MÁSTER DATA ENGINEERING ONLINE

DURACIÓN: 300 Horas

Modalidad: Online Teleformación

Público

Este máster está orientado a quienes desean trabajar con datos de forma profesional y necesitan comprender cómo se construyen, procesan y escalan los sistemas modernos de información. Máster orientado a perfiles técnicos, analistas, desarrolladores, científicos de datos en formación y profesionales que buscan especializarse en la creación de pipelines, el manejo de grandes volúmenes de datos y la integración de herramientas cloud. También resulta útil para quienes provienen de áreas de negocio y quieren adquirir una base sólida en ingeniería de datos.

Objetivos

El Máster Data Engineering tiene como objetivo ofrecer una visión completa del trabajo de un ingeniero de datos, desde la programación en Python hasta la gestión de bases de datos, modelos analíticos y plataformas cloud. Aprenderás a diseñar flujos de datos, optimizar procesos, gestionar orígenes diversos y crear modelos que permitan análisis fiables. Además, desarrollarás competencias prácticas en herramientas como SQL, Pandas, Power BI y Snowflake, integrando conceptos de automatización, calidad de datos y preparación para certificaciones.

TEMARIO

Módulo 0. Presentación al máster

Módulo 1. Fundamentos de Programación en Python para Data Engineering

1. **INTRODUCCIÓN A PYTHON APLICADO A LA INGENIERÍA DE DATOS**
2. **ENTORNO DE DESARROLLO Y BUENAS PRÁCTICAS (VIRTUAL ENVIRONMENTS, ESTRUCTURA DE PROYECTOS)**
3. **TIPOS DE DATOS, ESTRUCTURAS BÁSICAS Y CONTROL DE FLUJO**
4. **FUNCIONES, MÓDULOS Y REUTILIZACIÓN DE CÓDIGO**
5. **MANEJO DE FICHEROS Y SISTEMAS DE ARCHIVOS (CSV, JSON, PARQUET)**
6. **INTRODUCCIÓN A PANDAS PARA MANIPULACIÓN DE DATOS**
7. **LECTURA, LIMPIEZA Y TRANSFORMACIÓN BÁSICA DE DATOS**
8. **CONSUMO DE APIS Y MANEJO DE DATOS EXTERNOS**
9. **AUTOMATIZACIÓN DE TAREAS Y SCRIPTING**
10. **GESTIÓN DE ERRORES Y EXCEPCIONES**
11. **LOGGING Y TRAZABILIDAD DE PROCESOS**
12. **INTRODUCCIÓN A PRUEBAS BÁSICAS (TESTING)**
13. **BUENAS PRÁCTICAS DE CÓDIGO PARA PIPELINES PRODUCTIVOS**

MÓDULO 2. Modelado de Datos y SQL avanzado

1. **FUNDAMENTOS DE SQL**
    a. Estructura de una consulta SQL
    b. Selección, filtrado y ordenación de datos

   c. Operaciones de agregación
2. **CONSULTAS SQL COMPLEJAS**
   a. Joins (INNER, LEFT, RIGHT, FULL)
   b. Self joins y joins sobre múltiples tablas
   c. Subconsultas correlacionadas y no correlacionadas
   d. Common Table Expressions (CTE)
3. **FUNCIONES ANALÍTICAS Y WINDOW FUNCTIONS**
   a. Uso de OVER, PARTITION BY y ORDER BY
   b. Funciones de ranking y cálculo acumulado
   c. Análisis temporal y comparativas entre registros
4. **TRANSFORMACIÓN DE DATOS CON SQL**
   a. Limpieza y enriquecimiento de datos
   b. Cálculos derivados y métricas
5. **OPTIMIZACIÓN DE CONSULTAS SQL**
   a. Uso de índices
   b. Análisis de planes de ejecución
   c. Buenas prácticas de rendimiento
6. **MODELADO DE DATOS RELACIONAL**
   a. Diseño de esquemas y relaciones
   b. Normalización y desnormalización
   c. Claves primarias y foráneas
7. **MODELADO DIMENSIONAL**
   a. Conceptos de hechos y dimensiones
   b. Definición de granularidad
   c. Medidas y atributos
8. **ESQUEMA ESTRELLA**
9. **ESQUEMA COPO DE NIEVE**
10. **BUENAS PRÁCTICAS DE MODELADO ORIENTADO A ANALÍTICA Y REPORTING**
11. **CASOS PRÁCTICOS DE DISEÑO Y CONSULTA SOBRE MODELOS DE DATOS**

MÓDULO 3. Bases de Datos NoSQL y Modelos No Relacionales

1. **INTRODUCCIÓN A LAS BASES DE DATOS NOSQL**
2. **LIMITACIONES DE LOS MODELOS RELACIONALES TRADICIONALES**
3. **NOSQL VS SQL**
   a. Casos de uso
   b. Escalabilidad horizontal y vertical
   c. Consistencia, disponibilidad y tolerancia a particiones (CAP)
4. **TIPOS DE BASES DE DATOS NOSQL**
   a. Documentales
   b. Clave-valor

   c. Columnas
   d. Grafos
5. **FUNDAMENTOS DE MODELADO NOSQL**
   a. Diseño orientado a consultas
   b. Desnormalización y duplicación controlada
6. **INTRODUCCIÓN A MONGODB**
   a. Arquitectura y conceptos básicos
   b. Colecciones y documentos
   c. Tipos de datos
7. **CONSULTAS EN MONGODB**
   a. Filtros, proyecciones y ordenación
   b. Operadores
8. **FRAMEWORK DE AGREGACIÓN EN MONGODB**
9. **ÍNDICES Y RENDIMIENTO EN MONGODB**
10. **INTRODUCCIÓN A CASSANDRA**
    a. Arquitectura distribuida
    b. Modelo de datos basado en columnas
    c. Particiones y clustering
11. **MODELADO DE DATOS EN CASSANDRA**
12. **CONSULTAS Y PATRONES DE ACCESO EN CASSANDRA**
13. **CASOS DE USO REALES DE MONGODB Y CASSANDRA**
14. **INTEGRACIÓN DE BASES DE DATOS NOSQL EN PIPELINES DE DATOS**

MÓDULO 4. Procesamiento Distribuido de Datos con Apache Spark

1. **INTRODUCCIÓN AL PROCESAMIENTO DISTRIBUIDO DE DATOS**
2. **ARQUITECTURA DE APACHE SPARK**
   a. Driver, executors y cluster manager
   b. Modos de ejecución (local, cluster)
3. **CONCEPTOS FUNDAMENTALES DE SPARK**
   a. RDDs: características y casos de uso
   b. DataFrames y Datasets
4. **OPERACIONES EN SPARK**
   a. Transformaciones y acciones
   b. Lazy evaluation
5. **SPARK SQL**
   a. Ejecución de consultas SQL sobre DataFrames
   b. Integración con fuentes de datos
6. **PROCESAMIENTO BATCH CON SPARK**
   a. Lectura y escritura de grandes volúmenes de datos

   b. Transformaciones complejas y agregaciones
7. **OPTIMIZACIÓN BÁSICA EN SPARK**
   a. Particionado y paralelismo
   b. Caché y persistencia
8. **INTRODUCCIÓN AL PROCESAMIENTO EN STREAMING**
   a. Conceptos de Spark Structured Streaming
   b. Casos de uso habituales
9. **CASOS PRÁCTICOS DE PROCESAMIENTO DE GRANDES VOLÚMENES DE DATOS**
   a. Transformaciones batch
   b. Análisis y agregaciones sobre datasets masivos

MÓDULO 5. Módulo 5. Ingesta y Streaming de Datos con Apache Kafka

1. **INTRODUCCIÓN A LAS ARQUITECTURAS ORIENTADAS A EVENTOS**
2. **CONCEPTOS FUNDAMENTALES DE APACHE KAFKA**
   a. Brokers y cluster
   b. Zookeeper / modo KRaft (visión general)
3. **TOPICS, PARTICIONES Y REPLICACIÓN**
   a. Distribución y paralelismo
   b. Garantías de entrega
4. **PRODUCTORES Y CONSUMIDORES**
   a. Publicación de eventos
   b. Consumo y procesamiento de mensajes
5. **SERIALIZACIÓN DE DATOS**
   a. Formatos comunes (JSON, Avro)
6. **GESTIÓN DE OFFSETS Y GRUPOS DE CONSUMIDORES**
7. **KAFKA CONNECT**
   a. Arquitectura y conectores
   b. Integración con bases de datos y sistemas externos
8. **INTRODUCCIÓN AL CHANGE DATA CAPTURE (CDC)**
   a. Concepto y casos de uso
   b. Integración con Kafka
9. **CASOS DE USO DE STREAMING DE DATOS**
   a. Procesamiento en tiempo real
   b. Integración con sistemas batch
10. **BUENAS PRÁCTICAS Y PATRONES DE DISEÑO CON KAFKA**

MÓDULO 6. Orquestación y Automatización de Pipelines de Datos

1. **INTRODUCCIÓN A LA ORQUESTACIÓN DE PIPELINES DE DATOS**
2. **DISEÑO DE PIPELINES DE DATOS**
    a. Pipelines batch y near-real-time
    b. Buenas prácticas de diseño
3. **CONCEPTOS ETL VS ELT**
    a. Diferencias y casos de uso
    b. Impacto en arquitecturas modernas de datos
4. **INTRODUCCIÓN A HERRAMIENTAS DE ORQUESTACIÓN**
    a. Apache Airflow o herramienta equivalente
5. **CONCEPTOS CLAVE DE ORQUESTACIÓN**
    a. DAGs
    b. Tareas y operadores
    c. Dependencias entre tareas
6. **PLANIFICACIÓN Y SCHEDULING DE PROCESOS**
7. **GESTIÓN DE ERRORES Y REINTENTOS**
8. **MONITORIZACIÓN Y CONTROL DE EJECUCIONES**
9. **BUENAS PRÁCTICAS PARA PIPELINES PRODUCTIVOS**
10. **CASOS PRÁCTICOS DE AUTOMATIZACIÓN DE FLUJOS DE DATOS**

MÓDULO 7. Plataformas Cloud de Datos y Data Warehousing (Snowflake)

1. **INTRODUCCIÓN A LAS ARQUITECTURAS CLOUD DE DATOS**
2. **CONCEPTOS DE DATA WAREHOUSE MODERNO**
3. **ARQUITECTURA DE SNOWFLAKE**
    a. Almacenamiento, compute y servicios
4. **CARGA DE DATOS EN SNOWFLAKE**
    a. Stages internos y externos
    b. Formatos de datos
5. **SQL analítico en Snowflake**
    a. Consultas complejas y transformaciones
6. **OPTIMIZACIÓN Y RENDIMIENTO**
    a. Clustering
    b. Gestión de recursos
7. **TIME TRAVEL Y RECUPERACIÓN DE DATOS**
8. **STREAMS Y TASKS**
    a. Automatización de flujos de datos
9. **INTRODUCCIÓN A SNOWPARK**

   a. Procesamiento de datos con Python en Snowflake
10. **CASOS PRÁCTICOS DE USO DE SNOWFLAKE EN PIPELINES DE DATOS**

MÓDULO 8. Contenedores, Control de Versiones y DevOps para Data Engineers

1. **INTRODUCCIÓN A BUENAS PRÁCTICAS DEVOPS EN DATA ENGINEERING**
2. **CONTROL DE VERSIONES CON GIT**
   a. Repositorios
   b. Ramas y flujos de trabajo colaborativos
3. **BUENAS PRÁCTICAS DE VERSIONADO DE CÓDIGO DE DATOS**
4. **INTRODUCCIÓN A DOCKER**
   a. Conceptos básicos de contenedores
   b. Imágenes y contenedores
5. **CREACIÓN DE ENTORNOS REPRODUCIBLES CON DOCKER**
6. **DOCKER COMPOSE**
   a. Orquestación básica de servicios
   b. Entornos de desarrollo y pruebas
7. **INTRODUCCIÓN A CI/CD APLICADO A PROYECTOS DE DATOS**
   a. Automatización de pruebas y despliegues
8. **INTEGRACIÓN DE PIPELINES DE DATOS EN FLUJOS CI/CD**

MÓDULO 9. Consumo Analítico y Business Intelligence para Data Engineers

1. **ROL DEL DATA ENGINEER EN LA CAPA DE CONSUMO**
2. **ARQUITECTURA DE LA CAPA ANALÍTICA**
3. **PREPARACIÓN DE DATASETS PARA CONSUMO**
4. **INTEGRACIÓN CON HERRAMIENTAS DE BUSINESS INTELLIGENCE**
   a. Power BI u otras herramientas
5. **MODELOS SEMÁNTICOS Y DATASETS**
6. **BUENAS PRÁCTICAS PARA SERVIR DATOS A NEGOCIO**
   a. Rendimiento
   b. Calidad del dato
7. **GESTIÓN DE ACTUALIZACIONES Y REFRESCOS**
8. **CASOS PRÁCTICOS DE CONSUMO ANALÍTICO**

MÓDULO 10. Proyecto final integrador de Data Engineering

1. DEFINICIÓN DEL CASO PRÁCTICO
2. DISEÑO DE LA ARQUITECTURA DEL PIPELINE DE DATOS
3. INGESTA DE DATOS DESDE DISTINTAS FUENTES
4. PROCESAMIENTO Y TRANSFORMACIÓN DE DATOS
5. ALMACENAMIENTO Y MODELADO DE DATOS
6. AUTOMATIZACIÓN Y ORQUESTACIÓN DEL PIPELINE
7. PREPARACIÓN DE LA CAPA DE CONSUMO
8. CONTROL DE CALIDAD Y VALIDACIÓN DE DATOS
9. DOCUMENTACIÓN TÉCNICA DEL PROYECTO
10. PRESENTACIÓN FINAL DEL PROYECTO