Introducción al *business intelligence* y *big data*

Generando valor a partir de los datos

Josep Curto

PID_00242521

Universitat Oberta de Catalunya

Índice

Introducción

En los últimos años las empresas se han embarcado en un proceso de transformación digital de profundo calado dentro del marco de lo que se conoce como la cuarta revolución industrial, que está dando paso a una nueva manera de organizar los medios de producción. Las empresas se están transformando en "fábricas inteligentes" capaces de una mayor adaptabilidad a las necesidades y a los procesos de producción, así como a una asignación más eficaz de los recursos. De este modo han abierto la vía a una nueva revolución industrial que se ha llamado también transformación digital. No se trata solo de la digitalización de los procesos a través de la automatización, sino del uso de la información, y la tecnologías de la información (TI) y las comunicaciones con el objetivo de aumentar el valor para el cliente y la ventaja competitiva de la empresa. Por lo que TI ha pasado de estar en la periferia de la organización a estar en el centro, erigiéndose en uno de sus pilares. Esta progresiva transformación de base tecnológica se ha combinado con otros aspectos, como el advenimiento de las redes sociales, la democratización de Internet o el despliegue del Internet de las Cosas.

Lectura complementaria

Schwab, K. (2016). *The Fourth Industrial Revolution*. Davos: World Economic Forum

Internet de las Cosas

Internet de las Cosas hace referencia a la interconexión digital de objetos cotidianos con Internet. Nos referiremos a él por su acrónimo en inglés IoT, *Internet of Things*.

El resultado de esta tormenta perfecta en la que se hallan todas las organizaciones es una explosión del dato en volumen, velocidad y variedad. Y de forma natural, ha crecido la complejidad para capturar, procesar, almacenar, analizar y visualizar los datos.

Como consecuencia, han aparecido múltiples métodos, técnicas y tecnologías que buscan ayudar a las organizaciones a tomar mejores decisiones a partir de los datos y a extraer valor de los mismos. Estos métodos, técnicas y tecnologías para la captura, el procesamiento, el almacenamiento, la gestión y el análisis se han ido estructurando progresivamente en diferentes estrategias que conocemos como ***business intelligence***, ***business analytics*** y ***big data***.

Sin embargo, a medida que estas estrategias se han hecho conocidas, las organizaciones las han ido implementado con menos fortuna de la esperada, tal y como apuntan los estudios de Aberdeen Group, Dresner Advisory Services o Harvard Business Review. Si las herramientas han ido madurando a lo largo de los últimos años, ¿cómo es que las organizaciones siguen teniendo tantos problemas en la implantación de este tipo de proyectos? Existen, por lo tanto, todavía múltiples preguntas para cualquier empresa:

- ¿Qué es *business intelligence*?
- ¿Qué es *business analytics*?
- ¿Qué es *big data*?

Referencias bibliográficas

Michael Lock (2012). *Managing the TCO of BI: The Path to ROI is Paved with Adoption. Aberdeen Group*.
Howard Dresner (2015). *Wisdom of Crowds Business Intelligence Market Study*. Dresner Advisory Services
Donald A. Marchand; Joe Peppard (2013). *Why IT Fumbles Analytics*. Harvard Business Review.

- ¿Qué significa para mi organización?
- ¿Cuándo es relevante?
- ¿Está preparada mi organización?
- ¿Cómo desplegar con éxito este tipo de iniciativas?
- ¿Qué barreras presentan este tipo de proyectos?
- ¿Qué tecnologías existen dentro de *business intelligence*, *business analytics* y *big data*?

Respondiendo a las anteriores preguntas, el presente material busca capacitar a estudiantes y profesionales en el contexto del análisis de la información con el objetivo de desarrollar estrategias de negocio que incluyan *business intelligence*, *business analytics* y *big data*, en el seno de su propia organización. Y en consecuencia, poder detectar en la propia organización casos de uso y problemáticas que necesiten este tipo de enfoque.

Objetivos

Este material didáctico está dirigido a:

- Desarrolladores y consultores que quieren conocer *business intelligence* y *big data*.
- Desarrolladores y consultores que quieren ayudar en el desarrollo de estrategias de negocio que incluyan *business intelligence* y *big data*.
- Gestores que están interesados en la transformación digital de su organización y en la inclusión de *business intelligence* y *big data* como uno de sus pilares fundamentales.
- Estudiantes de cualquier disciplina, en especial los de formación no tecnológica en el ámbito de la empresa.

Y tiene los siguientes objetivos:

1. Entender los conceptos de *business intelligence* y *big data*, las situaciones en las que es necesario desplegar una solución de este tipo y las ventajas que proporciona.
2. Promover la necesidad: ¿por qué es necesario tener una estrategia de negocio que incluya *business intelligence* y *big data*?
3. Presentar y discutir las tecnologías que engloban *business intelligence* y *big data*.
4. Dar a conocer casos de uso y ejemplos.

Si bien la obra es autocontenida en la medida de lo posible, se introducirán los conceptos necesarios para el seguimiento del material.

1. Toma de decisiones orientada al dato

La gestión de una organización se fundamenta en tomar decisiones adecuadas respecto a clientes, productos, empleados, proveedores y procesos de negocio en todos sus departamentos, desde finanzas hasta *marketing*. Por lo tanto, es necesario tener mecanismos que den soporte a una toma de decisiones eficiente.

En los últimos años, ha emergido una nueva forma de competir que se fundamenta en tomar decisiones basadas en datos y evidencias dejando atrás la intuición. Esta forma de competir combina diferentes estrategias para generar valor de negocio: ***business intelligence***, ***business analytics*** y ***big data***. No es extraño que los CIOs de las principales empresas del mundo destaquen por quinto año consecutivo que su principal prioridad tecnológica son este tipo de iniciativas*.

CIO

Cuando hablamos de CIO, hacemos referencia al *Chief Information Office*, responsable de las tecnologías de la información en una organización.

***Building the Digital Platform: Insights From the 2016 CIO Agenda Report.* Gartner.**

Así, la explotación de la información en el contexto de las organizaciones ha pasado de ser una necesidad más a ser la prioridad de máxima relevancia. El objetivo es poder tomar mejores y más rápidas e informadas decisiones de negocio. ¿Qué significa tomar mejores decisiones? Consideremos el siguiente ejemplo.

> En el contexto actual, conocer al cliente es primordial. Se busca comprender patrones y comportamiento del cliente. Interesa, por lo tanto, conocer aspectos como: ¿cuál es la probabilidad de que un cliente pague cada mes por el servicio contratado?, ¿quién compró qué productos?, ¿qué productos son los mejores (por región, por canal, etc.)?, ¿qué objetos se tienden a comprar juntos?, ¿qué otros productos podemos recomendar a nuestros clientes?
>
> Amazon, conocida empresa de *e-commerce*, hace recomendaciones en tiempo real en las que combina el histórico de ventas de sus clientes y las preferencias mostradas cuando buscamos un determinado producto. Esta iniciativa fundamentada en el dato permite incrementar los ingresos por cliente.

Muchas organizaciones aún no han desplegado este tipo de iniciativas y, entre aquellas que lo han hecho, no todas han logrado alcanzar el éxito esperado. Este tipo de iniciativas son complejas puesto que suponen una transformación de gran calado en la organización. No solo se trata de implementar un sistema de información sino de cambiar la forma en la que opera una organización en cada uno de sus departamentos. Por ello, en este material empezaremos hablando de la inteligencia de negocio.

El mercado de *business intelligence* existe desde hace bastantes años y ha evolucionado hacia soluciones con mayores prestaciones, y podemos considerar que ha alcanzado una significativa madurez. Destacamos, por ejemplo, que:

- Se ha producido una consolidación en el mercado, mediante la compra de empresas pequeñas por parte de los principales agentes del mercado, para complementar su propuesta de valor (entre ellas destacamos SAP, IBM, Microsoft u Oracle).

- Han aparecido nuevas empresas con foco en la innovación que cubren nuevos nichos en el mercado de la inteligencia de negocio, como la visualización, el análisis predictivo, las *virtual appliances*, es decir, soluciones que combinan *hardware* y *software*, y/o el *real-time business intelligence* (entre ellas destacamos Tableau, Qlikview, o Yellowfin).

- Las principales soluciones BI *open source*, a saber, Pentaho, JasperSoft y Actuate, han sido adquiridas por Hitachi Data Systems, Tibco y OpenText respectivamente.

- Hemos asistido a la aparición de una nueva generación de soluciones enfocadas a la generación de valor de conjuntos de datos complejos, lo que frecuentemente se refiere como *big data*, que expande el valor de la inteligencia de negocio.

- Se han empezado a consolidar las propuestas de inteligencia de negocio vinculadas con *Cloud Computing* (entre ellas destacamos GoodData, Sisence o Iberinform).

Open source

Cuando hablamos de *open source* (o código libre), hacemos referencia a programas informáticos cuyo código libre está disponible para todo el mundo para ser revisado, modificado y/o mejorado.

Algunas de las herramientas en el contexto de la inteligencia de negocio acumulan diversos años de desarrollo y evolución, y están respaldadas por organizaciones que tienen un claro modelo de negocio y que generan sinergias entre ellas en forma de ecosistemas. Podemos encontrar tanto herramientas de bases de datos como de minería de datos. Tal es la madurez de dichas soluciones que es posible desarrollar e implementar proyectos de inteligencia de negocio para todo tipo de organizaciones, tanto pymes como grandes organizaciones.

En este módulo, se presentan los conceptos y diferencias entre ***business intelligence***, ***business analytics*** y ***big data***, y sus principales usos en la empresa.

1.1. ¿Qué es el *business intelligence*?

El contexto de la sociedad de la información ha propiciado la necesidad de tener mejores, más rápidos y más eficientes métodos para extraer y transformar los datos de una organización en información y distribuirla a lo largo de la cadena de valor.

Sin embargo, esta necesidad, que actualmente se considera crítica en la gran mayoría de empresas, no es nueva. En octubre de 1958 Hans Peter Luhn, investigador de IBM, en el artículo *A Business Intelligence System* acuñó un término que responde a esta problemática como la habilidad de aprehender las relaciones de hechos presentados de forma que guíen las acciones hacia una meta deseada.

Cadena de valor empresarial

Cuando hablamos de la cadena de valor empresarial, descrita y popularizada por Michael E. Porter en su obra *Competitive Advantage: Creating and sustaining superior performance*, hacemos referencia a un modelo teórico que permite describir las actividades que generan valor en una organización.

No es hasta 1989 que Howard Dresden, en dicho momento analista de Gartner, propone una definición formal del concepto.

Conceptos y métodos para mejorar las decisiones de negocio mediante el uso de sistemas de soporte basados en hechos.

Desde entonces, el concepto del que estamos hablando ha evolucionado aunando bajo su paraguas diferentes tecnologías, metodologías y términos. Es, por lo tanto, necesario establecer una definición formal de uso en el presente material.

Se entiende por *business intelligence* el conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite a los usuarios de una organización tomar mejores decisiones.

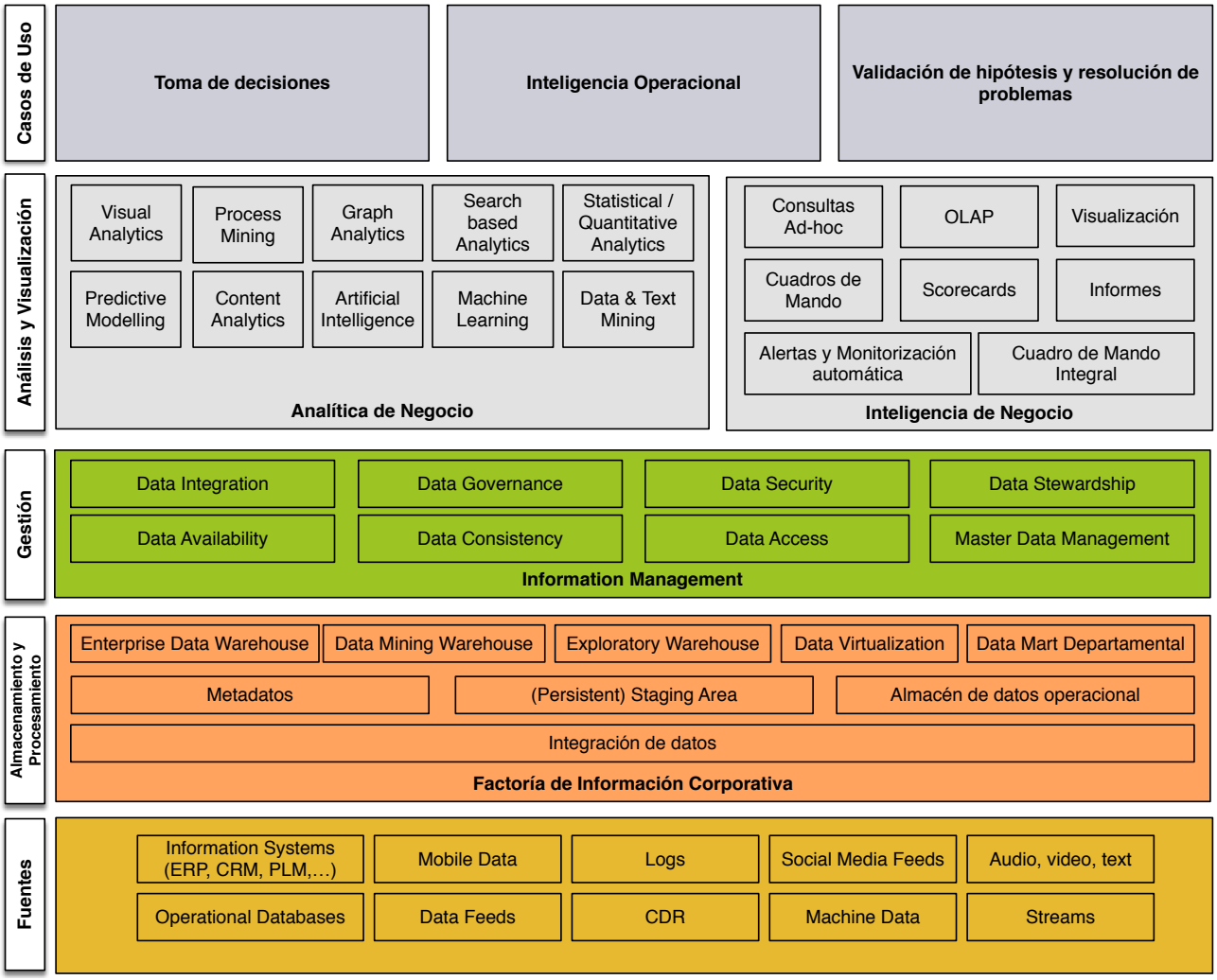
En esencia, mediante la inteligencia de negocio, podemos romper con la siguiente máxima:

> "Lo que no se define no se puede medir. Lo que no se mide no se puede mejorar. Lo que no se mejora, se degrada siempre."
>
> William Thomson

Dentro del contexto de la inteligencia de negocio se incluyen múltiples tecnologías. Algunas de ellas son: *Data warehouse*, *Reporting*, Análisis OLAP (*Online Analytical Processing*), Análisis Visual, Análisis predictivo, Cuadro de mando, Cuadro de mando integral, Minería de datos, Gestión del rendimiento, Previsiones, Reglas de negocio, *Dashboards*, Integración de datos –que incluye ETL (*Extract*, *Transform and Load*)–, etc.

La figura 1 representa los casos de uso, la composición tradicional de una plataforma de datos y el rol de la inteligencia de negocio.

Figura 1. Plataforma de datos



Fuente: Josep Curto

A lo largo de este material entraremos en detalle en algunas de ellas para tener claras las componentes mínimas que deben tener este tipo de sistemas.

1.2. Diferencias entre BI, BA y *big data*

Hemos comentado que las organizaciones tienen a su disposición diferentes estrategias para la explotación del dato. Frecuentemente se combinan entre ellas para responder a una necesidad de negocio, pero cada una de ellas tiene diferentes casos de uso. Para poder entender las diferencias, necesitamos definirlas también.

Primero definamos qué es *business analytics* (BA) o analítica de negocio.

> Se entiende por *business analytics* el conjunto de estrategias, tecnologías y sistemas que permiten analizar el rendimiento pasado de una organización para poder predecir comportamientos futuros, así como para detectar patrones ocultos en la información.

Actualmente también hablamos de ***data science*** como el siguiente paso a *business analytics*.

Data science

Cuando hablamos de *data science*, hacemos referencia a un campo multidisciplinar que busca generar conocimiento de datos complejos combinando algoritmos, técnicas y conocimientos de matemáticas/estadística, programación y negocio.

Ahora definimos qué es *big data*.

Se entiende por *big data* el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos, que frecuentemente está definida por volumen, velocidad y variedad del dato.

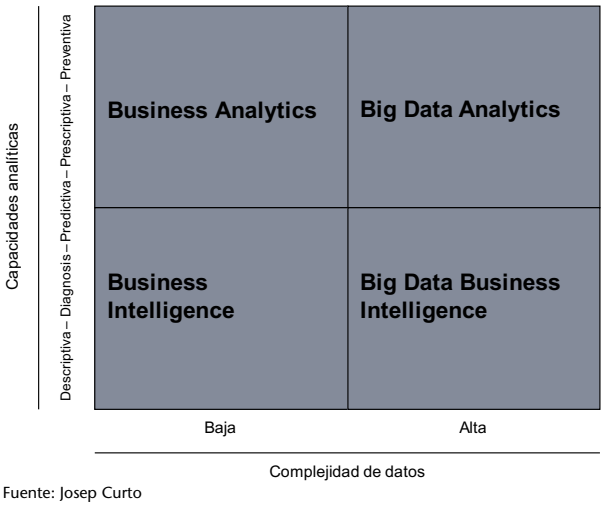
Vamos a comparar estas estrategias respecto a diferentes factores: herramientas, foco, uso, tipo del dato, complejidad del dato y alcance. Además, indicamos su nivel de madurez en el mercado. La tabla 1 describe las diferencias entre estas estrategias.

Tabla 1. Diferencias entre BI, BA y *big data*

Estrategia	***Business intelligence***	***Business analytics***	***Big data***
Madurez	Alta	Alta	Emergente
Herramientas	Consultas, Alertas, Reporting, OLAP, etc.	Clasificación, Clustering, Regresión, etc.	Machine learning, Deep Learning, Visualización, etc.
Foco	Qué y cómo pasó, cuántos, con qué frecuencia, cuál es el problema, qué es necesario hacer	Por qué está pasando, qué pasaría si todo continua igual, qué pasará a continuación, qué es lo mejor que puede pasar	Capturar, almacenar, procesar, analizar
Uso	Reactivo	Predictivo, Proactivo, Prescriptivo	Todos los anteriorees
Tipo de dato	Estructurado	Estructurado y Semi-estructurado	Todo tipo, principalmente no estructurado
Complejidad del dato	Baja	Baja/Media	Alta
Alcance	Dirección	Procesos	Vertical/Procesos

Aunque queda claro que estas estrategias son diferentes, lo normal es que se combinen en los proyectos de explotación del dato. La figura 2 permite identificar casos de uso respecto a la complejidad del dato y las capacidades analíticas que se deben desarrollar en la organización.

Figura 2. Combinación de estrategias



Fuente: Josep Curto

La lectura de esta gráfica se realiza a partir de sus ejes. Por ejemplo, cuando necesitemos desarrollar capacidades analíticas descriptivas y la complejidad de dato sea alta, combinaremos *big data* con *business intelligence* formando así un único sistema.

Una cadena de supermercados como, por ejemplo, Carrefour usa la inteligencia de negocio para comprender el rendimiento de cada uno de sus centros comerciales, usa la analítica de negocio para identificar los principales grupos de clientes (y las características que los definen) y usa *big data* para implementar y desplegar un sistema de recomendación de productos para incentivar las compras.

1.3. Beneficios

La implantación de estos sistemas de información proporciona diversos beneficios entre los que podemos destacar:

- Proporcionar una visión única, conformada, histórica, persistente y de calidad de toda la información relevante para la organización.

- Crear, manejar y mantener métricas, indicadores claves de rendimiento (*Key Performance Indicador* –KPI–) e indicadores claves de metas (*Key Goal Indicator* -KGI-) fundamentales para la empresa.

- Habilitar el acceso a información actualizada tanto a nivel agregado como en detalle.

- Reducir el diferencial entre los enfoques del departamento TI y el resto de departamentos a través de la implementación del proyecto, al comprender el departamento TI las necesidades de negocio.

- Mejor comprensión y documentación de los sistemas de información en el contexto de una organización, al identificar las fuentes relevantes de información para el negocio.

- Mejorar cómo se gestiona y compite la organización como resultado de ser capaces de:

– Diferenciar lo relevante sobre lo superfluo a través de la identificación de las métricas adecuadas de negocio.
– Acceder más rápido a información a partir de la automatización de la extracción y consolidación de datos.
– Tener mayor agilidad en la toma de decisiones.

- Crear un círculo virtuoso de la información: los datos se transforman en información que genera un conocimiento, que permite tomar mejores decisiones, que se traducen en mejores resultados y que generan nuevos datos.

1.4. ¿Cuándo es necesario?

Tal y como Thomas Davenport en su libro *Competing on Analytics* explica, está emergiendo una nueva forma de estrategia competitiva basada en el uso de la estadística descriptiva, modelos productivos y complejas técnicas de optimización, datos de alta calidad y una toma de decisiones basada en hechos. En dicho contexto, la inteligencia de negocio es el paso previo para dicha estrategia dado que ayuda a sentar las bases para su futuro despliegue.

Existen situaciones en las que la implantación de un sistema de *business intelligence* resulta adecuada. Destacamos, entre todas las que existen:

- La toma de decisiones se realiza de forma intuitiva en la organización.
- Identificación de problemas de calidad de información.
- Uso masivo de Excel como repositorios de información corporativos o de usuario. Lo que se conoce como Excel caos.
- Necesidad de cruzar información entre departamentos de forma ágil.
- Evitar silos de información.
- Las campañas de *marketing* no son efectivas por la información base usada.
- Existe demasiada información en la organización como para ser analizada de la forma habitual. Se ha alcanzado la masa crítica de datos.
- Es necesario automatizar los procesos de extracción y distribución de información.

En definitiva, los sistemas de *business intelligence* buscan responder a las preguntas:

- ¿Qué pasó?
- ¿Qué pasa ahora?
- ¿Por qué pasó?
- ¿Qué pasará?

Desplegar un proyecto de inteligencia de negocio en el seno de una organización no es un proceso sencillo. Las buenas prácticas indican que, para llegar a buen puerto, es necesario tener una estrategia de inteligencia de negocio que coordine las tecnologías, el uso, los procesos de madurez y la metodología que se va a usar.

1.4.1. ¿Cómo detectar que no existe una estrategia de gestión de datos?

Es posible detectar, a través de los siguientes puntos y percepciones, que no existe una estrategia definida en el seno de una organización:

- Los usuarios identifican el departamento de informática como el origen de sus problemas de inteligencia de negocio.
- La dirección considera que la inteligencia de negocio es otro centro de coste.
- El departamento de IT no comprende las necesidades de negocio y el sistema de inteligencia de negocio no ayuda a los usuarios de negocio.
- El sistema de BI se considera como un sistema de soporte bajo *help desk* en lugar de atender directamente a negocio.
- No se conoce la diferencia entre inteligencia de negocio y la gestión del rendimiento.
- No es posible medir el uso del sistema de inteligencia de negocio.
- No es posible medir el retorno de la inversión (*Return On Invest* –ROI–) del proyecto de *business intelligence*.
- Se considera que la estrategia para el *data warehouse* es la misma que para el sistema de inteligencia de negocio.
- No hay un plan para desarrollar, contratar, retener y hacer crecer el equipo de BI.
- Los directores de negocio desconocen si la empresa tiene una estrategia para el BI que puedan usar para comprender el rendimiento de sus unidades de negocio.
- No existe un responsable funcional, es decir, un director de inteligencia de negocio (o bien el asignado no es el adecuado).
- No existe un centro de competencia que permita definir la estrategia de inteligencia de negocio.
- Existen múltiples soluciones en la organización, distribuidas en diferentes departamentos, que repiten funcionalidad.
- No hay un plan de formación real y consistente de uso de las herramientas.
- Alguien cree que es un éxito que la información consolidada esté a disposición de los usuarios finales al cabo de dos semanas.
- Los usuarios creen que la información del sistema de inteligencia de negocio no es correcta.
- No existe una cultura analítica en la que el dato y los hechos son relevantes para tomar decisiones, sea cual sea el nivel de la organización.

El desarrollo de una estrategia de negocio es un proceso a largo plazo que incluye múltiples actividades, entre las que conviene destacar:

- Poner atención a las necesidades que requieren BI en la organización, porque se acostumbra a satisfacer a los usuarios o departamentos que gritan más fuerte, cosa que no significa que den mayor valor a la compañía. Por ejemplo, los departamentos de finanzas son un caso típico de baja atención en soluciones BI.
- Identificar qué procesos de negocio necesitan diferentes aplicaciones analíticas que trabajen de forma continua para asegurar que no existen silos de funcionalidad.
- Desarrollar un *framework* de métricas a nivel empresarial como el pilar de una gestión del rendimiento a nivel corporativo.

- Incluir los resultados de aplicaciones analíticas (minería de datos u otras) en los procesos de negocio con el objetivo de añadir valor a todo tipo de decisiones.
- Establecer los estándares de BI en la organización para racionalizar tanto las tecnologías existentes como las futuras adquisiciones.
- Revisar y evaluar el portafolio actual de soluciones en un contexto de riesgo / recompensas.
- Considerar inversiones tácticas cuyo retorno de inversión esté dentro de un período de tiempo de un año. Además, tener en cuenta los diferentes análisis de mercado, de soluciones e, incluso, el *hype cycle* de Gartner para conocer el estado del arte.
- Aprender de los éxitos y fracasos de otras empresas revisando casos de estudio y consultado a las empresas del sector para determinar qué ha funcionado y qué no.
- Crear un centro de competencia (o de excelencia) de BI (BICC). Tiene el objetivo de aunar conocimiento en tecnologías, metodologías, estrategia, con apoyo a nivel ejecutivo y con analistas de negocio implicados, y que tenga responsabilidad compartida en éxitos y fracasos.
- Alinear el departamento IT y el negocio en caso de no poder organizar un BICC, fundamental para trabajar como equipo integrado. El departamento de IT debe entender las necesidades y entregar la mejor solución ajustada a la necesidad particular y escalable a otras futuras.
- Evangelizar la organización.

1.4.2. *Business Intelligence Maturity Model*

Si bien el objetivo de este material no es dar pautas para definir una estrategia de *business intelligence* sino una introducción de conceptos, un buen punto de partida es identificar cuál es el grado de madurez de la organización respecto a la inteligencia de negocio.

El BIMM (*Business Intelligence Maturity Model*) es un modelo de madurez que permite clasificar nuestra organización desde el punto de vista del grado de madurez de implantación de sistemas *business intelligence* en la misma.

Veamos las fases:

Fase 1: No existe BI. Los datos se hallan en los sistemas de información operacionales, como la contabilidad, la facturación o la nómina, desperdigados en otros soportes o incluso solo contenidos en el *know-how* de la organización. Las decisiones se basan en la intuición, la experiencia, pero no en datos consistentes. El uso de datos corporativos en la toma de decisiones no ha sido detectado y tampoco el uso de una herramienta adecuada al hecho.

En esta fase, el sistema de administración comercial (el control de las ventas por vendedores, regiones, productos, clientes, precios, descuentos...) toma sus decisiones basadas en el conocimiento (experiencia e intuición) de cada uno de sus comerciales.

Fase 2: No existe BI, pero los datos son accesibles. No existe un procesado formal de los datos para la toma de decisiones, aunque algunos usuarios tienen acceso a información de calidad y son capaces de justificar decisiones con dicha información. Frecuentemente, este proceso se realiza mediante Excel o algún sistema simple para generar informes. Se intuye que deben existir soluciones para mejorar este proceso pero se desconoce la existencia del *business intelligence*.

En esta fase, el sistema de administración comercial ha identificado la necesidad de usar los datos para tomar mejores decisiones. Algunos comerciales toman sus decisiones basadas en datos del sistema de control de ventas y generan informes usando Excel.

Fase 3: Aparición de procesos formales de toma de decisiones basada en datos. Se establece un equipo que controla los datos y que permite hacer informes contra los mismos que permiten tomar decisiones fundamentadas. Los datos son extraídos directamente de los sistemas transaccionales sin procesos de calidad o preparación para el análisis automático y no existe un almacén único para los datos relevantes.

En esta fase, el sistema de administración comercial ha creado un equipo que prepara los informes, aunque el proceso sigue siendo manual y supone muchas horas de trabajo preparando los datos.

Fase 4: *Data warehouse*. El impacto negativo contra los sistemas transaccionales lleva a la conclusión de que un repositorio de datos es necesario para la organización. Se percibe el *data warehouse* como una solución deseada. El *reporting* sigue siendo personal.

En esta fase el sistema de administración comercial ha creado un repositorio único de calidad con los datos relevantes para el análisis.

Fase 5: *Data Warehouse* crece y el *reporting* se formaliza. El *data warehouse* funciona y se desea que todos se beneficien de él, de forma que el *reporting* corporativo se formaliza. Se habla de OLAP (análisis multidimensional), pero solo algunos identifican realmente sus beneficios.

En esta fase, el sistema de administración comercial extiende el repositorio único a otras áreas y se automatizan los informes.

Fase 6: Despliegue de OLAP. Después de cierto tiempo, ni el *reporting* ni la forma de acceso al *data warehouse* es satisfactoria para responder a preguntas sofisticadas. OLAP se despliega para dichos perfiles. Las decisiones empiezan a impactar de forma significativa en los procesos de negocio a lo largo de la organización.

En esta fase, el sistema de administración comercial empieza a usar el análisis multidimensional (OLAP) para algunos perfiles avanzados.

Fase 7: *business intelligence* se formaliza. Aparece la necesidad de implantar otros procesos de inteligencia de negocio, como *Data Mining*, *Balanced ScoreCard*... y procesos de calidad de datos impactan en procesos como *Customer Relationship Management* (CRM), *Supply Chain Management* (SCM)... Se ha establecido una cultura corporativa que entiende claramente las diferencias entre sistemas OLTP y DSS.

En esta fase, el sistema de administración comercial usa un sistema de inteligencia de negocio y se inicia la fase de implementar otros análisis más avanzados.

Existen otros modelos de madurez como el modelo Delta analítico de Thomas Davenport o el de TDWI*.

***https://tdwi.org/pages/maturity-model/analytics-maturity-model-assessment-tool.aspx**

2. Gestión del dato

En el apartado anterior, hemos introducido el concepto de inteligencia de negocio. Este concepto hace referencia, al mismo tiempo, tanto a un sistema de información como a una estrategia de negocio para mejorar la toma de decisiones. El despliegue de este tipo de estrategias/sistemas pasa por la gestión eficiente del dato.

En este apartado vamos a revisar qué significa gestionar el dato en el contexto de la inteligencia de negocio.

2.1. ¿Qué significa gestionar el dato?

Dentro de una organización, las personas que deben tomar decisiones tienen expectativas a ser cubiertas. Nos referimos a que, para tomar una decisión, el dato debe estar **disponible** y **accesible**, ser de **calidad**, **en momento adecuado**, **securizado** y **transformado en información**.

Esto significa:

- **Disponible**: el dato ha sido capturado y almacenado en un repositorio.
- **Accesible**: existe un mecanismo para el consumo de dato que habilita su acceso por terceros, tanto sistemas como personas.
- **Calidad**: el dato que se ha validado tiene el nivel de calidad suficiente para la toma de decisiones.
- **En momento adecuado**: se han tenido en cuenta las necesidades temporales de negocio para la disponibilidad y accesibilidad del dato.
- **Securizado**: el dato está protegido y solo pueden acceder aquellos que tienen permisos.
- **Información**: el dato se ha transformado en información a través del análisis.

Al final, gestionar el dato significar ser capaz de almacenar y procesar el dato de forma eficiente cumpliendo las expectativas anteriores para que responda a las necesidades actuales de una organización.

En los siguientes subapartados discutiremos los elementos de la inteligencia de negocio que permiten almacenar y procesar el dato de forma eficiente.

2.2. Almacenamiento del dato en BI

Como ya se ha comentado, un sistema de inteligencia de negocio está formado por diferentes elementos, pero de todas las piezas, la principal de ellas es el *data warehouse* o almacén de datos. Necesitamos definir este concepto.

> Un *data warehouse* es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, independientemente de cómo se vayan a utilizar posteriormente por los consumidores o usuarios; con las propiedades siguientes: estable, coherente, fiable y con información histórica.

Al abarcar un ámbito global de la organización y con un amplio alcance histórico, el volumen de datos puede ser muy grande (centenas de terabytes o incluso petabytes). Las bases de datos relacionales son el soporte técnico más comúnmente usado para almacenar las estructuras de estos datos y sus grandes volúmenes.

> El *data warehouse* de Wal-Mart guarda información de las transacciones de más de 100 millones de clientes y los datos logísticos de más de 25.000 proveedores. Ya en 1992, este almacén de datos llegó a tener más de 1 terabyte de información relevante para comprender el comportamiento de clientes y optimizar la relación con los proveedores.

Resumiendo, el *data warehouse* presenta las siguientes características:

- **Orientado a un tema**: organiza una colección de información alrededor de un tema central.
- **Integrado**: incluye datos de múltiples orígenes y presenta consistencia de datos.
- **Variable en el tiempo**: se realizan fotos de los datos basadas en fechas o hechos.
- **No volátil**: la información es persistente y solo de lectura para los usuarios finales.

Frecuentemente el *data warehouse* está constituido por una base de datos relacional (que guarda sus registros por filas), pero no es la única opción factible, también es posible considerar las bases de datos orientadas a columnas (que guardan los datos por columnas), basadas en lógica asociativa (que identifican conceptos de negocio usando lógica) o *appliances* especializadas (optimizadas para el rendimiento en el análisis).

2.2.1. *Data Warehousing*

Debemos tener en cuenta que en el contexto de un *data warehouse* existen otros elementos que se combinan para poder responder a las necesidades de negocio:

- ***Data Warehousing***: es el proceso de extraer y filtrar datos de las operaciones comunes de la organización, procedentes de los distintos sistemas de información operacionales y/o sistemas externos, para transformarlos, integrarlos y almacenarlos en un almacén de datos con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones de la organización.

- ***Data Mart***: es un subconjunto de los datos del *data warehouse* con el objetivo de responder a un determinado análisis, función o necesidad, y con una población de usuarios específica. Está pensado para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización. Por ejemplo, un posible uso sería para la minería de datos o para la información de *marketing*.

- ***Operational Data Store* (ODS)**: es un tipo de almacén de datos que proporciona solo los últimos valores de los datos y no su historial; generalmente, es además admisible un pequeño desfase o retraso sobre los datos operacionales. También podemos tener específicos para la minería de datos y la exploración del dato.

- ***Staging Area***: es el sistema que permanece entre las fuentes de datos operacionales y el *data warehouse* con el objetivo de:

  – Facilitar la extracción de datos desde fuentes de origen con una heterogeneidad y complejidad grande.
  – Mejorar la calidad de datos.
  – Ser usado como caché de datos operacionales con el que posteriormente se realiza el proceso de *data warehousing*.
  – Acceder en detalle a información no contenida en el *data warehouse*.

- **Procesos ETL**: es una tecnología de integración de datos basada en la consolidación de datos; tradicionalmente se usa para alimentar almacenes de datos de cualquier tipo: *data warehouse*, *data mart*, *staging area* y ODS. Usualmente se combina con otras técnicas de consolidación de datos. Esta tecnología permite extraer, transformar y cargar datos.

- **Metadatos**: son datos estructurados y codificados que describen características del proceso de *data warehousing* y de los diferentes elementos que se han considerado en la arquitectura del *data warehouse*.

Minería de datos

Cuando hablamos de minería de datos, o *data mining*, hacemos referencia al campo interdisciplinar con el objetivo general de predecir resultados y/o descubrir relaciones en los datos. Puede ser descriptivo, i.e. descubrir patrones que describen los datos, o predictivo, para pronosticar el comportamiento del modelo basado en los datos disponibles.

Caché

Cuando hablamos de caché, hacemos referencia al proceso de usar la memoria como almacenamiento temporal de datos.

La figura 3 resume las diferentes componentes que encontramos en el contexto del *data warehouse*.

Figura 3. Componentes del *data warehouse*

Fuente: Josep Curto

2.2.2. Elementos en data *warehousing*: hechos, dimensiones y métricas

Forma normal

Cuando hablamos de forma normal, hacemos referencia al proceso que consiste en designar y aplicar una serie de reglas para el diseño de base de datos con el objetivo de eliminar datos repetidos y tener integridad en los datos.

La estructura relacional de una base de datos operacional sigue las formas normales en su diseño. En un *data warehouse* no debe seguirse ese patrón de diseño. La idea principal es que la información sea almacenada de forma desnormalizada para optimizar las consultas. Para ello, debemos identificar en el seno de nuestra organización los procesos de negocio, las perspectivas de análisis para el proceso de negocio y medidas cuantificables asociadas a los mismos.

Es decir, se estructura el dato en procesos de negocio, vistas de análisis y las medidas para comprender su evolución. De esta manera hablaremos de:

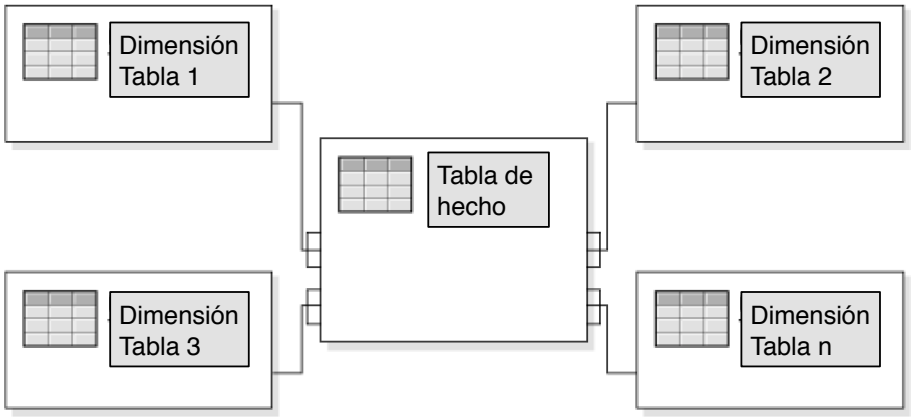
- **Tabla de hecho**: es la representación en el *data warehouse* de los procesos de negocio de la organización. Por ejemplo, una venta puede identificarse como un proceso de negocio de manera que es factible, si corresponde en nuestra organización, considerar la tabla de hecho ventas.
- **Dimensión**: es la representación en el *data warehouse* de una vista para un cierto proceso de negocio. Si regresamos al ejemplo de una venta, para la misma, tenemos el cliente que ha comprado, la fecha en la que se ha realizado. Estos conceptos pueden ser considerados como vistas para este proceso de negocio. Puede ser interesante recuperar todas las compras realizadas por un cliente. Ello nos hace entender por qué la identificamos como una dimensión.
- **Métrica**: son los indicadores de un proceso de negocio. Aquellos conceptos cuantificables que permiten medir nuestro proceso de negocio. Por ejemplo, en una venta tenemos el importe de la misma.

Existen principalmente dos tipos de esquemas para estructurar los datos en un almacén de datos:

- **Esquema en estrella**: consiste en estructurar la información en procesos, vistas y métricas recordando a una estrella (por ello el nombre). A nivel de diseño, consiste en una tabla de hechos (lo que en los libros encontraremos como *fact table*) en el centro para el hecho objeto de análisis y una o varias tablas de dimensión por cada punto de vista de análisis que participa de la descripción de ese hecho. En la tabla de hecho encontramos los

atributos destinados a medir (cuantificar): sus métricas. La tabla de hechos solo presenta uniones con dimensiones. La figura 4 ilustra este esquema.

Figura 4. Esquema en estrella



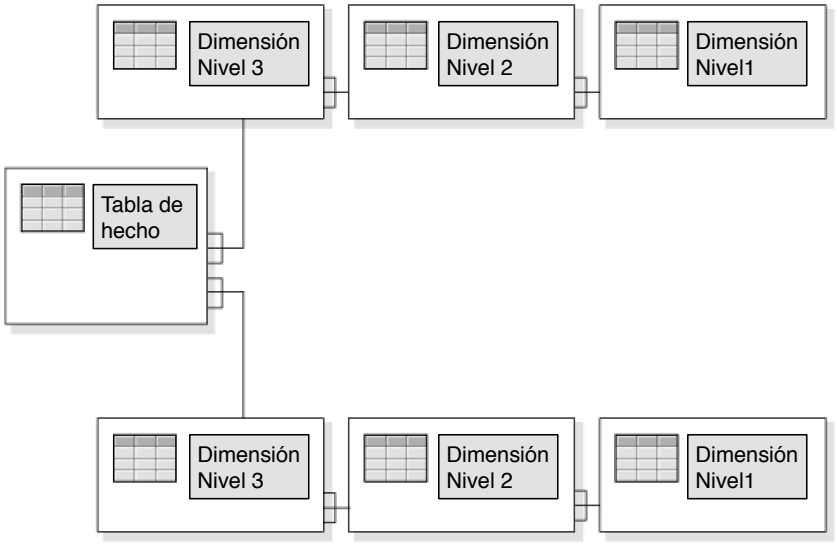
Fuente: Josep Curto

- **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 y aparecen nuevas uniones. Es posible distinguir dos tipos de esquemas en copo de nieve:

– **Completo**: en el que todas las tablas de dimensión en el esquema en estrella aparecen ahora normalizadas.
– **Parcial**: solo se lleva a cabo la normalización de algunas de ellas.

La figura 5 ilustra este esquema.

Figura 5. Esquema en en copo de nieve



Fuente: Josep Curto

Es conveniente profundizar en los conceptos de tabla de hecho, dimensión y métrica.

Tabla de hecho

Ya sabemos que una tabla de hecho representa un proceso de negocio. A nivel de diseño, es una tabla que permite guardar dos tipos de atributos diferenciados:

- Medidas del proceso/actividad/flujo de trabajo/evento que se pretende modelizar.
- Claves foráneas hacia registros en una tabla de dimensión (o en otras palabras, como ya sabemos, hacia una vista de negocio).

Existen diferentes **tipos de tablas de hecho**:

- ***Transaction Fact Table*** **(tabla de hechos de transacciones)**: representa eventos que suceden en un determinado espacio-tiempo. Se caracteriza por permitir analizar los datos con el máximo detalle. Por ejemplo, podemos pensar en una venta que tiene como resultado métricas como el importe de la misma.

- ***Factless Fact Tables/Coverage Table*** **(tablas de hechos sin medidas)**: son tablas que no tienen medidas, cosa que tiene sentido dado que representan el hecho que el evento suceda. Frecuentemente, a dichas tablas se añaden contadores para facilitar las consultas SQL. Por ejemplo, podemos pensar en la asistencia a un acto benéfico en el que, por cada persona que asiste, tenemos un registro, pero podríamos no tener ninguna métrica asociada más.

- ***Periodic Snapshot Fact Table*** **(tablas de hechos periódicas)**: son tablas de hecho usadas para recoger información de forma periódica a intervalos de tiempo regulares. Dependiendo de la situación medida o de la necesidad de negocio, este tipo de tablas de hecho son una agregación de las anteriores o están diseñadas específicamente. Por ejemplo, podemos pensar en el balance mensual. Los datos se recogen acumulados de forma mensual.

- ***Accumulating Snapshot Fact Table*** **(tablas de hecho agregadas)**: representan el ciclo de vida completo de una actividad o proceso, que tiene un principio y final. Se caracterizan por presentar múltiples dimensiones relacionadas con los eventos presentes en un proceso. Por ejemplo, podemos pensar en un proceso de matriculación de un estudiante y que recopila, durante su periodo de vida, datos que suelen sustituir los anteriores (superación y recopilación de asignaturas, por ejemplo).

Dimensión

Sabemos que una dimensión principalmente recoge los puntos de análisis de un hecho. Por ejemplo, una venta se puede analizar respecto del día de venta, producto, cliente, vendedor o canal de venta, entre otros.

Existen diferentes tipos de dimensiones:

- ***Slowly Changing Dimensions* (SCD)**: son dimensiones que tienen en cuenta la gestión de los cambios históricos en los datos. En función de las necesidades de negocio, el dato se borra, se modifica o se guarda para su comparación.

- **Degeneradas**: son dimensiones que solo tienen un atributo y frecuentemente se dejan en la tabla de hecho. Por ejemplo, el sexo de un paciente.

- ***Junk***: son dimensiones que contienen información volátil que se usa puntualmente y que no se guarda de forma permanente en el *data warehouse*.

- **Conformadas**: son dimensiones que se usan para compartir información entre tablas de hecho, lo que permite hacer consultas comunes y cruzar información. El ejemplo más fácil es la dimensión temporal.

- ***Bridge* (puente)**: permite definir relaciones entre tablas de hecho, necesarias para definir, por ejemplo, la relación entre un piloto y sus múltiples patrocinadores.

- ***Role-playing* (roles)**: que tienen asignado un significado. Por ejemplo, podemos tener la dimensión fecha, pero también fecha de entrega.

- **Alta cardinalidad o *monster***: que contienen una gran cantidad de datos difícilmente consultables en su totalidad. Una buena práctica es romper la dimensión en dos tablas: una que contenga los valores estáticos y otra que contenga los valores volátiles. Un ejemplo claro puede ser la información de cliente. Debemos ser conscientes de cuál es la información primordial del mismo respecto de la que solo se usa puntualmente en los informes u otros análisis.

Métricas

También podemos distinguir diferentes tipos de medidas, basadas en el tipo de información que recopilan así como su funcionalidad asociada:

- **Métricas**: valores que recogen el proceso de una actividad o los resultados de la misma. Estas medidas proceden del resultado de la actividad de negocio.

– **Métricas de realización de actividad (*leading*)**: miden la realización de una actividad. Por ejemplo, la participación de una persona en un evento.

– **Métricas de resultado de una actividad (*lagging*)**: recogen los resultados de una actividad. Por ejemplo, la cantidad de puntos de un jugador en un partido.

- **Indicadores clave**: entendemos por este concepto valores correspondientes que hay que alcanzar y que suponen el grado de asunción de los objetivos. Estas medidas proporcionan información sobre el rendimiento de una actividad o sobre la consecución de una meta.

– ***Key Performance Indicator* (KPI)**: Indicadores clave de rendimiento. Más allá de la eficacia, se definen unos valores que nos explican en qué rango óptimo de rendimiento nos deberíamos situar al alcanzar los objetivos. Son métricas del proceso.

– ***Key Goal Indicator* (KGI)**: Indicadores de metas. Definen mediciones para informar a la dirección general si un proceso TIC ha alcanzado sus requisitos de negocio y se expresan, por lo general, en términos de criterios de información.

Debemos añadir que existen también indicadores de desempeño. Los indicadores clave de desempeño (en definitiva, son KPI) definen mediciones que determinan cómo de bien se está desempeñando el proceso de TI para alcanzar la meta. Son los indicadores principales que indican si será factible lograr una meta o no, y son buenos indicadores de las capacidades, prácticas y habilidades. Los indicadores de metas de bajo nivel se convierten en indicadores de desempeño para los niveles altos.

2.3. Captura, transformación y gestión del dato en BI

En el anterior subapartado hemos discutido cómo se almacenan los datos en un sistema de inteligencia de negocio. Es importante recalcar que en la inteligencia de negocio se estructura de antemano el formato de análisis.

Tras crear un modelo que representa los procesos de negocio relevantes para la organización y las diferentes perspectivas de análisis, y haberlo implementado en el *data warehouse*, el siguiente paso es la carga de los datos. No solo se trata de cargar datos en el repositorio sino también de preocuparse por otros aspectos como cuál es la mejor forma de capturar el dato, qué tipo de transformaciones son necesarias para transformar los datos en información y qué acciones son necesarias para gestionar de forma eficiente el dato.

Para capturar, transformar y gestionar el dato de forma eficiente se usa la integración de datos, que proporciona una visión única de todos los datos de negocio se encuentren donde se encuentren.

Este subapartado se centrará en la integración de datos en general y en los procesos ETL (Extracción, Transformación y Carga) en particular, que es una de las tecnologías de integración de datos que se usa en los proyectos de implantación de *business intelligence*. El objetivo de este apartado es conocer las diferentes opciones de integración de datos en el ámbito de la inteligencia de negocio y, en particular, conocer el diseño de procesos ETL.

2.3.1. Integración de datos

La integración de datos incluye diversas componentes destinadas a la gestión eficiente del dato como pueden ser:

- Servicios de acceso/entrega de datos (vía adaptadores/conectores)
- Gestión de servicios
- *Data profiling* o perfilado de datos, que permite conocer y analizar los datos para conocer su estructura, contenido, relaciones y reglas que puedan derivarse de los datos
- *Data Quality* o calidad de datos, que permite conocer, analizar y gestionar el nivel de calidad de un conjunto de datos
- Procesos Operacionales
- Servicios de transformación: CDC, SCD, validación, agregación
- Servicios de acceso a tiempo real
- *Extract, Transform and Load* (ETL)
- *Enterprise Information Integration* (EII)
- *Enterprise Application Integration* (EAI)
- Capa de transporte de datos
- Gestión de metadatos

La figura 6 ilustra las componentes de integración dentro del contexto de la inteligencia de negocio.

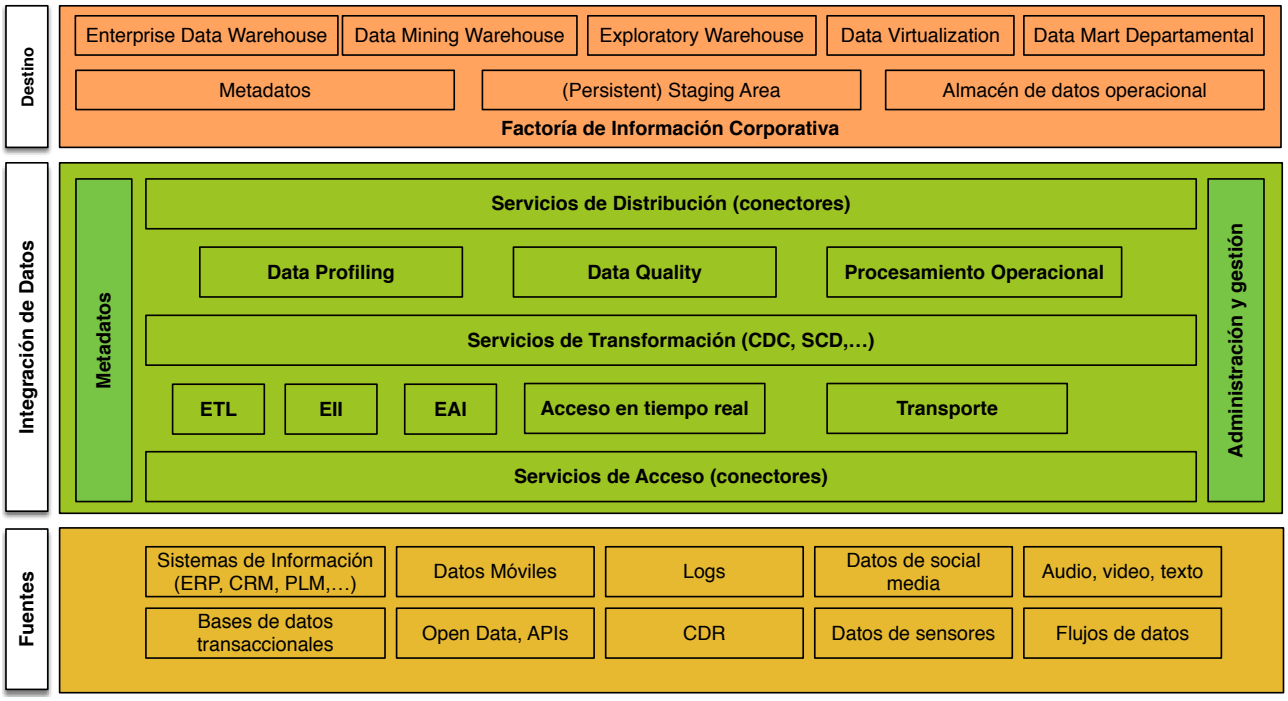
Las herramientas de integración de datos están evolucionando para incluir mayores prestaciones de curación y calidad de datos basadas en algoritmos analíticos automáticos o semi-automáticos. Por ejemplo, en entornos internacionales, dónde el salario de un empleado puede denominarse de forma diferente (*wages vs. salary*) e incluso incluir prestaciones diferentes, la herramienta de integración de datos puede reconocer que son el mismo concepto de negocio.

Curación de datos

Cuando hablamos de curación de datos, hacemos referencia al proceso que busca asegurar que los datos sean confiables y recuperables para fines de investigación futuros o reutilización.

El punto de partida adecuado es definir formalmente el concepto de integración de datos.

Figura 6. Componentes Integración de datos



Fuente: Josep Curto

Se entiende por integración de datos el conjunto de aplicaciones, productos, técnicas y tecnologías que permiten una visión única consistente de nuestros datos de negocio.

Respecto a la definición:

- Las aplicaciones son soluciones a medida que permiten la integración de datos en base al uso de productos de integración.
- Los productos comerciales desarrollados por terceros capacitan la integración mediante el uso de tecnologías de integración.
- Las tecnologías de integración son soluciones para realizar la integración de datos.

La integración de datos juega un papel cada vez más fundamental en las organizaciones puesto que en la actualidad el dato puede residir en plataformas internas (sistemas operacionales y decisionales, *cloud* privado) y externas (*internet of things*, *cloud* público, dispositivos móviles), y es necesario poder capturar y distribuir el dato a los sistemas de análisis necesarios.

Cloud computing

Cuando hablamos de *cloud computing* (o computación en nube), hacemos referencia a un término general para la prestación de servicios alojados a través de Internet. Tiene diferentes modalidades: privada (proporcionada por organización), pública (proporcionada por terceros) o híbrida (combinación de ambos).

Existen diferentes técnicas de integración de datos:

- **Propagación de datos**: consiste en copiar datos de un lugar de origen a un entorno destino local o remoto. Los datos pueden extraerse del origen

mediante programas que generen un fichero que debe ser transportado al destino, donde se utilizará como fichero de entrada para cargar en la base de datos de destino. Una aproximación más eficiente es descargar solo los datos que han cambiado en origen respecto a la última propagación realizada, generando un fichero de carga incremental que también será transportado al destino.

- **Consolidación de datos**: consiste en capturar los cambios realizados en múltiples entornos origen y propagarlos a un único entorno destino, donde se almacena una copia de todos estos datos. Ejemplos de esta son un *data warehouse* o un ODS, alimentado por varios entornos de producción.

- **Federación de datos**: proporciona a las aplicaciones una visión lógica virtual común de una o más bases de datos. Esta técnica permite acceder a diferentes entornos de origen de datos, que pueden estar en los mismos o en diferentes gestores de datos y máquinas, y crear una visión de este conjunto de bases de datos como si fuese, en la práctica, una base de datos única e integrada.

- **CDC (*Change Data Capture*)**: se utilizan para capturar los cambios producidos por las aplicaciones operacionales en las bases de datos de origen, de tal manera que pueden ser almacenados y/o propagados a los entornos destino para que estos mantengan la consistencia con los entornos origen. Existen cuatro técnicas principales: CDC por aplicación (la aplicación genera la actualización), CDC por *timestamp* (la base de datos genera el cambio basado en fechas), CDC por *triggers* (la base de datos genera el cambio en función de acciones de actualización de los datos que contiene) y CDC por captura de LOG (consiste en monitorizar los cambios a nivel del *log* de registros de cambios de la aplicación).

- **Técnicas híbridas**: la técnica elegida en la práctica para la integración de datos dependerá de los requisitos de negocio para la integración, pero también, en gran medida, de los requisitos tecnológicos y de las probables restricciones presupuestarias. A la práctica, se suelen emplear varias técnicas de integración constituyendo lo que se denomina una técnica híbrida.

2.3.2. ETL

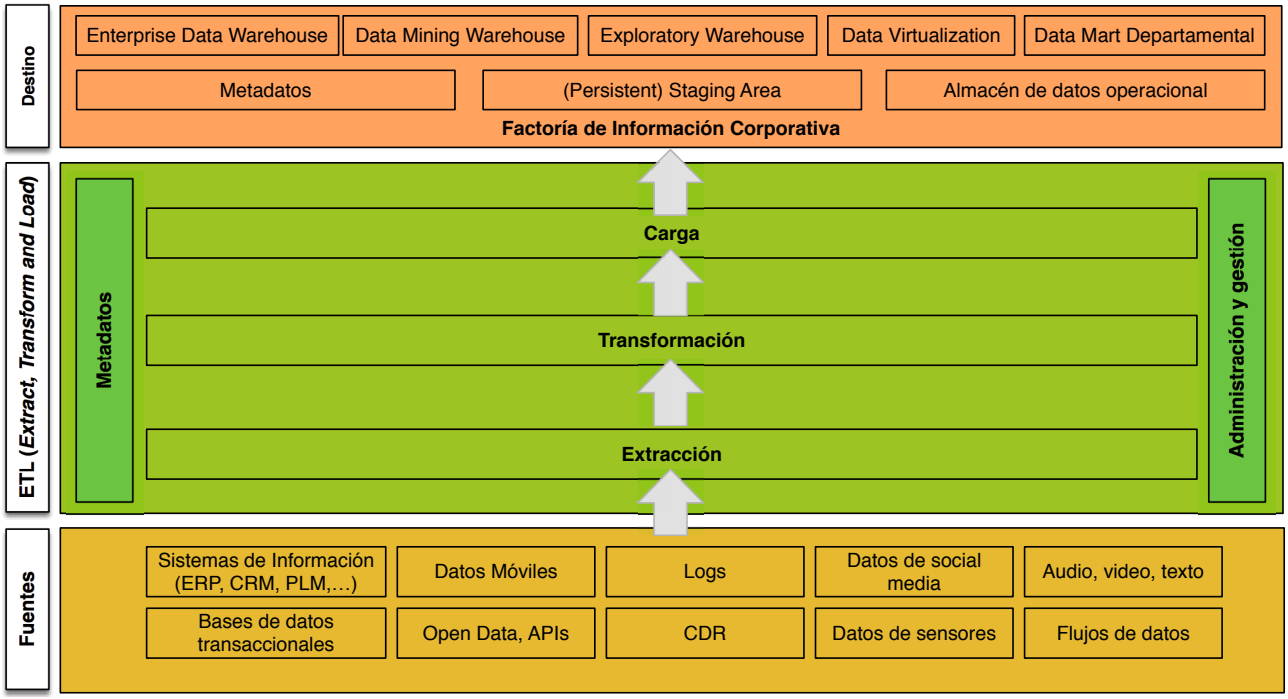
En el contexto de la inteligencia de negocio, dentro de todas las técnicas de integración de datos, las herramientas ETL han sido la opción usual para alimentar el *data warehouse*. La funcionalidad básica de estas herramientas está compuesta por:

- Gestión y administración de servicios
- Extracción de datos

- Transformación de datos
- Carga de datos
- Gestión de datos

La figura 7 ilustra las componentes de ETL dentro del contexto de la inteligencia de negocio.

Figura 7. Componentes ETL



Fuente: Josep Curto

ETL es una tecnología que permite extraer datos del entorno origen, transformarlos según nuestras necesidades de negocio para integración de datos y cargar estos datos en los entornos destino. Los entornos origen y destino son usualmente bases de datos y/o ficheros, pero en ocasiones también pueden ser colas de mensajes de un determinado *middleware*, así como ficheros u otras fuentes estructuradas, semiestructuradas o no estructuradas. Está basada en técnicas de consolidación.

Las herramientas de ETL en la práctica mueven o transportan datos entre entornos origen y destino, pero también documentan cómo estos datos son transformados (si lo son) entre el origen y el destino, almacenando esta información en un catálogo propio de metadatos; intercambian estos metadatos con otras aplicaciones que puedan requerirlos y administran todas las ejecuciones y procesos de la ETL: planificación del transporte de datos, *log* de errores, *log* de cambios y estadísticas asociadas a los procesos de movimiento de datos. Este tipo de herramientas suelen tener una interfaz gráfica de usuario y permiten diseñar y administrar y controlar cada uno de los procesos del entorno ETL.

¿Cómo funciona un proceso ETL?

- Se identifican los conjuntos de datos que se van a extraer.
- Se accede a las fuentes de datos de origen y se recuperan.
- Se realizan las transformaciones necesarias en los datos (por ejemplo, cambiar el formato de fecha).
- Se carga el dato transformado en el dato de destino (por ejemplo, el *data warehouse*).

Existen diferentes tipos:

- ETL de generación de código (mediante un lenguaje de programación)
- ETL basado en una herramienta especializada
- ETL integrado en la base de datos

3. Explotación del dato

Hemos visto que para generar valor en el contexto de la inteligencia de negocio es necesario almacenar el dato de forma adecuada, considerando solo la información que es relevante. El siguiente paso es explotar el dato almacenado. Lo que, en definitiva, nos permita tomar decisiones y llevar a cabo acciones informadas.

En este apartado vamos a revisar los siguientes enfoques para la explotación del dato: informes, OLAP y cuadros de mando.

3.1. Informes

El punto de entrada tradicional para una herramienta de inteligencia de negocio en el contexto de una organización es la necesidad de informes operacionales.

A lo largo de la vida de una empresa, la cantidad de datos que se generan por su actividad de negocio crece de forma exponencial y esa información se guarda tanto en las bases de datos de las aplicaciones de negocio como en ficheros en múltiples formatos.

Es necesario generar y distribuir informes para conocer el estado del negocio y poder tomar decisiones a todos los niveles: operativo, táctico y estratégico.

El primer enfoque es modificar las aplicaciones de negocio para que las mismas puedan generar los informes. Frecuentemente, el impacto en las aplicaciones es considerable, afectando tanto el rendimiento de los informes como de las operaciones que soporta la aplicación.

En el momento en que se busca una solución que permita generar informes sin impactar en el rendimiento de las aplicaciones de negocio, es cuando se considera el *data warehouse*.

Es necesario comentar que:

- Las herramientas de informes existen desde hace mucho tiempo y, por ello mismo, son soluciones maduras que permiten cubrir las necesidades de los usuarios finales respecto de los informes.

- Cada fabricante soporta la creación de todo tipo de informes; en función del enfoque, la dependencia de los usuarios finales respecto al departamento IT puede ser diferente.

- Las fuentes de origen de los informes son varias, desde el propio *data warehouse*, OLAP, metadatos u ODS.

3.1.1. Qué es un informe

Las herramientas de informes (o también llamadas de *reporting*) permiten responder principalmente a la pregunta **¿qué pasó?** Dado que esa es la primera pregunta que se formulan los usuarios de negocio, la gran mayoría de las soluciones de *business intelligence* del mercado incluyen un motor de generación de informes. Definamos primero qué es un informe.

> Un informe es un documento a través del cual se presentan los resultados de uno o varios procesos de negocio. Suele contener texto acompañado de elementos como tablas o gráficos para agilizar la compresión de la información presentada.

Los informes están destinados a usuarios de negocio que tienen la necesidad de conocer la información consolidada y agregada para la toma de decisiones.

> Imaginemos una empresa que tiene diversas tiendas distribuidas en una ciudad. El director de esta cadena necesita conocer el rendimiento de cada una de las tiendas para poder gestionarlas de forma eficiente. Por lo que necesita un informe en que el que se presenten y analicen los resultados tanto a nivel agregado (para conocer la relevancia de una tienda respecto del total) como a nivel de tienda (para conocer los pormenores). Es decir, estamos hablando de ventas, costes, clientes, productos vendidos, productos en el almacén, personal...

Ahora podemos definir formalmente las herramientas de *reporting*.

> Se entiende por plataforma de *reporting* aquellas soluciones que permiten diseñar y gestionar (distribuir, planificar y administrar) informes en el contexto de una organización o en una de sus áreas.

3.1.2. Tipos de informes

Existen diferentes tipos de informes en función de la interacción ofrecida al usuario final y la independencia respecto del departamento TI:

- **Estáticos**: tienen un formato preestablecido inamovible.

- **Paramétricos**: presentan parámetros de entrada y permiten múltiples consultas.

- ***Ad hoc***: son creados por el usuario final a partir de la capa de metadatos que permite usar el lenguaje de negocio propio. Esta capa es muy relevante porque oculta el lenguaje técnico a los usuarios de negocio y además permite aumentar el valor del informe añadiendo nueva información (como reglas de negocio o nuevas métricas).

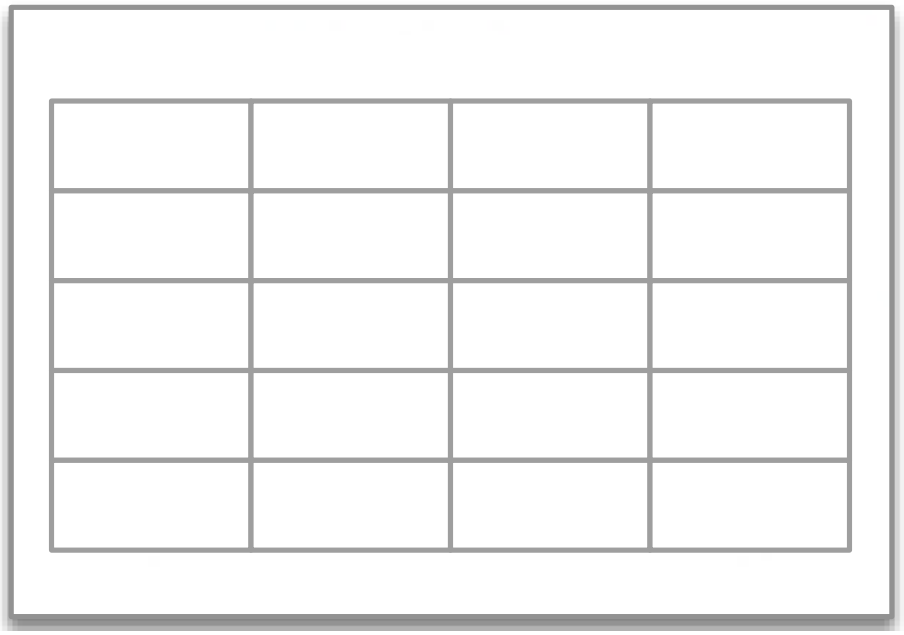
3.1.3. Elementos de un informe

Principalmente un informe puede estar formado por diversos elementos que representan tablas de hecho, dimensiones y métricas:

- **Texto**: describe el estado del proceso de negocio, proporciona las descripciones necesarias para entender el resto de elementos del informe, así como etiquetas (título) y/o metadatos (fecha de ejecución, fórmulas de cálculo...).

- **Tablas**: este elemento tiene forma de matriz (filas y columnas) y permite presentar una gran cantidad de información como ilustra la figura 8.

Figura 8. Tablas



Fuente: Josep Curto

- **Gráficos**: este elemento persigue el objetivo de mostrar información con un alto impacto visual que sirva para obtener información agregada con mucha más rapidez que a través de tablas.

- **Mapas**: este elemento permite mostrar información geolocalizada como se ilustra en la figura 9.

Figura 9. Mapa

Fuente: Josep Curto

- **Métricas**: este elemento permite conocer cuantitativamente el estado de un proceso de negocio. Son parámetros que se miden como, por ejemplo, las ventas de un acondicionador de pelo por región.

- **Alertas visuales y automáticas**: permiten definir avisos automáticos de los cambios de estado de un proceso de negocio. Estas alertas están formadas por elementos gráficos como fechas, iconos o colores resaltados, y deben estar automatizadas en función de reglas de negocio encapsuladas en el informe.

Vamos a entrar en detalle en alguno de estos elementos.

3.1.4. Tipos de métricas

Los informes incluyen métricas de negocio. Es por ello necesario definir los diferentes tipos de medidas existentes basadas en el tipo de información que recopilan así como la funcionalidad asociada:

- **Métricas**: valores que recogen el proceso de una actividad o los resultados de la misma. Estas medidas proceden del resultado de la actividad de negocio.
  - **Métricas de realización de actividad (*leading*)**: miden la realización de una actividad. Por ejemplo, la participación de una persona en un evento.

– **Métricas de resultado de una actividad (*lagging*)**: recogen los resultados de una actividad. Por ejemplo, la cantidad de puntos de un jugador en un partido.

- **Indicadores clave**: entendemos por este concepto valores correspondientes que hay que alcanzar y que suponen el grado de asunción de los objetivos. Estas medidas proporcionan información sobre el rendimiento de una actividad o sobre la consecución de una meta.

– ***Key Performance Indicator* (KPI)**: Indicadores clave de rendimiento. Más allá de la eficacia, se definen unos valores que nos explican en qué rango óptimo de rendimiento nos deberíamos situar al alcanzar los objetivos. Son métricas del proceso. Por ejemplo, la ratio de crecimiento de altas en un servicio.

– ***Key Goal Indicator* (KGI)**: Indicadores de metas. Definen mediciones para informar a la dirección general si un proceso TIC ha alcanzado sus requisitos de negocio. Por lo general, se expresan en términos de criterios de información. Si consideramos el KPI anterior, sería marcar un valor objetivo de crecimiento del servicio que se pretende alcanzar, por ejemplo, un 2 %.

3.1.5. Tipos de gráficos

En el proceso de confección de un informe, uno de los puntos más complicados es la selección del tipo de gráfico. Debemos empezar primero por la definición formal del concepto.

> Se entiende por gráfico la representación visual de una serie de datos.

El gráfico puede ser una herramienta eficaz ya que:

- Permite presentar la información de forma clara, sencilla y precisa.
- Facilita la comparación de datos y habilita destacar tendencias y diferencias.

El uso del gráfico va a depender del tipo de dato, que podemos clasificar en:

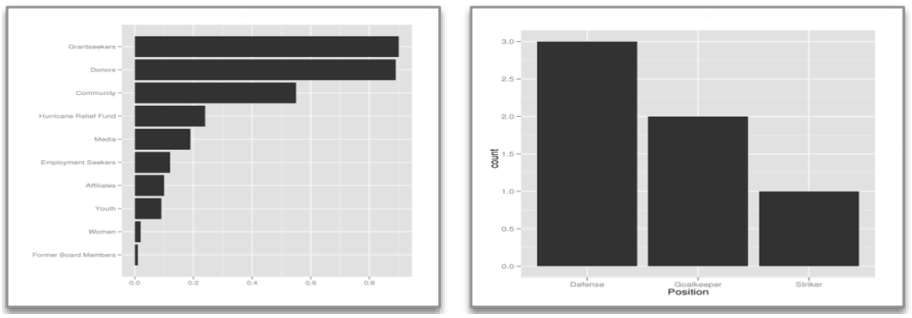
- **Cualitativos**: se refieren a cualidades o modalidades que no pueden expresarse numéricamente. Pueden ser ordinales (siguen un orden) o categóricos (sin orden).

- **Cuantitativos**: se refieren a cantidades o valores numéricos. Pueden ser discretos (toman valores enteros) o continuos (toman cualquier valor en un intervalo).

Revisemos ahora algunos de los tipos de gráficos más relevantes:

- **Gráficos de barras**: es una representación gráfica en un eje cartesiano de las frecuencias de una variable cualitativa o discreta. La orientación puede ser vertical u horizontal. Se pueden clasificar en sencillo (representa una única serie de datos), agrupado (contiene varias series de datos) o apilado (se divide en segmentos de diferentes colores o texturas, y cada uno de ellos representa una serie), como se ilustra en la figura 10.

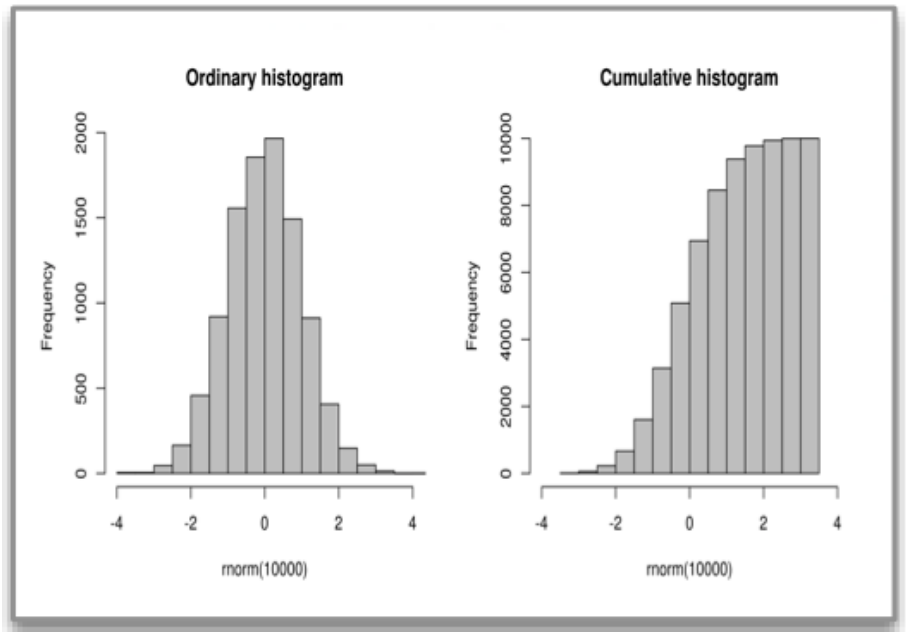
Figura 10. Gráfico de barras



Fuente: Josep Curto

- **Histograma**: se usa para representar las frecuencias de una variable cuantitativa continua. En uno de los ejes se posicionan las clases de la variable continua (los intervalos o las marcas de clase, que son los puntos medios de cada intervalo) y en el otro eje, las frecuencias. Existen también los histogramas bi-direccionales que contienen dos series de datos cuyas barras de frecuencias crecen en sentidos opuestos como se ilustra en la figura 11.

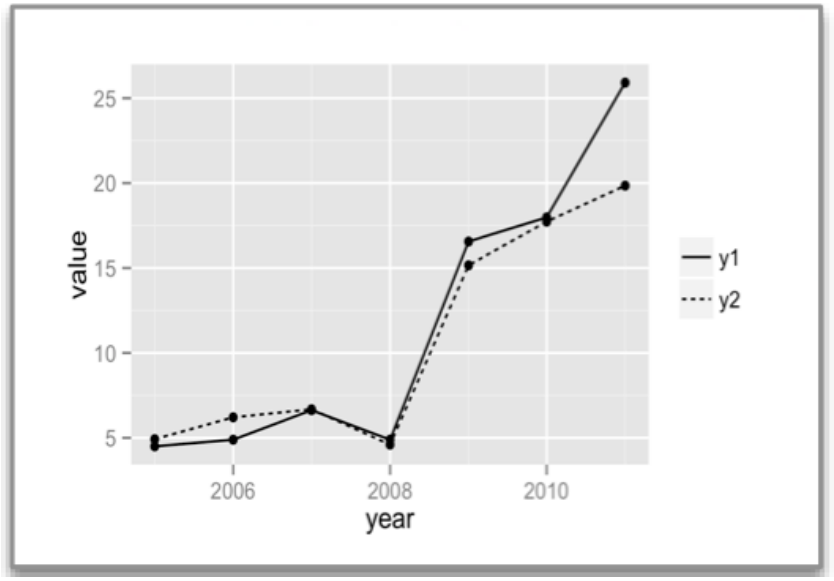
Figura 11. Histograma



Fuente: Josep Curto

- **Gráfico de líneas**: es una representación gráfica en un eje cartesiano de la relación que existe entre dos variables. Se suelen usar para presentar tendencias temporales como se ilustra en la figura 12.

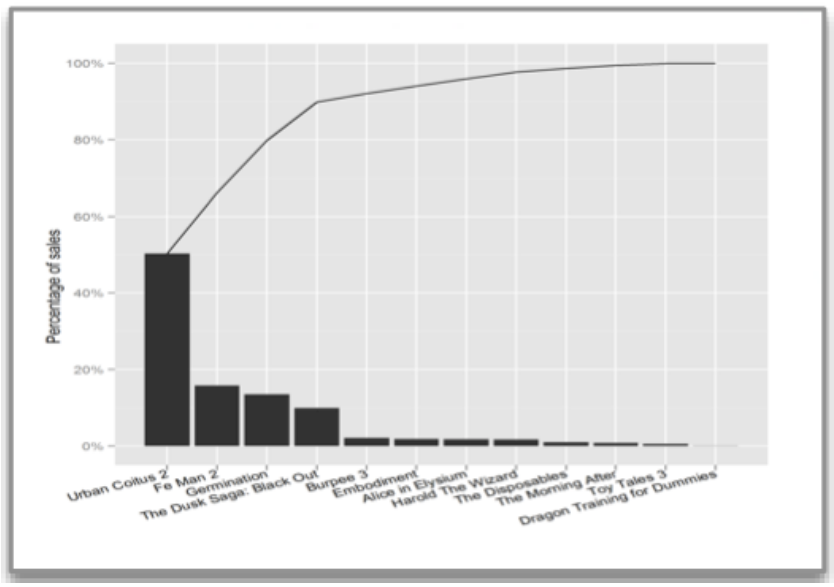
Figura 12. Gráfico de líneas



Fuente: Josep Curto

- **Gráfico de Pareto**: es un tipo de grafico de barras verticales ordenado por frecuencias de forma descendente, que identifica y da un orden de prioridad a los datos como se ilustra en la figura 13.

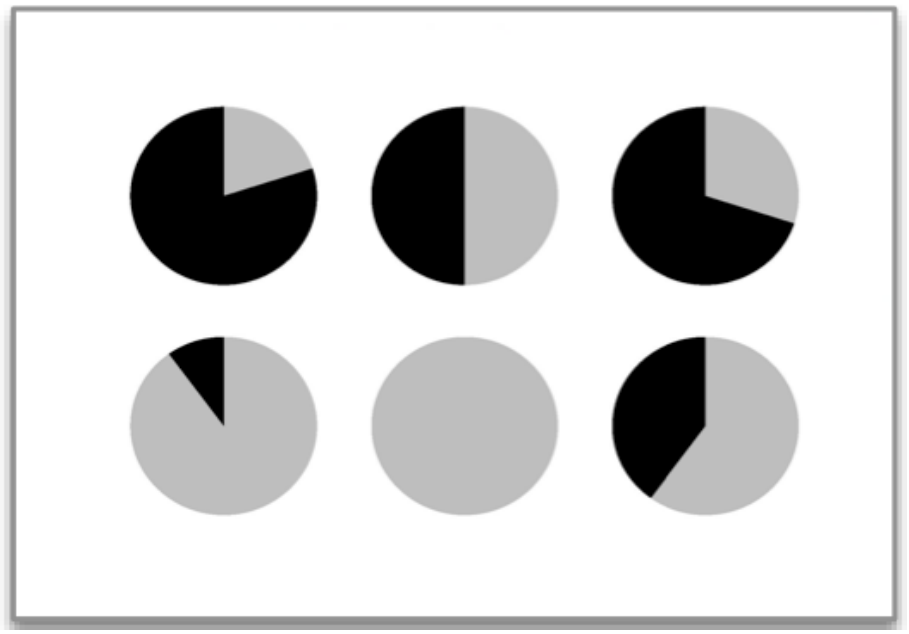
Figura 13. Gráfico de pareto



Fuente: Josep Curto

- **Gráfico de sectores**: es una representación circular de las frecuencias relativas de una variable cualitativa o discreta que permite, de una manera sencilla y rápida, su comparación como se ilustra en la figura 14.

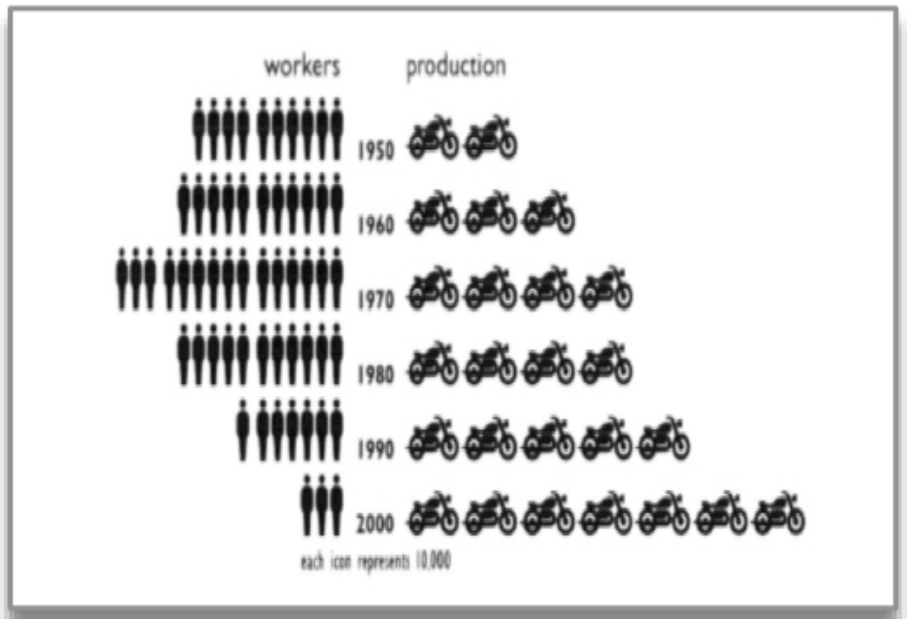
Figura 14. Gráfico de sectores



Fuente: Josep Curto

- **Pictograma**: es un gráfico que representa, mediante figuras o símbolos, las frecuencias de una variable cualitativa o discreta como se ilustra en la figura 15.

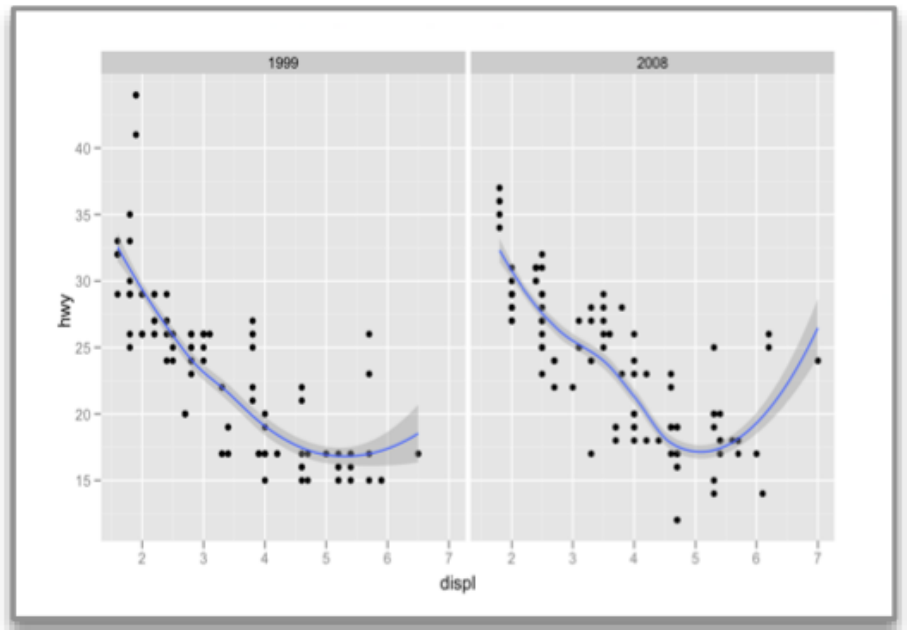
Figura 15. Pictograma



Fuente: Josep Curto

- **Gráfico de dispersión**: muestra en un eje cartesiano la relación que existe entre dos variables e informa del grado de correlación entre dos variables. El tipo de correlación se puede deducir según la forma de la nube de puntos, a saber: nula, lineal o no lineal como se ilustra en la figura 16.

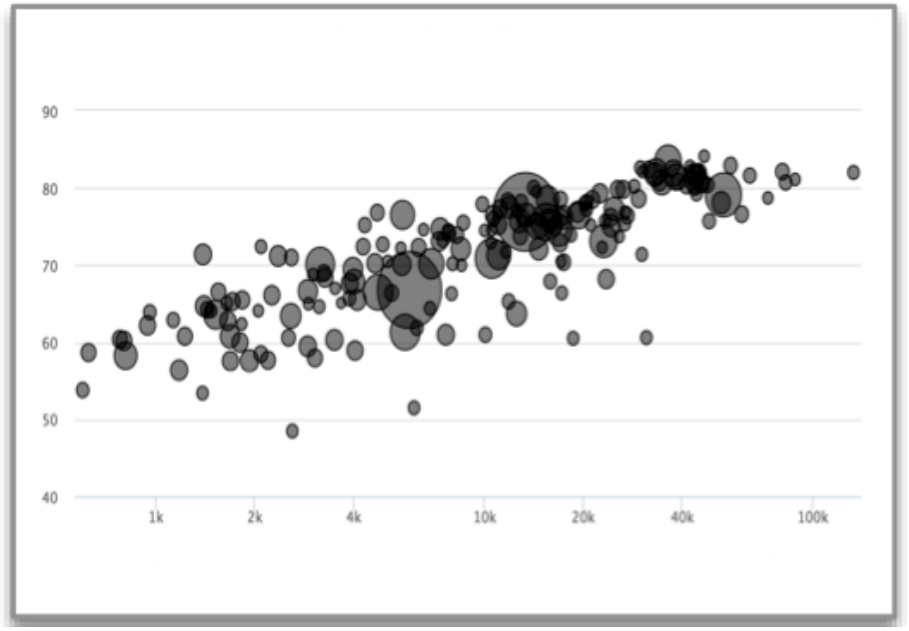
Figura 16. Gráfico de dispersión



Fuente: Josep Curto

- **Gráfico de burbujas**: es una variante del gráfico de dispersión al que se añade una tercera dimensión vinculada al tamaño de los puntos (que se convierten en burbujas) e incluso puede añadirse una cuarta vinculada con el color de cada burbuja. Por lo tanto, permite estudiar la relación de tres variables como se ilustra en la figura 17.

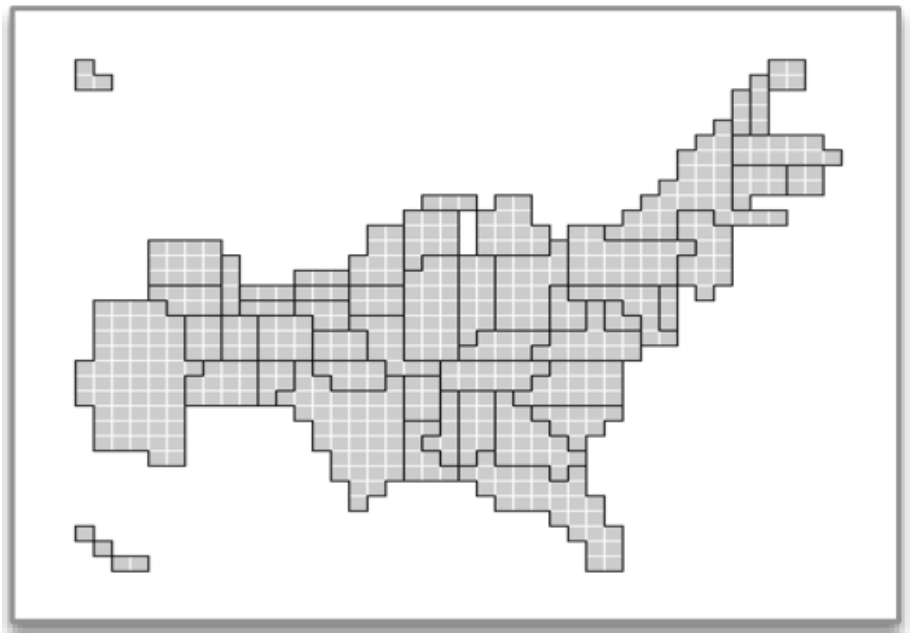
Figura 17. Gráfico de burbuja



Fuente: Josep Curto

- **Cartograma**: es un mapa en el que se presentan datos por regiones bien poniendo el número, bien coloreando las distintas zonas, en función del dato que representan como se ilustra en la figura 18.

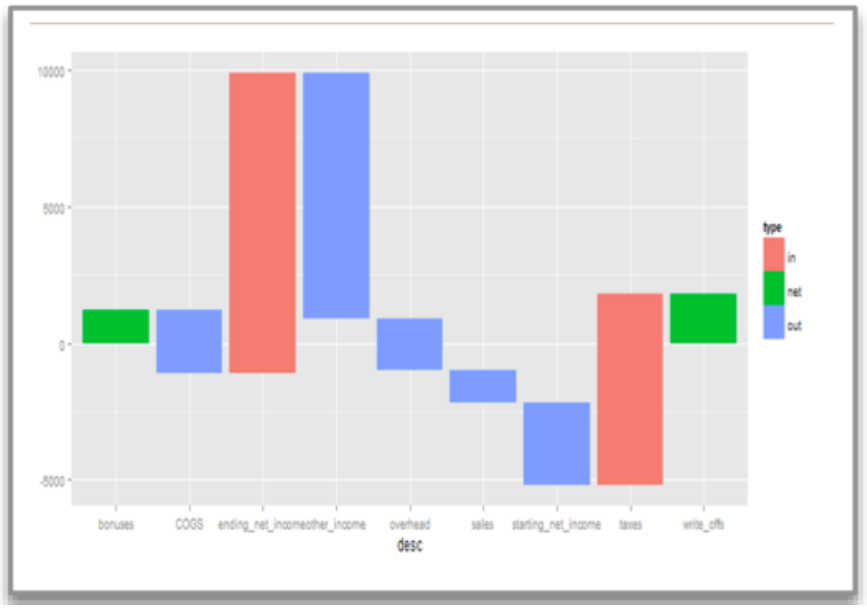
Figura 18. Cartograma



Fuente: Josep Curto

- **Gráficos en cascada**: es un tipo de gráfico normalmente usado para comprender cómo un valor inicial se ve afectado por una serie de cambios intermedios positivos y negativos como se ilustra en la figura 19.

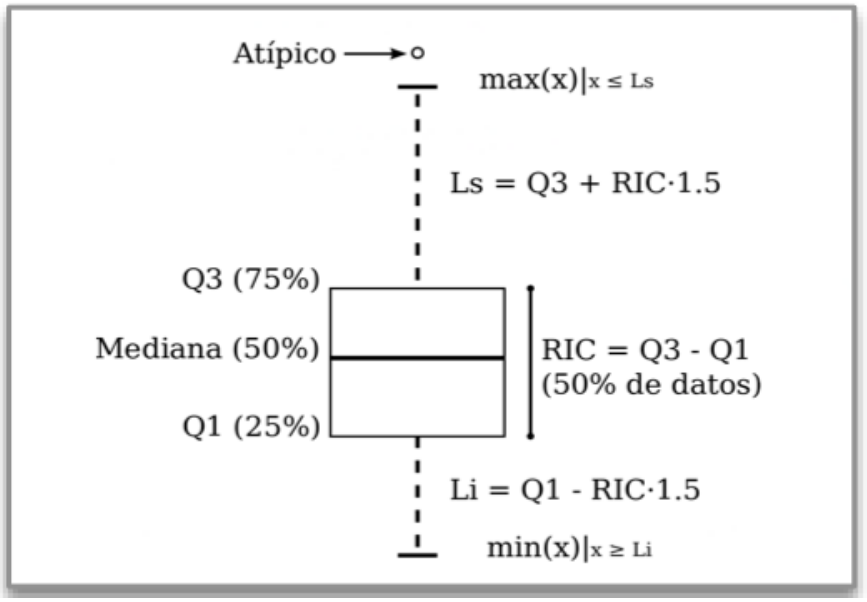
Figura 19. Gráfico en cascada



Fuente: Josep Curto

- **Diagrama de caja**: es un tipo de gráfico que utiliza los cuartiles para representar un conjunto de datos. Permite observar de un vistazo la distribución de los datos y sus principales características: centralidad, dispersión, simetría y tamaño de las colas como se ilustra en la figura 20.

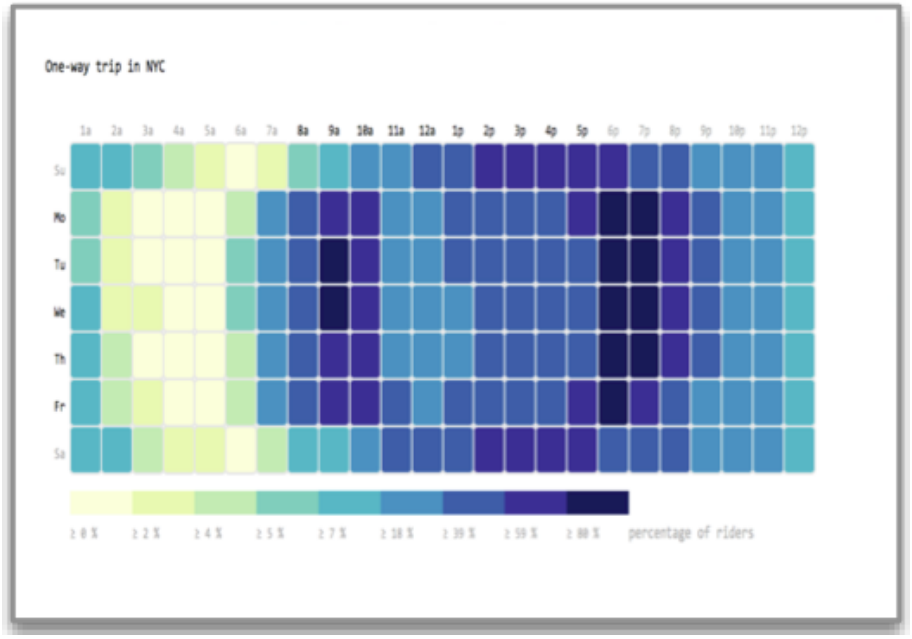
Figura 20. Diagrama de caja



Fuente: Josep Curto

- **Mapa de calor**: es una representación gráfica de los datos donde los valores individuales contenidos en una matriz se representan como colores, como se ilustra en la figura 21.

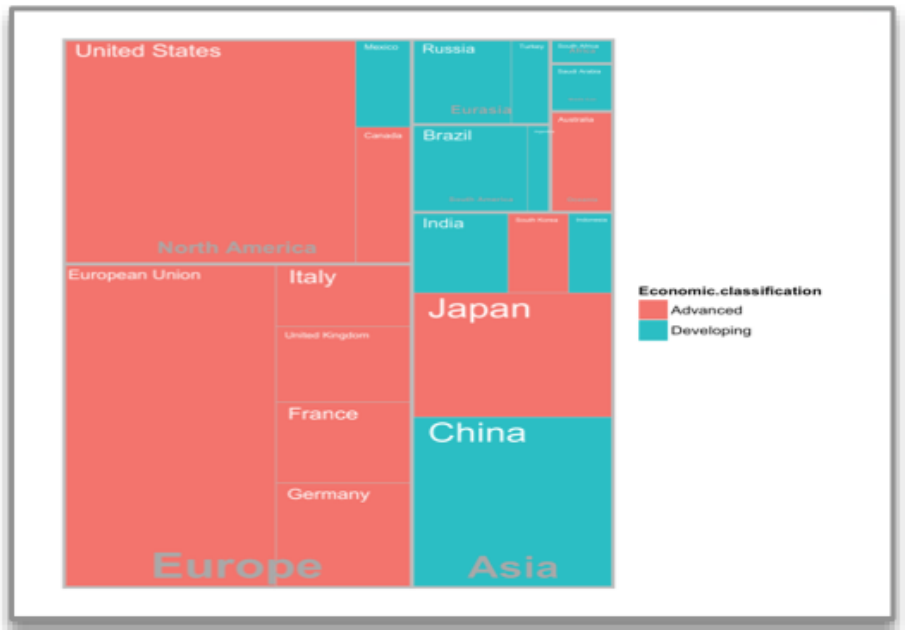
Figura 21. Mapa de calor



Fuente: Josep Curto

- ***Treemap***: es un método para la visualización de datos jerárquicos mediante el uso de rectángulos anidados y de diferentes tamaños como se ilustra en la figura 22.

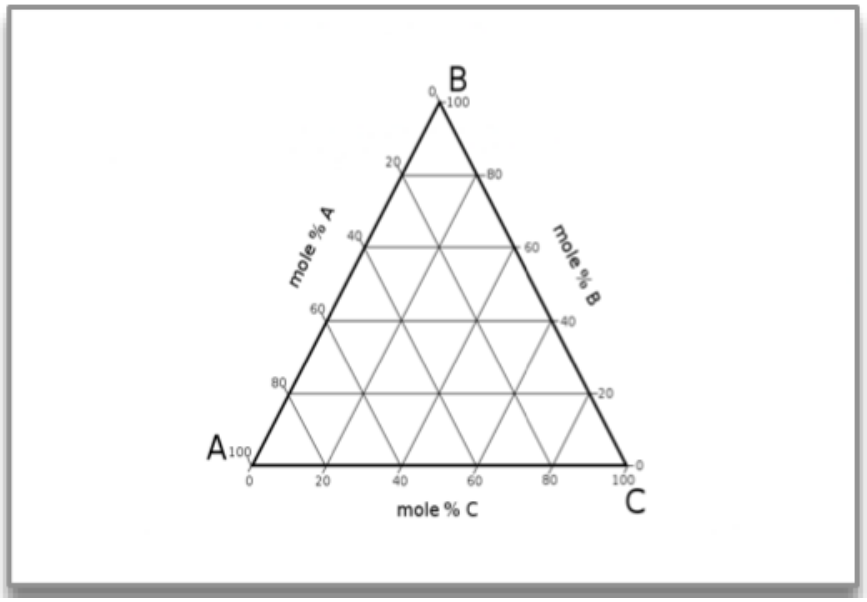
Figura 22. *Treemap*



Fuente: Josep Curto

- **Diagramas ternarios**: son usados para representar el porcentaje relativo de tres componentes donde el único requerimiento es que los tres componentes tienen que sumar un 100 % como se ilustra en la figura 23.

Figura 23. Diagrama ternario



Fuente: Josep Curto

Para escoger un gráfico es necesario seguir un proceso sistemático a través de una serie de preguntas:

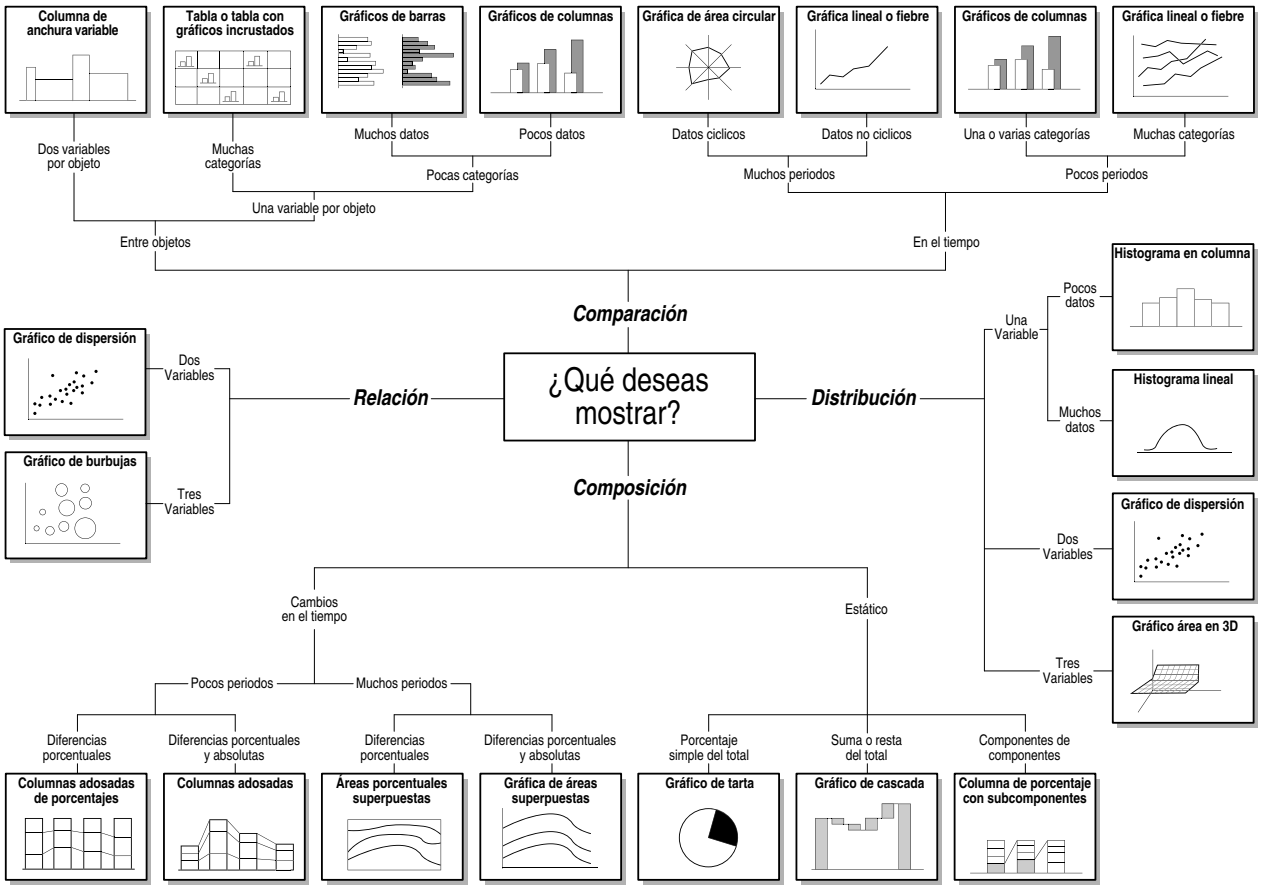
- **¿Qué se desea mostrar?** Tenemos diversas opciones: comparación, distribución, composición y relación.

- **¿Cómo es el dato?** Identificar el tipo de dato: cuantitativo o cualitativo. Así como el tipo de variable: continua o discreta.
- **¿Cuántas variables tenemos?** Podemos tener una o más de una, y tener la necesidad de trabajar con ellas.
- **¿Es estático o cambia en el tiempo?** Es decir, es necesario identificar si hay una dimensión de análisis temporal.
- **¿Depende de la región o del canal?** Es decir, es necesario identificar si hay una dimensión de análisis geográfico.

Se recomienda revisar los criterios presentados en Extreme Presentation* que se ilustran en la figura 24.

***http://extremepresentation.com**

Figura 24. Criterio de selección



Fuente: Extreme Presentation

3.1.6. Ciclo de vida de un informe

Como ya se ha definido, el objetivo de un informe es presentar los resultados de un área o proceso de negocio. En el momento de diseñar un informe, no solo es necesario tener en cuenta la forma y el contenido que tendrá, sino su ciclo de vida para que pueda continuar generando valor para la organización. Es por ello que debemos introducir lo que se conoce como el ciclo de vida de un informe, que se compone de las siguientes etapas:

- **Identificar**: consiste en determinar los aspectos de negocio relevantes para su comprensión e identificar las métricas que representan dichos aspectos y que son relevantes para la compañía y sus gestores.

- **Medir:** consiste en desarrollar o revisar los sistemas de información que recopilan la información necesaria para las métricas. Inicialmente, la compañía debería tener ya implementados estos sistemas, pero no es extraño encontrarse con la necesidad de habilitar este tipo de sistemas.

- **Revisar**: consiste en comprobar que el dato de los sistemas anteriores representa de forma efectiva, válida, completa y con calidad los procesos de negocio, por lo que el sistema de *reporting* posterior tendrá dichas características. En esencia, estamos hablando de gobernanza del dato.

Gobernanza de datos

Cuando hablamos de gobernanza de datos, hacemos referencia a un conjunto de estándares, procesos y políticas que rigen el desarrollo y la utilización de los datos a nivel corporativo.

- **Crear**: consiste en crear el informe y en habilitar su distribución a las partes interesadas.

- **Recopilar**: consiste en recopilar de forma continua el *feedback* por parte de los usuarios así como futuras necesidades.

- **Mejorar**: consiste en implementar en el sistema de *reporting* las mejoras recopiladas en el punto anterior. Estas mejoras pueden ser en forma, contenido, distribución, calidad del dato, etc.

En este ciclo es realmente importante detectar quién va a usar el informe y para qué. A cada usuario, le gusta trabajar la información de una manera y hace o no hace cosas a partir de dicha premisa.

3.2. OLAP

Los informes proporcionan una visión estática del rendimiento de la organización. Para algunos de los usuarios de negocio, normalmente analistas, esto no es suficiente. Necesitan crear sus propios análisis, filtrando la información, creando nuevas métricas, agregando la información respecto a las diferentes perspectivas de negocio, estableciendo relaciones entre hechos, etc.

Este tipo de usuario necesita lo que se conoce como OLAP (*Online Analytical Processing*), término acuñado por Edgar F. Codd. Una manera sencilla de entender qué significa este concepto es que se trata de una tecnología que permite el análisis multidimensional a través de tablas matriciales o pivotantes (como las tablas dinámicas de Excel). Si bien el término OLAP se introduce por primera vez en 1993, los conceptos base del mismo, como, por ejemplo, el análisis multidimensional, son mucho más antiguos.

A pesar de ser una tecnología que ya tiene más de cuatro décadas, sus características y su evolución han producido que la gran mayoría de soluciones del mercado incluyan un motor OLAP.

Es necesario comentar que las herramientas OLAP de los diferentes fabricantes, si bien son similares, no son completamente iguales dado que presentan diferentes especificaciones del modelo teórico.

3.2.1. OLAP como herramienta de análisis

OLAP forma parte de lo que se conoce como sistemas analíticos que permiten responder preguntas como **¿por qué paso?** Estos sistemas pueden encontrarse tanto integrados en sistemas de *business intelligence* como ser simplemente una aplicación independiente.

Es necesario, antes de continuar, introducir una definición formal de OLAP.

Se entiende por OLAP, o proceso analítico en línea, el método para organizar y consultar datos sobre una estructura multidimensional. A diferencia de las bases de datos relacionales, todas las potenciales consultas están calculadas de antemano, lo que proporciona una mayor agilidad y flexibilidad al usuario de negocio.

Una herramienta OLAP está formada por un motor y un visor. El motor es, en realidad, justo el concepto que acabamos de definir. El visor OLAP es una interfaz que permite consultar, manipular, reordenar y filtrar datos existentes en una estructura OLAP mediante una interfaz gráfica de usuario que dispone de funciones de consulta MDX y otras.

MDX

Cuando hablamos de MDX (*Multidimensional Expressions*), hacemos referencia a un lenguaje de consulta para bases de datos OLAP.

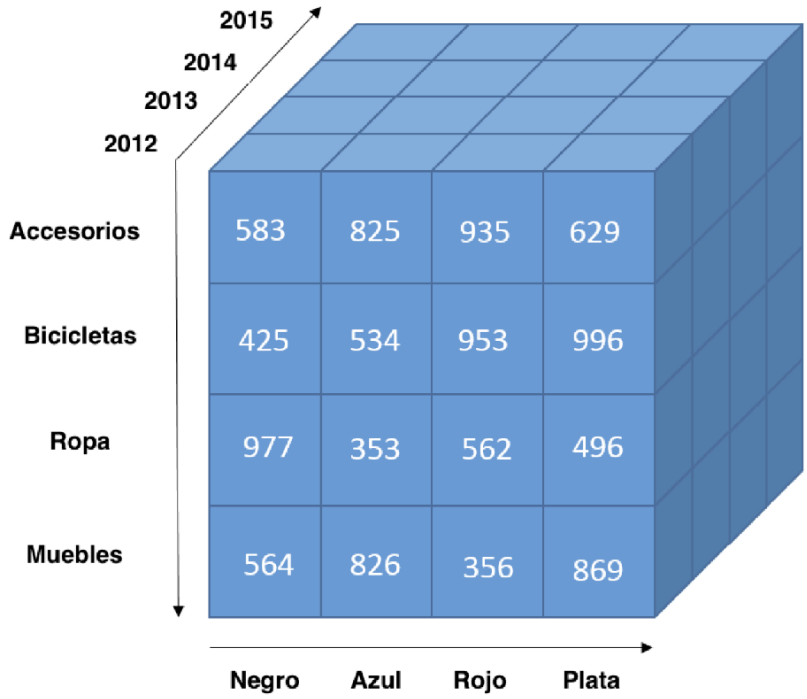
Las estructuras OLAP permiten realizar preguntas que serían sumamente complejas mediante el lenguaje de consulta a base de datos, conocido como SQL (*Structured Query Lenguage*).

Consideremos un ejemplo que nos permitirá entender la potencia de este tipo de herramientas.

Imaginemos que queremos responder a la siguiente pregunta: ¿cuál es el margen de beneficios de la venta de bicicletas en el año 2014? Si tenemos un análisis OLAP (a veces llamado cubo), como el del ejemplo, formado por el año, los productos y su gama de colores, la respuesta es la intersección entre los diferentes elementos. Cabe observar que una estructura de esta forma permite consultas mucho más completas como, por ejemplo, comparar el margen de beneficios de 2013 y 2014 entre diferentes productos, etc. Además, mediante el visor OLAP, proporcionan libertad a los usuarios finales para realizar dichas consultas de forma independiente al departamento de IT.

La figura 25 ilustra este ejemplo.

Figura 25. Cubo



Fuente: Josep Curto

3.2.2. Tipos de OLAP

Existen diferentes tipos de OLAP, que principalmente difieren en cómo se guardan los datos:

- **MOLAP (*Multidimensional OLAP*)**: es la forma clásica de OLAP y frecuentemente es referida con dicho acrónimo. MOLAP es una base de datos multidimensional (o más coloquialmente cubo). En definitiva, se crea un fichero que contiene todas las posibles consultas precalculadas. A diferencia de las bases de datos relacionales, estas formas de almacenaje están optimizadas para la velocidad de cálculo. También se optimizan a menudo para la recuperación a lo largo de patrones jerárquicos de acceso. Las dimensiones de cada cubo son típicamente atributos tales como período, localización, producto o código de la cuenta. La forma en la que cada dimensión será agregada se define por adelantado.

- **ROLAP (*Relational* OLAP)**: es una forma de OLAP donde el dato está en la base de datos relacional, solo en el momento de consulta, se recuperan y se construyen los resultados multidimensionales.

- **HOLAP (*Hybrid* OLAP)**: No hay acuerdo claro en la industria en cuanto a qué constituye el OLAP híbrido, exceptuando el hecho de que es una base de datos en la que los datos se dividen entre almacenaje relacional y multidimensional. Por ejemplo, para algunos vendedores, HOLAP consiste en utilizar las tablas relacionales para guardar las cantidades más grandes de datos detallados; utiliza el almacenaje multidimensional para algunos aspectos de cantidades más pequeñas de datos menos detallados o agregados.

- ***Extreme* OLAP**: este nombre se está empezando a usar en la industria para referirse a un motor OLAP que trabaja sobre alguna de las tecnologías *big data*.

- **DOLAP (*Desktop* OLAP)**: es un caso particular de OLAP ya que está orientado a equipos de escritorio. Consiste en obtener la información necesaria desde la base de datos relacional y guardarla en local. Las consultas y análisis son realizados contra los datos guardados en el escritorio.

- ***In-memory* OLAP**: este enfoque se fundamenta en el uso de la memoria del ordenador para crear la consulta OLAP. Al trabajar solo en memoria, se aceleran las operaciones de acceso y consulta.

Cada tipo tiene ciertas ventajas, aunque hay desacuerdo sobre las especificidades de las ventajas entre los diferentes proveedores:

- MOLAP es mejor en sistemas más pequeños de datos, es más rápido para calcular agregaciones y retornar respuestas, y necesita menos espacio de almacenaje. Últimamente, *in-memory* OLAP está apuntalándose como una opción para MOLAP.

- ROLAP se considera más escalable. Sin embargo, es difícil implementar eficientemente el pre-proceso de grandes volúmenes de datos, lo que hace que se deseche con frecuencia. De otro modo, el funcionamiento de las consultas podría no ser óptimo.

- HOLAP está entre los dos en todas las áreas, pero puede pre-procesar rápidamente y escalar bien.

- *Extreme* OLAP será usado cuando la empresa implemente proyectos de *big data* y esté interesada en tener disponible análisis OLAP sobre grandes volúmenes de datos.

Todos los tipos son, sin embargo, propensos a la explosión de la base de datos, exceptuando *Extreme* OLAP. Éste es un fenómeno que causa la cantidad extensa de espacio de almacenaje que es utilizado por las bases de datos OLAP cuando se resuelven ciertas, pero frecuentes, condiciones: alto número de di-

mensiones, de resultados calculados de antemano y de datos multidimensionales escasos.

Las últimas tendencias en OLAP incluyen la tecnología *in-memory* así como su adaptación a tecnologías *big data* como Apache Kylin* (creado por eBay) o Pinot** (creado por Linkedin).

*http://kylin.apache.org
*http://github.com/linkedin/pinot/

La dificultad en la implementación OLAP radica en la formación de las consultas, en la elección de los datos base y en el desarrollo del esquema. Como resultado, la mayoría de los productos modernos se proveen de bibliotecas enormes de consultas preconfiguradas. Otro problema que puede afectar al cubo es trabajar con datos de baja calidad o incompletos.

3.3. Cuadros de Mando

Tanto los informes como OLAP son herramientas que proporcionan información a los usuarios finales. La gran cantidad de información que normalmente incluyen estas herramientas las puede hacer inadecuadas para usuarios que necesiten tomar decisiones de forma rápida a partir de ellas o que dispongan de poco tiempo para hacer su propio análisis.

El cuadro de mando proviene del concepto francés *tableau de bord* y permite mostrar información consolidada a alto nivel. Se focaliza en:

- Presentar una cantidad reducida de aspectos de negocio.
- Uso mayoritario de elementos gráficos.
- Inclusión de elementos interactivos para potenciar el análisis en profundidad y la comprensión de la información consultada.

El cuadro de mando es una herramienta muy popular dado que permite entender muy rápidamente la situación de negocio y es muy atractiva visualmente. Por ello, todas las soluciones del mercado incluyen este tipo de soluciones. La oferta se diferencia principalmente en la facilidad del proceso de creación del cuadro de mando, en las opciones disponibles de visualización, y en la capacidad de trabajar con flujos continuos de datos y el reflejo de dichos cambios en tiempo real. Los cuadros de mando permiten el análisis visual de la información, lo que se conoce como *Visual Analytics*.

El cuadro de mando suele usarse también para la dirección por objetivos (DPO), que consiste en identificar las áreas clave para la organización y definir los resultados esperados para cada una de ellas y para cada uno de los puestos directivos. Para cada área y directivo, se establecen metas coordinadas y negociadas que se convierten en indicadores de metas que permiten seguir la evolución de la meta en un período de tiempo determinado. En definitiva, el cuadro de mando proporciona, en este escenario, soporte al establecimiento

de planes de acción para lograr los objetivos y controlar su marcha hacia los mismos.

Las últimas tendencias que están afectando a los cuadros de mando incluyen *Data Visualization* y *Data Storytelling*. La primera hace referencia a la inclusión de una mayor cantidad de elementos gráficos para la comprensión del dato y al uso de criterios para el empleo de dichos elementos. La segunda, a que la herramienta permite construir y explicar historias de negocio fundamentadas en datos y hechos para describir qué ha sucedido. No todas las herramientas del mercado incluyen esta tendencia y solo se encuentra en algunos productos innovadores.

Un cuadro de mando permite monitorizar los procesos de negocio dado que muestra información crítica a través de elementos gráficos de fácil comprensión. Este tipo de herramientas, cuya periodicidad de refresco suele ser cercana al tiempo real, es de gran utilidad para todos aquellos usuarios encargados de tomar decisiones diariamente.

Estos sistemas pueden encontrarse integrados en *suites* de *business intelligence* o ser simplemente aplicaciones independientes.

Es necesario, antes de continuar, introducir una definición formal de cuadro de mando.

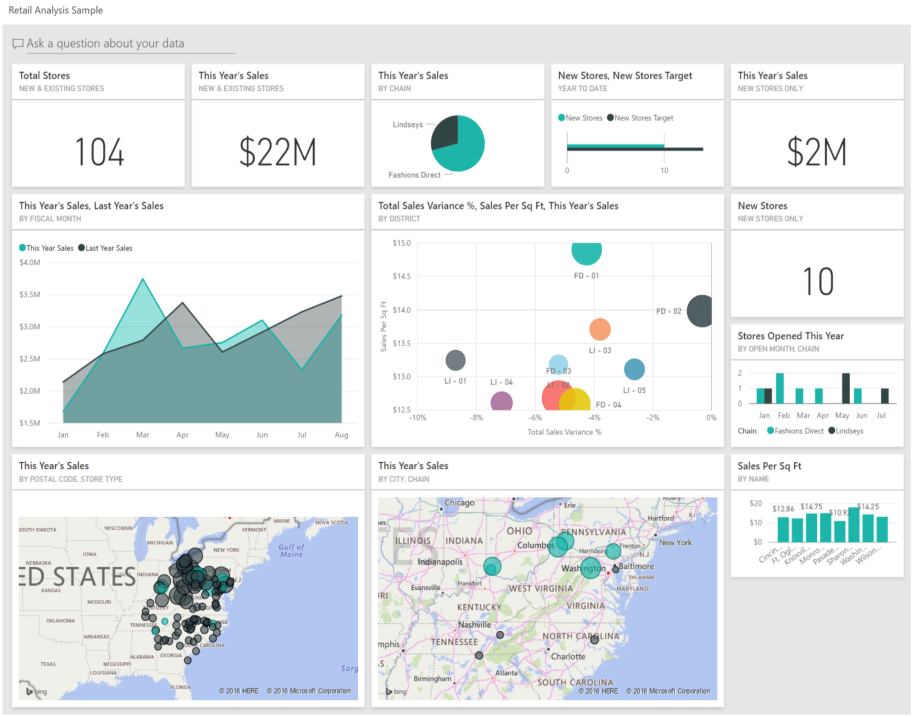
> Se entiende por cuadro de mando o *dashboard* el sistema que informa de la evolución de los parámetros fundamentales de negocio de una organización o de un área del mismo.

La información que se presenta en un cuadro de mando se caracteriza por:

- Usar diferentes elementos (gráficos, tablas, alertas...).
- Combinar los elementos de forma uniforme y precisa.
- Basar la información presentada en indicadores clave de negocio.
- Presentar las tendencias de negocio para propiciar la toma de decisiones.

¿A qué se parece un cuadro de mando? La figura 26 nos presenta un ejemplo que combina algunos de los diferentes elementos anteriores para el control del rendimiento de una compañía de *retail*. Este cuadro de mando permite conocer la evolución de las principales magnitudes financieras, pero al mismo tiempo aquellas específicas del negocio, como ventas por metro cuadrado o expansión en tiendas.

Figura 26. Cuadro de Mando



Fuente: Power BI (MSFT)

La tipología de usuarios que necesita estas herramientas es:

- Alta dirección, con el objetivo de comprender lo que sucede en el negocio
- Gerentes que deben monitorizar procesos de negocio
- Usuarios de negocio que necesitan poder hacer un análisis exploratorio del dato

Por lo tanto, el cuadro de mando aporta valor a nivel estratégico, táctico y operativo.

Principalmente, un cuadro de mando está formado por diversos elementos combinados. El cuadro de mando comparte la mayoría de los elementos de los informes a excepción del hecho que debe incluir **menús de navegación**, que facilitan al usuario final realizar operaciones con los elementos del cuadro de mando.

3.3.1. Proceso de creación de un cuadro de mando

El proceso de crear un cuadro de mando es un proceso iterativo que combina diversos pasos:

- Identificar la necesidad de negocio y los potenciales usuarios del cuadro de mando.

- Elegir los datos que se mostrarán en el cuadro de mando. En este punto, es necesario tener en cuenta las necesidades del usuario final, habiendo mantenido las reuniones necesarias para identificar los requisitos.

- Elegir el formato de presentación. A partir la información que se mostrará y las necesidades del cliente, es posible determinar qué tipo de elemento de un cuadro de mando es el más adecuado. Se recomienda realizar un boceto.

- Integrar, combinar datos y presentarlos conjuntamente. Una vez tenemos los diferentes elementos, se realiza un boceto con todos ellos.

- Planificar la interactividad del usuario.

- Implementación del cuadro de mandos. En este punto entra la herramienta seleccionada e incluye los siguientes pasos:

– Conseguir los datos y formatearlos para conseguir los KPIs.
– Formatear los elementos del cuadro de mando en función de las capacidades de la solución escogida.

3.3.2. *Dashboard* vs. *Balanced Scorecard*

Frecuentemente se confunde el cuadro de mando o *dashboard* con el cuadro de mando integral o *balanced scorecard*. La razón es la similitud de los nombres en castellano. Necesitamos definir este nuevo concepto.

> Se entiende por *balanced scorecard* el método de planificación estratégica basado en métricas y procesos ideado por los profesores Kaplan y Norton, que relaciona factores medibles de procesos con la consecución de objetivos estratégicos.

La teoría del *Balanced Scorecard* surgió en los años 90 como respuesta ante la necesidad de analizar las organizaciones desde un punto de vista diferente al financiero, que se estaba quedando obsoleto. El objetivo era establecer un nuevo modelo de medidas que permitiera conocer mejor las organizaciones.

Para ello, el instituto Nolan Norton patrocinó un estudio de un año cuyo objetivo era definir un *scorecard* corporativo en el que participaron varias compañías de múltiples sectores. De dicho estudio surgió el concepto de *Balanced Scorecard* que organizaba indicadores clave de negocio en cuatro grandes grupos o perspectivas: financiera, cliente, interna e innovación y aprendizaje.

Balanced refleja que los indicadores tratan de ser un equilibrio entre los objetivos a corto y largo plazo, entre las medidas financieras y las no financieras, entre los indicadores de retraso o liderazgo, y entre las perspectivas internas y externas.

Por lo que el *Balanced Scorecard* permite traducir la estrategia de la empresa a un conjunto comprensible de medidas de rendimiento que proporcionen el marco de medida estratégica y de sistema de gestión.

Un cuadro de mando integral está formado por los siguientes elementos:

- **Perspectiva**: punto de vista respecto del cual se monitoriza el negocio. Según esta metodología, toda empresa tiene cuatro perspectivas: financiera, de cliente, de procesos, y de aprendizaje y crecimiento. Si bien puede extenderse o reducirse en número de perspectivas. Vamos a detallar las perspectivas clásicas:

  - **Financiera**: permite medir las consecuencias económicas de las acciones tomadas en la organización. Incorpora la visión de los accionistas y mide la creación de valor de la empresa.
  - **Cliente**: refleja el posicionamiento de la empresa en el mercado o en los segmentos de mercado donde quiere competir.
  - **Interna**: pretende explicar las variables internas consideradas como críticas, así como definir la cadena de valor generado por los procesos internos de la empresa.
  - **Aprendizaje y crecimiento**: identifica la infraestructura que la organización debe construir para crear crecimiento y valor a largo plazo.

- **Objetivos**: que se deben cumplir en cada una de las perspectivas.

- **Líneas estratégicas**: engloban los objetivos que siguen una relación de causalidad.

- **Indicadores**: son principalmente KPIs.

- **Relaciones causa-efecto**: permiten comprender cómo la consecución de un objetivo impacta en otro.

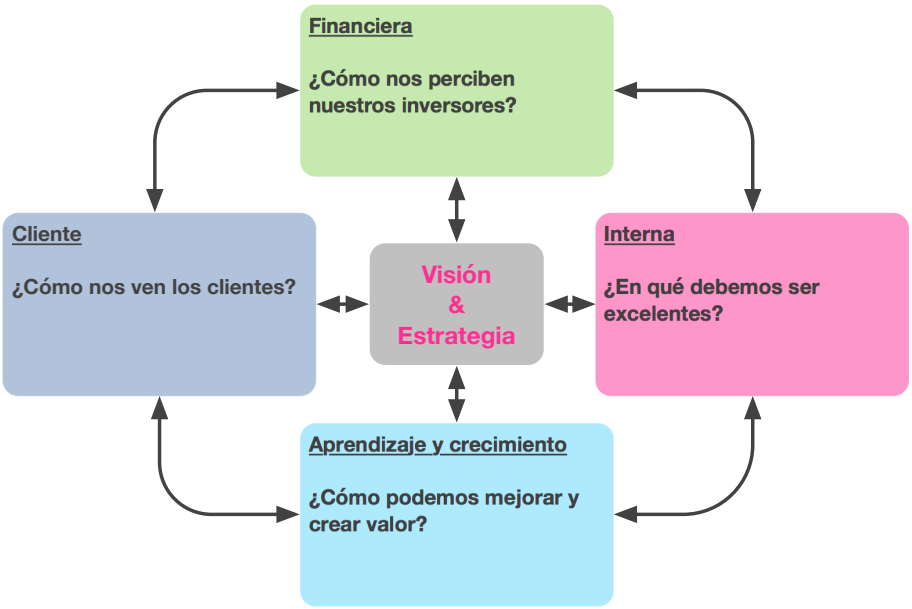
- **Planes de acción**: acciones que se realizan para la consecución de un objetivo.

- **Pesos relativos**: importancia de un objetivo dentro de una perspectiva o de una línea estratégica.

- **Matriz de impacto**: permite dirimir cómo un plan de acción afecta los objetivos y en la medida que lo hace.

La figura 27 representa las diferentes perspectivas tradicionales:

Figura 27. Perspectivas *Balanced ScoreCard*



Fuente: Josep Curto

El proceso de construcción de un cuadro de mando integral es:

- Definir las perspectivas de negocio. Frecuentemente, las perspectivas clásicas son suficientes para representar la estrategia.
- Definir para cada perspectiva los objetivos estratégicos.
- Definir para cada objetivo planes de acción para conseguir dichos objetivos.
- Definir indicadores para monitorizar la consecución de los objetivos.
- Definir las relaciones de causalidad entre los objetivos.
- Identificar las líneas estratégicas a las que pertenecen los objetivos estratégicos.

Este proceso se estructura a través de un mapa estratégico que podemos ver en la figura 28 que representa el proceso anterior.

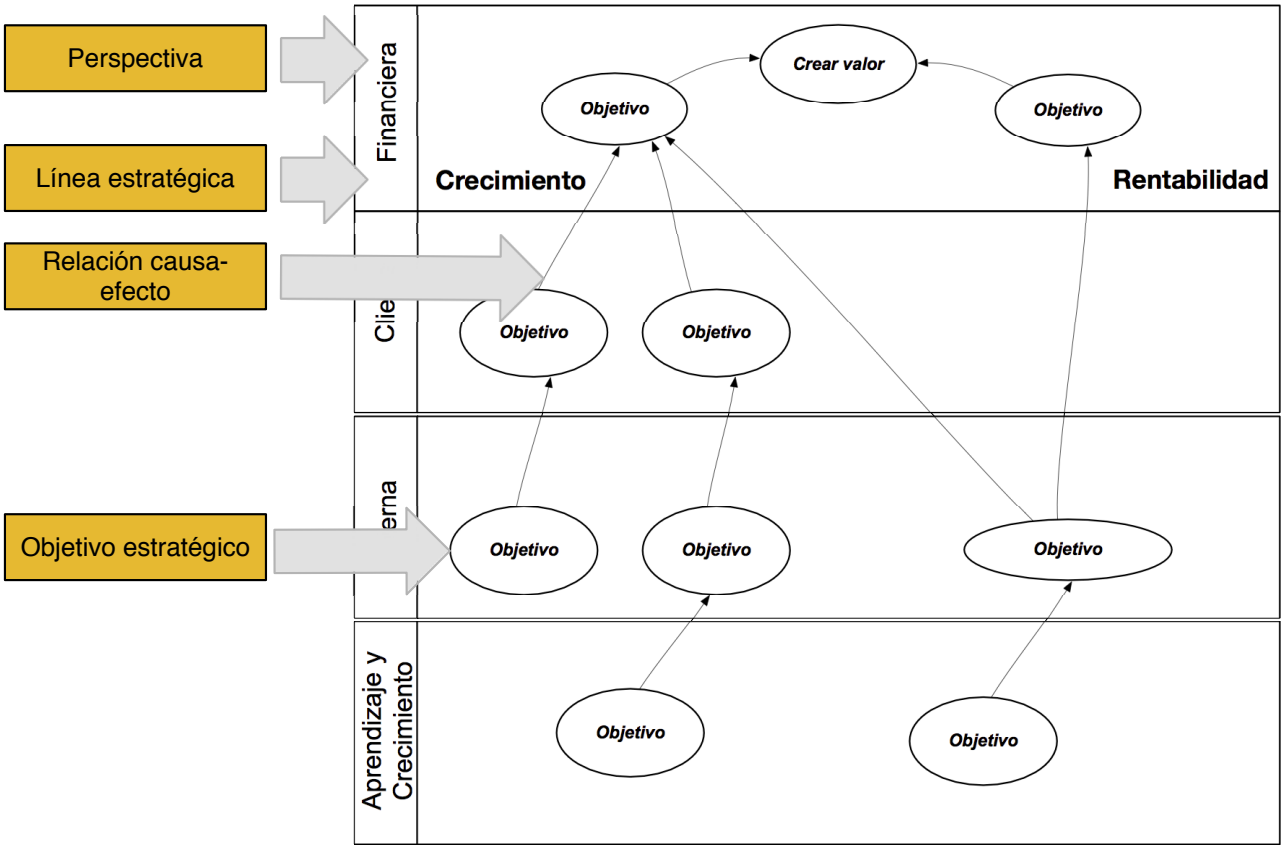
Un punto importante que cabe destacar es que un *Balanced ScoreCard* debe ser flexible y ágil, por lo que la recopilación de información debe llevarse a cabo de forma rápida, sencilla y en el tiempo oportuno para que las acciones que se deriven puedan tomarse de forma eficaz.

La implantación de un cuadro de mando integral proporciona los siguientes beneficios:

- Define y clarifica la estrategia.
- Suministra una imagen del futuro mostrando el camino que conduce a él.
- Comunica la estrategia a toda la organización.

- Permite alinear los objetivos personales con los departamentales.
- Facilita la vinculación entre el corto y el largo plazo.
- Permite formular con claridad y sencillez las variables más importantes objeto de control.
- Constituye un instrumento de gestión.
- Facilita el consenso en toda la empresa al explicitar el modelo de negocio de la organización y traducirlo en indicadores.
- Permite comunicar los planes de la empresa, aunar los esfuerzos en una sola dirección y evitar la dispersión. En este caso, el CMI actúa como un sistema de control por excepción.
- Permite detectar de forma automática desviaciones en el plan estratégico u operativo, e incluso indagar en los datos operativos de la compañía hasta descubrir la causa original que dio lugar a esas desviaciones.

Figura 28. Construcción *Balanced ScoreCard*



Fuente: Josep Curto

Por lo tanto, existen claras diferencias entre un cuadro de mando y un cuadro de mando integral que se recogen en la tabla 2.

Tabla 2. Diferencias entre Cuadro de Mando y Cuadro de Mando Integral

Característica	Cuadro de Mando	Cuadro de Mando Integral
Objetivo	Monitorizar un área de negocio y tomar decisiones operativas y/o tácticas	Definir la estrategia de una organización y enlazar la estrategia con la operativa a través de planes de acción
Elementos	Tablas, gráficos, listas, alertas, menús, mapas...	Perspectivas, objetivos, indicadores, metas...

Está ahora claro que son herramientas diferentes que responden a necesidades distintas y que es necesario no confundirlas. La confusión suele venir del hecho que un cuadro de mando puede soportar DPO, BSC y/o control estadístico.

4. ¿Qué es *business analytics*?

Como se ha comentado anteriormente, lo más importante para una organización no es ser capaz de almacenar o procesar datos, sino generar valor a partir de ellos. El valor toma la forma del análisis.

El análisis se concentra en dos grandes áreas: inteligencia de negocio, con enfoque a conocer el rendimiento pasado, y que ya hemos introducido, y *analítica de negocio*, con enfoque a predecir el rendimiento futuro y conocer patrones ocultos en el dato.

4.1. Definición de *business analytics*

Recuperamos la definición que ya hemos introducido.

> Se entiende por ***business analytics*** el conjunto de estrategias, tecnologías y sistemas para la identificación y comprensión de patrones, y el desarrollo de capacidades predictivas respecto del rendimiento de la organización.

Business analytics permite responder preguntas diferentes de las que la inteligencia de negocio como, por ejemplo:

- ¿Por qué pasó?
- ¿Que pasará?
- ¿Qué pasará si cambiamos X?
- ¿Qué patrones ocultan los datos que no hemos identificado?

Para poder responder a este tipo de preguntas, *business analytics* incluye diferentes tipos de análisis que revisaremos a continuación.

4.2. Tipos de *business analytics*

En la analítica de negocio tenemos diferentes tipos de análisis. Destacamos los siguientes, aunque no es una taxonomía exhaustiva ni exenta de solapamientos:

- **Análisis estadístico/cuantitativo**: rama de las matemáticas que investiga la recolección, el análisis, la interpretación y la presentación de datos de una muestra representativa; busca explicar las correlaciones y dependencias de un fenómeno físico o natural, de ocurrencia en forma aleatoria o condicional. El análisis cuantitativo es un conjunto de técnicas de análisis estadístico que puede incluir, entre otros, el análisis cuantitativo del comportamiento.

  El departamento comercial puede usar el análisis estadístico para analizar si hay una correlación entre las ventas y la época del año.

- **Minería de datos**: es una técnica que permite la extracción de información y conocimiento a partir del dato.

  El departamento comercial puede usar la minería de datos para agrupar los clientes fundamentándose en múltiples atributos al mismo tiempo como ventas, canal, perfil demográfico, etc.

- **Minería de textos**: es una técnica que permite la extracción de información y conocimiento a partir de texto.

  El departamento comercial puede usar la minería de texto para analizar las opiniones de los clientes compartidas en las redes sociales.

- **Minería de procesos**: es una técnica que permite el análisis de los procesos de negocio basado en *logs* de eventos.

  El departamento comercial puede usar la minería de procesos para analizar el rendimiento del proceso comercial desde que el cliente muestra interés hasta que se realiza la compra.

- ***Machine Learning***: conocida también como aprendizaje automático, es una rama de la informática que ha evolucionado desde el estudio y reconocimiento de patrones hacia la inteligencia artificial. Se fundamenta en diferentes tipos de algoritmos clasificados en aprendizaje supervisado, no supervisado y basados en refuerzos.

  Una organización puede usar *machine learning* para crear un sistema de recomendación de productos.

- **Inteligencia artificial**: es un área multidisciplinar que combina computación, matemáticas, lógica... y que busca el diseño de sistemas capaces de resolver problemas cotidianos por sí mismos, utilizando como paradigma la inteligencia humana.

  Una organización puede usar la inteligencia artifical para crear un agente de conversación (*chatbot*) que automatiza y simula interacciones humanas con los clientes para, por ejemplo, proporcionar una respuesta rápida a peticiones de información de producto o servicios.

Sistemas cognitivos

Dentro de los sistemas que buscan simular la inteligencia humana, destacan los sistemas cognitivos. La computación cognitiva hace referencia a *hardware* y *software* que simula el funcionamiento del cerebro humano para tomar decisiones. El aprendizaje se fundamenta en instrucciones y experiencia.

- **Analítica de contenidos**: es una técnica que permite la extracción de información y conocimiento de contenido, como pueden ser imágenes o videos. El foco no solo está en la extracción de valor, sino también en la composición automática de contenidos personalizados.

  Una organización puede usar la analítica de contenidos para personalizar las comunicaciones con los clientes.

- **Analítica de grafos**: es una técnica que permite la extracción de información y conocimiento de datos estructurados como un grafo.

  Una organización puede usar la analítica de grafos para detectar a los seguidores más relevantes de la organización en las redes sociales.

- **Analítica visual**: es una técnica que habilita la exploración de datos y la detección de patrones a través de técnicas de visualización.

  El departamento comercial puede usar la analítica visual para analizar el rendimiento de las diferentes zonas geográficas.

- **Modelización predictiva**: es una técnica para la representación de modelos mediante técnicas estadísticas o matemáticas (como ecuaciones diferenciales), que permite identificar representaciones y hacer predicciones.

  El departamento comercial puede usar la modelización predictiva para identificar los factores que inciden en la compra de los productos y servicios, y estimar qué va a suceder en periodos siguientes.

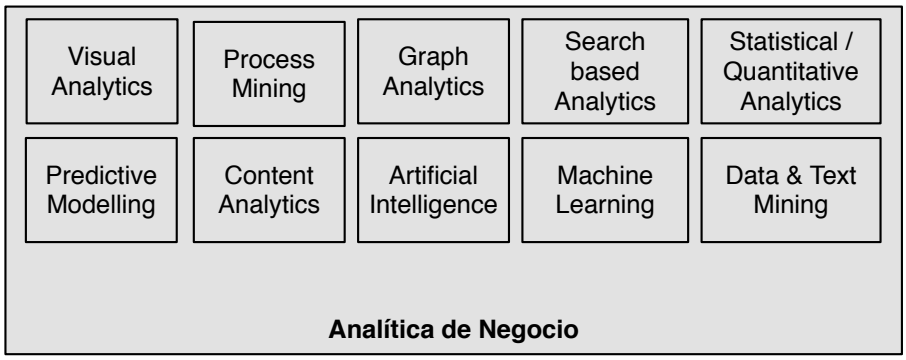
Ecuaciones diferenciales

Cuando hablamos de ecuaciones diferenciales, estamos haciendo referencia a una ecuación matemática que relaciona una función y sus derivadas.

La estadística, la inteligencia artificial, *machine learning* y la minería de datos y texto son disciplinas relacionadas y, en realidad, habilitan los casos de uso presentados en esta taxonomía.

La figura 29 ilustra las componentes en la analítica de negocio.

Figura 29. Analítica de negocio



Fuente: Josep Curto

En general, cuando hablamos de *business analytics* hacemos referencia a soluciones encapsuladas que solo es necesario parametrizar. Estas soluciones están

optimizadas para un sector y un proceso. Por ejemplo, podemos hablar de análisis de la tasa de abandono para el sector de las telecomunicaciones o la detección del fraude en el sector financiero.

Cuando hablamos de aplicar esta técnica –y para ello es necesario desarrollar una solución *ad hoc*–, es cuando estamos considerando lo que se conoce actualmente como *data science*.

4.3. Beneficios de *business analytics*

La implantación de *business analytics* proporciona diversos beneficios entre los que podemos destacar:

- Permitir a los usuarios de negocio plantear hipótesis y validarlas. Como, por ejemplo, cuál es el factor más relevante para el comportamiento de nuestros clientes.

- Poder trabajar con escenarios múltiples y comparar sus resultados. Como, por ejemplo, usando árboles de decisión para ver qué test es el más probable ante un cuadro de síntomas de un paciente.

- Poder encontrar y analizar patrones en los datos. Como, por ejemplo, usar la segmentación de clientes para definir diferentes estrategias de precios para los clientes.

- Poder automatizar tareas manuales basadas en reglas y patrones. Como, por ejemplo, la identificación de preguntas repetidas o similares en Quora.

- Poder hacer recomendaciones de productos y servicios de forma automática. Como, ejemplo, los sistemas de recomendación de Spotify o Netflix.

- Encontrar los motivos reales detrás de un suceso. Como, por ejemplo, en qué punto de la red y por qué ha fallado la retransmisión de un partido de fútbol *online*.

- Poder hacer acciones preventivas. Como, por ejemplo, en la detección de las áreas con mayor potencialidad de crímenes en Río durante los Juegos Olímpicos de 2016.

5. El nuevo contexto de negocio

El uso de dato para tomar mejores decisiones no es nuevo. De hecho, desde hace tiempo las organizaciones se han estado anclando en estrategias como la inteligencia de negocio y/o la analítica de negocio. Pero una serie de condiciones en el mercado han propiciado que sea necesaria una nueva estrategia para el análisis de datos: ***big data***.

En este apartado, nos centraremos en comprender cuáles son estas nuevas condiciones del mercado, qué ha cambiado de la naturaleza del dato y, por último, discutiremos por qué las tecnologías clásicas de *business intelligence*, basadas en el *data warehouse* y las bases de datos relacionales, muchas veces no son suficientes para manejar el nuevo entorno de datos y de negocio.

5.1. Qué ha cambiado desde el punto de vista de negocio

En las últimas décadas, las tecnologías de la información poco a poco han ido cogiendo mayor relevancia en las organizaciones. Se han transformado en un componente básico para las operaciones automatizando, por un lado, parte o incluso todo el proceso y, por otro, proporcionando soporte a las diferentes necesidades departamentales (desde finanzas hasta *marketing*).

Ésta ha sido una progresiva transformación digital y aún muchas empresas están en este proceso de profundo calado. En los últimos años, la transición se ha acelerado por diversos factores como la democratización de Internet, el advenimiento de las redes sociales, la emergencia de los dispositivos inteligentes y/o el despliegue del Internet de las Cosas. Y como resultado, las organizaciones se encuentran en un periodo de competitividad y evolución disruptiva basada en TI* y fundamentada principalmente en cuatro factores: el *cloud*, social, movilidad y analítica. Expliquemos estos factores tecnológicos:

***A este período, Gartner lo denomina el nexo de fuerzas; IDC lo explica como la tercera plataforma y Cognizant simplemente usa una palabra, SMAC, acrónimo de *Social, Mobile, Analytics* y *Cloud*.**

- ***Cloud***: hace referencia a tecnologías que permiten consumir recursos TI (desde almacenamiento hasta un CRM) gestionados por terceros y que frecuentemente comparten su uso.
- **Social**: hace referencia a las tecnologías que permiten facilitar las interacciones sociales.
- **Movilidad**: hace referencia a las tecnologías que habilitan acceso a información e interacciones con independencia de la localización.
- **Analítica**: hace referencia a aquellas tecnologías que maximizan la utilidad del dato.

CRM

es el acrónimo de *Customer Relationship Management*, que hace referencia a la gestión de la relación con clientes.

Estos cuatro factores provocan que los modelos de negocio sean diferentes o, incluso, que se generen nuevos. El *cloud* permite que tengamos flexibilidad en la implementación, en el despliegue, en la escalabilidad y en la globalización; lo social redefine la forma en la que interactuamos con clientes, empleados y proveedores; lo móvil amplía los canales de interacción y desdibuja el perímetro de lo que conocemos como empresa; y, por último, la analítica significa ya no solo que podemos conocer lo que pasa en la organización sino utilizar la información como fuente de ventaja competitiva. En definitiva, es una transformación de la relación entre personas, negocio y tecnología.

Como resultado, hemos asistido a la explosión de nuevas formas de acercarse al mercado y generar valor para el cliente y la organización. Por ejemplo, sabemos de empresas que han conseguido crear sistemas de recomendación para sus clientes, como Amazon, diseñar productos basados en preferencias, como Netflix, o identificar el riesgo crediticio basado en fuentes tan dispares de información como las redes sociales o las compras en eBay y Amazon, como Kreditech*. Aunque compañías como Facebook, Google o Netflix acaparan la atención por sus avances en el uso de las tecnologías de datos, la realidad es que estamos viviendo una revolución de amplio espectro y muchas otras empresas ya están apostando por la implementación de este tipo de proyectos.

***http://www.kreditech.com**

De hecho, es posible encontrar ejemplos en múltiples sectores; estas aplicaciones tienen múltiples formas y colores, y frecuentemente están profundamente especializadas. Por ejemplo, en el contexto de los *Massively Multiplayer Online Game* (MMOG), o videojuegos multijugador masivo en línea, empresas como Jagex* ya monitorizan las transacciones de micropagos y el funcionamiento de los sistemas que soportan las operaciones usando tecnologías de *big data*. En el sector del deporte, equipos como el FC Barcelona analizan grandes cantidades de datos en diferentes formatos (vídeos, estadísticas, datos geolocalizados, etc.) para comprender mejor el rendimiento propio como equipo y de forma individual, así como el de los equipos contrarios, y diseñar consecuentemente estrategias más eficientes para ganar. En el sector de la agricultura, *big data* permite mejorar la eficiencia de los sistemas de riego, al ser la pieza clave para integrar y analizar datos de estaciones meteorológicas, informes de plagas y enfermedades, sensores en plantas, bocas de riego y suelo de parcelas, y sistemas de información como ERP, como en el caso de la bodega Luna Beberide**.

ERP

es el acrónimo de *Enterprise Resource Planning*, que hace referencia a la gestión de los recursos de una organización.

***http://www.jagex.com**

****http://www.lunabeberide.es**

Pero, como ocurre cada vez que aparece una nueva tecnología innovadora y de vanguardia que tiene el potencial de transformar profundamente la sociedad, no resulta sencillo llevar a buen puerto la implementación. Y una primera pregunta surge: **¿en qué medida ha cambiado el dato?**

5.2. La naturaleza del dato

Tal y como hemos comentado, estamos viviendo una explosión en la complejidad del dato. Para entender esta complejidad es necesario hablar sobre la naturaleza, hacer un inciso sobre qué entendemos por las magnitudes físicas del dato y entrar en detalle en dos puntos cada vez más relevantes: dónde se encuentran los datos importantes para una organización y el crucial rol de los metadatos.

5.2.1. Las magnitudes físicas del dato

Hay tres magnitudes físicas del dato: volumen, velocidad y variedad.

- Cuando hablamos de **volumen**, hacemos referencia al tamaño del conjunto de datos creado diariamente. En apenas una década, las organizaciones han pasado de trabajar con Terabytes a tener que lidiar con Petabytes o magnitudes superiores.

- Cuando hablamos de **velocidad**, hacemos referencia tanto al procesamiento de datos como a su latencia. El primero hace referencia a la cantidad de datos en movimiento (medida en términos de Gigabytes o Terabytes por segundo). El segundo hace referencia a la suma de retardos temporales que se aplica a la ingestión de datos y el análisis de los mismos de forma separada o conjunta (medido en milisegundos). Esto implica tratar con datos desde procesos *batch* hasta en tiempo real y/o *streaming*.

- Cuando hablamos de **variedad**, hacemos referencia tanto a la cantidad de fuentes diferentes que se deben combinar (respecto a formatos como video, audio, texto...), como a la heterogeneidad del dato (siendo estos estructurados, semiestructurados o no estructurados).

byte

Cuando hablamos de byte, hacemos referencia a una unidad de medida de información digital. Hablaremos de múltiples bytes:
Gigabyte (GB) 10^9 bytes
Terabyte (TB) 10^{12} bytes
Petabyte (PB) 10^{15} bytes
Exabyte (EB) 10^{18} bytes
Zettabyte (ZB) 10^{21} bytes
Yottabyte (YB) 10^{24} bytes

Latencia

Cuando hablamos de latencia, hacemos referencia a la suma de retardos temporales en la captura, almacenamiento, procesamiento y análisis del dato.

Más allá de las magnitudes físicas es posible encontrar otras características como:

- **Veracidad**, que hace referencia a la incertidumbre en el dato producto de su baja calidad, la ambigüedad en su definición o simplificaciones en su modelización.

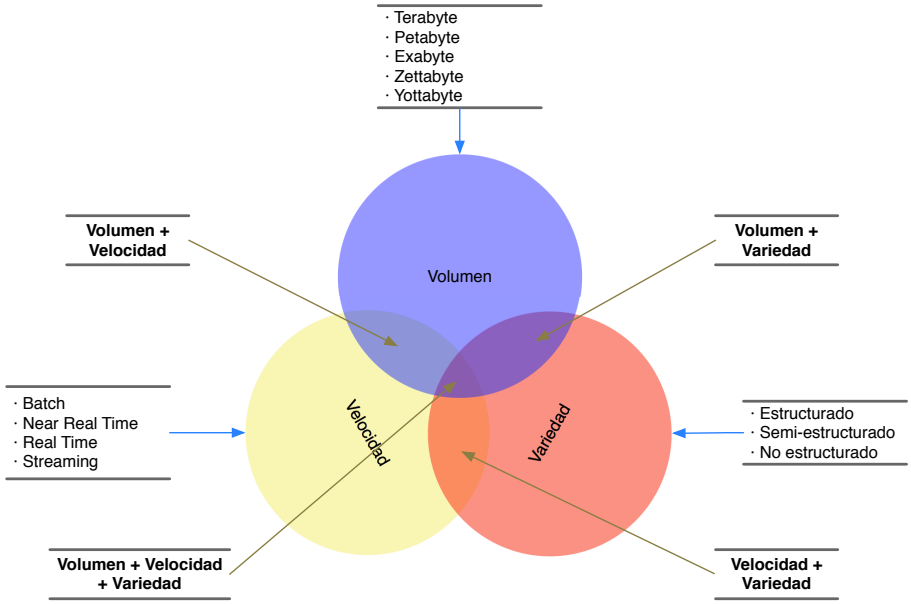
- **Variabilidad**, que hace referencia a que los flujos de datos pueden tener comportamientos erráticos o inconsistentes en ciertos períodos.

- **Vinculación**, que hace referencia a la dificultad de relacionar diferentes y dispares fuentes de datos.

Cabe comentar que no siempre nos encontramos con estas tres últimas características y dependen de la naturaleza del problema que se pretende resolver. Es por ello, que solo se habla de las tres primeras, conocidas como las 3 Vs.

Diferentes problemáticas de negocio tendrán asociadas diferentes combinaciones de estas magnitudes como se ilustra en la figura 30.

Figura 30. Magnitudes físicas del dato



Fuente: Josep Curto, adaptado de SmartDataCollective

Hay un último punto que debemos comentar respecto a la naturaleza del dato: el **valor**. Aunque hayamos discutido cómo podemos entender la complejidad del dato, lo importante no es si una organización es capaz de gestionar el dato en reposo, en movimiento y/o en sus múltiples formas y fuentes. Lo más relevante es cómo una organización es capaz de generar valor a partir del dato y qué impacto tiene para el negocio y para los clientes. En una primera instancia, este valor puede tomar diferentes formas:

1) Toma de decisiones: el uso del dato nos permite tomar mejores y/o más rápidas decisiones, lo que se traduce en que la organización es más competitiva en su respectivo mercado. Es decir, somos capaces de tomar decisiones informadas.

2) Ingresos: el uso del dato permite mejorar los ingresos en líneas de negocio o habilita la creación de nuevas.

3) Costes: el uso del dato permite optimizar nuestros procesos de negocio tanto a nivel de sistemas como de personas, lo que implica que podemos hacer más con menos.

A posteriori analizaremos más en profundidad los casos de uso para el dato.

5.2.2. ¿Dónde se encuentran los datos relevantes para el negocio?

Hemos hablado en el subapartado anterior sobre el aspecto más importante del dato: su valor. O lo que es lo mismo, que el dato sea relevante para el negocio. Ante el nuevo contexto, las organizaciones necesitan trabajar con el dato dejando atrás la noción de que la información de valor se encuentra tan solo en el seno de la organización. Este hecho obliga a pensar no solo en las magnitudes del dato sino también en el origen de partida de los datos. Por lo tanto, debemos hablar de datos internos y externos.

- **Datos internos**: hace referencia a datos que pertenecen a la organización. Dentro de los datos internos tenemos aquellos que ya existen o se crean en los propios sistemas de información de la organización (como pueden ser el ERP y/o el CRM), o bien que se están capturando y almacenando mediante mecanismos automáticos a través de diferentes estrategias como el *crowdsourcing*, sensores y/o dispositivos de monitorización (como un podómetro con localización geográfica).

Crowdsourcing

Cuando hablamos de *crowdsourcing*, hacemos referencia al proceso de obtener servicios, ideas y contenido a través de la participación de una gran masa de personas.

- **Datos externos**: hace referencia a datos de terceros y que deben ser conseguidos por la organización. Estas fuentes de datos pueden estar disponibles para su compra o ser de libre acceso. Los datos de libre acceso pueden ser, a su vez, de tres tipos: datos capturados mediante técnicas de *crowdsourcing*, datos de redes sociales (como pueden ser Facebook, Twitter o LinkedIn) y *open data*.

Open data

Cuando hablamos de *open data*, hacemos referencia a conjuntos de datos considerados un bien común y que, por ello, son gratuitos, accesibles y bien estructurados para su descarga y análisis. Las tipologías son múltiples: de transporte, financieros, meteorológicos, estadísticos, científicos, culturales y geolocalizados.

5.2.3. Metadatos: más allá del valor del dato

Es necesario hablar de una última potencial fuente de datos que, aunque podría incluirse dentro de la categoría de datos internos, no suele contemplarse dentro de las organizaciones. Estamos hablando del metadato y el valor asociado al mismo. Debemos definir qué es el metadato.

Se entiende por metadatos datos estructurados y codificados que describen características de un objeto, dato o proceso de negocio.

Referencia bibliográfica

Conesa, J.; Curto, J. (2012). *Introducción al business intelligence*. Barcelona: Editorial UOC.

Es decir, no es suficiente generar valor a partir del dato, sinó también a partir de los metadatos vinculados a dicho dato. Podemos hablar de tres grandes categorías de metadatos:

- **Técnicos**: describen los aspectos técnicos vinculados al dato. Como pueden ser las magnitudes del mismo o, por ejemplo, los derechos de propiedad.

- **Operacionales**: hacen referencia a los procesos de captura, transformación, almacenaje, análisis y visualización del dato, incluyendo las fórmulas de cálculo.

- **Atributos**: hacen referencia a los atributos que enriquecen la información sobre el dato. Por ejemplo, en una fotografía encontramos aspectos como el dispositivo con el que se realizó.

El metadato abre la puerta a una nueva gama de análisis del valor y, sobretodo, a comprender de una forma mucho más profunda lo que sucede en una organización. Información que no siempre es relevante para el usuario final, pero sí de suma importancia para el sistema que gestiona el dato.

> Consideremos el email, la llamada o el mensaje de texto que se usa para la comunicación tanto personal como profesional. En este caso, los metadatos son el horario, la fecha en que se envió, la duración, los agentes que participan en la conversación y la localización desde donde se conectó el usuario la última vez, entre otros. Esta información no revela el contenido de las comunicaciones, sino de las transacciones electrónicas mostrando sus patrones, relaciones y comportamientos.

Tras discutir la naturaleza del dato, otra pregunta natural surge: **¿por qué necesitamos una nueva tecnología para analizar el dato?**

5.3. Las limitaciones del *Data Warehouse*

Tradicionalmente, las organizaciones han abordado su necesidad de analizar datos y generar valor a través de dos sistemas interconectadatos: el *Data Warehouse* y la inteligencia de negocio, o *business intelligence* (BI). Sabemos que el *data warehouse* es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, mientras que la inteligencia de negocio es un conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a mejorar la toma de decisiones.

El *data warehouse* ha sido el componente principal para el almacenamiento de datos y el BI para su explotación. Sin embargo, a medida que las organizaciones han ido progresando en su transformación digital, la complejidad del dato se ha ido incrementado y nuevas necesidades han emergido. Éstas son algunas de ellas:

- La toma decisiones necesita integrar datos estructurados, semi-estructurados o no estructurados.

- La toma de decisiones necesita trabajar con estructuras de datos que no son persistentes en el tiempo.

- La toma de decisiones necesita considerar toda la información asociada a un proceso de negocio, lo que se traduce, para algunos de ellos, en grandes cantidades de información, no procesables de forma eficiente.

- La toma de decisiones debe realizarse en tiempo real acelerando la captura y el consumo del dato.

- La toma de decisiones debe fundamentarse en el uso de aplicaciones analíticas dónde el metadato del proceso juega un papel fundamental en la comprensión y el descubrimiento de lo sucedido.

- El dato se reutilizará para diferentes análisis y, por ello, se necesita guardar en bruto o aplicando el mínimo de transformaciones posible.

Una forma de entender este tipo de escenarios es comparar los casos de uso del almacén de datos respecto a los nuevos casos de uso tal y como se recoge en la tabla 3.

Tabla 3. *Data Warehouse* vs. Nuevos escenarios de uso del dato.

Factor	Data Warehouse	Nuevos escenarios de uso del dato
Fuentes de datos	Sistemas corporativos y transaccionales	Fuentes no tradicionales como: sensores, logs, videos, etc.
Volumen	Hasta 100 Terabytes	A partir de 100 Terabytes
Velocidad	*Batch* o procesos que no requieren respuesta inmediata	Respuesta inmediata
Variedad	Principalmente estructurada	De todo tipo
Veracidad	Organizada y de calidad	De calidad variable
Valor	BI y analítica	*Machine Learning*, *Deep Learning* y anteriores
Objetivo	Toma de decisiones	Múltiple, pero destaca la creación de productos y servicios de datos

La gran mayoría de implementaciones de *data Warehouse* han sido creados de forma optimizada para la generación de informes, cuadros de mandos, así como el análisis OLAP. Escenarios enfocados al análisis de rendimiento pasado de una organización y que deben estar fundamentados en información de calidad.

Los nuevos escenarios no han sido tratados por el almacén de datos e incluso no forman parte de sus capacidades y, por lo tanto, como respuesta a dicha necesidad, ha emergido una nueva generación de tecnologías y enfoques que amplía las capacidades de nuestra organización a nuevos casos de uso.

Estamos hablando, por lo tanto, de casos de uso diferentes, aunque complementarios, que es necesario comprender para identificar las capacidades que se deben desarrollar, así como los costes asociados. El *data Warehouse*, en definitiva, permite proporcionar respuestas a preguntas recurrentes en una organización (que son conocidas). Las nuevas tecnologías abren la puerta a escenarios flexibles de análisis en los que la estructura del dato y de las preguntas que cabe responder o bien no son conocidas, o bien cambian frecuentemente.

6. ¿Qué es *big data*?

En 2009, IDC estimó el tamaño de la información digital generada y guardada, a la que llamó el Universo Digital, en 0,8 Zettabytes (ZB) y predijo que para el año 2020 se llegarían a los 35 ZB. Posteriores estudios de la misma compañía han revisado la cifra al alza para dicho año y la han ajustado a 45 ZB, siendo de 8 ZB para el año 2015. Esta revisión en las predicciones ilustra la aceleración fruto de la aparición de cada vez más fuentes que producen y consumen datos, de una mayor incorporación de usuarios a Internet, del despliegue de una mayor cantidad de dispositivos inteligentes y del continuo desarrollo de soluciones y servicios digitales.

Lectura complementaria

Gantz, J.; Riensel, D. (2009). *As the Economy Contracts, the Digital Universe Expands*. New York: IDC

Esta explosión de datos está caracterizada por un crecimiento en las magnitudes físicas del dato: volumen, variedad y velocidad. Se crea un mayor volumen de datos, provenientes de una mayor variedad de fuentes, representados en múltiples formatos y que se deben capturar y consumir a una mayor velocidad. Este nuevo paradigma de los datos se conoce frecuentemente como ***big data***, si bien el nombre produce confusión teniendo en cuenta su referencia a solo una de las magnitudes (volumen). En esencia, estamos hablando de una explosión en la complejidad del dato.

6.1. Definición de *big data*

Se considera que *big data* es un concepto novel, dado que existen múltiples definiciones del mismo*.

*Tal y como se argumenta en el siguiente artículo académico: *Undefined By Data: A Survey of Big Data Definitions*. Fuente: http://arxiv.org/pdf/1309.5821v1.pdf. Y como también ha argumentado Timo Elliot, evangelista de SAP, en su blog: http://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html.

Es necesario comentar que podemos encontrar referencias a la problemática del dato en 2001 cuando Doug Layney** apuntó que el crecimiento de los datos en volumen, variedad y velocidad iba a propiciar la necesidad de invertir en nuevas tecnologías que permitirían capturar, extraer, procesar, guardar y analizar los datos en la nueva era. Pero los orígenes del término pueden encontrarse incluso antes, en la década de los 90, en las conversaciones dentro de la comunidad de Silicon Graphics dirigidas por el científico John Mashley en las que se analizaban las principales tendencias de futuro.

**En aquel momento, este analista pertenecía a MetaGroup, actualmente en Gartner.

El hecho de que existan múltiples definiciones complica su comprensión y la identificación de escenarios dentro de la propia organización. La gran mayoría de ellas incluyen lo que se conoce como las 3Vs del *big data*, que hemos

comentado en el anterior subapartado: volumen, velocidad y variedad, que son magnitudes físicas del dato. Aunque podemos encontrar otras definiciones que incluyen algunas más como, por ejemplo, la veracidad.

Por lo que, en aras de tener un enfoque pragmático, vamos a usar la siguiente definición.

Se entiende por ***big data*** el conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de conjuntos de datos complejos.

Es necesario recordar que, cuando hablamos de almacenamiento, nos referimos a los soportes físicos y de *software* que permiten guardar el dato en estructuras que representan su complejidad; que, cuando hablamos de procesamiento, nos estamos refiriendo a aquellas operaciones que permiten la ingestión, la transformación y la distribución del dato para adecuarlo al consumo; que, cuando hablamos de análisis, hacemos referencia a las técnicas aplicadas para generar valor; y que, cuando hablamos de visualización, hacemos referencia a los mecanismos de consumo de información.

6.2. Tipos de *big data*

La definición de *big data* enmascara, en cierta medida, las complejidades de lo que supone trabajar con datos extremos en términos de su volumen, velocidad y variedad. Ya hemos introducido en el subapartado 5.2 en qué consiste la nueva naturaleza del dato.

En este sentido y, con el objetivo de mejorar la comprensión de la definición de *big data*, es necesario hablar de las diversas tipologías de *big data* existentes.

6.2.1. Clasificación de NIST

De acuerdo al NIST y, en particular dentro de su grupo de trabajo de *big data**, una forma de categorizar *big data* es mediante las necesidades de negocio:

- El modelo de negocio no se puede representar mediante una estructura de datos relacional (es decir, mediante una base de datos relacional).
- El modelo de negocio necesita ser escalable por el crecimiento de datos respecto a su velocidad o volumen.

***http://bigdatawg.nist.gov**

NIST

NIST es el acrónimo de *National Institute of Standards and Technology*, una institución americana que estudia, define y promueve estándares tecnológicos.

La base de esta clasificación es poder identificar correctamente los diferentes escenarios que se derivan de estas condiciones, ya sea mediante recursos internos, ya sea mediante la ayuda de especialistas externos. La combinación de estos dos aspectos nos proporciona tres tipologías de *big data* y un escenario en el que no existe esta necesidad. Aunque es patente el valor que aporta *big data*, no todos los problemas de una organización son necesariamente un problema de datos. Los tipos disponibles se resumen en la tabla 4.

Estructura de datos relacional

Cuando hablamos de estructura de datos relacional, hacemos referencia a un tipo de base de datos que permite establecer interconexiones o relaciones entre los datos guardados en tablas.

Tabla 4. Tipos de *big data*. Fuente: NIST

Tipo	Descripción
Tipo 1	Donde una estructura de datos no relacional es necesaria para el análisis de negocio
Tipo 2	Donde es necesario aplicar estrategias de escalabilidad horizontal para procesar y analizar de forma eficiente el negocio
Tipo 3	Donde es necesario procesar una estructura de datos no relacional mediante estrategias de escalabilidad horizontal para procesar y analizar de forma eficiente el negocio

Escalabilidad

Cuando hablamos de escalabilidad, hacemos referencia a la habilidad de un sistema, red o proceso para reaccionar y adaptarse sin perder calidad, o bien manejar el crecimiento continuo de trabajo de manera fluida, o bien para estar preparado para hacerse más grande sin perder calidad en los servicios ofrecidos. La escalabilidad horizontal está fundamentada en el incremento de nodos del sistema, proceso o red. Mientras que la vertical consiste en añadir más recursos –memoria, disco duro y/o procesadores–.

Por lo que, para una determinada necesidad de negocio, es posible identificar si estamos en un escenario de *big data* o no, y si es necesario este tipo de tecnologías, hecho que cada vez más se erige como un punto relevante y de partida para la implementación de este tipo de proyectos. Esto se resumen en la tabla 5.

Tabla 5. Autoevaluación de la existencia de *big data*. Fuente: NIST

Volumen	Velocidad	Variedad	Escalabilidad horizontal	Estructura no relacional	Tipo de *Big data*
No	No	No	No	No	No
No	No	Sí	No	Sí	Sí, Tipo 1
No	Sí	No	Sí	Quizá	Sí, Tipo 2
No	Sí	Sí	Sí	Quizá	Sí, Tipo 3
Sí	No	No	Sí	Quizá	Sí, Tipo 2
Sí	No	Sí	Sí	Sí	Sí, Tipo 3
Sí	Sí	No	Sí	Quizá	Sí, Tipo 2
Sí	Sí	Sí	Sí	Sí	Sí, Tipo 3

Esta tabla permite evaluar una necesidad de negocio. Una reflexión interesante es que, dentro de una misma organización, pueden plantearse diferentes escenarios de *big data* para resolver diferentes necesidades de negocio, lo que, en definitiva, apunta que serán necesarias diferentes tecnologías de *big data*.

Caso: Infojobs

InfoJobs es uno de los principales portales de empleo en Europa a través de una plataforma *online* que conecta empresas (que publican ofertas) y personas (que buscan nuevas oportunidades para su carrera profesional).

La compañía estaba interesada en poder ofrecer recomendaciones analizando las trayectorias profesionales más habituales de sus usuarios (lo que también se conoce como

crecimiento profesional). Esto se traduce en analizar una gran cantidad de información que debe estructurarse como una red que relaciona candidatos (más de 4 millones), experiencias profesionales (más de 12 millones) y capacidades (más de 18 millones).

Usando la tabla anterior, estamos en una situación en la que priman el volumen y la variedad, y, por lo tanto, es un escenario *big data* de tipo 3.

6.3. ¿Cuándo es necesario *big data*?

La no existencia de una definición formal, una que permita distinguir de forma completamente precisa cuándo una organización está en una situación de necesidad de *big data*, ha generado barreras en la adopción de este tipo de tecnologías.

Hemos visto que hay escenarios en los que no es suficiente trabajar con un *Data Warehouse*. Tenemos también, para los tipos de *big data*, una clasificación que permite dirimir escenarios genéricos. Sin embargo, esto no es suficiente en el contexto de una organización dónde la experimentación no tiene un amplio margen.

En esta aproximación más pragmática a *big data*, las organizaciones están trabajando en cinco grandes categorías de casos de uso. Estos casos de uso son movimientos organizacionales de una estrategia de negocio enfocada a *big data*. Estos casos son:

- Toma de decisiones
- Operaciones e inteligencia operacional
- Validación de hipótesis y resolución de problemas
- Productos y servicios basados en datos
- Comercio de datos

Vamos a explicar en detalle cada uno de estos movimientos organizacionales.

6.3.1. Toma de decisiones

El primer caso de uso es la toma de decisiones. Esta aproximación consiste en la ampliación de las capacidades tradicionales de toma de decisiones mediante las tecnologías de *big data*. Lo que significa que los sistemas de inteligencia de negocio y almacenes de datos corporativos pueden alimentarse o combinarse con los repositorios de *big data*.

Caso: NH

NH es una cadena hotelera con más 400 hoteles en 25 países. Dentro de la estrategia de mejorar el servicio para sus clientes, la compañía selecciona cada año diversos hoteles sobre los que hará mejoras. Las mejoras que se realizan van desde la ampliación del personal hasta la creación de nuevas instalaciones. Para tomar la decisión de "dónde es necesario invertir en este período", NH se ha fundamentado tradicionalmente en dos fuentes de datos:

- Los datos financieros consolidados en el *data warehouse* de la compañía.
- Una serie de encuestas realizadas a los clientes para conocer su satisfacción en relación a los servicios e instalaciones del hotel. Estas encuestas no son exhaustivas y no cubren todos los hoteles ni todos los clientes por los costes asociados a su realización.

En los últimos años, en NH se han dado cuenta de que, para conocer la satisfacción del cliente así como las áreas de mejora, la información relevante se encuentra más allá del perímetro de la organización. Los clientes de NH comparten sus impresiones a través de diferentes canales como pueden ser TripAdvisor, Yelp o Expedia. Es decir, estamos hablando de fuentes de datos externas a la organización y, además, no estructuradas o con diferentes formatos.

El enfoque de la organización ha sido complementar la información financiera en el *data warehouse* con información externa que se almacena y se procesa con tecnologías de *big data* y minería de texto (para extraer los comentarios relevantes para la mejora de los hoteles), y que permite mejorar y complementar la toma de decisiones en un proceso ya existente.

6.3.2. Operaciones e inteligencia operacional

El segundo caso de uso son las operaciones y la inteligencia operacional, que suceden en tiempo real. Esta aproximación consiste en la aplicación de estas tecnologías en el ámbito de operaciones tanto para el control y el análisis de proceso de negocio como para el diseño e implementación de sistemas transaccionales. Este segundo escenario trasciende a la toma de decisiones y permite entender por qué las tecnologías *big data* están llamadas a ser muy relevantes dentro de las tecnologías de información. Es previsible que se integren de forma natural en múltiples aplicaciones.

Estamos hablando, por lo tanto, por un lado, de sistemas de inteligencia y detección de patrones en tiempo real y, por el otro, de sistemas operacionales que, o bien por sus necesidades en escalabilidad, o por su complejidad en el esquema de los datos ya no se fundamentan en tecnologías relacionales.

Caso: Santader/Caixabank

Uno de los puntos más relevantes para muchas organizaciones es cuando interaccionan con sus clientes. Es lo que llamamos momentos de la verdad. Entre estos momentos destaca ese en que el cliente se pone en contacto con una organización para la resolución de una incidencia. En épocas anteriores se ha automatizado el proceso (mediante sistemas de respuesta predefinida) o se ha dividido el servicio en diversas capas dentro y fuera de la organización para tener diferentes niveles de servicio.

En el contexto financiero son varias las entidades financieras españolas que ya usan *speech analytics* para entender las emociones de sus clientes durante sus interacciones en una llamada. Es posible, por lo tanto, detectar cuándo va a disminuir la satisfacción del cliente y actuar consecuentemente.

6.3.3. Validación de hipótesis y resolución de problemas

Uno de los escenarios más importantes es la validación de hipótesis y resolución de problemas. Este escenario consiste en encontrar soluciones para problemas de negocio que no han sido anteriormente abordados en una organización y para los cuales no hay preguntas predefinidas. Es decir, se busca

conocer qué ha sucedido, qué factores son los más relevantes y el porqué. Es necesario crear hipótesis y validarlas a través de la técnica más adecuada y eficiente. Este tipo de aplicación es el equivalente a tener a Sherlock Holmes en casa. Es, en definitiva, un entorno que debe ser lo suficientemente flexible para funcionar en diferentes escenarios de necesidades.

El resultado se puede ser una solución puntual o una propuesta que pase a convertirse en uno de los otros escenarios.

Caso: Sky

Sky es una empresa que produce y distribuye contenidos de video tanto en directo como bajo demanda, con presencia en diversos países europeos. La distribución de contenidos de video se realiza a través de redes informáticas conocidas como *content delivery networks*. Estas redes están formadas por múltiples elementos, tanto pertenecientes a la propia compañía como a terceros, que deben asegurar que la distribución se realiza manteniendo el nivel de calidad contratado por el cliente. Frecuentemente, el proceso de transmisión de video a través de la red se controla y se mide para asegurar su correcto funcionamiento. Este es el caso de Sky. Sin embargo, meses atrás tuvo un gran fallo en su *content delivery networks* durante la transmisión de la jornada futbolística en el fin de semana, que producía errores de acceso al sistema, congelación de la imagen o transmisión de imágenes en baja calidad. A pesar de tener un sistema de monitorización, Sky no conocía los motivos por los que había sucedido esta caída de calidad en el servicio.

El enfoque de la organización ha sido contratar a un experto para investigar lo sucedido. Lo que se traduce en este caso en trabajar con millones de registros en formato *log* e investigar las causas del error. Tras aplicar lo que se conoce como *root cause analysis* a un conjunto de datos almacenados en tecnologías de *big data* por su tamaño, se encontraron las razones del error y se propusieron una serie de mejoras para la red.

6.3.4. Productos y servicios de datos

El cuarto escenario de uso es la creación de productos y servicios basados en datos. El dato se transforma en la pieza angular para mejorar la experiencia de uso del producto y servicio, o para el diseño y despliegue del mismo.

Estamos hablando de modelos de negocio en los que el dato y algoritmos analíticos generan valor para el cliente y la organización, y por ello modifican todos los aspectos primordiales del modelo de negocio.

Caso: Nest

Nest es una empresa que produce dispositivos inteligentes, en particular, termostatos, detectores de humo y cámaras de vigilancia, entre otros. Fue adquirida por Google en 2014.

Los productos creados por Nest son un ejemplo de producto basado en datos y algoritmos. Por ejemplo, el termostato contiene diferentes tipos de sensores para detectar a las personas y los animales presentes en el hogar, así como la temperatura del hogar. Además, va registrando las preferencias de las personas que viven en casa. A saber, a qué hora están en casa y cuál es la temperatura que prefieren. Y lo combinan con información contextual, como dónde se encuentra la casa y de qué época del año se trata. Con todos estos datos, se crea un perfil de preferencias y, a partir de un cierto momento, el dispositivo empieza a trabajar de forma automática. Este proceso automático permite reducir el coste energético.

Existe también otro potencial beneficiario de estos datos, aunque en formato agregado: las empresas productoras y distribuidoras de energía. A partir de los datos de todos los

usuarios de Nest en aquellas zonas geográficas en las que ofrecen servicio, pueden conocer las necesidades energéticas y, por lo tanto, ajustar la demanda.

6.3.5. Comercio de datos

El último escenario de uso es el comercio de datos. El dato se prepara para su venta a terceros. Esto puede incluir diversos procesos como agregación, transformación y distribución del dato o, en el caso de contener información sensible, enmascarar dichos datos para que el conjunto final contenga datos anónimos. Este tipo de uso también puede derivar en tener que diseñar una plataforma *ad hoc*. El dato puede comercializarse en bruto o en la forma de conocimiento.

Caso: Vodafone/TomTom

Vodafone es una conocida compañía que proporciona servicios de telecomunicaciones a nivel mundial. TomTom es una compañía que ofrece productos y servicios de GPS.

Los servicios de TomTom permiten conocer la ruta óptima a un conductor. La calidad de este servicio depende de trabajar con datos actualizados incluyendo accidentes o atascos de tráfico. Por ello, TomTom compra datos de terceros a, por ejemplo, Vodafone.

En este caso particular, Vodafone comercializa los datos agregados, anónimos y geolocalizados de los usuarios de su red. En el caso de tener una gran acumulación de usuarios en un mismo lugar (y estar dicho lugar en una carretera), esto se traduce en una situación de atasco o accidente. Por lo que TomTom puede usar esta información para proponer una ruta alternativa y ofrecer una mejor experiencia de cliente.

7. Tecnologías de *big data*

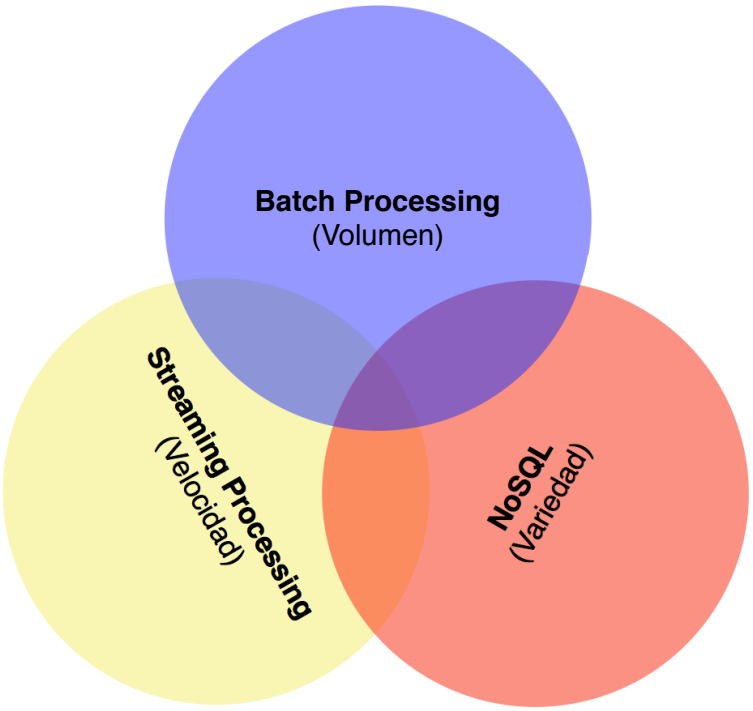
Un primer enfoque para pensar en las tecnologías de *big data* es recuperar las 3Vs presentadas en el subapartado 5.2. Diferentes problemáticas del dato necesitan diferentes paradigmas, tal y como apuntan Casado y Younas.

Lectura complementaria

Casado, R.; Younas, M. (2015). "Emerging trends and technologies in big data processing". *Concurrency and Computation: Practice and Experience*, (nº 27 (8), págs. 2078-2091).

Cuando el volumen es la principal problemática, se usan las tecnologías de *batch processing*; para la velocidad, las de *streaming processing*; y para las de variedad, *NoSQL*, como se ilustra en la figura 31.

Figura 31. Tecnologías de *big data*



Fuente: Casado, R.; Younas, M.

NoSQL

Cuando hablamos de NoSQL hacemos referencia a bases de datos no relacionales. NoSQL es el acrónimo de *Not Only SQL*. SQL es el acrónimo de *Structure Query Language* y hace referencia al lenguaje de consultas de bases de datos relacionales.

Hay una correspondencia directa entre la explosión en la problemática en el dato y la emergencia de una determinada tecnología. De esta forma, tenemos:

- **Tecnologías de procesamiento por lotes o *batch processing***: permiten resolver problemas vinculados con el volumen del dato.

Criteo* es una compañía que ofrece soluciones para *marketing* digital basadas en datos. Esta organización usa las tecnologías de procesamiento por lotes para consolidar datos y optimizar sus algoritmos de análisis de campañas de *marketing*.

*http://www.criteo.com

- **Tecnologías de procesamiento en flujo o (*streaming processing*)**: permiten resolver problemas vinculados con la velocidad del dato.

  Capital One* es una entidad bancaria que ofrece productos y servicios financieros a consumidores. Esta organización usa las tecnologías de procesamiento en flujo para monitorizar la actividad de sus clientes en tiempo real.

  *http://www.capitalone.com

- **NoSQL**: permiten resolver problemas vinculados con la variedad del dato.

  Metlife** es una entidad aseguradora con presencia internacional. Esta organización usa las tecnologías NoSQL para integrar todas las referencias de cliente en un único punto de acceso y tener una visión de 360 grados.

  **http://www.metlife.com

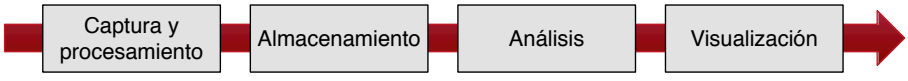
Esta aproximación a las tecnologías de *big data* está principalmente centrada en dos puntos: el almacenamiento y el procesamiento. En este material, vamos a ampliar los puntos que trataremos añadiendo el análisis y la visualización. El motivo detrás de este enfoque reside en que los cambios en las capas de procesamiento y almacenamiento influyen en el resto.

Cuando hablamos de tecnologías de *big data*, nos estamos refiriendo, en realidad, a una colección de componentes, plataformas y soluciones que cubren las diferentes necesidades para con el dato. Estas necesidades son:

- **Almacenamiento**: permitir el almacenamiento del dato conforme a las necesidades de negocio.
- **Procesamiento**: permitir la captura, la transformación y el movimiento del dato conforme a las necesidades de negocio.
- **Análisis**: permitir la generación de valor para el negocio a partir del dato.
- **Visualización**: permitir la presentación y comunicación de los resultados de acuerdo con las necesidades de negocio.

Estas necesidades se combinan siguiendo un flujo como representa la figura 32.

Figura 32. Flujo de *big data*



Fuente: Josep Curto

Muchas de las tecnologías de *big data* tienen origen *open source* para acelerar la innovación, lo que significa que podemos tener acceso a una versión *community* y, al mismo tiempo, diversos fabricantes ofrecen una plataforma de pago con diferentes componentes integradas y preparadas a nivel empresarial.

7.1. Almacenamiento

En las últimas décadas, las bases de datos relacionales han sido la opción de almacenamiento *de facto* para los sistemas de información. En algunos contextos con grandes necesidades de almacenamiento y procesamiento, como puede ser la meteorología, se ha trabajado con sistemas combinados de *hardware* y *software* optimizados para tareas intensivas en el dato conocidos como *High Performance Computing* (HPC). El enfoque de HPC se ha fundamentado principalmente en la escalabilidad vertical.

HPC

Cuando hablamos de HPC, hacemos referencia a la práctica de añadir capacidad de computación de forma que mejora el rendimiento de una estación de trabajo y hace posible abordar problemas complejos en la ciencia, ingeniería y/o negocios.

Con la emergencia de *big data* esto está cambiando de forma significativa, principalmente por diversos motivos:

- La tecnología relacional no es escalable para soportar el volumen de datos en el contexto de *big data*.
- La tecnología relacional es incompatible con los datos no estructurados, que cada vez son más relevantes para el negocio.
- La nueva tecnología no necesita HPC para ejecutarse, sino que puede trabajar con redes de ordenadores trabajando de forma combinada con prestaciones de computación menores individualmente, pero mayores colectivamente.

En el contexto de un proyecto de *big data* existen diferentes tecnologías de almacenamiento que habilitan estrategias eficientes y escalables tanto en coste como en respuesta a las necesidades de la naturaleza del dato. Una de las características de este tipo de sistemas es que proporcionan alta disponibilidad (*High Availability* o HA) y/o tolerancia a fallos (*Fault Tolerance* o FT). Aunque similares, no son lo mismo. Por un lado, HA implica tener un esquema en que los tiempos de caídas deben mantenerse muy cortos en un período anual. Por otro lado, FT hace referencia a un sistema donde no existe la posibilidad de perder ni un solo minuto de trabajo en producción, lo que implica tener infraestructura totalmente redundante.

Una de las técnicas usadas para la alta disponibilidad es la replicación que habilita la copia y el mantenimiento de los objetos en una base de datos distribuida. También se conoce como *sharding*. Usaremos indistintamente una palabra u otra.

La tabla 6 resume las diferentes opciones disponibles así como qué aporta cada una de ellas.

El sistema de archivos distribuido también ha sido adoptado por las bases de datos relacionales, lo que da lugar a poder trabajar en paralelo, que se conoce como *Massive Parallel Processing* (MPP). Tenemos ejemplos como: Teradata*, IBM Netezza, Pivotal Greenplum** u Oracle Exadata.

***www.teradata.com**
****http://greenplum.org**

Tabla 6. Tecnologías del almacenamiento.

Tecnología	Descripción	Características	Productos	Caso de uso
Sistema de archivos distribuido	Sistema que proporciona almacenamiento basado en la división de los datos en ficheros y servidores.	Proporciona redundancia y alta disponibilidad por replicación. Acceso secuencial de datos. Para minimizar las lecturas de búsqueda a disco, así como el procesamiento de muchos ficheros, este tipo de sistemas agregan los datos en ficheros de mayor tamaño.	Apache HDFS, Amazon S3 o Google File System	Archivado de conjuntos de datos. Almacenamiento de datos en bruto. Almacenamiento de bajo coste para largos períodos.
NoSQL	Sistema que proporciona almacenamiento basado en una ordenación/representación no relacional.	En general cumple: escalado horizontal en lugar de vertical; alta disponibilidad, consistencia eventual; BASE, no ACID y *autosharding*. Persistencia políglota. Consultas distribuidas.	MongoDB, Apache Cassandra, Riak, Redis, Neo4j o CouchDB	El modelo de negocio no puede representarse de forma relacional. El modelo de negocio evoluciona rápidamente y necesita una base de datos flexible en su modelo.
NewSQL	Sistema NoSQL que combina propiedades ACID.	Además de las características de NoSQL, incluye soporte para SQL y el uso de estructuras relacionales.	VoltDB, NuoDB, Google Spanner o CockroachDB	Sistemas OLTP con alto volumen de transacciones. Analítica en tiempo real.
In-Memory	Uso de la memoria del procesador para el almacenamiento de datos.	Reduce la latencia de acceso y de cálculo. Puede basarse en *grid* o base de datos.	HazelCast, Pivotal Gemfire, Aerospike, MemSQL o Altibase HDB	Analítica operacional. BI Operacional. *Streaming Analytics*.

Dentro de NoSQL, existen principalmente cuatro tipos de bases de datos:

- ***Key-Value Store***: el almacenamiento se fundamenta en el uso de parejas clave-objeto en las que no hay esquema alguno. Ejemplos:

  - Apache HDFS,
  - Riak (http://basho.com),
  - Voldemort(http://www.project-voldemort.com),
  - Redis (http://redis.io),
  - RocksDB (http://rocksdb.org) o
  - Amazon DynamoDB.

- **Bases de datos orientadas a columnas**: el almacenamiento del dato se realiza por columnas, no por filas. Ejemplos:

  - Apache Hbase (http://hbase.apache.org),
  - Apache Cassandra (http://cassandra.apache.org),
  - MonetDB (http://www.monetdb.org),
  - Druid (http://druid.io/),
  - Vertica,
  - Sybase IQ,
  - LucidDB o
  - Amazon SimpleDB.

- **Bases de datos de grafos**: Usa nodos y vértices para representar datos. Ejemplos:

  - Neo4J (http://neo4j.com),
  - HyperGraphDB (http://hypergraphdb.org),

Grafo

Cuando hablamos de grafo, hacemos referencia a un conjunto de objetos (llamados vértices o nodos) unidos por enlaces (llamados aristas o arcos). Un grafo permite estudiar las interrelaciones entre sus nodos.

– ArangoDB (http://www.arangodb.com),
– Ontotext GraphDB (http://ontotext.com) u
– OrientDB (http://orientdb.com).

- **Bases de datos orientadas a documentos**: el almacenamiento del dato se realiza como si fuera un documento semi-estructurado. Ejemplos:

– MongoDB (https://www.mongodb.org),
– CouchDB (http://couchdb.apache.org) o
– MarkLogic (http://www.marklogic.com).

Para algunas de las opciones disponibles, las distinciones entre las diferentes bases de datos se están diluyendo, ya sea porque una misma base de dato pasa a ser multi-NoSQL (soportando más de un tipo), o porque pertenece a varias de las categorías al mismo tiempo. Ejemplos: ArangoDB combina grafos, documentos y *key-value*; OrientDB combina grafos, documentos, objetos y *key-value*. En general, estamos hablando de una alta especialización en el caso de uso y, por lo tanto, de un escenario políglota en el almacenamiento del dato.

7.2. Procesamiento

La necesidad de procesar datos no es un aspecto nuevo para las organizaciones. En el pasado, esto se ha abordado usando técnicas de integración de datos, o *data integration*, como hemos explicado en anteriores apartados. Aunque existen muchas técnicas de integración, el procesamiento de *big data* se fundamenta principalmente en ELT (*Extract, Load, Transform*). Es decir, se pone el foco en guardar el dato en bruto, con el menor número de cambios, y el proceso de transformación se ejecuta en cada una de las bases de datos (sea cual sea su tipología).

Data integration

Cuando hablamos de *data integration*, hacemos referencia al conjunto de aplicaciones, productos, técnicas y tecnologías que permiten una visión única y consistente de nuestros datos de negocio.

En línea con los sistemas de almacenamiento, las principales aproximaciones para el procesamiento son:

- **Procesamiento de datos en paralelo**: lo que significa que un proceso se divide en múltiples tareas que se ejecutan en paralelo. Tradicionalmente, este enfoque se ha realizado con una única máquina con múltiples procesadores o núcleos.

- **Procesamiento de datos distribuidos**: lo que significa que el proceso se divide en múltiples tareas que se ejecutan en un clúster de máquinas conectadas en red siguiendo la filosofía "divide y vencerás".

Clúster

Cuando hablamos de clúster, hacemos referencia al conjunto de ordenadores conectados en red que trabajan de forma conjunta. Cada ordenador del clúster es llamado nodo. Si los ordenadores son heterogéneos, realizan tareas independientes o no están en la misma localización, hablamos de *grid*.

En el contexto de *big data*, para poder abordar las necesidades de trabajar con grandes volúmenes de datos y/o de capturarlos y consumirlos a diferentes velocidades (desde horas hasta por debajo del segundo), han emergido diferentes aproximaciones:

- **Procesamiento en modo *batch*, o por lotes**: el dato se procesa en modo *offline*. Su latencia puede ir desde minutos hasta horas. Antes de ser procesado, el dato se ha almacenado previamente. Apache MapReduce y Spark, este último con mejores prestaciones en términos de velocidad, permiten este tipo de procesamiento.

  Hulu* es un servicio de video en *streaming* con más de 5.5 millones de suscriptores y más de 20 millones de visitantes únicos por mes. Esta compañía usa MapReduce para procesar los *logs* resultado de la visualización de más de 400 millones de videos al mes. El objetivo es poder ofrecer un servicio de *streaming* con un nivel de calidad consistente. Es decir, siempre disponible, desde cualquier dispositivo y con el nivel de calidad de video adecuado al dipositivo.

  ***http://hulu.com**

- **Procesamiento en modo *real time*, o en tiempo real**: el dato se procesa en modo *online*. Su latencia está en el rango desde menos de un segundo hasta el minuto. Por ello, el dato se procesa en memoria en el momento de su captura, antes de almacenarlo. Hay dos tipos: procesamiento en flujo (*stream*), en el que el dato llega de forma continua, y procesamiento por intervalos o eventos (*event*). Apache Storm, Apache Flink y Spark permiten este tipo de procesamiento.

  MyFitnessPal* es un servicio que permite conocer el número de calorías consumidas y da soporte al tratamiento de dietas. Esta compañía usa Spark para limpiar, mejorar y complementar los datos específicos de comida introducidos por los usuarios con el objetivo de tener una base de datos de comida/calorías de máxima calidad en tiempo real. Es importante para este servicio que sea lo más cómodo y menos intrusivo para el usuario. Además, también se aprovecha de las capacidades de Spark para hacer recomendaciones.

  ***http://www.myfitnesspal.com**

El procesamiento por intervalos no es nuevo. Los sistemas CEP (*Complex Event Processing*) se han usado desde hace años en sectores como Banca o Energia para resolver esta necesidad. En este tipo de sistemas, lo relevante no es procesar todo el flujo de datos sino detectar aquellos subconjuntos que cumplen un patrón. El sistema monitoriza el flujo de datos y lo compara con los patrones definidos. Por ejemplo, para detectar el fraude en entidades financieras, lo revelante es detectar que un determinado cliente está realizando un conjunto de operaciones sospechosas de cometer un fraude.

CEP

Cuando hablamos de CEP, hacemos referencia al procesamiento de eventos en tiempo real que combinan múltiples fuentes y que se usa para inferir eventos o patrones que sugieren situaciones complicadas como oportunidades y/o amenazas.

En el ámbito del procesamiento hacemos referencia al paradigma SCV. La relación de las tres componentes de SCV se ilustra en la figura 33 y significa:

SCV

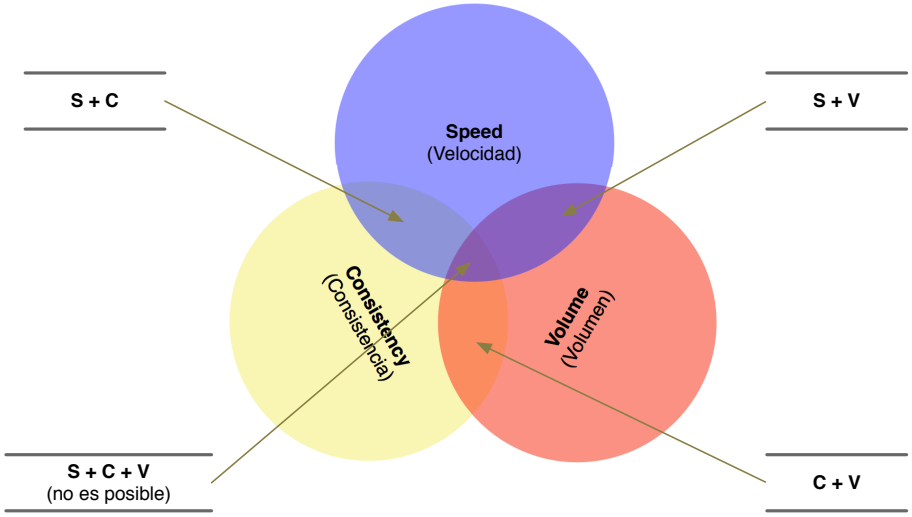
Cuando hablamos de SCV, hacemos referencia a *Speed*, *Consistency* y *Volume*. Es decir, a la velocidad de procesamiento, a la exactitud del dato y a la cantidad de datos procesados.

- Si se necesita S y C, no es posible procesar grandes volúmenes de datos porque retrasan el procesamiento.

- Si se necesita C y V, no es posible trabajar a una gran velocidad porque el procesamiento a gran velocidad requiere menores cantidades de datos.

- Si se necesita V y C, se consideran muestras (en lugar de trabajar con todo el conjunto de datos), lo que reducirá la consistencia.

Figura 33. SCV



Fuente: Josep Curto

Revisemos un ejemplo que necesita ambos procesamientos.

> Cuando una empresa se enfrenta al fraude, como puede ser en el sector de las finanzas, energético o *retail*, tiene diversas necesidades. Por un lado, necesita hacer un análisis forense de todo el histórico de transacciones para poder detectar nuevos patrones de fraude. Esta necesidad puede considerarse como un problema en el que prima la capacidad de trabajar con todo el historial y no la velocidad. Estamos ante una necesidad que puede cubrirse con el procesamiento y almacenamiento *batch*.
>
> Por otro lado, también existe otra necesidad una vez se han reconocido los patrones. Esta necesidad consiste en analizar el flujo de transacciones en tiempo real y detectar si se cumple alguno de los patrones. Aquí prima la velocidad de detección del evento. Estamos ante un escenario de procesamiento en *streaming*.

7.3. Análisis

La creciente complejidad en el dato ha permeado en la capa del análisis, lo que significa ajustar y modificar los diferentes tipos de análisis a la nueva naturaleza del dato.

Para distinguir claramente los cambios en procesamiento y almacenamiento, hemos separado el *data warehouse* y la integración de datos de la inteligencia de negocio, aunque, en general, no se conciben este tipo de sistemas sin estas componentes. Sin embargo, *big data* abre la puerta a una nueva combinación y de ahí la separación que estamos considerando, puesto que la arquitectura para el almacenamiento y el procesamiento de datos puede llegar a ser más compleja de lo que era antaño. Es necesario recordar los diferentes componentes de análisis de la inteligencia de negocio:

- **Informes**: documentos a través de los cuales se presentan los resultados de uno o varios procesos de negocio, que pueden distribuirse o simplemente estar disponibles para su acceso. Suelen contener texto acompañado de elementos como tablas o gráficos para agilizar la comprensión de la información presentada.

- **OLAP (*OnLine Analytical Processing*)**: método para organizar y consultar datos sobre una estructura multidimensional.

- **Cuadros de mando (o *dashboard*)**: sistema que informa de la evolución de los parámetros fundamentales de negocio de una organización o de un área del mismo a través de componentes visuales integradas.

- ***Scorecards***: tipo de cuadro de mando formado solo por listas de indicadores. A veces también toma la forma de informe.

- **Consultas *ad hoc***: método que ofrece autoservicio y exploración de datos a usuarios finales basados en metadatos de negocio.

- **Alertas y monitorización automática**: sistema para crear, gestionar y distribuir alertas críticas basadas en indicadores clave de negocio con foco en la gestión de excepciones.

Los componentes anteriores soportan el enfoque de planificación y control estratégico.

- **Cuadro de mando integral (o *balanced scorecard*)**: método de planificación estratégica, basado en métricas y procesos, ideado por los profesores Kaplan y Norton, que relaciona factores medibles de procesos con la consecución de objetivos estratégicos.

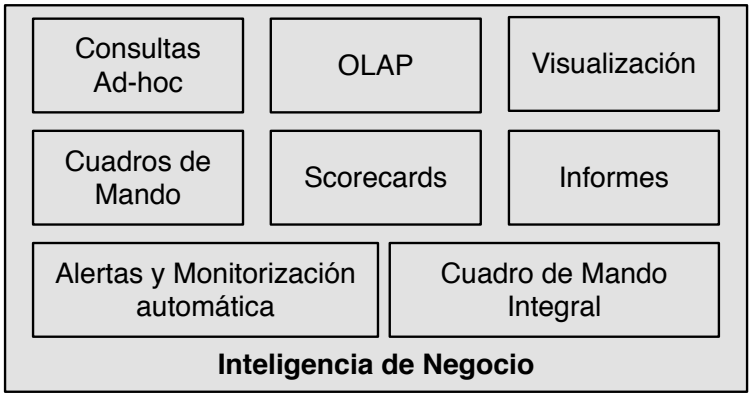
Una solución de inteligencia de negocio puede tener solo una o varias de estos componentes. Las soluciones más maduras de mercado suelen tenerlos todos en formato modular. La implementación de uno o más componentes en una organización debe depender de las necesidades de negocio en la organización y no de la plataforma, del proveedor seleccionado o de las preferencias del usuario de negocio o departamento.

Lectura complementaria

Kaplan, Robert S; Norton, D. P. (1996). *The Balanced Scorecard: Translating Strategy into Action*. Boston, MA.: Harvard Business School Press.

La figura 34 ilustra las componentes de análisis de la inteligencia de negocio.

Figura 34. Inteligencia de Negocio



Fuente: Josep Curto

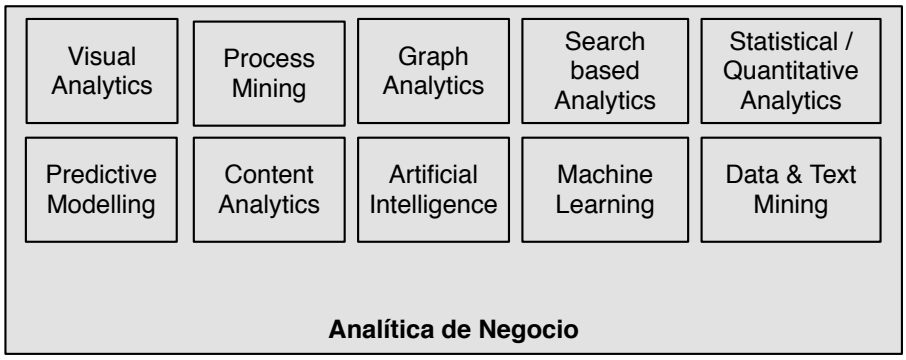
Consideremos otro ejemplo de cómo se combina con *big data*.

Jagex* es una compañía de videojuegos para móviles que soportan millones de usuarios jugando al mismo tiempo. Para esta compañía es absolutamente primordial comprender a sus clientes: aquellos que pagan, aquellos que se dan de alta, qué productos virtuales se compran y se usan en cada uno de los videojuegos, y poder analizar esta información tanto temporalmente como geográficamente. Para ello, se han combinado las capacidades de almacenamiento y procesamiento de *big data* con las capacidades de análisis en formato cuadro de mando e informes de la inteligencia de negocio para tener control del negocio.

***http://www.jagex.com**

En la analítica de negocio, hemos introducido los diferentes tipos de análisis existentes resumidos en la figura 35, que ilustra las componentes en la analítica de negocio.

Figura 35. Analítica de Negocio



Fuente: Josep Curto

Consideremos un ejemplo de cómo se combina con *big data*.

SuperCell* es una compañía de videojuegos para móviles que soportan millones de usuarios jugando al mismo tiempo. Entre sus éxitos destaca *Clash of Clans*. Esta compañía usa la combinación de tecnologías de almacenamiento de *big data* y analítica de negocio para validar hipótesis de negocio. Uno de los test A/B realizados ha sido para decidir si valía la pena añadir la conectividad de Facebook, conociendo si los usuarios usan esta posibilidad tanto para invitar a sus amigos como para compartir sus logros, y si esto incide en la retención del usuario.

***http://www.supercell.com**

Las taxonomías presentadas tienen una razón de ser. La principal diferencia de la inteligencia y la analítica de negocio tradicionales con respecto a *big data* es que cada una de las componentes se ha tenido que adaptar. Por un lado, en el contexto de la inteligencia de negocio esto sucede:

- A través de conectores para el uso de los sistemas de almacenamiento y procesamiento de *big data*, como fuentes de datos para el sistema de inteligencia de negocio.

- A través la adaptación de la tecnología a la complejidad del dato. Tenemos, por ejemplo, Apache Kylin*, creado por eBay, que proporciona OLAP para *big data*; Hue**, creado por Cloudera, que permite visualizar consultas *ad hoc* sobre Hadoop; o Caravel*** de Airbnb con foco en la exploración de datos.

***http://kylin.apache.org**
***http://gethue.com**
****http://github.com/airbnb/caravel**

Los fabricantes tradicionales como IBM, Microsoft, Microstrategy, Oracle o Information Builders también se están posicionando en este mercado, creando su propia propuesta integrada y/o a través de conectores específicos para su plataforma.

Por otro lado, en el contexto de la analítica, tenemos también que las soluciones y librerías ya existentes se están adaptando de una forma similar a la inteligencia de negocio. Adaptar, en este caso, se traduce en crear nuevas versiones del algoritmo que encapsula una cierta técnica para que pueda aplicarse a un conjunto de datos complejos, pueda escalar y, sobretodo, tenga sentido desde un punto de vista estadístico y matemático. Por ello, el gran cambio reside en la aparición de nuevas librerías de *machine learning*, *graph analytics* y *deep learning* adaptadas a *big data*.

Lectura complementaria

Leskovec, J.; Rajaraman, A.; Ullman, J. (2016). *Mining of Massive Datasets segunda edición*.

7.4. Visualización

Tradicionalmente, las componentes de inteligencia de negocio, como los cuadros de mando, informes y/o vistas OLAP, se han usado para presentar el resultado del análisis de la información. Con el advenimiento de *big data* y la combinación de tecnologías, este enfoque ya no es suficiente. Dos disciplinas han emergido para ayudar en la visualización de la información: *Data Visualization* (Visualización de datos) y *Data Storytelling* (historias fundamentadas en datos). Debemos comprender primero estos conceptos.

Se entiende por *Data Visualization* la representación de datos que explota las habilidades visuales para amplificar los procesos cognitivos.

Data Visualization persigue incrementar las capacidades exploratorias y explicativas, representar grandes volúmenes de datos y comprender las relaciones ocultas en los datos de forma visual. Ha aparecido una gran colección de librerías especializadas en este ámbito*. Entre estas librerías y herramientas destacan:

***http://selection.datavisualization.ch**

Lectura complementaria

Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Analytics Press.

- D3.js (http://d3js.org),
- Polimaps (http://polymaps.org),
- Processing.js (http://processingjs.org),
- Grafana (http://grafana.org),
- Tableau (http://www.tableau.com),
- Qliksense (http://www.qlik.com),
- CartoDB (http://cartodb.com) o
- Yellowfin (http://www.yellowfinbi.com).

Por otro lado, se entiende por *data storytelling* el método visual de presentar información para hacerla más comprensible y fácil de comprender.

En la actualidad, algunas herramientas propietarias, como las que ofrecen Tableau, Qliksense, Quadrigram*, Miso**, TimelineJS*** o Yellowfin, capacitan a las organizaciones para el uso de *data storytelling*, si bien también es posible crear de forma programática.

*http://www.quadrigram.com
**http://misoproject.com
***http://timeline.knightlab.com

Lectura complementaria

Segel, E.; Heer, J. (2010). *Narrative Visualization: Telling Stories with Data*, IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis).

Estas técnicas no solo tratan de usar la mejor representación para explicar lo que sucede, sino que, además, deben poderse conectar con las componentes de procesamiento y almacenamiento de *big data*. No solo se trata de tener la tecnología adecuada (escalable y adaptable a *big data*), sino de dominar la comunicación de la información. Como comenta Stephen Few, las capacidades para mostrar y explicar la información de forma efectiva no son intuitivas y es necesario aprender unos nuevos principios:

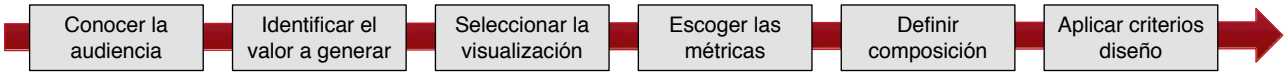
- Conocer la audiencia de la visualización. Esto incluye factores como el rol, el flujo de trabajo, el conocimiento técnico y de negocio de la audiencia.
- Determinar el valor que se quiere proporcionar a la audiencia. En este sentido tenemos dos grandes opciones. Tenemos ya una pregunta para responder, o bien estamos haciendo un análisis exploratorio. Esto incluye identificación de lo que es relevante, establecimiento de metas y expectativas.
- Seleccionar la visualización correcta. Esto incluye la elección del gráfico y/o la representación*, el alcance, horizonte temporal y el tipo de decisiones.

*http://extremepresentation.com

- Escoger las medidas adecuadas que deben siempre ayudar a tomar decisiones.
- Creación/Composición de la visualización, que debe tener en cuenta la forma, la estructura, la funcionalidad y los principios de diseño.
- Uso de criterios de diseño y presentación de información, como la elección de colores y tipografía.

Estos principios se ilustran en la figura 36.

Figura 36. Principios visualización



Fuente: Josep Curto

Resumen

En este módulo didáctico hemos presentado los conceptos de *business intelligence*, *business analytics* y *big data*, que fundamentalmente habilitan a las organizaciones para generar valor a partir de sus datos.

En relación a *business intelligence*, hemos presentado su definición, cuándo es necesario, qué beneficios aporta y las tecnologías que forman parte de esta estrategia. Hemos entrado en detalle en cómo se gestiona y cómo se explota el dato.

En relación a *business analytics*, hemos presentado su definición y una taxonomía de las tecnologías que forman parte de esta estrategia.

En relación a *big data*, hemos presentado su definición, los tipos que existen, qué beneficios aporta, cuándo es necesario aplicarla y las tecnologías que forman parte de esta estrategia. Y sobretodo se ha puesto de manifiesto cuán diferente es y por qué complementa a la inteligencia de negocio y la analítica de negocio.

También hemos ido introduciendo ejemplos. Tal y como se ha mostrado en estos materiales, existen multitud de organizaciones que ya han desplegado este tipo de sistemas de información y han conseguido rendimientos de la explotación de conjuntos de datos complejos.

Glosario

ACID Estándar *de facto* de las bases de datos relacionales. Es el acrónimo de *Atomicity* (atomicidad), *Consistency* (consistencia), *Isolation* (aislamiento) y *Durability* (durabilidad).

API Conjunto de subrutinas, funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro *software* como una capa de abstracción.

BASE Estándar para las tecnologías *big data*. Es el acrónimo de *Basically Available* (básicamente disponible), *Soft state* (estado blando) y *Eventual consistency* (consistencia eventual).

big data Conjunto de estrategias, tecnologías y sistemas para el almacenamiento, procesamiento, análisis y visualización de datos complejos.

business intelligence Conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización.

byte Unidad de medida de información digital.

CAP Estándar *de facto* de los sistemas distribuidos. Es el acrónimo de *Consistency* (consistencia), *Availability* (disponibilidad) y *Partition Tolerance* (tolerancia a la partición).

CEP Procesamiento de eventos en tiempo real que combina múltiples fuentes y que se usa para inferir eventos o patrones que sugieren situaciones complicadas, como oportunidades y/o amenazas.

clúster Conjunto de ordenadores conectados en red que trabajan de forma conjunta. Cada ordenador del clúster es llamado nodo.

CRM Acrónimo de *Customer Relationship Management*; hace hace referencia la gestión de la relación con clientes.

crowdsourcing Proceso de obtener servicios, ideas y contenido a través de la participación de una gran masa de personas.

cuadro de mando Sistema que informa de la evolución de los parámetros fundamentales de negocio de una organización o de un área del mismo.

data integration Conjunto de aplicaciones, productos, técnicas y tecnologías, que permiten una visión única y consistente de nuestros datos de negocio. También denominada como integración de datos.

data warehouse Repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, independientemente de cómo se vayan a utilizar posteriormente por los consumidores o usuarios; con las propiedades siguientes: estable, coherente, fiable y con información histórica.

ecuaciones diferenciales Ecuación matemática que relaciona una función y sus derivadas.

edge analytics Aplicaciones analíticas para IoT en las que ciertos algoritmos se ejecutan en los nodos de la red y no solo en el centro de datos.

ERP Acrónimo de *Enterprise Resource Planning*; hace referencia a la gestión de los recursos de una organización.

escalabilidad Habilidad de un sistema, red o proceso para reaccionar y adaptarse sin perder calidad, o bien manejar el crecimiento continuo de trabajo de manera fluida, o bien para estar preparado para hacerse más grande sin perder calidad en los servicios ofrecidos.

escalabilidad horizontal Escalabilidad fundamentada en el incremento de nodos del sistema, proceso o red.

escalabilidad vertical Escalabilidad fundamentada en añadir más recursos –memoria, disco duro y/o procesadores–.

estructura de datos relacional Tipo de base de datos que permite establecer interconexiones o relaciones entre los datos guardados en tablas.

grid Conjunto de ordenadores conectados en red que trabajan de forma conjunta, pero, a diferencia del clúster, los ordenadores son heterogéneos, realizan tareas independientes o no están en la misma localización.

latencia Suma de retardos temporales en la captura, almacenamiento, procesamiento y análisis del dato.

Internet de las Cosas Interconexión digital de objetos cotidianos con Internet. Nos referiremos a él por su acrónimo en inglés IoT, *Internet of Things*.

metadatos Datos estructurados y codificados que describen características de un objeto, dato o proceso de negocio.

NIST Acrónimo de *National Institute of Standards and Technology*, una institución americana que estudia, define y promueve estándares tecnológicos.

NPL Acrónimo de *Natural Processing Language*. Hace referencia a un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano.

NoSQL Acrónimo de *Not Only SQL*. Hace referencia a bases de datos no relacionales.

HPC Práctica de añadir capacidad de computación de forma que mejora el rendimiento de una estación de trabajo y hace posible abordar problemas complejos en la ciencia, ingeniería y/o negocios.

OLAP Método para organizar y consultar datos sobre una estructura multidimensional. Es el acrónonimo de *Online Analytical Processing* o proceso analítico en línea.

open data Conjuntos de datos considerados un bien común y que, por ello, son gratuitos, accesibles y bien estructurados para su descarga y análisis.

OWL Acrónimo de *Web Ontology Language*. Hace referencia a un estándar para el diseño de ontologías de modelos de datos.

PMML Acrónimo de *Predictive Model Markup Language*. Hace referencia a un estándar para el intercambio de datos entre organizaciones.

RIF Acrónimo de *Rule Interchange Format*. Hace referencia a un estándar para el intercambio de datos entre organizaciones.

SCV Hace referencia a *Speed*, *Consistency* y *Volume*. Es decir, a la velocidad de procesamiento, a la exactitud del dato y a la cantidad de datos procesados.

SLA Acuerdo que estipula el nivel de servicio, el soporte, posibles penalizaciones, el nivel de alta disponibilidad, tanto de *hardware* como de *software*, y el precio.

SQL Acrónimo de *Structure Query Language* y hace referencia al lenguaje de consultas de bases de datos relacionales.

taxonomía Clasificación u ordenación en grupos de cosas que tienen unas características comunes.

UIMA Acrónimo de *Unstructured Information Management Architecture*. Hace referencia a un estándar que permite la interoperabilidad de analítica de datos en información no estructurada.

variabilidad Hace referencia a que los flujos de datos pueden tener comportamientos erráticos o inconsistentes en ciertos períodos.

velocidad Hace referencia tanto al procesamiento de datos como a su latencia.

variedad Hace referencia tanto a la cantidad de fuentes diferentes que se deben combinar como a la heterogeneidad del dato.

veracidad Hace referencia a la incertidumbre en el dato producto de su baja calidad, la ambigüedad en su definición o simplificaciones en su modelización.

vinculación Dificultad de relacionar diferentes y dispares fuentes de datos.

volumen Tamaño del conjunto de datos creado diariamente.

XBRL Acrónimo de *eXtensible Business Reporting Language*. Hace referencia a un estándar para informes financieros.

Bibliografía

Corr, L.; Stagnitto, J. (2011). *Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema*. Leeds: DecisionOne Press

Davenport, T.H. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Boston: Harvard Business Review Press.

Davenport, T.H.; Harris, J.G. (2007). *Competing on Analytics: The New Science of Winning*. Nueva York: Harvard Business Press.

Davenport, T.H.; Kim, J. (2013). *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*. Boston: Harvard Business Review Press.

Erl, T.; Khattak, W.; Buhler. P. (2015). *Big Data Fundamentals: Concepts, Drivers & Techniques*. New Jersey: Prentice Hall

Fisher, T. (2009). *The Data Asset: How Smart Companies Govern Their Data for Business Success*. New Jersey: Wiley

Howson, C. (2013). *Successful Business Intelligence, Second Edition: Unlock the Value of BI & Big Data*. New York: McGraw-Hill Education

Kimball, R.; Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. New Jersey: Wiley.

Malcolm, F.; Roehrig, P.; Pring, B. (2014). *Code Halos: How the Digital Lives of People, Things, and Organizations Are Changing the Rules of Business*. New Jersey: Wiley

Foreman, J.W. (2013). *Data Smart: Using Data Science to Transform Information into Insight*. New Jersey: Wiley.

Redman, T. C. (2008). *Data Driven: Profiting from Your Most Important Business Asset*. Boston: Harvard Business Review Press.

Schmarzo, B. (2016). *Big Data MBA: Driving Business Strategies with Data Science*. New Jersey: Wiley.

Schmarzo, B. (2013). *Big Data MBA: Understanding How Data Powers Big Business*. New Jersey: Wiley.