PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

DESARROLLO DE UN MODELO DE CALIDAD DE DATOS APLICADO A UNA SOLUCIÓN DE INTELIGENCIA DE NEGOCIOS EN UNA INSTITUCIÓN EDUCATIVA: CASO LAMBDA

Tesis para optar por el Título de Ingeniero Informático, que presenta el bachiller:

Marshall André Fernández Sáenz

ASESOR: Mg. Abraham Eliseo Dávila Ramón
COASESOR: Mg. Cecilia Yanett García García

Lima, 6 de Abril del 2018

Dedicatoria

A mi familia, por haberme apoyado y permitido llegar al final de mi carrera. Principalmente a mi madre, por haber sido un apoyo durante todo mi proceso de formación, y a mi padre por ser un soporte espiritual y por haber dejado un legado que me permitió llegar hasta este punto de mi carrera.

Agradecimientos

A mi asesor y co-asesora por guiarme en el desarrollo de este trabajo y darme las facilidades para su culminación.

A la Pontificia Universidad Católica del Perú y a los profesores que he tenido a lo largo de la carrera por aportarme los conocimientos que me permitieron finalizar la carrera y desarrollarme profesionalmente.

A todas las personas involucradas por todo el apoyo brindado en este proyecto.

Reconocimiento

Este trabajo ha sido realizado dentro del proyecto ProCal-ProSer (ProCal-ProSer: Determinación de factores que influyen en la PROductividad y CALidad en organizaciones que desarrollan PROductos software y ofrecen SERvicios software utilizando como base normas ISO en pequeñas organizaciones.) financiado por Innóvate Perú bajo el Contrato 210-FINCYT-IA-2013 y parcialmente soportado por el Departamento de Ingeniería y el Grupo de Investigación y Desarrollo de Ingeniería de Software (GIDIS) de la Pontificia Universidad Católica del Perú.

Resumen

La Inteligencia de Negocios comprende una serie de tecnologías que permiten la extracción, transformación y carga de datos hacia Data Warehouse preparados para recibir información útil para altos mandos de la organización. En la actualidad, las soluciones basadas en Inteligencia de Negocios permiten que las organizaciones de todo tipo cuenten con apoyo en la toma de decisiones de gran impacto.

Sin embargo, en un contexto en el que los datos están en rápido y constante crecimiento, las soluciones de Inteligencia de Negocios son vulnerables a los problemas de consistencia que puedan presentar los datos, por lo que los procedimientos realizados para su transformación llevan a la obtención de reportes inconsistentes que generan problemas en el manejo de la institución. Bajo esta situación, se proponen diversas herramientas para garantizar la calidad de los datos, como la descrita en el actual documento.

En el presente proyecto de tesis se propuso el desarrollo de un Modelo de Calidad de Datos que permite la evaluación de los datos que guarda un Data Warehouse, así como sus métricas relacionadas. Este modelo permitió la realización de investigaciones de calidad sobre la base de los parámetros especificados en la familia de normas ISO/IEC 25000, enfocándose especialmente en la ISO/IEC 25012 y la ISO/IEC 25024. Asimismo, se planificó su aplicación práctica en los datos relacionados a la solución de Inteligencia de Negocios de una institución elegida, la institución educativa Edu.Lambda.

Este proyecto forma parte de la iniciativa del Grupo de Investigación y Desarrollo de Ingeniería de Software de la Pontificia Universidad Católica del Perú (GIDIS-PUCP), dentro del marco del proyecto [Pro]ductividad y [Cal]idad en [Pro]ductos software y [Ser]vicios software (ProCal-ProSer).

El presente trabajo consta de 4 capítulos; en el primero, se realiza la formulación del proyecto al detalle de la problemática, objetivos, resultados esperados, herramientas y alcances; en el segundo, se detallan los conceptos relacionados a la investigación, así como los avances que se han realizado en la misma línea de trabajo; en el tercero, se detalla el modelo de calidad propuesto y la aplicación del mismo sobre la solución de

Inteligencia de Negocios de la institución mencionada; y en el cuarto, se muestran las conclusiones, observaciones y recomendaciones relacionadas al proyecto.

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO:	**Desarrollo de un Modelo de Calidad de Datos aplicado a una solución de Inteligencia de Negocios en una Institución Educativa. Caso Lambda**
ÁREA:	Ingeniería de Software
ASESOR:	Mg. Abraham Eliseo DÁVILA RAMÓN Mg. Cecilia Yanett GARCÍA GARCÍA
ALUMNO:	Marshall André FERNÁNDEZ SÁENZ
CÓDIGO:	20101287
TEMA N°:	#674
FECHA:	San Miguel, 20 de Mayo de 2016

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
Facultad de Ciencias e Ingeniería
29 NOV. 2017
MIGUEL MEJIA PUENTE
DECANO

DESCRIPCIÓN

La Inteligencia de Negocios (o BI de Business Intelligence, como se le conoce en inglés) consiste en el conjunto de métodos y herramientas diseñados para realizar el análisis de los datos que permite la planificación y toma de decisiones en una organización. Actualmente, se puede decir que la Inteligencia de Negocios es muy usada en las organizaciones, a todo nivel, tanto para la parte operativa como para la gerencial.

Las soluciones de Inteligencia de Negocios permiten el diseño de las estructuras y procedimientos que permiten la transformación y almacenamiento de los datos, los que se basan en los conceptos de proceso extracción-transformación-carga (o ETL, de Extract-Transform-Load en inglés) y Data Warehousing.

En la institución educativa Lambda se ha implementado una solución de Inteligencia de Negocios que permite el procesamiento de los datos usados en las operaciones rutinarias, tanto académicas como administrativas, para obtener distintos reportes e indicadores que pueden ser analizados para tomar decisiones que impliquen cambios importantes en la organización. Esta implementación ha sido utilizada por menos de un quinquenio de años, tiempo en el cual se han establecido nuevas reglas de negocio y se han adoptado nuevas tecnologías en las operaciones de la organización. Estos cambios han configurado un escenario con datos inconsistentes o incompletos que han ido apareciendo en diversos momentos, según algunos reportes requeridos.

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

En este proyecto se propone desarrollar un modelo de calidad de datos almacenados en el Data Warehouse de la solución de Inteligencia de Negocios de la institución Lambda.

OBJETIVO GENERAL

Desarrollar un modelo de calidad de datos para la solución de Inteligencia de Negocios utilizada en la institución educativa Lambda basado en la serie de normas ISO/IEC 25000.

OBJETIVOS ESPECÍFICOS

Los objetivos específicos son:
OE1. Identificar los registros de datos relevantes para el análisis.
OE2. Seleccionar características y sub características de calidad relevantes para los datos identificados.
OE3. Definir métricas adecuadas para evaluar la calidad de datos.
OE4. Evaluar los resultados de la aplicación del modelo en la institución.

ALCANCE

El proyecto implica la definición y aplicación de un modelo de calidad de datos derivado de la serie de estándares ISO/IEC 25000, en particular la ISO/IEC 25012 que define un marco de calidad de producto software orientado a los datos. Con dicho modelo se evaluará los datos obtenidos de una solución de Inteligencia de Negocios (es decir, el Data Warehouse) de la institución educativa Lambda.
Para la definición del modelo y la evaluación de las métricas se utilizará los protocolos establecidos en el propio Estándar. Asimismo, para la evaluación se trabajará con la base de datos de Lambda que soporta los procesos operacionales de la Organización y de una solución existente de Inteligencia de Negocios cuyas reglas han variado en el tiempo. Además se circunscribirá al contexto académico.
Dado que el modelo de calidad se está realizando en el contexto de una organización específica, será único, por lo que su elaboración y los resultados dependerán mucho de este factor.

Máximo: 100 páginas

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

Contenido

Índice de Figuras

Índice de Tablas

1. Formulación del Proyecto

En este capítulo se presenta el proyecto, orientándolo a partir del contexto en que se realiza y el problema que debe ser abordado en el trabajo de tesis, así como los objetivos, el alcance y los parámetros que se seguirán para su realización.

1.1. Problemática Contextualizada

Toda organización, para su funcionamiento, requiere que en sus niveles estratégicos se tomen decisiones que sean útiles para su funcionamiento adecuado. Sin embargo, muchas de estas decisiones serían muy difíciles sin herramientas de Inteligencia de Negocios que apoyen al análisis e interpretación de la información que manejan los altos mandos.

Según predicciones realizadas, al 2017 la mayor parte de los analistas y usuarios en las organizaciones contarán con herramientas de Inteligencia de Negocios que den servicio en la misma organización para preparar información para su respectivo análisis (Gartner Inc., 2015), dado que la generación de reportes todavía es ampliamente considerada como un impulso de la innovación en las organizaciones (Forrester Consulting, 2014).

Se puede decir que en una organización, existen múltiples procesos que pueden manejarse mediante el uso de soluciones de Inteligencia de Negocios a todo nivel (Kimball & Ross, 2013).

En la parte educativa, existen múltiples procesos que pueden manejarse mediante el uso de una de estas soluciones, tanto al nivel directamente relacionado con la enseñanza como a un nivel administrativo para el soporte de las labores que realiza una organización (Kimball & Ross, 2013). Esto se justifica dado que gracias a las soluciones se cuenta con un apoyo para que una institución educativa aumente su efectividad y reduzca sus costos de operación, así como un respaldo para competir frente a otras instituciones similares (Qlik, 2016).

Actualmente las instituciones educativas requieren de medidas que les permitan hacer la evaluación de sus indicadores, principalmente para conocer la efectividad de su

forma de enseñanza y el impacto que pueden tener sobre su alumnado (Kimball & Ross, 2013).

En las instituciones educativas, la Inteligencia de Negocios es útil para su uso en la evaluación del rendimiento del alumnado y la obtención de datos sobre los estudiantes, docentes y los procedimientos administrativos de la institución. Los datos pueden ser clasificados en tres funciones primarias para una institución educativa: investigación, educación y servicios (N. A. H. M. Rodzi, M. S. Othman, & L. M. Yusuf, 2015).

Asimismo, es necesario enfocarse en la Calidad de Datos como una disciplina que permite garantizar que la información esté en condiciones adecuadas para apoyar al éxito de una organización, especialmente en un entorno en que la cantidad de datos utilizados va en aumento (ISO, 2008) y en el que es muy común que los datos sean de baja calidad (Redman, 2002), lo cual se refleja en la vulnerabilidad crítica de las soluciones de Inteligencia de Negocios a inconsistencias (Dorsett, 2014).

Específicamente, se puede observar el caso de la institución Edu.Lambda, elegida para su estudio. La necesidad de poder analizar rápidamente la información sobre toda la organización y sus actividades obligó al uso de diversas herramientas informáticas que permitan el procesamiento y la presentación de los datos de forma automática e intuitiva para los usuarios de este sistema. Se dotó a la empresa con una solución que puede realizar procesamiento OLAP (On-Line Analytic Processing, de Procesamiento Analítico en Línea en inglés), eliminando la dependencia de un área de informática y la réplica de esfuerzos.

Desde hace menos de 5 años hasta el día de hoy, Edu.Lambda trabajó y continúa trabajando con una solución de Inteligencia de Negocios ideada para procesar y obtener información respecto a sus temas operacionales y administrativos que es útil para que la alta dirección pueda tomar decisiones. Esta solución está conformada por un Data Warehouse que permite el almacenamiento de los datos procesados, y herramientas e infraestructura que soportan el diseño y ejecución de procesos ETL (Extract-Transform-Load, de Extracción-Transformación-Carga en inglés) y diversos reportes. Además, es mantenida por un equipo de trabajo, que se encarga de administrar el Data Warehouse, gestionar los procesos ETL que se encargan de su llenado y elaborar las aplicaciones que explotan los datos procesados.

A lo largo de este tiempo, dada la aplicación de nuevas tecnologías y reglas de negocio en la organización, la solución, así como los datos que se almacenaban, fueron creciendo a la par con todos los sistemas de la organización, lo cual genera el aumento de la complejidad de las fuentes de información utilizadas para alimentar a la solución de Inteligencia de Negocios. Debido a esto, siempre se busca que los datos utilizados sean de calidad, de tal modo que se mantenga su consistencia; sin embargo, en diversos momentos de tiempo surgieron diversas inconsistencias propias de estos cambios, las cuales se manifestaban en los reportes solicitados, un efecto propio de la pobre calidad de los datos, lo cual es manifestado por varios autores (Azvine, Cui, Nauck, & Majeed, 2006; Chaudhuri, Dayal, & Narasayya, 2011; Yeoh & Koronios, 2010).

Por esta razón, los encargados de la solución han considerado mitigar el riesgo de la posible reducción de la calidad de datos, debido a que esto puede ocasionar una pérdida de tiempo en la corrección de estos errores y puede llevar a la toma incorrecta de decisiones por parte de los directivos que utilizan la información, lo cual deriva en una pérdida de confianza hacia el equipo que mantiene la solución. Además, se busca una manera de contribuir a la mejora continua de la organización, al evaluar los datos de manera más precisa y garantizar su calidad de forma permanente, evitando inconsistencias futuras.

Es por este motivo que se llegó a la necesidad de abordar el problema del mejoramiento de Calidad de Datos en la solución de Inteligencia de Negocios de esta organización, utilizando métodos basados en ingeniería.

Actualmente, se tiene el estándar de calidad ISO/IEC 25000, conocido como SQuaRE; específicamente, se tiene la división ISO/IEC 25012, enfocada en la calidad de datos, que permite el manejo de requisitos y medidas de calidad de datos y la planificación de evaluaciones (ISO, 2008). En base a las normas mencionadas, el propósito de este proyecto es desarrollar un modelo de calidad de Datos que permita evaluar los datos del Data Warehouse de la institución Edu.Lambda.

1.2. Objetivo General

Desarrollar un modelo de calidad de datos para la solución de Inteligencia de Negocios utilizada en la institución educativa Edu.Lambda basado en normas internacionales ISO/IEC.

1.3. Objetivos Específicos

Los objetivos específicos son:

OE1. Identificar los registros de datos relevantes para el análisis.
OE2. Seleccionar características y sub características de calidad relevantes para los datos identificados.
OE3. Definir métricas adecuadas para evaluar la calidad de datos.
OE4. Evaluar los resultados de la aplicación del modelo en Edu.Lambda.

1.4. Resultados Esperados

Los resultados esperados son:

- Para OE1:

RE1. Lista de registros de datos identificados sobre los cuales se aplicará el modelo de calidad.

- Para OE2:

RE2. Modelo de Calidad de Datos a nivel de Características y Sub características para la institución educativa Edu.Lambda.

- Para OE3:

RE3. Documento de métricas utilizadas para la evaluación de calidad con sus definiciones y fórmulas respectivas.
RE4. Scripts para obtener los resultados de las métricas definidas.

- Para OE4:

RE5. Plan de evaluación de los datos de la institución educativa Edu.Lambda.

RE6. Reporte de la evaluación de calidad en la organización.

1.5. Herramientas, Métodos y Procedimientos

Según su necesidad para obtener los resultados esperados, las herramientas a utilizar se muestran en la Tabla 1.1.

Tabla 1.1. Herramientas a utilizar por resultado esperado a conseguir.
Cuadro diseñado por el autor.

Resultado Esperado	Herramienta
RE1. Lista de registros de datos identificados sobre los cuales se aplicará el modelo de calidad.	• Modelo de datos de la organización. • Método de clasificación manual para selección de registros. • Juicio experto.
RE2. Modelo de Calidad de Datos a nivel de Características y Sub características para la institución educativa Edu.Lambda.	• Norma ISO/IEC 25000. • Norma ISO/IEC 25010. • Norma ISO/IEC 25012. • Juicio experto. • Técnica de Grupo Nominal. • Método de Jim Brosseau.
RE3. Documento de métricas utilizadas para la evaluación de calidad con sus definiciones y fórmulas respectivas.	• Norma ISO/IEC 25024.
RE4. Scripts para obtener los resultados de las métricas definidas.	• Norma ISO/IEC 25024.
RE5. Plan de evaluación de los datos de la institución educativa Edu.Lambda.	• Norma ISO/IEC 25040.
RE6. Reporte de la evaluación de calidad en la organización.	• Norma ISO/IEC 25040. • Scripts para obtención de resultados de métricas.

1.6. Alcance y Limitaciones

1.6.1. Alcance

El proyecto de tesis abarca la definición y aplicación de un modelo de calidad de datos, tomando como base la familia de normas ISO/IEC 25000. Principalmente, se utilizó el estándar ISO/IEC 25012, el cual define un marco de calidad de producto de software orientado a datos.

Mediante el uso del modelo desarrollado se evaluaron los datos utilizados en la solución de Inteligencia de Negocios utilizada en la institución educativa Edu.Lambda almacenados en el Data Warehouse según lo mostrado en la Figura 1.1.

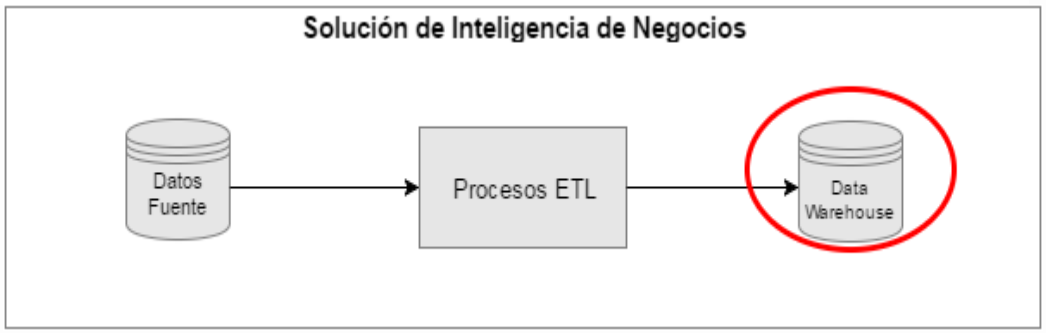


Figura 1.1. Datos elegidos para la medición de la calidad.
Figura diseñada por el autor.

1.6.2. Limitaciones

Las limitaciones de este proyecto son:

- Se consideraron exclusivamente las características definidas en la ISO/IEC 25012 y las métricas y metodologías consideradas en la familia de normas ISO/IEC 25000.

- Debido al tamaño de las bases de datos utilizadas, sólo se eligió una cantidad limitada de registros de datos para hacer la evaluación de la calidad. Los datos elegidos estarán asociados a la solución de Inteligencia de Negocios a evaluar, por ser el resultado de los procesos en el Data Warehouse y de acuerdo a lo obtenido según la clasificación a realizar para conocer su relevancia en conjunto con los expertos de la organización en mención.

- El trabajo se realizará respetando los parámetros del acuerdo de confidencialidad suscrito entre la institución educativa Edu.Lambda y el grupo GIDIS-PUCP.

1.7. Riesgos

Los riesgos identificados en el proyecto se muestran en la Tabla 1.2.

Tabla 1.2. Matriz de riesgos identificados.
Cuadro diseñado por el autor.

Riesgo	Impacto	Medidas correctivas
Modelo de datos utilizado sufre una caída o pérdida.	Pérdida de tiempo, demora en entrega de resultados.	Obtener una copia de los datos a la brevedad posible para realizar el trabajo fuera de línea.
Eliminación de accesos a los datos a evaluar.	Pérdida de tiempo, demora en entrega de resultados.	Obtener una copia de los datos a la brevedad posible para realizar el trabajo fuera de línea.
Pérdida de vigencia de las normas ISO/IEC utilizadas.	Pérdida de tiempo por búsqueda de nuevas normas actualizadas, replanteamiento del proyecto.	Actualizarse constantemente para conocer cambios en las normas ISO/IEC utilizadas.
Resultados erróneos en las métricas de calidad utilizadas.	Resultados no útiles.	Auditoría de los resultados, volver a obtener las métricas de ser necesario

1.8. Justificación

El presente proyecto implica la obtención de experiencia que puede ser útil para la definición y evaluación de modelos de calidad en instituciones educativas.

Este proyecto forma parte de la línea de investigación de Calidad de Producto del proyecto [Pro]ductividad y [Cal]idad en [Pro]ductos software y [Ser]vicios software (ProCal-ProSer) del Grupo de Investigación y Desarrollo en Ingeniería de Software de la Pontificia Universidad Católica del Perú (GIDIS-PUCP). En particular, se tiene previsto el desarrollo de modelos de calidad de datos usando como referencia normas ISO/IEC especialmente adaptadas para organizaciones educativas.

1.9. Viabilidad

1.9.1. Viabilidad Técnica

Para la preparación del proyecto, se han realizado ya las coordinaciones con la institución para tener el acceso a los datos a evaluar, dado que se tiene actualmente la autorización pertinente de las autoridades a cargo, por lo que técnicamente el proyecto será viable.

1.9.2. Viabilidad Temporal

Dadas las exigencias del grupo GIDIS-PUCP, se estima que el proyecto debe estar terminado en un máximo de 3 meses, por lo que estará finalizado para la fecha programada para su revisión.

1.9.3. Viabilidad Económica

El proyecto es viable económicamente dado que los recursos e información necesarios para realizar el desarrollo del modelo de calidad son brindados por el grupo GIDIS-PUCP.

2. Marco de Referencia

Para poder apoyar la necesidad de realizar este trabajo de tesis, se realiza una presentación de los conceptos utilizados en el trabajo, así como una revisión del estado de la práctica respecto a la calidad de datos en Inteligencia de Negocios y algunos detalles respecto a los parámetros legales que se están considerando en el trabajo de tesis.

2.1. Fundamentos de Inteligencia de Negocios

2.1.1. Inteligencia de Negocios

La Inteligencia de Negocios es el conjunto de herramientas y métodos que se utilizan para transformar datos operacionales en información útil para la toma de decisiones en una organización (Kimball & Ross, 2002). Esto permite que los datos emitidos permitan guiar a la organización hacia una meta deseada (Luhn, 1958).

2.1.2. Data Warehouse

Un Data Warehouse (Almacén de Datos en inglés) es una colección de datos orientados a un tema, integrados, no volátiles y variables en el tiempo (Inmon, 2005) que contiene información procesada para que sus usuarios tengan un fácil entendimiento de ella, sea consistente, soporte cambios en la organización, proteja la información y mejore la toma de decisiones (Kimball & Ross, 2002).

Para llenar esta colección se realiza la extracción, limpieza, verificación y envío de los datos fuente a un almacén dimensional que permite la consulta y el análisis para propósitos de toma de decisiones (Kimball & Caserta, 2004). El éxito de la implementación de un Data Warehouse depende de sus usuarios, quienes deben considerar que esta herramienta pueda satisfacer sus necesidades (Kimball & Caserta, 2004).

2.1.3. Modelamiento multidimensional

Se dice que una base de datos está modelada de forma multidimensional cuando la data es presentada en cubos (estructuras dimensionales para el procesamiento de datos) en lugar de un modelo relacional. (Kimball & Ross, 2002)

Un Data Warehouse puede estar modelado de esta manera si es que es manejado con esquemas de tipo "estrella", "copo de nieve" o un híbrido de ambos, consistente en una serie de "dimensiones" (entidades utilizadas para una mejor comprensión de los datos) enlazados a una o más "fact tables" (tablas de hechos en inglés, las cuales representan lo que se analizará), como se muestra en las Figuras 2.1 y 2.2. (Kimball & Ross, 2002).

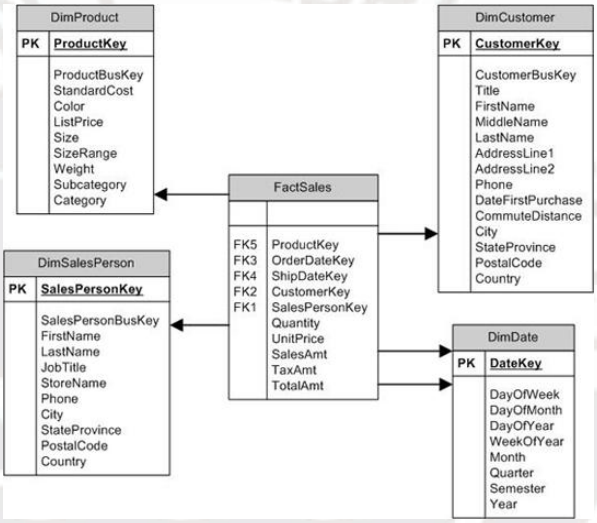


Figura 2.1. Uso de esquema tipo Estrella.
(Sheldon, 2008b)

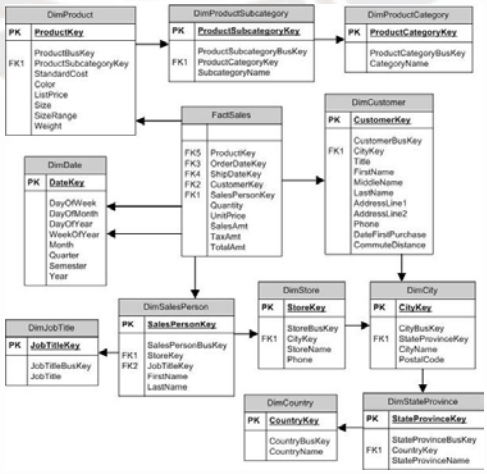


Figura 2.2. Uso de esquema tipo Copo de Nieve.
(Sheldon, 2008a)

2.1.4. Procesos ETL (Extract, Transform, Load)

Se conoce como proceso ETL a aquel que permite el procesamiento de los datos tomando como entrada los datos transaccionales encontrados en una base de datos fuente, y dando como salida la información que se cargará en el Data Warehouse.

El nombre que se les da a estos procesos es lo que define lo que harán: se empieza extrayendo los datos de los sistemas fuente y almacenando los datos de tal manera que estén listos para realizar una transformación (Kimball & Ross, 2002).

Posteriormente, se realiza la transformación de los datos, la cual se hace en dos partes: la limpieza, que se realiza para eliminar inconsistencias que haya habido en los datos fuente y normalizar la información, de tal manera que la información tenga calidad y sea consistente, y la transformación en sí, en la que se hacen los cambios a los datos que sean requeridos por las reglas de negocio y lo solicitado por los usuarios, así como la unión de los datos de múltiples fuentes (Kimball & Ross, 2002).

Finalmente, los datos son cargados en el Data Warehouse para su posterior presentación a los usuarios a través de la explotación mediante diversas herramientas (Kimball & Ross, 2002).

2.2. Conceptos de Calidad

2.2.1. Concepto de Calidad de Datos

La Calidad de Datos puede definirse como el estado de completitud, validez, consistencia, oportunidad y precisión, entre otras propiedades, que hace los datos apropiados para un uso específico al satisfacer las necesidades especificadas e implícitas bajo ciertas condiciones específicas (ISO, 2008; Roebuck, 2011). Por lo general, se puede decir que los datos son correctos si es que están aptos para ser usado en las operaciones, toma de decisiones y planificación del negocio (Redman, 2002).

Para garantizar la precisión de los datos: estos tienen que ser (Kimball & Caserta, 2004):

- Correctos (sus valores tienen que representar fielmente a los objetos que representan).
- No ambiguos (los valores solo pueden tener un significado).
- Consistentes (los valores deben guiarse por un único estándar general).
- Completos (los valores deben estar completos y no deben perderse durante su uso).

Se deben realizar procesos de evaluación y aseguramiento de calidad para garantizar estas características (Kimball & Caserta, 2004).

2.2.2. Serie de normas ISO/IEC 25000

La serie de normas ISO/IEC 25000, conocida como SQuaRE (Systems and Software Quality Requirements and Evaluation, en inglés) fue creada con el objetivo de mejorar y unificar las normas relacionadas a las ISO/IEC 9126 e ISO/IEC 14598, es decir, el proceso de especificación de requisitos de calidad de software y el proceso de evaluación de calidad de software (ISO, 2005).

Esta norma se divide en 5 divisiones, las cuales se presentan en la figura 2.3 y se presentarán a continuación (ISO, 2005).

Figura 2.3. Divisiones de la norma SQuaRE.
Traducido y adaptado de (ISO, 2005)

- División de Gestión de la Calidad: Define todos los estándares que tienen en común todas las normas en la serie SQuaRE.

- División de Modelo de Calidad: Define un modelo de calidad incluyendo características para calidad externa, interna y en uso. También provee sub características y guía en el uso del modelo.

- División de Medición de la Calidad: Define un modelo de referencia para la medición de la calidad de producto de software, así como definiciones matemáticas de medidas de calidad.

- División de Requerimientos de Calidad: Ayuda a la especificación de requerimientos de calidad.

- División de Evaluación de la Calidad: Aporta requerimientos, recomendaciones y guía para la evaluación del producto de software.

2.2.3. ISO/IEC 25010: Modelos de Calidad

La división ISO/IEC 25010 provee los parámetros básicos para la realización de un modelo de calidad.

Proporciona 3 modelos: el modelo de calidad de uso, el modelo de calidad de producto de software y el modelo de calidad de datos; este último se encuentra en la ISO/IEC 25012.

En la Figura 2.4 se muestra la estructura que utilizan los modelos de calidad mencionados según el estándar ISO/IEC 25010. En los modelos se evalúan una serie de características, las cuales son divididas en sub características y propiedades de calidad que posteriormente son medidas mediante su uso.

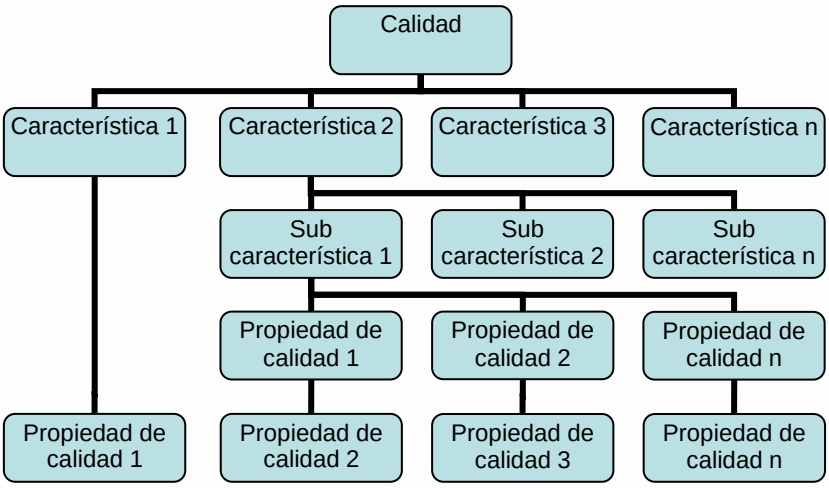


Figura 2.4. Estructura de los modelos de calidad del estándar ISO/IEC 25010. Traducido y adaptado de (ISO, 2011)

Asimismo, se define hacia dónde están dirigidos los modelos de calidad, lo cual se muestra en la Figura 2.5. Se discute los tres modelos de calidad propuestos y se indica hacia qué tipo de sistema debería estar enfocado, así como los factores que influyen sobre el modelo de calidad en uso.

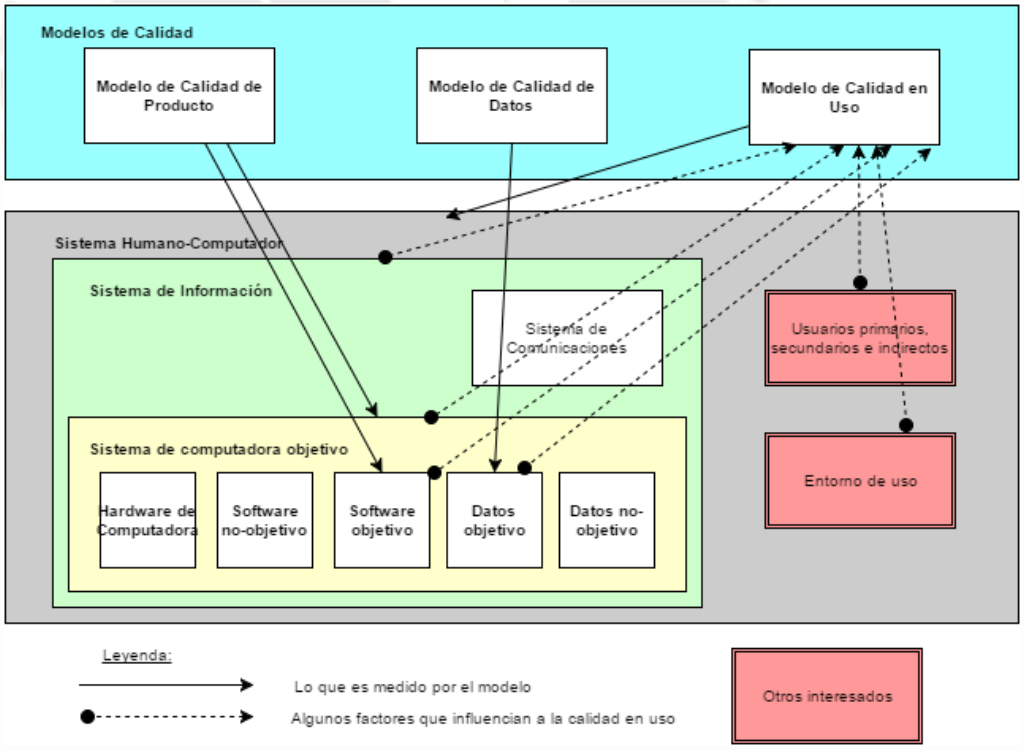


Figura 2.5. Objetivo de los modelos de calidad del estándar ISO/IEC 25010. Traducido y adaptado de (ISO, 2011)

2.2.4. ISO/IEC 25012: Modelo de Calidad de Datos

La norma ISO/IEC 25012:2008 es parte de la serie de normas ISO/IEC 25000, específicamente parte de la división 2501n correspondiente a Modelos de Calidad.

Este modelo está dirigido a los "datos objetivo", es decir, los datos que la organización decide analizar y validar a través del modelo (ISO, 2008).

El modelo proporciona un marco para la especificación y evaluación de requisitos de calidad de datos en términos de las propiedades del sistema en un entorno particular (Bevan, 2010).

En esta norma se definen dos puntos de vista respecto a la calidad de datos:

- Inherente: Cuando la calidad se refiere a los datos en sí mismos (Natale, 2010), indica si las características de calidad de los datos tienen el potencial de satisfacer las necesidades de los usuarios cuando los datos se usan bajo ciertas condiciones (Rafique, Lew, Qanber Abbasi, & Li, 2012). En esta parte se pueden evaluar los valores del dominio de datos y sus restricciones, las relaciones entre los valores de los datos (es decir, su consistencia) y la metadata correspondiente (ISO, 2008).

- Dependiente del sistema: Evalúa el grado en que la calidad de datos es alcanzada dentro de un sistema informático, por lo cual depende del entorno tecnológico en que se usen los datos (Natale, 2010). Esto se logra a partir de las capacidades del hardware y software que utilicen los datos (ISO, 2008)

En la Figura 2.6 se muestran las características de calidad que se siguen en la ISO/IEC 25012, las cuales serán detalladas a continuación (ISO, 2008; Natale, 2010). Estas características pueden pertenecer al punto de vista inherente, dependiente del sistema o a ambos.

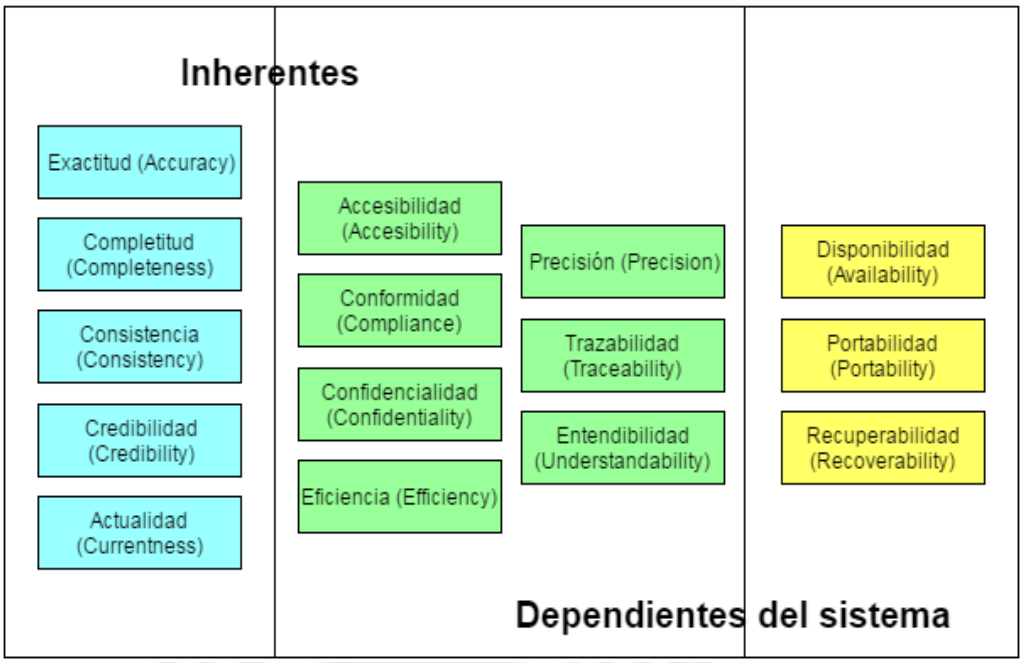


Figura 2.6. Características utilizadas para medir la calidad de datos en la ISO/IEC 25012:2008.
Traducido y adaptado de (*Características de Calidad de Producto de Datos en la ISO 25012*, 2016)

- Características Inherentes

  - Exactitud: Los datos representan de forma correcta el verdadero valor de los atributos de un concepto en un contexto específico de uso.

    Para este caso, la precisión puede ser:

    - Semántica: Cercanía de los datos a un juego de valores definidos en un dominio considerado semánticamente correcto (es decir, correctos en significado).

    - Sintáctica: Cercanía de los datos a un juego de valores definidos en un dominio considerado sintácticamente correcto (es decir, correctos en la forma en que se cumplen las reglas de sintaxis para los valores).

  - Completitud: Los datos tienen valores para todos los atributos e instancias esperadas en un contexto específico de uso.

  - Consistencia: Los datos están libres de contradicciones y son coherentes con el resto de datos en un contexto específico de uso.

- Credibilidad: Los usuarios consideran que los datos son creíbles y verdaderos en un contexto específico.
- Actualidad: Los datos tienen un tiempo adecuado para el contexto específico en que se utilicen.

- Características inherentes y dependientes del sistema

  - Accesibilidad: Se puede acceder a los datos en un contexto específico de uso, en especial por personas con discapacidades que requieran el uso de tecnología.
  - Conformidad: Los datos se adhieren a estándares, convenciones o normas en un contexto específico de uso.
  - Confidencialidad: Los datos sólo son accesibles e interpretables por los usuarios autorizados en un contexto específico de uso.
  - Eficiencia: Los datos pueden ser procesados y proporcionan el nivel de rendimiento esperado mediante el uso de una cantidad apropiada de recursos en un contexto específico de uso.
  - Precisión: Los datos son exactos o son discriminativos en un contexto específico de uso.
  - Trazabilidad: Los datos proporcionan la información necesaria para poder auditar los accesos y las modificaciones que se les han realizado en un contexto específico de uso.
  - Comprensibilidad: Los datos pueden ser leídos e interpretados por los usuarios, y son expresados en lenguajes apropiados, símbolos y unidades en un contexto específico de uso.

- Características dependientes del sistema

  - Disponibilidad: Los datos pueden ser recuperados por los usuarios autorizados o aplicaciones en un contexto específico de uso.

- Portabilidad: Los datos pueden ser instalados, reemplazados o movidos de un sistema a otro preservando la calidad existente en un contexto específico de uso.

- Recuperabilidad: los datos mantienen y preservan un nivel especificado de operaciones y de calidad, incluso en caso de fallo en un contexto específico de uso.

2.2.5. ISO/IEC 25024: Métricas para evaluar la calidad de datos

La norma ISO/IEC 25024:2015 define métricas de calidad de datos para poder evaluar el cumplimiento de las características definidas en el estándar ISO/IEC 25012 (ISO, 2015).

En la Tabla 2.1 se muestran las métricas definidas (ISO, 2015).

Tabla 2.1. Métricas de la ISO/IEC 25024.
Traducido y adaptado de (ISO, 2015).

Característica	Métrica
Exactitud	Exactitud sintáctica de los datos
	Exactitud semántica de los datos
	Aseguramiento de la Exactitud de los datos
	Riesgo de inexactitud de un conjunto de datos
	Exactitud del modelo de datos
	Exactitud de la metadata
	Rango de Exactitud de datos
Completitud	Completitud de registros
	Completitud de atributos
	Completitud de archivos de datos
	Completitud de valores de datos
	Registros vacíos en un archivo de datos
	Completitud del modelo de datos conceptual
	Completitud de los atributos del modelo de datos conceptual
	Completitud de la metadata
Consistencia	Integridad referencial
	Consistencia del formato de datos
	Riesgo de inconsistencia de datos
	Consistencia de la arquitectura
	Cobertura de la consistencia de valores de datos
	Consistencia semántica
Credibilidad	Credibilidad de los valores
	Credibilidad de las fuentes
	Credibilidad del diccionario de datos
	Credibilidad del modelo de datos
Actualidad	Frecuencia de actualización
	Oportunidad de la actualización
	Requisito de actualización de items

Característica	Métrica
Accesibilidad	Accesibilidad de usuario
	Accesibilidad de dispositivos
	Accesibilidad del formato de datos
Conformidad	Conformidad regulatoria de los valores y/o formato de datos
	Conformidad regulatoria debido a la tecnología
Confidencialidad	Uso de encriptación
	No vulnerabilidad
Eficiencia	Formato de ítems de datos eficiente
	Eficiencia usable
	Eficiencia en formato de datos
	Eficiencia en procesamiento de datos
	Riesgo de espacio desperdiciado
	Espacio ocupado por duplicación de registros
	Desfase temporal de actualización de datos
Precisión	Precisión de valores de datos
	Precisión del formato de datos
Trazabilidad	Trazabilidad de valores de datos (inherente)
	Trazabilidad del acceso a los datos
	Trazabilidad de valores de datos (dependiente del sistema)
Entendibilidad	Entendibilidad de símbolos
	Entendibilidad semántica
	Entendibilidad de datos maestros
	Entendibilidad de valores de datos
	Entendibilidad del modelo de datos
	Entendibilidad de la representación de datos
	Entendibilidad de datos maestros enlazados
Disponibilidad	Ratio de disponibilidad de datos
	Probabilidad de datos disponibles
	Disponibilidad de elementos de arquitectura
Portabilidad	Ratio de portabilidad de datos
	Portabilidad de datos prospectivos
	Portabilidad de elementos de arquitectura
Recuperabilidad	Ratio de recuperabilidad de datos
	Backup periódico
	Recuperabilidad de la arquitectura

2.2.6. ISO/IEC 25040: Proceso de Evaluación de Calidad

La norma ISO/IEC 25040:2002 brinda requerimientos y recomendaciones para la evaluación de la calidad de producto de software, complementándose con las normas de la línea ISO/IEC 2504n (ISO, 2002).

En la norma se especifica el proceso de evaluación de Calidad de Producto de Software, el cual se muestra en la Tabla 2.2.

Tabla 2.2. Proceso de Evaluación de Calidad.
Traducido y adaptado de (ISO, 2002).

Proceso de Evaluación de Calidad – ISO/IEC 25040	
Establecer los requerimientos de la evaluación	1. Establecer el propósito de la evaluación
	2. Obtener los requerimientos de calidad de productos de software
	3. Identificar partes del producto a ser incluidas en la evaluación
	4. Definir el rigor de la evaluación
Especificar la evaluación	1. Seleccionar métricas de calidad (módulos de evaluación)
	2. Definir criterios de decisión para métricas de calidad
	3. Establecer criterios de decisión para la evaluación
Diseñar la evaluación	1. Planificar actividades de evaluación
Ejecutar la evaluación	1. Realizar medición
	2. Aplicar criterios de decisión para métricas de calidad
	3. Aplicar criterios de decisión para la evaluación
Concluir la evaluación	1. Revisar los resultados de la evaluación
	2. Crear el reporte de evaluación
	3. Revisar la evaluación de calidad y proporcionar feedback a la organización
	4. Disponer de los datos de la evaluación

2.3. Herramientas para obtención de elementos de calidad

En esta parte se presentan las técnicas utilizadas para la obtención de los elementos a considerar en el modelo de calidad resultante.

2.3.1. Técnica de Grupo Nominal

La Técnica de Grupo Nominal es un método que permite realizar una lluvia de ideas en un grupo con la contribución de todos los participantes (American Society for Quality, 2004).

Para su realización, se deben seguir los siguientes pasos (American Society for Quality, 2004):

1. Indicar el propósito de la evaluación, la pregunta a responder y clarificar las dudas hasta que sea entendido.

2. Cada participante escribe las ideas que tiene sobre la pregunta propuesta en silencio, en un lapso de tiempo determinado (de 5 a 10 minutos).

3. Los participantes mencionan sus ideas en voz alta.

4. Las ideas de los participantes son discutidas en grupo una por una.

5. Se priorizan las ideas mediante votación o reducción de la lista de ideas.

2.3.2. Método de Análisis Comparativo de Jim Brosseau

Este método es utilizado para obtener los elementos de calidad en un contexto particular. El proceso para realizarlo tiene dos etapas, las cuales se detallan a continuación (Brosseau, 2010):

1. Se determina si existe el conocimiento que permita que los atributos estén dentro del alcance de la evaluación, a partir de lo cual las características son filtradas.

2. En base a los resultados del punto 1, se construye una matriz de los atributos restantes para comparar cada par de atributos de acuerdo a su importancia, a partir de la cual se toma un puntaje en base al cual se priorizan las características. Un ejemplo de esta matriz se muestra en la Tabla 2.3.

Tabla 2.3. Ejemplo de Matriz de atributos.
Adaptado de (Brosseau, 2010).

	Característica 1	**Característica 2**	**Característica 3**	**Puntaje final**
Característica 1		<	<	2
Característica 2			^	0
Característica 3				1

2.4. Ley de Protección de Datos Personales

La Ley de Protección de Datos Personales fue publicada en el año 2011 por el Gobierno del Perú, con la finalidad de salvaguardar los derechos de las personas cuyos datos hayan sido almacenados por diversas organizaciones (Ley de Protección de Datos Personales, 2011).

En esta ley, se definen los derechos que tiene el titular de los datos personales, así como las obligaciones tanto del titular como del administrador de un banco de datos, y

se establecen las sanciones para las personas u organizaciones que den un incorrecto uso a los datos (Ley de Protección de Datos Personales, 2011).

2.5. Trabajos relacionados a Calidad de Datos

En esta sección se mencionarán diversos avances realizados en Calidad de Datos, principalmente como propuestas y aplicaciones de diversos modelos de calidad.

2.5.1. Modelo de Calidad alineado a SQuaRE aplicado a portales Web (SPDQM)

En el año 2009, Moraga, Calero y Caro desarrollaron un modelo de calidad el cual está totalmente derivado del modelo planteado en la ISO/IEC 25012 y del Modelo de Calidad de Datos para Portales (PDQM, de Portal Data Quality Model en inglés) (C. Moraga, M. Á Moraga, C. Calero, & A. Caro, 2009).
Este modelo fue denominado SQuaRE-Aligned Portal Data Quality Model (SPDQM). Está conformado por 42 características tomadas a partir de los dos modelos anteriormente mencionados y de una revisión de la literatura revisada por los autores (C. Moraga et al., 2009).

El desarrollo de este modelo parte del hecho de que las personas que utilizan los datos emitidos por los portales Web necesitan tener seguridad de que la data tenga calidad suficiente para satisfacer sus necesidades, en un entorno en que estos portales, si bien es cierto dan acceso a cantidades ingentes de información, pueden también llevar a datos tanto correctos como incorrectos (C. Moraga et al., 2009). Por este motivo, se definió el objetivo del trabajo como la determinación de un conjunto de características organizadas y relevantes para los consumidores de datos cuando se evalúa su calidad en cualquier portal Web (C. Moraga et al., 2009).

La metodología para su realización implicó la combinación de las características de la ISO/IEC 25012 y el modelo PQDM, así como los resultados de una revisión de la literatura para obtener un juego inicial de características. Estos resultados fueron refinados mediante una evaluación de aplicabilidad en el contexto de portales web, resolución de conflictos y una selección final de características. Posteriormente se realizó la organización de las características, para las cuales también se combinaron las categorías de PQDM y la ISO/IEC 25012, lo que llevó al modelo SPDQM como resultado final (C. Moraga et al., 2009).

Este proyecto todavía está pendiente para su aplicación en la industria y su continuación por parte de los mismos autores; sin embargo, se han desarrollado herramientas basadas en este trabajo, tales como PoDQA (Portal Data Quality Assessment), que permite la evaluación de la calidad de datos en portales web (Enríquez de Salamanca, 2012).

Las características obtenidas en el modelo se muestran en la Tabla 2.4.

Tabla 2.4. Características del Modelo SPDQM.
Traducido de (C. Moraga et al., 2009)

Perspectiva	Categoría	Característica	Sub característica
Inherente	**Intrínseca**	Exactitud	
		Credibilidad	Objetividad
			Reputación
		Trazabilidad	
		Actualidad	
		Expiración	
		Completitud	
		Consistencia	
		Accesibilidad	
		Conformidad	
		Confidencialidad	
		Eficiencia	
		Precisión	
		Entendibilidad	
Dependiente del sistema	**Operacional**	Disponibilidad	
		Accesibilidad	Interactivo
			Facilidad de operación
			Soporte al cliente
		Verificabilidad	
		Confidencialidad	
		Portabilidad	
		Recuperabilidad	
	Contextual	Validez	Confiabilidad
			Alcance
		Valor agregado	Aplicabilidad
			Flexibilidad
			Novedad
		Relevancia	Novedad
			Oportunidad
		Especialización	
		Utilidad	
		Trazabilidad	
		Conformidad	
		Precisión	
	Representacional	Representación concisa	
		Representación consistente	

Perspectiva	Categoría	Característica	Sub característica
		Entendibilidad	Interpretabilidad
			Cantidad de datos
			Documentación
			Organización
		Atractivo	
		Legibilidad	
		Eficiencia	
		Efectividad	

2.5.2. Modelo de Calidad de datos para su aplicación en proyectos de Ciencia Ciudadana

Ente los años 2010 y 2012, Jane Hunter, Abdulmonem Alabri y otros autores desarrollaron un modelo de calidad de datos para su aplicación directa a un sistema de Ciencia Ciudadana (es decir, recolección de datos por parte de voluntarios como aporte a la investigación científica) desarrollado por Alabri, denominado CoralWatch (A. Alabri & J. Hunter, 2010; Hunter, Alabri, & Ingen, 2013).

Este modelo fue elaborado a partir del Ciclo de Gestión Total de la Calidad de Datos (TQDM, de Total Data Quality Management Cycle en inglés) del MIT; es decir, está diseñado para su aplicación luego de realizar mejoras a la calidad de los datos utilizados (Hunter et al., 2013).

Este proyecto se justificó debido a que un análisis de los datos empleados por CoralWatch encontró una gran cantidad de errores en los datos, sobre todo en datos geográficos (Hunter et al., 2013).

Para realizar este proyecto, se realizó la identificación de las dimensiones de calidad de datos, a partir de encuestas con los usuarios de CoralWatch. Posteriormente, se realizaron mediciones de calidad, las cuales fueron analizadas para encontrar discrepancias, y se implementaron herramientas que permiten tomar acciones para mejorar la calidad de datos (Hunter et al., 2013).

En este modelo, aparte de utilizar dimensiones comunes de calidad de datos, se han utilizado métricas de confianza social para conocer la reputación del sistema en base a los datos de sus usuarios (Hunter et al., 2013). Para esto, se calculó la reputación de

las entidades que organizan los datos de CoralWatch y se infirió la reputación a nivel de todo el programa (Hunter et al., 2013).

El modelo fue aplicado a partir de una mejora realizada a CoralWatch basada en los errores encontrados anteriormente, por lo cual se encontró que la calidad de los datos mejoró, principalmente en la reducción de los errores sintácticos en las entradas del sistema en un 70%. Sin embargo, los resultados también indicaron que se debe realizar un mejor manejo de las métricas de confianza social (Hunter et al., 2013).

Las dimensiones propuestas se muestran en la Tabla 2.5.

Tabla 2.5. Dimensiones propuestas en el Modelo para Medición de Calidad de Datos en Ciencia Ciudadana.
Traducido y adaptado de (Hunter et al., 2013)

Dimensiones
Accesibilidad
Cantidad apropiada de información
Credibilidad
Completitud
Representación concisa
Representación consistente
Facilidad de manipulación
Libre de error
Interpretabilidad
Objetividad
Relevancia
Reputación
Seguridad
Actualidad
Entendibilidad
Valor agregado

2.5.3. Extensión del Modelo de Calidad ISO/IEC 25012 de Irfan Rafique

En el año 2012, Irfan Rafique y otros autores extendieron el modelo ISO/IEC 25012 con dos nuevas características: valor agregado y adecuación representacional, creando un marco de trabajo para evaluar la calidad de información (Rafique et al., 2012).

En un contexto en que las aplicaciones Web se orientan más al uso de la información que almacenan, este trabajo parte del hecho de que las actualizaciones en las

tecnologías en línea hacen que las aplicaciones realizadas deban seguir requerimientos de calidad más sofisticados y consistentes, de tal manera que estén más orientados a la Calidad de Información (es decir, datos contextualizados y estructurados); asimismo, se buscaba integrar la calidad de información con la calidad de software de forma adecuada (Rafique et al., 2012).

El trabajo fue realizado a partir de una revisión de la literatura y de la ISO/IEC 25012. A partir de esto, se redefinieron las características existentes para que estén más orientadas a la información, y se agregaron las dos características antes mencionadas, dada su presencia en la literatura relacionada a la Calidad de Información (Rafique et al., 2012).

Este modelo aún está pendiente de tener una aplicación experimental o en la industria (Rafique et al., 2012).

Las características y sub características del modelo se muestran en la Tabla 2.6.

Tabla 2.6. Extensión del modelo ISO/IEC 25012 de Irfan Rafique.
Traducido y adaptado de (Rafique et al., 2012)

Características	Sub características	Perspectiva
Precisión de la Información	Correctitud	Inherente
	Credibilidad	Inherente
	Actualidad	Inherente
	Precisión	Ambas perspectivas
	Trazabilidad	Ambas perspectivas
Accesibilidad de la Información	Accesibilidad	Ambas perspectivas
Oportunidad de la información	Completitud	Inherente
	Entendibilidad	Ambas perspectivas
	Consistencia	Inherente
	Adecuación representacional	Dependiente del Sistema
	Valor Agregado	Dependiente del Sistema
Eficiencia	Eficiencia	Ambas perspectivas
Confidencialidad	Confidencialidad	Ambas perspectivas
Disponibilidad	Disponibilidad	Dependiente del Sistema
Portabilidad	Portabilidad	Dependiente del Sistema
Recuperabilidad	Recuperabilidad	Dependiente del Sistema

2.5.4. Modelo 3A de Calidad de Datos en Uso para Big Data

En el año 2015, Jorge Merino y otros autores propusieron un modelo de Calidad de Datos en Uso, denominado 3A, que se ajuste a las nuevas tendencias basadas en el uso

intensivo de Big Data (es decir, grandes cantidades de datos) (Merino, Caballero, Rivas, Serrano, & Piattini, 2015).

Este modelo partió del hecho de que modelos tradicionales, como la ISO/IEC 25012, no están adaptados para la evaluación de Big Data, debido al esfuerzo que implica la búsqueda y reparación de las inconsistencias, por lo que se busca hacer frente a los desafíos que conlleva este tipo de colecciones de datos. Asimismo, estos modelos permiten principalmente la evaluación de la calidad interna y externa de los datos, pero no de la calidad de uso (Merino et al., 2015).

Las características de este modelo están basadas en las ISO/IEC 25012 pero han sido reclasificadas en tres categorías: adecuación contextual, adecuación temporal y adecuación operacional, de tal manera que se ajusten mejor a los conceptos que abarca Big Data (Merino et al., 2015).

Las métricas propuestas para la evaluación de estas características fueron tomadas de la ISO/IEC 25024, dada su compatibilidad con la ISO/IEC 25012 (Merino et al., 2015).

En el trabajo se incluye un ejemplo de aplicación del modelo 3A sobre grandes cantidades de datos financieros, para los cuales la adecuación contextual era importante puesto que los datos debían ser lo más sólidos y válidos posible (Merino et al., 2015). Los resultados indicaron que si bien es cierto los resultados son aceptables, el requerimiento relativo a solidez y validez de datos no se cumplía (Merino et al., 2015). A partir de esto, se llega a la conclusión de que el manejo de la calidad de datos depende tanto de los datos en sí como de los procesos analíticos y las nuevas tecnologías que los soportan (Merino et al., 2015).

Las características utilizadas en el modelo se muestran en la Tabla 2.7.

Tabla 2.7. Reclasificación de las características según el modelo 3A.
Traducido de (Merino et al., 2015)

Característica	Adecuación Contextual	Adecuación Temporal	Adecuación Operacional
Exactitud	X	X	
Completitud	X		
Consistencia	X	X	
Credibilidad	X		
Actualidad		X	

Característica	Adecuación Contextual	Adecuación Temporal	Adecuación Operacional
Accesibilidad			X
Conformidad	X		
Confidencialidad	X		X
Eficiencia			X
Precisión			X
Trazabilidad			x
Entendibilidad	X		
Disponibilidad			X
Portabilidad			X
Recuperabilidad			X

2.5.5. Definición y aplicación de marco de trabajo de Calidad de Datos a Datos Gubernamentales Abiertos

En el año 2016, Antonio Vetrò y otros autores propusieron un marco de medición para datos gubernamentales abiertos (es decir, que pueden ser modificados libremente por cualquier persona) (Vetrò et al., 2016).

Este marco se justificó en el hecho de que la baja calidad de datos que puede tener una base de datos abierta puede hacer que no exista mucha confianza en este tipo de sistemas y que se realice un mayor gasto de recursos para el procesamiento de los datos utilizados (Vetrò et al., 2016).

Este modelo se basa en el SPDQM explicado en 2.4.1, y provee un conjunto de características para la evaluación de los datos, así como las métricas utilizadas para su evaluación (Vetrò et al., 2016).

Para su realización, se seleccionaron algunas características de SPDQM, las cuales fueron sometidas a una encuesta realizada a desarrolladores especializados en el uso de datos gubernamentales abiertos para conocer los principales problemas que se tiene en su uso. Estos datos fueron relacionados a las características elegidas para su clasificación.

Asimismo, se incluye una aplicación del marco, que implica la comparación entre los datos del proyecto Open Coesione (que contiene información unificada sobre proyectos realizados por el Gobierno Italiano) y otros proyectos separados realizados por la Fundación Social Europea, de tal manera que se puedan encontrar las buenas prácticas, debilidades y factores discriminantes en dos entornos distintos (Vetrò et al., 2016). Los resultados indicaron que Open Coesione manejaba mejores prácticas de

calidad de datos, y que las diferencias de varianza se encuentran principalmente en la entendibilidad de los datos, así como en la completitud, precisión, trazabilidad, actualidad y expiración, aunque hay similaridades en los niveles de trazabilidad, conformidad y expiración (Vetrò et al., 2016).

Adicionalmente, se proponen guías para la mejora de los aspectos de tiempo, trazabilidad, precisión, completitud, entendibilidad y conformidad.

La conclusión a la que llegaron los autores indica que las métricas mostraron los beneficios de una apertura de datos centralizada (el caso de Open Coesione) como ejemplo de buenos datos gubernamentales abiertos (Vetrò et al., 2016). Asimismo, se define como meta el uso del modelo en una herramienta que permita la medición automática de la calidad de datos para realizar su mejora rápida antes de que puedan ser utilizados por el público usuario (Vetrò et al., 2016).

Las características y métricas del marco se muestran en la Tabla 2.8.

Tabla 2.8. Características y métricas del Marco de Trabajo de Calidad de Datos de Vetrò.
Traducido de (Vetrò et al., 2016)

Característica	Métrica
Trazabilidad	Trazabilidad de creación
	Trazabilidad de actualizaciones
Actualidad	Porcentaje de filas actualizadas
	Demora en la publicación
Expiración	Demora luego de la expiración
Completitud	Porcentaje de celdas completas
	Porcentaje de filas completas
Conformidad	Porcentaje de columnas estandarizadas
	Conformidad con eGMS (e-Government Metadata Standard)
	Datos abiertos de Cinco Estrellas
Entendibilidad	Porcentaje de columnas con metadata
	Porcentaje de columnas con formato comprensible
Precisión	Porcentaje de celdas precisas
	Precisión en la agregación

3. Elaboración y Aplicación del Modelo de Calidad

En este capítulo se presentará el modelo de calidad propuesto como proyecto de tesis, así como los resultados de su aplicación en el contexto de la institución educativa Edu.Lambda.

3.1. Descripción de la Organización

La institución educativa Edu.Lambda es una organización dedicada a la enseñanza y a la investigación académica.

3.2. Solución de Inteligencia de Negocios a estudiar

La solución a estudiar ha sido creada hace menos de 5 años a partir de las necesidades de los usuarios que necesitaban una forma rápida de obtener indicadores que puedan ser analizados para el crecimiento y el manejo correcto de la organización.

Esta solución se compone de:

- Una serie de procesos ETL que realizan la transformación de los datos fuente en información útil para el análisis.

- El Data Warehouse organizacional, utilizado para almacenar la información procesada en los procesos ETL, y que es el objetivo del modelo de calidad elaborado. Este funciona como una base de datos relacional, pero usando conceptos típicos de Inteligencia de Negocios como tablas de hechos y dimensiones.

- Una herramienta OLAP que toma los datos del Data Warehouse y se utiliza para manejar, externamente al Data Warehouse, el procesamiento de los datos mediante la generación de reportes. Esta herramienta es la que realiza el mayor contacto con los usuarios, brindándoles acceso a la información utilizada para la toma de decisiones.

3.3. Repositorios elegidos para la evaluación

De acuerdo a las exigencias del trabajo se han elegido como repositorio de datos el Data Warehouse de la institución educativa Edu.Lambda, en el cual se almacena toda la información procesada para su análisis.

A partir de esto, se eligió el conjunto LAMBDA.WAREHOUSE, que es una muestra de datos de algunas tablas elegidas del Data Warehouse.

Para elegirlos, se han seguido los siguientes criterios:

- Los datos tienen que ser datos de suma importancia para la organización, y deben ser de alto valor para el personal de la organización, lo que fue corroborado con los expertos.

- Los datos deben corresponder a la salida de un proceso ETL.

- Los datos deben haber sido autorizados para su uso por las autoridades pertinentes en la institución.

3.4. Definición del Modelo de Calidad de Datos

3.4.1. Determinación de Características

Para la definición del modelo, se realizaron los siguientes pasos, según lo especificado en el Anexo C:

- Se utilizó la Técnica de Grupo Nominal con un grupo de expertos en BI de la organización, para obtener la valoración que ellos les dan a las diversas características contempladas en la ISO/IEC 25012. La plantilla utilizada para la técnica se muestra en el Anexo A, Los resultados se muestran en los Anexos B y C.

- Se utilizó el Método de Jim Brosseau para agregar el criterio del autor a la elección de las características, tomando como base los resultados obtenidos de

la Técnica de Grupo Nominal. Los resultados obtenidos se muestran en el Anexo C.

A partir de estos métodos, se ha obtenido el siguiente juego de características, basadas en la ISO/IEC 25012, las cuales se muestran en la Tabla 3.1.

Tabla 3.1. Características seleccionadas de la ISO/IEC 25012 para el estudio a realizar.
Cuadro diseñado por el autor.

Características obtenidas
Exactitud
Completitud
Consistencia
Credibilidad
Actualidad
Conformidad

3.4.2. Definición de Métricas

Las métricas a utilizar han sido obtenidas a partir de la ISO/IEC 25024, en la cual se sugiere un conjunto de métricas para cada característica. Estas métricas se muestran en la Tablas 3.2 a 3.9 (el código de métrica mostrado es el correspondiente a la ISO/IEC mencionada).

Tabla 3.2. Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad – Exactitud (Inherente).
Traducido y adaptado de (ISO, 2015).

Métrica	**Fórmula**	**Acciones**	**Intervalo de aceptabilidad**
ACC-I-1 Exactitud Sintáctica de Datos Ratio de cercanía a de los valores de datos a un conjunto de valores definidos en un dominio	X=A/B A = Número de ítems de datos que tienen valores relacionados sintácticamente exactos B = Número de ítems de datos para los cuales se requiere exactitud sintáctica	A: Revisar todos los valores para ver si siguen las reglas de sintaxis adecuadas (ej. Limites entre fechas, palabras bien escritas, etc) y es el mismo que una fuente validada. Requiere scripts de comparación entre valores fuente y destino. B: El total.	No definido. Entre 0 y 1 (ratio)

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
ACC-I-3 Aseguramiento de exactitud de datos Ratio de cobertura de medición para datos exactos	X=A/B A = Número de ítems de datos a los cuales se les midió su exactitud B = Número de ítems de datos para los cuales se requiere medir su exactitud	A: Revisar la cantidad de campos que se están midiendo. B: El total de datos a medir.	No definido. Entre 0 y 1 (ratio)
ACC-I-4 Riesgo de inexactitud de conjunto de datos El número de valores atípicos en los valores indica un riesgo de inexactitud para valores de datos en un conjunto de datos	X = A/B A = Número de valores de datos atípicos B= Número de valores de datos considerados en un conjunto de datos	A: Se colocan los posibles valores en una tabla estadística y se cuentan para observar los datos atípicos. Requiere script de comparación y análisis de resultados mediante las distribuciones sugeridas en la ISO/IEC 25024. B: El total.	Mejor mientras X sea más bajo
ACC-I-6 Exactitud de Metadata ¿La metadata describe los datos con la exactitud requerida?	X = A/B A = Número de metadata que proporciona la información requerida apropiada B= Número de metadata definido dentro de la especificación de requerimientos de datos	A: Conteo de la metadata presente del conjunto de datos (nombres, comentarios…) a partir de un diccionario de datos o de la metadata existente. B: El total	No definido
ACC-I-7 Rango de exactitud de datos ¿Los valores de datos están incluidos en los intervalos requeridos?	X = A/B A = Número de ítems de datos teniendo un valor incluido en un intervalo específico B= Número de ítems de datos para los cuales se define un intervalo de valores requerido.	A: Conteo de ítems de datos con valores en un intervalo de datos establecidos. Requiere script de comparación. B: Conteo de ítems de datos para los que se definió un intervalo de datos. Requiere script de comparación y la especificación de los requerimientos de los datos.	Definido en requisitos

Tabla 3.3. Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad - Completitud (Inherente)
Traducido y adaptado de (ISO, 2015).

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
COM-I-1 Completitud de Registros Completitud de ítems de datos de un registro dentro de un archivo de datos	X=A/B A = Número de ítems de datos con un valor no nulo asociado en un registro B = Número de ítems de datos del registro para los cuales se puede medir su completitud	A: Script para contar nulos en los datos. B: El total.	A: Script para contar nulos en los datos. B: El total.
COM-I-2 Completitud de Atributos Completitud de ítems de datos dentro de un archivo de datos	X=A/B A = Número de ítems de datos con un valor no nulo asociado para un ítem de datos específico B = Número de registros contados	A: Script para contar nulos en los datos. B: El total.	A: Script para contar nulos en los datos. B: El total.

Tabla 3.4. Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad – Consistencia (Inherente).
Traducido y adaptado de (ISO, 2015).

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
CON-I-1 Integridad Referencial Por cada valor de un atributo de una tabla existe el mismo valor del mismo atributo en una tabla diferente: es decir, hay un vínculo entre el mismo atributo representado en diferentes tablas y contienen los mismos valores	X=1-A/B A = Número de ítems de datos no consistentes en su valor B = Número de ítems de datos para los cuales la integridad referencial debe estar definida.	A: Script para analizar si los datos son iguales en tablas enlazadas por integridad referencial. B: El total de ítems de datos con integridad referencial definida.	No definido (entre 0 y 1)
CON-I-2 Consistencia de formato de datos Consistencia del formato de datos del mismo ítem de datos	X=A/B A = Número de ítems de datos donde el formato de todas las propiedades es consistente en diferentes archivos de datos A = Número de ítems de datos para los cuales la consistencia de formatos debe estar definida.	A: Script para comparar igualdad de formatos en todos los conjuntos de datos. B: El total de ítems de datos con formato definido.	No definido (entre 0 y 1)

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
CON-I-3 Riesgo de inconsistencia de datos Riesgo de tener inconsistencias debido a la duplicación de valores de datos	X = A/B A = Número de ítems de datos donde existe duplicación en valores B= Número de ítems de datos considerados	A: Número de ítems de datos duplicados. B: El total de ítems de datos.	Mejor mientras X sea más bajo
CON-I-4 Consistencia de la arquitectura Grado al que los elementos de la arquitectura tienen una correspondencia en elementos referenciados de la arquitectura	X = A/B A = Número de elementos en una arquitectura que tienen un elemento correspondiente referenciado en la arquitectura instalada B= Número de elementos de la arquitectura referenciada	A: Número de elementos que coinciden entre el diccionario de datos y el modelo de datos conceptual. B: El total de elementos en el modelo de datos.	No definido (entre 0 y 1)
CON-I-5 Cobertura de consistencia de valores de datos Cobertura de la medición de la consistencia de valores de datos	X = A/B A = Número de ítems de datos considerados en la medición de consistencia de valores de datos. B= Número de ítems de datos para los cuales se mide la consistencia.	A: Cantidad de ítems de datos (campos) a medir. B: Cantidad de ítems de datos (campos) para los cuales la consistencia está medida.	No definido (entre 0 y 1)
CON-I-6 Consistencia semántica Grado al cual se respetan las reglas semánticas	X = A/B A = Número de ítems de datos donde los valores son semánticamente correctos en el archivo de datos B= Número de ítems de datos para los cuales se han definido reglas semánticas	A: Conteo de ítems de datos que respetan las reglas semánticas (ejm, reglas, rangos de fechas...). Requiere script de comparación. B: El total de ítems con reglas definidas.	No definido (entre 0 y 1)

Tabla 3.5. Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad - Credibilidad (Inherente).
Traducido y adaptado de (ISO, 2015).

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
CRE-I-1 Credibilidad de valores Grado al cual los ítems de información son tomados como verdaderos, reales y creíbles	X=A/B A = Número de ítems de información cuyos valores son validados/certificados por un proceso específico. B = Número de ítems de información a ser validados/certificados.	A: Conteo de ítems de información que pasan por un proceso de validación. (Si se aplica ACC-I-4, se incluye este como proceso) B: Conteo de ítems de información que serán validados. Ambos requieren scripts de conteo y revisión del modelo de datos y del diccionario de datos.	No definido (entre 0 y 1)
CRE-I-3 Credibilidad de diccionario de datos Grado al cual el diccionario de datos proporciona información creíble	X = A/B A = Número de ítems de información en el diccionario de datos cuyos valores son validados/certificados por un proceso específico. B = Número de ítems de información en el diccionario de datos	A: Conteo de ítems en el diccionario de datos B: El total de ítems de información en DD.	No definido (entre 0 y 1)
CRE-I-4 Credibilidad de modelo de datos Grado al cual el modelo de datos proporciona información creíble	X = A/B A = Número de elementos de un modelo de datos con definición apropiada validada/certificada por un proceso específico B= Número de elementos del modelo de datos	A: Número de elementos que coinciden entre el diccionario de datos y el modelo de datos conceptual. B: El total de elementos en el modelo de datos.	No definido (entre 0 y 1)

Tabla 3.6. Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad - Actualidad (Inherente).
Traducido y adaptado de (ISO, 2015).

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
CUR-I-1 Frecuencia de actualización Grado al cual los ítems de datos son actualizados con la frecuencia requerida	X=A/B A = Número de ítems de datos actualizados con la frecuencia requerida. B = Número de ítems de datos que tienen un requerimiento de frecuencia de actualización.	A: Conteo de ítems de datos para saber si se actualizan con la frecuencia necesaria. Depende de los requerimientos y de las fechas de actualización. B: El total de datos (todos se actualizan cada cierto tiempo).	No definido (entre 0 y 1)
CUR-I-2 Oportunidad de la actualización Grado al cual los ítems de datos están actualizados a tiempo	X=A/B A = Número de ítems de datos actualizados a tiempo. B = Número de ítems de datos que requieren actualización	A: Conteo de ítems de información actualizados al día correspondiente. Requiere revisión de las fechas de actualización. B: El total de datos (todos se actualizan cada cierto tiempo).	No definido (entre 0 y 1)
CUR-I-3 Requisito de actualización de ítems Grado al cual el requisito de actualizar los ítems de datos frecuentemente existe	X = A/B A = Número de ítems de información con un requisito explícito de actualización. B = Número de ítems de información para los cuales se requiere un requisito de actualización.	A: Conteo de ítems con un requisito explícito de actualización (ver requerimientos) B: Conteo de ítems que se actualizan.	No definido (entre 0 y 1)

Tabla 3.7. Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad - Conformidad (Inherente).
Traducido y adaptado de (ISO, 2015).

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
CMP-I-1 Conformidad regulatoria de valores y/o formato Grado al cual los valores de datos y/o formatos cumplen con los estándares específicos, convenciones o regulaciones	X=A/B A = Número de ítems de datos que tienen valores y/o formato que se conforma a estándares, convenciones o regulaciones B = Número de ítems de datos que deben estar conformes a los estándares, convenciones o regulaciones debido a su valor.	A: Conteo de ítems de datos con formato/valores que respetan las leyes establecidas sobre los datos. Se requiere script de conteo y estudio de las herramientas utilizadas. B: El total de datos que se deben ajustar a alguna regulación.	No definido (entre 0 y 1)

Tabla 3.8. Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad - Conformidad (Dependiente del Sistema).
Traducido y adaptado de (ISO, 2015).

Métrica	Fórmula	Acciones	Intervalo de aceptabilidad
CMP-D-1 Conformidad regulatoria debido a la tecnología Grado al cual el ítem de datos cumple con estándares específicos, convenciones o regulaciones	X=A/B A = Número de ítems de datos conformes a los estándares, convenciones o regulaciones gracias a la tecnología B = Número de ítems de datos que deben estar conformes a los estándares, convenciones regulaciones gracias a la tecnología	A: Conteo de datos (incluyen documentos, modelo, diccionario) que respetan las leyes establecidas sobre los datos gracias a temas tecnológicos. Se requiere script de conteo y estudio de las herramientas utilizadas. P.ej. un ítem no cumple esta métrica si debido a una falla técnica no es conforme a la regulación. B: El total de datos que deben cumplir la conformidad.	No definido (entre 0 y 1)

*Considerar también estándares de Inteligencia de Negocios propuestos en la organización.

3.4.3. Requisitos para la evaluación automática o manual de las métricas

Para la evaluación de estas métricas, se requiere:

- Conceptos de la ISO/IEC 25024.

- Scripts de obtención de resultados de métricas, los cuales pueden aplicarse directa o indirectamente sobre el conjunto de datos.

3.5. Planificación de la evaluación

De acuerdo a la ISO/IEC 25040, se tienen siete restricciones, los cuales se explicarán a continuación para el contexto en que se encuentra el proyecto:

- **Necesidades específicas del usuario:** Los encargados de la solución, que tienen el papel de usuarios a su vez, tienen la expectativa de obtener una herramienta para la medición de la calidad de datos de su Data Warehouse de tal manera que puedan asegurarla permanentemente, evitando y eliminando inconsistencias y realizando procesos de mejora continua.

- **Recursos:** Según lo explicado en la parte de Viabilidad, en la parte técnica se utilizarán las herramientas que brinda la empresa previa autorización. Para la obtención de las normas, se recurrió al apoyo del grupo GIDIS-PUCP.

- **Cronograma:** Se tiene como fecha máxima para la entrega de los resultados de la evaluación el día 30 de noviembre de 2016.

- **Costo:** Las herramientas han sido brindadas por el grupo GIDIS-PUCP y se utiliza los programas licenciados por la organización, por lo cual no se tiene un costo monetario.

- **Entorno:** Se trabajará dentro de un entorno con gente especializada en temas de Inteligencia de Negocios y con herramientas preparadas para tal fin. El entorno debe dar prioridad al trabajo de la organización, por lo que la planificación y el acceso a ellos para consultas está restringido por el tiempo disponible de los encargados.

- **Herramientas y metodologías:** Las herramientas utilizadas para la evaluación se mencionarán en la parte 3.6 del documento.

- **Reportes:** Para las métricas utilizadas se presentarán reportes de su medición tanto a la empresa como al grupo GIDIS-PUCP. Además, al final de la evaluación se preparará un informe conteniendo las observaciones encontradas.

La evaluación será realizada según los pasos definidos en la ISO/IEC 25040.

El plan de evaluación completo se muestra en el Anexo D.

3.6. Definición de Instrumentos de Evaluación

Los instrumentos de evaluación utilizados son los siguientes, en conjunto con las normas ISO/IEC propuestas:

- Structured Query Language (SQL): Lenguaje de consulta usado masivamente para realizar operaciones sobre bases de datos relacionales. Esta herramienta permitirá la elaboración de los scripts utilizados para las consultas que permitan el conteo de las métricas de calidad.

- IBM SPSS Statistics 23: Herramienta de procesamiento de datos utilizada para manejo estadístico. Esta herramienta será utilizada para crear scripts complementarios a las consultas realizadas en SQL.

3.7. Evaluación de Métricas

A continuación se presentarán los resultados de la evaluación de métricas realizada para el presente trabajo.

3.7.1. Métricas relevantes seleccionadas

En esta sección se muestran las métricas más relevantes del total de métricas evaluadas: el total de su detalle se muestra en el Anexo E.

ACC-I-1 Exactitud Sintáctica de Datos

Cálculo de la métrica:

X = A/B
A = Cantidad de datos sintácticamente exactos = 66484083
B = Cantidad de datos que requieren exactitud sintáctica = 66484344
X = 0.99

Forma de cálculo:

Se utilizaron los siguientes scripts por cada atributo que requiera verificación de exactitud:

SELECT COUNT(ATRIBUTO)
FROM LAMBDA.WAREHOUSE
WHERE ATRIBUTO = CORRECTO

para los cuales se sumaron todos sus valores para llegar al resultado de la métrica.

Comentarios:

Debido a que la mayor parte de los datos está basada en integridad referencial, prácticamente no se tienen errores de sintaxis en la información.

ACC-I-4 Riesgo de inexactitud de conjunto de datos

Cálculo de la métrica:

X = A/B
A = Número de valores de datos que son atípicos = 457913
B = Número de valores de datos a considerar en un conjunto de datos = 22161448
X = 0.02
Para propósitos de estandarización, el resultado será variado para que exprese que sea mejor mientras sea más cercana a 1. El valor final sería:
X = 0.98

Forma de cálculo:

Esta métrica se aplicó a los valores continuos (que son indicadores) en el conjunto de datos.
Primero, se generó un script para obtener todos los valores mencionados:

SELECT INDICADORES
FROM LAMBDA.WAREHOUSE

Usando la herramienta IBM SPSS Statistics, se obtuvo el percentil 25 y 75 de cada indicador para el cálculo de la métrica según la ISO/IEC 25024 mediante la siguiente función:

```
NPAR TESTS
  /K-S(NORMAL)=var1 var2 var3…
  /MISSING ANALYSIS.
```

Posteriormente, para cada indicador, se hizo la siguiente consulta:

```
SELECT COUNT(INDICADOR)
FROM LAMBDA.WAREHOUSE
WHERE INDICADOR NOT BETWEEN LIMITEINFERIOR, LIMITESUPERIOR
```

de la cual se obtiene el conteo de los valores atípicos.
Para el cálculo de los límites inferior y superior, se obtuvo primero la longitud de caja de cada indicador, y posteriormente se le multiplicó por 1.5 según las reglas estadísticas.
Los límites se calcularon de la siguiente manera:

LIMITEINFERIOR = PERCENTIL25 – LONGITUD*1.5
LIMITESUPERIOR = PERCENTIL75 + LONGITUD*1.5

Comentarios:

El valor obtenido indica que la cantidad de datos atípicos es muy baja (aproximadamente 2% de los datos); sin embargo, teniendo en cuenta que se trata de una solución de Inteligencia de Negocios, es decir, con baja tolerancia a inconsistencias, estos datos deben ser analizados para ver si son erróneos.

COM-I-2 Completitud de Atributos

Cálculo de la métrica:

X = Promedio de A/B para cada atributo
A = Cantidad de valores no nulos para un atributo
B = Cantidad de registros evaluados

Forma de cálculo:

Para la evaluación de esta métrica, se calculó la proporción de valores no nulos para cada atributo, mediante este script:

```
SELECT COUNT(ATRIBUTO)/TOTAL
FROM LAMBDA.WAREHOUSE
```

Los resultados por atributo se muestran en la Tabla 3.10. Debido a la estructura que tiene un Data Warehouse, en la tabla se resaltan los atributos que corresponden a indicadores que se obtienen de los cálculos de los procesos utilizados para llenar la base de datos, debido a su importancia.

Tabla 3.9. Resultados de la medición de la métrica para cada atributo.
Cuadro diseñado por el autor.

Atributo	Métrica
C1	1
C2	1
C3	1
C4	1
C5	1
C6	1
C7	1
C8	1
C9	1
C10	1
C11	1
C12	1

Atributo	Métrica
C13	1
C14	1
C15	1
C16	1
C17	1
C18	1
C19	1
C20	1
C21	0.999996
C22	0.999996
C23	0.999996
C24	0.999996
C25	0.999996
C26	0.999996
C27	0.999996
C28	0.97035
C29	0.598285
C30	0.598285
C31	0.598285
C32	0.598285
C33	0.757206
C34	1
C35	1
C36	1
C37	1
C38	1
C39	1
C40	1
C41	1
C42	1
C43	1
C44	1
C45	1
C46	1

Atributo	Métrica
C47	1
C48	0.923701
C49	1
C50	1
C51	0.50965
C52	0.966816
C53	0.999996
C54	0.981985
C55	0.999996
C56	0.509428
C57	1

Indicadores (marcados en amarillo)

Menor valor = 0.509428

Mayor valor = 0.99

Promedio = 0.84

General:

Menor valor = 0.509428

Mayor valor = 1

Promedio = 0.947

Comentarios:

Principalmente los atributos que están completos son aquellos que son usados como llaves (37 de las 57 totales) y tienen la integridad referencial requerida para los datos.

Los atributos incompletos se encuentran básicamente en los indicadores sin integridad referencial utilizados como base para los cálculos de toma de decisiones, esto dado que no se siguen completamente los estándares requeridos para su llenado y se dejan como valores nulos.

Además, se consideraron los valores "sin dato" como valores nulos en el caso de los indicadores con integridad referencial y en las llaves.

Dado que se está evaluando parte de un sistema de apoyo a las decisiones, se puede decir que la cantidad de datos faltantes es relevante, especialmente en el caso de los indicadores, los cuales son pieza clave de todo Data Warehouse. Esto puede ocurrir

debido a problemas en la base de datos fuente, así como en el cálculo de los valores a almacenar.

CON-I-3 Riesgo de inconsistencia de datos

Cálculo de la métrica:

X = A/B
A = Número de ítems de datos en los que existen duplicación de valores = 62759763
B = Número de ítems de datos considerados = 62999443
X = 0.996
Para propósitos de estandarización, el resultado será variado para que exprese que sea mejor mientras sea más cercana a 1. El valor final sería:
X = 0.004

Forma de cálculo:

Se realizaron los siguientes scripts:

```
SELECT COUNT(DISTINCT ATRIBUTOS) FROM LAMBDA.WAREHOUSE
SELECT COUNT(ATRIBUTOS) FROM LAMBDA.WAREHOUSE
```

Luego, al resultado del segundo script se les restó los resultados del primer script para obtener los campos repetidos. Estos valores fueron sumados, dando los valores de A y B requeridos para el cálculo de la métrica.
No se han considerado valores nulos.

Comentarios:

El resultado ocurre debido a la cantidad de campos con constraints de integridad referencial (altamente usado en este Data Warehouse), lo cual provoca que los valores estén casi fijos y se repitan. Debido a esto, el riesgo de duplicación de datos es alto, especialmente si se hacen consultas basadas en agrupación.
Es importante notar que los registros (es decir, las tuplas de campos) nunca se duplican.

CON-I-6 Consistencia semántica

Cálculo de la métrica:

X = A/B
A = Cantidad de datos que cumplen con las reglas semánticas = 18115685
B = Cantidad de datos totales con requerimiento de reglas semánticas = 25660624
X = 0.7

Forma de cálculo:

Se han contado los datos que no siguen las reglas semánticas dentro del conjunto de datos.
Los scripts utilizados dependen de los atributos que se han seguido. Para cada atributo, se obtuvo la cantidad de datos que no cumplían con estas reglas, para luego ser restados al total, a partir del siguiente script:

```
SELECT COUNT(ATRIBUTO)
FROM LAMBDA.WAREHOUSE
WHERE ATRIBUTO = CORRECTO
```

Comentarios:

Si bien es cierto la proporción de datos que cumplen las reglas semánticas es alta, también se puede decir que en muchos casos estas reglas no se cumplen, sobre todo en datos de gran antigüedad.

CRE-I-1 Credibilidad de valores

Cálculo de la métrica:

X = A/B
A = Cantidad de datos verificados por un proceso específico = 45910725
B = Total de datos por ser verificados = 66484344
X = 0.69

Forma de cálculo:

Esta métrica se calculó en base a dos scripts, basados en un año X a partir del cual la data es considerada como creíble.
Para datos posteriores al año X, se calculó la credibilidad de los valores de tres atributos de datos a partir de un indicador que indicaba si los valores eran creíbles o no. Se utilizó el siguiente script:

```
SELECT COUNT(ATR1, ATR2, ATR3)
FROM LAMBDA.WAREHOUSE
WHERE IND = 0
AND ANHO >= X
```

Para datos anteriores al año X, se consideró que ningún dato era creíble, por lo que se calculó el total de los datos.

```
SELECT COUNT (ATR)
FROM LAMBDA.WAREHOUSE
WHERE ANHO < X
```

El resultado de ambos scripts fue sumado y fue dividido entre el total de datos para obtener la métrica.

Comentarios:

Este resultado ocurre principalmente porque los encargados de la solución evaluaron con anterioridad la confiabilidad de los datos. Como se puede observar, los datos no confiables se concentran en los datos más antiguos (anteriores al año X), lo que provoca que, en total, sólo 69% de los datos sean confiables, algo excesivo para un sistema de soporte a las decisiones. Sin embargo, es aceptado, dado que los datos más antiguos son usados principalmente con propósitos históricos.
En el caso de los datos posteriores al año X, esto ocurre debido a la complejidad de los procesos que derivan en dichos indicadores.

CUR-I-2 Oportunidad de la actualización

Cálculo de la métrica:

X=A/B
A = Total de datos actualizados a tiempo = Ninguno (fecha no cumplida)
B = Total de datos que requieren actualización = Todos
X = 0

Forma de cálculo:

Para el cálculo, se revisan los logs de los procesos ETL y se verifica si el proceso de actualización del conjunto de datos se corrió la cantidad de veces requeridas para que esté actualizada en el momento oportuno.

Resultado:

Debido a no estar actualizado a la fecha, se presume que hay errores en los procesos ETL que provocan que los datos no estén actualizados a la fecha.

CMP-I-1 Conformidad regulatoria de valores y/o formato

Cálculo de la métrica:

X = A/B
A = Items de datos que tienen valores y/o formato conforme a estándares = =65317952
B = Total de datos que deben tener valores y/o formato conforme a estándares = 66484344
X = 0.98

Forma de cálculo:

Para el cálculo de esta métrica, se consideró que los datos deben cumplir los siguientes estándares:

- Ley de Protección de Datos Personales.
- Nombres en mayúscula, sin acentos.
- Los campos que son llaves deben tener constraints de integridad referencial.

Para que un valor ingrese en la métrica, todos estos estándares deben ser cumplidos; de lo contrario, no será contado.

Dado que todos los atributos cumplen en la actualidad la ley de protección de Datos Personales, se han evaluado los otros dos estándares definidos.

Para el análisis del primer estándar, se usó el primer script, por cada atributo que tenga nombres:

```
SELECT ATRIBUTO, NOMBREATRIB
FROM LAMBDA.WAREHOUSE wh
LEFT JOIN LAMBDA.DIMENSION dm
ON (wh.ATRIBUTO = dm.ATRIBUTO)
WHERE upper(dm.NOMBREATRIB) <> dm.NOMBREATRIB
```

La cantidad de registros emitidos se le resta al total de registros del repositorio de datos.

Para el análisis del segundo estándar, se verificó si los campos en mención poseen constraints que definan la integridad referencial hacia otra tabla.

El cálculo provino de la multiplicación: AtributoConIntegridad * CantidadDeDatos.

Comentarios:

El resultado indica que los procesos ETL utilizados funcionan correctamente respecto al cumplimiento de las convenciones establecidas como estándares para el manejo de los datos de la organización.

3.7.2. Resumen de resultados de métricas

Los resultados obtenidos se muestran totalmente en la Tabla 3.11. Los valores en que se encontraron problemas se encuentran resaltados en amarillo.

Tabla 3.10. Compilación de resultados de las métricas.
Cuadro diseñado por el autor.

Característica	Métrica	Resultado	Medición	Promedio de característica
Exactitud	ACC-I-1	0.99	Mejor si es más cercano a 1	0.99
	ACC-I-3	1	Mejor si es más cercano a 1	
	ACC-I-4	0.98	Mejor si es más cercano a 0 (Estandarizada)	
	ACC-I-6	1	Mejor si es más cercano a 1	
	ACC-I-7	1	Mejor si es más cercano a 1	
Completitud	COM-I-1	0.97	Mejor si es más cercano a 1	0.97
	COM-I-2	**0.947 (General)** 0.84 (Indicadores)	Mejor si es más cercano a 1	
	COM-I-3	1	Mejor si es más cercano a 1	
Consistencia	CON-I-1	1	Mejor si es más cercano a 1	0.78
	CON-I-2	1	Mejor si es más cercano a 1	
	CON-I-3	0.004	Mejor si es más cercano a 0 (Estandarizada)	
	CON-I-4	0.975	Mejor si es más cercano a 1	
	CON-I-5	1	Mejor si es más cercano a 1	
	CON-I-6	0.7	Mejor si es más cercano a 1	
Credibilidad	CRE-I-1	0.69	Mejor si es más cercano a 1	0.9
	CRE-I-3	1	Mejor si es más cercano a 1	
	CRE-I-4	1	Mejor si es más cercano a 1	
Actualidad	CUR-I-1	0	Mejor si es más cercano a 1	0.33
	CUR-I-2	0	Mejor si es más cercano a 1	
	CUR-I-3	1	Mejor si es más cercano a 1	
Conformidad	CMP-I-1	0.98	Mejor si es más cercano a 1	0.99
	CMP-D-1	1	Mejor si es más cercano a 1	

Para un mejor entendimiento de las características, se muestran gráficas que representan la medición de cada característica (tanto por métrica como por promedio), en las Figuras 3.1 y 3.2, las cuales muestran el estado de los datos evaluados.

Figura 3.1. Gráfico de radar de las métricas evaluadas.
Gráfico diseñado por el autor.

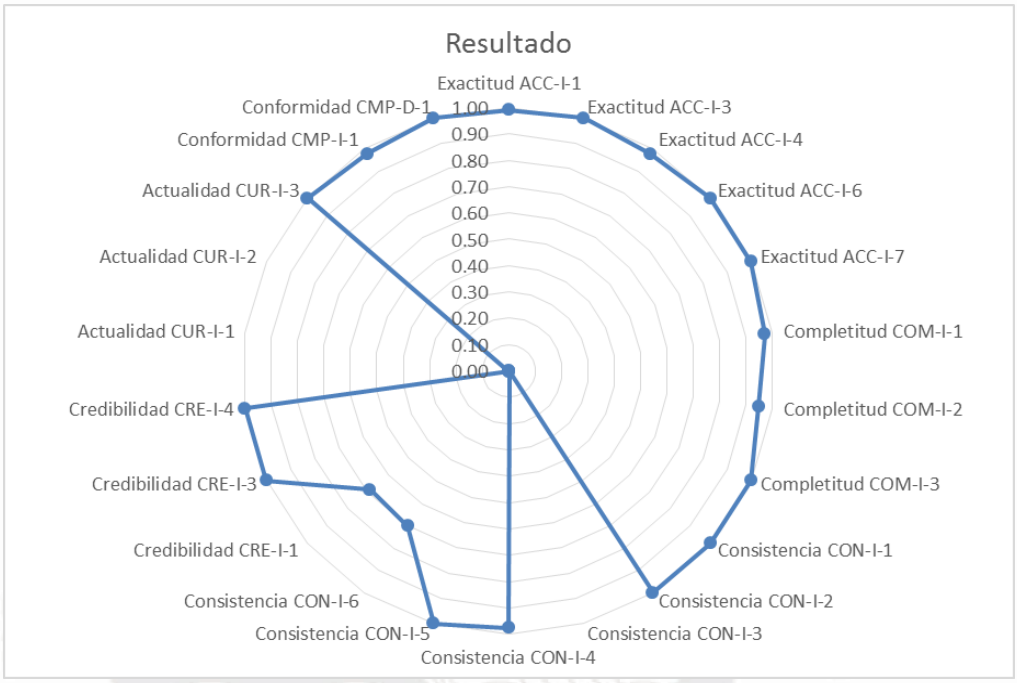
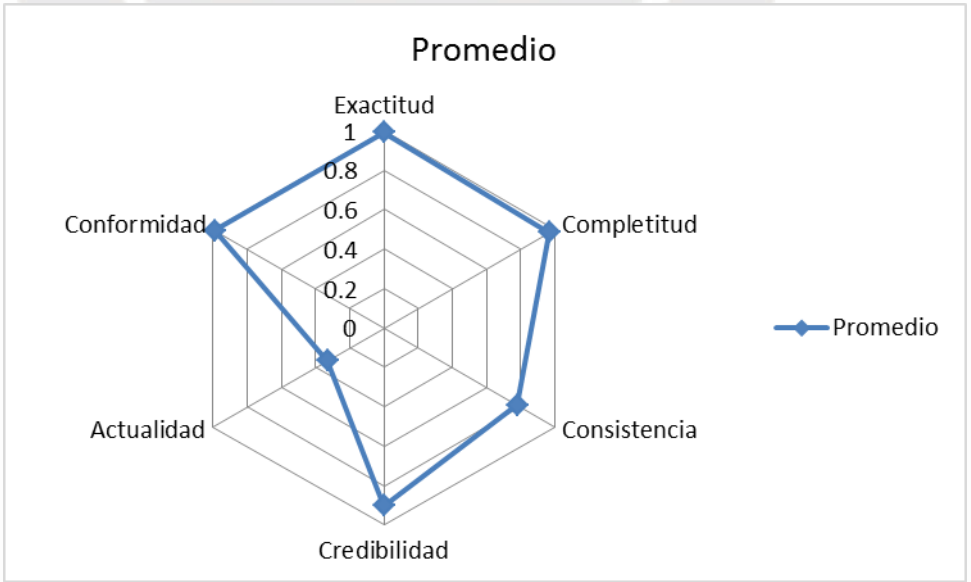


Figura 3.2. Gráfico de radar de las características evaluadas (basado en el promedio de las métricas).
Gráfico diseñado por el autor.



3.8. Análisis de Resultados y Conclusiones de Calidad de Datos

En base a los resultados, se puede decir que a nivel general los datos tienen un nivel de calidad bastante aceptable por parte de los encargados de la solución, especialmente en lo que respecta a su conformidad, exactitud, completitud y credibilidad, pero teniendo problemas en su consistencia y actualidad.

A partir de las métricas obtenidas, se puede decir que se siguen ciertos estándares reglamentarios en la organización. Esto se refleja principalmente en las métricas relacionadas a metadata, diccionario de datos y modelo de datos, las cuales han obtenido resultados óptimos.

Respecto a los datos en sí, se observa que son muy exactos y que existirían pocos casos atípicos que deben ser analizados. Se puede decir, además, que los datos están completos en su mayor parte, aunque esto es en parte debido a que el conjunto de datos posee gran cantidad de llaves; los indicadores tienen un menor grado de completitud, lo cual es crítico en soluciones de Inteligencia de Negocios.

Los datos son consistentes en su gran mayoría (aunque existen fallos en su consistencia semántica). Debido a la cantidad de llaves puede ocurrir un riesgo altísimo de inconsistencia de datos debido a su repetición, un caso que puede manifestarse en caso de un mal uso de los datos.

Los datos son creíbles en gran parte, aunque existe un porcentaje considerable de los datos que está indicado como no confiable, según los criterios establecidos por la organización para el contexto del trabajo.

El punto más bajo de la calidad de los datos está en su actualidad, la cual es directamente afectada por las caídas de los procesos ETL, lo cual provoca que no se actualicen ni en las fechas ni con la frecuencia requerida, a pesar de existir requerimientos de actualización.

Debido a que los datos corresponden a una solución de inteligencia de negocios, y por ende, a un sistema de soporte a las decisiones, las fallas mencionadas son extremadamente críticas, debido a la alta importancia que tienen los datos sobre la organización y a las consecuencias negativas que puede traer su uso incorrecto sobre las políticas de la organización.

4. Conclusiones, Observaciones y Recomendaciones

4.1. Conclusiones

En el presente trabajo se logró desarrollar un modelo de calidad de datos para una solución de Inteligencia de Negocios de una institución educativa, mediante el uso de la técnica de grupo nominal y el método de Brosseau sobre la base de las características definidas en la ISO/IEC 25012, de tal manera que se tuvo en cuenta aquellos atributos que deben tener los datos utilizados.

Se identificó el repositorio de datos que contenía los datos relevantes para realizar el modelo de calidad de datos y la posterior evaluación de calidad del Data Warehouse.

Se seleccionaron métricas tomando como base la ISO/IEC 25024 sobre el contexto específico de la organización.

Se realizó la evaluación de calidad de datos siguiendo las pautas dadas por la norma ISO/IEC 25040, utilizando diversas herramientas y scripts que sirvieron para evaluar los datos de manera rápida.

Se logró analizar los resultados obtenidos, de tal modo que se pudieron realizar reportes que sirven a la organización como una manera de conocer el estado de la calidad de sus datos, así como una fuente de experiencia en el campo.

4.2. Observaciones

El trabajo de selección de características fue muy fiable, debido a que las dinámicas se realizaron con personal experimentado en el análisis y uso de herramientas de Inteligencia de Negocios, lo cual llevaba a un debate para resolver ciertas diferencias respecto a los requisitos de calidad que deben tener los datos. El personal, sin embargo, nunca realizó este tipo de evaluaciones y no tenía claro la definición exacta de las características que se debían buscar.

Además, a pesar que la calidad esté en buen nivel, las fallas encontradas traen un alto riesgo de inconsistencias. Este tipo de solución es fuertemente dependiente de la

calidad de información utilizada tal como se señala en la literatura (Wieder & Ossimitz, 2015), lo cual aumenta el impacto de un posible error en los datos.

Durante la evaluación se observó que de las 22 métricas, 11 obtuvieron una puntuación de 1, 3 obtuvieron una puntuación menor de 0.005 y el resto obtuvo una puntuación mayor a 0.68 (y menor a 1); lo que se justifica en alguna medida por tratarse de mediciones en el contexto de Inteligencia de Negocios.

Como parte de un resultado complementario, en base a este trabajo se desarrolló el artículo *Data quality applied to an academic business intelligence solution: Lesson learned* que fue aceptado y presentado en la conferencia *2017 IEEE Colombian Conference on Communications and Computing (COLCOM)* desarrollada en Cartagena, Colombia entre los días 16 y 18 de Agosto de 2017. El artículo está disponible en la base de datos indexada IEEE Xplore, en el siguiente enlace: http://ieeexplore.ieee.org/document/8088200/.

4.3. Recomendaciones

A partir de los resultados del trabajo, y teniendo en cuenta las posibilidades obtenidas, se puede recomendar lo siguiente a la organización:

- Realizar las gestiones para el análisis de los posibles fallos y riesgos para evitar problemas futuros.

- Utilizar el modelo como base para ampliar la medición a otros conjuntos de datos, y eventualmente a toda la solución, de tal manera que se pueda garantizar la calidad de los datos.

- Preparar un entorno y un plan de evaluación para poder facilitar revisiones posteriores de calidad, así como evitar discordancias debido al posible cambio no anticipado de los datos.

Además, se puede recomendar lo siguiente como trabajos futuros:

- Derivar nuevos modelos de calidad de datos, los cuales pueden ser un replanteamiento del modelo realizado, o generalizado tanto para instituciones educativas (a nivel local) como para soluciones de Inteligencia de Negocios.

- Derivar modelos enfocados a la calidad interna y en uso de los datos, según lo especificado en la ISO/IEC 25012.

- Derivar modelos relacionados a la calidad de las herramientas que consumen los datos utilizados en la solución, tanto externa, interna y en uso.

- Derivar modelos enfocados relacionados a la calidad de los procesos ETL que realizan las transformaciones requeridas para obtener los datos almacenados.

- Elaborar herramientas que usen como base el modelo realizado (o los modelos futuros propuestos) para conseguir realizar la evaluación de forma más rápida y automatizada. Esta herramienta podría unirse con soluciones de Data Mining o de Inteligencia de Negocios para poder conocer la calidad de los datos de forma histórica.

5. Referencias Bibliográficas

A. Alabri, & J. Hunter. (2010). Enhancing the Quality and Trust of Citizen Science Data. En e-Science (e-Science), 2010 IEEE Sixth International Conference on (pp. 81–88). https://doi.org/10.1109/eScience.2010.33

American Society for Quality. (2004). Nominal Group Technique (NGT) - ASQ. Recuperado el 26 de septiembre de 2016, a partir de http://asq.org/learn-about-quality/idea-creation-tools/overview/nominal-group.html

Azvine, B., Cui, Z., Nauck, D., & Majeed, B. A. (2006). Real Time Business Intelligence for the Adaptive Enterprise. Presentado en The 3rd IEEE International Conference on E-Commerce Technology, 2006. The 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services, San Francisco.

Bevan, N. (2010). Los nuevos modelos de ISO para la Calidad y la Calidad en uso del software. En Calidad del producto y proceso software. RA-MA S.A. Editorial y Publicaciones.

Brosseau, J. (2010). Software quality attributes: Following all the steps. Clarrus Consulting Group Inc.

C. Moraga, M. Á Moraga, C. Calero, & A. Caro. (2009). SQuaRE-Aligned Data Quality Model for Web Portals. En 2009 Ninth International Conference on Quality Software (pp. 117–122). https://doi.org/10.1109/QSIC.2009.23

Características de Calidad de Producto de Datos en la ISO 25012. (2016). Recuperado a partir de http://iso25000.com/images/figures/ISO_25012.png

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. Communications of the ACM, 54(8), 88–98.

Dorsett, P. (2014, abril 10). Data Consistency is Key to Analytics. Recuperado el 17 de junio de 2017, a partir de https://www.silvon.com/blog/data-consistency/

Enríquez de Salamanca, J. (2012). Portal Data Quality Assessment Tool. Recuperado el 21 de junio de 2016, a partir de http://podqa.webportalquality.com/PrincipalIng.aspx#Definicion

Forrester Consulting. (2014). Boost The Value Of Your Enterprise BI With Advanced Search. Recuperado a partir de http://go.thoughtspot.com/rs/thoughtspot/images/ThoughtSpot%20Forrester%20White paper%20.pdf

Gartner Inc. (2015, enero 27). Gartner Says Power Shift in Business Intelligence and Analytics Will Fuel Disruption. Recuperado el 15 de abril de 2016, a partir de http://www.gartner.com/newsroom/id/2970917

Hunter, J., Alabri, A., & Ingen, C. (2013). Assessing the quality and trustworthiness of citizen science data. Concurrency and Computation: Practice and Experience, 25(4), 454–466.

Inmon, W. H. (2005). Building the Data Warehouse. John Wiley & Sons, Inc. Recuperado a partir de https://books.google.com.pe/books?id=QFKTmh5IFS4C

International Organization for Standardization. (2002). ISO/IEC 25040:2002-Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Evaluation process.

International Organization for Standardization. (2005). ISO/IEC 25000:2005 -Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE.

International Organization for Standardization. (2008). ISO/IEC 25012:2008 - Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model.

International Organization for Standardization. (2011). ISO/IEC 25010:2011- Systems and software engineering —Systems and software Quality -Requirements and Evaluation (SQuaRE) — System and software quality models.

International Organization for Standardization. (2015). ISO/IEC 25024:2015-Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality.

Kimball, R., & Caserta, J. (2004). The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data. John Wiley \& Sons.

Kimball, R., & Ross, M. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons, Inc.

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Wiley Publishing.

Ley de Protección de Datos Personales, 29733 § (2011). Recuperado a partir de http://www.pcm.gob.pe/transparencia/Resol_ministeriales/2011/ley-29733.pdf

Luhn, H. P. (1958). A business intelligence system. IBM J. Res. Dev., 2(4), 314–319.
Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2015). A Data Quality in Use model for Big Data. Future Generation Computer Systems. https://doi.org/10.1016/j.future.2015.11.024

N. A. H. M. Rodzi, M. S. Othman, & L. M. Yusuf. (2015). Significance of data integration and ETL in business intelligence framework for higher education. En 2015 International Conference on Science in Information Technology (ICSITech) (pp. 181–186). https://doi.org/10.1109/ICSITech.2015.7407800

Natale, D. (2010). ISO/IEC 25012 Modelo de Calidad de Datos y Data Governance. En Calidad del producto y proceso software. RA-MA S.A. Editorial y Publicaciones.

Qlik. (2016). Education BI Solution | Business Intelligence for Education. Recuperado el 21 de abril de 2016, a partir de http://global.qlik.com/us/explore/solutions/industries/public-sector/education

Rafique, I., Lew, P., Qanber Abbasi, M., & Li, Z. (2012). Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model. En International Conference on Information and Systems Engineering (pp. 523–528). waset.org.

Redman, T. (2002). A Long, strange trip ahead: Process management and Data quality. Recuperado a partir de http://web.mit.edu/tdqm/www/iqc/ICIQ02-PMDQRedman.pdf

Roebuck, K. (2011). Data Quality: High-Impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Pty Limited. Recuperado a partir de https://books.google.com.pe/books?id=b5aUZwEACAAJ

Sheldon, R. (2008a). Uso de un esquema tipo Copo de Nieve. Recuperado a partir de https://www.simple-talk.com/iwritefor/articlefiles/592-image003.jpg

Sheldon, R. (2008b). Uso de un esquema tipo Estrella. Recuperado a partir de https://www.simple-talk.com/iwritefor/articlefiles/592-image002.jpg

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. Government Information Quarterly. https://doi.org/10.1016/j.giq.2016.02.001

Wieder, B., & Ossimitz, M.-L. (2015). The Impact of Business Intelligence on the Quality of Decision Making – A Mediation Model. Conference on ENTERprise Information Systems/International Conference on Project MANagement/Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2015 October 7-9, 2015, 64, 1163–1171. https://doi.org/10.1016/j.procs.2015.08.599

Yeoh, W., & Koronios, A. (2010). Critical Success Factors for Business Intelligence Systems. Journal of Computer Information Systems, 50(3), 23–32.