CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

CIENCIA DE DATOS
BIG DATA: Explotación de Datos

PRESENTACIÓN

La disrupción tecnológica y los modelos derivados, han creado una especialidad que cada vez cobra más relevancia en el mundo moderno: La Ciencia de Datos. El Data Scientist es un profesional capaz de analizar realidades complejas y resolver problemas de negocio a través de la explotación de grandes volúmenes de datos, siendo capaz de ser el puente y traductor entre las áreas técnicas de la empresa y la alta gerencia.

La definición de Big Data es que son datos que contienen una mayor variedad, que llegan en volúmenes crecientes y con más velocidad. Esto también se conoce como las tres Vs. En pocas palabras, los grandes datos son conjuntos de datos más grandes y complejos, especialmente de nuevas fuentes de datos. Crear valor a partir de los datos es privilegio de muy pocos.

Los datos masivos tienen varias características además de ser muy abundantes, entre ellas, heterogeneidad, complejidad, desestructuración, falta de completitud, y tener potencial para ser erróneos. Por esta razón, al diseñar procesos para gestionar datos, debemos tener en cuenta todos estos aspectos con el objetivo de garantizar y preservar su calidad, así como la extracción de información útil y sin errores, lo cual garantizará la fiabilidad de los datos y por ende del análisis resultante.

Hadoop y Spark, ambos desarrollados por Apache Software Foundation, son frameworks de código abierto ampliamente utilizados para arquitecturas de big data. Cada marco contiene un extenso ecosistema de tecnologías de código abierto que preparan, procesan, administran y analizan grandes conjuntos de datos.

Apache Hadoop, es una utilidad de software de código abierto que permite a los usuarios administrar grandes conjuntos de datos (desde gigabytes hasta petabytes) al habilitar una red de computadoras (o "nodos") para resolver problemas de datos extensos e intrincados. Es una solución altamente escalable y rentable que almacena y procesa datos estructurados, semiestructurados y no estructurados (por ejemplo, registros de flujo de clics de Internet, registros de servidores web, datos de sensores de IoT, etc.).

Apache Spark, que también es de código abierto, es un motor de procesamiento de datos para grandes conjuntos de datos. Al igual que Hadoop, Spark divide tareas grandes en diferentes nodos. Sin embargo, tiende a funcionar más rápido que Hadoop y utiliza memoria de acceso aleatorio (RAM) para almacenar en caché y procesar datos en lugar de un sistema de archivos. Esto permite que Spark maneje casos de uso que Hadoop no puede.

PUBLICO OBJETIVO

El curso está dirigido a estudiantes UNI que quieran aprender y aplicar la ciencia de datos en su desarrollo profesional.

En general, este curso te permite enriquecer tu curriculum vitae (CV) en un mundo cada vez más competitivo que está incluyendo entre las habilidades comunes, las de científico de datos, con lo que podrás conseguir mejoras de desarrollo profesional.

CARACTERISTICAS

- Horas síncronas: 24 (6 sesiones)
- Horas asíncronas: 36 (6 por semana)
- Material del curso: Formato digital
- Certificado del curso: Si cumple los criterios de éxito

HERRAMIENTAS Y SOFTWARE

- Apache Spark
- Jupyter Notebooks
- Google Colaboratory
- Kaggle

REQUISITOS

Los postulantes a este curso deben estar cursando por lo menos el VI ciclo y tener conocimiento en:

- Estadística
- Programación con Python

- Lenguaje SQL
- Fundamentos de Machine Learning

METODOLOGÍA

En el desarrollo del curso se aplicará el aprendizaje colaborativo, el autoaprendizaje y el "aprender haciendo". Las técnicas que se usarán son: Método de casos, Método de proyectos, debate y el ABP.

MERCADO LABORAL

Si bien la ciencia de datos sigue siendo un nuevo campo profesional, los empleadores reconocen cada vez más el valor de los profesionales con esta experiencia. Hoy en día, encontrará científicos de datos que trabajan en una variedad de organizaciones, incluidas nuevas empresas tecnológicas, agencias gubernamentales, grandes empresas e instituciones de investigación.

Al finalizar este módulo tendrás los conocimientos fundamentales con el objetivo de convertirte en un científico de datos, las áreas en las que se está aplicando son muchas y cada día va en crecimiento, a continuación, tienes una lista casos en que se aplica la ciencia de datos:

- Ciberseguridad: identificación de ciberamenazas
- Finanzas: detección de fraudes
- Seguros: cálculo de primas
- Medicina: detección de tumores y búsqueda de tratamientos
- Industria: mantenimiento predictivo o la salud de las máquinas
- Marketing: clasificación de los clientes y las audiencias
- Buscadores: reconocimiento de imágenes
- Automatización: coches que conducen solos y artistas digitales
- Energía: asegurando el suministro

Existen muchos otros campos del conocimiento donde el científico de datos tiene mucho que aportar.

CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

CONTENIDO

El contenido del taller se detalla a continuación:

SESIÓN	DETALLE
1	**Introducción al Big Data** ▪ Conceptos y terminología Big Data ▪ Introducción a Hadoop y Spark ▪ Introducción Scala ▪ Practica 1
2	**Arquitectura Hadoop** ▪ Servidores ▪ Job Tracker ▪ Load Balance ▪ Yarn ▪ Practica 2
3	**MapReduce** ▪ Mapper ▪ Reducer ▪ Practica 3
4	**Spark** ▪ Ventajas ▪ Componentes ▪ Arquitectura ▪ Practica 4
5	**Programación Spark** ▪ Rdd & Agregaciones ▪ Merging And Cross Joints ▪ Practica 5
6	**Scala** ▪ Map Collection ▪ Transformaciones ▪ Practica 6

CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

EVALUACION

- Tres practicas calificadas: PC1, PC2 y PC3
- Promedio de prácticas: PP = (PC1+PC2+PC3)/3
- Un examen individual (cuestionario en línea): EF
- Promedio final (PF) = (PP+EF)/2

CRITERIO DE EXITO

Los criterios de éxito son los siguientes:

- Asistencia mínima a las clases síncronas de 80%
- Obtener mínimo 14 como promedio final en las evaluaciones