IFCD66. Data Scientist (Analista de datos masivos y científico de datos).

Código del curso: PSIFCD66 **Horas:** 310

Descripción del curso:

Acerca de este curso

Cod: PSIFCD66

1. Sistemas de apoyo a la toma de decisiones y gestión de datos

Caracterización de la aplicación del lenguaje Python:

Lenguaje Python.
Ejecución de programas Python.
Objetos en Python.
Tipos numéricos y dinámicos.
Gestión de cadenas de texto: listas, diccionarios, tuplas y ficheros.
Sentencias Python: asignaciones, expresiones e imprimir resultados.
Tests de variables, reglas de sintaxis.
Bucles for y while.

Interpretación la aplicación de protocolos API:

Uso de APIs remotas.
Integración de las aplicaciones con APIs remotas.
Ejemplos de aplicación de APIs remotas en lenguaje Python.

Programación de un algoritmo modular en lenguaje Python:

Programación de módulos.
Fundamentos de programación de clases.
Utilización de APIs e integración con aplicaciones Python.
Distinción de los conceptos Cloud básicos.
Principios de computación en la nube (Cloud Computing).
Ingeniería de servicios: software as a service, Platform as a service, Infrastructure as a Service.
Ejemplos de aplicaciones relevantes en la industria.

Uso de BBDD NoSQL y nuevos modelos de datos (estructurados y no estructurados):

Fundamentos del paradigma NoSQL.
Distribución de los datos y procesamiento en paralelo.
Principales modelos de datos en el mundo NoSQL: clave-valor, orientación a documentos, grafos de propiedad, grafos de conocimiento.

Conocimiento del almacenamiento Big Data y las herramientas de procesamiento masivo:

Aplicaciones basadas en la gestión y el análisis de grandes volúmenes de datos.
Fundamentos arquitectónicos de los sistemas distribuidos.
Principales arquitecturas de referencia.
Nuevos modelos de datos.
Sistemas de ficheros distribuidos.
Document stores.
Bases de datos de grafos.

Evaluación de las metodologías y técnicas aplicadas en la resolución de problemas y justificación de los planteamientos, decisiones y propuestas realizadas:

Sistemas de soporte a la toma de decisiones.
Análisis de los datos: análisis descriptivo, predictivo y prescriptivo.

Casos de uso: gestión y análisis de grandes volúmenes de datos.

Identificación de los factores clave de un problema complejo en el contexto de un proyecto de analítica:

Contexto de la sociedad /economía de los datos y el paradigma de las aplicaciones orientadas a los datos.
Fundamentos de bases de datos relacionales: lenguaje SQL.
Necesidad de un cambio de paradigma: NoSQL. El principio ‘one size does not fit all’.
Principales modelos de datos en el mundo NoSQL: Key-Value, Documento-oriented, Property Graphs y Knowledge Graphs.
Fundamentos arquitectónicos: sistemas distribuidos, escalabilidad, paralelismo. Principales arquitecturas de referencia (shared nothing, shared disk, shared memory).

Distinción y aplicación de los nuevos modelos de datos:

Sistemas de archivos distribuidos: conceptos y principios (distribución, replicación, particionamiento horizontal vs. Vertical, formatos de archivos especializados).
Conocimiento y utilización deHadoop File System (HDFS), Apache Avro, Apache Parquet, Key-value stores: Apache HBase.
Document stores: conceptos y principios (mecanismos de réplica, sharding, consultas espaciales) Inmersión a MongoDB y el Aggregation Framework.
Graph databases: property y knowledge graphs. Conceptos y principios Modelización en grafo, consultas regulares. Introducción a Neo4j y Cypher.
Knowledge graphs. Conceptos y principios: el paradigma open / linked data, RDF y SPARQL. Introducción a GraphDB.

Identificación y análisis de problemas complejos en el área de análisis de datos y planteamiento de soluciones:

Principales conceptos de los flujos de procesamiento de datos en sistemas de gran volumen.
Fases principales de la gestión de grandes volúmenes de datos y retos asociados.

Roles del ingeniero de datos en las fases principales de la gestión de datos.
Limitaciones principales de los modelos tradicionales de gestión de los datos.
Nuevos modelos de datos.

Planificación y ejecución de un trabajo de análisis de datos con una propuesta metodológica:

Definición de un conjunto de datos de partida y una serie de necesidades de negocio que requieran una agregación de los datos, una captura de datos externa, un proceso ETL, análisis de datos y una visualización final de los resultados obtenidos.
Implementación de un sistema de archivos distribuido.
Uso de Hadoop para almacenar un conjunto de datos de actividad de red social. Almacenamiento de un conjunto de datos en un entorno HDFS.
Modelización de grafos: almacenar un conjunto de datos en una base de datos documental u orientada a grafos.

Elección de un repositorio adecuado para los datos del problema y definición de una estrategia de almacenamiento:

Ciclo de vida de los datos: diseño de bases de datos, gestor de los flujos de datos, arquitectura de los sistemas de extracción, carga y transformación de los datos y sistemas de almacenamiento y procesamiento distribuido.
Gestión de los datos: límites del modelo relacional y distribución de los datos.

2. Gestión y procesamiento de datos

Evaluación crítica de las metodologías y técnicas a aplicar en la resolución de problemas y justificación de los planteamientos, decisiones y propuestas realizadas

Fundamentos de gestión de los datos para un proyecto con múltiples fuentes de entrada de datos
Técnicas de organización de modelos de datos desde un punto de vista lógico y físico

Identificación de los flujos de datos y ETL (Extract Transform Load)

Fundamentos de Data Warehousing y Business Intelligence
Conceptos de OLAP y extracción de información
Proceso ETL: extracción, transformación y carga de los datos
Tipos de flujos y operaciones
Data cleaning
Data quality
Ejemplos de aplicaciones

Diseño de un proceso ETL y un modelo de análisis multidimensional.

Modelización multidimensional
DFM: Dimensional Fact Model
Esquema en estrella y derivados
Operadores OLAP
Implementación de cubos y operadores OLAP en entornos relacionales
Herramientas de modelización multidimensional

Diseño de una carga de datos a un repositorio NoSQL y análisis de los datos básico utilizando Spark

Diseño, implementación y mantenimiento de soluciones Fecha Lake. Conceptos y principios (schema-on-write vs. schema-on-read). Modelización y gobernanza de datos
Conceptos y principios de procesamiento distribuido de datos (soluciones declarativas vs. no declarativas)
Modelos de procesamiento distribuido de datos: Basados en disco y basados en memoria principal
MapReduce y a Apache Spark
Procesamiento de datos en tiempo real (streaming). Conceptos y principios (modelos, ventanas

temporales, consultas temporales). Lenguajes de consultas sobre streams. Introducción a herramientas streaming: Apache Kafka, Apache Spark Streaming
Arquitecturas BigData: Lambda, Kappa y orquestadores. Herramientas de gestión de workflows: Apache Airflow

Identificación de los factores clave de un problema complejo en el contexto de un proyecto de analítica.

Proyecto de diseño e implementación ETL con herramientas NoSQL
Proceso de incorporación de datos batch con herramientas Apache.
Análisis de datos y extracción de datos para modelo de negocio a partir del conjunto de datos con Spark
Análisis de datos con Apache Spark
Lectura y exportación de datos
Revisión de la calidad de los datos
Filtros y transformaciones de los datos
Procesamiento de los datos para obtener resúmenes y agrupaciones
Combinaciones, particiones y reformulación de los datos.
Configuración, monitorización y gestión de los errores de las aplicaciones Spark

3. Aprendizaje automático y visualización

Identificación de los fundamentos de análisis de datos y aprendizaje automático (Machine Learning)

Tipología de tareas y algoritmos de aprendizaje (supervisado, no supervisado, semisupervisado)
Métodos principales de aprendizaje
Validación y evaluación de resultados

Distinción de los métodos clasificadores.

Modelos predictivos
Métodos no supervisados. Agrupamiento jerárquico. Agrupamiento particional (k-means y derivados).
Reducción de la dimensionalidad (PCA y otros)
Métodos supervisados. K-NN. Árboles de decisión. SVM. Redes neuronales
Validación y evaluación de resultados

Aplicación de las técnicas de aprendizaje automático y la integración de diversas fuentes de datos

Análisis de sentimientos y polaridad sobre el conjunto de tweets recogidos.
Construcción de un análisis de perfiles mediante el uso de algoritmos de agrupamiento no supervisados (clustering).
Implementación de un análisis de polaridad (sentimiento analysis) sobre el conjunto de mensajes recogidos.
Implementación de dos enfoques alternativos para poder comparar el rendimiento obtenido: Aproximación basada en diccionarios. Aproximación en vectorización (Word2Vec) y uso de un modelo supervisado de aprendizaje automático.

Diseño, desarrollo y evaluación de los métodos de aprendizaje automático.

Procesamiento de datos
Fundamentos de aprendizaje automático
Tipología de tareas y algoritmos de aprendizaje
Validación y evaluación de resultados

Diseño y desarrollo de dashboards.

Principios de visualización de datos.
Diseño de paneles de control y dashboards para definir alarmas y transmitir resultados
Integración de la visualización con herramientas de análisis y consultas de datos
Documentación visual y escrita de los resultados de los proyectos de analítica de datos para audiencias

no especializadas

Utilización de una herramienta de visualización de datos para el diseño y carga de datos a un panel de control

Herramientas de visualización de datos: Grafana, MS PowerBar, Tableau
Visualización de consultas de negocio y panel de control de resultados en herramientas de visualización de datos

Elección, aplicación y evaluación de la calidad de un algoritmo de aprendizaje automático para un problema dado a partir de un conjunto de datos.

Procesamiento de textos (NLP)
Análisis de polaridad basado en diccionarios
Análisis basado en modelos predictivos supervisados
Extracción de características (Word2Vec)

¿Qué aprenderás?

Extraer el conocimiento de utilidad para un propósito en concreto a partir de grandes volúmenes de datos de diferentes fuentes disponibles en formato digital.