ACADEMIA INTEGRAL
CENTRO DE NUEVAS TECNOLOGIAS

LA FORMACIÓN ES LA CLAVE DEL ÉXITO

Guía del Curso

IFCT165PO BIG DATA PARA INGENIERÍAS

Modalidad de realización del curso: A distancia y Online

Titulación: Diploma acreditativo con las horas del curso

OBJETIVOS

Este CURSO IFCT165PO BIG DATA PARA INGENIERÍAS le ofrece una formación especializada en la materia dentro de la Familia Profesional de Informática y comunicaciones. Con este CURSO IFCT165PO BIG DATA PARA INGENIERÍAS el alumno será capaz de desenvolverse dentro del Sector y conocer las tecnologías disponibles para realizar estrategias de Big Data para Ingenierías, realizar un desarrollo con Spark y Hadoop y analizar datos con Pig Hive e Impala.

CONTENIDOS

UNIDAD DIDÁCTICA 1. INTRODUCCIÓN

1. ¿Qué es Big Data?
2. Paradigmas de procesamiento en Big Data
3. Las 8 V de Big Data (Volumen, Volatilidad, Variedad, Valor, Velocidad, Variabilidad, Veracidad, Validez)

UNIDAD DIDÁCTICA 2. BATCH PROCESSING

1. MapReduce
2. - Entorno MapReduce

3. - Función Map y función Reduce
4. - Flujo de datos
5. - Características de MapReduce
6. - Uso de MarpReduce
7. - Ventajas e inconvenientes de Map Reduce
8. - Ejercicios y ejemplos con MapReduce
9. Hadoop
10. - Entorno Hadoop
11. - Almacenamiento: HDFS
12. - Características de HDFS
13. Apache Hadoop YARN
14. - Funciones de Framework computacionales
15. - YARN: El gestor de recursos del cluster
16. - Conceptos de Apache Spark
17. - Ejecución de Computational Frameworks en YARN
18. - Exploración de las aplicaciones de YARN Applications a través de la Web UIs y de Shell
19. Agregación de los logs de YARN
20. - Configuración de Hadoop y registros de Daemon
21. - Localizar configuraciones y aplicar cambios de configuración
22. - Gestión de instancias de Role y añadir servicios
23. - Configuración del servicio HDFS
24. - Configuración de los logs de Hadoop Daemon
25. - Configuración del servicio YARN
26. Obtención de datos en HDFS
27. - Ingestión de datos desde fuentes de recursos externos con Flume
28. - Ingestión de datos desde bases de datos relacionales con Sqoop
29. - REST Interfaces
30. - Buenas prácticas para la importación de datos
31. Planificación de un cluster Hadoop
32. - Consideraciones generales de planificación
33. - Elección correcta de Hardware
34. - Opciones de Virtualización
35. - Consideraciones de red
36. - Configuración de nodos

37. Instalación y configuración de Hive, Pig e Impala
38. Clientes Hadoop incluidos en Hue
39. - ¿Qué es un cliente de Hadoop?
40. - Instalación y configuración de clientes Hadoop
41. - Instalación y configuración de Hue
42. - Autorizaciones y autenticación Hue
43. Configuración avanzada de un cluster
44. - Parámetros avanzados de configuración
45. - Configuración de puertos Hadoop
46. - Configuración de HDFS para la organización en rack
47. - Configuración de HDFS para obtención de alta disponibilidad
48. Seguridad Hadoop
49. - ¿Por qué es importante la seguridad en Hadoop?
50. - Conceptos del sistema de seguridad de Hadoop
51. - Qué es Kerberos y cómo funciona
52. - Securización de un clúster Hadoop Cluster con Kerberos
53. - Otros conceptos de seguridad
54. Gestión de recursos
55. - Configuración de cgroups con Static Service Pools
56. - El Fair Scheduler
57. - Configuración de Dynamic Resource Pools
58. - Configuraciones de CPU y memoria YARN
59. - Impala Query Scheduling
60. Mantenimiento de un cluster
61. - Chequeo del estado de HDFS
62. - Copia de datos entre clústers
63. - Añadir y eliminar de nodos en el clúster
64. - Rebalanceo del Cluster
65. - Directorio de Snapshots
66. - Actualización del clúster
67. Solución de problemas y monitorización de un cluster
68. - Sistema general de monitorización
69. - Monitorización de clústers Hadoop
70. - Solución de problemas habituales en el clúster de Hadoop

71. - Errores habituales en la configuración

UNIDAD DIDÁCTICA 3. CIENCIA DE DATOS

1. Data Science
2. - Que hacen los data scientists, herramientas y procesos que utilizan
3. - Aplicación de lo aprendido en módulo 2: Uso de Hue
4. Apache Spark
5. - Cómo trabaja Apache Spark y que capacidades nos ofrece
6. - Que formatos de ficheros populares puede usar Spark para almacenar datos
7. - Que lenguajes de programación puedes utilizar para trabajar con Spark
8. - Cómo empezar a utilizar PySpark y Sparklyr
9. - Cómo comparar PySpark y Sparklyr
10. Machine Learning
11. - ¿Qué es machine learning?
12. - Algunos conceptos y términos importantes
13. - Diferentes tipos de algoritmos
14. - Librerías que se utilizan
15. Apache Spark MLlib
16. - Que capacidades de machine learning nos proporciona MLlib
17. - Cómo crear, validar y utilizar modelos de machine learning con MLlib
18. - Ejecución de trabajos Apache Spark
19. - Cómo un trabajo de Spark se compone de una secuencia de transformaciones seguida de una acción
20. - Cómo Spark utiliza la ejecución lenta
21. - Cómo Spark divide los datos entre las particiones
22. - Cómo ejecuta Spark operaciones limitadas y grandes
23. - Cómo Spark ejecuta un trabajo en tareas y fases

UNIDAD DIDÁCTICA 4. DESARROLLO PARA SPARK Y HADOOP

1. Datasets y Dataframes
2. Operaciones en Dataframe
3. Trabajar con Dataframes y Schemas

4. Crear Dataframes a partir de Data Sources
5. Guardar DataFrames en Data Sources
6. DataFrame Schemas
7. Rapidez y lentitud de ejecución
8. Análisis de datos con consultas de DataFrame
9. - Consultar DataFrames con el empleo de expresiones de columna
10. - Agrupación y agregación de consultas
11. - Unión de DataFrames
12. RDD
13. - Introducción RDD
14. - RDD Data Sources
15. - Creando y guardando RDDs
16. - Operaciones con RDDs
17. Transformación de datos con RDDs
18. - Escritura y paso de funciones de transformación
19. - Ejecuciones de transformación
20. - Conversión entre RDDs y DataFrames
21. Agregación de datos con Pair RDDs
22. - Key-Valué Pair RDDs
23. - Mal-Reduce
24. - Otras operaciones Pair RDD
25. Consulta y vistas de tablas con Spark SQL
26. - Datasets y DataFrames
27. - Creación de Datasets
28. - Ejecución y guardado de Datasets
29. - Operaciones de Dataset
30. Creación, configuración y ejecución de aplicaciones Spark
31. - Creación de una aplicación Spark
32. - Compilar y ejecutar la aplicación
33. - Application Deployment Mode
34. - La interfaz Spark Application Web UI
35. - Configuración de las propiedades de la aplicación
36. Procesamiento distribuido
37. - Apache Spark en un Clúster

38. - Particiones RDD
39. - Ejemplo: Particionamiento en consultas
40. - Etapas y Tareas
41. - Planificación de tareas de ejecución
42. Persistencia de datos distribuidos
43. - Persistencia en Datasets y DataFrames
44. - Persistencia en niveles de almacenamiento
45. - Visualización de RDDs persistentes
46. Patrones comunes al procesar datos con Spark
47. - Casos comunes de uso de Spark
48. - Algoritmos de iteración en Apache Spark
49. - Machine Learning
50. Spark Streaming: Introducción a DStreams
51. - Vista general de Spark Streaming
52. - DStreams
53. - Desarrollo de aplicaciones en Streaming
54. Spark Streaming: procesamiento de múltiples lotes
55. - Operaciones Multi-Batch
56. - Time Slicing
57. - Operaciones de estado
58. - Operaciones Sliding Window
59. - Vista previa: Streaming estructurado
60. Apache Spark Streaming: Data Sources
61. - Vista general de Streaming Data Source
62. - Apache Flume y Apache Kafka Data Sources
63. - Ejemplo: uso de un Kafka Direct Data Source

UNIDAD DIDÁCTICA 5. ANÁLISIS DE DATOS

1. Introducción a Pig
2. - ¿Qué es Pig?
3. - Características de Pig
4. - Casos de empleo de Pig
5. - Interacción con Pig

6. Análisis de datos básico con Pig
7. - Sintaxis Pig Latin
8. - Carga de datos
9. - Tipos simples de datos
10. - Definición de campos
11. - Datos de salida
12. - Vistas y esquemas
13. - Filtrado y ordenación de datos
14. - Funciones habituales
15. Procesado de datos complejos con Pig
16. - Formatos de almacenamiento
17. - Tipos de datos complejos y anidados
18. - Agrupaciones
19. - Funciones predefinidas para datos complejos
20. - Iteración de datos agrupados
21. Operaciones con multiconjuntos de datos con Pig
22. - Técnicas para combinar conjuntos de datos
23. - Unión de conjuntos de datos con Pig
24. - Conjunto de operaciones
25. - División de conjuntos de datos
26. Troubleshooting y optimización de Pig
27. - Troubleshooting en Pig
28. - Inicio de sesión
29. - Empleo de UI web Hadoop
30. - Muestreo de datos y depuración
31. - Visión general del rendimiento
32. - Comprensión del plan de ejecución
33. - Consejos para mejorar el rendimiento de Jobs en Pig
34. Introducción a Hive e Impala
35. - ¿Qué es Hive?
36. - ¿Qué es Impala?
37. - ¿Por qué utilizar Hive e Impala?
38. - Schema y almacenamiento de datos
39. - Comparación entre Hive y bases de datos tradicionales

40. - Casos de uso
41. Consultas con Hive e Impala
42. - Tablas y bases de datos
43. - Sintaxis básica en consultas Hive e Impala
44. - Tipos de datos
45. - Empleo de Hue para ejecutar consultas
46. - Empleo de Beeline (la Shell de Hive)
47. - Empleo de la Shell de Impala
48. Administración de datos
49. - Almacenamiento de datos
50. - Creación de bases de datos y tablas
51. - Carga de datos
52. - Alteración de bases de datos y tablas
53. - Simplificación de consultas con vistas
54. - Almacenamiento de resultados de consultas
55. Almacenamiento y datos de rendimiento
56. - Partición de tablas
57. - Carga de datos en tablas particionadas
58. - Cuándo utilizar el particionamiento
59. - Elección de formato de almacenamiento
60. - Gestión de metadatos
61. - Control de acceso a datos
62. Análisis de datos relacional con Hive e Impala
63. - Unión de conjuntos de datos
64. - Funciones predefinidas habituales
65. - Agregaciones y Windowing
66. Datos complejos con Hive e Impala
67. - Datos complejos con Hive
68. - Datos complejos con Impala
69. Análisis de texto con Hive e Impala
70. - Empleo de expresiones regulares
71. - Procesamiento de texto con SerDes en Hive
72. - Análisis de los sentimientos y N•Grams
73. Optimización Hive

74. - Rendimiento de las consultas
75. - Bucketing
76. - Indexación de datos
77. - Hive en Spark
78. Optimización de Impala
79. - Ejecución de consultas
80. - Mejorar el rendimiento de Impala
81. Extendiendo Hive e Impala
82. - Customizar SerDes y formatos de fichero en Hive
83. - Transformación de datos con Scripts personalizados en Hive
84. - Funciones definidas por el usuario
85. - Consultas parametrizadas
86. - Comparación entre MapReduce, Pig, Hive, Impala, y bases de datos relacionales. ¿Cuál elegir?

C/ San Lorenzo 2 - 2
29001 Málaga

Tlf: 952 215 476
Fax: 951 987 941

www.academiaintegral.com.es
E-mail: info@academiaintegral.com.es

ACADEMIA INTEGRAL
CENTRO DE NUEVAS TECNOLOGIAS