Infraestructura de Big Data para el Proceso de Evaluación de Integridad de Ductos en la Industria Petrolera

Big Data Infrastructure for Pipeline Integrity Assessment Processes in the Oil and Gas Industry

Gustavo Calzada-Orihuela[1], Gustavo Urquiza-Beltrán[1]
Jorge A Ascencio-Gutiérrez[2] y René Santaolaya Salgado[3]

[1]Centro de Investigación en Ingeniería y Ciencias Aplicadas. Universidad Autónoma del Estado de Morelos
Av. Universidad No. 1001, Col Chamilpa, Cuernavaca, Morelos, C. P. 62209
[2]Universidad Politécnica de Quintana Roo
Av. Arco Bicentenario, Mza. 11, Lote 1119-33, Km. 255. Cancún, Quintana Roo, México. C.P. 77500
[3]Departamento de Ciencias Computacionales. Centro Nacional de Investigación y Desarrollo Tecnológico CENIDET
Interior Internado Palmira S/N, Col. Palmira, Cuernavaca, Morelos, C.P. 62490

[*]gurquiza@uaem.mx

PALABRAS CLAVE:

Almacén de Datos, Big Data, Big Data Analytics, Depósito de Datos, Evaluación de Integridad, Evaluación de Riesgo, Minería de Datos, Optimización, Riesgo.

RESUMEN:

El petróleo es un material esencial para la vida cotidiana. La sociedad y economía están íntimamente vinculadas a los combustibles fósiles. Sin embargo, para que el petróleo pueda ser utilizado debe ser procesado previamente y transformado en sus productos derivados, como la gasolina, keroseno, aceites, gases, entre otros. Para lo cual se tiene que transportar el petróleo crudo desde los yacimientos hasta las diferentes estaciones de procesamiento. El transporte de hidrocarburos se realiza por diferentes métodos, pero el uso de los ductos es el más común alrededor del mundo debido a su confiabilidad y efectividad. Y a pesar de ser un sistema considerado como seguro, no es infalible y en ocasiones, puede llegar a fallar, provocando pérdidas económicas, daños ambientales y pérdidas humanas. Para prevenir estos siniestros, la industria petrolera continuamente está invirtiendo recursos y esfuerzos en el desarrollo de Algoritmos de Evaluación de Riesgo para prevenir las fallas en los ductos. Estos algoritmos se basan en los datos e información relacionada con el sistema de ductos, con el fin de dar luz por medio de proyecciones y estimaciones hacia el futuro. Parte de estos esfuerzos se llevan a cabo en el desarrollo e investigación de los modelos de riesgo, sin embargo, la industria está mirando a las nuevas tecnologías computacionales para obtener el mayor beneficio posible de sus datos. Big Data, así como en otras industrias y disciplinas, es un conjunto de herramientas que están abriendo caminos en la ciencia donde antes no los había. Por lo que este proyecto tiene el propósito de integrar elementos del Big Data para aplicarlos en los procesos de Evaluación de Riesgo en la industria petrolera para poder optimizar los procesos de toma de decisiones por medio de la estructuración y explotación de datos.

KEYWORDS:

Data Warehouse, Big Data, Big Data Analytics, Data Repository, Integrity Assessment, Risk

ABSTRACT:

Oil is an essential material in the daily life. Society and economy are intimately linked to the fossil fuel. Nonetheless, in order to use the oil, it needs to be previously processed and transformed into oil derivatives, such as gasoline, kerosene, oils, gas, among others. For which the crude oil has to be transported form the wells to the different processing stations. Oil is transported by several

Assessment, Data Mining, Optimization, Risk

methods, however, pipelines are the most common method used worldwide due to its reliability and effectiveness. Regardless pipeline systems are considered safe, they are not flawless and might fail, promoting economic loses, environmental damages and human loss. In order to prevent these failures, oil industry is continuously investing resources and efforts in the development of Risk Assessment Algorithms to prevent them. These Algorithms usually are based on data related to the pipeline systems to shed some light creating projections and estimations towards the future. Some of these efforts are carried to the development and research risk models, nonetheless, the industry is focusing on new computer technologies to obtain the greater outcome from the data. Big Data is a computational tool set which is creating paths in science where there were none. This is the reason why this project has the purpose of integrate Big Data elements and apply them into the oil industry's Risk Assessments in order to optimize decision making process through structuration an intelligent data exploitation.

• **Recibido: 31 de julio de 2018** • **Aceptado: 10 de abril de 2019** • **Publicado en línea: 31 de octubre de 2019**

1. INTRODUCCIÓN

En la industria petrolera así como en la mayoría de las industrias, el cambio es un factor constante que no puede ni debe evitarse. El ambiente económico, social, científico, industrial entre otros, cambia y nada se queda igual, por lo que las organizaciones compiten en una dinámica que exige una alta flexibilidad y adaptabilidad, así como los equipos de trabajo, para responder óptimamente. Para ello, se utilizan diversas herramientas técnicas y tecnológicas donde las Tecnologías de Información se han convertido en un agente fundamental en este continuo proceso de cambios.

En las últimas décadas, los datos y la información se han convertido en un activo extremadamente valioso para las organizaciones a nivel mundial. Y aunque el procesamiento de datos no es algo nuevo, el nacimiento de tecnologías como Big Data, ha creado un potencial al acceso de múltiples fuentes de datos, transmisión y análisis de información, combinando disciplinas como la Ingeniería del petróleo, Geología, Ciencias Computacionales, Finanzas, entre otras, para la explotación inteligente de datos y extracción de conocimiento.

Muchas empresas, organizaciones y centros de investigación alrededor del mundo, están invirtiendo múltiples esfuerzos y recursos para obtener mecanismos y tecnologías para el procesamiento inteligente de sus datos. En el reporte del Foro Económico Mundial en 2012 [1], se reconoce la amplia aplicación e importancia de los sistemas de procesamiento y análisis de datos masivos en diferentes áreas de la sociedad, como la Educación, la Salud, Agricultura, Servicios Financieros, Competitividad del Mercado, entre otras. Por su parte, la Organización de las Naciones Unidas ha estado trabajando a través de Global Pulse, en la vinculación de datos provenientes de dispositivos móviles para identificar posibles situaciones que requieran acciones oportunas por parte de los distintos gobiernos y evitar crisis sociales o económicas. La industria petrolera por supuesto que también está cimentando su futuro sobre el uso de la información implementando Sistemas de Información (SI) en los diversos niveles de la organización [2].

En respuesta a las necesidades y oportunidades tecnológicas el Centro de Investigación en Ingeniería y Ciencias Aplicadas (CIICAp), de la Universidad Autónoma del Estado de Morelos (UAEM) con la colaboración del Instituto de Ciencias Físicas (ICF) de la Universidad Nacional

Autónoma de México (UNAM), se está desarrollando un trabajo de investigación cuyo enfoque es la implementación de una infraestructura computacional basada en componentes de Big Data para contener los datos asociados a los sistemas de ductos con la finalidad de generar respuestas eficientes por medio de la generación de conocimiento a través de la disposición de datos utilizando un Almacén de Datos (DW) constituido por diferentes depósitos para facilitar el almacenamiento histórico y dimensional para procesos analíticos.

2. LOS DATOS EN LA INDUSTRIA

Alrededor del mundo, los procesos de toma de decisiones son elementos críticos y fundamentales para las organizaciones y el desarrollo de distintas industrias, para lo cual se requiere información que al ser procesada por mecanismos analíticos se encuentren indicadores clave o factores de interés según cada organización. En la industria petrolera, este proceso pieza compleja que posee un riesgo latente para la industria, la población y el medio ambiente.

Desde el punto de vista organizacional, los diferentes niveles de las inteligencias de una empresa (niveles de generación y aplicación del conocimiento) conforman la Inteligencia Organizacional, que es definida por [3] como "el ensamblaje colectivo de beneficios con valor agregado derivados de los bienes intangibles de la organización". Estos bienes pueden ser los datos consumidos y generados dentro de una organización. La Inteligencia de Negocios (BI) es un conjunto de métodos técnicos y tecnológicos cuyo objetivo es la explotación inteligente de los datos. Y aunque su origen se asocia con aspectos administrativos, financieros y de marketing, ha generado un gran interés en otras industrias y disciplinas como las ciencias e ingenierías para expandir el conocimiento y su valor en los procesos durante la generación, procesamiento y almacenamiento masivo de datos.

De acuerdo al trabajo de Michalewicz [4], para que las industrias y organizaciones se adapten continuamente a los ambientes cambiantes, requieren de sistemas de Inteligencia de Negocios Adaptativos que contengan los siguientes tres componentes:

- Un componente para hacer predicciones.
- Un componente para tomar decisiones lo más óptimas posibles.
- Un componente para adaptar el módulo de predicción ante los cambios del ambiente.

El valor agregado del proceso de BI incluye:

- La velocidad a la que los reportes pueden ser generados.
- La precisión del contenido
- El grado del formato amigable con el usuario.

De acuerdo a Sabherwal [5] el concepto de Conocimiento es intrínsecamente diferente de Información. El conocimiento en un área lo define como "el conjunto de creencias justificadas acerca de las relaciones entre conceptos relevantes a esa área particular". En la Figura 1 se muestran las diferentes capas para el proceso de toma de decisiones basadas en conocimiento [5].

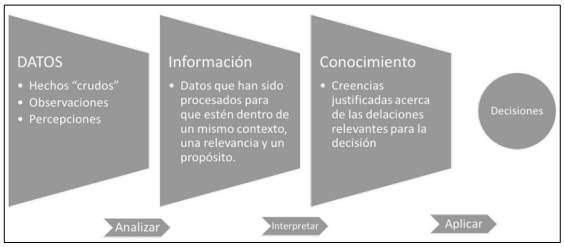

Figura 1. Capas de interacción entre Datos Información y Conocimiento.

Los datos generados en los niveles operativos generalmente se les conocen como datos "crudos", es decir que carecen de procesamiento analítico y que por sí mismos, no poseen un significado. El siguiente nivel se caracteriza por un procesamiento de los

datos crudos o de niveles inferiores con el objetivo de obtener un significado o información de ellos. En los niveles superiores, los datos crudos ya no son visibles en su forma original, sino que son procesados previamente para convertirlos no sólo en información sino en conocimiento con valor para la empresa, estructurando procedimientos para encontrar, de alguna forma, patrones que posteriormente se pueden convertir en proyecciones para responder y mitigar futuras situaciones antes de que éstas se presenten además de la creación de diversas estrategias, como la optimización en la localización de pozos de extracción, nuevas metodologías para el taladrado direccional, procesamiento de datos intensivos, procesamiento de datos para la Evaluación de Riesgo en Ductos e Instalaciones, entre otros.

Sin embargo, la gran cantidad de datos en las organizaciones generó grandes retos para los sistemas computacionales tradicionales por lo que metodologías y tecnologías como la BI, minería de datos, Internet de las Cosas (IoT) e incluso la Inteligencia Artificial, se fusionaron para dar pie a lo que se conoce como Big Data. De acuerdo a Lohr [2], un artículo del New York Times, el Big data puede utilizarse en campos tan variados como ciencias y deportes, publicidad y salud pública. Así mismo, el Centro de Investigación en Economía y Negocios en Inglaterra, predice que la mayoría de los sectores beneficiados de los análisis de Big Data serán los servicios financieros, sectores públicos, ventas y manufactura [6]. Así mismo, en [7] se describe una implementación de un Sistema de Soporte de Decisiones (DDS) para responder al riesgo de las fallas generadas en los ductos de agua en los sistemas de distribución australiana.

Las compañías del petróleo están conscientes del poder de la información y la ventaja competitiva que provee, tanto en aspectos económicos de la organización, como en el dominio tecnológico y beneficios para la sociedad. Por lo que está mostrando un gran interés por aprovechar las tecnologías emergentes para explotar de la mejor forma posible los datos que se producen día a día y obtener una ventaja por medio del análisis del pasado, el estudio del presente y la proyección del futuro.

3. EL RIESGO EN LA INDUSTRIA DEL PETRÓLEO

El riesgo es un término que puede adoptar diferentes significados dependiendo del dominio que lo describa y los factores que lo compongan. Por ejemplo, el riesgo de desarrollo de cáncer depende del historial médico y la genética de un paciente, la identificación de factores de riesgo de desempleo puede expresarse basándose en el comportamiento y actividad de las redes sociales, además, las proyecciones del riesgo de una crisis social se puede medir por medio del comportamiento del precio de arroz en zonas de riesgo, así como la detección de nuevos yacimientos petroleros de acuerdo a patrones de datos geológicos.

En México, la Norma Oficial Mexicana [8] es el instrumento oficial que dictamina los elementos, procedimientos, estándares y regulaciones para los procedimientos de los modelos de evaluación de Riesgo de los ductos para el trasporte de petróleo, ya sea en un sistema terrestre o marino. Este documento define al Riesgo de falla en los ductos como “una combinación de la probabilidad de que ocurra un evento y su consecuencia”. En el trabajo de Muhlbauer [9], el Riesgo se define como un comportamiento teorizado del ducto durante un tiempo. Esto significa que, utilizando metodologías y datos, se crean modelos para describir el comportamiento del ducto con la finalidad de predecir futuros incidentes.

3.1. Evaluación de integridad de ductos

Los sistemas de ductos es el método de trasporte de hidrocarburos muy utilizado para distribuir el petróleo y otros de sus subproductos a través de distintos tipos de regiones geográficas. De acuerdo a Cosham [10] y [11], la aceptación del uso de ductos se debe la buena reputación histórica que tiene comparado con las estadísticas de otros medios de trasporte, como el ferrocarril, camiones o buques. Sin embargo, los sistemas de ductos también pueden fallar. Por lo que la industria petrolera dirige sus esfuerzos en la investigación, elaboración e implementación de métodos para detectar posibles fallas y su conocer su posible severidad.

La validación de los niveles de Riesgo depende de las reglas y normas propias de las compañías, sin embargo éstas, normalmente están guiadas por las regulaciones oficiales, como la Sociedad Americana de Ingenieros Mecánicos (ASME), la Sociedad Americana del Petróleo o la Norma Oficial 027 en México.

3.2. Análisis de datos en la industria petrolera

Como en muchas industrias, instituciones científicas, centros de investigación y empresas en general, la industria del petróleo se enfrenta a distintos retos tecnológicos y se ayuda de los Sistemas Computacionales para mejorar su rendimiento y el conocimiento del ambiente en el que se desenvuelve. Los analistas, expertos y responsables de los sistemas de distribución de hidrocarburos están en la continua búsqueda de herramientas que puedan asistirles a tomar mejores y más oportunas decisiones para favorecer y proteger al negocio, al medio ambiente y a la sociedad. Sin embargo, el reto al que se están enfrentando es que entre más sistemas se implementen, más datos se generan.

El Proceso de Evaluación de Riesgo o de Integridad de Ductos, define las reglas y estándares para controlar, administrar, procesar y almacenar los datos y mecanismos que se utilizarán para determinar los valores de Riesgo y la Administración o Control de Riesgo. Aunque no existe un proceso único para llevar a cabo estas actividades, muchas organizaciones petroleras agrupan las actividades principales en: Modelado del Riesgo, Colección y Preparación de Datos, Segmentación de Ducto, Evaluación de Riesgo y por último la Administración de Riesgo.

Una de las finalidades fundamentales de este trabajo es asistir al proceso de toma de decisiones en la industria petrolera por medio del soporte al proceso Administración de Riesgo por medio de la Evaluación de Riesgo de los Ductos. Para lograr esto, se tuvieron que abordar varias áreas técnicas y tecnológicas de las Ciencias de la Computación para el desarrollo de una propuesta de infraestructura informática.

4. BIG DATA Y BIG DATA ANALYTICS

Es inevitable, los datos se han convertido en una parte fundamental de nuestras operaciones diarias. Desde mensajes de texto, correos electrónicos, llamadas por Internet, contenido web, video conferencias, contenido audiovisual, bases de datos estructuradas y no estructuradas, contenido en la nube, sensores de todo tipo que miden y registran lo que sucede día a día, segundo a segundo. Los datos van de un lado a otro incesantemente, siendo procesados, analizados, estudiados, explotados, almacenados y retransmitidos. La cantidad y complejidad de los datos han alcanzado un punto tal que han superado las infraestructuras de soporte, procesamiento y almacenamiento, así como la capacidad de las organizaciones para usarlos [12].

En el Foro Económico Mundial de 2012, realizado en la ciudad de Davos en Suecia, se reconoció ante múltiples países el gran impacto que tiene la información en la sociedad moderna, declarando que los datos son una nueva clase de bien económico, una entidad comerciable con un alto impacto en la economía mundial, tanto en las organizaciones tradicionales como en las empresas digitales.

El término de Big Data ha crecido y está siendo usado para describir el fenómeno del incremento en el volumen, la complejidad y la disparidad de los datos. De acuerdo a [12], la sociedad está convencida de que existe valor en cada parte o pedazo de datos, siempre y cuando se sepa la forma de cómo analizarlo. Este ha creado la tendencia de cruzar referencias de datos masivamente, para lo cual se necesitan motores analíticos más poderosos para procesar una cantidad mucho mayor de datos.

De acuerdo al trabajo de Dumbill [13], el término de "Big" no es necesariamente por el tamaño, sino se relaciona al fenómeno que estamos tratando de registrar y entender, por lo que la perspectiva del Big Data, la cuestión es qué y no porqué, es decir que es más importante qué se desea analizar y no el porqué, ya que comúnmente se desconoce en las primeras fases de desarrollo.

4.1. Procesos generales de Big Data

El diseño de una solución de Big Data normalmente comienza desde el análisis de los datos de salida mientras que los procesos de operación comienzan desde los sistemas de datos fuente. Una solución de Big Data debe de identificar los datos de estos sistemas fuente y extraerlos periódicamente para posteriormente cargarlos en un sistema de almacenamiento objetivo, por ejemplo, un Almacén de Datos.

El proceso de Extracción de los datos forma parte de una entidad denominada como ETL. Esta entidad, que es un conjunto de motores de procesamiento de datos, se encarga de realizar la conexión con los sistemas fuente y extraer todos los datos de interés para después almacenarlos temporalmente en una base de datos denominada Staging, de donde serán tomados para ser transformados, estandarizados y "limpiados" de acuerdo a las normas para la administración de los datos, las cuales deben de cumplir con el formato requerido por los sistemas de salida.

Aunque no existe un método general para la construcción de una solución de Big Data, se puede resumir en los siguientes pasos:

1. Determinar un equipo de trabajo conformado por líderes del negocio y líderes de las Tecnologías de la Información para buscar la colaboración y acuerdos entre los aspectos tecnológicos y económicos,
2. Estudiar los requerimientos y necesidades particulares de la organización o departamento para el cual se desarrollará la solución.
3. Estudiar la factibilidad, técnica-tecnológica y económica, del diseño y construcción de la solución de Big Data.
4. Estudiar y determinar los datos de salida y los formatos necesarios para ajustar la infraestructura a las necesidades de la organización y/o departamento.
5. Diseño e implementación de los sistemas de almacenamiento, que contendrán los datos que serán utilizados para los análisis.
6. Determinación e identificación de los sistemas fuente y los datos de interés.
7. Análisis, diseño y construcción de mapeos entre los distintos sistemas de almacenamiento para identificar las relaciones entre unos y otros.
8. Diseño y construcción de los procesos de Extracción, Transformación y Carga de Datos, que serán los responsables de colectar, procesar y transmitir los datos hasta los sistemas objetivo.

De acuerdo a Almeida [14], el arquitecto o ingeniero de Big Data debe incluir en su diseño el tipo y objetivos del almacenamiento, las reglas de administración, el mapeo y la relación entre depósitos de datos, así como el análisis detallado de los requerimientos del sistema. Los procesos de recolección, transmisión y almacenamiento de datos generan costos y es necesaria la inversión de recursos, por lo que los equipos o responsables de desarrollo deben de considerar opciones económicas y tomar decisiones sobre qué es fundamentalmente necesario para la organización y analizar el costo-beneficio de la solución, así como las limitaciones del presupuesto, factores tecnológicos y del personal necesario.

4.2. Arquitectura y almacenamiento de datos

Para las tecnologías de información emergentes como la Inteligencia de Negocios (BI), Big Data e Internet de las Cosas (IoT), los medios y estructuras de almacenamiento son una pieza clave para su funcionamiento y el aprovechamiento de la información generada. Una de las estructuras más populares es el Almacén de Datos o Data Warehouse (DW), que es un sistema que Bernabeu [15], define como "un sistema que recupera y consolida datos periódicamente desde sistemas fuente en un depósito de datos dimensional o normalizado". Así mismo, Bill Inmon lo define como "una colección de datos orientada, integrada, no volátil y variante en el tiempo que apoya a las decisiones de la gerencia" [16] mientras que Ralph Kimball lo define como "un sistema que extrae, limpia, se ajusta y entrega datos fuente en un almacén dimensional de datos y después ayuda e implementa consultas y análisis con el propósito de toma de decisiones" [17].

Un almacén de datos usualmente contiene datos históricos recolectados desde otros sistemas, combinándolos con datos procesados y condensados. Las características principales de un almacén de datos son:

- Orientado a un tema.
- Integrado.
- Persistente.
- No volátil.
- Variante en el tiempo.
- Contiene datos consolidados a detalle.

Es común, y generalmente recomendado, que un DW sea diseñado analizando el objetivo al que se desea atender para que los datos con los que se va a poblar sean identificados y preparados desde los procesos de extracción, que posteriormente se adecúen al formato específico dictaminado por las reglas de Transformación.

4.3. Arquitectura del almacén de datos

La configuración lógica y físicamente el DW es un punto fundamental en su desarrollo y operación. Las organizaciones y los equipos responsables deben de tomar en cuenta los factores que intervienen en el funcionamiento y primordialmente en el objetivo del DW. De acuerdo a Bernabeu [15] los Almacenes de Datos pueden desarrollarse generalmente considerando dos arquitecturas principales: la Arquitectura del flujo de datos y la Arquitectura del sistema. De forma general, la Arquitectura de Flujo de Datos es la configuración de los depósitos de datos que trabajan dentro del DW y muestra el mecanismo de flujo de los datos a través de los depósitos dentro del almacén.

La Arquitectura de Flujo de Datos es un elemento esencial en el desarrollo de un DW de las primeras cosas que se necesitan definir en el proceso de desarrollo y construcción de un Almacén de Datos ya que es en esta fase donde se determinan los componentes o entidades que constituirán el DW. El trabajo de Bernabeu [15] y Kimball [17] se proponen cuatro Arquitecturas de Flujo de Datos:

- La Arquitectura DDS simple, que está compuesta principalmente tres elementos, el Staging, un Depósito de Datos Dimensional y un proceso ETL.
- La Arquitectura NDS + DDS es una combinación del Staging, un proceso de ETL, un Depósito de Datos Normalizados y un Depósito de Datos Dimensional.
- La Arquitectura ODS + DDS, donde se sustituye el NDS por un Depósito de Datos Operacional, que contiene una versión actual de los datos, no almacena los históricos.
- La Arquitectura de un Almacén de Datos "Federado" (FDW) es una estructura que "consiste en diversos Almacenes de Datos con una capa de extracción de datos por encima de ellos".

Una vez que se ha determinado los depósitos de datos que se necesitan desarrollar, se puede diseñar el mecanismo ETL para poblarlos. 4.3.1. Depósitos de datos Los Depósitos de Datos consisten en el conjunto de Bases de Datos estructurados en el formato particular requerido para atender los objetivos de la solución informática y forman parte de los procesos del DW. Los depósitos de datos se pueden clasificar en diferentes tipos basados en su accesibilidad y en su formato de datos. Desde el punto de vista del Formato de Datos, los Depósitos de Datos se pueden clasificar en cuatro categorías:

- Un "Staging" o base de datos de "escalón".
- Un Depósito de Datos Normalizado (NDS).
- Un Depósito de Datos Operacional (ODS).
- Un Depósito de Datos Dimensional (DDS).

Para el desarrollo de este proyecto, se utilizó una Arquitectura NDS + DDS por lo que significa que el Almacén de Datos es una combinación de un Depósito "Staging", un Depósito de Datos Normalizado, un Depósito de Datos Dimensional y sistemas ETL para extraer, procesar y alojar los datos utilizados en la industria petrolera con el objetivo de asistir al proceso de Evaluación de Integridad de Ductos. El Big Data tiene un enorme potencial, pero conlleva varios retos técnicos, tecnológicos, legales y administrativos que las organizaciones tienen que tomar muy en serio. El uso eficiente del conjunto de entidades pertenecientes al Big Data tiene la capacidad de transformar la productividad, plusvalía, ventaja competitiva, conocimiento, nuevas líneas, campos de investigación, soluciones heurísticas, entre otras.

5. RESULTADOS

La industria petrolera nacional, así como la internacional, existe una constante preocupación por mantener "seguros" sus sistemas de distribución de hidrocarburos. Sin embargo, determinar la seguridad de un ducto no generalmente no se responde con un sí o un no, más bien se define por medio del análisis de los datos asociados al ducto y a los datos estadísticos y probabilísticos del comportamiento de operación y entorno de los sistemas. Por lo que el interés de este trabajo, se enfocó en la construcción de un modelo de Flujo de Datos a partir de las variables o datos que participan en el Algoritmo de Evaluación de Riesgo o Integridad de Ductos. Los datos de entrada se refieren a todos los datos asociados al ducto, como su diámetro, su espesor de pared, la profundidad en la que está enterrado, el producto que transporta, la probabilidad de daño por factores externos, etc. Éste modelo resultante se encarga de procesar los datos de entrada u origen para que sean procesados y transformados en un conjunto de valores de riesgo para cada segmento del ducto.

5.1. Generación de del Diccionario de Datos

Un aspecto fundamental para el desarrollo de la infraestructura de Big Data y de la aplicación DSS fue la generación del Diccionario de Datos. Este elemento es esencial ya que se documentan todos los datos que participan en el modelo para el que

es construido y durante su operación. En este documento se detallan los tipos, estructura, nombres, definiciones, así como la granularidad y atomicidad de cada uno de los datos.

5.2. Diseño e implementación del Almacén de datos

El Almacén de Datos (DW) es una parte fundamental y está compuesto, en su forma básica, por tres elementos principales: los Sistemas Fuente, los procesos de ETL y los sistemas de almacenamiento objetivo. En la Figura 2 se muestra el esquema inicial del DW para este trabajo, sin embargo, es un diseño que evolucionó a una arquitectura NDS + DDS.

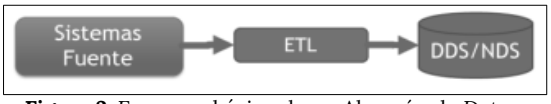

Figura 2. Esquema básico de un Almacén de Datos.

Los sistemas fuente en la mayoría de los casos no son controlables desde la perspectiva del diseño del DW, sino que se preparan los mecanismos de extracción para ajustarse a las necesidades de los sistemas de almacenamiento objetivo. En este caso, el diseño se ajusta a las entradas del Algoritmo de Evaluación de Integridad, independientemente del formato de la fuente de datos. Por esta razón, se construyó el diseño un sistema objetivo con aproximadamente quinientas entradas distribuidas en diferentes tablas de datos.

De acuerdo a la literatura, el personal técnico responsable del diseño y construcción del DW debe de considerar algunos puntos claves o etapas para considerar los factores alrededor de la Solución. En la Figura 3 se muestra el Proceso que se siguió para la Implementación de DW.

La Selección de Indicadores es una actividad en donde se analizaron y seleccionaron los datos considerados de "Salida". Para este trabajo, estos datos representan los datos que son utilizados y generados por el Algoritmo de Evaluación de Riesgo, por lo que los indicadores atienden esta necesidad particular y por lo tanto deben de estar todos los datos necesarios y en el formato adecuado para que Algoritmo pueda "alimentarse".

5.3. Arquitectura NDS + DDS

Para que el DW sea operacional, primero se definió la arquitectura con la que se va a construir ya que es un elemento responsable del correcto almacenamiento de los datos según el objetivo para el que el DW es diseñado. En la Figura 4 se muestra la arquitectura completa del DW.

En este tipo de arquitectura, no sólo trabajan los Depósitos de Datos, sino que para que se necesitan más elementos intermedios que colaboren y cumplan funciones independientes. El primer elemento diseñado fue el DDS, que es la base de datos que se considera va a ser utilizada para el análisis de datos y desde la cual se considera que el Algoritmo de Evaluación de Riesgo será "alimentado". El Sistema DDS deberá almacenar los datos con las especificaciones de es te Algoritmo por lo que es el último que recibirá datos, pero es el primero en ser diseñado y construido. El segundo elemento que fue construido fue el NDS, que es la base de datos destinada a almacenar principalmente los datos generales e históricos de forma normalizada, lo que significa que deberá de reducirse totalmente la redundancia de los datos, por lo que se estructura en tablas relacionales independientes.

El elemento Staging o base de Datos de Escalonamiento consiste en un sistema que fue diseñado para almacenar los datos de forma temporal cuando éstos son extraídos desde los Sistemas Fuente. De esta forma, los sistemas de origen ocupados en una sola

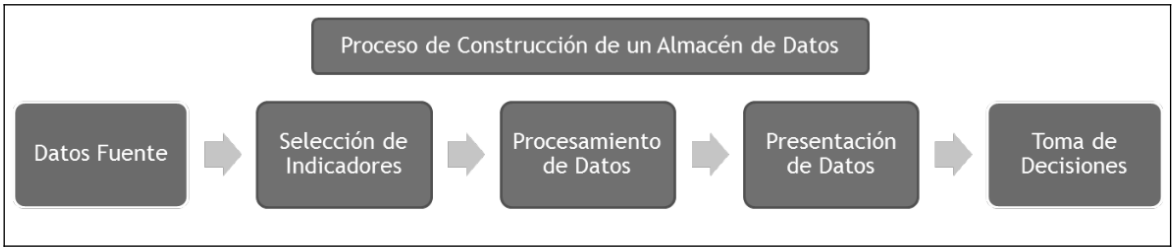

Figura 3. Proceso general para la construcción de un Almacén de Datos.

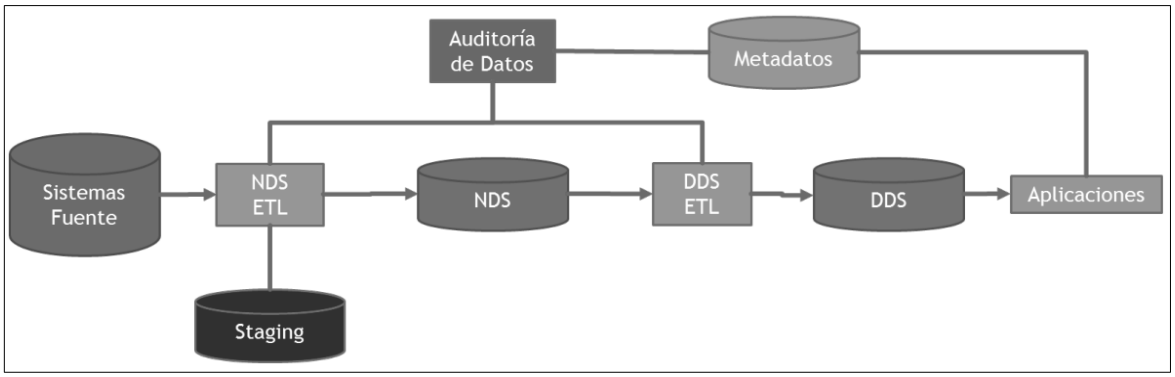

Figura 4. Diagrama de la Arquitectura NDS + DDS.

ocasión por extracción, liberando la carga de trabajo y desocupándolos para después procesar los datos de forma interna en el DW. Esta base de datos es relevante en este desarrollo, y en soluciones similares, ya que se encarga de recibir los datos en su forma más "cruda" o más directa desde los sistemas fuente, por lo que se convierte en el primer filtro de datos, a este mecanismo es el primer sistema de Extracción, que usualmente recolecta los datos de forma periódica y por lotes.

Una vez que los datos están almacenados en el Staging, se tienen que cargar en el NDS para su posterior utilización, este mecanismo intermediario se le denomina como "NDS ETL", cuyo objetivo es Extraer los datos desde el Staging para después procesarlos y cargarlos en el NDS. De esta forma, los datos de los sistemas fuente, ya han sido transformados para ajustarse a las necesidades iniciales del DW. Una vez que los datos están en el NDS, se utiliza otro mecanismo denominado como DDS ETL para extraer los datos desde el NDS y almacenarlos en el DDS, asegurándose en que estos datos reciban el procesamiento adecuado para cumplir las reglas del DDS. Las reglas y formatos en todos los mecanismos de transformación y limpieza son conformados dentro de las entidades como Metadatos y el Diccionario de Datos. La entidad de Auditoría de datos registra los que no cumplen con las reglas del Metadatos con el fin de que los usuarios estén enterados de estas variaciones.

5.4. Arquitectura del NDS

El NDS es un sistema de almacenamiento Normalizado, por lo cual adquiere una estructura relacional tradicional. En esta estructura, se definió una tabla central general que se encarga de almacenar los datos de los valores de riesgo de falla (RoF), mientras que se relaciona directamente con las tablas que registran los datos destinadas a la Probabilidad de Falla (PoF) y la Consecuencia de Falla (CoF). El NDS almacena los datos sobre las condiciones del ducto, así como los datos de reportes o estudios realizados al ducto, de forma permanente. En la Figura 5, se muestra un ejemplo de la estructura de la arquitectura del NDS.

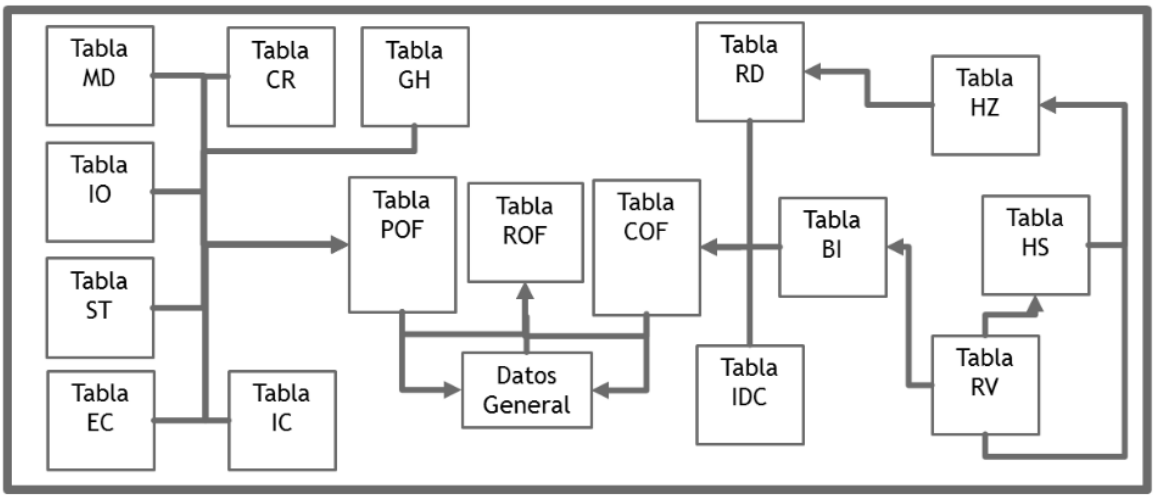

Figura 5. Esquema ejemplo de la arquitectura del NDS.

El NDS cuenta con doce tablas relacionadas con la Consecuencia de Falla y veinticuatro tablas relacionadas con la Probabilidad de Fallas, además de tres tablas que controlan los datos de la segmentación del ducto. Dando un total de cuarenta y dos tablas que conforman el Depósito completo.

5.5. Arquitectura del DDS

El Depósito Dimensional de Datos o DDS, es sistema de almacenamiento que tiene la característica de no ser normalizado, lo que significa que no cumple con las Normas de Normalización tradicionales de una base de datos relacional. Este depósito trabaja con tablas de "Hechos" y tablas "Dimensión", en vez de la relación Padre-Hijo. Cada tabla Dimensión posee datos específicos a cierto rubro y las tablas Hechos poseen datos que son procesados por medio de algún mecanismo, como sumatorias, promedios, agregaciones, entre otras cosas. Este tipo de estructuras pueden generar mucha redundancia, ya que los datos pueden repetirse en diferentes tablas Dimensión, sin embargo estos depósitos son utilizados particularmente para optimizar mecanismos analíticos y presentación de datos.

En este trabajo se utilizó un esquema tipo "Constelación" o "Galaxy", que se caracteriza por tener tablas con relaciones en diferentes subniveles, por ejemplo, la tabla de hechos de RoF almacenará datos sobre el cálculo del Riesgo, pero depende de los datos de la Tabla de hechos de PoF y CoF, que a su vez dependen de las tablas de dimensión de la Amenazas y los Impactos, y estas dependerán de las tablas dimensión en subniveles que almacenan datos más específicos. Es decir que hay tablas de hechos que no dependen inmediatamente de otras tablas dimensión, sino dependen de otras tablas de hechos. En la Figura 6 se muestra un ejemplo de la estructura de la arquitectura del DDS.

5.6. Staging

Ya que se trata de un sistema de almacenamiento temporal que deposita los datos que han sido extraídos de los Sistemas Fuente para procesarlos de forma local, antes de ser cargados a los demás sistemas de almacenamiento. Su estructura general es tipo estrella, donde las tabla central relaciona los datos del segmento del ducto y las tablas a su alrededor relacionan los datos asociados a cada elemento del Algoritmo de Evaluación de Riesgo.

Procesos de Extracción, Transformación y Carga de Datos

En cada nivel de procesamiento, los datos se deben ajustan a las necesidades de cada

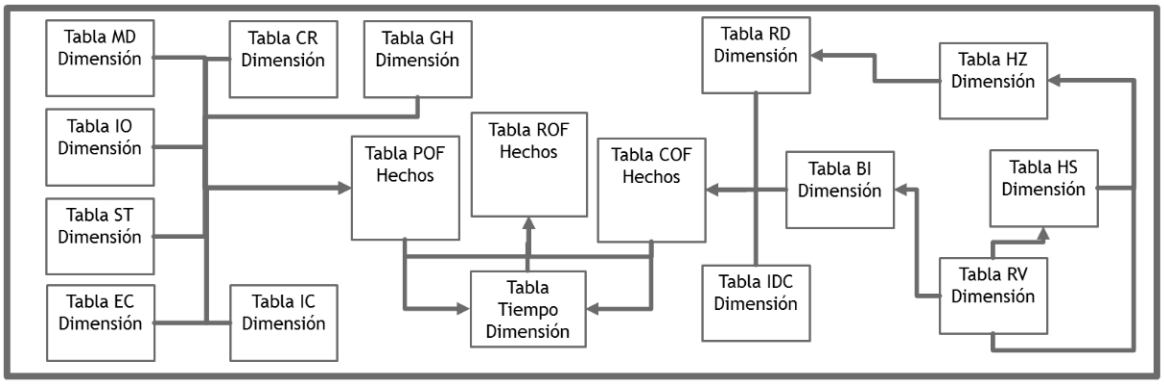

Figura 6. Esquema ejemplo de la Arquitectura del DDS.

depósito. Para esto se realizaron mapeos. Los mapeos, básicamente son tablas que relacionan los registros de origen con los registro destino, así como los procesos o cambios que se necesitan para cumplir con las reglas del DW.

Los mapeos son relaciones fundamentales para describir el comportamiento de los datos entre los depósitos. Los mapeos se clasifican en los siguientes:

- Mapeo Sistemas Fuente – Staging. Éste relaciona los datos de origen con los del Staging. Esta relación de datos es muy importante ya que es pieza clave para vincular los datos provenientes de los sistemas externos con los formatos necesarios para el DW.
- Mapeo Staging – NDS. Este mapeo relaciona los datos que se encuentran en el Staging y el formato además de destino en NDS.
- Mapeo NDS – DDS. Este relaciona los datos que se encuentran distribuidos en el NDS para transmitirlos al DDS.

Uno de los objetivos de este proyecto de proveer un desarrollo que posea la ventaja de que se atiendan necesidades específicas del usuario para facilitar el proceso de toma de decisiones y la generalización del uso del resultado del proyecto en todas las divisiones o áreas de la industria. Así mismo, implementando técnicas y metodologías modernas que faciliten el acceso a la información.

6. CONCLUSIONES

La información se ha convertido en un activo fundamental de las organizaciones y con ella, han surgido numerosas herramientas para el control, transmisión, almacenamiento y análisis de datos, como el Big Data y los Sistemas de Soporte de Decisiones. En este trabajo se considera que los elementos propuestos contribuyen a la unificación de las entidades de la industria petrolera mexicana para optimizar el tratamiento y almacenamiento de datos para favorecer a la industria por medio del análisis y minería de datos con el fin de llevar a nuevos conocimientos y la respuesta pronta ante un potencial riesgo, evitando desastres antes de que éstos sucedan.

La utilización de la arquitectura NDS + DDS puede favorecer a la industria mexicana e internacional para homogeneizar sus procesos de evaluación y administración de riesgo y con esto, tener la oportunidad de aprovechar el conocimiento generado. De esta forma, la industria puede estar mejor capacitada para la realización de análisis complejos que les provean información valiosa para determinar estrategias y toma de decisiones.

Desde el punto de vista del Big Data, la principal aportación de este trabajo es el diseño y construcción de una arquitectura informática que combina Depósitos de Datos para asistir al proceso de Evaluación de Riesgo de Ductos en la industria petrolera, proveyendo la oportunidad para centralizar los datos de distintos ductos y sus administradores para centralizar los datos de forma unánime, además de la posibilidad de implementar nuevos procesos de análisis de datos. Facilitando la explotación de datos y el descubrimiento de conocimiento.

Se considera que es este proyecto abre áreas de oportunidad para implementar otras estructuras, arquitecturas y tecnologías para poder aprovechar de mejor manera el tratamiento de los datos y por ende, la explotación y utilización optimizada de la información para el bien de la industria y la sociedad mexicana.

La implementación de esta infraestructura de tratamiento de datos puede adaptarse a los diferentes modelos de Evaluación de Riesgo debido a que no sólo existe uno y de esta forma, diversificar los resultados de los análisis generados.

REFERENCIAS

[1] World Economic Forum. Big Data, Big Impact: New Possibilities for International Development [Internet]. Ginebra: World Economic Forum; 2012 [citado 2024 Jun 14]. Disponible en: https://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf.

[2] Lohr S. The Age of Big Data. The New York Times [Internet]. 2012 [citado 2014 Dic 8]. Disponible en: https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

[3] Liebowitz J. Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management. Boca Raton: Auerbach Publications; 2006.

[4] Michalewicz Z, Schmidt M, Chiriac C. Adaptive Business Intelligence. Berlin: Springer; 2006. doi: https://doi.org/10.1007/978-3-540-32929-9.

[5] Sabherwal R, Becerra I. Business Intelligence – Practices, Technologies, and Management. Hoboken: John Wiley & Sons; 2011. .

[6] Centre for Economics and Business Research. Data quality: Unlocking the value of big data. London: Centre for Economics and Business Research; 2012. p. 7-26.

[7] Mongolia M, Burn S. Decision Support System for Water Pipeline Renewal Prioritization. Journal of Information Technology in Construction. 2006; 11, p.237-256. Disponible en: https://www.itcon.org/paper/2006/18.

[8] Secretaría de Energía. Administración de la integridad de ductos de recolección y transporte de hidrocarburos. Norma Oficial Mexicana NOM-027-SESH-2010 [Internet]. Mexico City: Secretaría de Energía; 2010. Disponible en: https://vlex.com.mx/vid/integridad-ductos-recoleccion-hidrocarburos-222735622.

[9] Muhlbauer K. Enhanced Pipeline Risk Assessment, Part 1, Probability of Failure Assessments, Revision 2.1 [Internet]. 2006 [citado 2024 Jun 14]. Disponible en: https://www.pipelinerisk.com/downloads.

[10] Cosham A, Hopkins P. The assessment of Corrosion in Pipelines – Guidance in the Pipeline Defect Assessment Manual (PDAM). Presentado en el International Colloquium 'Reliability of High Pressure Steel Pipelines'; 2003; Prague, Czech Republic.

[11] Hopkins P, Goodfellow G, Ellis R, Haswell J, Jackson N. Pipeline Risk Assessment: New Guidelines. Presentado en: WTIA/APIA Welded Pipeline Symposium; 2009; Australia.

[12] Slack E. What is Big Data? [Internet]. 2012 [citado 2014 Dic 8]. Disponible en: http://www.storageswitzerland.com/Articles/Entries/2012/8/3_What_is_Big_Data.html.

[13] Dumbill E. A Revolution that will transform how we live, work and think. Big Data [Internet]. 2013 [citado 2024 Jun 14];1(2).

[14] Almeida F. The main challenges and issues of big data management. Porto: Facultad de Ingeniería de la Universidad de Porto – Calistru Catalin, Centro de Innovación y Desarrollo, ISPGaya; 2012.

[15] Bernabeu R. Hefesto v 1.1. Córdoba, Argentina; 2009.

[16] Inmon WH. Building The Data Warehouse. Hoboken: Wiley; 2005.

[17] Kimball R, Caserta J. The Data Warehouse ETL Toolkit. Hoboken: Wiley; 2004.