***BIG DATA · DATA ENGINEERING · DATA SCIENCE · DATA ANALYTICS***

**Caso Práctico: Ordenamiento de Hoteles**

Los conceptos de ***data engineering***, ***data science*** y ***data analytics*** están íntimamente relacionados con el concepto de ***big data***.

Veamos un caso práctico de ordenamiento de hoteles, donde trataremos de dar una mirada más concreta y práctica, y no tan teórica. También aparecerán herramientas que están enunciadas al final de la PPT y que serán combinadas con el caso práctico planteado.

*Imaginemos a un usuario que está realizando una búsqueda de hoteles para irse de vacaciones.*

En un mundo donde no era tan cambiante y no existían redes sociales, ni blogs, ni el desarrollo de nuevas tecnologías tan constantes, lo que terminábamos “grabando” como información era la confirmación de la reserva de un hotel con todos los datos del usuario.

Con el concepto de ***big data***, lo que tratamos de hacer es guardar primero todo tipo de información del acceso y navegación que hace un usuario en Internet que va más allá del sitio donde entre.

*Este usuario tendrá la capacidad de seleccionar precios, cantidad de estrellas del hotel, cuáles son los hoteles más valorados no sólo por él sino por varios usuarios, disponibilidad, ubicación, etc.*

Como mencionamos anteriormente, lo que buscamos con el concepto de ***big data*** es reunir información previa importante para quien está buscando el hotel que no sólo dependa de información breve.

Por ejemplo, podemos grabar: por todos los hoteles previos que el usuario navegó, por todos los hoteles donde hizo click en el detalle del hotel donde el usuario pasó, por todos los lugares donde hizo click en una foto, información que se tenga de blogs donde el usuario accede que tenga que ver con turismo, información de redes sociales y distintas fuentes de datos a la que la aplicación o sitio web pueda acceder (por convenios que compartan información)… Acá aparecen los conceptos de ***data lakes*** y ***data engineering***.

***Data engineering*** trata de resolver los cuellos de botella que puedan existir entre toda la información que deseamos almacenar de distintas fuentes de datos. Una de las diferencias entre un ***data warehouse*** y el concepto de ***data lakes*** es que éste último logra almacenar información que no sea solamente estructurada. Por eso podemos usar, por ejemplo, *Kafka*, que permite facilitar el acceso relativamente fácil en tiempo real de información (recordemos: clicks, navegación en detalles de hoteles, redes sociales, blogs, etc.), inclusive también de eventos vía streaming. La plataforma *Kafka* podría estar soportada, por ejemplo, en una arquitectura *cloud*.

Entonces, lo que detallamos anteriormente se puede unir con el concepto de ***data engineering***, con la utilización de *Kafka*, con el concepto de ***data lakes*** y también con la utilización de *HFS* de *Hadoop*, que es un framework utilizado para la implementación de ***big data***.

*Sigamos con el ejemplo del hotel…*

*Supongamos que ya tenemos la información variada (estructurada y no estructurada) que necesitamos para armar un “ordenamiento previo” para lo que el usuario desea elegir.*

Pasamos entonces a lo que sería una etapa de ***data science***: tenemos personas dedicadas a decir “bueno, ahora tengo toda la información disponible, pero no toda esa información me sirve, y además está desordenada y viene de distintos lugares…”.

Entonces, por ejemplo, podemos aplicar el concepto de ***machine learning*** y, mediante una herramienta como *Spark* (un motor que nos permite realizar modelos matemáticos y de análisis), obtener información que necesitamos para el ordenamiento de los hoteles. También podemos utilizar un lenguaje como *Python* para el desarrollo de toda esta funcionalidad.

Este proceso me permitiría ya tener un listado de hoteles que da con el perfil del usuario y todo su *background* de gusto o preferencias…

Todo parecería ya cerrar y terminar con todo el proceso, pero no es tan así… Si nosotros obtenemos un algoritmo que ya nos permita tener los 100 mejores hoteles donde el usuario podría decidir, pero ese algoritmo tarda 1 minuto en desplegarse, no creemos que una persona se quede esperando, por más que la información que se está mostrando sea la correcta o la óptima.

Acá entonces, y finalmente, deberíamos pensar qué herramientas de ***data analytics*** podríamos utilizar para que los algoritmos o métodos matemáticos propuestos por los ***data science*** puedan mostrar la información y puedan refinarla de forma que el producto final sea de agrado del usuario, que es quien tiene que terminar decidiendo la compra.