La ciencia de datos en una organización de E&P

Cómo armar un HUB de ciencia de datos en la práctica

Por ***Daniel Yankelevich*** (Practia)

La ciencia de datos, entendida como el análisis sistemático de datos utilizando algoritmos y técnicas automáticas para extraer información no evidente, ha ganado su lugar entre las metodologías de trabajo aceptadas y exitosas en la industria de petróleo y gas. Existen numerosos casos y ejemplos de aplicación de ciencia de datos tanto para explicar fenómenos como para la construcción de modelos predictivos, que abarcan situaciones tan distintas como la predicción de fallas en equipos de perforación, el mantenimiento predictivo de instalaciones de superficie, la simulación dinámica de los yacimientos, la identificación de causas de pérdida de producción, la optimización de flotas y la optimización de yacimientos, por mencionar solo algunos[1].

La aceptación de estas técnicas es impulsada por una serie de factores, entre ellos los principales son los siguientes:

- La baja de costos del procesamiento de información, que hace económicamente factible realizar cálculos complejos y, por lo tanto, permite realizar experimentos con datos y análisis estadísticos sumamente sutiles y avanzados.
- El desarrollo de nuevas técnicas y algoritmos más eficientes, sobre todo para la gestión y análisis de datos no estructurados. A la fecha, se estima que una pequeñísima parte (aproximadamente el 1%) de los datos no estructurados son utilizados en la toma de decisiones del negocio, y existe un consenso en que el 99% restante permitiría agregar muchísimo valor en forma de

Muchas organizaciones están incorporando la ciencia de los datos en sus operaciones. En este trabajo presentamos el aprendizaje y las conclusiones de la introducción de esta ciencia en una organización de E&P mediante la creación de un "HUB de ciencia de datos", los desafíos que esto plantea y recomendaciones basadas en la experiencia.

ahorro de costos y nuevas oportunidades[2].

- La proliferación de enormes cantidades de datos. A modo de ejemplo, se estima que un pozo genera en la actualidad entre 1 y 15 terabytes de datos, o 3 megabytes por metro perforado[1,3]. Si bien en las áreas de E&P siempre se trabajó con grandes cantidades de datos, por lo general, se trataba de datos correspondientes a una categoría, que eran analizados en forma compartimentada y muy específica (por ejemplo, la sísmica). El cruce de datos y el uso de información heterogénea presenta un nuevo desafío.

Muchas organizaciones han incorporado, o están por hacerlo, prácticas de ciencias de datos en sus operaciones. En este trabajo presentaremos el aprendizaje y las conclusiones de la introducción de ciencia de datos en una organización de exploración y producción de hidrocarburos mediante la creación de un "HUB de ciencia de datos" (en adelante, HUB), los desafíos que esto plantea y algunas recomendaciones basadas en la experiencia.

Por qué un HUB

Una vez resueltas las primeras resistencias y objeciones, la primera decisión en muchas organizaciones es la incorporación de personas formadas en ciencia de datos (los científicos de datos) para crear áreas que realicen el trabajo y armen las prácticas dentro de la organización. De hecho, la profesión de "científico de datos" fue catalogada como "la profesión más sexy del siglo XXI". Se publicaron varios artículos, en los que se analiza la formación y el perfil adecuado de este científico de datos[4]. Pareciera que, si se incluyen a los equipos científicos de datos, el problema está resuelto.

En la práctica, en organizaciones de un cierto tamaño, la incorporación de uno, dos o cinco expertos es solo un aspecto y una faceta para lograr la adopción efectiva de nuevas técnicas en la organización. No se trata solo de perfiles técnicos o personas brillantes: se trata de conformar una estructura efectiva, que permita:

- Cambiar la organización para que acepte una nueva forma de trabajo. La aparición de nueva información y nuevas técnicas influye directamente en la forma en que se toman las decisiones. El "siempre se hizo así" es cuestionado de una forma distinta y es necesario realizar un fuerte trabajo de gestión del cambio en la organización.
- Gestionar la demanda, ya que a medida que se realicen proyectos, aparecerán más áreas y actores con interés en realizar pruebas, prototipos o proyectos. Gestionar esta demanda no es solo gestionar y priorizar pedidos, por ejemplo, muchas veces los pedidos llegan en forma muy cruda y requieren un trabajo previo antes de considerarse como tarea realizable.
- Armar los casos de negocio, no solo para ayudar a convencer de la utilidad de esta aproximación a los problemas, sino también para medir el impacto y el retorno de la inversión realizada en forma realista.
- Brindar soporte a las áreas que están experimentando o trabajando con datos y requieren ayuda o, por el contrario, ofrecen su información para enriquecer al resto de la organización.
- Ejecutar proyectos de diferentes características y estilos: no es lo mismo una prueba de concepto con datos de una base relacional que un proyecto de reconocimiento de imágenes en un área remota. La capacidad de ejecutar proyectos (en forma interna, mediante subcontrataciones, o con un esquema de *sourcing*) es clave a la hora de dar respuesta rápida al negocio.

Para esto es necesario crear una estructura que cuente con cargos técnicos especializados y además con los siguientes aspectos:

- Metodología de trabajo, incluyendo procesos específicos.
- Gobierno, tanto de los algoritmos como de los datos y de las estructuras asociadas.
- Gestión del conocimiento que permita aprender y replicar lo aprendido a lo largo del tiempo (algo muy difícil de hacer si los proyectos de ciencia de datos son gestionados como proyectos aislados y no existe una estructura que los trascienda en el tiempo).

El HUB en la práctica

El caso al que hace referencia este trabajo se trata de la implementación de un HUB de ciencia de datos para el área de E&P de una empresa líder en la explotación de petróleo y gas. El HUB armado empezó a funcionar con siete roles, no todos ocupados por profesionales *full time*. De hecho, algún rol fue tomado por personal que ya se encontraban trabajando en gestión de datos (data management) y otro por gerencia, solo algunos eran científicos de datos en modalidad *full time*.

Adicionalmente, el equipo tuvo experiencia de armado de estructuras similares en otras organizaciones, lo que se toma como validación adicional para los conceptos trabajados en este documento.

Al momento de configurar un HUB, las discusiones sobre misión, visión y objetivos suelen dominar la agenda, pero luego aparecen disyuntivas mucho más concretas y pragmáticas.

Una discusión es sobre la estructura organizacional del HUB. Un modelo posible es un HUB centralizado, que concentre los proyectos e iniciativas de ciencia de datos. Las ventajas son evidentes: permite concentrar el conocimiento, facilita el apalancamiento de la inversión, evita la proliferación de herramientas y tecnología, y evita repetir proyectos o iniciativas. Por otro lado, cuenta con varias desventajas: se convierte en un cuello de botella para la ejecución de proyectos, se constituye como único defensor del tema en la organización, lo que indefectiblemente demora el cambio cultural, y requiere conocimiento de varios negocios (ya que debe ejecutar proyectos para todas las áreas). En el otro extremo, un modelo distribuido permite que las diferentes áreas desarrollen sus propios equipos, con altísima especialización en el negocio, y el HUB funciona solamente garantizando estándares, herramientas comunes y comunicando experiencias y conocimiento. En la actualidad, el mercado se ha decantado en forma bastante unánime por modelos híbridos. En estos modelos, el HUB funciona como un integrador, garantizando el uso de estándares y proveyendo conocimiento y herramientas, pero a la vez es una "primera línea de defensa", al tomar la ejecución de proyectos y al agregar profesionales y especialistas a proyectos de otras áreas. Esta visión "mixta" de un HUB, que además de proveer metodología, unifica forma de trabajo y garantiza estándares, que ayuda y resuelve problemas concretos de diferentes áreas, permite obtener las ventajas de ambos modelos si su gestión es adecuada.

Otro problema que suele aparecer en forma muy temprana es la existencia de silos de información en la organización. De hecho, en alguna bibliografía los silos se incluyen entre los principales inconvenientes a la hora de implantar una práctica de ciencia de datos.

Los silos pueden describirse como compartimientos estancos de datos, integrados por negocio o área funcional. Determinada área integra información de diferentes sistemas, bases de datos e incluso datos externos, pero solo a los fines de resolver la operatoria de su propio negocio, y en la mayoría de los casos se trata de datos transaccionales. Una enorme ventaja de los proyectos de ciencia de datos es que logran integrar datos de distintas fuentes y formatos, de esa manera encuentran relaciones y correlaciones entre los datos que no se había detectado con anterioridad. Para que esto suceda, es necesario romper los silos.

Una de las consecuencias de los silos es el gap semántico. Esto se da cuando áreas distintas usan el mismo

concepto para nombrar entidades o datos diferentes. Un ejemplo tradicional para explicar el gap semántico es la noción de "cliente". Pensamos que para una organización el concepto de "cliente" debería ser bastante claro, pero suele suceder que lo que es un cliente para el área de Marketing no es lo mismo que el cliente para el área de Ventas, ni hablar de áreas administrativas (donde el cliente solo aparece cuando se factura o se toman datos formales), para el área de Logística, y así sucesivamente. Es más, los datos que cada área considera relevantes sobre esa entidad "cliente" son distintos y, en muchos casos, la información no es consistente. Esto se multiplica por numerosas entidades y conceptos, y al tratarse de conceptos técnicos el problema puede ser aún mayor.

Las soluciones, en muchos casos existentes en la bibliografía de gobierno de datos, para resolver los problemas de los silos son, por ejemplo los registros dorados o validados. Sin embargo, una vez destruir los silos, puede ser un problema que tome años en ser resuelto y requiere un esfuerzo significativo. En nuestro caso, hemos encontrado que es mucho más efectivo pensar en silos permeables, es decir, no ponerse como objetivo destruir los silos, sino lograr que puedan conectarse e intercambiar datos (a veces resolviendo las inconsistencias o gap semánticos, a veces usando esquemas estadísticos o evitando datos de baja calidad). No es sencillo encontrar una forma sistemática de definir y trabajar con silos permeables, pero el cambio de foco ayuda a establecer objetivos más cercanos y alcanzables y evitar la idea: "en esta organización eso no se puede hacer".

Otro desafío es el dimensionamiento del equipo. Si se dimensiona para picos de demanda, se construye una estructura demasiado grande que, a su vez, será exigida para responder por un gasto mayor. Si se subdimensiona, se corre el riesgo de no poder responder en tiempo y forma y perder la oportunidad de un cambio. En nuestra experiencia, una estrategia inteligente de *sourcing*, que combine agilidad con disponibilidad de capacidades, suele brindar la mejor respuesta.

Finalmente, es necesario definir el alcance del HUB en la organización. En la sección anterior se presentaron varios roles y ventajas de tener un HUB. Sin embargo, una organización puede decidir cubrir solo una parte de esos roles y responsabilidades. En la figura 1 se presentan las diferentes capacidades que deben tener los profesionales del equipo, y en rojo se encuadran los roles cubiertos por el HUB en el caso de discusión. Este tipo de análisis permite entender claramente qué tareas quedarán fuera de las capacidades del HUB. No es razonable pensar que el HUB puede cumplir con todo lo que se espera si los perfiles no cubren todas las capacidades necesarias.

Figura 1. **Áreas de conocimiento cubiertas por el HUB de datos.**

Enfoque metodológico

El enfoque metodológico es parte integral del armado de un HUB de ciencia de datos, que complementa el equipo de especialistas y el grupo profesional. En los proyectos intensivos en datos, la metodología de trabajo no suele ser la misma que en otros proyectos tecnológicos, de sistemas o de software. De hecho, existe una dinámica particular en el uso de datos que llevó, históricamente, a proponer metodologías específicas para minería de datos o para analytics. Por ejemplo, CRISP-DM[5] fue un marco metodológico que se utilizó como referencia obligada durante años. Recientemente se empezó a hablar de DataOps[6] como una forma distinta de encarar los proyectos de datos. La filosofía detrás de DataOps[7] es muy similar a lo que fue la filosofía detrás de las metodologías ágiles de desarrollo de software. De hecho, hay un Manifiesto DataOps muy parecido al Manifiesto Ágil que dio origen a la tendencia de metodologías ágiles.

En lo concreto, la aplicación de una metodología se caracteriza por las prácticas que usa y sostiene. Por ejemplo, en el caso de metodologías ágiles, los *daily meetings* son una práctica que define y caracteriza la forma de trabajo. En el caso del HUB de datos, existen diferentes prácticas que caracterizan la forma de trabajo y es preciso considerarlas específicamente.

En primer lugar, la forma de análisis de los problemas debe ser centrada en datos, y no centrada en procesos o tecnología. Esto quiere decir que es necesario entender, documentar y poner foco en la creación, la transformación y el uso de los datos, vistos como elementos dinámicos y vivos (no como algo estático transformado por un proceso externo). El dato es el que se mueve y cambia y esos cambios son los que transforman al dato en información y conocimiento. Este cambio de punto de vista lleva a modificar, por ejemplo, las técnicas de documentación, pasando de historias de usuarios a recorridos del dato (o *data journeys*).

La automatización de tareas y procesos es otra parte clave de la metodología. Los proyectos de datos suelen tener mucho trabajo de preparación de datos (algunas fuentes aseguran que es de un 80% el tiempo dedicado a preparar y transformar datos, antes de proceder al análisis), este

Figura 2. **La metodología pensada como forma de trabajo, es parte fundamental del HUB, no sólo la gente o la tecnología[5].**

tiempo incluye carga, transformación, cambio de formato, integración y ajuste. La mayor parte de estas tareas puede automatizarse. El hecho de automatizar procesos de análisis y de carga y transformación de datos asegura un ahorro de costos y un aumento de velocidad, además de definir una forma de trabajo y poner el foco de los científicos de dato donde agregan más valor.

Finalmente, el gobierno de datos[8] desempeña un papel fundamental en la definición de la metodología. El gobierno no debe ocuparse solamente de controlar riesgos, mejorar la calidad del dato y asegurar su disponibilidad y resguardo, también debe garantizar que su acceso sea posible en forma y en tiempo adecuados para el análisis que se deba realizar. Si dedicamos más tiempo a controlar que a facilitar el uso de los datos, no hacemos gestión de datos, sino *compliance*.

Las claves

A partir de la experiencia de implantación y del intercambio con otros profesionales que han realizado imple-

mentaciones similares, podemos identificar una serie de claves a la hora de concretar una estructura de trabajo en ciencias de datos. Este aprendizaje se puede resumir en cinco claves o puntos principales:

Gestión del cambio: no es posible exagerar su importancia. La clave del éxito o del fracaso de un HUB dedicado a ciencia de datos estará en gran medida definido por su capacidad para modificar la forma de trabajo y la toma de decisiones en la organización. Esto debe encararse como una tarea activa y explícita del HUB que forma parte de su responsabilidad.

Evitar la torre de marfil: el riesgo de que otras áreas vean al grupo como académicos "en una torre de marfil" puede atentar contra la efectividad del HUB. Por este motivo es fundamental que el grupo interactúe frecuentemente con los equipos de negocio y muestre su predisposición a resolver los problemas que las otras áreas tienen.

Manejar y dominar múltiples tecnologías y herramientas: trabajar con una única herramienta puede llevar a que se elijan los problemas para los que la herramienta es buena y se descarten aquellos para los que no sirve. A la fecha, no existe una única herramienta o tecnología que abarque todos los problemas de manera efectiva. Además, la selección de una sola herramienta toma tiempo y requiere un nivel de madurez de la disciplina que aún no se ha logrado.

Priorizar el valor y no lo interesante: el HUB no debe elegir los proyectos porque sean más interesantes o respondan a sus prioridades, sino por la prioridad que estos tienen para el negocio. El valor agregado debe tenerse en cuenta y ser el principal indicador a la hora de seleccionar proyectos.

Foco en el cliente: el grupo del HUB debe funcionar como un área de servicio, que trabaja siempre en función de los intereses y pedidos de un cliente. Mantener el foco en el cliente facilita tanto la gestión del cambio como la adecuada priorización, y permite evitar el síndrome de la torre de marfil.

Conclusiones

En este artículo hemos presentado las lecciones aprendidas en la implementación de un HUB como forma de introducir el uso de ciencia de datos en una organización, específicamente en el área de E&P.

Hemos hecho referencia al cambio cultural requerido para que una organización pueda incorporar las prácticas de ciencia de datos, pero también para que pueda capturar el valor de estas prácticas y del cambio que la adopción implica. Tomar decisiones basadas en datos no es solo un problema de algoritmos o de técnicas estadísticas, incluye aspectos organizacionales profundos y, por lo tanto, el cambio cultural es una de las claves. Una forma de iniciar este cambio y de ver la reacción de la organización frente a la propuesta es iniciar con proyectos pequeños, un piloto o una primera iteración de un proyecto más grande. El hecho de iniciar con proyectos y resultados, aunque sean pequeños, en vez de con promesas y una mayor estructura, evita crear objeciones innecesarias hasta contar con evidencia concreta del valor que estas técnicas pueden agregar a la organización.

Junto a esta primera aproximación, y en paralelo al trabajo en pilotos, es conveniente generar demanda en forma inicial y entender la situación de la organización en cuanto a su madurez. La ejecución de talleres específicos de generación de demanda y de priorización de iniciativas permiten lograr estos objetivos con un costo muy bajo, tanto en tiempo como en recursos. Estos talleres de descubrimiento permiten discutir, relevar y clasificar iniciativas, además, forman parte de la metodología que un HUB en su etapa inicial puede utilizar. No hemos hablado de las etapas o estados por los que pasa un HUB, pero resulta claro que las capacidades y objetivos de una primera etapa no son los mismos que debe tener un HUB maduro que ya ha logrado varios éxitos.

Como consideración final, es importante remarcar que no es conveniente tener toda la estructura armada para empezar a trabajar con el HUB. En algunas organizaciones se inicia el trabajo seleccionando herramientas, proveedores, definiendo procesos, para luego abrirse a la organización. La visión recomendada en este trabajo (que responde a la idea de hacer foco en el cliente) es empezar con algunas iniciativas para aprender y construir sobre lo aprendido.

Referencias

1. Mark Mills, SHALE 2.0 *Technology and the Coming Big-Data Revolution in America's Shale Oil Fields, Manhattan Institute Research, N° 16, Mayo 2015.*
2. Andres Brun, Monica Trench, Thijs Vermaat, *Why oil and gas companies must act on analytics*, McKinsey&Company, October 2017.
3. David Wethe, "Better Fracking Through Sound-Sensing Fiber Optics", *Bloomberg*, July 11, 2013.
4. *Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review*. Thomas H. Davenport, D. J. Patil. Octubre, 2012.
5. Shearer C., el modelo CRISP-DM: el nuevo plan para la minería de datos, almacenamiento de los datos J (2000); 5:13-22.
6. "What is DataOps (data operations)? - Definition from WhatIs.com". *Search Data Management. Retrieved 2017-04-05.*
7. "DataOps - It's a Secret". www.datasciencecentral.com. *Retrieved* 2017-04-05.
8. *CEB IT Leadership Council for Midsized Companies* – "Data Governance: Step-by-Step Guide".