

# Data Discovery Paraguay

Ariel Hernán Landaida Duarte, Iván Ariel Caceres Cañete

Data Discovery

## 1 Business Intelligence

Business Intelligence está ganando popularidad en las organizaciones. Los sistemas de Business Intelligence son desarrollados para ayudar a las organizaciones a entender sus clientes, las operaciones y el rendimiento. Sistemas de BI actúan como unidades de medida para una negocio o una organización. Inteligencia deficiente conduce a la toma de decisiones pobres. BI tiene problemas como la calidad de los datos, que una organización tiene que superar, con el fin de mejorar su toma de decisiones. Se ha hecho un esfuerzo serio para entender la causa del análisis de negocios de mala calidad en los últimos años. La mayoría de las organizaciones y los analistas coinciden en que la principal razón para informes erróneos son los datos operativos que se utilizan para generar estos informes como datos lleno de errores, duplicidades e incoherencias. Este trabajo de investigación explora los conceptos de BI, las características de calidad de datos y los problemas de calidad de datos, tipos de herramientas de BI y mecanismos de entrega y las características principales de la arquitectura de BI.

La inteligencia de negocios (BI) se define como la capacidad de una organización para llevar a cabo todos sus procesos y capacidades y luego convertirlos en conocimiento, en última instancia, la obtención de información correcta para las personas adecuadas, en el momento adecuado, a través del canal adecuado. Esto produce grandes cantidades de información que pueden conducir al desarrollo de nuevas oportunidades para la organización. Cuando estas oportunidades han sido identificados y una estrategia se ha aplicado de manera efectiva, BI puede proporcionar a una organización una ventaja competitiva en el mercado, y la estabilidad en el largo plazo (dentro de su industria)[17]. Tecnologías de BI ofrecen vistas históricas, actuales y predictivas de las operaciones de negocios. Funciones comunes de tecnologías de BI reportan, procesamiento analítico en línea, análisis, minería de datos, minería de procesos, procesamiento de eventos complejos, gestión del rendimiento de negocio, el BI tiene como objetivo apoyar una mejor toma de decisiones empresariales, por lo tanto un sistema de BI puede ser llamado un sistema de apoyo a las decisiones (DSS). Herramientas de BI son un tipo de software de aplicación diseñada para recuperar, analizar y reportar datos. Las herramientas generalmente leen los datos que se han almacenado previamente, a menudo, aunque no necesariamente, en un almacén de datos. BI ha experimentado un alto crecimiento y las tecnologías de BI han ganado mucha popularidad. Según Hancock y Toren, 2006, BI es "un conjunto de conceptos, métodos y tecnologías para convertir los datos separados en una organización en i En un entorno de Business Intelligence, datos de diversas fuentes se extraen, transforman y se cargan (ETL) en un almacén de datos empresariales (EDW) y, a partir EDW, se utilizan para la generación de los informes a través de la organization. Procesos BI y sus diversas etapas son mostrados en la figura 1. Calidad de los datos juega un papel importante y crítico en el éxito de BI, ya que la mala calidad de datos puede afectar las decisiones de negocio en todos los niveles de la organización y esos datos de mala calidad también obstaculiza el crecimiento de la organización. Información útil para

mejorar el rendimiento del negocio". El hecho es que el negocio corre sobre los datos; actúa como combustible para el motor de la industria corporativa. Una empresa no puede entender sus clientes, proveedores, competidores o su propia gente, los procesos y resultados de la empresa sin buenos datos. Por lo tanto, los negocios y de TI deben trabajar juntos para asegurar que los datos de alta calidad.

#### TIPOS DE PROBLEMAS DE DATOS

Hay muchos tipos de problemas de datos que afectan al negocio y a sus procesos; por lo general se pueden dividir en dos categorías: datos defectuosos y datos inconsistentes. La mayoría de las empresas no toman las medidas necesarias para descubrir y tratar estos temas hasta que tenga un impacto negativo en su negocio.

##### A. Datos defectuosos

###### (i) Tipos de datos defectuosos

Hay muchos tipos de datos defectuosos. Datos inexactos, incompletos, no disponible o obsoletos se define como datos defectuosos. Por ejemplo, los datos introducidos manualmente es a menudo lleno de errores de ortografía, errores tipográficos, y tal vez falsificado. A veces la gente olvida rellenar todos los campos o introducir los datos correctos en los campos incorrectos. Datos defectuosos también pueden aparecer cuando un sistema se mueve desde una plataforma a otra, una aplicación vieja se sustituye por una nueva, o los datos se mueven de una aplicación a otra sobre una base automatizada y programada.

###### (ii) Corregir los datos defectuosos

El problema con los datos defectuosos es que es difícil encontrarlos una vez que se introducen en el sistema. La mejor manera es evitar que los datos defectuosos se introduzcan, para ello las empresas deben invertir en sistemas que validan y corrigen los datos en la fuente. Para solucionar cuando ya están en el sistema de datos defectuosos, la empresa debe invertir en herramientas de perfilado de datos y de limpieza para limpiar y validar conjuntos de datos antes de que se carguen en los almacenes de datos para su posterior uso en BI.

##### B. Datos incoherentes

(i) Datos externos La inconsistencia en los datos se produce en el tiempo como los datos están duplicados, ya sea o fuera de fecha. Por ejemplo, se degrada de datos de clientes con el tiempo como la gente se casa, el divorcio, morir, mover o cambiar sus nombres, por lo que, esto se traduce en datos inconsistentes.

(ii) Datos internos Internamente, las compañías de datos se fragmentan(sucursales) para adaptarse a sus necesidades, en gran parte debido a la forma en que se organizan y también porque carecen de la gestión de datos centralizada. Cada analista, departamento o división ve el mundo a través de su propia lente y por lo tanto define y modela los datos de manera diferente.

**CARACTERIZACIÓN DE CALIDAD DE LOS DATOS EN BI** Calidad de los datos puede ser medido y cuantificado de acuerdo a varios parámetros. El trabajo previo ofrece diferentes clasificaciones de las dimensiones de calidad de datos. Mediante el análisis de estas clasificaciones, es posible definir un conjunto básico de las dimensiones de calidad de datos incluyendo la precisión, exhaustividad, coherencia, oportunidad, interpretabilidad y accesibilidad, que representa las dimensiones consideradas por la mayoría de los autores.

##### A. BI-Dimensiones específicas de calidad de datos

**Compleitud:** se ocupa de asegurar "¿Está toda la información necesaria disponible? ¿Están faltando algunos valores de datos, o están en un estado inservible? Por ejemplo, el nombre de un cliente y apellido son obligatorios, pero el segundo nombre es opcional; por lo que un registro puede considerarse completa, incluso si un segundo nombre no está disponible. Las preguntas que usted puede preguntarse: ¿Está toda la información necesaria disponible? ¿Algunos valores de los datos tienen elementos faltantes? ¿O están en un estado inutilizable?

Consistencia: Coherencia significa que los datos en todos los sistemas refleja la misma información y están en sintonía con los demás en toda la empresa. Por ejemplos, el estado del empleado es despedido, pero el estado de pago está activo. Las preguntas que usted puede preguntarse: ¿Los datos son los mismos a través de los conjuntos de datos? ¿Hay alguna ocurrencia distinta de la misma instancia de datos que proporciona información contradictoria?

Conformidad: Conformidad significa que los datos siguen el conjunto de definiciones de datos estándar como tipo de datos, el tamaño y el formato. Por ejemplo, la fecha de nacimiento del cliente está en el formato "dd/mm/ yyyy" Las preguntas que usted puede preguntarse: ¿Los valores de los datos cumplen con los formatos especificados? Si es así, no todos los valores de los datos cumplen con los formatos? El mantenimiento de la conformidad con los formatos específicos es importante.

Precisión: ¿Los objetos de datos representan con precisión los valores del "mundo real" que se espera del modelo? Ejemplo: Dirección de un empleado en la base de datos de los empleados es la dirección real. La pregunta que usted puede preguntarse: ¿Hay errores de ortografía en los nombres de productos, personas, direcciones, e incluso datos desactualizados?

Integridad: Integridad significa validez de los datos a través de las relaciones y garantiza que todos los datos en una base de datos se pueden remontar y se conectan a otros datos. Por ejemplo, en una base de datos de clientes, debe haber un cliente válido, direcciones y la relación entre ellos. Si existe una relación de datos de direcciones sin un cliente los datos no son válidos y se considera un registro huérfano. Pregúntese: ¿Hay importantes datos faltantes en los vínculos de relación? La incapacidad para enlazar registros relacionados entre sí en realidad puede introducir la duplicación a través de sus sistemas.

Puntualidad: Puntualidad se refiere a si la información está disponible cuando se espera y necesita. Esto se refleja en: Las empresas que están obligados a publicar sus resultados trimestrales dentro de un marco de tiempo determinado. Sistemas de crédito comprueban en tiempo real sobre la actividad de la cuenta de tarjeta de crédito. La puntualidad depende de las expectativas del usuario. Disponibilidad de datos online podría ser necesaria para el sistema de asignación de habitaciones en un hotel, pero actualizar los datos cada noche podría ser perfectamente aceptable para un sistema de facturación.

Algunas otras características de calidad de datos son el Aseguramiento, Entrega, Apertura / Transparencia, Responsabilidad.(from reference [1])

## References

- [1] N Kumari. Business intelligence in a nutshell. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(4):969–975, 2013.