ENAE
BUSINESS SCHOOL

TEMA 3: Definición y Conceptos Básicos

MÓDULO /CURSO

Fundamentos de Data Science y Big Data

PROFESOR:

Mario De Felipe

ÍNDICE TEMÁTICO

CONTENIDOS

Objetivos:

- Aclarar la terminología: Business Intelligence, Business Analytics y Data Science.
- Definir la ciencia de los datos.
- Entender los tipos de análisis que podemos hacer sobre los datos y sobretodo, el valor que aportan cada uno de ellos.
- Revisar la arquitectura de datos necesaria para poder llevarla a cabo.

Contenido:

Presentar los conceptos de *Data Science, Business Analytics, Business Intelligence y Big Data* como las técnicas orientadas a implementar en la medición del negocio para contribuir a la mejora del crecimiento empresarial. Exponer sus características principales y definir los diferentes tipos de analítica que podemos hacer.

TEMA 2: DEFINICIÓN Y CONCEPTOS BÁSICOS

LA CUESTIÓN TERMINOLÓGICA

Antes de comenzar a adentrarnos en la ciencia de los datos me gustaría aclararos de forma breve la cuestión terminológica de varios conceptos alrededor de esta disciplina que, seguramente estáis familiarizados con ellos, pero no tenéis del todo claro cómo se relacionan los unos con los otros y en qué se diferencian. Estamos hablando de los conceptos de Business Analytics, Business Intelligence, Data Science y Big Data.

En la actualidad hablamos de *Business Analytics* o Analítica de Negocio, como una evolución de *Business Intelligence* o Inteligencia de Negocio tradicional. Aunque como muchos de vosotros conoceréis, hay bastante confusión con ambos términos, y más aún, si encima los contraponemos al de *Data Science o Ciencia de los Datos*. Gran parte de esta confusión, la tenemos los profesionales que, bien por razones de marketing o bien por abuso del lenguaje, usamos estos términos según creemos conveniente, ya que no existe ningún dogma o convenio académico al respecto[1]. De esta forma, os vais a encontrar con diferentes significados bajo el mismo término.

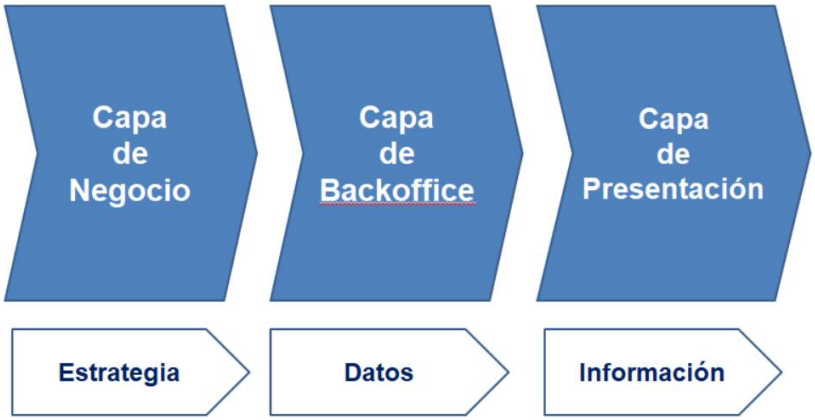
Al igual que hizo Nietzchse en su "Genealogía de la Moral" (tranquilos que no me voy a poner tan filosófico), vamos a intentar entender los conceptos, conociendo las diferentes acepciones que se han utilizado.

Términos complementarios

Es muy común hablar de los términos de *Business Intelligence*, *Business Analytics* y *Data Science* como distintas disciplinas o roles profesionales que intervienen en el proceso de generación o extracción de conocimiento en las diferentes etapas o capas de dicho proceso.

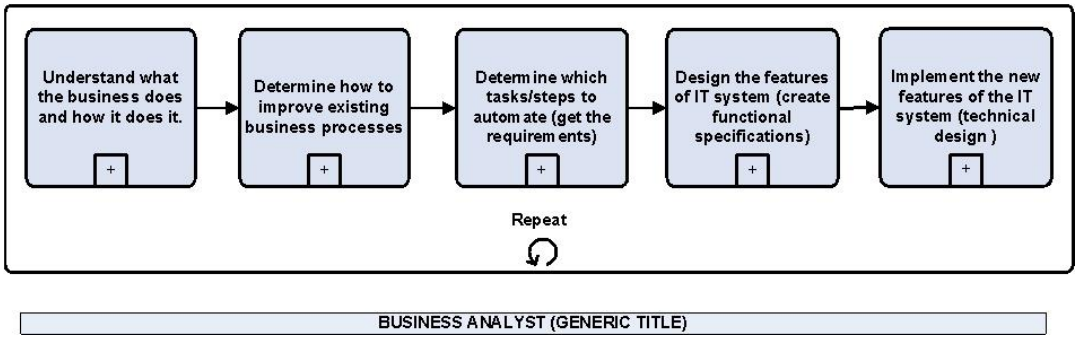
La arquitectura de un Sistema de Inteligencia de Negocio consta de 3 capas:

[1] Curioso es que la llamada Ciencia de los Datos carezca de cuerpo normativo como disciplina científica. En realidad, más que una ciencia deberíamos hablar de un conjunto de disciplinas: Matemática, Estadística, Inteligencia Artificial, Programación, Bases de Datos, Ciencias Empresariales, etc. Únicamente está "normalizada" por el estándar CRISP-DM, sobre la cual desarrollaremos nuestra analítica de negocio en este tema.



En esta arquitectura, podemos encajar:

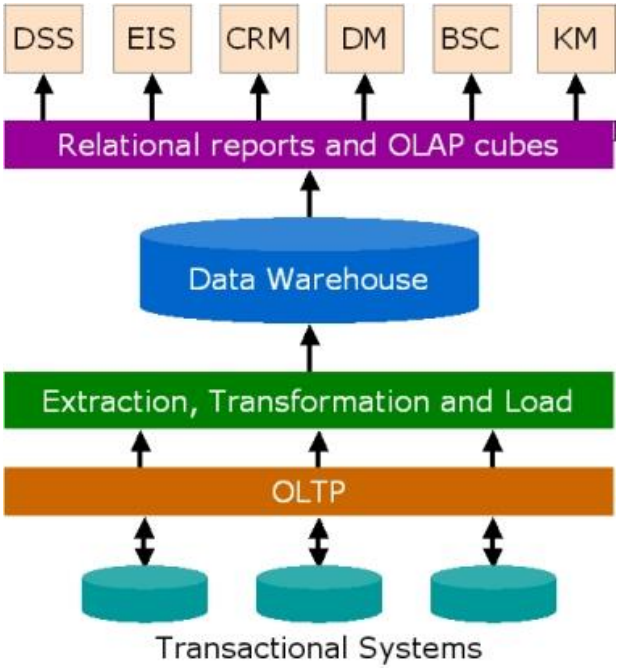
- **Business Analytics.** Definición y diseño de la **capa de negocio**. Se asocia típicamente a la persona encarga de realizar el análisis funcional, detectar el problema de negocio e intentar dar una solución analítica; es decir, resolverlo con información. Es su responsabilidad también definir el modelo de negocio, así como *la capa de presentación* de la aplicación. Actuará de enlace entre negocio, arquitectos de sistema, ingenieros de datos y científicos de datos, validando que se cumplen los objetivos de negocio.



- **Business Intelligence**. Traduce el modelo de negocio a un modelo de datos. Se encarga de diseñar e implementar el grueso de la **capa de backoffice**: desde el Data Warehouse o Datamart[2] y de crear los procesos de extracción, transformación y carga (ETL)[3], para finalmente diseñar la **capa de presentación** con informes o cuadros de mando.

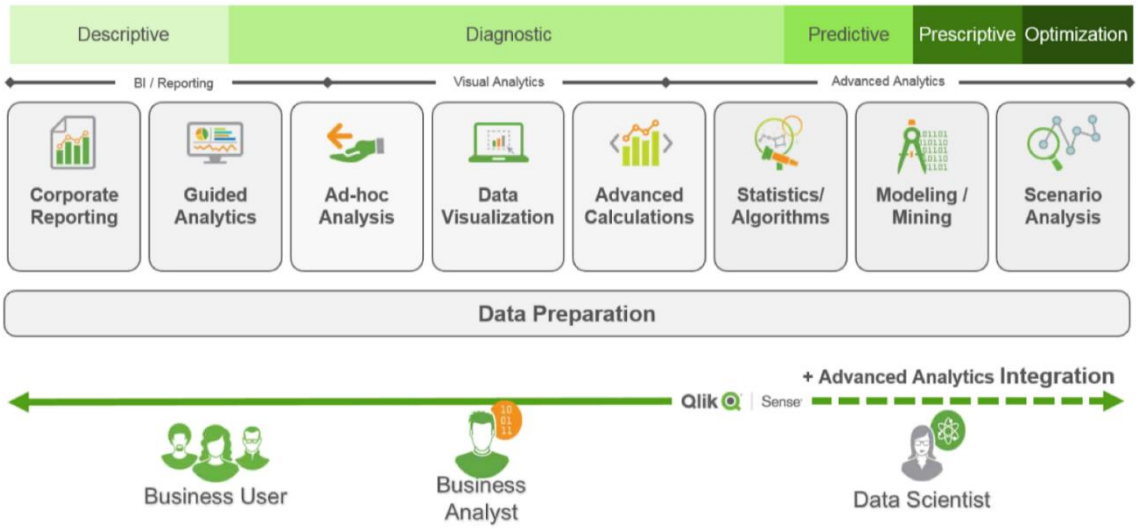
[2] En el Business Intelligence tradicional, únicamente hablamos de repositorios de datos estructurados y relacionales: Data Warehouse o Datamart. No se habla de Data Lakes que es una técnica asociada al Big Data. Cuando hablamos de expertos en Data Lake se suele hablar de Big Data Architect.

[3] De igual forma, en el Business Intelligence tradicional sólo se habla de ETL y no de ELT, típico de entornos Hadoop y NoSQL. En estos casos, se habla de Ingenieros de Datos o *Data Engineer*.



- **Data Science**. Se ocupa básicamente de un punto muy concreto de back office como es el procesado de los datos. Básicamente su misión es desplegar modelos y algoritmos para desplegar analítica avanzada: predictiva, prescriptiva y modelos de optimización.

Me parece muy ilustrativa la siguiente imagen donde en función del tipo de analítica, podéis distinguir el rol o disciplina de análisis de datos:



Términos contrapuestos

También es bastante frecuente que os encontréis con que muchos autores hablan de *Business Intelligence* y *Business Analytics* como técnicas bastante diferenciadas.

	Business Intelligence	Advanced Analytics
Orientation	Rearview	Future
Types of questions	What happened When, who, how many	What will happen? What will happen if we change this one thing? What's next?
Methods	Reporting (KPIs, metrics) Automated Monitoring/Alerting (thresholds) Dashboards Scorecards OLAP (Cubes, Slice & Dice, Drilling) Ad hoc query	Predictive Modeling Data Mining Text Mining Multimedia Mining Descriptive Modeling Statistical / Quantitative Analysis Simulation & Optimization
Big Data	Yes	Yes
Data types	Structured, some unstructured	Structured and Unstructured
Knowledge Generation	Manual	Automatic
Users	Business Users	Data scientists, Business analysts, IT, Business Users
Business Initiatives	Reactive	Proactive

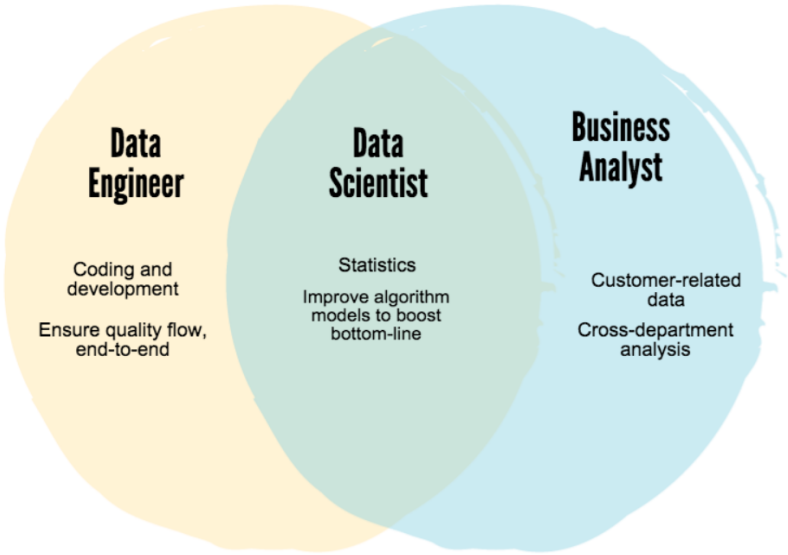
Como podéis ver en la imagen, se diferencia el BI del BA básicamente en los siguientes aspectos:

- **Orientación**. Se dice que el BI mira hacia el pasado (describe qué ocurrió y por qué), mientras que el BA mira hacia el futuro (qué es probable que ocurra y qué acciones puedo tomar).
- **Técnicas empleadas**. Frente al reporting y a los cuadros de mando tradicionales del BI, se habla de los modelos matemáticos y minería de datos.
- **Naturaleza de los datos**. BI procesa datos estructurados frente a BA que es capaz de tratar datos no estructurados.
- **Creación de conocimiento**. En el BI se dice que la creación de conocimiento es manual, dado que depende de que un especialista defina nuevos KPIs que aporten nuevo conocimiento. En el BA, un algoritmo de aprendizaje no supervisado por ejemplo, es capaz de descubrir por sí sólo reglas o patrones en los datos.
- **Iniciativas de Negocio**. Los sistemas de BI tradicionales son reactivos, en el sentido de que únicamente diagnostican lo que ha sucedido y dejan a la interpretación del usuario el descubrimiento de porqué y las acciones que debe tomar para ello. Con BA, se pueden realizar modelos de optimización que incluso propongan al usuario acciones a tomar en función del análisis de los datos.

También podemos encontrar una confrontación muy similar de los términos de BI y DS:

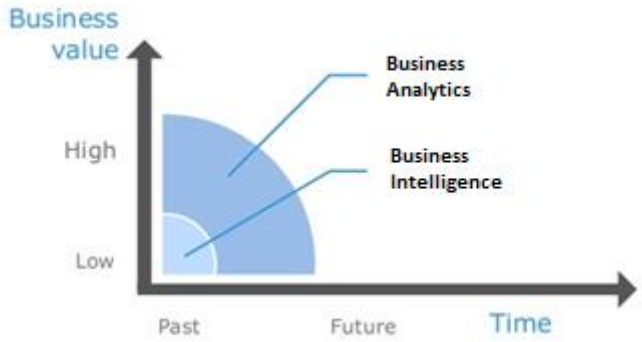
Area	BI Analyst	Data Scientist
Focus	Reports, KPIs, trends	Patterns, correlations, models
Process	Static, comparative	Exploratory, experimentation, visual
Data sources	Pre-planned, added slowly	On the fly, as-needed
Transform	Up front, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on load	Schema on query
Analysis	Retrospective, Descriptive	Predictive, Prescriptive

Y es que, en esta corriente de términos contrapuestos, podemos encontrar que se distingue DS vs BA, básicamente separando las tareas más "matemáticas" de las de "negocio":



Términos evolutivos

En último lugar, nos encontramos con la aproximación de *Business Analytics* como evolución de *Business Intelligence*.

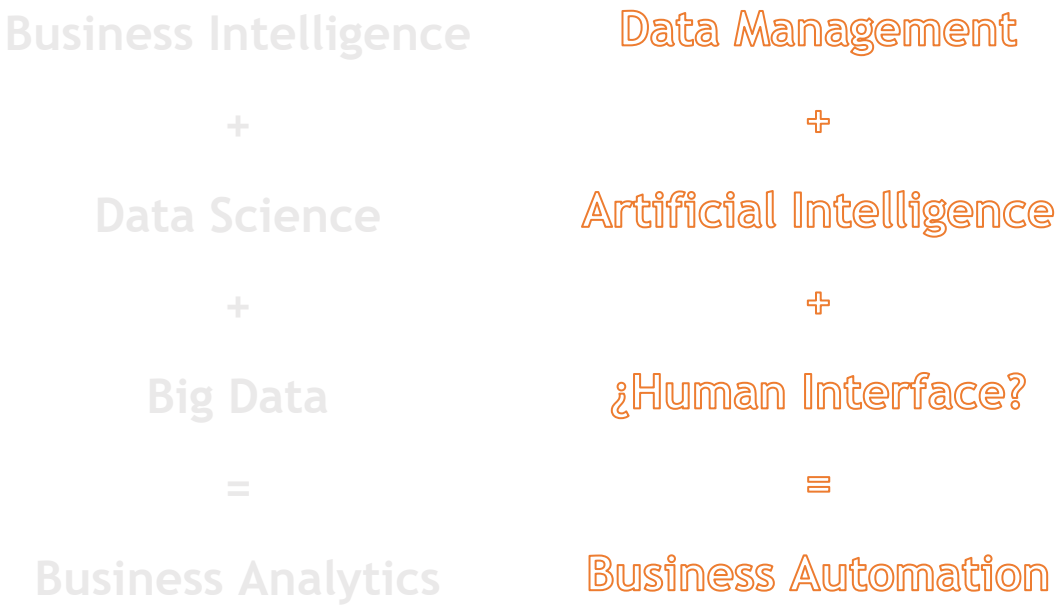


Y es que el mercado empuja. Los grandes fabricantes de software de Business Intelligence (Microsoft, Oracle, IBM, SAP, etc.) "tradicional" se han visto presionados por diferentes soluciones que vienen de cubrir las nuevas necesidades del negocio respecto al análisis de datos: como el Big Data (Hadoop y NoSQL), visualización (Tableau), autoservicio (Qlik) y analítica avanzada (R, Python). En la actualidad, todos ellos están incorporando soluciones a suite /portfolio para cubrir estas nuevas funcionalidades y así convertirse de nuevo en soluciones "full stack". El libre mercado y el capitalismo mandan, y las tendencias tecnológicas hacen (re)evolucionar al software. Esto realmente es lo que ha hecho que pasemos de tener soluciones de Inteligencia de Negocio a disponer de soluciones de Analítica de Negocio.

Como podéis deducir entonces, en cuanto a lo que software comercial se refiere, los mismos fabricantes de soluciones de Inteligencia de Negocio son los mismos que nos ofrecen soluciones de Analítica de Negocio.

Lo que os puedo trasmitir bajo mi punto de vista, es que a lo largo de mi carrera profesional lo percibo como un salto evolutivo: en relativamente poco tiempo hemos empezado a hablar en las organizaciones de proyectos que, bien por costes, bien por viabilidad técnica, no podíamos plantearnos salvo en la gran empresa.

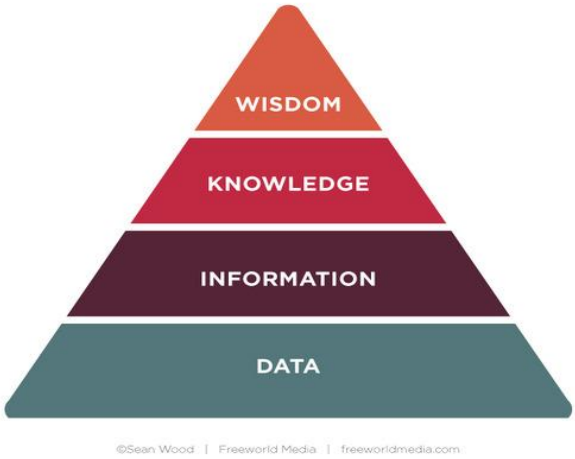
Hoy día podemos afirmar que la **analítica de negocio** que se realiza en las compañías es la **unión de análisis y procesamiento de datos estructurados** basado en informes y cuadros de mando (*Business Intelligence*), **análisis de datos estructurados y no estructurados de cualquier volumen, naturaleza y complejidad tanto real-time o batch** (*Big Data*) y **aplicación de modelos estadísticos y matemáticos para la predicción y la detección de patrones** (*Data Science*).



Para finalizar, como podéis ver en la figura de arriba, no me he podido resistir a compartiros lo que para mí será el siguiente paso: con la extensión y crecimiento de la tecnología de la Inteligencia Artificial, podremos automatizar los procesos de negocio que ahora dependen de la analítica para su realización. Pero eso será otra historia y quién sabe si nuestro mundo se parecerá más a Matrix o al de Wall-e.

ENTONCES, ¿QUÉ ES LA CIENCIA DE LOS DATOS?

Hablamos de la **Ciencia de los Datos** como la disciplina que utiliza métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sean estructurados o no estructurados.



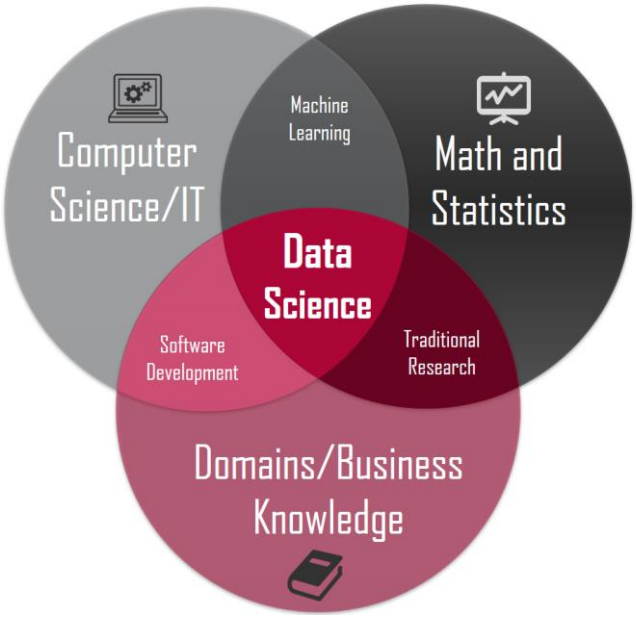
Por tanto, la Ciencia de los Datos está íntimamente relacionada con otros conceptos que hemos visto:

- **Big Data,** como sabemos, constituye la materia prima (datos) y la herramienta (tecnología), facilitadores de la Ciencia de la Datos para poder crear conocimiento.
- La **Inteligencia de Negocio** hace referencia a las técnicas, herramientas y estrategias que persiguen la creación de conocimiento con el fin de optimizar de la toma de decisiones. La Ciencia de los Datos en este sentido sería la parte habilitadora de la creación de conocimiento para la Inteligencia de Negocio, pero al mismo tiempo, la Ciencia de los Datos no cubre un único propósito como la mejora de la toma de decisiones sino que tiene un carácter general, con lo cual podemos afirmar que su misión es crear valor a partir de los datos.

La Ciencia de los Datos es la aplicación del método científico[4] al mundo del análisis de los datos. La Ciencia de los Datos surge como necesidad para dotar de rigor y objetividad a la toma de decisiones en cualquier organización. Daos cuenta de que disponemos de las herramientas y de la tecnología (BI y Big Data) pero eso no implica necesariamente que sepamos sacarle provecho. De igual manera que los griegos distinguían entre el *areté* y el *tekné*, nosotros hablamos de la Ciencia de los Datos como los procedimientos que nos permiten usar estas herramientas y tecnologías para intentar sacar conocimiento.

[4] Lo discutiremos ampliamente en el TEMA 3.

A su vez, la Ciencia de los Datos comprende tres grandes disciplinas:



- **Ingeniería de Datos**. Para el manejo, limpieza, almacenamiento y preparación de los mismos.
- **Aprendizaje Automático**. Para la creación y validación de modelos son necesarias Matemáticas, Probabilidad y Estadística.
- **Operaciones TIC**. Para la implementación del modelo y la comunicación de los datos son necesarias Programación en lenguajes informáticos y su despliegue en los Sistemas Informáticos Correspondientes.

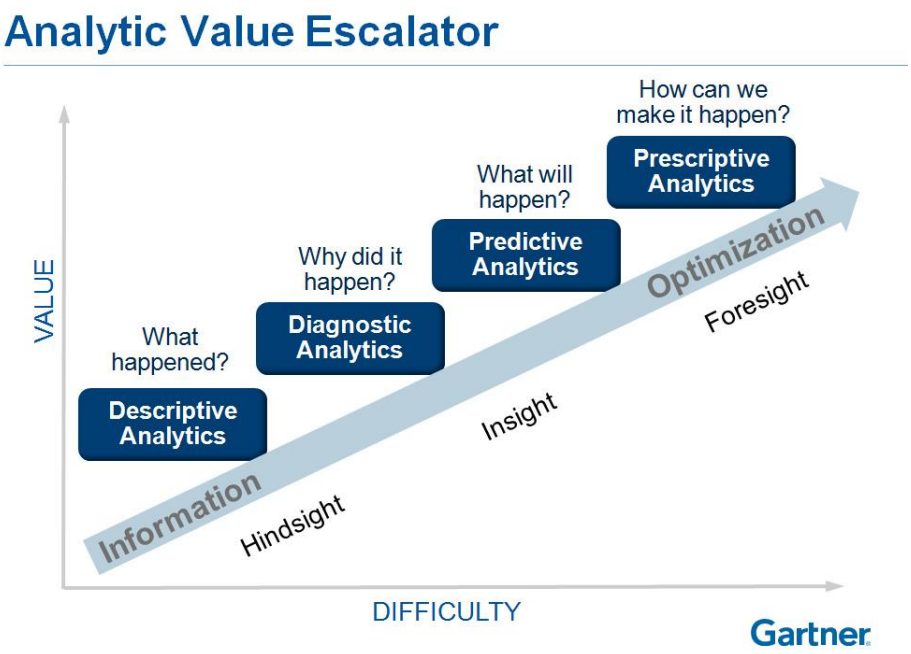
Con esto quiero que os deis cuenta de la magnitud del reto que tenemos por delante en este máster: a lo largo de los diferentes módulos vais adquirir los conocimientos y las competencias necesarios para formaros en estas 3 grandes disciplinas, que son bastante diferentes entre sí: vais a aprender Informática de Sistemas y Gestión (para la arquitectura y gestión de los datos), Ciencias Exactas (para los modelos de aprendizaje automático) y Programación (para el desarrollo e implantación de dichos modelos).

Es muy difícil que podamos ver todo esto en un solo módulo, por lo que lo que vamos a hacer en este módulo es presentaros los conceptos y técnicas más importante, que veáis todo el ciclo completo, todos los procesos de la ciencia de datos e ir presentándoos las herramientas y técnicas más importantes.

Vamos a comenzar entendiendo los diferentes tipos de analítica que podemos hacer en la ciencia de datos para a continuación ver los sistemas y la arquitectura que necesitaremos.

DISTINTOS PROBLEMAS, DISTINTAS SOLUCIONES: TIPOS DE ANALÍTICA DE NEGOCIO

En diciembre de 2012, en el *summit* de *Gartner* fue cuando se habló del "*Analytic Value Scalator*", que distinguía los diferentes tipos en función del valor que aportaban a negocio.



Fue un momento emocionante porque primera vez se ponía de manifiesto (para todo el mundo, desde una autoridad reconocida en el mercado como lo es *Gartner*) que el verdadero "destino final" de la analítica de negocio no era sólo ser soporte a la toma de decisiones sino realizar la toma de decisiones en sí. A partir de ahí, todos tuvimos claro que el escenario cambiaba radicalmente. Ya no valía únicamente utilizar la tecnología del pasado (nuestro BI tradicional con el que nos encontrábamos tan cómodos), sino que se avecinaba una nueva era y habría que experimentar con tecnologías emergentes (*Hadoop, cloud, NoSQL*,…) para intentar llegar antes que los demás. Y por si eso no fuera suficiente, la Inteligencia Artificial tarde o temprano aparecería en escena. Daba un poco de vértigo, pero el reto era francamente apasionante.

Volviendo al *Analytic Value Escalator*, daos cuenta de que podemos establecer tres clasificaciones de los Tipos de Analítica de Negocio:

- En función **del momento temporal** que queremos medir, evaluar o cuantificar.
- En función **del valor que aporta al negocio**, al decisor o usuario consumidor de la analítica.
- En función **del problema de negocio** o la pregunta que resuelven.

Clasificación Temporal	Clasificación por Valor para el Negocio	Clasificación por Problema de Negocio
• Hindsight • Insight • Foresight	• Information • Performance • Optimization	• Análisis Descriptivo • Análisis Diagnóstico • Análisis Predictivo • Análisis Prescriptivo

Clasificación Temporal

Atendiendo al momento temporal en que suceden los eventos que queremos medir:

- **Análisis Histórico (*Hindsight*)**. Analizamos eventos pasados e intentamos extraer conclusiones de ellos. En palabras del historiador Pierre Vilar: "*Hay que comprender el pasado para comprender el presente*".
- **Análisis de la Situación actual (*Insights*).** Diagnósticar el estado actual. Conocer qué está ocurriendo y por qué.
- **Análisis de la Previsión Futura (*Foresights*).** Predecir qué es probable que ocurra en base al conocimiento aprendido.

Clasificación por Valor para el negocio`

En función de lo que aportan al negocio podemos hablar de:

- **Conocimiento (Information)**. Poner a disposición de toda la organización el máximo de información posible para poder extraer conocimiento de ella.
- **Decisión (Performance).** Apoyar la toma de decisiones de forma objetiva en base a información con el fin de lograr el máximo rendimiento.

- **Acción (Optimization).** Optimizar y mejorar las decisiones y las acciones tomadas anticipándose a situaciones futuras y evaluando la mejor alternativa posible

Clasificación por Problema de negocio

Finalmente, tenemos la clasificación que ya conocéis, que viene determinada en función a la pregunta de negocio que resuelve:

- **Análisis Descriptivo o *Descriptive Analytics***: **¿Qué ha pasado?** Cuantificar y describir los eventos que han ocurrido o están sucediendo.
- **Análisis de Diagnóstico o *Diagnostic Analytics***: **¿Por qué ha pasado?** Analizar las causas de forma cuantitativa que subyacen a un evento pasado o presente.
- **Análisis Predictivo o *Predictive Analytics***: ¿**Qué es problable que ocurra?** Estimar el valor de una variable/medida en base a información actual o pasada.
- **Análisis Prescriptivo o *Prescriptive Analytics***: **¿Cuál es la mejor acción que puedo tomar ante esto que es probable que ocurra o que está ocurriendo?** Automatizar procesos de negocio mediante la generación automática de decisiones.

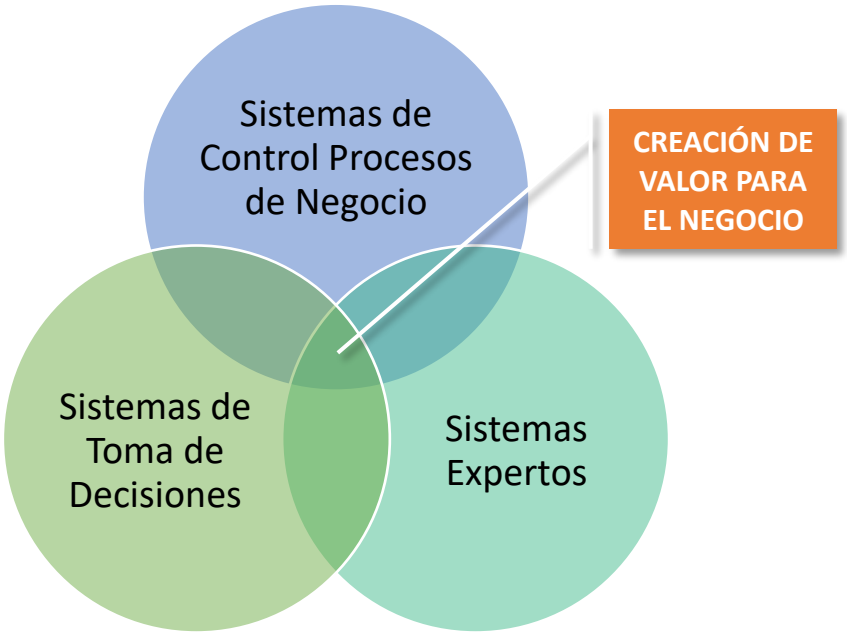
Es muy importante determinar qué tipo de analítica se requiere para solucionar el problema de negocio que se plantea porque eso determinará el diseño tanto del sistema como del modelo y la elección de la/s herramientas/s.

LOS SISTEMAS DE ANALÍTICA DE NEGOCIO: TIPOS Y ARQUITECTURA

En la práctica, para poder desplegar cualquier tipo de analítica de las que veíamos anteriormente, necesitamos implementar un sistema tecnológico que automatice todas las tareas asociadas al análisis de datos. Esto es lo que denominaremos Sistema de Analítica de Negocio (SAN).

Podemos caracterizar y agrupar los diferentes tipos de Sistemas de Analítica de Negocio en función de los 3 criterios o variables por los cuales definíamos los distintos tipos de analítica:

- Dimensión temporal
- Valor / Beneficio que aportan
- Problema o pregunta que resuelven



Como se puede ver en la figura anterior, se habla de 3 grandes familias de sistemas de analítica de negocio:

	Sistemas de Control de Procesos de Negocio	Sistemas de Toma de Decisiones	Sistemas Expertos
Análisis Temporal	Hindsight Insight	Insight Foresight	Foresight
Valor para el Negocio	Información	Rendimiento	Optimización
Problema de Negocio	Descriptivo Diagnóstico	Diagnóstico Predictivo	Predictivo Prescriptivo

- **Sistemas de Control de Procesos de Negocio**. Son aquellos que permiten medir el desempeño o rendimiento de uno o varios procesos de negocio concretos. El objetivo es detectar si el proceso cumple los objetivos y detectar comportamientos o sucesos

anómalos o perjudiciales para la compañía. Se basan en la definición de métricas y dimensiones a partir del procesado de datos históricos.

- **Sistemas de Toma de Decisiones**. Son los que permiten a un decisor recabar la información que necesita ante un problema de negocio, analizarlo y evaluar posibles opciones en base a datos. Para ello intervienen los tipos de analítica descriptiva, diagnóstica y predictiva.
- **Sistemas Expertos**. Son aquellos que en base al análisis de datos proponen tomar una decisión o llevar a cabo una acción o tarea. Básicamente tratan de anticipar tanto posibles problemas o situaciones de riesgo como oportunidades o ventajas competitivas a partir de simulaciones / predicciones realizadas en los datos, a los cuales responde bien con acciones programadas o aprendidas.

Ejemplos de Sistemas de Control de Procesos de Negocio

Según el área funcional o departamento propietario de la información que va a desplegar la analítica podemos encontrar diferentes tipos de sistemas:

- Comercial / Ventas
- Administración / Finanzas.
- Producción
- Marketing
- IT
- Calidad
- Compras / Logística / Almacén
- Dirección
- Etc.

Como ejemplo os dejo un enlace a un cuadro de mando de marketing:

https://webapps.qlik.com/marketing360/index.html#/web

También podríamos hablar de manera global, de un **Cuadro de Mando Corporativo**, o lo que es lo mismo, un sistema que englobe a todas las áreas funcionales de la compañía. De igual manera que un *datawarehouse* engloba a todos los posibles *datamarts*, el Cuadro de Mando Corporativo engloba a todos los posibles Cuadros de Mando departamentales.

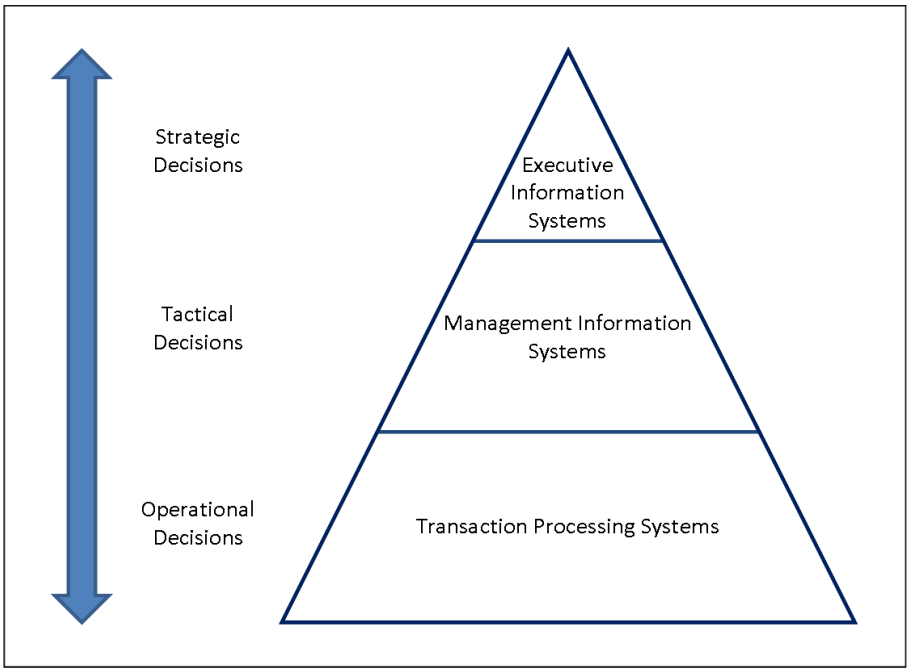
En realidad, generalizando podemos no sólo limitarnos a crear un sistema por departamento, sino que tenemos la posibilidad de definirlo para funcionalidades más concreta. De esta forma podemos hablar de:

- Sistema de Análisis de la Cartera de Clientes.
- Estudio de Proveedores.
- Sistema de Gestión de Tesorería.

- Sistema de Control Presupuestario.
- Sistema de Control de Inventario.
- Etc.

Ejemplos de Sistemas de Toma de Decisiones

Como sabéis, podemos categorizar los diferentes sistemas que componen el Sistema de Información Corporativo en función del tipo de decisión:



Esto nos lleva a concluir que distintos tipos decisión, nos llevan inevitablemente también a diferentes tipos de Sistemas de Inteligencia de Negocio:

- **Sistemas Operativos** para Decisiones Operativas. Para las decisiones de los empleados que están el nivel más ligado a la actividad directa de la compañía o jerárquicamente el de menos responsabilidad.
- **Sistemas Tácticos** para Decisiones Tácticas. Son aquellos destinados a los mandos intermedios y que se centran principalmente en la planificación, en el control y el seguimiento de los procesos.
- **Sistemas Estratégicos** para Decisiones Estratégicas. Es la información que necesitan los directivos al más alto nivel de la compañía para poder guiarla en el medio-largo plazo.

Un buen ejemplo de ello es el **Cuadro de Mando Integral (CMI) o Balanced Scorecard (BSC)**[5]. En 1992, Robert Kaplan y David Norton publicaron en la Hardvard Business Review un sistema gerencial para evaluar el seguimiento de la compañía. A grandes rasgos consiste en:

[5] http://es.wikipedia.org/wiki/Cuadro_de_mando_integral

- Identificar los principales objetivos que debe cumplir la compañía. Ejemplo: Aumentar las Ventas.
- Determinar para cada objetivo un método para cuantificarlo: establecer una o varias métricas para medir el valor del objetivo. Ejemplo: La métrica para el objetivo "Aumentar las Ventas" podría ser "Facturación en Euros Año Actual contra el Año Anterior".
- Definir una meta para cada objetivo, para que el valor de la métrica nos indique si estamos cumpliéndola o no. Ejemplo: "Crecer un 10%".
- Agrupar dichos objetivos en cuatro perspectivas: Financiera, Clientes, Procesos Internos y Recursos. Ejemplo: El objetivo "Aumentar las Ventas" lo incluiríamos en la perspectiva Financiera.
- Definir las dependencias o relaciones existentes entre los objetivos.

Ejemplos de Sistemas Expertos

En la actualidad es prácticamente imposible que hablemos de un Sistema Experto de propósito general; es decir, que haya un sistema al cual le planteemos cualquier problema de negocio y sea capaz de analizarlo, modelarlo y automatizar las tareas necesarias para resolverlo.

Lo más habitual es que en base a un problema de negocio u objetivo que nos planteemos, desarrollemos un sistema que nos permita resolverlo; es decir, una relación 1:1. Otra cosa diferente es que la tecnología que usemos (infraestructura, software) luego sea reaprovechable en otros casos.

Algunos ejemplos que podemos encontrar:

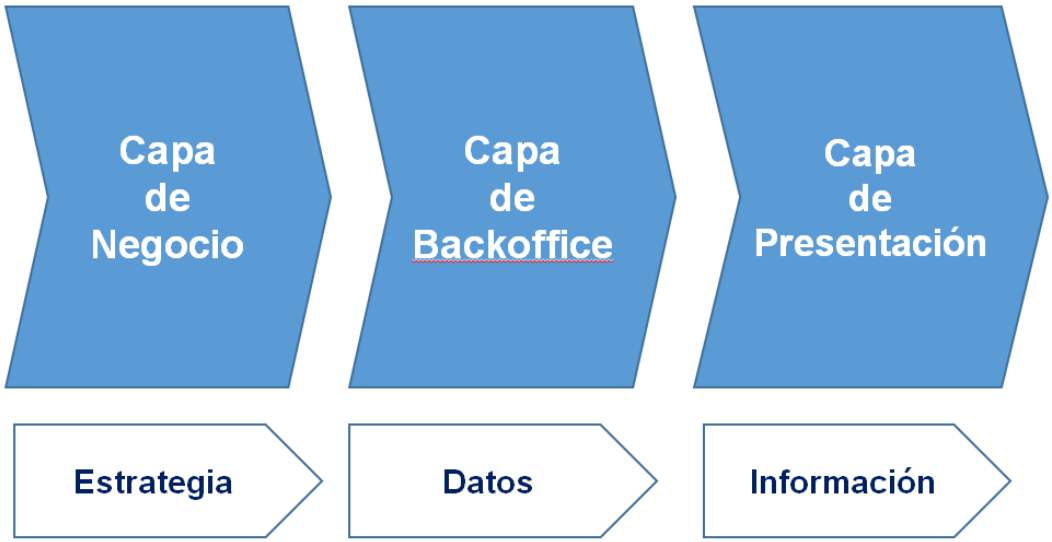
- **Marketing Automation**. En base al análisis de nuestros clientes y potenciales, realizar campañas específicas por cualquier medio digital (email, web, redes sociales, app, etc.) personalizando la oferta y además pudiendo hacer un seguimiento de la misma. Os paso un ejemplo de Marketo:

https://www.youtube.com/watch?v=j8gP7JLJ0-w

- **Reaprovisionamiento de Tiendas.** Estimar las ventas de uno o varios productos de una tienda, para calcular cuando podría haber una rotura de stock y así reaprovisionarla. Detectar stock ocioso y moverlo a tiendas donde puede dársele salida.
- **Optimización de precios**. Cálculo automático del precio en función de la previsión de la demanda. Como ejemplo,
- **Optimización del riesgo**. Como sabéis muchos de vosotros, la minimización del riesgo y la prevención del fraude es uno de los grandes caballos de batalla del mundo financiero y asegurador. En la actualidad, gracias al *machine learning* se pueden bloquear transacciones fraudulentas o avisar al responsable o al empleado de un posible riesgo de una operación de un cliente concreto.

ARQUITECTURA DEL SISTEMA DE ANALÍTICA DE NEGOCIO

Como ya hemos visto, la arquitectura de un Sistema de Inteligencia de Negocio está formada por las siguientes capas:



Esta arquitectura es perfectamente válida cuando hablamos de los sistemas que van a dar soporte a la Ciencia de los Datos y que de forma generalizada hablaremos como Sistemas de Analítica de Negocio, y además necesaria para realizar cualquiera de los tipos de análisis que vimos en el apartado anterior.

Permitidme que revisemos las 3 capas del sistema, aplicando los conceptos a la analítica de negocio.

Capa de Negocio

Si hay algo que me gustaría que se os quedara claro en esta parte del módulo, es que la Analítica de Negocio ha de solucionar un problema de negocio. Resolver dicho problema, como hemos comentado en apartados anteriores, debe suponer un beneficio para la compañía. Por tanto, un Sistema de Negocio debe definirse por:

- **Objetivo**: Problema o necesidad de negocio a solucionar.
- **Beneficio**: Retorno esperado de la resolución del problema.
- **Forma de presentación y comunicación**: forma en que se espera consumir el resultado del sistema.
- **Plazo:** Fecha máxima en la que se espera tener el sistema en producción.
- **Presupuesto y Recursos:** Partidas económicas y personal destinado para ejecutar la implantación.

En el próximo tema, examinaremos con más detalle estas cuestiones.

Capa de Backoffice

La capa de datos del Sistema de Analítica de Negocio se puede descomponer en:

- **Fuentes de datos**. Como ya discutimos en el apartado 1, el punto de origen de los datos es básicamente una de las grandes diferencias entre la Inteligencia de Negocio tradicional y la Analítica de Negocio actual. El objetivo de negocio determinará las fuentes de datos a utilizar y en función de las características de cada una de ellas, nos determinará la complejidad del problema y la tecnología a utilizar. Fijaos, desde un punto de vista de negocio podemos caracterizar las fuentes de datos por 3 variables:

    - Internas o Externas. Las fuentes de datos internas suelen ser más fáciles de acceder, más fiables (entendiendo fiabilidad como la confianza del negocio en la información que se puede extraer de ella) y presentar un menor coste de la información. Por otra parte, la información de fuentes externas nos pueden aportar información imposible de obtener de fuente de datos propias (mayor conocimiento).
    - Estructuradas o No Estructuradas. Las fuentes de datos estructuras permiten producir información con mayor rapidez y fiabilidad, frente a las no estructuradas que potencialmente pueden aportarnos un conocimiento nuevo con datos que antes no podríamos explotar. Las fuentes de datos estructuradas requieren de una tecnología madura, más asequible y reconocida que las no estructuradas, mucho más especializada.
    - Hombre o Máquina. Otro factor es la posibilidad de acceder a información directamente generadas por máquinas. Esto elimina la subjetividad humana de la ecuación y nos da una mayor seguridad o confianza en la información extraída a partir de estos datos. Sin embargo, no es posible obtener información "no humana" de todos los procesos de negocio y la adquisición de datos (por mucho que sean máquinas) puede no ser trivial ni tampoco asequible.

- **Ingesta o Captura**. Es el procedimiento por el cual se realiza la captura de los datos de las diferentes fuentes. Desde un punto de vista de negocio, lo que determina el proceso de ingesta es conocer la frecuencia con la cual negocio quiere disponer del dato final:

- Tiempo Real (o similar). Es una de las mayores oportunidades / necesidades de negocio, pero hay un precio a pagar. Tener la información actualizada prácticamente al instante supone un esfuerzo mayor tanto en recursos de infraestructura como de ingesta y, por tanto, ha de valorarse muy bien con el beneficio esperado. Podemos utilizar herramientas como **Amazon Kinesis, Flume o Kafka**.
- Períodico (Batch). Realizar cargas de información con cierta frecuencia, simplifica mucho el proceso y por tanto nos permite optar por técnicas más asequibles y conocidas. En este caso podemos emplear un amplio abanico de tecnologías, desde software libre como **Talend o Pentaho** hasta especializado como **Informatica o Denodo**.

- **Almacenamiento**. Como veréis a lo largo de este máster, para entrenar modelos de *machine learning* confiables es necesario una cantidad indecente de datos. Si determináis que para solucionar el problema de negocio que se plantea, es necesario un modelo de *machine learning*, en ese caso es imprescindible contar con datos suficientes y por tanto, acumular la mayor cantidad de datos posible. Según la naturaleza de los datos a analizar, tendremos que usar diferentes tecnologías de almacenamiento:

  - Estructurados. Los datos estructurados provenientes de bases de datos o en formato tabular son fácilmente almacenables en repositorios convencionales. Se usan las **tecnologías de *Datawarehousing*** que visteis en el módulo anterior: **Oracle, Teradata, Microsoft SQL Server o PostgreSQL**.
  - Semiestructurados. Son datos como los ficheros XML o JSON que aunque en parte tienen cierta lógica, tiene parte de contenido no estructurado. Para procesarlos es muy habitual utilizar las **tecnologías NoSQL** como **MongoDb, Cassandra, DynameDB o BigTable**.
  - No Estructurados. Si requerimos procesar datos sin ningún tipo de estructura como documentos de texto, imágenes, audio o vídeo, entonces deberemos optar por **sistemas de almacenamiento basados en ficheros**, como son los sistemas basados en **Hadoop**, como **Cloudera, Hortonworks o MapR**.

  Observad que independientemente de que el consumo de la información sea en tiempo real o de forma periódica, siempre es recomendable que almacenos los datos procesados para guardarlos para un posterior uso. Además, daos cuenta también que existe numerosa información que es volátil, no perdura siempre (externa, provenientes de fuentes web, redes sociales o interna, como la de sensores, telemetría o móviles) y que por tanto es conveniente persistir.

- **Procesamiento o Modelado**. Como ya vimos en el apartado anterior, el objetivo de negocio determina el tipo de análisis a realizar y esto a su vez nos definirá las técnicas y tecnologías a emplear:

  - Análisis Descriptivo. Podemos abordarla con las técnicas de Inteligencia de Negocio tradicional. La **tecnología OLAP** y el modelado dimensional de los datos nos van a permitir saber qué ha ocurrido mediante la definición de dimensiones y de métricas que agregan eventos o transacciones de negocio.

- Análisis Diagnóstico. Además de los cubos OLAP (donde la información reside agregada, una o varias métricas por una o varias dimensiones de análisis), necesitamos la posibilidad de profundizar en los datos y bajar a un mayor nivel de detalle, para poder entender la información que nos muestra. Para ello se emplean bien **técnicas de consultas directas** (con datos previamente filtrados) o las **técnicas *in-memory*** (donde los datos residen a máximo nivel de detalle en memoria y las agregaciones se producen en la medida que las solicita el usuario). La mayoría de softwares de BI cuentan con esta funcionalidad, pero son conocidos como especialistas en tratamiento de datos *in-memory* **Qlik, Tableau, Tibco o SAP Hana**.
- Análisis Predictivo. Cuando queramos estimar el valor de una variable, detectar patrones o reglas en un conjunto de datos o extraer información de datos no estructurados, tenemos que utilizar técnicas de procesamiento más allá del BI tradicional. Aunque los grandes fabricantes de software como IBM, SAS, o Microsoft cuentan con soluciones de análisis predictivo dentro de sus suites de BI, la tendencia del mercado es **utilizar software específico** para realizar esta tarea: **R, Python, KNIME, Rapidminer o Alteryx** son algunos de los más utilizados. También es importante destacar la tendencia en el **uso de** las **APIs** que proveedores como **Amazon** o Google ofrecen con "paquetes encapsulados" donde solucionan problemas de negocio como pueda ser el reconocimiento de voz, imágenes, traducción, etc.
- Análisis Prescriptivo. El sistema ha de permitir la programación de reglas de negocio sobre la variable estimada para disparar una serie de acciones configuradas mediante **workflows** o flujos de trabajo. Como las acciones a desencadenar pueden afectar a diferentes sistemas y tecnologías, la clave de los sistemas prescriptivos es la **integración,** con lo que ya no sólo depende del propio sistema sino del resto que tenga capacidad de ser interactuado vía API por ejemplo. Algunos softwares comerciales que ya lo están haciendo son **SAS**, **IBM** o **Salesforce**.

Capa de Presentación

Como parte de la capa de negocio, hemos de definir claramente la forma en que el sistema va a comunicarse e interactuar con el usuario consumidor de la información.

Las mismos tipos de interfaz de cualquier sistema de Inteligencia de Negocio son perfectamente válidos:

- **Informes o *Reporting***. Los informes son, han sido y serán la forma más común de comunicar información estática. Además es el formato más económico de generación y distribución de la información. Los usuarios de negocio están habituados a consumir la información en este formato y eso genera menos resistencia al uso de nuevas herramientas o plataformas. Excel y PDF son los formatos más utilizados para consumir la información y la práctica totalidad de softwares de BI que utilicemos nos permitirán generar este tipo de formatos.
- **Cuadros de Mando o *Dashboards*.** Cuando queremos presentar la información de forma muy visual, agregada y con un interfaz interactivo y dinámico, la mejor forma de representación son los cuadros de mando. Tecnologías como **Qlik, Tableau o Microsoft PowerBI** son las más populares para este tipo de interfaces.

Además de eso, están cobrando especial importancia:

- **Mensajería instantánea**. Se está popularizando el uso de aplicaciones que emulan un interfaz conversacional con los usuarios: los ***chatbots***. **Microsoft, Telegram, Skype, Google o Amazon** permiten desarrollar bots que, integrados con nuestros sistemas de analítica de negocio interactuar con los usuarios y proporcionarles la información que necesita. Podéis ver un ejemplo aquí:

  https://www.youtube.com/watch?v=9urHjGQoiTw

- **Alertas**. El sistema permite al usuario bien programarlo o bien de forma automática que le envíe una alerta cuando se cumple una determinada condición. Por ejemplo, cuando se prevee que el precio desciende por debajo de un determinado umbral se le envía al usuario un email, un alerta vía app móvil, una ventana emergente en su navegador web o un simple SMS.
- **Asistentes domóticos**. **Amazon Echo, Google Home o Samsung Otto** son dispositivos conocidos como "altavoces inteligentes", pero que cuentan con capacidad para reconocimiento de voz y desencadenar determinadas tareas simples. Aunque todavía no es una tecnología madura, sin duda será un interfaz de comunicación muy popular.

Lo realmente importante es tener en cuenta las preferencias de los usuarios para lograr nuestro objetivo, y por tanto será importante tener en cuenta el dispositivo que van a utilizar (móvil u ordenador) y la mejor manera de comunicar la información que vamos a producir (texto, visualización, imagen, audio o vídeo).

A modo de resumen, en la tabla siguiente podéis ver el tipo de tecnología a utilizar por cada uno de los pasos

Negocio	• **Objetivos** => Determina el Tipo de Análisis de Negocio a utilizar • **Beneficios** • **Plazo, Presupuesto y Recursos**
Fuentes de datos	• **Internas** (ERP, CRM, MRP, ECM) / **Externas** (Redes Sociales, Open Data, Proveedores de Datos) • **Estructuradas** (Bases de Datos, ficheros Excel) / **No Estructuradas** (documentos de texto, correos electrónicos, imágenes, audio, vídeo, etc.) • **Hombre** (introducción manual de datos en cualquier sistema) o **Máquina** (Sensores, Telemetría, Logs de servicios)
Captura	• **Real Time** => Amazon Kinesis, Flume, Kafka • **Batch** => Tecnología BI tradicional: Informatica, Talend, Pentaho, Microsoft SQL Server, etc.
Almacenamiento	• Datos **estructurados** => Tecnología ***Datewarehouse*** **BI tradicional**: Oracle, Microsoft, Teradata, PostgreSQL. • Datos **semiestructurados** (ficheros JSON, XML, ..) => Tecnología **No SQL**: MongoDB, Cassandra, Amazon DynamoDB, Google BigTable • Datos **no estructurados** (texto, audio, imágenes o vídeo) => Tecnología **basada en ficheros (Hadoop)**: Cloudera, Hortonworks, MapR.
Procesamiento	• Analítica **Descriptiva** => Tecnología **OLAP BI tradicional:** Microstrategy, Business Objects, Microsoft SQL Server. • Analítica **Diagnóstica** => Tecnología **Direct Query** o **In-Memory**: Tableau, Qlik, Microsoft PowerBI. • Analítica **Predictiva** => Tecnología **específica** como R, Python, Rapidminer o Knime. • Analítica **Prescriptiva** => **Integración**. APIs. IBM,SAS, Salesforce, Amazon o Google.
Consumo	• Interfaces **Visuales** => ***Reporting*** o ***Dashboards*** : Qlik, Tableau o Microsoft • Interfaces **Conversacionales** => ***Chatbots*** o **Alertas** con Google, Amazon o Samsung

MERCADO ACTUAL DE SOLUCIONES DE ANALÍTICA DE NEGOCIO

Tras entender y examinar las diferentes tecnologías que componen un Sistema de Analítica de Negocio, el siguiente paso es conocer las soluciones que el mercado nos ofrece. Para abordar este reto, lo conveniente es atender a lo que los analistas de mercado TIC nos dicen. Existen numerosas consultoras que se dedican a realizar por nosotros estos estudios, entre ellas podemos encontrar: Gartner, IDC, Forrester, Dresner o BARC.

De entre todos ellos, vamos a utilizar uno de los más famosos estudios de Gartner que abarcan las tecnologías que empleamos, que es el denominado ***"Magic Quadrant for Business Intelligence and Analytics Platforms".*** Cubren todo el proceso de analítica descriptiva y diagnóstica tradicional. Tal y como comentamos en el apartado anterior, desde la captura en modo batch, procesamiento OLAP o in memory (capa de datos) y la generación de informes o dashboards (capa de presentación)

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

Source: Gartner March (2023)

https://www.qlik.com/es-es/gartner-magic-quadrant-business-intelligence

Esto es tiene que dar una idea de las soluciones comerciales que existen y su grado de madurez e implantación en otras compañías. Es una buena guía cuando no controlamos o no disponemos de una tecnología en nuestras compañías, pero lo cual no quiere decir que lo tomemos como dogma de fe. Lo mejor en estos casos es probarlo: desarrollar pilotos y hacer una prueba de concepto que nos permita evaluarlo objetivamente.

RESUMEN

Ya sabemos qué es la ciencia de datos, tal y como adelantamos en el tema anterior, comprendemos mucho mejor dónde se ubica la ciencia de datos con respecto a otras disciplinas y otros conceptos que se manejan en las organizaciones: Big Data, Business Intelligence o Business Analytics.

Hemos acordado que la Ciencia de los Datos será el conjunto de procesos que nos permite sacar valor de los datos. Dicho valor hemos visto que vendrá determinado por el tipo de analítica que empleemos y que en función de esta analítica necesitaremos un sistema más o menos sofisticado que soporte todos los procesos que tendremos que llevar a cabo sobre los datos.

Sabemos pues ahora que la herramienta principal de la ciencia de datos son los Sistemas de Analítica de Negocio, que como hemos visto son una evolución de los Sistemas de Inteligencia de Negocio clásicos, que incorporan las tecnologías Big Data para aportarnos mayor funcionalidad y casos de uso.

Además, la Ciencia de los Datos se fundamenta en 3 disciplinas: Arquitectura y Gestión del Dato, Ciencias Exactas y Programación.

ENAE BUSINESS SCHOOL

C.I.F. G-30164099