo, extraer todos los registros de clientes que se agregaron o cambiaron después de la última fecha de extracción.
- **Transformar (Transform):** Cambiar la forma y el contenido de los datos para adaptarlos a la estructura del Data Warehouse a cargar. Por ejemplo, buscar por nombre del país para obtener la clave numérica.
- **Cargar (Load):** Almacenar los datos en el Data Warehouse objetivo.

En la transformación de esos datos se pueden realizar muchas operaciones como ser integrar datos entre distintas fuentes, verificar el dominio de los datos, agregar información a los datos, eliminar datos redundantes, etc. Es en esta etapa donde se validan los datos y se limpian (Data cleansing).[19, 26]

2.2.2. Procesos de ETL con otros usos

Los procesos ETL no son solo usados para la carga de un Data Warehouse, son usados también para cargar datos en cualquier SI. Estos son de gran utilidad ya que tienen muchas componentes para limpieza de datos.

Capítulo 3

Trabajos existentes en evaluación de herramientas de calidad de datos

Como punto de partida para evaluar las herramientas de calidad de datos existentes se analizaron algunos trabajos de evaluación previos. Se tuvieron en cuenta tanto herramientas desarrolladas en la industria como herramientas desarrolladas en la academia, con el fin de tener un panorama lo más amplio posible. A continuación se detallan las evaluaciones encontradas.

3.1. Evaluaciones de herramientas desarrolladas en la industria

Se tuvieron en cuenta dos evaluaciones hechas sobre herramientas desarrolladas en la industria. Una de las evaluaciones es un artículo [24] que propone una metodología para evaluar estas herramientas y realiza la evaluación. La otra evaluación tomada en cuenta es la realizada por la consultora Gartner [23].

El artículo mencionado [24] propone una metodología para evaluar las herramientas de calidad de datos aplicable en un dominio de bases de datos CRM ("Customer Relationship Management") con muchas fuentes y grandes cantidades de datos. En particular se usa una empresa de electricidad como aplicación directa. La idea propuesta es evaluar la adecuación de las herramientas de calidad de datos a ciertas tareas y comparar entre diferentes herramientas usándolas para una misma tarea. Para esto se definen tres dimensiones a las que se le debe asignar una puntuación: el nivel de disponibilidad (mide si un determinado criterio puede ser evaluado con la herramienta), el nivel de facilidad para evaluar un criterio, y el nivel de la salida de calidad. Asimismo como no todas las tareas requieren las mismas funcionalidades, se define una manera de obtener una puntuación usando ponderaciones en las tres dimensiones mencionadas. El artículo propone los criterios a tener en cuenta en la evaluación: criterios generales, normalización de direcciones, duplicación, reglas de consistencia, archivos referencia, reportes, integración, y estadísticas. Dentro de cada criterio se proponen varias características a analizar. Para poder medir estos cri-

terios se tienen que definir ciertos indicadores o métricas específicos al proyecto. La medición del nivel de la salida de calidad se hace usando una matriz de resultados esperados y obtenidos. En ésta se ingresan los resultados obtenidos con la herramienta de calidad de datos (datos correctos e incorrectos) y los esperados para un indicador a medir. Usando la matriz mencionada se calcula una puntuación para ese indicador, que sería la medición del nivel de la salida de calidad.

En la comparación hecha por la consultora Gartner [23] se comparan varios proveedores de herramientas de calidad de datos. En esta comparación se dejan afuera algunas herramientas de calidad de datos, el criterio utilizado para esto es en base a lo que ofrecen las mismas. Como mínimo para ser incluidas deben ofrecer las siguientes funcionalidades: análisis de datos y visualización, parseo de datos, matcheo de datos, estandarización y limpieza, y monitoreo. Asimismo se evalúa la capacidad de los proveedores de estar presentes en el mercado de herramientas de calidad. Esto se hace teniendo en cuenta el servicio o producto que ofrecen, la fortaleza económica del proveedor, el precio y la capacidad de vender, cuestiones de marketing y la experiencia del usuario. Además se analiza la completitud de la visión, esto es la manera que tienen los proveedores de cumplir en cierto grado con el liderazgo ante nuevas direcciones del mercado, la innovación, las necesidades de los clientes, y fuerzas competitivas. En esta evaluación para cada conjunto de herramientas se detallan las fortalezas y las precauciones a tener en cuenta.

Como podemos ver son evaluaciones muy distintas las que se hacen en el artículo [24] y en la consultora Gartner: el artículo tiene un enfoque más riguroso ya que propone un método de evaluación en base a ciertos criterios, mientras que en el estudio de la consultora Gartner se hace más hincapié en el proveedor en sí, considerando por ejemplo cuestiones de marketing. La consultora Gartner realiza un análisis de mercado para las herramientas de calidad de datos, analizando todas las existentes (que entren en el criterio) y catalogándolas. Sin embargo, también se hace un análisis de las prestaciones de calidad que ofrecen a la hora de incorporar o dejar de lado herramientas en el estudio. En el artículo, donde se propone un framework para analizar herramientas, se hace un análisis más a fondo de las prestaciones de calidad que brinda, definiendo criterios y características dentro de estos a considerar.

Los proveedores tenidos en cuenta en la evaluación de la consultora Gartner son: Ataccama, Datactics, DataMentors, Human Inference, IBM, Informatica, Information Builders/iWay, Innovative Systems, Oracle, Pitney Bowes Software, RedPoint (DataLever), SAP, SAS/DataFlux, Talend, Trillium Software, y Uniserv.

Mientras que en el artículo [24] se evaluó una herramienta del proveedor Informatica (Data Quality), además al momento de la escritura del mismo estaba en progreso la evaluación del proveedor SAS DataFlux y se preveía que el próximo a evaluar fuera IBM.

3.2. Evaluaciones de herramientas desarrolladas en la academia

Existen algunas comparaciones hechas de las herramientas existentes en el ámbito académico. Para esto se consideraron dos artículos con el fin de lograr un acercamiento a las mismas y lo que brindan.

La herramienta para limpieza de datos "AJAX" [27, 32], la cual es extensible y flexible, tiene como objetivo separar el nivel lógico del físico en cuanto a la limpieza de datos. Se entiende por nivel lógico el diseño del flujo de limpieza de datos y la especificación de las operaciones de limpieza a realizar. Por otra parte, nivel físico se le llama a la implementación. La idea de esta herramienta es ir haciendo transformaciones en los datos y que en ese proceso se vayan limpiando los mismos. El proceso de limpieza se especifica organizando las transformaciones en un gráfico de flujo de datos lineal, donde la salida de una transformación es la entrada de la siguiente. Aparentemente este proyecto esta discontinuado.

Otro ejemplo de herramienta de limpieza de datos basada en lenguaje declarativo es "FraQL" [27, 32]. El lenguaje utilizado es una extensión de SQL basado en un modelo orientado a objetos. Soporta la especificación de las transformaciones a realizar tanto a nivel de esquema como a nivel de datos, el usuario puede definir sus propias funciones para estandarizar y normalizar valores. Soporta la detección y eliminación de duplicados, así mismo tiene un método para resolver las contradicciones. También tiene la funcionalidad de rellenar valores incompletos en los datos, y eliminación de tuplas invalidas.

"Potter's Wheel" [27, 32] es una herramienta interactiva de limpieza de datos que integra transformación de datos y detección de errores usando una interfaz del estilo de una planilla de cálculo. Se le permite al usuario definir su dominio y algoritmos para que se cumplan las restricciones del dominio. También permite que el usuario muestre los resultados que quiere con ejemplos de datos y de ahí se infiere el domino sin que se tenga que especificar con anterioridad. El dominio que se deduce puede ser usado para detectar discrepancias en los datos. La especificación del proceso de limpieza se hace interactivamente, los resultados inmediatos que se obtienen le permiten al usuario ir refinando el proceso de limpieza.

"ARKTOS" [27, 32] es una herramienta de extracción, transformación y carga de datos (ETL), los autores consideran que la limpieza de datos se hace en el mismo proceso de ETL. Se especifica un metamodelo permitiendo modelar el proceso completo de ETL. Las operaciones de limpieza se declaran en sentencias SQL, cada sentencia se asocia con un error particular y una política que especifica el comportamiento en caso de error. Seis tipos de errores pueden ser detectados y tratados en esta herramienta: violaciones de claves primarias, violaciones de unicidad, violaciones de referencias, existencia de nulos, no concordancia con el dominio, y no concordancia con el formato. Las posibles políticas en caso de error son: ignorar, borrar, escribir a un archivo, o insertar en una tabla. El éxito de la limpieza se puede medir para cada operación de limpieza ejecutando una sentencia que cuente la cantidad de tuplas que violan las reglas de limpieza.

"IntelliClean" [27, 32] es una herramienta basada en reglas con foco en la eliminación de duplicados, consiste en tres pasos. En el paso de pre-procesamiento se eliminan errores sintácticos, los valores son estandarizados según un formato y se chequea la consistencia de las abreviaciones utilizadas. El paso de proce-

samiento representa la evaluación de las reglas de limpieza sobre ítems de condiciones que especifican medidas a tomar bajo ciertas circunstancias. Hay cuatro tipos distintos de reglas. Reglas de identificación de duplicados que especifican bajo qué condiciones las tuplas son consideradas como duplicadas. Reglas de unión de datos que especifican cómo van a ser tratadas las tuplas duplicadas. Reglas de actualización que especifican cómo van a ser actualizadas las tuplas en situaciones particulares. Por último las reglas de alertas que especifican cuándo el usuario va a ser notificado. En los dos primeros pasos se registran las acciones tomadas, para en el último paso (paso de verificación y validación) ser analizadas por un humano y posiblemente corregir las acciones a tomar.

En el artículo [27] se hace una comparación de estas herramientas mencionadas teniendo en cuenta las anomalías detectadas por las mismas. Además se detalla una breve indicación de los métodos y técnicas usadas para definir cada una de esas anomalías. Las anomalías que se tienen en cuenta son: errores lexicográficos, errores del formato de dominio, irregularidades, violación de restricciones, valores no considerados, tuplas no consideradas, duplicados, y tuplas inválidas.

3.3. Conclusiones

Las evaluaciones sobre herramientas de la academia estudiadas nos dicen que los desarrollos en la academia de herramientas de calidad de datos son bastante puntuales, algunos solo cubren algunos aspectos del proceso de aseguramiento de la calidad, como por ejemplo las herramientas que solo son para limpieza de datos. Por otro lado, muchos de estos proyectos son solo un prototipo o fueron discontinuados.

En cuanto a las evaluaciones encontradas de las herramientas comerciales de calidad de datos son bastante vagas. Si bien la consultora Gartner realiza un estudio muy detallado de cada conjunto de herramientas que brinda un proveedor y realiza una comparación de cada proveedor, es enfocado en el marketing de dicho proveedor. De todas maneras es un buen punto de partida para tener una rápida visualización de todo lo que existe en el mercado y las virtudes y carencias que puede tener cada proveedor. A su vez el artículo estudiado [24] si bien es una propuesta de evaluación de las herramientas la misma está enfocada en un dominio específico. Además esa evaluación se hace en base a los resultados que debería obtener y lo que obtengo con la herramienta, con lo cual es bastante subjetiva al dominio con el cual se prueben las herramientas. Asimismo en este artículo no se revelan los datos obtenidos para las herramientas analizadas ya que son confidenciales. Igualmente sirve como puntapié inicial a la hora de pensar una evaluación de herramientas de calidad.

Para lograr un mejor análisis de las herramientas de calidad de datos no tan enfocado desde el ámbito del marketing y que no sea basado en un dominio específico, se hace necesario hacer una evaluación propia de herramientas focalizada fuertemente en sus funcionalidades y alcance.

Capítulo 4

Evaluación de herramientas de calidad de datos

Con la motivación de realizar una evaluación de las herramientas de calidad de datos no tan enfocada en el marketing, y que a su vez sea lo más general posible, surge esta evaluación de herramientas comerciales. A continuación se detalla la evaluación realizada.

4.1. Clasificación de herramientas

Hoy en día existe un creciente uso de herramientas para analizar, limpiar y monitorear los datos de las empresas, de la misma manera existe un crecimiento en las ofertas de herramientas con estos fines. Dentro de la variada gama existente podemos clasificar las herramientas según cuan abarcadoras del proceso de aseguramiento de calidad son. Existen herramientas que no son específicas de calidad de datos pero atacan algunos problemas de calidad de datos permitiendo aplicar reglas de limpieza sobre estos, ejemplos de estas son algunas herramientas de ETL. Otras herramientas que en general son una familia de herramientas independientes (o en algunos casos están integradas en una sola) son más amplias, en el sentido que abarcan más etapas de la gestión de la calidad, esto es análisis de datos (data profiling), limpieza y enriquecimiento de los datos, y monitoreo de la calidad. Asimismo existen herramientas para calidad de datos que permiten definir reglas de limpieza pero que son específicas para algunos tipos de datos, como ser direcciones, teléfonos, mail, etc. Por otra parte hay herramientas que atacan solo algunas de las tareas de gestión de calidad de datos, como pueden ser herramientas de análisis de datos.

Entonces, a efectos de proponer una clasificación se definen los siguientes tipos de herramientas que abordan el problema de calidad de datos:

1. *Embebidas en herramientas más generales:*

   Dentro de este grupo de herramientas se encuentran las que no son específicas de calidad de datos, pero atacan algunos problemas de ésta, como por ejemplo las herramientas de ETL.

2. *Tratamiento integral de la gestión de la calidad:*

Estas son en general familias de herramientas las cuales abarcan varias etapas de la gestión de calidad, realizando un tratamiento integral de la gestión de calidad.

3. *Generales para algunas de las tareas de gestión de calidad de datos:*

   Son herramientas que atacan solo algunas tareas de la gestión de calidad de datos. Ejemplos de estas son herramientas que solo realizan análisis de datos.

4. *De limpieza para un dominio específico:*

   Herramientas de limpieza aplicables solo para algunos dominios específicos, por ejemplo herramientas que solo atacan limpiezas de direcciones, mail, teléfonos, etc.

Esta clasificación propuesta es aplicable a nivel de familia de herramientas, ya que en muchos casos son varias herramientas que en su conjunto atacan el problema de calidad de datos, y se sugiere usarlas en conjunto.

Otra posible clasificación, ya a nivel de herramienta es según el tipo de tarea de gestión de calidad de datos en la que se focaliza:

1. *Análisis de datos:*

   Herramientas que analizan los datos y sacan estadísticas de los mismos, viendo por ejemplo duplicados, valores nulos, entre otros.

2. Limpieza de datos:

   En este tipo entran las herramientas que se focalizan en limpiar los datos una vez detectados los problemas de calidad, ya sea de dominios específicos o más generales.

3. Monitoreo de la calidad:

   Son herramientas que se concentran en monitorear la calidad de los datos.

En la Tabla 4.1 se muestra para cada proveedor las familias de herramientas que éste ofrece. Llamamos familia de herramientas a un conjunto de herramientas que están pensadas para ser usadas juntas. Cada familia de herramientas se clasifica según cuan abarcadoras son del proceso de gestión de calidad (Clasificación 1 en la tabla). Dentro de cada familia de herramientas se encuentran las herramientas específicas, las cuales se clasifican según qué tareas de la gestión de calidad atacan (Clasificación 2 en la tabla).

Proveedor	Familia de herramientas	Clasificación 1	Nombre de herramienta	Clasificación 2
Oracle	Oracle Enterprise Data Quality	Tratamiento integral de la gestión de calidad	Enterprise Data Quality Profile and Audit	Análisis de datos
			Oracle Enterprise Data Quality Parsing and Standardization	Limpieza de datos
			Oracle Enterprise Data Quality Match and Merge	Limpieza de datos
	Oracle Data Profiling and Data Quality for Data Integrator	Generales para algunas de las tareas de gestión de calidad de datos	Oracle Data Profiling	Análisis de datos
			Oracle Data Quality	Limpieza de datos
IBM		Tratamiento integral de la gestión de calidad	InfoSphere Information Analyzer	Análisis de datos
			InfoSphere Data Stage and QualityStage	Limpieza de datos
Microsoft	Data Quality Services	Generales para algunas de las tareas de gestión de calidad de datos	Data Quality Services	Análisis y limpieza de datos
Human Inference	Easy Data Quality	De limpieza para un dominio especifico	Easy Data Quality (para Pentaho)	Limpieza de datos
Talend	Talend Open Studio for Data Management	Embebidas en herramientas más generales	Talend Open Studio for Data Management	Análisis, limpieza de datos y monitoreo de la calidad
	Talend Open Studio for Data Quality	Generales para algunas de las tareas de gestión de calidad de datos	Talend Open Studio for Data Quality	Análisis de datos
	Talend Enterprise Data Quality	Tratamiento integral de la gestión de calidad	Talend Enterprise Data Quality	Análisis y limpieza de datos, y monitoreo de la calidad
	Talend Open Studio for Data Integration	Embebidas en herramientas más generales	Talend Open Studio for Data Integration	Limpieza de datos
Melissa Data	Contact Zone	De limpieza para un dominio específico	Contact Zone	Limpieza de datos
Pentaho	Pentaho Data Integration	Embebidas en herramientas más generales	Pentaho Data Integration	Limpieza de datos
Human Inference	Data Cleaner		Data Cleaner	Análisis de datos

Tabla 4.1: Clasificación de herramientas

En la tabla 4.2 podemos ver qué dimensiones de calidad ataca cada herramienta, y a su vez a qué tipo de campos se aplican estas. En el caso de algunas herramientas, se aplican para cualquier tipo de campo, mientras que otras son más específicas y se aplican solo para algunos. Este análisis se pudo realizar para la mayoría de las herramientas, sin embargo algunas herramientas no están en él ya que no se contaba con la información necesaria.

Proveedor	Nombre herramienta	Dimensiones	Dimensión aplicada a
Oracle	Oracle Data Quality	Exactitud (Correctitud semántica y sintáctica)	nombres y direcciones
		Correctitud semántica	códigos postales y direcciones
		Frescura (actualidad)	Cualquier campo que se tenga una referencia anterior
		Consistencia (integridad de dominio, integridad intra-relación, integridad referencial)	cualquier campo
IBM	InfoSphere Data Stage and QualityStage	Unicidad (Duplicación)	cualquier campo
		Exactitud (Correctitud semántica y sintáctica)	cualquier campo que se tenga referencia
Microsoft	Data Quality Services	Unicidad (Duplicación)	cualquier campo
		Exactitud (Correctitud semántica y sintáctica)	cualquier campo que se tenga referencia
Human Inference	Easy Data Quality	Exactitud (Correctitud semántica y sintáctica)	Direcciones
		Exactitud (Correctitud semántica y sintáctica)	Nombres
		Exactitud (Correctitud semántica y sintáctica), Completitud (Densidad)	Teléfonos
		Exactitud (Correctitud semántica y sintáctica)	Mail
		Unicidad (Duplicación)	para cualquier registro
Talend	Talend Open Studio for Data Quality	Unicidad (Duplicación), Exactitud (Correctitud sintáctica)	
		Consistencia (integridad de dominio, integridad intra-relación, integridad referencial)	cualquier campo
	Talend Open Studio for Data Integration	Unicidad (Duplicación)	cualquier campo
		Exactitud (Correctitud sintáctica)	cualquier campo
		Consistencia (integridad de dominio, integridad intra-relación, integridad referencial)	cualquier campo
Melissa Data	Contact Zone	Exactitud (Correctitud semántica y sintáctica)	nombres, direcciones, mail, teléfonos
		Unicidad (Duplicación)	direcciones
Pentaho	Pentaho Data Integration	Unicidad (Duplicación)	cualquier campo
		Exactitud (Correctitud sintáctica)	cualquier campo
		Consistencia (integridad de dominio, integridad intra-relación, integridad referencial)	cualquier campo

Tabla 4.2: Dimensiones atacadas por las herramientas

4.2. Análisis general de las principales herramientas consideradas

A continuación comentamos en detalle algunas herramientas de algunos proveedores.

Oracle ofrece por un lado una familia de productos empresariales de calidad de datos: "Oracle Enterprise Data Quality Profile and Audit", "Oracle Enterprise Data Quality Parsing and Standardization", "Oracle Enterprise Data Quality Match and Merge", "Oracle Enterprise Data Quality Product Data Parsing and Standardization", "Oracle Enterprise Data Quality Product Data Match and Merge" [10]. Las tres primeras herramientas son orientadas para la calidad de datos en general ya sea de compradores, empleados, productos, etc. Mientras que las últimas dos herramientas son específicas para la calidad de los productos de las empresas. La herramienta "Oracle Enterprise Data Quality Profile and Audit" debe ser usada como el primer paso de análisis de datos para definir las métricas de calidad claves, detectar datos faltantes, valores incorrectos, duplicados e inconsistencias. Los resultados de estos análisis son presentados en un tablero donde se puede monitorear y revisar la calidad contra las métricas definidas. Como segundo paso se puede utilizar la herramienta "Oracle Enterprise Data Quality Parsing and Standardization" la cual sirve para buscar y estandarizar datos, como ser crear un atributo a partir de varios campos, etc. La herramienta "Oracle Enterprise Data Quality Match and Merge" sirve para detectar campos iguales, permitiendo detectar duplicados e integrar esos datos. Mientras que el producto "Oracle Enterprise Data Quality Product Data Parsing and Standardization" permite estandarizar la información de los productos clasificándoles en categorías y estandarizar los atributos y descripciones, el producto "Oracle Enterprise Data Quality Product Data Match and Merge" permite detectar registros duplicados y definir con cual quedarse. Por otra parte Oracle tiene dos herramientas más: "Oracle Data Profiling" y "Oracle Data Quality", las cuales son herramientas de distribución gratuita. La primer herramienta sirve para analizar los datos y sacar estadísticas de los mismos, incluyendo Time Series que sirve para evaluar y monitorear los datos a través del tiempo. En ella se pueden hacer análisis de los datos personalizados. A nivel de campo se pueden ver la cantidad de nulos, cantidad de repetidos, patrones de los datos, mínimo y máximo cuando corresponda, longitud máxima y mínima de un campo, dependencias entre campos, claves. A nivel de fila se puede ver la fila mínima, la fila máxima, filas duplicadas y también posibles claves. La segunda herramienta, como bien se puede visualizar en la tabla de clasificación de herramientas 4.1, es específica para una tarea de la gestión de calidad, limpieza de datos. A su vez en la tabla que se clasifican las herramientas según las dimensiones de calidad 4.2 se pueden ver algunas dimensiones asociadas a ciertos componentes.

Por su parte IBM ofrece dos herramientas: "InfoSphere Information Analyzer", "InfoSphere QualityStage". El producto "InfoSphere Information Analyzer" sirve para la primera instancia del proceso de aseguramiento de calidad, que es el análisis de los datos. Permite analizar columnas, tablas y así como análisis cruzado con otras tablas. A nivel de columnas en esta herramienta se puede definir umbrales para detectar si una columna contiene valores nulos, si contiene duplicados, o si contiene valores constantes, así mismo a nivel de columna per-

mite inferir el tipo de información que contiene cada campo (por ejemplo si son direcciones, tarjetas de crédito, etc.). También a nivel de tabla permite definir un umbral para detectar presencia de clave primaria, además de poder definir qué tamaños de muestra se van a usar entre otras opciones. A nivel de tablas cruzadas permite detectar campos que tengan el mismo dominio.[4] Para atacar el problema de limpieza de datos se tiene el producto "InfoSphere Quality Stage" el cual puede ser usado en conjunto con el producto "Infosphere DataStage" que es una herramienta para integración de datos. Se proponen 4 pasos: entender los objetivos de la organización y como ellos proponen los requerimientos, comprender y analizar la naturaleza de los datos de las fuentes, diseño y construcción de los trabajos que van a realizar la limpieza, y por último evaluar los resultados. Para la limpieza de datos, se proponen tres pasos: estandarizarlos (correcciones de formatos), encontrar datos en otras fuentes para chequear su veracidad, y finalmente decidir qué datos sobreviven. Esta herramienta sirve también para la detección de duplicados, análisis de la correctitud sintáctica y semántica, las reglas pueden ser aplicadas a nivel de campo. Asimismo se puede enriquecer los datos con información de otras fuentes.

El proveedor Talend ofrece varias opciones, tanto herramientas gratuitas como herramientas pagas [16]. La herramienta gratuita que ofrece se llama "Talend Open Studio for Data Quality". Esta herramienta es sólo para el análisis de datos, en la misma se pueden realizar varios análisis y a distintos niveles. Nos permite hacer análisis a nivel de columnas pudiendo sacar estadísticas de valores nulos, duplicados, únicos, vacíos, y por defecto. También se pueden hacer análisis a nivel de tabla, viendo las mismas estadísticas que para las columnas pero con un conjunto de éstas. Además a nivel de tabla se pueden definir reglas para detectar anomalías, viendo las dependencias funcionales, esto es ver qué columnas determinan a otras (por ejemplo con campos como ciudad y país). Asimismo ofrece un análisis de tabla definiendo reglas de negocio con SQL. Esta misma herramienta permite hacer análisis de redundancia, comparando columnas iguales en distintas tablas, o viendo si coinciden las claves foráneas de una tabla con las claves primarias de otra tabla y viceversa. Por otra parte Talend tiene su herramienta paga "Talend Open Studio for Data Management" que es una plataforma que ya incluye todo: análisis de datos, limpieza, estandarización, parseo y matcheo de datos). Esta herramienta incluye las mismas funcionalidades de análisis de datos que "Talend Open Studio for Data Quality" con alguna opción más, como guardar el histórico de los análisis, generar reportes, entre otros. Además incluye limpieza y reportes.

A su vez, el proveedor Talend cuenta con la herramienta "Talend Open Studio for Data Integration". Si bien esta herramienta es de ETL, la misma puede ser utilizada para realizar limpieza de datos sobre algunas dimensiones (tal como se detalla en la Tabla 4.2)

Melissa Data provee varias soluciones para distintas herramientas de ETL [7]. Provee "Contact Zone" que ya viene integrado con "Pentaho Data Integration", la herramienta de ETL de Pentaho. La misma nos ofrece algunos componentes extras de calidad de datos para agregarlos al ETL. Uno de los componentes es "Contact Verify" el cual hace chequeos sobre campos de nombres, direcciones, mail, teléfono. Además este componente enriquece los datos agregando información, por ejemplo para las direcciones agrega la latitud y longitud, para los teléfonos los define en un cierto formato, etc. Para hacer los chequeos de los

campos se puede hacer localmente bajando los archivos de datos que se van a usar como referencia. Otra opción es conectarse al *cloud* de Melissa Data. Se cuenta también con la posibilidad de extraer reportes en base a estos análisis. Otro componente es el "MD Ip Locator" que dada una IP me devuelve la dirección de donde se encuentra. "MD Matchup" es un componente que sirve para unir dos fuentes de datos, hacer chequeos de duplicaciones, y enriquecer los datos. "MD Personator" sirve para verificar datos de personas y compañías, como son nombres, direcciones. "MD Presort" sirve para agrupar los destinatarios por códigos postales, lo cual abarata los costos de envío del correo. "MD Smart Moves" actualiza los datos de direcciones con el official USPS National Change of Address Data.

Human Inference ofrece su plugin para la herramienta "Pentaho Data Integration" denominado "Easy Data Quality" [25]. Ese plugin permite hacer validaciones y limpieza de direcciones, mail, teléfonos, y nombres. Esas validaciones y limpiezas se hacen invocando servicios de Human Inference remotamente o localmente si se tiene instalado Human Inference.

"Data Quality Services" [8] es una herramienta de Microsoft que posibilita detectar duplicados, limpiar, validar, unir, y enriquecer los datos usando un referencial. Asimismo posibilita el monitoreo para chequear que se esté haciendo lo diseñado para calidad de datos. En esta herramienta se propone construir el proceso de calidad basado en el conocimiento de los datos.

El proveedor Pentaho [11] ofrece análisis de datos y manejo de calidad de los datos, a través de los componentes mencionados anteriormente de los proveedores Melissa Data y Human Inference. Estos componentes corren sobre la herramienta de integración de datos "Pentaho Data Integration".

4.3. Conclusiones

Como se puede ver en la evaluación hecha se cuenta con una gran cantidad de herramientas. Asimismo es clara la tendencia a incorporar herramientas específicas para la calidad de datos de un sistema. Cabe recalcar que casi todos los proveedores de herramientas de ETL y DataWarehouse incorporaron una herramienta específica para calidad de datos, o la incluyeron en alguna de sus herramientas existentes. Esto denota que cada vez más se está haciendo hincapié en analizar la calidad de los datos antes de usarlos, así como en limpiarlos y monitorear la calidad de los datos. Si bien se está volviendo cada vez más importante utilizar una herramienta de calidad de datos, aún las herramientas existentes tienen varias carencias. En muchas ocasiones esa herramienta de calidad de datos sirve simplemente para realizar un análisis general de los mismos, otras herramientas cuentan además con limpieza de datos. Además, el análisis de datos propuesto consiste principalmente en estadísticas sobre los datos.

Por otra parte podemos ver que la gran mayoría de las herramientas de calidad analizadas atacan las mismas dimensiones dejando de lado algunas, por ejemplo la dimensión completitud, factor cobertura, no es atacado por ninguna herramienta. El factor densidad tampoco es cubierto por la gran mayoría de las herramientas. A su vez la dimensión frescura con sus factores actualidad, y edad tampoco es cubierta por la gran mayoría de las herramientas. El factor contradicción de la dimensión unicidad tampoco es atacado por las herramientas

analizadas. Como estos, existen muchas otras dimensiones en la literatura de Calidad de Datos que no podrían ser manejados con estas herramientas.

Otra carencia detectada es un enfoque más formal, riguroso y exacto a la hora de medir la calidad. Es clara la falta de una herramienta que haga una medición sistemática y objetiva de la calidad de los datos y que no sean solo observaciones generales sobre los datos. Con un enfoque más formal se puede tener métricas de calidad definidas claramente a la hora de analizar los datos, lo cual permite tener medidas exactas de calidad para poder mejorar la toma de decisiones con respecto a la gestión de la misma.

A su vez, es notoria la poca integración de herramientas de calidad de datos en los procesos de ETL. Solo dos herramientas se encontraron (de los proveedores Melissa Data y Human Inference), que se integran a la herramienta de ETL "Pentaho Data Integration". Estas dos herramientas pueden ser incluidas en el proceso de ETL pero tienen la desventaja que son limitadas para ciertos tipos de datos y ciertas dimensiones. Si bien los datos pueden ser analizados independientemente de si van a pasar por un proceso de ETL o no, es de gran interés en los procesos de ETL poder analizar la calidad de los datos con el fin de tomar decisiones y hasta realizar limpiezas sobre los mismos previo a ser cargados.

Capítulo 5

Herramienta desarrollada

En este capítulo se describirá el análisis, diseño e implementación de la solución propuesta.

En la sección 5.1 se describirá el análisis de los requerimientos que debe cumplir la solución, una descripción general de la solución propuesta que incluye el alcance definido.

En la sección 5.2 se hará un análisis de los casos de uso que debe incluir la solución, los componentes que implementan esos caso de usos, y las bases de datos utilizadas para almacenar la información en cada uno de los componentes.

Por último en la sección 5.3 se describen las decisiones de implementación tomadas, las tecnologías usadas y los problemas encontrados.

5.1. Descripción de la solución propuesta

En esta sección se describen los requerimientos funcionales como los no funcionales que debe cumplir la solución. También se describe en términos generales la solución propuesta, los actores involucrados y la interacción de estos con la solución.

A continuación haremos una breve descripción del término ESB (Enterprise Service Bus o Bus de servicios empresariales) el cual será usado en esta sección. El bus de servicios empresariales (ESB de ahora en más) es una infraestructura que facilita la arquitectura orientada al servicio. Este suministra interfaces de programación (API's) para el desarrollo de servicios y facilita su interacción. Técnicamente, ESB es una columna vertebral de mensajería la cual contiene protocolos de conversión, transformación de formato y ruteo [21].

5.1.1. Requerimientos

Uno de los requerimientos deseables es el de crear una solución que sea integrable a un proceso de ETL que permita medir la calidad de los datos procesados, y tomar decisiones dentro del propio ETL en base a las mediciones. Asimismo se desea reutilizar las conexiones a bases de datos que poseen las herramientas de ETL.

Por otra parte, es de gran interés que las mediciones se hagan en base a un enfoque más riguroso y exacto, para lo cual se requiere poder definir el

Modelo de Calidad previo a la evaluación. Es importante tener en cuenta al definir el Modelo de Calidad el atributo al cual se le va a evaluar la calidad. Con el fin de reutilizar información se desea tener almacenadas las dimensiones más comúnmente usadas, con sus respectivos factores y respectivas métricas, de forma de brindar esas opciones a la hora de construir el Modelo de Calidad.

Asimismo se desea que las mediciones se hagan en base a servicios ya existentes, provistos por conocidas herramientas de calidad de datos. Es deseable generalizar el acceso a estos servicios a través de un ESB, para esto la solución debe consumir los servicios expuestos por el ESB en base a las métricas y los atributos a evaluar.

5.1.2. Solución y su interacción con usuarios y sistemas

Con el fin de cumplir con los requerimientos planteados en la sección 5.1.1 es necesario desarrollar un nuevo componente o un plugin para una herramienta de ETL. Con este objetivo se analizaron dos herramientas de ETL: "Pentaho Data Integration" y "Talend Open Studio for Data Integration" (de ahora en más Pentaho y Talend), a los efectos de definir sobre cual implementar la solución. Se seleccionaron solo estas dos herramientas por ser ambas libres y por tener conocimientos previos de uso de las mismas. Luego para decidir cuál de estas dos herramientas usar se tuvo en cuenta la documentación existente y los desarrollos hechos en ambas. Por este motivo se decidió usar Pentaho ya que cuenta con mucha documentación de cómo implementar un plugin y hay varios desarrollos hechos e incorporados en la herramienta. En resumen, la solución propuesta es el desarrollo de un prototipo en forma de plugin para Pentaho.

Como se puede ver en la figura 5.1.1 el plugin podrá analizar la calidad de los datos de una fuente determinada consultando servicios expuestos a través del bus de servicios. El ESB a su vez consumirá servicios de calidad de datos expuestos por diferentes herramientas. Se prevé que los diferentes servicios de las herramientas estén catalogados según cuales son mejores dentro de los que implementan la misma métrica, así como cual es mejor dependiendo del dominio. Esa información se almacena en la base de datos de "Configuración".

En base a la solución planteada se distinguen tres roles fundamentales para el uso de la misma, que se detallan a continuación:

- Usuario de ETL: Es el usuario con conocimientos de ETL y más específicamente con conocimientos de Pentaho.
- Técnico en Servicios: Es el administrador del ESB, que se encargará de la integración de nuevos servicios al ESB.
- Experto en Calidad de Datos: Es el experto en calidad de datos.

Cada uno de estos actores interactúan de manera distinta con la solución, lo cual se puede ver en la figura 5.1.1, y en distintos momentos. El "Experto en Calidad de Datos" va a ser el encargado en definir las dimensiones predefinidas así como el Modelo de Calidad en una primera instancia. Luego de que esto esté definido el "Usuario de ETL" va a poder hacer uso de la solución para realizar mediciones de calidad en su ETL. El "Técnico en Servicios" en cualquier momento podrá agregar nuevos servicios al ESB.

El alcance del desarrollo se ve también en la figura 5.1.1, es lo que contiene el recuadro azul. Esto significa que el alcance incluye el desarrollo del plugin

mencionado, que el mismo consuma servicios expuestos por el ESB, y guarde los datos en base de datos tanto para el Modelo de Calidad como para las mediciones realizadas.

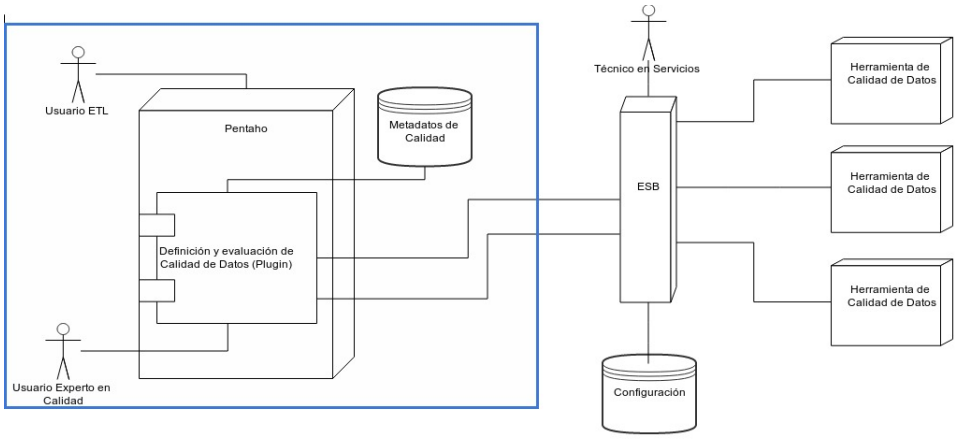

Figura 5.1.1: Principales roles y componentes de la solución

5.2. Análisis de casos de uso y de componentes

En base a la solución planteada en la sección 5.1.2 y al alcance definido, se distinguen cinco casos de uso los cuales están implementados en tres componentes distintos. A continuación se definen los casos de uso detectados, los componentes a los que se corresponden y las base de datos utilizadas en cada componente.

5.2.1. Descripción de Casos de Uso

En esta sección se describen los casos de uso llevados a cabo por los actores del sistema relacionados con el plugin desarrollado.

5.2.1.1. Definir Dimensiones Predefinidas

En este caso de uso el actor con el rol "Experto en Calidad de Datos" va a definir las dimensiones, factores y métricas generales que piense que se pueden aplicar a varios casos particulares. Esta carga se puede realizar una sola vez.

En la figura 5.2.1 se puede ver el diagrama de flujo para este caso de uso. Como se ve consta de ingresar los nombres de la dimensión, el factor y la métrica (para la métrica también se puede ingresar una descripción), luego de esto se puede o aceptar los datos ingresados en cuyo caso se guardan los datos, o se puede cancelar en cuyo caso termina el caso de uso (no guardando nada).

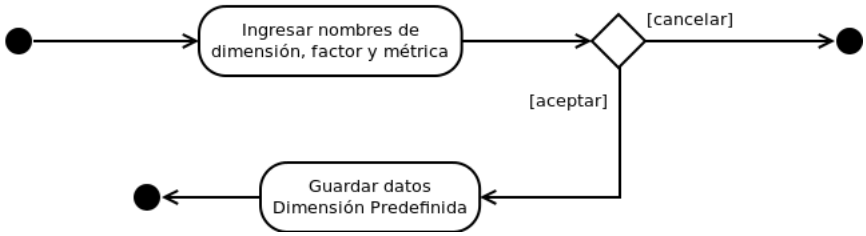

Figura 5.2.1: Flujo del caso de uso Definir Dimensiones Predefinidas

5.2.1.2. Definir el Modelo de Calidad

Este caso de uso le permite al actor con el rol "Experto en Calidad de Datos" definir un nuevo Modelo de Calidad el cual está compuesto por un conjunto de dimensiones, factores, métricas, atributos, y servicios. Para las dimensiones, factores y métricas se pueden usar las Dimensiones Predefinidas en su totalidad o parcialmente (por ejemplo usar la dimensión y asociarle un nuevo factor con su respectiva métrica). A su vez se permite guardar como Dimensiones Predefinidas las dimensiones, factores y métricas nuevas que se generen.

En la figura 5.2.2 se puede ver el diagrama de flujo para este caso de uso. El primer paso en este caso de uso es definir si se va a agregar un nuevo Modelo de Calidad o usar uno existente. Si se decide agregar uno nuevo se debe ingresar el nombre para el mismo, si se decide usar uno existente se debe seleccionar el nombre del modelo. Luego de esto se pasa a agregar los datos asociados al modelo. Los primeros datos a agregar son la dimensión, el factor y la métrica asociados, para esto se debe definir si usar las Dimensiones Predefinidas o si se van a agregar nuevos datos. Como paso siguiente se debe agregar la información del atributo a evaluar, para esto lo primero es agregar la conexión a la base de datos donde está el atributo. Esto se hace ingresando una nueva conexión o seleccionado una existente. Luego de completada la conexión se debe seleccionar el esquema y la tabla, donde se encuentra el atributo, para seleccionar después el atributo en sí. El siguiente paso es seleccionar el servicio que implementa la métrica (asociada a un factor y dimensión) que se eligió. Una vez ingresados todos estos datos se puede ingresar otro conjunto de datos al modelo o no agregar más, en la figura 5.2.2 se ve como "Seleccionar Agregar Dimensión". Si se decide seguir agregando se debe volver al punto de ingresar o seleccionar las Dimensiones Predefinidas. Si no se desean agregar más se puede aceptar en cuyo caso se guardan los datos del modelo con sus datos asociados, o se puede cancelar en cuyo caso se acaba el caso de uso (no guardando nada).

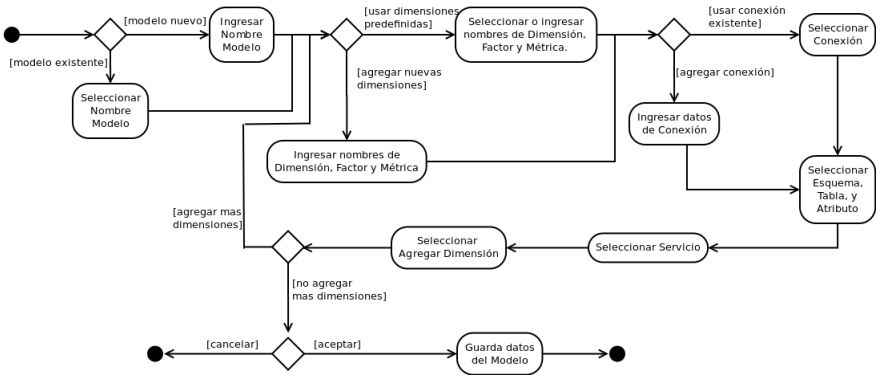

Figura 5.2.2: Flujo de caso de uso "Definir el Modelo de Calidad"

5.2.1.3. Ver Modelo de Calidad

Este caso de uso permite ver toda la información que se le ha agregado a un Modelo de Calidad, esto es de utilidad para el actor con el rol "Experto en Calidad de Datos". Se listan los modelos existentes, se selecciona uno y se pueden ver los datos asociados al mismo. Si se elige ver los datos asociados estos se mostraran en una tabla con la información correspondiente.

En la figura 5.2.3 se puede ver el flujo de este caso de uso. El mismo consta en primer lugar de seleccionar un Modelo de Calidad, luego seleccionar Ver Datos y como resultado se muestran los datos asociados al modelo.

Figura 5.2.3: Flujo de caso de uso "Ver Modelo de Calidad"

5.2.1.4. Eliminar Modelo de Calidad

Este caso de uso sirve para eliminar totalmente un Modelo de Calidad y su información asociada. Este caso también es de utilidad para el actor con el rol "Experto en Calidad de Datos".

En la figura 5.2.4 se muestra como es el flujo en este caso de uso. Es muy simple y similar al caso de uso "Ver Modelo de Calidad". La única diferencia es que se selecciona eliminar modelo y se eliminan los datos del mismo.

Figura 5.2.4: Flujo de caso de uso "Eliminar Modelo de Calidad"

5.2.1.5. Evaluar la Calidad de Datos

Este caso de uso permite que el actor con el rol "Usuario de ETL" seleccione las métricas a usar, definidas previamente en el Modelo de Calidad, para evaluar

la calidad de los datos dentro del flujo de ETL. Para que se pueda usar este caso de uso debe estar definido el Modelo de Calidad. En este momento el usuario obtiene las mediciones de calidad tanto en el flujo del ETL, como en la base de metadatos de calidad.

El flujo de este caso de uso se observa en la figura 5.2.5. Como se aprecia lo primero es seleccionar un Modelo de Calidad y una de las métricas asociadas al mismo (solo deben aparecer los modelos definidos en el caso de uso "Definir Modelo de Calidad de Datos"), el segundo paso depende de si la métrica seleccionada está asociada a algún atributo de los de entrada del flujo de ETL o no. Si está asociada (esto pasa si está definido así en el Modelo de Calidad) se muestran los atributos asociados a la misma. Luego se debe seleccionar uno de los atributos y agregarlo para ser evaluado (en la figura se ve como "Seleccionar Agregar Métrica"). Seguido de esto se pueden seguir agregando métricas y atributos para ser evaluados o no. Si se desean agregar más se debe volver al paso de seleccionar el modelo y la métrica. Si no se desean agregar más métricas se puede aceptar en cuyo caso se guardan los datos de las mediciones realizadas en base a los datos de entrada del flujo de ETL. En caso contrario se puede cancelar en cuyo caso se acaba el caso de uso no evaluando ningún atributo.

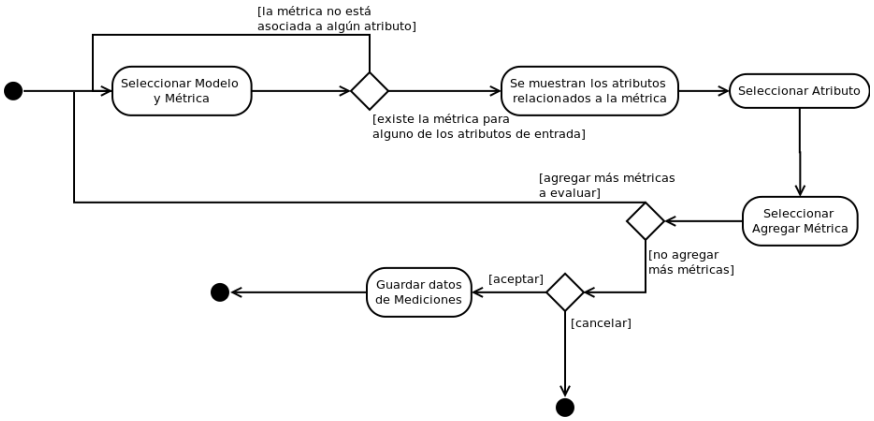

Figura 5.2.5: Flujo del caso de uso "Evaluar la Calidad de Datos"

5.2.2. Componentes y etapas de la solución

Los casos de uso mencionados dan lugar a tres componentes dentro de la solución, estos son:

- Definición de Dimensiones Predefinidas
- Definición del Modelo de Calidad
- Evaluación de la Calidad de Datos

A continuación describiremos que casos de uso están dentro de estos componentes, así como los actores que los van a usar y las bases de datos relacionadas.

La figura 5.2.6 muestra los componentes que incluye el plugin, como interactúan los actores con los mismos y el almacenamiento de la información en las respectivas bases de datos. Una vez más nos centraremos en el alcance definido, o sea en lo incluido en el recuadro azul.

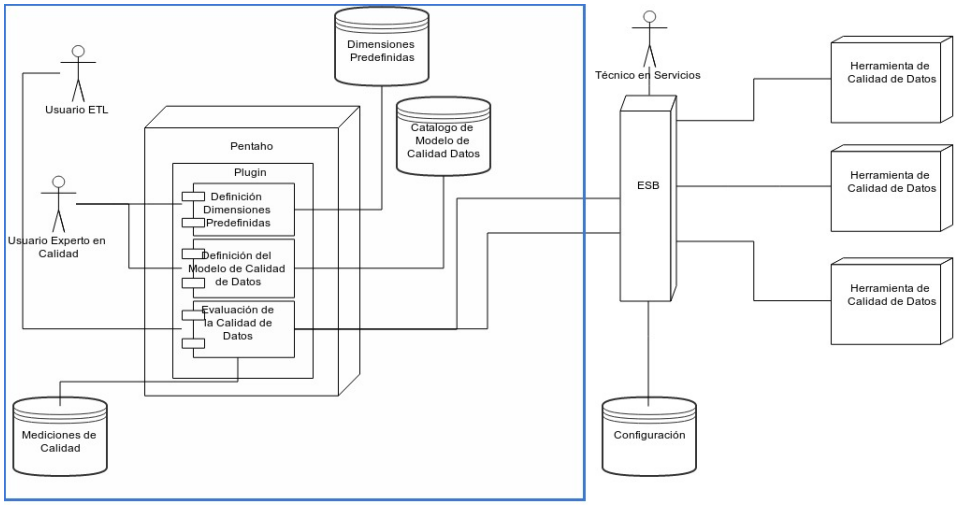

Figura 5.2.6: Diagrama de despliegue

5.2.2.1. Definición de Dimensiones Predefinidas

El Componente "Definición de Dimensiones Predefinidas" surge del caso de uso con su mismo nombre.

Este componente debe ser usado como primer paso de uso del plugin, y puede volver a ser usado en cualquier momento que se quiera agregar una dimensión predefinida, con sus factores y métricas asociadas. Pero puede no volver a usarse. Su utilidad es que a la hora de definir el Modelo de Calidad contar con las dimensiones, factores, y métricas precargadas.

La información ingresada en este componente va a ser guardada en una base de datos específica que en la figura 5.2.6 se ve como "Dimensiones Predefinidas".

5.2.2.2. Definición del Modelo de Calidad

Este componente contiene el caso de uso "Definición del Modelo de Calidad", el de "Ver Modelo de Calidad" y el caso de uso "Eliminar Modelo de Calidad".

La sección de este componente de "Definición del Modelo de Calidad" va a ser usada una vez que se haya usado el componente "Definición de Dimensiones Predefinidas" (con que se haya usado una sola vez alcanza).

La información del Modelo de Calidad creado o modificado se guarda en la base de datos "Catálogo de Modelos de Calidad de Datos" como se puede ver en la figura 5.2.6. Esta misma base de datos es la que se consulta cuando se está en la sección de "Ver datos de un Modelo de Calidad" y en la sección "Eliminar Modelo de Calidad" se elimina físicamente, de esa misma base de datos, los datos del modelo y los datos asociados al mismo.

5.2.2.3. Evaluación de la Calidad de Datos

Este componente está asociado al caso de uso "Evaluación de la Calidad de Datos".

Este debe ser utilizado en último lugar, esto es una vez que se haya definido un Modelo de Calidad para el caso particular que se está trabajando.

Los datos de las mediciones de calidad se guardan en la base de datos "Mediciones de Calidad", además de devolverse en el flujo del ETL.

Por otro lado, el componente "Evaluación de la Calidad de Datos" es el que interactúa con el ESB, consumiendo servicios del mismo.

5.2.3. Bases de datos utilizadas

La solución propuesta utiliza tres bases de datos (figura5.2.6), estas son:

- Dimensiones Predefinidas
- Catálogo de Modelo de Calidad de Datos
- Mediciones de Calidad

En la figura 5.2.7 se observa el esquema de la base de datos "Dimensiones Predefinidas". Este esquema es bien sencillo, consta de las métricas con su factor asociado y la dimensión asociada al mismo. Las claves de estas tres tablas son auto generadas.

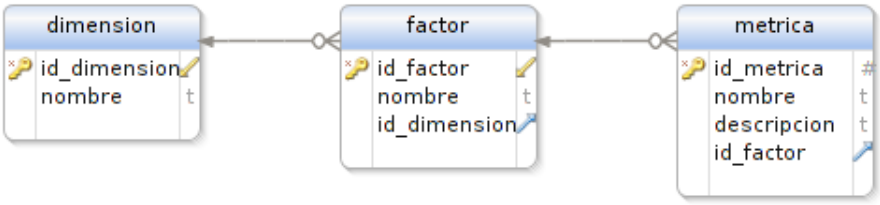

Figura 5.2.7: Esquema Dimensiones Predefinidas

Con respecto a la base de datos "Catálogo de Modelos de Calidad de Datos" se puede ver en la figura 5.2.8 que la misma contiene modelos los cuales pueden tener varias métricas instanciadas asociadas. Una métrica instanciada contiene una referencia al atributo al cual se va a aplicar la métrica, y una referencia a la métrica a aplicar. La métrica tiene un servicio asociado que es el que implementa la misma. A su vez la métrica tiene un factor asociado, el cual tiene su dimensión asociada. Respecto al atributo se guarda la información necesaria para identificar el atributo, esto es la base de datos, el esquema y la tabla a la que pertenece, también se guarda que campo de esa tabla es la clave primaria de la misma. Una vez más todas las claves de estas tablas son auto generadas.

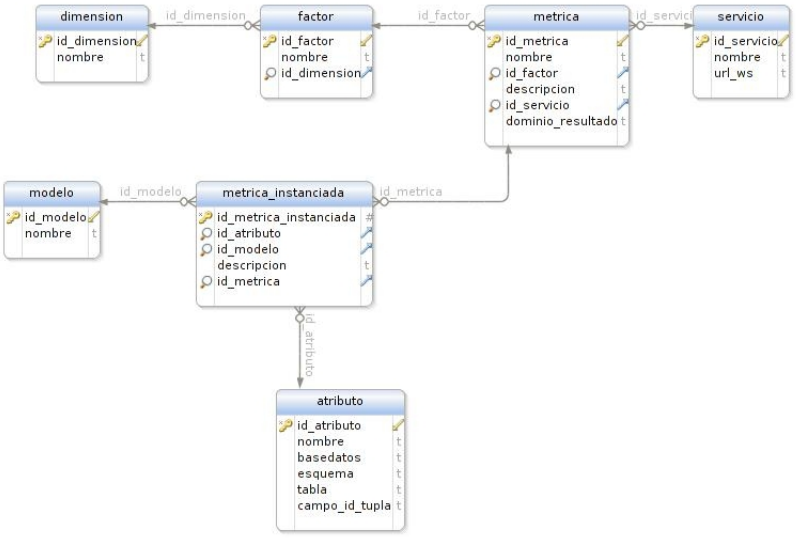

Figura 5.2.8: Catálogo de Calidad

En la figura 5.2.9 se puede ver el esquema de la base de datos que guarda las mediciones realizadas, se guarda el identificador de la métrica instanciada para poder sacar a que métrica corresponde el valor, además se guarda el valor del identificador de la tupla para la que se hizo la medición.

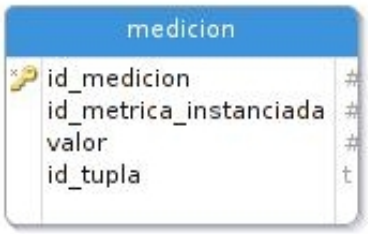

Figura 5.2.9: Esquema de Mediciones

5.3. Implementación

En esta sección se describirán las decisiones de implementación tomadas (las más importantes), las tecnologías utilizadas y algunos detalles a resaltar del desarrollo de un plugin para Pentaho.

Para ver en detalle el uso de los distintos componentes del plugin desarrollado se debe ir al manual de usuario (Apéndice A) .

5.3.1. Decisiones de implementación y dificultades encontradas

En el desarrollo de la solución se tomaron algunas decisiones que valen la pena destacar. Una de las decisiones tomadas fue desarrollar un mismo plugin que contenga tres componentes, en vez de desarrollar tres plugins distintos. Los componentes desarrollados se corresponden a los detallados en la sección 5.2.2. Como se observa en la figura 5.3.1 los tres componentes desarrollados se corresponden a un "Step" o paso de Pentaho. Estos "Steps" fueron integrados a la carpeta "Transform" porque conceptualmente era la que mejor se ajustaba a lo desarrollado.

Otra de las decisiones fue que el componente "Evaluación de la Calidad de Datos", además de almacenar las mediciones en la base de datos por defecto, tiene como salida la información de las mediciones realizadas. Esto se decidió porque se vio que era de gran interés tomar decisiones en el mismo flujo de ETL en base a las mediciones realizadas, de esta manera se evita tener que ir a consultar la base de datos de mediciones para obtener esa información.

En el componente "Definir Modelo de Calidad de Datos" cuando se define una nueva conexión a base de datos se utiliza el componente ya existente en Pentaho para este fin. Además en ese mismo componente cuando se muestran los atributos a seleccionar se decidió que mostrara el nombre de los mismos y el tipo de datos para poder identificar mejor los atributos.

Asimismo se tomó la decisión de en los componentes persistir los datos a la base de datos en el momento que se ejecutan las transformaciones ("Steps").

Respecto a la granularidad de las mediciones se decidió hacerlas a nivel de celda (o sea, por cada tupla) para poder incorporarlas en el ETL.

Por otra parte, en la sección 5.2.3 se detallaron tres bases de datos distintas, sin embargo para facilitar los reportes y las consultas se decidió que las bases de datos "Catálogo de Modelo de Calidad de Datos" y "Mediciones de Calidad" estuvieran dentro de una misma base de datos pero en distintos esquemas. Esto facilita el accionar a la hora de poder generar reportes de las mediciones hechas para una cierta métrica (asociada a un factor y este a una dimensión), obteniendo los datos en un sola consulta a la base de datos.

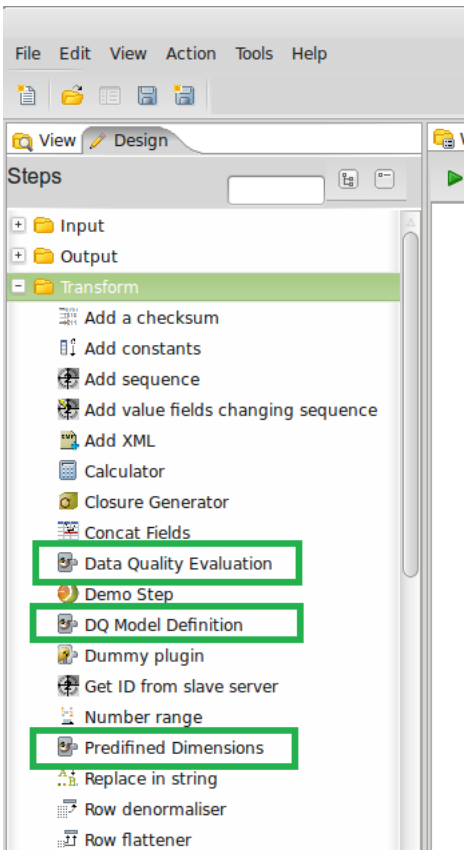

Figura 5.3.1: "Steps" desarrollados

El plugin desarrollado es un prototipo y como tal tiene algunas limitaciones que se detallarán a continuación. Una de las limitaciones es que el plugin permite evaluar solamente atributos que estén almacenados en bases de datos. Otra limitante es que el motor de base de datos donde se almacena la información es fijo (PostgreSQL) y para cambiar de motor se deben hacer cambios a nivel de código. Por otra parte, las bases de datos mencionadas en la sección 5.2.3 deben ser creadas previo al uso del plugin.

Otra dificultad fue que no se contaba con la parte del ESB desarrollada como para poder probar el plugin, sumado a esto no se pudo contar con servicios de calidad de datos que puedan ser consultados directamente. Se hicieron varios intentos para acceder a los webservices de "Easy DQ" pero no recibimos respuesta por parte de la empresa. Estas dificultades hicieron que se desarrollaran a modo de ejemplo tres webservices que implementan tres métricas, para poder probar el funcionamiento del plugin.

5.3.2. Tecnologías utilizadas

El plugin desarrollado como ya se mencionó es para la herramienta Pentaho Data Integration, la cual está desarrollada en el lenguaje de programación Java, por lo tanto nuestro plugin está desarrollado en ese mismo lenguaje. Asimismo

la interfaz de usuario se hizo usando SWT (The Standard Widget Toolkit) [15] que es lo que se usa en Pentaho Data Integration para la presentación.

Como motor de base de datos para almacenar la información creada en el uso del plugin se utilizó PostgreSQL [13]. Esto significa que las tres bases de datos mencionadas en la sección 5.2.3 están almacenadas en PostgreSQL. En el ejemplo de uso los datos a los cuales se les va a evaluar la calidad están almacenados en MySQL [9].

Para el manejo de control de versiones se utilizó Git [3], y para el respaldo del código se usó BitBucket [1].

Para los webservices de ejemplo se utilizó el entorno de desarrollo Jboss Developer Studio [14], el cual ya viene con un servidor Jboss integrado para poder exponer los webservices desarrollados. Los webservices implementados son SOAP.

Con el fin de hacer reportes de mediciones realizadas se utilizó Pentaho Report Designer [12].

5.3.3. Particularidades de Pentaho

Para desarrollar los "Steps" que componen al plugin desarrollado se implementaron las cuatro interfaces que son obligatorias de implementar para desarrollar un "Step".[2] Cada una de estas tiene una responsabilidad definida. A continuación se muestran las interfaces con sus respectivas responsabilidades:

- StepMetaInterface: Las principales responsabilidades de esta son: mantener la configuración del "step", validar la configuración del "step", proveer acceso a las clases del "step", serializar la configuración del "step", realizar cambios en el diseño de la tupla que se está procesando.
- StepDialogInterface: La principal responsabilidad de este es construir y mostrar la pantalla de configuración del "step".
- StepInterface: Esta interfaz es la encargada de procesar las tuplas.
- StepDataInterface: La responsabilidad de esta interfaz es proveer almacenamiento para el procesado de las tuplas.

Es obligatorio implementar estas cuatro interfaces si se quiere desarrollar un "Step", y cada una de las clases que lo implementa debe seguir la notación existente para ello. Estas cuatro clases son las que hacen todo el manejo de la configuración del "step" o sea de los datos ingresados para el mismo.

Capítulo 6

Ejemplo de Uso

A continuación se detallará un ejemplo de uso completo para el plugin desarrollado de forma de mostrar cómo debería usarse el mismo.

Para mostrar el uso del plugin desarrollado, se utilizó la base de datos de ejemplo de Pentaho Data Integration "sakila" [20]. Esta base está pensada para dar soporte en el proceso de negocio de alquiler de DVD a una cadena.

La cadena está compuesta por muchas tiendas que son atendidas por distintos empleados. Los clientes se registran en una tienda de la cadena, pero pueden alquilar DVD en cualquiera de las tiendas de la cadena. Los alquileres deben ser pagos por los clientes, y pueden ser pagos en cualquier momento.

En esta base hay cuatro grandes categorías que ayudan a entender mejor el modelo de datos:

- **Películas:** Comprende la tabla de Películas ("film") disponibles y otras tablas auxiliares que proveen información de las mismas como son actores, categoría y lenguaje ("actor", "category", "language").
- **Tiendas:** Se tiene la tabla de tiendas ("store") y las tablas relacionadas que contienen los empleados y el inventario ("staff", "inventory").
- **Clientes:** Incluye la tabla de clientes ("customer") y las tablas relacionadas con el alquiler y el pago ("rental", "payment").
- **Ubicación:** comprendiendo las tablas de países, ciudades y direcciones ("country", "city", "address"), las cuales se usan para las direcciones normalizadas de los clientes, los empleados y las tiendas de la cadena.

Se puede ver en la figura 6.0.1 el diagrama de la base de datos que da soporte a este negocio.

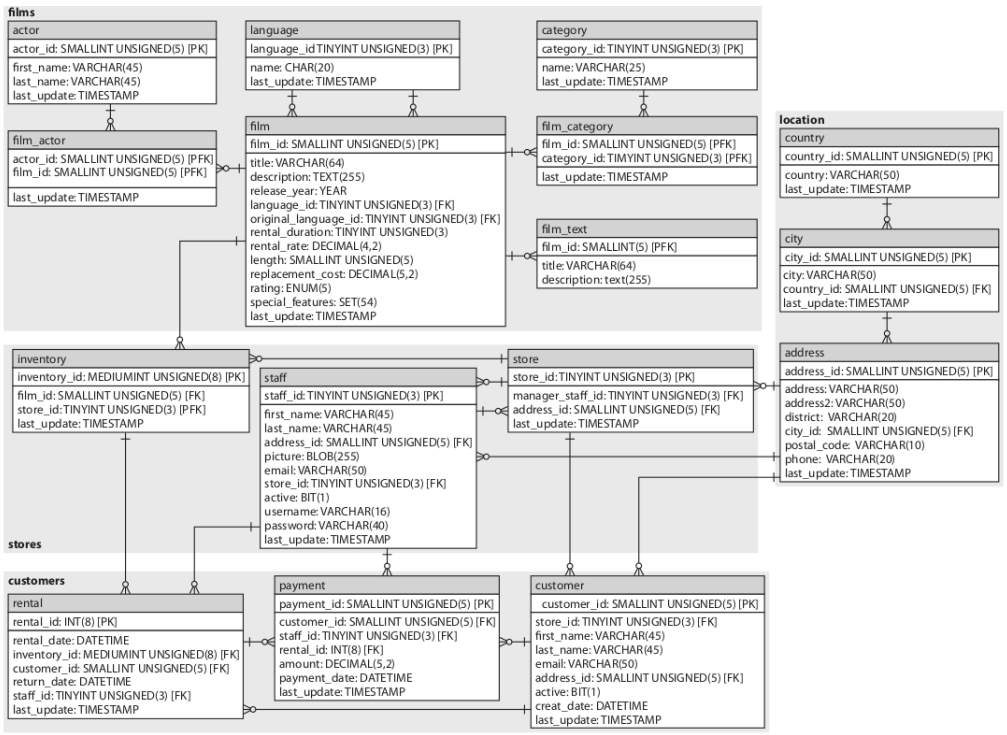

Figura 6.0.1: Base de datos Sakila

A partir de la base de datos de Sakila es de interés hacer un modelo dimensional focalizado en el proceso de negocio de alquiler como se puede ver en la figura 6.0.2 el mismo está diseñado con un esquema estrella.

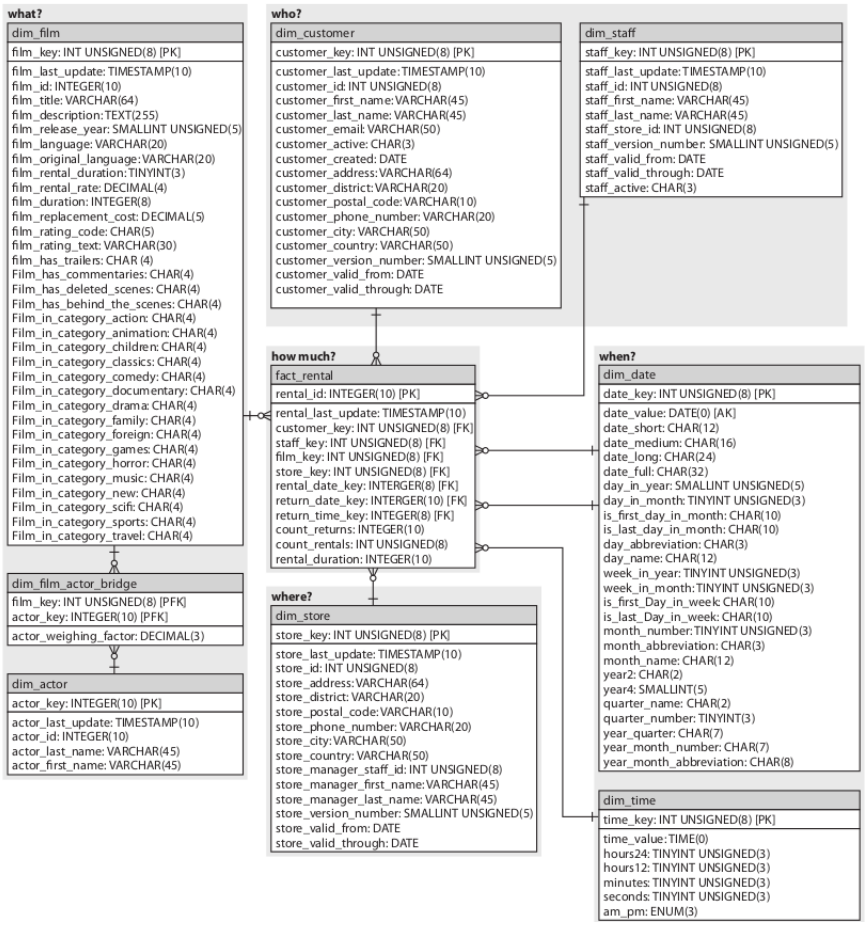

Figura 6.0.2: Base de datos del Data Warehouse de Sakila

6.1. Tabla de hechos

Podemos ver que se tiene una tabla de hechos "fact_rental" la cual contiene algunas medidas o indicadores para evaluar el proceso de negocio de alquiler como son "count_rentals" para contar la cantidad de alquileres, "count_returns" para contar la cantidad de devoluciones. Asimismo contiene las claves de las dimensiones asociadas, las cuales proveen el contexto para el estado del negocio en el momento que se sacaron estas mediciones. Esta tabla corresponde directamente a la tabla "rental" del esquema de base de datos sakila (de donde es sacada la información).

6.2. Tablas de dimensiones

Como se puede ver al ser un esquema estrella cada dimensión se corresponde a una sola tabla de dimensión. Todas las tablas de dimensiones tienen de nombre: dim_-nombre_dimension- donde -nombre-dimension- es un nombre descriptivo para la dimensión en juego.

Como se puede ver en la figura 6.0.2 se sugiere que las dimensiones están organizadas en cuatro grupos que contienen dimensiones relacionadas conceptualmente (más un quinto grupo que es "how much?" que refiere a la tabla de hechos).

- **¿Quién? ("who?"):** En este grupo se encuentran las dimensiones relacionadas a los clientes y a los empleados ("dim_customer" y "dim_staff") que participan en el alquiler. Se tienen en estas algunos campos para poder tener el historial: %_version_number, %_valid_from, y %_valid_through.
- **¿Cuándo? ("when?"):** Este grupo contiene las tablas de dimensión que marcan en el tiempo cuando un alquiler o una devolución ocurre.
- **¿Dónde? ("where?"):** Acá se tiene la dimensión de la tienda ("dim_store") donde fue alquilado un DVD, al igual que las tablas en el grupo de ¿quién? se tienen los mismos campos para guardar el historial de cambios.
- **¿Qué? ("what?"):** Este grupo contiene las dimensiones del actor y de la película ("dim_actor" y "dim_film") que se está alquilando. Solo tiene relación con la tabla de hechos la tabla de película dado a que es la que tiene relación directa con el DVD alquilado. A su vez como una película tiene muchos actores se tiene una tabla de relación entre la película y los actores ("dim_film_actor_bridge").

La información para las tablas de dimensiones es sacada de la respectiva tabla de la base de datos Sakila, por ejemplo la información para la tabla de dimensión de las tiendas ("dim_store") es sacada de la tabla de las tiendas ("store").

6.3. Uso del plugin en la carga del modelo dimensional

Para cargar los datos en la base de datos de la figura 6.0.2 se utilizan procesos de ETL en Pentaho Data Integration.

En algunos de estos procesos se añadió el uso del plugin desarrollado con el fin de evaluar la calidad de los datos y poder tomar decisiones de si son cargados o no.

Se hicieron algunas propuestas de mejora de los datos:

1. Para poder sacar reportes del negocio se quiere que los países (tanto el del cliente que alquila como el de la tienda) sean datos correctos, para esto se impone que los mismos estén en la lista de países de la norma ISO 3166 [6].
2. Se precisa que los clientes tengan mail asociado para poder enviarles promociones a los clientes que más alquilan.
3. Es deseable que los teléfonos de los clientes sean datos correctos, esto es que sean números y de largo 8, para poder hacer campañas por teléfono.

Para cumplir con las mejoras mencionadas se definió un Modelo de Calidad "Alquiler de Películas" con tres métricas (asociadas a sus respectivas dimensiones) sobre los campos país, mail y teléfono. A continuación se detallan los pasos realizados en los tres casos.

6.3.1. Mejora de datos 1 - países

Para esta mejora se definió en el Modelo de Calidad la dimensión *exactitud* con el factor *correctitud sintáctica* y con la métrica *booleano de correctitud sintáctica*. Como atributo se seleccionó el atributo "country" de la tabla país ("country") de la base de datos Sakila tal como se ve en la figura 6.0.1. El servicio elegido asociado sirve para comparar el nombre del país con la lista de países de la norma ISO 3166, si no se encuentra una coincidencia exacta se devuelve 0 si lo encuentra devuelve 1.

Luego de definido el Modelo de Calidad a usar se pasó a incorporar la evaluación de la calidad de datos en la transformación correspondiente de Pentaho Data Integration. Cada tabla de dimensión es cargada con una transformación y existe otra transformación para cargar la tabla de hechos. Asimismo se tiene una sub transformación que carga las direcciones de manera desnormalizada.

Nuestro componente de evaluación fue agregado a esa sub transformación que se usa para cargar las direcciones desnormalizadas como se puede ver en la figura 6.3.1. Esta sub transformación es usada por la transformación que carga los clientes y por la que carga las tiendas.

Los componentes agregados a la misma son el componente de evaluación de calidad de datos "Data Quality Evaluation", el componente de filtrado de tuplas "Filter rows" y el componente de guardado en archivo de texto "Text file output". La idea de agregar estos componentes es poder definir según el dato que venga si para nosotros es un dato correcto sintácticamente o no. Si es un dato correcto sintácticamente (el valor de la evaluación retornó 1) entonces lo vamos a agregar a la salida, en caso contrario se va a guardar toda la información de la tupla correspondiente en un archivo de texto para analizarlo posteriormente.

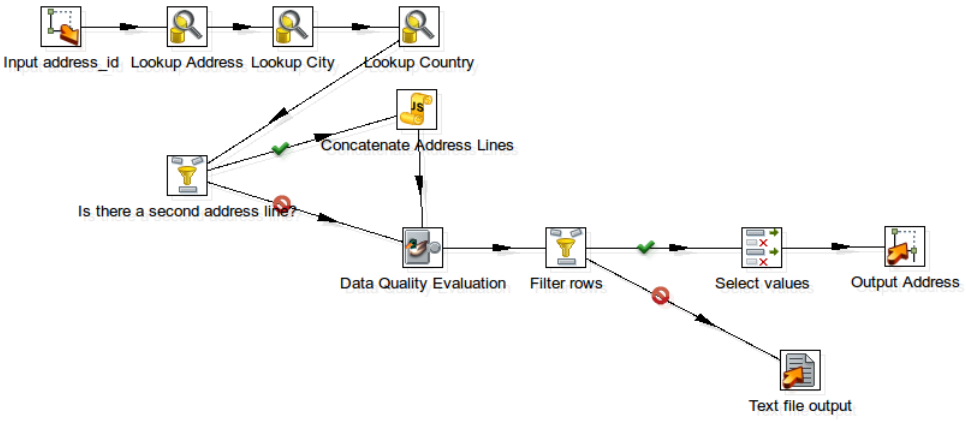

Figura 6.3.1: Carga de direcciones desnormalizadas

6.3.2. Mejora de datos 2 - email

Para esta mejora de datos se agregó al Modelo de Calidad la dimensión *completitud* con su factor *densidad* y con la métrica *booleano de valor nulo*. Como atributo se seleccionó el campo "email" de la tabla de clientes ("customer") de la base de datos Sakila (ver figura 6.0.1). El servicio asociado seleccionado

chequea que el campo pasado por parámetro no sea nulo ni vacío, si es nulo o vacío devuelve 1, en caso contrario devuelve 0.

Una vez agregada esta métrica al Modelo de Calidad se pasó a incorporar los componentes para la evaluación de datos en la transformación correspondiente. Para este caso se usó la transformación que carga los clientes, como se ve en la figura 6.3.2 al igual que en el caso anterior los componentes agregados son "Data Quality Evaluation", "Filter rows", "Text file output". El uso de cada uno es el mismo que en el caso anterior, en el componente "Data Quality Evaluation" se evalúa la calidad del dato que se va a cargar (en este caso el mail) si este nos devuelve 0 (significa que no es nulo) se carga en la tabla de la dimensión correspondiente, en caso contrario se guarda en un archivo de texto toda la tupla para su posterior análisis.

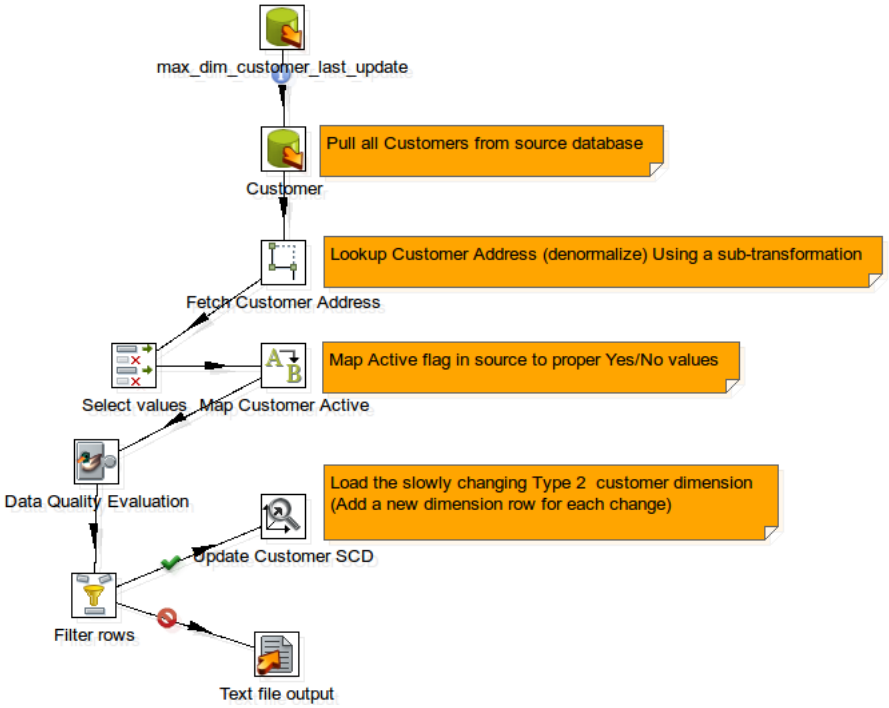

Figura 6.3.2: Carga de clientes

6.3.3. Mejora de datos 3 - phone

Con el fin de implementar esta mejora se definió en el Modelo de Calidad la dimensión *exactitud* con el factor *correctitud sintáctica* y con la métrica *booleano de correctitud sintáctica teléfono*. Como atributo se seleccionó el atributo "phone" de la tabla de direcciones ("address") de la base de datos Sakila tal como se ve en la figura 6.0.1. El servicio asociado elegido se fija si el teléfono está compuesto de números y si tiene largo ocho, si se cumplen estas dos condiciones devuelve 1 en caso contrario devuelve 0.

Para evaluar la calidad de los teléfonos se precisa agregar a la transformación de carga de clientes esta métrica, esto se hace de la misma manera que la mejora presentada en la sección 6.3.2.

En la figura 6.3.3 se puede ver los datos ingresados en el componente “Data Quality Evaluation” de la transformación que se muestra en la figura 6.3.2. Se agregaron dos métricas las dos métricas de los dos atributos a evaluar (“phone” y “email”) en esta transformación.

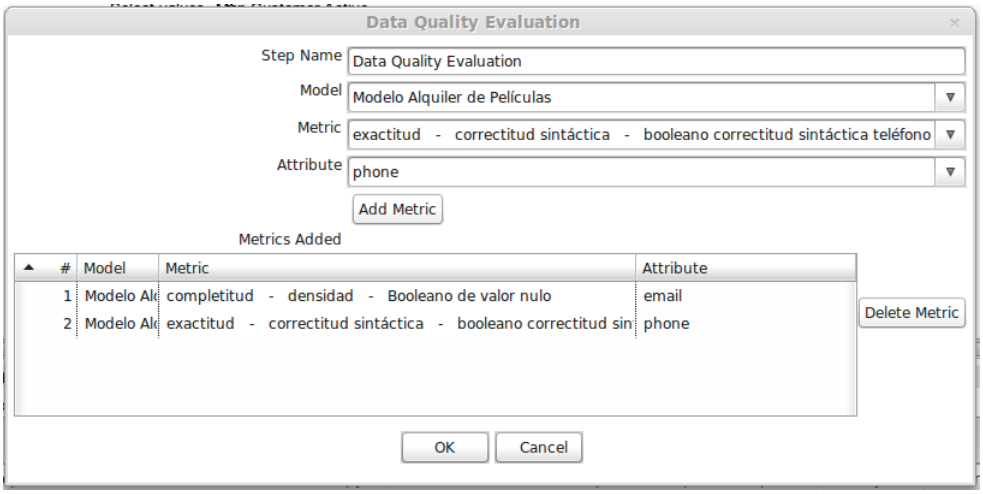

Figura 6.3.3: Datos ingresados en el componente “Data Quality Evaluation”

6.4. Reportes sobre las mediciones

Como ya se explicó todas las mediciones son guardadas en una tabla a partir de la cual es posible posteriormente sacar reportes sobre la misma.

Con los tres ejemplos de uso que se mostraron se pueden sacar reportes para ver la cantidad de tuplas rechazadas y la cantidad de tuplas correctas como se ve en la figura 6.4.1.

Atributo	Métrica	Valor	Cantidad
email	Booleano de valor nulo	0	502
email	Booleano de valor nulo	1	63
phone	booleano correctitud sintáctica teléfono	0	182
phone	booleano correctitud sintáctica teléfono	1	383
country	booleano de correctitud sintáctica país	0	34
country	booleano de correctitud sintáctica país	1	565

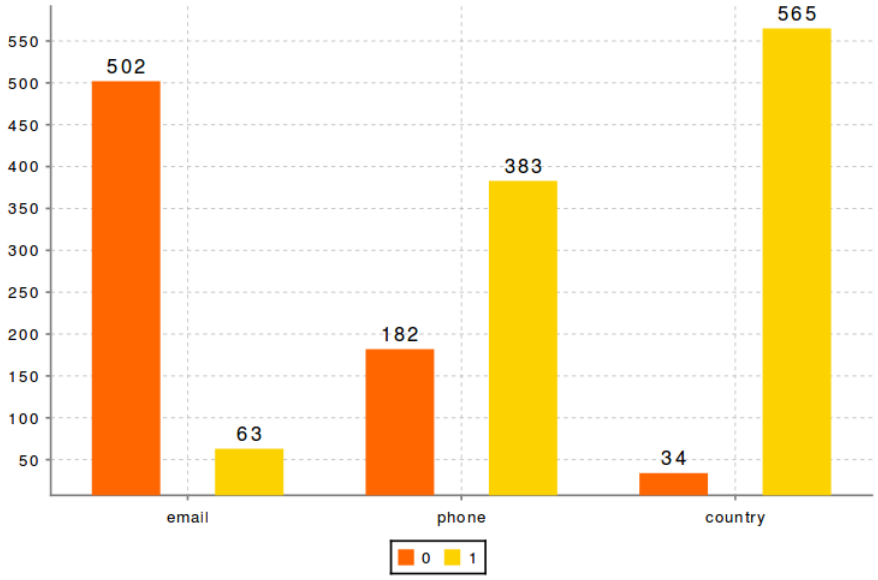

Figura 6.4.1: Reporte de Mediciones

Capítulo 7

Conclusiones y Trabajo a futuro

En este capítulo se presentan las conclusiones obtenidas en base al trabajo realizado, así como posibles mejoras y líneas de trabajos a futuro que se podrían desarrollar.

7.1. Conclusiones

En primera instancia se realizó la evaluación de herramientas de calidad de datos cumpliendo con uno de los objetivos de este proyecto. En esta evaluación se analizaron muchas herramientas de calidad de datos, herramientas libres y gratuitas así como herramientas no libres y comerciales, herramientas desarrolladas en la academia, etc. Con la variada oferta de herramientas existente se hizo necesario definir nuevas clasificaciones para éstas, y así entender mejor que ofrece cada una. La evaluación realizada confirmó la sospecha de que hace falta una herramienta de evaluación de calidad de datos más completa y basada en un enfoque más riguroso y estricto, motivando la realización de un prototipo de herramienta con esas características.

En una segunda etapa del proyecto se desarrolló un prototipo de herramienta que permite evaluar la calidad de los datos en los procesos ETL, cumpliendo así con el objetivo principal del proyecto. Esta herramienta incluye un componente para definir previamente el Modelo de Calidad aplicado al caso de estudio en cuestión. Esto permite tener un marco riguroso y estricto definido previo a la evaluación de los datos. Además la herramienta desarrollada prevé la reutilización de las definiciones de calidad de datos, permitiendo definir en primer lugar las dimensiones, factores, y métricas utilizados comúnmente. Por otra parte, la herramienta es un plugin para la herramienta "Pentaho Data Integration" (herramienta de ETL), permitiendo agregar la evaluación de la calidad como un paso más en el proceso ETL. Este paso tiene como salida además de los datos de entrada, los resultados de la medición realizada, esto permite tomar decisiones en base a las medidas de calidad obtenidas, lo cual es de gran utilidad para mejorar la calidad de los datos finales. Si bien, no se contó con la parte del ESB exponiendo servicios, se pudo probar correctamente el consumir servicios a través de los webservices desarrollados.

La evaluación de herramientas arrojó que existen dos herramientas integrables en "Pentaho Data Integration" pero como se mencionó previamente las mismas son para un dominio muy específico. Ambas herramientas ("Contact Zone" y "Easy Data Quality") se focalizan en la validación y corrección de nombres, direcciones, teléfonos y mail, esto demuestra que son de uso acotado. En contraste con esto, la herramienta desarrollada es mucho más general ya que no está focalizada en determinados tipos de datos, sino que permite definir cualquier Modelo de Calidad. Además, la herramienta desarrollada permite definir el Modelo de Calidad de forma rigurosa para cada caso específico. Por último, esta herramienta genera una base de metadatos de calidad, fundamental para el monitoreo y la gestión de la calidad.

7.2. Trabajo a futuro

Dado a que se desarrolló un prototipo de herramienta existen muchas mejoras técnicas posibles que le agregarían valor a la herramienta. Por otra parte, existen varios lineamientos de trabajos a futuro, que en conjunto con la herramienta desarrollada harían más completa la solución.

7.2.1. Mejoras y profundización

Dentro de las mejoras técnicas que se le podrían hacer a la herramienta desarrollada sería brindar la posibilidad de almacenar la información en otros manejadores de bases de datos, y que no sea solo en PostgreSQL. Esto sería dándole la opción de elegir donde almacenar los datos al usuario en cuestión.

Otra mejora es que el mismo plugin se encargue de crear los esquemas de las bases de datos usadas, si estos no fueron creados aún. Se deberían crear los esquemas la primera vez que se precisan los mismos, esto es por ejemplo crear el esquema de Dimensiones Predefinidas cuando se va a definir por primera vez una dimensión en el componente de Dimensiones Predefinidas.

Además, otra mejora es poder evaluar datos que provengan de otras fuentes (y no solo de bases de datos), por ejemplo datos que están almacenados en un archivo. Sería deseable que se pudiera seleccionar de qué tipo de fuente se van a extraer los datos a evaluar y realizar las configuraciones necesarias para definir el Modelo de Calidad, realizar las evaluaciones, y que estas tengan una traza al atributo evaluado.

Otro punto es que sería deseable poder realizar bajas y modificaciones para las dimensiones predefinidas. Si bien hoy en día se pueden dar de alta dimensiones predefinidas, no es posible modificarlas (una vez que fueron almacenadas en base de datos) ni darlas de baja. Además puede ser de interés poder visualizar que datos están cargados como Dimensiones Predefinidas.

Para la definición del Modelo de Calidad se permite dar de baja un modelo en su totalidad, pero no se permite dar de baja algunos datos del mismo, ni modificarlos (una vez que fueron guardados en base de datos). Por esto, sería deseable dar de baja o modificar el contenido del modelo, por ejemplo dar de baja la asociación de una dimensión, factor y métrica con un atributo, o modificarla cambiando el atributo.

El manejo de errores podría mejorarse, atrapando las excepciones de base de datos, tanto al insertar como al consultar y transformarlas en un mensaje

amigable para el usuario.

7.2.2. Líneas de trabajo a futuro

Una de las líneas de trabajo a futuro es la de incorporar nuevos componentes al plugin desarrollado que le agreguen valor. Uno de estos componentes posibles es para limpieza de datos, una vez definido el Modelo de Calidad, luego de evaluada la calidad se podría limpiar los datos que no cumplen con los valores esperados de calidad. Esta limpieza se podría realizar de la misma manera que la evaluación, consultando servicios expuestos por el ESB, y que este a su vez consulte servicios de herramientas de calidad de datos existentes.

Otra de las líneas de trabajo a futuro es completar la solución propuesta con el ESB exponiendo servicios de las herramientas de calidad de datos y el plugin desarrollado consumiendo servicios del mismo.

Bibliografía

[1] Bitbucket. https://bitbucket.org/, accedida Noviembre 2013.

[2] Extending pentaho. http://docs.huihoo.com/pentaho/pentaho-business-analytics/4.8/pdi_embed_extend_guide.pdf, accedida Diciembre 2013.

[3] Git. http://git-scm.com/, accedida Noviembre 2013.

[4] Ibm - infosphere information analyzer. http://publibfp.boulder.ibm.com/epubs/pdf/c1937960.pdf, accedida Mayo 2013.

[5] Information quality. http://www.informationquality.org/newspdf/Study-Medication%20errors%20harm%201.5M%20a%20year.pdf, accedida Octubre 2013.

[6] Iso. http://www.iso.org/iso/country_codes.htm, accedida Marzo 2014.

[7] Melissa data - contact zone. http://www.melissadata.com/dqt/contact-zone/index.htm, accedida Mayo 2013.

[8] Microsoft - data quality services. http://msdn.microsoft.com/en-us/library/ff877917.aspx, accedida Mayo 2013.

[9] Mysql. http://www.mysql.com/, accedida Noviembre 2013.

[10] Oracle enterprise data quality. http://www.oracle.com/us/products/applications/master-data-management/oracle-enterprise-data-quality-ds-430148.pdf, accedida Mayo 2013.

[11] Pentaho - data integration. http://www.pentaho.com/product/data-integration, accedida Setiembre 2013.

[12] Pentaho report designer. http://community.pentaho.com/projects/reporting/, accedida Marzo 2014.

[13] Postgresql. http://www.postgresql.org/, accedida Noviembre 2013.

[14] Red hat jboss developer studio. https://www.jboss.org/products/devstudio.html, accedida Marzo 2014.

[15] The standard widget toolkit. http://www.eclipse.org/swt/, accedida Diciembre 2013.

[16] Talend - herramientas de calidad. http://www.talend.com/, accedida Mayo 2013.

[17] Donald Ballou, Richard Wang, Harold Pazer, and Giri Kumar Tayi. Modelling information manufacturing systems to determine information product quality. *Management Science*, 44:4, 1998.

[18] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer, 2006.

[19] R. Bouman and J. van Dongen. *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL*. Wiley, 2010.

[20] M. Casters, R. Bouman, and J. van Dongen. *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*. IT Pro. Wiley, 2010.

[21] D.A. Chappell. *Enterprise Service Bus: Theory in Practice*. Theory in practice. O'Reilly Media, 2004.

[22] Harold L. Pazer Donald P. Ballou. Modeling completeness versus consistency tradeoffs in information decision contexts. *Journal on Management of Information Systems*, 15:240–243, 2003.

[23] Inc Gartner. Magic quadrant for data quality tools. http://www.gartner.com/technology/reprints.do?id=1-1BO662V\&ct=120809\&st=sb, accedida Mayo 2013.

[24] Virginie Goasdoué, Sylvaine Nugier, Dominique Duquennoy, and Brigitte Laboisse. An evaluation framework for data quality tools. Technical report, A.I.D., EDF R&D, 2007.

[25] Human Inference. Easy data quality. http://www.easydq.com/, accedida Mayo 2013.

[26] W.H. Inmon, D. Strauss, and G. Neushloss. *DW 2.0: The Architecture for the Next Generation of Data Warehousing: The Architecture for the Next Generation of Data Warehousing*. The Morgan Kaufmann series in data management systems. Elsevier Science, 2010.

[27] Heiko Müller and Johann-Christoph Freytag. Problems, Methods and Challenges in Comprehensive Data Cleansing. Technical Report HUB-IB-164, Humboldt-Universität zu Berlin, Institut für Informatik, 2003.

[28] Felix Naumann, Johann-Christoph Freytag, and Ulf Leser. Completeness of information sources. Proc. of the Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), Italy, 2003.

[29] Verónika Peralta. *Data Quality Evaluation in Data Integration Systems*. PhD thesis, Universidad de Versalles, Francia y Universidad de la República, Uruguay, 2006.

[30] T.C. Redman. *Data quality for the information age*. Artech House Telecommunications Library. Artech House, Incorporated, 1996.

[31] Monica Scannapieco and Tiziana Catarci. Data quality under a computer science perspective. *Archivi & Computer*, 2:1–15, 2002.

[32] Alfonso Vicente. Relevamiento de herramientas de limpieza de datos. 2006.

[33] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal on Management of Information Systems*, 12:5–34, 1996.

Apéndice A

Manual de Usuario

En este capítulo se muestra como se usan los tres componentes desarrollados, una vez que se tiene todo el ambiente instalado. Esto es el plugin desarrollado incorporado a "Pentaho Data Integration", PostgreSQL instalado y con los esquemas de base de datos creados.

A.1. Definición de Dimensiones Predefinidas

Para usar este componente lo primero que se debe hacer es crear una nueva Transformación en Pentaho Data Integration. Esto se hace seleccionando dentro de Pentaho Data Integration a "File"->"New"->"Transformation".

En la transformación creada se debe seleccionar el paso de Definición de Dimensiones Predefinidas como se ve en la figura A.1.1. Se selecciona el "Step" con nombre "Predefined Dimensions" y se lo arrastra a la transformación creada.

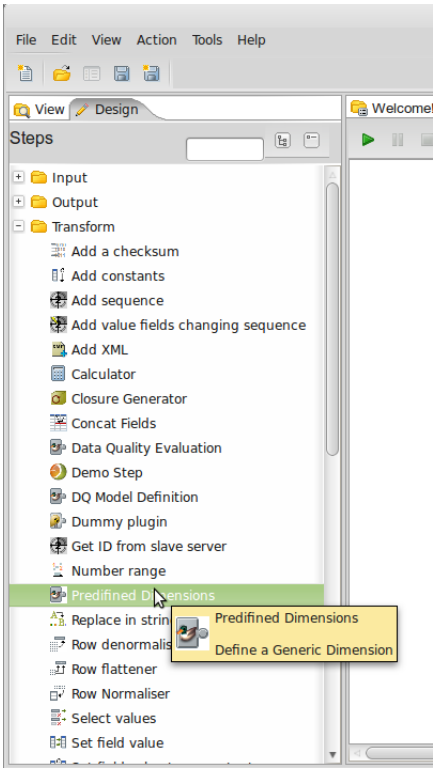

Figura A.1.1: Seleccionar componente de Definición de Dimensiones Predefinidas

En la figura A.1.2 se puede ver el componente ya seleccionado y arrastrado a la transformación creada para ser usado.

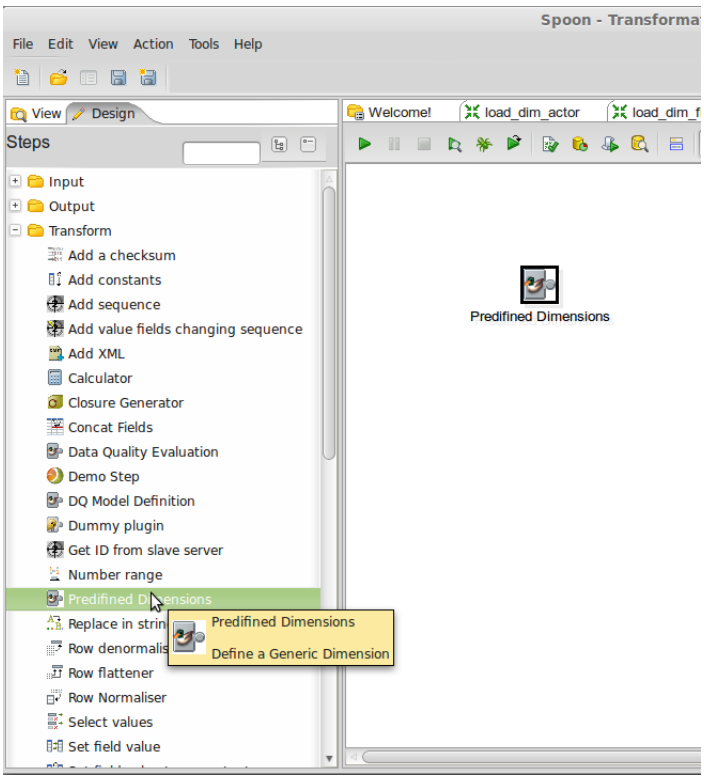

Figura A.1.2: Componente de Dimensiones Predefinidas en la Transformación

Una vez que se eligió el componente y se encuentra el mismo en la transformación, se pasa a configurarlo. Para esto se debe seleccionar el mismo y se mostrará la pantalla de configuración que se ve en la figura A.1.3. En esta pantalla aparecen los campos a completar: “Step Name” para cambiarle el nombre al paso, “Dimension” donde se debe poner el nombre de la dimensión, “Factor” donde se debe poner el nombre del factor, “Metric” donde se debe agregar el nombre de la métrica, y por último “Metric Description” para agregarle si se quiere una descripción a la métrica. Una vez que se hayan completado todos los campos se debe presionar el botón “Ok”.

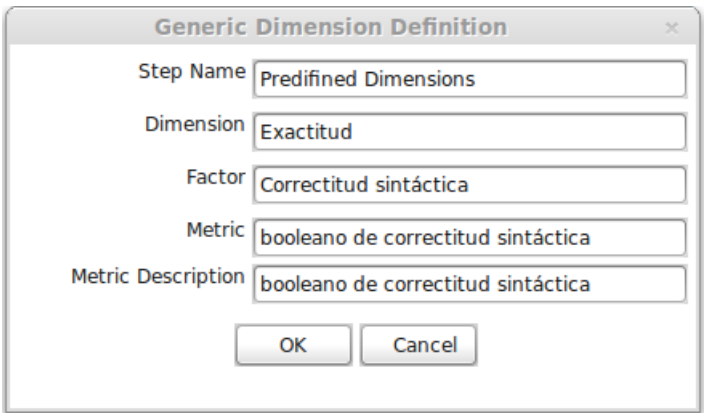

Figura A.1.3: Configuración del componente

Luego de configurar el componente se debe ejecutar la transformación para que los datos se guarden en la base de datos correspondiente. Esto se hace como se observa en la figura A.1.4 seleccionando el botón de ejecutar (botón que está seleccionado en la figura). Si no hay errores en los datos ingresados se muestra un tic arriba del componente en verde. Si se desea agregar otro conjunto de dimensión, factor y métrica se puede agregar otro componente de este tipo a la transformación, o elegir el mismo componente y cambiarle los datos ingresados.

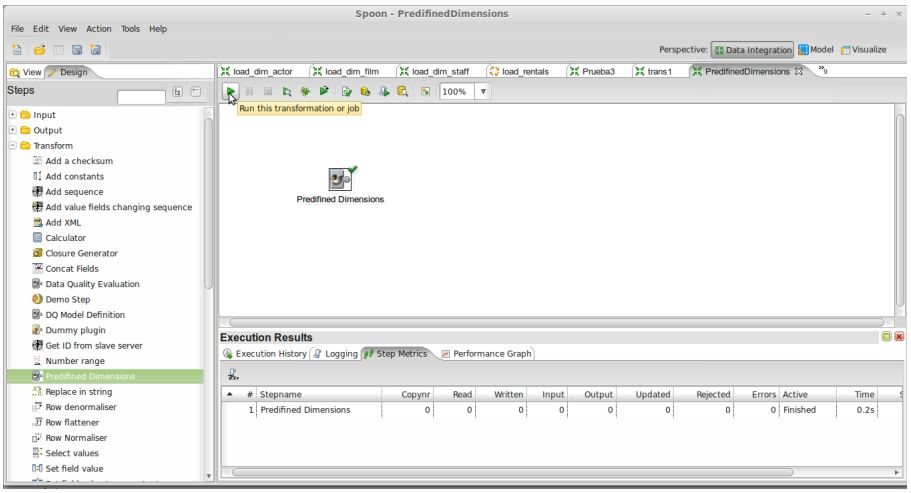

Figura A.1.4: Ejecutar la transformación

A.2. Definición del Modelo de Calidad

Lo primero que se debe hacer es crear una nueva Transformación en Pentaho Data Integration. Esto se hace seleccionando dentro de Pentaho Data Integration a “File”->”New”->”Transformation”.

Una vez en la nueva Transformación creada se debe seleccionar el paso de Definición del Modelo de Calidad. Esto se hace como se muestra en la figu-

ra A.2.1 seleccionando a la vista “Design” y luego a “Transform”, ahí se va a *encontrar* el componente “DQ Model Definition”.

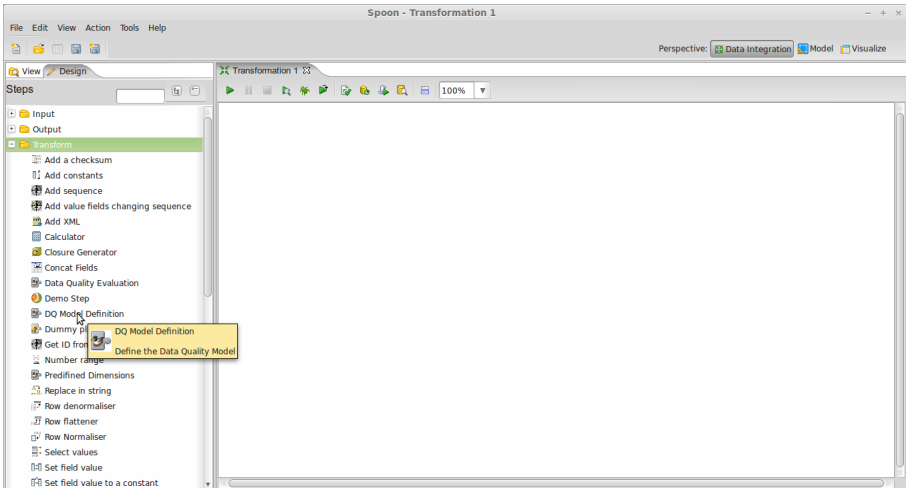

Figura A.2.1: Seleccionar componente de Definición del Modelo de Calidad

El mismo se debe seleccionar y arrastrar hacia la nueva transformación como se muestra en la figuraA.2.2

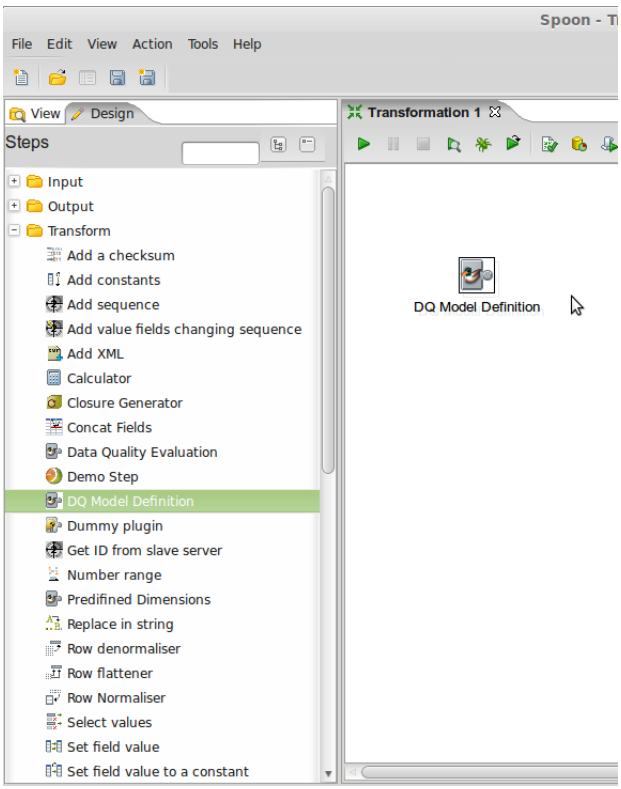

Figura A.2.2: Componente agregado a la transformación

Luego de que ya se tiene en la transformación el componente, se pasa a configurarlo, para lo cual se selecciona el mismo (dentro de la transformación) y se muestran los campos a completar. Como se puede ver en la figura A.2.3 se tienen dos pestañas, la primera (que es la que está seleccionada) es donde se van a agregar nuevos Modelos o modificar Modelos ya existentes. La segunda pestaña es para visualizar los Modelos existentes con sus respectivas configuraciones.

A.2.1. Pestaña de agregar o modificar Modelos

En la pestaña de agregar o modificar Modelos ("Add/Modify Model"), se tienen que completar varios campos como se ve en la figura A.2.3. A continuación se van a detallar cada uno de los grupos de campos a completar.

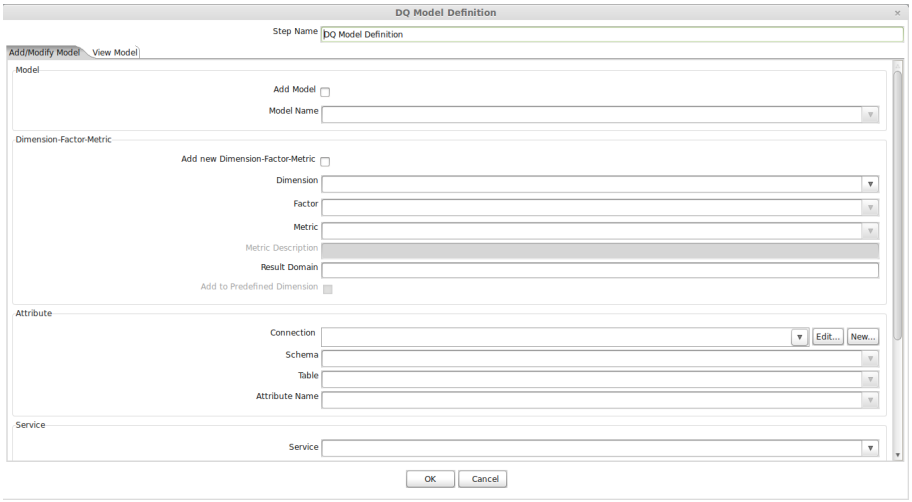

Figura A.2.3: Vista de componente de Definición del Modelo de Calidad

A.2.1.1. Datos del Modelo

La primer sección de campos son los que están recuadrados en "Model" estos son los campos relativos al Modelo.

Si se selecciona "Add Model" significa que se va a agregar un nuevo modelo para lo cual se debe escribir un nombre en "Model Name", en cambio si no se selecciona "Add Model" se debe elegir en "Model Name" uno de los modelos ya existentes.

A.2.1.2. Datos de Dimensiones, Factores y Métricas

En la sección recuadrada como "Dimension-Factor-Metric" se van a agregar los datos de una dimensión, un factor y una métrica.

En esta sección hay dos opciones se puede agregar un nuevo conjunto Dimensión-Factor-Métrica o se pueden usar los ya existentes en la base de Dimensiones Predefinidas. Cabe destacar que también se puede seleccionar por ejemplo solo la Dimensión y a esa agregarle un factor o métrica. Esta decisión de si agregar todo nuevo o usar la información preexistente (total o parcialmente) en las dimensiones predefinidas se toma seleccionando "Add new Dimension-Factor-Metric". Si se selecciona, entonces no se va a usar la información preexistente, en caso contrario se puede usar la información ya existente total o parcialmente.

En los campos "Dimension", "Factor", y "Metric" van los nombres de esos campos como se aprecia en la figura A.2.4. Si se está agregando nueva información se debe llenar el campo "Metric Description" en el cual se espera una breve descripción de la métrica. Asimismo cuando son datos nuevos se puede elegir agregarlos a la base de Dimensiones Predefinidas. Para la métrica se debe indicar siempre el dominio del resultado, este se detalla en el campo "Result Domain".

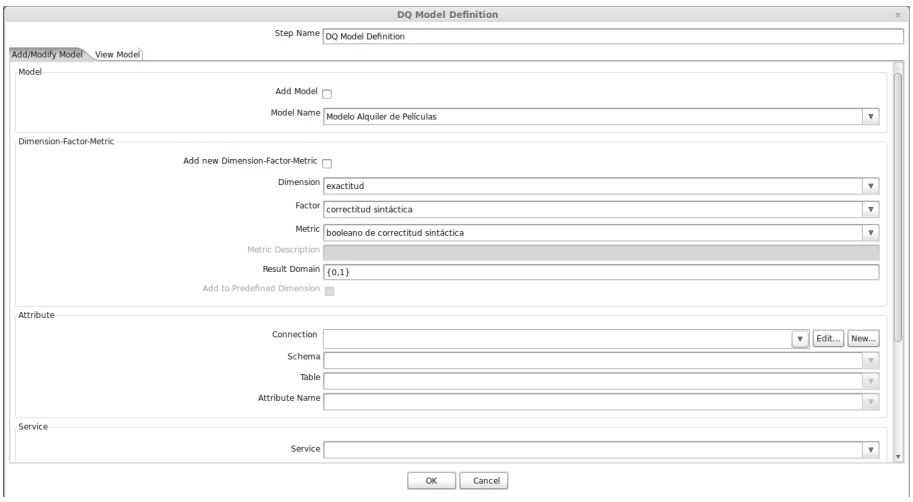

Figura A.2.4: Campos del grupo "Dimension-factor-metric"

A.2.1.3. Datos del Atributo

El siguiente recuadro es "Attribute" que es donde se va a almacenar la información correspondiente al atributo al que se le va a evaluar la calidad, teniendo en cuenta la dimensión-factor-métrica definidos arriba.

El primer campo que se debe completar es "Connection" el cual es la conexión a la base de datos donde se encuentra el atributo. Para lo cual se puede seleccionar una conexión ya definida, o de lo contrario se puede realizar una nueva conexión. Al seleccionar el botón "New" se va a mostrar el componente de definición de conexiones a base de datos de Pentaho. Como muestra la figura A.2.5 se debe seleccionar primero un tipo de conexión en "Connection Type" y luego se deben completar los campos relacionados a la conexión en "Settings". Además se le debe agregar un nombre a la conexión en "Connection Name". Luego de completados todos los datos se puede probar la definición de la conexión, seleccionando el botón "Test". Aquí se muestra un mensaje indicando si la conexión fue exitosa o si hubo algún error. Una vez terminado esto se selecciona el botón "OK" para volver a la definición del Modelo de Calidad.

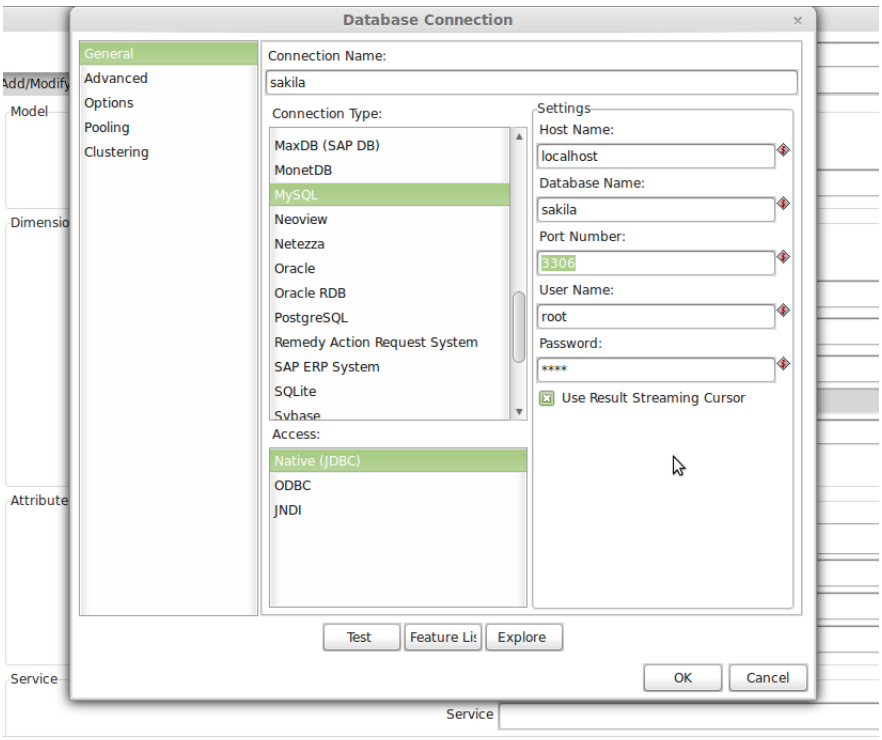

Figura A.2.5: Definición de la conexión del atributo

Luego de configurada la conexión se usa la misma para desplegar la información de los campos del atributo que faltan. De esta manera en el campo "Schema" se muestran los esquemas disponibles para la base de datos definida, se debe seleccionar un esquema para pasar al próximo campo.

En el campo "Table" se muestran las tablas disponibles para la conexión definida y el esquema seleccionado previamente.

Por último se debe seleccionar el nombre del atributo que se desea usar. En base a la información seleccionada previamente se muestran los atributos de la tabla con sus respectivos tipos de datos entre paréntesis, (figuraA.2.6).

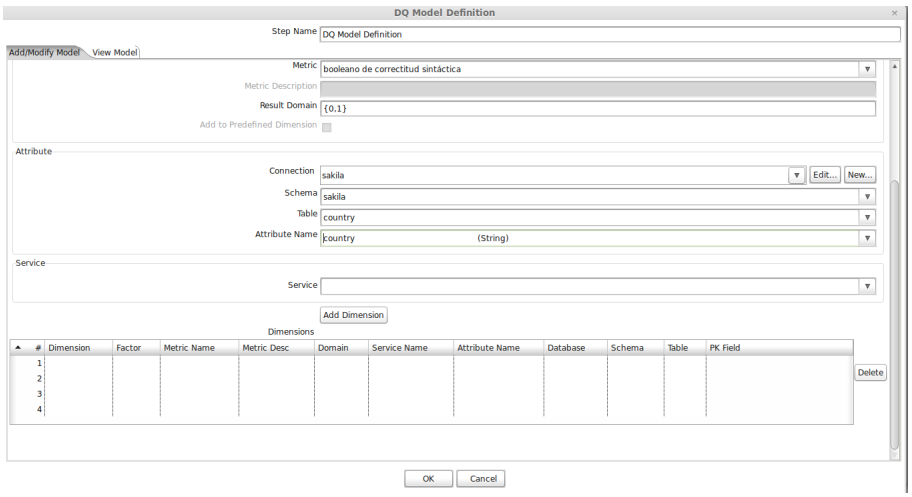

Figura A.2.6: Definición del atributo

A.2.1.4. Datos del servicio

Como se ve en la figura A.2.6 el siguiente recuadro a completar es "Service" el cual tiene la información del servicio que se quiere usar. Estos servicios ya estarían precargados en base a los servicios que se tienen disponibles, por lo tanto aquí solo se selecciona el servicio que queremos usar.

A.2.1.5. Agregar datos y visualización en grilla

Luego de haber completado todos los datos correspondientes a la dimensión, el factor y la métrica a utilizar, así como los datos del atributo a evaluar y del servicio a usar se debe asociar esta información al modelo. Esto se realiza seleccionando el botón "Add Dimension". Al asociar la información al modelo se agrega a la grilla que se encuentra abajo como se observa en la figura A.2.7. Si se quieren agregar más datos al modelo se vuelve a llenar los campos a partir del recuadro de "Dimension-Factor-Metric" y se selecciona una vez más el botón "Add Dimension" para agregar a la grilla.

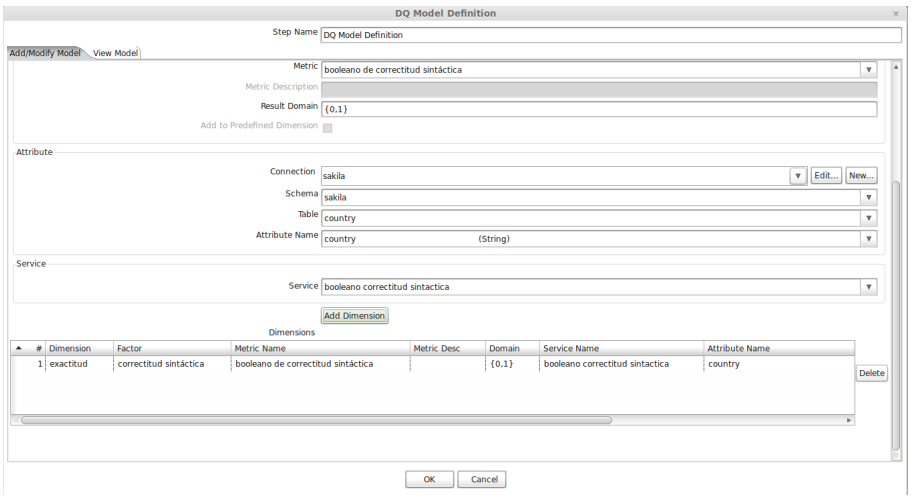

Figura A.2.7: Agregar información al modelo

También se permite agregar más información al modelo en una instancia posterior, para esto se debe guardar la transformación hecha en Pentaho Data Integration seleccionando el botón "OK". Cuando se quiera agregar más información se abre la transformación correspondiente y se selecciona el componente, ahí se podrá visualizar en la grilla la información de ese modelo almacenada temporalmente (si aún no se corrió la transformación).

Para almacenar el Modelo con su información se debe ejecutar la transformación hecha, esto se hace seleccionando el botón de ejecutar como se muestra en la figura A.2.8. Si todo funciona correctamente se observa un tic verde sobre el componente.

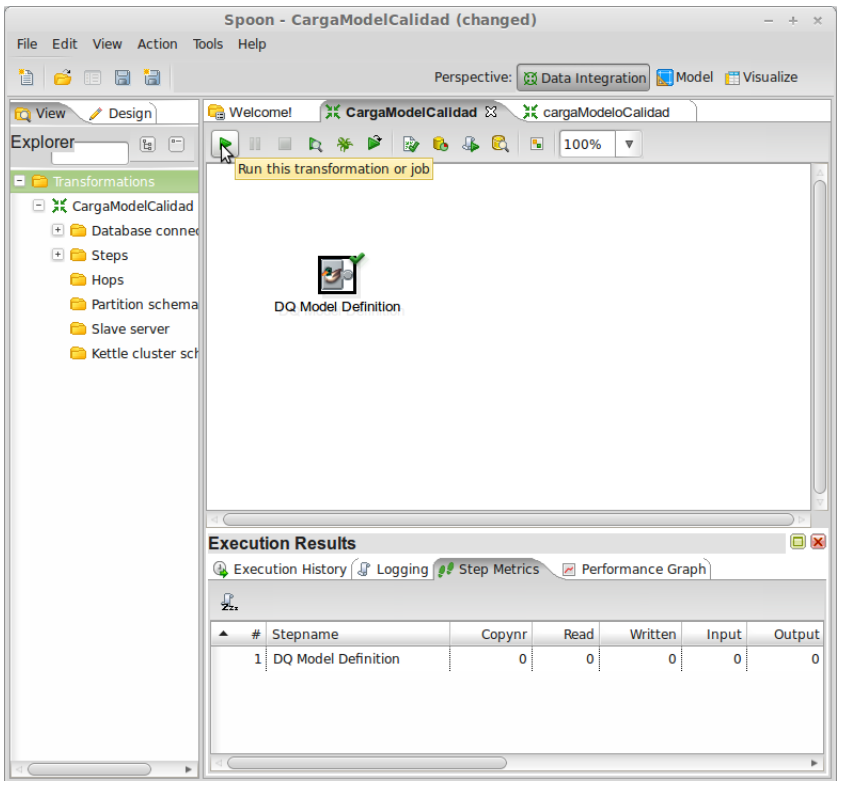

Figura A.2.8: Correr la transformación

A.2.2. Pestaña ver Modelos

Una vez agregado un modelo y ejecutada la transformación se puede visualizar la información de ese Modelo en la pestaña de "View Model", (figura A.2.9). Para esto se debe seleccionar la pestaña "View Model" y posteriormente se debe seleccionar un Modelo en el combo "Model Name".

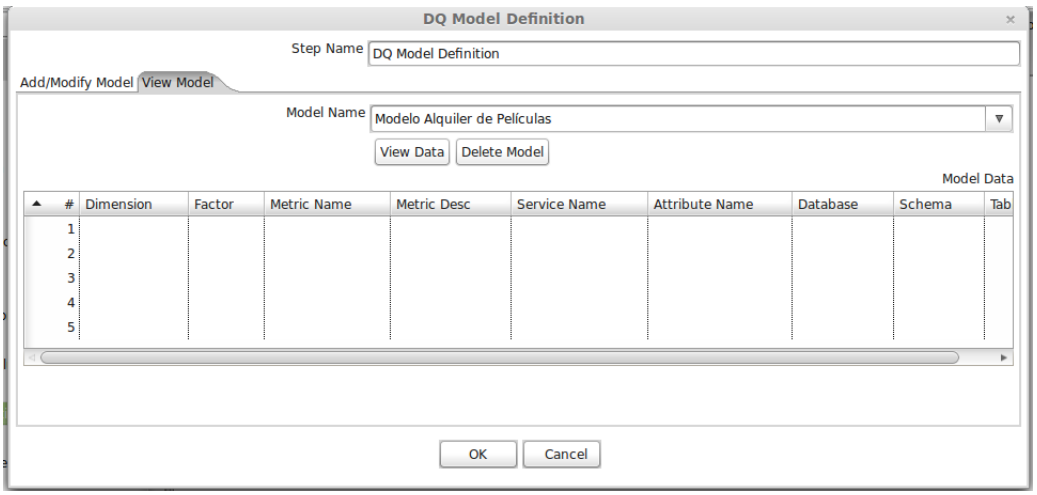

Figura A.2.9: Ver Modelos

Luego de esto se puede seleccionar el botón "View Data" para ver la información del Modelo seleccionado como se ve en la figura aparece la información relacionada con el Modelo seleccionado en la grilla de abajo.

Asimismo se puede borrar el modelo entero esto es presionando el botón "Delete Model". Cuando se selecciona el botón "Delete Model" se borra el Modelo de la base de datos directamente.

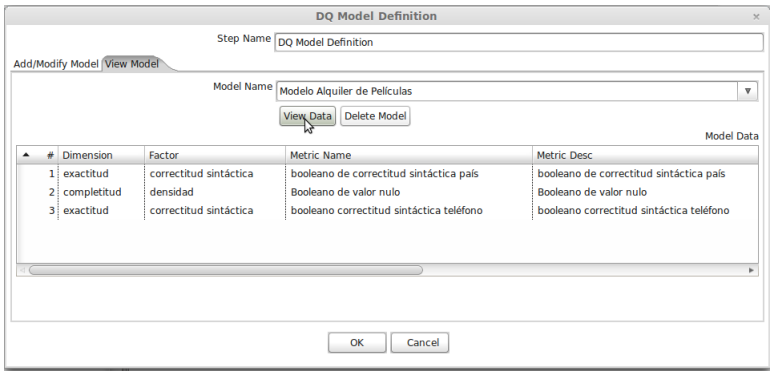

Figura A.2.10:

A.3. Evaluación de Calidad de Datos

El principal uso del componente de "Evaluación de Calidad de Datos" es para incorporarlo en un flujo de ETL.

En la figura A.3.1 se muestra cual es el componente que se debe seleccionar de la lista de "Steps". Luego de seleccionarlo se debe arrastrarlo hacia la transformación donde se desea agregar el componente.

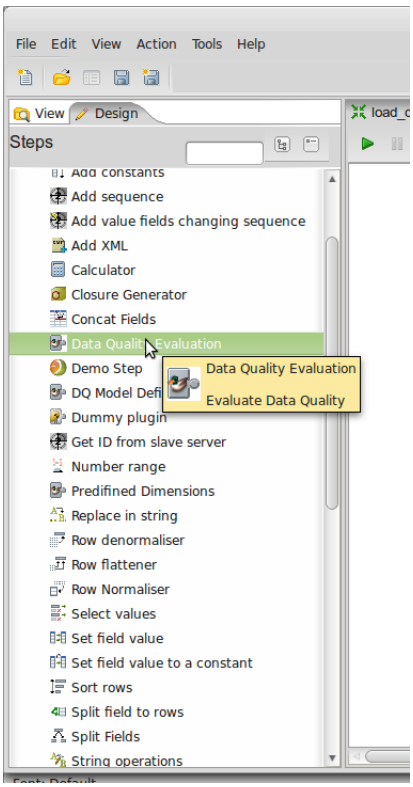

Figura A.3.1: Seleccionar componente de Evaluación de Calidad de Datos

En la figura A.3.2 se muestra el componente agregado a una transformación de carga de clientes, este componente es agregado al flujo de datos previo a que se almacenen los datos.

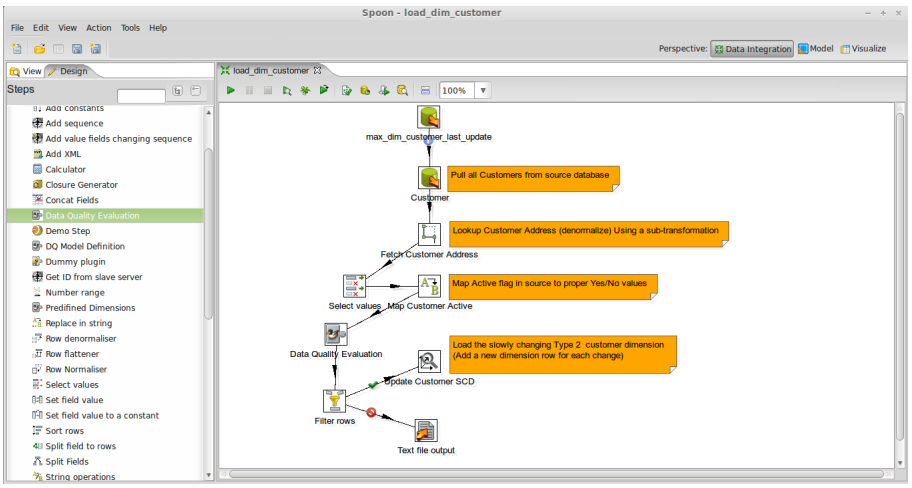

Figura A.3.2: Componente de Evaluación agregado a transformación de carga de clientes

Una vez agregado el componente al flujo de ETL, se debe configurar el mismo seleccionándolo ("Data Quality Evaluation") en la transformación. Una vez seleccionado aparece la pantalla de configuración, (figura A.3.3). En esta pantalla se debe seleccionar primero el Modelo en el campo "Model", luego seleccionar la dimensión, factor y métrica en el campo "Metric" y por último seleccionar el atributo asociado en el campo "Attribute". Cabe destacar que si en el flujo de entrada no hay ningún atributo para el que se haya definido una medición en el Modelo de Calidad no se mostrará nada.

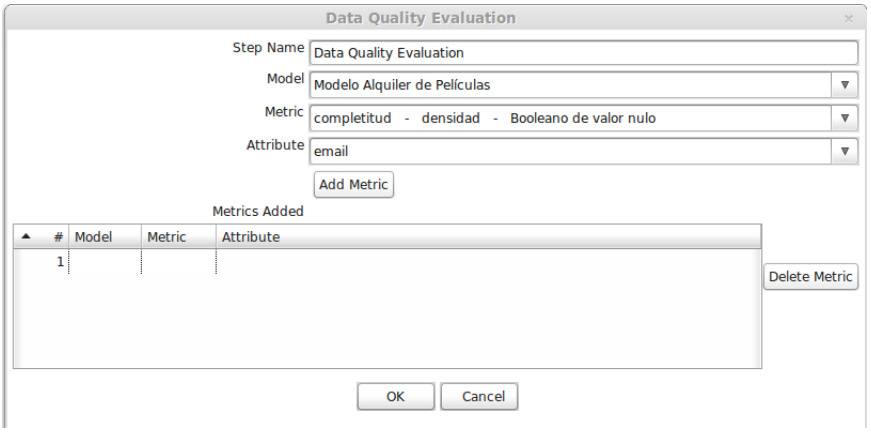

Figura A.3.3: Configuración de componente de Evaluación

Luego se debe seleccionar el botón "Add Metric" como se muestra en la figura A.3.4, al seleccionar el botón se agrega en la grilla de abajo el Modelo, la métrica y el atributo seleccionados.

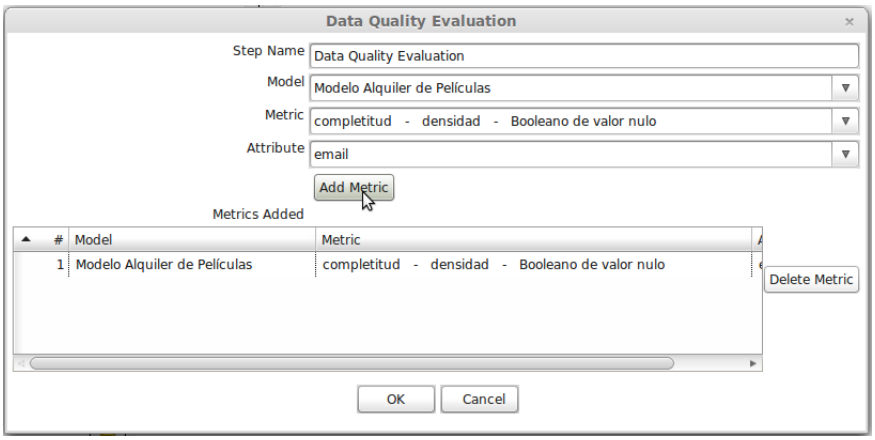

Figura A.3.4: Métricas y atributos agregados para evaluar

En la figura A.3.5 se muestra que se pueden tener varios atributos a evaluar al mismo tiempo. La información de esas mediciones va a formar parte de la salida junto con los datos de la tupla de entrada de este componente.

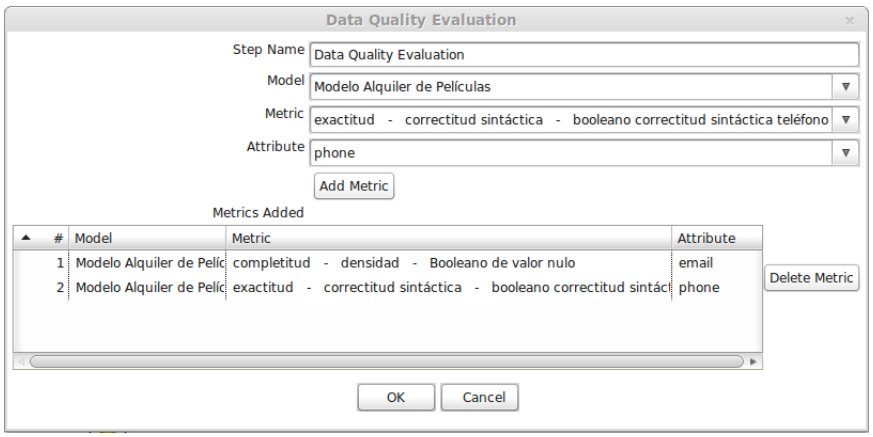

Figura A.3.5: Evaluación de Calidad para dos atributos