Master Big Data y Ciencia de los Datos

Duración: 300 h

Modalidad: Teleformación

Introducción al científico de los datos.

Hace ya algunos años Jeff Hammerbacher y DJ Patil acuñaron el término "científico de datos" para referirse a aquel profesional multidisciplinar capaz de abordar proyectos de extremo a extremo en el ámbito de ciencia de los datos, lo que incluye almacenar y limpiar grandes volúmenes de datos, explorar conjuntos de datos para identificar información potencialmente relevante, construir modelos predictivos y construir una historia alrededor de los hallazgos derivados de tales modelos. Un científico de datos es capaz de definir modelos estadísticos y matemáticos que sean de aplicación en el conjunto de datos, lo que incluye aplicar su conocimiento teórico en estadística y algoritmos para encontrar la solución óptima a un problema.

En la actualidad el científico de datos representa uno de los perfiles profesionales con más demanda en el ámbito empresarial, institucional y gubernamental alrededor del mundo, así como una de las profesiones en el entorno de big data y data science mejor pagadas. La creciente demanda de tales profesionales hace necesaria la creación de espacios de formación dirigidos al diseño y puesta en marcha de itinerarios formativos integrales que cubran todas aquellas áreas de conocimiento asociadas a este perfil profesional.

Con el fin de cubrir esta necesidad hemos diseñado el "Plan de formación integral para el científico de datos" basado en el conocido como "Metromap" de Swami Chandrasekaran.

Metodología: El plan integral de formación que presentamos se basa en una metodología 100% online. Los niveles de habilidades se presentan mediante casos prácticos y casos de uso asociados a situaciones reales. Los alumnos disponen de tutoría online durante un período de 12 meses a partir de su

matriculación, así como actualizaciones de contenidos gratuitas durante ese periodo.

Los contenidos se han preparado para su presentación progresiva durante un período de 3 meses, si bien los alumnos dispondrán de acceso a los contenidos sin límite de tiempo. En cada nivel se presentará una prueba inicial o pre-evaluación y finalizará con una evaluación que deberá superar con una puntuación al menos del 80% para pasar al siguiente nivel.

Los contenidos se presentarán mediante video tutoriales, documentos teóricos para descarga y contenidos interactivos.

TEMARIO

Tema 1. FUNDAMENTOS

1. ¿Qué es ciencia de datos?
2. Matemáticas para la ciencia de datos. Introducción al álgebra lineal y matrices
3. Fundamentos de bases de datos y álgebra relacional
4. Ciclo de vida de un proyecto de ciencia de datos
5. Teorema CAP
6. Tratamiento de datos
7. OLAP
8. Modelos de datos multidimensionales
9. ETL. Extracción, transformación y carga
10. Reporting vs Business Intelligence vs Analytics
11. Formatos de intercambio de datos. JSON y XML
12. Introducción a las bases de datos NoSQL
13. Panorama de soluciones de bases de datos

Tema 2.ESTADÍSTICA

1. Conceptos fundamentales de estadística
2. Histogramas
3. Teoría de probabilidades
4. Teorema de Bayes
5. Variables aleatorias
6. Distribución de probabilidades. CDF
7. Distribuciones continuas
8. Asimetría estadística
9. Análisis de varianza. ANOVA

10. Función de densidad de probabilidad
11. Teorema del límite central
12. Método Monte Carlo
13. Regresión
14. Covarianza
15. Correlación. Coeficiente de correlación de Pearson
16. Causation (Causalidad)
17. Least Squares Fitting
18. Distancia euclídea

Tema 3.PROGRAMACIÓN

1. Programación Python para ciencia de datos
2. Programación R para ciencia de datos

Tema 4.MACHINE LEARNING Y DEEP LEARNING

1. ¿Qué es Machine Learning?
2. Variables numéricas y categóricas. Datos cualitativos y cuantitativos
3. Tipos de aprendizaje:
4. Conceptos. Inputs y atributos.
5. Datos de entrenamiento y prueba
6. Clasificadores (Classifier)
7. Predicción
8. Lift. Overfitting
9. Bias y Varianza
10. Árboles y clasificación
11. Ratio de clasificación
12. Árboles de decisión
13. Boosting
14. Clasificadores Naive-Bayes
15. K-Nearest Neighbor
16. Regresión
17. Clustering
18. Redes neuronales

19. ¿Qué es Deep Learning?
20. Análisis de sentimientos
21. Filtrado colaborativo
22. Tagging (Etiquetado)

Tema 5.TEXT MINING Y NLP

1. ¿Qué es text mining y procesamiento de lenguaje natural (NLP)?
2. Corpus
3. Named Entity Recognition (Reconocimiento de Entidades Nombradas)
4. Análisis de textos
5. UIMA. Aplicaciones para la administración de información no estructurada
6. Matriz de términos del documento
7. Frecuencia y peso de términos
8. SupportVector Machines. Máquina de Vectores de Soporte
9. Reglas de asociación
10. Market BasketAnalysis
11. Extracción de características
12. Apache Mahout
13. Weka
14. NLTK. Natural LanguageToolKit
15. Clasificación de texto
16. Mapeo de vocabulario

Tema 6.VISUALIZACIÓN

1. Exploración de datos con R
2. VisualizacionUnivariate, Bivariate y Multivariate
3. ggplot2
4. Tipos de gráficos
5. D3.js
6. Tableau

Tema 7. BIG DATA

1. Fundamentos de Big Data
2. Map Reduce
3. Hadoop y su ecosistema
4. HDFS
5. Principios de la replicación de datos
6. Distribuciones. Hadoop con Cloudera y HortonWorks
7. Arquitectura de un clúster Hadoop. Nodos. NameNode y DataNodes
8. Job Tracker y TaskTracker
9. Programación de tareas MapReduce
10. YARN
11. Carga de datos en HDFS con Sqoop
12. Carga de datos no estructurados con Flume
13. SQL con PIG
14. Data Warehouse con Hive
15. Weblog con Chukwa
16. Apache Mahout
17. Serialización e intercambio de datos para Hadoop con Avro
18. Coordinación de servicios en el clúster con Zookeeper
19. Hadoop en tiempo real con Storm
20. R y Hadoop
21. Conceptos básicos de Cassandra
22. Introducción a MongoDB
23. Introducción a Neo4.js
24. Apache Spark

Tema 8. DATA INGESTION

1. ¿Qué es Data Ingestion?
2. Formatos de datos
3. Data Discovery
4. Fuentes de datos y Adquisición
5. Data integration y Data Fusion
6. Google OpenRefine

7. Herramientas ETL

Tema 9 DATA MUNGING

1. ¿Qué es Data Munging
2. Dimensionality & Numerosity Reduction
3. Normalización
4. Data Scrubbing
5. Tratamiento de valores perdidos (MissingValues)
6. Unbiased Estimators
7. Binning Sparse NumericalValues
8. Extracción de características
9. Denoising
10. Samplig
11. Stratified Sampling
12. Principal ComponentAnalysis

Tema 10. HERRAMIENTAS

1. Microsoft Excel como herramienta para la analítica
2. Elastic Stack
3. Apache Impala / Kudu