MASTER	Master en Data Science para Finanzas
ASIGNATURA	*Extracción, transformación y carga - ETL*
Nº de ECTS	**3**
Nº de horas docentes	**22.5h (3 ECTS, 15 sesiones)**
Nº de horas actividades académicas dirigidas	
Profesor responsable de la asignatura	**Gabriel Antonio Valverde Castilla**
Curso académico	**2020 / 2021**
Cuatrimestre	**1º Cuatrimestre**

1.- DESCRIPCIÓN GENERAL DE LA ASIGNATURA Y OBJETIVOS DE DOCENCIA:

El objetivo de la asignatura de ETL es múltiple, por un lado se pretende introducir diferentes conceptos y terminologías referentes a DWH y su extensión el Data Lake, por otro lado se pretenden abarcar el proceso de gestión de los datos detallando enfoques estándar y prácticos de procesos de implementación de una solución completa de ETL/ELT, haciendo hincapié en la mejora de la limpieza y calidad del dato. Finalmente, buscando minimizar el tiempo dedicado a las tareas de tratamiento, limpieza y mejora de grandes volúmenes de datos introducimos las herramientas y conocimiento técnicos necesarios para trabajar con Big Data de manera robusta, automática y científica, como son sistemas de bases de datos NOSQL (HIVE, CASSANDRA, MONGODB), programación distribuida (SPARK, HADOOP)

Extracción, tratamiento y carga – ETL (Horas Lectivas/Horas autónomas)

- Primera parte (6h): Introducción a la gestión de información
  - DWH: ODS, OLAP, Data Mart
  - Data Modeling: ER, DM, KV,…
  - Metadata
  - Data Lake (El nuevo enfoque)
  - Integración de Datos y ETL
  - ETL vs ELT
- Segunda parte (6,5h/20h): Data Quality & Data Cleaning, Feature engineering
  - Qué entendemos por Data Quality
  - Métricas DQ
  - DQ Continuum
  - Qué entedemos Data Cleaning
  - Feature Engineering
  - Técnicas y algoritmos de DQ, DC y FE
- Tercer parte (10h/30h): Herramientas de programación y Big Data (Spark)
  - General: Control de versiones (Git), Terminal y Docker, Expresiones regulares
  - Herramientas en R: Tidyverse, vtreat, lubridate, dplyr, datatable, sparklyr, leaflet
  - Herramientas en Python: pandas, pyspark
  - Sistemas de almacenamiento distribuido: HDFS
  - NOSQL: Hive, Cassandra, mongoDB
  - Herramientas ELT: Flume, Kafka, Sqoop
  - Spark (incluido Mlib, spark streaming, sparkSQL)

2.- FORMA DE EVALUACIÓN PREVISTA:

Participación y asistencia	10%
Actividades académicas dirigidas	50%
Prueba objetiva final	40%

Nota: para aprobar la asignatura será imprescindible obtener al menos un 5 en examen final; las actividades académicas dirigidas no serán reevaluables. La asistencia a clase es obligatoria, admitiéndose hasta un 20% de ausencias sin justificación; será criterio del profesor admitir o no la justificación; una asistencia menor del 80% supondrá la pérdida del derecho a examen en convocatoria ordinaria.

PROGRAMA DETALLADO		
Nº de sesión	**Detalle del contenido docente: temas, casos prácticos, actividades académicas dirigidas que se verán en dicha sesión,…**	**Lecturas recomendadas o referencias bibliográficas relativas a los conceptos-temas desarrollados en la sesión**
1	Data Warehouse concepts (DWH) Práctica 1. Instalar las herramientas utilizadas durante el curso y realizar una primera prueba con docker	Apuntes. Documentación propia https://kitematic.com/ http://spark.rstudio.com/ https://medium.com/@GalarnykMichael/install-spark-on-windows-pyspark-4498a5d8d66c
2	DWH. Práctica 2. Elegir una empresa del Ibex 35 y plantear un posible esquema de su estructura de datos y flujo de información.	Ponniah http://www.expansion.com/mercados/cotizaciones/indices/ibex35_I.IB.html
3	ETL Practica 3. Realizar extracción de distintas fuentes de información.	https://es.wikipedia.org/wiki/YAML https://pyyaml.org/wiki/PyYAMLDocumentation https://cran.r-project.org/web/packages/yaml/yaml.pdf Generar metadatos de estas fuentes, definir ficheros yaml, json que contengan la estructura de estos metadatos, crear un ejemplo y generar código R/Python de lectura de esta información
4	Data Lake. Introducción al Paradigma Big Data Práctica 4. Calcular el	https://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf Joyanes

	número de registros sobre datos de SP500 con Spark.	Marr SP 500. Datos financieros e identificación de grupos.
5	ETL vs ELT Práctica 5. Plantea ejemplos en los que un proceso de ELT mejora el proceso ETL.	https://blog.panoply.io/etl-vs-elt-the-difference-is-in-the-how https://www.softwareadvice.com/resources/etl-vs-elt-for-your-data-warehouse/ Dull
6	MapReduce Hadoop Distributed File System (HDFS) Práctica 6. Descarga datos con R o Python de una página web o mediante una API. Cargar ficheros en el sistema HDFS	Datos NBA Shakespeare SP 500. Datos financieros. distribuido. Lee desde R y Python. Thilina Gunarathne Prajapati
7	Transformación: ¿Qué es Data Quality? Métricas de Data Quality Práctica 7: Detecta posibles errores en el sistema de datos. Desarrolla código que pueda solventar algunos de estos errores en la lectura de los datos	Data Gathering, Delivery, Monitoring, Storage, Integration, Retrieval, Manning
8	Transformación de datos R, Python, Spark Practica 8: Análisis exploratorio de la información, generación de variables y transformaciones.	Tidyverse, dplyr, pandas, data.table, sparklyr, reshape,… Dada la información inicial se define el proceso de transformaciones sobre los datos para obtener información veraz en una estructura robusta.
9	ELT: Transformación en paradigma BigData, NOSQL Práctica 9: Crear una tabla en Hive con datos generados por HDFS, utilizar esquemas Hive,	https://cwiki.apache.org/confluence/display/Hive/LanguageManual www.mongodb.com/ Definiremos los pipeline clásicos de carga de la información transformada sobre el sistema, técnicas de logs y definición de alertas ante errores.

	sparklyr y pyspark.	
10	Spark Avanzado Práctica 10: Aplicar las nuevas mejoras sobre el código realizado.	Trabajaremos con técnicas para mejorar el rendimiento, generaremos UDFS para aplicar código Python a través del clúster Estableceremos trucos y buenas prácticas para trabajar con cantidades masivas de información con código en producción
11	Data Quality Continuum. Redefiniendo Data Quality Práctica 11: Define un proceso de validación iterativo en la línea del Data Quality Continuum y genera un conjunto de funciones que lo validen.	Trabajaremos sobre la codificación y desarrollo de código automático para la transformación y evolución del proceso de medición de la calidad y su mejora.
12	Data Quality Continuum. Redefiniendo Data Quality Práctica 12: Generación de paquetes y automatización de código	Hadley Wickham
13	Spark Streaming Práctica 13: Crea un proceso de datos continuo usando spark streaming.	Karau Katsov
14	Introducción a Spark Machine Learning: MLlib Práctica 14: Reducción dimensional y predicción de churn	Dado un conjunto de datos no estructurados realizaremos las tareas oportunas para generar variables, seleccionarlas y aplicar un modelo de predicción. Datos de telefonía
15	Modelos de grafos con Spark: Sistema de recomendación Practica 15: Dado un conjunto de datos en forma de grafos realiza un análisis exploratorio utilizando la librería de Spark	

INFORMACION ADICIONAL

Bibliografía básica	• R for data science. Wickham and Grolemund. O'Reilly, 2017.ISBN 978 1 491 91039 9 • Big Data Analytics with R and Hadoop (by Vignesh Prajapati) • Spark In Action (Petar Zečević and Marko Bonaći) • Git: http://mgaitan.github.io/intro-git/index.html#/step-20 • Data Warehousing Fundamentals. Paulraj Ponniah • Big Data, Análisis de grandes volúmenes de datos en organizaciones. Luis Joyanes Aguilar
Bibliografía Complementaria	• Big data en la práctica : cómo 45 empresas exitosas han utilizado análisis de big data para ofrecer resultados extraordinarios. Bernard Marr • "The Data Lake Debate: Pro is Up First", Dull, Tamara, smartdatacollective.com, March 20, 2015. • "Hadoop: The Definitive Guide 3rd Edition" (O'Reilly) • "Hadoop MapReduce v2 Cookbook, 2nd Edition" (Packt), Thilina Gunarathne •"Hadoop in practice, 2nd edition" (Manning) Alex Holmes •"Big Data Analytics with R and Hadoop" (Packt), Prajapati • " Fast Data Processing with Spark (Holden Karau) • In-Stream Big Data Processing (Llya Katsov) • Spark Workshop Hosted by Stanford ICME (Reza Zadeh, Matei Zaharia, Ion Stoica) • Machine Learning with Spark (Nick Pentreath) • Spark GraphX in Action (Michael S. Malak and Robin East) • Advanced R, Hadley Wickham
Actividades Complementarias	- Instalación de pyspark y sparklyr - Instalación de Docker - Gestión de librerías - Curso breve de git y bash (documentación facilitada) - Se establece una práctica final a modo examen donde el alumno presentará su proyecto ETL, sobre información preferiblemente válida para su TFM.
Localización del profesor	**gvalverd@ucm.es**, **gabriel.valverde@cunef.edu**