# Trabajo 1 Arquitectura Big Data y Data Science

## Caso

Su grupo forma parte de un equipo de arquitectura de sistemas en una empresa llamada XYZ Company, dedicada a la venta de productos de toda índole a través de distintos portales web de e-commerce de los cuales esta empresa es la propietaria. Por ejemplo, tienen un sitio web dedicado a la venta de zapatos, otro dedicado a muebles de hogar, entre otros tipos de productos.

En el último año, los dueños de la empresa han evaluado distintas alternativas para aumentar los ingresos generados ya que se ha visto un estancamiento en los ingresos y un aumento en los costos, y luego de varios meses, han decidido promocionar la venta de sus productos mediante publicidad contextual, es decir, por ejemplo, si un usuario está navegando por el sitio de zapatos, le podría aparecer publicidad de productos de su sitio ropa.

A ustedes se les ha encargado proponer una arquitectura donde puedan obtener los datos generados por los sitios web (compras, clics, etc.), información de redes sociales, entre otros para poder disponibilizarlo a científicos de datos (existen 2 en la empresa), quienes puedan generar un modelo de recomendación de productos, que debe ser desplegado cuando el usuario está navegando por alguno de los sitios en tiempo real.

También debe desarrollar modelos y reportes para probar que la arquitectura implementada cumple con las expectativas.

La tasa de crecimiento de información de compras es de 7 GB/día, y actualmente ya cuenta con información histórica de un total de 5 TB para analizar. La tasa de clics entre todos los sitios es de 2 GB/día (proveniente de datos desde Google Analytics), con información histórica de 2.3 TB. Por último, los datos de redes sociales, aunque no se sabe a priori cuanta información diaria se recolecta, la historia ya rescatada es de 23 TB, los cuales consideran fotos, videos, post, etc.

Esta arquitectura debe ser escalable en el caso que a posterior aparezcan nuevos orígenes de datos o haya que cambiar los actuales, o también en caso que se decida cambiar un componente por otro. Y también debe desplegar información analítica para medir la efectividad de estas promociones con el fin de que los dueños puedan evaluar si la meta se está cumpliendo.

La infraestructura actual de la empresa es Cloud (Microsoft Azure), la cual sólo se utiliza para tener los sitios y bases de datos de ventas. Pero además cuenta con servidores on-premise subutilizados (actualmente se paga por servidores que no se utilizan). Estos servidores on-premise se descomponen en lo siguiente:

1. 8 máquinas con 64GB y 2 TB de almacenamiento.
2. Sistema Operativo de las máquinas: Red Hat Enterprise Linux 6.8
3. Todas las máquinas cuentan con acceso a Internet mediante un router (que comparten las 8 máquinas), cada máquina cuenta con su propia IP pública.

## Restricciones

1. El presupuesto total para este proyecto es de US$ 300.000
   1. El presupuesto debe ser considerado para el levantamiento de recursos tecnológicos como de recursos humanos.
2. La empresa no ha adquirido adicionalmente licencias de software de ningún tipo, por lo que debe ser considerado dentro del presupuesto si dentro de la arquitectura hay componentes propietarios.
3. La infraestructura on-premise actual, debe ser considerada como un sistema de respaldo es caso que la nube pueda perder la conexión.

Ustedes envían una propuesta inicial, pero es rechazada por esta incompleta o no considerar todos los puntos mencionados. Tienen una oportunidad de ajustarla y volverla a enviar.

## Entregable

Debe ajustar y corregir la propuesta inicial que envió, para entregar algo más completo. Esos ajustes deberían contener al menos:

* Verificar si la arquitectura propuesta es consistente o debe ser cambiada, con más o menos tecnologías, o removiendo algunas de las propuestas.
* Ajustar los costos y tiempos del proyecto, considerando las restricciones.
* Completar el recurso humano necesario para este proyecto (está incompleto).
* Completar cuántos modelos y reportes añadiría al entregable del proyecto.

Fecha de entrega: Viernes 30 de junio a las 23:59 a los correos [n.campos.rojas@gmail.com](mailto:n.campos.rojas@gmail.com)