SENATI

PLAN DE TRABAJO DEL ESTUDIANTE

RDA
RECURSO DIDÁCTICO PARA EL APRENDIZAJE

DATOS DEL ESTUDIANTE

Apellidos y Nombres:	Pachas Saaveda Jesús Omar	ID:	001514827
Dirección Zonal/CFP:	Ica – Ayacucho / CFP Chincha		
Carrera:	Ingeniería de software con inteligencia artificial	Semestre:	VI
Curso/ Mód. Formativo:	Big Data y Análisis De Datos		
Tema de Trabajo Final:	Fundamentos del Big Data y el Análisis de Datos		

1. INFORMACIÓN

- **Identifica la problemática del caso práctico propuesto.**

La empresa DataCorp, debido a un aumento en la cantidad de datos generados a partir de redes sociales, compras en línea, sensores IoT en tiendas físicas y registros de clientes entre otros, presenta diversos problemas, como:

- Problemas de almacenamiento.
- Largos tiempos de procesamiento en el análisis de datos.
- Dificultad para extraer información relevante y útil.

- **Identifica propuesta de solución y evidencias.**

Propuestas de solución frente a la problemática presentada:

- Identificar la causa que genera datos masivos.
- Identificar datos duplicados o poco relevantes.
- Implementación de una base de datos NoSQL para manejar de manera más optima los datos masivos.
- Describir la arquitectura Big Data que se utilizará en la solución.

- **Respuestas a preguntas guía**

Durante el análisis y estudio del caso práctico, debes obtener las respuestas a las interrogantes:

Pregunta 01:	**¿Cuáles son las principales fuentes de datos masivos en la empresa y cómo pueden ser gestionadas de manera eficiente?**

Principales fuentes de datos masivos en DataCorp:

- Redes sociales
- Compras en línea
- Sensores IoT en tiendas físicas

Estos datos pueden ser gestionados adecuadamente implementando una base de datos NoSQL, que permita un almacenamiento flexible y escalable para grandes volúmenes de información.

Pregunta 02:	**¿Qué ventajas y desventajas tienen los diferentes sistemas de almacenamiento de datos masivos para la empresa?**

La clave para gestionar adecuadamente grandes volúmenes de datos, está en usar almacenamiento escalable, procesamiento en tiempo real y por lotes, junto con herramientas analíticas para extraer valor práctico de los datos.

Ventajas:

- Permite guardar grandes volúmenes de datos en formato estructurado o no estructurado.
- Centraliza la información para su posterior procesamiento.
- Fácil integración con sistemas de procesamiento por lotes y en tiempo real.

Desventajas:

- Puede ser costoso si no es escalable.
- Almacenamiento rígido si no se utiliza tecnología flexible (ej. bases NoSQL).
- Riesgo de cuellos de botella si no se optimiza correctamente.

Pregunta 03:	**¿Cuál es el framework de procesamiento de datos más adecuado para los objetivos de DataCorp y por qué?**

El framework de procesamiento de datos más adecuado para DataCorp es Apache Spark, debido a su capacidad para procesar grandes volúmenes de datos de manera rápida y eficiente gracias a su procesamiento en memoria. Además, Spark soporta tanto el procesamiento por lotes como en tiempo real, lo que permite a la empresa analizar datos provenientes de redes sociales, compras en línea y sensores IoT de forma simultánea.

Pregunta 04:	**¿Cómo se pueden aplicar herramientas de análisis de Big Data para mejorar la toma de decisiones empresariales en la empresa?**

Las herramientas de análisis de Big Data pueden mejorar la toma de decisiones empresariales en la empresa al permitir recopilar, procesar y analizar grandes volúmenes de datos provenientes de diversas fuentes en tiempo real.

Pregunta 05:	**¿Cuáles son los principales desafíos éticos y legales en la gestión de datos masivos dentro de DataCorp?**

Los principales desafíos éticos y legales dentro de DataCorp incluyen:

- **Cumplimiento legal:** Asegurar que la empresa cumpla con normativas como el GDPR u otras leyes locales de protección de datos para evitar sanciones legales.
- **Transparencia y equidad:** Mantener claridad sobre cómo se utilizan los datos y prevenir sesgos en los algoritmos que puedan generar decisiones injustas o discriminatorias.
- **Seguridad de la información:** Implementar medidas robustas para proteger los datos contra accesos no autorizados, pérdidas o manipulaciones maliciosas.

2. PLANIFICACIÓN DEL TRABAJO

- **Cronograma de actividades:**

N°	ACTIVIDADES	CRONOGRAMA					
		12/08/2025					
01	Identifica la problemática del caso práctico propuesto	X					
02	Identifica propuesta de solución y evidencias	X					
03	Respuestas a preguntas guía	X					
04	Desarrollo de la propuesta solución	X					

- **Lista de recursos necesarios:**

1. MÁQUINAS Y EQUIPOS	
Descripción	**Cantidad**
Pc laboratorio	
Laptop	

2. HERRAMIENTAS E INSTRUMENTOS	
Descripción	**Cantidad**
Canva	
Apuntes de clases	

3. MATERIALES E INSUMOS	
Descripción	**Cantidad**
Intertet	
Páginas web	

3. DECIDIR PROPUESTA

- **Describe la propuesta determinada para la solución del caso práctico**

PROPUESTA DE SOLUCIÓN

1. **Video 01:**

BIG DATA:

Big Data es un conjunto de grandes cantidades de datos, estructurados y no estructurados que, por su tamaño y velocidad de generación no pueden ser gestionados ni analizados con las herramientas tradicionales de bases de datos.

Hadoop:

Hadoop es un framework de código abierto desarrollado por la Apache Software Foundation que permite el almacenamiento y procesamiento distribuido de grandes volúmenes de datos en clusters de computadoras utilizando un modelo sencillo de programación.

Donde se usa hadoop:

Hadoop se utiliza en sectores que generan grandes volúmenes de datos, como bancos, telecomunicaciones, empresas de tecnología, salud, gobiernos y en general cualquier organización que requiera procesar información a gran escala.

Desarrollo y uso:

Hadoop está desarrollado principalmente en Java, por lo que este es el lenguaje más utilizado para interactuar directamente con su núcleo. Sin embargo, gracias a su ecosistema y a las herramientas que lo complementan, también se puede usar con otros lenguajes:

- **Python:** Muy usado con Pydoop o mediante frameworks como Apache Spark sobre Hadoop, ideal para análisis de datos y machine learning.
- **R:** Se utiliza en análisis estadístico y científico de datos sobre Hadoop, especialmente con librerías como RHadoop.
- **C++:** Puede usarse en casos específicos a través de pipes y extensiones de MapReduce para tareas de alto rendimiento.

Las 3 V's del Big Data:

. Volumen
Se refiere a la gran cantidad de datos que se generan constantemente, estos volúmenes de datos puede ir desde terabytes hasta petabytes o exabytes. La clave del Big Data es que no basta con almacenar esa

información, sino que se deben contar con herramientas y tecnologías capaces de procesarla y extraer valor de ella.

. Velocidad
Es la rapidez con la que los datos se generan, transmiten y procesan. Hoy en día, la información llega en tiempo real como por ejemplo: streams, notificaciones, GPS).

.Variedad
La variedad hace referencia a la diversidad de fuentes y tipos de datos que existen. Antes, la mayoría de los datos eran estructurados (tablas y bases de datos relacionales). Hoy en día, además de esos, también encontramos datos semiestructurados (archivos JSON, XML, correos electrónicos) y datos no estructurados (videos, audios, imágenes, publicaciones en redes sociales).

2. **Video 02:**

Apache Spark:

Apache Spark es un framework de procesamiento de datos distribuidos diseñado para manejar grandes volúmenes de información de manera rápida y eficiente. Funciona en memoria, lo que significa que, en lugar de escribir y leer constantemente desde el disco como hacen otras tecnologías, mantiene los datos en la memoria RAM durante el procesamiento

Cluster:

Un clúster es un conjunto de computadoras interconectadas que trabajan como un solo sistema para procesar grandes volúmenes de datos. Dentro de este clúster, el Driver actúa como el nodo maestro encargado de coordinar la ejecución de las tareas, mientras que los Workers son los nodos de trabajo que realizan el procesamiento real sobre los datos.

Lenguajes de programación:

Apache Spark se puede usar en varios lenguajes de programación:

- **Java** → También soportado de forma oficial, útil para entornos empresariales que ya trabajan con este lenguaje.
- **Python** (PySpark) → Uno de los más usados por su simplicidad y la facilidad para integrarse con librerías de análisis de datos y Machine Learning.
- **R (SparkR)** → Orientado a la estadística y el análisis de datos, permite a los usuarios de R aprovechar la escalabilidad de Spark.

3. **Video 03:**

Hadoop vs Spark

Hadoop y Apache Spark son dos de las tecnologías más importantes en el mundo del Big Data. Ambos permiten trabajar con grandes volúmenes de datos

que no podrían manejarse con sistemas tradicionales, pero lo hacen de formas distintas.

Principales similitudes:

. Procesamiento de grandes volúmenes de datos:
Ambos están diseñados para trabajar con Big Data, gestionando conjuntos de datos demasiado grandes para sistemas tradicionales.

. Arquitectura distribuida:
Hadoop y Spark utilizan un clúster de computadoras para dividir el trabajo en múltiples nodos, lo que permite escalabilidad y tolerancia a fallos.

. Uso en entornos empresariales:
Se aplican en áreas como análisis de datos, machine learning, procesamiento de logs, IoT y analítica en tiempo real o batch, dependiendo de la necesidad.

. Compatibilidad con múltiples lenguajes:
Tanto Hadoop como Spark pueden usarse con Java, Python y Scala, lo que los hace flexibles para distintos perfiles de desarrolladores y científicos de datos.

Principales diferencias:

1. **Velocidad de procesamiento:**

- **Hadoop**: Procesa datos en disco (lectura y escritura constante), lo que lo hace más lento.
- **Spark**: Procesa datos principalmente en memoria (RAM), lo que lo hace hasta 100 veces más rápido en algunas tareas.

2. **Procesamiento de datos:**

- **Hadoop:** Está orientado a procesamiento batch (por lotes).
- **Spark:** Soporta batch y streaming en tiempo real, lo que lo hace más versátil.

3. **Facilidad de uso:**

- **Hadoop:** Requiere escribir programas en Java MapReduce, que suelen ser más complejos.
- **Spark:** Ofrece APIs más simples en Scala, Python, Java y R, con librerías integradas como MLlib (machine learning) y GraphX (gráficos).

4. **Uso de recursos:**

- **Hadoop:** Consume menos memoria pero depende mucho del disco.
- **Spark**: Requiere más memoria RAM, pero aprovecha mejor los recursos del sistema.

4. EJECUTAR

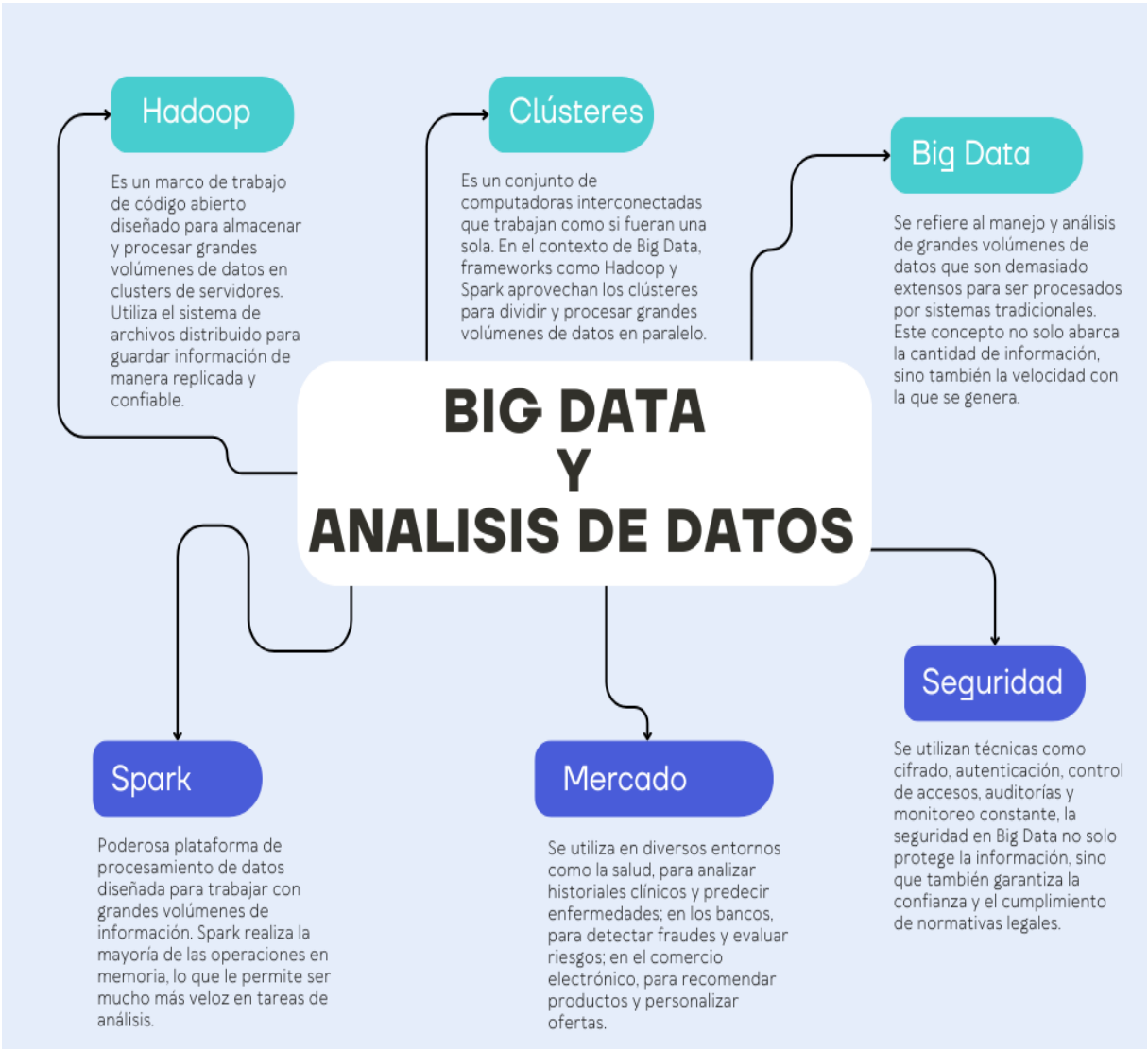
- **Resolver el caso práctico, utilizando como referencia el problema propuesto y las preguntas guía proporcionadas para orientar el desarrollo.**
- **Fundamentar sus propuestas en los conocimientos adquiridos a lo largo del curso, aplicando lo aprendido en las tareas y operaciones descritas en los contenidos curriculares.**

INSTRUCCIONES: Ser lo más explícito posible. Los gráficos ayudan a transmitir mejor las ideas. Tomar en cuenta los aspectos de calidad, medio ambiente y SHI.

OPERACIONES / PASOS / SUBPASOS	NORMAS TÉCNICAS - ESTANDARES / SEGURIDAD / MEDIO AMBIENTE

DIBUJO / ESQUEMA / DIAGRAMA DE PROPUESTA

(Adicionar las páginas que sean necesarias)



5. CONTROLAR

- **Verificar el cumplimiento de los procesos desarrollados en la propuesta de solución del caso práctico.**

EVIDENCIAS	CUMPLE	NO CUMPLE
• ¿Se identificó claramente la problemática del caso práctico?	☐	☐
• ¿Se desarrolló las condiciones de los requerimientos solicitados?	☐	☐
• ¿Se formularon respuestas claras y fundamentadas a todas las preguntas guía?	☐	☐
• ¿Se elaboró un cronograma claro de actividades a ejecutar?	☐	☐
• ¿Se identificaron y listaron los recursos (máquinas, equipos, herramientas, materiales) necesarios para ejecutar la propuesta?	☐	☐
• ¿Se ejecutó la propuesta de acuerdo con la planificación y cronograma establecidos?	☐	☐
• ¿Se describieron todas las operaciones y pasos seguidos para garantizar la correcta ejecución?	☐	☐
• ¿Se consideran las normativas técnicas, de seguridad y medio ambiente en la propuesta de solución?	☐	☐
• ¿La propuesta es pertinente con los requerimientos solicitados?	☐	☐
• ¿Se evaluó la viabilidad de la propuesta para un contexto real?	☐	☐

6. VALORAR

- **Califica el impacto que representa la propuesta de solución ante la situación planteada en el caso práctico.**

CRITERIO DE EVALUACIÓN	DESCRIPCIÓN DEL CRITERIO	PUNTUACIÓN MÁXIMA	PUNTAJE CALIFICADO POR EL ESTUDIANTE
Identificación del problema	Claridad en la identificación del problema planteado.	3	
Relevancia de la propuesta de solución	La propuesta responde adecuadamente al problema planteado y es relevante para el contexto del caso práctico.	8	
Viabilidad técnica	La solución es técnicamente factible, tomando en cuenta los recursos y conocimientos disponibles.	6	
Cumplimiento de Normas	La solución cumple con todas las normas técnicas de seguridad, higiene y medio ambiente.	3	
PUNTAJE TOTAL		**20**	


SENATI