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

ANEXOS

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

Anexos

Contenido

Anexo A: Cuestionario utilizado para obtención de datos

En esta parte se muestra la plantilla brindada a los participantes de la entrevista.

¿Qué características les parecen las más importantes?
Pongan una puntuación a cada característica de acuerdo a su importancia.
Puntuar del 1 (menos importante) al 5 (más importante)

Característica	Ptje.
Exactitud (I): Los datos representan el valor verdadero de un atributo esperado.	
Completitud (I): Los datos asociados con una entidad objetivo tienen los valores esperados para todas las propiedades de una entidad asociada.	
Consistencia (I): Los datos tienen atributos libres de contradicciones y con coherentes con otros datos en un contexto específico de uso.	
Credibilidad (I): Los datos tienen atributos considerados como verdaderos y creíbles por los usuarios.	
Actualidad (I): Los datos tienen atributos de la edad correcta.	
Accesibilidad (I/DS): Los datos pueden ser accedidos en un contexto específico de uso, particularmente por personas que requieran tecnologías de apoyo o configuraciones especiales debido a una discapacidad.	
Conformidad (I/DS): Los datos tienen atributos que se adhieren a los estándares, convenciones o regulaciones vigentes relacionadas a calidad de datos.	
Confidencialidad (I/DS): Los datos solo son accesibles e interpretables por usuarios autorizados en un contexto específico de uso.	
Eficiencia (I/DS): Los datos pueden ser procesados y brindan los niveles esperados de rendimiento usando los tipos y cantidades apropiadas de recursos.	
Precisión (I/DS): Los datos son exactos o brindan discriminación.	
Trazabilidad (I/DS): Los datos permiten auditar el acceso y los cambios hechos a los datos.	
Entendibilidad (I/DS): Los datos pueden ser leídos e interpretados por los usuarios, y se expresan en un lenguaje apropiado.	
Disponibilidad (DS): Los datos pueden ser obtenidos por usuarios autorizados y/o aplicaciones.	
Portabilidad (DS): Los datos tienen atributos que le permiten ser instalados o movidos de un sistema a otro preservando la calidad existente.	

Anexo B: Datos obtenidos del cuestionario

En esta parte se muestran los datos provistos por los participantes en la entrevista.

Característica	Participante 1	Participante 2	Participante 3
Exactitud	5	5	5
Completitud	4	4	4
Consistencia	5	5	5
Credibilidad	5	5	5
Actualidad	5	5	5
Accesibilidad	1	1	2
Conformidad	5	5	5
Confidencialidad	5	5	5
Eficiencia	1	3	4
Precisión	3	4	5
Trazabilidad	1	4	4
Entendibilidad	4	3	4
Disponibilidad	4	4	4
Portabilidad	1	2	4
Recuperabilidad	5	5	5

Anexo C: Obtención de Características del Modelo de Calidad

En este anexo se detallarán el proceso utilizado para obtener las características relevantes para la calidad, así como se indicarán las características resultantes.

AC.1. Proceso de obtención

Para obtener estas características, se siguió una combinación de la Técnica de Grupo Nominal y el método de Jim Brosseau.

AC.1.1. Detalle de la Técnica de Grupo Nominal

Para esta parte, se siguió el proceso especificado en. Se trabajó con un grupo de 3 participantes, todos expertos en Inteligencia de Negocios en la organización.

Se hizo que los participantes respondieran a la siguiente pregunta: **¿Qué características les parecen las más importantes?**

Cada participante eligió sus ideas en base a dar una calificación a cada característica, de tal manera que se pueda tener una medida cuantificable que pueda ser utilizada luego en la discusión.

Posteriormente, los participantes discutieron entre ellos la importancia de las características, indicando su justificación.

Para finalizar, los participantes indicaron cuáles eran las características más importantes para ellos, con lo que se obtuvo la lista final utilizada para el modelo de calidad. Estos resultados se entregaron de forma cuantificada, la cual se muestra en la Tabla AC.1.

Tabla AC.1. Resultados cuantificados de la metodología, con características elegidas resaltadas.
Cuadro diseñado por el autor.

Característica	Participante 1	Participante 2	Participante 3	Promedio
Exactitud	5	5	5	5
Completitud	4	4	4	4
Consistencia	5	5	5	5
Credibilidad	5	5	5	5
Actualidad	5	5	5	5
Accesibilidad	1	1	2	1.33333333
Conformidad	5	5	5	5
Confidencialidad	5	5	5	5

Característica	Participante 1	Participante 2	Participante 3	Promedio
Eficiencia	1	3	4	2.66666667
Precisión	3	4	5	4
Trazabilidad	1	4	4	3
Entendibilidad	4	3	4	3.66666667
Disponibilidad	4	4	4	4
Portabilidad	1	2	4	2.33333333
Recuperabilidad	5	5	5	5

La lista de características elegidas se muestra en la tabla AC.2.

Tabla AC.2. Características obtenidas por la Técnica de Grupo Nominal.
Cuadro diseñado por el autor.

Características obtenidas
Exactitud
Consistencia
Credibilidad
Actualidad
Conformidad
Confidencialidad
Recuperabilidad

AC.1.2. Detalle del Método de Jim Brosseau

Para agregar el criterio de evaluación del autor, se utilizó el método de Jim Brosseau como una manera formal de seleccionar las características. Se realizó a partir de las características en la Tabla AC.1, y su aplicación se muestra en la Tabla AC.3.

Tabla AC.3. Método de Brosseau.
Cuadro diseñado por el autor.

Característica	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	S
Exactitud (C1)	<	<	<	<	<	<	<	<	<	<	<	<	<	<	14
Completitud (C2)		^	^	<	<	<	<	<	<	<	<	<	<	<	11
Consistencia (C3)			^	<	<	<	<	<	<	^	^	<	<	<	10
Credibilidad (C4)				<	<	<	<	<	<	<	<	<	<	<	13
Actualidad (C5)					<	<	<	<	<	<	<	<	<	<	10
Accesibilidad (C6)						^	^	^	^	^	^	^	^	^	0
Conformidad (C7)							^	<	<	<	<	<	<	<	8
Confidencialidad (C8)								<	<	<	<	<	<	<	10
Eficiencia (C9)									<	<	^	^	<	<	5
Precisión (C10)										<	<	^	<	<	6
Trazabilidad (C11)											^	<	<	^	4
Entendibilidad (C12)												^	<	<	6

Característica	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	S
Disponibilidad (C13)													<	<	6
Portabilidad (C14)														^	2
Recuperabilidad (C15)															1

A partir de los resultados del método, se observó lo siguiente:

- Las características importantes suelen ser del tipo inherente.
- Una característica obtuvo el menor puntaje general (Accesibilidad).

Por este motivo, y mediante conversaciones con los expertos en Inteligencia de Negocios, se comparó esta lista con la lista inicial obtenida de la técnica de grupo nominal, con lo cual se obtuvo la lista mostrada en la Tabla AC.4.

La característica de confidencialidad fue eliminada debido a que no hay acceso a los datos requeridos para su medición, previo acuerdo con los encargados de la solución.

Tabla AC.4. Características obtenidas mezclando el método de Brosseau y la Técnica de Grupo Nominal.
Cuadro diseñado por el autor.

Características obtenidas
Exactitud
Completitud
Consistencia
Credibilidad
Actualidad
Conformidad

AC.2. Definición de las Características

Para conocer la definición de las características, referirse a los conceptos especificados en 2.2.4 y en la ISO/IEC 25012 (International Organization for Standarization, 2008).

Anexo D: Plan de Evaluación

En este anexo se especifican los parámetros de la evaluación según lo especificado en la ISO/IEC 25040.

AD.1. Requerimientos de evaluación

Para obtener los requerimientos, se han tenido en cuenta los siguientes puntos:

Propósito de la evaluación: Asegurar la calidad del producto.

Interesados:

- Encargados de la solución de Inteligencia de Negocios
- Evaluador (el autor, también es usuario de la solución)

Los requerimientos de evaluación serán los siguientes:

- Se debe asegurar la calidad del conjunto de datos acordado con los encargados de la solución.
- Las características y métricas deben ser evaluadas en conjunto con los encargados.
- Las mediciones deben ser documentadas para mostrar a los encargados.
- Se realizará un reporte final de la evaluación.
- Se debe seguir el acuerdo de confidencialidad para la evaluación en todo momento.

Partes del producto a incluir:

- Data Warehouse de la institución educativa Edu.Lambda.

AD.2. Especificación de la evaluación

AD.2.1. Especificación de Métricas

Las métricas se encuentran especificadas en las Tablas 3.2 a 3.9 del documento principal.

AD.2.2. Planificación de la evaluación

Teniendo en cuenta las restricciones de tiempo del proyecto, y los recursos facilitados tanto por la organización como por el grupo GIDIS-PUCP, se tiene el plan de trabajo mostrado en la tabla AD.1:

Tabla AD.1. Plan de evaluación.
Cuadro diseñado por el autor.

Actividad / Descripción	Recursos	Fecha de inicio	Fecha de Fin
Identificación de Registros de Datos Identificar los registros sobre los que se aplicará el modelo de calidad a utilizar.	• Aplicación de gestión de base de datos. • Coordinación con expertos.	09/08/16	08/09/16
Definición de Características de Calidad Escoger las características que deben cumplir los datos y que deben ser evaluadas.	• Norma ISO/IEC 25012. • Coordinación con expertos.	08/08/16	16/09/16
Definición de Métricas y Scripts de Calidad Elegir las métricas para la medición de las características y elaborar los scripts que permiten obtener las entradas requeridas para las métricas.	• Norma ISO/IEC 25024. • Aplicación de gestión de base de datos. • Aplicación de medición estadística. • Coordinación con expertos.	16/09/16	30/09/16
Planificación de la Evaluación Realizar la planificación y acordar la forma en que se va a realizar con los encargados de la solución.	• Norma ISO/IEC 25040. • Coordinación con expertos.	30/09/16	07/10/16
Proceso de Evaluación Evaluar la calidad de datos aplicando los scripts respectivos para evaluar las métricas.	• Norma ISO/IEC 25040. • Scripts de consulta de datos.	10/10/16	21/10/16
Análisis de resultados y conclusiones Realizar el estudio de los resultados de las métricas y obtener conclusiones que lleven a la mejora de la calidad.	• Aplicación de medición estadística.	21/10/16	27/10/16

Anexo E: Detalle total de Métricas evaluadas

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-3 Aseguramiento de exactitud de datos

Cálculo de la métrica:

X = A/B
A = Campos de datos evaluados
B = Total de campos de datos

Forma de cálculo:

Como se ha evaluado el 100% del conjunto de datos posible, X = 1.

Comentarios:

Esta métrica indica que la totalidad del conjunto de datos está siendo evaluado, dando sustento a las métricas de exactitud.

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.
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.

ACC-I-6 Exactitud de Metadata

Cálculo de la métrica:

X = A / B
A = Número de metadata que proporciona la información requerida apropiada = 58
B= Número de metadata definido dentro de la especificación de requerimientos de datos = 58
X = 1

Forma de cálculo:

Se observaron los diferentes campos que forman parte de la metadata del conjunto de datos y se indicó si apoyaban al entendimiento de los datos.

Comentarios:

La metadata colocada en el conjunto de datos describe adecuadamente los diversos campos y el conjunto de datos elegido, lo cual indica que los estándares relacionados a la metadata son adecuadamente seguidos.

ACC-I-7 Rango de exactitud de datos

Cálculo de la métrica:

X = A / B
A = Cantidad de datos que se encuentran dentro del intervalo de datos requerido = 3499176
B = Cantidad de datos que tienen un requerimiento de intervalo de datos = 3499176
X = 1

Forma de cálculo:

Para calcular los resultados de esta métrica, se siguieron los siguientes scripts:

```
SELECT COUNT(ATRIBUTO)
FROM LAMBDA.WAREHOUSE
WHERE ATRIBUTO >= LIMINFRANGO and ATRIBUTO <= LIMSUPRANGO
```

Comentarios:

Los atributos con rango de datos definido cumplen con el indicador, lo cual indica que en la organización se tienen claro los intervalos en que debe encontrarse la data para su utilización e interpretación adecuada.

COM-I-1 Completitud de Registros

Cálculo de la métrica:

Xpromedio = Promedio(X)
X = A/B (para cada registro)
A = Conteo de valores no nulos en un registro (depende del registro)
B = Total de valores evaluados por registro = 57
Xpromedio = 0.97

Forma de cálculo:

Se utilizó el siguiente script, para cada atributo en un registro:

```
SELECT CASE ATRIBUTO
IF SINDATO THEN 0
IF NULL THEN 0
ELSE 1
END
FROM LAMBDA.WAREHOUSE
```

Posteriormente, los valores obtenidos para cada atributo son sumados, y se obtiene el promedio de todos los atributos.
Además, se consideraron los valores "sin dato" como valores nulos.

Comentarios:

La gran mayoría de los atributos en los registros están llenos, esto debido a que la mayoría son atributos con integridad referencial definida (llaves) e indicadores provenientes del resultado de procesos ETL. La reducción ocurre debido a los indicadores definidos, los cuales no están completos en su totalidad.

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 AE.1. 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 AE.1. 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.

COM-I-3 Completitud de metadata

Cálculo de la métrica:

X = A/B
A = Cantidad de metadata completa = 58
B = Cantidad de metadata que requiere completitud = 58
X = 1

Forma de cálculo:

Se observaron los diferentes campos que forman parte de la metadata del conjunto de datos y se indicó si estaban completos y no tenían valores nulos.

Comentarios:

La metadata está completa de acuerdo a los requerimientos establecidos por la organización, esto debido a la necesidad de completar los datos para conocer su significado.

CON-I-1 Integridad Referencial

Cálculo de la métrica:

X = 1-A/B

A = Número de campos que no cumplen la integridad referencial = 0

B = Número de campos para los cuales debe estar definida la integridad referencial = 43156504

X = 1

Forma de cálculo:

Para cada atributo, se realiza el siguiente script por cada atributo:

```
SELECT ATRIBUTO FROM
LAMBDA.WAREHOUSE
WHERE ATRIBUTO not in
(SELECT ATRIBUTO FROM LAMBDA.DIMENSION)
```

Comentarios:

La integridad referencial es parte de los estándares del Data Warehouse, por lo que se coloca la metadata pertinente al momento de su mantenimiento, lo que hace que esté completa.

CON-I-2 Consistencia de formato de datos

Cálculo de la métrica:

X = A/B
A = Campos con el mismo formato estandarizado = 57
B = Campos que requieren tener el formato estandarizado = 57
X = 1

Comentarios:

Dado que se sigue un esquema típico de una solución de inteligencia de negocios (fact table y dimensiones), los constraints provocan que el formato de los datos esté estandarizado en todo el conjunto de datos.

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-4 Consistencia de la arquitectura

Cálculo de la métrica:

X = A/B
A = Número de elementos que coinciden con la arquitectura instalada = 39
B = Número de elementos de la arquitectura instalada = 40
X = 0.975

Forma de cálculo:

Para la medición, se contaron los elementos (campos y tablas) que cumplen con los estándares de arquitectura seguidos para el modelo de base de datos de esta solución.
El conteo se realizó de forma manual.

Comentarios:

Dado que el conjunto de datos está basado en una tabla de hechos en conjunto con sus dimensiones, se puede decir que en su mayor parte se cumplen los estándares de arquitectura realizados para una solución basada en los esquemas estrella y copo de

nieve, así como la organización que pide la metodología planteada para su mantenimiento.

CON-I-5 Cobertura de consistencia de valores de datos

Cálculo de la métrica:

X = A/B
A = Campos de datos evaluados
B = Total de campos de datos

Forma de cálculo:

Como se ha evaluado el 100% del conjunto de datos posible, X = 1.

Comentarios:

Esta métrica indica que la totalidad del conjunto de datos está siendo evaluado, dando sustento a las métricas de consistencia.

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.

CRE-I-3 Credibilidad de diccionario de datos

Cálculo de la métrica:

X = A / B
A = Items en el diccionario de datos los cuales son validados por un proceso específico = 57
B = Total de Items en el diccionario de datos = 57
X = 1

Forma de cálculo:

Se han visto los valores en el diccionario de datos de la organización, los cuales fueron verificados con los encargados para saber si fueron validados.

Resultado:

La métrica indica que el diccionario de datos es validado cada vez que se realiza una modificación, de acuerdo con los estándares organizacionales definidos.

CRE-I-4 Credibilidad de modelo de datos

Cálculo de la métrica:

X = A / B

A = Items en el modelo de datos los cuales son validados por un proceso específico = 57

B = Total de items en el modelo de datos = 57

X = 1

Forma de cálculo:

Se ha revisado el modelo de datos y se consultó con los encargados si estaban validados. Asimismo, se vio qué ítems son validados por un proceso ETL.

Comentarios:

La métrica indica que el modelo de datos es validado cada vez que se realiza un pase para su modificación, de acuerdo con los estándares organizacionales definidos.

CUR-I-1 Frecuencia de actualización

Cálculo de la métrica:

X=A/B

A = Total de datos que son actualizados con la frecuencia requerida = Ninguno (se actualizó una cantidad de veces inferior a la debida)

B = Total de datos que tienen una frecuencia requerida de actualización = Todos

X = 0

Forma de cálculo:

Para el cálculo, se revisan los logs de los procesos ETL y se verifica si la actualización se hizo en las fechas requeridas.

Resultado:

Debido a no estar actualizado a la fecha, se presume que hay errores en los procesos ETL que provocan que los datos no se actualicen con la frecuencia que deberían.

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.

CUR-I-3 Requisito de actualización de ítems

Cálculo de la métrica:

X=A/B
A = Total de datos que tienen un requisito explícito de actualización = Todos
B = Total de datos que necesitan un requisito explícito de actualización = Todos
X = 1

Forma de cálculo:

Se revisaron los datos que tienen un requisito de actualización. Para el caso de este Data Warehouse, existe un calendario con las fechas de obtención de ETL, en el cual se tienen las fechas en que los datos deben actualizarse.

Comentarios:

Se manejan fechas para la actualización de los datos, lo cual indica que todos los datos en el conjunto (y en general en el Data Warehouse) requieren actualización cada cierto tiempo, lo cual es revisado de manera frecuente; esto indica la existencia de estándares organizacionales para el manejo de las actualizaciones.

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
```

y se le resta al total.
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.

CMP-D-1 Conformidad regulatoria debido a la tecnología

Cálculo de la métrica:

X = A/B

A = Cantidad de ítems de datos que están conformes a regulación debido a la tecnología = Todos

B = Cantidad de ítems de datos que deben estar conformes a regulación debido a la tecnología = Todos

X = 1

Forma de cálculo:

Para el cálculo de esta métrica, se contaron todos los elementos que no cumplen con la regulación debido a fallas tecnológicas y se le restó esta cantidad al total. El conteo se realizó de forma manual, dado que principalmente la conformidad a nivel tecnológico ocurre a través de los formatos y la configuración de los campos del conjunto de datos.

Comentarios:

La evaluación indica que la tecnología no está provocando fallas que afecten a la conformidad de los datos almacenados, con lo cual se puede decir que la configuración física de los datos es confiable.

Anexo F: Reporte de la evaluación

En el presente anexo se presenta un reporte realizado para la presentación formal de los resultados de la evaluación. La estructura del reporte fue adaptada a partir de la ISO/IEC 25062.

1. Información inicial

1.1. Nombre del producto: Data Warehouse de la Institución Educativa Lambda

1.2. Versión: 1.0

1.3. Responsable de la evaluación: Marshall Fernández Sáenz

1.4. Fechas de Evaluación: Del 10 de Octubre de 2016 al 21 de Octubre de 2016

1.5. Fecha de Elaboración del Reporte: 28 de Octubre de 2016.

1.6. Responsable del reporte: Marshall Fernández Sáenz

1.7. Teléfono de contacto: 952332772

1.8. E-mail de contacto: marshall2004_33@hotmail.com

1.9. Nombre de la Institución: Lambda

2. Resumen ejecutivo

2.1. Nombre y descripción del producto:

Data Warehouse de la Institución Educativa Lambda.

En esta base de datos se almacenan los datos que son útiles para su análisis por parte de los ejecutivos de la organización. Esta base de datos es alimentada por una serie de procesos ETL utilizados para la transformación de los datos.

2.2. Resumen de los métodos utilizados

Se elaborará una serie de scripts de consulta sobre los datos. Estos scripts serán ejecutados sobre un conjunto de datos que es parte de la base de datos mencionada.

2.3. Resultados cuantificables

Se mostrarán resultados a partir de la medición de cada métrica, los cuales se expresan en valores entre 0 y 1.

2.4. Justificación y naturaleza de la prueba

La justificación y naturaleza se encuentra en la problemática y en la justificación del documento principal del documento de tesis (partes 1.1 y 1.8).

En resumen, la justificación es la obtención de experiencia útil para la evaluación de calidad, así como el aseguramiento permanente de la calidad del Data Warehouse mencionado.

3. Introducción

3.1. Descripción completa del producto

En el Data Warehouse de la Institución Educativa Lambda se almacenan los datos que son útiles para su análisis por parte de los ejecutivos de la organización. Esta base de datos es alimentada por una serie de procesos ETL utilizados para la transformación de los datos.

Se evaluó un conjunto de datos elegidos del Data Warehouse, elegido en base a su relevancia para la organización y de acuerdo a lo acordado con los encargados de la solución.

Esta solución es utilizada tanto para la revisión de datos, tanto por los usuarios organizacionales como por diversas aplicaciones que permiten el acceso de los ejecutivos a los datos almacenados mediante el consumo de la información del Data Warehouse.

3.2. Objetivos de la prueba

Tesista/Evaluador:

- Conclusión de los estudios de pregrado, y obtención de experiencia respecto a calidad.

Grupo de Investigación GIDIS (incluye al Asesor y Coasesor):

- Realizar un modelo de calidad de datos para el Data Warehouse de la Institución Educativa Lambda.
- Realizar un aporte a la investigación realizada por el proyecto ProCal-ProSer.
- Aportar al aprendizaje del tesista.

Encargados del Data Warehouse:

- Asegurar permanentemente la calidad de los datos almacenados.

4. Método

4.1. Participantes:

- Evaluador

4.2. Pasos a realizar

El método utilizado para la evaluación consiste en los siguientes pasos:

- Obtener las características de calidad que deben cumplir los datos almacenados.
- Definir las métricas que sean relevantes para la evaluación de las características.
- Elaborar y ejecutar scripts que permitan la obtención de los parámetros utilizados para el cálculo de las métricas.
- Realizar el análisis de las métricas obtenidas para llegar a las conclusiones del trabajo.

5. Métricas

Las métricas evaluadas se muestran en la tabla AF.1 y fueron las siguientes:

Tabla AF.1. Métricas evaluadas.
Cuadro diseñado por el autor.

Característica	Métrica
Exactitud	ACC-I-1
	ACC-I-3
	ACC-I-4
	ACC-I-6
	ACC-I-7

Característica	Métrica
Completitud	COM-I-1
	COM-I-2
	COM-I-3
Consistencia	CON-I-1
	CON-I-2
	CON-I-3
	CON-I-4
	CON-I-5
	CON-I-6
Credibilidad	CRE-I-1
	CRE-I-3
	CRE-I-4
Actualidad	CUR-I-1
	CUR-I-2
	CUR-I-3
Conformidad	CMP-I-1
	CMP-D-1

El detalle de las métricas se puede encontrar en el Anexo E.

6. Resultados

Los resultados de la medición se muestran en la tabla AF.2 y fueron los siguientes:

Tabla AF.2. Resultados de la medición realizada.
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	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	

Característica	Métrica	Resultado	Medición	Promedio de característica
	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 la obtención de estos datos, se ejecutaron los scripts sobre un conjunto de datos establecido entre el evaluador y los encargados de la solución.