**UNIVERSIDAD EXTERNADO DE COLOMBIA**

**FACULTAD DE ECONOMÍA**

**PROGRAMA DE CURSO**

**2022-II**

|  |  |
| --- | --- |
| **IDENTIFICACIÓN GENERAL DEL CURSO** | |
| **Nombre de la materia** | Big Data |
| **Profesor** | Julio Antonino Zainea Maya |
| Correo de contacto |  |
| Horario de atención a estudiantes |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **A diligenciar por la Secretaría académica** | | | | | | | |
| Código de la materia |  | | | | | | |
| Número de créditos |  | Grupo |  | Salón |  | Semestre | **III** |
| Horario de clase | Sábados de 7:00 am a 10:00 am | | | | | | |

|  |
| --- |
| Fundamentación |
| La analítica del Big Data es un campo orientado al análisis, procesamiento y almacenamiento de grandes colecciones de datos que, con frecuencia, provienen de diversas fuentes. Particularmente, aborda distintos requerimientos, como la combinación de múltiples conjuntos de datos no relacionados, el procesamiento de grandes cantidades de datos sin estructurar y la recopilación de información con plazos de tiempo definidos. En los entornos Big Data la analítica por lo general es aplicada usando tecnologías y plataformas distribuidas y altamente escalables. El termino de Big Data en muchos sentidos es relativo. Lo que para una entidad pueda ser Big Data, para otra no lo es. El término debe utilizarse asociado a una o más de las siguientes características: Volumen, Velocidad, Variedad, Veracidad y Valor. Están interrelacionadas de muchas maneras. La analítica del Big Data es una tendencia tecnológica aún no madura lo que se evidencia por su vaga definición, el limitado número de aplicaciones en producción en la industria, el comercio y el gobierno, o las barreras de seguridad y de privacidad sin resolver para su aplicación generalizada. A pesar de esto, como tendencia tecnológica, es un tema importante de explorar y entender por parte de todo analista de inteligencia de negocios. |
| Objetivo |
| El objetivo principal de este curso es proporcionar a los estudiantes herramientas para entender qué es y hacia dónde va esta tendencia tecnológica y el tipo de herramientas que existen actualmente para ayudar al análisis de datos que se acogen a esta categoría. |
| Competencias |
| Las competencias que desarrollarán los estudiantes son:  Contextual:  Comprender el entorno cambiante de la tecnología asociada a la Big Data, ubicar cuál es el estado del arte y visualizar las alternativas que se desprenden de las tendencias tecnológicas actuales y que puedan responder a las necesidades que atiende la Inteligencia de Negocios.  Comunicativa:  Comunicar con propiedad a los equipos de ciencia de datos y de ingeniería las necesidades que requiere el área de Inteligencia de Negocios respecto a tecnología.  Comprender los artículos divulgativos y académicos de tendencia tecnológica relacionada con Big Data y entender si atienden a las necesidades de su labor profesional.  Cognitiva:  Dimensionar la carga de trabajo y tiempo de desarrollo que implica la inteligencia de negocios en un ambiente que maneja Big Data.  Diferenciar diferentes tipos de datos de gran tamaño y las alternativas tecnológicas disponibles para explotarlos. |
| Estructura temática |
| Por supuesto, aquí tienes el temario corregido:  Semana 1: Introducción a Big Data y GCP  - Introducción a Big Data y su importancia.  - Conceptos fundamentales de datos estructurados y no estructurados.  - Presentación de Google Cloud Platform (GCP) para Big Data.  Semana 2: Small Data (Optimización de la ingesta de datos)  - Estrategias de optimización en la ingesta de datos.  - Taller: Simulación de ingesta y procesamiento de gran cantidad de datos.  Semana 3: Medium Data (Manipulación de datos en memoria y paralelización)  - Manipulación de datos más grandes que la memoria RAM.  - Paralelización de tareas.  - Taller: Procesamiento de datos con Dask.  Semana 4: Procesamiento de Medium Data en Python (Polars y Pandas 2.0)  - Procesamiento de datos a gran escala en Python.  - Estrategias viables: muestreo en la lectura, reducción dimensional, procesamiento estratificado, computación distribuida, manipulación en chunks.  - Uso de Polars y Pandas 2.0.  - Taller: Procesamiento de datos mediante Polars.  Semana 5: ETL e Ingestión de Datos (Velocidad y optimización)  - Tópicos relacionados con la velocidad en ETL e ingestión de datos.  - Taller: Simulación de computación distribuida utilizando el paquete Parallel.  Semana 6: Desarrollo de Estudio de Caso  Semana 7: Computación Distribuida (Hadoop, Spark)  - Conceptos básicos de computación distribuida.  - Análisis de datos con computación distribuida en Python.  - Taller: Computación distribuida en Python con PySpark.  Semana 8: Machine Learning con PySpark  - Machine Learning con PySpark.  - Taller: Práctica de Machine Learning en PySpark.  Semana 9: Uso de Herramientas de Big Data en Google Cloud Platform y Azure  - Introducción y práctica con herramientas de Big Data en GCP y Azure.  - Taller: Práctica en Databricks y DataProc.  Semana 10: Hive y Consultas en Hive  - Introducción a Hive y consultas SQL en el contexto de Big Data.  - Taller: Computación distribuida con SQL.  Semana 11: Introducción a MongoDB en Google Cloud Platform  - Introducción a MongoDB y su uso en GCP.  - Taller: Casos de uso en Azure o GCP.  Semana 12: Exposición de Estudio de Caso |
| Metodología |
| El curso comienza con una introducción básica al Big Data y discute los desafíos técnicos, conceptuales y éticos asociados. Los puntos fuertes y las limitaciones de la explotación de Big Data se discuten en profundidad utilizando ejemplos del mundo real.  Los estudiantes desarrollan y presentan un concepto de Big Data para un caso específico del mundo real.  El programa incorpora talleres con el propósito de brindar a los estudiantes una experiencia inicial y práctica en la manipulación y análisis de estructuras de datos extensas y complejas. Asimismo, se abordan el uso de la nube y las herramientas relacionadas para aplicar estas técnicas. |
| Resultados de aprendizaje |
| Al final de la materia el estudiante estará en capacidad de distinguir si su labor profesional tiene desafíos relacionados con Big Data y el tipo de tecnologías disponibles en el mercado para enfrentarlos. |
| Evaluación |
| Para evidenciar los aprendizajes alcanzados se realizarán talleres y quices. Los porcentajes de cada uno dentro de la nota final serán los siguientes:   * Talleres (2), semanas 4-5 y semana 8-9 35% * Desarrollo de estudio de caso (dos) 35% * Tareas, quices y controles de lectura 30%   Los talleres son prácticas que demuestran la adquisición de los conocimientos y habilidades en los espacios académicos.  Los estudios son sobre temáticas del interés de los estudiantes, ojalá con datos reales. |
| Bibliografía |
| * Warden, P. (2011). "Big Data Glossary: A Guide to the New Generation of Data Tools." O'Reilly Media. * Schmarzo, B. (2015). "Big Data MBA: Driving Business Strategies with Data Science." Wiley. 1ra Edición. * Franco Galeano, M.I. (2018). "Big Data Processing with Apache Spark: Efficiently tackle large datasets and big data analysis with Spark and Python." Packt Publishing. * Holden, K., Konwinski, A., Wendell, P., & Zaharia, M. (2015). "Learning Spark: Lightning-Fast Big Data Analysis." O'Reilly Media. * Singh, P. (2018). "Machine Learning with PySpark: With Natural Language Processing and Recommender Systems." Apress. 1ra edición. * Guío Español, M., Tamayo Uribe, H., & Gómez Ayerbe, P. (2021). "Marco ético para la Inteligencia Artificial Colombia." Consejería Presidencial para Asuntos Económicos y Transformación Digital. Colombia. * Khattak, W., Buhler, P., & Erl, T. (2016). "Big Data Fundamentals: Concepts, Drivers & Techniques." Pearson Education. 1ra Edición. * Roldán, R. C., Martín, E. M., & Rodríguez, A. R. (2019). "Big Data con Python: Recolección, almacenamiento y proceso." Alpha Editorial. * Marin, I., Shukla, A., & Sarang, V. K. (2019). "Big Data Analysis with Python: Combine Spark and Python to unlock the powers of parallel computing and machine learning." Packt Publishing Ltd. |