Máster en Business Analytics y Big Data
Edición 2015

Asignatura: EXTRACTORES DE DATOS
Módulo: ADQUISICIÓN DE DATOS
Coordinador: Javier Alba, fjavieralba@gmail.com

OBJETIVOS

El objetivo general del módulo es tener una visión lo más completa posible de los distintos métodos de adquisición de datos en entornos Big Data.

Dichos datos suelen orígenes, formatos y protocolos de acceso distintos. Los datos pueden ser públicos o privados. Los orígenes pueden ser ficheros de log de aplicación, eventos generados por dispositivos físicos ó software, APIs, páginas web… y un largo etcétera.

Debido a esta heterogeneidad de orígenes, formatos y métodos de acceso, este módulo pone énfasis en dar una visión general de distintas técnicas y herramientas de adquisición de datos, y aplicarlas en algunos casos prácticos que pueden ser habituales en proyectos de Big Data.

Como **objetivo global**, se persigue que el alumno sea capaz de entender los distintos métodos de adquisición y transporte de datos desde distintos orígenes y que adquiera una visión más o menos completa de las tecnologías actuales orientadas a este cometido, que le será de gran ayuda cuando tenga que enfrentarse al diseño de arquitecturas orientadas a la adquisición y procesamiento de datos a gran escala.

Los **objetivos concretos** del módulo son los siguientes resultados del aprendizaje:

1. Conocer los factores más relevantes a la hora de diseñar una estrategia de adquisición de datos en entornos Big Data
2. Practicar la recolección y transformación de datos heterogéneos con FLUME
3. Conocer qué es KAFKA y cómo puede tener un papel relevante en una arquitectura de recolección de datos heterogéneos, tanto para cargas de trabajo batch como flujos de datos continuos.
4. Practicar la recolección de datos desde bases de datos relacionales a HDFS utilizando SQOOP
5. Conocer técnicas básicas para recolectar datos de webs (crawling, scraping) y de APIs sociales (Twitter)

METODOLOGÍA

El módulo consiste en 2 sesiones presenciales, en las que se combinará la presentación de conceptos del profesor con ejercicios prácticos.

- Inicialmente el profesor explicará algunos conceptos teóricos utilizando una presentación. Posteriormente, se harán ejercicios prácticos en los que se haga uso de las herramientas o técnicas de extracción de datos explicadas.
- Para poder realizar la parte práctica, es necesario que cada alumno cuente con un portátil. El profesor facilitará una máquina virtual para la realización de los ejercicios prácticos.
- La evaluación se realizará en base a dos aspectos: los ejercicios realizados en clase por cada alumno, y una **práctica final individual**

PROGRAMA

Módulo 1: Introducción a la extracción de datos en entornos Big Data

Actividades: En este módulo se explican conceptos básicos y se da una visión general de la adquisición de datos en entornos Big Data.
Es un módulo algo más teórico que el resto, pero también habrá algunos ejercicios prácticos de técnicas básicas de transferencia de datos a HDFS.

Materiales: Portátil del alumno y máquina virtual proporcionada por el profesor

Módulo 2: Recolección de datos con FLUME

Actividades: En este módulo se explica en detalle FLUME, una tecnología orientada a la recolección de datos, frecuentemente utilizada en proyectos Big Data. Se realizarán ejercicios prácticos para familiarizarse con FLUME.

Materiales: Portátil del alumno y máquina virtual proporcionada por el profesor

Módulo 3: KAFKA: una tecnología pensada para la recolección de datos

Actividades: En este módulo se explica qué es Apache Kafka y cómo encaja en una arquitectura de recolección de datos heterogéneos. Se realizarán ejercicios prácticos para familiarizarse con las APIs de "producers" y "consumers" de Kafka.

Materiales: Portátil del alumno y máquina virtual proporcionada por el profesor

Módulo 4: Recolección de datos de bases de datos con SQOOP

Actividades: En este módulo se explica un caso muy típico en proyectos big data: la adquisición de datos de Bases de Datos relacionales a Hadoop. Para ello se explica la tecnología SQOOP y se realizan ejercicios prácticos.

Materiales: Portátil del alumno y máquina virtual proporcionada por el profesor

Módulo 5: Técnicas de extracción de datos de la web: Crawling, Scraping y APIs sociales

Actividades: En este módulo se explica una fuente de datos muy distinta a las vistas hasta ahora: Internet. En muchas ocasiones es necesario mezclar datos que las compañías poseen internamente, con datos externos, como por ejemplo contenidos de webs o datos obtenidos de APIs de redes sociales como Twitter. Se harán algunos ejercicios prácticos, para aprender algunas de estas técnicas.

Materiales: Portátil del alumno y máquina virtual proporcionada por el profesor

MATERIALES

No será necesario más material que el portátil que cada alumno debe llevar a clase.
El profesor proporcionará una máquina virtual sobre la que realizar los ejercicios prácticos.

EVALUACIÓN

Niveles de consecución de los objetivos

Objetivo específico	*Nivel alto*	*Nivel medio*	*Nivel bajo*
O1 – Fuentes de Datos	Ejercicios de clase sobre transferencias simples a HDFS 100% completados.	Ejercicios de clase sobre transferencias simples a HDFS 80% completado.	Comprensión de conceptos teóricos explicados
O2 – FLUME	Ejercicios de clase sobre adquisición de datos con Flume 100% completados. Ejercicio sobre Flume de práctica final 100% completado.	Ejercicios de clase sobre adquisición de datos con Flume 100% completados.	Comprensión de conceptos teóricos explicados en clase
O3 – KAFKA	Ejercicios de clase sobre Kafka 100% completados.	Ejercicios de clase sobre Kafka 80% completados.	Comprensión de conceptos teóricos explicados
O4 – SQOOP	Ejercicios de clase sobre adquisición de	Ejercicios de clase sobre adquisición de	Comprensión de conceptos teóricos

	datos con SQOOP 100% completados. Ejercicio sobre SQOOP de práctica final 100% completado	datos con SQOOP 100% completados	explicados
O5 – Web Data	Ejercicios de clase sobre crawling & scraping 100% completados. Ejercicio sobre scraping de práctica final 100% completado	Ejercicios de clase sobre crawling & scraping 100% completados.	Comprensión de conceptos teóricos explicados

Modelo de evaluación

Elemento	*Peso*
Completar ejercicios prácticos en cada sesión	50%
Práctica final individual	50%

PROFESORADO

Javier Alba es consultor Big Data en Pivotal. Durante los últimos años ha desempeñado labores de desarrollo de software tanto en grandes corporaciones como en start-ups. Ha estado siempre ligado a desarrollos de back-end, en los que la adquisición, procesamiento, almacenaje y explotación de datos ha sido una constante.