Comparacion entre Data Warehouse y Data Lake

Willian Huillca Umpiri, Juan Carlos Panty, and Adnner Esperilla Ruiz
Universidad Privada de Tacna \Facultad de Ingenieria \Escuela Profesional de Ingenieria de Sistemas

Resumen

Esta investigación de corte teórico plantea una reflexión en torno a la necesidad de contemplar el empleo del storytelling como técnica comunicativa para presentar los contenidos en el ámbito del Periodismo de datos, frente a la tendencia académica mayoritaria centrada en aspectos exclusivamente técnicos. En esa línea, se promueve un abordaje crítico en torno al triángulo que conforman el Periodismo de datos, el storytelling y el Periodismo narrativo, al tiempo que se establece una revisión de aquellos géneros periodísticos que, de acuerdo con el criterio de los autores, pueden encajar mejor en dicha combinación. Por tanto, el objetivo central es el de establecer un nuevo modelo para acompañar a líneas de tiempo, infografías, gráficos y mapas interactivos como recursos habituales del Periodismo de datos de contenidos textuales que complementen, refuercen y contribuyan a la inteligibilidad y el atractivo de esos elementos.
Palabras clave: Storytelling; Periodismo de datos; Periodismo narrativo.

Abstract

This theoretical research raises a reflection on the need to contemplate the use of storytelling as a technique communication to present the contents in the field of Data journalism, facing the majority academic trend focused on exclusively technical aspects. In that line, it promotes a critical approach around the triangle that make up the Data journalism, storytelling and narrative journalism, at the same time that a review of those journalistic genres that, according to the authors' criteria, they may fit better in said combination. Therefore, the central objective is to establish a new model to accompany timelines, infographics, interactive graphics and maps as usual resources of Journalism of textual content data that complement, reinforce and contribute to the intelligibility and attractiveness of those elements.

Keywords: Storytelling; Data journalism; Narrative journalism.

I. INTRODUCCIÓN

En las empresas donde se va trabajar con cantidades de datos empieza a ser la norma más que la excepción era necesario buscar una solución más eficiente de almacenar y procesar grandes volúmenes de información. Los modelos tradicionales corporativos de Business Intelligence (BI) han operado a través del almacenamiento de la información crítica para el negocio en los Data Warehouse corporativos (Enterprise DW o EDW). Estas soluciones han sido, y son, formas eficientes de manejar volúmenes relativamente altos de datos en un formato estructurado, en filas y columnas organizados en tablas, con un esquema relacional.

El término Data Warehouse fue acuñado en un artículo de diario de Sistemas de IBM de 1988 en el que analizaba las necesidades de las organizaciones de mantener sus bases de datos operacionales así como la necesidad de proporcionar capacidades de acceso y análisis a los usuarios finales.

Generalmente los departamentos de Business Intelligence (BI) creaban subconjuntos de la información contenida en el DW para hacerla accesible a ciertas partes del negocio en respuesta a las necesidades específicas de información de sus procesos de negocio. A esta estructura de datos ad hoc se la denomina Data Mart. Es habitual que una empresa cuente con diferentes Data Marts para diferentes unidades de negocio con distintas estructuras de datos. De esta manera se hacia accesible en un formato más eficiente la explotación de los datos de interés para cada unidad de negocio.

Pero en los últimos tiempos se ha incorporado un nuevo concepto, nos referimos al Data Lake. Este se centra más en integrar, gestionar y distribuir todos los datos en el mínimo tiempo que sea posible. A continuación, te mostramos las diferencias entre Date Lake vs Data Warehouse.

Un Data Lake conserva todos los datos, no sólo los que podrían utilizarse actualmente, sino también aquellos que podrían necesitarse en un futuro. Por otra parte, está el Data Warehouse que estudia muy bien qué datos incluir, cuáles son las fuentes de los datos. Además, se necesita dedicar tiempo para entender el negocio y así seleccionar y perfilar los datos necesarios. El Data Warehouse al final, contiene un modelo de datos altamente estructurado, diseñado para la generación de informes. El Data Lake utiliza un hardware muy diferente al del Data Warehouse. En el Data Lake, se almacenan datos tanto estructurados como no estructurados y la ampliación a terabytes y petabytes es mucho más económico que en el caso del Data Warehouse. Es por eso, que en este último se mira tanto qué datos son necesarios para conservar, y cuales

eliminar, ya que supone un costoso almacenamiento.

II. OBJETIVOS

A. General:

El objetivo principal es liderar el cambio dentro de un mercado en continuo y repentino crecimiento, a través de la adopción de nuevas tecnologías que incluyen Big Data, procesos de Inteligencia Artificial, Machine Learning y los cada vez más solicitados Data Lakes empresariales. Durante mucho tiempo el Data Warehouse se ha ocupado de facilitar el acceso a dichos datos, aunque se originen de diferentes fuentes, y también de convertir esos datos en información relevante. Dicha información ha sido vital para analizar el mercado, los clientes y la competencia. A partir de ahí, se han podido tomar las decisiones más adecuadas.

III. MARCO TEÓRICO

La narración de datos es una metodología para comunicar información, adaptada a una audiencia específica, con una narrativa convincente. Son los últimos diez pies de su análisis de datos y posiblemente el aspecto más importante.

A. Que es Data Lake

El Data Lake y el Data Warehouse son similares a primera vista: ambos sistemas están diseñados para permitir el almacenamiento de una gran cantidad de datos.

Para entender mejor las diferencias entre un Data Lake y un Data Warehouse, necesitamos examinar más de cerca sus utilidades principales.

El Data Lake es un repositorio compartido que le permite adquirir y almacenar grandes cantidades de datos procedentes de sistemas heterogéneos en formato nativo, es decir, datos en bruto estructurados, semiestructurados y no estructurados. La adquisición puede provenir de sistemas heredados, como CRM y ERP, o de fuentes externas, como feeds, Internet de las Cosas y datos de redes sociales.

El propósito de un Data Lake es, por lo tanto, proporcionar una visión no necesariamente refinada de los datos para apoyar las actividades de Data Discovery, lo que lo convierte en un sistema adecuado para usuarios expertos.

Por el contrario, el Data Warehouse está destinado a lograr una visión única de la empresa a través de herramientas de análisis empresarial y de análisis de Big Data. Una visión controlada y certificada a través de procesos especiales de ingesta, destinados a almacenar únicamente los datos procesados para un propósito o proceso de negocio específico.Una de las principales fortalezas de los Data Lakes es la capacidad de almacenar cualquier tipo de datos. Esta característica es aún más evidente cuando los datos se adquieren con una frecuencia horaria o diaria, a través de estructuras en árbol (pensemos en una estructura de sistema de archivos en carpetas y subcarpetas organizadas por año, mes, día y, si es necesario, hora). En un Data Lake, la historización y posterior recuperación de los datos se puede llevar a cabo sin pérdida alguna del rendimiento, a diferencia de lo que podría ocurrir con los Data Warehouses para una enorme cantidad de datos.

B. Elementos Claves

- El llamado Data Storytelling no es más que un enfoque estructurado sobre cómo comunicamos insights a partir de los datos, e involucra una combinación de tres elementos: datos, visualización y narrativa.
  Narrativa + Datos = podremos explicar qué ha pasado y por qué un insight puede ser importante. Necesitaremos contexto para entender las conclusiones por completo.
  Visualización + Datos = Enlighten. Cuando añadimos una visualización a nuestros datos, podemos iluminar a nuestra audiencia con insights que no habrían visto de otra manera.
  Narrativa + Visualización = Engagement. La combinación perfecta para lograr ese interés e incluso para entretener a nuestra audiencia.

C. Principales diferencias entre Data Lake y Data Warehouse

- Las características específicas que distinguen a un Data Lake de un sistema tradicional de Data Warehouse son muchas, dependiendo del tipo de datos adquiridos y de la estructura de los mismos. A continuación, resumiremos las principales diferencias y analizaremos las más importantes..[**?**]

	DATA LAKE	DATA WAREHOUSE
Estructura de los datos	*Brutos (estructurados, semiestructurados y no estructurados)*	*Estructurados, procesados*
Finalidad de los datos	*Por definir, definida* ***Nota: Es posible que haya datos cuyo propósito no se haya definido (para uso futuro)***	*Definida*
Esquema	*On Read*	*On Write*
Usuarios	*Data Scientists*	*Usuarios empresariales*
Accessibilidad	*Gran accesibilidad y fácil actualización*	*Acceso y actualizaciones más complicadas y costosas*
Almacenamiento	*Almacenamiento distribuido y costes limitados (potencialmente ampliable a la nube)*	*Costes y revisión de costosos procesos de ingesta*

Figura 1: Modelo conceptual. Izquierda: enfoque analítico visual basado en datos convencional. Derecha: diseño orientado al aprendizaje, narración de datos enfoque para apoyar la creación de sentido.

Desde un punto de vista técnico quizá sea una definición muy vaga, sin embargo resulta muy intuitiva para entender algunos de los puntos clave para diferenciar un Data Lake de un Data Warehouse (y por supuesto un Data Mart).

En primer lugar la información llega al Data Lake tal y como viene de la fuente original (raw data), sin procesos intermedios de transformación. Esta filosofía implica la segunda característica del Data Lake, su capacidad para recoger los datos de diversas fuentes sin preocuparse de la estructura o la ausencia de estructura del dato que le llega, se lo traga todo por decirlo de alguna manera.

Otra característica de los Data Lakes es su flexibilidad ya que los datos están en formato raw mientras que un Data Warehouse ha realizado un proceso de transformación y adaptación (ETL) a una determinada estructura antes de guardar los datos. Este es un punto fundamental en la diferencia entre los modelos de un DW y un Data Lake.

Tradicionalmente los Data Warehouse han trabajado en una arquitectura denominada schema-on-write, el fundamento detrás del modelo ETL de la carga de datos en un Data Warehouse. Este modelo obliga a la empresa a definir un modelo de datos y crear un marco analítico previo a la carga de ningún dato, es decir, necesitamos definir que vamos a querer hacer con los datos antes de cargarlos en la base de datos. Evidentemente la definición no es inamovible pero el esfuerzo en tiempo y dinero para cambiar el esquema de un Data Warehouse es mucho mayor.

La filosofía de la ingesta de datos de un Data Lake se basa en otro modelo de arquitectura denominado schema-on-read. En esta alternativa se sigue otra secuencia diferente a la anterior, es decir, en lugar de marcar la estructura de los datos en la entrada a la base de datos es cuando se quieren usar los datos cuando se aplica el proceso de transformación de los datos.

Precisamente al estar el dato en formato bruto (raw data) tenemos la posibilidad de poder adaptarnos para prácticamente cualquier proceso analítico. De esta manera podemos dar respuesta a las necesidades de un usuario típico de negocio a la vez que le damos solución a las mucho más complejas y exigentes necesidades de un científico de datos (Data Scientist).

D. Ingesta de Datos

- En esta parte del proceso se definen las reglas mediante las cuales se va a realizar la transferencia de datos a un Data Lake. Una solución de gestión de un Data Lake debe proporcionar control sobre como los datos son ingestados en función de su fuente de origen, el momento en que llegan y donde queremos guardarlos en el Data Lake.

  Generalmente la ingestión de datos se realiza en una tabla gigante que se organiza mediante etiquetas de metadatos. Cada dato pieza de dato que llega se aloja en una celda de esa enorme tabla sin importar donde está esa celda, de donde viene el dato o su formato.

  Normalmente es en esta fase donde se aplica cierta lógica de Gobierno del Dato. Siguiendo la lógica de la arquitectura de un Data Lake los datos son alojados previamente en un área de staging antes de decidir incorporarlos al Data Lake.

E. ¿Almacenamiento y Retención de datos?

- Según parece en el modelo del DW grandes cantidades de espacio de almacenamiento pueden desperdiciarse debido a un problema denominado sparce table, o traducido de andar por casa, tablas poco densas. La cuestión es que en un Data Warehouse si tenemos que almacenar una tabla que combina datos de dos fuentes diferentes., una que tenga 200 filas y la otra con 400 campos. Para ser capaz de combinarlas tendríamos que añadir 400 columnas en la tabla original de 200 filas. Las filas originales no contendrían datos para esas nuevas columnas y las filas de la segunda fuente no tendrían datos de las 200 columnas originales. El resultado: un montón de celdas vacías. En un Data Lake, al no necesitar tratar los datos en la ingesta, cada dato ocuparía su celda y no habría desperdicio de espacio. Además al separar el almacenamiento del procesado se puede pagar por espacio de almacenamiento a un coste inferior.

F. ¿Procesamiento de los datos?

- Airbnb

  Al guardar los datos tal y como le llegar el usuario de los datos en un Data Lake tiene que procesarlos cuando quiere acceder a ellos. Este es la clave de la flexibilidad de este tipo de almacenamiento de datos, cada usuario puede aplicar su estandarización y transformación de los datos según sus necesidades.

  Perfectamente se podría realizar sobre el Data Lake un proceso de ETL para cargar los datos que se precisen en un Data Warehouse y proporcionar acceso a la información en un modelo combinado de almacenamiento en el Data Lake y consumo en un Data Mart a partir del Data Warehouse.

IV. ANÁLISIS

- Comprensión ágil de la información: las representaciones gráficas permiten ver grandes cantidades de datos de forma clara y coherente, lo que facilita la extracción de conclusiones e insights.

- Crear un nuevo lenguaje de negocio para contar la historia a otros: una vez que hemos descubierto nuevos insights, el siguiente paso es comunicarlos a través de gráficos simples o visualizaciones elaboradas para lograr engagement.

- Encontrar relaciones y patrones dentro de los activos digitales: descubrir tendencias dentro de los datos nos puede dar una ventaja competitiva, como detectar puntos clave que están afectando a la calidad del producto o solucionar problemas antes que se vuelvan más complicados.

V. CONCLUSIONES

- En conclusión, el Data Storytelling empleado de manera correcta, mejorará la comunicación entre grupos dentro de una organización, ya que se logrará transmitir las metas y objetivos claros; pero adicionalmente, es importante contar con las personas con las habilidades de comunicación indicadas y poder interpretar los datos encontrados. Esta es una habilidad que debemos desarrollar todos nosotros ya que en el futuro será un requisito mínimo requerido por todas las empresas que buscan candidatos potenciales para formar parte de equipo de colaboradores.

Los datos de cualquier organización son una propiedad muy valiosa. La seguridad de los datos confidenciales es siempre un gran desafío para una organización en cualquier nivel. Las bases de datos son un objetivo favorito de los atacantes debido a sus datos. Hay muchas maneras en que una base de datos puede ser comprometida. Existen varios tipos de ataques y amenazas a partir de los cuales se debe proteger una base de datos. Para asegurar los datos que consideraciones debemos tener en cuenta, se mencionan en este documento y todas las técnicas que se utilizan recientemente para la seguridad de la base de datos.

Para persuadir al cliente lo mejor es afectar a sus emociones, y esto se consigue mediante la narrativa en la visualización de datos. Además nos asegura la atención de los receptores, pues toda historia bien construida con un principio y un desarrollo, hace que inconscientemente el público desee saber la conclusión del relato.