Alternativas de herramientas para facilitar el monitoreo de la calidad de datos

Tool alternatives to facilitate the monitoring of data quality

DOI: 10.46932/sfjdv3n2-026

Received in: February 15th, 2022
Accepted in: March 1st, 2022

Monica Rosa Lopez-Guayasamin
Institución: Universidad Nacional de Colombia Sede Manizales, Facultad de Administración, Colombia
Dirección: a 62-338,, Cra. 25 #62236, Manizales, Caldas, Colômbia
Correo electrónico: mrlopezg@unal.edu.co

Nestor Dario Duque-Mendez
PhD. Ingeniería
Institución: Universidad Nacional de Colombia Sede Manizales, Facultad de Administración, Colombia
Dirección: a 62-338,, Cra. 25 #62236, Manizales, Caldas, Colômbia
Correo electrónico: ndduqueme@unal.edu.co

RESUMEN
Las empresas están manejando grandes volúmenes de datos, los cuales deben ser almacenados de manera óptima en una infraestructura adecuada y que sea escalable. Sin embargo, el gran problema que enfrentan las organizaciones es el control y monitoreo de la calidad de los datos, pues siendo la calidad un tema tan relevante, aún hay asuntos por enfrentar acordes con la importancia que merece. El trabajo presentado en este artículo ofrece un recorrido sobre experiencias en herramientas que permiten a las empresas monitorear y controlar la calidad de sus datos; para lo cual se evaluaron diferentes plataformas tecnológicas. Teniendo en cuenta que la calidad de los datos incorpora diferentes dimensiones que facilitan este análisis, se incorpora en este trabajo una revisión exploratoria de las mismas, se definen Completitud y Validez como las dimensiones a trabajar en este reto. Se inicia una exploración de calidad de datos con herramientas libres; luego se continua el ejercicio con una herramienta de analítica, muy poderosa y de fácil uso como SPSS Modeler y por último se trabaja sobre una herramienta licenciada DQS Data Quality Server que facilita el ejercicio de definición de reglas y su aplicabilidad en las organizaciones.

Palabras clave: calidad de datos, dimensiones, visualización, herramientas.

ABSTRACT
Companies are managing large volumes of data, which must be optimally stored in adequate infrastructure and scalable. However, the big problem for using these data is the visualization for controlling the quality of the same; since quality is such a relevant issue, there are still issues to be faced following the importance it deserves.The work presented in this article offers a journey on experiences in tools that allow companies to monitor and control the quality of their data; for which different technological platforms were evaluated. Taking into account that the quality of the data incorporates different dimensions that facilitate this analysis, an exploratory review of the same is incorporated in this work, Completeness and Validity are defined as the dimensions to be worked on in this challenge. A data quality scan is started with free tools; Then the exercise is continued with a very powerful and easy-to-use analytical tool such as SPSS Modeler and finally, we work on a DQS Data Quality Server licensed tool that facilitates the exercise of defining rules and their applicability in the company.

Keywords: data quality, dimensions, visualization, tools.

1 INTRODUCCIÓN

El problema de calidad de datos se ha trabajado en las empresas, pero en ocasiones no ha tenido la suficiente importancia hasta que la consolidación de los datos y el uso de estos, no han dado las señales esperadas por los negocios. Uno de los puntos importantes y que ocupa mucho tiempo en un estudio analítico es la revisión de la información, pues es ahí donde el analista de datos se enfrenta a dos retos importantes en este proceso: uno eliminar los datos con problemas, y el otro es limpiar la data, proceso en el cual se pueden perder señales importantes de estos datos.

En las empresas, los procesos deben garantizar calidad y confiabilidad en la información de la cual son responsables; sin embargo, uno de los puntos más complicados es como responder ante esta tarea de manera oportuna y con las herramientas adecuadas. Por lo tanto, el reto es contar con una herramienta que permita de manera automática hacer el seguimiento y control de la calidad de los datos y que dicho control pueda ser monitoreado en el tiempo.

Este documento, comparte la experiencia que se ha tenido con diferentes herramientas tecnológicas para facilitar el monitoreo de la calidad de datos sobre un grupo de atributos los cuales son definidos de manera importante para el negocio. Se contribuye a la empresa definiendo una forma de como implementar control y seguimiento a la calidad mediante un ejercicio articulado y ordenado que inicia desde la parametrización de los datos, el análisis de la información en las dimensiones de calidad de Completitud y Validez, la consolidación y almacenamiento de los procesos de calidad y por último la visualización del comportamiento de la misma.

2 REVISIÓN BIBLIOGRÁFICA

Para la revisión bibliográfica asociada a la calidad de datos, se aplica la ecuación definida en la tabla 1 donde se trabajan palabras asociadas al concepto calidad, problemas y métricas. Se aclara que los artículos seleccionados fueron consultados de las bibliotecas digitales como son: IEEE, Springer, Science direct, Scielo, MINTIC entre otros.

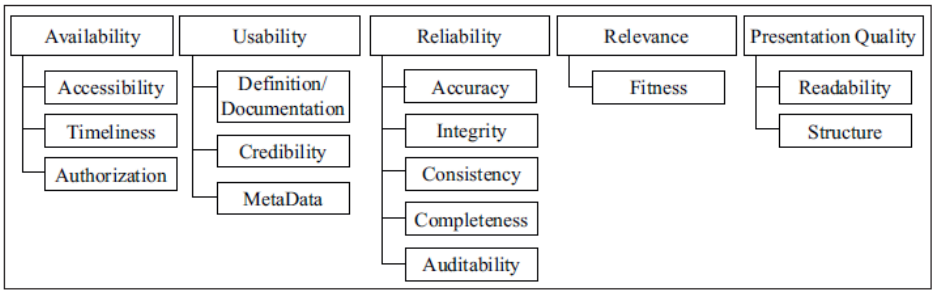
Tabla 1 Ecuación bibliográfica para calidad de datos.

COMPONENTES (AND)		
PALABRAS CLAVE (OR)	Calidad de datos	Métricas de calidad de datos
	Calidad	metric*
	quality	ind*
	problems	
ECUACIÓN DE BUSQUEDA		
(Calidad OR quality OR problems) AND (metric* OR ind*))		
Fuente: Elaboración propia		
Criterios:		
Año de publicación: Mayor a 2015		
Tipo de publicación: Articulo científico o libros		
Idioma: Inglés o español		

Mediante esta ecuación se identifican artículos asociados a la temática, y aunque este proceso exige una mayor depuración de los artículos, finalmente se encontraron más elementos de estudio que favorecieron el proceso investigativo.

Existen muchas definiciones asociadas al concepto calidad de datos, según Power Data " Es la cualidad de un conjunto de información recogida en una base de datos, un sistema de información o un data warehouse que reúne entre sus atributos la exactitud, completitud, integridad, actualización, coherencia, relevancia, accesibilidad y confiabilidad necesarias para resultar útiles al procesamiento, análisis y cualquier otro fin que un usuario quiera darles" [1]. De acuerdo al autor Cai [2] en la figura 1 se muestra un estándar universal de calidad de datos en dos capas y que puede ser aplicado a cualquier conjunto de datos, donde las dimensiones relevantes para dicho análisis son: Disponibilidad, Facilidad de uso, Confiabilidad, Importancia y Calidad de presentación.

Figura 1 Un estándar universal de calidad de big data de dos capas para la evaluación. [2]



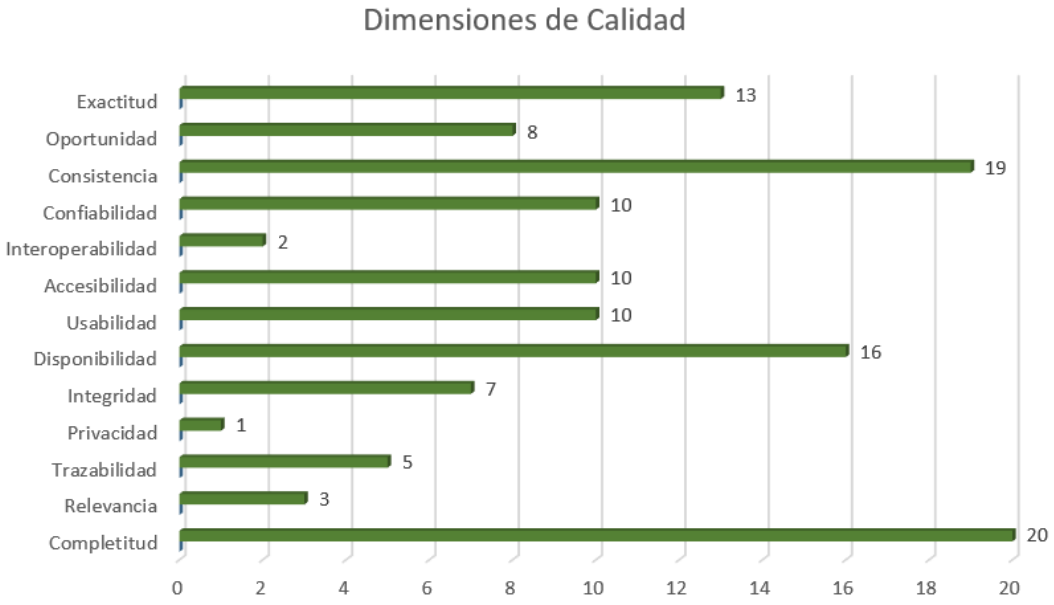
Por otro lado, en la revisión de la literatura se encuentra que los estudios de calidad de datos a nivel mundial se ha centrado en temas puntuales como Calidad en Datos Abiertos [3], [4], [5], Evaluación de Calidad [6], Riesgos en Calidad de Datos [7], Calidad en Big Data [8], Herramientas de Calidad [9], [10], Problemas Calidad de Datos [11], Calidad Datos Gobierno [12], [13], [14], Métricas de Calidad [15], [13], [16], [17] entre otros.

En la figura 2 se presenta un consolidado de las temáticas abordadas en los artículos que hicieron parte de este análisis asociado al tema de calidad de datos.

Fig. 2 Revisión Bibliográfica Calidad de Datos

Otro tema que también se aborda paralelo a esta revisión bibliográfica es el interés de la comunidad en los problemas de calidad de datos particularmente en los datos abiertos y cuáles son las dimensiones más aplicadas y abordadas por los autores siendo esto el punto de referencia para identificar que las dimensiones mas mencionadas o trabajadas en la comunidad son completitud y consistencia como se pueden evidenciar en la figura 3 donde dichas dimensiones son referenciadas por la cantidad de artículos aquí asociados.

Figura 3 Dimensiones de Calidad en Open Data. Fuente: Elaboración Propia



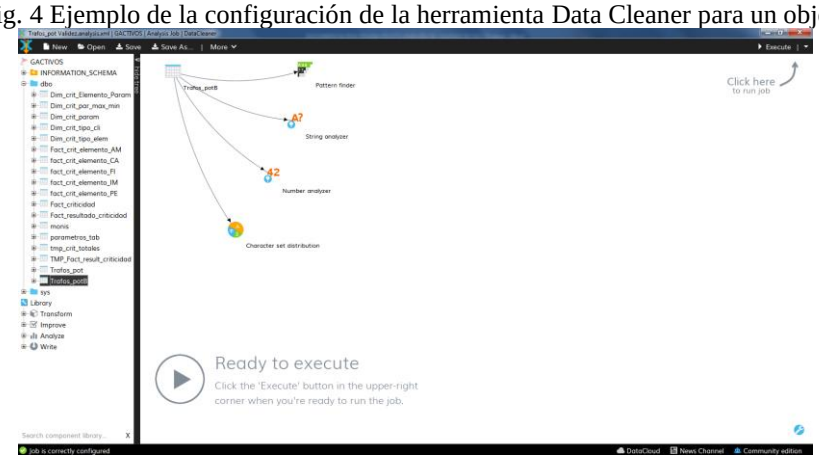
3 METODOLOGÍA

Según [18] Rivadera la metodología Kimball es de gran aceptación en las empresas ya que se pueden implementar repositorios de datos con temas específicos. El presente trabajo utilizó esta metodología para apoyar la consolidación de los ejercicios de calidad de datos asociados a este trabajo.

En la figura 2 se ilustran las principales etapas de Kimball definidas en la metodología y se mapea con las diferentes etapas que normalmente se realizan en un proceso de calidad. A continuación se explica como se realiza el mapeo entre las dos metodologías:

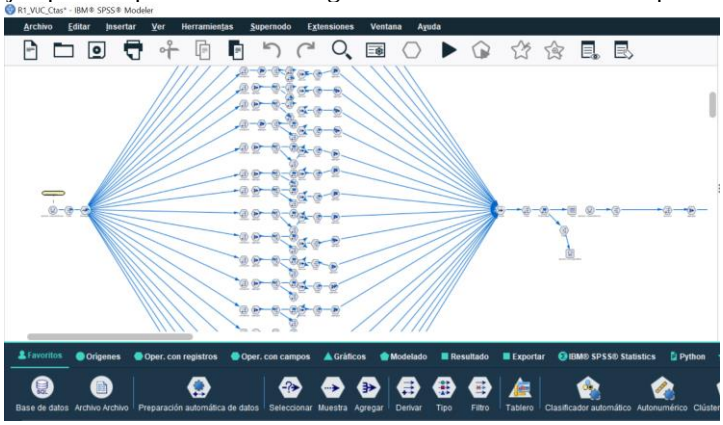
- En la metodología Kimball la etapa de planificación es donde se define el alcance de la solución y se definen las bases de la información que es relevante para el ejercicio. Si consideramos que desde calidad lo primero que debemos realizar es identificar las necesidades de calidad en la información, estamos asociando de manera adecuada estos dos pasos.
- La fase de análisis de requerimientos definida en Kimball requiere un acercamiento y un conocimiento del negocio de manera profunda donde se pueda identificar las necesidades de elementos analíticos o relevantes a incorporar en el modelamiento de negocio. La siguiente etapa de calidad es identificar las fuentes de información a las cuales se les debe procesar calidad de datos, incluyendo los atributos relevantes para el proceso etapa que se incorpora de manera natural con las tareas que se realizan en la etapa de requerimientos.
- La fase de Diseño de Arquitectura es donde se identifica a nivel de tecnología la plataforma adecuada para el proceso de modelamiento. Si mapeamos esta etapa con calidad, es cuando definimos con que herramienta incorporamos calidad que para efectos del artículo se trabajaron en diferentes etapas del tiempo Data cleaner, Spss Modeler y DQS Data Quality Server.
    - Data Cleaner. Software de licenciamiento corto de forma libre en el que la configuración define las conexiones a las bases de datos fuente y la inclusión de las variables a revisar en cada modelamiento. En la Figura 4 se visualiza un ejemplo de configuración con la herramienta Data Cleaner mediante el uso de los objetos Patter Finder (Identificación de patrones), String Analizer (Identificación de cadenas), Number Analizer (Identificación de patrones en números), Character set distribution (Patrones en cadenas) que hacen parte de la dimensión Validez

Fig. 4 Ejemplo de la configuración de la herramienta Data Cleaner para un objeto



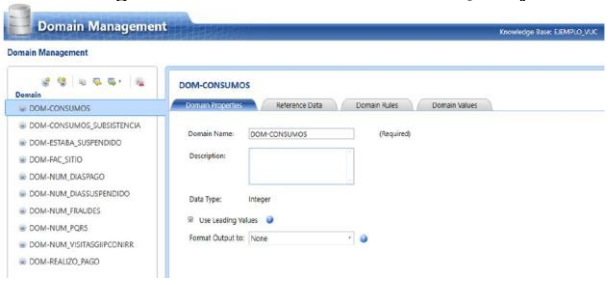
- Spss Modeler. Software licenciado de análisis de texto y minería de datos de IBM, la configuración y modelamiento de los ejercicios de validez se realizan en la plataforma utilizando los componentes básicos para la transformación de los datos. La Figura 5 permite visualizar un ejemplo del uso de componentes como Derivar (transforma valores), Agregar (implementa funciones), Filtro (excluye datos), Fundir (consolida datos) para poder modelar la data original y transformarla de acuerdo con las reglas definidas en validez.

Fig. 5 Ejemplo de Spss Modeler configurando un modelo de validez para un objeto



- DQS Data Quality Services software licenciado que permite que un administrador de datos mantenga la calidad de los datos mediante la administración y configuración de variables. La configuración se hace por medio de la definición de dominios, como se aprecia en la Figura 6, en donde se parametrizan reglas por cada uno, lo cual facilita el modelamiento de los datos para que al ejecutar el modelo se obtengan las respuestas acordes a las expectativas del usuario.

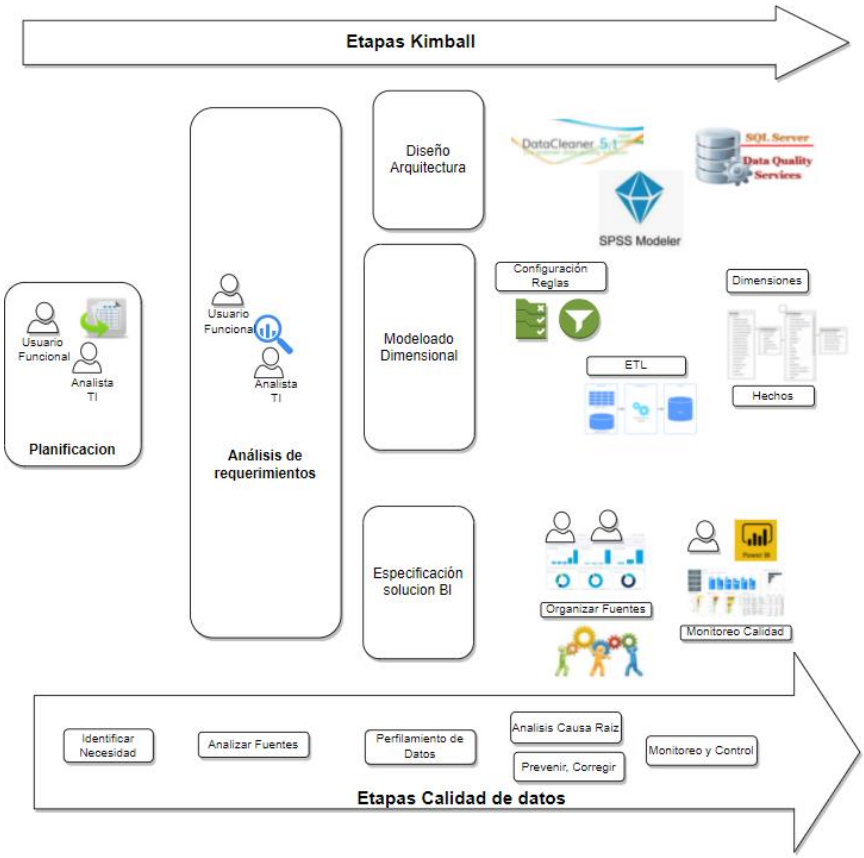
Fig. 6 Ejemplo de la configuración en la Plataforma Data Quality Server para un objeto



- La fase Modelado dimensional es donde se definen las estructuras y la forma en que se deberá procesar la extracción, transformación y carga de datos de los sistemas transaccionales al nuevo modelo. De igual forma lo que se defina en la metodología de calidad en como se deberá analizar y almacenar los resultados del perfilamiento de los datos debería incorporarse en las estructuras que se modelen en esta fase.

- Especificación solución BI es la etapa desde Kimball donde se podrá visualizar y monitorear los resultados requeridos por el negocio. En esta fase por lo general se incorporan herramientas tecnológicas adicionales que permitirán mediante procesos de visualización tener una mirada mas completa de la solución a nivel de usuario final. Desde Calidad esta es la parte donde lo que se ha procesado se puede materializar en reportes que permitan al usuario final identificar causa – raíz de calidad de datos, el usuario podrá identificar elementos fundamentales para tomar decisiones de como prevenir y corregir datos (cuando sea necesario), lo anterior debido a que una solución BI facilitaría el monitoreo y control de la calidad de los datos.

Fig. 7 Etapas de Kimball vs Etapas de Calidad



4 EXPERIMENTACIÓN

En esta sección se exponen 3 escenarios de calidad con las plataformas tecnológicas definidas con el fin de dar claridad sobre cada uno de los pilotos realizados.

Inicialmente se debe aclarar que para cualquiera de los pilotos, el primer paso es definir la estructura con la cual se debe trabajar el piloto de calidad de datos. En esta fase se debe identificar de todos los atributos asociados al objeto cuales serían los atributos a los cuales se les aplicaría calidad de datos, y deberá ser el usuario final quien determine que peso deberá darle a cada atributo para evaluar su

calidad. Como premisa la sumatoria de los pesos de los atributos de un objeto finalmente deberá dar un 100 porciento.

Por otro lado, deberá ser el usuario quien defina del 100 porciento de la calidad del dato que porcentaje considera deberá dar a la dimensión completitud y cual porcentaje deberá asignar a la dimensión validez, teniendo en cuenta que estos son los atributos que finalmente se evaluarán en estos pilotos.

4.1 HERRAMIENTA TECNOLÓGICA DATA CLEANER APLICADO EN UN CASO DE NEGOCIO PARA DISTRIBUCIÓN.

Para este piloto, el negocio de distribución define que los objetos a analizar son algunos elementos que contienen la topología de la red de la empresa, como se muestra en la figura 8 donde los elementos escogidos son “conductor primario” o un “reconectador de 13,2 y 33”. Si revisamos el objeto conductor primario se evidencia que se escogen 10 atributos cuyo peso en porcentaje asignado a cada uno es 10 y la sumatoria de todos los pesos para este objeto da un 100 porciento. En el caso del objeto “reconectador de 13,2 y 33” se escogen 8 atributos, donde el peso dado a cada uno de ellos es 12,5 y la sumatoria de todos los pesos para ese objeto y corresponde a un 100 porciento.

Fig.8 Parametrización de porcentajes de variables en Excel

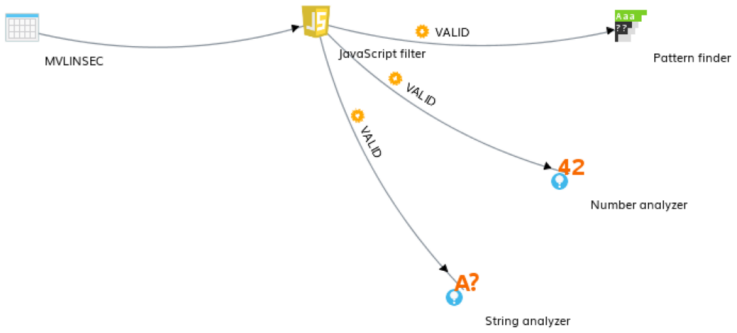
DESCRIPCION CLASIFICACION (ACTIVO)		ATRIBUTO DEL ACTIVO	
DESCRIPCIÓN	PONDERACIÓN (%)	NOMBRE DEL ATRIBUTO	PONDERACIÓN (%)
CONDUCTOR PRIMARIO	8,33%	TENSION_CONN	10%
		NUMERO_FASES	10%
		MATERIAL	10%
		AISLAMIENTO	10%
		TIPO_AISLAMIENTO	10%
		CIRCUITO_CONN	10%
		LOCALIZACION_CONN	10%
		USO	10%
		CALIBRE	10%
		LONGITUD	10%
RECONECTADOR, 13,2 y 33	8,33%	NUMERO_FASES	12,5%
		CORR_NOM	12,5%
		TIPO_GES	12,5%
		TIPO_AISLAMIENTO	12,5%
		CORR_INT	12,5%
		BIDIRECCIONAL	12,5%
		I_NOMCOR	12,5%
		FASES_CONN	12,5%

La anterior definición es importante porque son las bases sobre las cuales se calcula el comportamiento de los datos. Por otro lado, con la herramienta Data Cleaner se procede a configurar el modelo a trabajar dependiendo de las dimensiones de calidad que se requiera, para el ejercicio se configura el modelamiento para la dimensión validez. Una vez los resultados del modelamiento se ejecuten estos deberán ser exportados a archivo .dat para luego ser consolidados en Excel y posteriormente se procede con la visualización para el usuario final en Excel.

En la figura 9 se visualiza como desde la plataforma Data Cleaner se realiza el montaje del modelamiento del comportamiento de validez de los datos de la entidad mvlinsec que finalmente tiene la información del conductor primario y el reconectador. En este modelo se utiliza el componente pattern

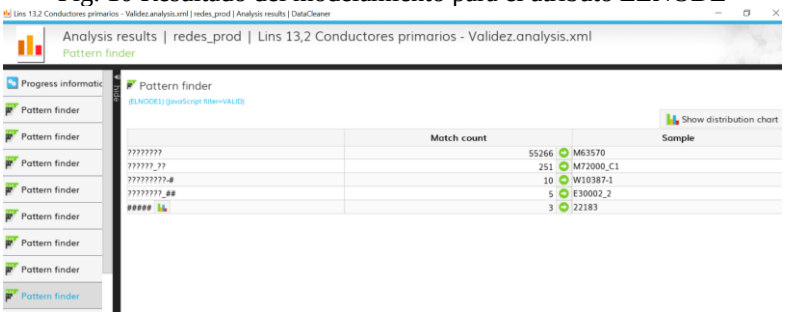
Finder para revisar patrones de comportamiento de los atributos y los componentes Number analyzser y String analyzer se utilizan para analizar la consistencia de los atributos numéricos y tipo cadena del objeto.

Fig. 9 Configuración de modelo de validez asociado a la estructura mvlinsec



Una vez ejecutado el anterior proceso se obtiene el resultado de las variables programadas de manera visual en la herramienta como ejemplo se evidencia que en la variable ELNODE se encuentra un patrón atípico que corresponde a 3 registros de información que los diferencia de los otros patrones como es el caso ejemplo 22183, en la figura 10 se puede visualizar este comportamiento.

Fig. 10 Resultado del modelamiento para el atributo ELNODE



Para terminar de entender el caso, en la figura 11 se visualiza los datos que contienen el patrón analizado que corresponde a atributos numéricos con 5 números, donde el usuario al visualizar los datos podrá ver los registros cuyo atributo ELNODE contengan 5 números. Esto le permite al usuario identificar los registros que tienen problema de calidad y poder tomar decisiones adecuadas para el ajuste de la calidad de los datos desde la fuente de información.

Fig. 11 Resultado del modelamiento para el atributo ELNODE con patron ####

Records (3) View detailed rows Save dataset

CODE	DESCRIPTIO	ADDRESS	ASSEMBLY	USER_	PHASES	METERCODE	CUSTOMER_0	CUSTOMER_1	FPARENT	XPOS1	YPOS1	XPOS2	YPOS2	ELNOD
57019	USO	<null>	<null>	REPOSICI...	6	<null>	<null>	<null>	HER23L14	1160600.5	1030190.1	1160616.8	1030150.1	22183
57049	USO	<null>	<null>	MANTENI...	7	<null>	<null>	<null>	ROS23L14	1157882.3	1027510.4	1157913	1027494	15266
57489	USO	<null>	<null>	REPOSICI...	15	<null>	<null>	<null>	AZA23L18	1174983.7	1050463.4	1174905.2	1050485.3	40640

4.2 HERRAMIENTA TECNOLÓGICA SPSS MODELER APLICADA EN UN CASO DE NEGOCIO COMERCIAL

En esta ocasión el negocio comercial decide que es importante aplicar calidad de datos en una estructura que contiene la consolidación de variables relevantes para el negocio, cuya información es tomada directamente del sistema comercial y posteriormente es almacenada como copia en un repositorio de datos alterno. Dicha información se encuentra en una estructura se llama “Caracterizacion de Cuentas”, igualmente el usuario funcional debe decidir a cuales atributos considera les debe aplicar calidad de datos, decidir a cuales les debe aplicar un peso o porcentaje que permita definir cual atributo es más relevante. En este ejercicio cada atributo tiene un peso del 4% y la sumatoria de los atributos deberán dar el 100% de la estructura, lo cual se muestra en Figura 12 donde se presenta la distribución de porcentajes para valorar la calidad de datos de acuerdo a lo referenciado por el usuario final.

Fig. 12 Distribución de porcentajes de valoración para la estructura “Caracterizacion de Cuentas”

DESCRIPCION DE LA CLASIFICACION (ACTIVO)		ATRIBUTO DEL ACTIVO	
DESCRIPCION	PONDERACION (%)	NOMBRE DEL ATRIBUTO	PONDERACION (%)
Caracterizacion de Cuentas	100%	AÑO	4%
		cant_pqrs	4%
		cantidad_visitas_giip	4%
		cantidad_visitas_giip_con_irr	4%
		cart_30	4%
		cart_60_90_120	4%
		consumo_subsistencia_kwh	4%
		consumos_kvarh	4%
		consumos_kwh	4%
		dias_pago	4%
		dias_suspendido	4%
		estaba_suspendido	4%
		facturacion_kvarh	4%
		facturacion_kwh	4%
		facturacion_sitio	4%
		MES	4%
		niu	4%
		realizo_pago	4%
		tiene_consumo_kwh_no_cero	4%
		valor_compensado	4%
		valor_contribucion	4%
		valor_pago	4%
		valor_subsidio	4%
		valor_total_facturado	4%

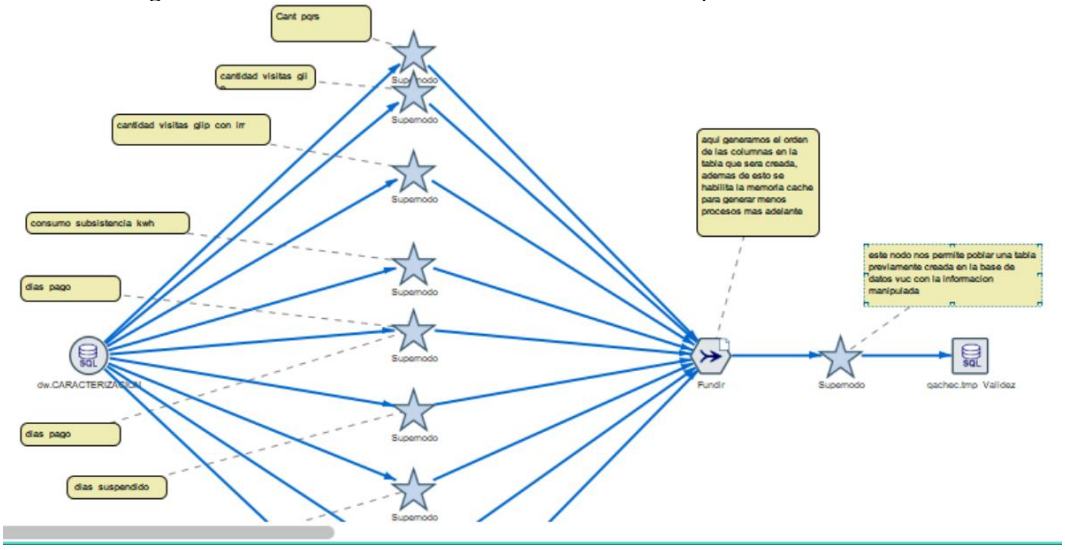
En este nuevo ejercicio de calidad, se utiliza la herramienta Spss Modeler donde lo primero que hay que definir es cual sería la variable y el caso de estudio para analizar calidad. Para este ejercicio se define que la dimensión a explorar de calidad es Validez y que la regla que se desea aplicar en el modelamiento esta definida por rangos Bueno, Malo, Atípico los cuales finalmente son definidos por el usuario final. En la figura 13 se puede visualizar la configuración de criterios de validez para las variables “cantidad_pqrs” y “cantidad_visitas_giip” definidos por el usuario final.

Fig. 13 Configuración de criterios de validez para atributos de “caracterizacion de cuentas”

cant_pqrs				cantidad_visitas_giip			
BUENO	MALO	ATIPICO		BUENO	MALO	ATIPICO	
Mayor a 0 y menor que 4	Menor a 0	Mayor a 4		Mayor = 0 y Menor = 2	Menor 0	Mayor a 3	

En la herramienta SPSS Modeler se debe estructurar un modelo que incorpore funciones que permitan la configuración de cada una de las variables a revisar como se muestra en la figura 14 donde cada estrella corresponde a un super nodo, e internamente tiene la configuración de funciones requerido para analizar cada caso y gestionar su comportamiento, por ultimo la funcion fundir consolida todos los resultados, mediante otro super nodo se configuran los cálculos finales por cada atributo, sus valoraciones y porcentajes para finalmente en la función registrar base de datos se incorpora en una estructura temporal todo el resultado del ejercicio calculado.

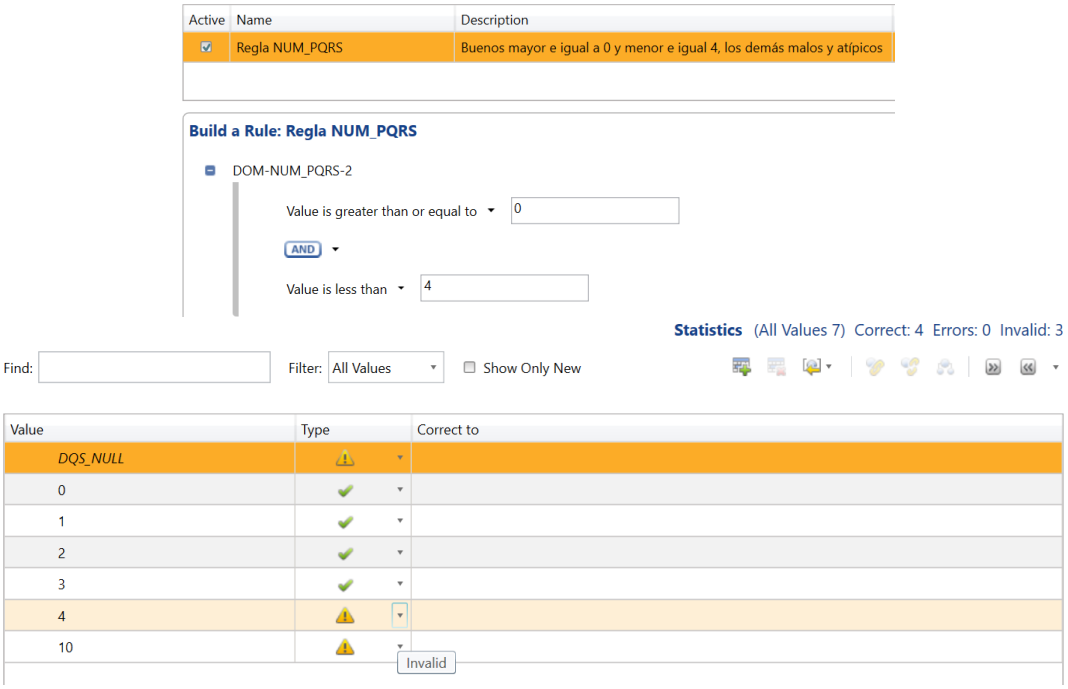
Fig. 14 Modelamiento de variables en SPSS Modeler para el caso comercial



4.3 HERRAMIENTA TECNOLÓGICA DQS DATA QUALITY SERVER APLICADA EN UN CASO DE NEGOCIO COMERCIAL

Para la implementación del proyecto de calidad con esta plataforma se respetaron las mismas condiciones del ejercicio anterior, lo que significa que se toman como referencia la misma estructura, variables, porcentajes y pesos asociados por el usuario final. Al cambiar la plataforma de modelamiento es necesario configurar en la herramienta las reglas necesarias para cumplir con los criterios definidos por el usuario, por ejemplo para este ejercicio la regla asociada cantidad pqrs indica que los valores entre 0 y 4 son correctos de lo contrario son malos o atípicos. Una forma de ver cono se configuran estos valores se presenta en la figura 15, donde se visualiza como se parametriza la regla y en la otra grafica se evidencia la ejecución de la regla en algunos casos.

Fig. 15 configuración de variables en DQS para el caso num_pqrs



Una vez configuradas las variables en Data Quality Server, se procede a configurar un proceso de ETL mediante la herramienta Visual Studio Integration Services, donde se hace un llamado a la ejecución de la base de conocimiento configurada en DQS mediante la cual se ejecutan los dominios anteriormente configurados y cuyo resultado se lleva a una tabla temporal donde finalmente reposa la información procesada. Lo anteriormente descrito se observa en la Figura 16.

Fig. 16 Proceso de ETL para llevar ejecutar la operación de limpieza DQS y llevar el resultado a una estructura temporal

Posteriormente se realiza la consolidación de los resultados en estructuras finales en un Datamart de calidad, donde se consolida en estructuras de dimensiones y hechos los resultados históricos de la información procesada por este modelamiento, la cual posteriormente deberá ser utilizada para su visualización mediante herramientas de usuario final.

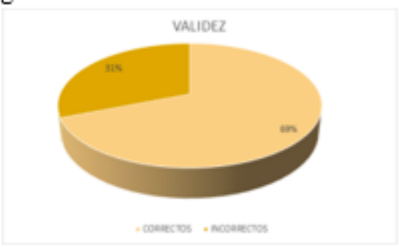
5 RESULTADOS

5.1 HERRAMIENTA TECNOLÓGICA DATA CLEANER APLICADA EN UN CASO DE NEGOCIO DISTRIBUCIÓN

Después de realizar la consolidación de datos en excel, se procedió a integrar los resultados de la ejecución de la calidad de datos. Inicialmente este procesamiento se realizó como una fase exploratoria, se usó la herramienta Excel para consolidar datos mediante tablas y gráficos dinámicos. En las Figuras 17 y 18 se puede visualizar el resultado de los cálculos generados por la plataforma Data Cleaner que luego fueron procesadas en Excel para las dimensiones Completitud y Validez aplicado a los datos de distribución.

Fig. 17 visualización dimensión Completitud

Fig. 18 visualización dimensión Validez



Como resultado de este ejercicio se puede concluir que siendo la herramienta Data Cleaner facilitadora para la generación de resultados en las dimensiones completitud y validez, el ejercicio asociado a la consolidación y cálculos de valores para identificar resultados en las dimensiones de validez y completitud de los objetos es muy manual y operativa pues todo finalmente se procesa en Excel.

5.2 HERRAMIENTA TECNOLÓGICA SPSS MODELER APLICADA EN UN CASO DE NEGOCIO COMERCIAL

Utilizando la herramienta Spss Modeler se tiene la ventaja de poder consolidar de manera automática todos los resultados asociados al proceso de calidad de datos y generarlos en una estructura final de datos que finalmente es la que se utiliza para visualización de resultados. Adicionalmente se cuenta con el beneficio de poder programar una ejecución periódica del modelo implementado de tal forma que sus resultados se almacenen de manera histórica y se pueda tener una trazabilidad en el tiempo de dicha información. Para la visualización de la calidad de los datos de tal forma que el usuario final pueda hacer uno monitoreo y control de los mismos se utiliza la plataforma Power BI donde la fuente de información es la consolidación de data en el tiempo realizado con la ejecución del modelo implementado. Un ejemplo de este resultado se registra en las Figuras 19 y 20 donde se muestra el reporte inicial consolidado de datos, y el reporte de la dimensión Validez de algunas variables.

Fig. 19 Visualización de datos comercial modelado con la herramienta SPSS Modeler

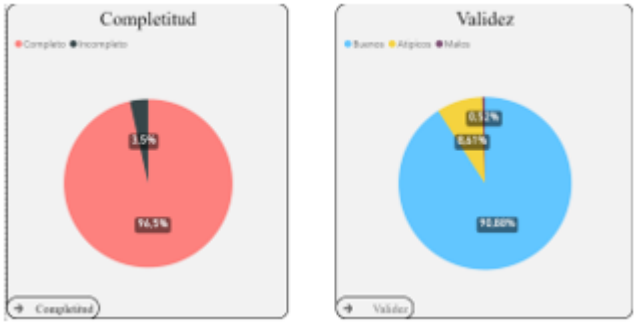
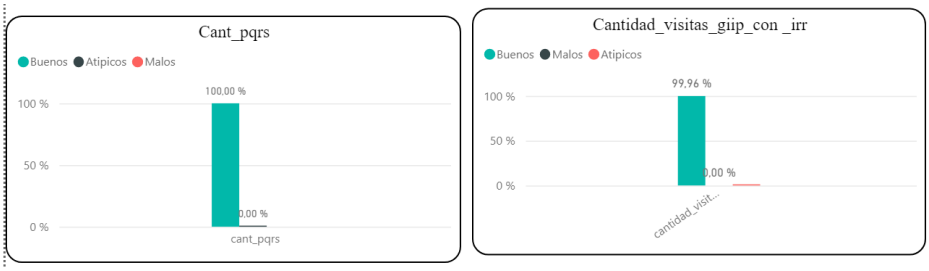


Fig. 20 Visualización datos comercial dimensión Validez modelado con la herramienta SPSS Modeler



Particularmente, en este ejercicio se puede concluir que la herramienta permite de manera flexible configurar las dimensiones que se requieran para el modelamiento de calidad de datos desde que se identifica la fuente, se procesa la información, se transforma y se consolida en estructuras históricas para su revisión y monitoreo de calidad. Por otra parte los modelamientos que se realicen con la plataforma se pueden ejecutar cada periodo de tiempo que el usuario requiera con lo cual se puede tener mayor flexibilidad en la gestión de la calidad de los datos, y por ultimo el ejercicio de visualización de los datos en un dashboard como Power BI facilita la interacción del usuario final con los resultados de calidad definidos por ellos. El uso de la herramienta y de la forma de estructurar el proyecto de calidad en esta plataforma es muy completo, solo que el único problema es que es una herramienta algo costosa a nivel empresarial.

5.3 HERRAMIENTA TECNOLÓGICA DQS DATA QUALITY SERVER APLICADA EN UN CASO DE NEGOCIO COMERCIAL

Para la implementación de este ejercicio, se mantuvo la filosofía utilizada en el ejercicio anterior. Las variables estaban ya definidas y los criterios o reglas a implementar también. El punto adicional en este ejercicio fue gestionar las reglas en el historial de ejecuciones del proceso de calidad de datos desde DQS a los objetos de comercial. Para la visualización y segmentación de la información procesada en calidad, se utilizó la herramienta Power BI para los resultados los cuales se pueden observar en la Figura 21 donde se aprecia la forma de filtrar y visualizar el resultado de la calidad de datos en el tiempo por atributo, fecha u objeto, y en la Figura 21 se visualiza un ejemplo de medición de variables de comercial

Fig. 19 Consolidado datos *comercial* - origen DQS

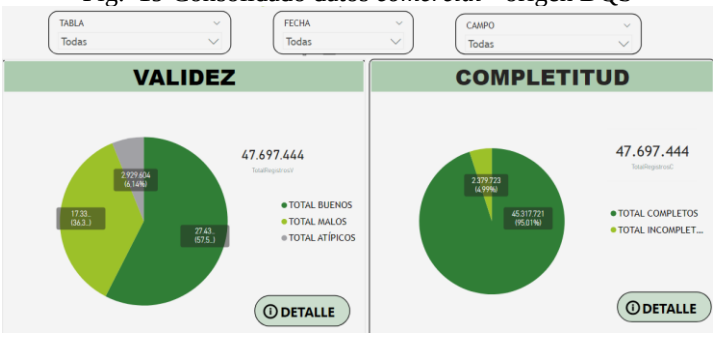


Fig. 20 Visualización dimensión Validez - origen DQS

Como conclusión en este ejercicio, se encuentra que la herramienta DQS es una herramienta poderosa para la gestión de datos como lo es Spss Modeler, En este ejercicio se encuentra que la herramienta también debe ser licenciada para la gestión de la información y un punto que se considera importante a tener en cuenta es que no se pueden trabajar dimensiones que incorporen análisis de integridad entre diferentes atributos y objetos lo cual es una debilidad para ejercicios de calidad ya que en algunas ocasiones se debe realizar este tipo de estudios. Por otra parte, los modelamientos que se realicen con la plataforma se pueden ejecutar cada periodo de tiempo que el usuario considere necesario.

6 CONCLUSIONES

Este trabajo recopila un ejercicio de aplicación de diferentes herramientas para identificar la calidad de datos. Es importante resaltar que estos ejercicios se han realizado en diferentes momentos del tiempo desde el 2018 a la fecha y han requerido un esfuerzo en aprendizaje de las herramientas y plataformas tecnológicas aquí mencionadas. En las diferentes etapas se ha ido mejorando en la visualización de la calidad de datos de un objeto en un periodo de tiempo, pero es importante destacar que dicha mejora se debe a que las herramientas actuales han facilitado este proceso.

Como aprendizaje se reconoce que las plataformas libres si bien facilitan las actividades, no siempre vienen con los componentes requeridos para trabajar de manera completa en un ejercicio y normalmente toca ajustar con nuevas tareas los resultados entregados por dichas plataformas. Este ejercicio exige algunas configuraciones y un poco de tiempo adicional para la entrega de resultados.

Por otro lado, herramientas como SPSS Modeler son una ventaja para cualquier empresa, ya que aparte de facilitar la generación de modelos analíticos permiten cerrar el ciclo completo de consolidación de calidad de datos. Pero este tipo de herramientas tienen un alto costo en licenciamiento, por lo cual no es viable su implementación en cualquier empresa.

Finalmente, con la herramienta DQS se tiene una ventaja importante frente a las demás: la configuración de variables es parametrizada de forma aislada en una base de conocimiento, la cual es procesada posteriormente con un ejercicio de ETL que se adecua perfectamente a la necesidad de calidad del usuario y permite ser aplicada en muchos dominios, si se requiere. Esta independencia en la parametrización garantiza que el ajuste de variables para su ejecución facilita a los procesos de Tecnología ser más oportunos en dicha tarea.

Con este trabajo se puede continuar explorando en otras dimensiones de calidad ya que la herramienta DQS facilita la implementación de nuevos dominios de acuerdo con unas definiciones y estándares identificados por el usuario final.

RECONOCIMIENTOS

A la Universidad Nacional de Colombia Sede Manizales, Institución que ha apoyado este proceso de investigación y a la Central Hidroeléctrica de Caldas CHEC S,A. E,S,P empresa que ha facilitado los datos, el conocimiento y los recursos para la realización de este trabajo.

REFERENCIAS

[1] PowerData, "Calidad de Datos. Cómo impulsar tu negocio con los datos." https://www.powerdata.es/calidad-de-datos (accessed May 19, 2019).

[2] L. Cai and Y. Zhu,8 "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, no. November, 2015, doi: 10.5334/dsj-2015-002.

[3] S. Sadiq and M. Indulska, "Open data: Quality over quantity," *Int. J. Inf. Manage.*, vol. 37, no. 3, pp. 150–154, 2017, doi: 10.1016/j.ijinfomgt.2017.01.003.

[4] V. Estrada-Galinanes and K. Wac, "Visions and Challenges in Managing and Preserving Data to Measure Quality of Life," *2018 IEEE 3rd Int. Work. Found. Appl. Self* Syst.*, pp. 92–99, 2019, doi: 10.1109/fas-w.2018.00031.

[5] I. Mergel, A. Kleibrink, and J. Sörvik, "Open data outcomes: U.S. cities between product and process innovation," *Gov. Inf. Q.*, vol. 35, no. 4, pp. 622–632, 2018, doi: 10.1016/j.giq.2018.09.004.

[6] H. H. Ahmed, "Data quality assessment in the integration process of linked open data (LOD)," in *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, 2018, doi: 10.1109/AICCSA.2017.178.

[7] A. Colborne and M. Smit, "Identifying and mitigating risks to the quality of open data in the post-truth era," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 2588–2594, 2018, doi: 10.1109/BigData.2017.8258218.

[8] P. Zhang, F. Xiong, J. Gao, and J. Wang, *Data Quality in Big Data Processing: Issues, Solutions and Open Problems*. .

[9] H. Wahl, "LEIWI – A Tool to Compute the Quality of Life Using Open Data," no. Iscit, pp. 116–120, 2018.

[10] W. Xia, Z. Xu, and C. Mao, "User-driven filtering and ranking of topical datasets based on overall data quality," *Proc. - 2017 14th Web Inf. Syst. Appl. Conf. WISA 2017*, vol. 2018-Janua, no. 1, pp. 257–262, 2018, doi: 10.1109/WISA.2017.24.

[11] A. Nikiforova, "Open Data Quality Evaluation: A Comparative Analysis of Open Data in Latvia," *Balt. J. Mod. Comput.*, vol. 6, no. 4, pp. 363–386, 2018, doi: 10.22364/bjmc.2018.6.4.04.

[12] R. Machova and M. Lnenicka, "Evaluating the Quality of Open Data Portals on the National Level," *J. Theor. Appl. Electron. Commer. Res.*, vol. 12, no. 1, pp. 21–41, 2017, doi: 10.4067/S0718-18762017000100003.

[13] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: Definition and application to Open Government Data," *Gov. Inf. Q.*, vol. 33, no. 2, pp. 325–337, 2016, [Online]. Available: http://dx.doi.org/10.1016/j.giq.2016.02.001.

[14] A. Whitmore, "Using open government data to predict war: A case study of data and systems challenges," *Gov. Inf. Q.*, vol. 31, no. 4, pp. 622–630, 2014, doi: 10.1016/j.giq.2014.04.003.

[15] A. Abella and M. O. C. De-pablos-heredero, "INDICADORES DE CALIDAD DE DATOS ABIERTOS : EL CASO DEL PORTAL DE DATOS ABIERTOS DE BARCELONA Open data quality metrics : Barcelona open data portal case."

[16] C. Batini and M. Scannapieca, *Data-Centric Systems and Applications: Data Quality Concepts,*

Methodologies and Techniques. 2006.

[17] A. Abella, M. Ortiz-De-urbina-criado, and C. De-Pablos-heredero, "Meloda 5: A metric to assess open data reusability," *Prof. la Inf.*, vol. 28, no. 6, pp. 8–10, 2019, doi: 10.3145/epi.2019.nov.20.

[18] G. R. Rivadera, "La metodología de Kimball para el diseño de almacenes de datos (Data warehouses)," pp. 56–71.