UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN
CARRERA DE LICENCIATURA EN CIENCIA DE LA COMPUTACIÓN

TRABAJO DE DIPLOMA

Análisis de la calidad de datos en fuentes de la suite ABCD

Diplomante: Yordan Andreu Alvarez

Tutores: MSc. Lisandra Díaz De la Paz
Lic. Juan Luis García Mendoza

Santa Clara
2014-2015

El que suscribe, Yordan Andreu Alvarez, hago constar que el trabajo titulado ***"Análisis de la calidad de datos en fuentes de la suite ABCD"*** fue realizado en la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de los estudios de la especialidad de Licenciatura en Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____

Firma del Autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma del Tutor Firma del Jefe del Laboratorio

“Lo sabio es la meta del alma humana, y a medida que se avanza en sus conocimientos, va alejando a su vez el horizonte de lo desconocido”

Heráclito

DEDICATORIA

A mis padres, hermana y familia: por su apoyo, paciencia y comprensión toda mi vida para la realización de mis metas.

AGRADECIMIENTOS

- A mis padres, por soportarme abnegadamente en estos cinco años de carrera.
- A mi pequeña hermana Daniela, por ser la mejor hermana del mundo, el cariño que me tiene y la admiración que me guarda.
- A todo el "clan" de primas y primos, Sandra (mi jimagua), Claudia, Yasel y Ernesto, por su tolerancia.
- A mis abuelos y bisabuela, por haberme cuidado incondicionalmente como si fueran mis padres cando mi mamá no estuvo presente.
- A mis tíos, Yaquelín, Abel, Herminia, Argelia, Andrés, Jaime y Yuni, por su cuidado y darme todos los gustos.
- A mis tutores Lisandra y Juan Luis, por las horas de dedicación, por su ayuda, por sus ingeniosas ideas, por brindar sus conocimientos y tener una solución ante cualquier dificultad.
- A Lill Dianet y Margarita, por su apoyo y ayuda incondicional en todo momento.
- A mis amigos de la universidad, Guillermo, Osmani, Pavel, Pedro, Omar, Pablo y Mario, por compartir conmigo momentos importantes, ¡al fin nos vamos a graduar!
- A todos mis compañeros del aula, con los cuales compartí buenos momentos.
- A todos mis profesores, los cuales me han brindado su experiencia, permitiéndome ganar en conocimientos.
- A la profesora Gladys Casas Cardoso, por haber brindado su ayuda desinteresada en parte de este trabajo.
- Al compañero Amed Abel Leiva Mederos, principal consultante sobre el dominio de la aplicación, y a todas las personas encuestadas, que de forma voluntaria cooperaron con la recolección de los datos necesarios.
- A todas aquellas personas que han sido testigos de mi formación y han contribuido a lograr esta meta.

A todos, mis más sinceros agradecimientos

RESUMEN

Dado que los datos de baja calidad pueden tener graves consecuencias en los resultados finales de los análisis de datos, se está reconociendo cada vez más la importancia de la veracidad como dimensión de calidad de datos y como **"*V*"** de *big data*. La veracidad de los datos es directamente proporcional al valor que posee cada dato, así como las decisiones que se toman a partir de estos. En las 18 bibliotecas de la UCLV se cuenta actualmente con el sistema ABCD para el trabajo con documentos bibliográficos. Debido a un proceso de migración efectuado con motivo de trasladar los documentos bibliográficos desde el sistema Quipusnet (anteriormente usado para el trabajo con estos documentos) hacia este nuevo sistema, se produjeron diferentes errores que degradan la calidad de dicha suite. Esta investigación se centra fundamentalmente en el análisis de calidad de datos en fuentes de la suite ABCD, tomando como método de apoyo la aplicación de las tres primeras fases que componen la metodología total de calidad de datos (TDQM). Además, la aplicación de encuestas a una muestra de especialistas que laboran con este sistema, en aras de detectar dimensiones de calidad de datos, problemas y causas que conllevan a los mismos en este caso de estudio para una futura mejora.

Palabras claves: *big data*, calidad de datos, dimensiones de calidad de datos, problemas de calidad de datos, veracidad.

ABSTRACT

As low quality data can have serious consequences on the final results of the data analysis, is increasingly recognizing the importance of veracity as data quality dimension and "V" big data. The veracity of the data is directly proportional to the value held by each data item and the decisions that are taken from these. In the 18 libraries UCLV it currently has the ABCD system for work with bibliographic documents. Because a migration process made on the occasion of moving documents from the Quipusnet bibliographic system (formerly used to work with these documents) to the new system, there were different errors that degrade the quality of that suite. This research focuses mainly on the analysis of quality data sources ABCD suite, on the method of supporting the implementation of the first three phases that make up the overall data quality methodology (TDQM). In addition, the use of surveys to a sample of specialists working with this system, in order to detect data quality dimensions, problems and causes that lead to the same in this case study for future improvement.

Keywords: *big data*, data quality, data quality dimensions, data quality issues, veracity

TABLA DE CONTENIDOS

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

Los datos, son realidades concretas que se encuentran en el mundo; hechos atómicos que por sí solos carecen de significado (Batini and Scannapieco, 2006), y su calidad, puede influir sustancialmente en tres vectores estratégicos de las organizaciones: interacción con el cliente, configuración de recursos y aprovechamiento del conocimiento (Venkatraman and Henderson, 1998). Por consiguiente, los datos son un recurso importante para la competitividad organizacional (Redman, 1998).

Baja calidad de datos afecta la reputación de la empresa o entidad poseedora de los datos al tomar decisiones basadas en datos erróneos y no fiables, propicia la acumulación de errores, lo cual trae consigo el aumento de los costos y esfuerzos por mejorar la calidad. Por otra parte datos con elevada calidad conllevan al éxito del negocio y a la satisfacción de los clientes.

Debido a la tendencia actual adquirida por la mayoría de las empresas modernas, de almacenar cantidades crecientes de información, la gestión de la calidad de datos se ha convertido en un proceso sumamente importante (Moges *et al.*, 2013). Con el incremento del número de usuarios, el auge de las redes sociales, internet de las cosas (IoT, por sus siglas en inglés), contenidos multimedia, bases de datos relacionales, bases de datos analíticas, bases de datos NoSQL, GPS, archivos de texto, aplicaciones móviles, sensores, entre otras fuentes de datos; se produce en la actualidad un flujo enorme de datos, con una variedad de formatos: estructurados, semi-estructurados, no estructurados o mixtos (Hashem *et al.*, 2014). Lo cual trae consigo la creación de dicha información a un ritmo igual al de su almacenamiento (o grabación) (R.L. Villars, 2011), denominado *big data*, convirtiéndose en una tendencia ampliamente reconocida (Hashem *et al.*, 2014).

Por su actualidad y gran alcance *big data* ha despertado la atención de las empresas, el gobierno, la academia y la industria (Hashem *et al.*, 2014), pero trabajar con ellos haciendo uso de la mayoría de los sistemas de gestión de bases de datos relacionales se hace una tarea difícil, ya que requieren *softwares* diseñados para trabajar paralelamente ejecutándose posteriormente en decenas, cientos o incluso miles de servidores. Lo anterior trae consigo la existencia de desafíos relacionados con los datos para las organizaciones, donde se encuentra el reto de gestionar grandes cantidades de datos (*big data*), los cuales son cada vez mayores; por ejemplo, *Facebook*, sitio web de redes sociales, es el hogar de 40 mil millones de fotos, y *Wal-Mart*, multinacional de grandes almacenes, maneja más de 1 millón de transacciones de los clientes cada hora,

alimentando las bases de datos estimadas en más de 2,5 *petabytes* de información (Demirkan and Delen, 2013).

Una forma de tratar lo anteriormente mencionado incluye el uso de la computación en la nube, concepto conocido también bajo los términos servicios en la nube, informática en la nube, nube de cómputo o nube de conceptos, del inglés *cloud computing*, la cual es una rápida y creciente tecnología que permite ofrecer servicios de computación que incluyen un gran número de computadoras conectadas en tiempo real en una red de comunicación como internet.

En nuestro país, la finalidad de la gestión de recursos de información en bibliotecas consiste en proporcionar mecanismos que permitan a las mismas adquirir, producir y transmitir al menor costo posible, datos e información con actualidad suficiente. Esta modalidad de servicio facilita al usuario su uso y permite satisfacer sus necesidades de información sin requerir su presencia física en la biblioteca (Calero *et al.*, 2011).

Dentro de estos mecanismos se encuentra la suite ABCD (Automatización de Bibliotecas y Centros de Documentación), la cual es una red de fuentes de información dinámica y descentralizada, basada en todo lo que sea posible recuperar y extraer de información y conocimiento, para el apoyo de la toma de decisiones. Es una aplicación *web*, de código abierto, que comprende las principales funciones de una biblioteca: adquisición, catalogación, préstamos y administración de bases de datos. Por orientaciones del Ministerio de Educación Superior (MES), este sistema se adoptó en las 18 bibliotecas de la Universidad Central "Marta Abreu" de Las Villas (UCLV), desde el año 2011 hasta la fecha, para llevar a cabo diferentes tareas de gestión bibliotecaria. Dicha adopción trajo consigo la migración de los datos bibliográficos desde el sistema que llevaba a cabo dicha tarea con anterioridad (*Quipusnet*), hacia las bases de datos que conforman la suite ABCD. La migración no pudo realizarse de forma automatizada, lo que trajo consigo la aparición de diferentes errores e inconsistencias en los datos que degradan la calidad de los procesos que brinda dicho sistema. Todos estos elementos exponen la **situación problemática** de la actual investigación; tomando como **objeto de estudio** el análisis de la calidad de datos en fuentes de la suite de Automatización de Bibliotecas y Centros de Documentación (ABCD).

Debido a la importancia que conlleva el proceso de calidad de datos como desarrollador de productos de información (PI) de calidad, en unión con el auge alcanzado por lo que se denomina actualmente como la era de *big data*, poseedora de nuevos desafíos para el manejo de

la calidad de datos; y a modo de solución al problema de la investigación se establece como **objetivo general**: Aplicar las tres primeras fases de la metodología para la gestión total de la calidad de datos en fuentes de la suite de Automatización de Bibliotecas y Centros de Documentación (ABCD) para identificar dimensiones y problemas de calidad de datos, y las causas que conllevan a los mismos; mediante la caracterización del sistema ABCD y el procesamiento estadístico de encuestas aplicadas a especialistas.

Para dar cumplimiento a este objetivo general se proponen los siguientes **objetivos específicos**:

1. Describir los principales algoritmos y herramientas que se utilizan en cada una de las fases de la metodología para la gestión total de la calidad de datos.
2. Caracterizar *big data* de acuerdo a sus clasificaciones, su relación con la nube informática y las principales herramientas que se utilizan en cada etapa.
3. Analizar la similitud entre las arquitecturas de un almacén de datos y de *big data* con la nube, para determinar donde es más apropiado ubicar la calidad de datos.
4. Caracterizar el sistema ABCD en cuanto a su misión, arquitectura, módulos de trabajo, fuentes de datos y alcance que posee.
5. Procesar estadísticamente las encuestas aplicadas a los especialistas que trabajan con el sistema ABCD para identificar dimensiones y problemas de calidad de datos existentes, y las causas que conllevan a los mismos.

Basándose en dichos objetivos específicos se tomaron en cuenta las siguientes **preguntas de investigación** para el desarrollo de los mismos:

1. ¿Cuáles son los principales algoritmos, técnicas y herramientas actuales que se utilizan en cada una de las fases de la metodología para la gestión total de la calidad de datos?
2. ¿Cuáles son las características de las principales técnicas y tecnologías que se utilizan para capturar, almacenar, distribuir, procesar y analizar los datos de tipo *big data*?
3. ¿Qué similitud existe entre las arquitecturas de un almacén de datos y de *big data* con la nube, para determinar dónde ubicar la calidad de datos?
4. ¿Qué caracteriza al sistema ABCD instalado en las bibliotecas de la UCLV?
5. ¿Cuáles son las dimensiones y problemas de calidad de datos, y las causas que conllevan a los mismos que se detectan al procesar estadísticamente las encuestas aplicadas a especialistas que trabajan con el sistema ABCD?

A raíz de los principales problemas que se han detectado en fuentes de la suite ABCD mediante entrevistas realizadas al administrador de este sistema; y debido a que en la base de datos NoSQL CDS/ISIS (encargada del almacenamiento de datos para el proceso de catalogación) no pueda hacerse uso de ninguna herramienta de perfilado, lo que conlleva a la necesidad de realizar encuestas a especialistas para determinar cuáles son las principales dimensiones, problemas de calidad y las causas que conllevan a los mismos, se encuentra la principal **justificación de la investigación**.

El presente trabajo de diploma se encuentra estructurado en tres capítulos.

En el capítulo uno se realiza una revisión bibliográfica sobre aspectos generales de calidad de los datos. Se exponen las principales definiciones del término calidad de datos desde el punto de vista de diversos autores. Se presenta el ciclo de vida de la calidad de datos a través de la caracterización de las fases que componen la metodología para la gestión total de la calidad de datos (TDQM, por sus siglas en inglés), tales como: definir, medir, analizar y mejorar. Además, se analiza el papel que juegan diferentes aspectos dentro de cada una de ellas como son: las dimensiones de calidad de datos (primera etapa), las métricas de calidad de datos (segunda etapa), el perfilado de datos y la auditoría de datos (tercera etapa), y finalmente la limpieza de datos (cuarta etapa).

En el capítulo dos se exponen elementos fundamentales sobre *big data.* Se presentan definiciones de big data de acuerdo al criterio de diversos autores. Se muestran cinco categorías que clasifican a big data según la fuente de datos, formato del contenido, almacenamiento, organización y procesamiento de los datos. Además, se presentan las principales características de la nube informática y su relación con big data. Se describen las fases por las cuales transita toda solución big data, así como herramientas y tecnologías usadas en cada fase. Finalmente, se propone una arquitectura integradora que vincula los términos mencionados anteriormente con el proceso de calidad de datos.

En el capítulo tres se realiza un análisis de la calidad de datos en *big data*, comenzando por la descripción del caso de estudio correspondiente al trabajo con fuentes de datos del sistema ABCD así como los antecedentes que conllevaron a su utilización en las bibliotecas de la UCLV. Posteriormente se realiza la descripción del instrumento utilizado para el análisis de calidad apoyado en la aplicación de encuestas a profesionales que trabajan directamente con este sistema

en búsqueda de dimensiones, principales problemas, y las causas que conllevan a los mismos. Finalmente se exponen los resultados una vez procesadas las encuestas estadísticamente.

CAPÍTULO 1. CONSIDERACIONES GENERALES SOBRE CALIDAD DE DATOS

El presente capítulo tiene como objetivo realizar un análisis de los principales aspectos relacionados con el proceso de gestión de calidad de datos, comenzando primeramente por las principales definiciones atribuidas al término por diferentes autores. Seguido de esto, a través del uso de la metodología TDQM (*Total Data Quality Management*) para llevar a cabo el proceso de calidad de datos a través de sus principales fases: definir, medir, analizar, mejorar; se realiza una completa caracterización de las mismas y el papel que juegan diferentes aspectos dentro de cada una de ellas como son: las dimensiones de calidad de datos (primera etapa), las métricas de calidad de datos (segunda etapa), el perfilado de datos y la auditoría de datos (tercera etapa), y finalmente la limpieza de datos (cuarta etapa).

1.1 Definición de calidad de datos

Es difícil dar una definición universal de lo que significa calidad (Bobrowski *et al.*, 1999). La noción de calidad se utiliza en el lenguaje cotidiano como un rasgo intangible, algo que puede ser subjetivamente juzgado, pero a menudo no es medido exactamente. Términos como buena o mala calidad son intensamente difusos y utilizados sin la intención de ser una ciencia exacta. Por otra parte, la calidad posee carácter multidimensional, e incluye un objeto de interés; el punto de vista de ese objeto y los atributos de calidad atribuidos al objeto. Esto puede hacer del término **"**calidad**"**, un concepto confuso. Con el fin de discutir la calidad, poder evaluarla y mejorarla, esta tiene que ser definida (Wingkvist *et al.*, 2010).

La calidad aplicada a los datos tiene varias definiciones (Chapman, 2005). Una de las más aceptadas para el término **"**calidad de datos**"** está dada por (Juran and Gryna, 1980), y retomada por (Wang and Strong, 1996, Francalanci and Pernici, 2004, Chapman, 2005, Shankaranarayanan and Cai, 2006, Batini *et al.*, 2009, Sadiq *et al.*, 2011, Salomone *et al.*, 2011, Moges *et al.*, 2013, Yeganeh *et al.*, 2014), quienes coinciden en definirla como **"**datos adecuados para el uso**"** de los consumidores de datos. Lo cual significa que el usuario es el encargado de evaluar el nivel de calidad de un conjunto de datos usados para una determinada tarea realizada en un contexto específico, de acuerdo con un conjunto de criterios, determinando de esta manera si dichos datos pueden ser utilizados para ese propósito (Strong *et al.*, 1997). Por consiguiente, al evaluar la calidad de datos de las fuentes encontradas, la atención debe centrarse en el análisis de cómo la calidad de datos actual afecta negativamente el análisis posterior de los datos. Un

problema crucial es determinar cuáles son los criterios de calidad de datos que deben ser considerados para la selección de los datos en cuestión, en función del uso que se les da. Aún más, cuando se trata con ambas fuentes de datos internas y externas, se debe ser capaz de determinar un intercambio satisfactorio entre la cantidad y la calidad de datos recuperados (Abelló *et al.*, 2013).

La calidad de datos se ha estudiado ampliamente y desde diferentes puntos de vista durante las últimas décadas. El Instituto de Tecnología de Massachusetts (MIT, por sus siglas en inglés) ha sido pionero en investigaciones de calidad de datos desde que el programa de gestión total de calidad de datos (TDQM) fue iniciado en 1992 (Madnick *et al.*, 2009).

1.2 Ciclo de vida de la calidad de datos (Metodología TDQM)

La metodología TDQM fue la primera metodología publicada en la literatura de calidad de datos (Wang, 1998). Esta metodología es el resultado de investigaciones académicas, pero ha sido exhaustivamente usada como guía para iniciativas de reingeniería de datos organizacionales (Batini *et al.*, 2009). TDQM utiliza una versión adaptada del ciclo propuesto por W. E. Deming en el año 1986. Este ciclo se ha convertido en un tema clave en la literatura para el manejo total de la calidad (TQM, por sus siglas en inglés) (Wijnhoven *et al.*, 2007), y se identifica por cuatro pasos principales: planificar, hacer, verificar y actuar (Deming, 1986).

Años más tarde, Wang (1998) realiza una analogía entre la gestión total de la calidad (TQM) de productos físicos manufacturados y la gestión total de la calidad de datos (TDQM); donde argumenta que la fabricación de un producto puede ser vista como un sistema de procesamiento que actúa sobre las materias primas para producir productos físicos. Análogamente, la fabricación de información puede ser vista como un sistema de procesamiento que actúa sobre los datos en bruto para producir productos de información (ver Tabla 1) (Koronios *et al.*, 2005).

	Fabricación de Productos	**Fabricación de Información**
Entrada	Materiales sin pulir	Datos en bruto
Proceso	Línea de ensamblaje	Sistema de Información
Salida	Productos físicos	Productos de Información

Tabla 1 - Fabricación de productos contra fabricación de información. [Fuente: (Wang, 1998)]

Usando esta analogía, Wang propuso un sistema de fabricación de información, adaptando el método de Deming para: definir, medir, analizar y mejorar los productos; para aplicarlos luego al entorno de fabricación. La Figura 1 ilustra el ciclo TDQM para una mejora continua y entrega de productos de información de alta calidad (Koronios *et al.*, 2005).

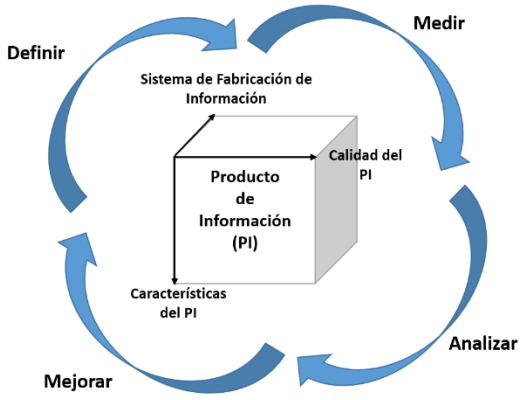


Figura 1 - Ciclo de la metodología TDQM. [Basado en: (Wang, 1998)]

1.2.1 Fase de definición

La fase de definición se identifica por tres pasos (Wijnhoven *et al.*, 2007):

1) Determinar las características de los productos de información

Para describir las características de un producto de información, Wang (1998) utiliza dos niveles; en el nivel más alto se describen las características generales del producto de información, y en un nivel más bajo se describe cada atributo del producto de forma individual. La Figura 2 muestra la diferencia entre las características de los productos de información y los atributos de dichos productos (Wijnhoven *et al.*, 2007).

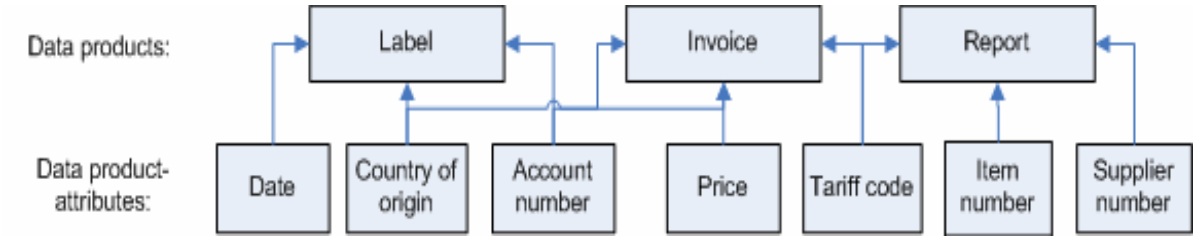


Figura 2 - Datos del producto y datos de los atributos del producto. [Fuente: (Wijnhoven *et al.*, 2007)]

Debido a que no todos los productos de información y atributos pueden ser evaluados en un instante de tiempo, es útil centrarse en el producto de información o atributo más importante a la vez, que pueden ser aquellos cuyos problemas de calidad tienen mayor impacto (Wijnhoven *et al.*, 2007).

2) Determinar los requisitos para los productos de información

La metodología TDQM propone determinar cuáles de las dimensiones de calidad asociadas a los productos de información (o atributos) (Wang and Strong, 1996) son las más importantes, atendiendo a preguntas sobre:

- El nivel de percepción de calidad en una dimensión, y
- El nivel esperado de calidad en una dimensión.

3) Determinar el proceso de fabricación de la información

El proceso de fabricación de la información consiste en flujos de datos que van desde el proveedor hasta los datos del usuario, incluidas ciertas actividades de procesamiento y controles de calidad (Wijnhoven *et al.*, 2007). Según Wang (1998), el conocimiento del proceso de fabricación sirve como base para comprender mejor por qué ciertas dimensiones de calidad son importantes y cómo se dividen las responsabilidades sobre este proceso. Además, los procesos y almacenamientos redundantes pueden detectarse en este paso, los cuales a menudo conducen fácilmente a errores e inconsistencias. Tener un proceso claramente definido, también ayuda en la gestión de calidad de datos con la codificación de los procesos, haciéndolos independientes a la persona y confiables. Una versión simplificada de un proceso de fabricación de información es mostrado en la Figura 3 (Wijnhoven *et al.*, 2007).

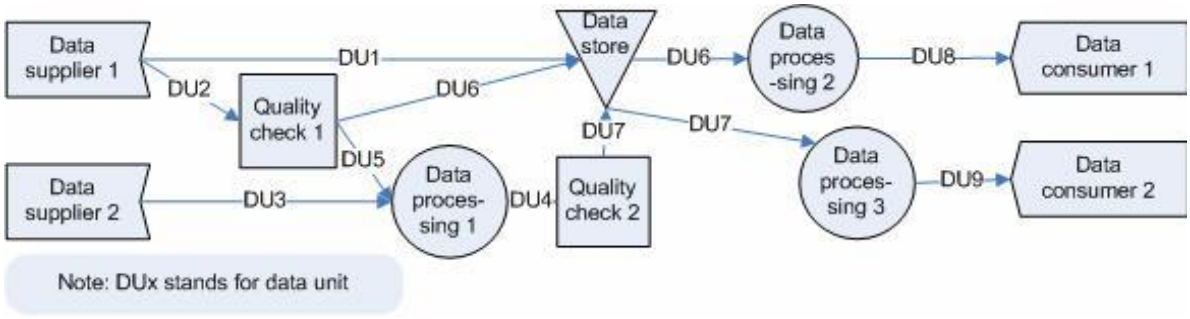


Figura 3 - Ejemplo de un Sistema de Fabricación de Información. [Fuente: (Wijnhoven *et al.*, 2007)]

1.2.2 Fase de medición

La fase de medición determina la calidad de las dimensiones identificadas en la fase de definición. Esta fase se basa en dos pasos fundamentales (Wijnhoven *et al.*, 2007):

1) Seleccionar las métricas adecuadas

Para evaluar una misma dimensión de calidad existen varias métricas, pero cuál escoger depende del contexto de análisis (caso de estudio). En la determinación de una métrica para una dimensión de calidad de datos, se debe tener en cuenta qué hay detrás de las reglas de negocio, normas ISO (*International Organization for Standardization*) y leyes que pudieron haber contribuido a la importancia de las dimensiones identificadas (Wijnhoven *et al.*, 2007). Tres formas de medición de dimensiones de calidad fueron identificadas por (Pipino *et al.*, 2002):

- Relación simple,
- Operación de mínimo (min) o máximo (max), y
- Media ponderada.

Debido a que pueden ser difícil de determinar, las siguientes pautas (ver Tabla 2) pueden ser usadas para la selección de métricas de calidad de las dimensiones elegidas en la fase anterior (Kovac *et al.*, 1997).

R	¿Es una métrica Razonable (***R**easonable*)?
U	¿Es una métrica Comprensible (***U**nderstandable*)?
M	¿Es una métrica Medible (***M**easurable*)?
B	¿Es una métrica Creíble (***B**elievable*)?
A	¿Es una métrica Alcanzable (*Achievable*)?

Tabla 2 - Pautas RUMBA para las métricas. [Fuente: (Kovac *et al.*, 1997)]

2) Medir y presentar los datos

Una vez que se determinan las métricas, la medición puede llevarse a cabo. Si hay por ejemplo 20.000 productos en la base de datos, se necesita un tamaño de muestra de aproximadamente 400 productos para alcanzar una fiabilidad de un 95% con una posible variación de un 5% (Moore and McCabe, 1989). Para mostrar los resultados, varios tipos de gráficos se pueden utilizar, por ejemplo, para identificar qué dimensiones son las causantes de que ocurran problemas en los

resultados, se puede hacer uso de los diagramas de Pareto (Zahedi, 1995). En estos gráficos el eje horizontal muestra los errores en las diferentes dimensiones, y el eje vertical, muestra el porcentaje de los errores en comparación con el número total de errores. Esta es una manera fácil de mostrar que las dimensiones causan la mayoría de problemas (Wijnhoven *et al.*, 2007).

1.2.3 Fase de análisis

El objetivo de esta fase es encontrar las causas fundamentales que conllevan a los problemas de calidad en las diferentes dimensiones identificadas (Wang, 1998). Tres métodos para encontrar dichas causas se han hecho populares en la actualidad: los diagramas de causa y efecto (CED, por sus siglas en inglés), los diagramas de interrelación (ID, por sus siglas en inglés), y los árboles de realidad actual (CRT, por sus siglas en inglés). Según (Doggett, 2005), el método CED es la herramienta más fácil para identificar las causas fundamentales. Este paso debe responder a las interrogantes ¿cuál es el problema?, ¿cuándo pasó?, ¿por qué pasó? , y ¿cómo afecta a los objetivos generales de la empresa? (Wijnhoven *et al.*, 2007).

1.2.4 Fase de mejora

Para la fase de mejora, (Wang, 1998) propone cuatro pasos principales, dentro de los cuales se encuentra lo que se conoce como proceso de limpieza de datos, para lo cual existe un conjunto de técnicas y algoritmos que se emplean para mejorar la calidad de los datos sin afectar su naturaleza:

1) Generación de soluciones:

Para este paso (Wang, 1998) propone el uso de la matriz de análisis de la fabricación de información diseñada por (Ballou *et al.*, 1998) y presentada en la Figura 3, desde la cual se puede obtener la falta de controles de calidad o procesos aparentes, o identificar posibles procesos redundantes y fuentes de datos redundantes.

2) Selección de soluciones:

Al seleccionar las soluciones, es necesario saber qué impacto podrían tener sobre las eficiencias para el manejo de los datos, el costo de diseñar e implementar una solución, los recursos requeridos y la reducción del riesgo. Las prioridades para las soluciones pueden establecerse cuando la importancia de cada criterio de evaluación es conocido (Zahedi, 1995).

3) Desarrollo de un plan de acción:

La solución seleccionada tendrá que ser transmitida en acciones concretas. En este paso se debe asegurar que todas las acciones sean ejecutadas. Un plan de acción del proyecto se puede configurar para realizar un seguimiento de todas las acciones y sus estados (Wijnhoven *et al.*, 2007).

4) Control de progresos:

En este paso, con el chequeo de progresos, se puede conocer si ocurren deficiencias al llevar a cabo las acciones implementadas para el plan de acción desarrollado, donde en caso de fallo se comenzaría a ejecutar nuevamente el ciclo que encierra la metodología o se espera un tiempo en caso que el progreso sea satisfactorio, o sea, la mayoría de los datos estén limpios (Wijnhoven *et al.*, 2007).

El propósito de la metodología TDQM es entregar productos de información de alta calidad a los consumidores de información. Su objetivo es facilitar la aplicación de la política general de la calidad de datos de una organización expresada formalmente por los altos directivos (Wang *et al.*, 1995).

1.3 Dimensiones de calidad de datos

Es posible que se desee tener en cuenta únicamente los atributos específicos con cierta relevancia determinada por usuarios, especialistas del negocio o consumidores de datos, dependiendo del contexto particular que se esté analizando; por ejemplo, la calidad de datos en el contexto de los sistemas de información se relaciona con los beneficios que puede brindar a una organización. Como se ha analizado, la calidad de datos depende de varios aspectos, por lo tanto, con el fin de obtener una medida precisa de calidad de datos, se debe elegir qué atributos conviene considerar y la cantidad de cada uno, contribuyendo a la calidad como un todo (Bobrowski *et al.*, 1999).

1.3.1 Definición y principales técnicas de identificación de dimensiones de calidad de datos

Las identificación de dimensiones de calidad de datos en un contexto específico, juega un papel fundamental en la primera fase de definición de la metodología TDQM (ver epígrafe 1.2).

Los problemas de calidad de datos no pueden abordarse eficazmente sin identificar las dimensiones de calidad de datos pertinentes. Así, el primer objetivo de toda investigación o

estudio práctico sobre calidad de datos, es determinar las características de los datos que son importantes para los consumidores de datos, o que resultan convenientes para los mismos (Wang and Strong, 1996). Mientras que el término **"**adecuación al uso**"** capta la esencia de la calidad de datos, desde hace tiempo se ha reconocido que los datos son mejor descritos o analizados a través de múltiples atributos o dimensiones (Tayi and Ballou, 1998, Shankaranarayanan and Cai, 2006, Maydanchik, 2007). Sin embargo, a pesar de un amplio debate en la literatura de calidad de datos, no existe un único conjunto definido de dimensiones de calidad de datos, porque dependen del contexto en que se estén analizando (Moges *et al.*, 2013).

Diferentes estudios han analizado la calidad de datos desde cierta perspectiva para una tarea específica. Por ejemplo, (Zhu and Gauch, 2000) evaluaron la calidad de datos de una página *Web* en términos de un marco de calidad de datos comprendido en las siguientes seis dimensiones de calidad: actualidad, disponibilidad, relación información-ruido, autoridad, popularidad, y la cohesión; midiendo las dimensiones a través de las propiedades de las páginas *Web*. Del mismo modo, (Chen and Tseng, 2011) evaluaron diferentes dimensiones de calidad de datos con el fin de calcular la calidad de la revisión de productos en línea realizada por los clientes; adoptando varias definiciones de diferentes dimensiones de calidad de datos para el análisis de la calidad de los comentarios de productos en línea. Por ejemplo, definieron la objetividad, como el grado en que un elemento de información es parcial; la cantidad adecuada de los datos, como la medida en que el volumen de información en una revisión es suficiente para la toma de decisiones; y la integridad, como la medida en que la información en una crítica es completa y abarca diversos aspectos de un producto. Además, identificaron objetividad y la cantidad adecuada de la información, como dimensiones de calidad de datos eficaces en la identificación de la calidad para la revisión de un producto; e ineficaces para una completa evaluación a la hora de medir la calidad de una revisión por los clientes u otras partes (Moges *et al.*, 2013).

Por otro lado, hay una serie de estudios que identifican y definen las dimensiones de calidad de datos independientemente de la utilización de los datos, con el fin de facilitar la aplicabilidad general y la comparabilidad de sus dimensiones (Moges *et al.*, 2013).

En este sentido, (Wand and Wang, 1996) basaron su definición de calidad de datos, auxiliados en la vista interna de los sistemas de información (producción de datos y los procesos de diseño del sistema), ya que este punto de vista es independiente del contexto. Este enfoque permite que una serie de definiciones de dimensiones de calidad de datos sean comparables entre aplicaciones. En

primer lugar, identificaron varios criterios de un sistema del mundo real para posteriormente ser representado adecuadamente por un sistema de información. En base a estos criterios, definieron cuatro deficiencias nombradas: representación ambigua, representación incompleta, estados sin sentido, y deficiencias de operación. Sobre la base de estas deficiencias, resumieron diferentes aspectos de calidad de datos en dimensiones completas, sin ambigüedades, con sentido, y corregidas. Además, en el mismo estudio, se clasifican las diferentes dimensiones de calidad de datos de la literatura como vista interna (diseño u operación relacionada) o vista externa (uso o el valor relacionado), por medio del cual ambos puntos de vista se vuelven más refinados como cualquiera de los sistemas o datos relacionados con dimensiones de calidad. Dentro de la visión interna, la exactitud o precisión, el tiempo de vida o la actualidad, la fiabilidad, la integridad y la consistencia, son definidas como datos relacionados, mientras que la fiabilidad es definida también como una dimensión relacionada con el sistema. Por otra parte, en la vista externa, el tiempo de vida, la relevancia, el contenido, la importancia y la cantidad, son definidas como datos relacionados, mientras que el tiempo de vida, la flexibilidad, el formato y la eficiencia son definidas también como dimensiones relacionadas con el sistema (Moges *et al.*, 2013).

Del mismo modo, (Wang and Strong, 1996) analizan diversas dimensiones de calidad de datos desde varias perspectivas del usuario final, pero sin tener en cuenta el uso de los datos, realizando una encuesta a gran escala para determinar y clasificar las dimensiones. Su análisis se inició mediante la recopilación de información de los usuarios, en relación con diversos descriptores de calidad de datos, que resultaron en más de 100 *ítems* y en donde las dimensiones fueron agrupadas en 20 categorías (Moges *et al.*, 2013); las cuales fueron agrupadas posteriormente en las siguientes cuatro categorías de calidad de datos más generales: intrínseca (grado en el que los valores de los datos están en conformidad con los valores reales o verdaderos), contextual (la medida en que los datos del usuario son aplicables a una tarea), representacional (grado en que los datos se presentan de una manera comprensible), y la accesibilidad (grado en que los datos están disponibles o accesibles) (Wang and Strong, 1996).

Para identificar las dimensiones de calidad de datos más a menudo recurrentes y sus definiciones, en el presente estudio se ha adoptado el marco de calidad de datos propuesto por (Wang and Strong, 1996). Este marco se reconoce por intentar establecer un equilibrio entre la consistencia teórica y la viabilidad. Además, se ha encontrado que el marco es aplicable a diversos dominios (Eppler and Wittig, 2000). La estructura del marco es jerárquica, y organiza

aspectos de calidad de datos en quince dimensiones para comprender las cuatro categorías principales de calidad de datos (Moges *et al.*, 2013). La Figura 4 proporciona una visión general de las dimensiones de calidad de datos consideradas en el estudio de (Wang and Strong, 1996).

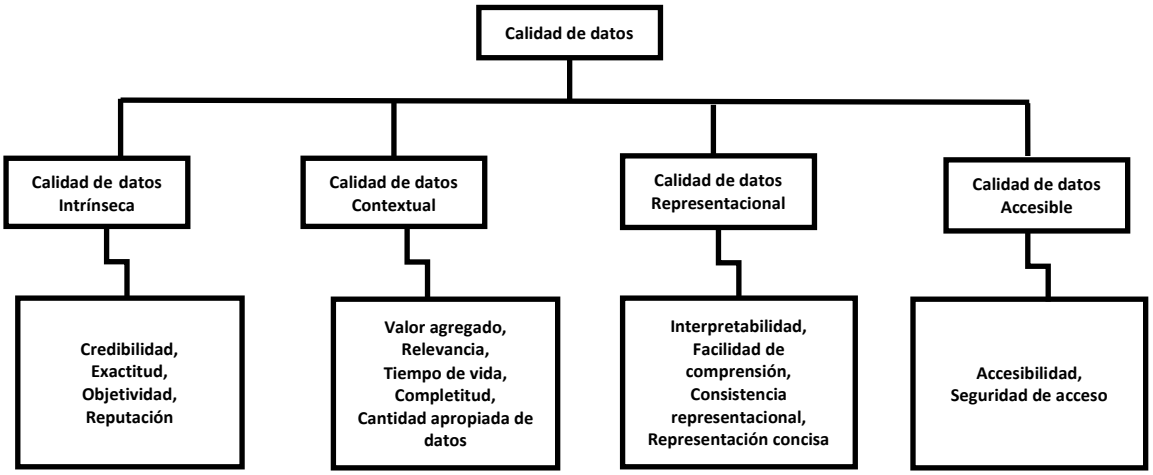


Figura 4 - Esquema conceptual de calidad de datos. [Fuente: (Wang and Strong, 1996)]

1.3.2 Categorías para dimensiones de calidad de datos: Intrínseca y Contextual

La calidad de datos se puede medir a través de muchas dimensiones como la exactitud, la completitud, el tiempo de vida, la relevancia, la objetividad, la credibilidad, entre otras (Wang, 1998, Eppler and Wittig, 2000). Algunas de estas dimensiones (por ejemplo, exactitud y objetividad) se prestan para ser medidas haciendo uso de la medición objetiva, que es intrínseca a los datos en sí, independientemente del contexto en el que se utilicen. Sin embargo, hay dimensiones de calidad de datos que no se pueden medir objetivamente. Por ejemplo, las dos dimensiones, relevancia y credibilidad (Fisher *et al.*, 2003, Watts and Zhang, 2004), tienden a variar con el contexto de uso. La relevancia de datos depende principalmente de la tarea que se desarrolle, ya que los datos que son altamente relevantes para una tarea pueden ser irrelevantes para otra; por ejemplo, diferentes datos son requeridos en tareas de ventas a la hora de realizar las hojas de balance, siendo estos datos irrelevantes para las tareas de *marketing*. Para entender los efectos contextuales de la calidad de datos, es importante tener en cuenta factores relacionados con el uso de los datos. En este sentido, varios factores como la relevancia de los datos en una tarea, la capacidad del usuario para entenderla, y la claridad de la tarea, afectan la

usabilidad de dichos datos (Watts *et al.*, 2009). Desde esta perspectiva de uso, la evaluación de la calidad de datos tiende a ser contextual (Moges *et al.,* 2013).

Los investigadores (Fisher *et al.*, 2003) y (Shankaranarayanan and Zhu, 2012) identificaron el impacto de la experiencia de los usuarios (principiante o experto en el dominio), el tipo de tarea (simple o compleja) que estos llevan a cabo, y las limitaciones de tiempo que presentan para desarrollarlas, como el posible uso de información relacionada con la calidad de datos. Sus resultados indican que cuando aumenta el nivel de experiencia de los usuarios y la complejidad de la tarea que realizan disminuye, se utiliza con más frecuencia información sobre calidad de datos en tareas para la toma de decisiones. Asimismo, otros autores como (Price and Shanks, 2011) investigaron el impacto que poseen diferentes estrategias para la toma de decisiones sobre el uso de información relacionada con calidad de datos.

En general, todos estos estudios ilustran la existencia de factores que pueden afectar la evaluación de la calidad de datos, y estimulan a investigar si otros factores, como la existencia de equipos de especialistas en calidad datos o el tamaño de las organizaciones, podrían impactar en dicha evaluación (Moges *et al.*, 2013).

1.3.3 Categorías para dimensiones de calidad de datos: Representacional y Accesible

Las dimensiones de calidad de datos mencionadas con más frecuencia en las categorías de calidad de datos representación y acceso son: consistencia representacional, facilidad de comprensión, accesibilidad y seguridad. La consistencia representacional y la facilidad de comprensión evalúan la representación y la comprensión de datos respectivamente. Los problemas típicos como el uso de diferentes formas de difusión o propagación de datos, diferentes formatos y diferentes nombres para las columnas o filas similares, son tratados por la dimensión consistencia representacional. Por otro lado, las dos últimas dimensiones de calidad de datos evalúan, respectivamente, la facilidad de acceso y la seguridad de los datos. La dimensión accesibilidad, por ejemplo, se refiere a la solicitud y entrega del tiempo de salida. Por ejemplo, los datos pueden ser clasificados como inaccesible si la brecha entre la entrada y el tiempo de entrega de salida es demasiado grande (Strong *et al.*, 1997).

1.3.4 Dimensiones de calidad de datos más comunes

La literatura proporciona una minuciosa clasificación relacionada con dimensiones de calidad de datos; sin embargo, hay una serie de discrepancias en la definición de la mayoría de dimensiones

debido a la naturaleza contextual del término calidad. Las seis clasificaciones que generalmente coinciden en la mayoría de los artículos y estudios realizados sobre el tema de dimensiones de calidad de datos son proporcionadas por (Jarke *et al.*, 1995, Redman, 1996, Wand and Wang, 1996, Wang and Strong, 1996, Bovee *et al.*, 2001, Naumann, 2002). Mediante el análisis de estas clasificaciones, es posible definir un conjunto básico de dimensiones de calidad de datos, incluyendo la exactitud, la completitud, la consistencia y el tiempo de vida, que constituyen el centro de atención de la mayoría de los autores (Scannapieco and Catarci, 2002).

Sin embargo, no existe ningún acuerdo general en el cual conjuntos de dimensiones definan la calidad de los datos, o en el cual se defina el significado exacto de cada dimensión. Las diferentes definiciones dadas en la literatura por los diferentes autores se discuten seguidamente (Batini *et al.*, 2009).

1.3.4.1 Exactitud

Varias definiciones son proporcionadas para el término exactitud. Wang and Strong (1996) definen exactitud como **"***el grado en que los datos son correctos, confiables y certificados***"**. Ballou and Pazer (1985) especifican que los datos son exactos, cuando los valores de los datos almacenados en la base de datos se corresponden con los valores reales. Según (Redman, 1996), la exactitud se define como una medida de la proximidad de un valor de datos *v*, a algún otro valor *v´*, que se considera correcto. En esta definición, de forma general, dos tipos de exactitud se pueden distinguir, la sintáctica y la semántica. Las metodologías de calidad de datos sólo tienen en cuenta la exactitud sintáctica y la definen como la cercanía de un valor *v*, a los elementos del dominio de definición correspondiente *D*. En la exactitud sintáctica, no interesa la comparación de *v* con su valor en el mundo real *v´*; más bien, lo que se busca es comprobar si *v* es uno de los valores de *D*, o cuan cerca está de los valores en *D*. Por ejemplo, *v* = **"***Jean***"** se considera sintácticamente preciso incluso si *v'* = **"***John***"** (Batini *et al.*, 2009).

1.3.4.2 Completitud

La completitud es definida como el grado en el que un conjunto de datos dado, incluye datos que describen el conjunto correspondiente de objetos del mundo real (Batini *et al.*, 2009).

La Tabla 3 presenta aportes de varias investigaciones que proporcionan varias definiciones de la dimensión completitud. Al comparar estas definiciones, se puede observar que hay un acuerdo substancial en el resumen de definiciones de dicha dimensión. A pesar de esto, las definiciones

difieren en el contexto al que se refieren, por ejemplo, sistemas de información en (Wand and Wang, 1996), almacenes de datos en (Jarke *et al.*, 1995), y entidades en (Bovee *et al.*, 2001).

Referencia	Definición
(Wand and Wang, 1996)	Capacidad de un sistema de información para representar todos los estados significativos de un sistema del mundo real
(Wang and Strong, 1996)	Grado en el que los datos son de suficiente amplitud, profundidad y alcance para la tarea en cuestión
(Redman, 1996)	Grado en el cual son incluidos valores en una colección de datos
(Jarke et al., 1995)	Porcentaje de información del mundo real entrado en fuentes de datos y/o almacenes de datos
(Bovee et al., 2001)	Información en la que se tiene todas las partes requeridas de una descripción de entidades
(Naumann, 2002)	Relación entre el número de valores no nulos en una fuente y el tamaño de la relación universal
(Liu and Chi, 2002)	Todos los valores que se supone sean recopilados de acuerdo con una teoría de recopilación

Tabla 3 - Definiciones proporcionadas para la dimensión completitud. [Fuente: (Batini et al., 2009)]

En el área de investigación de bases de datos relacionales, la completitud a menudo se relaciona con el significado de los valores nulos. Un valor nulo tiene el significado general de un valor perdido, o sea, un valor que existe en el mundo real, pero no está disponible en una colección de datos. Con el fin de caracterizar la completitud, es importante entender por qué el valor no se encuentra. Un valor puede faltar ya sea porque existe pero no se conoce, porque no existe, o porque no se sabe si existe (Atzeni and De Antonellis, 1993). A modo de ejemplo, en la Tabla 4 se muestra la tabla Persona con los atributos nombre, apellido, fecha de nacimiento, y correo electrónico, donde, para las tuplas con id = 2, 3 y 4 el valor del correo electrónico es nulo, sin embargo no se consideran casos de incompletitud (Batini *et al.*, 2009).

ID	Nombre	Apellido	Fecha de nacimiento	Correo electrónico
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	Null
3	Anthony	White	01/01/1936	Null
4	Marianne	Collins	11/20/1995	Null

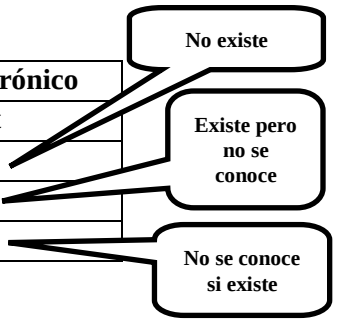


Tabla 4 - Valores nulos y datos completos. [Fuente: (Batini *et al.*, 2009)]

1.3.4.3 Consistencia

La dimensión consistencia o coherencia se refiere a la violación de las reglas semánticas definidas sobre los elementos de un conjunto de datos. Con referencia a la teoría relacional, las restricciones de integridad son un tipo de dichas reglas semánticas. En el ámbito estadístico, las ediciones de datos son las reglas semánticas típicas que permiten comprobaciones de la consistencia de datos. En la teoría relacional, dos categorías fundamentales para las restricciones de integridad se pueden distinguir, nombradas: restricciones de intra-relación y restricciones de inter-relación. Las restricciones de intra-relación definen el rango de valores admisibles para el dominio de un atributo. Ejemplos de esto son ***"la edad debe oscilar entre 0 y 120 años"***, o ***"si trabajo en años es inferior a 3, entonces salario no puede ser mayor de 25.000 pesos al año"***. Por otro lado, las restricciones de inter-relación implican atributos de diferentes relaciones (Batini *et al.*, 2009).

En la literatura existe una gran variedad de información sobre las bases de datos consistentes, vinculadas a la teoría relacional y a las restricciones de integridad. Por ejemplo, (Arenas *et al.*, 1999) toman en consideración el problema de la caracterización lógica de la noción de respuesta consistente en una base de datos relacional, el cual se puede violar dadas las restricciones de integridad. A razón de esto, los autores proponen un método para calcular respuestas consistentes, demostrando su solidez e integridad. Según (Calı *et al.*, 2004), las restricciones de integridad también se han estudiado como facilitadoras de la integración de datos.

En el área estadística, los datos de los cuestionarios de un censo tienen una estructura que corresponde a la de un esquema de cuestionario. Las reglas semánticas, llamadas ediciones, pueden ser definidas en el esquema de cuestionario para especificar el conjunto correcto de respuestas. Tales reglas denotan típicamente condiciones de error. Por ejemplo, una edición podría ser: si el estatus marital es ***"casado"*** la edad no debe ser inferior a 14 años. Después de la detección de registros erróneos, el acto de restaurar los valores correctos se llama imputación (Fellegi and Holt, 1976).

1.3.4.4 Dimensiones relacionadas con tiempo: Actualidad, Volatilidad, Tiempo de vida

Un aspecto importante en los datos es su actualización en el tiempo. Las dimensiones principales relacionados con el tiempo que se proponen en la literatura son actualidad, volatilidad y tiempo de vida. En la Tabla 5 se comparan las definiciones propuestas por varios autores en la literatura

para estas tres dimensiones de tiempo. Wand and Wang (1996) y (Redman, 1996) ofrecen definiciones muy similares sobre tiempo de vida y actualidad. Por otra parte, (Wang and Strong, 1996) y (Liu and Chi, 2002) asumen el mismo significado para el tiempo de vida, mientras que (Bovee *et al.*, 2001) proporcionan una definición del tiempo de vida en términos de la actualidad y la volatilidad. La definición de la actualidad expresada en (Bovee *et al.*, 2001) corresponde al tiempo de vida definido por (Wang and Strong, 1996) y (Liu and Chi, 2002). Esta comparación muestra que no hay acuerdos en el resumen de definiciones para las dimensiones relacionadas con el tiempo; típicamente, la actualidad y el tiempo de vida se utilizan a menudo para referirse al mismo concepto (Batini *et al.*, 2009).

Referencia	Definición
(Wand and Wang, 1996)	El tiempo de vida se refiere solamente a la demora entre el cambio de un estado del mundo real y la resultante modificación del estado del sistema de información
(Wang and Strong, 1996)	El tiempo de vida es el grado en el que la edad de los datos es apropiada para la tarea en cuestión
(Redman, 1996)	La actualidad es el grado en el que un dato se encuentra actualizado. El valor de un dato se encuentra actualizado si este es correcto, a pesar de posibles discrepancias que puedan existir causadas por cambios de retrazos del tiempo al valor correcto
(Jarke *et al.*, 1995)	La actualidad describe el momento en el cual la información es introducida en las fuentes y/o almacenes de datos La volatilidad describe el período de tiempo durante el cual la información es válida en el mundo real
(Bovee *et al.*, 2001)	El tiempo de vida tiene dos componentes: la actualidad y la volatilidad La actualidad es una medida de cuan antigua es la información, basada en cuánto tiempo hace desde que esta fue grabada La volatilidad es una medida de la inestabilidad de la información, o sea, la frecuencia del cambio de un valor para un atributo de entidad
(Naumann, 2002)	El tiempo de vida es la edad promedio de los datos en una fuente
(Liu and Chi, 2002)	El tiempo de vida es el grado en el que los datos se encuentran suficientemente actualizados para una tarea

Tabla 5 - Definiciones existentes sobre dimensiones relacionadas con el tiempo. [Fuente: (Batini et al., 2009)]

1.4 Métricas de calidad de datos

Cada dimensión de calidad de datos identificada en la fase de definición de la metodología TDQM (ver epígrafe 1.3), posee una o varias métricas (formas de evaluación) asociadas, definidas en la fase de medición de dicha metodología.

La clave para la medición es el desarrollo de métricas de calidad de datos. Estas métricas pueden ser las medidas básicas de calidad de datos tales como la exactitud de datos, el tiempo de vida, la completitud y la consistencia (Ballou and Pazer, 1985, Wang and Lee, 1998). En la cuenta de base de datos del cliente, las métricas de calidad de datos pueden ser diseñadas para realizar un seguimiento a los datos, por ejemplo (Wang, 1998):

- El porcentaje de los códigos postales incorrectos de direcciones de clientes, que se encuentran en una cuenta de cliente seleccionada al azar (libre de error).
- Un indicador de cuando los datos de la cuenta del cliente se actualizaron por última vez (el tiempo de vida o la actualidad para la comercialización de bases de datos y fines reglamentarios).
- El porcentaje de cuentas inexistentes o el número de cuentas con valores ausentes en el campo de código de la industria (completitud).
- El número de registros que violan la integridad referencial (consistencia).

1.4.1 Definición y tipos de métricas de calidad de datos

El nivel de calidad de datos en las organizaciones puede ser evaluado mediante el uso de cuestionarios o métricas (Moges *et al.*, 2013). Las métricas de calidad de datos son una forma de calcular **"***las puntuaciones de calidad***"** (o medidas) para los datos, las cuales son los resultados de aplicar dichas métricas. Las métricas de calidad de datos se suelen utilizar para definir los criterios de aceptabilidad de los datos y para realizar la corrección y mejora de los mismos (Emran *et al.*, 2012).

Pipino *et al.* (2002) desarrollaron métricas de calidad de datos basados en tres formas funcionales: la relación simple, operación de min/max, y la media ponderada.

La primera de estas mide la relación existente entre los resultados deseados con los resultados totales. A pesar de que la mayoría de personas tiende a medir las excepciones o errores, una forma preferida para realizar la medición es el número de resultados no deseados, dividido por el total de los resultados y restado con 1. Esta relación simple se adhiere a la convención de que 1 representa el resultado más deseable y 0 el resultado menos deseable (Ballou *et al.*, 1998, Ballou and Pazer, 1985, Huang *et al.*, 1998, Redman, 1996). Aunque una relación que ilustre los resultados indeseables brinde la misma información de una que ilustre los resultados más deseables, (Pipino *et al.*, 2002) sugieren manejar preferiblemente la relación que muestra

resultados positivos, ya que esta forma es útil para comparaciones longitudinales que ilustran las tendencias de mejora continua. Según (Pipino *et al.*, 2002), muchas dimensiones tradicionales de calidad de datos, tales como la exactitud, la completitud y la consistencia toman esta forma de medición.

La dimensión completitud se puede ver desde diversos puntos de vista, lo que lleva a diferentes métricas. En el nivel más abstracto, se puede definir el concepto de completitud de esquema como el grado en el cual las entidades y atributos se encuentran presentes en el esquema. En el nivel de datos, se puede definir la completitud de columnas como una función que evalúa los valores perdidos en una columna de una tabla. Esta medición corresponde a la integridad de columnas de Codd's (Codd, 1982), que evalúa los valores perdidos. Un tercer tipo se denomina completitud poblacional, donde si una columna debe contener al menos una ocurrencia de 50 estados, por ejemplo, pero sólo contiene 43 de estos, entonces tenemos incompletitud poblacional. Cada uno de los tres tipos (completitud de esquema, completitud de columnas, y completitud poblacional) pueden ser medidos tomando la relación de la cantidad de *ítems* incompletos entre el número total de elementos y substraerle 1 a esta división (Pipino *et al.*, 2002).

La dimensión consistencia se puede ver como la consistencia de valores redundantes a través de las tablas. Las restricciones de integridad referencial de Codd's son una instancia de este tipo de consistencia. Al igual que con las dimensiones discutidas previamente, una métrica para medir la consistencia puede ser la relación de violaciones de un tipo de consistencia específica entre el número total de comprobaciones de consistencias menos uno (Pipino *et al.*, 2002).

Las variables individuales se pueden medir usando una relación simple. Para manejar dimensiones que requieran la agregación de múltiples indicadores de calidad de datos (variables), la operación de mínimo o máximo puede ser aplicada, calculando el valor mínimo (o máximo) entre los valores normalizados de los indicadores individuales de calidad de datos. El operador mínimo (min) es conservador, en el sentido de que asigna a la dimensión un valor total no mayor que el valor de su indicador de calidad de datos más débil (evaluado y normalizado entre 0 y 1) (Pipino *et al.*, 2002).

Dos ejemplos interesantes de dimensiones que pueden hacer uso del operador mínimo son la credibilidad y la cantidad adecuada de los datos. El operador máximo resulta útil en métricas más complejas aplicables a las dimensiones del tiempo de vida y la accesibilidad (Pipino *et al.*, 2002).

El tiempo de vida refleja cómo los datos están actualizados con respecto a una tarea que esté utilizándolos. Una métrica para medir el tiempo de vida en general fue propuesta por (Ballou *et al.*, 1998), quienes sugieren medir el tiempo de vida como el máximo entre dos términos: 0, y 1 menos la relación de la actualidad entre la volatilidad. Esta métrica depende del contexto en el que se vaya a aplicar. Si la dimensión del tiempo de vida no es crítica para la tarea en cuestión, entonces una medida de sensibilidad más liberal se puede aplicar. Inversamente, si la dimensión es muy crítica, una medida de sensibilidad conservadora se sugiere (Moges *et al.*, 2013). Aquí, la actualidad se define como la edad más el tiempo de entrega menos el tiempo de entrada. La volatilidad se refiere a la longitud del tiempo de los datos que sigue siendo válida; el tiempo de entrega se refiere a cuando los datos son entregados al usuario; el tiempo de entrada se refiere a cuando se reciben los datos por el sistema; y la edad se refiere a la edad de los datos cuando se reciben primeramente por el sistema (Pipino *et al.*, 2002).

En dicha métrica, un exponente se puede utilizar como un factor de sensibilidad, elevando el valor máximo a este exponente. El valor del exponente es una tarea dependiente y refleja el juicio del analista. Por ejemplo, supongamos que la calificación del tiempo de vida sin utilizar el factor de sensibilidad (equivalente a un factor de sensibilidad igual a 1) es 0,81. El uso de un factor de sensibilidad igual a 2 debería entonces de producir una calificación para el tiempo de vida de 0,64 (un factor de mayor sensibilidad refleja el hecho de que los datos se conviertan más rápidos en menos tiempo) y 0,9 cuando el factor de sensibilidad es 0,5 (un factor de menor sensibilidad refleja el hecho de que los datos pierden el tiempo de vida a un menor ritmo) (Pipino *et al.*, 2002).

Para el caso multivariable, una alternativa al operador mínimo es el uso de la media ponderada de variables. Si una empresa, por ejemplo, tiene una buena comprensión de la importancia de cada variable para la evaluación global de una dimensión, entonces el uso de una media ponderada de las variables es lo adecuado. Para asegurar la calificación esta es normalizada, cada factor de ponderación debe estar entre cero y uno, y los factores de ponderación deben añadir a uno (Pipino *et al.*, 2002).

Más reciente, en (Fisher *et al.*, 2009) proponen una métrica para la exactitud cambiando la escala de razón simple a un vector de aproximación, que incluye porcentajes, una medida de aleatoriedad, y una distribución de probabilidad. La métrica combina una relación simple, mediante el cálculo del número de celdas con errores entre el número total de celdas, con una

medida de aleatoriedad, calculada mediante el algoritmo Lempel-Ziv[1] (L-Z, por sus siglas en inglés) para medidas de complejidad. Este algoritmo se utiliza para diferenciar si los errores en una base de datos son aleatorios o sistemáticos en la naturaleza. Una vez determinada la aleatoriedad de los errores, una distribución de probabilidad se utiliza para ayudar a abordar diversas cuestiones de gestión. La métrica se basa en la suposición de que el valor corresponde a la gama de validez posible (Moges *et al.*, 2013).

Del mismo modo, (Sessions and Valtorta, 2009) sugieren un enfoque de medición para la exactitud usando redes bayesianas[2].

Las métricas discutidas anteriormente pueden ser para tareas independientes o tareas dependientes (Moges *et al.*, 2013). Las métricas para tareas independientes, reflejan estados de los datos sin el conocimiento del contexto de su aplicación, y pueden aplicarse a cualquier conjunto de datos, independientemente de la tarea en cuestión. Las métricas para las tareas dependientes, las cuales incluyen reglas de organización de negocios, regulaciones de compañías y del gobierno, y restricciones establecidas por administradores de base de datos, se desarrollan en contextos de aplicación específicos (Pipino *et al.*, 2002).

1.4.2 Métricas para dimensiones de calidad de datos más comunes

Una cantidad considerable de métricas han sido y están siendo desarrolladas por diferentes investigadores y organizaciones de calidad de datos, y muchas de las dimensiones son multivariantes en la naturaleza. La elección de las variables o componentes específicos para la medición puede ser más difícil que definir la métrica en general, que a menudo se reduce a la forma de relación. Aunque caen dentro de una categoría dimensional específica, la medida para evaluar una dimensión determinada puede variar de una organización a otra (Lee *et al.*, 2006).

Tomando en cuenta todo lo explicado (ver subepígrafe 1.4.1), a continuación se muestra una tabla resumen, que acapara las métricas propuestas por diferentes autores, para las dimensiones de calidad de datos más comunes (ver Tabla 6).

[1] Lempel-Ziv es un algoritmo para la comprensión de pérdidas de datos. De hecho, no es un único algoritmo, sino toda una familia de algoritmos, derivadas de los dos algoritmos propuestos por Jacob Ziv y Abraham Lempel en 1978 (Zeeh, 2003).

[2] Una red bayesiana o red de creencia, es un modelo gráfico probabilístico que representa un conjunto de variables aleatorias y sus dependencias condicionales a través de un grafo acíclico dirigido (DAG, por sus siglas en inglés) (Sessions and Valtorta, 2009).

Dimensión	Métrica de calidad de datos		Referencia
	Tarea independiente	Tarea dependiente	
Exactitud	$1 - \frac{\text{número de unidades de datos con error}}{\text{total de unidades de datos}}$		(Pipino et al., 2002, Lee et al., 2006)
	$f(\text{porcentaje de exactitud, medida de aleatoriedad con } L - Z, \text{ distribución de probabilidad})$		(Fisher et al., 2009)
	$\text{acercamiento a las redes bayesianas}$		(Sessions and Valtorta, 2009)
	$\frac{\text{número de unidades de datos correctos}}{\text{total de unidades de datos}}$		(Pipino et al., 2002, Batini et al., 2009)
Completitud	$1 - \frac{\text{número de artículos incompletos}}{\text{total de artículos}}$		(Codd, 1982, Pipino et al., 2002)
	$\frac{\text{número de valores no nulos}}{\text{total de valores}}$		(Batini et al., 2009)
	$1 - \frac{\text{proporción de violaciones de un tipo específico de consistencia}}{\text{total de verificaciones de consistencia}}$		(Pipino et al., 2002)
Consistencia	$\frac{\text{número de valores consistentes}}{\text{total de valores}}$		(Batini et al., 2009)
Actualidad	$(\text{tiempo de entrega} - \text{tiempo de entrada}) + \text{edad}$		(Pipino et al., 2002, Batini *et al.*, 2009, Moges *et al.*, 2013)
Volatilidad	$\text{longitud del tiempo por la cuál los datos siguen siendo válidos}$		(Pipino et al., 2002, Batini et al., 2009, Moges et al., 2013)
Tiempo de vida	$Q_{act} = e^{-\text{disminución}(A)*\text{edad}(W*A)}$		(Pipino et al., 2002, Lee et al., 2006)
		$max(1 - \frac{actualidad}{volatilidad}, 0)^s$	(Pipino *et al.*, 2002, Batini *et al.*, 2009, Moges *et al.*, 2013)

Tabla 6 - Métricas de calidad de datos de la literatura [Basado en: (Moges *et al.*, 2013)].

Algunas connotaciones importantes para las tareas independientes no mencionadas con anterioridad a tener en cuenta para un mejor entendimiento de la Tabla 6 son: en el caso de la exactitud, *f* indica *"función de"*. En el caso del tiempo de vida, *Qact* es el nivel de actualidad de los datos, *A* es un atributo, *W* es un valor de atributo, edad se refiere a la diferencia entre el

momento en que la calidad de datos es evaluada y el momento de adquisición de los datos; y la disminución se refiere a la velocidad promedio de descenso del tiempo de vida de los valores del atributo, para el atributo bajo consideración (Moges *et al.*, 2013).

Cualquiera sea la naturaleza de las métricas de calidad de datos, estas son implementadas como parte de un nuevo sistema de fabricación de información o como un complemento de rutinas de utilidad en un sistema existente. Con las métricas de calidad de datos, las medidas pueden ser obtenidas a lo largo de varias dimensiones para el análisis (Pipino *et al.*, 2002).

1.5 Perfilado de datos y Auditoría de datos

Una vez terminada la fase de medición de la calidad de datos (ver epígrafe 1.4), con los resultados obtenidos de dicha fase se investigan las principales causas que conllevan a los problemas actuales de calidad de datos en la fase de análisis.

Investigar cuáles son las características de los datos y qué provecho se puede extraer de estos, se ha convertido en un objetivo prioritario en empresas e instituciones que manejan grandes volúmenes de información y se enfrentan a fuertes competencias de mercado. En este marco adquiere especial atención los procesos de perfilado de datos, los cuales mediante la utilización de métodos estadísticos y de minería de datos, brindan amplios conocimientos sobre las fuentes de datos (González, 2010).

El perfilado de datos se puede definir como el proceso de análisis de las fuentes de datos con respecto al dominio[3] de calidad de datos (Marshall, 2007), para identificar y priorizar los problemas de calidad que puedan presentar los datos. Los reportes del perfilado de datos sobre la completitud de conjuntos de datos y registros de datos, organizan dichos problemas por importancia; por ejemplo, las salidas a la distribución de los problemas de calidad de datos en un conjunto de datos, y las listas de valores perdidos en los registros existentes (Olson, 2003). La identificación de problemas de calidad de datos antes de iniciar un proyecto de limpieza de datos es crucial para asegurar la entrega de información precisa (Barateiro and Galhardas, 2005).

Por otro lado, la auditoría de datos se define como el proceso de gestionar cómo los datos se ajustan a los propósitos definidos por la organización, estableciendo políticas para gestionar los criterios de datos para la empresa (Daza *et al.*, 2012).

[3] El dominio de calidad de datos es una aplicación o uso de datos que imponen un conjunto de reglas de calidad, es decir, la especificación de uno o más problemas de calidad que no deben existir en un conjunto de datos (Barateiro and Galhardas, 2005).

En orden para el analista, determinar el alcance de las causas subyacentes principales de los problemas de calidad de datos, y planificar el diseño de las herramientas que se pueden utilizar para hacer frente a dichos problemas, se consideran dos aspectos de suma importancia para entender los problemas de calidad de datos más comunes; con el propósito de que la clasificación formada sea útil para las comunidades de almacenes de datos y calidad de datos (Pandey, 2014).

1.5.1 Causas y problemas que conllevan a una mala calidad de datos

La taxonomía que se presenta a continuación para los problemas de calidad que pueden presentar los datos se basa en lo expresado por (Kim *et al.*, 2003, Müller and Freytag, 2005, Oliveira *et al.*, 2005, Rahm and Do, 2000). En esta taxonomía se dividen los problemas de calidad de datos a nivel de esquema y de instancia. Los problemas en el nivel de esquema pueden evitarse con un diseño de esquema mejorado y no dependen de los contenidos reales de los datos. Por otro lado, los problemas en el nivel de instancia están relacionados con el contenido de los datos, solo que aquí no pueden ser evitados con una mejor definición de esquema, como en los lenguajes de definición de esquema, ya que estos no son lo suficientemente potentes para especificar todas las limitaciones de los datos requeridos. En esta taxonomía, los problemas con los datos en el nivel de instancia incluyen todos los problemas de datos que no son problemas en el nivel de esquema, por lo que la taxonomía es completa en este punto (Barateiro and Galhardas, 2005).

1.5.1.1 Problemas de calidad de datos a nivel de esquema

Los problemas de calidad de datos en el nivel de esquema se pueden prevenir con un diseño de esquema mejorado, y una correcta traducción del esquema e integración de todos los datos manejados. Los Sistemas Gestores de Bases de Datos Relacionales (RDBMS, por sus siglas en inglés) proporcionan mecanismos importantes para asegurar definiciones de esquema. Por tanto, se pueden distinguir entre problemas de calidad de datos a nivel de esquema que pueden ser evitados mediante un RDBMS, y problemas de calidad de datos a nivel de esquema que no pueden ser evitados mediante un RDBMS (Barateiro and Galhardas, 2005).

1) Evitados por los RDBMS

Durante la fase de diseño de bases de datos, las restricciones de integridad pueden ser definidas para especificar diferentes condiciones que deben satisfacer los objetos en la base de datos (Türker and Gertz, 2001). En este sentido, SQL (ISO and ANSI, 1999) proporciona un lenguaje

para apoyar la especificación de las siguientes restricciones de integridad declarativas: (i) restricción no nulo, para evitar que una columna tome un valor nulo; (ii) restricción predeterminada, para especificar el valor predeterminado de una columna dada; (iii) restricción único, para definir que una columna o un conjunto de columnas deben tener valores únicos dentro de una tabla; (iv) restricción de clave principal, que corresponde a la restricción no nulo y a la restricción único; (v) restricción de integridad referencial, para especificar atributos cuyos valores deben coincidir con los valores de una clave principal (o única) de una tabla externa; (vi) restricción de verificación, para especificar una condición definida por el usuario; (vii) restricción de dominio, para definir valores restringidos del dominio de columna; (viii) restricción de afirmación, define una restricción de integridad tabla-independiente. SQL (1999), también ofrece una definición procedural para las restricciones de integridad usando disparadores (*triggers*), que son procedimientos invocados por el RDBMS en respuesta a eventos específicos de bases de datos (Barateiro and Galhardas, 2005).

La siguiente lista resume los problemas de calidad de datos que pueden ser evitados con una definición adecuada de restricción de integridad (Barateiro and Galhardas, 2005):

- **Datos perdidos:** Datos que no han sido llenados. Una restricción no nula puede evitar este problema.
- **Tipo de datos incorrecto:** Violación del tipo de datos de una restricción; por ejemplo, la edad del empleado es ***"xy"***. Las restricciones de dominio pueden evitar este problema.
- **Valor de datos incorrecto:** Violación del rango de datos de una restricción; por ejemplo, si la edad de un empleado debe pertenecer al rango [18, 65], entonces una edad de 15 años es un valor de datos incorrecto. Las restricciones de verificación y dominio se utilizan para evitar este problema.
- **Datos colgantes:** Los datos que en una tabla no tienen equivalente en otra tabla; por ejemplo, un identificador de departamento que no exista en una tabla Departamento, y sin embargo hay una referencia a este valor en una tabla Empleado. Este problema se aborda mediante restricciones de integridad referencial (es decir, claves o llaves foráneas).
- **Duplicado exacto de datos:** Los diferentes registros tienen el mismo valor en un campo (o una combinación de campos) para los cuales la duplicación de valores no está permitida; por ejemplo, un número de seguro social de un empleado. Las restricciones de claves únicas y primarias pueden evitar duplicados exactos.

- **Restricciones genéricas de dominio:** Registros o valores que violan una restricción de dominio de un atributo; por ejemplo, dependencias de atributos o un número máximo predefinido de filas. Las restricciones de afirmación y de dominio pueden evitar este problema. Sin embargo, estas características no son proporcionadas por la mayoría de los RDBMS (por ejemplo, Oracle 9i) que manejan este problema de datos utilizando los *triggers*.

2) No evitados por los RDBMS

Los siguientes problemas de calidad de datos no pueden ser manejados por las restricciones de integridad de los RDBMS (Barateiro and Galhardas, 2005):

- **Datos categóricamente incorrectos:** Un valor de categoría que está fuera del rango de dicha categoría, por ejemplo, países y estados respectivos. El uso de un nivel de abstracción mal, por ejemplo, ***"****alimento congelado****"*** o ***"****pizza congelada****"*** en lugar de ***"****alimentos****"***, también se considera un tipo de datos incorrecto categóricamente.
- **Datos desfasados temporalmente:** Los datos que violan una restricción temporal, que especifica el instante de tiempo o intervalo en el que los datos son válidos; por ejemplo, el salario de un empleado ya no es válido cuando este empleado es ascendido.
- **Datos espaciales inconsistentes:** Inconsistencias entre los datos espaciales (por ejemplo, las coordenadas, las formas) cuando están almacenados en múltiples campos; por ejemplo, las coordenadas rectangulares de un punto en una tabla deben combinarse para producir un rectángulo cerrado.
- **Conflictos de nombre:** El mismo nombre de campo se utiliza para diferentes objetos (homónimos), o se utilizan nombres diferentes para el mismo objeto (sinónimos).
- **Conflictos estructurales:** Diferentes representaciones de esquema del mismo objeto en diferentes tablas o bases de datos; por ejemplo, una dirección se puede representar en un campo de forma libre o descomponerse en los campos calle, ciudad, estado, etc.

Aunque, en el nombre y en las estructuras pueden producirse conflictos dentro de un mismo esquema de datos, con frecuencia surgen en un escenario multi-esquema (Barateiro and Galhardas, 2005).

1.5.1.2 Problemas de calidad de datos a nivel de instancia

Los problemas con los datos en el nivel de instancia se refieren a errores e inconsistencias en los datos que no son visibles o evitados a nivel de esquema. Se debe tener en cuenta que las instancias de datos también reflejan los problemas a nivel de esquema (por ejemplo, un registro con un valor nulo en un campo requerido). Los problemas de datos a nivel de instancia se pueden dividir en problemas de un único registro y problemas de múltiples registros (Barateiro and Galhardas, 2005).

1) Registro simple

Los problemas con los datos en un registro individual o simple se refieren a uno o varios atributos de un único registro. En otras palabras, este tipo de problemas se relaciona con una sola entidad y no depende de otra información almacenada en la base de datos (Barateiro and Galhardas, 2005):

- **Falta de datos en un campo no nulo:** Atributos que son rellenados con algún valor ficticio; por ejemplo, un número de seguro social 99999 es un valor indefinido utilizado para superar la restricción no nula.
- **Datos erróneos:** Datos que son válidos, pero no se ajustan a la entidad real; un ejemplo de datos erróneos es 31 que indica la edad de un empleado cuando este en realidad tiene 30 años.
- **Errores ortográficos:** Palabras mal escritas en campos de bases de datos; por ejemplo, ***"Jhon Stevens"*** en lugar de ***"John Stevens"***.
- **Valores implícitos:** Existencia de datos extraños en algún campo de datos; un ejemplo común de valores implícitos es la inserción de un título en un campo de nombre, por ejemplo, ***"Presidente John Stevens"***.
- V**alores de campos perdidos:** Datos que son almacenados en el campo equivocado. Por ejemplo, el valor ***"Portugal"*** ubicado en el atributo ciudad.
- **Datos ambiguos:** Datos que pueden ser interpretados en más de una manera, con diferentes significados. La existencia de datos ambiguos puede ocurrir debido a la presencia de abreviaturas o contextos incompletos, como los siguientes:
    - **Abreviaturas:** El nombre abreviado ***"J. Stevens"*** puede ser ampliado de diferentes formas, como: ***"John Stevens"***, ***"Jack Stevens"***, ***"Jeff Stevens"***, etc.

- **Contextos incompletos:** El nombre de la ciudad ***"Miami"*** puede ser soportado para el estado de la Florida, o el estado de Ohio.

2) Registros múltiples

Los problemas con los datos en múltiples registros no pueden ser detectados considerando cada registro por separado, ya que como el nombre lo indica, se está haciendo referencia a más de un registro. Se debe tener en cuenta que pueden ocurrir múltiples problemas de registro entre los registros que pertenecen al mismo conjunto de entidades (o tabla), o a diferentes conjuntos de entidades (correspondientes a diferentes tablas o incluso a diferentes bases de datos) (Barateiro and Galhardas, 2005):

- **Registros duplicados:** Registros que representan la misma entidad real y no contienen información contradictoria; por ejemplo, los siguientes registros de empleados se consideran duplicados: Empleado1 (Nombre = ***"John Stevens"***, Dirección = ***"223, Primera Avenida, Ciudad de Nueva York"***, Nacimiento = 01/01/1975); Empleado2 (Nombre = ***"J. Stevens"***, Dirección = ***"223, 1st Avenue, New York"***, Nacimiento = 01/01/1975).
- **Contradicción de registros:** Registros que representan la misma entidad real y contengan algún tipo de información contradictoria; por ejemplo, si se consideran los registros Empleado1 y Empleado2, pero con la siguiente información: Empleado1 (Nombre = ***"John Stevens"***, Dirección = ***"223, Primera Avenida, Ciudad de Nueva York"***, nacimiento = 01/01/1975); Empleado2 (Nombre = ***"John Stevens"***, Dirección = ***"223, Primera Avenida, Ciudad de Nueva York"***, nacimiento = 01/01/1965).
- **Datos no estandarizados:** Diferentes registros que no utilizan las mismas representaciones de datos, invalidando así su comparación:
  - **Transposición de palabras:** En un solo campo, las palabras pueden aparecer con diferentes ordenamientos; por ejemplo, los nombres ***"John Stevens"*** y ***"Smith, Jack"*** no utilizan la misma regla de ordenación.
  - **Formato de codificación:** Uso de diferentes formatos de codificación, por ejemplo, ASCII, UTF-8.

- **Formato de representación:** Diferentes representaciones de un formato para la misma información; un ejemplo es el formato de moneda que puede ser € 10.5, 10.5 €, etc.
- **Unidad de medida:** Diferentes unidades utilizadas en registros distintos; por ejemplo, las distancias dadas en cm y pulgadas.

1.5.2 Algoritmos o métodos empleados en el perfilado de datos

Uno de los inconvenientes, mencionado con anterioridad, que pueden presentar los datos almacenados se da cuando una misma entidad del mundo real se almacena más de una vez de diferentes formas. El proceso para detectar este tipo de problemas se conoce como registro de enlace (*record linkage)* en el área de la estadística; radicalización de base de datos (*database hardening)* en el área de la inteligencia artificial; mezclar-depurar (*merge-purge)*, desduplicar datos (*data deduplication)* o identificación de instancias (*instance identification)* en el área de las bases de datos; otros nombres como resolución correferencial (*coreference resolution)* y detección de registros duplicados (*duplicate record detection)* también se usan con frecuencia (Amón *et al.*, 2012).

Este proceso fue identificado inicialmente por (Dunn, 1946). Algunos fundamentos probabilísticos fueron desarrollados posteriormente por (Newcombe and Kennedy, 1962), y formalizados por (Fellegi and Holt, 1976), como una regla de decisión probabilística. Algunas mejoras de este modelo han sido propuestas por (Winkler and Census, 1993). Esencialmente, el proceso es como sigue; dado un conjunto *R* de registros: (i) se define un umbral en el intervalo cerrado [0,1], (ii) se compara cada registro con los demás registros y (iii) si la similitud entre una pareja de registros es mayor o igual que el umbral, se supone que son duplicados y se considera que son representaciones de una misma entidad del mundo real (Amón *et al.*, 2012).

A consecuencia de esto, se han desarrollado funciones de similitud para la detección de duplicados; algunas de ellas son de tipo fonético como *Soundex* (Raghavan and Allan, 2004), *Metaphone* (Philips, 1990), *Double Metaphone* (Philips, 2000), *Onca* (Gill, 1997) y *Nysiis* (Taft, 1970). Estas técnicas, se basan en la forma como se pronuncian las palabras en un idioma en particular y no están orientadas al idioma español. Otro tipo de funciones de similitud se basan en emparejamiento de patrones. En esta última categoría, se encuentran técnicas como

Levenshtein, *Brecha Afín*, *Smith-Waterman*, *Jaro*, *Jaro-Winkler*, *Bi-grams*, *Tri -grams*, *Monge-Elkan* y *SoftTF-IDF* (Amón *et al.*, 2012).

En los gestores de datos el intento más generalizado de buscar cadenas similares (no exactas) está relacionado con la utilización de la función *Soundex* (cadena) (Gálvez, 2006). *Soundex* es un algoritmo de codificación fonética, que convierte una palabra en un código (Else, 2002). El código *Soundex* consiste en sustituir las consonantes de la palabra afectada por un número; si es necesario se agregan ceros al final del código para conformar un código de 4 dígitos. *Soundex* elige la clasificación de los caracteres con base en el lugar de articulación de la lengua inglesa. La Tabla 7 presenta las equivalencias usadas por *Soundex* (Amón *et al.*, 2012).

Dígito	**Caracteres**
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Tabla 7 - Equivalencias *Soundex* [Fuente: (Amón *et al.*, 2012)]

Debido a que en el idioma inglés, las letras *A, E, I, O, U, H, W* e *Y* no hacen diferenciación fonética, son descartadas. Adicionalmente existen otras reglas complementarias para la codificación de letras dobles (si el texto tiene letras dobles, estas se deben tratar como una sola) y para letras con el mismo código, las cuales al realizar la operación de descarte quedan una lado de la otra (si el texto tiene diferentes letras una al lado de la otra que tienen el mismo número en la guía de codificación *Soundex*, estas se deben tratar como una sola letra), entre otras. Por ejemplo: Giraldo se codifica *G643* (G, 6 por la R, 4 por la L, 3 por la D, las otras letras se descartan). Juan se codifica *J500* (J, 5 por la N, las otras letras se descartan, y se agregan dos ceros) (Amón *et al.*, 2012).

Las limitaciones del algoritmo *Soundex* han sido documentadas en (Patman and Shaefer, 2003, Stanier, 1990) y han dado lugar a varias mejoras, pero ninguna orientada hacia el idioma español. Es claro, que *Soundex* no está orientado al español ya que ni siquiera contempla el juego de caracteres (***"ñ"***, ***"ll"***). Asimismo, la dependencia de la letra inicial, la agrupación por punto de

articulación del idioma inglés y el estar limitado a cuatro caracteres implica que no es eficiente para detectar errores ortográficos comunes en el idioma español (Amón *et al.*, 2012).

En el caso del emparejamiento de patrones, la distancia de edición entre dos cadenas fue introducida por Levenshtein (Levenshtein, 1965) como la cantidad mínima de operaciones de inserción, eliminación y sustitución que hay que hacer para transformar una cadena en la otra. La función así definida se demuestra que constituye una métrica y ha tenido algunas variaciones en el tiempo. Levenshtein consideró todas las operaciones con costo unitario, en trabajos posteriores se planteó que era posible que cada una de las operaciones de edición tuvieran costos diferentes (Hall and Dowling, 1980), por lo que se adicionó una nueva operación: la transposición, intercambio de caracteres adyacentes que es un error frecuente en las personas que teclean rápido (Zobel and Dart, 1996). No se han encontrado en la literatura heurísticas para determinar los costos de las operaciones, aunque en (Scherbina, 2005, Smith and Waterman, 1981) se señala que, a partir de experimentos, se obtienen buenos resultados cuando los costos no son unitarios sino cuando tienen valor 2.

La métrica de *Smith Waterman* (Smith and Waterman, 1981), utilizada inicialmente para buscar alineación entre moléculas, utiliza la idea de la distancia de edición, y propone utilizar costos negativos en las operaciones de inserción y eliminación, y costos positivos para las operaciones de sustitución y para cuando haya coincidencia. Aquí se toma el costo de las operaciones de inserción y eliminación en función del tamaño del ***"hueco"*** que se produce al hacer una de estas operaciones.

La métrica de *Jaro* (Jebamalar-Tamilselvi and Saravanan, 2008) es usada comúnmente para buscar similitudes entre nombres, en sistemas de enlaces de registros (Winkler, 2006). Se basa en el conteo de los caracteres comunes de ambas cadenas, el número de transposiciones y el empleo de una expresión donde intervienen estos valores y la longitud de las cadenas.

La métrica *Q-Gram* (Ukkonen, 1992) y algunas de sus variantes (Li *et al.*, 2007) se basan en la determinación de subcadenas de longitud q (habitualmente se usa *q* igual a 1, 2 o 3). Se buscan las subcadenas comunes en las dos cadenas pues se plantea que, si dos cadenas son similares, comparten un alto número de *q-gram*.

Las métricas basadas en *tokens* dividen las cadenas en partes (*tokens*), a partir de espacios en blanco, signos de puntuación, etc. Entre las métricas basadas en *tokens* se encuentra la similitud de *Monge-Elkan* (Monge and Elkan, 1996), que se calcula como la similitud máxima promedio

entre una pareja de *tokens*. Una variante se presenta en (Gelbukh *et al.*, 2009) donde se utiliza la media aritmética generalizada en lugar del promedio. Otra métrica basada en *tokens* es la *similaridad del coseno* (Cohen *et al.*, 2003), en la que a cada palabra se da un peso en dependencia de la frecuencia relativa de aparición en la cadena, formándose vectores con estos valores y buscándose el coseno del ángulo que existen entre estos vectores. Estas técnicas son muy utilizadas en la recuperación de la información (Rema *et al.*, 2008, Chaudhuri *et al.*, 2009).
Estas distancias, incluyendo las de edición, no tienen en cuenta de manera explícita los errores tipográficos. Considerando que la operación más frecuente dentro de los errores tipográficos es la sustitución, pudiera mejorarse el proceso de estandarización de cadenas de caracteres teniendo en cuenta la posición de los caracteres en el teclado al calcular la distancia entre dos cadenas (Porrero, 2011).

1.5.3 Herramientas comúnmente usadas para el perfilado de datos

Las herramientas de calidad de datos tienen como objetivo detectar y corregir problemas de datos que afectan la precisión y la eficiencia de las aplicaciones de análisis de datos (Barateiro and Galhardas, 2005). Es comúnmente aceptado que las herramientas de calidad de datos se pueden agrupar de acuerdo con la parte del proceso de calidad de datos que estas cubren (Olson, 2003). El siguiente conjunto de herramientas comerciales y de investigación implementan técnicas para el perfilado de datos (ver Tabla 8) (Barateiro and Galhardas, 2005).

Herramienta	**Comercial (C)/Investigación (I)**
dfPower	C
ETLQ	C
Trillium	C
Migration Architect	C
WizWhy	C
WizRule	C
DataStage	C
Firstlogic	C
Hummingbird ETL	C
Informatica ETL	C
Sagent	C
SQL Server 2000 DTS	C
SQL Server 2005	C
SQL Server 2008	C
Data Cleaner	C
Oracle	C
Oracle 10g	C

Profiler	C
Sunopsis	C
Talend	C
IBM Ascential	C
Potter´s Wheel	I
Ken State University Tool	I

Tabla 8 - Herramientas comerciales y de investigación para la implementación de técnicas de perfilado de datos. [Basado en: (Barateiro and Galhardas, 2005)]

A continuación y tomando en cuenta las herramientas anteriormente mencionadas, se muestra en la Tabla 9 un resumen de los principales rasgos con los que cuentan las herramientas de calidad de datos, analizando diferentes funcionalidades que poseen las mismas.

Herramienta	Fuentes de datos	Capacidad de extracción	Capacidad de carga	Actualizaciones incrementales	Interfaz	Repositorio de metadatos	Técnicas de rendimiento	Versiones	Biblioteca de funciones	Lenguaje vinculante	Depuración	Manejo de errores	Linaje de datos
dfPower	Varias	Y	Y	-	G	Y	Y	-	-	-	-	-	-
ETLQ	Varias	Y	Y	-	G	Y	Y	Y	Y	N	-	-	-
Migration Architect	Varias	-	-	-	G	Y	-	-	-	-	-	-	-
Trillium	Varias	Y	Y	-	G	Y	Y	Y	Y	N	Y	-	Y
WizWhy	DB,FF	-	-	-	G	-	Y	-	-	-	-	-	-
WizRule	DB,FF	-	-	-	G	-	Y	-	-	-	-	-	-
SQL Server 2008	Varias	Y	Y	-	G	-	-	-	Y	N	-	-	-
DataStage	Varias	Y	Y	-	G	Y	Y	Y	Y	Y	Y	Y	Y
Firstlogic	DB,FF	Y	Y	-	G	Y	Y	-	Y	Y	-	-	-
Hummingbird ETL	Varias	Y	Y	Y	G	Y	Y	Y	Y	N	Y	M	Y
Informatica ETL	Varias	Y	Y	Y	G	Y	Y	Y	Y	Y	Y	Y	Y
Sagent	Varias	Y	Y	X	G	Y	Y	X	Y	N, SQL	-	-	-
SQL Server 2000 DTS	Varias	Y	Y	X	G	X	-	-	Y	N	X	X	X

SQL Server 2005	Varias	Y	Y	-	G	-	-	-	Y	N	-	-	-
Oracle	-	-	-	-	G	-	-	-	-	-	-	-	-
Oracle 10g	-	-	-	-	G	-	-	-	-	-	-	-	-
DataCleaner	-	-	-	-	G	-	-	-	-	-	-	-	-
Profiler	-	-	-	-	G	-	-	-	-	-	-	-	-
Sunopsis	DB,FF	Y	Y	Y	G	Y	Y	Y	X	SQL	Y	Y	X
Talend	-	-	-	-	G	-	-	-	-	-	-	-	-
IBM Ascential	-	-	-	-	G	-	-	-	-	-	-	-	-
Potter´s Wheels	Y	-	ODBC	-	G	Y	-	-	-	-	Y	-	-
Ken State University Tool	-	-	X	-	NG	-	-	-	-	-	-	Y	-

Tabla 9 - Funcionalidades generales de herramientas de calidad de datos comerciales y de investigación usadas en el perfilado de datos. Y: soportado; X: no soportado; -: información desconocida; N: nativo; DB: bases de datos relacionales; FF: archivos planos; G: gráfica; NG: no gráfica; M: manual. [Basado en: (Barateiro and Galhardas, 2005)]

Algunas connotaciones importantes referentes a los rasgos de las herramientas se muestran a continuación para un mejor entendimiento de la Tabla 9 (Barateiro and Galhardas, 2005):

- **Fuentes de datos:** Capacidad de extraer datos desde diferentes y heterogéneas fuentes. Las fuentes de datos pueden ser bases de datos relacionales, archivos planos, archivos XML, hojas de cálculo, sistemas legados y paquetes de aplicaciones (como SAP), fuentes basadas en internet y software EAI (*Enterprise Application Integration*).
- **Capacidad de extracción:** El proceso de extracción de datos desde fuentes de datos debe proporcionar las siguientes capacidades importantes: (i) la posibilidad de programar las extracciones por hora, intervalo o evento; (ii) un conjunto de reglas para seleccionar datos de la fuente y (iii) la capacidad para seleccionar y combinar registros de múltiples fuentes.
- **Capacidad de carga:** El proceso de carga de datos dentro del sistema destino debe ser capaz de: (i) cargar datos dentro de múltiples tipos de sistemas destinos; (ii) cargar datos

dentro de los sistemas destinos heterogéneos en paralelo; (iii) actualizar y añadir datos dentro de las fuentes de datos del destino y (iv) crear automáticamente las tablas del destino.

- **Actualizaciones incrementales:** Capacidad de actualizar de forma incremental destinos de datos, en lugar de reconstruirlos cada vez desde cero.
- **Interfaz:** Algunos productos ofrecen un entorno de desarrollo visual integrado que hace que sean más fáciles de usar. Estas herramientas gráficas permiten al usuario definir los procesos de calidad de datos modelados como flujos de trabajo usando una interfaz de *"arrastrar y soltar"*.
- **Repositorio de metadatos:** Repositorio que almacena los esquemas de datos e información, sobre el diseño de los procesos de calidad de datos. Esta información se consume durante la ejecución de los procesos de calidad de datos.
- **Técnicas de rendimiento:** Conjunto de características para acelerar los procesos de limpieza de datos y para garantizar la escalabilidad. Algunas técnicas importantes para mejorar el rendimiento son: la separación, el procesamiento en paralelo, los hilos, el *clustering* y el balanceo de carga.
- **Versiones:** Mecanismo de control de versiones con opciones de control estándar (por ejemplo, el registro de entrada, salida), que permite a los desarrolladores realizar un seguimiento de las diferentes versiones de desarrollo del código fuente.
- **Biblioteca de funciones:** Conjunto de funciones predefinidas, tales como convertidores de tipos de datos y funciones de normalización, que abordan problemas específicos de calidad de datos. Una característica importante es la posibilidad de ampliar la biblioteca de funciones con nuevas funciones. Esto se puede lograr a través de un lenguaje de programación o mediante la adición de funciones externas de una librería de enlace dinámico (DLL, por sus siglas en inglés).
- **Lenguaje vinculante:** Soporte de lenguaje de programación integrado para desarrollar nuevas funciones, con el fin de ampliar la biblioteca de funciones. Este soporte puede variar desde un lenguaje de programación existente (como *C* o *Perl*) a uno nativo.
- **Depuración y rastreo:** Rastreo de gestiones que documentan la ejecución de los programas de calidad de datos con información útil (por ejemplo, tiempos de inicio y final de ejecución de rutinas importantes, el número de registros de entrada y salida).

- **Detección y manejo de excepciones:** Las excepciones son el conjunto de registros de entrada para los cuales falla la ejecución por parte de los procesos de calidad de datos. El manejo de excepciones puede ser manual mediante una interfaz de usuario dada, o automático, al ignorar/borrar registros de excepción, o informar de ellos en un archivo de excepción o tabla.
- **Linaje de datos:** El linaje de datos o procedencia de datos, identifica el conjunto de elementos de datos fuente que produjo un elemento de datos dado (Cui, 2001).

1.6 Limpieza de datos

La existencia de problemas de calidad de datos, comúnmente llamados ***"datos sucios"***, degrada significativamente la calidad de la información, con un impacto directo en la eficiencia de la empresa que maneja dicha información (Barateiro and Galhardas, 2005). Una vez detectados dichos problemas en la fase de análisis de datos (ver epígrafe 1.5), entra en vigor el proceso de limpieza de datos con el objetivo de proporcionarle a los datos una mayor ***"limpieza"*** para un buen desempeño posterior con los mismos.

La limpieza de datos (también conocida como limpiador o fregador de datos) es el acto de detectar, eliminar y/o corregir ***"datos sucios"***, con el objetivo de obtener datos de alta calidad. La limpieza de datos no tiene solamente como objetivo limpiar datos, sino también dar consistencia a diferentes conjuntos de datos que hayan sido combinados a través de bases de datos independientes (Barateiro and Galhardas, 2005).

Este proceso de limpieza es una tarea crucial en diversos escenarios de aplicación. Dentro de una sola fuente de datos (por ejemplo, un listado de clientes), es importante su uso para corregir los problemas de integridad, estandarizar valores, rellenar los datos que faltan y consolidar ocurrencias duplicadas. La construcción de un almacén de datos (DW, por sus siglas en inglés) (Chaudhuri and Dayal, 1997) requiere un importante paso denominado ETL (Extracción, Transformación y Carga), proceso que se encarga de extraer información de las fuentes de datos operacionales, transformar dicha información y posteriormente cargarla en el esquema de datos del almacén. Varios procesos de migración de datos (por ejemplo, cuando se interrumpe un paquete de software) tienen como objetivo convertir los datos legados, almacenados en fuentes con un cierto esquema, en las fuentes de datos destino cuyo esquema es distinto y predefinido (Carreira and Galhardas, 2004). En el área TDQM la limpieza de datos adquiere una relevancia

especial para la comunidad científica y de negocios; en este marco dicho proceso de limpieza es aquel que trata de ***"precisar el grado de corrección en los datos y mejorar su calidad"*** (Fox *et al.*, 1994).

Dependiendo del contexto en el que se aplique la limpieza de datos, este proceso puede ser conocido bajo diferentes nombres; por ejemplo, cuando se detectan y eliminan registros duplicados dentro de un solo archivo, un registro de enlace (área estadística) o el proceso de eliminación de duplicados (área de base de datos) se llevan a cabo. En el contexto de almacenes de datos, los procesos ETL encierran tareas de limpieza de datos, y algunos autores como (Kimbal *et al.*, 1998) designan un almacenamiento específico para los datos, denominado área de organización o preparación de los datos (*data staging area*), en la arquitectura de almacenes de datos, para la recopilación de los resultados intermedios de las transformaciones de limpieza de datos (Barateiro and Galhardas, 2005).

1.6.1 Algoritmos o métodos empleados en la limpieza de datos

Los métodos para llevar a cabo la limpieza de datos están estrechamente vinculados al área en que se aplica y el paso del proceso que se esté realizando. Sin embargo, se destacan algunos métodos generales, como los siguientes:

Para la detección de errores en (Marcus and Maletic, 2005) se señalan como muy usados *los métodos estadísticos* que aunque simples y rápidos, pueden generar muchos falsos positivos; *los métodos de agrupamiento basados en distancias,* cuya principal desventaja radica en la complejidad computacional; *los métodos basados en patrones y en reglas de asociación* que, a partir del análisis de los registros que incumplen los patrones y las reglas descubiertas, detectan posibles errores.

También en (Müller and Freytag, 2003) se describe el *parsing* (Raman and Hellerstein, 2001), *las transformaciones a nivel de esquemas e instancias* (Sattler and Schallehn, 2001), *el reforzamiento de las restricciones de integridad* (Suzanne *et al.*, 2001), *el método de las vecindades ordenadas* (Lee *et al.*, 1999, Hernández and Stolfo, 1998) y otros (Ananthakrishna *et al.*, 2002, Bilenko and Mooney, 2003, Lehti and Fankhauser, 2005, Monge and Elkan, 1997, Arasu *et al.*, 2009) que utilizan diferentes enfoques para realizar la eliminación de duplicados.

1.6.2 Herramientas comúnmente usadas para la limpieza de datos

Varias aplicaciones sofisticadas de software están disponibles para limpiar datos utilizando funciones específicas, normas y tablas de consulta. En el pasado, esta tarea se realizaba manualmente y por lo tanto se encontraba sujeta a errores humanos. El siguiente conjunto de herramientas comerciales y de investigación (ver Tabla 10) aplican técnicas de limpieza de datos (Barateiro and Galhardas, 2005).

Herramienta	**Comercial (C)/Investigación (I)**
DataBlade	C
dfPower	C
ETLQ	C
ETI*DataCleanser	C
Firstlogic	C
NaDIS	C
QuickAddress Batch	C
Sagent	C
Trillium	C
WizRule	C
WizWhy	C
DataFusion	C
Hummingbird ETL	C
Informatica ETL	C
SQL Server 2000 DTS	C
SQL Server 2005	C
SQL Server 2008	C
Sunopsis	C
Centrus Merge/Purge	C
ChoiceMaker	C
DeDupe	C
DoubleTake	C
Identity Search Server	C
MatchMaker	C
Merge/Purge Plus	C
WizSame	C
DataStage	C
PureName PureAddress	C
Integrity	C
Oracle	C
Oracle 10g	C
SSA-Name/Data Clustering Engine	C
d. Centric	C
reUnion and MasterMerge	C
PureIntegrate	C
TwinFinder	C
Data Tools Twins	C

NoDupes	C
DeDuce	C
DataCleaner	C
Pentaho Data Integrator	C
RapidMiner	C
IntelliClean	I
Flamingo Project	I
TranScm	I
Potter´s Wheel	I
Clio	I
Ajax	I
Arktos	I
FraQL	I

Tabla 10 - Herramientas comerciales y de investigación para la implementación de técnicas de limpieza de datos. [Basado en: (Barateiro and Galhardas, 2005)]

Siguiendo la misma idea del perfilado de datos, y tomando en cuenta las herramientas citadas en la Tabla 10, a continuación se muestra un resumen (ver Tabla 11) de los principales rasgos con los que cuentan las herramientas de calidad de datos, analizando diferentes funcionalidades que poseen. Varias de las herramientas no han vuelto a ser incluidas debido a su anterior inclusión en la Tabla 9 por la capacidad de realizar ambos procesos (perfilado y limpieza de datos).

Herramienta	Origen de datos	Capacidad de extracción	Capacidad de carga	Actualizaciones incrementales	Interfaz	Repositorio de metadatos	Técnicas de rendimiento	Versiones	Biblioteca de funciones	Lenguaje vinculante	Depuración	Manejo de errores	Linaje de datos
DataFusion	DB	Y	DB	Y	G	-	Y	Y	Y	N	Y	X	-
Centrus Merge/Purge	DB	-	-	-	G	-	-	-	-	-	-	-	-
DataBlade	Informix	-	Informix	-	G	-	-	-	-	-	Y	X	X
ChoiceMaker	DB,FF	-	-	-	G	Y	Y	-	Y	N	Y	Y	Y
ETI*Data Cleanser	Varias	-	-	-	G	Y	Y	-	Y	Y	-	Y	-
DeDupe	DB	-	-	-	G	-	-	-	-	-	-	-	-
NaDIS	-	X	-	X	G	X	-	-	X	X	-	X	X
QuickAddress Batch	ODBC	X	-	X	G	X	-	-	X	X	-	X	X
DoubleTake	ODBC	-	-	-	G	-	-	-	-	Y	-	-	-

Identity Search Server	DB	-	-	Y	G	Y	-	-	-	-	-	-	-
MatchMaker	DB	-	-	-	G	-	-	-	-	-	Y	-	-
Merge/Purge Plus	-	-	-	-	-	-	-	-	-	-	-	-	-
WizSame	DB,FF	-	-	-	G	-	Y	-	-	-	-	-	-
PureName PureAddress	-	-	-	-	-	-	-	-	-	-	-	-	-
Integrity	-	-	-	-	-	-	-	-	-	-	-	-	-
SSA-Name/Data Clustering Engine	-	-	-	-	-	-	-	-	-	-	-	-	-
d. Centric	-	-	-	-	-	-	-	-	-	-	-	-	-
PureIntegrate	-	-	-	-	-	-	-	-	-	-	-	-	-
TwinFinder	-	-	-	-	-	-	-	-	-	-	-	-	-
Data Tools Twins	-	-	-	-	-	-	-	-	-	-	-	-	-
NoDupes	-	-	-	-	-	-	-	-	-	-	-	-	-
DeDuce	-	-	-	-	-	-	-	-	-	-	-	-	-
IntelliClean	DB	-	DB	-	NG	-	-	-	-	-	Y	X	-
Flamingo Project	DB	-	DB	-	NG	-	-	-	-	-	-	-	-
TranScm	Y	-	-	-	G	-	-	-	Y	N	-	-	-
Clio	DB, XML	X	-	X	G	Y	Y	X	-	-	Y	-	-
Ajax	DB, FF	Y	DB	X	NG	X	Y	X	Y	Java	Y	Y	Y
Arktos	JDBC	-	JDBC	-	G	-	-	X	-	-	Y	Y	-
FraQL	-	-	-	-	NG	-	-	-	Y	N	-	-	-
Pentaho Data Integration	Varias	Y	Y	-	G	-	-	-	-	-	-	-	-
RapidMiner	-	-	-	-	G	-	-	-	-	-	-	-	-

Tabla 11 - Funcionalidades generales de herramientas de calidad de datos comerciales y de investigación usadas en la limpieza de datos. Y: soportado; X: no soportado; -: información desconocida; N: nativo; DB: bases de datos relacionales; FF: archivos planos; G: gráfica; NG: no gráfica; Informix: Información mixta. [Basado en: (Barateiro and Galhardas, 2005)].

1.7 Conclusiones parciales

En este capítulo se muestran los principales enfoques que influyen en el proceso de gestión de la calidad de datos, haciendo énfasis fundamentalmente en las fases para la realización de este proceso bajo la guía de la metodología TDQM. Con el uso de esta guía, es válido mencionar que una adecuada selección de dimensiones de un contexto específico, permite determinar si los datos que se manejan en dicho proceso pueden ser usados para el propósito que se tenga en mente para ellos. Posteriormente el uso de métricas de calidad de datos, para cuantificar la calidad de los datos, el análisis de los mismos para la detección de posibles inconsistencias y la limpieza de estos, vienen a complementar el ciclo con el propósito de obtener datos de alta calidad que influyan de una manera positiva en los vectores estratégicos de toda organización. Este estudio teórico apunta a consolidar los factores principales del proceso de calidad de datos a través del uso de la metodología TDQM, la cual sirve de guía para asegurar la calidad de datos en casos de estudio reales.

CAPÍTULO 2. ASPECTOS FUNDAMENTALES DE *BIG DATA*

El presente capítulo tiene como objetivo realizar un análisis de los principales elementos relacionados con *big data*, viendo para esto diferentes definiciones dadas a esta terminología, así como clasificaciones y características de los mismos, como son las fuentes u origen de los datos, el formato del contenido de los datos, lugar de almacenamiento de los datos, formas de organización de los datos y como son procesados para un uso posterior. Además, se analiza la informática de la nube definiendo para esto sus principales características, la forma en que esta trabaja con *big data* y la relación existente entre ambos. Se investigan las principales fases que componen generalmente toda arquitectura *big data* así como las herramientas, técnicas y tecnologías que son usadas en cada una de las fases, haciendo énfasis en las principales características que hacen de estas, un conjunto útil para el trabajo con *big data*. Finalmente, se propone una arquitectura integradora donde se vincula los términos mencionados anteriormente con el proceso de calidad de datos.

2.1 Definición de *big data*

En los últimos años la manera en la que los usuarios interactúan con la tecnología ha cambiado de manera radical debido a la constante evolución de esta. Revoluciones tecnológicas como la web 2.0, *blogs*, foros de opinión, redes sociales, multimedia, dispositivos móviles, entre otras, facilitan la conectividad y la generación de grandes cantidades de información que hasta hace muy poco eran impensables. La magnitud del fenómeno es tal que los datos generados durante dos días en 2011, por ejemplo, fueron más que los acumulados desde el origen de la civilización hasta principios de 2003 (Lyman and Varian, 2004). Y no solo la sociedad de consumo ha avanzado tecnológicamente; campos como la ciencia, medicina o la economía también requieren cada vez más tratar con grandes cantidades de datos.

Varias discusiones han existido entre la industria y la academia sobre la definición de *big data* (Grobelnik, 2012, Team, 2011), puesto que no existe un concepto único, debido a que con el paso del tiempo nuevas propiedades o características se han ido incorporando al conjunto de definiciones dadas a esta terminología.

El término *big data* es relativamente nuevo en las tecnologías de la información (IT, por sus siglas en inglés) y en los negocios, sin embargo, varios investigadores y profesionales han

utilizado el término en la literatura varios años atrás, por ejemplo, (Cox and Ellsworth, 1997) se refirieron a *big data* como **"***grandes volúmenes de datos científicos para visualizar***"**.

Doug Laney, analista del grupo META (actualmente *Gartner*), en un informe de investigación en el año 2001, definió los desafíos y oportunidades que trae consigo el aumento en los datos con un modelo de tres **"***Vs***"**: volumen (*volume*), velocidad (*velocity*), y variedad (*variety*) (Laney, 2001). Aunque tal modelo no se utilizó originalmente para definir *big data*, Laney y muchas otras empresas, incluidas IBM y algunos departamentos de investigación de *Microsoft*, se mantuvieron utilizando este modelo para describir *big data* dentro de los siguientes diez años (Beyer, 2011).

En 2010, *Apache Hadoop* definió *big data* como ***"****conjuntos de datos que no pueden ser capturados, gestionados y procesados por computadoras dentro de un margen aceptable***"**. Sobre la base de esta definición, en mayo de 2011, *McKinsey & Company*, una agencia global de consultoría, anunció a *big data* como la próxima frontera para la innovación, la competencia y la productividad, definiéndolo como **"***conjuntos de datos que no pueden ser adquiridos, almacenados y gestionados por softwares clásicos de base de datos relacionales***"**. Esta definición incluye dos connotaciones: (i) volúmenes de conjuntos de datos que cumplen con los estándares de *big data*, los cuales se encuentran en continuo cambio, y pueden crecer con el tiempo o con los avances tecnológicos; y (ii) volúmenes de conjuntos de datos que cumplen con los estándares de *big data* en diferentes aplicaciones difiriendo una de otra. En la actualidad, *big data* generalmente oscila entre varios *terabytes* a varios *petabytes* (Manyika *et al.*, 2011). En este mismo año, un informe de la Corporación Internacional de Datos (IDC, por sus siglas en inglés), uno de los líderes más influyentes de *big data* y sus campos de investigación, definió el término como ***"****una nueva generación de tecnologías y arquitecturas, diseñada para extraer económicamente valor de grandes volúmenes de una amplia variedad de datos, permitiendo la captura de alta velocidad, el descubrimiento, y el análisis de dichos datos***"** (Gantz and Reinsel, 2011).

En el año 2012, Gartner (Beyer and Laney, 2012) y el Instituto Nacional de Estándares y Tecnologías (NIST, por sus siglas en inglés) (NIST, 2012) reiteran nuevamente el modelo de tres ***"****Vs***"** dado por Laney (2001) para definir *big data*. Las ***"****Vs***"** pertenecientes a este modelo son totalmente independientes al resto de ***"****Vs***"** que con los años han ido surgiendo, en el cual:

- **Volumen**: siendo quizás la característica que se asocia con mayor frecuencia al término *big data*, el volumen hace referencia a las cantidades masivas de datos que las

organizaciones intentan aprovechar para mejorar la toma de decisiones (Moreno, 2014). Existen varios factores que contribuyen al incremento en el volumen de datos: transacciones de datos a través del paso de los años, constante intercambio de archivos proveniente de las redes sociales, incremento en la cantidad de dispositivos de recopilación de información, son solo algunas de estas (Syed *et al.*, 2013). El volumen, presenta con respecto a las demás ***"Vs"*** el mayor reto para las estructuras convencionales de las tecnologías de la información, debido a que muchas compañías poseen grandes cantidades de datos archivados, pero no cuentan con la capacidad de procesarlos (Syed *et al.*, 2013). Una de las ventaja de reunir grandes cantidades de datos incluye la creación de información y patrones ocultos a través del análisis de datos (Hashem *et al.*, 2014). El aumento en los volúmenes de datos se encuentra en un continuo crecimiento a un ritmo sin precedentes, no obstante, lo que constituye un volumen verdaderamente ***"alto"*** varía en función del sector e incluso de la ubicación geográfica, y es más pequeño que los *petabytes* y *zetabytes* a los que a menudo se hace referencia (Moreno, 2014).

- **Variedad**: la variedad tiene que ver con gestionar la complejidad de múltiples tipos de datos (Moreno, 2014). Estos tipos de datos incluyen vídeo, imagen, texto, audio, y registros de datos, ya sea en formato estructurado, semi-estructurado o no estructurado (Hashem *et al.*, 2014). Las organizaciones necesitan integrar y analizar datos de un complejo abanico de fuentes de información tanto tradicional como no tradicional, procedentes tanto desde dentro como fuera de las empresas (Moreno, 2014). Según algunas estimaciones, el 80 por ciento de los datos de una organización no es numérico, sin embargo, este porcentaje todavía se debe incluir en los análisis y toma de decisiones. Un uso común de gran procesamiento de datos es tomar los datos no estructurados y extraer sus significados ordenadamente, ya sea luego para el consumo de humanos o como entrada estructurada de una aplicación (Syed *et al.*, 2013).
- **Velocidad**: la velocidad se refiere a la rapidez con la que son transferidos los datos (Hashem *et al.*, 2014). Aunque los ciclos de negocio se han acelerado hoy en día, no todos los datos de una organización tienen la misma urgencia de análisis asociada. La clave para entender en qué punto del espectro de la velocidad es necesario trabajar (desde el procesado en lote hasta el flujo de datos continuo) está asociada a los requerimientos de los procesos y los usuarios, ello lleva consigo, la perspectiva de los

datos en movimiento. La velocidad afecta la latencia: el tiempo de espera entre el momento en el que se crean los datos, se captan y están accesibles. Hoy en día, los datos se generan de forma continua a una velocidad sumamente rápida, provocando que a los sistemas tradicionales les resulte imposible captarlos, almacenarlos y analizarlos. Para los procesos en los que el tiempo resulta fundamental, por ejemplo, la detección de fraude en tiempo real o el marketing ***"instantáneo"***, ciertos tipos de datos deben analizarse en tiempo real para que resulten útiles para el negocio (Moreno, 2014).

Las restantes ***"Vs"*** no son usadas para definir *big data*, sino vistas como retos o desafíos para las tres ***"Vs"*** principales, y usadas mayormente en el ámbito de los negocios. Con respecto al surgimiento de la cuarta y quinta ***"V"*** existen diferentes opiniones. Autores como (McAfee, 2012, Syed *et al.*, 2013, Tee, 2013, Moreno, 2014, Saha and Srivastava, 2014, Bahrami and Singhal, 2015) y empresas como IBM (IBM, 2013) opinan que la veracidad (*veracity*) es la cuarta **"***V***"** de *big data*. Otros como (Kaisler *et al.*, 2012, Chen *et al.*, 2014, Hashem *et al.*, 2014, Mayer-Schönberger and Cukier, 2013) y empresas como la IDC (Gantz and Reinsel, 2011) opinan que el valor (*value*) es la cuarta **"***V***"**. Lo cierto es que el orden no es lo importante sino lo que ambas significan:

- **Veracidad**: la veracidad se refiere directamente a inconsistencias y problemas de calidad de datos (Saha and Srivastava, 2014), es decir, a la incertidumbre de los datos (Moreno, 2014). La veracidad hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos. Esforzarse por conseguir datos de alta calidad es un requisito importante y un reto fundamental de *big data*, pero incluso los mejores métodos de limpieza de datos no pueden eliminar la imprevisibilidad inherente de algunos datos relacionados con el tiempo, el costo o las futuras decisiones de compra de un cliente. La necesidad de reconocer y planificar la incertidumbre es una dimensión de *big data* que surge a medida que los directivos intentan comprender mejor el mundo incierto que les rodea (Moreno, 2014).
- **Valor:** en el contexto de *big data*, el valor hacer referencia a los beneficios que se desprenden del uso del mismo, tales como reducción de costos, eficiencia operativa, mejoras de negocio, etc. Esta *V* indica el problema más crítico en *big data*, referente a cómo descubrir valores en grandes conjuntos de datos con varios tipos y rápida generación (Chen *et al.*, 2014, Mayer-Schönberger and Cukier, 2013).

Aunque estas sean las cinco ***"Vs"*** que con mayor frecuencia se mencionan en la literatura, han ido surgiendo otras como la visibilidad (*visibility*) y la variabilidad (*variability*), que en conjunto con las cinco definidas anteriormente son conocidas como las siete ***"Vs"*** de *big data*; incluso otra característica, que a pesar de no representar una "*V*" es tomada en cuanta, denominada complejidad (*complexity*):

- **Visibilidad**: una vez procesados los datos, estos necesitan ser presentados de una manera legible y accesible; es aquí donde la visualización o visibilidad juega su papel fundamental. Las visualizaciones pueden contener decenas de variables y parámetros, muy lejos de las variables *x* e *y* de un gráfico de barras estándar. Encontrar una manera de presentar la información que haga que los resultados sean claros es uno de los desafíos de *big data* (Mcnulty, 2014).
- **Variabilidad**: la variabilidad se refiere a los datos cuyo significado está en constante cambio. Este es particularmente el caso cuando la recolección de datos se basa en el procesamiento del lenguaje (Mcnulty, 2014). Las tecnologías que componen una arquitectura *big data* deben ser flexibles a la hora de adaptarse a nuevos cambios en el formato de los datos, tanto en la obtención como en el almacenamiento, y su procesado. La rápida evolución en la tecnología es considerada como una constante, de manera que los nuevos sistemas deben estar preparados para admitirla (Serrat Morros, 2013).
- **Complejidad**: la complejidad mide el grado de interconexión (posiblemente muy grande) y la interdependencia en estructuras *big data*, de manera tal que un pequeño cambio (o una combinación de pequeños cambios) en uno o pocos elementos de dichas estructuras pueden producir cambios muy grandes, o pequeños cambios que dominen en cascada al sistema y de forma sustancial afecten su comportamiento (Kaisler *et al.*, 2012).

En términos generales *big data* puede ser considerado como una tendencia en el avance de las tecnologías, que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) (Moreno, 2014), con los cuales no pueden realizarse operaciones de almacenamiento, procesamiento, análisis y visualización, utilizando la mayoría de tecnologías tradicionales. *Big data* puede contar con una o múltiples características de las definidas anteriormente (***"Vs"***) (Bahrami and Singhal, 2015), por ejemplo, el almacenamiento y la computación de datos proveniente de medios sociales cuenta con grandes cantidades de datos

(volumen) para ser transferidos en un tiempo de respuesta específico (velocidad), para los cuales puede no interesar las características de variedad ni la veracidad de estos. Otro ejemplo lo podemos encontrar en una empresa que lleve a cabo tareas de *marketing*, para la cual es muy importante que todos sus datos (variedad) posean cierta calidad (veracidad) en vistas a una entrega final de productos de calidad, con los cuales no se incurriría en gastos mayores (valor) por posibles devoluciones de los mismos, elevando así la eficiencia operativa de dicha empresa. En este ejemplo el volumen y la velocidad no son características significativas. La Figura 5 resume las características más usadas en la literatura para referirse al término *big data*, resaltando la ***"V"*** de veracidad (color rojo) por la importancia que representan para esta investigación.

Figura 5 - 7 *"Vs"* de *big data.*

2.2 Categorías para la clasificación de *big data*

Según Hashem et al. (2014), para comprender mejor las características de *big data*, debe ser clasificado en diferentes categorías. La clasificación se basa en cinco aspectos: (i) fuentes de datos, (ii) formato del contenido, (iii) almacenamiento de datos, (iv) organización de datos, y (v) procesamiento de datos (ver Figura 6).

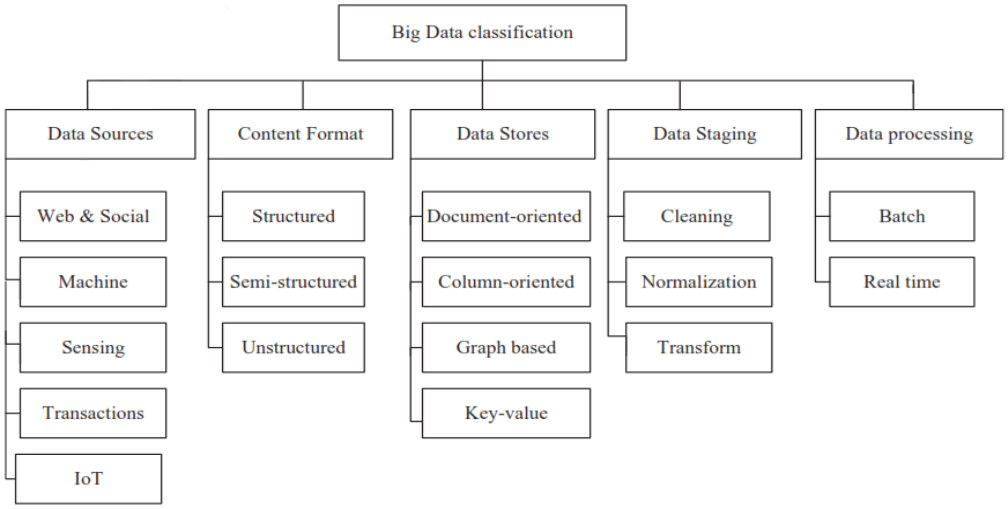


Figura 6 - Clasificaciones de *big data*. [Fuente: (Hashem *et al.*, 2014)]

Cada una de estas categorías tiene sus propias características y complejidades. Rangos de datos desde estructurados a no estructurados son almacenados en varios formatos (Hashem *et al.*, 2014). La categoría más popular es el almacenamiento de datos, la cual presenta un gran número de variedades de base de datos para tal objetivo (Hurwitz, 2013). Como resultado de la amplia variedad de fuentes de datos, los datos capturados difieren en tamaño con respecto a la redundancia, la consistencia y el ruido, etc. (Hashem *et al.*, 2014).

2.2.1 Fuentes de datos (*data source*)

Las fuentes de datos representan las diferentes procedencias de los datos. *Big data* posee una amplia gama de fuentes tales como (ver Figura 6):

- **Web y redes sociales**: los medios sociales son la fuente de información generada a través de la localización de recursos uniforme (URL, por sus siglas en inglés) para compartir o intercambiar información e ideas en comunidades y redes virtuales, tales como proyectos de colaboración, *blogs*, *microblogs*, *Facebook*, *Twitter*, etc. (Hashem *et al.*, 2014).
- **Datos generados por máquinas**: información que se genera de forma automática a partir de hardware o software, procedente de computadoras, dispositivos médicos, u otras máquinas, sin intervención humana (Hashem *et al.*, 2014).
- **Sensores**: dispositivos de detección que existen para medir cantidades físicas y convertirlas en señales (Hashem *et al.*, 2014).

- **Transacciones**: las transacciones de datos, como datos financieros y de trabajo, comprenden un evento que involucra una dimensión de tiempo para describir los datos (Hashem *et al.*, 2014).
- **IoT**: representa un conjunto de objetos que son únicamente identificados como parte de internet. Estos objetos incluyen teléfonos inteligentes, cámaras digitales, tabletas, entre otros. Cuando estos dispositivos se conectan entre sí a través de internet, permiten que más procesos inteligentes y servicios soporten necesidades básicas, económicas, de salud, etc. Un gran número de dispositivos conectados a internet ofrecen muchos tipos de servicios y produce enormes cantidades de datos e información (Rao *et al.*, 2012).

2.2.2 Formato del contenido (*content format*)

El formato del contenido indica los diferentes tipos en que pueden estar representados los datos. *Big data* trabaja con los siguientes tipos de datos (ver Figura 6):

- **Estructurados**: datos que a menudo son manejados con SQL, lenguaje de programación creado para la gestión y consulta de datos en RDBMS. Los datos estructurados son fáciles de entrar, consultar, almacenar y analizar en los RDBMS. Ejemplos de datos estructurados incluyen números, palabras y fechas (Hashem *et al.*, 2014).
- **Semi-estructurados**: datos que no siguen un sistema de base de datos convencional. Los datos semi-estructurados pueden estar en forma de datos estructurados que no se encuentran organizados en modelos de bases de datos relacionales, tales como tablas. La captura de datos semi-estructurados para el análisis es diferente de la captura de un formato de archivo fijo. Por lo tanto, la captura de datos semi-estructurados requiere el uso de normas complejas que permiten decidir dinámicamente el proceso siguiente después de la captura (Franks, 2012).
- **No estructurados**: los datos no estructurados, tales como mensajes de texto, información de ubicación, contenido multimedia y datos de redes sociales, son datos que no siguen un formato específico. Teniendo en cuenta que el tamaño de este tipo de datos se encuentra en continuo aumento debido al uso de teléfonos inteligentes, la necesidad de analizar y comprender estos datos se ha convertido en un desafío (Hashem *et al.*, 2014).

2.2.3 Almacenamiento de datos (*data stores*)

El almacenamiento de datos incluye todos aquellos lugares donde pueden ser almacenados los datos para su utilización. Dichos lugares indican diferentes bases de datos o sistemas de ficheros distribuidos para realizar esta tarea en *big data* (ver Figura 6):

- **Orientado a documentos**: el almacenamiento de datos orientado a documentos está diseñado principalmente para almacenar y recuperar colecciones de documentos, y apoyar a las formas complejas de datos en varios formatos estándar, como JSON, XML, y binario (por ejemplo, PDF y MS Word). Un almacenamiento de datos orientado a documentos es similar a un registro o fila de una base de datos relacional, pero más flexible y puede recuperar documentos según su contenido (por ejemplo, *MongoDB*, *SimpleDB* y *CouchDB*) (Hashem *et al.*, 2014).
- **Orientado a columnas**: el almacenamiento de base de datos orientado a columnas almacena su contenido en columnas además de las filas, con valores de atributos que pertenecen a las mismas columnas almacenadas de forma contigua. Este almacenamiento es diferente al de los sistemas tradicionales de bases de datos que almacenan filas enteras una tras otra (Abadi *et al.*, 2009), tales como *BigTable* (Chang et al., 2008).
- **Basado en grafos**: una base de datos basada en grafos, como *Neo4j*, está diseñada para almacenar y representar datos que utilizan un modelo de grafo con nodos, aristas y propiedades relacionadas entre sí a través de las relaciones (Neubauer, 2010).
- **Llave-valor**: el almacenamiento de datos basado en llave-valor está diseñado para almacenar y acceder a los datos, cuyos elementos se encuentran separados en tuplas (pares llave-valor), y diseñado para trabajar con grandes cantidades de datos (Seeger and Ultra-Large-Sites, 2009). *Dynamo* es un buen ejemplo de un sistema de almacenamiento llave-valor de alta disponibilidad, el cual es utilizado por *Amazon* en algunos de sus servicios (DeCandia *et al.*, 2007). Otros ejemplos de almacenamiento llave-valor son *Apache Hbase*, *Apache Cassandra*, y *Voldemort*. *Hbase* almacena datos en tablas, filas y celdas. Las filas son ordenadas por llaves de filas, y cada celda de una tabla es especificada mediante una llave de fila, una llave de columna, y una versión, con el contenido incluido como un arreglo de bytes no-interpretado (Hashem *et al.*, 2014).

2.2.4 Organización de los datos (*data staging*)

La organización de los datos incluye diferentes preprocesamientos que son aplicados a los datos una vez estos son almacenados (ver Figura 6):

- **Limpieza**: proceso de identificación de datos incompletos e incorrectos en general (Rahm and Do, 2000).
- **Transformación**: proceso de transformación de datos en una forma adecuada para su análisis (Hashem *et al.*, 2014).
- **Normalización**: proceso de estructuración del esquema de base de datos para minimizar la redundancia (Quackenbush, 2002).

2.2.5 Procesamiento de los datos (*data processing*)

El procesamiento de los datos incluye diferentes procesos que son aplicados a los datos para la búsqueda de resultados (ver Figura 6):

- ***Batch* o Lotes**: sistemas basados en *MapReduce* los cuales han sido adoptados por muchas organizaciones en los últimos años para la corrida de trabajos por lotes (Chen *et al.*, 2012). Dichos sistemas permiten el escalado de aplicaciones a través de grandes grupos de máquinas comprendidas en miles de nodos (Hashem *et al.*, 2014).
- **Tiempo real**: una de las herramientas *big data* más famosa y poderosa basada en procesos en tiempo real es el sistema de *streaming* escalable *S4* (Neumeyer *et al.*, 2010). *S4* es una plataforma de computación distribuida que permite a los programadores desarrollar convenientemente aplicaciones para el procesamiento de flujos continuos de datos ilimitados. *S4* es una plataforma parcialmente tolerante a fallos, escalable y de propósito general (Hashem *et al.*, 2014).

2.3 La nube informática

Varias soluciones tradicionales han surgido para hacer frente a *big data* como la supercomputación, computación distribuida, computación paralela, entre otras. Sin embargo, la escalabilidad es importante en *big data*, la cual puede ser soportada por servicios en la nube. La computación en la nube tiene varias capacidades para el tratamiento de grandes cantidades de datos que están relacionadas con el manejo de *big data*, además de soportar dos grandes cuestiones de *big data*, el almacenamiento y el tratamiento (computación) de estos. La nube

informática ofrece un conjunto de recursos (almacenamiento y tratamiento) que pueden añadirse en cualquier momento para el manejo de *big data;* estas características permiten a la nube convertirse en una tecnología emergente para hacer frente a dichos datos (Bahrami and Singhal, 2015).

2.3.1 Definición y principales características de la nube informática

La computación en la nube es un modelo que permite, convenientemente, el acceso ubicuo a la red bajo demanda de un conjunto compartido de recursos informáticos configurables (por ejemplo, redes, servidores, almacenamiento, aplicaciones y servicios) que pueden ser rápidamente distribuidos y liberados con un mínimo esfuerzo de gestión o una interacción proveedora de servicios (Mell and Grance, 2011).

Las principales características de la computación en la nube fueron definidas por NIST (Liu *et al.*, 2011). A continuación se muestra un resumen de estas (Bahrami and Singhal, 2015):

- **Servicios elásticos en demanda**: esta característica muestra los siguientes puntos: (i) un modelo económico de computación en la nube que permite a los consumidores solicitar servicios necesarios (máquinas de cómputo y/o dispositivos de almacenamiento). La solicitud requiere que los servicios prestados puedan ser escalables rápidamente hacia arriba o abajo en la demanda; (ii) esto es responsabilidad de las máquinas las cuales no requieren ningún humano para controlar los servicios solicitados. La arquitectura de la nube gestiona las solicitudes en demanda (aumento o disminución de solicitudes de servicios), la disponibilidad, la asignación, suscripción y facturas del cliente. Esta característica es interesante para las empresas de nueva creación, ya que permite a una empresa iniciar con datos tradicionales o bases de datos tradicionales (en particular, la puesta en marcha de negocios) y aumentar sus conjuntos de datos a *big data* a medida que reciben las peticiones de los clientes o sus datos crecen durante la marcha de los negocios.
- **Conjunto de recursos**: un proveedor de la nube proporciona una reserva de recursos (por ejemplo, máquinas de computación, dispositivos de almacenamiento y de red) a los clientes. La arquitectura de la nube gestiona todos los recursos disponibles a través de los gerentes globales y locales para diferentes sitios y localidades, respectivamente. Esta característica permite que grandes cantidades de datos (*big data)* puedan ser distribuidos

en diferentes servidores, lo cual no es posible realizar en modelos tradicionales, como los sistemas de supercomputación.

- **Accesibilidad de servicios**: un proveedor de la nube ofrece todos los servicios a través de redes de banda ancha (a menudo a través de internet). Los servicios que ofrece están disponibles a través de un modelo basado en la web o aplicaciones de clientes heterogéneos (Singhal, 2013). El modelo basado en la *web* podría ser una interfaz de programación de aplicaciones (API), servicios *web* como *Web Service Description Language* (WSDL), etc. Por otro lado, las aplicaciones de clientes heterogéneos son proporcionadas por los proveedores. Los clientes pueden ejecutar aplicaciones en los sistemas clientes heterogéneos, como *Windows*, *Android* y *Linux*. Esta característica permite a los socios de las empresas contribuir con *big data*. Dichos socios pueden proporcionar aplicaciones de *software* en la nube, infraestructura o datos. Por ejemplo, varias aplicaciones de diferentes sitios pueden conectarse a un dato individual o a almacenes de datos múltiples para capturar, analizar o procesar datos *big data*.
- **Medición de servicios**: los proveedores de la nube cobran a sus clientes por una capacidad de medición proporcionada por la facturación para un abonado, basada en el modelo de pago por uso. Esta característica gestiona todos los precios del servicio en la nube, las suscripciones y la medición de los servicios utilizados. Esta capacidad permite a una organización pagar por el tamaño actual de los conjuntos de datos, y luego pagar más cuando el tamaño de dicho conjunto aumenta. Este servicio permite a los clientes comenzar con una baja inversión.

2.3.2 Arquitectura de la nube

La tecnología de la computación en la nube puede ser proporcionada por un proveedor, la cual permite a los departamentos de tecnología de información centrarse en sus desarrollos de *software*, en lugar de preocuparse por el mantenimiento de *hardware*, seguridad, recuperación, sistemas operativos y actualizaciones de *software*; además, si un departamento establece un sistema de computación en la nube en su organización, este puede ayudarles a manejar grandes volúmenes de datos (*big data*) (Bahrami and Singhal, 2015).

La arquitectura de un sistema de computación en la nube es específica del sistema y los requisitos de cada componente y sub-componentes en general. La arquitectura de la nube permite

a sus proveedores analizar, diseñar, desarrollar e implementar datos *big data*. Dichos proveedores ofrecen servicios a través de diferentes capas en los sistemas de computación en la nube. Las principales categorías se dividen en cuatro capas de servicio: Infraestructura como un Servicio, Plataforma como un Servicio, *Software* como un Servicio e Inteligencia de Negocios (IaaS, PaaS, SaaS, y BI, por sus siglas en inglés respectivamente) y otras capas de servicios asignadas a las capas principales, como se muestra en la Figura 7, tales como Datos como un Servicio (DaaS, por sus siglas en inglés) asignado a la capa de IaaS (Bahrami and Singhal, 2015).

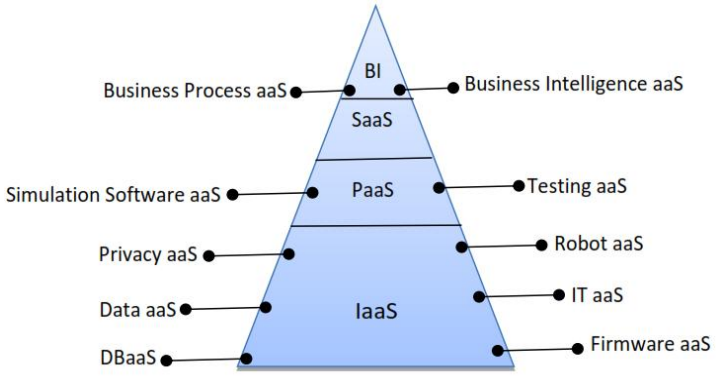


Figura 7 - Servicios en la nube. [Fuente: (Bahrami and Singhal, 2015)]

1) Capas de servicios principales

La Tabla 12 muestra una breve descripción de las cuatro capas de servicios principales que conforman la arquitectura de la nube informática.

Nombre del servicio	**Descripción**	**Rol del servicio en *big data***
Infraestructura como un Servicio (IaaS)	Cubre varios servicios como el *firmware*, *hardware*, utilidades, datos, bases de datos, recursos e infraestructura. Permite a los clientes instalar sistemas operativos, recibir infraestructura presupuestada, y desarrollar aplicaciones de *software* necesarias	Almacenamiento de *big data*. Acceso a recursos de *hardware* para *big data*
Plataforma como un Servicio (PaaS)	Ofrece una aplicación de *software* para la entrega de productos en la nube. Permite al desarrollador centrarse en la creación de aplicaciones de *software*, sin tener que preocuparse por el mantenimiento del sistema operativo. Ofrece servicios a los programadores de *software* para el	Manejo, tratamiento y análisis de *big data*

	desarrollo y despliegue de sus aplicaciones con una abstracción en la capa de *hardware*	
Software como un Servicio (SaaS)	Proporciona aplicaciones en la nube a través de la red y no requiere que los clientes instalen aplicaciones en sus equipos locales	Captura de *big data*
Inteligencia de Negocios (BI)	Proporcionar modelos analíticos necesarios para clientes de la nube	Análisis de *big data*

Tabla 12 - Capas de servicios principales en la arquitectura de la nube.

2) Otras capas de servicios

La Tabla 13 muestra otras capas de servicios, las cuales son relacionadas con las capas principales de servicios de la nube informática analizadas en la tabla anterior.

Nombre del servicio	**Relación con**	**Descripción**	**Rol del servicio en *big data***
Proceso de Negocios como un Servicio (BPaaS)	BI	Ofrece soporte de herramientas automatizado	Análisis de *big data*
Inteligencia de Negocios como un Servicio (BIaaS)	BI	Proporciona enfoques integrados para soportes de mantenimiento	Análisis de *big data*
Simulación de *Software* como un Servicio (SimSaaS)	SaaS	Ofrece la simulación de servicios con un modelo de configuración MTA	Análisis de *big data*
Pruebas como un Servicio (TaaS)	SaaS	Ofrece entornos de pruebas de *software*	Prueba de herramientas *big data*
Robot como un Servicio (RaaS)	PaaS	Proporciona servicios de computación orientados a la robótica	Funcionamiento de *big data*
Privacidad como un Servicio (PaaS)	PaaS	Ofrece un marco de trabajo para la preservación de la privacidad de datos junto con una vista de aplicaciones prácticas	Privacidad en *big data*
Tecnología de la Información como un Servicio (ITaaS)	IaaS	Permite la adquisición de recursos para departamentos de IT	Mantenimiento de *big data*
Hardware como un Servicio (HaaS)	IaaS	Ofrece la integración transparente de *hardware* remoto distribuido a través de múltiples ubicaciones geográficas dentro de un sistema operativo.	Captura y mantenimiento de *big data*

Base de Datos como un Servicio (DBaaS)	IaaS	Ofrece: (1) Un enfoque de cargas de trabajo para contrataciones múltiples (2) Un algoritmo de partición de datos basado en grafos (3) Un esquema de seguridad ajustable	Almacenamiento de *big data*
Datos como un Servicio (Daas)	IaaS	Permite el análisis de las principales preocupaciones para los datos como un servicio	Almacenamiento de *big data*
Big Data como un Servicio (BDaaS)	Todas la capas	Generación de servicios para *big data*	Generar *big data*

Tabla 13 - Otras capas de servicios en la arquitectura de la nube. [Fuente: (Bahrami and Singhal, 2015)]

2.3.3 Relación entre *big data* y los servicios en la nube

Big Data proporciona a los usuarios la posibilidad de utilizar cómodamente la computación para procesar consultas distribuidas a través de múltiples conjuntos de datos y devolver conjuntos resultantes de una manera oportuna. La computación en la nube proporciona el motor subyacente para llevar a cabo dicha tarea mediante el uso de plataformas de procesamiento de datos distribuidos como *Hadoop* (Hashem *et al.*, 2014).

El uso de la computación en la nube en *big data* es mostrado en la Figura 8. Primeramente varias fuentes de datos de gran tamaño en la nube y la *web* se almacenan en una base de datos distribuida con tolerancia a fallos y se procesan a través de un modelo de programación para grandes conjuntos de datos con un algoritmo distribuido en paralelo en un clúster. Finalmente la visualización de datos permite ver los resultados analíticos presentados a través de diferentes gráficos para la toma de decisiones (Hashem *et al.*, 2014).

Big data utiliza la tecnología de almacenamiento distribuido basado en la nube informática, en lugar de utilizar un almacenamiento local de un ordenador o dispositivo electrónico. La evaluación de *big data* es impulsada por aplicaciones de rápido crecimiento basadas en la nube, desarrolladas utilizando tecnologías virtualizadas; por tanto, la computación en la nube no sólo ofrece facilidades para el tratamiento y procesamiento de *big data*, sino también sirve como un modelo de servicio (Hashem *et al.*, 2014).

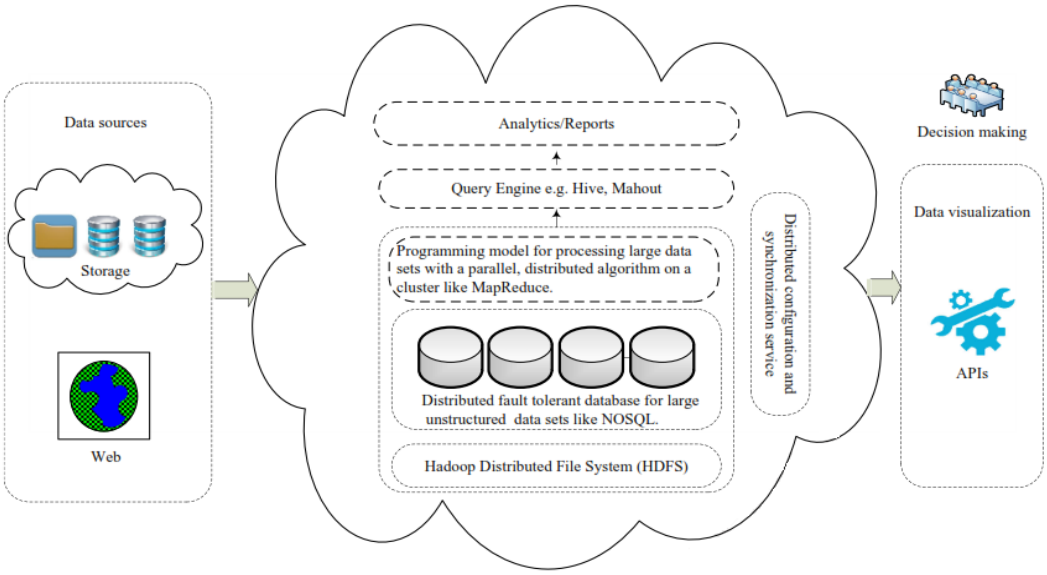


Figura 8 - Uso de la nube informática en *big data*. [Fuente: (Hashem *et al.*, 2014)]

Talia (2013), discutió la complejidad y la variedad de tipos de datos y su capacidad de procesamiento para realizar análisis de grandes conjuntos de datos; señalando que la infraestructura de la nube informática puede servir como una plataforma eficaz para abordar el almacenamiento de datos (necesario para realizar el análisis de *big data*). La computación en la nube está correlacionada con un nuevo modelo para la provisión de infraestructura informática y con un método de procesamiento de datos para todos los tipos de recursos disponibles en la nube a través del análisis de datos. Varias tecnologías basadas en la nube tienen que hacer frente a este nuevo entorno, pues lidiar con *big data* para el procesamiento simultáneo se ha convertido cada vez más complicado (Ji *et al.*, 2012).

MapReduce es un buen ejemplo para el procesamiento de *big data* en un entorno de la nube (Dean and Ghemawat, 2008); permitiendo el procesamiento de grandes cantidades de datos almacenados en paralelo en un clúster de computación, el cual exhibe un buen rendimiento en entornos de sistemas distribuidos, como la potencia de los ordenadores, el almacenamiento y las redes de comunicación. Bollier and Firestone (2010) hicieron hincapié en la capacidad de computación en clústeres para proporcionar un contexto hospitalario para el crecimiento de datos. Sin embargo, (Miller, 2013) sostuvo que la falta de disponibilidad de datos es costosa, ya que los usuarios descargan más decisiones a los métodos de análisis; donde un uso incorrecto de estos métodos o debilidades inherentes en los mismos puede producir decisiones equivocadas y costosas. Los RDBMS son considerados una parte de la arquitectura de la nube informática

actual y juegan un papel importante para garantizar la transición fácil de aplicaciones, desde las infraestructuras empresariales antiguas a las nuevas arquitecturas de infraestructura de la nube. La presión para que las organizaciones adopten rápidamente e implementen tecnologías, como la computación en la nube, para abordar el reto que constituye el trabajo con *big data* y el procesamiento de demandas, conlleva riesgos inesperados y consecuencias (Hashem *et al.*, 2014).

2.4 Fases de *big data* y principales herramientas, técnicas y tecnologías utilizadas en estas

Las principales fases que generalmente componen una arquitectura *big data* son: recolección o captura de datos, almacenamiento y distribución de datos, procesamiento de datos, análisis de datos y visualización de datos. Estas fases no son nuevas, ya que son algo generalizadas en las soluciones de inteligencia de negocios que existen hoy en día. Sin embargo, debido a nuevas necesidades y en función del grado de detalle que se dé a la hora de solucionar un determinado problema *big data*, cada una de estas fases ha ido adaptándose y aportando nuevas tecnologías, al igual que han abierto nuevas oportunidades para el trabajo y uso de los datos de este tipo (Morros and Picañol, 2013).

En la Figura 9 se puede observar el flujo de cierta información en una arquitectura *big data*, con orígenes de datos diversos, por ejemplo, bases de datos, documentos o datos recibidos en *streaming;* los cuales son capturados y almacenados a través de las capas de recolección y almacenamiento de datos, con herramientas específicamente desarrolladas para dichas funciones. Los datos recibidos pueden almacenarse, procesarse, analizarse y visualizarse tantas veces como haga falta y lo requiera el caso de uso específico (Morros and Picañol, 2013).

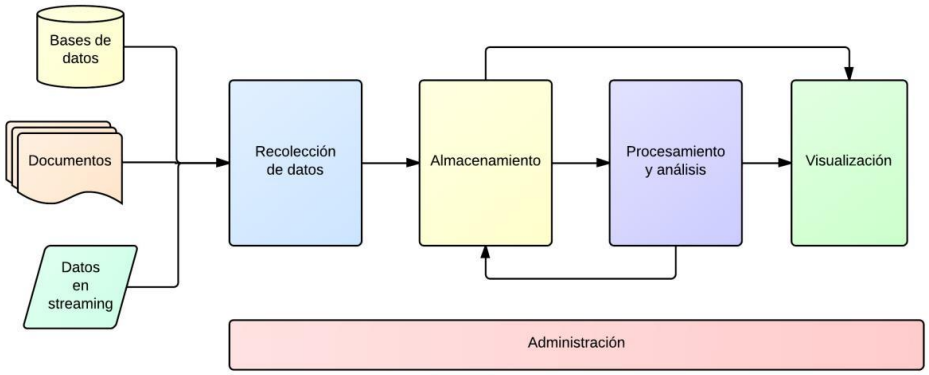


Figura 9 - Arquitectura *big data* por fases. [Fuente: (Morros and Picañol, 2013)]

2.4.1 Recolección de datos

En esta etapa el sistema debe conectarse a las fuentes de información y extraerlas. Las herramientas de recolección de datos pueden dividirse en dos grupos, dependiendo de cómo se conecten al origen de los datos (Morros and Picañol, 2013).

1) *Batch* o por lotes

Se conectan de manera periódica a las fuentes de datos buscando nueva información. Generalmente se usan para conectarse a sistemas de ficheros o bases de datos, buscando cambios desde la última vez que se conectaron. Una herramienta para migrar datos periódicamente, dígase: una vez al día, desde una base de datos a otra es un ejemplo de recolección de datos por lotes (Morros and Picañol, 2013).

2) *Streaming* o por transmisión en tiempo real

Están conectadas de manera continua a las fuentes de datos, descargando información cada vez que éstas transmiten. Se acostumbran a usarse para monitorización de sistemas (para aumentar la seguridad y la detección de fallos), conjuntos de sensores o para conectarse a redes sociales y descargar información en tiempo real (Morros and Picañol, 2013).

Las herramientas para la recolección de datos son las que más han evolucionado gracias a la aparición de *Hadoop* y a la popularización de sistemas de almacenamiento NoSQL; todo ello fruto de la necesidad de tratar datos no estructurados. Una de las ventajas que ofrecen las herramientas de recolección de datos es la flexibilidad que tienen; tanto a la hora de configurarse y adaptarse a distintos orígenes y destinos de datos, como para trabajar independientemente del sistema donde estén montadas (es decir, no necesitan *Hadoop* de manera imperiosa). En función del tipo y origen de datos se encuentran varias herramientas que forman parte de esta fase; por ejemplo, para el caso de datos no estructurados, como pueden ser ficheros *logs*, las herramientas más utilizadas son *Flume* y *Chukwa*. Por otro lado, para la captura de datos provenientes de una base de datos relacional *Sqoop* es la herramienta más utilizada. Todas estas herramientas forman parte del ecosistema *Hadoop*. Otra de las herramientas interesantes a mencionar en esta primera fase, aunque no se encuentre dentro del ecosistema *Hadoop* es *Storm*, el cual es un sistema de procesamiento de eventos en *streaming* que permite manejar datos en tiempo real. Actualmente las herramientas han evolucionado de manera que muchas de ellas ya pueden usarse de ambas

formas. En esta etapa, los datos pueden sufrir algún tipo de proceso o cambio si la aplicación así lo requiere, por ejemplo, el filtrado de información no deseada o el formato con el que se guardarán finalmente en el sistema de almacenamiento (Morros and Picañol, 2013).

2.4.2 Almacenamiento y distribución

La fase de almacenamiento tiene, a grandes rasgos, dos elementos básicos: sistemas de ficheros y bases de datos. Los sistemas de tratamiento de la información se centran principalmente en las bases de datos, pero debido a que en los sistemas *big data* se busca la mayor variedad posible, las bases de datos tradicionales acostumbran a ser poco flexibles, por lo que los sistemas de ficheros han cobrado mayor importancia (Morros and Picañol, 2013).

1) Sistemas de ficheros y sistemas de ficheros distribuidos

Los sistemas de ficheros son una parte fundamental en una arquitectura *big data*, ya que varias herramientas están construidas sobre ellos. Además, el hecho de trabajar con datos no estructurados los hace aún más importantes ya que son el medio principal para trabajar con este tipo de información. Adicionalmente, un objetivo que buscan los sistemas *big data* es la escalabilidad (*scalability*), es decir, un sistema que pueda variar su tamaño (ya sea aumentándolo o disminuyéndolo) según las necesidades y de manera que no afecte el rendimiento general de todo el sistema. Esta necesidad fue la que motivó la aparición de los sistemas de ficheros distribuidos, que consisten en una red o clúster de ordenadores (o nodos) interconectados entre sí y configurados para tener un sólo sistema de ficheros lógico. Un ejemplo de estos ficheros y uno de los más usados es el sistema de ficheros distribuidos de *Hadoop* (HDFS, por sus siglas en inglés), el cual está diseñado especialmente para ejecutarse en *hardware* asequible o de bajo costo, y para ser tolerante a fallos (Morros and Picañol, 2013).

En la Figura 10 se puede observar un ejemplo simplificado del funcionamiento de un sistema de ficheros distribuido, en el cual se tiene un directorio con cuatro ficheros (FicheroA, FicheroB, FicheroC y FicheroD) que el usuario, al conectarse al sistema y entrar en el directorio donde estos se encuentran, visualiza los ficheros como si estuvieran todos almacenados en un mismo ordenador (sistema lógico). La realidad es que cada fichero está físicamente almacenado en un nodo u ordenador distinto a los demás (sistema físico) (Morros and Picañol, 2013).

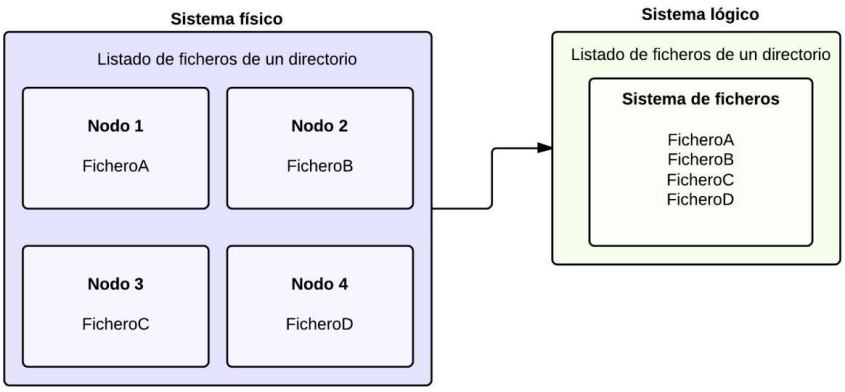


Figura 10 - Funcionamiento de un sistema de ficheros distribuido. [Fuente: (Morros and Picañol, 2013)]

2) Bases de datos

Las bases de datos continúan estando presentes en los sistemas de explotación de la información, especialmente las bases de datos relacionales, establecidas como un nuevo paradigma en la década de los años 70 y consolidadas gracias a la facilidad de conceptualizar los problemas. Estas bases de datos siguen el modelo relacional establecido por Peter Chen en el año 1976, que permite interconectar los datos almacenados en diferentes tablas, mediante relaciones entre los mismos. Con la aparición de las bases de datos relacionales también surgió el lenguaje de consultas estructurado (SQL, por sus siglas en inglés) como una especificación de lenguaje para trabajar con estas bases de datos. El lenguaje SQL permite realizar consultas muy sencillas, similares al lenguaje humano, lo que lo hace accesible a usuarios no expertos. Además, se nutre de características del álgebra y el cálculo relacional para recuperar de forma sencilla información de interés (Morros and Picañol, 2013).

A pesar de ser sistemas muy fáciles de usar y rápidos en la ejecución de consultas (gracias a la creación de índices), los RDBMS tradicionales tienen ciertos impedimentos que hacen que estén empezando a perder rendimiento en problemas *big data;* puesto que cuando la información almacenada supera varios límites (normalmente alrededor de *terabytes*), mantener la información estructurada tiene un coste en la creación y mantenimiento de índices y en el rendimiento de las consultas. Además, son bases de datos poco flexibles ya que cuando se crea su estructura es bastante conflictivo realizar cambios en esta (como añadir nuevas columnas a una tabla o cambiar el tipo de una columna) (Morros and Picañol, 2013).

Con la aparición de los primeros problemas *big data*: empresas que basan su actividad en internet y redes sociales; ha aumentado la popularidad de las llamadas bases de datos NoSQL

(*Not-only SQL*). Las principales compañías que promueven estas bases de datos son *Amazon*, *Google*, *Facebook* y *Twitter*. Estas bases de datos no siguen el modelo relacional, por tanto no usan el lenguaje SQL. Las mismas aportan más flexibilidad en el trabajo con *big data* al no requerir estructuras fijas como las tablas. Otra ventaja de estos sistemas es que responden a las necesidades de escalabilidad, ya que al no tener que mantener índices para los datos, el volumen de información que almacenan siempre crece de forma horizontal (en bases de datos SQL el mantenimiento de índices hace que el volumen de información crezca de manera exponencial al añadir nuevos datos). Algunos ejemplos de sistemas NoSQL son *MongoDB* (orientado a documentos y basado en ficheros JSON o BSON), *Riak* (basado en el modelo llave-valor), *eXist* (basado en ficheros XML), *BigTable* de *Google* (orientado a columnas), *Dynamo* de *Amazon* (orientado a columnas), *Cassandra* (orientado a columnas), etc. (Morros and Picañol, 2013).

Independientemente del paradigma de base de datos, cada vez es más frecuente que los sistemas se adapten para funcionar con los sistemas distribuidos, obteniendo una mayor escalabilidad. Los sistemas NoSQL acostumbran a ser en este sentido más adaptables a los sistemas distribuidos, permitiendo una mayor flexibilidad en la configuración de máquinas (con hardware más sencillo en lo relativo a prestaciones); mientras que los sistemas SQL requieren infraestructuras más especializadas (y a la vez más costosas) para trabajar con los SGBDR tradicionales. En la Tabla 14 se presenta una pequeña lista con las ventajas y desventajas de los sistemas SQL y NoSQL (desde el punto de vista *big data*), reflejando la aplicabilidad de cada modelo en cada caso. El modelo relacional con SQL es más adecuado en casos con menos volumen de información, fáciles de conceptualizar y con la necesidad de obtener un tiempo de respuesta reducido. Por su parte, los sistemas NoSQL son más adecuados en ocasiones donde se necesita tener un volumen de datos mayor, una escalabilidad horizontal y más flexibilidad y variedad en los tipos de datos (Morros and Picañol, 2013).

Sistemas	**Ventajas**	**Desventajas**
	Velocidad	Escalabilidad no horizontal
SQL	Facilidad de uso	Requiere de hardware especificado
	Mayor volumen	Más lento
	Más variabilidad	Más complejo
NoSQL	Flexible en la configuración del hardware	
	Más variedad	

Tabla 14 - Ventajas y desventajas entre sistemas SQL y NoSQL. [Fuente: (Morros and Picañol, 2013)]

Una forma más eficiente de realizar las funcionalidades de esta fase radica en intentar aprovechar lo mejor de los dos paradigmas. En estos casos se crea un sistema de almacenamiento (ya sea un sistema de ficheros distribuido o una base de datos NoSQL) para almacenar la información no estructurada en grandes volúmenes de datos y, posteriormente, se almacenan los resultados de los procesos y análisis realizados sobre estos datos en un sistema SQL, obteniendo una mayor velocidad de respuesta al consultar los resultados (Morros and Picañol, 2013).

2.4.3 Procesamiento y Análisis

Una vez que se tienen los datos almacenados, los siguientes pasos en una arquitectura *big data* consisten en explotar la información (procesarla y analizarla) para llegar a los resultados deseados. Las herramientas de análisis y procesamiento de información han evolucionado considerablemente, especialmente aquellas que trabajan sobre datos no estructurados. La necesidad de crear nuevas aplicaciones y que éstas estén adaptadas a los sistemas de almacenamiento más recientes (NoSQL y sistemas de ficheros distribuidos) ha promovido la creación de nuevos paradigmas de programación para el procesamiento de datos que intentan ofrecer un acercamiento a una solución para *big data*. Estos paradigmas han terminado caracterizando las arquitecturas *big data*, adaptando el resto de las fases para funcionar de forma óptima. Los dos paradigmas que se centran en el desarrollo de aplicaciones para el procesamiento de datos *big data* son *MapReduce* y las llamadas *Massive Parallel Procesing* (MPP), ambas con aspectos en común pero bien diferenciadas. En el caso del análisis de datos las herramientas desarrolladas permiten el manejo, tratamiento, administración y la creación de consultas en grandes volúmenes de datos distribuidos en forma de tablas relacionales. Como ejemplos de estas herramientas se pueden mencionar los motores de consultas *Hive* (para almacenes de datos) y *Pig* (para paralelizar flujos de datos); *Mahout* (para realizar *clustering*, algoritmos de regresión o implementar modelos estadísticos sobre los datos de salida ya procesados), *Oozie* (para la planificación de flujos de trabajo), *Datameer* (integración, almacenamiento, análisis y visualización), *Cascading* (para crear y ejecutar datos complejos que procesan flujos de trabajos de clústeres de *Hadoop*), entre otras (Morros and Picañol, 2013).

2.4.4 Visualización

La fase de visualización es la que menos ha cambiado a la hora de solucionar un determinado problema *big data*. En esta fase los resultados a visualizar del procesamiento y análisis se llevan

a cabo a través del uso de herramientas que permiten la visualización de datos como *Hue* (desarrollada por *Cloudera*), *Pentaho Big Data Analytics*, *WYSIWYG Infographic Designer*, entre otras. En la Figura 11 se observan varias imágenes obtenidas a partir de la herramienta de visualización de la suite Pentaho (Morros and Picañol, 2013).

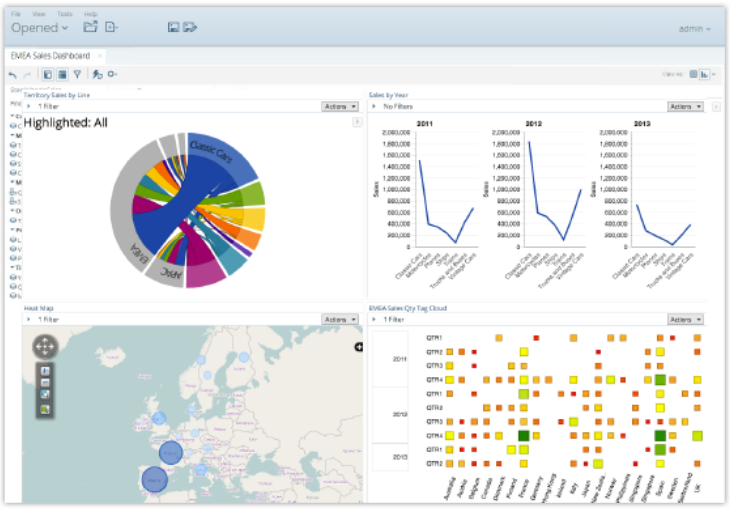


Figura 11 - Herramienta de visualización de Pentaho. [Fuente: (Morros and Picañol, 2013)]

Existe otro conjunto variado de herramientas que no son clasificadas en ninguna de las fases de una arquitectura *big data*, sin embargo, también son consideradas herramientas esenciales, ya que ofrecen soporte y funcionalidades a los desarrolladores de aplicaciones *Hadoop* en cualquiera de las fases definidas, y están pensadas principalmente para la comunicación y sincronización de procesos. Como ejemplos de estas herramientas se pueden mencionar *Avro* (para la serialización de datos), *Zookeeper* (gestionar y administrar coordinaciones entre procesos en sistemas distribuidos) y *Apache Solr* (motor de búsqueda) (Morros and Picañol, 2013).

2.5 Arquitectura integradora (*big data*, nube informática y proceso de calidad de datos)

La Figura 12, tiene como objetivo principal representar la ubicación más adecuada del proceso de calidad de datos (basado en la metodología TDQM descrita en el capítulo 1), en una arquitectura que incluya el flujo de datos en *big data* (ver epígrafe 2.4) y su tratamiento en la nube informática (ver epígrafe 2.3).

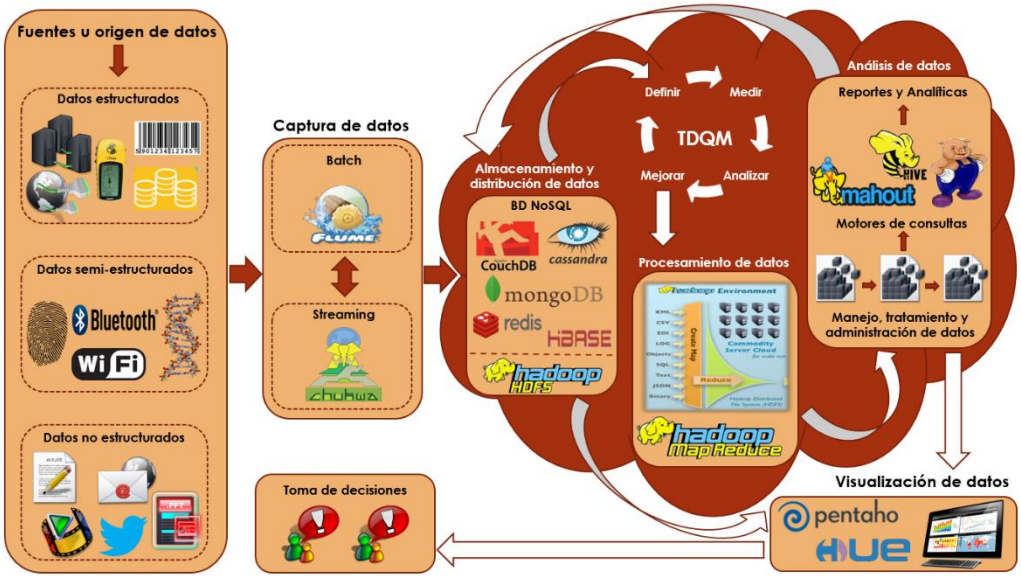


Figura 12 - Arquitectura integradora (*big data*, nube informática y proceso de calidad de datos).

El flujo de toda información *big data* comienza a partir de las fuentes u orígenes de datos, dentro de las cuales se pueden encontrar uno o todos los tipos de datos que maneja *big data*. Posteriormente, entra en función la fase de captura o recolección, por la cual pasan los datos antes de ser almacenados. Seguido a esto, la necesidad de almacenar, procesar y analizar grandes cantidades de datos ha traído consigo que muchas organizaciones e individuos adopten la nube informática para la realización de dichas tareas (Huan, 2013), debido al gran número de aplicaciones científicas para extensos experimentos desplegadas actualmente en la nube, y en constante crecimiento, debido a la falta de instalaciones informáticas o tecnologías disponibles en servidores locales, los altos costos, y el aumento en el volumen de datos producidos y consumidos por dichos experimentos (Pandey, 2013). Además, los proveedores de servicios de la nube han comenzado a integrar marcos de trabajo para el procesamiento de datos en paralelo en sus servicios, ayudando al usuario en accesos a sus recursos y despliegue de sus programas (Warneke, 2009). El proceso de calidad de datos fue ubicado en una etapa de pre-procesamiento, con el objetivo de realizar las funciones de procesamiento y análisis con datos de calidad, evitando así la inclusión de posibles errores o inconsistencias en los resultados finales, los cuales son visualizados para culminar con una correcta toma de decisiones. En cada una de las fases destacar el uso de varias herramientas claves que permiten realizar de forma eficiente las tareas a desarrollar en cada una de ellas.

2.6 Conclusiones parciales

En este capítulo se muestran los aspectos fundamentales que conforman *big data* y como ha llegado a convertirse en un reto para organizaciones que intentan corregir diversos problemas internos, aplicando soluciones *big data* a sus casos de estudio reales para llevar a cabo una correcta toma de decisiones. El uso de *big data* ha permitido a los investigadores descubrir nuevos conocimientos referentes al tratamiento de grandes volúmenes de información, la transferencia de datos a alta velocidad y el trabajo con variada información. También su uso representa una novedosa solución computacional ante la incapacidad humana de procesar grandes volúmenes de datos de manera equivalente a su crecimiento continuo. *Big data* posee diversos campos abiertos a investigación, dentro de los cuales se encuentra la calidad de datos como uno de los principales desafíos para la entrega de productos de información de calidad. Por otra parte, una solución *big data* implica la integración de diversos componentes que en conjunto forman el ecosistema necesario para analizar estas grandes cantidades de datos. En este sentido, *Hadoop* es la solución que más éxito y repercusión está teniendo, no sólo porque su utilización en compañías que ofrecen distribuciones *big data* sino porque cuenta con un ecosistema muy completo (en continuo crecimiento), poseedor de herramientas adaptables a cada una de las fases que conforman una arquitectura *big data*.

CAPÍTULO 3. ANÁLISIS DE LA CALIDAD DE DATOS EN EL CASO DE ESTUDIO ABCD

El presente capítulo tiene como objetivo realizar el análisis de la calidad de datos en el caso de estudio ABCD, iniciando con una caracterización de dicho sistema, viendo para esto su principal misión, visión, bases de datos que lo conforman, y los principales módulos en los que se encuentra estructurado. Además, se analiza el instrumento utilizado para llevar a cabo dicho análisis de calidad, el cual se basa en la aplicación de encuestas a especialistas que trabajan directamente con este sistema. Finalmente, con los resultados obtenidos del procesamiento de las encuestas aplicadas, se realizan varios análisis que conllevan a la detección de las principales dimensiones de calidad de datos relacionadas con el contexto para este caso de estudio, un levantamiento de los problemas de calidad de datos que poseen los módulos de este sistema, así como las causas fundamentales que traen consigo la aparición de los mismos.

3.1 Descripción general del ABCD

El sistema ABCD, está dirigido a la gestión integrada de procesos de bibliotecas y operación automatizada en línea así como *off-line*. Su arquitectura tecnológica está basada en servicios, utilización de estándares web, protocolos abiertos, innovación e integración con nuevas metodologías y tecnologías de información y comunicación, accesibilidad a usuarios y buenas prácticas de seguridad según los patrones internacionales (de Smet and Spinak, 2009). Su misión consiste en servir de soporte a la gestión bibliotecaria, integrando en una plataforma los procesos de adquisición, selección, procesamiento y circulación (o préstamos); además gestiona colecciones y maneja clases de usuarios.

Su desarrollo, en marcha desde el año 2007, está asegurado por el Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud (BIREME) con el apoyo financiero de la *Vlaamse Interuniversitaire Raad* (Consejo Interuniversitario Flamenco, VLIR, Bélgica). ABCD es de hecho una modernización del sistema gestor de bases de datos orientado a documentos textuales *CDS/ISIS* (*Centralized Documentation System/Integrated Set of Information Services*), originalmente desarrollado y mantenido por la Organización Internacional del Trabajo (OIT) en los años 60 y reanudado más tarde por la Organización de Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO); además, muy utilizado para la gestión de documentación sobre todo en los países en desarrollo (de Smet and Spinak, 2009).

La flexibilidad y versatilidad son la vanguardia de los criterios en los que se desarrolla este sistema. La flexibilidad, por ejemplo, se ilustra por el hecho de que, en principio, pero también prácticamente, cualquier estructura bibliográfica puede ser gestionada por el software, o incluso creada por el mismo. Las estructuras no bibliográficas, por otro lado, se pueden crear siempre y cuando la información sea principalmente información textual, ya que esta es la limitante planteada por la tecnología de base de datos subyacente, en este caso, la base de datos textual CDS/ISIS. ABCD es llamado una *suite* de programas, ya que cuenta con varios módulos los cuales pueden cooperar, pero también pueden existir sin la necesidad de otro (Figura 13). ABCD consta de cinco módulos principales que ofrecen las funciones de catalogación, préstamos, adquisiciones, registro de publicaciones periódicas (SeCSWeb), y administración de préstamos avanzados (EmpWeb). Además existen otros como difusión selectiva de la información (*Content Management System*, CMS), generación de estadísticas, gestión de usuarios, diccionario y catálogo accesible al público (*Open Public Access Catalog*, OPAC) (de Smet and Spinak, 2009).

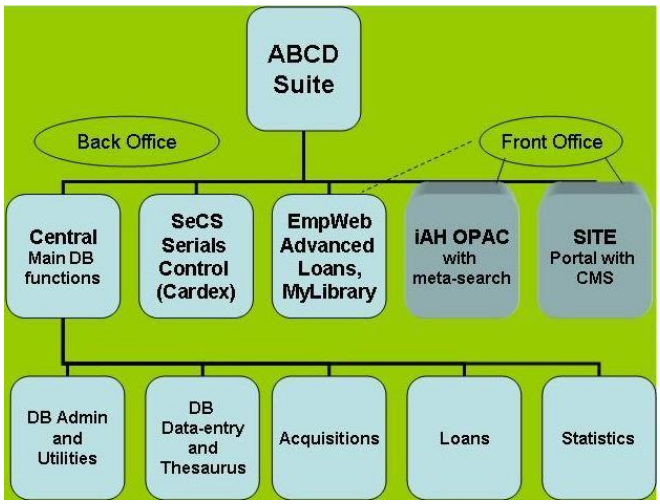


Figura 13 - *Suite* ABCD. [Fuente: (de Smet and Spinak, 2009)]

ABCD se utiliza en un entorno cliente/servidor, y sus funciones son accesibles a través de un navegador. Está desarrollado en los lenguajes de programación web HTML, PHP, JavaScript y XML, a los cuales se suma el WXIS, un servidor de bases de datos de la familia CDS/ISIS desarrollado por BIREME. Para el cumplimiento de normas en materia de información documental, ABCD viene pre-configurado para trabajar con algunos estándares bibliográficos tales como: MARC21, INTERMARC, UNIMARC, CEPAL, AGRIS, LILACS, entre otros;

facilitando así la comunicación y el intercambio de datos bibliográficos con otros sistemas de información, y la migración de bases de datos desde CDS/ISIS y sus variantes (WinISIS, MicroISIS, OpenISIS, CISIS, JAVA-ISIS, J-ISIS, etc.) y/o *Excel*. El 3 de diciembre de 2009, la versión 1.0 del sistema ABCD, fue lanzada con una base de datos en formato bibliográfico MARC21, como formato universal, con la posibilidad para todos los usuarios de construir todo tipo de bases de datos tan estructuradas como el mismo MARC21 (de Smet and Spinak, 2009).

3.1.1 Antecedentes y uso del ABCD en la UCLV

En nuestro país el sistema ABCD se ha adoptado en diferentes centros bibliotecarios por orientaciones del MES. En nuestra provincia este sistema es usado por la Universidad Central "Marta Abreu" de Las Villas (UCLV), cuya instalación se encuentra en la biblioteca central de la misma, la cual sirve como servidor principal, permitiendo que mediante la web, las restantes 18 bibliotecas ubicadas en cada facultad de la universidad, se conecten al mismo para la realización y gestión de diferentes actividades vinculadas a procesos bibliográficos.

Antes de ser adoptado el sistema ABCD por las bibliotecas de la UCLV, estas realizaban el tratamiento de datos bibliográficos con el sistema *Quipusnet*, herramienta desarrollada por el grupo Chasqui perteneciente al Centro de Documentación de Información Científico Técnica (CDICT) y la colaboración de estudiantes de la carrera Ciencia de la Computación de la UCLV. Dicha herramienta guardaba toda su información en el gestor de bases de datos *Microsoft SQL Server* y los datos seguían un formato bibliográfico no estandarizado creado por especialistas del MES. A partir del año 2011 los datos bibliográficos contenidos en el sistema *Quipusnet*, fueron sometidos a un proceso de migración hacia ABCD, donde el MES decide utilizar el formato estandarizado MARC21. La migración no pudo llevarse a cabo de manera automática, porque no existían herramientas computacionales capaces de hacerlo y el proceso era muy engorroso. Por tanto, se decide comenzar desde cero y entrar manualmente todos los datos almacenados en *Quipusnet* hacia ABCD. La entrada manual de datos al sistema trajo consigo la aparición de varios problemas de calidad en las fuentes de la suite ABCD, la cual cuenta con la base de datos ISIS NoSQL orientada a documentos y denominada MARC, que almacena datos semi-estructurados (metadatos bibliográficos de catalogación). Además, cuenta con 10 bases de datos relacionales que manejan datos estructurados sobre: adquisiciones y copias, proveedores,

sugerencias, órdenes de compra, usuarios de préstamos, objeto de préstamo, reservaciones, suspensiones y multas, transacciones de préstamo y usuarios del sistema.

En el resto del país el sistema ABCD es usado por otras universidades dentro de las que se destacan las cinco universidades que forman parte del proyecto Red-TIC, tal es el caso de Pinar del Rio, Holguín, Camagüey, la Universidad de Ciencias Informáticas (UCI) y la UCLV; dentro de su visión futura se encuentra ser el sistema de gestión bibliotecaria del MES, integrando mediante la búsqueda federada todos los centros de educación superior del país. Dicha integración ha comenzado a dar sus primeros pasos en la provincia de Villa Clara, la cual incluye al centro ***"Manuel Fajardo"***, el cual realiza los procesos de gestión bibliotecaria de forma manual (tradicional) y la Universidad de Ciencias Pedagógicas ***"Félix Varela"*** que lo hace mediante el uso de la herramienta *WinISIS*.

3.2 Descripción del instrumento utilizado para el análisis de calidad de datos

Como método de apoyo para llevar a cabo el proceso de gestión de calidad de datos en el sistema ABCD, se siguen las tres primeras fases que conforman la metodología TDQM (ver epígrafe 1.2).

Para la identificación de dimensiones y problemas de calidad de datos, así como las causas que conllevan a los mismos en fuentes de datos *big data*, se utiliza como instrumento la entrevista estructurada apoyada en la aplicación de encuestas impresas a diferentes especialistas que manejan e interactúan con el sistema ABCD, buscando con estas específicamente:

- Analizar la información obtenida para captar la percepción del usuario respecto a varios aspectos relevantes sobre calidad de datos, específicamente, dimensiones de calidad de datos, problemas de calidad de datos y las causas que conllevan a los mismos.
- Medir indicadores de calidad de datos para el caso de estudio del sistema ABCD.
- Aplicar acciones de mejora para la calidad de datos en el sistema ABCD.

Los dos últimos puntos, correspondientes a medir y aplicar acciones de mejora, quedan abiertos para futuras investigaciones, pues en este trabajo de diploma se realiza solamente el análisis descriptivo de los datos recogidos en las encuestas.

La realización de una encuesta puede ser vista como una serie de etapas sucesivas que en conjunto permiten obtener datos acerca de la población objetivo de la misma. De forma general las encuestas son caracterizadas en tres etapas: etapa de diseño, levantamiento de datos, y

procesamiento de datos. Consecuentemente, una encuesta que presente problemas en una de sus etapas, eventualmente afectará alguna de las etapas siguientes, pudiendo llegar a alterar los resultados del proceso consolidados en las bases de datos definitivas. Visto de esta manera una encuesta no mejora su calidad trabajando solo el producto final, sino que resulta necesario además construir un concepto de calidad de mayor complejidad, que permita comprender y trabajar las particularidades de cada etapa en pos de una mejor evaluación global (Estadísticas, 2011).

3.3 Etapa de diseño

El diseño de la encuesta se estructura en tres secciones, como se muestra en la Tabla 15.

Sección	**Descripción**	**N^{o} de preguntas**
1 - Datos del encuestado	Esta parte de la encuesta consiste en una caracterización de los usuarios encuestados respecto a su nivel educacional, ocupación laboral, experiencia con el ABCD, rol que ocupan en dicho sistema, entre otros aspectos	10
2 - Dimensiones de calidad de datos	Esta sección indaga en la percepción que poseen los encuestados sobre diferentes dimensiones de calidad de datos, adoptadas en este contexto, para un mejor trabajo con los datos del ABCD	1
3 - Problemas y causas	Esta sección pretende realizar un levantamiento de los posibles problemas de calidad de datos que presentan los módulos principales que conforman el sistema ABCD, así como las causas que conllevan a los mismos	3
Total de preguntas		**14**

Tabla 15 - Estructura general de la encuesta para la evaluación de la calidad de datos en el sistema ABCD.

El diseño de la sección 1 se basa en la entrevista realizada a uno de los jefes de servicio del sistema ABCD. Para el diseño de la sección 2, se utiliza como referente el marco de trabajo propuesto por Wang and Strong (1996) (ver subepígrafe 1.3.1), del cual se toman solo 14 dimensiones (el tiempo de vida se trata mediante las dimensiones actualidad y volatilidad) del total de 15 que conforman dicho marco y se adicionan las siguientes 9 dimensiones: precisión, consistencia, actualidad, volatilidad, unicidad, utilidad, preservación, comodidad y conformidad, las cuales se consideran representativas teniendo en cuenta la opinión de especialistas en el caso

de estudio analizado. La sección 3 se elabora tomando en cuenta algunos de los diferentes problemas y causas tratados en la taxonomía de errores analizada en el subepígrafe 1.5.1.

Finalmente, la encuesta queda conformada en un total de seis páginas, y cuenta con varios tipos de preguntas dentro de las que se encuentran: preguntas de selección individual y múltiple, en otras se utiliza la escala de *Likert* (1 al 7) y algunos casos con preguntas abiertas (ver Anexo 1). Algunas recomendaciones empleadas en esta etapa de diseño fueron:

- El tipo y número de preguntas que contenga la encuesta dependerá de las variables a analizar. No se debe incorporar preguntas que no tengan relación directa. Se recomienda que sea breve, usando vocabulario sencillo y con una secuencia de preguntas de acuerdo a la lógica del entrevistado (Arias Chalico *et al.*, 2002).
- En lo posible, las preguntas deben ser de respuesta cerrada, pues su procesamiento es más sencillo. Las preguntas abiertas son muy útiles para detectar opiniones, percepciones y preferencias; para estas variables no se recomiendan las preguntas de respuesta cerrada (Arias Chalico *et al.*, 2002).
- Para formular el cuestionario definitivo es necesario probar uno preliminar con una pequeña muestra de la población destino, pues con ello se logra reconocer la variabilidad de respuestas posibles o situaciones no consideradas originalmente (Arias Chalico *et al.*, 2002).
- Cuando se trabaje con muestras muy grandes y con preguntas complejas, es conveniente que las preguntas tengan instrucciones precisas de aplicación que queden resaltadas en el cuestionario (Arias Chalico *et al.*, 2002).

3.4 Levantamiento de datos

El proceso de levantamiento de datos en el terreno se efectuó en la biblioteca central de la UCLV y estuvo apoyado y coordinado por el jefe de servicios del sistema ABCD, además estuvo supervisado por dos encuestadores en las diferentes áreas donde estas fueron aplicadas.

3.4.1 Definición de la muestra

El universo de este estudio se compuso de 44 especialistas que laboran directamente con el sistema ABCD realizando tareas de catalogación, préstamos, adquisiciones, consultas, entre otras. Como representativa de este universo, se propuso una muestra independiente con 16 casos

respectivamente, lograda en su totalidad. La cual representa aproximadamente el 36.4% de la población.

3.4.2 Aplicación de la encuesta

En general, no se registraron problemas durante la aplicación de las encuestas, solo es válido mencionar que si bien la acogida fue positiva por la mayoría de los usuarios, se observaron algunos rechazos, principalmente por el estado anímico de algunos de estos, la extensión de la encuesta, o cierto desconocimiento del tema, por lo que prefirieron no responder algunas preguntas de la misma. Si bien ello no implicó una dificultad para el trabajo de terreno, se deja constancia de dicha situación, por su posible impacto en la evaluación final de los resultados.

3.5 Procesamiento de datos

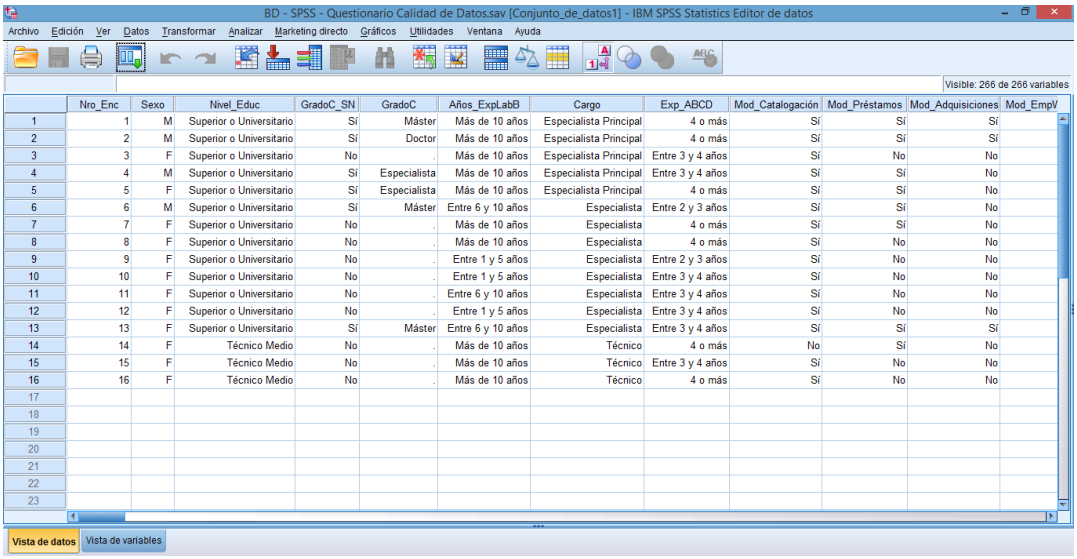
Una vez concluida la etapa de levantamiento de datos, se procedió de manera progresiva, a la digitalización de encuestas y posteriormente a su procesamiento.

3.5.1 Digitalización de los resultados (base de datos)

La base de datos para el almacenamiento de la información obtenida una vez aplicada las encuestas, se construyó con el uso del software *IBM SPSS Statistics* v.21, el cual posee una amplia gama de funcionalidades para el procesamiento estadístico de datos.

Cada sección del cuestionario cuenta con un conjunto de variables (nominales u ordinales), que responden a cada una de las preguntas de la encuesta. Cada una de las columnas del SPSS representa una variable, donde en casos de preguntas con respuesta múltiple, por ejemplo, los módulos con los que ha trabajado cierta persona, así como las bases de datos que ha consultado, se deben usar tantas variables como valores esta pueda tomar. La base de datos quedó finalmente compuesta por un total de 266 variables.

Al ingresar los datos se registró primero el número consecutivo de los cuestionarios, pues esto facilita las posteriores búsquedas en los datos originales. Además, el uso de códigos para ingresar las respuestas fue realizado con ayuda de etiquetas, facilitando la comprensión de los resultados a otros usuarios que harán uso posterior de la base de datos e impidiendo que con el paso del tiempo se olviden sus significados. La Figura 14 muestra una pequeña parte de la base de datos confeccionada en el SPSS.



	Nro_Enc	Sexo	Nivel_Educ	GradoC_SN	GradoC	Años_ExpLabB	Cargo	Exp_ABCD	Mod_Catalogación	Mod_Préstamos	Mod_Adquisiciones	Mod_EmpV
1	1	M	Superior o Universitario	Sí	Máster	Más de 10 años	Especialista Principal	4 o más	Sí	Sí	Sí	
2	2	M	Superior o Universitario	Sí	Doctor	Más de 10 años	Especialista Principal	4 o más	Sí	Sí	Sí	
3	3	F	Superior o Universitario	No	.	Más de 10 años	Especialista Principal	Entre 3 y 4 años	Sí	No	No	
4	4	M	Superior o Universitario	Sí	Especialista	Más de 10 años	Especialista Principal	Entre 3 y 4 años	Sí	Sí	No	
5	5	F	Superior o Universitario	Sí	Especialista	Más de 10 años	Especialista Principal	4 o más	Sí	Sí	No	
6	6	M	Superior o Universitario	Sí	Máster	Entre 6 y 10 años	Especialista	Entre 2 y 3 años	Sí	Sí	No	
7	7	F	Superior o Universitario	No	.	Más de 10 años	Especialista	4 o más	Sí	Sí	No	
8	8	F	Superior o Universitario	No	.	Más de 10 años	Especialista	4 o más	Sí	No	No	
9	9	F	Superior o Universitario	No	.	Entre 1 y 5 años	Especialista	Entre 2 y 3 años	Sí	No	No	
10	10	F	Superior o Universitario	No	.	Entre 1 y 5 años	Especialista	Entre 3 y 4 años	Sí	No	No	
11	11	F	Superior o Universitario	No	.	Entre 6 y 10 años	Especialista	Entre 3 y 4 años	Sí	No	No	
12	12	F	Superior o Universitario	No	.	Entre 1 y 5 años	Especialista	Entre 3 y 4 años	Sí	No	No	
13	13	F	Superior o Universitario	Sí	Máster	Entre 6 y 10 años	Especialista	Entre 3 y 4 años	Sí	Sí	Sí	
14	14	F	Técnico Medio	No	.	Más de 10 años	Técnico	4 o más	No	Sí	No	
15	15	F	Técnico Medio	No	.	Más de 10 años	Técnico	Entre 3 y 4 años	Sí	No	No	
16	16	F	Técnico Medio	No	.	Más de 10 años	Técnico	4 o más	Sí	No	No	
17												
18												
19												
20												
21												
22												
23												

Figura 14 - Fragmento de la base de datos en el SPSS de la encuesta aplicada.

3.5.2 Análisis estadístico e informe de los resultados

3.5.2.1 Análisis de frecuencia

Primeramente se realizó un análisis de frecuencia buscando posibles inconsistencias que pudieron haber ocurrido con el almacenamiento de la información en la base de datos. En la Tabla 16 se muestran solo algunas variables con sus análisis de valores válidos y perdidos en este proceso.

Statistics

		Género	Módulo Catalogación	BD Adquisiciones y Copias	Completitud	Exactitud	Entrada manual de datos	Migración de datos
N	Valid	16	16	16	16	16	16	15
	Missing	0	0	0	0	0	0	1

Tabla 16 - Valores válidos y perdidos para algunas variables de la base de datos.

Como se observa en la Tabla 16, los resultados no indican pérdidas de gran valor, solo algunos casos que arrojaron valores pequeños correspondientes a campos sin llenar en las encuestas. A continuación se muestran algunas de las tablas de frecuencia con sus gráficas asociadas correspondientes a las variables *Género* (Tabla 17), *Módulo Catalogación* (Tabla 18) y *BD Adquisiciones y Copias* (Tabla 19).

Género

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	12	75,0	75,0	75,0
	M	4	25,0	25,0	100,0
	Total	16	100,0	100,0	

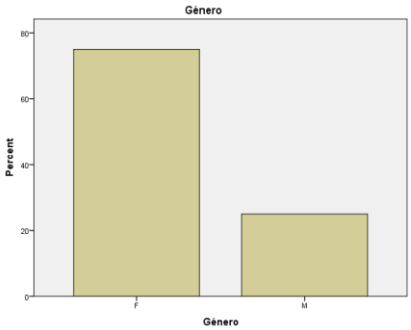


Tabla 17 - Frecuencia de la variable *Género*.

Módulo Catalogación

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	1	6,3	6,3	6,3
	Sí	15	93,8	93,8	100,0
	Total	16	100,0	100,0	

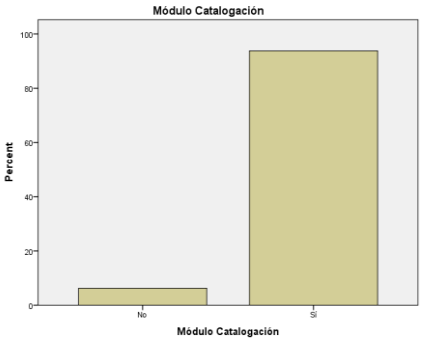


Tabla 18 - Frecuencia de la variable *Módulo Catalogación*.

BD Adquisiciones y Copias

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	6	37,5	37,5	37,5
	Sí	10	62,5	62,5	100,0
	Total	16	100,0	100,0	

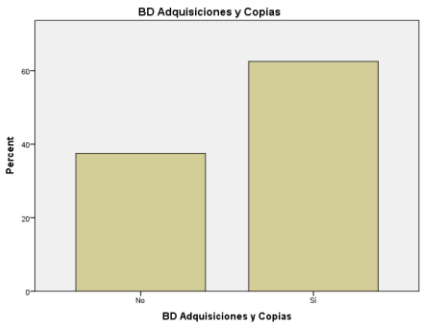


Tabla 19 - Frecuencia de la variable *BD Adquisiciones y Copias*.

En los tres casos anteriores las tablas muestran la frecuencia de los resultados para cada una de las variables así como los porcientos asociados a estos valores.

3.5.2.2 Análisis descriptivo para dimensiones de calidad de datos

Una vez efectuado el análisis de frecuencia inicial, se procede a realizar un análisis descriptivo para detectar las dimensiones con mayores grados de significación en el contexto analizado como muestra la Tabla 20.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Completitud	16	7	7	7,00	,000
Exactitud	16	7	7	7,00	,000
Precisión	15	7	7	7,00	,000
Consistencia	16	7	7	7,00	,000
Actualidad	16	3	7	**6,56**	1,209
Volatilidad	16	4	7	**6,63**	,885
Credibilidad	16	7	7	7,00	,000
Objetividad	16	4	7	**6,81**	,750
Reputación	15	7	7	7,00	,000
Valor agregado	15	7	7	7,00	,000
Relevancia	15	7	7	7,00	,000
Cantidad apropiada de datos	15	4	7	**6,73**	,799
Interpretabilidad	15	7	7	7,00	,000
Facilidad de comprensión	16	7	7	7,00	,000
Consistencia representacional	16	7	7	7,00	,000
Representación concisa	15	7	7	7,00	,000
Accesibilidad	15	7	7	7,00	,000
Unicidad	16	7	7	7,00	,000
Seguridad	15	7	7	7,00	,000
Utilidad	16	7	7	7,00	,000
Preservación	16	7	7	7,00	,000
Comodidad	16	7	7	7,00	,000
Conformidad	16	7	7	7,00	,000
Valid N (listwise)	14				

Tabla 20 - Análisis descriptivo para dimensiones de calidad de datos.

Con los resultados de este análisis se llega a la conclusión que el conjunto de dimensiones en su totalidad presentan altos niveles de significación dados por los encuestados, lo que indica que todas las dimensiones son consideradas importantes en el trabajo con los datos. Los casos que se encuentran por debajo de la media (resaltados en color rojo), no influyen a la hora de determinar

que estas dimensiones son menos significativas que otras ya que la diferencia entre estos es mínima (0.44 en el caso más distante, el cual se refiere a la dimensión actualidad).

3.5.2.3 Análisis descriptivo para la detección de problemas de calidad de datos

El análisis descriptivo para la detección de problemas de calidad de datos se realizó por separado en cada uno de los módulos principales del sistema ABCD. Las conclusiones para este análisis se basan en el valor de la media de cada uno de los problemas, tomando en consideración solo los valores mayores que cero (ordenados descendentemente), los cuales indican cuales son los problemas mayores y menores que afectan un módulo determinado, ya que los problemas con media igual cero no son representativos en los módulos.

La Tabla 21 muestra primeramente los resultados del análisis que indica la cantidad de encuestados que utilizan cada módulo.

Módulo	**Cantidad de especialistas por módulo**
Catalogación	**16**
Préstamos	8
Adquisiciones	3
EmpWeb	**13**
SeCSWeb	2

Tabla 21 - Cantidad de especialistas por módulo.

Como se puede observar, la Tabla 21 indica que el módulo más usado por los especialistas es el de Catalogación (color rojo) con el total de la muestra seleccionada, seguido del módulo EmpWeb (color azul) con un total de 13 especialistas.

1) Módulo Adquisición

La Tabla 22 muestra el análisis descriptivo realizado para la detección de problemas de calidad de datos en el módulo Adquisición.

Descriptive Statistics

	N	Mean
MA_Errores ortográficos	16	**,13**
MA_Inseguridades en los datos	16	**,06**
MA_Datos desfasados temporalmente	16	**,06**
MA_Diversidad en las fuentes de datos	16	**,06**
MA_Datos no estandarizados	16	**,06**
MA_Campos que no siguen un formato estandarizado	16	,00
MA_Datos perdidos	16	,00
MA_Valores implícitos	16	,00
MA_Falta de datos en un campo no nulo	16	,00
MA_Conflictos estructurales	16	,00
MA_Conflictos de nombre	16	,00
MA_Valor de datos incorrecto	16	,00
MA_Tipo de datos incorrecto	16	,00
MA_Datos colgantes	16	,00
MA_Campos multipropósitos	16	,00
MA_Valores ficticios en los campos	16	,00
MA_Inexactitud en los datos	16	,00
MA_Metadatos no claros	16	,00
MA_Valores nulos o vacíos	16	,00
MA_Formato de datos incompatibles	16	,00
MA_Violaciones de restricciones de integridad	16	,00
MA_Datos con ruido (erróneos, incorrectos)	16	,00
MA_Inconsistencias entre diferentes copias del mismo dato	16	,00
MA_Datos ambiguos	16	,00
MA_Datos duplicados	16	,00
MA_Datos incompletos, ausentes	16	,00
Valid N (listwise)	16	

Tabla 22 - Análisis descriptivo. Detección de problemas en el módulo Adquisición.

Los resultados anteriores arrojaron un total de cinco problemas que afectan este módulo. La Figura 15 muestra los resultados obtenidos, indicando los problemas más representativos del mismo.

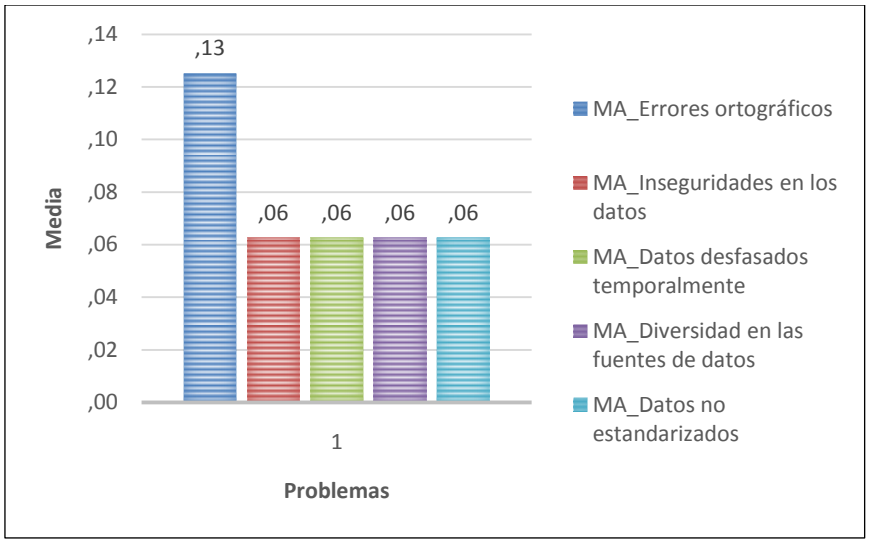


Figura 15 - Problemas más representativos del módulo Adquisición.

2) Módulo Catalogación

La Tabla 23 muestra el análisis descriptivo realizado para la detección de problemas de calidad de datos en el módulo Catalogación.

Descriptive Statistics

	N	Mean
MC_Errores ortográficos	16	,94
MC_Datos incompletos, ausentes	16	,88
MC_Inseguridades en los datos	16	,81
MC_Datos perdidos	16	,75
MC_Conflictos de nombre	16	,69
MC_Inconsistencias entre diferentes copias del mismo dato	16	,69
MC_Datos duplicados	16	,69
MC_Datos con ruido (erróneos, incorrectos)	16	,69
MC_Valores nulos o vacíos	16	,63
MC_Diversidad en las fuentes de datos	16	,63
MC_Datos no estandarizados	16	,63
MC_Datos colgantes	16	,56
MC_Inexactitud en los datos	16	,56
MC_Metadatos no claros	16	,56
MC_Falta de datos en un campo no nulo	16	,50
MC_Tipo de datos incorrecto	16	,50

MC_Campos multipropósitos	16	,50
MC_Datos desfasados temporalmente	16	,50
MC_Violaciones de restricciones de integridad	16	,44
MC_Valores implícitos	16	,38
MC_Valor de datos incorrecto	16	,38
MC_Valores ficticios en los campos	16	,38
MC_Formato de datos incompatibles	16	,31
MC_Campos que no siguen un formato estandarizado	16	,19
MC_Conflictos estructurales	16	,19
MC_Datos ambiguos	16	,13
Valid N (listwise)	16	

Tabla 23 - Análisis descriptivo. Detección de problemas en el módulo Catalogación.

Los resultados anteriores arrojaron que el módulo para la catalogación de datos bibliográficos al ser el más usado y conocido por los especialistas, presenta la mayor cantidad de problemas. La Figura 16 muestra los resultados obtenidos del análisis anterior, indicando los diez problemas más representativos de este módulo.

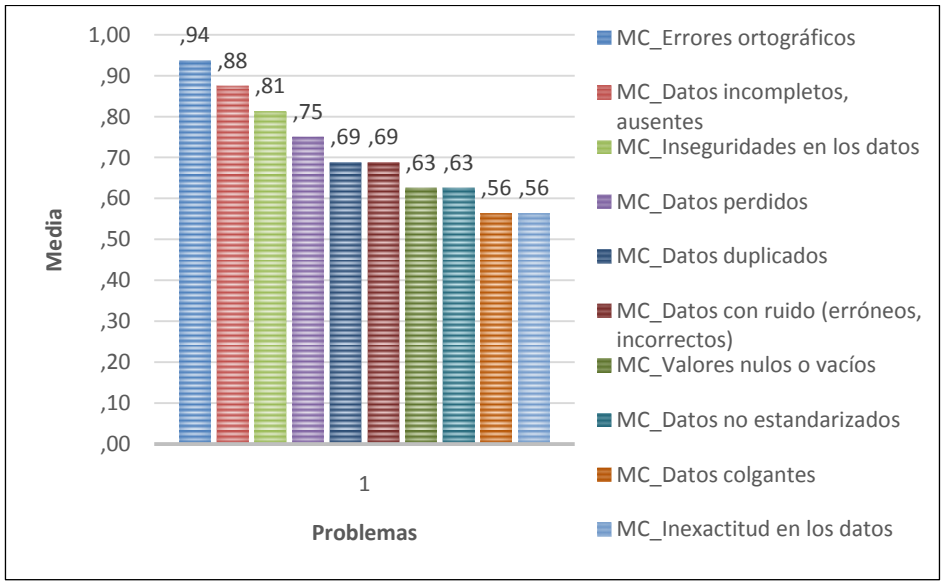


Figura 16 - Problemas más representativos del módulo Catalogación.

3) Módulo Préstamos

La Tabla 24 muestra el análisis descriptivo realizado para la detección de problemas de calidad de datos en el módulo Préstamos.

Descriptive Statistics

	N	Mean
MP_Inexactitud en los datos	16	**,13**
MP_Inseguridades en los datos	16	**,13**
MP_Errores ortográficos	16	**,06**
MP_Violaciones de restricciones de integridad	16	**,06**
MP_Datos desfasados temporalmente	16	**,06**
MP_Campos que no siguen un formato estandarizado	16	,00
MP_Datos perdidos	16	,00
MP_Valores implícitos	16	,00
MP_Falta de datos en un campo no nulo	16	,00
MP_Conflictos estructurales	16	,00
MP_Conflictos de nombre	16	,00
MP_Valor de datos incorrecto	16	,00
MP_Tipo de datos incorrecto	16	,00
MP_Datos colgantes	16	,00
MP_Campos multipropósitos	16	,00
MP_Valores ficticios en los campos	16	,00
MP_Metadatos no claros	16	,00
MP_Valores nulos o vacíos	16	,00
MP_Formato de datos incompatibles	16	,00
MP_Datos con ruido (erróneos, incorrectos)	16	,00
MP_Inconsistencias entre diferentes copias del mismo dato	16	,00
MP_Diversidad en las fuentes de datos	16	,00
MP_Datos ambiguos	16	,00
MP_Datos duplicados	16	,00
MP_Datos incompletos, ausentes	16	,00
MP_Datos no estandarizados	16	,00
Valid N (listwise)	16	

Tabla 24 - Análisis descriptivo. Detección de problemas en el módulo Préstamos.

Los resultados anteriores arrojaron un total de cinco problemas que afectan este módulo. La Figura 17 muestra los resultados obtenidos, indicando los problemas más representativos del mismo.

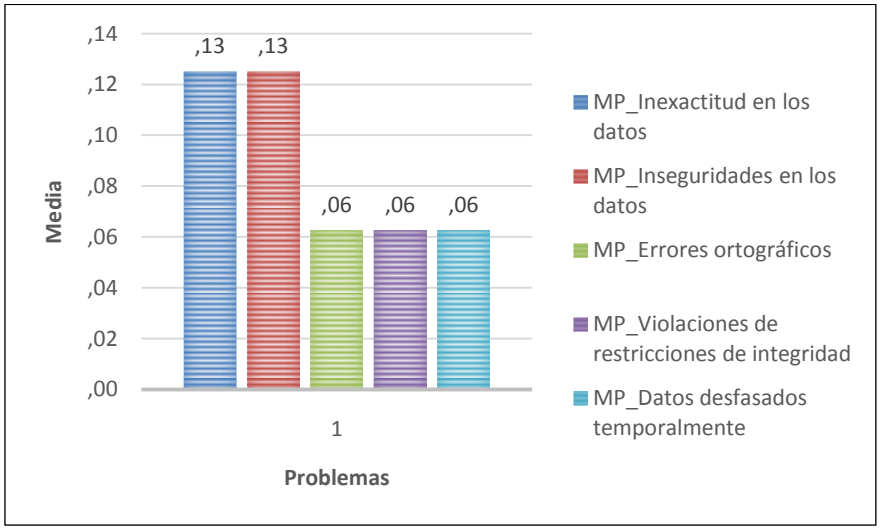


Figura 17 - Problemas más representativos del módulo Préstamos.

4) Módulo SeCSWeb

La Tabla 25 muestra el análisis descriptivo realizado para la detección de problemas de calidad de datos en el módulo SeCSWeb.

Descriptive Statistics

	N	Mean
MSW_Errores ortográficos	16	**,13**
MSW_Inseguridades en los datos	16	**,06**
MSW_Diversidad en las fuentes de datos	16	**,06**
MSW_Datos perdidos	16	**,06**
MSW_Conflictos de nombre	16	**,06**
MSW_Tipo de datos incorrecto	16	**,06**
MSW_Valores ficticios en los campos	16	**,06**
MSW_Datos incompletos, ausentes	16	**,06**
MSW_Campos que no siguen un formato estandarizado	16	,00
MSW_Valores implícitos	16	,00
MSW_Falta de datos en un campo no nulo	16	,00
MSW_Conflictos estructurales	16	,00

MSW_Valor de datos incorrecto	16	,00
MSW_Datos colgantes	16	,00
MSW_Campos multipropósitos	16	,00
MSW_Inexactitud en los datos	16	,00
MSW_Metadatos no claros	16	,00
MSW_Valores nulos o vacíos	16	,00
MSW_Formato de datos incompatibles	16	,00
MSW_Violaciones de restricciones de integridad	16	,00
MSW_Datos con ruido (erróneos, incorrectos)	16	,00
MSW_Inconsistencias entre diferentes copias del mismo dato	16	,00
MSW_Datos desfasados temporalmente	16	,00
MSW_Datos ambiguos	16	,00
MSW_Datos duplicados	16	,00
MSW_Datos no estandarizados	16	,00
Valid N (listwise)	16	

Tabla 25 - Análisis descriptivo. Detección de problemas en el módulo SeCSWeb.

Los resultados anteriores arrojaron un total de ocho problemas que afectan este módulo. La Figura 18 muestra los resultados obtenidos, indicando los problemas más representativos del mismo.

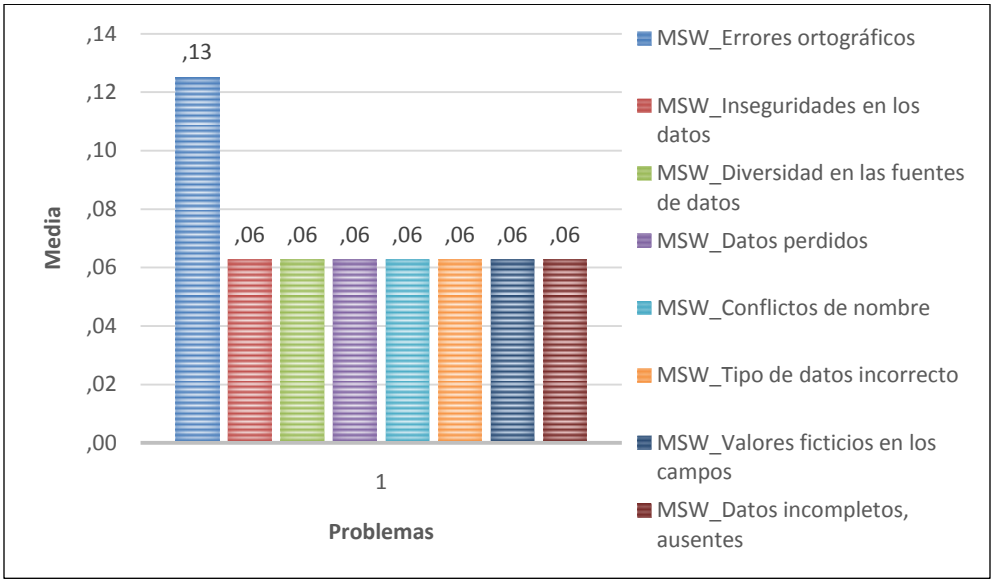


Figura 18 - Problemas más representativos del módulo SeCSWeb.

5) Módulo EmpWeb

La Tabla 26 muestra el análisis descriptivo realizado para la detección de problemas de calidad de datos en el módulo EmpWeb.

Descriptive Statistics

	N	Mean
MEW_Errores ortográficos	16	**,56**
MEW_Datos duplicados	16	**,50**
MEW_Datos perdidos	16	**,44**
MEW_Datos incompletos, ausentes	16	**,44**
MEW_Conflictos de nombre	16	**,38**
MEW_Inseguridades en los datos	16	**,38**
MEW_Falta de datos en un campo no nulo	16	**,31**
MEW_Tipo de datos incorrecto	16	**,31**
MEW_Metadatos no claros	16	**,31**
MEW_Datos con ruido (erróneos, incorrectos)	16	**,31**
MEW_Datos desfasados temporalmente	16	**,31**
MEW_Datos no estandarizados	16	**,25**
MEW_Valores implícitos	16	**,25**
MEW_Valor de datos incorrecto	16	**,25**
MEW_Inexactitud en los datos	16	**,25**
MEW_Inconsistencias entre diferentes copias del mismo dato	16	**,25**
MEW_Datos ambiguos	16	**,25**
MEW_Datos colgantes	16	**,19**
MEW_Campos multipropósitos	16	**,19**
MEW_Valores ficticios en los campos	16	**,19**
MEW_Valores nulos o vacíos	16	**,19**
MEW_Violaciones de restricciones de integridad	16	**,13**
MEW_Campos que no siguen un formato estandarizado	16	**,06**
MEW_Conflictos estructurales	16	**,06**
MEW_Formato de datos incompatibles	16	**,06**
MEW_Diversidad en las fuentes de datos	16	**,06**
Valid N (listwise)	16	

Tabla 26 - Análisis descriptivo. Detección de problemas en el módulo EmpWeb.

Los resultados anteriores arrojaron que este módulo, al ser el segundo más usado por los especialistas (junto al de Catalogación analizado anteriormente) para la realización de préstamos avanzados, presenta la mayor cantidad de problemas. La Figura 19 muestra los resultados obtenidos del análisis anterior, indicando los diez problemas más representativos de este módulo.

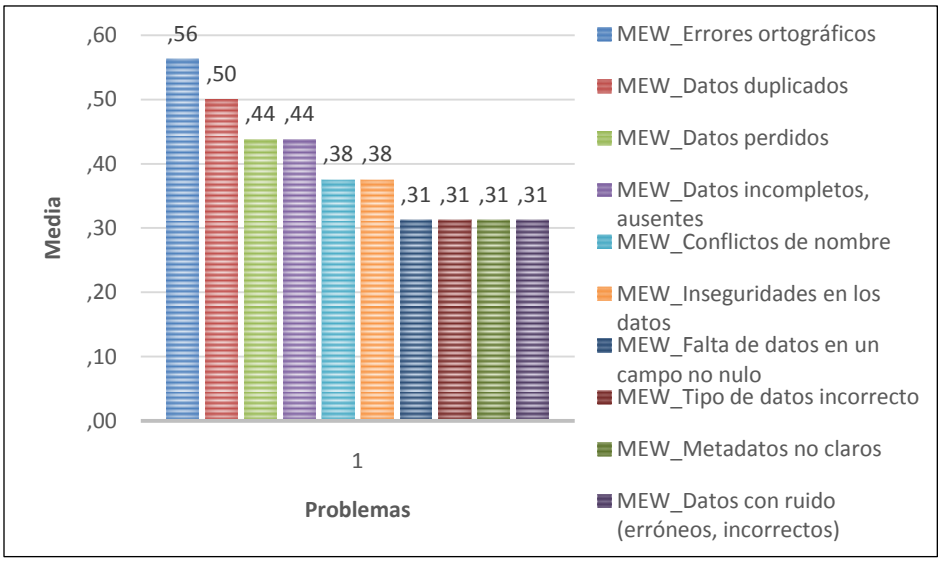


Figura 19 - Problemas más representativos del módulo EmpWeb.

6) Problemas más frecuentes

Una vez analizado los diferentes problemas que afectan cada módulo del sistema ABCD, la Tabla 27 muestra los más frecuentes o los más repetidos por módulos.

Problema	**Cant. Módulos**
Errores ortográficos	5
Inseguridades en los datos	5
Diversidad en las fuentes de datos	4
Datos desfasados temporalmente	4
Datos incompletos, ausentes	3
Datos perdidos	3
Conflictos de nombre	3
Datos no estandarizados	3
Inexactitud en los datos	3
Tipo de datos incorrecto	3
Violaciones de restricciones de integridad	3

Valores ficticios en los campos	3
Inconsistencias entre diferentes copias del mismo dato	2
Datos duplicados	2
Datos con ruido (erróneos, incorrectos)	2
Valores nulos o vacíos	2
Datos colgantes	2
Metadatos no claros	2
Falta de datos en un campo no nulo	2
Campos multipropósitos	2
Valores implícitos	2
Valor de datos incorrecto	2
Formato de datos incompatibles	2
Campos que no siguen un formato estandarizado	2
Conflictos estructurales	2
Datos ambiguos	2

Tabla 27 - Problemas más frecuentes en los módulos.

La Figura 20 muestra los resultados obtenidos con anterioridad (hasta al menos 3 módulos), donde se puede observar como problemas que afectan el total de módulos los errores ortográficos y las inseguridades en los datos.

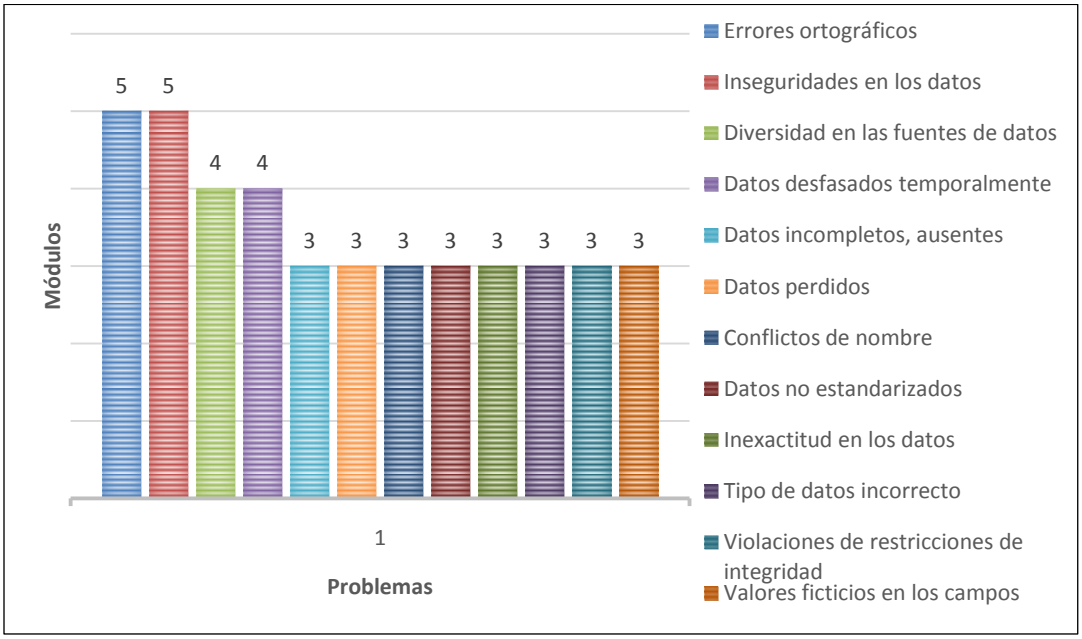


Figura 20 - Problemas más frecuentes en los módulos.

3.5.2.4 Análisis descriptivo para las causas que conducen a problemas de calidad de datos

Una vez analizado el conjunto de problemas que afectan cada uno de los módulos del sistema ABCD, en la Tabla 28 se muestran las causas posibles que conllevan a la aparición de dichos problemas, las cuales fueron ordenadas descendentemente de acuerdo al valor de la media, el cual indica causa mayor para los valores mayores de 2.5 (color rojo), media para los valores entre 1.5 y 2.5 (color verde) o baja para los valores menores de 1.5 (color azul).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Incorrecta aplicación de RCA	16	3	3	**3,00**	,000
Entrada manual de datos	16	2	3	**2,94**	,250
Migración de datos	15	2	3	**2,87**	,352
Información redundante	14	0	3	**2,71**	,825
Metadatos incompletos	14	0	3	**2,64**	,929
Desactualizaciones de datos	15	0	3	**2,53**	1,060
Cambios no capturados	14	0	3	**2,21**	1,311
Complejidad	14	0	3	**2,14**	1,027
Procesamiento de datos	15	0	3	**2,07**	1,280
Limpieza masiva de datos	15	0	3	**2,00**	1,309
Realciones entre las tablas	14	0	3	**1,79**	1,188
Integración de datos	13	0	3	**1,31**	1,109
Actualizaciones del sistema	14	0	3	**1,21**	1,369
Automatización de procesos	13	0	3	**1,15**	1,214
Purga de datos	15	0	3	**,93**	1,280
Diferentes formatos de codificación	15	0	3	**,67**	1,234
Valid N (listwise)	11				

Tabla 28 - Análisis descriptivo de causas que conllevan a problemas de calidad de datos.

3.6 Conclusiones parciales

En este capítulo se realizó el análisis de calidad de datos en el caso de estudio ABCD, aplicando las tres primeras fases correspondientes a la metodología TDQM. Para llevar a cabo dicho análisis se utilizó como instrumento la entrevista estructurada, apoyada por la aplicación de encuestas a una muestra de 16 especialistas que trabajan directamente con este sistema. Posteriormente con los resultados obtenidos del procesamiento de encuestas, se detectaron diversas anomalías y problemas de calidad en dicho sistema; para lo cual se hizo un

levantamiento de estos que incluyó primeramente la detección de dimensiones. El resultado de este primer análisis arrojó que todas las dimensiones que abarca la encuesta son altamente importantes para los especialistas. En un segundo análisis se detectaron los problemas que más afectan los módulos principales del ABCD para el cual los resultados arrojaron que los errores ortográficos y las inseguridades en los datos se encuentran en los cinco módulos de este sistema; así como el problema de diversidad en las fuentes de datos y los datos desfasados temporalmente que se encuentran en al menos 4 de estos. El último análisis arrojo como resultado que las causas altas que conllevan a la aparición de estos problemas se encuentran en la incorrecta aplicación de las reglas de catalogación angloamericanas, entrada manual de datos, migración de datos, información redundante, metadatos incompletos y desactualizaciones de datos.

CONCLUSIONES

1. Se describieron los principales algoritmos y herramientas que se utilizan en cada una de las fases de la metodología TDQM, la cual sirvió de guía para analizar la calidad de datos en casos de estudio reales.
2. Se describió *big data* de acuerdo a sus clasificaciones, su relación con la nube informática y las principales herramientas que se utilizan en cada etapa, permitieron conocer como ha llegado a convertirse en un reto para organizaciones que intentan corregir diversos problemas internos, aplicando soluciones *big data* a sus casos de estudio reales para llevar a cabo una correcta toma de decisiones.
3. Se diseñó una arquitectura integradora que muestra la relación existente entre las fases de *big data* con la nube y el proceso de calidad de datos, de manera que este último fue ubicado en la etapa de pre-procesamiento de datos para aprovechar las ventajas de las tecnologías de procesamiento en paralelo de big data y garantizar que los datos estén limpios antes de pasar a la etapa de análisis de datos.
4. Se caracterizó el sistema ABCD en cuanto a su misión, arquitectura, módulos de trabajo, fuentes de datos y alcance que posee, lo cual permitió realizar un mejor diseño de las encuestas aplicadas a los especialistas.
5. Se procesaron estadísticamente las encuestas aplicadas, las cuales arrojaron los siguientes resultados:
    a. Se identificó que todas las dimensiones que abarca la encuesta son altamente importantes para los especialistas, en correspondencia al trabajo con los datos de la suite ABCD.
    b. Se identificaron como los módulos más conocidos y utilizados por los especialistas el de Catalogación y EmpWeb.
    c. Se detectaron los problemas que más afectan los módulos principales del ABCD donde los errores ortográficos y las inseguridades en los datos se encuentran en los cinco módulos de este sistema.
    d. Se identificaron como causas altas que conllevan a la aparición de problemas en los datos la incorrecta aplicación de las reglas de catalogación angloamericanas, entrada manual de datos, migración de datos, información redundante, metadatos incompletos y desactualizaciones de datos.

RECOMENDACIONES

1. Continuar perfeccionando el diseño de las encuestas aplicadas trabajando en las particularidades de cada etapa en pos de obtener una mejor evaluación global.
2. Aplicar la encuesta diseñada en otras bibliotecas y centros de documentación del país para obtener resultados más verídicos.
3. Implementar herramientas de perfilado y limpieza de datos basadas en los resultados arrojados en las encuestas, con el propósito de entregar datos de mejor calidad y de esta manera completar el ciclo correspondiente a la metodología TDQM para el análisis de calidad de datos en la automatización de bibliotecas y centros de documentación.

REFERENCIAS BIBLIOGRÁFICAS

ABADI, D. J., BONCZ, P. A. & HARIZOPOULOS, S. 2009. Column-oriented database systems. *Proceedings of the VLDB Endowment,* 2**,** 1664-1665.

ABELLÓ, A., DARMONT, J., ETCHEVERRY, L., GOLFARELLI, M., MAZÓN LÓPEZ, J. N., NAUMANN, F., PEDERSEN, T. B., RIZZI, S., TRUJILLO MONDÉJAR, J. C. & VASSILIADIS, P. 2013. Fusion cubes: towards self-service business intelligence.

AMÓN, I., MORENO, F. & ECHEVERRI, J. 2012. Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español. *Revista Ingenierías Universidad de Medellín,* 11**,** 127-138.

ANANTHAKRISHNA, R., CHAUDHURI, S. & GANTI, V. Eliminating Fuzzy Duplicates in Data Warehouses. Proceedings of the 28th VLDB Conference, 2002 Hong Kong,China.

ARASU, A., CHAUDHURI, S. & KAUSHIK, R. Learning String Trasnformations From Examples. VLDB' 09, August 24-28 2009 Lyon, France.

ARENAS, M., BERTOSSI, L. & CHOMICKI, J. Consistent query answers in inconsistent databases. Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 1999. ACM, 68-79.

ARIAS CHALICO, T., RIEGELHAUPT, E. & FAO, R. 2002. Guía para encuestas de demanda, oferta y abastecimiento de combustibles de madera. FAO.

ATZENI, P. & DE ANTONELLIS, V. 1993. *Relational database theory*, Benjamin-Cummings Publishing Co., Inc.

BAHRAMI, M. & SINGHAL, M. 2015. The Role of Cloud Computing Architecture in Big Data. *Information Granularity, Big Data, and Computational Intelligence.* Springer.

BALLOU, D., WANG, R., PAZER, H. & TAYI, G. K. 1998. Modeling information manufacturing systems to determine information product quality. *Management Science,* 44**,** 462-484.

BALLOU, D. P. & PAZER, H. L. 1985. Modeling data and process quality in multi-input, multi-output information systems. *Management science,* 31**,** 150-162.

BARATEIRO, J. & GALHARDAS, H. 2005. A Survey of Data Quality Tools. *Datenbank-Spektrum,* 14**,** 48.

BATINI, C., CAPPIELLO, C., FRANCALANCI, C. & MAURINO, A. 2009. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR),* 41**,** 16.

BATINI, C. & SCANNAPIECO, M. 2006. *Data quality: concepts, methodologies and techniques,* Springer.

BEYER, M. 2011. Gartner says solving big data challenge involves more than just managing volumes of data. Gartner. *2011) http://www. gartner. com/it/page. jsp*.

BEYER, M. A. & LANEY, D. 2012. The importance of 'big data': a definition. *Stamford, CT: Gartner*.

BILENKO, M. & MOONEY, R. J. 2003. On evaluation and training-set construction for duplicate detection. *Proc. of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation***,** 7-12.

BOBROWSKI, M., MARRÉ, M. & YANKELEVICH, D. 1999. Measuring data quality. *Universidad de Buenos Aires. Report***,** 99-002.

BOLLIER, D. & FIRESTONE, C. M. 2010. *The promise and peril of big data*, Aspen Institute, Communications and Society Program Washington, DC, USA.

BOVEE, M., SRIVASTAVA, R. & MAK, B. A conceptual framework and belief-function approach to assessing overall information quality. Proceedings of the 6th International Conference on Information Quality, 2001.

CALERO, E., GUEVARA, D., CABALLERO, D. & FRÓMETA, M. 2011. Desarrollo del ABCD en la biblioteca del Centro Provincial de Información de Ciencias Médicas Camagüey.

CALı, A., CALVANESE, D., DE GIACOMO, G. & LENZERINI, M. 2004. Data integration under integrity constraints. *Information Systems,* 29**,** 147-163.

CARREIRA, P. & GALHARDAS, H. Efficient development of data migration transformations. Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 2004. ACM, 915-916.

CODD, E. F. 1982. Relational database: a practical foundation for productivity. *Communications of the ACM,* 25**,** 109-117.

COHEN, W. W., RAVIKUMAR, P. & FIENBERG, S. E. A Comparison of String Distance Metrics for Name-Matching Tasks. Proceedings of II Web, 2003. 73--78.

COX, M. & ELLSWORTH, D. Managing big data for scientific visualization. ACM Siggraph, 1997. 21.

CUI, Y. 2001. *Lineage tracing in data warehouses.* Stanford University.

CHANG, F., DEAN, J., GHEMAWAT, S., HSIEH, W. C., WALLACH, D. A., BURROWS, M., CHANDRA, T., FIKES, A. & GRUBER, R. E. 2008. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS),* 26**,** 4.

CHAPMAN, A. D. 2005. *Principles of data quality,* GBIF.

CHAUDHURI, S. & DAYAL, U. 1997. An overview of data warehousing and OLAP technology. *ACM Sigmod record,* 26**,** 65-74.

CHAUDHURI, S., GANTI, V. & XIN, D. 2009. *Exploiting web search to generate synonyms for entities,* Madrid, Spain, ACM.

CHEN, C. C. & TSENG, Y.-D. 2011. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems,* 50**,** 755-768.

CHEN, M., MAO, S. & LIU, Y. 2014. Big data: A survey. *Mobile Networks and Applications,* 19**,** 1-39.

CHEN, Y., ALSPAUGH, S. & KATZ, R. 2012. Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads. *Proceedings of the VLDB Endowment,* 5**,** 1802-1813.

DAZA, A., DE LA TORRE, P., ZEPEDA, V. V. & VILLEGAS, C. M. 2012. Hacia un Modelo de Madurez para la Gestión de Calidad de Datos en Inteligencia de Negocios. *Links*.

DE SMET, E. & SPINAK, E. 2009. The abc of ABCD: the Reference Manual.

DEAN, J. & GHEMAWAT, S. 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM,* 51**,** 107-113.

DECANDIA, G., HASTORUN, D., JAMPANI, M., KAKULAPATI, G., LAKSHMAN, A., PILCHIN, A., SIVASUBRAMANIAN, S., VOSSHALL, P. & VOGELS, W. Dynamo: amazon's highly available key-value store. ACM SIGOPS Operating Systems Review, 2007. ACM, 205-220.

DEMING, W. E. 1986. Out of the crisis, Massachusetts Institute of Technology. *Center for advanced engineering study, Cambridge, MA,* 510.

DEMIRKAN, H. & DELEN, D. 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems,* 55**,** 412-421.

DOGGETT, A. M. 2005. Root cause analysis: A framework for tool selection. *Quality Management Journal,* 12**,** 34.

DUNN, H. L. 1946. Record linkage. *American Journal of Public Health and the Nations Health,* 36**,** 1412-1416.

ELSE, W. I. 2002. *The Complete Soundex Guide: Discovering the Rules Used by the Census Bureau and the Immigration and Naturalization Service when Those Organizations Indexed Federal Records; an Unofficial Supplement to National Archives' Federal Population Census and Immigrant and Passenger Arrivals Publications and More,* Closson Press.

EMRAN, N. A., ABDULLAH, N. & MUSTAFA, N. 2012. A Review of Failure Handling Mechanisms for Data Quality Measures.

EPPLER, M. J. & WITTIG, D. Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years. IQ, 2000. 83-96.

ESTADÍSTICAS, I. N. D. 2011. *Informe de calidad de datos. VI Encuesta Nacional Urbana de Seguridad Ciudadana*

FELLEGI, I. P. & HOLT, D. 1976. A systematic approach to automatic edit and imputation. *Journal of the American Statistical association,* 71**,** 17-35.

FISHER, C. W., CHENGALUR-SMITH, I. & BALLOU, D. P. 2003. The impact of experience and time on the use of data quality information in decision making. *Information Systems Research,* 14**,** 170-188.

FISHER, C. W., LAURIA, E. J. & MATHEUS, C. C. 2009. An accuracy metric: Percentages, randomness, and probabilities. *Journal of Data and Information Quality (JDIQ),* 1**,** 16.

FOX, C., LEVITIN, A. & REDMAN, T. 1994. The notion of data and its quality dimensions. *Information processing & management,* 30**,** 9-19.

FRANCALANCI, C. & PERNICI, B. Data quality assessment from the user's perspective. Proceedings of the 2004 international workshop on Information quality in information systems, 2004. ACM, 68-73.

FRANKS, B. 2012. *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics,* John Wiley & Sons.

GÁLVEZ, C. 2006. Identificación de nombres personales por medios de sistemas de codificación fonética. *Encontros Bibli,* Segundo Semestre**,** 105-116.

GANTZ, J. & REINSEL, D. 2011. Extracting value from chaos. IDC Iview. Framingham: IDC Goto-Market Services.

GELBUKH, A., JIMENEZ, S., BECERRA, C. & GONZÁLEZ, F. Generalized Monge-Elkan method for approximate text string comparision. 10th International Conference on Computational Linguisticsand Intelligent Text Processing, 2009. 559-570.

GILL, L. 1997. OX-LINK: the Oxford medical record linkage system. Citeseer.

GONZÁLEZ, Y. I. 2010. Descubrimiento de llaves foráneas en colecciones de datos: reglas de inclusión. *Serie Científica,* 3.

GROBELNIK, M. 2012. Big data tutorial.

HALL, P. & DOWLING, G. 1980. Approximate string matching. *ACM Computing Surveys,* 12**,** 381-402.

HASHEM, I. A. T., YAQOOB, I., ANUAR, N. B., MOKHTAR, S., GANI, A. & KHAN, S. U. 2014. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems,* 47**,** 98-115.

HERNÁNDEZ, M. A. & STOLFO, S. J. 1998. Real world Data is Dirty: Data Cleansing and The Merge-Purge Problem. *Journal of data mining and Knowlwgde Discovery,* 2.

HUANG, K.-T., LEE, Y. W. & WANG, R. Y. 1998. *Quality information and knowledge*, Prentice Hall PTR.

HURWITZ, J. N., A. ; HALPER, F. ; KAUFMAN, M. 2013. Big data for dummies.

IBM 2013. *What is big data? Bringing big data to the enterprise,* http://www-01.ibm.com/software/data/bigdata/.

ISO, I. O. F. S. & ANSI, A. N. S. I. 1999. *ISO International Standard: Database Language SQL – part 2*.

JARKE, M., LENZERINI, M., VASSILIOU, Y. & VASSILIADIS, P. 1995. *Fundamentals of Data Warehouses*, Springer Verlag.

JEBAMALAR-TAMILSELVI, J. & SARAVANAN, V. 2008. A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse. *International Journal of Computer Science and Network Security*, 8.

JI, C., LI, Y., QIU, W., AWADA, U. & LI, K. Big data processing in cloud computing environments. Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on, 2012. IEEE, 17-23.

JURAN, J. M. & GRYNA, F. M. 1980. *Quality planning and analysis*, McGraw-Hill.

KAISLER, S., ARMOUR, F., ESPINOSA, J. A., MONEY, W. & WASHINGTON, G. 2012. Big Data: Issues and Challenges Moving Forward.

KIM, W., CHOI, B.-J., HONG, E.-K., KIM, S.-K. & LEE, D. 2003. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7**,** 81-99.

KIMBAL, R., REEVES, L., ROSS, M. & THORNTHWAITE, W. 1998. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons.

KORONIOS, A., LIN, S. & GAO, J. A data quality model for asset management in engineering organisations. IQ, 2005.

KOVAC, R., LEE, Y. W. & PIPINO, L. Total Data Quality Management: The Case of IRI. IQ, 1997. 63-79.

LANEY, D. 2001. 3-d data management: Controlling data volume, velocity and variety. META Group, Research Note.

LEE, M. L., LU, H., LING, T. W. & KO, Y. T. 1999. Cleansing data for mining and datawarehousing. 10.

LEE, Y. W., PIPINO, L. L., FUNK, J. D. & WANG, R. Y. 2006. Journey to Data Quality. *The MIT Press, ISBN,* 262122871**,** 36.

LEHTI, P. & FANKHAUSER, P. 2005. A Precise Blocking Method for Record Linkage. *Lecture Notes Computer Science,* 3589**,** 210-220.

LEVENSHTEIN, V. I. 1965. Binary Code capable of correcting deletions, insertions and reversal. *Soviet Physics Doklady,* 163**,** 845-848.

LI, C., WANG, B. & YANG, X. VGRAM: Improving Performance of Approximate Queries on String Collections Using Variable-Length Grams. *In:* ACM, ed. VLDB'07, September 23-28 2007 Vienna, Austria.

LIU, F., TONG, J., MAO, J., BOHN, R., MESSINA, J., BADGER, L. & LEAF, D. 2011. NIST Cloud Computing Reference Architecture. *NIST Special Publication,* 292-500.

LIU, L. & CHI, L. Evolutionary data quality. Proceedings of the 7th international conference on information quality (IQ), MIT, Cambridge, USA, 2002.

LYMAN, P. & VARIAN, H. 2004. How much information 2003?

MADNICK, S. E., WANG, R. Y., LEE, Y. W. & ZHU, H. 2009. Overview and framework for data and information quality research. *Journal of Data and Information Quality (JDIQ),* 1**,** 2.

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C. & BYERS, A. 2011. Big data: The next frontier for innovation, competition, and productivity. Report, McKinsey Global Institute.

MARCUS, A. & MALETIC, J. I. 2005. *Cap. 2 A Prelude to Knowledge Discovery. The Data Mining and Knowledge Discovery Handbook*, Springer.

MARSHALL, B. 2007. Data quality and data profiling: a glossary of terms.

MAYDANCHIK, A. 2007. *Data quality assessment*, Technics publications.

MAYER-SCHÖNBERGER, V. & CUKIER, K. 2013. *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt.

MCAFEE, A. B., ERIK 2012. Big data: the management revolution. 60-66.

MCNULTY, E. 2014. *Understanding Big Data: The Seven V's,* http://www.dataconomy.com/understanding-big-data/.

MELL, P. & GRANCE, T. 2011. The NIST definition of cloud computing.

MILLER, H. 2013. Big-Data in Cloud Computing: A Taxonomy of Risks.

MOGES, H.-T., DEJAEGER, K., LEMAHIEU, W. & BAESENS, B. 2013. A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management,* 50**,** 43-58.

MONGE, A. E. & ELKAN, C. P. 1996. The field matching problem: Algorithms and applications. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,* 267--270.

MONGE, A. E. & ELKAN, C. P. 1997. An efficient domain-independent algorithm for detecting approximately duplicate database tuples. *Proceedings of the SIGMOD 1997 workshop on data mining and knowledge discovery*.

MOORE, D. S. & MCCABE, G. P. 1989. *Introduction to the Practice of Statistics*, WH Freeman/Times Books/Henry Holt & Co.

MORENO, J. P. 2014. Una aproximación a Big Data. An approach to Big Data. *Revista de Derecho de la UNED (RDUNED),* 471-506.

MORROS, R. S. & PICAÑOL, J. S. 2013. *Big data: Análisis de herramientas y soluciones*.

MÜLLER, H. & FREYTAG, J.-C. 2005. *Problems, methods, and challenges in comprehensive data cleansing*, Professoren des Inst. Für Informatik.

MÜLLER, H. & FREYTAG, J. C. 2003. Problems, Methods, and Challenges in Comprehensive Data Cleansing. *Technical Report HUB-IB-164, Humboldt University Berlin*

NAUMANN, F. 2002. *Quality-driven query answering for integrated information systems*, Springer Science & Business Media.

NEUBAUER, P. 2010. Graph databases, NOSQL and Neo4j.

NEUMEYER, L., ROBBINS, B., NAIR, A. & KESARI, A. S4: Distributed stream computing platform. Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, 2010. IEEE, 170-177.

NEWCOMBE, H. B. & KENNEDY, J. M. 1962. Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM,* 5**,** 563-566.

NIST 2012. *Big Data Working Group (NBD-WG),* http://bigdatawg.nist.gov/home.php.

OLIVEIRA, P., RODRIGUES, F., HENRIQUES, P. & GALHARDAS, H. A taxonomy of data quality problems. 2nd Int. Workshop on Data and Information Quality, 2005. 219-233.

OLSON, J. E. 2003. *Data quality: the accuracy dimension*, Morgan Kaufmann.

PANDEY, R. K. 2014. Data Quality in Data warehouse: problems and solution.

PATMAN, F. & SHAEFER, L. 2003. Is Soundex Good Enough for You? On the Hidden Risks of Soundex-Based Name Searching© 2001-2003 Language Analysis Systems. Inc.

PHILIPS, L. 1990. Hanging on the metaphone. *Computer Language,* 7.

PHILIPS, L. 2000. The double metaphone search algorithm. *C/C++ users journal,* 18**,** 38-43.

PIPINO, L. L., LEE, Y. W. & WANG, R. Y. 2002. Data quality assessment. *Communications of the ACM,* 45**,** 211-218.

PORRERO, B. L. 2011. *Limpieza de datos: Remplazo de valores ausentes y estandarización*.

PRICE, R. & SHANKS, G. 2011. The impact of data quality tags on decision-making outcomes and process. *Journal of the Association for Information Systems,* 12**,** 1.

QUACKENBUSH, J. 2002. Microarray data normalization and transformation. *Nature genetics,* 32**,** 496-501.

R.L. VILLARS, C. W. O., M. EASTWOOD 2011. Big data: what it is and why you should care. *IDC*.

RAGHAVAN, H. & ALLAN, J. Using soundex codes for indexing names in ASR documents. Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004, 2004. Association for Computational Linguistics, 22-27.

RAHM, E. & DO, H. H. 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.,* 23**,** 3-13.

RAMAN, V. & HELLERSTEIN, J. M. 2001. Potter's Wheel: An Interactive Data Cleaning System. *The VLDB Journal***,** 381-390.

RAO, B., SALUIA, P., SHARMA, N., MITTAL, A. & SHARMA, S. Cloud computing for Internet of Things & sensing based applications. Sensing Technology (ICST), 2012 Sixth International Conference on, 2012. IEEE, 374-380.

REDMAN, T. C. 1996. Data quality for the information age. Boston: Norwood: Artech House.

REDMAN, T. C. 1998. The impact of poor data quality on the typical enterprise. *Communications of the ACM,* 41**,** 79-82.

REMA, A., VIJIL, C., PRASAD, M. D. & RAGHURAM, K. 2008. *Rule based synonyms for entity extraction from noisy text*, Singapore, ACM.

SADIQ, S., YEGANEH, N. K. & INDULSKA, M. 20 years of data quality research: themes, trends and synergies. Proceedings of the Twenty-Second Australasian Database Conference-Volume 115, 2011. Australian Computer Society, Inc., 153-162.

SAHA, B. & SRIVASTAVA, D. Data quality: The other face of Big Data. Data Engineering (ICDE), 2014 IEEE 30th International Conference on, 2014. IEEE, 1294-1297.

SALOMONE, S., HYLAND, P. & MURPHY, G. D. Perceptions of data quality dimensions and data roles. 25th ANZAM Conference, 2011, 2011.

SATTLER, K. U. & SCHALLEHN, E. 2001. A Data Preparation Framework based on a Multidatabase Language. *International Database Engineering Applications Symposium (IDEAS).*

SCANNAPIECO, M. & CATARCI, T. 2002. Data quality under a computer science perspective. *Archivi & Computer,* 2**,** 1-15.

SCHERBINA, A. 2005. Clustering of Web Access Sessions *Lecture Notes in Computer Science*.

SEEGER, M. & ULTRA-LARGE-SITES, S. 2009. Key-Value stores: a practical overview. *Computer Science and Media, Stuttgart*.

SERRAT MORROS, R. 2013. Big Data: análisis de herramientas y soluciones.

SESSIONS, V. & VALTORTA, M. 2009. Towards a method for data accuracy assessment utilizing a Bayesian network learning algorithm. *Journal of Data and Information Quality (JDIQ),* 1**,** 14.

SHANKARANARAYANAN, G. & CAI, Y. 2006. Supporting data quality management in decision-making. *Decision Support Systems,* 42**,** 302-317.

SHANKARANARAYANAN, G. & ZHU, B. Data quality metadata and decision making. System Science (HICSS), 2012 45th Hawaii International Conference on, 2012. IEEE, 1434-1443.

SINGHAL, M. 2013. A client-centric approach to interoperable clouds. *Int. J. Soft Comput. Softw. Eng.[JSCSE],* 3**,** 3-4.

SMITH, T. F. & WATERMAN, M. S. 1981. Identification of common molecular subsequences. *Journal Molecular Biology***,** 195-197.

STANIER, A. 1990. How accurate is Soundex matching. *Computers in Genealogy,* 3**,** 286-288.

STRONG, D. M., LEE, Y. W. & WANG, R. Y. 1997. Data quality in context. *Communications of the ACM,* 40**,** 103-110.

SUZANNE, M. E., BRANDT, S. M., ROBINSON, J. S., SUTHERLAND, I., BISBY, F. A., GRAY, W. A., JONES, A. C. & WHITE, R. J. 2001. Adapting integrity enforcement techniques for data reconciliation. *Inf. Syst.,* 26**,** 657-689.

SYED, A. R., GILLELA, K. & VENUGOPAL, C. 2013. The Future Revolution on Big Data. *Future,* 2.

TAFT, R. L. 1970. *Name search techniques,* Bureau of Systems Development, New York State Identification and Intelligence System.

TALIA, D. 2013. *Clouds for scalable big data analytics*.

TAYI, G. K. & BALLOU, D. P. 1998. Examining data quality. *Communications of the ACM,* 41**,** 54-57.

TEAM, O. 2011. Big data now: current perspectives from OReilly Radar. *OReilly Media*.

TEE, J. 2013. *Handling the four 'V's of big data: volume, velocity, variety, and veracity*, TheServerSide.com.

TÜRKER, C. & GERTZ, M. 2001. Semantic integrity support in SQL: 1999 and commercial (object-) relational database management systems. *The VLDB Journal,* 10**,** 241-269.

UKKONEN 1992. Approximate string matching with q-grams and maximal matches. *Theoretical Computer Science***,** 191-211.

VENKATRAMAN, N. & HENDERSON, J. C. 1998. Real strategies for virtual organizing. *MIT Sloan Management Review,* 40**,** 33.

WAND, Y. & WANG, R. Y. 1996. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM,* 39**,** 86-95.

WANG, R. & LEE, Y. 1998. Integrity Analyzer: A Software Tool for Total Data Quality Management. Cambridge Research Group, Cambridge, MA.

WANG, R. Y. 1998. A product perspective on total data quality management. *Communications of the ACM,* 41**,** 58-65.

WANG, R. Y., STOREY, V. C. & FIRTH, C. P. 1995. A framework for analysis of data quality research. *Knowledge and Data Engineering, IEEE Transactions on,* 7**,** 623-640.

WANG, R. Y. & STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems***,** 5-33.

WATTS, S., SHANKARANARAYANAN, G. & EVEN, A. 2009. Data quality assessment in context: A cognitive perspective. *Decision Support Systems,* 48**,** 202-211.

WATTS, S. & ZHANG, W. 2004. *Knowledge adoption in online communities of practice*, System d'Information et Management.

WIJNHOVEN, F., BOELENS, R., MIDDEL, R. & LOUISSEN, K. 2007. Total Data Quality Management: A Study of Bridging Rigor and Relevance.

WINGKVIST, A., ERICSSON, M., LINCKE, R. & LOWE, W. A metrics-based approach to technical documentation quality. Quality of Information and Communications Technology (QUATIC), 2010 Seventh International Conference on the, 2010. IEEE, 476-481.

WINKLER, W. E. 2006. Overview of record linkage and current research directions *Technical Report RR2006/02*.

WINKLER, W. E. & CENSUS, U. S. B. O. T. 1993. Improved decision rules in the Fellegi-Sunter model of record linkage.

YEGANEH, N. K., SADIQ, S. & SHARAF, M. A. 2014. A framework for data quality aware query systems. *Information Systems*.

ZAHEDI, F. 1995. *Quality information systems*, Boyd & Fraser Publishing Co.

ZEEH, C. The lempel ziv algorithm. URL: http://w3studi. informatik. uni-stuttgart. de/~ zeehca/Seminar/LempelZivReport. pdf, 2003.

ZHU, X. & GAUCH, S. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000. ACM, 288-295.

ZOBEL, J. & DART, P. Phonetic String Matching: Lessons from Information Retrieval. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 1996.

ANEXOS

Anexo 1. Cuestionario para evaluar la calidad de datos en el sistema de Automatización de Bibliotecas y Centros de Documentación (ABCD).

Sección 1: Preguntas generales sobre datos del encuestado

Marque con una cruz (X)

1. Género: _____ Femenino _____ Masculino

2. Nivel educacional alcanzado:
Nota: ***Marque solamente el más elevado.***
_____ Secundaria Básica
_____ Técnico Medio
_____ Preuniversitario
_____ Superior o Universitario

2.1 ¿Qué título o diploma académico obtuvo en el último nivel aprobado?
(Ejemplo: Técnico Medio en Bibliotecología, Lic. en Ciencias de la Información, Ing. Informático, etc.)

_____

3. ¿Posee algún grado científico?
Nota: ***En caso afirmativo, marque y especifique en qué especialidad (Ejemplo: Especialista en Recursos Humanos, Máster en Ciencias de la Computación, etc.)***
_____ Sí _____ No

3.1 Grado científico:
_____ Especialista _____ Máster _____ Doctor

3.2 Especialidad: _____

4. Años de experiencia laboral en la biblioteca:
_____ Menos de un año
_____ Entre 1 y 5 años
_____ Entre 6 y 10 años
_____ Más de 10 años

5. ¿Qué responsabilidad/cargo u ocupación tiene en la biblioteca?
_____ Especialista
_____ Especialista Principal
_____ Técnico
_____ Otro (Especificar) _____

6. Experiencia con el ABCD:

_____ Menos de un año
_____ Entre 1 y 2 años
_____ Entre 2 y 3 años
_____ Entre 3 y 4 años
_____ 4 o más

6.1 Marque el o los módulos del ABCD con los que ha trabajado:

_____ Módulo de Catalogación
_____ Módulo de Préstamos
_____ Módulo de Adquisiciones
_____ Módulo EmpWeb
_____ Módulo SecsWeb

6.2 Marque cuál o cuáles bases de datos del ABCD ha consultado:

_____ MARC
_____ Adquisiciones y Copias
_____ Proveedores
_____ Sugerencias
_____ Órdenes de Compras
_____ Usuarios de Préstamos
_____ Objeto de Préstamo
_____ Reservaciones
_____ Transacciones de Préstamo
_____ Suspensiones y Multas
_____ Usuarios del Sistema

7. ¿Con qué frecuencia consulta el sistema ABCD?

_____ Ocasionalmente
_____ Una vez al mes
_____ Una vez a la semana
_____ Varios días en la semana
_____ Diariamente

8. ¿Cuál es el rol principal que usted ocupa en el sistema ABCD?

_____ Jefe de Servicio
_____ Responsable de la Catalogación
_____ Responsable de la Adquisición
_____ Responsable de los Préstamos
_____ Administrador del Sistema
_____ Otro (Especificar) _____

9. El papel principal de usted en relación a las fuentes de datos del ABCD es:
***Nota:* Puede marcar más de una opción en caso de realizar diferentes operaciones.**

_____ Insertar datos manualmente
_____ Consultar datos (Búsquedas)
_____ Editar datos
_____ Importar datos
_____ Exportar datos
_____ Eliminar datos
_____ Crear relaciones entre los datos
_____ Otro (Especificar) _____

10. ¿Conoce usted las reglas de catalogación angloamericanas que se aplican a la descripción bibliográfica de los distintos tipos de documentos del ABCD?
_____ Sí _____ No

10.1 En caso afirmativo especifique cómo aplica dichas reglas:
_____ Las conoce de memoria _____ Todas _____ Algunas
_____ Consulta los documentos donde se definen estas reglas
_____ Consulta a cierta persona que las conoce de memoria
_____ Utiliza alguna herramienta informática que las trae implementadas
_____ Otra vía (Especificar) _____

Sección 2: Conocimiento sobre calidad de datos

1. Cuando piensa en el término "*calidad de datos*" ¿Qué atributos o dimensiones considera usted necesarios para la realización de su trabajo con los datos del ABCD? y ¿Qué nivel de importancia le atribuye usted a dichos datos en correspondencia con las dimensiones?
***Nota:* Marque con una cruz (un solo valor) el nivel de importancia que le otorga a cada dimensión dentro del rango de 1 a 7. Donde 1 significa MENOR GRADO DE IMPORTANCIA, 4 MEDIO y 7 MÁXIMO.**

- ***Opcionalmente puede agregar otra(s) dimensión(es) de calidad al final de la tabla. Para ello debe especificar su definición y nivel de importancia.***

Dimensión	Definición	1	2	3	4	5	6	7
Exactitud	Grado en el cual los datos son certificados, libre de errores, correctos y confiables							
Precisión	Grado en el cual los datos tienen el nivel de detalle requerido							
Completitud	Grado en el cual no existen valores ausentes o perdidos, cubren las necesidades de las tareas, y son lo suficientemente amplios y profundos para una tarea determinada							
Consistencia	Grado en el cual el formato y contenido de los datos es el mismo a través de múltiples sistemas fuentes							
Actualidad	Medida de cuan actuales son los datos, basado en cuánto tiempo hace desde que estos fueron almacenados							

Volatilidad	Frecuencia con la que puede cambiar el valor de un dato							
Credibilidad	Grado en el cual los datos son considerados verdaderos y creíbles							
Objetividad	Grado en el cual los datos son objetivos, imparciales y basados en hechos							
Reputación	Grado de aceptación en el cual el uso de datos de una fuente concreta por parte de otras proporciona una evidencia, ya sea en sentido positivo o negativo, de la reputación de esa fuente concreta							
Valor agregado	Grado en el cual los datos son beneficiosos y proporcionan ventajas de su uso. Los datos adicionan valores a las operaciones							
Relevancia	Grado en el cual los datos son aplicables y útiles para una tarea determinada							
Cantidad apropiada de datos	Grado en el cual el volumen de información es adecuado para una tarea determinada							
Interpretabilidad	Grado en el cual los datos están en un lenguaje apropiado y las definiciones son claras							
Facilidad de comprensión	Grado en el cual los datos son comprendidos fácilmente							
Consistencia representacional	Grado en el cual los datos son continuamente representados en el mismo formato							
Representación concisa	Grado en el cual los datos son representados de forma compacta, bien organizados y estéticamente agradables							
Accesibilidad	Grado en el cual los datos están disponibles o son fáciles y rápidos de recuperar							
Unicidad	Grado en el cual cada valor es único en su dominio. No existen duplicados							
Seguridad	Grado en el cual el acceso a los datos está restringido apropiadamente para mantener su seguridad							
Utilidad	Grado en el cual cada dato del sistema debe satisfacer ciertos requerimientos de los usuarios							
Preservación	Grado en el cual se preservan los datos ante cambios de hardware, software, formatos y procesos a los que los datos son sometidos (migración)							
Comodidad	Grado en el cual los caminos de navegación (relaciones entre las tablas) resultan difíciles. Número de caminos de navegación perdidos/interrumpidos							
Conformidad	Grado en el cual los datos son almacenados siguiendo un formato estandarizado							

Sección 3: Problemas de calidad de datos

1. ¿Cuáles son los principales problemas de calidad de datos que presentan los diferentes módulos que conforman el ABCD?

Nota: ***Puede marcar varios de estos problemas y relacionarlos con el o los módulos afectados. En los casos que reconozca un problema y no sepa el módulo afectado, marcar la opción NSCM o la opción NRUP en caso que considere que ningún módulo presenta dicho problema.***

- ***Opcionalmente, puede agregar otro(s) problema(s) al final de la tabla.***

MA*: Módulo Adquisición,* ***MC****: Módulo Catalogación,* ***MP****: Módulo Préstamos,* ***MEW****: Módulo EmpWeb,* ***MSW****: Módulo SecsWeb,* ***NSCM****: No Se Conoce el Módulo,* ***NRUP****: No Representa Un Problema*

Problema	MA	MC	MP	MEW	MSW	NSCM	NRUP
Datos no estandarizados							
Datos incompletos, ausentes							
Datos duplicados							
Datos ambiguos							
Diversidad en las fuentes de datos							
Datos desfasados temporalmente							
Inseguridades en los datos							
Inconsistencias entre diferentes copias del mismo dato							
Datos con ruido (erróneos, incorrectos)							
Violaciones de restricciones de integridad							
Formato de datos incompatibles							
Valores nulos o vacíos							
Metadatos no claros							
Inexactitud en los datos							
Valores ficticios en los campos							
Campos multipropósitos							
Datos colgantes							
Tipo de datos incorrecto							
Valor de datos incorrecto							
Conflictos de nombre							
Conflictos estructurales							
Falta de datos en un campo no nulo							
Errores ortográficos							
Valores implícitos							
Datos perdidos							
Campos que no siguen un formato estandarizado							

2. ¿Existe en su departamento algún equipo de especialistas en calidad de datos?

_____ Sí _____ No

2.1 Como realizan la gestión de la calidad de datos:

_____ De forma manual

_____ Utilizando alguna herramienta computacional (Especificar) _____

3. A continuación se muestran una serie de causas probables que pudieran conllevar a la aparición de problemas de calidad de datos en el ABCD. Escriba un valor del 1 al 3, donde 1 representa una causa **BAJA**, 2 **MEDIA** y 3 **ALTA** para cuantificar la magnitud del problema. Marque **NCC** si considera que el problema No Constituye una Causa.

Nota: ***Puede agregar otra(s) causa(s) que considere conlleve(n) a la aparición de problemas de calidad.***

Causa	1	2	3	NCC
Entrada manual de datos: introducir datos en el sistema de forma manual				
Migración de datos: transferencia de materiales digitales de un origen de datos a otro				
Integración de datos: proceso de combinar datos que residen en diferentes fuentes y permitirle al usuario final tener una vista unificada de todos sus datos.				
Incorrecta aplicación de las reglas de catalogación angloamericanas				
Uso de diferentes formatos de codificación: CEPAL, MARC 21, entre otros				
Procesamiento de datos: cambio en los programas responsables del procesamiento regular de datos (recolección y manipulación de elementos de datos para producir información significativa)				
Limpieza masiva de datos: uso de reglas automatizadas de limpieza de datos para hacer correcciones en masas sin tener en cuenta las particularidades de cada dato				
Purga de datos: eliminación rutinaria de datos antiguos del sistema para dar paso a datos más nuevos				
Cambios no capturados: se han producido diferentes cambios en la biblioteca, pero estos no han sido capturados en el sistema				
Actualizaciones del sistema: actualizaciones en el software del sistema cada cierto tiempo				
Complejidad: complejidad en los sistemas de almacenamiento				
Metadatos incompletos: existencia de gran cantidad de conocimiento sobre los datos en la mente del personal que trabaja con estos, en lugar de existir información sobre los datos bibliográficos (metadatos) de los documentos				
Automatización de procesos: con el progreso de la tecnología, cada vez más tareas son automatizadas sin intervención humano				
Información redundante: información duplicada en las bases de datos del sistema				
Relaciones entre las tablas: creación manual de relaciones entre las tablas las cuales pueden llevar a la aparición de valores colgantes en las mismas				
Desactualizaciones de datos: falta de actualización en todas las réplicas de datos del sistema (Ejemplo: los diferentes ejemplares de un libro determinado)				