**PREWORK**

Sesión 01

**Introducción al Big Data, las 5 V's y el ambiente Apache**

**🔎 Introducción**

El día de hoy aprenderás los alcance y concepto del Big Data, y las diferencias que guarda con respecto al análisis de datos estático, además de la distinción que existe a la hora de desarrollar consultas entre Big Data y lo que hemos venido haciendo para bases de datos relacionales clásicas. La sesión del día de hoy se centrará en definir conceptualmente las bases sobre las cuales estaremos trabajando a lo largo del módulo en una primera parte, y en la segunda hablaremos sobre la definición para tu proyecto de Big Data, con el cual trabajarás a lo largo de todo el curso.

**🎯 Objetivos**

* Comprender que es el Big Data
* Aprender las diferencias entre NoSQL y SQL: los administradores de bases de datos dedicados a Big Data
* Desarrollar un breve repaso de SQL
* Comprender algunas aplicaciones del Big Data

📋 **Instrucciones**

Revisa los recursos que se enlistan a continuación para que puedas seguir los ejemplos y realizar todos los retos de la sesión.

Bienvenido a tu Prework #01 . A continuación te presentamos la siguiente guía de materiales para que organices tu tiempo y puedas revisar todos los recursos al máximo:

| **Materiales** | **Tiempo sugerido (minutos)** |
| --- | --- |
| Tema 01. La era de los Macrodatos | 10 |
| Video 01. Que es Big Data: Tipos de Datos, Dimensiones, Aplicaciones, Beneficios | 25 |
| Tema 02. Repaso de SQL | 15 |
| Tema 03. Aplicaciones del Big Data | 15 |
| Actividad. Cuestionario sobre Big Data | 10 |

Recuerda que es muy importante completar tu Prework para familiarizarte con los conceptos, retos y ejercicios que abordaremos en el Work. Toma nota de lo que consideres y anota tus dudas si es necesario para que las resolvamos.

¡Nos vemos en el Work!

**📚 Tema 01.** La era de los Macrodatos

| **El Big Data, engranaje de cambio e innovación**  **“Los patrones ocultos en grandes cantidades de datos podrían salvarle la vida a un paciente enfermo de un hospital, o inclusive, cambiar al mundo”**   * **Atul Butte, Stanford University** |
| --- |

¿Por qué hablamos tanto de la era de los Macrodatos? Como podrás haberte dado cuenta, en el siglo XXI, es normal que en la mayoría de los hogares contemos con al menos:

* Un smartphone o tablet
* Una smart TV
* Una Computadora Personal

Estos 3 aparatos, por sí mismos, con base en tus patrones de comportamiento, pueden generar datos para aquellos servicios que tengas contratados, donde, por ejemplo:

* En tu teléfono, seguramente de forma continua usas Instagram, Twitter o Facebook para poder socializar.
* En tu TV, posiblemente veas YouTube, Netflix o Prime para poder entretenerte un rato.
* En tu PC, navegas por la web, haces tu tarea, trabajas a distancia y usas varios programas como Office, tu antivirus o alguno otro en especial.



Cada vez que usas alguno de esos servicios, empiezas generar datos de uso, los cuales son sumamente preciados por las empresas para poder averiguar tus patrones de consumo y como estos se empatan con el resto de usuarios (y de tal forma, el servicio podrá generarte publicidad personalizada, mejorar tu experiencia de uso, o simplemente comprender la distribución de minutos y horas con las que usas ciertas aplicaciones o servicios). Todos estos datos que generas, pueden ser de distinta índole.

**Datos Estructurados**

Cualquier dato que se pueda almacenar, acceder y procesar en forma de formato fijo se denomina datos "estructurados". A lo largo del tiempo, el talento en ciencias de la computación ha logrado un mayor éxito en el desarrollo de técnicas para trabajar con este tipo de datos (donde el formato es bien conocido de antemano) y también en obtener valor de ellos. Sin embargo, hoy en día, estamos previendo problemas cuando el tamaño de dichos datos crece en gran medida, los tamaños típicos están creciendo en... zettabytes.

Por ejemplo, una tabla de los salarios de una n cantidad de empleados como la que se muestra a continuación es un claro ejemplo de Datos Estructurados:

| **ID de empleado** | **Nombre** | **Sexo** | **Área de adscripción** | **Salario en pesos mexicanos** |
| --- | --- | --- | --- | --- |
| 8754 | Juan Lopez | Male | Contabilidad | 650000 |
| 8796 | Jaffet Chavez | Male | Estadística | 650000 |
| 1589 | Marisol Rodriguez | FeMale | Estadística | 500000 |
| 9857 | José Bonilla | Male | Dirección | 500000 |
| 2584 | Mireya Guadarrama | Female | Dirección | 550000 |

**Datos No Estructurados**

Cualquier dato cuya forma o estructura sea desconocida se clasifica como datos no estructurados. Además de que el tamaño es enorme, los datos no estructurados plantean múltiples desafíos en términos de su procesamiento para obtener valor de ellos. Un ejemplo típico de datos no estructurados es una fuente de datos heterogénea que contiene una combinación de archivos de texto simples, imágenes, videos, etc. Hoy en día, las organizaciones tienen una gran cantidad de datos disponibles, pero desafortunadamente, no saben cómo obtener valor de ellos, ya que estos datos están en su forma sin procesar o en formato no estructurado.

**Datos Semi Estructurados**

Los datos semiestructurados pueden contener ambas formas de datos. Podemos ver los datos semiestructurados como una forma estructurada, pero en realidad no están definidos con p. Ej. una definición de tabla en DBMS relacional. Un ejemplo de datos semiestructurados es un dato representado en un archivo XML.

**📹 Video 01.** Qué es Big Data: Tipos de Datos, Dimensiones, Aplicaciones, Beneficios

A continuación, puedes ver el siguiente vídeo en YouTube que ofrece una clara introducción conceptual al tema.



El link del vídeo está justo [aquí.](https://www.youtube.com/watch?v=l7-SKbcv6is)

❓ **Actividad.** Cuestionario sobre Big Data

1. ¿Qué son los datos estructurados?

1. **Datos comprendidos y bien organizados en variables por filas y columnas**
2. Es una forma de programación basada en datos
3. Es un método para comprender SQL

2. ¿Qué son los datos no estructurados?

1. **Son los más abundantes. Requieren un alto poder de computo para discernir entre lo que es útil y lo que no.**
2. Son conjuntos de datos sin una estructura clara
3. Integración de datos masivos con .Py

3. ¿Qué son los datos semi estructurados?

1. **Una combinación de datos estructurados como no estructurados**
2. Los datos generados por las máquinas
3. La data del gobierno