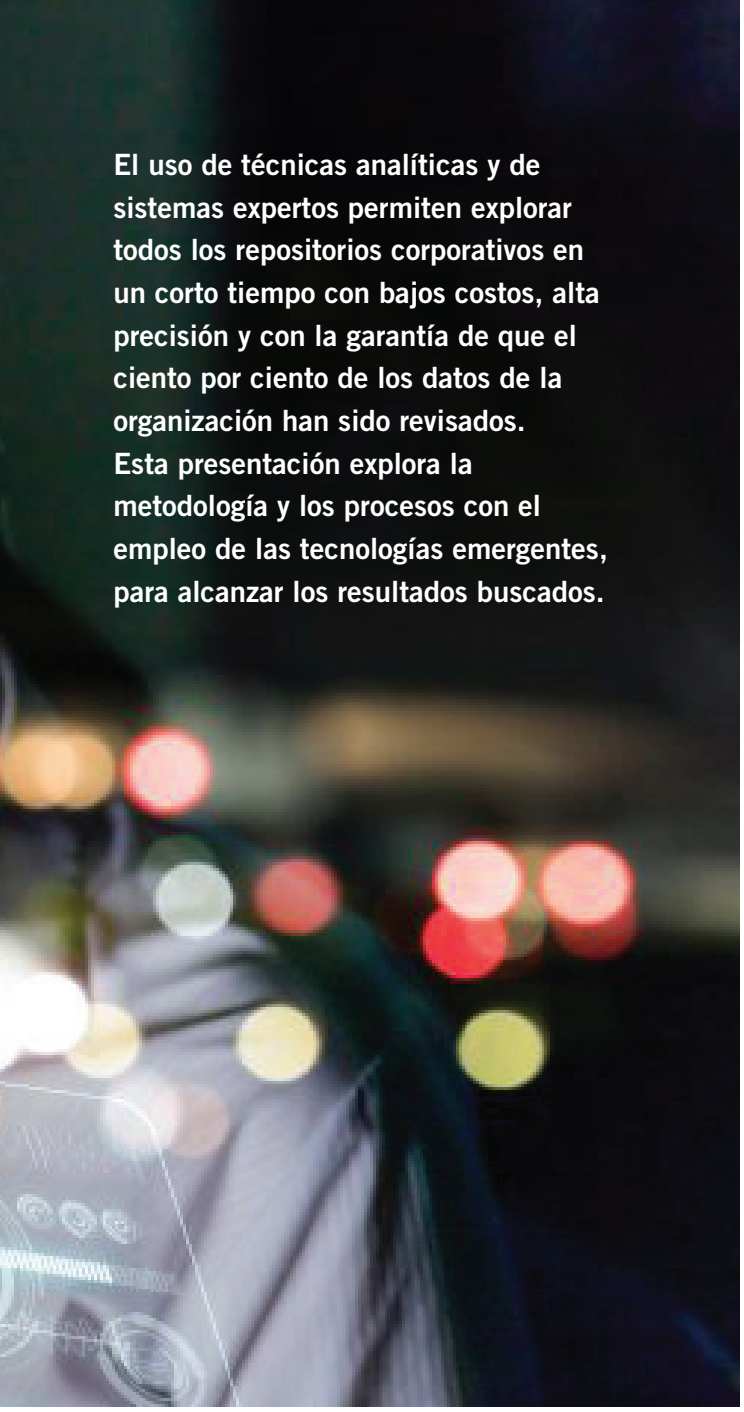
Estrategia de datos, *data management*, gobierno y gobernanza de datos, *master data management*.

Estrategias de analítica avanzada para la generación de nuevas oportunidades

Por ***Iván Rodrigo Plata Arango, Juan Pablo Pretelt y Christian Osorio Cabrera*** (Ecopetrol); ***Dirk Elric Adams*** (Kadme AS); ***Mario Gómez, Andrés Ulloa, Luis Molina, Mauricio Hoyos*** y ***Nicolas Balaguera*** (Wavecomm)



El uso de técnicas analíticas y de sistemas expertos permiten explorar todos los repositorios corporativos en un corto tiempo con bajos costos, alta precisión y con la garantía de que el ciento por ciento de los datos de la organización han sido revisados. Esta presentación explora la metodología y los procesos con el empleo de las tecnologías emergentes, para alcanzar los resultados buscados.

Se discute la estrategia para la implementación de procesos de automatización de un sistema ciento por ciento en un ambiente en la nube para el manejo de datos físicos y digitales en una empresa de clase mundial. La presentación se centra de manera novedosa en el manejo del gobierno del dato, la calidad de la información y la explotación de todo el contenido disponible con técnicas y sistemas similares a los utilizados bajo el estándar de revolución industrial 4.0.

En una etapa temprana se desarrolló el primer sistema de lenguaje natural con técnicas de generación e integración de una solución de interpretación y reconocimiento basada en técnicas de inteligencia artificial que permita la interpretación en lenguaje natural de las consultas típicas realizadas por una empresa petrolera, para la optimización de las búsquedas que se realizarán en un sistema integral como parte de un proyecto de buscador inteligente para bases de datos petrotécnicas.

Para lograr estos objetivos se buscó la implementación de un sistema que permitiera la evolución hacia la big data y la inteligencia artificial, y que estuviera diseñado con los servicios que brindar la nube.

Parte de la solución es permitir que sea la tecnología la que gobierne el dato y que controle la comparación y el poblamiento de atributos y que este proceso no dependa, en lo posible, de la intervención humana. Un sistema con estas características maximiza el uso de la información y optimiza los tiempos y recursos humanos invertidos.

Derivado del éxito temprano de esta funcionalidad de NLP, el sistema se extiende a otras funcionalidades que incluyen funciones analíticas extendidas con la identificación de patrones e imágenes e identificación y aumento de metadatos de modo cada vez más automatizado, tanto en la extracción directa de contenido (por ejemplo, un título o nombre de elemento) como en la extracción indirecta (por ejemplo, clasificación documental) derivadas en mayor proporción del análisis global del documento.

El resultado final es un sistema holístico que pueda automatizar los procesos de carga de archivos con oportunidad y calidad; que permita la identificación de palabras, elementos y frases; poblar con metadatos el sistema; y generar índices que posibiliten búsquedas complejas de palabras, términos e imágenes.

Desarrollo técnico

El proyecto comenzó como la implementación de un sistema de Lenguaje Natural en español (NLP), que incluyó el entrenamiento y la implementación del modelo de búsqueda en Lenguaje Natural para un sistema implementado en las vicepresidencia de Exploración, Desarrollo y proyectos de Ecopetrol, con el uso de los motores nativos de procesamiento de lenguaje natural del sistema (NLP Server) y modelos/categorías extraídas de conversaciones comunes de los usuarios especializados de Exploración y Desarrollo de Campos.

Esta etapa permite, mediante el entrenamiento de un sistema de aprendizaje profundo tipo ANN (Artificial Neuronal Network), el reconocimiento de sentencias básicas de consulta en el idioma español del personal de Exploración y Desarrollo de Campos, la generación automática de metadatos relacionados con esas sentencias, su respectivo filtrado y clasificación a través de categorías que generen los insumos suficientes para perfilar la posterior elaboración de vectores de búsqueda que son reinyectados al sistema con el objetivo de mejorar la experiencia del usuario final en su interacción con la interfaz web de consultas, luego de validaciones y pruebas del modelo implementado, se ha realizado la implementación dentro del ambiente de la plataforma para usar dentro de las interacciones con los usuarios finales.

Las fases de implementación fueron las siguientes:

Fase I: entrenamiento e implementación del Modelo Corpus Lingüístico Online.

Fase II: validación del Modelo por parte de usuario expertos.

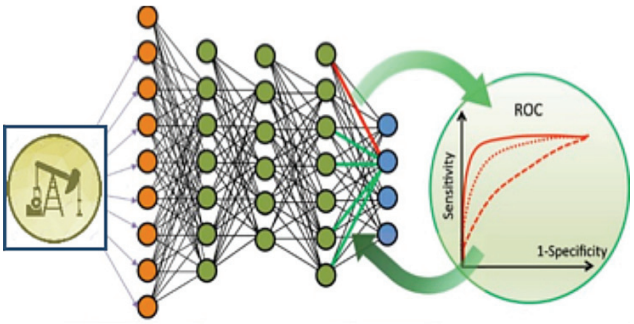


Figura 1. **Ilustración Modelo Red Neuronal.**

Fase III: conexión y presentación de resultados.

Fase IV: integración al modelo.

El proceso de entrenamiento del sistema es continuo. En la figura 2 se muestra un flujo de trabajo.

Gracias a este entrenamiento constante, se fortalecen las taxonomías y los corpus lingüísticos correspondientes, lo que permitió identificar dentro de las dinámicas de búsqueda, nuevas necesidades que demandaban información que no estaba presente en las tablas base, ni en los títulos o en las características de los documentos-objetos de esta búsqueda.

Con esta nueva necesidad latente, se opta por la extracción automática de nueva información de los archivos existentes; para ese fin se usan herramientas de OCR que, al ser combinadas con los patrones identificados, permiten la generación y la depuración de nueva información que puede ser aprovechada por personal que no tenía conocimiento de las fuentes de información originales.

Un ejemplo es la creación de un mapa de gradientes geotérmicos basado en información de pozos, donde la información de profundidad *versus* temperatura está implícitamente descrita en múltiples fuentes (ejemplo, encabezados de perfiles de pozos, documentos de perforación y mapas históricos, entre otros), pero no están explícitamente consolidados para su posterior estudio geográfico.

Otro ejemplo del uso de la metodología consiste en atender una necesidad de los operadores que requiera el cuerpo del registro de lodos de la empresa. Esta información no existe de manera explícita, para obtenerla se recurre a técnicas de IA que busca información de parámetros, como descripciones de gases, o realiza búsquedas indirectas de los registros, que permite rescatar la información "perdida" (Figura 3).

De acuerdo con lo descripto en los ejemplos implementados, se valida la hipótesis de generación de nuevo conocimiento tomando como insumo la información existente, lo que permite ampliar la visión con el uso de más herramientas de IA para poder identificar y extraer nueva información. En la siguiente etapa se toman como referencia las imágenes incluidas en las fuentes de información para buscar el mismo objetivo.

Las tecnologías de visión artificial abren el panorama a nuevas identificaciones, tomando como punto de partida imágenes que se consignaron en documentos electrónicos o digitalizados que no han sido tenidos en cuenta anteriormente en el análisis de estos. Este proceso se inicia con una identificación de modelos gráficos objeto de búsqueda, se realiza la identificación de los posibles patrones de estos objetos (modelos) y se hace un entrenamiento asociado a los objetos identificados.

Una vez consolidado el modelo entrenado, se ejecuta sobre los documentos por analizar una lectura automática para extraer las imágenes de valor y que, a la luz del modelo entrenado, tengan unos valores de confianza altos.

Las imágenes de valor encontradas en los documentos, se asocian a una categoría de datos accesible para los usuarios del proceso generando un nuevo valor a la información.

Finalmente, al combinar búsquedas de texto e imágenes se amplía la probabilidad de encontrar información asociada a los patrones parametrizados en el sistema; gra-

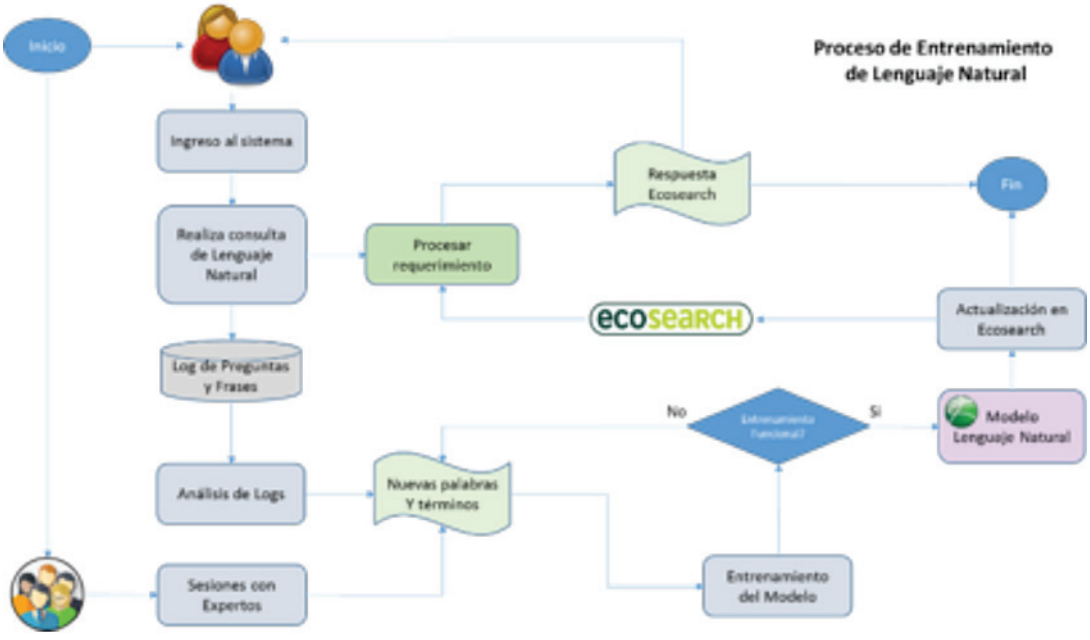


Figura 2. **Proceso de entrenamiento.**

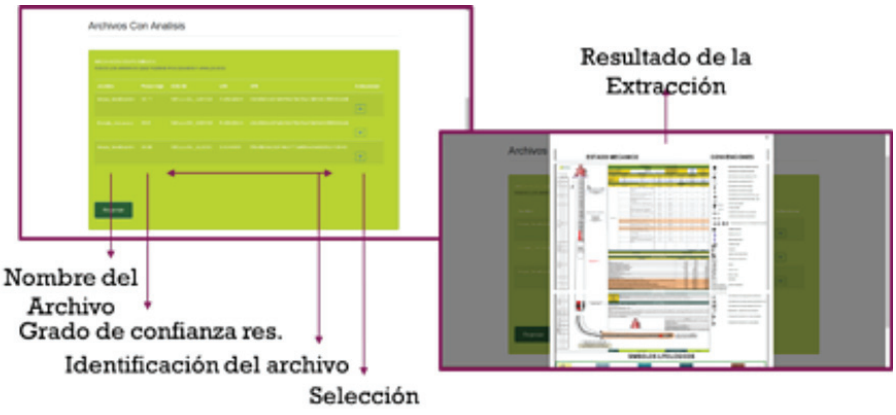


Figura 3. **Ejemplo para un caso de uso "archivos de estados mecánicos de pozos".**

cias al uso de Lenguaje Natural, términos equivalentes y búsqueda por imágenes, se abordan los insumos de información en tres frentes lo que permite extraer valor y generar conocimiento partiendo de los repositorios de datos corporativos que nos permiten mejorar en los siguientes frentes de trabajo:

- Alimentación y extracción de metadatos.
- Extracción, clasificación y procesamiento de imágenes.
- Clasificación de imágenes (imágenes de valor dentro de documentos, pero nunca clasificados ni tipificados).
- Entrenamientos con las posibles formas o patrones según relevancia temporal o estructural.
- Búsquedas por asociación en documentos nuevos.
- Inclusión de imágenes dentro de una metaclase específica, se usan tecnologías de inteligencia artificial para disponibilidad de datos que no están a la vista.
- Reconocimiento de imágenes sin discriminar el origen del dato.
- Identificación de búsquedas de imágenes a través de textos (patrones de textos), con el fin de mejorar el desempeño por combinación de técnicas.
- Enriquecimiento de metadata.

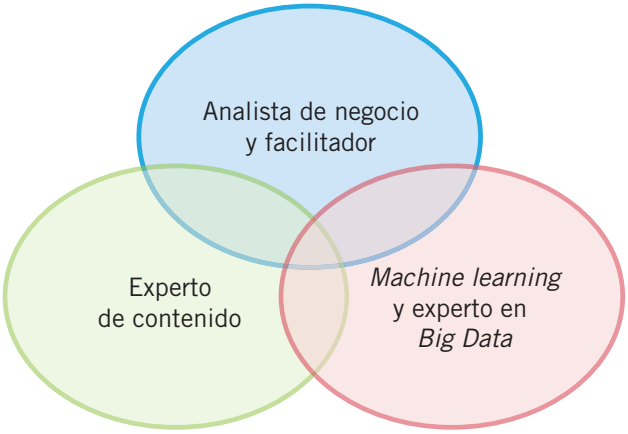


Figura 4. **Equipo de trabajo.**

En la figura 4 se muestran los componentes y la configuración general necesarios para ejecutar este tipo de proyectos.

Una primera etapa para el desarrollo y la implementación de un sistema de estas características requiere la consideración de los siguientes aspectos:

- Uso de herramientas públicas – Minimizar el uso de plataformas de desarrollo propietarias. Una de las premisas de un desarrollo de este tipo es de mantener un sistema independiente de las plataformas de desarrollo propietarias de los proveedores de "hosting". Esta revisión es fundamental, ya que un desarrollo de este tipo requiere de un balance entre el costo de implementarlo con herramientas propias versus herramientas provistas por el "hosting". Esto a su vez se tiene que balancear contra los costos de "escape" en caso de que se quiera migrar de "hosting" a otro.
- Usar una combinación de sistemas de analítica propios con servicios contratados a terceros para maximizar el uso de sistemas ya consolidados, pero buscando una racionalización de costos al optimizar el uso de estas herramientas y guardar los metadatos generados.
- Tipo de estrategia de manejo de datos. Un balance entre ejecutar las funciones analíticas básicas de modo masivo *versus* tener funciones que serán ejecutadas cuando se las requieran.
- Arquitectura para implementar. La arquitectura de la solución propuesta tiene que brindar la posibilidad de manejar grandes cantidades de datos, permitir el procesamiento distribuido de datos, brindar los más altos estándares de calidad, persistencia, escalabilidad, costo-eficiencia.

Finalmente, el esfuerzo técnico y la metodología implementada buscan apoyar los siguientes puntos clave:

- Uso de imágenes etiquetadas como referencia.
- Entrenar al modelo para que aprenda la diferencia entre texto, tablas e imágenes.
- Entrenar al modelo para que extraiga la información de documentos, imágenes, PDF escaneados, etc.
- Las imágenes extraídas se clasifican en clases con significado para el negocio.

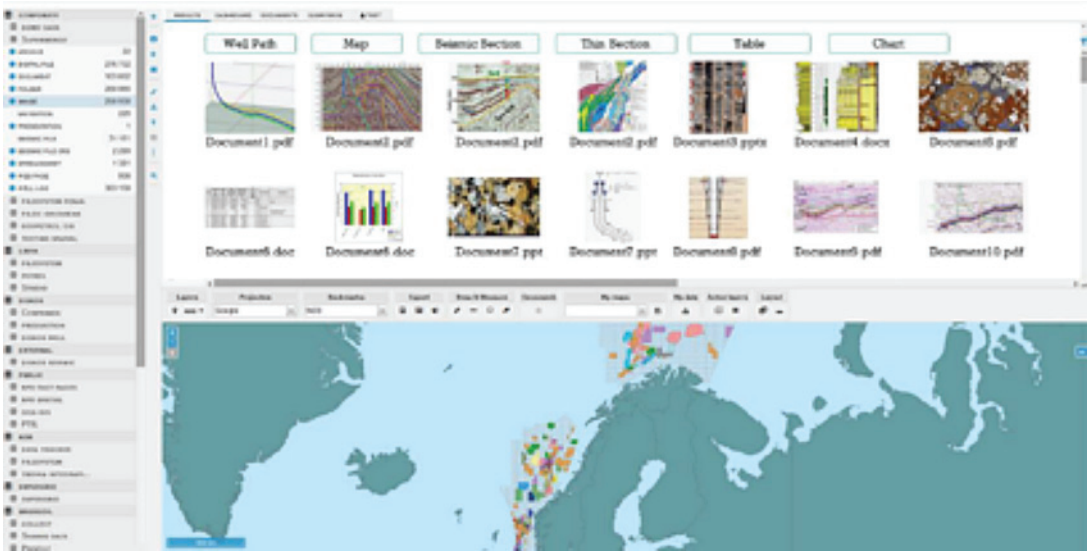


Figura 5. **Ejemplo de clasificación de imágenes y su georreferenciación.**

Resultados obtenidos

En el desarrollo de los primeros clientes se logró validar la importancia de los siguientes aspectos del producto:

- Sistema tan agnóstico como sea posible de la plataforma que lo contiene (sistema independiente del hosting)
- Búsqueda multidimensional en lenguaje natural dentro del contenido y contexto de documentos e imágenes en varios formatos (NLP).
- Un sistema entrenable con un corpus de "conocimiento" y entrenamiento avanzado que puede mejorar los procesos en la medida que se siga entrenando y, al mismo tiempo, se adapte con facilidad para ser entrenado en nuevas funciones de extracción de metadatos e imágenes y generación de metadatos.
- Facilitar la "democratización de los datos" mediante el acceso al Datalake de los nuevos datos y productos generados que centralice toda la información relevante con el uso de formatos abiertos y que permita su libre operación en un entorno tecnológicamente seguro, robusto y flexible. El sistema otorga acceso dentro de la empresa, así como a entidades externas autorizadas (socios, proveedores, reguladores, etc.) con datos mejorados, correctamente clasificados e identificados y rápidamente asequibles sin los costos asociados de hacer estos procesos a mano por especialistas.
- Eliminar la necesidad de que profesionales costosos y con tiempo escaso tengan que hacer la labor de búsqueda, limpieza y clasificación de información y puedan apoyarse en un sistema que les permita solicitar estos servicios de manera automática,

En la clasificación de imágenes se buscan varias ganancias de gran valor y son las siguientes:

- Generación de modelos de imágenes para clasificar las figuras extraídas en grupos empresariales.
- Visualización rápida para los usuarios.
- Presentación de las cantidades y la ubicación de las imágenes en el contexto de los documentos padres.
- Georreferenciación de los documentos y de las imágenes clasificadas (Figura 5).

Conclusiones

El uso de técnicas de big data e inteligencia artificial para los procesos de clasificación, extracción de metadatos, procesamiento de lenguaje y otras técnicas resulta en grandes beneficios para toda la cadena de valor, sin limitarse a formas más simples y rápidas de resolver los problemas relacionados con los datos y la información; la adopción de tecnologías que sean más eficientes y menos costosas.

La tecnología actual proporciona una plataforma mucho más avanzada para la gestión de datos e información que puede producir conocimiento con la interacción humana y las tecnologías de inteligencia artificial; acceso a datos ricos, organizados y de alta calidad que permiten maximizar los resultados y minimizar los costos y los riesgos en el proceso de toma de decisiones.

Además permite:

- Automatización en los procesos de clasificación, identificación y búsqueda de contenido.
- Reasignación de especialistas en buscar información y generar conocimiento.
- Ubicación y aprovechamiento de información "perdida", que un sistema experto puede conseguir y presentar en un menor tiempo.
- Aumento en la riqueza del ecosistema de datos.
- Producción de herramientas que puedan ser escaladas para procesos más avanzados, como la digitalización automática y la generación de contenido de manera semiautomática e incluso completamente automática una vez que el modelo esté cien por ciento entrenado.
- Acortar los tiempos desde la recepción de la información hasta disponibilidad al menor tiempo posible (años, meses,semanas, días, horas).