ESPECIALIZACIÓN EN INTELIGENCIA DE DATOS APLICADA
PROPUESTA DE CONTENIDOS

1-	**DATOS GENERALES DE LA ACTIVIDAD CURRICULAR**
1.1 Título del Curso	EXTRACCIÓN, PREPARACIÓN Y ALMACENAMIENTO DE LOS DATOS
1.2 Tipo de Curso[1]	OBLIGATORIO

2-	**COMPOSICION DEL EQUIPO DOCENTE**
2.1 Responsable a cargo de la actividad curricular	Dra. Agustina Buccella
2.2 Docentes	

3-	**CARGA HORARIA**				
Carga horaria teórica	15 hs				
Carga horaria práctica	25 hs				
Carga horaria total	40 hs				
Distribución horaria semanal	Lu	Ma	Mie	Jue	Vie
Fecha de inicio sugerida					

4-	**BREVE RESUMEN DE CONTENIDOS (hasta 400 palabras)**
Captura de datos. Fuente de datos. Captura de datos en la Web. Almacenamiento en bases de datos relacionales y no relacionales. Repositorios NoSQL. Depósitos de datos. Limpieza de datos. Lagos de datos.	

5-	**CONOCIMIENTOS PREVIOS REQUERIDOS**
Cursos aprobados: B1, B2, B3	

6-	**OBJETIVOS**
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 procesamiento de los datos. La asignatura tendrá por objetivo comprender y aplicar, con las tecnologías brindadas, las técnicas utilizadas en las actividades principales de las metodologías de Big Data que hoy son relevantes.	

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

[1] Corresponde Básico, Obligatorio, Optativo.

PROGRAMA ANALÍTICO:

Unidad I: IDENTIFICACIÓN Y RECOLECCIÓN DE DATOS
Ingesta de datos. Tipos de datos fuente. Extracción de datos de la Web (Web Scraping, Web Crawling). Patrones de ingesta de datos en el contexto de Big Data.

Unidad II: PREPARACIÓN DE DATOS
Calidad de Datos. Limpieza de Datos. Ruido y detección de anomalías en los datos. El proceso ETL y ELT. Integración de los datos.

Unidad III: ALMACENAMIENTO DE DATOS
Tipos de almacenamiento. ACID-CAP-BASE. NoSQL, Distribuidas. Depósitos de Datos vs Lago de Datos. Conceptos del almacenamiento de grandes volúmenes de datos.

Tecnologías para Big Data:
Unidad I: Librerías de python (urllib, beautifulsoup,pandas).

Unidad II: Librerías de python (pyspark, matplitlib,numpy).

Unidad III: BD Relacionales MySQL, MariaDB orientada a columnas.

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, los/as estudiantes deben 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 al menos un trabajo práctico de elaboración individual o grupal. Las clases se centran en atención de consultas y resolución de ejercicios con participación activa de estudiantes. Se plantean 3 trabajos prácticos para plasmar los conceptos teóricos vistos. El Práctico 1 está orientado al análisis de las fuentes de información disponibles y las formas de extraerlas, analizado su disponibilidad, formatos, estructuras, etc. Luego el Práctico 2 posee ejercicios prácticos de preparación de los datos para la búsqueda de anomalías, nulos y/o información inconsistente y las formas de cómo tratarlos. Por último, en el Práctico 3 se trabaja con los diferentes sistemas de almacenamiento de datos para analizar en forma práctica las formas de guardar la información analizada en el práctico anterior junto con los objetivos, ventajas y desventajas de cada uno.

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

9- MODALIDAD DE EVALUACIÓN Y CONDICIONES DE ACREDITACIÓN[2]

Mediante la realización de los trabajos prácticos propuestos en clase y la elaboración de un trabajo final que deberá ser realizado una vez entregado al finalizar el curso (cuyo tema y fecha de entrega se acordará con los participantes).

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

[2] 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.

1. Big Data Fundamentals: Concepts, Drivers & Techniques. Thomas Erl, Wajid Khattak and Paul Buhler. 1st Edición. ISBN: 978-0134291079. Prentice Hall 2016
2. 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
3. Python for data analysis (1st. ed.). McKinney Wes. O'Reilly Media, Inc. ISBN: 9781449319793. 2012
4. Mining the Social Web. Matthew A. Russell. 1st Edición.ISBN: 9781449388348. O'Reilly Media. 2011

Herramientas:

https://spark.apache.org/docs/latest/api/python/
https://docs.python.org/3/library/urllib.html
https://beautiful-soup-4.readthedocs.io/en/latest/
https://pandas.pydata.org/
https://matplotlib.org/
https://mariadb.com/kb/en/mariadb-columnstore/

11- INFRAESTRUCTURA E INSUMOS REQUERIDOS[3]

Proyector multimedia, pantalla, acceso a Internet

12 – OTRA INFORMACIÓN RELEVANTE

[3]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.