Módulo 1. El camino a la implementación del *Big Data Analytics*

Unidad 1.1. Hacia una organización *data-driven*

1.1.1 El contexto empresarial y las organizaciones *data-driven*

Cada vez es más necesario, por parte de las empresas, tomar decisiones con base en datos. A lo largo de la historia, los CEO (*chief executive officer*) de las organizaciones han tomado muchas decisiones a partir de corazonadas o lecciones aprendidas de su experiencia. Sin embargo, la complejidad, la competitividad y el dinamismo de la economía actual hacen que cada vez se sumen más variables que intervienen en el éxito o fracaso de los proyectos. Es por esto que es necesario apoyarse en los datos para minimizar el componente de incertidumbre y poder tomar las decisiones con mayores probabilidades de éxito.

Hoy, cada vez más, está en boga la transformación de las empresas en organizaciones denominadas *data-driven*, pero ¿a qué se refiere esto? Se refiere a contar con las herramientas, los recursos y las habilidades, pero, sobre todo, a construir una cultura interna organizacional que actúe con base en los datos. Esta construcción tiene que desplegarse a lo largo y a lo ancho de toda la organización, no solo en áreas de *marketing* o de riesgo. Hoy esta construcción cultural en las organizaciones es una de las ventajas competitivas más importantes que poseen.

En este marco, cuanta más información, mejor, siempre que se posean los sistemas y recursos necesarios para saber qué hacer con ella. Poseer grandes cantidades de datos requiere mucha inversión en almacenamiento, mantenimiento y administración, así como los recursos humanos adecuados para su explotación. Igualmente, todos estos costos se ven licuados una vez que se comienza a optimizar la experiencia de los clientes y aumentan los retornos de las campañas de *marketing*.

Para que las empresas creen verdaderas organizaciones basadas en datos, deben centrarse en cómo los utilizarán de una manera única para su marca y sus necesidades comerciales. Es muy fácil que usted sea imaginativo sobre lo que le gustaría hacer, pero muy pocas organizaciones realmente tienen los medios para hacer todo lo que quieren con los datos que tienen disponibles. Es por esto que la mejor manera de empezar es centrándose en ese 20 % de su negocio que genera el 80 % de sus utilidades.

Finalmente, las compañías que no están basadas en datos no serán competitivas en el futuro cercano. Esto se transformará en un factor de supervivencia. La información se esparce a pasos cada vez más grandes. Hace algunos años hubiera sido imposible pensar en un curso como este en el que se aprenden conocimientos de alto valor desde el sillón del *living* con solo una *notebook*, auriculares y conexión a wifi. En los próximos años, las ideas y conocimientos que las empresas necesitarán para generar una experiencia de cliente realmente buena

serán más accesibles y se seguirán recopilando grandes cantidades de datos en tiempo real. Ciertamente, estamos mejor con más datos, pero en el futuro las compañías que tengan éxito no lo harán solo con base en una gran cantidad de datos: los aprovecharán combinándolos con aprendizaje automático e inteligencia artificial como el combustible final para potenciar su negocio basado en la experiencia.

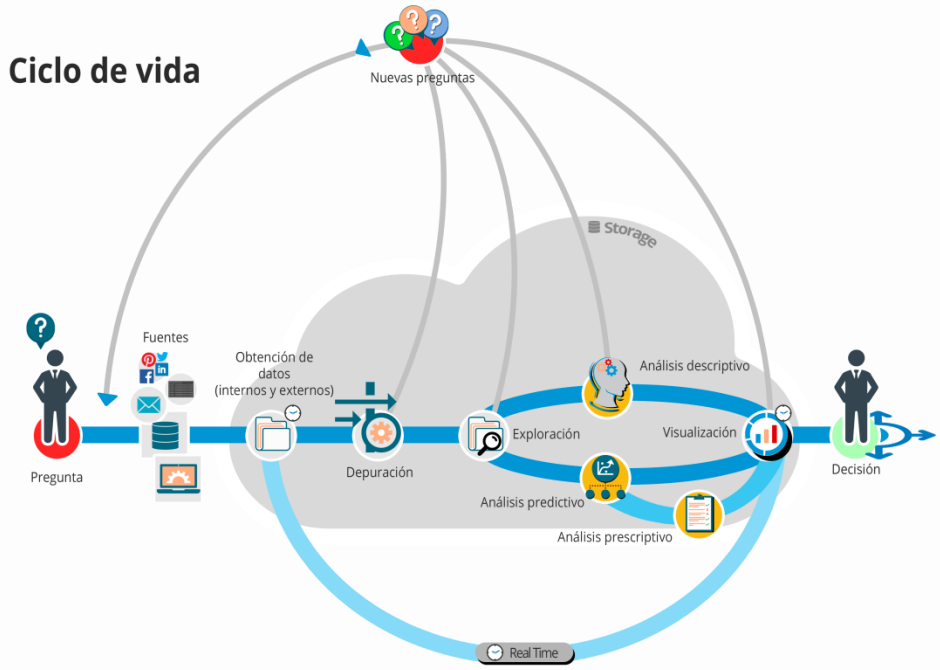
1.1.2 Ciclo de vida de los datos

Vamos a comenzar tratando de entender a grandes rasgos todo el flujo por el que pasan los datos desde que son generados hasta que dan sus frutos, con soporte en la toma de decisiones.

1. Análisis interno

Al inicio de toda solución, proceso o iniciativa informática o de cualquier naturaleza, siempre hay una persona con una necesidad. Se identifica un problema que se debe resolver, caracterizado por la incertidumbre, que es el motor de una iniciativa de cambio y evolución.
Así, una vez planteada una **pregunta o necesidad** por una o más personas en el ámbito de una organización que decide actuar con información y disponer recursos para ello, nace el ciclo de vida del dato, que es el combustible esencial para convertir una empresa en *data-driven*.

Figura 1. Ciclo de vida



Fuente: Panel.es, 2018, https://www.panel.es/wp-content/uploads/2018/04/ciclo_de_vida.png

Asimismo, en esta primera etapa se deben establecer aquellos KPI (*key performance indicator*) objetivo que permitirán comprender, en última instancia, si los *outputs* del proceso repagaron la inversión en tiempos y recursos que se abocaron a llevar a cabo este proceso. Es una

manera de fijar expectativas y de definir el mínimo valor esperado para poder llevar adelante el proyecto.

2. Recolección y filtrado de datos

En segundo lugar, es necesario **identificar las fuentes** en donde se generan estos datos y generar los procesos necesarios para poder hacer la extracción de ellos. Hay una innumerable cantidad de fuentes de información de las que disponen hoy las organizaciones. Para citar algunos ejemplos: CRM (*customer relationship management*), como por ejemplo, Oracle Siebel o Salesforce; ERP (*enterprise resource planning*), como por ejemplo, SAP (Systems, Applications, Products in Data Processing) y Microsoft Dynamics, interacciones en redes sociales, transacciones de e-commerce, Google Analytics de la web empresarial, envíos de *email marketing*, etcétera.

Dependiendo de la industria, habrá un sinfín de fuentes. Para citar algunos ejemplos: en las grandes *telcos* (compañías de telecomunicaciones), la información de las redes, el tráfico de datos, la geolocalización de móviles, las reproducciones de video, etcétera; en empresas financieras, las transacciones de tarjeta de crédito en distintos canales; en los bancos, los movimientos de Banelco/Link y del *homebanking*; en empresas IOT (*internet of things*), se puede disponer de datos no estructurados, como grabaciones de video, datos de fluctuaciones de temperatura, geolocalización de automóviles/camiones, etcétera.

La selección de los datos y de las fuentes dependerá de la naturaleza del problema planteado en la etapa anterior y de los objetivos que se hayan fijado como deseados. Los datos se someten a una serie de filtros en donde se descartan todos aquellos datos que están corruptos o que no aportan al análisis para alcanzar los KPI establecidos.

3. Obtención y almacenamiento de datos

En esta etapa se lleva a cabo la extracción de los datos y su posterior transformación en estructuras organizadas con criterio para hacerlas comprensibles e identificables por parte de los analistas. Aquí es donde entran tecnologías como Hadoop y Spark, o las más tradicionales, como el concepto de EDW (*enterprise data warehouse*), donde se almacenan grandes volúmenes de información. La diferencia entre ambos abordajes es el tipo de procesamiento, el costo, el mantenimiento, el volumen y el tipo de datos que puede almacenar.

La elección de un *framework* u otro va a depender de la organización, de sus estructuras de datos y de los tipos de usos que se le va a dar a la información.

4. Validación y depuración de los datos

Esta es la primera etapa de la explotación de los datos. Para llevar adelante cualquier análisis, primero es necesario realizar un abordaje exploratorio de las bases de datos. Esto implica validar todas las variables relevantes que se van a utilizar en el análisis que se vaya a llevar a cabo. Más adelante, nos vamos a centrar en qué tipos de usos analíticos se le

pueden dar a los datos, pero, sea cual fuere el seleccionado, la limpieza y depuración de los datos es un paso necesario en todos ellos.

Del **análisis exploratorio** de todas las variables o atributos disponibles, el analista además ya comienza a encontrar patrones que pueden de ser de gran utilidad en el momento del modelado. El analista descarta aquellas variables que no son de utilidad, toma acciones de limpieza, depuración y transformación de las que no tengan una calidad óptima y selecciona un conjunto de variables relevantes para el estudio que se quiere llevar adelante.

5. Análisis de los datos

Hay tres tipos de análisis que pueden realizarse con los datos que sirven de soporte a la toma de decisiones. La elección de uno u otro va a depender del problema planteado y del perfil del analista que esté llevando adelante la tarea. Más adelante, vamos a profundizar en este punto. En el cuadro a continuación, se explica cada uno de ellos de manera detallada.

Tabla 1. Tipos de análisis de datos para la toma de decisiones

Tipo de análisis	Definición	Técnicas
Análisis descriptivo	Es el tipo de análisis que se utiliza cuando se tiene un gran conjunto de datos sobre eventos del pasado. Para que estos datos cobren un significado que sume valor para la organización, es necesario procesarlo, resumirlo y expresarlo de una manera comprensible que ayude a la audiencia a comprender qué paso. Es el tipo de análisis mas básico y más rápido, y posee bajo grado de complejidad. Es muy utilizado por las áreas de planning y reporting.	- Tablas de frecuencias - Tablas de contingencia - Estadísticos descriptivos - Coeficiente de correlación - Gráfico de barras - Gráfico de sectores - Gráfico de líneas - Gráfico de dispersión - Histograma - Diagrama de cajas
Análisis predictivo	El análisis predictivo es la aplicación de técnicas y modelos matemáticos y estadísticos sobre una gran volumen de datos históricos para predecir con cierta probabilidad que va a suceder en el futuro.	- Técnicas de Machine Learning : - Regresión Lineal - Regresión logística - Arboles de decisión - Random Forest - Deep Learning
Análisis prescriptivo	Este tipo de análisis va más alla de predecir el futuro, recomienda distintos escenarios o rutas de acción que puede seguir una empresa, y cuantifica el efecto de cada una de ellos. De esta manera, los tomadores de decisiones, dependiendo el contexto, seleccionan una u otra alternativa con un bajo riesgo de incertidumbre y conociendo de antemano las implicancias.	Combina técnicas de la analíticas descriptiva y predictiva, en conjunto con un amplio conocimiento del negocio para realizar las recomendaciones de mayor ROI. Asimismo utiliza otras técnicas como encuestas, técnicas de simulación y de optimización.

Fuente: elaboración propia.

6. Visualización de los datos

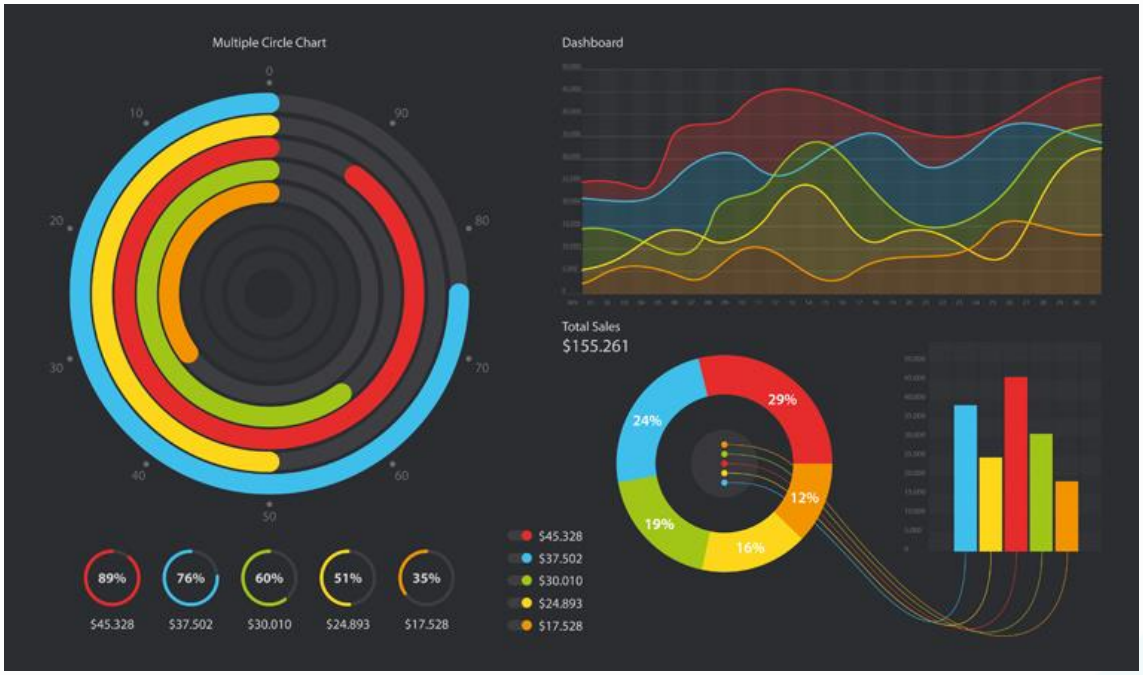
Con la *visualización de datos*, hablamos de la representación gráfica de la información analizada, utilizando elementos visuales como cuadros, gráficos y mapas. Estas herramientas de visualización de datos proporcionan una forma accesible de ver y comprender tendencias, valores atípicos y patrones en los datos, y disponen de un sinfín de recursos que atraen la atención del público, como formas, colores y movimientos (Tableau, s. f.).

La visualización de los datos es un arte que permite mantener los ojos de la audiencia en lo importante, lo cual lo enfoca y le permite comprender e internalizar más rápidamente el mensaje.

Se puede haber consumido horas de análisis, pero, si después no se le agrega un condimento de visualización que le permita al público comprender los resultados y "comprarlos", puede que todo el esfuerzo no haya servido para nada.

Hay varios *softwares* disponibles en el mercado que son muy simples y dinámicos, como por ejemplo, Tableau, PowerBI y QlikView, pero también se pueden aprovechar los recursos y complementos que tiene Excel en materia de visualización, o las librerías de gráficos de Python: Matplotlib, Seaborn. ggplot, Bokeh, pygal, Plotly, geoplotlib, entre otros.
A continuación y solo a manera ilustrativa, se observa un ejemplo de tablero de visualización de datos creado en Tableau:

Figura 2. Visualización de datos en gráficos radiales en Tableau



Fuente: Tableau, 2018, https://www.analiticaweb.es/tableau-visualizacion-datos-graficos/

7. Decisión de negocio y retroalimentación

En la etapa final del ciclo de vida, se encuentra el decisor, quien toma la información presentada, la comprende, la incorpora y determina un curso de acción que minimice los riesgos y maximice los resultados. Como *feedback* de este proceso, la mayoría de las veces surgen nuevas preguntas que retroalimentan nuevamente el circuito y generan que nuevas fuentes, nuevos datos y nuevos análisis se requieran para seguir sumando valor e incrementando la eficacia de las decisiones tomadas.

Cabe mencionar que, siempre que se implementan técnicas avanzadas de análisis de datos, suelen compararse con las tasas de efectividad que tenían los mismos procesos, pero con las fuentes de información anteriores. La diferencia sustantiva de una técnica respecto de la otra es la que brinda el aval para el desarrollo y la inversión en la exploración avanzada de datos.

1.1.3 *Big data* y *big data analytics*

Big data

Antes de comenzar a adentrarnos en la parte analítica del *big data*, es importante comenzar preguntándonos qué es *big data* y qué no lo es. El *big data* se puede conceptualizar como todo conjunto de datos que cumple con tres características necesarias y suficientes para poder definirlo así. Esta definición es de Gartner (2001) y se sigue tomando de referencia en la actualidad.

Estas características son conocidas como las *3 V* del *big data*:

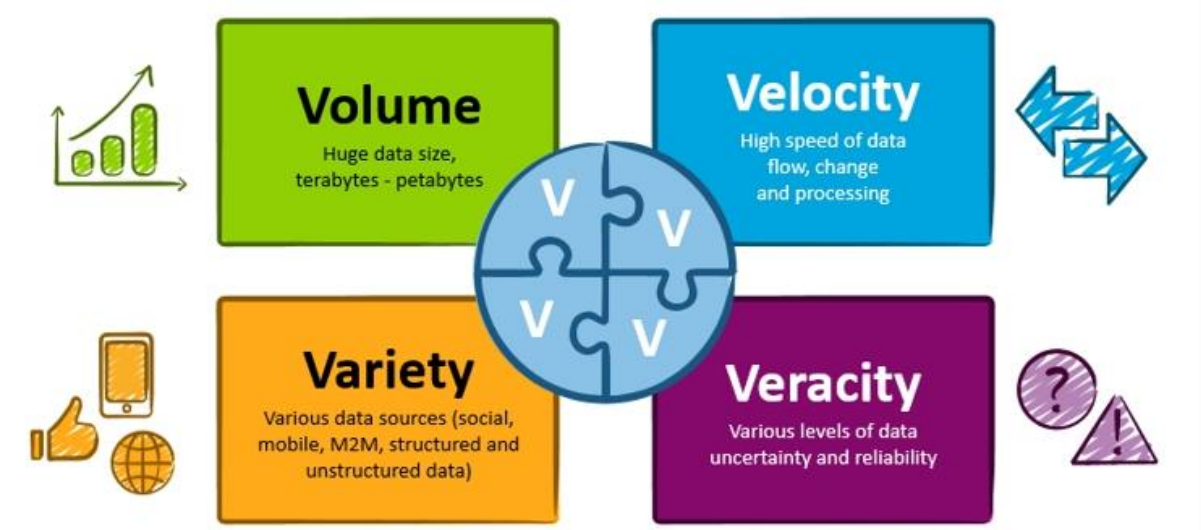
- **Volumen**: se refiere al tamaño de los datos. Se puede considerar como *big data* a un conjunto masivo de datos que exceden la capacidad de almacenamiento de un sistema de base de datos convencional.
- **Velocidad**: tiene que ver con el ritmo en el que se levantan, procesan y reenvían estos datos. La tendencia hoy va hacia la disponibilidad de la información *real time*, es decir, en tiempo real, para poder tomar decisiones más oportunas en el momento en que un hecho está sucediendo.
- **Variedad**: alude a la diversidad de fuentes de donde pueden venir los datos, por ejemplo, sistemas transaccionales, publicaciones en redes sociales, búsquedas en un navegador de *e-commerce*, visitas a páginas web, reproducciones de video, e incluso las imágenes que se "levantan" de una cámara de seguridad.

IBM (International Business Machines) incorporó esta definición, pero, años más tarde, propuso agregar una cuarta *V*:

- **Veracidad**: esta es una característica muy importante que ha llevado a las empresas incluso a crear áreas cuya tarea principal sea velar por la calidad de los datos que son "levantados" por los sistemas. Un modelo, por ejemplo, creado a partir de datos que se han roto o que dejaron de traer información o que, por alguna razón, traen información incorrecta, no arroja buenos resultados, lo que puede derivar en tomar

decisiones incorrectas.

Figura 3- Las cuatro *V* del *big data*



Fuente: Infodiagram, 2019, https://blog.infodiagram.com/2019/01/big-data-presentation-appealing-diagrams-ppt.html/bigdata15

Algunos incluso le sumaron una quinta *V* al concepto: el **valor**, en alusión a que el *big data* tiene que aportar conocimiento de valor para apoyar la toma de decisiones. A nuestro criterio, esta última *V* se aleja del concepto de *big data* e ingresa en el marco de otra disciplina que le da el sentido a esta tecnología detrás de los datos. Es la disciplina que acarrea los conocimientos necesarios para resolver los desafíos planteados por el *big data*.
Ahora que disponemos de toda información… ¿qué hacemos con ella? Aquí entra la ciencia de datos (*data science*) para combinar los conocimientos y la metodología de materias como la informática, la matemática y la estadística, y, al mismo tiempo, empaparse del conocimiento del negocio, para poder resolver los problemas que se plantean en el seno empresarial. Se centra en la extracción de los datos que brinda el *big data* y en su procesamiento para extraer los conocimientos que den soporte a la toma de decisiones más eficaces.

Big data analytics

El *big data analytics* es la disciplina que utiliza técnicas analíticas avanzadas en conjuntos de datos muy grandes y diversos que incluyen datos estructurados, semiestructurados y no estructurados, de diferentes fuentes y en diferentes tamaños, desde terabytes hasta zettabytes (IBM, s. f.).

Entre las distintas técnicas que se utilizan en este campo, se encuentran: *sentiment analysis*, *machine learning*, *clustering*, minería de datos, *deep learning*, estadísticas y procesamiento de lenguaje natural. El uso de estas técnicas permite identificar *insights* que de otra manera serían imposibles de encontrar entre la gran magnitud de datos que generan las organizaciones.

La analítica del *big data* es el soporte esencial de las organizaciones para identificar nuevas y mejores oportunidades en un contexto cada vez más competitivo. Al mismo tiempo, no solo deriva en acciones más estratégicas y exitosas, sino que también mejora las utilidades, reduce los costos de procesos ineficientes y genera un mayor valor para el cliente mejorando la *user experience*.

Las principales ventajas de la explotación de los datos a través de la analítica de *big data* son:

I. **Toma de decisiones de *marketing***: la gran velocidad y la capacidad de almacenamiento de Hadoop combinadas con la posibilidad de analizar nuevas fuentes de datos permiten que las empresas puedan tomar mejores decisiones, de manera más rápida y oportuna respecto de las campañas que se envían a sus clientes.

II. **Nuevos productos y servicios**: con la posibilidad de medir las necesidades y la satisfacción del cliente a través de la analítica, las empresas están en condiciones de ampliar sus negocios y ser más asertivas en sus ofertas de productos y servicios para los clientes.

III. **Mejora en los procesos**: con la analítica de *big data*, también se puede tener un mejor panorama acerca de cómo están funcionando los procesos internos de la compañía, como por ejemplo, tiempos de respuesta de reclamos, tiempos de logística de productos, nivel de calidad de lotes de producción, cantidad de quiebres de *stock*, etcétera.

IV. **Prevención de fraude**: sobre todo en empresas financieras, con la posibilidad de medir la información en tiempo real, se pueden crear alertas que permitan frenar posibles casos de fraude en el momento en que están ocurriendo.

V. **Fidelización de clientes y prevención de *churn***: con toda la información histórica del comportamiento de los clientes que se acumula en las bases de datos de las empresas, se pueden elaborar tendencias que permitan detectar cuándo un cliente se está desviando de su comportamiento normal esperado. Esto le da la posibilidad a las empresas de accionar sobre él con políticas de fidelización que previenen que se vaya a la competencia (PowerData, s. f.).

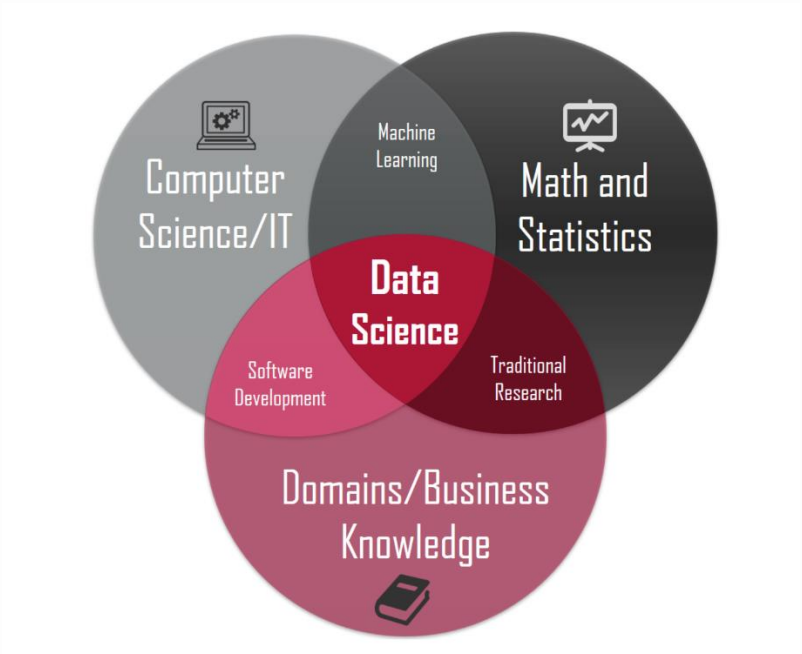
Unidad 1.2. La gestión de los datos organizacionales

1.2.1 Data science

En los últimos años, se han realizado grandes inversiones en infraestructura empresarial que han mejorado la capacidad de recopilar datos en toda la empresa. Prácticamente todos los aspectos del negocio ahora están abiertos a la recopilación de datos: operaciones, manufactura, gestión de la cadena de suministro, comportamiento del cliente, rendimiento de la campaña de *marketing*, procedimientos de flujo de trabajo, etcétera. Al mismo tiempo, ahora está disponible la información sobre eventos externos, como tendencias del mercado, noticias de la industria y movimientos de los competidores. Esta amplia disponibilidad de datos ha llevado a un creciente interés en los métodos para extraer información útil y conocimiento de los datos, el ámbito de la ciencia de datos.

Data science, o ciencia de datos, se encarga de estudiar los datos con un abordaje metodológico que se apoya en otras ciencias, como la matemática, la estadística y la informática.

Figura 4. *Data science*



Fuente: Neoland, 2019, https://www.neoland.es/blog/el_futuro_del_data_science

El científico de datos es el encargado de aplicar sus conocimientos en materia de inteligencia artificial y *machine learning* para explotar el gran volumen de datos que tienen las organizaciones y generar algoritmos que puedan traducir toda la complejidad a un lenguaje más simple, por ejemplo, en *scores* de clientes o clústeres. Entonces, el objetivo general de todo científico de datos consiste en transformar la información en una estructura comprensible y útil para el negocio. Otra cualidad importante del científico de datos es que complementa sus conocimientos técnicos con la capacidad de comprensión e interpretación del negocio en el

que se desenvuelve. Este conocimiento le permite ser más estratégico y preciso al momento de seleccionar los datos y poner en práctica las técnicas de *machine learning*.

Probablemente, las aplicaciones más amplias de las técnicas de minería de datos se encuentran en el *marketing* para tareas, como el *marketing* dirigido, *online advertising* y recomendaciones para *cross-selling*. La minería de datos se utiliza para la gestión general de la relación con el cliente para analizar su comportamiento con el fin de gestionar su fidelización y maximizar el valor esperado (NPV). La industria financiera utiliza la minería de datos para la calificación crediticia (*scoring*) y acciones comerciales, y en las operaciones a través de la detección de fraudes y la gestión de la fuerza laboral. Los principales *retailers* del mundo, desde Walmart hasta Amazon, aplican la minería de datos en todos sus negocios, desde el *marketing* hasta la gestión de la cadena de suministro. Muchas empresas se han diferenciado estratégicamente de la ciencia de datos, a veces hasta el punto de convertirse en empresas de minería de datos.

Nos parece importante aclarar que hay cierta confusión sobre los conceptos *ciencia de datos* y *minería de datos*, que, a menudo, se usan indistintamente. En un alto nivel, la ciencia de datos es un conjunto de principios fundamentales que guían la extracción de conocimiento de los datos. La minería de datos, o *data mining*, es la extracción de conocimiento de los datos a través de tecnologías que incorporan estos principios. Como término, *ciencia de datos* a menudo se aplica de manera más amplia que el uso tradicional de *minería de datos*, pero las técnicas de minería de datos proporcionan algunas de las ilustraciones más claras de los principios de la ciencia de datos.

Para concretizar todo lo conceptualizado anteriormente, creemos útil enumerar las responsabilidades cotidianas de un *data scientist* o científico de datos en un ámbito empresarial. Algunas de sus tareas más representativas son:

- extracción, limpieza y procesamiento de grandes volúmenes de datos;
- predicción de problemas de negocios;
- desarrollo de modelos de aprendizaje automático y métodos analíticos;
- implementación de *data mining* utilizando métodos de última generación;
- presentación de los resultados de manera clara;
- monitoreo de la eficiencia de los algoritmos implementados.

1.2.2 Data analytics

Data analytics es la ciencia de obtener ideas de fuentes de información sin procesar. En este rol es fundamental el conocimiento del negocio para detectar las tendencias y métricas relevantes para la organización. Su dinamismo está al día con las variaciones en el entorno organizacional, con las decisiones que se generan en áreas de *marketing*, siguiendo el impacto de las campañas y monitoreando que los modelos de inteligencia artificial implementados sigan operando dentro de los rangos esperados. No aplica algoritmos para sus análisis, pero sí se apoya en la estadística para demostrar o refutar teorías.

El perfil de un analista de datos permite evaluar propuestas para proyectos de minería de datos. Por ejemplo, si un colega de negocio o consultor propone mejorar una aplicación comercial particular mediante la explotación de conocimiento desde los datos, debe poder evaluar la propuesta sistemáticamente y decidir si es sólida o defectuosa. Esto no significa que podrá saber si realmente tendrá éxito (en los proyectos de *data mining*, muchas veces es necesario abocarse a la tarea y simplemente intentarlo), pero debería ser capaz de detectar defectos obvios, suposiciones poco realistas y piezas faltantes de información.

Las responsabilidades de un experto en *data analytics* son:

- identificar cualquier problema de calidad de datos en la adquisición de datos;
- sugerir acciones de mejora y oportunidades de negocio con base en los análisis de la información;
- mapear las fuentes de información donde hay oportunidades de extraer datos de valor al negocio;
- coordinar con los ingenieros para recopilar nuevos datos y ponerlos a disposición en ambientes de Hadoop, Spark o EDW, según sea el caso;
- realizar análisis estadísticos de datos comerciales;
- documentar los tipos y la estructura de los datos comerciales;
- "levantar" la necesidad de generar modelos predictivos o clústeres con los *data scientist*;
- monitorear la correcta implementación de los modelos.

1.2.3 SQL y Python como herramientas para la explotación de datos

Todas estas tareas analíticas para sacar provecho de los datos pueden realizarse con muchas herramientas que existen en el mercado. Las hay licenciadas y *open source*, y utilizan distintos lenguajes, algunos son propios y otros, multiplataforma. Para este curso, vamos a hacer alusión a dos. En primer lugar, SQL (Structured Query Language), que es un lenguaje propio de las bases de datos relacionales. SQL "brinda la posibilidad de realizar consultas con el objetivo de recuperar información de las bases de datos de manera sencilla" (Plasencia Prado, s. f., https://devcode.la/blog/que-es-sql/). Es un lenguaje más orientado a análisis descriptivos y manipulación de información estructurada, de fácil aprendizaje y muy versátil, que se utiliza en múltiples plataformas.

Por otro lado, tenemos Python, que se ha convertido en uno de los lenguajes de programación interpretados más populares. En los últimos 10 años, Python ha pasado de ser un lenguaje informático científico innovador a uno de los lenguajes más importantes para la ciencia de datos, el aprendizaje automático y el desarrollo de *software* en general en la academia y la industria. El soporte mejorado de Python para bibliotecas (como pandas y *scikit-learn*) lo ha convertido en una opción popular para las tareas de análisis de datos. Combinado con la fuerza general de Python para la ingeniería de *software* de propósito general, es una excelente opción como lenguaje principal para crear aplicaciones de datos.
Entre las principales ventajas de este lenguaje de programación para la ciencia de datos, tenemos:

- **Simplicidad.** Python es un lenguaje "fácil de usar y toma menos tiempo en la codificación. Tampoco hay limitación para el procesamiento de datos. Puede calcular datos en cualquier tipo de equipo y entorno" (Universidad Internacional de Valencia, 2019, https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/).

- "**Compatibilidad.** Hadoop es la plataforma de *big data* de código abierto más popular actualmente y la compatibilidad inherente de Python es" (Universidad Internacional de Valencia, 2019, https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/) una de las razones más importantes para posicionarlo como uno de los favoritos entre los lenguajes de programación.

- "**Facilidad de aprendizaje.** En comparación con otros idiomas, Python es fácil de aprender incluso para [principiantes]" (Universidad Internacional de Valencia, 2019, https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/). Las principales razones son: "cuenta con amplios recursos de aprendizaje, [posee] un código legible y se rodea de una gran comunidad" (Universidad Internacional de Valencia, 2019, https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/), que sirve de importante soporte en el momento en que surgen inconvenientes con el código.

- **Variedad de paquetes.** "Python tiene un poderoso conjunto de paquetes para una amplia gama de necesidades de análisis y ciencia de datos" (Universidad Internacional de Valencia, 2019, https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/). Algunos de los paquetes más populares son: "NumPy, Pandas, Scipy, Scikit-learn, PyBrain, Tensorflow, Cython PyMySQL, BeautifulSoup o iPython" (Universidad Internacional de Valencia, 2019, https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/).

- **Visualización de datos.** "Python para *big data* ha ido mejorando su oferta en esta materia" (Universidad Internacional de Valencia, 2019, https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/). Algunas librerías son: Matplotlib, Seaborn, Pygal, Bokeh y Plotly.

1.2.4 Otras disciplinas complementarias

La gestión de los datos en una organización implica varias funciones que trabajan de manera sincronizada para asegurar la disponibilidad de los recursos de datos necesarios para que los científicos y analistas de datos tengan a su disposición para poder explotarlos y aportar valor a la toma de decisiones empresariales.
Según el Data Management Association Internacional (2007), las nueve funciones de la gestión de datos son:

1. **Gobierno de datos o *data governance***: se encarga de la planeación, supervisión y

control en la gestión y uso de datos.

2. **Arquitectura, análisis y diseño de datos**: se encarga de la modelación y especificación de datos. Define la arquitectura de las bases de datos y la integración de las fuentes y la arquitectura de almacenamiento de estas.

3. **Gestión de la base de datos**: su función es el diseño de la base de datos, sus implementación y soporte. Brinda respaldo y recuperación de los datos en caso de incidentes.

4. **Gestión de la seguridad de los datos**: asegura la confidencialidad y el control de acceso a información sensible por parte de los usuarios.

5. **Gestión de la calidad de los datos**: define, mejora y controla la calidad de los datos.

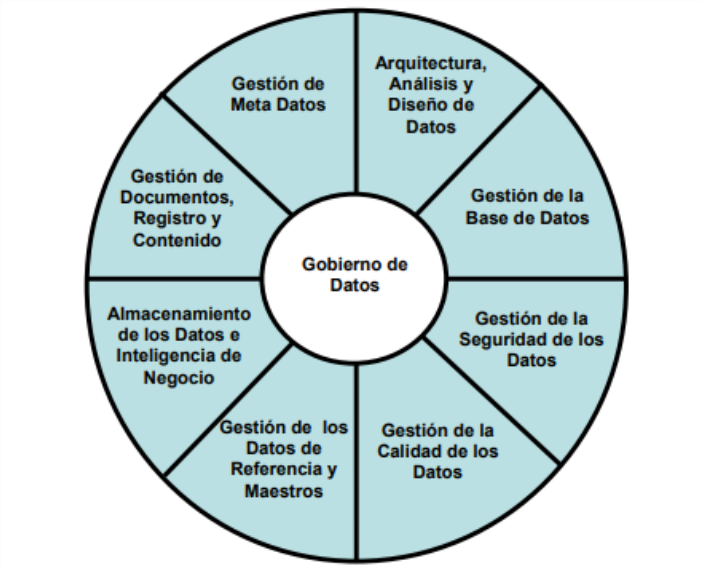
6. **Gestión de datos de referencia y maestros**: administra versiones maestras y copias de seguridad de los datos.

7. **Gestión del almacenamiento de datos e inteligencia de negocio**: se encarga de realizar informes y análisis de las estructuras de datos existentes. Implementa *data warehouses* y *data marts*, aplicaciones analíticas, y da soporte a profesionales del negocio en el adecuado uso de las bases de datos.

8. **Gestión de documentos, registro y contenido**: gestiona datos fuera de las bases de datos, en las fuentes no migradas.

9. **Gestión de metadatos**: "levanta" los requerimientos de metadatos e integra, controla y proporciona los metadatos. Administra y distribuye los glosarios y repositorios de metadatos.

Figura 5. Gobierno de datos



Fuente: Data Management Association Internacional, 2007, p. 8.

1.2.5 *Big data* y ética

Cada día que pasa, se acumulan en el mundo miles de petabytes de información provenientes de redes sociales, dispositivos móviles, electrónicos, sensores, redes, etcétera, que están a disposición de las grandes empresas. El *big data* creció de manera muy vertiginosa y las regulaciones en materia de privacidad y confidencialidad de los datos no han alcanzado a tapar todos los grises y a cubrir todos los aspectos para proteger la identidad de las personas en la era digital.
Son numerosos los desafíos que se deberán enfrentar en relación con el *big data* respecto de la privacidad de las personas. A medida que la cantidad de datos crece, surgen dilemas relacionados con la ética (NIC Argentina, 2018).

> Todos nuestros datos quedan registrados en el mundo digital y esta información puede ser utilizada para diferentes funciones con las cuales hemos acordado o aceptado sin tener el conocimiento adecuado para hacerlo: desde campañas publicitarias cada vez más [*targueteadas*]… a políticas públicas orientadas a mejorar la calidad de vida de las personas. (NIC Argentina, 2018, https://nic.ar/es/enterate/novedades/que-es-big-data).

En las manos de las organizaciones, está el poder de proteger la identidad y privacidad de las personas. La norma argentina aún no se ha puesto al día como algunos países de la Unión Europea que han ido perfeccionando su regulación. Gran parte de los datos que proporcionamos a los usuarios dentro del territorio se van fuera del país. Ejemplo de esto es Google y Facebook: en la letra chica de las condiciones que se aceptan cuando se navega en sus aplicaciones, se menciona que, ante cualquier litigio referido a la privacidad de los datos, los tribunales competentes son los de California.

Pero no todo el riesgo recae en las prácticas comerciales de las organizaciones que utilizan la información de las personas para realizar campañas con base en perfilados de clientes y modelos de predicción que se aprovechen de las necesidades de las personas para sacar una pequeña tajada de sus bolsillos. Existen cada vez más riesgos de seguridad informática. Los *hackers* o *cibercriminales* cuentan con cada vez más recursos para obtener información sensible de las personas que podrían implicar vulnerar tanto sus bienes a través de prácticas fraudulentas como su intimidad.

Lo que sí podemos afirmar es que el *big data* está modificando la manera en que conocemos el mundo y requiere simultáneamente políticas y prácticas acordes que resguarden y protejan nuestros derechos.

Referencias

Data Management Association Internacional. (2007). *Entidad de Conocimiento para la Gestión de Datos (DAMA-DMBOK©). Marco de Trabajo Funcional.* Recuperado de https://dama.org/sites/default/files/download/DI_DAMA_DMBOK_es_v2_0.pdf

Infodiagram. (2019). 4 Vs of Big Data – 4 Pieces Central Puzzle [Imagen]. Recuperado de https://blog.infodiagram.com/2019/01/big-data-presentation-appealing-diagrams-ppt.html/bigdata15

International Business Machines. (s. f.). Analítica de Big Data. Recuperado de https://www.ibm.com/analytics/es/es/hadoop/big-data-analytics/

Neoland. (2019). [Imagen sin título sobre *data science*]. Recuperado de https://www.neoland.es/blog/el_futuro_del_data_science

NIC Argentina. (2018). ¿Qué es Big Data? Recuperado de https://nic.ar/es/enterate/novedades/que-es-big-data

Panel.es. (2018). Ciclo de la vida [Imagen]. Recuperado de https://www.panel.es/wp-content/uploads/2018/04/ciclo_de_vida.png

Plasencia Prado, C. E. (s. f.). ¿Qué es y por qué aprender SQL? Recuperado de https://devcode.la/blog/que-es-sql/

PowerData. (s. f.). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. Recuperado de https://www.powerdata.es/big-data

Tableau. (s. f.). Guía de visualización de datos para principiantes: definición, ejemplos y recursos de aprendizaje. Recuperado de https://www.tableau.com/es-mx/learn/articles/data-visualization

Tableau. (2018). [Imagen sin título sobre visualización de datos en gráficos radiales]. Recuperado de https://www.analiticaweb.es/tableau-visualizacion-datos-graficos/

Universidad Internacional de Valencia. (2019). Python para big data: motivos para elegirlo. Recuperado de https://www.universidadviu.com/python-para-big-data-motivos-para-elegirlo/