MÁSTER EN DATA SCIENCE

Presentación de la asignatura

Obtención de Datos

Universidad Rey Juan Carlos

Cuestiones generales

- Cuestiones generales de la asignatura
  - Créditos: 3 ECTS (24 horas presenciales + 6 horas de tutorías).
  - Horario: Jueves (Móstoles) / Viernes (Ferraz), ambas de 17:00 a 19:00.
  - 12 semanas de clase.

Profesorado

- Paloma Cáceres García de Marina
  - Doctora en Informática por la URJC.
  - Campus de Móstoles, Edificio Departamental II, despacho 118.
  - paloma.caceres.garciademarina@gmail.com
  - paloma.caceres@urjc.es

- Almudena Sierra Alonso
  - Doctora en Informática por la UPM.
  - Campus de Móstoles, Edificio Departamental II, despacho 111.
  - asierraalonso@gmail.com
  - almudena.sierra@urjc.es

Enfoque de la asignatura

- Un proceso básico de **big data** pasa por las etapas de extracción, transformación y carga: **Proceso ETL**.

- Es un proceso necesario cuando se quiere trabajar con datos procedentes de diferentes fuentes tanto homogéneas como heterogéneas para su posterior consumo. Por ello el proceso consta de las siguientes etapas:
  - **Extract**: Extracción de datos de calidad procedentes tanto de fuentes homogéneas como heterogéneas
  - **Transform**: Aplicación de un conjunto de reglas de transformación para "traducir" los datos y hacerlos homogéneos. Por ejemplo, datos con la misma unidad de medida (km, metros, hectómetros,…).
  - **Load**: Carga de los datos con aseguramiento de la calidad y corrección, con el gasto mínimo de recursos.

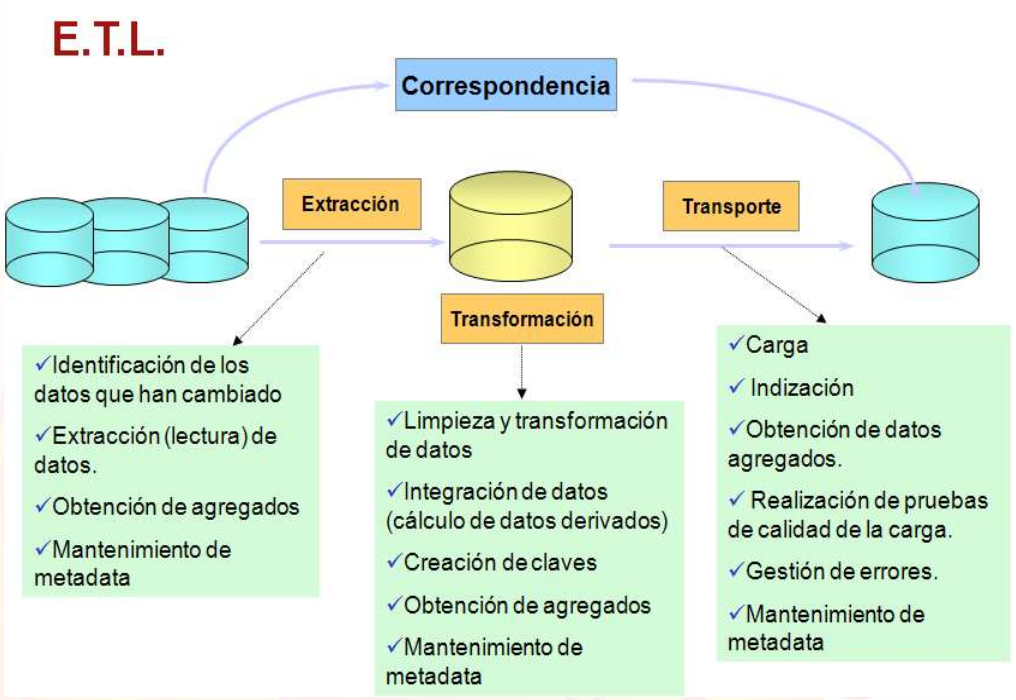
- ETL, el gran coste del Big Data
  - http://www.channelbiz.es/2015/07/15/etl-el-gran-coste-del-big-data/
  - Julio 2015

Enfoque de la asignatura

Un tercio de los profesionales de BI (Business Intelligence) gastan entre un 50% y un 90% de su tiempo en limpiar los datos que luego van a analizar.

Además, para el 97% de los encuestados los procesos ETL son críticos en sus tareas de Business Intelligence.

[Encuesta Xplenty, Julio 2015, https://www.xplenty.com/]



http://carlosproal.com/dw/dw05.html

Presentación de la asignatura

- Dentro de un proceso básico de ETL, nuestra asignatura se centra en la obtención de datos procedentes de la Web o de ficheros con diferentes formatos.

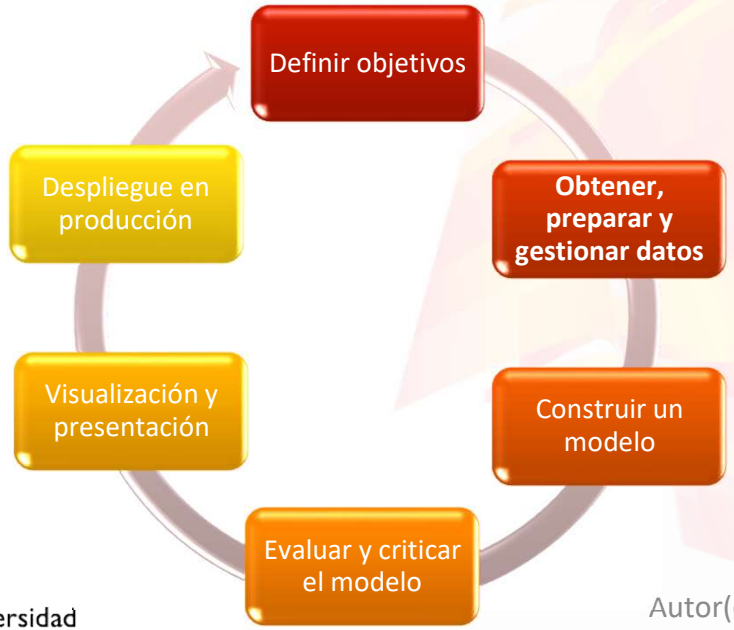
- También veremos, aunque con menos énfasis:
  - Transformación de datos
  - Almacenamiento de datos

- Por lo tanto, nos aprovisionaremos de mecanismos que nos permitan automatizar procesos que trabajen en la web, en los que en la medida de lo posible no estén involucrados directamente los seres humanos, con la finalidad de obtener datos para su posterior consumo.

Objetivos de la asignatura

- **Objetivo de la asignatura**
  - Aprender métodos y técnicas para obtener datos procedentes de múltiples y diferentes fuentes (y transformarlos y almacenarlos).

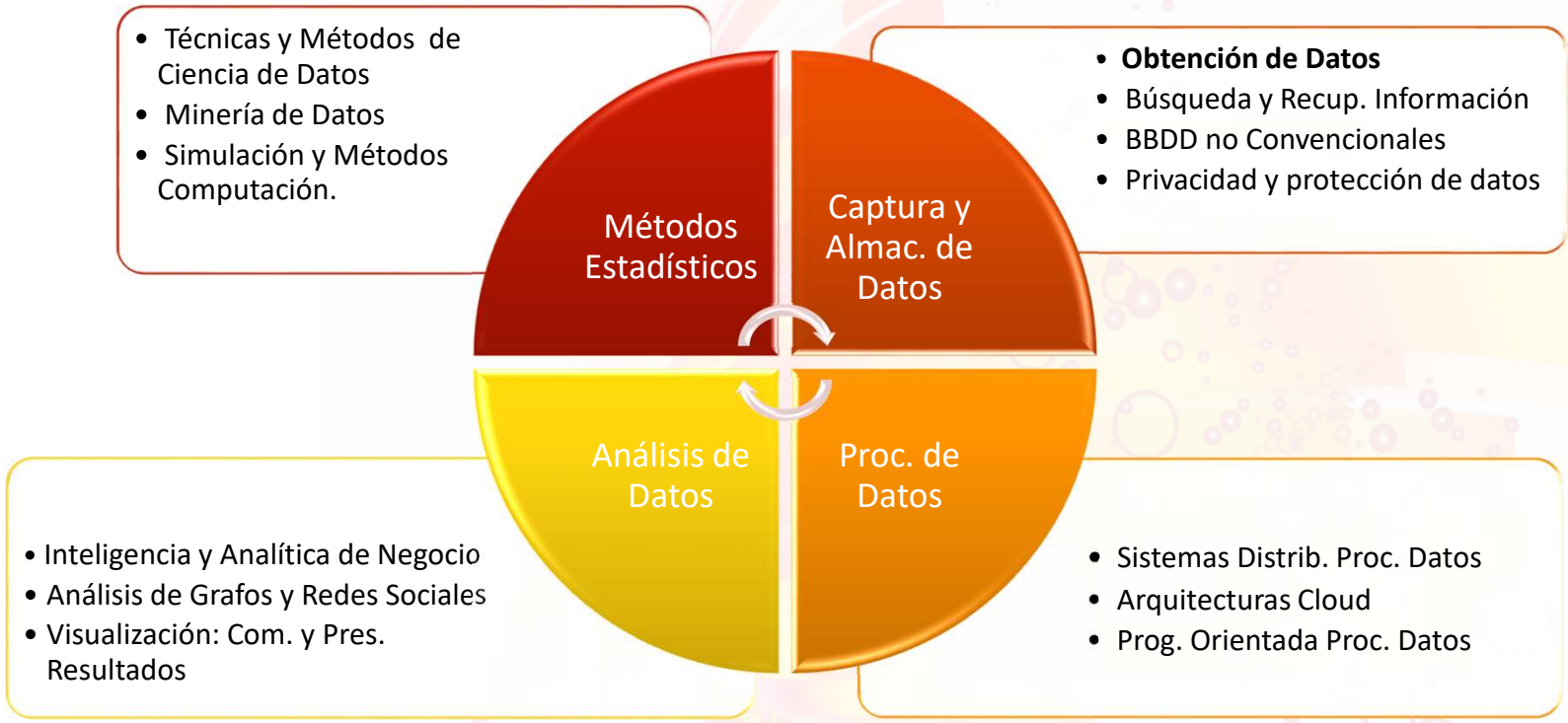
- **Ciclo de vida de proyecto de DS**



Obtener, preparar y gestionar los datos

- Definir el programa de obtención
  - Entorno, lenguaje y librerías
  - Implementación
- Definir las transformaciones
  - Transformar el modelo conceptual en el modelo destino (de un diagrama de clases UML a tripletas RDF,….).
  - Homogeneización de los datos.
  - Almacenamiento intermedio (selección de formatos de destino)

Tipos de tareas en el ciclo de proyecto



Temario de la asignatura

- Tema 1. Representación de datos
- Tema 2. Datos en la Web
- Tema 3. Datos semánticos y enlazados

Temario de la asignatura

- **Tema 1. Representación de Datos (y extracción)**
  - En este tema se abordan los formatos más habituales en que se pueden encontrar los datos.
  - Objetivo: Poder identificar la ubicación de los datos que queremos obtener en función del formato en el que estén los datos. Aprender los mecanismos de extracción de Xpath.
  - Duración aproximada: 6 horas de clase.
  - Formatos:
    - HTML (HTML5), XML, JSON, YAML, HDF5.
  - Contenidos:
    - Datos estructurado y más
    - Lenguajes de marcado
    - Namespaces
    - DTD y XML Schema
    - XPath

Temario de la asignatura

- **Tema 2. Datos en la Web**
  - En este tema se aborda la obtención de datos en la web, propiamente dicha.
  - Objetivo: Ser capaz de obtener datos residentes en la web, procesándolos con Python, a partir del formato concreto en el que se encuentren dichos datos.
  - Duración aproximada: 12 horas de clase.
  - Entorno de trabajo:
    - Anaconda
  - Contenidos:
    - Web scraping
    - APIs de datos
    - Bibliotecas

Temario de la asignatura

- **Tema 3. Datos semánticos y enlazados**
  - En este tema se tratarán conceptos relativos a la web semántica y los datos enlazados.
  - Objetivo: Ser capaz de comprender las características de la web semántica, datos enlazados y las técnicas relacionadas.
  - Duración aproximada: 6 horas de clase.
  - Contenidos:
    - Ontologías
    - Metadatos
    - Datos enlazados
    - Tecnologías semánticas