GUION DEL CURSO

DENOMINACIÓN DE LA ESPECIALIDAD: ANALISTA DE DATOS BIG DATA CLOUDERA

CÓDIGO DE LA ESPECIALIDAD: IFCT35

CÓDIGO DEL CURSO: 24/1513

FECHA DE INICIO: 13/11/2024 **FECHA DE FINAL:** 09/01/2025 **HORARIO:** 09:00 h a 14:00 h

CALENDARIO (días no lectivos): 6, 23, 24, 25, 26, 27, 30 y 31 de diciembre de 2024, 1, 2, 3 y 6 de enero de 2025.

DURACIÓN (horas totales): 150

OBJETIVO GENERAL DEL CURSO:

Extraer, manejar, almacenar, buscar y visualizar grandes volúmenes de datos de diferentes tipos relacionándolos entre sí para obtener información relevante mediante herramientas y soluciones Cloudera.

RELACIÓN DE MÓDULOS (objetivo, contenido y duración):

Módulo 1. Denominación: INTRODUCCIÓN A SQL.

Objetivo:

Conocer los fundamentos y dominar las acciones operaciones necesarias que permitan interactuar con bases de datos relacionales utilizando el lenguaje SQL.

Contenido teórico/práctico:

- Introducción a las BBDD:
  Conceptos básicos de bases de datos.
  Utilidad y relevancia del lenguaje SQL.
- Creación y estructura de una base de datos:
  Estructura básica de una base de datos.
  Creación de una base de datos usando SQL.
- Realización de Consultas básicas:
  Búsqueda y recuperación de datos básicos.
  Manejo de consultas simples.
- Mantenimiento de la Base de datos
  Copia de tablas y modificaciones de columnas.
  Índices y restricciones.
  Eliminación o modificación de filas de la tabla de datos.
  Creación de objetos de BBDD (tablas, índices, vistas…)
- Utilización de Funciones:
  Uso y tipos de funciones.
  Funciones predeterminadas.
- Exportación e importación de datos:
  Exportación de datos, consultas y utilidades.
  Importación de datos y archivos de datos
  Importación con sentencias y utilidades
- Utilización de Joins para la extracción de datos
  Definición y tipo de Joins
  Uso de joins para combinar datos de múltiples tablas
- Ejecución y diseño de Subconsultas:
  - Tipos de subconsultas
    - Subconsultas generales y básicas.
    - Subconsulta no correlacionada y correlacionada.
    - Modificación de la tabla con subconsultas.
- Realización de operaciones avanzadas:
  Creación de Vistas.
  Gestión de transacciones.

- Utilización del diccionario de Metadatos:
  Uso de metadatos para obtener información sobre los objetos de la base de datos.

Duración: 50 horas

Módulo 2. Denominación: Fundamentos de Cloudera Apache Hadoop.

Objetivo:

Aprender los conceptos básicos de Big Data, conocer de la arquitectura de Hadoop e Identificar los componentes de Big Data y su integración en el ecosistema de Hadoop.

Contenido teórico/práctico:

- Identificación y asimilación de conceptos
  Introducción a Big Data
  La motivación por Apache Hadoop
  Conceptos básicos de Hadoop
- Visión del Ecosistema de Hadoop
  Soluciones del ecosistema de Hadoop
  Aplicaciones comunes y usos especiales de Hadoop.
- Gestión de Hadoop
  - Uso de herramientas para la gestión del almacenamiento
  - Ejecución de aplicaciones con procesamiento distribuido

Duración: 25 horas

Módulo 3. Denominación: ANALYZING WITH CLOUDERA DATA WAREHOUSE.

Objetivo:

Aplicar las habilidades tradicionales de análisis de datos e inteligencia empresarial a grandes volúmenes de datos conociendo las herramientas que se necesitan para acceder, manipular, transformar y analizar datos de diferentes tipos utilizando SQL y lenguajes de scripting.

Contenido teórico/práctico:

- Fundamentos para el Análisis de Big Data
  Visión General del Análisis de Big Data
  Almacenamiento de Datos: HDFS
  Procesamiento Distribuido de Datos: YARN,
  MapReduce y Spark
  Procesamiento y Análisis de Datos: Hive e Impala
  Integración de Bases de Datos: Sqoop
  Otras Herramientas de Datos
- Introducción a Hive e Impala
  ¿Qué es Hive?
  ¿Qué es Impala?
  ¿Por qué utilizar Hive e Impala?
  Schema y almacenamiento de datos
  Comparación entre Hive y bases de datos tradicionales
  Casos de uso
- Consultas con Hive e Impala
  Tablas y bases de datos
  Sintaxis básica en consultas Hive e Impala
  Tipos de datos
  Empleo de Hue para ejecutar consultas
  Empleo de Beeline (la Shell de Hive)
  Empleo de la Shell de Impala
- Operadores comunes y funciones integradas
  Operadores
  Funciones escalares
  Funciones de agregado
- Administración de datos
  Almacenamiento de datos

Creación de bases de datos y tablas
Carga de datos
Modificación de bases de datos y tablas
Simplificación de consultas con vistas
Almacenamiento de resultados de consultas

- Almacenamiento de datos y rendimiento
  Particionamiento de tablas
  Carga de datos en tablas particionadas
  Cuándo utilizar particionamiento
  Selección del formato de archivo
  Uso de los formatos de archivo Avro y Parquet
- Trabajando con múltiples Datasets
  UNION y Joins
  Manejo de valores NULL en Joins
  Joins avanzados
- Funciones analíticas y Windowing
  Utilización de funciones analíticas comunes
  Otras funciones analíticas
  Sliding Windows
- Datos complejos
  Datos complejos con Hive
  Datos complejos con Impala
- Análisis de texto
  Empleo de expresiones regulares con Hive e Impala
  Procesamiento de texto con SerDes en Hive
  Análisis de Sentimiento y n-grams en Hive
- Optimización de Apache Hive
  Comprender cómo se ejecutan las consultas
  Optimización basada en Costes y Estadísticas
  Bucketing
  Optimizaciones de ficheros ORC
  Indexación de datos
  Hive en Spark
- Optimización de Apache Impala
  Cómo Impala ejecuta las consultas
  Mejorar el rendimiento de Impala
- Extendiendo Hive e Impala
  Customizar SerDes y formatos de archivo en Hive
  Transformación de datos con Scripts personalizados en Hive
  Funciones definidas por el usuario
  Consultas parametrizadas
- Selección de la Mejor Herramienta para cada Tarea
  Comparación entre MapReduce, Hive, Impala, y bases de datos relacionales ¿Cuál elegir?
- Introducción a CDP Public Cloud Data Warehouse

Duración: 75 horas

Módulo 4: Prevención de Riesgos Laborales (transversal)

<u>METODOLOGÍA DIDÁCTICA (máquinas virtuales, conexiones remotas, versiones software, dominios, suscripciones, …):</u>

Para cada alumno se incluirá una máquina virtual en remoto, como mínimo, de las siguientes características:

- Memoria RAM 40 GB
- 2 Discos SSD
- Procesador Intel AMD (24 cores)

El curso usará a VMware Virtual Machine (VM), las cuales se irán configurando según las necesidades durante el curso según marca Cloudera:

- VMware Workstation 17.6
- VirtualBox v7.1.4-165100-Win
- Navegador Chrome Versión 130.0.6723.59 (Build oficial) (64 bits)

<u>MATERIAL DIDÁCTICO DEL ALUMNO (fungible):</u>

A cada alumno se le entregará una mochila, un block de notas y un bolipen.

MEDIOS DIDÁCTICOS DEL ALUMNO:

Se entregará un ejemplar para cada alumno asistente al curso y un ejemplar de muestra para el CRN Getafe.

Manuales oficiales de Cloudera en formato electrónico descargable en idioma inglés.

- Learning SQL by Alan Beaulieu (3ª Edición) ISBN 978-1-492-05761-1
- Cloudera_Essentials_for_Apache_Hadoop (Material oficial de Cloudera)
- Analyzing with Cloudera Data Warehouse (Material oficial de Cloudera)

CERTIFICACIÓN OFICIAL DE FABRICANTE (si procede):

Certificación a la que se opta: La ejecución y financiación del programa formativo incluye la presentación de los alumnos que han realizado el curso con aprovechamiento el siguiente examen de certificación oficial o el que lo sustituya actualizado al momento de su impartición:

CDP Data Analyst CDP-4001

Entrega y recepción del Voucher: Los asistentes a la formación de que hayan resultado APTOS en las pruebas finales recibirán un correo electrónico con las instrucciones para llevar a cabo la realización del examen incluido en el curso.

La validez del Voucher será de un año desde la finalización del curso.

Realización del examen: Las instrucciones incluyen los enlaces para reservar fecha, se puede modificar hasta 48 horas antes. Los exámenes de Certificación los podrán realizar en cualquier ordenador con acceso a internet, webcam y micrófono (son válidos los incluidos en los portátiles). En el momento de realizar la prueba deben de tener algún documento que acredite su identificación (DNI, pasaporte…). Sólo está incluido en el programa el primer intento.

El examen es en idioma inglés.

Resultado del examen: El examen se califica inmediatamente después de la presentación y se le enviará por correo electrónico un informe de resultados el mismo día.

Entrega de acreditación a alumno: Si pasa el examen, recibirá un segundo correo electrónico a los pocos días de su examen con su certificado digital en formato PDF, el número de licencia, una actualización del perfil de Linkedin, y un enlace para descargar los logotipos de CCA para su uso en su personal garantía de negocio y perfiles en redes sociales.

SISTEMA DE EVALUACIÓN (instrumentos, criterios y momentos):

La superación del programa y, en consecuencia, la obtención del Diploma acreditativo de la CAM, están sujetas al cumplimiento de las siguientes condiciones que serán transmitidas al alumno/a al inicio de la formación.

El criterio de evaluación del alumnado es continuo, teniendo especial importancia para ello la asistencia a las clases, la puntualidad, la resolución de cada uno de los ejercicios de los distintos módulos y la actitud proactiva.

El instructor evaluará las competencias y comportamientos de los asistentes: puntualidad, asistencia, responsabilidad en la ejecución de las prácticas, perseverancia, comunicación, trabajo en equipo, proactividad y resolución correcta de las prácticas durante el curso.

El criterio del instructor en base a los aspectos descritos puntuará de 1 a 10 y ponderará sobre la nota final un 30%

A la finalización del curso se realizará una prueba tipo test para valorar el aprendizaje de los asistentes, este criterio puntuará de 1 a 10 y ponderará un 70% sobre la nota final.

Los asistentes serán alumnos APTOS si la nota final del curso es igual o superior a 5.