

**MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN  
PROPUESTA DE CURSO DE POSGRADO**

<b>1- DATOS GENERALES DE LA ACTIVIDAD CURRICULAR</b>	
1.1 Título del Curso	Big Data: Procesos, componentes y herramientas
1.2 Área temática <sup>1</sup>	INGENIERÍA DE SOFTWARE

<b>2- COMPOSICIÓN DEL EQUIPO DOCENTE</b>	
2.1 Responsable a cargo de la actividad curricular	Dra. AGUSTINA BUCCELLA
2.2 Docentes	

<b>3- CARGA HORARIA</b>					
Carga horaria teórica	30 hs				
Carga horaria práctica	30 hs				
Carga horaria total	60 hs				
Distribución horaria semanal	Lu	Ma	Mie	Jue	Vie
Fecha de inicio sugerida					

<b>4- BREVE RESUMEN DE CONTENIDOS (hasta 400 palabras)</b>
<p>Big Data es un campo dedicado a tareas de análisis, procesamiento y almacenamiento de grandes conjuntos de datos que se originan desde diferentes fuentes. Generalmente se utilizan las soluciones de Big Data cuando las técnicas tradicionales son insuficientes para estas tareas. Existen en la actualidad muchas metodologías propuestas para llevar a cabo un proceso de desarrollo para Big Data. En general, todas ellas convergen en tres grandes campos de trabajo involucrando la limpieza y/o preprocesamiento, almacenamiento y la analítica de los datos. Los dos primeros campos pueden ser intercambiados dependiendo del enfoque utilizado. Es decir, para realizar un desarrollo orientado al análisis de los datos, se debe primero seleccionar la/s fuentes de información que puedan resultar útiles, procesarlas para que tengan sentido, y luego almacenarlas en algún tipo de repositorio, en la forma de depósitos de datos o lagos de datos. La característica principal de estos repositorios es que actúan como centros de información en donde se vuelcan, en algún formato específico, los datos de todas las fuentes que se deseen explotar en un proceso de extracción, transformación y carga (ETL). Luego, la analítica de datos (data analytics) se dedica al proceso de crear información desde los datos fuente por medio de la contextualización, análisis y gobernanza de datos. A su vez, en los últimos años han surgido una serie de estándares que normalizan aspectos como las arquitecturas de referencia, la interoperabilidad, terminología común, servicios requeridos, etc. Estos esfuerzos de estandarización han sido llevados a cabo por el comité ISO/IEC JTC 1/SC 42 (Artificial intelligence) y por el grupo de trabajo del NIST (Big Data Public Working Group - NBD-PWG). Se han presentado muchas aplicaciones de Big Data en la actualidad en dominios específicos, todas ellas realizando análisis interesantes que permiten explotar la información de manera de clasificarla, relacionarla y/o predecir nuevos comportamientos o sucesos.</p> <p>De esta manera, este curso abarca cada una de las actividades involucradas en el campo de Big Data</p>

<sup>1</sup> Corresponde a uno de los siguientes tópicos: Algoritmos y Lenguajes; Teoría de la Computación; Ingeniería de Software, Bases de Datos y Sistemas de Información; Arquitecturas, Sistemas Operativos y Redes.

comenzando desde aquellas actividades ya conocidas en otras áreas de la ingeniería de software y las bases de datos, para ver su adaptación y extensión en la gestión y explotación de grandes volúmenes de datos.

**Contenidos Mínimos:** Concepto y terminología de Big Data. Características, Ejemplos de Aplicación, Ciclo de Vida. Arquitectura y Estándares. Almacenamiento de grandes volúmenes. Diseño e implementación de Lagos de Datos. Procesamiento y análisis de datos. Técnicas de análisis de grandes volúmenes de datos.

#### 5- CONOCIMIENTOS PREVIOS REQUERIDOS

Se requieren conocimientos en depósitos y minería de datos, como también conocer conceptos principales de las arquitecturas de Software.

#### 6- OBJETIVOS

El objetivo general del curso es conocer y profundizar la base metodológica para el desarrollo de aplicaciones de Big Data junto con las tendencias tecnológicas y conceptuales basadas en el almacenamiento y analítica de los datos. Los tópicos se centrarán en paradigmas de desarrollo bien establecidos, analizando los estándares actuales y a su vez las tendencias futuras. La asignatura tendrá por objetivo el comprender y aplicar - en lo posible - los conceptos y técnicas desde la ingeniería de software y las bases de datos basadas en las actividades principales de las metodologías de Big Data que hoy son relevantes.

##### Objetivos Específicos

- Entender aspectos básicos de las actividades involucradas en todo proceso de Big Data.
- Conocer los esfuerzos de estandarización actuales sobre varios aspectos de Big Data.
- Conocer y comprender las particularidades de las metodologías de Big Data así como aspectos centrados en su diseño e implementación a nivel arquitectural y tecnológico.
- Entender factores de riesgos, ventajas y desventajas de cada una de las actividades involucradas
- Conocer y aplicar metodologías de Big Data basadas en dominios de la realidad

#### 7- CONTENIDOS (organizados en unidades, ejes, módulos, otros)

##### Unidad I: INTRODUCCIÓN A BIG DATA

Concepto de Big Data y Analíticas. Terminología. Características: Volumen, Velocidad, Variedad, Veracidad, Valor. Ciclo de Vida. Analítica de Datos: descriptiva, diagnóstica, predictiva y prescriptiva. Ejemplos. Casos de Estudio

##### Unidad II: ESTÁNDARES EN BIG DATA

Introducción a los estándares existentes. Familia de estándares ISO/IEC JTC 1/SC 42 (Artificial intelligence). Familia de estándares propuestos por el NIST (Big Data Public WorkingGroup - NBD-PWG). Vocabulario. Arquitectura de referencia. Casos de Uso.

##### Unidad III: ALMACENAMIENTO DE DATOS

Conceptos del almacenamiento de grandes volúmenes de datos. Tipos de fuentes y tipos de almacenamiento. NoSQL (basadas en documentos, clave-valor, familia de columnas, grafos). New SQL, Distribuidas. Depósitos de Datos. Lago de Datos. Diseño de los Lagos de Datos. Tecnologías asociadas.

##### Unidad IV: PROCESAMIENTO Y ANÁLISIS DE DATOS

Conceptos y tipos de procesamiento: paralelo, distribuido, batch (map-reduce), transaccional, tiempo real. Técnicas de Análisis: cuantitativo, cualitativo, minería, estadístico, machine learning, análisis semántico y visualización. Ejemplos. Aplicaciones. Tecnologías asociadas.

**Tecnologías para Big Data:**

**Unidad III:** HDFS, Apache YARN, MongoDB, Redis, Cassandra, OrientDB.

**Unidad IV:** Apache Spark. Optimus. Hadoop-MapReduce, Spark MLib, Pandas, NumPy, Matplotlib

**8- PROPUESTA DIDÁCTICA (metodología de trabajo de clases teóricas y prácticas)**

La asignatura se organiza en torno a clases teórico/prácticas. Las mismas inician desde conceptos teóricos generales hasta aquellos más específicos y complejos. Para las clases, el alumno debe disponer con anterioridad de la documentación básica que se explicará, así como de los ejercicios o prácticas asociadas. En general cada material teórico brindado posee una guía de los temas que se van dictando en cada clase, así como de la bibliografía utilizada.

Para fomentar la participación y facilitar la comprensión de temas, se propone realizar al finalizar cada contenido, ejercicios prácticos de laboratorio en donde los alumnos puedan aplicar conceptos vistos y participar activamente.

En cuanto a la práctica, cada unidad temática tiene asociado uno o más trabajos prácticos de elaboración individual o grupal. Las clases se centran en atención de consultas y resolución de ejercicios con participación activa de estudiantes.

A su vez, como los alumnos deben presentar reportes teórico/prácticos, los trabajos prácticos realizados conducen a la elaboración de dichos entregables.

Como herramienta principal para la comunicación entre docentes-alumnos se utilizará el sistema PEDCO (Plataforma de Educación a Distancia) basado en el ambiente educativo virtual Moodle. En la misma se crean foros para permitir a los alumnos interactuar entre ellos (con los docentes como tutores) y resolver ejercicios de mayor nivel de complejidad.

**9- MODALIDAD DE EVALUACIÓN Y CONDICIONES DE ACREDITACIÓN<sup>2</sup>**

Certificación de Postgrado: 80% de asistencia, realización de los trabajos prácticos propuestos en clase y elaboración de un trabajo final que involucra tareas de investigación y que deberá ser realizado una vez terminado el curso (cuyo tema y fecha de entrega se acordará con los participantes). La nota de los trabajos obligatorios y del trabajo final debe ser no menos a 7 (siete) puntos.

**10- BIBLIOGRAFÍA DE LECTURA OBLIGATORIA CORRESPONDIENTE A CADA UNIDAD Y GENERAL**

---

<sup>2</sup> Son condiciones mínimas para la aprobación de todos los cursos: cumplir con un mínimo del 80% de asistencia a las clases, realizar las tareas y aprobar las evaluaciones que se hayan propuesto en el programa, con una calificación no menor a 7 (puntos). Los trabajos de evaluación pautados y la calificación de los alumnos deberán realizarse dentro de los 60 días posteriores a la finalización del curso.

Big Data Fundamentals: Concepts, Drivers & Techniques. Thomas Erl, Wajid Khattak and Paul Buhler. 1st Edición. ISBN: 978-0134291079. Prentice Hall 2016

Big Data Analytics: A Hands-On Approach. Arshdeep Bahga and Vijay Madiseti. Published by Arshdeep Bahga & Vijay Madiseti. ISBN-10: 9781949978001. 2019

Big Data for Dummies. Judith Hurwitz, Alan Nugent, Fern Halper and Marcia Kaufman. ISBN: 978-1118504222. 1st Edition. Editor: For Dummies. 2013.

The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science By Alex Gorelik. ISBN: 978-1491931554. O'Reilly Media, Inc. 2019

McKinney Wes. 2012. Python for data analysis (1st. ed.). O'Reilly Media, Inc.

A. Vaisman and E. Zimnyi. Data Warehouse Systems: Design and Implementation, 1st ed. Springer Publishing Company, Incorporated, 2016.

C. Quix and R. Hai. Data Lake. Springer International Publishing, 2018, pp. 1–8. <https://doi.org/10.1007/978-3-319-63962-87-1>

Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques (3rd. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Revistas actuales para analizar en el transcurso del dictado

### **Estándares:**

Familia de estándares ISO/IEC JTC 1/SC 42 (Artificial intelligence) - <https://www.iso.org/committee/6794475.htm>

Familia de estándares propuestos por el NIST (Big Data Public WorkingGroup - NBD-PWG) - <https://bigdatawg.nist.gov/>

### **Herramientas:**

<https://spark.apache.org/>

<https://hadoop.apache.org/>

<https://redis.io/>

<https://www.mongodb.com/>

<http://cassandra.apache.org/>

<https://orientdb.com/>

<https://pandas.pydata.org/>

<https://hi-optimus.com/>

<https://numpy.org/>

<https://matplotlib.org/>

<b>11- INFRAESTRUCTURA E INSUMOS REQUERIDOS<sup>3</sup></b>
---

Proyector multimedia, pantalla, acceso a Internet
---

<b>12 – OTRA INFORMACIÓN RELEVANTE</b>
--

---

<sup>3</sup> Deberá constar aquí si la realización del curso requiere contar con instalaciones especiales (laboratorio, sala de informática, equipamiento audiovisual, etc). Explicitar si se estima que el curso debe tener un número máximo determinado de asistentes para poder ser dictado.