Principios fundamentales del *big data*

Big data e inteligencia artificial

CONTENIDO

Objetivos

- Entender el concepto de *big data*.
- Aprender cómo trabajar con *big data*.

¿Qué es el *big data*?

El *big data* es el análisis masivo de datos. Hace referencia a una cantidad de datos sumamente grande, que no se puede procesar con aplicaciones tradicionales.

También, este término, se refiere a las nuevas tecnologías que hacen posible el almacenamiento y el procesamiento de estas cantidades de datos.

¿De dónde proceden esos datos?

En los últimos años la información disponible ha crecido de manera exponencial *(figura 1)*. El límite superior de procesamiento ha ido creciendo a lo largo de los años. Se estima que el mundo almacenó unos 5 *zettabytes* en 2014. Si se pone esta información en libros, convirtiendo las imágenes y todo eso a su equivalente en letras, se podría hacer 4500 pilas de libros que lleguen hasta el sol.

Figura 1. ¿De dónde vienen los datos? [4].

Principalmente, los datos se están generando de las siguientes formas:

- **Datos generados por personas:** mandar un *email*, contestar a un *Whatsapp* o escribir en redes sociales son solo algunos de los ejemplos.
- **Datos generados entra máquinas:** los contadores de la luz de las viviendas, los teléfonos móviles o los GPS de los vehículos. Estos se comunican con otros dispositivos a través de otros aparatos a los que transmiten los datos que van recogiendo.
- **Biométricas:** son los datos que tienen como origen sensores de huellas dactilares, escáneres de retina, lectores de ADN, sensores de reconocimiento facial o reconocimiento de voz.
- ***Marketing* en la *web*:** los movimientos en la *web* están sujetos a mediciones, como el rastreo del movimiento del cursor por parte de los usuarios en una *web*.
- **Transacciones de datos:** algunos ejemplos, están en las transacciones electrónicas entre bancos, la reserva de un billete de avión o la compra de productos a través de un una *web*.

Tipo de datos

Los datos, se pueden clasificar según su estructura:

- **Datos estructurados:** son los que se pueden almacenar en tablas, tienen una determinada longitud y formato. Por ejemplo, los datos numéricos, las fechas o las cadenas de caracteres.
- **Datos no estructurados:** se trata de datos en su forma original, según fueron escogidos. No poseen un formato especifico. Por ejemplo, los videos, *emails*, textos, documentos, etc.

Características principales

Las características principales del *big data*, se agrupan en lo que se llama "las 7 'v'".

- **Velocidad:** velocidad en el análisis de los datos. Ejecución de algoritmos cada vez más complejos en menos tiempo.
- **Variedad:** los datos provienen de numerosas fuentes y con diferentes formatos.
- **Valor:** una gran cuantía de datos frecuentemente extrae pequeñas informaciones de valor.
- **Variabilidad:** los macrodatos, se generan en un entorno cambiante por lo que la información varía mucho.
- **Volumen:** la cantidad de datos que se está generando es muy grande, ya que la sociedad está interconectada.
- **Veracidad:** saber la fiabilidad de la información recogida es importante para obtener unos datos de calidad.

- **Visualización:** convertir cientos de hojas de información en un único gráfico que muestre claramente unas conclusiones predictivas.

Gestión del dato

- **Captura de información:** dónde está la información que se necesita y cómo capturarla, son dos de los puntos que se deben plantear.

  Para capturarla, existen métodos como el *web scraping* (técnica para extraer información de sitios *web*), mediante la utilización de API o de otros servicios de SW diseñados para capturar grandes cantidades de información.

- **Almacenamiento:** después de la captura del dato se necesitará guardarlo. Se podrá optar por hojas de cálculo para información estructurada y por sistemas NoSQL, que permiten el almacenamiento de información no estructurada, de forma más flexible y rápida.
- **Tratamiento:** el tratamiento dependerá del tipo de información y del uso que se quiera hacer de ella. Por lo que, se podrá tener desde tratamientos sencillos, hasta sistemas predictivos muchos más complejos. Se puede extraer conocimiento y buscar patrones de comportamiento.
- **Puesta en valor:** los datos por sí mismos no aportan conocimiento, sin un análisis y un tratamiento adecuado, no sirven para nada. El valor no estará en los propios datos, sino en la relación de estos entre sí. El valor, puede ser una visualización en un gráfico, donde se haga un análisis predictivo, una recomendación de un artículo relacionado en un portal de comercio electrónico, un cliente que adquiera un producto concreto, etc.

Beneficios del *big data*

- Es un conjunto de tecnologías muy ágiles y flexibles.
- Es escalable por lo que, si se tienen más datos, se pueden poner más máquinas e incluir nuevas analíticas y más información. Así, la infraestructura puede crecer al mismo tiempo que la empresa o proyecto.
- Es rápido y económico respecto a la infraestructura de almacenamiento. Por ejemplo, a través de servicios de computación en la nube, se puede pagar únicamente por lo que se consume.
- La mayoría de las aplicaciones pertenecen a la comunidad *open source*, *software* de código abierto o que forma parte del dominio público. Lo que supone un ahorro económico evidente. También se tiene a disposición una gran comunidad de desarrolladores que trabajan diariamente en mejorar sistemas y procesos de estos recursos.
- Para el final, la gran ventaja: a través del *big data* y habiendo seguido correctamente el ciclo de gestión de información, que culmina con la puesta en valor de los datos, se habrá conseguido un objetivo previamente marcado.
  - Ganar más dinero o ahorrarlo a través de mejoras de costes.
  - Adquisición de nuevos clientes o retención de los actuales
  - Avanzar en una investigación científica.
  - Obtener una ventaja competitiva respecto a los competidores anticipándose a lo que deparará el mercado a través de análisis predictivos.
  - Optimizar los procesos de producción.

Principios básicos

Para descubrir cómo construir adecuadamente los sistemas de datos, se debe volver a los principios, en el nivel más fundamental, ¿qué hace un sistema de datos?

Se puede comenzar con una definición intuitiva: un sistema de datos responde preguntas basadas en información que se adquirió en el pasado hasta el presente. Así que un perfil de red social responde preguntas como ¿cuál es el nombre de esta persona? y ¿cuántos amigos tiene esta persona? Una página *web* de cuenta bancaria responde preguntas como ¿cuál es mi saldo actual? y ¿qué transacciones han ocurrido recientemente en mi cuenta?

Muchos de estos sistemas de big data fueron iniciados por Google o Amazon, que creó una innovadora distribución de almacén de clave / valor llamado *Dynamo*. La comunidad de código abierto respondió en los años siguientes con *Hadoop, HBase, MongoDB, Cassandra, RabbitMQ* y otros proyectos.

Las propiedades por las que deberían esforzarse los sistemas de big data tienen que ver tanto con la complejidad como con la escalabilidad. Un sistema *big data* no solo debe funcionar bien y ser eficiente en recursos, sino cumplir con las siguientes propiedades:

- Robustez y tolerancia a fallos
- Lecturas y actualizaciones de baja latencia
- Escalabilidad
- Generalización
- Extensibilidad
- Mínimo mantenimiento
- Consultas *had oc*

Herramientas para trabajar con *big data*

A continuación, se muestra una selección de herramientas open source, para trabajar con *big data*.

Hadoop

Hadoop es la herramienta estándar para almacenar grandes volúmenes de datos de forma distribuida y redundante en un clúster de máquinas.

Hadoop originalmente fue diseñada como un *framework* completo, incluyendo también capacidades de procesado a través del paradigma *MapReduce*, pero hoy en día se emplea principalmente como almacenamiento: las tareas de procesado se relegan a herramientas más flexibles a través de YARN.

YARN es un sistema de gestión de recursos del clúster. Desde este punto de vista, se puede entender *Hadoop* como el "sistema operativo" de un clúster *big data*, permitiendo así desplegar sobre él aplicaciones de procesado más avanzadas.

Otra de las ventajas de *Hadoop* es que, al haberse convertido en un estándar, está disponible en prácticamente todos los proveedores de servicios *cloud*, que facilitan su configuración y despliegue. Igualmente, a través de estas plataformas *cloud* resulta sencillo redimensionar el clúster según las necesidades de cada momento, resultando así muy eficiente en costes *hardware*.

Elasticsearch

Elasticsearch continúa destacando como una de las bases de datos NoSQL más populares. Inicialmente, concebida como una herramienta para realizar búsquedas complejas en grandes volúmenes de documentos en texto, hoy día *Elasticsearch* también se emplea frecuentemente para trabajar con datos en otros formatos, como información estructurada o geolocalizaciones.

Su principal ventaja es la capacidad de responder a consultas en tiempo real, incluso si requieren de filtros complejos o búsquedas aproximadas en texto. Para soportar esta funcionalidad, *Elasticsearch* se despliega sobre un clúster de máquinas que reparten los datos con redundancia y comparten el trabajo de las consultas. Con la infraestructura apropiada, *Elasticsearch* es capaz de responder consultas en milisegundos, habilitando así aplicaciones de análisis en tiempo real.

Apache Spark

Apache Spark se puede considerar la evolución del paradigma *MapReduce* nativo de *Hadoop*, permitiendo ejecutar tareas de procesamiento de datos hasta 100 veces más rápido que su predecesor, gracias a un uso efectivo de la memoria RAM de las máquinas del clúster.

Las tareas de procesado en *Spark* pueden programarse empleando los lenguajes más habituales en el sector, como son *Scala, Java, Python* o R.

Apache Kafka

Si *Hadoop* es el cimiento de las soluciones *big data*, se podría decir que *Apache Kafka* es el pegamento. Kafka es un sistema de colas de mensajes redundante y distribuido, que permite implementar de manera fiable las comunicaciones entre los diferentes módulos o componentes de una solución *big data*.

El uso habitual de Kafka es conectar los diferentes elementos de la solución: un módulo que captura datos de una red social y debe transmitirlos a una base de datos para su almacenamiento o un modelo de aprendizaje automático que debe hacer predicciones y enviarlas a través de un servicio de mensajería. Kafka mantiene un historial de todos los mensajes, de modo que, si cualquier parte de la solución sufre un fallo, el sistema puede reiniciarse sin que se haya perdido ningún mensaje.

Python

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional.

R

R es un lenguaje y un entorno para gráficos y computación estadística.

R proporciona una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento, etc.), técnicas gráficas y es altamente extensible.

MongoDB

Es una base de datos NoSQL orientada a documentos. Esto quiere decir que, en lugar de guardar los datos en registros, guarda los datos en documentos. Estos documentos son almacenados en BSON, que es una representación binaria de JSON.

Una de las diferencias más importantes con respecto a las bases de datos relacionales, es que no es necesario seguir un esquema. Los documentos de una misma colección (concepto similar a una tabla de una base de datos relacional), pueden tener esquemas diferentes.

¿Qué conocimientos se necesitan para ser *big data analyst*?

A continuación, se puede observar una lista de *skills* y conocimientos necesarios, y muy recomendables, para optar a un puesto de *big data analyst*:

- **Conocimientos y habilidades de programación:** hay que estar cómodo programando y, como mínimo, es bueno saber R, *Python* y *Java*. Posiblemente, se acabe ampliando el espectro y añadiendo herramientas a la caja de herramientas de programación, por ejemplo, con C++, Ruby, SQL, Hive, SAS, SPSS, MATLAB, etc.
- **Conocimientos de matemáticas y estadística:** principalmente, conocimientos en álgebra lineal y resumen de estadísticas, distribución de probabilidad, variables aleatorias, marco de prueba de hipótesis.
- **Habilidades con *machine learning*:** esto ayuda a gestionar estructuras de datos complejas y patrones de aprendizaje que son demasiado difíciles de manejar utilizando el análisis de datos tradicional.
- **Habilidades de manejo de datos:** para permitir un consumo más conveniente de estos.
- Habilidades de comunicación y de visualización de datos.

Además, es muy conveniente tener **conocimientos y experiencia en almacenamiento de datos**, tanto si se habla de sistemas de bases de datos relacionales como no relacionales; y también ciertos conocimientos, o familiaridad, con frameworks como *Apache Spark, Apache Storm, Apache Samza* o *Hadoop*.

Sin duda, lo que necesita un *big data analyst* es tener buen conocimiento del dominio (de negocio) en el que se está trabajando. Es frecuente disponer de buenos analistas en negocios y estadística, pero que no son expertos programando; y de igual manera, existen muy buenos programadores que no tienen conocimientos más allá.

Por eso, un valor muy grande del buen analista de *big data* es que conozca no solo los aspectos técnicos, y que domine la estadística y las matemáticas, sino que también tenga visión de negocio, que sepa qué objetivos se persiguen y cómo puede utilizar los datos para contribuir a alcanzarlos.

Bibliografía

[1] N., Wilkins, *Artificial intelligence: the ultimate guide to ai, the internet of things, machine learning, deep learning + a comprehensive guide to robotics*.

[2] A., Campbell, *Python data analysis: comprehensive guide to data science, analytics and metrics with python: 4*.

[3] R., Hurley, *Data science: a comprehensive guide to data science, data analytics, data mining, artificial intelligence, machine learning, and big data*.

[4] Facultad de Estudios Estadísticos de la Universidad Complutense de Madrid, *¿Qué es big data?* 2020 [En línea] Disponible en: https://www.masterbigdataucm.com/que-es-big-data/