Marco de Referencia para la Evaluación de la Calidad de Datos: Un enfoque basado en un Scorecard (Investigación en Curso)

J. Mauricio Pinto V., Lorena G. Gómez M. & Miguel Ángel Pérez G.

Tecnológico de Monterrey - Campus Monterrey
Monterrey, NL México
mauricio.pinto@xenxar.com,lgomez@itesm.mx,maperez@itesm.mx
http://www.itesm.edu

Resumen La Calidad de Datos (DQ[1]) es un área de investigación relativamente nueva y que ha suscitado mayor interés, debido a la importancia de los datos electrónicos. La DQ es un concepto multifacético y existen en la literatura bases suficientes para su caracterización. Los pilares de la DQ son las Dimensiones, que pueden aplicar al esquema y a los datos.
Las dimensiones han sido utilizadas hasta ahora como la forma para evaluar la DQ, sin embargo están orientadas a calificar los datos sin tomar en cuenta el contexto en que son usados. El objetivo de este artículo es exponer la investigación en curso, que propone un modelo para la evaluación de Datos, basado en un *Scorecard*[2]. Los *Scorecards* han sido usados ampliamente en otras áreas para la medición de variables de desempeño no financieras y críticas, expresados en términos del contexto de negocio.
Se pretende definir un conjunto de KPI[3] de DQ, que puedan ser evaluados y validados, a través de un caso de Estudio aplicado en área de la Salud.

Palabras Clave: Data,Information, Data Quality, Information Quality, Database, Data Quality Dimensions, Scorecard

1 Introducción

Actualmente los datos electrónicos juegan un rol primordial en todas las actividades de la Sociedad de la Información[4]. Las Tecnologías de la Información y Comunicaciones, son ya una parte imprescindible de las empresas, organizaciones e individuos, donde fluyen los datos a una velocidad y cantidad cada vez mayor.

Varios fenómenos han suscitado el interés en los datos. Entre estos el advenimiento de la computadora en el siglo pasado, el abaratamiento y empoderamiento de las

[1] En adelante, se usará el acrónimo "DQ" para referirse a "Calidad de Datos."
[2] Tablero de Mando.
[3] Key Performance Indicators.
[4] Término acuñado por Peter Drucker en 1969

M.A. Cruz-Chávez (Ed): CICos 2011, ISBN. 978-607-00-5091-6. pp. 301 – 321, 2011

computadoras personales en los años ochenta y la interconectividad regional y global en el final del siglo veinte. El crecimiento y la complejidad de las organizaciones, han puesto al descubierto la importancia de los datos, como un activo y un recurso imprescindible para las operaciones de estas.

Este dramático crecimiento de las organizaciones ha ido a la par con el crecimiento de la cantidad de datos recopilados, transportados y almacenados. Esto ha creado problemas y nuevos retos, que permitan garantizar que el activo más importante —los datos—, tengan un nivel aceptable para el uso que se les quiera dar. La calidad de datos puede degradarse rápidamente conforme crece su volumen y más aun con el intercambio incontrolado que permite la interconectividad en la Infósfera.

La norma ISO 9000 define la Calidad como "El grado en el que un conjunto de características inherentes cumple con los requisitos". Wikipedia define Dato como "una representación simbólica: numérica, alfabética, algorítmica etc.; un atributo o una característica de una entidad". También define Información como "Un conjunto organizado de datos procesados, que constituyen un mensaje que cambia el estado de conocimiento del sujeto o sistema que recibe dicho mensaje".

Los datos no tendrían utilidad alguna si no son procesados para obtener la Información. La Información es entonces el recurso y activo que permitirá tomar una decisión en un contexto determinado. Es decir, tanto los Datos como la Información, deben satisfacer (o exceder) los requisitos de quien haga uso de estos. J. M. Juran sostiene que existirá un alto nivel de Calidad de Datos "si estos se ajustan a lo que se espera de su uso en las operaciones, toma de decisiones y planeación" [10] .

Para que los datos sean un instrumento útil en la toma de decisiones, deben cumplir criterios de calidad. La DQ reside en el nivel de confianza que los consumidores de información depositen en esta. Si la confianza es baja, quiere decir que la información como un recurso de entrada no esta teniendo el nivel de calidad requerido. Es imperativo que la información cuente con un nivel de calidad aceptado por los consumidores, porque es la base en la que se sustentarán las decisiones en todos los niveles.

Según un reporte del "The Datawarehouse Institute" [3], se estima que los problemas relacionados con la DQ, causan una pérdida a las empresas estadounidenses de seiscientos billones de dólares cada año. Además de esto en la literatura revisada se encontraron efectos negativos e incidentes desastrosos, cuyo origen esta directamente relacionado con la DQ.

Uno de los incidentes más llamativos y conocidos es el del transbordador espacial Challenger [6], donde según las investigaciones se encontraron al menos diez categorías de DQ que tuvieron relación causal con la tragedia. Otro evento conocido es el del "Y2K", que fue relacionado con el cambio de dígito para el año dos mil y cuyo costo para resolverlo se estima en un trillón y medio de dólares [4]. Estos son solo algunos casos que ponen de manifiesto, lo crítico del tema de la DQ. Se pueden encontrar en la literatura, más casos documentados (ver [5] [11]).

La DQ tiene dos áreas principales, que son la Evaluación y la Mejora. La evaluación tiene que ver con el diagnóstico y la detección de las causas raíz de la DQ. La mejora es el proceso por el cual se pueden eliminar o mitigar los eventos indeseados que están generando una baja Calidad de Datos. Estos son procesos complementarios y comprenden en su totalidad la visión holística de la DQ. El área que se pretende abordar en este artículo, es la Medición de la DQ.

En la literatura revisada, se encontró coincidencia con un elemento clave para la Evaluación DQ: Las dimensiones. Como se dijo anteriormente, la DQ se aborda con una perspectiva de múltiples facetas que son las dimensiones. Estas son criterios que califican un aspecto del conjunto de datos a evaluar. Este aspecto puede ser intr´ınseco o extr´ınseco a los datos y se instrumenta a través de una o más métricas asociadas a cada dimensión.

En la literatura se encuentran una lista extensa de dimensiones a considerar, sin embargo existe consenso en las dimensiones principales. Estas son: Exactitud, Completud, Vigencia y Consistencia[1]. Se debe hacer hincapié en que, si bien son cuatro las dimensiones principales, no se puede generalizar su uso para todas las actividades de DQ, menos a´un para todos los dominios de problema. Es menester realizar una selección objetiva y subjetiva de las dimensiones a considerar para cada caso[1].

Los nombres de las dimensiones son intuitivos e indican una cualidad respecto de un conjunto de datos a evaluar, a saber:

- **Exactitud:** Que el conjunto de datos sea exacto en su contenido.
- **Completud:** Que el conjunto de datos este completo.
- **Vigencia:** Que el conjunto de datos este vigente en relación al tiempo.
- **Consistencia:** Que el contenido del conjunto de datos sea coherente a su dominio.

La DQ es por definición un nivel que refleja la calificación de las dimensiones que la componen. Se pueden describir tres momentos en la Evaluación de la DQ: La definición de métricas, que consiste en establecer el mecanismo, la escala o metodología para obtener una medida.

La medición, que es el proceso de obtener las métricas de un conjunto de datos determinado.

La Evaluación, que el proceso de comparar las medidas obtenidas contra un criterio de calidad o dimensión.

Se plantea el *Scorecard*, dado que es un modelo que ha dado buenos resultados en la industria y es ampliamente aceptado en la comunidad académica.

El objetivo es ligar la Evaluación de la DQ a otras perspectivas del desempeño organizacional, e instrumentarla mediante unos KPI específicos para la DQ, que le den el contexto de la Evaluación en función de su contribución al desempeño del negocio.

Otro beneficio del enfoque basado en *Scorecard*, es que brinda un instrumento para la medici´on continua y con tendencia al tiempo real. En el caso de la DQ, se tendr´a un monitoreo continuo del comportamiento de los KPI relacionados con esa perspectiva. Esto dará un enfoque más proactivo, diferente a la forma convencional que es más bien reactiva o en demanda. Elevar la eficiencia en el proceso de Evaluación de la DQ, significará una mejora sustantiva en el proceso de Mejora de la DQ.

En este artículo, se presenta la Construcción del Marco de Referencia para la Evaluación de la DQ, basado en el enfoque de *Scorecard* y KPI. Para la validación de este modelo se diseñará además una prueba piloto, en una entidad pública local.

En la siguiente sección se desarrollarán los conceptos que fundamentan este artículo, además del estado del arte de la Calidad de Datos. Posteriormente se

exponen las motivaciones para realizar esta investigación, así como la definición de la Problemática, objetivos, metodología y los resultados preliminares. Al final del artículo se presentan las conclusiones.

2 Calidad de Datos

La Calidad no es un concepto nuevo y ha ido evolucionando para convertirse en un elemento clave, a lo largo de todas las actividades de la sociedad. La Norma ISO 9000 define a la Calidad como "el grado en el que un conjunto de características inherentes cumple con los requisitos". Para poder aseverar que un servicio o producto es de Calidad, se deberá hacer una evaluación, que nos permita conocer el grado de desviación de las características esperadas, para luego analizar las causas y ejecutar acciones preventivas o correctivas.

Una búsqueda del término Data Quality en Google, arroja 434,000,000 resultados. Esto da cuenta de la importancia que ha adquirido esta temática. El interés no solo es en la comunidad de negocios, sino que también la academia ha ido tomando este tema que ha adquirido mayor connotación en la última década.

Una de las razones es la tasa de crecimiento sostenido de los datos a nivel mundial. Según un estudio de la Universidad de California en Berkley el mundo produce de 1 a 2 billones de bytes anualmente.

A continuación se darán las bases para poner en contexto el concepto de Calidad en los Datos y como se caracterizan estos para poder generar un modelo de evaluación y mejora. También se expondrá la importancia de los datos como recurso y activo, para luego analizar la DQ desde la perspectiva de Sistemas. Por último se explicarán las dimensiones de la DQ, que son la base para cualquier actividad que este relacionada con la DQ.

2.1 Datos: Activo Empresarial

Los activos empresariales son diversos por su naturaleza, pueden ser tangibles e intangibles. Entre algunos ejemplos están los recursos humanos, los financieros, la marca y muchos otros. La característica común a los activos empresariales es que tienen un valor reconocido y controlado por la organización. Este valor es otorgado, por la capacidad de que estos recursos y activos puedan coadyuvar, a que la organización que los controla y posee pueda usarlos para lograr sus objetivos y metas [8]. Es deseable entonces que estos activos sean gobernados para su óptimo aprovechamiento y la posibilidad de elevar aun más el nivel de calidad de estos activos.

Partiendo de este precepto, podemos aseverar que los Datos son un activo. Los Datos son un activo intangible con un valor inherente. En los activos tangibles se pueden evidenciar visualmente su estado de aprovechamiento y funcionamiento, en los datos esto representa una barrera para la toma de conciencia de la importancia que tienen. Si bien los datos esta almacenados en medios físicos, esto no demuestra que estos estén en óptimo estado. Los datos deben ser tratados con el mismo cuidado que un activo tangible [8].

Los datos son un activo de vital importancia para el funcionamiento de la empresa y la sociedad. Pero para que los datos puedan ser aprovechados necesitan tener una alta calidad. El tener un déficit de calidad en los datos, como en cualquier recurso de la empresa, significaría que podría ser la causante de que no se logren los objetivos en mayor o menor medida. Fisher, afirma que existe una clara relación entre el logro de objetivos de una empresa y como administran sus datos[7].

La toma de conciencia de que los Datos son un activo y recurso importante para la organización, obliga a que se tomen todos las medidas necesarias para que su gobernancia y calidad estén garantizadas. De otra forma existirá una incertidumbre que podría ser la raíz de problemas difíciles de diagnosticar y dependiendo la criticidad podrán ir de sencillos a fatales.

2.2 Datos vs. Información vs. Conocimiento

Los datos son la representación de eventos de diversa naturaleza como texto, sonidos, imagen, etc. La palabra dato deriva del Latín datum, que significa hecho. El dato representa una abstracción del mundo real, limitada a las características o propiedades de interés. Representarlas todas o demasiadas sería impráctico y costoso. No representar las suficientes o representarlas de manera errónea, causaría una distorsión entre lo representado y la realidad.

Otra características de los datos es que pueden representar el hecho mismo o inclusive a la estructura de si mismos. A esto se lo conoce como metadata o "los datos de los datos". Ambos son importantes porque de nada serviría tener un contenido, sin tener la estructura del esquema de este.

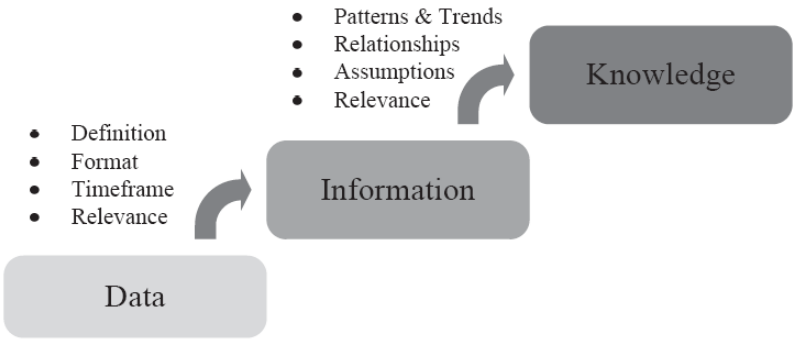

Fig. 1. Data, Information and Knowledge. [8]

Un dato por si solo no es más que una representación simbólica y para que este tenga significado debe tener un contexto. A partir del dato como unidad atómica, se construye la Información y el Conocimiento. Esta relación jerárquica hace que la Calidad de los Datos se reflejen en niveles superiores. La información es el producto de un procesamiento de los datos en un contexto determinado, que lo convierten en un instrumento de toma de decisiones.

Se debe aclarar, que para esta investigación se hará uso indistinto del término datos e información. Es usual que el término Datos sea asociado con la manipulación de bajo nivel de estos e Informaci´on cuando se enfoca hacia la gestión de sistemas.

El Conocimiento es la Información puesta en perspectiva, con el reconocimiento, entendimiento y familiaridad de una situación con su complejidad [8].

El Conocimiento también implica el plantear teorías y premisas acerca de las causas raíz del evento. El conocimiento se construye en base a la significancia de la información y es también un recurso y activo valioso para las organizaciones. Las disciplinas que se encargan de esto son:

- **Data Management**, que se encarga de la gestión de los Datos.
- **Knowledge Management**, que se encarga de gestión del conocimiento.

Ambas disciplinas gozan de aceptación en la industria y en la academia y se encargan de sentar y formalizar las bases de la práctica. En esta investigación se tomará solo en cuenta la Calidad de los Datos, que es la base para la Información y Conocimiento de Calidad. La Calidad del Conocimiento escapa al alcance de este estudio.

2.3 Historia de la Calidad de Datos

Es probable que la DQ haya estado presente desde que se inicio el procesamiento automatizado de datos, mediante las primeras computadoras, talvez no con el mismo término acuñado, pero con iniciativas para poder lograr un nivel de certidumbre y control de error.

Segun Batini [1], la DQ empezó curiosamente con los Estadísticos en los años sesenta a raíz de las investigaciones de la duplicidad de datos en los que se conocían en esos años como como Databanks [9]. Posteriormente, en los años ochenta los investigadores de la Alta Administración empezaron a manejar el concepto de “Sistemas de Manufactura de Datos”, talvez motivados por la ola de éxito en la implementación de Sistemas de Calidad en las empresas.

Es en los años noventa, cuando la computación personal y las interconectividad está en su auge y los grandes sistemas mainframes son desplazados, cuando los investigadores en Ciencias de la Computación empiezan a abordar el tema de la DQ. Es en esta etapa donde se se proponen las bases para la definición, medición y mejora del nivel de la DQ, en bases de datos relacionales, en sistemas legacy y en los nacientes Datawarehouses.

La historia de la DQ va de la mano con la historia de la computación y es una disciplina eminentemente multidisciplinaria. Sirven a la DQ: la Minería de Datos, la representación del Conocimiento, los Sistemas de Información Gerencial, la Estadística, la Integración de Datos entre otras.

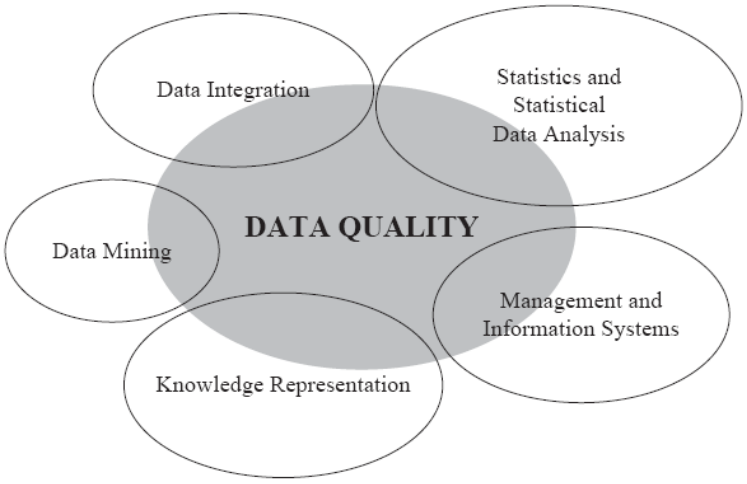

Fig. 2. Areas de Investigación relacionadas con DQ. [1]

2.4 Definición de DQ

> Data are of high quality if they are fit for their intended use in operations, decision-making, and planning. (Juran, 1964) [10].

Como sucede con el concepto de Calidad, cuando esta se aplica a Datos existirán muchas definiciones, algunas convergentes otras divergentes. No obstante, se pueden encontrar muchos puntos comunes entre los diferentes autores que se revisaron en la literatura.

La definición de Juran, se podría contar entre las primeras que trataron de conceptualizar la DQ. En su definición, Juran que venia del área de la Calidad, plasmó un concepto clave que perdura hasta nuestro días es la de "Fit for intented use" [10], es decir que "sea adecuado para el uso deseado". English, por otro lado plantea que el concepto de "fitness for use" de los Datos es limitado, ya que los datos podrían ser útiles no solo para el uso "deseado", sino también para otros usos potenciales [4].

Mcgilvray amplia este concepto y plantea la DQ como un nivel o grado de confianza que los datos ofrecen, para el o los usos en los que se quieran aprovechar [12]. De manera simple "los datos correctos, en el lugar correcto, en el tiempo correcto, para la persona correcta y que sean usados para el mejor desempeño del negocio".

Un punto importante para recalcar es que la Calidad de Datos esta relacionada con su uso y aprovechamiento, por tanto no se puede excluir al consumidor de datos, ya que este forma parte de la evaluación y mejora [15]. A esto se podría añadir que el consumidor es parte de la evaluación y el sistema del cual forma parte, es decir la organización.

Otro enfoque es el de asumir que los datos de una base de Datos no tienen valor ni calidad actual [2]. El potencial de los Datos puede ser solo evidenciado, cuando

alguien más los usa para algo de utilidad [15]. Este punto es importante, porque nos hace reflexionar acerca de que la DQ no es algo estático y que es menester para los custodios hacer posible que los datos sean relevantes para una amplia audiencia y que su potencial de uso sea incrementado de manera sistemática. Esto solo puede ser logrado a través de un proceso formal de Evaluación y Mejora de la DQ.

Redman agrega a lo ya expuesto, que para cumplir con la cualidad de fitness for use los datos deben ser accesibles, exactos, oportunos, completos y consistentes con otras fuentes, relevantes, integrales, deben proveer un nivel de detalle, para poder ser fáciles de leer e interpretar [14]. En esta definición, ya se hace alusión a la multidimensionalidad de la DQ, dado que no existe un ´unico criterio para poder definirla. Se hace hincapi´e, en que la DQ es más que solo la exactitud[5] (que es solo una dimensión) y no solo es un problema tecnológico [6]. Su dominio va más allá de esto.

Con este preámbulo podemos concluir que la DQ:

- Es multidimensional.
- Es el nivel de confianza que se atribuye a un conjunto de datos para sus consumidores.
- Es el grado de satisfacer o exceder los requerimientos de un consumidor de Datos
- Es un concepto holítisco que no solo engloba a los datos, sino a los consumidores de estos.
- No es estática y puede cambiar a través del tiempo y uso de los datos.
- No solo es exactitud, sino se deben considerar las dimensiones adecuadas a cada problemática particular.
- No solo es un problema tecnológico o de procesos.
- Se evidencia cuando alguien accede a un dato que tiene un uso potencial, para un fin útil.

2.5 Tipos de Datos

Los datos son una abstracción de objetos del mundo real, que pueden ser representados de manera digital para que puedan ser almacenados, recuperados, procesados, aprovechados y eventualmente desechados. Esta representación puede ser de virtualmente cualquier objeto que pueda ser caracterizado por al menos un dato y su esquema. La representación del objeto es el dato y el esquema es la estructura de esta. Esta se puede considerar un tipo de clasificación. En la figura 3, se muestra la recopilación que hace Batini de la clasificación de los datos.

[5] En la literatura se encuentra que la exactitud fue una de las primeras dimensiones o criterios a ser considerados para definir la DQ.

[6] La Tecnología es una herramienta necesaria para abordar las problemáticas de DQ, pero no es una panacea

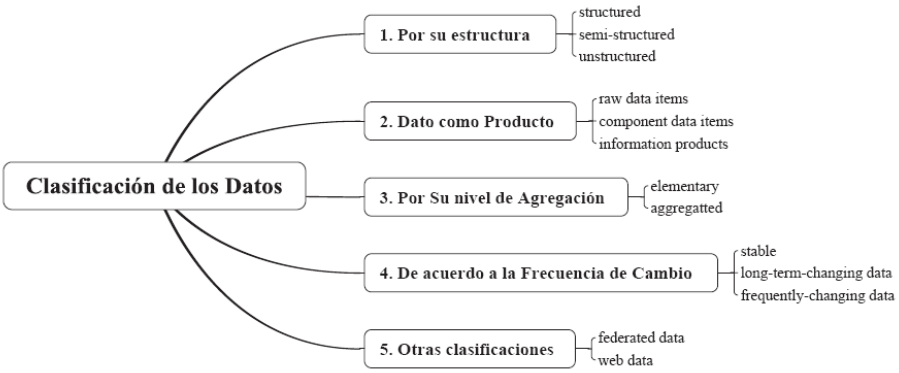

Fig. 3. Clasificación de los Datos [1].

Como se puede observar existen diversas formas de clasificar a los datos. Esto ayuda a determinar el tipo de datos con el que se piensa realizar la actividad de DQ. Para esta investigación, se tomarán en cuenta los siguientes tipos:

- Por su Estructura:
  - Estructurados, que comprenden las bases de datos relacionales, ya sean OLTP[7] o Datawarehouse.
- Dato como Producto:
  - Raw Data Items, unidades atómicas de información.
  - Component data items, datos intermedios para la contrucción un in IP[8].
  - Information products, productos de información para el consumo final (toma de decisiones).
- Por su nivel de Agregación:
  - Elementary, datos con bajo o nulo nivel de agregación.
  - Aggregatted, datos con agregación media y alta.
- Por la frecuencia de cambio en el tiempo:
  - stable, datos con tasa nula de cambios
  - long-term-changing data, datos con tasa media de cambios.
  - frequently-changing data, datos con tasa alta de cambios.

2.6 Dimensiones de la Calidad de Datos

Como ya se mencionó, la DQ es un concepto que tiene múltiples facetas. Esta facetas son conocidas como dimensiones y son la base para cualquier actividad práctica o de investigación en el área de DQ. Cada faceta caracteriza un criterio de calidad de los datos de forma cualitativa. Cada dimensión puede ser asociada o no, a una o más métricas que le dan el car´acter cuantitativo. Las dimensiones pueden ser en

[7] Online Transaction Processing, se refiere a las Bases de Datos Operacionales o de Transacción.
[8] Information Product.

referencia a los datos o al esquema. Sin embargo, se ha dado mayor importancia a evaluar la DQ de los datos, soslayando el esquema [1]. El DQ del esquema tiene una gran importancia, porque muchos de los problemas de DQ pueden ser evitados en la etapa de diseño de los datos.

A continuación, se dar´an las definiciones de las dimensiones principales y más aceptadas.

- **Exactitud**. La exactitud es una de las primeras dimensiones que apareció en la literatura de DQ [9]. La explicación de esto radica en que en general se asocia la exactitud a datos sintácticamente correctos. Y la definición de la exactitud es la distancia de un valor υ a un valor υ_, es decir el número de cambios que se requieren para poder acercarse al valor exacto. un ejemplo es el de escribir Jvier en lugar de Javier. la Distancia de un dato a otro es 1, dado que existe un carácter distinto.

  Batini clasifica a la exactitud en dos tipos [1]:

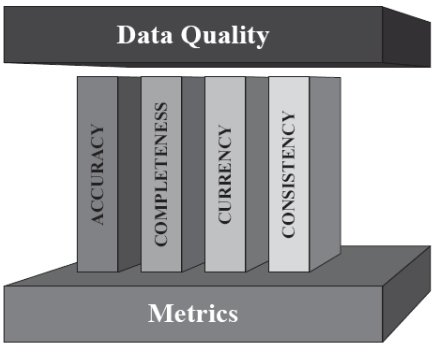

Fig. 4. Dimensiones de la Calidad de Datos. (Pinto, 2011)

- **Sintáctica**. se define como la cercanía de un valor υ a los elementos correspondientes a un dominio de definiciones D. Se la expresa en forma de distancia y tiende a coincidir con la semántica, dado que al corregir un error sintáctico existe una alta probabilidad de que se corrija el semántico también.
- **Semántica**. coincide con el concepto de Correctitud y se refiere a que el contenido de los datos es correcto. Es más complejo de calcular que la sintáctica, dado que se debe realizar, primero una identificación del objeto y segundo una proceso de decisión para seleccionar el correcto. La comparación se puede realizar con otras bases de datos. Su calificación puede ser binaria:
- <si, no>o <correcta, incorrecta>. La duplicación es un problema típico de exactitud, que puede ser controlado mediante las restricciones de integridad referencial en el esquema.

Completud. Wang define la Completud como “el grado en el que los datos tienen la suficiente amplitud, profundidad y alcance para la tarea que tienen que cumplir” [16]. Existen tipos de Completud, de acuerdo al alcance:

- De Esquema
- De Columna
- De Población

En los datos relacionales la completud tiene que ver con el uso o no de campos NULL. Este es un detalle que se debe cuidar desde la etapa del diseño, porque puede haber incongruencias por el uso indiscriminado de campos NULL. Batini menciona dos enfoques para el diseño lógico [1]:

- ***Closed World Assumption***. donde se modela con valores NULL.
- ***Open World Assumption.*** donde se modela sin valores NULL.

Algunas de las dimensiones como el caso de Completud, tienen que ver más con las etapas de diseño conceptual y l´ogico. Prevenir siempre será la mejor estrategia para tener Datos de Alta Calidad y Completos.

Consistencia. Tiene que ver con la violación de reglas semánticas definidas contra un conjunto de datos, que pueden ser tuplas de tablas de bases de datos relacionales, o registros en un archivo[1]. La forma de contener los errores por inconsistencia, es definir Restricciones de Integridad, que son propiedades que deben ser satisfechas por todas las instancias del esquema de la Base de Datos. Estas restricciones puede ser:

- de Intrarelación: Ej. Sexo es M o F.
- de Interrelación: Ej. CP IFE debe ser igual a CP SAT.

Dimensiones relacionadas con el Tiempo El tiempo es una dimensión universal, así como el espacio. Los eventos suceden en el tiempo y este juega un papel importante en la DQ. Los datos registran los eventos, los abstraen de la realidad y los almacenan. El tiempo en que ocurren también es registrado. Calidad es que los datos estén disponibles de forma oportuna y que la latencia entre el momento que ocurre el evento y el momento en que se registra y presenta, debe ser el requerido por el consumidor de datos. No cumplir con la restricción de tiempo hará que la DQ se degrade, aun y cuando las demás dimensiones tenga una calificación aceptable. A continuación se dan las características más importantes, que están en función del tiempo, según Batini [1].

- **Vigencia**, la prontitud en que los datos son actualizados.
- **Volatilidad**, la frecuencia con la que cambian los datos.
- **Oportunidad**, si los datos están disponibles en el tiempo requerido.

Otras dimensiones. Las dimensiones descritas anteriormente son las principales. No obstante, en la literatura se encuentran otras que son para propósitos específicos. Las dimensiones pueden ser usadas de acuerdo al dominio y deben ser seleccionadas cuidadosamente, de manera que el modelo de DQ sea consistente y objetivo. Algunos dominios tienen su dimensiones ad hoc, como es el caso de los Sistemas de Archivo Documental, Sistemas Geográficos y Geoespaciales, Financieros, Estadísticos y otros. Otras dimensiones puede ser:

- Interpretabilidad
- Accesibilidad

- Credibilidad
- Objetividad
- Reputación
- Veracidad

2.7 Evaluación de la Calidad de Datos

A lo largo de los a˜nos la Calidad ha evolucionado y se han generado diversos modelos para medirla, evaluarla y mejorarla. La DQ también ha ido evolucionando y robusteciendo con metodologías, técnicas y herramientas que permitan evaluar y mejorar la DQ, de una manera más eficiente y eficaz.

El objetivo primario de una metodología para la evaluación de la DQ, es brindar un diagnóstico preciso de la situación de los Sistemas de Información, en relación a la DQ [1]. Los salidas de la evaluación de DQ se traducen en:

- **Mediciones**, obtenidas de las fuentes y flujos de datos
- **Costos**, que se atribuyan a la baja DQ.
- **An´alisis Comparativo**, del nivel de DQ contra el criterio de Expertos, *benchmarks* o mejores prácticas.

El proceso de evaluación tiene tres procesos principales:

1. Selección, Clasificación y Medición de Dimensiones y Métricas relevantes (Fase Objetiva).
2. Evaluación Subjetiva, realizada por expertos.
3. Análisis Comparativo entre la Evaluación Objetiva y Subjetiva.

Esta Investigación se basará en el modelo de Evaluación de DQ presentado por Batini [1]. Las fases de este modelo están ilustradas en la figura 5.

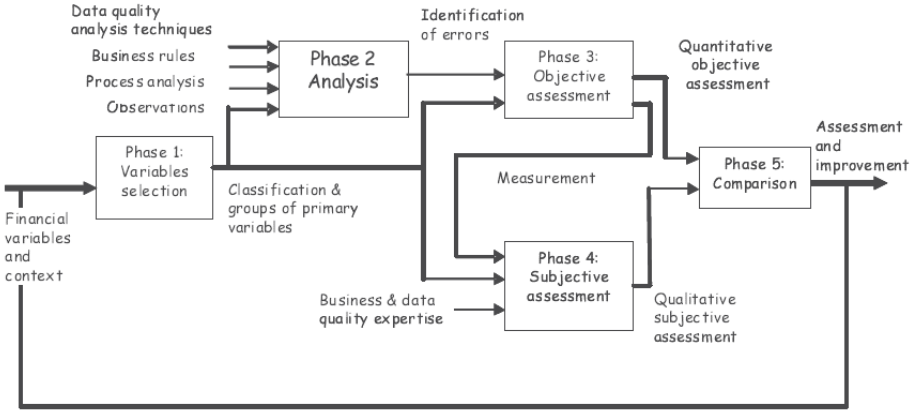

Fig. 5. Fases principales de la Metodología de Evaluación de DQ, presentada en [1]

3 Balanced Scorecard

El *Balanced Scorecard* (BSC) es un modelo ampliamente aceptado en la industria y en la academia. A través de los años ha ido madurando y se ha convertido en un estándar de facto para la planeación estratégica de todo tipo de empresas.

El BSC provee un framework para seleccionar múltiples medidas de desempeño relacionadas con los objetivos estratégicos. Este proporciona medidas tradicionales financieras, integrando además medidas "no financieras" en tres perspectivas adicionales (ver figura 6), que son:

- Cliente
- Procesos Internos
- Aprendizaje e Innovación

Kaplan y Norton, los autores de este modelo, tuvieron la intención de que el BSC sirviera como una herramienta de comunicación y alineación estratégica y como una herramienta de monitoreo.

En el nivel más amplio el BSC, se puede usar como un Marco de Referencia para un Sistema de Planeación Estratégica, que enlaza la estrategia de largo plazo con las acciones en el corto plazo. Un principio fundamental del BSC es que alinea a toda la organización hacia su estrategia, que es formulada en el nivel jerárquico de la organización y es bajado hacia los niveles táctico y estratégico, de manera que haya coherencia en los objetivos, medidas y metas de todo la organización.

Al ser una herramienta de Gestión del desempeño estratégico, se genera como un reporte estructurado semiestándar, que hoy en día esta soportado por muchas herramientas disponibles en el mercado. El espíritu del BSC, es que el nivel estratégico y táctico pueda dar seguimiento a las actividades de los niveles operativos y así monitorear las consecuencias de sus acciones. Esto se hace en función de metas pactadas previamente que buscan mejorar los KPI o Indicadores Clave de Desempeño.

Las multiperspectiva del BSC, le da flexibilidad para ser adoptado por otras iniciativas de Monitoreo del desempeño estratégico especifico. Una de las adopciones más exitosas es la de la adopci´on el modelo BSC para la Gestión Estratégica de las Tecnologías de Información y Comunicaciones.

Este modelo será el Marco Conceptual para el Desarrollo de *Scorecard* de DQ y lo extenderá en el sentido de que la DQ está presente en todas las perspectivas del BSC.

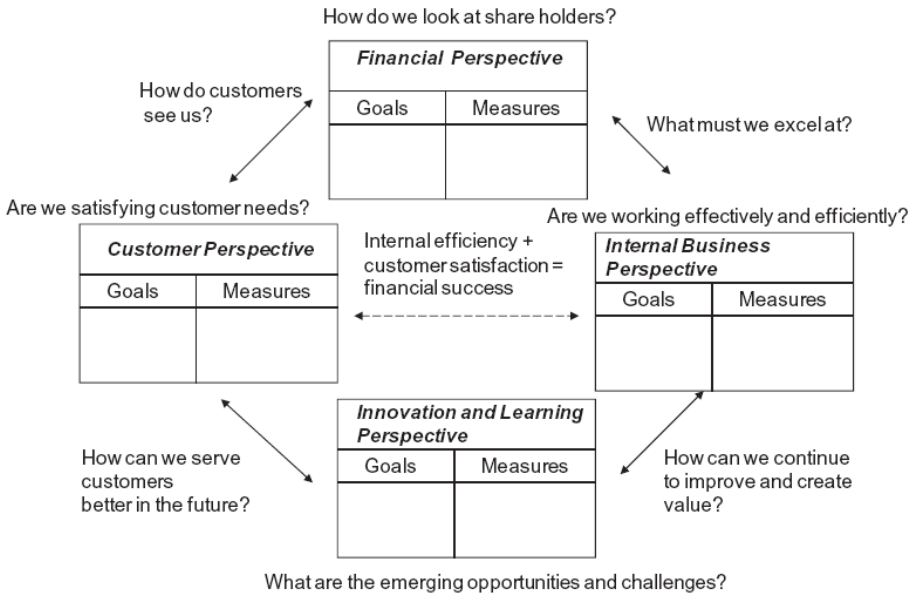

Fig. 6. Modelo Gen érico de Kaplan y Norton (1992).

4 KPI

Pese a que el concepto de KPI es muy conocido, se tiene a veces una concepción errónea de su propósito. El objetivo de un KPI, como lo dice su nombre, es el de proporcionar una medición de un indicador crítico para el desempeño de la organización. La confusión radica en que existen distintos tipos de indicadores. Parmenter nombra los siguientes:

- KRI (Key Result indicators), que expresa resultados clave en alguna de las perspectivas del Scorecard.
- RI (Result Indicator), resultados de lo que se ha logrado.
- PI (Performace Indicator), expresa lo que se tiene que hacer.
- KPI (Key Performance Indicator), expresa lo que se tiene que hacer, para lograr un incremento dram ático en el desempeño.

Las organizaciones ha hecho una mezcla de estos indicadores y han utilizado el término KPI de manera indistinta.

Los KPI representan un conjunto de medidas, que están enfocadas a aquellos aspectos del desempeño organizacional que son más críticos. Estas medidas pueden ser obtenidas de la situación actual o proyectadas a un escenario en el futuro. Los KPI no son variables desconocidas para el negocio, pero se requiere un ejercicio formal y met´odico para estructurarlos y obtener sus beneficios.

Para efectos de esta investigación, diferenciaremos a los KPI de otros indicadores basados en la características que se mencionan en [13]:

- No son financieras y se debe evitar expresarlas en términos monetarios.
- Se miden frecuentemente (24/7, diaria o semanal).
- Son indicadores que causan la participación de un Mando Gerencial o Directivo.
- Indican claramente la acción que se requiere, para mitigar o eliminar el efecto no deseado.
- Otorgan responsabilidad por un equipo.
- Su impacto es significativo.
- Instan a tomar una acción apropiada (correctiva o preventiva)

Los KPI están ligados a una perspectiva de un *Scorecard* (BSC) y generalmente se instrumentan diez a veinte indicadores. Parmenter [13] recomienda que existe una proporción de '10/80/10', es decir diez para KRI, ochenta para RI y PI y diez para KPI.

5 La Investigación

La presente investigación se basa en dos cuerpos de conocimiento:

- Evaluación de la DQ.
- Scorecards y KPI.

Ambos cuerpos de conocimiento, engloban teorías, marcos de referencia, casos de estudio, metodologías, técnicas y herramientas en los que se basará la construcción del modelo, resultado de esta investigación.

Como se puede observar en la figura 7, la Evaluación de Datos tiene dos elementos a considerar: Las dimensiones y las Métricas de Calidad de DQ. Estos son los elementos fundamentales y multidimensionales sobre los que se construye el concepto de DQ. De forma an´aloga las perspectivas y los KPI, son los elementos fundamentales para la construcción de un Scorecard.

De la intersección de estos elementos se generarán dos conceptos nuevos:

- Modelo de un *Scorecard* para la evaluación de DQ.
- Definición de KPI para la Evaluación de DQ.

Ambos productos forman parte del constructo final, que es el Marco de Referencia para la Evaluación de la DQ. De estos también se derivan productos prácticos, a saber:

- Arquitectura de un *Scorecard* para la evaluación de la DQ.
- Herramientas, Metodologías y técnicas para la instrumentación de KPI.

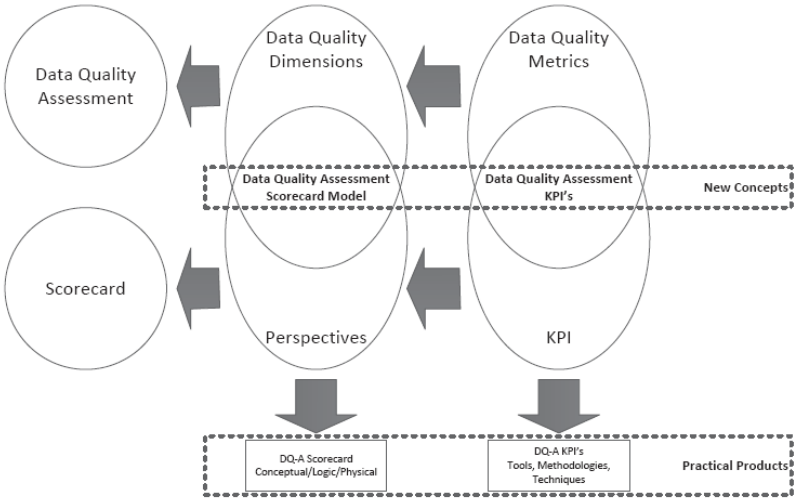

Fig. 7. Esquema de los productos de la Investigación.[Pinto, 2011]

5.1 Definición del Problema

Como se ha expuesto anteriormente, la baja Calidad de Datos es la causa de problemas críticos para empresas, organizaciones e individuos. Los datos son la materia prima para otros sistemas como los operativos y los estratégicos. Los datos se transforman en información y en conocimiento, que es usado como instrumento para la toma de decisiones. Por lo tanto, conocer el Nivel de Calidad de los Datos es prioritario y es el primer paso hacia la mejora continua del activo más importante: Los Datos.

A lo largo de los años la Calidad ha evolucionado y se han generado diversos modelos para medirla, evaluarla y mejorarla. La DQ también ha ido evolucionando y robusteciendo con metodologías, técnicas y herramientas que permitan evaluar y mejorar la DQ, de una manera más eficiente y eficaz.

La evaluación de DQ se ha realizado hasta ahora en aspectos inherentes a los datos, es decir solo considerando los criterios de calificación del conjunto y sus características sintácticas y semánticas intrínsecas, pero sin considerar el impacto de estos en el contexto en que son utilizados.

Para ejemplificar esta limitación, en la tabla 1 se muestra una Medición de una Dimensión de DQ, realizada a un Conjunto de Datos de "Pacientes".

Los valores numéricos expresan el porcentaje de cumplimiento respecto de la dimensión, de manera cuantitativa.

Desde el punto de vista de la Dimensión de Completud, el valor 100%significa que existe un valor para cada campo de la variable que satisface el criterio de DQ. Sin embargo esta información no expresa nada relacionado con el contexto, ¿De qué le serviría a la organización saber que solo tienen registrado el 60% de

Tabla 1. Ejemplo de Evaluación de DQ a un Dataset de Pacientes.

DataField	Completeness
FirstName	**98** %
LastName	**97** %
Address	**60** %
Birthday	**65** %
Age	**30** %
CURP	**35** %

las direcciones de sus pacientes? ¿Es bueno o malo? ¿Significa esto algún tipo de perdida en términos financieros o no financieros?.

Es ahí donde se encuentra la brecha, entre la medición de las características de los datos per sé y la interpretación en función de otro contexto, en este caso el del beneficio para el desempeño de la empresa. La investigación que se presenta en este artículo, pretende cubrir esta brecha y para esto se usará un enfoque novedoso basado en un Scorecard[9] y KPI.

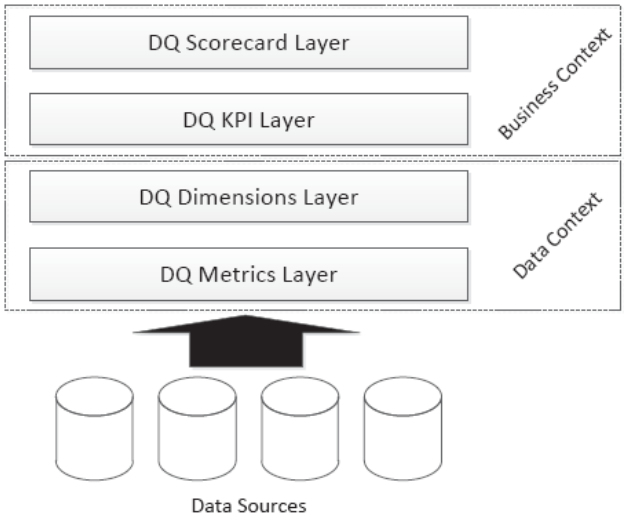

Fig. 8. Capas de Abstracción. (Pinto, 2011)

El Marco de Referencia propuesto, pretende dar otra capa de abstracción a la que ya existe, usando para esto KPI's específicos para la Evaluación de la Calidad de Datos. Los KPI por definición, son indicadores de variables clave para el negocio, que pueden ser actuales o futuros. La latencia de los KPI tiende al tiempo real y esta expresada en términos de los elementos que son críticos al negocio. Los KPI le dan un contexto de evaluación a la DQ, más cercano a las reglas del negocio y a la mejora de su desempeño. Se espera que este enfoque mejore la eficacia y la eficiencia en el

[9] Scorecard: Tablero de Control o Mando.

proceso de evaluación de la DQ, en contraste de la evaluación de la DQ basada solo en las dimensiones.

5.2 Objetivo

Basados en la problemática descrita proponemos los siguientes objetivos:

- Adaptar el *Balanced Scorecard* considerando la perspectiva de DQ.
- Construir un Marco de Referencia para la Evaluación de DQ, basado en un *Scorecard*.
- Instrumentar el Scorecard de DQ propuesto.
- Diseñar un Caso de Estudio para validar el Marco de Referencia Propuesto

5.3 Pregunta de Investigación

La pregunta de investigación es: *¿Existe un modelo que evalúe y haga un monitoreo continuo de la Calidad de Datos, desde el contexto de su contribución de valor al desempeño empresarial?*

5.4 Metodología

Como se observa en la figura 9, la metodología a seguir tiene tres fases:

1. **Fase Exploratoria**, que consiste en la revisión de la literatura y estado del arte de la DQ, con énfasis en la Evaluación. También se revisan los modelos de *Scorecard* y KPI.
2. **Fase de Construcción del Modelo**, en esta fase se generan los elementos del Marco de Referencia, se realizan prueba de concepto y se integra el modelo a validar.
3. **Fase de Evaluación del Modelo**, una vez integrado el modelo se realizará la validación mediante un caso de estudio.

5.5 Resultados Preliminares

Se inició realizando una Prueba de Concepto Inicial de Evaluación DQ, utilizando para este fin una herramienta de DQ con Arquitectura Abierta. Las pruebas se realizaron contra una base de datos relacional de prueba. Una vez realizado el *deployment* y la puesta a punto de la Herramienta DQ se realizó el cargado de *metadata* y se inicio con la ejecución de las tareas de análisis básicas.

Para esto se implementó la arquitectura propuesta, de manera parcial. La arquitectura que se usará para la instrumentación del Modelo, se basa en la implementada en una arquitectura abierta para la Gestión de la Información y se puede ver en la figura 10. El principal análisis con el que se inicio fue el Análisis de Columnas y el establecimiento de Reglas, para poder generar validaciones mas

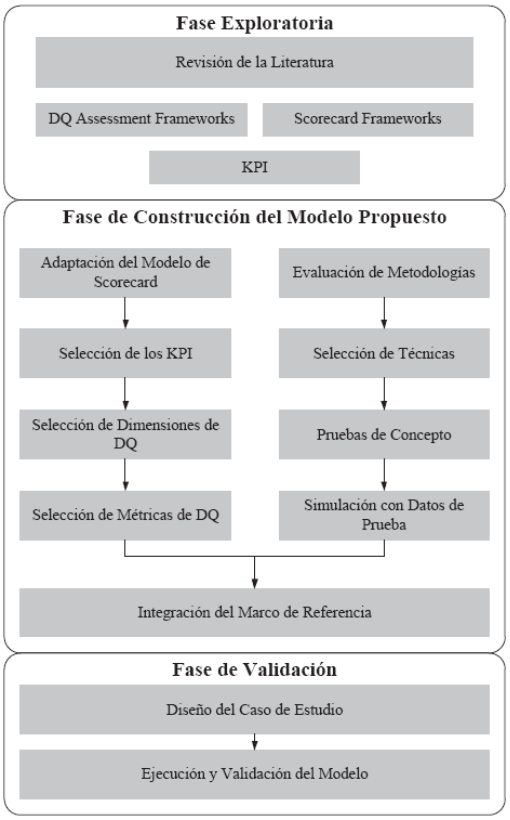

Fig. 9. Metodología de la Investigación.

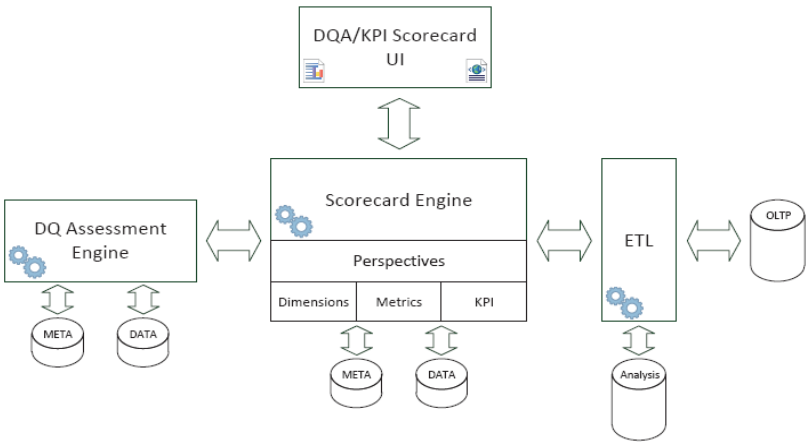

Fig. 10. Arquitectura Propuesta.

complejas. Se probó también la generación de línea base, para poder hacer un monitoreo continuo de la base de datos de prueba.

La razón por la que se selecciono la Herramienta de DQ de Arquitectura abierta, es porque se generará otra capa de Servicio SaaS[10] para el Scorecard. La Herramienta de DQ permite la interfaz a través de SOA y esto permitirá realizar una prueba de concepto, creando un WebService que solicite los servicios de DQ de la herramienta y los envíe a la siguiente capa: la del *Scorecard*.

Conclusiones

La DQ es un tema actual y de alta prioridad, para individuos y organizaciones. Los datos de alta calidad, contribuyen a que las empresas logren sus objetivos, ayudándoles a producir mejores productos y servicios, mejorando la relación con sus clientes y colaboradores. Es por eso que se debe procurar el establecimiento de procesos formales de Evaluación y Mejora de la DQ, para garantizar que los datos sean exactos, estén completos, sean consistentes y lleguen con oportunidad a quien los requiera.

La propuesta de un Modelo de Evaluación de la DQ basado en un *Scorecard*, busca incorporar este elemento clave a la estrategia organizacional, de tal forma que se pueda monitorear el impacto, los beneficios y permita conocer las causas raíz de una baja DQ.

La Evaluación convencional de la DQ, solo se orienta a evaluar los datos solo desde la perspectiva de los mismos datos y no da cuenta del valor que esto representa para el negocio. El llevar la DQ a nivel de evaluación estratégica clave, como lo es el BSC, permitirá difundir la conciencia de la importancia de mantener un nivel alto en la DQ.

Referencias

[1]. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Springer-Verlag New York, Inc. (2006)

[2]. Dalcin, E.C., University of, S.: Data quality concepts and techniques applied to taxonomic databases. University of Southampton, Southampton (2005)

[3]. Eckerson, W.W.: Data quality and the bottom line:achieving business success through a commitment to high quality data. Tech. rep., THE DATA WAREHOUSING INSTITUTE (2002)

[4]. English, L.P.: Improving data warehouse and business information quality : methods for reducing costs and increasing profits. Wiley, New York (1999)

[5]. English, L.P.: Information quality applied : best practices for improving business information, processes, and systems. Wiley, Indianapolis, Ind. (2009)

[6]. Fisher, C.W., Kingma, B.R.: Criticality of data quality as exemplified in two disasters. INFORMATION AND MANAGEMENT 39(2), 109–116 (2001)

[7]. Fisher, T.: The data asset : how smart companies govern their data for business success. John Wiley & Sons, Hoboken, N.J. (2009)

[8]. International, D.: The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK. Technics Publications, LLC (2009), 1593444

[10] Software as a Service.

[9]. Ivanov, K., Stockholm. Tekniska, h., Stockholm. Universitetet. Institutionen f¨or informationsbehandling, A.D.B.: Quality control of information: on the concept of accuracy of information in data-banks and in management information systems. THE ROYAL INSTITUTE OF TECHNOLOGY AND THE UNIVERSITY OF STOCKHOLM, Stockholm, SWEDEN (1972)
[10]. Juran, J.M.: Managerial breakthrough; a new concept of the manager's job. McGraw-Hill, New York (1964)
[11]. Loshin, D.: Enterprise knowledge management: the data quality approach. Morgan Kaufmann Publishers Inc. (2001)
[12]. McGilvray, D.: Executing data quality projects : ten steps to quality data and trusted information. Morgan Kaufmann, Burlington, MA (2008)
[13]. Parmenter, D.: Key performance indicators : developing, implementing, and using winning KPIs. John Wiley & Sons, Hoboken, N.J. (2010)
[14]. Redman, T.: Data quality: the field guide. Digital Press (2000)
[15]. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Commun. ACM 40(5), 103–110 (1997)
[16]. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. 12(4), 5–33 (1996)