Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

DEFINICIÓN DE MÉTRICAS PARA ESTABLECER EL NIVEL DE CALIDAD DE DATOS NECESARIO EN UN AMBIENTE ANALÍTICO

Luis Antonio Gálvez Dávila
Asesorado por el Ing. Sergio Alonzo

Guatemala, noviembre de 2011

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DEFINICIÓN DE MÉTRICAS PARA ESTABLECER EL NIVEL DE CALIDAD DE DATOS NECESARIO EN UN AMBIENTE ANALÍTICO

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

LUIS ANTONIO GÁLVEZ DÁVILA
ASESORADO POR EL ING. SERGIO ALONZO

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, NOVIEMBRE DE 2011

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Murphy Olympo Paiz Recinos
VOCAL I	Ing. Alfredo Enrique Beber Aceituno
VOCAL II	Ing. Pedro Antonio Aguilar Polanco
VOCAL III	Ing. Miguel Ángel Dávila Calderón
VOCAL IV	Br. Juan Carlos Molina Jiménez
VOCAL V	Br. Mario Maldonado Muralles
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Murphy Olympo Paiz Recinos
EXAMINADOR	Ing. Marlon Antonio Pérez Türk
EXAMINADORA	Inga. Sonia Yolanda Castañeda Ramírez
EXAMINADORA	Inga. Floriza Ávila Pesquera de Medinilla
SECRETARIA	Inga. Marcia Ivónne Véliz Vargas

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DEFINICIÓN DE MÉTRICAS PARA ESTABLECER EL NIVEL DE CALIDAD DE DATOS NECESARIO EN UN AMBIENTE ANALÍTICO

Tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha enero de 2011.

Luis Antonio Gálvez Dávila

ACTO QUE DEDICO A:

Dios	Porque con su amor he logrado ser perseverante y gracias a su voluntad estoy hoy terminando esta empresa.
Mis padres	Por regalarme su sabiduría. Por confiar en mí y darme todo el apoyo necesario para culminar este camino de conocimientos, guiándome en cada una de las etapas. Por heredarme con amor esta plataforma en la que apoyaré mis principios y visión para hacer de Guatemala un mejor país para vivir.
Mis hermanos	Porque cada uno ha sabido tenderme la mano en los momentos bajos, y ha disfrutado a mi lado los altos.
Mi novia	Karen Cerón, por recordarme siempre que mi propósito de vida debe ser agradar a Dios. Por inspirarme.
Asesor	Sergio Alonzo, por transmitirme su experiencia y conocimientos, consejos y sabiduría a lo largo de mi corta carrera profesional, y su ayuda para terminar este trabajo final de graduación.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

GLOSARIO

Almacenamiento Digital	Información que puede ser transformada a *bits* y almacenada en un dispositivo electrónico.
Base de Datos Relacional	Una base de datos es un conjunto de datos almacenados en algún medio. Se convierte en relacional cuando los datos contenidos en la misma están relacionados de alguna manera entre sí.
Bit	Es la medida más pequeña que actualmente se utiliza para almacenar información. Es representada en informática por medio de un cero o un uno.
DBA	Por sus siglas en inglés *Database Administrator*, es el usuario administrador de la base de datos.
Dimensión	Perspectiva o punto de vista desde el cual puede ser analizado un sujeto de estudio. Se utilizan en ambientes analíticos de información para poner a las métricas en contexto.

ETL	(*Extraction, Transform and Load*) Significa extraer, transformar y cargar. Es el proceso por medio del cual los datos son transformados y trasladados desde un sistema transaccional hacia un sistema estructuralmente analítico.
KPI	(*Key Performance Indicator*) Representa a los indicadores clave para medición de rendimientos. Estos indicadores son una herramienta indispensable para orientar la administración de un negocio hacia estrategias.
Métrica	Es la información cuantitativa en un modelo estrella que puede ser sujeto de operaciones aritméticas y cálculos complejos de análisis. Las métricas alimentan a los KPIs.
Modelo Estrella	Estructura de la información en los modelos analíticos. Fue un aporte de Ralph Kimball, y sus elementos importantes están divididos en cubos o tablas de hecho y dimensiones.
Datos Nulos	Cuando un campo en un atributo de la base de datos viene vacío o sin valor alguno.
Tabla de Hechos o Cubo	Una tabla de hechos representa a la información que es sujeto de análisis y que puede ser vista desde uno o varios puntos de vista. Por lo general son métricas o cantidades.

RESUMEN

Debido al creciente uso de información digital en las empresas, para sobrellevar el día a día de sus procesos de negocio, este trabajo expone una forma práctica de medir el nivel de calidad de los datos en una base de datos empresarial. Los datos juegan un papel muy importante en el flujo general de procesos de una empresa. Cuando los datos no cumplen con un nivel de calidad adecuado el impacto puede ser alarmantemente negativo.

Hay estudios dedicados a comprobar y medir dicho impacto, pero de lo que trata esta investigación, es de establecer un método práctico y estructurado para realizar la medición de la calidad de datos.

Las empresas necesitan llevar un control de calidad de datos para establecer si los mismos poseen el nivel adecuado de calidad para su tipo especial de negocio. Esto también permite accionar los procesos necesarios para que alcancen el nivel adecuado, si el nivel encontrado resulta deficiente. El problema es que no se tiene una metodología establecida para medir el nivel de calidad de datos, ni las métricas definidas para hacerlo.

En los resultados de esta investigación se presenta una estructura flexible con sus métricas relevantes asociadas para que, dependiendo del tipo de negocio, sea posible tener la noción más clara de cuál es el nivel de calidad que poseen los datos actualmente.

OBJETIVOS

General

Definir una metodología estructurada para establecer las métricas relevantes en la medición del estado o nivel de calidad que un conjunto de datos empresariales posee. Para luego establecer (según la metodología definida) el nivel de calidad de ese conjunto de datos dentro de un ambiente analítico que soporte decisiones efectivas.

Específicos

1. Elegir un adecuado conjunto de datos que sirva como fuente de la metodología a establecer.
2. Perfilar el conjunto de datos elegido (en el inciso anterior) y encontrar las reglas de datos necesarias del negocio que servirán para efectuar el control de calidad del conjunto de datos.
3. Controlar la calidad de los datos fuente en base a las reglas de datos encontradas para establecer el porcentaje de cumplimiento.
4. Definir las métricas relevantes en este estudio, el nivel de calidad del conjunto de datos sometido a análisis en base a las reglas de datos y el cumplimiento de las mismas.

INTRODUCCIÓN

La humanidad se encuentra actualmente en la era de la información y los datos por consiguiente juegan un rol muy importante en las empresas. Los negocios que se dedican a la venta de productos y/o servicios encuentran en los datos a un aliado bastante útil, con ellos pueden llevar el control de clientes, productos, fechas, proveedores, canales de distribución y muchas otras características propias de cada negocio.

Que los datos tengan un nivel de calidad debajo de lo deseado puede repercutir en pérdidas seguras, por ejemplo, en el siguiente caso: un repartidor recibió una orden de entrega con la dirección errónea. Por la confusión él no podrá entregar dicha orden a tiempo, si es que finalmente la llega a entregar. El cliente, por consiguiente, obtendrá una sensación de insatisfacción por el mal servicio, y eventualmente hasta puede exigir la devolución de su dinero y nunca más volver a comprar en ese negocio.

La calidad de datos pretende asegurar que los datos de una empresa, recolectados a lo largo de la historia (transacciones y operaciones históricas), puedan servir de forma efectiva en el momento que se necesite de ellos. Esta investigación busca encontrar una forma que guíe y, a la vez, facilite la medición de la calidad de datos en una empresa. Se analizaran temas como *data profiling* y *data rules, c*ómo establecer métricas de medición para diferentes ambientes de negocio en los que puedan habitar los conjuntos de datos medidos.

La investigación centra su línea en explicar los conceptos necesarios para que, conforme se avance en la lectura, se pueda ir comprendiendo cada uno de los capítulos sin perder de vista el tema central que es el de encontrar una metodología para establecer el nivel de calidad de un conjunto de datos.

Primero se darán las directrices que expliquen cómo se escogieron los datos de estudio, esto es importante para que la investigación fluya sin presentar percances fuera de lo considerado normal. Entonces, se procederá a encontrar un conjunto de datos acorde para alcanzar los objetivos de esta investigación.

Luego, se continúa con el proceso de perfilado de datos que, básicamente, ayuda a tener un mejor y más claro conocimiento acerca del estado y estructura actual de los datos. Comparando el estado actual de los datos con las reglas de negocio de la empresa, a la cual pertenecen los datos, se obtienen las normas a las que, de acuerdo con las necesidades del negocio, deberían regirse los datos estudiados, a estas se les denomina reglas de datos.

Por último, para establecer el nivel de calidad en los datos estudiados, se procede a monitorear y auditar si los datos cumplen o no con las reglas de datos afines al negocio que fueron antes generadas. Por medio de esta auditoría y el porcentaje de cumplimiento obtenido, por los datos de estudio, se establece el nivel de calidad.

Una vez realizado el proceso para encontrar el nivel de calidad de un conjunto de datos, también se comprueba y define la metodología propuesta para alcanzar dicho nivel, como parte de los objetivos de este trabajo de investigación.

1. MARCO TEÓRICO

1.1. Calidad de datos

Para entrar en contexto, cuando se hace referencia al término de calidad de datos en esta investigación, se está aludiendo a datos almacenados digitalmente en una base de datos relacional relevantes a un ambiente de negocio. El propósito de establecer el nivel de calidad de los datos, antes mencionado, es poder medir y controlar otro aspecto del negocio que es susceptible de ser mejorado.

1.1.1. Concepto de calidad de datos

Según (Oracle warehouse builder 10gR2: transforming data into quality information, 2006), la calidad de datos es un término que abarca dos aspectos generales: el estado de los datos, es decir, que sean completos, exactos y relevantes; y el conjunto de procesos para lograr ese estado. El objetivo es disponer de datos libres de duplicación, errores de escritura, omisiones y variaciones innecesarias, para tener datos que se ajusten a una estructura definida.

Una parte significativa de la calidad de datos trata con información de clientes, nombres y direcciones, debido tanto a su importante papel en los procesos de negocio, como a su carácter altamente dinámico. Los nombres y direcciones están en todas partes, tienden a existir en casi todas las fuentes y, a veces, son datos únicamente de identificación.

La mayoría de las aplicaciones dependen en gran medida de coincidencias en los nombres y direcciones, ya que un identificador único común no suele estar disponible en todos los sistemas. Por consiguiente, cualquiera de los datos disponibles se utiliza para determinar si diferentes individuos, empresas u otro tipo de registros en realidad son uno mismo.

Pero los nombres y direcciones suelen contener ruido en sus datos, ya que, a menudo, entre ellos se incluyen apodos, abreviaturas, errores de ortografía, mecanografía y redacción.

Además, los nombres y las direcciones constantemente se deterioran con el tiempo, a medida que la gente se muda de vivienda o cambian sus nombres parcial o completamente.

El enfoque acerca de la calidad de los datos referentes a los nombres y direcciones, a veces, causa una idea errónea referente a que la calidad de datos se trata sólo de garantizar que los nombres y las direcciones tengan la calidad necesaria, para que el correo postal sea entregado a quien se debe o en donde se debe. Por ende, el pensamiento es - si su empresa no envía facturas u órdenes, por correo, a los clientes, la calidad de los datos no es tan importante -.

La apreciación anterior es incorrecta por dos razones:

- La corrección y normalización de los nombres y las direcciones no es la meta final, o al menos no es la única. En realidad, el objetivo final es identificar y hacer coincidir a los clientes de forma fiable, basándose en el nombre y los datos de la dirección.

- La calidad de los datos, sin duda, no se limita a los nombres y direcciones. Todos los otros datos, como los de productos, proveedores, etc. se beneficiarán de una completa normalización e integración de los datos

1.1.2. Beneficio de la calidad de datos

La gestión de la calidad en los datos organizacionales (Monitoring data quality performance using data quality metrics, 2006) es a menudo introducida como reacción a problemas agudos atribuidos a fallos de los datos. Por ello, se busca una forma de rastrear el problema y ver el impacto de estos fallos sobre el negocio.

Este enfoque reactivo se puede caracterizar por una carrera para identificar, evaluar y comprar soluciones técnicas que pueden (o no) hacer frente a la manifestación de estos fallos sin tomar en cuenta que una solución de raíz, aislando las causas y eliminando desde la fuente la introducción de datos defectuosos.

En organizaciones más cuidadosas, el modelo comercial para la mejora de calidad de datos ha sido desarrollado como resultado de la evaluación del impacto de cómo la pobre calidad de los datos afecta al alcance de los objetivos trazados por el negocio. Y la revisión de cómo los enfoques de la gestión de calidad de datos puede beneficiar a la organización, desde un punto de vista holístico, es decir, con enfoque global.

Evidentemente, la calidad de datos no es un esfuerzo realizado una única vez. Los acontecimientos y cambios que permiten a los datos defectuosos introducirse en el entorno del negocio no son únicos. Siempre hay nuevas e insidiosas formas que puede influir negativamente en la calidad de los datos.

Los equipos de gestión de calidad de datos deben no sólo hacer frente a fallos agudos en los datos, sino también revisar el estado de calidad actual de la base de datos, de modo que se pueda identificar los puntos críticos y determinar objetivos de mejora.

Algunas ideas implicadas con las organizaciones:

- Formalizar las expectativas de calidad de los datos, como un medio para medir la conformidad de los datos hacia esas expectativas.

- Comparar los niveles de calidad de los datos con sus respectivos referentes y proporcionar un mecanismo para identificar las fugas de aceptación y generar el análisis de y determinación de las causas de los fallos.

- Establecer el máximo nivel de efectividad y comunicar a las comunidades de clientes el nivel de confianza que deben tener sus datos. Lo que requiere el establecimiento de un medio para la medición, monitoreo y seguimiento de la calidad de los datos.

La capacidad de motivar una actitud hacia la mejora de la calidad de datos, como motor de una creciente productividad, muestra un nivel de madurez en la organización y presenta a la información como un fin y recompensa por la proactividad en la gestión del cambio en el personal.

El siguiente paso lógico después de darse cuenta que las lagunas de información corresponden en la baja de rendimiento del negocio, es identificar los beneficios en la productividad que se derivan de la buena gestión de calidad en los datos.

Algo interesante es que estas actividades son las dos caras de la misma moneda. Las dos dependen fundamentalmente de un proceso de determinación del valor agregado de la mejora en la calidad de datos, en función de la conformidad con las expectativas de negocio y cómo esas expectativas se miden en relación con las reglas de calidad de los datos.

Si el éxito del negocio es cuantificable y la dependencia de la empresa por los datos de alta calidad puede establecerse, entonces las mejoras en la información deberían reflejar activamente mesurables mejoras de rendimiento en el negocio.

Con lo anterior se concluye que los parámetros utilizados para medir el nivel de calidad de datos realmente se pueden utilizar, en un nivel más alto, como indicadores para medir el rendimiento general de la empresa.

1.1.3. Perspectivas de la calidad de datos

El enfoque de (Olsen, 2003) acerca de algunos conceptos de calidad de datos es el siguiente.

Los datos son de calidad si reúnen los requisitos previstos para su uso. Les falta calidad en la medida en que estos no cumplan con esos requisitos. En otras palabras, la calidad de datos depende tanto del uso previsto de los datos como también dependen ellos mismos. Para satisfacer el uso previsto, los datos deben ser precisos, oportunos, pertinentes, completos, entendibles y confiables, o alguna combinación inteligente de estas características dependiendo del negocio.

Algunos casos prácticos ayudarán a comprender la noción de la calidad de datos, en función del uso para el que fueron previstos. A continuación se explorarán ejemplos de los aspectos anteriormente mencionados y la relación con la integridad y calidad de los datos.

1.1.3.1. Precisión

Se toma la existencia de una base de datos que contiene nombres, direcciones, números de teléfono y correos electrónicos de los médicos de una ciudad X. En esta base de datos se han detectado una serie de errores: algunos registros están malformados, otros faltan y muchos son obsoletos. Si se compara la base de datos con la población real de los médicos, se espera un 85% de precisión.

Si estos datos se van a utilizar para notificar a los médicos de una nueva ley sobre el suicidio asistido, sin duda se consideraría de mala calidad. De hecho, sería peligroso usarla para ese propósito.

Ahora bien, si fuese utilizada por un fabricante nuevo de dispositivos quirúrgicos para encontrar clientes potenciales, entonces sería considerada de alta calidad. Toda empresa estaría encantada de tener una base de datos de clientes potenciales que posee el 85% de precisión. A partir de ella, podría llevarse a cabo una campaña de tele-mercadeo, para identificar prospectos para estos dispositivos, con una tasa de éxito totalmente aceptable.

La misma base de datos posee una calidad relativamente baja o relativamente alta, para un uso puede que tenga baja calidad en los datos y para otro, una calidad alta, tal como se observa en el ejemplo anterior.

1.1.3.2. Puntualidad

Considere la posibilidad de una base de datos con información de ventas de una división de sociedad. Esta base de datos contiene tres años de movimientos históricos. Sin embargo, la base de datos es lenta en completar las ventas de cada mes.

Algunas unidades deben presentar su información de forma inmediata, mientras que otras tardan varios días para enviar la información.

Asimismo, hay una serie de correcciones y ajustes en el flujo entrante. Entonces, por un lapso, al final del período contable, el contenido está incompleto. Sin embargo, todos los datos están correctos al finalizar el período.

Si esta base de datos se utilizará para calcular los montos de ventas que se vencen el día 15 del siguiente mes, los datos serían de mala calidad a pesar de que los datos tengan siempre la precisión requerida. En todo caso, los datos no están lo suficientemente a tiempo para el uso previsto.

Sin embargo, si esta base de datos se utilizará para analizar las tendencias históricas y así tomar decisiones en la modificación de los territorios, sería de excelente calidad, siempre y cuando el usuario supiera cuándo todas las ventas y los cambios se incorporan. Esperar para que todos los datos ingresen al sistema no es un problema, porque el uso previsto es el de tomar decisiones a largo plazo.

1.1.3.3. Pertinencia

Considere la posibilidad de una base de datos de inventario que contiene números de partes, lugares de distribución, cantidades a la mano y otra información. Sin embargo, no contiene información del origen de las partes.

Si una parte fue comprada a diferente tipo de proveedores, una vez que las piezas son recibidas y puestas en el estante, no hay ningún registro del proveedor del cual provenían las partes.

La información contenida en la base de datos es siempre precisa y se mantiene actualizada. Para las transacciones de inventario normal y toma de decisiones, la base de datos es, sin duda, de alta calidad. Aunque, si un proveedor informa que en uno de sus envíos figuran piezas defectuosas, esta base de datos no ayuda a identificar si se tiene en *stock* cualquiera de esas partes o no. La conclusión será que la base de datos es de mala calidad, ya que no contiene un elemento relevante y útil de la información. Sin esa información, la base de datos presenta una calidad deficiente para el uso previsto.

1.1.3.4. Exhaustividad

Una base de datos contiene información acerca de las reparaciones en maquinaria de capital. Sin embargo, es un hecho conocido que, a veces, las reparaciones se llevan a cabo y la información sobre la reparación simplemente es ingresada oportunamente a la base de datos.

Esta situación anómala se considera el resultado de la falta de compromiso por parte del departamento de reparación y la falta de supervisión por parte de sus jefes. Se estima que la cantidad de información faltante es del 5% del total de reparaciones.

Esta base de datos es, probablemente, una base de datos de buena calidad para evaluar la salud general del equipo. El equipo que requiere una gran cantidad de gastos de mantenimiento puede ser identificado a partir de estos datos. A menos que los datos que faltan estén desproporcionadamente sesgados, los registros se pueden utilizar para todas las decisiones ordinarias.

Sin embargo, si se trata de usar como una base para evaluar la información, la convierte en una base de datos de baja calidad. Las transacciones faltantes podrían utilizarse para controlar las garantías de una buena parte del equipo reparado, pero, dada la falta de información no es posible hacerlo con precisión.

1.1.3.5. Comprensibilidad

Considere la posibilidad de una base de datos con las órdenes de ventas de un conjunto de clientes. Una práctica para el manejo de quejas y devoluciones es crear un "ajuste" para respaldar la orden original y luego escribir una nueva orden para la información corregida, si procede. Este procedimiento asigna números nuevos para el ajuste y las órdenes de reposición.

Para el departamento de contabilidad, se trata de una base de datos de alta calidad. Todos los números salen en los reportes. Para un analista de inteligencia de negocios que trata de determinar tendencias de crecimiento de pedidos, por regiones, se trata de una base de datos de mala calidad.

Si el analista supone que cada número representa una orden distinta, su análisis estará del todo mal. Alguien tendría que explicarle las prácticas y métodos necesarios para desentrañar los datos y así llegar a los números reales, si eso fuera posible incluso después de los hechos, es decir, que la base de datos no permite una lectura precisa a partir de su misma organización e información contenida.

1.1.3.6. Confianza

Una nueva aplicación es implementada y se utiliza para determinar la cantidad y la fecha de los pedidos de piezas para la maquinaria antigua y el tiempo de servicio de las máquinas utilizadas actualmente.

La aplicación original tenía un error de programación que incorrectamente hacía un pedido de 10 veces la cantidad realmente necesaria. El error fue detectado hasta que un gran pedido fue enviado y hubo un gran revuelo en el departamento de compras. Luego fue corregido y el problema ya no se repitió La base de datos no se equivocó nunca, fue la aplicación en la orden de compra la que estaba errónea, porque fue registrada, en la base de datos, tal como la aplicación mandó.

Debido al temor de que se repitiera el incidente, el jefe de mantenimiento optó por no utilizar la aplicación, ni la información dentro de la base de datos.

Ahora ordena las partes sobre la base de una pequeña hoja de cálculo que él construyó para mantener gran parte de la misma información, a pesar de que, a menudo, faltan transacciones y no siempre sabe cuándo las partes nuevas llegan al inventario.

A menos que su confianza en la aplicación original se recupere, la base de datos es de mala calidad, aunque sea del todo exacta. No sirve para cumplir su propósito como consecuencia de una falta de credibilidad.

1.1.4. Antecedentes

En la búsqueda de metodologías actuales, para medir el nivel de calidad de un conjunto de datos, se encontraron dos distintas. Cada una con sus ventajas y desventajas, y su propia forma de identificar las reglas adecuadas del negocio, para que la auditoría entregue un resultado más apegado al referente, pero, ambas ejecutando dicha auditoría de manera manual. Lo anterior, dependiendo del tamaño del conjunto de datos a analizar, podría representar demasiado tiempo invertido en el proceso.

Durante la presente investigación, fueron seleccionadas las mejores prácticas de ambas formas para la identificación de reglas del negocio, que pudieran ser aplicadas seguidamente al monitoreo de los datos, es el caso de *data rules*, sin embargo, se implementó una mejora en la metodología utilizando procesos de *data profiling,* para llevar a cabo la auditoría y así ganar tiempo en entrega de los resultados; incluso automatizando dichos procesos para que las alertas de incumplimiento saltasen a la vista, reduciendo aún más el tiempo de identificación de errores.

A continuación se ampliará y detallará qué es y para qué sirven los procesos de perfilación de datos.

1.2. Perfilación de datos

El proceso de realizar la perfilación de datos permite conocer ciertas características de los datos del negocio que, a simple vista, pasan inadvertidas. Por ejemplo, patrones o comportamientos en los datos que únicamente pueden ser encontrados por el poder de procesamiento actual con el que se cuenta en la Era de la Información.

Otra característica relevante que puede ser descubierta por la operación de perfilación de datos es la forma en que los usuarios encargados de registrar los datos al sistema usan el aplicativo concebido para este fin, es decir, errores frecuentes, malas prácticas o incluso buenas, etc.

Más adelante, estos patrones o comportamientos de los datos, o prácticas de los usuarios, aunados a las necesidades del negocio, pueden derivar reglas de datos como se verá conforme se avance en la lectura.

1.2.1. Conceptualización

Una de las características (Oracle data quality option, 2010) más importantes en cuanto a la manipulación y administración de datos en el proceso de control de calidad es la capacidad de evaluar el estado actual de los datos de un sistema u organización. A esto proceso se le conoce comúnmente como perfilación de datos.

El proceso de perfilación de los datos nos permite obtener estadísticas de la forma y estructura actual de los datos y metadatos, ya sea de una o varias tablas sometidas al análisis. Nos indica sobre posibles relaciones entre tablas o restricciones que puedan ser aplicadas en determinadas columnas.

Según Soto (2008) “por perfilado de datos se entiende el análisis de los datos de los sistemas para entender su contenido, estructura, calidad y dependencias”.

La perfilación de datos es un paso fundamental en el proceso de determinación de la calidad de los datos, ya que a la hora de plantear un análisis de los datos fuente, sucede en muchas ocasiones que, realmente, no se sabe qué preguntar, ni en dónde pueden residir los problemas con determinada base de datos.

Lo que resume de buena forma (Soto, 2008) “*you don´t know what you don’t know”*, es decir, desconocemos lo que no sabemos. La perfilación entrega un análisis exacto de la estructura y el modelo de los datos de estudio, siendo más precisos y permitiendo actuar sin solamente basar nuestro análisis en las experiencias del DBA. Ya que estos se basarían en suposiciones. El proceso de perfilación con una herramienta especializada entregará “hechos sobre los cuales construir el diseño”.

Fernández (2010) expresa que el proceso de perfilación de datos es una de las primeras tareas que se suelen abordar en la gestión de calidad de los datos, y consiste en realizar un primer análisis sobre los datos de origen, normalmente sobre tablas, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad, haciendo consultas a nivel de tabla, columna, relaciones entre columnas e incluso relaciones entre tablas.

Por medio de este proceso, será posible obtener informes detallados en los cuales se podrá analizar la estructura y formato de cada uno de los registros, número de registros para cada tipo de formato, comprobación de integridad referencial, dependencias entre tablas, etc.

Quizá la característica más útil del proceso de perfilación de datos es la que permite, por medio de patrones establecidos y dominios de información, encontrar anomalías en los datos y generar reglas de datos para que sean corregidos, lo cual producirá datos más limpios.

El proceso de corrección de datos debe de estar integrado perfectamente con las rutinas de extracción, transformación y carga (ETL) de los datos, para la construcción de un ambiente analítico que entregue calidad de información. Adicionalmente, esta integración con los procesos de ETL permite que la obtención y publicación de datos de alta calidad se efectúe de forma ágil y dinámica cumpliendo con los requerimientos del negocio.

A continuación se presenta un análisis del estado actual del mercado de las herramientas más emblemáticas y especializadas para llevar a cabo el proceso de perfilación.

Es importante conocer las características y ventajas que nos presenta cada herramienta, ya que, de acuerdo con la herramienta que se escoja se podrá obtener una mejor perspectiva general del estado de los datos de estudio. Entre más detalles proporcione la herramienta de análisis y más información se tenga a disposición, mejores serán los resultados de la medición del nivel de calidad de los datos.

1.2.2. Beneficios

- Lo que se espera de un sistema es que, en su diseño, incluya validaciones y protecciones contra errores de ingresos de datos. En la realidad, hay una gran brecha entre los procesos de ingreso de información al sistema y los controles de calidad, porque cada sistema deja ingresar errores. La perfilación de los datos le permite a las empresas ahorrar dinero y tiempo, al detectar los errores antes de que ellos representen algún costo adicional o pérdida de clientes.

- Los recursos humanos pueden asignarse no a buscar errores sino a corregirlos, luego de que la perfilación de datos haga el trabajo de análisis e identificación de los mismos.

- Se puede medir certeramente el grado de corrupción de los datos. Con este conocimiento una empresa puede tomar la mejor decisión en cuanto a las medidas de corrección y el esfuerzo necesario para componer la información errónea que es considerada de baja calidad.

- Es necesario identificar y eliminar datos redundantes u obsoletos de la base de datos. El espacio de almacenamiento físico necesario para el sistema se reduciría, conduciendo de forma directa a ahorro de tiempo y dinero en el manejo y administración de los datos.

- Los resultado de la perfilación de datos puede ser usado para incrementar la efectividad en los procesos ETL, para la generación de un *data warehouse,* por ejemplo.

- El proceso de perfilación de datos puede también ser usado para evaluar el desempeño de las aplicaciones y el funcionamiento de sus filtros.

- La generación de meta-data es un resultado directo del perfilado de datos.

1.2.3. Herramientas

1.2.3.1. *Informatica 9*

De acuerdo con Fernández (2010), en el mercado de productos para integración de datos, el pionero es *Informatica*, primer proveedor independiente de software de integración de datos. Su herramienta más reconocida es "*Informatica power center*", que ya tiene un largo recorrido y hasta hace unos años era un modelo en el mundo de la integración.

Este proveedor también dispone de otros productos orientados a propósitos más específicos que, a la vez, se integran dentro de la plataforma, siempre en el marco de la integración de datos.

La plataforma "*Informatica 9*" fue pensada para soportar el ciclo de vida completo de la integración de datos, que consta de cinco pasos principales:

- Acceso
- Detección
- Limpieza
- Integración

En donde lo que compete a esta investigación es la detección y limpieza.

1.2.3.2. *Microsoft SQL Server 2008*

Fernández (2010) opina que uno de los nuevos aportes a *Microsoft SQL server* 2008 en "*Integration services*" (herramienta de ETL) es su capacidad de perfilación de datos con "*Data profiletask*".

"Data profiletask" funciona seleccionando sobre una tabla, en una base de datos (únicamente *SQL server 2000* o superior), las opciones de perfilado que se quiera realizar sobre los datos de la tabla y almacenando en un fichero XML los resultados, cuando se ha ejecutado la misma.

Un buen avance podría ser que fuera posible integrar este servicio con bases de datos de diferentes proveedores, pero no lo es. Simultáneamente, pueden seleccionarse hasta ocho tipos de perfilado, tres con varias columnas a la vez y cinco tomando una sola columna.

Perfilación de varias columnas:

- Claves candidatas, columnas con potencial para ser llaves primarias
- Dependencia funcional, dependencia entre columnas
- Inclusión de valores, columnas con potencial para ser llaves foráneas

Perfilación a nivel de una columna:

- Longitud de valores (distribución)
- Porcentaje de valores nulos
- Patrones expresados mediante expresiones regulares
- Estadísticas de columna: mínimo, máximo, media o desviación estándar

1.2.3.3. *Oracle DQ for data integrator and Oracle data profiling 11g*

"*Oracle data quality for data integrator and Oracle data profiling 11g*" (Oracle data quality for data integrator and Oracle data profiling 11g, 2010) presenta un conjunto de poderosas herramientas y soluciones efectivas para llevar a cabo la perfilación y la gestión de calidad de datos.

Recolecta los metadatos y los datos fuente, para luego analizar y tabular la información en estadísticas compresibles, tales como: la longitud de los atributos, máximos y mínimos, valores de distribución, patrones, tipos de datos, y otros.

Automáticamente, aplica técnicas avanzadas de perfilación para identificar problemas potenciales con los datos como: fallos en las direcciones, en los códigos de productos, errores de ortografía, duplicaciones de datos, etc.

La interfaz de usuario provee un adecuado y útil aprovechamiento del detalle captado de la información, permitiendo analizar, tanto por estadísticas, como por el detalle de los datos perfilados.

Detecta y presenta atributos que son potencialmente escogidos para ser una llave primaria dentro de la tabla, también se puede detectar el grado de unicidad de un atributo, dependencias funcionales, duplicidad y otras inconsistencias. La detección de llaves foráneas también puede ser de gran utilidad.

Características:

- Entero soporte para gestión de calidad de datos
- Corrección y estandarización de nombres y direcciones
- Proceso de identificación automático de duplicados
- Funciones pre-construidas con extensas reglas personalizadas por región
- Personalización de reglas de datos
- Completa integración con "*Oracle data integrator 11g*"

Beneficios:

- Todos los proyectos son gestionados desde una misma herramienta
- Poderosa validación que realiza la corrección de nombre y direcciones
- Emparejador de registros para detectar y manejar automáticamente los duplicados
- Proceso completamente adaptable de gestión de la calidad de datos y auditoría de cumplimiento de las métricas establecidas

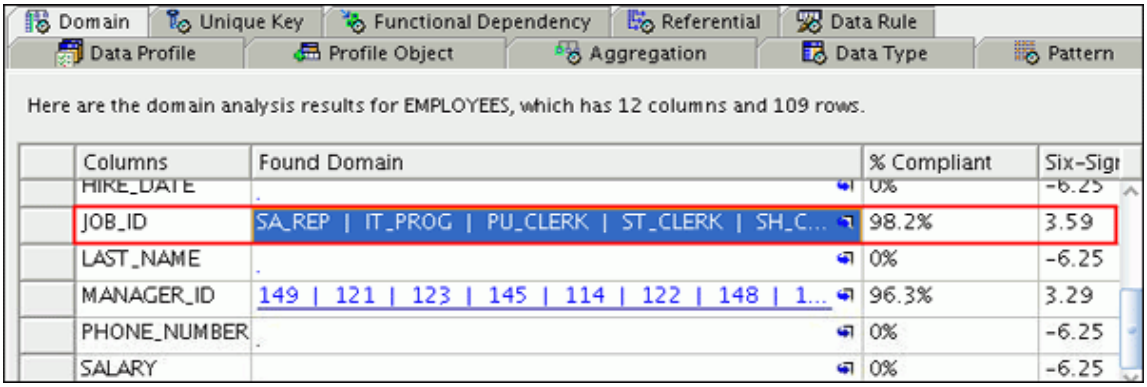
Las anteriores funciones facilitarán la detección e implementación de reglas de datos para el posterior monitoreo de cumplimiento de la calidad de la información.

1.2.4. Funcionamiento

El funcionamiento de *data profiling* permite profundizar en algunas características de los datos que a simple vista no son observables. En consecuencia, luego de haber escogido el conjunto de datos para analizar con la ayuda del perfilado, presenta un panorama amplio de los datos y sus características, para la generación de las *data rules*.

En la siguiente figura se analizó una base de datos de ejemplo. Lo que se buscaba era ejemplificar el análisis de cumplimiento de dominios de datos, que es una de las funcionalidades de los procesos de *data profiling.*

Figura 1. **Ejemplo de *data profiling***



Fuente: elaboración propia.

El perfilador de datos, en este caso "*Oracle warehouse builder*", no solo verifica si los datos pertenecen a un dominio definido, sino que propone dominios con base en los porcentajes actuales de los datos.

Para el caso del ejemplo, la herramienta hace una recomendación de dominio para el campo JOB_ID con un cumplimiento primario del 98.2%. En el caso de estudio se explicará con mayor detalle el accionar del análisis de *data profiling,* y se conocerá, mediante datos reales, el rol que juega este análisis en la metodología seleccionada.

1.3. *Data rules*

Este capítulo describe las reglas de datos y sus aplicaciones, es decir, la forma de diseñar y derivarlas a partir del proceso de perfilación.

1.3.1. Conceptualización

Reglas de datos son definiciones de valores de datos válidos y sus relaciones. Pueden ser aplicadas a tablas, vistas, vistas materializadas y tablas externas. Ellas determinan la forma legal establecida, dentro de un negocio de cómo deberían encontrarse los datos, puesto que las reglas de datos son esenciales para determinar el nivel de calidad de datos.

Después de realizar la perfilación de datos, las reglas de datos se pueden generar automáticamente, a partir de la información descubierta. La perfilación de datos incluso puede ser puesta a prueba, una vez estén establecidas las reglas de datos, para auditar el cumplimiento. La auditoría de datos utiliza las reglas de datos para comprobar el cumplimiento de las mismas, y generar estadísticas acerca de los datos que no cumplen con las normas definidas.

La utilización de estrategias de corrección y limpieza de datos también dependen de ellas, porque automáticamente se puede asociar las correcciones una vez se establezca que un conjuntos de valores no cumplieron con las reglas.

1.3.2. Tipos de reglas de datos

- Lista de dominio: define la lista de valores que un atributo puede tener. Por ejemplo, el campo “SEXO” únicamente podría tener entre sus valores “M” o “F”.

- Lista de patrones de dominio: define una lista de patrones que un atributo debería de cumplir. Los patrones se definen con expresiones regulares.

- Rango de dominio: define un rango de valores en el que un atributo debería encontrarse. Un ejemplo, el valor para el atributo SUELDO puede estar entre 100.00 y 10,000.00.

- Formato común: define un formato conocido, dentro del cual podría ajustarse un atributo. Este tipo de regla tiene muchos subtipos: número de teléfono, dirección IP, número de seguro social, dirección de vivienda, dirección de correo electrónico. Cada tipo tiene formatos predefinidos pero se pueden generar aún más.

- No se permiten valores nulos: especifica que el atributo no puede tener valores nulos.

- Dependencia funcional: define que los datos, en la tabla, pueden ser normalizados sí son relacionados con otras tablas, de las cuales dependan.

- Llave candidata: define si una columna o grupo de columnas pueden identificar de forma única a la tabla.

- Referencial: define el tipo de relación (de unos a muchos) que un valor debe o puede tener con otro valor, en la misma o en diferente tabla.

- Nombre y dirección: utiliza patrones establecidos para evaluar si un conjunto de campos pertenecen a la descripción de un nombre propio o dirección física de una vivienda o empresa.

- Personalizada: aplica una expresión SQL que especifique los parámetros de cumplimiento.

1.4. Uso de reglas de datos

Además de las reglas de datos, derivadas de resultados del proceso de perfilación, se pueden definir las reglas con otros tipos de análisis de información. Es factible, inclusive, enlazar una regla a varias tablas de datos, un objeto puede contener cualquier número de reglas de datos si así se requiere.

Algunos usos comunes que se les da a las reglas de datos:

- En procesos de perfilación
- En esquemas de corrección
- Para limpieza de datos
- En procesos de auditoría y monitoreo
- Para determinar la calidad de datos

2. CASO DE ESTUDIO

2.1. Comparación, antecedentes y propuesta

En este capítulo se analizan y se conocen a fondo las características de los datos seleccionados para el estudio. El procedimiento propuesto para realizar dicho análisis es el proceso de perfilado de datos.

El proceso de perfilación de datos ha existido desde que existen los repositorios de datos, incluso los datos no estructurados digitalmente. Este proceso consiste en la obtención, ordenamiento, tabulación y generación de estadísticas acerca de las características de una selección de datos. La diferencia entre los procesos antiguos y los actuales es el poder de procesamiento digital con el que se cuenta en estos tiempos.

Y dado ese poder de procesamiento obtenido de los computadores actuales, han surgido herramientas diseñadas específicamente con el propósito de facilitar los procesos de generación de estadísticas o *data profiling*. Entre otras funciones, el aumento en el rendimiento de dichos procesos permite el análisis de mayores volúmenes de datos, que en décadas pasadas era, prácticamente, impensable de realizar.

Lo propuesto en este estudio es aprovechar estas herramientas para realizar el proceso de perfilación de forma completa en el universo total de información, y no sobre una pequeña muestra estadística. Si el estudio se hace sobre una muestra, el resultado es real para esa muestra y es posible generalizar ese resultado para la población total en estudio.

Pero si el estudio se realiza sobre el universo de la información se puede decir que el resultado obtenido estará apegado a la realidad total del mencionado universo, es decir, a la población del estudio.

2.2. Proceso de perfilación y generación de reglas de datos

2.2.1. Elección del conjunto de datos para el estudio

Para analizar un conjunto de datos y establecer si poseen un nivel de calidad aceptable, según las reglas de negocio de una empresa, se debe primero obtener una muestra representativa de datos del negocio.

El enfoque de la presente investigación se encamina a determinar si los datos son aptos para realizar análisis de información, es decir, que la riqueza analítica que presenta la información sea amplia y abundante para tomar decisiones acertadas sobre ella. Partiendo de la anterior premisa, las tablas seleccionadas para el estudio deben ser aquellas que presenten hechos medibles y cuantificables, y que a través de la información contenida en ellas se pueda sumar valor al negocio tomando mejores y más rápidas decisiones.

En un *data warehouse*, la plataforma analítica por excelencia, se busca que la consistencia y la integridad de los datos sean óptimas para presentar de forma correcta los resúmenes de información requeridos. En un ambiente transaccional, los datos se ingresan tal cual la persona operativa de la empresa logra captarlos, por lo mismo, estos datos tienden a incurrir en muchos errores.

El conjunto de datos de estudio pertenece a una empresa dedicada a brindar soporte técnico a sus clientes. Entre las técnicas utilizadas para recopilar los datos para la investigación, se aplicó la entrevista personal, con los usuarios tomadores de decisiones, para conocer más en detalle el negocio, y establecer lo que ellos esperaban obtener en los reportes analíticos de su información. Con base en la información recabada, se llevaron a cabo reuniones con el usuario administrador de la base de datos (DBA), para determinar qué tablas eran las representativas, según los requerimientos de los usuarios tomadores de decisiones.

La tabla seleccionada para el proceso de perfilado y análisis de resultados contiene datos en los cuales se registran los incidentes de soporte. Según el emparejamiento de los requerimientos de los analistas de información y el conocimiento técnico del DBA, administrador del modelo de datos transaccional, se concluyó que hay una tabla en la cual se almacena el núcleo relevante del negocio, la cual fue seleccionada para el estudio.

Tabla I. **Características de la tabla seleccionada para el estudio**

Nombre:	INCIDENTE
Tamaño:	25,958 registros
Columnas:	19
Tipo:	Tabla de hechos
Uso:	Transaccional
Versión DB:	Oracle 10.2.0.4

Fuente: elaboración propia.

Hay distintos métodos estadísticos para seleccionar una muestra representativa, dado un universo de datos; sin embargo, en este estudio, se contempló la posibilidad de analizar el universo completo de los datos, ya que la cantidad total de registros resultaba perfectamente manejable, para cumplir con los objetivos propuestos. Si los datos hubiesen sido tabulados y ordenados a mano, se tendría la necesidad de reducir la muestra de estudio, para generar resultados en menor tiempo, pero, el hecho de poder utilizar una herramienta de perfilación como es "*Oracle warehouse builder*", el rendimiento puede ser maximizado y, por ello, se puede analizar un número mayor de datos y entregar los resultado en menor tiempo.

2.2.2. Perfilación de datos y análisis de resultados

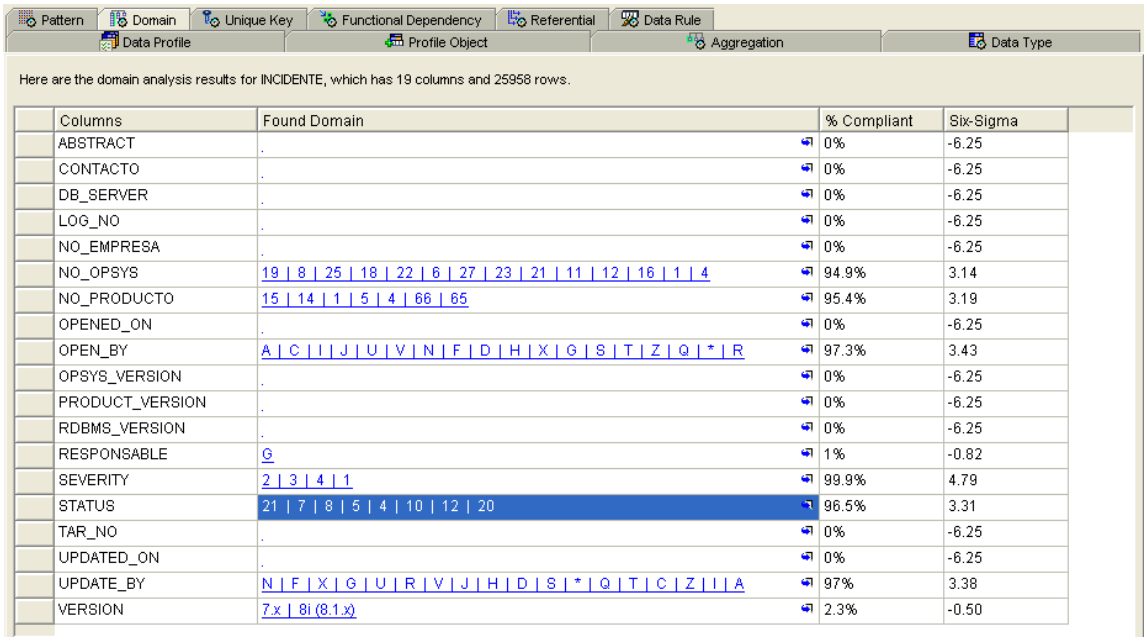
Una vez establecida la tabla fuente de estudio, el primer paso es perfilar los datos contenidos en ella, con lo cual será posible hacer resúmenes estadísticos de las características de los datos para analizar y conocer en detalle su estado, desde una perspectiva más profunda de análisis.

- Dominio de los datos

El dominio de los datos representa el conjunto de valores que puede tomar un campo. Con este análisis se generan muchas reglas de datos relacionadas de forma directa con las reglas de negocio. Es muy útil determinar el dominio de los datos para establecer niveles de calidad según el cumplimiento de estas reglas de datos generan en la perfilación y su relación directa con los requerimientos explícitos del negocio.

La perfilación de datos, para el nivel de dominios de datos, realizado en la tabla de estudio, produjo los siguientes resultados:

Figura 2. **Perfilado de datos para encontrar dominios de valores**



Columns	Found Domain	% Compliant	Six-Sigma
ABSTRACT	.	0%	-6.25
CONTACTO	.	0%	-6.25
DB_SERVER	.	0%	-6.25
LOG_NO	.	0%	-6.25
NO_EMPRESA	.	0%	-6.25
NO_OPSYS	19 | 8 | 25 | 18 | 22 | 6 | 27 | 23 | 21 | 11 | 12 | 16 | 1 | 4	94.9%	3.14
NO_PRODUCTO	15 | 14 | 1 | 5 | 4 | 66 | 65	95.4%	3.19
OPENED_ON	.	0%	-6.25
OPEN_BY	A | C | I | J | U | V | N | F | D | H | X | G | S | T | Z | Q | * | R	97.3%	3.43
OPSYS_VERSION	.	0%	-6.25
PRODUCT_VERSION	.	0%	-6.25
RDBMS_VERSION	.	0%	-6.25
RESPONSABLE	G	1%	-0.82
SEVERITY	2 | 3 | 4 | 1	99.9%	4.79
STATUS	21 | 7 | 8 | 5 | 4 | 10 | 12 | 20	96.5%	3.31
TAR_NO	.	0%	-6.25
UPDATED_ON	.	0%	-6.25
UPDATE_BY	N | F | X | G | U | R | V | J | H | D | S | * | Q | T | C | Z | I | A	97%	3.38
VERSION	7.x | 8i (8.1.x)	2.3%	-0.50

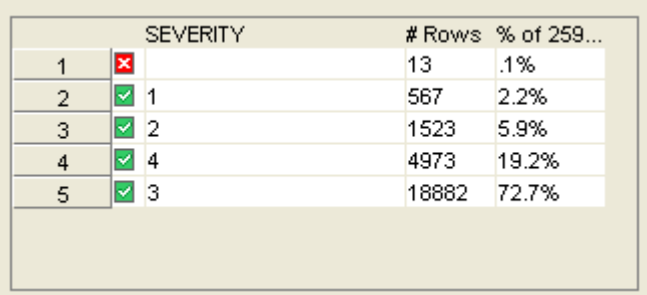
Fuente: elaboración propia.

En la figura 2, se muestran los dominios actuales con los que cuenta cada uno de los atributos de la tabla. Este es diferente al dominio ideal con el que debería de contar y que debe de ser definido por las reglas del negocio.

Cuando se ha estudiado el dominio requerido de un campo, se da paso a la generación de las llamadas reglas de datos, que servirán, específicamente, para llevar el control de que tanto un atributo cumple con lo requerido para su negocio.

Un ejemplo del dominio actual de un campo es el que se presenta a continuación para el campo *SEVERITY* dentro de la tabla estudiada:

Figura 3. **Dominio y porcentaje asociado a los datos**



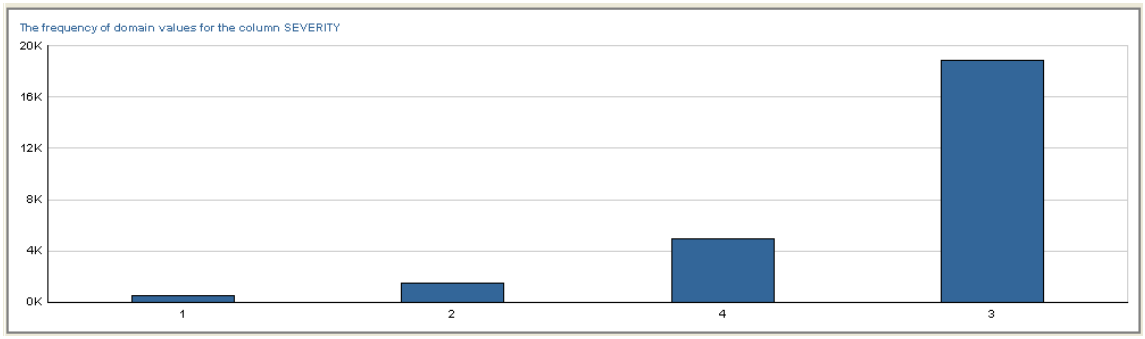
	SEVERITY	# Rows	% of 259...
1		13	.1%
2	1	567	2.2%
3	2	1523	5.9%
4	4	4973	19.2%
5	3	18882	72.7%

Fuente: elaboración propia.

La figura anterior muestra el dominio actual de un campo, el número de tuplas que cumplen con cada uno de los elementos del dominio y el porcentaje que representa ese número para la totalidad de registros de la tabla.

Observando la figura anterior de manera gráfica, es decir, desde otra perspectiva, se puede realizar el análisis respectivo para la generación de las reglas de datos con mayor claridad e infiriendo, de mejor forma, acerca de las mismas.

Figura 4. **Dominio de los datos representado gráficamente**



Fuente: elaboración propia.

- Patrones en los datos

Se efectuó también la perfilación sobre los datos para encontrar patrones dentro de los atributos de la tabla estudiada:

Figura 5. **Patrones descubiertos en los datos**

Pattern | Domain | Unique Key | Functional Dependency | Referential | Data Rule

Data Profile | Profile Object | Aggregation | Data Type

Here are the pattern analysis results for INCIDENTE, which has 19 columns and 25958 rows.

Columns	Dominant Character Pattern	% Compliant	Dominant Word Pattern	% Compliant	Common Format	% C
ABSTRACT	^(A{3}bA{9})$	0.4%	^(A+(b)+A+(b)+A+(b)...	6.8%	IP Address	0%
CONTACTO	^(Aa{5}bAa{5})$	2.5%	^(A+a+(b)+A+a+)$	42.1%	.	0%
DB_SERVER	.	0%	.	0%	.	0%
LOG_NO	^(9{5})$	74.7%	^(9+)$	100%	.	0%
NO_EMPRESA	^(9{4})$	50.2%	^(9+)$	100%	.	0%
NO_OPSYS	^(9{2})$	60.2%	^(9+)$	100%	.	0%
NO_PRODUCTO	^(9)$	76%	^(9+)$	100%	.	0%
OPENED_ON	^(9{2}/9{2}/9{2})$	100%	^(9+/9+/9+)$	100%	.	0%
OPEN_BY	^(A)$	84.9%	^(A+)$	84.9%	.	0%
OPSYS_VERSION	^(9.9)$	6.2%	^(9+)$	11%	IP Address	0.2%
PRODUCT_VERSION	^(9.9.9)$	12.3%	^(9+.9+.9+.9+.9+)$	21.1%	IP Address	16%
RDBMS_VERSION	^(9.9.9)$	11.8%	^(9+.9+.9+)$	17.7%	IP Address	12.3

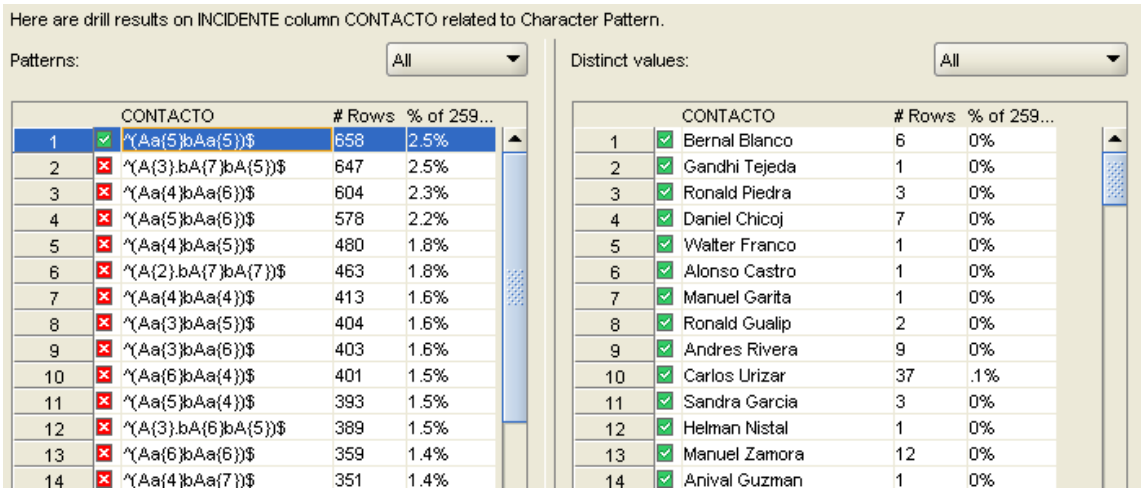
Fuente: elaboración propia.

Cuando la cardinalidad de los datos es muy alta y no se pueden determinar dominios enumerando los posibles valores, se puede optar por definir patrones de cumplimiento. Los patrones son estructuras a las que deben apegarse los valores de un atributo para ser considerados como válidos. Son, en cierto modo, una máscara que permite reglamentar dominios muy amplios de datos.

Al igual que en los dominios de valores; para este caso, se estudió cada atributo para obtener distintas posibilidad de patrones.

Queda a criterio de cada analista la elección del método que aplicará la regla de datos, si por medio de dominios o de patrones. A continuación, se muestra una figura en donde se establece el número de tuplas que cumplen con cada patrón y su porcentaje en referencia a la totalidad de registros de la tabla; el atributo en cuestión, para este caso, es *contacto*:

Figura 6. **Patrones y porcentaje asociado**



	CONTACTO	# Rows	% of 259...
1	^(Aa{5}bAa{5})$	658	2.5%
2	^(A{3}.bA{7}bA{5})$	647	2.5%
3	^(Aa{4}bAa{6})$	604	2.3%
4	^(Aa{5}bAa{6})$	578	2.2%
5	^(Aa{4}bAa{5})$	480	1.8%
6	^(A{2}.bA{7}bA{7})$	463	1.8%
7	^(Aa{4}bAa{4})$	413	1.6%
8	^(Aa{3}bAa{5})$	404	1.6%
9	^(Aa{3}bAa{6})$	403	1.6%
10	^(Aa{6}bAa{4})$	401	1.5%
11	^(Aa{5}bAa{4})$	393	1.5%
12	^(A{3}.bA{6}bA{5})$	389	1.5%
13	^(Aa{6}bAa{6})$	359	1.4%
14	^(Aa{4}bAa{7})$	351	1.4%

	CONTACTO	# Rows	% of 259...
1	Bernal Blanco	6	0%
2	Gandhi Tejeda	1	0%
3	Ronald Piedra	3	0%
4	Daniel Chicoj	7	0%
5	Walter Franco	1	0%
6	Alonso Castro	1	0%
7	Manuel Garita	1	0%
8	Ronald Gualip	2	0%
9	Andres Rivera	9	0%
10	Carlos Urizar	37	.1%
11	Sandra Garcia	3	0%
12	Helman Nistal	1	0%
13	Manuel Zamora	12	0%
14	Anival Guzman	1	0%

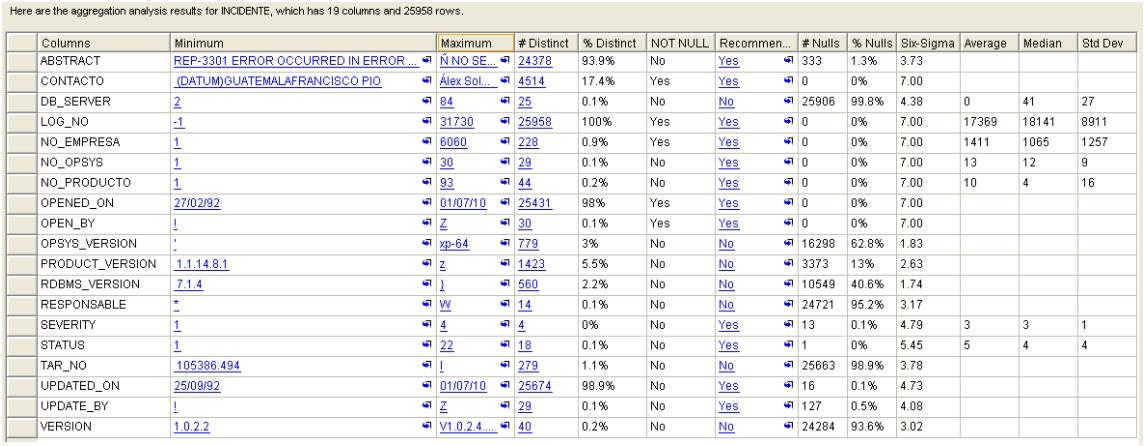
Fuente: elaboración propia.

Para el caso de la columna CONTACTO, el dominio de posibles valores para ese campo es muy amplio ya que contiene nombres de personas. En este caso, el análisis de patrones queda perfecto, ya que se puede establecer un patrón de nombre propio (por ejemplo) y monitorear todo aquel valor que no cumpla dicho patrón para este tipo de atributos.

- Resumen estadístico de características

El último análisis de perfilación que se practicó fue el encargado de presentar detalles estadísticos por medio de la agregación de cada uno de los diferentes campos de la tabla.

Figura 7. **Resumen estadístico de características**



Here are the aggregation analysis results for INCIDENTE, which has 19 columns and 25958 rows.

Columns	Minimum	Maximum	# Distinct	% Distinct	NOT NULL	Recommen...	# Nulls	% Nulls	Six-Sigma	Average	Median	Std Dev
ABSTRACT	REP-3301 ERROR OCCURRED IN ERROR ...	Ñ NO SE...	24378	93.9%	No	Yes	333	1.3%	3.73			
CONTACTO	(DATUM)GUATEMALAFRANCISCO PIO	Álex Sol...	4514	17.4%	Yes	Yes	0	0%	7.00			
DB_SERVER	2	84	25	0.1%	No	No	25906	99.8%	4.38	0	41	27
LOG_NO	-1	31730	25958	100%	Yes	Yes	0	0%	7.00	17369	18141	8911
NO_EMPRESA	1	6060	228	0.9%	Yes	Yes	0	0%	7.00	1411	1065	1257
NO_OPSYS	1	30	29	0.1%	No	Yes	0	0%	7.00	13	12	9
NO_PRODUCTO	1	93	44	0.2%	No	Yes	0	0%	7.00	10	4	16
OPENED_ON	27/02/92	01/07/10	25431	98%	Yes	Yes	0	0%	7.00			
OPEN_BY	!	Z	30	0.1%	Yes	Yes	0	0%	7.00			
OPSYS_VERSION	'	xp-64	779	3%	No	No	16298	62.8%	1.83			
PRODUCT_VERSION	1.1.14.8.1	z	1423	5.5%	No	No	3373	13%	2.63			
RDBMS_VERSION	7.1.4	]	560	2.2%	No	No	10549	40.6%	1.74			
RESPONSABLE	*	W	14	0.1%	No	No	24721	95.2%	3.17			
SEVERITY	1	4	4	0%	No	Yes	13	0.1%	4.79	3	3	1
STATUS	1	22	18	0.1%	No	Yes	1	0%	5.45	5	4	4
TAR_NO	105386.494	]	279	1.1%	No	No	25663	98.9%	3.78			
UPDATED_ON	25/09/92	01/07/10	25674	98.9%	No	Yes	16	0.1%	4.73			
UPDATE_BY	!	Z	29	0.1%	No	Yes	127	0.5%	4.08			
VERSION	1.0.2.2	V1.0.2.4....	40	0.2%	No	No	24284	93.6%	3.02			

Fuente: elaboración propia.

En el análisis de agregación se encuentran variables tales como: el número de valores nulos de cada campo y su porcentaje referido al total de registros de la tabla, los valores mínimos y máximos, el número de valores distintos por campo.

Lo anterior pretende representar otra perspectiva del análisis de la tabla INCIDENTE para que la generación de reglas de datos sea más fluida y con información fehaciente del estado de los datos estudiados.

Según los requerimientos planteados por el negocio, dueño de la información de estudio, se necesita tener la mayor cantidad de datos que sea posible captar, desde su sistema aplicativo transaccional. Asimismo, se establece que la dimensión de calidad de datos que requiere ser monitoreada es la Exhaustividad.

Según las estadísticas encontradas por el proceso de perfilación se definió una serie de campos que deben ser sometidos a la regla de que no permite vacíos de información en dichos campos.

Los campos y las reglas generadas para ellos se describen en la siguiente sección.

2.2.3. Generación de reglas de datos asociadas con el negocio

Todas las reglas de datos generadas fueron derivadas de la dimensión de calidad de datos: Exhaustividad. Básicamente, se puede decir que es la misma regla para muchos atributos, a continuación se presenta la lista de estos:

- INCIDENTES.*DB_SERVER*, no permite datos nulos
- INCIDENTES.*PRODUCT_VERSION*, no permite datos nulos
- INCIDENTES.*OPEN_BY*, no permite datos nulos
- INCIDENTES.NO_EMPRESA, no permite datos nulos
- INCIDENTES.CONTACTO, no permite datos nulos
- TECNICO.AREA, no permite datos nulos

- TECNICO.DISPONIBLE, no permite datos nulos
- TECNICO.TECNICO, no permite datos nulos
- EMPRESA.NOMBRE, no permite datos nulos
- EMPRESA.PESO_TECNICO, no permite datos nulos
- EMPRESA.DIRECCION, no permite datos nulos
- EMPRESA.TELEFONO, no permite datos nulos
- EMPRESA.*STATUS*, no permite datos nulos

2.3. Discusión de resultados

A. El proceso de perfilación de datos permite conocer a fondo los datos de estudio con base en las estadísticas. Se generan, luego de la tabulación y el ordenamiento automatizado por algoritmos computacionales de los datos de estudio.

B. Los avances en recursos de procesamiento informático permiten que el proceso de perfilación sea cada vez más rápido, otorgando una forma sencilla y ágil de conocer el estado de grandes volúmenes de información para que el analista se dedique de lleno al análisis, sin perder recursos, procesando la información y recabando las estadísticas a mano.

C. Los requerimientos de análisis del negocio pueden ser expresados como reglas de negocio, y con la ayuda de las estadísticas generadas en la perfilación, relacionando directamente estas reglas de negocio con los datos estudiados. Lo que da origen a las reglas de datos, que luego serán auditadas para verificar su cumplimiento y lograr establecer niveles de calidad en los datos de estudio.

D. Según la naturaleza de las reglas de negocio obtenidas de los requerimientos de los usuarios tomadores de decisiones, se define que la dimensión de calidad asociada a dichas reglas es la Exhaustividad. Que indica que mientras más información se obtenga en la captura de datos, mayor será la calidad de dicho datos; de ahí las reglas generadas al final del inciso anterior.

3. DEFINICIÓN DE MÉTRICAS PARA ESTABLECER EL NIVEL DE CALIDAD DE DATOS NECESARIO EN UN AMBIENTE ANALÍTICO

3.1. Definición

Para la elaboración del capítulo de aporte se tomó como referencia el marco teórico y el caso de estudio analizado, de tal forma que se facilite el total entendimiento de la propuesta hecha en esta investigación. Se pretende demostrar una forma estructurada de medir la calidad de datos en un conjunto determinado de tablas de una base de datos.

3.2. Justificación

El objetivo de mantener almacenados los datos históricos de una empresa se resume en poder disponer de un historial de información que otorgue riqueza de análisis, por ejemplo, en comparaciones con información actual. Es por ello que si estos no mantienen un nivel de calidad aceptable, únicamente se está incurriendo en elevados costos para mantenerlos guardados. Y pudiera darse el caso de incurrir en costos de oportunidad para el negocio, si estos datos con baja calidad fueran utilizados para tomar alguna decisión importante. La propuesta es definir una metodología que mida certeramente el nivel de calidad en un conjunto de datos, para que, teniendo dicho parámetro se puedan tomar decisiones más efectivas sobre lo que los datos representan, -el negocio -.

3.3. Desarrollo

El desarrollo del presente capítulo se llevó a cabo sobre determinadas tablas de la base de datos de producción de una reconocida empresa dedicada a brindar servicios de soporte en línea.

Esta empresa centra su negocio en atender llamadas de sus clientes que necesitan soporte o consultoría técnica. Se analizaron tres importantes tablas de la base de datos y se establecieron los atributos que, según los usuarios del sistema al cual pertenecen las tablas, se necesita sean de calidad.

3.3.1. Tablas analizadas

A continuación se presenta la descripción de las tablas que fueron analizadas:

Tabla II. **Tamaño de la tabla EMPRESA**

Entidad:	EMPRESA
No. Registros:	1,245

Fuente: elaboración propia.

Descripción: en la entidad referida en la Tabla II se guarda la información de las empresas a las que se presta el servicio de soporte. La importancia de sus datos radica en las políticas que se resumen en el slogan: “conozca a su cliente para servirle mejor*”.*

Tabla III. **Tamaño de la tabla TECNICO**

Entidad:	TECNICO
No. Registros:	35

Fuente: elaboración propia.

Descripción: la Tabla III se refiere a la entidad que guarda el registro de los técnicos que prestan actualmente, o prestaron en algún tiempo pasado, el servicio de soporte para algún cliente. Su importancia es justificada porque, de los datos almacenados en la misma se hacen los cálculos para el pago de planilla, según el rendimiento de cada técnico, se otorga cierta bonificación monetaria al final de cada mes.

Tabla IV. **Tamaño de la tabla INCIDENTES**

Entidad:	INCIDENTES
No. Registros:	26,030

Fuente: elaboración propia.

Descripción: en la entidad referida en la Tabla IV se guardan los incidentes de soporte por los cuales los clientes han requerido servicio. Es necesaria la calidad de sus datos porque es la bitácora de órdenes de soporte y su información indica la eficiencia del servicio prestado, además del historial y comportamientos de ciertos clientes y sus productos adquiridos para los cuales se contrató el soporte.

3.3.2. Reglas de datos derivadas

Las reglas de datos requeridas son todas derivadas de la dimensión de calidad de datos: exhaustividad. Básicamente, se puede decir que es la misma regla para muchos atributos, en la Tabla V se podrá observar en detalle.

Tabla V. **Reglas de datos asociadas a las tablas seleccionadas**

Entidad	**Atributo**	**Regla derivada**
INCIDENTES	*DB_SERVER*	No permite nulos.
INCIDENTES	*PRODUCT_VERSION*	No permite nulos.
INCIDENTES	*OPEN_BY*	No permite nulos.
INCIDENTES	NO_EMPRESA	No permite nulos.
INCIDENTES	CONTACTO	No permite nulos.
TECNICO	AREA	No permite nulos.
TECNICO	DISPONIBLE	No permite nulos.
TECNICO	TECNICO	No permite nulos.
EMPRESA	NOMBRE	No permite nulos.
EMPRESA	PESO_TECNICO	No permite nulos.
EMPRESA	DIRECCION	No permite nulos.
EMPRESA	TELEFONO	No permite nulos.
EMPRESA	*STATUS*	No permite nulos.

Fuente: elaboración propia.

El objetivo trazado con la regla de datos derivada “No permite nulos*”* es contar con la mayor cantidad de información posible para un registro, ya que en cualquiera de las tablas analizadas, el no contar con información en alguno de los campos listados provoca que los análisis sean incompletos. Como el nombre de la dimensión lo indica, el análisis pierde en Exhaustividad en las propiedades que describen alguno de esos registros.

3.3.3. Cumplimiento de reglas de datos, auditoría

La auditoría realizada en la investigación, compete a las tablas que fueron perfiladas y a los atributos a los que fueron asignadas reglas de datos. Los datos pertenecientes a los atributos en cuestión fueron tabulados para establecer el cumplimiento de su regla asociada. Las siguientes tablas muestran la auditoría realizada en cada atributo que se le asoció con una regla de datos:

- Entidad: INCIDENTES

Tabla VI. **Auditoría de cumplimiento para el atributo *DB_SERVER***

Atributo:	*DB_SERVER*
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	26030
Incumplimientos:	25906
% Cumplimiento:	0.2%

Fuente: elaboración propia.

Tabla VII. **Cumplimiento para el atributo *PRODUCT_VERSION***

Atributo:	*PRODUCT_VERSION*
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	26030
Incumplimientos:	3410
% Cumplimiento:	86.9%

Fuente: elaboración propia.

Descripción: este campo indica la versión del producto respaldado por el servicio de la empresa que presentó el problema.

Tabla VIII. **Auditoría de cumplimiento para el atributo *OPEN_BY***

Atributo:	*OPEN_BY*
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	26030
Incumplimientos:	0
% Cumplimiento:	100%

Fuente: elaboración propia.

Descripción: con este atributo queda registrado el código del técnico que atendió el soporte por primera vez y creó el incidente.

Tabla IX. **Auditoría de cumplimiento para el atributo NO_EMPRESA**

Atributo:	NO_EMPRESA
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	26030
Incumplimientos:	0
% Cumplimiento:	100%

Fuente: elaboración propia.

Descripción: este campo registra el código de la empresa que se está atendiendo en ese incidente en concreto.

Tabla X. **Auditoría de cumplimiento para el atributo CONTACTO**

Atributo:	CONTACTO
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	26030
Incumplimientos:	25974
% Cumplimiento:	0.2%

Fuente: elaboración propia.

Descripción: este campo nombra al contacto dentro de la empresa cliente que llamó para pedir el servicio de soporte.

- Entidad: TECNICO

Tabla XI. **Auditoría de cumplimiento para el atributo AREA**

Atributo:	AREA
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	35
Incumplimientos:	1
% Cumplimiento:	97.1%

Fuente: elaboración propia.

Descripción: este campo guarda el área de soporte al cual pertenece el técnico y por ende el tipo de soporte que se está prestando.

Tabla XII. **Auditoría de cumplimiento para el atributo DISPONIBLE**

Atributo:	DISPONIBLE
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	35
Incumplimientos:	4
% Cumplimiento:	88.6%

Fuente: elaboración propia.

Tabla XIII. **Auditoría de cumplimiento para el atributo TECNICO**

Atributo:	TECNICO
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	35
Incumplimientos:	0
% Cumplimiento:	100%

Fuente: elaboración propia.

Descripción: este campo nombra al técnico dentro de la empresa y lo identifica para llevar el control de los soportes que han solventado.

- Entidad: EMPRESA

Tabla XIV. **Auditoría de cumplimiento para el atributo NOMBRE**

Atributo:	NOMBRE
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	1245
Incumplimientos:	0
% Cumplimiento:	100%

Fuente: elaboración propia.

Tabla XV. **Auditoría de cumplimiento para el atributo PESO_TECNICO**

Atributo:	PESO_TECNICO
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	1245
Incumplimientos:	348
% Cumplimiento:	72.00%

Fuente: elaboración propia.

Descripción: este campo indica el nivel de esfuerzo requerido para satisfacer los requerimientos de soporte de esta empresa.

Tabla XVI. **Auditoría de cumplimiento para el atributo DIRECCION**

Atributo:	DIRECCION
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	1245
Incumplimientos:	54
% Cumplimiento:	95.70%

Fuente: elaboración propia.

Descripción: este campo guarda la dirección de la empresa cliente, es importante ya que los técnicos hacen visitas presenciales en algunos soportes.

Tabla XVII. **Auditoría de cumplimiento para el atributo TELEFONO**

Atributo:	TELEFONO
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	1245
Incumplimientos:	65
% Cumplimiento:	95.80%

Fuente: elaboración propia.

Descripción: este campo registra el teléfono de planta de la empresa cliente, o bien el del contacto más frecuente.

Tabla XVIII. **Auditoría de cumplimiento para el atributo *STATUS***

Atributo:	*STATUS*
Regla de datos	No permite datos nulos.
Dimensión:	Exhaustividad
Registros:	1245
Incumplimientos:	519
% Cumplimiento:	58.30%

Fuente: elaboración propia.

Descripción: este campo indica el estado actual de licenciamiento de soporte para una empresa.

3.3.4. Métricas y nivel de calidad de datos

El propósito de esta investigación es definir las métricas necesarias para evaluar el nivel de calidad de datos dentro de una empresa, en un ambiente analítico. Una vez descritas las distintas dimensiones de la calidad de datos en el marco teórico, se debe encontrar una forma de poder medir y establecer un nivel de calidad para cada una de ellas. La forma encontrada, para medir esas dimensiones, fue asociar las reglas de datos generadas, por medio de las políticas del negocio, a la dimensión que corresponda y auditar su cumplimiento.

Todas las reglas pertenecen a una única dimensión: la exhaustividad. El nivel de calidad se medirá por tabla, de acuerdo con el promedio correspondiente al porcentaje de cumplimiento obtenido por cada uno de sus campos que tenían una regla de datos asociada.

El cálculo del nivel de calidad de la primera tabla (INCIDENTES) se realiza promediando los niveles de cumplimiento de reglas de datos de cada uno de los atributos que tenía alguna regla asociada.

$$(0.2 + 86.9 + 100 + 100 + 0.2) / 5 = 57.46\%$$

Tabla XIX. **Porcentaje promedio de cumplimiento para la tabla INCIDENTES**

Entidad:	INCIDENTES
No. Registros:	57.46%

Fuente: elaboración propia.

De la misma forma se realiza con la siguiente tabla (TECNICO) para obtener el porcentaje unificado de calidad para esta tabla.

(97.1 + 88.6 + 100) / 3 = 95.23%

Tabla XX. **Porcentaje promedio de cumplimiento para la tabla TECNICO**

Entidad:	TECNICO
No. Registros:	95.23%

Fuente: elaboración propia.

Finalmente, se realiza el cálculo para la última de las tablas analizadas (EMPRESA), para obtener el nivel de calidad en la dimensión de exhaustividad.

(100 + 72 + 95.7 + 94.8 + 58.3) / 5 = 84.16%

Tabla XXI. **Porcentaje promedio de cumplimiento para la tabla EMPRESA**

Entidad:	EMPRESA
No. Registros:	84.16%

Fuente: elaboración propia.

El análisis de resultados ha entregado lo que se estableció en los objetivos de esta investigación, el estado o nivel de calidad con el que cuentan tres distintas tablas.

Si se determinara que un nivel aceptable de calidad para cumplir con las necesidades del negocio es de 80%, entonces, se concluiría que únicamente dos de tres tablas analizadas cuentan con un nivel de calidad, por encima de lo aceptable para el negocio estudiado.

3.4. Metodología

3.4.1. Título

Metodología de medición de calidad de datos utilizando perfilado de datos y auditorías de reglas de datos asociadas con el negocio.

3.4.2. Descripción

Basa los resultados en la buena elección de los datos de estudio, el análisis de los mismos, por medio de *data profiling*, la extracción de reglas del negocio y derivación a *data rules,* y el monitoreo de cumplimiento de las reglas en los datos actuales. El resultado estará representado en un valor numérico porcentual que indicará el nivel de calidad de los datos de estudio.

3.4.3. Métodos

- Elección del conjunto de datos de estudio
- Proceso de perfilación de los datos de estudio
- Análisis de resultados estadísticos del proceso de perfilación
- Comparación de resultados y estado actual de los datos contra la reglas del negocio estudiado
- Derivación de reglas de datos con base en las reglas del negocio

- Medición de cumplimiento del conjunto de datos de estudio contra las reglas de datos derivadas del paso anterior
- Establecimiento del nivel de calidad de los datos estudiados

3.4.4. Comentario

Esta metodología fue puesta a prueba en el caso de estudio de la presente investigación y aprobada por el asesor designado del proyecto, de acuerdo con el análisis y los resultados obtenidos.

CONCLUSIONES

1. El conjunto de datos elegido como fuente de la investigación metodológica fue acorde con el primer objetivo planteado, ya que permitió el correcto desarrollo de la búsqueda y presentación de la metodología.

2. La perfilación de los datos permitió la obtención de un conocimiento más adecuado, a cerca del conjunto de datos a estudiar. Por medio de este proceso, fue posible obtener las reglas de datos.

3. La auditoría de cumplimiento de reglas de datos permitió que se clarificara el camino óptimo para determinar el nivel de calidad del conjunto de datos escogido. Este paso fue obligado incluirlo en la metodología presentada.

4. Las métricas que pueden responder a la evaluación del nivel de calidad de un conjunto de datos son los porcentajes de cumplimiento de las reglas de datos asociadas a los mismos. Estas métricas fueron definidas y respecto de las mismas fueron evaluados los datos, estableciendo el porcentaje de cumplimiento y resumiéndolo en su nivel de calidad.

5. Aunque existan cinco o más dimensiones desde las cuales estudiar la calidad de datos, cada una de estas solo se podrá medir si las políticas de la empresa permiten generar reglas de datos que puedan ser relacionadas con las dimensiones.

6. La dimensión de exactitud, por su naturaleza, es muy difícil de medir ya que se deberá comprobar campo por campo si su información es verdadera, de acuerdo con el referente.

7. La perfilación de datos permite obtener una visión más amplia del estado de los datos y ayuda a la generación de reglas de datos. La mayoría de reglas generadas, en esta investigación, fueron identificadas en esta etapa del proceso.

8. El análisis dejó tres porcentajes de calidad diferentes para cada una de las tablas. Pero, corresponde a cada empresa, según su tipo de negocio, determinar si dicho porcentaje alcanza o no un nivel de aceptabilidad para sus necesidades de análisis.

9. Las métricas pueden variar para cada evaluación de calidad, si los datos difieren y, más aún, sí el tipo de negocio al que pertenecen los datos es otro.

RECOMENDACIONES

1. La empresa que prestó los datos de análisis, para el mejoramiento de la calidad de sus datos, debe colocar filtros en la aplicación que capta los datos de la tabla INCIDENTES, específicamente, en los campos *DB_SERVER* y CONTACTO. Ya que estos son los responsables directos que del porcentaje de calidad de esa tabla, que es casi del 50%.

2. Las empresas guatemaltecas que manejen sistemas para toma de decisiones deben de realizar una evaluación constante de la calidad de su información, ya que una rápida estrategia de identificación de incumplimiento de reglas de datos puede ayudar, considerablemente, en un mejor desarrollo y productividad de la empresa.

3. Las y los estudiantes de carreras afines a la informática y futuros administradores de bases de datos, deben profundizar en los temas de *data governance* y *data quality*. Por alguna razón válida, en los países líderes en tecnología de punta, toman muy en cuenta estos aspectos para tenerlos en consideración a la hora de implementar o dar mantenimiento a sistemas, no solo analíticos, sino también transaccionales.

4. Asimismo, los docentes encargados de las actualizaciones de los currículos de estudios, de las diferentes casas de estudios y de las carreras afines a la informática, deben incluir activamente los temas del punto anterior, para cumplir con sus propias tareas derivadas de su quehacer profesional y para dotar a los estudiantes de las habilidades para utilizar las herramientas actualizadas y que están disponibles en el mercado, para formar profesionales competentes y con una actuación de capacidad que conlleve el éxito en el ejercicio profesional.

BIBLIOGRAFÍA

1. BATINI, Carlo; SCANNAPIECA, Monica. *Data quality: concepts, methodologies and techniques.* Italia: Springer, 2006. 262 p.

2. FERNÁNDEZ, Carlos. [en línea]. 17 de junio de 2010. Disponible en Web: <http://www.dataprix.com/category/integracion-datos/perfilado-datos/>.

3. GÁLVEZ, Luis. *Oracle data quality option.* Guatemala: Datum, 2010. 7 p.

4. *Integración y Calidad de Datos.* [en línea]. 17 de julio de 2008. Disponible en Web: <http://integracionycalidad.blogspot.com/2008/07/migraciones-fusiones-y-adquisiciones.html>.

5. LINDSEY, Ed. *Three-dimensional analysis: data profiling techniques.* New York: Ed Lindsey Publications, 2008. 240 p.

6. LOSHIN, David. *Monitoring data quality performance using data quality metrics.* Los Angeles: Informática, 2006. 22 p.

7. MACMILLAN, Palgrave. *Information technology and the structure of the multinational enterprise.* Illinois: Ltd. Reddy, S. B., 1994. 325 p.

8. MAYDANCHIK, Arkady. *Data quality assessment.* New York: Techniques Publications, 2007. 312 p.

9. OLSON, Jack E. *Data quality: the accuracy dimension.* San Francisco: Morgan Kaufmann, 2003. 293 p.

10. *Oracle data quality for data integrator and oracle data profiling 11g.* San Francisco: Oracle, 2010. 3 p.

11. PETER, Rob; CORONEL, Carlos. *Database systems: design, implementation, and management.* 6a. ed. Toronto: Course Technology, 2004. 794 p.

12. ROCHNIK, Nikolai. *Oracle warehouse builder 10gR2: transforming data into quality information.* San Francisco: Oracle, 2006. 16 p.

13. REDAMN, Thomas C. *Data quality: the field guide.* San Rafael: Digital Press, 2001. 256 p.

14. TJOA, A Min; TRUJILLO, Juan. *Data warehousing and knowledge discovery.* Copenhagen, Dinamarca: Springer Science & Business, 2005. 538 p.

15. TRACTINSKY, Noam; JARVENPAA, S. *Information systems design decisions in a global versus domestic context.* Minnesota: Management Information System Quarterly, 1995. 507 p.

16. TRKMAN, P.; MCCORMACK, Kevin. *The impact of business analytics on supply chain performance.* Ljubijana: Decision Support Systems. 2010. 10 p.

17. WANG, Richard Y. *Data quality.* Los Angeles: Springer, 2001. 167 p.

18. WANG, Richard Y. *Information quality.* Los Angeles: Sharpe, 2005. 256 p.