softserve

Los 5 mejores libros para principiantes e intermedios que quieran dominar el Big Data

Ciudad de México, 04 de junio de 2024.- El mundo de la **creación de datos** es alucinante. En 2018, el mundo almacenaba, gestionaba y consumía 33 zettabytes de ellos (1 ZB equivale a **10^{21} bytes**); y ahora se prevé alcanzar un volumen anual de **175 zettabytes para el 2025**, lo que supone una generación de aproximadamente **491 petabytes diarios** a nivel internacional (1 PB equivale a **10^{15} bytes)**, de acuerdo con un análisis de Deloitte. En el 2024, se estima que **cada día se crean entre 380 y 390 petabytes** de datos.

Para entender mejor estos números, debemos partir de un ejemplo que comparte **Daniel Restrepo Hincapié, ingeniero sénior de Big Data en SoftServe**, compañía global de tecnologías de la información (TI) líder en desarrollo de *software* y consultoría: "*Si ver una hora de video 4K consume hasta 14 gigabytes de datos, **1 zettabyte sería almacenar casi 1,100 millones de videos**. Entonces, si alguien empieza a ver videos 4K de una hora sin parar, le llevará casi toda la vida útil del universo terminarlos*". El espacio digital se expande a un ritmo sin precedentes, desde usos comunes hasta la **generación masiva de información** que traerá una utilización extendida de los **modelos de lenguaje grande** (LLM, por sus siglas en inglés) o de la **inteligencia artificial** (IA).

Sin embargo, a pesar del volumen de datos creados, **sólo se analiza una quinta parte** de ellos según un estudio de Seagate; lo que significa que **hay un enorme valor que se queda sin tocar**. Tal brecha empieza a reducirse con los servicios que constantemente se desarrollan, actualizan y lanzan en la industria; abriendo **enormes oportunidades en el mercado laboral "*techie*"**. El mercado de Big Data ha **crecido 5.3 veces** en los últimos siete años, alcanzando un valor de **829 mil millones de euros para 2025** según la Comisión Europea. "*A medida que aumente la demanda, habrá una gran necesidad de **especialistas cualificados en Big Data**. Si estás pensando en hacer carrera dentro de este campo tan dinámico, **¡la oportunidad es hoy!***"; subraya el experto.

Para quienes se sientan atraídos por esta **oportunidad laboral**, Daniel Restrepo comparte una lista de **cinco libros que guiarán a principiantes e intermedios** por los principios fundamentales y las técnicas más vanguardistas de Big Data, con el fin de avanzar en sus conocimientos y **prosperar profesionalmente**:

1. Lectura obligada - *Fundamentos de ingeniería de datos: planificar y construir sistemas de datos robustos*

"*Fundamentals of Data Engineering: Plan and Build Robust Data Systems*" es más que un libro: es todo **un viaje al corazón de la ingeniería de datos**, dirigido por los expertos Joe Reis y Matt Housley. En él, se parte de que el camino hacia la comprensión de los datos siempre comienza con lo básico e implica lidiar con datos sin procesar ni refinar, lo que puede resultar abrumador para los principiantes. Sin embargo, los conocimientos y experiencia de los autores guían a los lectores por la vasta información; enseñándoles cómo **planificar, diseñar y construir sistemas de datos** que almacenarán ideas y conocimientos valiosos.

A medida que se explora más a fondo el texto, se muestran los **principios fundamentales de una arquitectura de datos eficaz** y una visión completa del panorama de la ingeniería de datos, llegando hasta conceptos clave de nivel medio-alto. No sólo cubre aspectos centrales, sino que también profundiza en las **tendencias emergentes** que se vislumbran en el horizonte, destacando nuevas perspectivas y desafíos bajo la mirada de los fundamentos de **Azure Data y los servicios en la nube**.

2. Progresando paso a paso - *Ingeniería de datos con Python*

"*Data Engineering with Python*" es una guía bien elaborada para navegar por las complejidades del **diseño e implementación de *pipelines*** (conjuntos de procesos automatizados que permiten la transferencia de datos de una fuente a un destino específico) y **conectividad** de las bases de datos; enseñando al lector el arte y la ciencia de **diseñar, orquestar y gestionar** sofisticados *pipelines* (canalizaciones, tuberías o cauces) de datos.

Mientras uno avanza la lectura, se introduce en las **técnicas ETL** (*Extract, Transform, Load*), fundamentales para la ingeniería de datos; aportando una habilidad crítica para **convertir datos sin procesar** en perspectivas

significativas. **Python**, con su gran cantidad de bibliotecas y herramientas, emerge como el héroe del libro que también **ofrece implementos para agilizar el flujo**. Por otro lado, va más allá de la conexión de bases de datos, al explorar **la personalización de las canalizaciones** y poniendo de relieve la flexibilidad de Python, valiéndose de **ejemplos claros.** Los capítulos finales entregan una buena colección de conocimientos técnicos y dejan una **fuerte sensación de empoderamiento**.

3. ¿Tienes dudas? Abre tu perspectiva con *The Datapreneurs*

"*The Datapreneurs, the promise of AI and the Creators Building Our Future*" (Los Datapreneurs, la promesa de la inteligencia artificial y los creadores que construyen nuestro futuro) proyecta el futuro más allá de los algoritmos y las máquinas, **dándole su lugar a las personas** que les dan vida. En este libro, Bob Muglia desentraña cómo la simbiosis entre **el ingenio humano y los datos digitales** es la piedra angular que conduce a **la nueva era de la tecnología** que es la inteligencia artificial. Sus capítulos ofrecen un viaje por **la evolución de la IA**, obteniendo el lector una comprensión exhaustiva de su nacimiento y la hoja de ruta hacia el futuro.

Este material no hay que leerlo como una perspectiva única, sino más bien como **una colección de voces** que incluyen a expertos del sector y líderes de opinión. Mediante conversaciones, presenta una visión profunda de los **beneficios y riesgos potenciales asociados a la inteligencia artificial**; que son un reflejo del **poder transformador** de las tecnologías basadas en datos, poniendo sobre la mesa **cuestiones éticas y sociales**.

4. Aquí está lo bueno - "*Aprender Spark*" (2ª edición)

"*Learning Spark*" se adentra en el corazón de **la manipulación de datos** y el descubrimiento de conocimientos, a través de **conceptos esenciales y aplicaciones prácticas** que encarna Apache Spark (un *framework* de computación en clúster *open-source* desarrollado por la Universidad de California y lanzado en 2014). El viaje comienza con una exploración de sus cimientos, donde los autores revelan la red interconectada del marco Spark; exponiendo la esencia de los **RDD (*Resilient Distributed Datasets*)** y de la resiliencia del *framework* con su **arquitectura distribuida**.

El texto despliega también **los estándares** API DataFrame, Dataset y Spark SQL, la transmisión estructurada y **el "arte alquímico" del aprendizaje automático** con MLlib. Por otro lado, los autores comparten la tradición de **desplegar aplicaciones Spark**, una especie de "rito de paso" para todo profesional del *framework*; aparte del conocimiento para garantizar que cada chispa de datos encienda las llamas de la información con la **máxima eficiencia**. Al respecto, Daniel sugiere añadir a la biblioteca el libro **"*Spark Cookbook*" de O'Reilly**, para descubrir **atajos potentes, técnicas y prácticas** que le harán la vida más fácil a los interesados.

5. Explorando la caja de Pandora - *Diseño de aplicaciones intensivas en datos*

"*Designing Data-Intensive Applications*" está hecho para navegar por los **complejos sistemas** de grandes volúmenes de datos, mediante **ejemplos del mundo real y estudios de casos**. Sirve como una guía para identificar y analizar los componentes básicos en la construcción de **sistemas de datos a gran escala**, concebidos para dar soporte a los mercados mundiales; y profundiza en tres principios clave que cimentan tales sistemas: **fiabilidad, escalabilidad y mantenibilidad**.

El libro teje **puentes entre la teoría y la práctica**. Mediante **aplicaciones reales**, los lectores son testigos de la acción de los **principios del diseño de datos**. Además, proporciona una **comprensión profunda** de los sistemas a un nivel técnico más alto. Sus casos prácticos no sólo muestran el camino para crear **sistemas fiables, escalables y mantenibles** de datos; sino que también ofrecen inspiración y conocimientos para lograr obras maestras de ingeniería. Es como embarcarse en un gran viaje por la **comprensión del Big Data**.

¿Por qué consultar estos libros?

Daniel Restrepo comenta que estos cinco textos se volvieron **guías invaluables para su desarrollo como experto en Big Data**, bajo la premisa de que "**la práctica hace al maestro**". "*Si te interesa dedicarte al campo del Big Data, estos libros que me ayudaron mucho en mi carrera inicial pueden servirte como* ***punto de partida****. No estoy diciendo que se convertirán en la fórmula mágica para cualquiera que quiera seguir la misma ruta profesional, pero podrían darle* ***un buen comienzo*** *en el mundo de los datos. Recuerda que tú eres el dueño de tu propio ritmo y dirección*"; concluye el **ingeniero sénior de Big Data en SoftServe**.

Acerca de SoftServe

SoftServe es una autoridad digital que asesora y proporciona servicios tecnológicos de vanguardia. Como la mayor empresa global de TI con raíces ucranianas, ofrece soluciones de desarrollo de software y consultoría. Con más de 11,000 empleados en 50 centros, oficinas y ubicaciones de clientes en todo el mundo, SoftServe es una de las mayores compañías de desarrollo de software de Europa Central y Oriental. Sus sedes centrales se encuentran en Lviv (Ucrania) y Austin (Texas, EE.UU.). Cuenta con centros de desarrollo en Ucrania, Polonia y Bulgaria, y en 2022 comenzó a operar en Rumanía, México, Chile y Colombia.

Para mayor información, visita www.softserveinc.com.

O síguenos en:

Facebook: @SoftServeInc

Twitter: @SoftServeInc

LinkedIn: @softserve
Blog: www.softserveinc.com/en-us/blog