I. IDENTIFICACIÓN

1. Asignatura : **Electiva III - Big Data**
2. Código :**7.2.B**
3. Horas semanales : 5 horas
4. Total de horas : 80 horas

II. JUSTIFICACIÓN

En la actualidad, las aplicaciones y dispositivos conectados constantemente a las redes de telecomunicaciones generan una cantidad de datos sin precedentes, a partir de los cuales se podría extraer información valiosa de utilidad tanto para las empresas que buscan maximizar sus ganancias, para las instituciones del estado que buscan mejorar la calidad de vida de los ciudadanos, así como para las distintas ramas de la ciencia que buscan correlacionar datos para encontrar soluciones a los problemas de la humanidad.

El volumen y diversidad de los datos generados superan las capacidades de las tecnologías tradicionales de almacenamiento y procesamiento de datos, por lo que en la actualidad existe mucho esfuerzo de investigación dedicado a encontrar nuevas tecnologías que sean capaces de atacar esta problemática, a la cual se denomina Big Data.

El énfasis del curso será el revisar los conceptos, técnicas y herramientas para la captura, el almacenamiento, el procesamiento y la generación de información a partir de datos, que por su volumen y complejidad estructural, son considerados Big Data.

III. OBJETIVOS GENERALES

1. Conocer los desafíos que presenta el procesamiento de datos en volúmenes considerados Big Data.
2. Conocer y aprender el uso de las herramientas de procesamiento y almacenamiento masivamente paralelo existentes para resolver la problemática Big Data.
3. Conocer como las herramientas orientadas a Big Data resuelven internamente los desafíos que esta problemática plantea.
4. Aprender a aplicar algoritmos y conceptos de Data Mining eficientemente sobre infraestructuras orientadas a Big Data.
5. Conocer los beneficios comerciales y usos prácticos que puede aportar el análisis de datos a escala de Big Data.

IV. OBJETIVOS ESPECÍFICOS

A. Conocimientos

1. Plataformas y sus arquitecturas para almacenamiento viable de datos a escala Big Data.
2. Herramientas para el procesamiento de datos almacenados en plataformas Big Data.
3. Paradigmas de programación aplicables a plataformas orientadas a Big Data.
4. Algoritmos de Data Mining típicamente aplicados a datos a escala Big Data.
5. Implementación de algoritmos de Data Mining sobre plataformas Big Data.
6. Conocimiento que puede extraerse desde las distintas fuentes de datos Big Data, y sus aplicaciones.

B. Habilidades

1. Evaluar y seleccionar las herramientas y plataformas más adecuadas según los requerimientos de almacenamiento, análisis de datos, integración, costos y gobierno de la infraestructura.
2. Dimensionar, costear y diseñar la arquitectura Big Data en función a los requerimientos de almacenamiento y procesamiento.
3. Diseñar procesos de obtención de datos y almacenamiento masivo en plataformas orientadas a Big Data.
4. Diseñar y programar procesos de análisis de datos a escala Big Data con altos niveles de rendimiento.
5. Diseñar esquemas de almacenamiento de datos a escala Big Data con altos niveles de seguridad de la información.
6. Diagnosticar y corregir problemas de rendimiento en procesos de análisis en la plataforma Big Data.
7. Detectar vulnerabilidades de seguridad en la plataforma Big Data, y proponer acciones paliativas.

C. Competencias

1. Capacidad de aplicar los conocimientos en la práctica.
2. Disposición para el trabajo en equipo.
3. Capacidad de abstracción, análisis y síntesis y presentaciones orales.
4. Habilidades para buscar, procesar y analizar información procedente de fuentes diversas.
5. Capacidad para identificar, plantear y resolver problemas
6. Capacidad de comunicación oral y escrita.

V. PRE – REQUISITO

1. Ingeniería de Software I.
2. Base de Datos III.

VI. CONTENIDO

6.1. Unidades programáticas

1. Desafíos de Big Data
2. Limpieza e integración de datos
3. Plataformas propias y en la nube
4. Bases de datos modernas
5. Plataformas de cómputo distribuido
6. NoSQL
7. Algoritmos de *Data Mining* sobre infraestructuras orientadas a *Big Data*

6.2. Desarrollo de las unidades programáticas

1. Desafíos de *Big Data*
    - 1.1. Características de los datos
    - 1.2. Características de las operaciones ejecutadas sobre los conjuntos de datos
    - 1.3. Limitaciones de las bases de datos convencionales
    - 1.4. Alternativas a las bases de datos convencionales
        - 1.4.1. Bases de datos columnares
        - 1.4.2. Bases de datos NoSQL
        - *1.4.3.*Sistemas de archivos distribuidos
2. Limpieza e integración de datos
    - 2.1. Importancia de la "limpieza de datos" (Data Cleansing)
    - 2.2. Técnicas y herramientas para extracción, transformación y carga de datos en repositorios orientados a Big Data, con capacidades de cómputo distribuido
3. Plataformas propias y en la nube
    - 3.1. Comparativa entre adopción de plataforma Big Data propia y montaje en la nube
    - 3.2. Herramientas y técnicas para la migración paulatina de infraestructura propia a la nube y viceversa
4. Bases de datos modernas
    - 4.1. Bases de datos columnares
    - 4.2. Bases de datos NoSQL
    - 4.3. Bases de datos distribuidas
    - 4.4. Bases de datos de tipo clave-valor
5. Plataformas de cómputo distribuido
    - 5.1. Paradigmas y patrones de diseño para el uso de herramientas de programación de procesos analíticos distribuidos.
        - 5.1.1. Map Reduce
        - 5.1.2. Hadoop
        - 5.1.3. Spark
        - 5.1.4. Hive
    - 5.2. Arquitectura y algoritmos de las herramientas de cómputo y almacenamiento distribuido.
6. Datos no estructurados
    - 6.1. Características de los datos no estructurados
    - 6.2. Casos de estudio de datos no estructurados
    - 6.3. Análisis de las operaciones frecuentes sobre datos no estructurados
    - 6.4. Herramientas para el almacenamiento y análisis eficiente de datos no estructurados
7. Algoritmos de *Data Mining* sobre infraestructuras orientadas a Big Data
    - 7.1. Implementación de algoritmos característicos de Data Mining (machine learning, clustering, etc) sobre infraestructuras orientadas a Big Data.

VII. ESTRATEGIAS METODOLÓGICAS

1. Exposición de fundamentos teóricos en clase por parte del profesor.
2. Trabajos de investigación y evaluación de herramientas y técnicas a desarrollarse fuera del horario de clases.
3. Trabajos de diseño y programación de interfaces de sistemas web a desarrollarse fuera del horario de clases.
4. Evaluación de trabajos en laboratorio, validaciones de caja negra y análisis de código fuente.

A. Medios Auxiliares

1. Pizarras acrílicas.
2. Marcadores.
3. Borrador de pizarra acrílica.
4. Computadoras.
5. Proyectores multimedia.
6. Parlantes para multimedia.

7. Plataforma virtual "EDUCA".
8. Sala de laboratorio equipada para las prácticas.
   8.1. Computadoras en red.
   8.2. Sistemas operativos Linux, Windows.
   8.3. Acceso a internet.

VIII. EVALUACIÓN

1. Para evaluar la asignatura se tienen en cuenta lo siguiente:
   1. Exámenes parciales de teoría con un % asignado.
   2. Examen final de teoría con un % asignado.
   3. Algunos trabajos prácticos tienen un % asignado.
2. Las calificaciones se basan en el reglamento de la Universidad.
3. Es imprescindible la entrega de todas las prácticas para poder calcular la nota de prácticas.

IX. BIBLIOGRAFÍA

A. Básica

- Nathan Marz, James Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning (2015).
- Hrushikesha Mohanty, Prachet Bhuyan, Deepak Chenthati. Big Data: A Primer (Studies in Big Data). Springer (2015).
- Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. Advanced Analytics with Spark: Patterns for Learning from Data at Scale. O´Reilly (2015).
- Daniel T. Larose, Chantal D. Larose. Data Mining and Predictive Analytics. Wiley (2015).
- Tom White. Hadoop: The Definitive Guide. O´Reilly (2015).

B. Complementaria

- Holden Karau. Learning Spark: Lightning-Fast Big Data Analysis. O´Reilly (2015).
- W.H. Inmon. Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault. Morgan Kaufmann (2015).
- Joel Grus. Data Science from Scratch: First Principles with Python. O´Reilly (2015).
- Aravind Shenoy. Hadoop Explained. Packt Publishing (2014).
- Foster Provost. Data Science for Business: What you need to know about data mining and data-analytic thinking. O´Reilly (2013).
- Bernard Marr. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. Wiley (2015).

C. Enlaces Web

- http://spark.apache.org/
- https://hadoop.apache.org/
- http://research.google.com/archive/mapreduce.html
- https://www.r-project.org/
- https://www.vertica.com/
- http://greenplum.org/
- https://hive.apache.org/