

10º Congreso de Exploración y Desarrollo de Hidrocarburos
VI Jornadas de Geotecnología: Crear y Administrar Ámbitos Colaborativos
Subtema: Integración de datos y aplicaciones

GESTIÓN DE LA CALIDAD DE LOS DATOS PARA UNA MEJORA CONTINUA UN CASO PRÁCTICO EN E&P

Ana Docampo¹, Rodolfo S. Figueroa¹, Fernando Spasoff¹, Matías Miranda¹, Lucas Mattar², Cesar Villegas²

1: YPF S.A., ana.docampo@ypf.com, rodolfo.figueroa@ypf.com,

fernando.spasoff@ypf.com, matias.miranda@ypf.com

2: Halliburton, lucas.mattar@halliburton.com, cesarluis.villegasmolina@halliburton.com

Palabras claves: gestión información, calidad datos, integración datos

ABSTRACT

Gestión de la calidad de los datos para una mejora continua. Un caso práctico en E&P.

El mundo desarrollado ha pasado de una economía industrial a una economía de la información.

Las empresas compiten sobre la capacidad de absorber y responder a la información, tomando decisiones a partir de ésta, no sólo de fabricar y distribuir productos. Por lo tanto, los datos son un activo valioso y necesario y la calidad de los mismos es crítica para el éxito.

En particular, la toma de decisiones técnicas y estratégicas en la industria del Petróleo & Gas se basan en el análisis de grandes volúmenes de datos de variadas características: tipos, orígenes, granularidad, temporalidad, entre otras. Estos datos generalmente están dispersos en diversas plataformas, lo cual puede ser redundante, complejo y costoso.

En Data Management de YPF se inició un proyecto de integración de datos entre plataformas de interpretación lo que generó la necesidad de avanzar en la calidad de datos a nivel general con el fin de asegurar que los proyectos integrales resultantes sean consistentes desde la fuente misma de los datos, y de esta manera cumplir con una de las principales premisas de nuestra actividad que es la de facilitar al usuario información consolidada de la mejor calidad posible.

Para lograr un nuevo nivel de calidad de los datos, se propuso medir, mejorar y monitorear la calidad de los datos en el tiempo. Esto implicó delimitar un alcance dentro del amplio concepto de Calidad, descomponer la calidad en características observables, definir -entre la gran variedad existente - los atributos de estas características de manera de generar métricas estandarizadas y sencillas.

Estos conceptos se implementaron técnicamente logrando formalizar las actividades de calidad de datos, obteniendo las métricas, clasificando e identificando en detalle los problemas de calidad, automatizando tareas repetitivas y generando la oportunidad de extender el concepto a otros dominios de datos de una forma sencilla.

INTRODUCCIÓN

YPF S.A. es una empresa operadora integrada de energía con larga historia en la industria y con altos niveles -en particular en E&P- de actividad. Basta mencionar que su base de datos maestra de pozos supera los 48.000 identificadores y que en los últimos 5 años tuvo un promedio superior a 700 pozos/año terminados en una vasta dispersión geográfica. Estos valores son un índice para

inferir el enorme volumen de información técnica que se genera. Para los procesos de análisis, estudio, y toma de decisiones, la calidad e integración de los datos es uno de los principales retos en el área de data management.

Dentro del ambiente técnico de E&P, y en particular para el área de G&G, existen diferentes tipos de repositorio de información:

- Bases de datos (BD) corporativas que funcionan como repositorio oficial del dato fuente y permiten el resguardo de la información en su estado original a través del tiempo, algunas desarrolladas por YPF y otras de origen comercial cuyos modelos se adaptan a las necesidades de la compañía.
- BD relacionadas a aplicaciones de interpretación. Estas por lo general con diferentes tecnologías, modelo de datos y flujos de trabajo. La compañía cuenta con cuatro distintas plataformas completas de interpretación y con el licenciamiento de gran parte de las aplicaciones que existen en el mercado.

Esta diversidad de aplicaciones permite aprovechar las bondades de cada software y explotar su potencial al máximo dependiendo de las necesidades del proyecto de negocio en el que se trabaje, sin embargo, también acarrea múltiples desafíos relacionados con:

- Mantener actualizadas y sincronizadas las BD de interpretación con los repositorios generales de información técnica. Destacando en este punto que estas últimas deben contener la totalidad de los datos registrados en campo, mientras que las primeras solo el dato que el intérprete necesita para realizar su trabajo, por lo cual la actualización entre estos ambientes no es lineal.
- Conservar la información sincronizada entre los diferentes ambientes/BD de interpretación que se utilicen. Dentro de un mismo proyecto un grupo multidisciplinario de intérpretes pueden utilizar diferentes plataformas de interpretación, dependiendo de la etapa del flujo que se esté trabajando.
- Usuarios que no comparten la misma información de referencia ni las interpretaciones de otros sectores, dificultando esto el trabajo colaborativo y pudiendo llevar a tomar decisiones sin disponer de todos los datos disponibles en determinado momento.

En particular el presente trabajo se divide en dos partes describiendo: (a) un proyecto crítico para el negocio que exigía una compleja y rápida integración de información entre ambientes de interpretación que, debido a los fuertes requerimientos respecto de la calidad y la necesidad gestionarla en forma continua, derivó en otra iniciativa de (b) formalización de la gestión de calidad de datos.

a) Respetto de la Integración de la información de G&G**METODOLOGÍA**

Los requerimientos de integración de información y actualización en las BD corporativas y proyectos de interpretación eran de tal complejidad que surgió la necesidad del utilizar alguna herramienta que permitiera controles automáticos para facilitar la sincronización.

Una vez definida esta necesidad de integración de datos se enmarco el alcance del proyecto, estableciendo principalmente que datos era critico tener sincronizados y en que bases residían.

Las principales tareas que se desarrollaron para esta implementación fueron:

- Diagrama de datos y aplicaciones tomando los datos que se consideraron necesarios para la sincronización.
- Definición de la fuente maestra de cada dato y los lugares donde debían replicarse.
- Criterios de búsqueda y/o macheo de la información. En los casos donde el identificador de pozo (uwi) estaba disponible esto facilitaba el trabajo, pero en los casos donde no se podía usar se utilizó el criterio de ubicación del pozo.
- Estudio de cada modelo de datos a conectar, para identificar campos necesarios de macheo y campos deseados de comparación y/o actualización.
- Unificación de sistemas de coordenadas dentro del integrador para que cualquier comparación que contenga información de superficie, se hiciese utilizando los mismos parámetros.
- Generación de virtual data bases, mediante las cuales se leen los datos en las diferentes fuentes.
- Definición de queries dentro del integrador, seleccionando el origen, el destino y el criterio de búsqueda o comparación.

RESULTADOS

- A partir de la ejecución de las primeras comparaciones, eran necesarios ajustes para calibrar los resultados, y poder obtener conclusiones certeras, por ejemplo en los casos de comparación por localización geográfica, para el header del pozo, se usó un criterio de +-50 metros, para identificar pozos iguales con diferente uwi.
- Los resultados de las comparaciones siempre necesitan un análisis de los casos que no matchearon para su ajuste, antes de ejecutar las acciones de actualización, sincronización, limpieza, etc.
- En aquellos casos donde las comparaciones resultaban en una relación de 1 a muchos, puede ser conveniente, reformular el criterio de búsqueda.
- Algunos resultados de comparaciones podían visualizarse en un mapa y facilitar así el trabajo de análisis de resultados.

- Para los queries referidos a logs se definió una lista muy reducida de mnemónicos más importantes o básicos para cada pozo. El tiempo de procesamiento de estos queries era considerablemente más largo que con otro tipo de datos.

CONCLUSIONES

- Al integrar las aplicaciones de interpretación se obtuvo más que una sincronización, se visualizaron errores en datos que de otra manera serían muy difíciles de identificar, dando inicio al proyecto de control de calidad.
- A partir de los resultados de las comparaciones en la fase de integración de datos, se podían ejecutar acciones correctivas de manera automática, o manual. Para este caso particular, solo cuando el dato ameritaba una transferencia total se hacía en automático, cuando era una modificación parcial se ejecuta manualmente.

b) Respetto de la Integración de la información de G&G

METODOLOGÍA

El primer paso para avanzar fue focalizarnos determinando el siguiente objetivo: “definición e implementación de un marco metodológico que facilite la obtención de métricas cuantitativas de calidad”. Estas permiten ser comparadas con valores de referencia, resguardadas durante los períodos de observación y posibilitan con el transcurso del tiempo la obtención de estadísticas factibles a utilizar en procesos de mejora continua.

Basándonos en un estándar como el ciclo del TDQM (Total Data Quality Management, Figura 1), podemos ver que nos centramos en las dos primeras etapas:

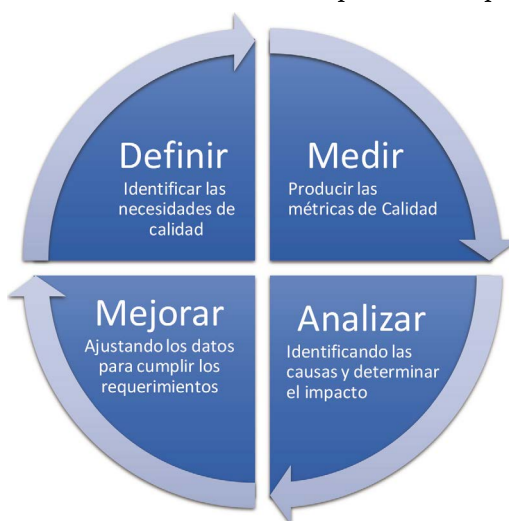


Figura 1. Ciclo TDQM - Total Data Quality Management, MIT, R. Wang

- Definir los requerimientos de la calidad de datos (DQ) en el entorno del negocio
- Medir el estado de dichos requerimientos

Fundamento en la disciplina

Si bien todos tenemos en mente alguna definición aproximada de DQ, en general está dada por nuestra experiencia personal en relación con la gestión de la información y/o tecnología.

Revisando la amplia bibliografía al respecto, encontraremos que la calidad de datos no es solo “datos sin errores”. La mayoría de los expertos toman aspectos más amplios para definirla, por ejemplo:

- ..” Es el conjunto de características que hacen que la información tenga más valor para los usuarios. Es el grado con el que los productos de datos satisfacen las necesidades y requisitos de los clientes”
- .. “Cumplir de forma consistente con las expectativas de los trabajadores del conocimiento y los clientes finales”
- .. “La calidad de los datos es la aptitud o idoneidad de los datos para cumplir con los requerimientos del negocio”

De manera de entender en forma más práctica y clara cuáles de todas esas facetas se relacionan el objetivo de nuestra iniciativa, tomamos una clasificación existente para la Calidad de los Datos que la divide en dos perspectivas:

- **Subjetivas.** Las que reflejan las percepciones, necesidades y experiencias de los stakeholders, los involucrados en el ciclo de vida del dato. Muchas veces sucede que los gestores del dato perciben alta calidad y los consumidores no, debido a la dificultad de uso para los procesos del negocio. Estas situaciones pueden ser dadas por recursos ajenos a la calidad del dato en sí: entorno de trabajo, hardware, software, etc.
- **Objetivas:** Las que se pueden reflejar en métricas obtenidas desde los datos. Se sub-clasifican en:
 - Independientes de la tarea: métricas que reflejan el estado de la data sin el conocimiento del contexto o la aplicación. Puede ser aplicada a cualquier set de datos
 - Dependientes de la tarea: incluyen las reglas del negocio, regulaciones de entes de control, etc.

Claramente nuestro objetivo se encuadra en la perspectiva Objetiva. Este orden conceptual identifica unívocamente qué parte del DQ estamos cubriendo.

De las perspectivas a las métricas

La técnica utilizada para llegar a las métricas es la descomposición la calidad de los datos en características observables. Estas, en su mayor nivel de detalle deben hacer referencia a campos, registros, o conjuntos de datos de las BD factibles a ser evaluadas y medidas con ciertos criterios. Estas características (figura 2) se denominan dimensiones.

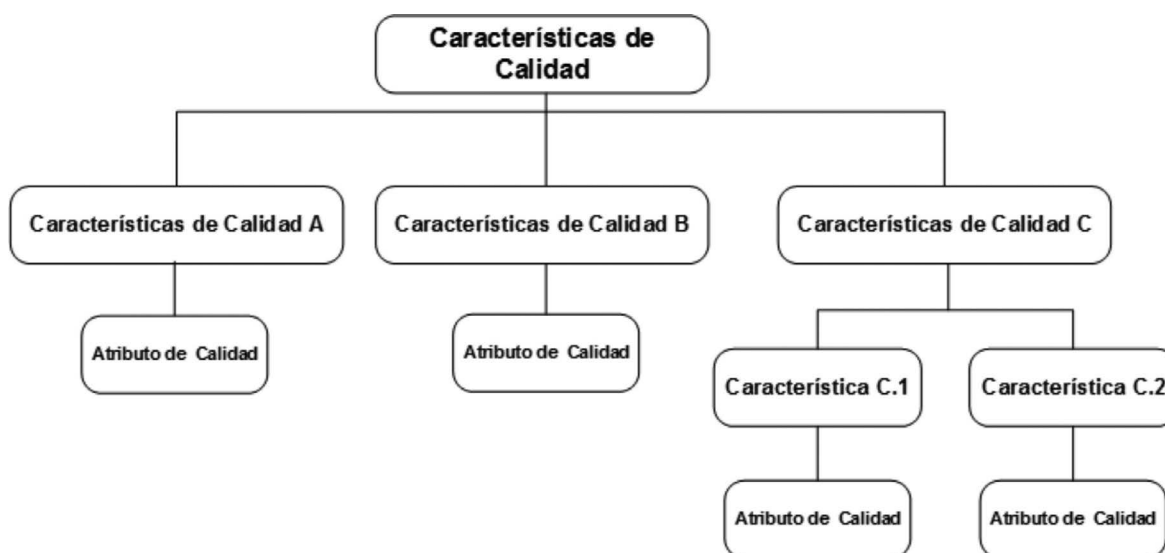


Figura 2. Descomposición de las características de calidad para obtener las Dimensiones.

Existen variadas clasificaciones de dimensiones para el DQ citadas en los textos de los expertos (Ballou and Pazer 1985, Wang et al. 1995, Wand and Wang 1996, Wang and Strong 1996, Haug et al. 2009, etc.). Basadas en estas, se han desarrollado múltiples variantes. Analizando varias de las listas propuestas en la literatura, se puede denotar que, si bien algunos de sus nombres coinciden, muchas veces el significado o interpretación asignado puede variar considerablemente.

A modo de ejemplo, podemos ver algunas definiciones de dimensiones de distintos autores (figuras 3 y 4).

Dimensions	Definition
Accessibility	The extent to which data is available, or easily or quickly retrievable
Appropriate Amount	The extent to which the volume of data is appropriate for the task at hand
Believability	The extent to which data is regarded as true and credible
Completeness	The extent to which data is not missing and is of sufficient breadth and depth of the task at hand
Concise Representation	The extent to which data is compactly represented
Consistent Representation	The extent to which data is presented in the same format
Ease of Manipulation	The extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	The extent to which data is correct and reliable
Interpretability	The extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial
Relevancy	The extent to which data is applicable and helpful for the task at hand
Reputation	The extent to which data is highly regarded in terms of its source or content
Security	The extent to which access to data is restricted appropriately to maintain its security
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand
Understandability	The extent to which data is easily comprehended
Value added	The extent to which data is beneficial and provides advantages from its use

Figura 3. Wang and Strong List of Data Quality Dimensions [Wang 1996]. IJDMS, Vol.4, No.2, April 2012

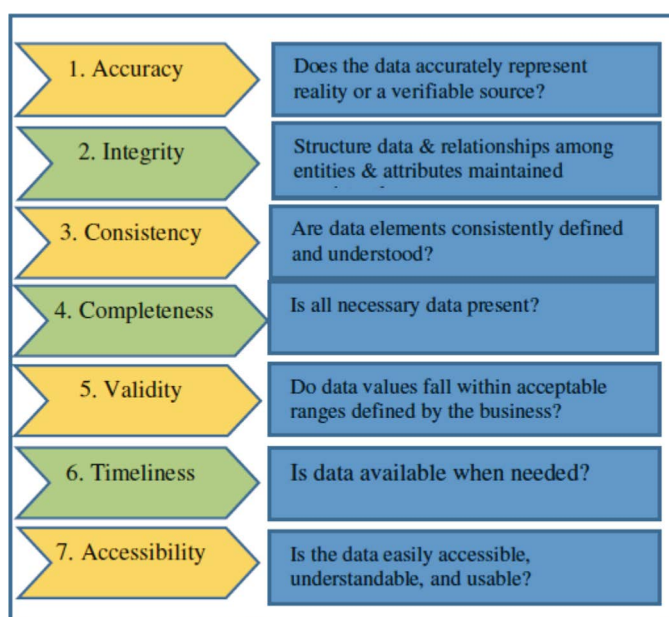


Figura 4. Noraini Abdullah et al. DQ in Big Data: A Review. Int. J. Advance Soft Compu. Appl, Vol. 7 2015

No existe una definición formal de conjuntos de dimensiones asociadas a una industria en particular. En general, estas derivan de los requerimientos de calidad que deben cumplir los datos que son utilizados para soportar con éxito los procedimientos del negocio.

En nuestro caso, la aproximación para definir el primer conjunto de dimensiones fue analizar los requerimientos ya conocidos de calidad de datos un ámbito de E&P y evaluar si las características de los mismos podían relacionarse con las dimensiones habitualmente más utilizadas en el DQ, las que se pueden encontrar de algunas publicaciones que resumen las más citadas en papers, casos y estudios publicados (ej. Figura 5).

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Figura 5. Wang et al. "Cited data quality dimensions".

Nuestro análisis derivó en la definición de un primer conjunto de seis dimensiones: Completitud, Unicidad, Consistencia, Sincronización, Validez y Correctitud, que cubren los requerimientos iniciales de calidad.

Cada dimensión se explica mediante una descripción que incluye mínimamente:

- Definición
- Descripción de la métrica a utilizar
- Unidad
- Alcance
- Pseudo código descriptivo de la regla que se aplicará sobre los datos para la evaluación.

Este es el punto donde se encuentran las dimensiones y consultas (reglas) aplicadas a los repositorios de datos con el criterio de las dimensiones, quienes le dan un orden para poder ser gestionadas dentro de un marco conceptual.

Dado que una misma dimensión aplicada a distintos repositorios o ambientes de datos puede tener distintas reglas (consultas a las BD, repositorios), se generaron múltiples sub-dimensiones cada una heredando las propiedades originales, salvo el Pseudo código que determina cómo se aplica la regla. Luego, la consolidación de las métricas de estas sub-dimensiones, permite tener un valor promedio de la Dimensión en general.

RESULTADOS

Basados en la metodología de aproximación a la mejora de la Calidad mediante su descomposición en dimensiones observable y medibles, lo que se obtiene es un marco de trabajo alineado con el estándar del ciclo TDQM anteriormente mencionado, donde los componentes se relacionan de esta manera:

1. Definir necesidades. Releva requerimientos y convierte en características de DQ a considerar

REQUERIMIENTOS DQ DEL NEGOCIO \ CARACTERÍSTICAS DE CALIDAD NECESARIAS	Característica A	Característica B	Característica C	...	Característica X
Requerimiento 1		●			●
Requerimiento 2		○	◇		
Requerimiento 3		●	●		
⋮	◇			○	
Requerimiento N	●	○			

Relación: ● Fuerte ○ Directa ◇ Baja

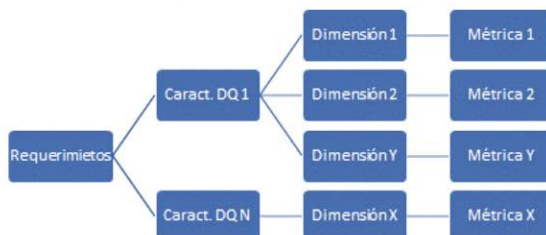
2. Medir necesidades. Las características se relacionan con las dimensiones definidas, obteniendo las métricas

Dimensión A			
	SubDim.A.1	Hallazgos A.1	Métrica A.1
	SubDim.A.2	Hallazgos A.2	Métrica A.2
	SubDim.A.n	Hallazgos A.n	Métrica A.n
Dimensión B			
	SubDim.B.1	Hallazgos B.1	Métrica B.1
	SubDim.B.2	Hallazgos B.2	Métrica B.2
	SubDim.B.n	Hallazgos B.n	Métrica B.n
Dimensión N			
	SubDim.N.1	Hallazgos N.1	Métrica N.1
	SubDim.N.2	Hallazgos N.2	Métrica N.2
	SubDim.N.n	Hallazgos N.n	Métrica N.n

4. Mejorar la calidad de los datos en base a las métricas y hallazgos, de manera que cumplan los requerimientos del negocio



3. Analizar resultados e impacto en los requerimientos



El próximo paso fue la implementación de un piloto de manera de testear la consistencia del modelo y la factibilidad de resolverlo técnicamente. Algunas de las tareas y ejecutadas fueron:

- Se relevó un conjunto de requerimientos sobre un determinado ambiente de datos para el piloto
- Se analizaron distintas dimensiones obtenidas de la literatura, contrastándolas con las necesidades
- Se definió un conjunto de seis dimensiones, especificando sus características y asegurando que cubrieran los actuales requerimientos y variantes de estos
- Por cada repositorio involucrado, se crearon las sub-dimensiones (aprox. 30) necesarias para obtener las métricas de cada regla en particular
- Se seleccionó una plataforma técnica que soportara conectividad con diferentes repositorios de datos, que tuviera herramientas que permitan ejecutar las reglas mediante consultas a los repositorios (BD planillas, etc), que permitiera automatizar tareas repetitivas, resguardar y visualizar hallazgos y resultados, hacer búsquedas en los mismos y que fuera factible de incorporar futuras funcionalidades mediante un desarrollo.
- Se implementaron las dimensiones y se corrieron las reglas

Los resultados obtenidos fueron:

- Una metodología que guía los pasos para avanzar sobre la calidad de datos
- Documentación de cada regla mediante el desarrollo de las dimensiones y métricas
- Un repositorio de reglas: Una herramienta dedicada al DQ que puede integrar ordenadamente todas las reglas factibles a definir. (Figura 6)
- La ejecución desatendida (programada) de complejas consultas de datos que permite:
 - o La obtención de los hallazgos (anomalías) que cada regla detecta
 - o La obtención de métricas de DQ
 - o La recepción de los hallazgos en forma automática (Figura 7)
- La posibilidad de contar con métricas que consoliden las mismas dimensiones en distintos repositorios (Ej: consistencia de un repositorio o de varios repositorios)
- La factibilidad de contar con un registro histórico de las métricas que permitan aplicar metodologías de análisis para la prevención de las anomalías (estadísticas, causa raíz, etc)
- La posibilidad de incrementar complejidad y cantidad de dimensiones/reglas sin requerir recursos extras para su gestión integral
- Contar con un motor de búsqueda de errores (simil Google): por patrones de texto sobre los identificadores (nombres de pozos, uwis, etc) de los registros identificados por las reglas que buscan las anomalías.
- Analizar los hallazgos encontrados, accediendo geográficamente o por distintos filtros y/o taxonomías. (Figuras 8 y 9)

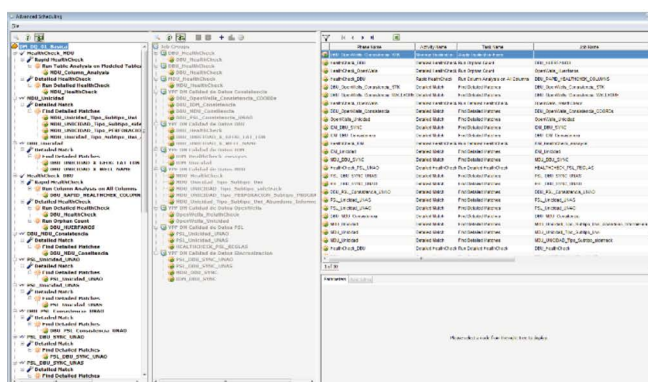


Figura 6. Reglas asociadas a Dimensiones. Jobs programados



Figura 7. Informe resultados de calidad enviados automáticamente por mail

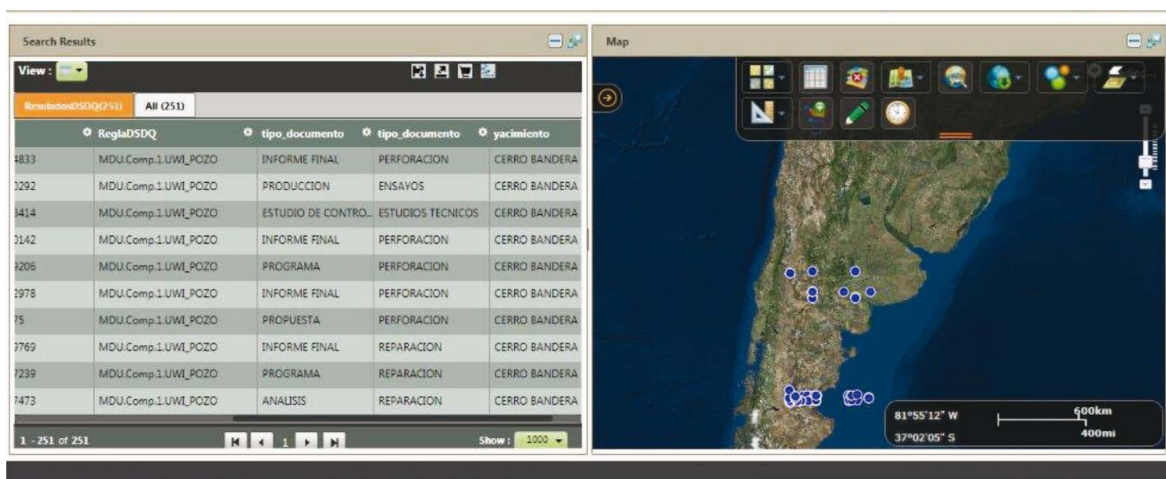


Figura 8. Búsqueda de errores y visualización geográfica

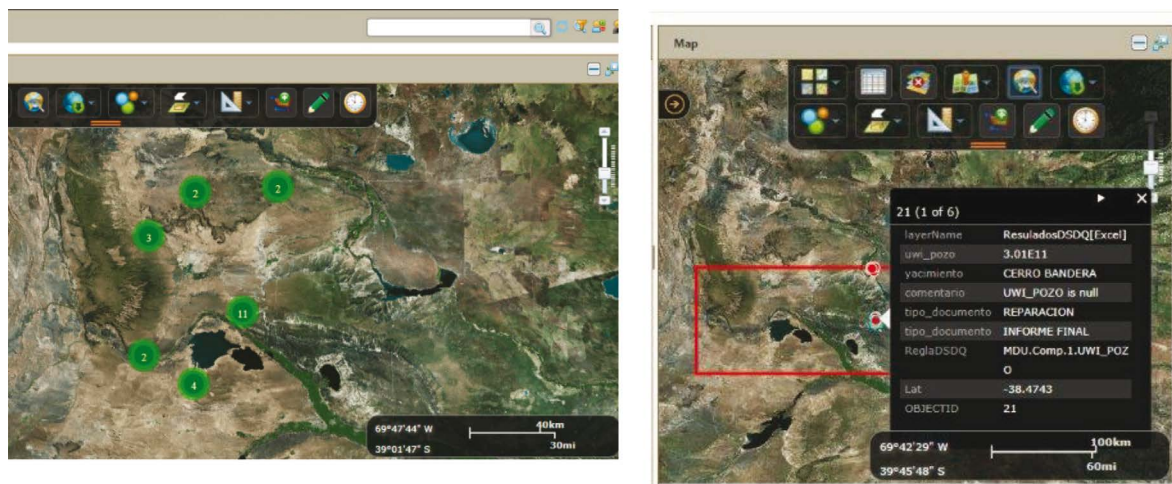


Figura 9 Búsqueda de errores y visualización geográfica

CONCLUSIONES

Del modelo:

- Enmarcar las actividades de Calidad de Datos en conceptos metodológicos permite crecer ordenadamente en complejidad logrando resultados consistentes y controlados.
- El modelo es genérico y puede ser aplicado a datos de distintas disciplinas
- La utilización de Dimensiones fuerza a realizar un análisis profundo de las reglas que se están ejecutando y el tipo de métricas que se obtendrá.
- El primer grupo de dimensiones se puede obtener de la experiencia de los gestores de la información.
- La asociación de los requerimientos del negocio con las características que el DQ debe cumplir, permite relevar necesidades puntuales.
- El modelo de Dimensiones es genérico e incremental, lo que permite que diferentes ámbitos de datos (ej: el Transaccional y el de Real Time) compartan iguales subconjuntos de dimensiones (con distintas reglas), y sus métricas puedan ser consolidadas

De la implementación:

- No es condición necesaria contar con una plataforma única para la implementación del modelo, pero el hecho de tenerla simplifica el desarrollo, evita todo tipo de problemas que generan las interfaces entre componentes, mantiene la coherencia y asegura resultados consistentes.
- Automatizar el proceso incrementa exponencialmente la ejecución de reglas y obtención de métricas, liberando recursos para el análisis, corrección y evaluación de otras necesidades de calidad.
- Implementar la solución técnica permite obtener los registros de datos a corregir, sin requerir el expertise para entender la complejidad de las reglas.
- Salvo casos muy puntuales, esta implementación puede cubrir un muy amplio espectro de la calidad de datos utilizados por procesos de distinta naturaleza.

REFERENCIAS CITADAS

Leo, L., Pipino, L. Yang, W. L., & Wang, R. Y. (2002).
Data quality assessment, Communications of the
ACM, Vol. 45, No. 4ve.

<http://iso25000.com/> Normas SQuaRE

Data Quality in Big Data: A Review, Noraini Abdul-

lah, Saiful Adli Ismail, Siti Sophiayati and Suriani Mohd Sam, Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, November (2015)

DAMA-DMBOK Guide 2009, Chapter 12