

# BIG DATA AVANZADO\_

## Modalidad:

e-learning con una duración 56 horas

## Objetivos:

1. Garantizar la integridad y calidad de los datos en sistemas distribuidos.
2. Gestionar sistemas de almacenamiento distribuidos y asegurar la tolerancia a fallos.
3. Aplicar herramientas de Big Data como MapReduce, Pig, Hive y Oozie en la automatización de procesos.
4. Evaluar los factores de riesgo.
5. Monitorear y optimizar entornos Big Data utilizando herramientas especializadas.
6. Implementar modelos de Inteligencia de Negocios (BI) y procesos Knowledge Discovery in Databases (KDD) en la toma de decisiones
7. Validar técnicas de Big Data en escenarios de negocios reales BI

## Contenidos:

### TEMA 1

#### 1. GESTIÓN DE SOLUCIONES CON SISTEMAS DE ALMACENAMIENTO Y HERRAMIENTAS DEL CENTRO DE DATOS PARA LA RESOLUCIÓN DE PROBLEMAS

##### 1.1. Conceptos básicos de Big Data

###### 1.1.1. Características de los sistemas de almacenamiento distribuidos

###### 1.1.2. Principios de tolerancia a fallos en sistemas de almacenamiento

##### 1.2 Procesamiento de datos en entornos Big Data

###### 1.2.1 Fundamentos de computación distribuida y paralela

## 1.2.2 Paradigmas de procesamiento de datos masivos

## TEMA 2

### 2. ANALÍTICA DE BIG DATA EN LOS ECOSISTEMAS DE ALMACENAMIENTO

#### 2.1. Analítica de Big Data en los ecosistemas de almacenamiento

##### 2.1.1. Conceptos de análisis de datos a gran escala

##### 2.1.2. Técnicas de procesamiento y análisis en Big Data

#### 2.2. Big Data y Cloud: conceptos y sinergia

##### 2.2.1. Fundamentos de Cloud Computing

##### 2.2.2. Integración conceptual de Big Data con tecnologías Cloud

#### 2.3. Estrategias y metodologías para la resolución de problemas en entornos de Big Data

## TEMA 3

### 3. GESTIÓN DE SISTEMAS DE ALMACENAMIENTO Y ECOSISTEMAS BIG DATA

#### 3.1. Teoría avanzada de computación distribuida y paralela

##### 3.1.1. Teorema CAP y sus implicaciones

##### 3.1.2. Modelos de consistencia en sistemas distribuidos

##### 3.1.3. Sincronización y consenso distribuido

#### 3.2 Arquitectura y diseño de sistemas de almacenamiento distribuidos

##### 3.2.1. Arquitectura y diseño de sistemas de almacenamiento distribuidos

- 3.2.2. Arquitectura detallada de HDFS
- 3.2.3. Comparación con otros sistemas de almacenamiento distribuido
- 3.2.4. Estrategias de replicación y consistencia
- 3.3. Ecosistemas Big Data: componentes y funcionalidades
  - 3.3.1. Fundamentos de procesamiento distribuido de datos
  - 3.3.2. Principios de consulta y análisis de datos masivos
  - 3.3.3. Conceptos de ingesta y exportación de datos en sistemas Big Data

## TEMA 4

### 4. TEORÍA DE LA AUTOMATIZACIÓN DE TRABAJOS EN ENTORNOS BIG DATA

- 4.1. Conceptos de orquestación de flujos de trabajo
- 4.2. Apache Oozie: arquitectura y funcionalidades
- 4.3. Apache Airflow: arquitectura y ventajas
- 4.4. Comparación entre Oozie y Airflow
- 4.5. Lenguajes de consulta para Big Data: principios y conceptos
  - 4.5.1. HiveQL: sintaxis y funcionalidades
  - 4.5.2. Pig Latin: constructos y transformaciones de datos
  - 4.5.3. Optimización de consultas en HiveQL y Pig Latin
- 4.6. Tendencias y evolución en ecosistemas Big Data
  - 4.6.1. Procesamiento en tiempo real y arquitecturas lambda

4.6.2. Machine Learning a gran escala

4.6.3. Gobernanza de datos y cumplimiento normativo

## TEMA 5

### 5. GENERACIÓN DE MECANISMOS DE INTEGRIDAD DE LOS DATOS. COMPROBACIÓN DE MANTENIMIENTO DE SISTEMAS DE FICHEROS

5.1. Conceptos de calidad de datos en sistemas Big Data

5.1.1. Desafíos de calidad de datos en Big Data

5.1.2. Técnicas de evaluación de calidad de datos

5.1.3. Mejores prácticas para mantenimiento de calidad de datos

5.2. Comprobación de la integridad de datos en sistemas de ficheros distribuidos

5.2.2. Procesos de verificación de integridad en HDFS

5.2.3. Recuperación de datos en caso de corrupción

5.3. Movimiento de datos entre clústeres

5.3.1. Uso de DistCp para copia distribuida

5.3.2. Estrategias para transferencias eficientes

5.3.3. Consideraciones de seguridad en el movimiento de datos

5.4. Actualización y migración de datos

5.4.1. Planificación de actualizaciones de Hadoop

5.4.2. Estrategias de migración de datos

#### 5.4.3. Validación post-migración

### 5.5. Gestión de metadatos en sistemas Big Data

#### 5.5.1. Importancia de los metadatos en Big Data

#### 5.5.2. Herramientas de gestión de metadatos

#### 5.5.3. Implementación de catálogos de datos

## TEMA 6

# 6. MONITORIZACIÓN, OPTIMIZACIÓN Y SOLUCIÓN DE PROBLEMAS

## 6.1. Herramientas de monitorización

### 6.1.1. Principios de monitorización de trabajos y recursos

### 6.1.2. Conceptos de monitorización de clústeres

## 6.2. Análisis de logs e históricos: teoría y mejores prácticas

### 6.2.1. Tipos de logs en ecosistemas Hadoop

### 6.2.2. Técnicas de análisis de logs

### 6.2.3. Mejores prácticas en gestión de logs

## 6.3. Principios de optimización del rendimiento en sistemas Big Data

### 6.3.1. Optimización de configuración de Hadoop

### 6.3.3. Optimización de aplicaciones Spark

### 6.3.4. Optimización de consultas en Hive e Impala

## 6.4. Metodologías de resolución de problemas en entornos Big Data

### 6.4.1 Enfoque sistemático para el troubleshooting

- 6.4.2. Herramientas de diagnóstico en Hadoop
- 6.4.3. Escenarios comunes de problemas y sus soluciones
- 6.4.4. Prácticas preventivas y mantenimiento proactivo

## TEMA 7

### 7. VALIDACIÓN DE TÉCNICAS BIG DATA EN LA TOMA DE DECISIONES EN INTELIGENCIA DE NEGOCIOS (BI)

- 7.1. Modelos de Inteligencia de negocios BI
  - 7.1.1. Evolución de BI en la era del Big Data
  - 7.1.2. Arquitecturas de BI para Big Data
  - 7.1.3. Casos de uso de Big Data en BI
- 7.2. Knowledge Discovery in Databases)
  - 7.2.1. Selección de datos
  - 7.2.2. Limpieza de datos
  - 7.2.3. Transformación de datos
  - 7.2.4. Minería de datos
  - 7.2.5. Interpretación y evaluación de datos
- 7.3. Implantación de modelos de inteligencia de negocios BI
  - 7.3.1. Arquitecturas para BI en tiempo real
  - 7.3.2. Integración de Big Data con herramientas tradicionales de BI
- 7.4. Técnicas de validación de modelos Big Data

- 
- 7.4.1. Validación cruzada en entornos distribuidos
  - 7.4.2. Técnicas de remuestreo para Big Data
  - 7.4.3. Validación temporal para modelos de series temporales
  - 7.4.4. Métricas de evaluación para modelos de Big Data