

**Proyecto/Guía docente de la asignatura**

<b>Asignatura</b>	Arquitecturas Big Data		
<b>Materia</b>	Tecnologías de Gestión de Información		
<b>Módulo</b>			
<b>Titulación</b>	Máster en Ingeniería Informática		
<b>Plan</b>	639	<b>Código</b>	54933
<b>Periodo de impartición</b>	1º Cuatrimestre	<b>Tipo/Carácter</b>	Obligatoria
<b>Nivel/Ciclo</b>	Máster	<b>Curso</b>	1º
<b>Créditos ECTS</b>	3		
<b>Lengua en que se imparte</b>	Castellano		
<b>Profesor/es responsable/s</b>	Anibal Bregón Bregón Fernando Díaz Gómez Miguel Ángel Martínez Prieto		
<b>Datos de contacto (E-mail, teléfono...)</b>	Escuela de Ingeniería Informática Plaza de la Universidad 1, 40005 Segovia anibal@infor.uva.es fdiaz@infor.uva.es migumar2@infor.uva.es		
<b>Departamento</b>	Informática (ATC, CCIA, LSI)		
<b>Fecha de revisión por el Comité de Título</b>	15 / 09 / 2023		



## 1. Situación / Sentido de la Asignatura

### 1.1 Contextualización

Durante la última década hemos vivido un cambio sin precedentes en el desarrollo de sistemas avanzados para la gestión y explotación de datos en entornos empresariales, pero también de nuestra vida cotidiana. El volumen de los datos crece de forma exponencial, año tras año. Además, la forma que adoptan estos datos ya no se adapta, tan fácilmente, a las soluciones utilizadas que se han utilizado durante los últimos 50 años en el mundo informático. Las bases de datos relacionales, y las arquitecturas desarrolladas en torno a ellas, han cedido el paso a nuevas soluciones que permiten almacenar, procesar, explorar y poner en explotación grandes volúmenes de datos (estructurados, semi-estructurados y no estructurados), que pueden consumirse casi al mismo tiempo en el que se generan. Esta situación se refiere bajo el término “Big Data” y abre numerosos desafíos, tecnológicos y de negocio.

La asignatura Arquitecturas Big Data ofrece al alumno los conocimientos fundamentales para comprender el reto que supone emprender un proyecto Big Data, proporcionando una revisión general de su arquitectura y de las tecnologías necesarias para satisfacer los diferentes tipos de necesidades que se presentan en un entorno. Más concretamente, se presenta la noción de Data Lake, como elemento vertebrador de un proyecto Big Data, y se utiliza el *framework* Apache Hadoop, como referencia tecnológica para la implementación de este tipo de soluciones informáticas. Los cuatro bloques temáticos de la asignatura plantean un análisis general del Data Lake y de cada una de las capas que lo componen, así como de la tecnología con la que se implementan, dentro del ecosistema Apache Hadoop. A lo largo de la asignatura, aprenderemos a identificar y caracterizar fuentes de datos (*raw data*) de interés en el ámbito de un proyecto, así como a realizar todos los pasos necesarios para transformarlas en las representaciones (*smart data*) que se requieren en el entorno de negocio en el que se desarrolla el proyecto. Además, se ofrece una aproximación a las bases de datos NoSQL y a algunos de los sistemas más utilizados en este nuevo paradigma, que encaja a la perfección con las necesidades de explotación que plantea un entorno Big Data.

Todos los contenidos de la asignatura estarán disponibles a través del Campus Virtual de la Universidad de Valladolid: <https://campusvirtual.uva.es/>.

### 1.2 Relación con otras materias

- I+D+i en Informática
- Supercomputación y BigData
- Técnicas escalables de análisis de datos en entornos Big Data: Clasificadores.
- Técnicas escalables de análisis de datos en entornos Big Data: Regresión y descubrimiento de conocimiento.

### 1.3 Prerrequisitos

Se recomienda que el alumno, en sus estudios de Grado, haya adquirido un mínimo de competencias en relación con el uso, configuración y administración, y conocimiento de los lenguajes de programación utilizados en sistemas operativos, sistemas distribuidos y sistemas de bases de datos.



## 2. Competencias

### 2.1 Generales

CG2. Capacidad para la dirección de obras e instalaciones de sistemas informáticos, cumpliendo la normativa vigente y asegurando la calidad del servicio.

CG6. Capacidad para la dirección general, dirección técnica y dirección de proyectos de investigación, desarrollo e innovación, en empresas y centros tecnológicos, en el ámbito de la Ingeniería Informática.

CG8. Capacidad para la aplicación de los conocimientos adquiridos y de resolver problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y multidisciplinares, siendo capaces de integrar estos conocimientos.

### 2.2 Específicas

CEG1. Capacidad para la integración de tecnologías, aplicaciones, servicios y sistemas propios de la Ingeniería Informática, con carácter generalista, y en contextos más amplios y multidisciplinares.

CET1. Capacidad para modelar, diseñar, definir la arquitectura, implantar, gestionar, operar, administrar y mantener aplicaciones, redes, sistemas, servicios y contenidos informáticos.

CET5. Capacidad para analizar las necesidades de información que se plantean en un entorno y llevar a cabo en todas sus etapas el proceso de construcción de un sistema de información.

### 3. Objetivos

- Comprender las características de un proyecto Big Data, su ciclo de vida y las necesidades funcionales que plantea.
- Conocer el concepto de Data Lake, así como las características básicas y responsabilidades de sus componentes arquitectónicos.
- Conocer los fundamentos del *framework* Apache Hadoop y de las tecnologías que ofrece para abordar la construcción de un Data Lake.
- Conocer las responsabilidades de la capa de almacenamiento de datos y su utilización mediante el sistema de archivos HDFS.
- Conocer las responsabilidades de la capa de computación y su utilización mediante modelos de computación como MapReduce o Spark.
- Conocer las responsabilidades de la capa de gestión de datos y su utilización mediante herramientas de captura (Flume y Kafka), transformación (Hive y Oozie) y carga (Sqoop).
- Ser capaz de entender los retos que supone el almacenamiento de Big Data y como este almacenamiento pasa por utilizar técnicas de distribución de datos.
- Ser capaz de entender las capacidades específicas de los modelos principales de almacenamiento NoSQL y sus diferencias respecto al modelo relacional.
- Ser capaz de identificar un problema Big Data y elegir el mejor modelo de almacenamiento NoSQL para afrontarlo.
- Ser capaz de utilizar algunas de las bases de datos NoSQL más demandadas en los escenarios Big Data actuales.

#### 4. Contenidos y/o bloques temáticos

##### Bloque 1: Introducción

Carga de trabajo en créditos ECTS: 0,6

##### a. Contextualización y justificación

¿Qué es Big Data? La sencillez de esta pregunta esconde la complejidad de una respuesta que depende, fuertemente, del contexto en el que plantee. En esta asignatura, el Big Data se entiende como parte del contexto tecnológico de una organización y, por tanto, tenemos que ser capaces de comprender las implicaciones que tiene en dicho contexto y los recursos de los que disponemos para poder explotarlo en el entorno de negocio de la organización.

El Data Lake es la piedra angular en el proceso de cambio que requieren las organizaciones para poder utilizar las grandes colecciones de datos (Big Data) de las que disponen. No es un sistema gestor de bases de datos, ni un sistema de archivos, ni tampoco un entorno de computación. Sin embargo, da cabida a todas estas piezas, además de a todas aquellas que garantizan el auto-servicio de sus usuarios, entendiéndose como tal a todos los miembros de la organización que utilicen el Big Data de la organización para satisfacer sus (diferentes) necesidades.

En este bloque se establecen los fundamentos necesarios para abordar el resto de la asignatura. Para ello, se planteará una pequeña motivación a las necesidades tecnológicas que requiere la gestión de Big Data en una organización y se justificará la necesidad del Data Lake para satisfacerlas. A continuación, se analizará la arquitectura del Data Lake, estableciendo la relación entre sus capas, analizando sus responsabilidades y los servicios que presta cada una de ellas. Finalmente, se planteará una sencilla metodología de ciclo de vida que establece los pasos a realizar para llevar a cabo un proyecto Big Data.

En la segunda parte de este bloque se introducirá el *framework* Apache Hadoop, como tecnología de referencia para la construcción de un Data Lake. También se describirán las plataformas que ofrecen esta tecnología, centrándonos en la que utilizará en la asignatura: Cloudera, y en las herramientas que esta proporciona.

##### b. Objetivos de aprendizaje

- Comprender qué es Big Data y por qué su gestión es un reto.
- Conocer el concepto de Data Lake, sus capas y cada una de sus responsabilidades.
- Aprender una metodología de ciclo de vida para la realización de proyectos Big Data.
- Conocer los fundamentos de Apache Hadoop y las tecnologías que ofrece.
- Familiarizarse con la plataforma Cloudera y los servicios que ofrece.

##### c. Contenidos

**Big Data.** Conceptos Básicos; Data Lake.

**Metodología de Ciclo de Vida.**

**Apache Hadoop.** Arquitectura; Plataformas; Tecnologías.

---

#### d. Métodos docentes

---

1. **Lección magistral:** exposición de la teoría y resolución de ejercicios propuestos.
2. **Realización de prácticas guiadas y libres de laboratorio.**
3. **Método de proyectos y aprendizaje por tareas y exposición del trabajo realizado por parte del alumno.**
4. **Estudio autónomo por parte del alumno**, incluyendo realización de problemas, consulta bibliográfica y realización de prácticas y pruebas de evaluación.
5. **Sesiones de tutorías (grupales o individuales), seguimiento y evaluación.**

---

#### e. Plan de trabajo

---

Los contenidos se desarrollarán en forma de clase magistral y, a continuación, los alumnos trabajarán en la resolución de supuestos teórico-prácticos que les ayuden a asentar los conocimientos adquiridos en un contexto de aplicación. Además, se llevará a cabo una actividad de aprendizaje basado en proyectos, en la que los alumnos trabajarán en equipo para abordar el desarrollo de un pequeño proyecto Big Data.

---

#### f. Evaluación

---

Véase apartado 7 de la guía docente.

---

#### g Material docente

---

La bibliografía correspondiente a la asignatura está disponible en la lista "Arquitecturas Big Data [MII]": [https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC\\_UVA/lists/5317780570005774](https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC_UVA/lists/5317780570005774).

---

##### g.1 Bibliografía básica

---

- GORELIK, A. *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science*. 2019. O'Reilly.
- SHARMA, B. *Architecting Data Lakes*. O'Reilly, 2018.
- WHITE, T. *Hadoop: The Definitive Guide*. 4th Ed. O'Reilly Media. 2015
- SINGH, C. y KUMAR, M. *Mastering Hadoop 3: Big data processing at scale to unlock unique business insights*. Packt Publishing, 2019.

---

##### g.2 Bibliografía complementaria

---

- KIMBALL, R. y CASERTA, J. *The Data Warehouse ETL Toolkit*. John Wiley&Sons, 2004.
- MARZ, N. y WARREN, J. *Big Data principles and best practices of scalable realtime data systems*. Manning, 2015.

---

##### g.3 Otros recursos telemáticos (píldoras de conocimiento, blogs, videos, revistas digitales, cursos masivos (MOOC), ...)

---

- KREPS, J. *Questioning the Lambda Architecture*. O'Reilly, 2014. <https://www.oreilly.com/ideas/questioning-the-lambda-architecture>.



PRESS, G. *A Very Short History of Big Data*. 2013. <https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/>.

MATTHEWS, K. *The difference between a data swamp and a data lake? 5 signs*. 2019. <https://www.information-age.com/data-swamp-data-lake-123481597/>.

#### h. Recursos necesarios

Aula con pizarra y cañón de proyección. Laboratorio con pizarra y cañón de proyección. Ordenadores y software adecuado. Despacho o seminario para tutorías.

Campus Virtual con foros de discusión y herramientas para la entrega y evaluación de ejercicios resueltos por los alumnos.

#### i. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
0,6	Semanas 1 y 2

## Bloque 2: Almacenamiento y Computación

Carga de trabajo en créditos ECTS: 0,8

### a. Contextualización y justificación

Las capas de almacenamiento y computación proporcionan (sobre la infraestructura del Data Lake) la escalabilidad necesaria para preservar y poder procesar colecciones de datos de forma efectiva.

En cuanto al almacenamiento, en este bloque se presentarán las características principales del sistema HDFS, su arquitectura y las funcionalidades básicas soportadas. Una vez introducidas las nociones fundamentales de HDFS, es importante analizar el papel que juega la capa de almacenamiento, en el contexto del desarrollo de proyectos Big Data, destacando su rol vertebrador dentro de la pila Hadoop. A su vez, la capa de almacenamiento (HDFS), junto con otros componentes de integración de datos, proporcionan a los usuarios y/o aplicaciones Big Data, la infraestructura básica sobre la que se asienta la capa superior.

En lo relativo al procesamiento, este bloque presentará los fundamentos de MapReduce y Spark, que, dentro de la pila Hadoop, son los motores de procesamiento orientados al procesamiento *batch* e interactivo (*in-memory*), respectivamente. En ambos casos, estamos ante soluciones que implementan modelos de computación distribuidos, en el que el cómputo se mueve a los datos, y no al revés, como se ha hecho de forma tradicional. La presentación de esta capa se realizará de una forma más somera, pero estableciendo los aspectos fundamentales de estos modelos de computación de cara a comprender como se explotan en la capa de Gestión de Datos.

### b. Objetivos de aprendizaje

- Conocer los principios de diseño y la arquitectura del sistema de ficheros HDFS.
- Identificar el papel de la capa de almacenamiento soportada por HDFS en una arquitectura Big Data basada en Hadoop.
- Conocer la arquitectura básica del sistema HDFS.
- Entender el funcionamiento de los distintos modelos de computación para Big Data en función del tipo de procesamiento llevado a cabo: procesamiento *batch* o interactivo.
- Conocer las restricciones y exigencias a las que están sometidos los programas que trabajan con Big Data.

### c. Contenidos

**Almacenamiento.** Introducción; HDFS.

**Computación.** Introducción; MapReduce; Spark.

### d. Métodos docentes

1. **Lección magistral:** exposición de la teoría y resolución de ejercicios propuestos.
2. **Realización de prácticas guiadas y libres de laboratorio.**
3. **Método de proyectos y aprendizaje por tareas y exposición del trabajo realizado por parte del alumno.**



4. **Estudio autónomo por parte del alumno**, incluyendo realización de problemas, consulta bibliográfica y realización de prácticas y pruebas de evaluación.
5. **Sesiones de tutorías (grupales o individuales), seguimiento y evaluación.**

#### e. Plan de trabajo

---

Los contenidos se desarrollarán en forma de clase magistral y, a continuación, los alumnos trabajarán en la resolución de supuestos teórico-prácticos que les ayuden a asentar los conocimientos adquiridos en un contexto de aplicación. Además, se llevará a cabo una actividad de aprendizaje basado en proyectos, en la que los alumnos trabajarán en equipo para abordar el desarrollo de un pequeño proyecto Big Data.

#### f. Evaluación

---

Véase apartado 7 de la guía docente.

#### g Material docente

---

La bibliografía correspondiente a la asignatura está disponible en la lista "Arquitecturas Big Data [MII]": [https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC\\_UVA/lists/5317780570005774](https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC_UVA/lists/5317780570005774).

#### g.1 Bibliografía básica

---

SINGH, C. y KUMAR, M. "Mastering Hadoop 3: Big data processing at scale to unlock unique business insights". Packt Publishing, 2019.  
WHITE, T. "Hadoop: The Definitive Guide". 4th Ed. O'Reilly Media. 2015

#### g.2 Bibliografía complementaria

---

BENGFORT, B., Y KIM, J. "Data Analytics with Hadoop: An Introduction for Data Scientists". Ed.. O'Reilly Media. 2016

#### h. Recursos necesarios

---

Aula con pizarra y cañón de proyección. Laboratorio con pizarra y cañón de proyección. Ordenadores y software adecuado. Despacho o seminario para tutorías.  
Campus Virtual con foros de discusión y herramientas para la entrega y evaluación de ejercicios resueltos por los alumnos.

#### i. Temporalización

---

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
0,8	Semanas 2, 3 y 4

### Bloque 3: Gestión de Datos

Carga de trabajo en créditos ECTS: 0,8

#### a. Contextualización y justificación

La capa de gestión de datos proporciona los mecanismos necesarios para llevar a cabo el ciclo de vida del Big Data dentro del Data Lake, apoyándose en los recursos que ofrecen las capas inferiores. Por lo tanto, se encarga de la ingesta del *raw data*, de gestionar su almacenamiento (en las diversas zonas que mantiene), de su transformación en *smart data* y de su posterior carga en sistemas de explotación de datos. Además, proporcionan mecanismos de exploración, fundamentales para la actividad del científico de datos. La capa de gestión de datos asume también la responsabilidad del auto-servicio.

En este bloque se proporcionan los fundamentos sobre los que se construyen los *dataflows* habituales de un proyecto Big Data (inspirados en los procesos ETL, utilizando en entornos de inteligencia de negocio) y se presentan las herramientas más utilizadas para estos propósitos en el ecosistema Apache Hadoop. Todo este conocimiento será puesto en práctica a través de pequeños supuestos que permitirán explorar los aspectos principales de cada herramienta, así como encajar cada uno de ellos en el ámbito del proyecto Big Data que se desarrolla en la asignatura.

#### b. Objetivos de aprendizaje

- Conocer los componentes del ecosistema Hadoop que soportan la ingestión de datos en un Data Lake (Apache Flume, Apache Kafka)
- Comprender los esquemas de procesamiento para la ingesta de datos (*batch*, macro-*batch*, interactivos)
- Comprender la complejidad de transformar *raw data* en *smart data*.
- Conocer los principios fundamentales de Apache Hive y aprender a utilizarlos para implementar procesos de transformación de datos en el ámbito de un Data Lake.
- Conocer los principios fundamentales de Apache Oozie y aprender a utilizarlos para construir los *dataflows* requeridos en los proyectos Big Data.
- Conocer los principios fundamentales de Apache Sqoop y aprender a utilizarlos para transferir datos entre el Data Lake y un entorno de explotación externo.

#### c. Contenidos

**Extracción.** Introducción; Apache Flume. Apache Kafka

**Transformación.** Introducción; Apache Hive; Apache Oozie.

**Carga.** Introducción; Apache Sqoop.

#### d. Métodos docentes

1. **Lección magistral:** exposición de la teoría y resolución de ejercicios propuestos.
2. **Realización de prácticas guiadas y libres de laboratorio.**
3. **Método de proyectos y aprendizaje por tareas y exposición del trabajo realizado por parte del alumno.**



4. **Estudio autónomo por parte del alumno**, incluyendo realización de problemas, consulta bibliográfica y realización de prácticas y pruebas de evaluación.
5. **Sesiones de tutorías (grupales o individuales), seguimiento y evaluación.**

#### e. Plan de trabajo

---

Los contenidos se desarrollarán en forma de clase magistral y, a continuación, los alumnos trabajarán en la resolución de supuestos teórico-prácticos que les ayuden a asentar los conocimientos adquiridos en un contexto de aplicación. Además, se llevará a cabo una actividad de aprendizaje basado en proyectos, en la que los alumnos trabajarán en equipo para abordar el desarrollo de un pequeño proyecto Big Data.

#### f. Evaluación

---

Véase apartado 7 de la guía docente.

#### g Material docente

---

La bibliografía correspondiente a la asignatura está disponible en la lista "Arquitecturas Big Data [MII]": [https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC\\_UVA/lists/5317780570005774](https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC_UVA/lists/5317780570005774).

##### g.1 Bibliografía básica

---

CAPRIOLO, E., WAMPLER, D., RUTHERGLEN, J. *Programming Hive*. 1st Ed. O'Reilly Media. 2012.

WHITE, T. *Hadoop: The Definitive Guide*. 4th Ed. O'Reilly Media. 2015

SINGH, C. y KUMAR, M. *Mastering Hadoop 3: Big data processing at scale to unlock unique business insights*. Packt Publishing, 2019.

##### g.2 Bibliografía complementaria

---

Mendelevitch, O.; Stella. C. y Eadline, D. *Practical Data Science with Hadoop® and Spark: Designing and Building Effective Analytics at Scale*. Addison-Wesley Professional. 2016.

BENGFORT, B., Y KIM, J. "Data Analytics with Hadoop: An Introduction for Data Scientists". Ed.. O'Reilly Media. 2016

##### g.3 Otros recursos telemáticos (píldoras de conocimiento, blogs, videos, revistas digitales, cursos masivos (MOOC), ...)

---

THUSOO, A.; SARMA, J.S.; JAIN, N.; SHAO, Z.; CHAKKA, P.; ANTHONY, S.; LIU, H.; WYCKOFF, P. Y MURTHY, R. *Hive: a Warehousing Solution over a Map-Reduce Framework*. Proceedings of the VLDB Endowment VLDB Endowment, 2(2), 1626-1629, 2009.

#### h. Recursos necesarios

---

Aula con pizarra y cañón de proyección. Laboratorio con pizarra y cañón de proyección. Ordenadores y software adecuado. Despacho o seminario para tutorías.

Campus Virtual con foros de discusión y herramientas para la entrega y evaluación de ejercicios resueltos por los alumnos.



### i. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
0,8	Semanas 4, 5 y 6





## Bloque 4: NoSQL

Carga de trabajo en créditos ECTS: 0,8

### a. Contextualización y justificación

Existen dos grandes formas de escalar bases de datos relaciones: el escalado vertical y el escalado horizontal. El escalado vertical se centra en aumentar las capacidades de escalabilidad añadiendo recursos hardware de más capacidad, lo cual acaba llegando a un punto a partir del cual no es posible (o no es económicamente viable) escalar. En cambio, el escalado horizontal se centra en aumentar el número de máquinas y distribuir los datos entre éstas. Sin embargo, esto genera problemas adicionales relacionados con el particionado de los datos. Veremos como el uso de *sharding* permite tener escalabilidad horizontal y mejora el rendimiento, pero aún así puede tener problemas debido a su complejidad. Además, si el *sharding* se utiliza con replicación basada en la localización, se mejora la escalabilidad y la disponibilidad, pero aparecen nuevos problemas que pueden comprometer la consistencia de los datos. Estas ideas darán pie al teorema CAP, que será el teorema central de este bloque.

El teorema CAP, nombre que viene de las siglas de sus 3 propiedades (*Consistency, Availability y Partition tolerance*), establece que sólo podemos tener 2 de esas 3 propiedades al mismo tiempo. Esto tuvo como resultado bases de datos con propiedades ACID relajadas, conocidas como propiedades BASE (*Basically Available, Soft-state, Eventually consistent*). BASE es una alternativa flexible a ACID para aquellos almacenes de datos que no requieren una adherencia estricta a las transacciones. Todas estas ideas dieron lugar a un tipo de bases de datos que siguen las propiedades BASE: las bases de datos NoSQL.

NoSQL define un conjunto de bases de datos no relacionales, que siguen las propiedades BASE, que escalan horizontalmente y que son baratas y fáciles de implementar. Existen 4 grupos principales de bases de datos NoSQL: clave-valor, orientadas a columnas, orientadas a documentos y basadas en grafos. En este bloque se presentarán las ideas fundamentales de estos 4 grupos de bases de datos NoSQL.

### b. Objetivos de aprendizaje

- Conocer que es una base de datos NoSQL, sus principales características, y saber en qué consisten el teorema CAP y las propiedades BASE.
- Conocer los principales tipos de bases de datos NoSQL y ser capaz de determinar qué base de datos NoSQL se ajusta mejor a cada problema particular.
- Conocer las características básicas de HBase, así como su modelo de datos, su arquitectura y su manejo.
- Conocer las características básicas de Cassandra, así como su modelo de datos, su arquitectura y su manejo.

### c. Contenidos

**NoSQL.** El teorema CAP; Bases de datos NoSQL; Taxonomía.

**HBase.** Modelo de datos; Modelo físico; Arquitectura.

**Cassandra.** Modelo de datos; Arquitectura.

#### **d. Métodos docentes**

---

1. **Lección magistral:** exposición de la teoría y resolución de ejercicios propuestos.
2. **Realización de prácticas guiadas y libres de laboratorio.**
3. **Método de proyectos y aprendizaje por tareas y exposición del trabajo realizado por parte del alumno.**
4. **Estudio autónomo por parte del alumno**, incluyendo realización de problemas, consulta bibliográfica y realización de prácticas y pruebas de evaluación.
5. **Sesiones de tutorías (grupales o individuales), seguimiento y evaluación.**

#### **e. Plan de trabajo**

---

Los contenidos se desarrollarán en forma de clase magistral y, a continuación, los alumnos trabajarán en la resolución de supuestos teórico-prácticos que les ayuden a asentar los conocimientos adquiridos en un contexto de aplicación. Además, se llevará a cabo una actividad de aprendizaje basado en proyectos, en la que los alumnos trabajarán en equipo para abordar el desarrollo de un pequeño proyecto Big Data.

#### **f. Evaluación**

---

Véase apartado 7 de la guía docente.

#### **g Material docente**

---

La bibliografía correspondiente a la asignatura está disponible en la lista "Arquitecturas Big Data [MII]": [https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC\\_UVA/lists/5317780570005774](https://buc-uva.alma.exlibrisgroup.com/leganto/public/34BUC_UVA/lists/5317780570005774).

##### **g.1 Bibliografía básica**

---

- TIWARI, SHASHANK, "*Professional NoSQL*", Wiley/Wrox. 2011.
- STRAUCH, CHRISTOF, "*NoSQL Databases*". Stuttgart Media University. 2012.
- VAISH, G., "*Getting Started with NoSQL*". Packt Publishing. 2013.
- BREWER, E. "*Towards Robust Distributed System*". Symposium on Principles of Distributed Computing (PODC). 2000.
- LARS, G., "*HBase: The Definitive Guide*", O'Reilly Media. Agosto de 2011
- CARPENTER, J., HEWITT, E. "*Cassandra: The Definitive Guide*", 2nd Edition, O'Reilly Media. 2016.

##### **g.2 Bibliografía complementaria**

---

- Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. 2006. Bigtable: a distributed storage system for structured data. In Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7 (OSDI '06), Vol. 7. USENIX Association, Berkeley, CA, USA, 15-15
- Eric A. Brewer. CAP twelve years later: How the "rules" have changed, February 2012
- S. Gilbert and N. Lynch. Perspectives on the CAP Theorem, February 2012
- Dan Pritchett. 2008. BASE: An Acid Alternative. Queue 6, 3 (May 2008), 48-55. DOI=10.1145/1394127.1394128



- HBase in Action, por Nicholas Dimiduk and Amandeep Khurana. Noviembre de 2012. ISBN 978161729052

#### h. Recursos necesarios

Aula con pizarra y cañón de proyección. Laboratorio con pizarra y cañón de proyección. Ordenadores y software adecuado. Despacho o seminario para tutorías.

Campus Virtual con foros de discusión y herramientas para la entrega y evaluación de ejercicios resueltos por los alumnos.

#### i. Temporalización

CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
0,8	Semanas 6, 7 y 8





## 5. Métodos docentes y principios metodológicos

La concreción de los métodos docentes se ha especificado en cada uno de los bloques temáticos.



## 6. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES PRESENCIALES o PRESENCIALES A DISTANCIA <sup>(1)</sup>	HORAS	ACTIVIDADES NO PRESENCIALES	HORAS
Clases teórico-prácticas (T/M)	18	Estudio y trabajo autónomo individual	20
Laboratorios (L)	12	Estudio y trabajo autónomo grupal	25
Total presencial	<b>30</b>	Total no presencial	<b>45</b>
TOTAL presencial + no presencial			<b>75</b>

(1) Actividad presencial a distancia es cuando un grupo sigue una videoconferencia de forma síncrona a la clase impartida por el profesor para otro grupo presente en el aula.

## 7. Sistema y características de la evaluación

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Resolución de problemas y exámenes	40%	
Actividad de aprendizaje basado en proyectos	50%	
Participación en foros y otros medios participativos	10%	

### CRITERIOS DE CALIFICACIÓN

- **Convocatoria ordinaria:**

Para aprobar la asignatura será necesario obtener una calificación total de 5 puntos, al promediar las notas obtenidas en cada una de las partes consideradas. No obstante, se establece un mínimo en la actividad de aprendizaje basado en proyectos, en la que será necesario obtener una nota igual o mayor a 5 puntos para optar a aprobar la asignatura. En el caso de suspender el proyecto, la asignatura estará directamente suspensa (con independencia de las notas obtenidas en la resolución de ejercicios y problemas).

- **Convocatoria extraordinaria:**

En la convocatoria extraordinaria se mantienen los criterios de calificación anteriores. Cabe destacar que se preservará la calificación obtenida por los alumnos en cada parte aprobada en la convocatoria ordinaria, de forma que sólo tendrán que superar aquellas partes suspensas. En el caso de no haber aprobado la actividad de aprendizaje basado en proyectos, los alumnos resolverán un nuevo proyecto, cuyo alcance y criterios de calificación serán equivalentes a los de la convocatoria ordinaria.

(\*) Se entiende por convocatoria extraordinaria la segunda convocatoria.

**Art 35.4 del ROA 35.4. La participación en la convocatoria extraordinaria no quedará sujeta a la asistencia a clase ni a la presencia en pruebas anteriores, salvo en los casos de prácticas externas, laboratorios u otras actividades cuya evaluación no fuera posible sin la previa realización de las mencionadas pruebas.**

<https://secretariageneral.uva.es/wp-content/uploads/VII.2.-Reglamento-de-Ordenacion-Academica.pdf>

## 8. Consideraciones finales