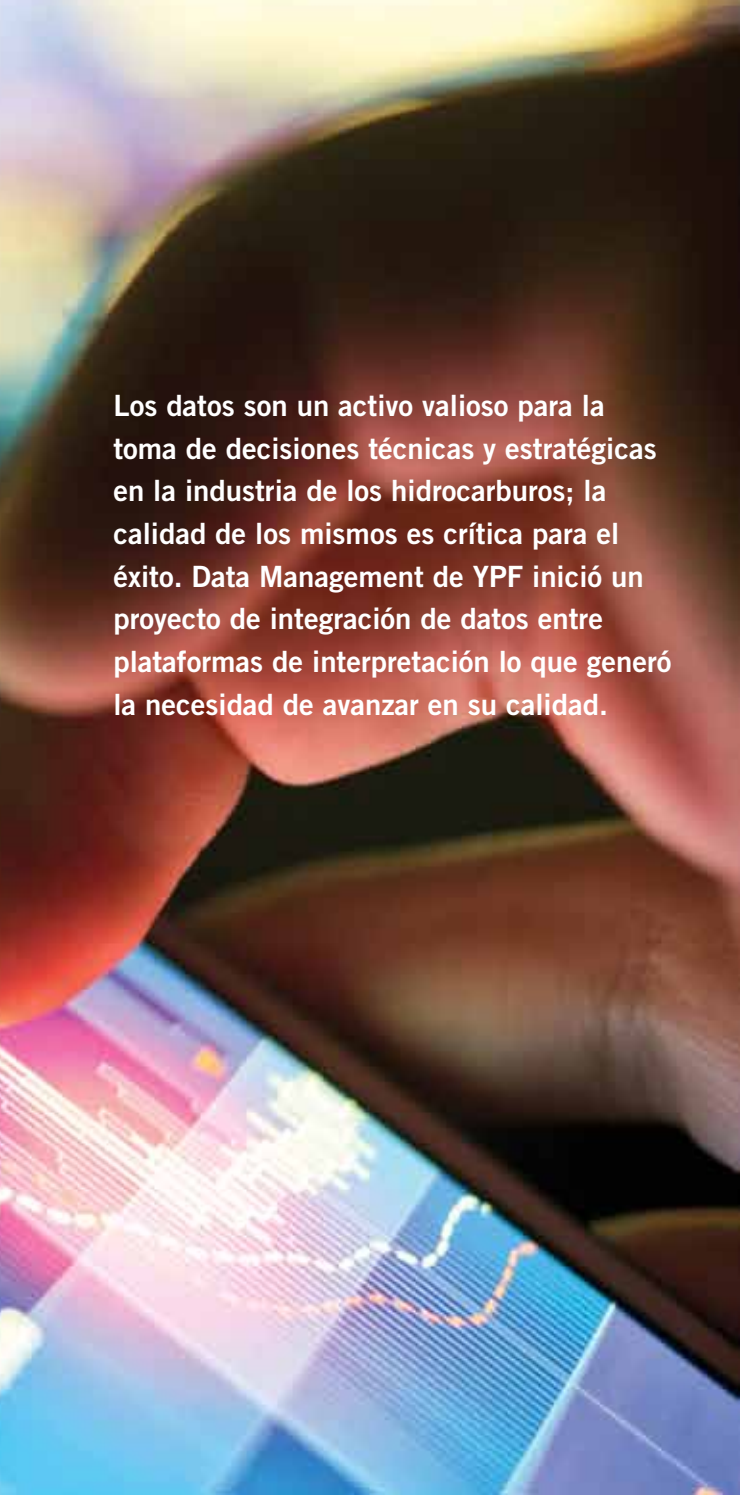
Gestión de la calidad de los datos para una mejora continua

Un caso práctico de E&P

Por ***Ana Docampo, Rodolfo Figueroa, Fernando Spasoff*** y ***Matías Miranda*** (YPF);
Lucas Mattar y César Villegas (Halliburton)

Este trabajo fue presentado en las *VI Jornadas de Geotecnología*, en el marco del *10° Congreso de Exploración y Desarrollo de Hidrocarburos del IAPG* (Noviembre de 2018, Mendoza).

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 esta, no solo de fabricar y distribuir productos. Por lo tanto, los datos son un activo valioso y necesario, y su calidad es crítica para el éxito.



Los datos son un activo valioso para la toma de decisiones técnicas y estratégicas en la industria de los hidrocarburos; la calidad de los mismos es crítica para el éxito. Data Management de YPF inició un proyecto de integración de datos entre plataformas de interpretación lo que generó la necesidad de avanzar en su calidad.

En particular, la toma de decisiones técnicas y estratégicas en la industria del Petróleo & Gas se basa en el análisis de grandes volúmenes de datos de variadas características: tipos, orígenes, granularidad y temporalidad, entre otras. Estos datos generalmente están dispersos en varias 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, esto 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 de los datos y poder cumplir con una de las principales premisas de nuestra actividad: 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, se logró formalizar las actividades de calidad de datos, se obtuvieron las métricas, se clasificó e identificó en detalle los problemas de calidad, se automatizaron tareas repetitivas y se generó 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 cinco 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. En general con diferentes tecnologías, modelo de datos y flujos de trabajo. La compañía cuenta con cuatro 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 según las necesidades del proyecto de negocio en el que se trabaje; sin embargo, también acarrea múltiples desafíos relacionados con los siguientes aspectos:

- Mantener actualizadas y sincronizadas las BD de interpretación con los repositorios generales de información técnica. Es de destacar 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 multidisci-

plinario de intérpretes pueden utilizar diferentes plataformas de interpretación, según la etapa del flujo en que esté trabajando.
- Usuarios que no comparten la misma información de referencia ni las interpretaciones de otros sectores, lo cual dificulta el trabajo colaborativo y la toma de decisiones, al no acceder a todos los datos disponibles en determinado momento.

Este trabajo se divide en dos partes: (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 de gestionarla en forma continua, derivó en otra iniciativa de (b) formalización de la gestión de calidad de datos.

Respecto 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, se estableció qué datos era critico tener sincronizados y en qué 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 que se debía 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 realizara con los mismos parámetros.
- Generación de virtual data bases mediante las que 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 fueron 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 m, 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.

Conslusiones

- 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, en cambio, cuando era una modificación parcial, se ejecuta manualmente.

Respecto de la Integración de la información de G&G

Metodología

El primer paso para avanzar fue focalizarnos en 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:
- Definir los requerimientos de la calidad de datos (DQ) en el entorno del negocio.
- Medir el estado de esos 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.

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

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".

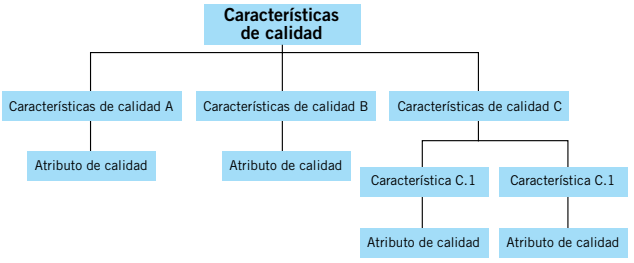


Figura 2. **Descomposición de las características de calidad para obtener las dimensiones.**

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 stackeholders, 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 subclasifican 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 cubrimos.

Dimensiones	Definición
Accesibilidad	La medida en que los datos están disponibles o son fácil o rápidamente recuperables
Cantidad apropiada	La medida en que el volumen de datos es apropiado para la tarea en cuestión
Credibilidad	La medida en que los datos son considerados verdaderos y creíbles
Completitud	La medida en que no faltan datos y tiene la amplitud y profundidad de la tarea en cuestión
Representación concisa	La medida en que los datos están representados de manera compacta
Representación concisa	La medida en que los datos son presentados en el mismo formato
Facilidad de manipulación	La medida en que los datos son fáciles de manipular y aplicar a diferentes tareas
Libre de error	La medida en que los datos son correctos y confiables
Interpretabilidad	La medida en que los datos están en lenguajes apropiados, símbolos y unidades, y las definiciones son claras
Objetividad	La medida en que los datos son imparciales y sin prejuicios
Relevancia	La medida en que los datos son aplicables y útiles para la tarea en cuestión
Reputación	La medida en que los datos son muy estimados en términos de su fuente o contenido
Seguridad	La medida en que el acceso a los datos está restringido apropiadamente para mantener su seguridad
Oportunidad	La medida en que los datos están lo suficientemente actualizados para la tarea en cuestión
Comprensibilidad	La medida en que los datos son fácilmente comprendidos
Valor agregado	La medida en que los datos son beneficiosos y otorgan ventajas al que los usa

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

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 de ser evaluadas y medidas con ciertos criterios. Estas características (Figura 2) se denominan dimensiones.

Existen varias clasificaciones de dimensiones para el DQ citadas en los textos de los expertos (Ballou y Pazer, 1985; Wang *et al.,* 1995; Wand y Wang, 1996; Wang y Strong, 1996; Haug *et al.,* 2009). Sobre su base se han desarrollado múltiples variantes. Se analizan varias de las listas propuestas en la literatura, se puede denotar que, si bien algunos de sus nombres coinciden, muchas veces el significado o la interpretación asignada puede variar considerablemente.

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

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 de un ámbito de E&P y evaluar si sus características podían relacionarse con las dimensiones más utilizadas en el DQ, las que se pueden encontrar en algunas publicaciones que resumen las más citadas en *papers*, casos y estudios publicados (Figura 5).

Este 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 que se utilizará,

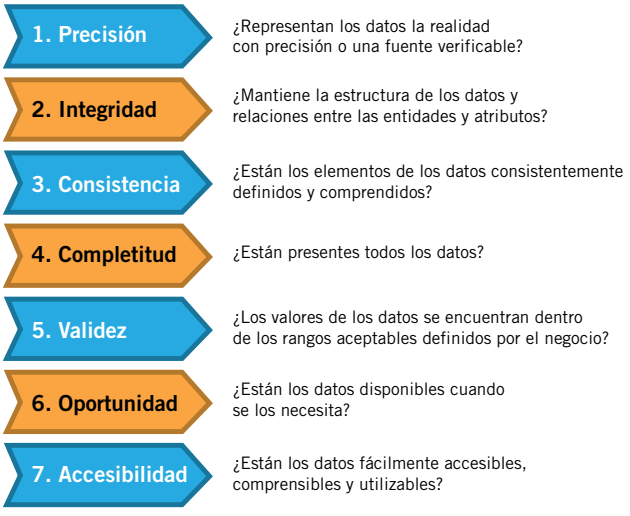


Figura 4. **Noraini Abdullah *et al.*, DQ in Big Data: A Review. Int. J. Advance Soft Compu. Appl, Vol. 7, 2015.**

Dimensión	# citado	Dimensión	# citado	Dimensión	# citado
Precisión	25	Formato	4	Comparablidad	2
Confibilidad	22	Interpretabilidad	4	Concisión	2
Oportunidad	19	Contenido	3	Imparcialidad	2
Relevancia	16	Eficiencia	3	Informatividad	2
Completitud	15	Importancia	3	Nivel de detalle	2
Circulación	9	Suficiencia	3	Cuantitatividad	2
Consistencia	8	Usabilidad	3	Alcance	2
Flexibilidad	5	Utilidad	3	Comprensibilidad	2
Precisión	5	Claridad	2		

Figura 5. **Wang y otros. "Dimensiones de calidad de los datos citados".**

- 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, que 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 subdimensiones, cada una hereda las propiedades originales, salvo el pseudocódigo que determina cómo se aplica la regla. Luego, la consolidación de las métricas de estas subdimensiones, 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 mencionado, donde los componentes se relacionan según el esquema de la página 76.

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 ejecutadas fueron las siguientes:

- 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 se aseguró que cubrieran los requerimientos actuales y sus variantes.
- Por cada repositorio involucrado se crearon las subdimensiones (aproximadamente 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,

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

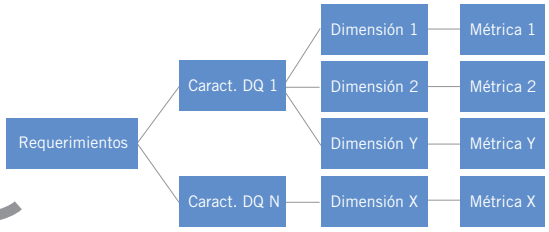
Características de calidad necesarias / Requerimiento DQ del negocio	Característica A	Característica B	Característica C	...	Característica X
Requerimiento 1		●			●
Requerimiento 2		●	●		
Requerimiento 3		●	●		
...	●			●	
Requerimiento N	●	●			

Relación: ● Fuente ● 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

3. Analizar resultados e impacto en los requerimientos



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

resguardar y visualizar hallazgos y resultados, hacer búsquedas 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 las 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:
  - La obtención de los hallazgos (anomalías) que cada regla detecta.
  - La obtención de métricas de DQ.
  - La recepción de los hallazgos en forma automática (Figura 7).

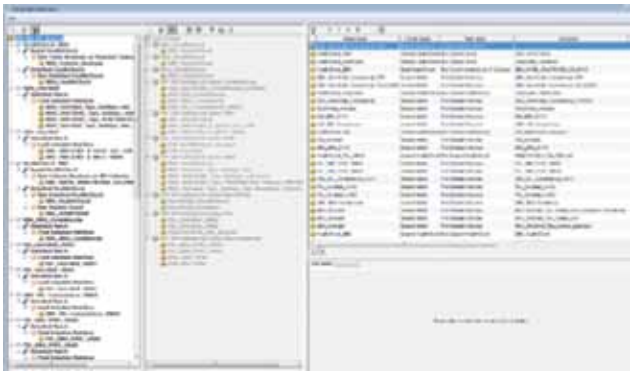


Figura 6. **Reglas asociadas a dimensiones. Jobs programados.**

Figura 7. **Informe resultados de calidad enviados automáticamente por mail.**

- La posibilidad de contar con métricas que consoliden las mismas dimensiones en distintos repositorios (por ejemplo, consistencia de un repositorio o de varios repositorios).
- La factibilidad de contar con un registro histórico de las métricas que permita aplicar metodologías de análisis para la prevención de las anomalías (estadísticas, causa raíz, etc.).
- La posibilidad de incrementar la complejidad y la cantidad de dimensiones/reglas sin requerir recursos extras para su gestión integral.
- Contar con un motor de búsqueda de errores (símil 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, al acceder geográficamente o por distintos filtros y/o taxonomías (Figuras 8 y 9).

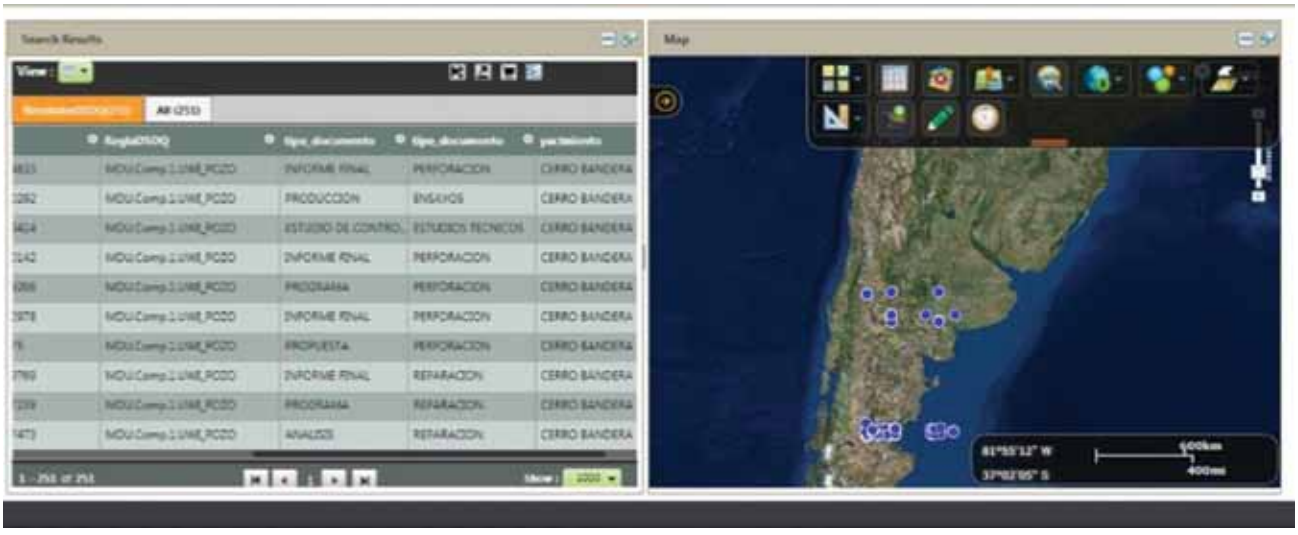


Figura 8. **Búsqueda de errores y visualización geográfica.**

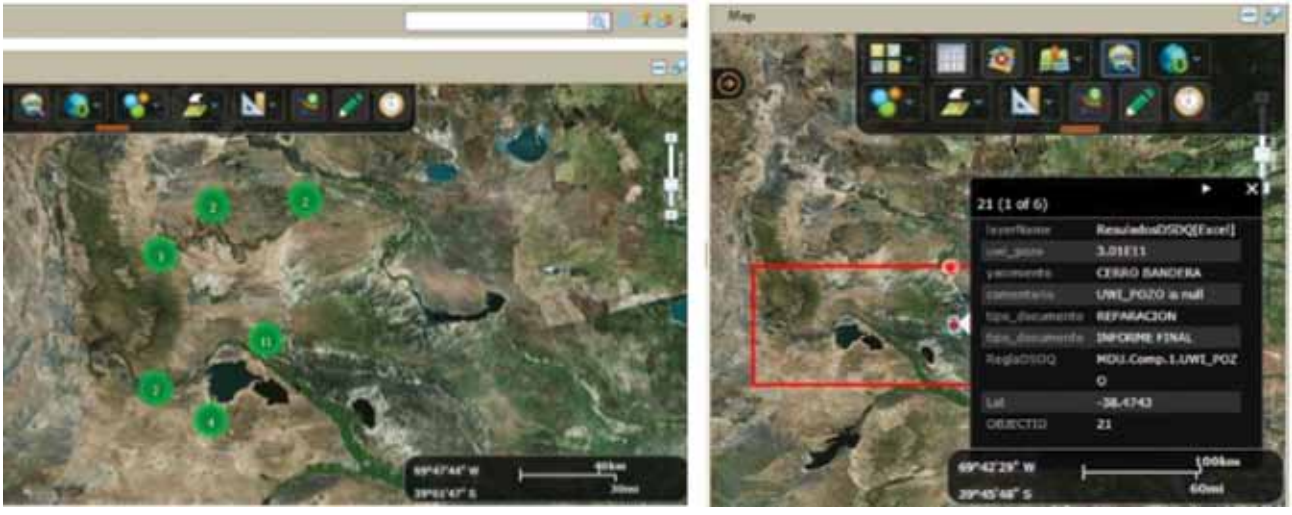


Figura 9. **Búsqueda de errores y visualización geográfica.**

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 (por ejemplo, el Transaccional y el de Real Time) compartan iguales subconjuntos de dimensiones (con distintas reglas), y que sus métricas puedan ser consolidadas.

De la implementación:

- No es condición necesaria contar con una plataforma única para la implementación del modelo, pero 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, la corrección y la evaluación de otras necesidades de calidad.
- Implementar la solución técnica permite obtener los registros de datos que se deben corregir, sin requerir del expertise para entender la complejidad de las reglas.
- Excepto casos muy puntuales, esta implementación puede cubrir un amplio espectro de la calidad de datos utilizados por procesos de distinta naturaleza.

Bibliografía consultada

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

Normas SQuaRE, http://iso25000.com/

Data Quality in Big Data: A Review, Noraini Abdullah, 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.