Pronóstico de la demanda en empresas retail

Técnica basada en Business Intelligence y Machine Learning

Raúl Benítez - Alberto Garcete
Tutores: PhD. Diego P. Pinto Roa - Ing. Aditardo Vázquez

Universidad Nacional de Asunción - Facultad Politécnica

Agosto 2018

Definición

El primero que acuñó el término fue Howard Dresner en 1989, quién fue consultor de Gartner Group. Dresner utilizó el término para describir un conjunto de conceptos y métodos que mejoran la toma de decisiones, partiendo de la información disponible acerca de los hechos

Objetivos

Según lo expuesto en la definición, Business Intelligence tiene los siguientes objetivos principales:

- Convertir datos en información, información en conocimiento, y el conocimiento en planes operativos o estratégicos.
- Facilitar la disponibilidad de información a los usuarios de negocios, que les ayude a tomar decisiones más rápidamente.
- Apoyar de forma sostenible y continua a las organizaciones, para mejorar su competitividad ante el entorno de negocios cambiante de forma que puedan adaptarse a él.
- Ante la cantidad de información que va creciendo, disponer de más tiempo para analizarla, en lugar de gastar mucho tiempo en prepararla, organizarla y estructurarla.
- Permitir a las organizaciones dirigir de mejor forma, decidir, medir, gestionar y optimizar el alcance de la eficiencia y los resultados financieros.
- Disminuir sustancialmente la incertidumbre que existe ante la toma de decisiones respecto a un plan estratégico.

Fuentes de Información

Las fuentes de información representan el origen de los datos con las cuales se alimenta de información al datawarehouse. Estas pueden provenir de diferentes sistemas como:

- Sistemas operacionales o transaccionales, que incluyen aplicaciones desarrolladas a medida, ERP, CRM, SCM, etc.
- Sistemas de información departamentales: previsiones, presupuestos, hojas de cálculo, etc.
- Fuentes de información externa, en algunos casos comprada a terceros. Las fuentes de información externas podrían ser importantes para enriquecer la información acerca de los clientes. En algunos casos es interesante incorporar información referente, por ejemplo, a población, número de habitantes, etc.

Fuentes de Información

También las fuentes de información son usualmente heterogéneas, pueden contener los siguientes tipos de datos:

- Estructurados: almacenados en las bases de datos.
- Semi estructurados: son formatos entendibles por los computadores como HTML tabulado, Excel, CSV u otros que pueden ser obtenidos mediante técnicas estándar de extracción de datos.
- No estructurados: son formatos no legibles para computadoras como Word, HTML no tabulado, PDF, etc. que pueden obtenerse mediante técnicas avanzadas de extracción de datos.

ETL - Extracción, Transformación y Carga

La extracción, transformación y carga, comúnmente abreviado por las siglas ETL (*del inglés "Extract, Transform and Load"*) es un tipo de integración de datos que consiste en todo el proceso que se realiza entre las fuentes de información y el área de presentación de los datos. Es utilizado para extraer los datos de los sistemas de origen, transformarlos en función a los requerimientos del negocio y cargar los datos en el entorno de destino.

ETL - Extracción, Transformación y Carga

Extracción

La extracción es el primer paso en el proceso de obtención de los datos, recupera los datos físicamente de las distintas fuentes de información. En este punto se dispone de los datos en bruto. El principal objetivo es extraer aquellos datos de los sistemas transaccionales que son necesarios y prepararlos para el resto de los subprocesos de ETL. Para ello se deben determinar las mejores fuentes de información, las de mejor calidad. Para tal finalidad, se debe analizar las fuentes disponibles y escoger aquellas que sean mejores

ETL - Extracción, Transformación y Carga

Limpieza

Este proceso recupera los datos en bruto y comprueba su calidad, elimina los duplicados y, cuando es posible, corrige los valores erróneos y completa los valores vacíos, es decir se transforman los datos -siempre que sea posible- para reducir los errores de carga. En este momento se disponen de datos limpios y de alta calidad. Los sistemas transaccionales contienen datos que no han sido depurados y que deben ser limpiados. Algunas causas que provocan que los datos estén "sucios" son:

- Valores por defecto, Ausencia de valor, Campos que tienen distintas utilidades, Valores contradictorios.
- Uso inapropiado de los campos, Re utilización de claves primarias, Selección del primer valor de una lista.
- Problemas de carga de antiguos sistemas o de integración entre sistemas.

ETL - Extracción, Transformación y Carga

Transformación

La transformación de los datos se realiza partiendo de los datos una vez limpios, se transforman los datos de acuerdo a las reglas y necesidades del negocio. El resultado de este proceso es la obtención de datos limpios, consistentes, sumarizados y útiles.
La transformación incluye:

- Cambios de formato.
- Sustitución de códigos.
- Valores derivados y agregados.

ETL - Extracción, Transformación y Carga

Integración

Este proceso valida que los datos que cargamos en el datawarehouse son consistentes con las definiciones y formatos del datawarehouse; los integra en los distintos modelos de las distintas áreas de negocio que hemos definido en el mismo. Estos procesos pueden ser complejos.

ETL - Extracción, Transformación y Carga

Actualización

Este proceso es el que nos permite añadir los nuevos datos al datawarehouse, determina la periodicidad con el que haremos nuevas cargas de datos al datawarehouse.

Datawarehouse

Bill Inmon definió las características que debe cumplir un datawarehouse:

- Orientado a un área: cada parte del datawarehouse está construida para resolver un problema de negocio. Por ejemplo: entender los hábitos de compra de clientes, analizar la calidad de los productos, analizar la productividad de una línea de fabricación.
- Integrado: la información debe ser transformada en medidas comunes, códigos comunes y formatos comunes para ser útil. Por ejemplo: la moneda en que están expresadas los importes es común.
- Indexado en el tiempo: se mantiene la información histórica. Ejemplo: analizar la evolución de las ventas en los periodos deseados.
- No volátil: los usuarios no la mantienen como lo harían en los entornos transaccionales. No se ve actualizado continuamente, sino periódicamente de forma preestablecida. La información se almacena para la toma de decisiones

Datawarehouse

Ralph Kimbal define los objetivos que debería cumplir un datawarehouse:

- El alcance de un datawarehouse puede ser a nivel de departamento o corporativo.
- El datawarehouse no es sólo información sino también las herramientas de consulta, análisis y presentación de la información.
- La información del datawarehouse es consistente.
- La calidad de información en el datawarehouse es el motor de business reengineering.

Datawarehouse

Se debe tener en cuenta que existen otros elementos en el contexto de un datawarehouse:

- Datawarehousing: es el proceso de extraer y filtrar datos de las operaciones, procedentes de los distintos sistemas de información operacionales y sistemas externos, para transformarlos, integrarlos y almacenarlos en un almacén de datos, con el fin de acceder a ellos para dar soporte al proceso de toma de decisiones.
- Data Mart: es un subconjunto de datos del datawarehouse cuyo objetivo es responder a un determinado análisis.
- Operational Data Store (ODS): es un tipo de almacén de datos que proporciona sólo los últimos valores de los datos y no su historial.

Datawarehouse

Elementos de un datawarehouse

- **Tabla de hecho**: es la representación en el datawarehouse de los procesos de negocios de la organización. A nivel de diseño es una tabla que permite guardar dos tipos de atributos diferenciados:
  - Medidas del proceso de trabajo que se pretende modelar.
  - Claves foráneas hacia registros de una tabla de dimensión.
- **Dimensión**: es la representación en el datawarehouse de una vista para un cierto proceso de negocio.
- **Métrica**: son los indicadores de negocio de un proceso de negocio. Aquellos conceptos cuantificables que permiten medir el proceso de negocio.

Datawarehouse

Tipos de esquemas

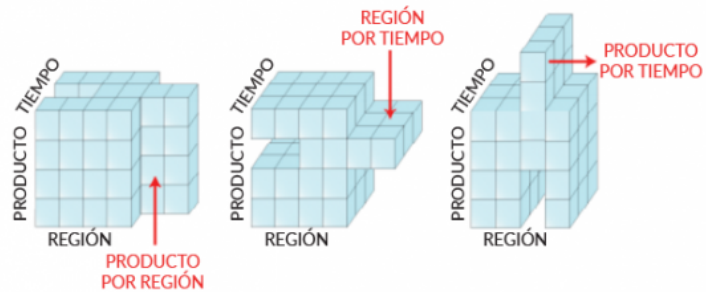
- **Esquema en estrella:** consiste en estructurar la información en procesos, vistas y métricas recordando a una estrella. A nivel de diseño, consiste en una tabla de hechos en el centro para el hecho objeto de análisis, y una o varias tablas de dimensión por cada punto de vista del análisis que participan en la descripción de ese hecho.
- **Esquema en copo de nieve:** es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas. Por esta razón la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas.

OLAP

Una definición formal de OLAP sería:
"*Se entiende por OLAP o proceso analítico en línea, al método ágil y flexible para organizar datos, especialmente metadatos, sobre un objeto o jerarquía de objetos como en un sistema u organización multidimensional, y cuyo objetivo es recuperar y manipular datos y combinaciones de los mismos a través de consultas o incluso informes*"

OLAP

Una representación gráfica del OLAP también conocida como cubo. En el ejemplo el cubo tiene 3 dimensiones (Tiempo, Producto y Región) sobre las cuales se pueden realizar consultas, ej.: Monto total de venta del producto “A” en el año “B” en la región “C”. Los cubos OLAP también permiten representar jerarquías, en el caso de la dimensión Tiempo la jerarquía podría estar compuesta por año, semestre, mes, semana y día.



OLAP

Existen distintos tipos de OLAP, las cuales difieren principalmente en la forma de guardar los datos:

- MOLAP (Multidimensional OLAP): es la forma tradicional del OLAP, accede directamente sobre una base de datos multidimensional, que utiliza estructuras de datos optimizadas para la recuperación de los mismos, es eficaz en los tiempos de respuestas de las consultas.
- ROLAP (Relational OLAP): accede directamente a las bases de datos relacionales que almacenan los datos base y las tablas dimensionales como tablas relacionadas.
- HOLAP (Hybrid OLAP): es una combinación de las dos anteriores, permite almacenar parte de los datos en una base de datos multidimensional y otra parte en una relacional. En la base de datos relacional se guardan grandes cantidades de información detallada, mientras que en la multidimensional se almacenan datos menos detallados o agregados.

Herramientas de BI

Las principales herramientas de Business Intelligence son:

- *Generadores de informes*: Utilizadas por desarrolladores profesionales para crear informes estándar para grupos, departamentos o la organización.
- *Herramientas de usuario final de consultas e informes*: Empleadas por usuarios finales para crear informes para ellos mismos o para otros; no requieren programación.
- *Herramientas OLAP*: Permiten a los usuarios finales tratar la información de forma multidimensional para explorarla desde distintas perspectivas y periodos de tiempo.

Herramientas de BI

- *Herramientas de Dashboard y Scorecard*: Permiten a los usuarios finales ver información crítica para el rendimiento con un simple vistazo utilizando iconos gráficos y con la posibilidad de analizar información detallada e informes.
- *Herramientas de planificación, modelización y consolidación*: Permite a los analistas y a los usuarios finales crear planes de negocio y simulaciones con la información de Business Intelligence. Pueden ser para elaborar la planificación, los presupuestos, las previsiones. Estas herramientas proveen a los dashboards y scorecards con los objetivos y umbrales de las métricas.
- *Herramientas datamining*: Permiten a estadísticos o analistas de negocio crear modelos estadísticos de las actividades de los negocios. Datamining es el proceso para descubrir e interpretar patrones desconocidos en la información mediante los cuales resolver problemas de negocios. Los usos más habituales del datamining son: segmentación, venta cruzada, sendas de consumo, clasificación, previsiones, optimizaciones, etc.