



# *Título de Máster en Big Data y Data Science: Ciencia e Ingeniería de Datos*

---

Título Propio de la Universidad Autónoma de Madrid

**Escuela Politécnica Superior**

**Universidad Autónoma de Madrid**

## Estructura

Fundamentos: sistemas y arquitecturas (Online)	3 ECTS
Fundamentos: lenguajes (Online)	3 ECTS
Estadística	3 ECTS
Ciclo de vida analítico del dato	4 ECTS
Ecosistema Spark	3 ECTS
Infraestructura de Big Data	6 ECTS
Fundamentos de análisis de datos	5 ECTS
Seguridad en Big Data, privacidad y protección de datos	3 ECTS
Fuentes de datos y aprovisionamiento	4 ECTS
Bases de datos NoSQL (Online)	3 ECTS
Explotación y visualización	3 ECTS
Tecnologías en Inteligencia Artificial	4 ECTS
Aplicaciones de IA en texto, imagen y audio	4 ECTS
<b>Trabajo Fin de Máster</b>	12 ECTS
	60 ECTS

## Programa

### Fundamentos: sistemas y arquitecturas (ONLINE)

1. Sistema Operativo Linux
  - 1.4. Conceptos generales de Linux
  - 1.5. Comandos, variables de entorno, scripts
  - 1.6. Control y planificación de procesos.
  - 1.7. Sistemas de almacenamiento y sistemas de ficheros
  - 1.8. Administración básica de Linux
2. Redes de comunicación
  - 2.1. Componentes y tipos de redes.
  - 2.2. Conceptos básicos: Direcciónamiento IP, máscara de red, puerta de enlace, servidor de nombres (DNS), direccionamiento dinámico DHCP.
  - 2.3. Configuración de una red TCP/IP en Linux.
  - 2.4. Acceso remoto a equipos y ficheros: ssh, ftp
  - 2.5. Conceptos básicos de seguridad: Claves pública y privada, VPN.
3. Máquinas Virtuales
  - 3.1. Concepto de virtualización
  - 3.2. Tipos de virtualización de plataforma
  - 3.3. Instalación y gestión de una Máquina Virtual
  - 3.4. Creación de máquinas virtuales con Vagrant
  - 3.5. Infraestructura como Servicio (IaaS): máquinas virtuales bajo demanda y con capacidades actualizables en tiempo real
4. Cluster de ordenadores
  - 4.1. Multicomputador
  - 4.2. Clusters de ordenadores
  - 4.3. Construir, desplegar y gestionar un cluster
  - 4.4. Planificación y balanceo de tareas.
  - 4.5. Tipos de Cluster y aplicaciones: HPC, Hadoop

### Fundamentos: lenguajes (ONLINE)

1. Python
  - 1.1. Introducción a python
    - 1.1.1. Instalación
    - 1.1.2. Intérpretes: python, ipython, notebooks
    - 1.1.3. Diferencias entre 2.7 y 3.0
  - 1.2. Tipos básicos: cadenas, listas, diccionarios, tuplas, etc.
  - 1.3. Funciones, funciones lambda e imports.
  - 1.4. Sentencias de control e iteración
    - 1.4.1.1. Loops e ifs
    - 1.4.1.2. Algunas formas de utilizar programación funcional: map, reduce.
  - 1.5. Entrada y salida de ficheros
  - 1.6. Programación orientada a objetos en python
  - 1.7. Librerías: numpy, matplotlib, pandas, etc.

2. R
  - 2.1. Introducción a R
  - 2.2. Objetos y atributos en R
  - 2.3. Vectores
  - 2.4. Arrays
  - 2.5. Listas
  - 2.6. Data frames
  - 2.7. Lecturas de ficheros
  - 2.8. Funciones y sentencias de control
  - 2.9. Gráficas
  - 2.10. Depuración y medición de tiempos
  - 2.11. Notebooks

### **Bases de datos NoSQL (ONLINE)**

1. Introducción a las bases de datos NoSQL
  - 1.1. ¿Qué son?
  - 1.2. Tipos de BBDD NoSQL
  - 1.3. Ventajas y desventajas
2. Base de Datos MongoDB
  - 2.1. Introducción
  - 2.2. Organización de los datos
  - 2.3. Manejo básico de los datos
  - 2.4. Métodos básicos de agregación
  - 2.5. MapReduce
  - 2.6. Aggregation Framework
  - 2.7. Uso de índices
3. Base de Datos Redis
  - 3.1. Estructuras de datos
  - 3.2. Programación en Lua
  - 3.3. Bibliotecas de Lua y depuración de scripts Lua
4. Base de Datos Cassandra
  - 4.1. Introducción
  - 4.2. Cassandra Query Language
5. Base de datos Neo4j
  - 5.1. Introducción
  - 5.2. Lenguaje de consulta Cypher

### **Ciclo de vida analítico del dato**

1. Proyecto Apache Hadoop
  - 1.1. HDFS
  - 1.2. Modelo de Programación MapReduce
  - 1.3. Desarrollo con Pig
2. Persistencia
  - 2.1. Persistencia en entornos Hadoop:
    - 2.1.1. Hive
    - 2.1.2. Hbase
  - 2.2. Otros formatos de persistencia: Avro, Parquet, ORC
3. Análisis
  - 3.1. Análisis de datos estáticos en entorno Hadoop
  - 3.2. Análisis de datos en vuelo: Storm
4. Explotación
  - 4.1. Buscadores: SolR, ElasticSearch
  - 4.2. Visualizadores asociados: Kibana

5. Arquitectura en entornos Big Data
  - 5.1. Terminología y Conceptos
  - 5.2. Arquitecturas de Referencia
  - 5.3. Diseño y representación de arquitecturas

## **Estadística**

1. Tema 1: Introducción
  - 1.1. ¿Qué es la estadística?
  - 1.2. Modelo estadístico
  - 1.3. Método estadístico
  - 1.4. Algunas herramientas de análisis de datos mediante estadística
2. Tema 2: Descripción de los datos
  - 2.1. Descripción de una variable
  - 2.2. Descripción multivariante
3. Tema 3: Modelos en estadística
  - 3.1. Probabilidad y variables aleatorias
  - 3.2. Modelos univariantes de distribución de probabilidad
  - 3.3. Modelos multivariantes de distribución de probabilidad
4. Tema 4: Inferencia Estadística
  - 4.1. Estimación puntual
  - 4.2. Estimación por intervalos
  - 4.3. Estimación bayesiana
  - 4.4. Contraste de hipótesis

## **Fundamentos de análisis de datos**

1. Introducción al aprendizaje automático
  - 1.1. Tipos de aprendizaje automático, conceptos básicos, tipos de atributos
  - 1.2. Flujo de un proyecto de aprendizaje automático
  - 1.3. Validación de modelos: tasas de error, matriz de confusión, curvas ROC y validación cruzada
  - 1.4. Regresión lineal, regresión logística
  - 1.5. Vecinos próximos en clasificación y regresión
  - 1.6. Sesgo y varianza. Maldición de la dimensionalidad
2. Preprocesado de datos
  - 2.1. Construcción de la base de datos; tratamiento de múltiples fuentes
  - 2.2. Preparación y auditoría de la base de datos
  - 2.3. Distribución de las variables.
  - 2.4. Reducción de la dimensionalidad
  - 2.5. Información no estructurada; casos prácticos.
3. Aprendizaje automático
  - 3.1. Clasificación y regresión con máquinas de vectores soporte
  - 3.2. Conjuntos de clasificadores y árboles de decisión
  - 3.3. Clustering

## **Ecosistema Spark**

1. Fundamentos de Spark
  - 1.1. Introducción: arquitectura y organización
  - 1.2. Datos en Spark: Resilient Distributed Datasets (RDDs)
  - 1.3. Flujo de un programa spark
  - 1.4. Entrada y salida de datos
  - 1.5. Transformaciones
  - 1.6. Persistencia
  - 1.7. Acciones

- 1.8. Variables compartidas: broadcast y acumuladores
- 2. Tuning en Spark
  - 2.1. Vista general de las APIs ofrecidas por Spark: Scala, Java, Python, R
  - 2.2. SparkR: paralelización de DataFrames de R
- 3. Spark SQL
  - 3.1. Introducción a DataFrames
  - 3.2. Fuentes de datos: Hive, JDBC/ODBC, etc.
  - 3.3. API de DataFrames
- 4. Procesamiento de grafos vía Spark
  - 4.1. Introducción general a las operadores sobre grafos
  - 4.2. Grafos en Spark: GraphX
  - 4.3. Paquetes adicionales para Spark: GraphFrames
  - 4.4. Algoritmos de grafos sobre GraphFrames
- 5. Procesado en tiempo real: Spark Streaming
  - 5.1. Spark Streaming clásico: Discretized Streams (DStreams)
  - 5.2. Operaciones con DStreams: estado, robustez, ventanas
  - 5.3. Streaming estructurado
  - 5.4. Operaciones sobre streaming estructurado: flujos, ventanas, entregas
  - 5.5. Fuentes de datos para streaming: Kafka
  - 5.6. Machine Learning sobre datos en streaming
- 6. Machine Learning: Spark ML
  - 6.1. Aprendizaje supervisado: clasificación y regresión
  - 6.2. Aprendizaje no supervisado
  - 6.3. Creación de pipelines de aprendizaje automático

## **Explotación y visualización**

- 1. Introducción, importancia de la visualización
- 2. Visualización de datos
  - 2.1 Cantidad
  - 2.2 Distribuciones
  - 2.3 Proporciones
  - 2.4 Asociaciones variables cuantitativas
  - 2.5 Series temporales
  - 2.6 Tendencias
  - 2.7 Datos geoespaciales
- 3. Herramientas de visualización
  - 3.1 CartoDB
  - 3.2 Watson Analytics
  - 3.3 MapBox
- 4. Tableau
  - 4.1 Acceso a datos desde Tableau
  - 4.2 Hojas y Dashboards
  - 4.3 Patrones temporales
  - 4.4 Información espacial y geográfica
  - 4.5 Gráficos interactivos, filtros
- 5. Grafana
  - 5.1 Conceptos básicos: roles, datasources y dashboards
  - 5.2 Templating y automatización
  - 5.3 Alerting y plugins

## Infraestructura para Big Data

1. Arquitecturas para tratar grandes volúmenes de información
  - 1.1. Arquitecturas de referencia para Hadoop
  - 1.2. Instalación y configuración de un cluster Hadoop
  - 1.3. Uso de herramientas de planificación y gestión.
  - 1.4. Hadoop Hortonworks
2. Supervisión y mantenimiento de un cluster para Big Data
  - 2.1. Estado de HDFS y copia de datos
  - 2.2. Añadir y quitar nodos
  - 2.3. Balanceo del cluster
  - 2.4. Impacto de la red de comunicación en un cluster Big Data.
  - 2.5. Copias de seguridad
3. Evaluación de prestaciones y optimización en un caso práctico
  - 3.1. Benchmarking y tuneado de parámetros
  - 3.2. Caso práctico: procesamiento Big Data de datos de red
4. Infraestructura para otros entornos Big Data: Ecosistema Spark
  - 4.1. Arquitectura de un sistema Spark. Estructura interna y flujos de datos
  - 4.2. Modos de ejecución en Spark: local vs cluster. Ámbitos y contextos: driver, ejecutores
  - 4.3. Gestión de memoria
  - 4.4. Modos de ejecución en cluster. Comunicación entre nodos.
  - 4.5. Ciclo de vida de un programa Spark
  - 4.6. Configuración
  - 4.7. Ejecución de tareas: spark-submit (dependencias, etc), REPL (spark-shell, pyspark), notebooks
  - 4.8. Acceso a datos: E/S (local, HDFS, etc).
  - 4.9. Interfaces de monitorización y análisis: Spark UI, Spark Master, Spark History Server
5. Virtualización de infraestructura
  - 5.1. Infraestructura local vs Cloud
  - 5.2. Infraestructura como Servicio (IaaS)
  - 5.3. Cloud privado: Propuestas Openstack y OpenNebula
  - 5.4. Cloud público: Propuestas de IBM Softlayer, Amazon EC2, Rackspace, Google Cloud y Microsoft Azure
  - 5.5. Cloud público vs Cloud privado
  - 5.6. Prácticas: Despliegues Cloud privado
  - 5.7. Prácticas: Monitorizando en el Cloud público
  - 5.8. Prácticas: Comparativa de rendimiento en el Cloud
6. Virtualización basada en contenedores
  - 6.1. Diseño de aplicaciones en contenedores
  - 6.2. Gestión de imágenes y versiones
  - 6.3. Orquestación y Comunicación
  - 6.4. Seguridad
7. Plataformas como servicio (PaaS): IBM Cloud
  - 7.1. Concepto de Plataforma como Servicio
  - 7.2. Almacenamiento en entornos Cloud
  - 7.3. Utilidades y nuevos modelos de consumo de servicios
  - 7.4. Ejercicios prácticos con IBM Cloud
8. Tendencias HW y SW para Big Data: computación cuántica

## Seguridad en Big Data, privacidad y protección de datos

1. Introducción al concepto de privacidad
2. Tecnologías criptográficas para la protección de la privacidad
  - 2.1. Fundamentos criptográficos de la protección de la información
  - 2.2. Problema de la gestión de la identidad digital mediante certificados digitales: el estándar X.509
  - 2.3. Definición de los conceptos de trazabilidad, enlazado, anonimato y pseudo-anonimato: firmas grupales
  - 2.4. Navegación anónima: introducción a las redes de mezcla de tráfico, *onion routing* y ofuscación de tráfico
3. Privacidad como control estadístico del acceso a datos
  - 3.1. Descripción de datos y metadatos que permiten la re-identificación de individuos
  - 3.2. Valor y riesgo asociado a los datos y metadatos accesibles en abierto (Open Data)
4. Marco jurídico de protección de datos y normas de transferencia internacionales
  - 4.1. Introducción al Reglamento General de Protección de Datos (RGPD).
  - 4.2. Privacy Shield Framework (transferencia de datos entre EE.UU. y UE).
  - 4.3. Privacy Directive ('cookie law')
5. Implicaciones ético-legales de la inclusión de Big Data en la toma de decisión: gobernanza y rendición de cuentas (8 horas)
  - 5.1. La gobernanza de los algoritmos como problema epistémico vs. moral o como problema técnico vs. de gestión o legal.
  - 5.2. Actual marco legal europeo.
  - 5.3. Introducción al debate sobre Algorithmic Fairness, Accountability and Transparency (FAT\*) a través de distintos casos de estudio.
  - 5.4. Consideraciones sobre el conjunto de datos y la selección de modelos de aprendizaje automático: riesgos de manipulación, problemas de predicción y path dependency.
6. Sesgos en IA

## Tecnologías en Inteligencia Artificial

1. Redes neuronales clásicas
2. Redes neuronales profundas
3. TensorFlow y Keras
  - 3.1. Posicionamiento de TensorFlow en el mercado: Introducción y casos de éxito
  - 3.2. Arquitectura: Tensores, Grafos, Diagrama de flujo de datos
  - 3.3. TensorFlow 2.0 y Keras
  - 3.4. Aplicaciones prácticas utilizando TensorFlow - Keras: perceptron, redes neuronales y CNN
  - 3.5. Técnicas de reducción de overfitting
4. Predicción de energías renovables con series temporales
  - 4.1. Introducción
  - 4.2. Análisis de series temporales: visualización, tendencia, estacionalidad y estacionariedad.
  - 4.3. Predicción de energías renovables usando modelos clásicos autorregresivos: AR, MA, ARMA, ARIMA, SARIMA.
  - 4.4. Predicción de energías renovables usando modelos de series temporales con entradas exógenas.
5. Impacto social: computación cognitiva
  - 5.1. Bases de la Computación Cognitiva. IBM Watson.
  - 5.2. Estrategia de Soluciones Cognitivas de IBM.
  - 5.3. Infraestructura para Soluciones Cognitivas.
  - 5.4. Servicios cognitivos a través de IBM Bluemix.

## **Fuentes de datos y aprovisionamiento**

1. Fuentes de datos y descubrimiento
  - 1.1. Internet de las cosas
    - 1.1.1. Conceptos y Escenarios.
    - 1.1.2. IoT Foundation en IBM Cloud.
  - 1.2. Industria 4.0
    - 1.2.1. Conceptos y Escenarios
    - 1.2.2. Casos de uso y de negocio
  - 1.3. Reconocimiento biométrico
    - 1.3.1. Introducción al Reconocimiento Biométrico de Personas
    - 1.3.2. Huella dactilar e iris
    - 1.3.3. Reconocimiento Facial y Particularidades del Reconocimiento Biométrico Conductual
    - 1.3.4. Prácticas de Reconocimiento Biométrico
  - 1.4. Big Data en Biomedicina
    - 1.4.1. Diversidad e integración de información biomédica
    - 1.4.2. Disponibilidad y acceso a bases de datos biomédicas
    - 1.4.3. Herramientas básicas de tratamiento masivo de datos para el diagnóstico y prevención de enfermedades
    - 1.4.4. Dispositivos móviles y big-data biomédico
2. Aprovisionamiento
  - 2.1. Tecnologías para captura y modificación de datos:
    - 2.1.1. Ecosistema Hadoop (Sqoop, Flume)
    - 2.1.2. Apache Nifi
  - 2.2. Otros ecosistemas: Apache Kafka

## **Aplicaciones de IA en texto, imagen y audio**

1. Multimedia (imagen, video)
  - 1.1. Introducción al tratamiento de imagen y vídeo
  - 1.2. Extracción de características en señales visuales: descriptores en imagen y vídeo
  - 1.3. Descriptores globales: color, puntos de interés
  - 1.4. Descriptores a nivel de región segmentada: color, puntos de interés, textura, forma
  - 1.5. Descriptores de movimiento: movimiento global, trayectorias
  - 1.6. Aplicaciones en imágenes I: búsqueda global por color y puntos de interés
  - 1.7. Aplicaciones en imágenes II: búsqueda en imágenes segmentadas
  - 1.8. Aplicaciones en vídeo
2. Multimedia (audio)
  - 2.1. Extracción de características en señal de voz: detección de voz, tono fundamental, espectro y envolvente espectral, espectrogramas
  - 2.2. Extracción de características en señal musical: detección de notas musicales simultáneas (multipitch), ritmo, armonía, cromagramas, detección de música, localización de fragmentos musicales
  - 2.3. Aplicaciones en voz I: reconocimiento de voz y detección de palabras clave
  - 2.4. Aplicaciones en voz II: detección de hablante, detección de idioma, reconocimiento de emociones
  - 2.5. Aplicaciones sobre audio broadcast: segmentación y separación de hablantes y eventos acústicos, búsquedas, indexación y búsquedas sobre audio.
3. Análisis de textos
  - 3.1. Introducción al procesado de texto
  - 3.2. Análisis de caracteres: frecuencia, complejidad de compresión
  - 3.3. Análisis de tokens: tokenización, n-gramas, stop-words
  - 3.4. Análisis léxico: lexemas, categorías gramaticales, reglas generativas

- 3.5. Análisis sintáctico: árboles de parsing, gramáticas formales y probabilísticas
- 3.6. Análisis semántico: ontologías, wordnet, embeddings semánticos
- 3.7. Aplicaciones del análisis de texto
- 4. Procesamiento de Lenguaje Natural
  - 4.1. Introducción al PLN y lingüística computacional
  - 4.2. Tokenización, splitter, lematización, Pos Tagging
  - 4.3. Ambigüedad léxica y semántica
  - 4.4. Pipelines de PLN: freeling y Spacy
  - 4.5. Análisis semántico
  - 4.6. Aplicaciones basadas en PLN
  - 4.7. Herramientas de análisis del sentimiento
  - 4.8. Herramientas de detección de entidades
  - 4.9. Herramientas de Chatbot
  - 4.10. Herramientas de Speech2text