

# Interpretación de Big Data

El uso de big data puede conducir a ganancias significativas para su organización como por ejemplo la reducción de costes operativos, mejora de la toma de decisiones y desarrollo de nuevos productos innovadores. Aunque tiene que saber cómo identificar, recopilar, analizar y usar de forma efectiva sus datos para cosechar los beneficios. En este curso, va a aprender el proceso de análisis de datos. Este curso también introduce los métodos analíticos básicos y avanzados más habituales, incluyendo minería de datos. Este curso también cubre algunas de las herramientas de big data más habituales y sus usos asociados, y algunos retos que hay que tener en cuenta al realizar actividades de análisis de big data en su organización.

## Tabla de contenido

1. [Interpretación de Big Data](#)
2. [Big Data y el proceso de análisis de datos](#)
3. [Big Data y Business Intelligence](#)
4. [Análisis básicos de Big Data](#)
5. [Análisis avanzados para Big Data](#)
6. [Herramientas de almacenamiento, gestión, limpieza y minería de datos](#)
7. [Herramientas de análisis, visualización e integración de datos](#)
8. [Retos del análisis de Big Data](#)

## Interpretación de Big Data

*[Título del curso: Interpretación de Big Data]* En el mundo de big data, tiene que conocer cómo identificar, recopilar y analizar datos para usarlos de forma efectiva. En este curso, va a aprender el proceso de análisis de datos y métodos de análisis básicos y avanzados comunes, incluyendo minería de datos. También descubrirá algunas de las herramientas de big data más habituales y sus usos asociados, y algunos retos a tener en cuenta al emprender actividades de análisis de big data en su organización.

## Big Data y el proceso de análisis de datos

*[Título del tema: Big Data y el proceso de análisis de datos]* ¿Qué cree que es "big data"? Evoca caudales de números, cifras y caracteres que representan infinitas cantidades de datos.

Aunque big data no es solo recopilación de datos. Es lo que se hace con los datos una vez recopilados. Trata sobre la aplicación de business intelligence para organizarlos, revisarlos y analizarlos para que tengan sentido. Esto ayuda a transformar datos en valiosas perspectivas que puede usar para impulsar reducciones de costes, decisiones empresariales inteligentes y desarrollo de productos innovadores.

Sabemos que el filtrado de grandes cantidades de datos es bastante difícil. Ese es el motivo por el que necesita un proceso efectivo, no solo para simplificar sus decisiones, sino para mejorar también sus habilidades de análisis de datos.

Comience realizando preguntas de definición. Haga preguntas relevantes, medibles y concisas. Cuando planifique sus preguntas, también resulta útil pensar sobre posibles soluciones. Esta planificación previa le ayudará a expresar mejor las preguntas para poder distinguir fácilmente soluciones posibles de soluciones inviables. Por ejemplo, ¿emplear a menos personas afectará a la calidad del trabajo realizado?

El siguiente paso consiste en establecer prioridades de medición...decidiendo qué medir y cómo medirlo.

Por ejemplo, si reduce el personal, tendrá que conocer cuántos miembros del personal emplea, cuál es el coste de ello, y considerar los efectos de una reducción. Esto se encarga del qué.

¿Y del cómo? Tendrá que saber: ¿Qué le costará esta reducción de personal a largo plazo? ¿Cuál es la unidad en la que medirá? Quizás es la moneda que está utilizando para medir el ahorro de costes. ¿Existen otros factores como por ejemplo beneficios para el personal?

Cuando establezca sus prioridades de medición, es el momento de recopilar datos. Empiece con datos de bases de datos existentes antes de usar otras fuentes. ¿Y si está en un equipo? Use un sistema de almacenamiento de archivos colaborativo. Asegúrese de registrar todos sus datos por fecha e incluya referencias.

A continuación, es el momento de unir fuerzas...para analizar sus datos. El análisis ayuda a medir si recopiló suficiente información. Considere el uso de diferentes formas de representación de datos. Las tablas dinámicas ayudan a encontrar correlaciones o variaciones entre lotes de datos.

Al final debe interpretar sus resultados. Si los datos responden a su pregunta inicial y le ayudan a defenderse contra objeciones anticipadas, pregunte cómo. Pregúntese si su conclusión está limitada de alguna forma, y si se han considerado todas las perspectivas. Si su conclusión responde positivamente a estas preguntas y consideraciones, tiene un buen inicio. Cuando haya interpretado los resultados de su análisis, puede usarlos para decidir la acción que hay que tomar.

Estos pasos para ordenar montañas de datos pueden ayudarle a tomar mejores decisiones comparativas para su negocio.

## **Big Data y Business Intelligence**

*[Título del tema: Big Data y Business Intelligence]* Trabajar con big data puede ser a veces como navegar en un barco, por la noche, en una tormenta, y sin saber el camino a seguir. Aunque si lo hace bien, big data, y los análisis correctos, puede ofrecer grandes perspectivas en varias relaciones de datos.

Como marino, existen varias herramientas de navegación que puede usar. Se encontrará que existen muchas herramientas y métodos de análisis de datos, aunque normalmente estas se encontrarán en cuatro categorías.

La mejor forma de predecir el futuro es basarse en el comportamiento pasado. Esa es la base del análisis predictivo. Utiliza big data para identificar patrones o comportamientos históricos para predecir el futuro.

Por ejemplo, los bancos usan análisis predictivo para medir el riesgo crediticio y medir las opciones de actividades fraudulentas antes de que se produzcan. La industria minorista es otro buen ejemplo. Usemos la función: Los clientes que compraron esto también compraron de Amazon...utiliza un historial de compras pasadas del cliente para predecir lo que ellos u

otros clientes pueden querer comprar en el futuro.

Pasemos a los análisis prescriptivos. Normalmente, cuando se pone enfermo va al médico. En función de sus síntomas, prescribe medicación específica para "solucionar" su problema. Así es como funcionan los análisis prescriptivos. De la misma forma que el médico receta una medicación, el análisis predictivo sugiere posibles soluciones...o acciones basadas en un problema específico. Aunque este tipo de análisis es probablemente el mejor método de análisis, apenas es utilizado por las empresas.

Puede comparar el análisis de diagnóstico con los 5 ¿Por qué? de Six Sigma, que hace incidencia en el por qué de un problema. Cada vez que responda a un por qué, como "¿por qué ocurrió?", vuelva a preguntar por qué hasta que conozca la causa raíz. Ya que es bastante laborioso, este método no es utilizado por muchas empresas de forma consistente. Normalmente, los resultados se compilarán en un panel analítico.

Finalmente, tenemos el análisis descriptivo. A veces conocido como minería de datos, ayudan a la identificación de patrones y comportamientos que ofrecen una perspectiva adicional para describir lo que ocurrió en el pasado. Saber lo que pasó en el pasado, en la mayoría de los casos, es un buen indicador de lo que ocurrirá en el futuro. Resulta útil mostrar datos históricos para cosas como informes financieros y de ventas de la empresa. Los análisis de big data, independientemente de su categoría, mejoran la comprensión de lo que está sucediendo ahora y son un buen indicador de lo que debe ocurrir a continuación.

## **Análisis básicos de Big Data**

*[Título del tema: Análisis básicos de Big Data]* Con big data, los métodos analíticos básicos resultan adecuados cuando se dispone de grandes cantidades de datos diferentes. O si tiene datos de los que desea conocer su valor. Los métodos analíticos básicos incluyen estadísticas y visualizaciones simples.

Imagine que le piden hacer una revisión de la optimización del motor de búsqueda para un blog que no tiene ningún tema o asunto específico. Existen artículos sobre comida, ropa de tendencia, eventos y productos sin ningún orden en particular. Desean conocer los temas, artículos o palabras clave que generan mayor tráfico. ¿Cómo abordaría esta revisión?

Un método que puede usar es la fragmentación. Como su nombre indica, este método implica desglosar los datos en segmentos más pequeños, lo que facilita su comprensión y uso. Para determinar los temas que suponen el mayor tráfico en el blog, puede filtrar por palabras clave para cada tema y representar sus resultados en un gráfico. También podría usar este método para encontrar las palabras clave que sean las más o las menos utilizadas, independientemente del tema. Por lo tanto, utilizará este método para centrarse en responder preguntas específicas en ciertas áreas.

Las cosas van bien, y el blog ha levantado expectación por el lanzamiento de un nuevo producto del que tienen los derechos exclusivos de cobertura. El día que se lanza el producto, espera que el tráfico del blog aumente y decide monitorizarlo en tiempo real para comprobar el progreso. Esto es lo que llamaría monitorización básica. Crea grandes conjuntos de datos de varias fuentes a través de la web; piense en tweets, imágenes de instagram o incluso en comentarios de texto simple. Pero tenga cuidado. Ya que se están generando grandes cantidades de diferentes tipos de datos, no debe usar este método con demasiada frecuencia a menos que sea necesario.

¿Pero qué ocurre si en el día que se produce el lanzamiento apenas hubiera tráfico en el

blog? La identificación de anomalías se encarga de esto. Este método implica buscar irregularidades en sus datos para darle una pista sobre lo que podría estar mal. Especialmente cuando ciertas cosas no están como esperaba que estuviesen. Si sabe que el tráfico se incrementó a medida que se acercaba la fecha de lanzamiento, ¿qué podría haber pasado al tráfico del blog ese día? Podría fijarse en las palabras clave que se están usando y quizás en aquellas que deberían haberse utilizado para explicar esta desviación. El uso de estos métodos para comprender grandes bloques de datos puede resultar útil al intentar determinar lo que es valioso en su negocio.

## **Análisis avanzados para Big Data**

*[Título del tema: Análisis avanzados para Big Data]* Cualquier problema complejo suele requerir una solución especializada. Lo mismo se aplica a análisis de big data complejos; requiere métodos analíticos sofisticados para datos estructurados y no estructurados. Los análisis avanzados se utilizan mejor para encontrar patrones en datos complejos, realizar previsiones y para necesidades de procesamiento avanzado de eventos. Si hubiera alguna forma de mirar al futuro, ¿no querría hacerlo? Esta es la base del modelado predictivo. Se aplican algoritmos a datos estructurados y no estructurados, por separado o en conjunto, para predecir futuros resultados. Por ejemplo, una empresa de cable podría usar este modelo para prever quién es más probable que cancele su servicio. ¿Y el método de análisis de texto? Se utiliza principalmente para entender datos no estructurados. Implica examinar texto fragmentado, sacando información útil y disponiéndola lógicamente. A continuación, puede usar esta información estructurada como la entrada para todo tipo de análisis, como predicción del fraude, o tendencias del cliente. La minería de datos en análisis avanzados implica identificar patrones de grandes conjuntos de datos. Cuando usa técnicas de minería de datos, desea saber si los datos se van a usar para clasificación o predicción. La clasificación trata sobre clasificar datos en conjuntos. Por ejemplo, una empresa podría estar interesada en las características de los empleados que agotan sus días por enfermedad y aquellos que no toman ninguno. ¿Y la predicción? Eso es autoexplicativo. Aborda la predicción de un valor variable no discreto o continuo. Una empresa puede intentar predecir quién responderá al incentivo de ganar días de vacaciones extra. La minería de datos usa varios algoritmos. Uno de ellos es un árbol de clasificación. Usa variables predictivas para clasificar variables categóricas dependientes. ¿El resultado final? Un "árbol" con nodos interrelacionados que puede utilizarse para formar reglas "si-entonces". Si un cliente estuvo en una empresa de cable durante más de 10 años, entonces es más probable que permanezca fiel. La regresión logística también aborda la clasificación. Propone una fórmula que prevé la probabilidad de una instancia como una función de las variables independientes. Otro algoritmo que realmente es un algoritmo de software que consiste en nodos de entrada y salida, además de capas ocultas, es la red neuronal. Aquí cada elemento se pondera, por lo que cuando se introducen los datos, el algoritmo modifica los elementos para llegar a un criterio de parada específico. Las redes neuronales pueden compararse con el funcionamiento de los cerebros de animales, el algoritmo se basa en prueba y error. Otra técnica es la de K-vecinos más cercanos, que es un ejemplo de clúster. Cuando se

implementa, encuentra grupos de registros que son similares y calcula la diferencia entre dichos registros y puntos de datos históricos. El registro se asigna a la clase del vecino más cercano en un conjunto de datos.

Por lo tanto, estas son técnicas de análisis avanzado que tiene que conocer para enfrentarse al juego de big data.

## **Herramientas de almacenamiento, gestión, limpieza y minería de datos**

*[Título del tema: Herramientas de almacenamiento, gestión, limpieza y minería de datos]*

A medida que continúa la evolución de big data, aparecen más herramientas. Software libre o propietario, la decisión es suya. No hay conjuntos de herramientas buenos o malos, sino que nos centraremos en algunas de las herramientas de almacenamiento, gestión, limpieza y minería de datos de big data más habituales disponibles.

Si piensa en almacenar y gestionar big data, tendrá que seleccionar un proveedor de almacenamiento que pueda manejar todas sus herramientas analíticas. También deben ofrecerle una plataforma para almacenar y consultar sus datos.

Hadoop es una de ellas. Es un software libre que puede almacenar conjuntos de datos considerables que puede aumentar o reducir cuando lo necesite. Además, puede gestionar una cantidad casi ilimitada de tareas simultáneamente, debido a su potencia de procesamiento.

Si necesita ayuda para construir un hub de datos para su negocio, podría intentarlo con Cloudera. Básicamente, es Hadoop con alas. Sigue siendo software libre, aunque es más una solución empresarial para que las empresas gestionen su sistema Hadoop; realizan el duro trabajo de Hadoop por usted.

Si no quiere usar bases de datos relacionales, MongoDB es una buena forma de almacenar y gestionar datos. Se suele utilizar para aplicaciones que admiten vistas singulares que abarcan numerosos sistemas. Resulta útil para organizar datos no estructurados o continuamente alterados.

Aunque los conjuntos de datos son amplios y variados, y con frecuencia "sucios". Puede usarla para enjuagar y centrifugar antes de extraer información.

OpenRefine es un software fácil de utilizar para la limpieza de datos que ayuda a organizar sus datos no estructurados, incluso si estuvieran muy desordenados. Esta herramienta puede manejar fácilmente grandes conjuntos de datos.

Otra herramienta de limpieza útil es DataCleaner. Le ahorra todo el trabajo de convertir conjuntos de datos no estructurados en datos útiles y legibles.

Una vez que haya limpiado sus datos, puede empezar la minería, filtrando sus datos para buscar patrones que le ayuden con la previsión.

Para análisis predictivos, puede usar RapidMiner para detectar información útil dentro de su base de datos. Además, si ya tuviera su propio algoritmo, puede integrarlo a través de su interfaz de programación de aplicaciones, o API.

El IBM SPSS Modeler es el más adecuado para grandes empresas. Sus soluciones de minería de datos incluyen análisis de texto, gestión y optimización de decisiones, además de análisis de entidades. Puede usarse en casi cualquier tipo de base de datos y puede integrarse con otros productos IBM SPSS.

Si no puede determinar adecuadamente cómo analizar big data, no resulta útil tener una

abundancia de puntos de datos. Ahí es donde Teradata entra en acción. Proporcionan varias soluciones y servicios para empresas que quieren empezar estar orientadas a datos. Esta lista cada vez más creciente de herramientas le permitirá desenmarañar sus datos y comprenderlos para utilizarlos de la mejor forma para su negocio.

## **Herramientas de análisis, visualización e integración de datos**

*[Título del tema: Herramientas de análisis, visualización e integración de datos]*

Diferentes trabajos requieren diferentes herramientas. Esto también se cumple con big data. Junto con herramientas de almacenamiento, gestión, limpieza y minería, también existen herramientas para análisis, visualización e integración de datos.

¿Qué es análisis de datos? Implica diseccionar los datos y patrones que descubrió en la minería. Y la evaluación de la importancia de los patrones que no reconoció anteriormente.

El análisis de datos permite encontrar en sus datos respuestas a preguntas concretas.

Qubole es una solución para grandes empresas. Hace uso de varios motores de procesamiento de datos, incluyendo Presto y Hive. Este programa es muy accesible y flexible.

Puede importar sus datos con la interfaz fácil de usar de BigML, y obtener previsiones a partir de ellos. Incluso le permitirán usar sus modelos de análisis predictivo.

Si le gusta más lo visual, entonces sabe lo aburrido que puede ser fijarse en números. Aquí pueden ayudarle herramientas de visualización de datos. Transforman todos sus datos en formas visuales más inteligentes y fáciles de utilizar, lo que hace innecesaria la codificación.

Con Tableau, crea dichas visualizaciones de datos, como mapas y gráficos. Incluso puede usar su conector web para proyectar sus datos en vivo de forma más visual.

Si prefiere trabajar con mapas, entonces tiene que utilizar CartoDB. Permite jugar con conjuntos de datos de muestra para que pueda aprender cómo situar sus datos de localización en una forma visual más adecuada.

Además está Chartio. En solo algunos pasos, puede incorporar fuentes de datos y ejecutar consultas online directamente desde un navegador. Con su plataforma, puede crear paneles y formatearlos en archivos PDF para enviarlos por correo.

Las plataformas de integración de datos permiten tomar datos extraídos de una plataforma y compartirlos con otra.

La plataforma Blockspring permite tomar una simple fórmula de hoja de cálculo de Google y aplicarla a programas de terceros. Por ejemplo, puede comprobar los seguidores, Tweet o incluso conectarse a otras herramientas, entre otras cosas, desde esa hoja.

Pentaho, por contra, permite integrar herramientas usando una interfaz de usuario básica de arrastrar y soltar. Además, aunque no va a necesitar ninguna codificación, permiten la integración de big data y también ofrecen servicios de análisis.

Aunque para poder hacer todo eso, primero necesita datos. Aquí es donde la recopilación o extracción de datos entra en juego. Significa recopilar información no estructurada, como un blog o álbum de fotos, y transformarla en algo estructurado, como una hoja de cálculo. Después, puede analizar y usar los datos para tomar decisiones importantes.

En la extracción de datos, Import.io es la mejor herramienta. Toma páginas web y después de un sencillo clic en su IU, transforma dichas páginas en hojas de cálculo fáciles de

utilizar.

Con un gran número de herramientas a su disposición, ya sea más visual o no, ahora la comprensión de los datos creados resulta mucho más sencilla.

## **Retos del análisis de Big Data**

*[Título del tema: Retos del análisis de Big Data]* Independientemente de lo buena que sea su solución, siempre existirán dificultades que influirán sobre ella. El análisis de big data no es una excepción.

Aprovecharse de las ventajas de big data puede aportar grandes beneficios para la organización en relación con la reducción de costes y tiempo, mejorando la toma de decisiones y el desarrollo de productos y servicios. Aunque si no se aplica el método y los recursos correctos, o no se utilizan métodos efectivos de análisis e interpretación, las buenas intenciones pueden convertirse en un riesgo para la organización.

Ya sabemos que puede ser difícil comprender big data, ¿cómo puede saber qué usar y cómo usarlo? Los casos de negocios orientados a datos suelen requerir una forma diferente de pensamiento, más original. Por lo tanto, la identificación y uso de los datos correctos puede resultar difícil, especialmente si se utilizan métodos de pensamiento más tradicionales.

Ya que la mayoría de las empresas no saben lo que supone descifrar big data, no tienen sistemas adecuados para gestionarlo. El acceso y conectividad inadecuados es un gran problema en el almacenamiento y gestión de sus datos.

Además de no tener la infraestructura para gestionar los datos, las empresas no disponen de las personas correctas para intentarlo. Tampoco ayuda que aunque big data esté evolucionando con bastante rapidez, no sea un tema específico que pueda encontrarse a nivel de universidad. Este es el motivo por el que resulta difícil encontrar las personas correctas que pueden comprender todos los datos. Si no conoce demasiado sobre big data, ¿cómo sabe qué buscar en un experto de big data? No tener personas correctas para sus datos puede suponer un enorme retroceso.

Todo se complica con sistemas de colaboración inadecuados. La propiedad de los datos suele segmentarse por departamento o función. Aunque para analizar big data, tiene que poder acceder a todo independientemente del departamento o función. Si no puede, sus datos podrían distorsionarse.

Muchas empresas siguen haciendo uso de bases de datos estándar mientras tantean big data. Aunque tantear no es suficiente. La tecnología está evolucionando muy rápido, y su empresa puede quedarse atrás con bastante rapidez si no sigue el ritmo de la tecnología actual. Ese es el motivo por el que debe aplicar un sistema que pueda evolucionar y adaptarse a medida que la tecnología sea más eficiente.

La siguiente pregunta sería ¿cuán seguros están los datos de su empresa? Si no dispone de las personas, sistemas o herramientas correctos para trabajar con big data, ¿cómo podrá mantenerlos seguros? A medida que crece una empresa, adquiere más información. Eso significa que existe más información que puede estar comprometida.

Por lo tanto, en lugar de precipitarse a ciegas en el mundo del análisis de big data, asegúrese de conocer estos retos para evitar dañar a su empresa en el proceso.