Guía de Estudio Ingeniería

TEMA 1. INTRODUCCIÓN A LAS TECNOLOGÍAS BIG DATA	1.1. Las Tres V's del Big Data	**Volumen, Velocidad y Variedad** son las características principales del Big Data. Entenderlas es clave para identificar un proyecto de Big Data.	
		Volumen: Se refiere a la gran cantidad de datos que se generan y necesitan ser procesados.	
		Velocidad: La rapidez con la que se generan y procesan los datos.	
		Variedad: La diversidad de los tipos de datos, estructurados y no estructurados.	
	1.2. Definición de Big Data y sus Aplicaciones	Big Data son conjuntos de tecnologías diseñadas para manejar grandes volúmenes de datos heterogéneos.	
		Apache Kafka: Es un sistema de mensajería que permite manejar flujos de datos en tiempo real.	
TEMA 2. HDFS Y MAPREDUCE	2.1. HDFS (Hadoop Distributed File System)	HDFS es un sistema de archivos distribuido que permite el almacenamiento y procesamiento de grandes volúmenes de datos a través de múltiples máquinas.	
		Comandos HDFS relevantes:	**mkdir**: Crea un nuevo directorio en HDFS.
			rm: Elimina archivos en HDFS, y puede hacerlo de manera recursiva para carpetas.
			chmod: Cambia los permisos de un archivo en HDFS.
	2.2. MapReduce y Apache Spark	**MapReduce**: Es un paradigma de programación para procesar datos distribuidos en clústeres. Se basa en las funciones Map y Reduce.	
		Apache Spark: Mejora las limitaciones de MapReduce al permitir operaciones en memoria y un procesamiento más rápido.	
		Funciones Map y Reduce: Spark permite realizar operaciones de mapeo y reducción en paralelo, distribuyendo tareas entre nodos.	
TEMA 3. SPARK I	3.1. Aprendizaje Automático en Big Data	Los algoritmos de Machine Learning se benefician de Spark al permitir múltiples iteraciones sobre los datos.	
		RDD (Resilient Distributed Dataset): Es una estructura de datos clave en Spark que permite almacenar y procesar datos en paralelo.	
	3.2. Métodos Principales en Spark ML	El método **fit** es el principal en los estimadores de Spark ML para ajustar los modelos a los datos.	
		Transformers y **Pipelines**: Después del entrenamiento, los modelos utilizan estas interfaces para procesar nuevos datos.	
TEMA 4. SPARK II	4.1. DataFrames y SQL en Spark	**DataFrames**: Estructuras de datos distribuidas que permiten manipular datos de forma eficiente en Spark.	
		Spark SQL: Interfaz que permite ejecutar consultas SQL sobre DataFrames.	
TEMA 5. SPARK III	5.1. Spark MLlib y Structured Streaming	**MLlib**: Biblioteca de aprendizaje automático de Spark que incluye algoritmos como regresión, clasificación y clustering.	
		Spark Structured Streaming: Módulo para procesar flujos de datos en tiempo real.	
TEMA 6. APACHE KAFKA	6.1. Introducción a Apache Kafka	Kafka es una plataforma de mensajería distribuida que permite la publicación, suscripción y procesamiento de flujos de datos en tiempo real.	
		Producers y Consumers: Los productores envían mensajes a Kafka, mientras que los consumidores los leen para su procesamiento.	
TEMA 7. HIVE E IMPALA	7.1. Apache Hive y Apache Impala	**Hive**: Herramienta para manejar datos en HDFS usando SQL.	
		Impala: Motor de consultas SQL que permite ejecutar consultas en HDFS con alta eficiencia.	
TEMA 8. CLOUD COMPUTING I	8.1. Introducción a Cloud Computing	**Cloud Computing**: Modelo de computación que permite el acceso a recursos de computación y almacenamiento a través de la nube.	
		Ventajas del Cloud Computing: Escalabilidad, reducción de costos y flexibilidad en el acceso a los recursos.	
	8.2. Servicios en la Nube	**Tipos de Nube**: Pública, privada, híbrida.	
		Servicios en la Nube: IaaS, PaaS, SaaS.	
		Microsoft Azure: Plataforma en la nube que ofrece una amplia gama de servicios.	
TEMA 9. CLOUD COMPUTING II	9.1. Amazon Web Services (AWS)	**AWS**: Plataforma de servicios en la nube que ofrece soluciones de computación, almacenamiento, bases de datos y más.	
		Servicios de AWS: Computación, almacenamiento, redes, bases de datos, seguridad y más.	
TEMA 10. CLOUD COMPUTING III	10.1. Google Cloud Platform (GCP)	**GCP**: Plataforma de servicios en la nube de Google que ofrece soluciones de computación, almacenamiento, bases de datos, y análisis de Big Data.	
		Servicios de GCP: Computación, almacenamiento, redes, bases de datos, y machine learning.	