Técnicas de unificación de datos para la visualización de grandes volúmenes de datos

Lilia Palomo, Norma Lesca y Laura Sánchez Piccardi
Facultad de Matemática Aplicada - Universidad Católica de Santiago del Estero
lilia.palomo@ucse.edu.ar – norma.lesca@gmail.com – lsanchezpiccardi@gmail.com

RESUMEN

En la actualidad, la incorporación de las Tecnologías de la Información y la Comunicación (TIC) en el ámbito de empresas y organizaciones como un elemento clave para mejorar su competitividad e impulsar su crecimiento económico, ha producido un crecimiento considerable en el volumen de datos generados por diferentes sistemas y actividades, lo que hace dificultoso mantenerse al día con los resultados y genera la necesidad de modificar, optimizar y desarrollar métodos y modelos de almacenamiento y tratamiento de datos que suplan las falencias que presentan las bases de datos y los sistemas de gestión de datos tradicionales.

Aunque, se pueda contar con grandes repositorios de datos flexibles y multi-estructurados, que están disponibles a costos bajos o gratuitos y listos para ser explotados por procesos informacionales, la gestión de múltiples tipos de datos (incluidos los datos estructurados, semi-estructurados y no estructurados) se vuelve compleja, por la variedad o existencia de diferentes tipos y fuentes de datos. A lo que se suma, la necesidad de las organizaciones de integrar y analizar datos en un complejo abanico de fuentes de información tradicional y no tradicional, que son producidos tanto internamente como por fuera de la empresa.

El proceso de entrega (*delivery*) de datos para la toma de decisiones, continúa desempeñando un papel de innovación y, por lo general, se confunde con extracciones manuales de datos, repeticiones costosas de procesos, informes propensos a errores y una integridad de datos no mensurable. Pero en realidad facilita la comprensión de los datos al transformarlos en información útil y ayuda a las organizaciones, a responder preguntas esenciales para la toma de decisiones que le permitan obtener ventajas competitivas y mejorar su posición en el mercado.

En este orden de ideas, cabe agregar como bien sostienen Jerry Held, Michael Stonebraker, Thomas H. Davenport, Ihab Ilyas, Michael L. Brodie, Andy Palmer, y James Markarian, 2016, que "la unificación de datos es una estrategia emergente, que cataloga el conjunto de datos, combina los datos de toda la empresa y los publica, para facilitar su consumo". Es decir, se trata de tener el manejo y control de la información, a fin de tener asegurada una vista única de los datos, que provienen de fuentes funcionalmente distintas (bases corporativas, bases propias, sistemas externos, etc.), ya que los usuarios finales no tienen la necesidad de aprender a utilizar diferentes sistemas de acceso y manipulación de los datos. El uso de la unificación de datos como una estrategia *frontend* puede acelerar el suministro de datos altamente organizados en sistemas, como ETL (Extract, Transform and Load) y MDM (Master Data Management) y lagos de datos (data lake), aumentando el valor de estos sistemas y los conocimientos que permiten. [1] De lo expuesto precedentemente, se desprende el propósito del presente proyecto como una investigación exploratoria de las estrategias del proceso de unificación y delivery de grandes volúmenes de datos organizacionales.

Palabras clave: Unificación de datos, BigData, Integración de datos, Armonización de datos, Entrega de datos unificados.

CONTEXTO

El presente artículo presenta una línea de estudio, ante la necesidad de las organizaciones modernas de contar con un buen proceso de unificación de datos que les permita obtener el mayor provecho de ellos para mejorar la toma de decisiones.

Integra una línea del trabajo de investigación de cátedra, "Aproximación teórica de las estrategias de delivery de datos unificados del ámbito organizacional", que promueve la interacción vertical y horizontal, a partir de las asignaturas de Programación I, Estructuras de Datos, Análisis Numérico, Sistemas de Información e Ingeniería de Software, en el marco de la carrera de Ingeniería en Informática de la facultad de Ciencias para la Innovación y el Desarrollo perteneciente a la Universidad Católica de Santiago del Estero.

Posibilita a los docentes obtener resultados que puedan ser aplicados en las aulas con el objetivo de promover la innovación de los contenidos de las cátedras y de las prácticas profesionales.

1. INTRODUCCIÓN

En la era de Big Data todas las organizaciones, independientemente de su tamaño o industria, tienen dificultades para gestionar de manera efectiva grandes volúmenes de datos internos y externos. Cuando los datos se encuentran en diferentes bases de datos y otros repositorios, las empresas se ven obligadas a establecer múltiples integraciones e interfaces de usuario para construir una visión holística del rendimiento del negocio. El proceso de extracción e integración de datos se vuelve altamente complejo, y hasta puede tornarse completamente inmanejable debido a un ecosistema demasiado complicado, de herramientas de Business Intelligence (BI) superpuestas y sistemas de Planificación de Recursos Empresariales (ERP) dispares [2].

En consecuencia, las organizaciones no pueden construir una base sólida para la estrategia de datos a largo plazo sin arriesgarse a una reducción de la alineación entre TI y los objetivos de la línea de negocio. Los riesgos operativos y analíticos incluyen [3]:

- Rendimiento operacional reducido.
- Implementaciones de TI reactivas y costosas.
- Falta de nuevos conocimientos.
- Informes y análisis caóticos
- Baja flexibilidad y tiempo de reacción a los cambios en el negocio.
- Problemas de calidad de datos no resueltos.
- Incapacidad para satisfacer las necesidades del cliente.
- Datos incorrectos y decisiones costosas.

Actualmente, la visión de casi todas las grandes organizaciones es maximizar el uso de sus activos de información para generar una ventaja competitiva. Sin embargo, solo unas pocas organizaciones modernas, realmente pueden sacar el máximo provecho de sus datos. A lo que se suma una tensión constructiva, ya que los consumidores de datos exigen un autoservicio instantáneo y siempre activo.

Hasta ahora, gran parte del enfoque, se ha relacionado con el uso de tecnologías de almacenamiento y herramientas analíticas para lograr este objetivo. Sin embargo, una pieza del rompecabezas que frecuentemente se pasa por alto, es el aprovechamiento de nuevos métodos para administrar los datos que conectan los sistemas de almacenamiento con los usos posteriores, como el análisis, ya que sin datos completos y limpios, el análisis se vuelve incompleto, inexacto e incluso engañoso [4].

En el ámbito del BI, el sofisticado trabajo de diseño de datos posibilita un análisis profundo. Por otro lado desde el entorno del workflow del proceso del Big Data, el almacenamiento de datos se hace más barato y más escalable.

Para llegar a una buena visualización, es

necesario entregar datos desde el origen hasta el punto de consumo. En el medio de este contexto, los procesos anticuados (tradicionales) pueden estorbar, por cuanto los datos se extraen manualmente y la falta de conexión puede producir errores. Además, las restricciones de ancho de banda retrasan la entrega del resultado de la consulta y reducen la puntualidad. Las evaluaciones cualitativas se transforman en hechos cuantitativos.

Entonces, el proceso semi-elaborado también puede llevar a la comprensión, pero hay que hacer una limpieza de los datos que tienen una confianza más débil, tarea que involucra mucho esfuerzo mal dirigido y ad-hoc.

Un buen Business Intelligence, puede ser una buena idea e idealmente conduce a un buen resultado. Pero cuando la estrategia se basa en las opciones de almacenamiento de datos, el flujo de entrada está restringido, la visibilidad se reduce y, en el peor de los casos, los resultados del negocio están predeterminados.

El solo hecho de que se hayan incorporado os datos no significa que se puedan operar o interpretar por un analista: los datos se encuentran en silos funcionales en diferentes ubicaciones de almacenamiento, o en silos lógicos dentro de un lago de datos no tan fácilmente accesible. La realidad es que los datos en un almacén de datos no pueden abrirse paso fácilmente, completamente curados e integrados a los consumidores. Para cuando se vuelve utilizable, a menudo se ha vuelto obsoleto [5].

En medio de esta situación, se encuentra como solución, el delivery, que sería la entrega de datos conectados o interrelacionados. Es hora de unificar los datos para que los usuarios comerciales puedan obtener el mayor provecho de los mismos.

Según señala, Michael Collins (2017), la tecnología de "unificación de datos", aprovecha las técnicas de aprendizaje automático y es una reinvención de las capacidades tradicionales de gestión de datos, como las que se encuentran en los procesos de MDM y ETL, para cumplir los requisitos de la era de Big Data [6].

Las técnicas tradicionales de gestión de datos son adecuadas cuando los conjuntos de datos son estáticos y relativamente pocos, pero fracasan en entornos de gran volumen y complejidad. Esto se debe en gran medida a sus enfoques descendentes y basados en reglas, que a menudo requieren un esfuerzo manual significativo para construir y mantener.

La tecnología de unificación de datos invierte este modelo, centrándose en la conexión y el dominio de conjuntos de datos mediante el uso de aprendizaje automático guiado por humanos, que aprovecha las señales en los datos para determinar cómo debe integrarse. El uso de la automatización guiada por la inteligencia humana para integrar y dominar los conjuntos de datos genera beneficios sustanciales en cuanto a velocidad, escala y flexibilidad del modelo de datos, al tiempo que garantiza los más altos niveles de precisión y confianza en los resultados. En su nivel más fundamental, la unificación de datos trae la promesa del aprendizaje automático a la preparación de conjuntos de datos a escala.

Sobre la base de lo expuesto, la unificación de datos beneficia a las organizaciones que buscan maximizar la entrega de valor derivado de los datos. Específicamente, la unificación de datos:

- Reduce los esfuerzos ad-hoc y únicos de integración de datos.
- Acelera la velocidad para obtener una idea.
- Mejora la confianza del usuario en los datos (al establecer una fuente única de verdad).
- Introduce flexibilidad en los workflow de datos para adaptarse al cambio.

La unificación de datos establece un modelo organizacional escalable con el crecimiento de los datos, a diferencia de "la vieja manera", una alternativa manual ad-hoc, que no lo hace.

Con la unificación de datos las personas dentro de la organización tendrán más posibilidades

de integrar datos en sus proyectos, compartir resultados y mantener los datos actualizados. Este ciclo de datos disponibles es clave para la innovación e intercambio de conocimientos.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación del presente trabajo tiene como eje central el estudio y análisis de las técnicas de unificación de datos para la entrega y visualización de grandes volúmenes de datos.

Se pueden mencionar los siguientes supuestos que dan estructura a la temática o campo de estudio del proyecto:

- Existen capacidades técnicas reales y potenciales que permiten aprovechar el avance del sector de servicios de informática e información, por cuanto, en las últimas décadas, la Argentina se posicionó entre los más dinámicos de la región [7].
- El mercado general de Big Data en Argentina muestra una demanda escasa y de baja complejidad, explicado en parte por la falta de conocimiento sobre el tema, problemas institucionales y limitaciones de infraestructura [8].
- Existen tres tipos de problemas asociados al Big Data: los tecnológicos, que se relacionan con el almacenamiento, la seguridad y el análisis de los volúmenes crecientes de datos; los comerciales, que se corresponden al valor añadido generado; y los sociales, relacionados con la privacidad de la información personal [9].
- Desde un punto de vista académico, Big Data genera estos retos que se vinculan a su vez a tres cambios de paradigma: mayor importancia de la disponibilidad y acceso de los datos; aceptación de niveles de imprecisión y desorden en los datos; centrarse más en las correlaciones, en vez de buscar constantemente la causalidad [10].

Se trabajará en la tipificación de técnicas de unificación de grandes volúmenes de datos organizacionales para posibilitar su entrega y visualización.

3. RESULTADOS ESPERADOS

El objetivo de esta línea de investigación plantea realizar un estudio y análisis del proceso de unificación de datos como estrategia para la recopilación, integración, preparación y entrega de los datos organizacionales.

Para lograr ese objetivo, este trabajo se centrará en los objetivos específicos:

- Investigar los diferentes enfoques técnicos disponibles actualmente para lograr el estado final deseado de conjuntos de datos limpios, precisos y consolidados
- Identificar las limitaciones existentes en las soluciones tradicionales para unificar grandes volúmenes y variedades de datos.
- Desarrollar una aproximación teórica para la recopilación, integración y preparación de datos organizacionales que posibilite una entrega eficiente de los mismos.

Las actividades que se llevarán a cabo son las siguientes:

- Estudio del proceso de unificación de datos.
- Análisis de las metodologías de preparación de datos variables.
- Análisis y estudio de herramientas de software y hardware.
- Comparación de los recursos existentes en el mercado.
- Selección de estrategias para la combinación de unificación y preparación de grandes volúmenes de datos.
- Identificación de los métodos de entrega eficiente de datos.

Se espera que los resultados de esta investigación se incorporen a los contenidos

de las cátedras relacionadas y al espacio curricular correspondiente.

Los resultados esperados respecto a la formación de recursos humanos son hasta el momento la consolidación del grupo de investigación, la formación de nuevos investigadores y la motivación y entrenamiento en investigación de los estudiantes de grado.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo está conformado por tres docentes de la carrera de Ingeniería en Informática, 2 dos con dedicación simple y una semiexclusiva.

El grupo hace difusión y formación de recursos humanos desde las asignaturas: Programación I, Estructura de Datos y Análisis Numérico, Sistemas de Información e Ingeniería de Software.

Asimismo, se considera de gran interés la incorporación de becarios, para motivar a los alumnos de la carrera de Ingeniería en Informática a realizar su trabajo final de grado en el área de este proyecto.

5. BIBLIOGRAFIA

1] Jerry Held, Michael Stonebraker, Thomas H. Davenport, Ihab Ilyas, Michael L. Brodie, Andy Palmer, and James Markarian, 2016. "Getting Data Right Volume and Variety". Copyright © 2016 Tamr, Inc.. O'Reilly Media, Inc. Disponible en: https://www.tamr.com/landing-pages/getting-data-right

2] 1010DATA, 2016. "Data Unification". Copyright © 2016 1010data Inc. Disponible en: https://www.1010data.com/media/1316/1010data_whitepaper_data_unification.pdf

3] Daily Data News. "What is Data Unification?", 2017. Disponible en: https://dailydatanews.com/2017/12/04/data-unification

4] Andy Oram, 2017. "Agile Data Mastering". Copyright © 2018 O'Reilly Media. Disponible en: https://www.tamr.com/landing-pages/agile-data-mastering-report-3

5] Toph Whitmore, 2017. **"**Connected Data Delivery: Combining Data Unification and Data Preparation". Copyright © 2017 Blue Hill Research. Disponible en: https://www.tamr.com/whitepaper/connected-data-delivery-combining-data-unification-data-preparation/

6] Michael Collins, 2017. "Data Unification: A New Path to Digital Transformation". Disponible en: https://www.enterprisetech.com/2017/08/09/data-unification-new-path-digital-transformation/

7] BARLETTA, PEREIRA, ROBERT & YOGUEL, 2013. "Argentina: Dinámica reciente del sector de software y servicios informáticos", Revista CEPAL N° 110.

8] MALVICINO & YOGUEL, 2015. "Descubriendo Big Data en Argentina. Encuesta Digital 2014". AGRANDA 1era ed. 44va Jornadas de Informática (JAIIO).

9] NUNAN, DI-DOMENICO, 2013. "Market research and the ethics of big data". International journal of market research, v. 55, n. 4, pp. 505-520. Disponible en: http://dx.doi.org/10.2501/IJMR-2013-015

10] MAYER-SCHÖNBERGER, CUKIER 2013. "Big data. La revolución de los datos masivos". Madrid: Turner.