

Conceptos y aplicaciones de big data**Año 2024****Carrera/ Plan:**

Licenciatura en Informática Plan2021/Plan 2015/Plan 2012
Licenciatura en Sistemas Plan2021/Plan 2015/Plan 2012
Analista en Tecnologías de la Información y la Comunicación
Plan2021/Plan 2017

Año: 4to (Lic. en Informática y Lic. en Sistemas)
3ro (Analista en TIC)

Régimen de Cursada: Semestral

Carácter (Obligatoria/Optativa): Optativa

Correlativas: Programación Concurrente (Lic. en Informática y
Lic. en Sistemas)

Programación Concurrente ATIC (Analista en TIC)

Profesor/es: Dr. Lic. Waldo Hasperué

Hs. semanales teoría: 2 hs

Hs. semanales práctica: 4 hs

FUNDAMENTACIÓN

La tecnología actual permite el almacenamiento de grandes volúmenes de información para ser analizados en búsqueda de patrones, características y conocimiento. A medida que la dimensión y el volumen de la información crece, llega un punto donde no es posible utilizar las estrategias convencionales de análisis de datos ya que aparecen problemas adicionales. Estos problemas básicamente son dos: por un lado la RAM no es suficiente para el manejo eficaz de un gran volumen de datos y por otro lado es necesaria una gran cantidad de tiempo para la ejecución de los algoritmos, la cual muchas veces no es aceptable.

El término Big Data surge ante la necesidad de resolver problemas de análisis de datos cuando el volumen a analizar es enorme y no puede realizarse con las estrategias convencionales. En el manejo de Big Data deben tenerse en cuenta los siguientes tres conceptos: Volumen, Variedad y Velocidad, todos ellos tienen un fuerte impacto en la manera en que debe ser procesada la información.

Un problema puntual de Big Data es cuando se poseen varios sensores que generan un continuo flujo de datos (data streaming) y que por razones técnicas es imposible de almacenar la gran cantidad de datos que proveen dichos sensores. Para resolver estos problemas uno de los enfoques que se está utilizando es el de implementar estrategias que permitan el análisis del flujo de datos. Este enfoque propone que cada uno de los datos que se recibe del flujo se pueda utilizar solo una vez, para ello la estrategia a ejecutar trabaja de la siguiente manera: se recibe un dato, se usa para llevar a cabo la tarea y finalmente se descarta el dato y no se vuelve a analizar.

Big Data se refiere a las técnicas de software y hardware que permiten realizar análisis de enormes cantidades de datos heterogéneos y que necesitan ser analizados en tiempo real o muy próximo a éste. Existen numerosos procesos y fuentes de datos que dan origen a este tipo de información desestructurada y seguramente continuará esa tendencia cuando se explote la información producida por el área conocida como Internet de las cosas (Internet of Things, o IoT) fundamentado en la creciente necesidad de las personas por conectarse a través de dispositivos.

Las técnicas desarrolladas para problemas de Big Data pueden ser aplicadas a distintas áreas tales como: análisis de información de ADN, smart cities para el desarrollo sustentable y alta calidad de vida, análisis de resultados de simulaciones, análisis de Grandes Datos provenientes de información geográfica, detección de fraudes online y cyber ataques, recomendaciones personalizadas a los clientes, etc.

A raíz de esto, surgen desafíos importantes en el aprovechamiento de la gran cantidad de datos, incluyendo retos en las capacidades del sistema y en el diseño de algoritmos.

Los objetivos de este curso son presentar al alumno las tecnologías actuales del diseño e implementación de aplicaciones en Big Data profundizando los conocimientos presentados con experimentación con trabajos en máquina.

OBJETIVOS GENERALES

- Estudiar los conceptos y fundamentos de Big Data.
- Analizar los principales problemas en las aplicaciones de Big Data
- Estudiar los frameworks actuales para el desarrollo de soluciones en Big Data
- Resolver problemas de Big Data utilizando procesamiento de flujos de datos.

COMPETENCIAS

- LI-CE4- Planificar, dirigir, realizar y/o evaluar proyectos de relevamiento de problemas del mundo real, especificación formal de los mismos, diseño, implementación, prueba, verificación, validación, mantenimiento y control de calidad de sistemas de software/sistemas de información que se ejecuten sobre equipos de procesamiento de datos, con capacidad de incorporación de tecnologías emergentes del cambio tecnológico. Capacidad de análisis, diseño y evaluación de interfases humano computador y computador-computador.

- LS-CE1- Planificar, dirigir, realizar y/o evaluar proyectos de relevamiento de problemas del mundo real. Especificación formal, diseño, implementación, prueba, verificación, validación, mantenimiento y control de calidad de sistemas de software que se ejecuten sobre sistemas de procesamiento de datos, con capacidad de incorporación de tecnologías emergentes del cambio tecnológico. Capacidad de análisis, diseño y evaluación de interfases humano computador y computador-computador.

CONTENIDOS MINIMOS (de acuerdo al Plan de Estudios)

- Conceptos de Big Data.
- Aplicaciones de Big Data sobre Cloud.
- Herramientas: Hadoop, Mapreduce, Spark

PROGRAMA ANALÍTICO

A. Fundamentos de Big Data

- Definición y dimensiones en Big Data.
- Aplicaciones de Big Data.
- Modelos de datos y modelos de procesamiento en Big Data
- Ética, seguridad, privacidad en Big Data.
- Casos de uso. IoT

B. Framework Hadoop MapReduce

- Paradigma MapReduce.
- Ecosistema Hadoop.
- Etapas map, reduce, shuffle y sort.

C. Framework Spark

- API de spark.
- Transformaciones y acciones.
- DAG de ejecución.
- Spark SQL

D. Algoritmos de Machine Learning con Big Data

- Big Data sobre documentos escritos en lenguaje natural.
- Machine Learning. MLlib. TensorFlow.

E. Spark Streaming

- Procesamiento de flujos de datos
- Tipos de ventanas temporales
- Armado de modelos de datos

BIBLIOGRAFÍA

- Judith Hurwitz, Alan Nugent, Dr. Fern Halper and Marcia Kaufman. Big Data for dummies. John Wiley & Sons, Inc. ISBN 978-1-118-50422-2. 2013.
- Karau, H.; Konwinski, A.; Wendell, P. & Zaharia, M. Learning Spark. O'Reilly Media, Inc. ISBN: 978-1-449-35862-4. 2015.
- Tom White. Hadoop: The Definitive Guide. Hadoop: The Definitive Guide. ISBN 978-1-449-38973-4. 2011.
- Gerard Maas & François Garillot. Stream Processing with Apache Spark. O'Reilly Media, Inc. ISBN: 978-1-491-94424-0. 2019.
- Charu C. Aggarwal. Data streams: models and algorithms. Springer US. ISBN 978-0-387-28759-1. 2007

BIBLIOGRAFÍA COMPLEMENTARIA

- Soumendra Mohanty, Madhu Jagadeesh and Harsha Srivatsa. Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics. Apress. ISBN 978-1430248729. 2013
- James A. Scott. Getting Started with Apache Spark. MapR Technologies, Inc., 2015.
- O'Reilly Media, Inc. Big Data Now. O'Reilly Media, Inc. ISBN 978-1-449-35671-2. 2012
- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. Cambridge University Press. ISBN 978-1107015357. 2011
- Kord Davis with Doug Patterson. Ethics of Big Data. O'Reilly Media, Inc. ISBN 978-1-449-31179-7. 2012
- Leskovec, J. et al. (2014). Mining of massive datasets. Cambridge University Press.
- Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity and Variety, META Group Original Research Note.

- Shanmuganathan, S. (2014). From data mining and knowledge discovery to big data analytics and knowledge extraction for applications in science.
- Kaisler, S. et al. (2013). Big data: Issues and challenges moving forward. In System Sciences (HICSS), 46th Hawaii International Conference on (pp. 995-1004). IEEE.
- Chalmers, S. et al. (2013). Big Data-State of the Art.
- Bajpayee, R. et al. (2015). Big Data: A Brief investigation on NoSQL Databases.
- Akay, A. et al. (2015). Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care. Biomedical and Health Informatics, IEEE Journal.

METODOLOGÍA DE ENSEÑANZA

1) Modalidad presencial

La asignatura se estructura con clases teórico-prácticas y prácticas experimentales.

- Las clases teórico-prácticas son dictadas por el profesor de la asignatura y son obligatorias para la promoción.
- Las explicaciones de práctica son introductorias al trabajo en Laboratorio, para facilitar la utilización del equipamiento y software por los alumnos. Se desarrollan en las clases teórico-prácticas.
- Se analizarán en clase los diferentes frameworks y paradigmas para el desarrollo de aplicaciones en Big Data.
- Se propone la resolución de problemas de Big Data, utilizando los frameworks estudiados. Analizar las soluciones en función del rendimiento en tiempo y calidad de las respuestas. En principio con información textual.
- Los trabajos se pueden realizar individualmente o en grupo de 2 personas. Las consultas y correcciones son realizadas en forma presencial o por medio de la plataforma de Educación a Distancia de la UNLP (IDEAS).

2) Modalidad no presencial

El alumno tendrá todo el material del curso a disposición por medio del entorno IDEAS.

Los alumnos deben aprobar las mismas entregas de los trabajos experimentales que los alumnos en modalidad presencial.

EVALUACIÓN

1) Modalidad presencial:

Los alumnos deberán tener más del 70% de asistencia a clase y aprobar los diferentes trabajos prácticos. Estos TPs deberán ser entregados acorde a un cronograma de fechas que la cátedra publicará oportunamente.

Además deberá aprobar un examen individual de manera escrita.

2) Modalidad no presencial

Deben cumplir y aprobar las entregas de los mismos trabajos experimentales que en la modalidad presencial. Además deberán defender las entregas en un coloquio en una fecha a convenir.

Deben rendir un examen final, que puede incluir un trabajo final experimental, pero necesariamente contendrá preguntas de la Teoría presencial.

CRONOGRAMA DE CLASES Y EVALUACIONES

Clase	Fecha	Contenidos/Actividades
1	13/08	Presentación de la materia. Conceptos de Big Data. Definición. Alcance. Tecnologías.
2	20/08	Ecosistema Hadoop. HDFS.
3	27/08	MapReduce. Introducción al paradigma. Etapas Map y Reduce, Shuffle y Sort.
4	03/09	MapReduce. Múltiples Jobs. Tipos y Formatos. Combiners.
5	10/09	Entendiendo y visualizando el DAG
6	17/09	Spark. Funcionamiento y uso. Conceptos. RDDs.
7	24/09	Spark. Acciones y transformaciones
8	01/10	Algoritmos iterativos en Spark
9	08/10	SparkSQL. Persistencia
10	15/10	Stream processing. Introducción al paradigma. Spark streaming. Modelos con ventana temporal de datos.
11	22/10	Spark streaming. DStream.
12	29/10	Spark streaming. Creación de un modelo de los datos a partir de un flujo. Persistencia temporal
13	05/11	MLlib. Algoritmos básicos.
14	12/11	MLlib. Sistemas de recomendación

Evaluaciones previstas	Fecha
Trabajo integrador 1	10/09
Trabajo integrador 2	08/10
Trabajo integrador 3	26/11
Examen escrito	03/12

Contacto de la cátedra (mail, sitio WEB, plataforma virtual de gestión de cursos):

Toda la información de la cátedra, durante la cursada, se refleja en el entorno virtual de enseñanza y aprendizaje IDEAS (<https://ideas.info.unlp.edu.ar>), donde los alumnos tendrán acceso a través de un nombre de usuario y una clave, para trabajar con sus docentes y compañeros de curso.

Los alumnos podrán consultar temas específicos de práctica o teoría mediante la herramienta de mensajería provista en el curso sobre IDEAS.

**Facultad de
Informática**



**UNIVERSIDAD
NACIONAL
DE LA PLATA**

Los docentes de la cátedra atenderán consultas durante las clases sincrónicas.

Email de contacto: whasperue@lidi.info.unlp.edu.ar