### Ciencias de Datos con Python

**Área:** Sistemas Paralelos, Distribuidos y Redes de Computadoras

**Tipo:** Curso de Ampliación

**Unidades:** 5 Unidades Crédito

**Horas académicas:** 70

**Código:** --------

**Fundamentación:**

Los seres humanos estamos creando y almacenando datos constantemente en cantidades astronómicas (estructurados, no estructurados y semi-estructurados) que tomarían demasiado tiempo y sería muy costoso cargarlos a una base de datos relacional para su análisis. Tampoco pueden ser procesados y analizados utilizando procesos o herramientas tradicionales. Esta explosión de "Grandes Volúmenes de Datos" está transformando la manera en que se conduce una investigación, dado que plantea una nueva forma de abordarlos mediante el uso de técnicas, habilidades y recursos adecuados para su procesamiento, esto debido a la complejidad e intensidad del cómputo necesario.

En este contexto, la Ciencia de Datos es un área interdisciplinaria y tecnológica que utiliza herramientas, métodos y técnicas de solución y optimización, tanto en hardware como en software, que permiten el manejo y procesamiento de Grandes Volúmenes de Datos. Esto se lleva a cabo mediante cómputo intensivo (algoritmos y programas), y manipulando los datos de forma distribuida en nodos de un clúster logrando así un alto paralelismo en el procesamiento. Como consecuencia se obtienen drásticas reducciones de los tiempos de solución sin sacrificar la confiabilidad, integridad y seguridad de los datos.

Python es uno de los lenguajes más populares en el área de Ciencia de Datos. Provee herramientas robustas que dan soporte a todo el *pipeline* de la Ciencia de Datos. En este curso, el participante debe demostrar habilidades para el uso de librerías en Python usadas para la manipulación y visualización de datos, tales como **numpy**, **pandas**, **matplotlib** y conocimiento de herramientas de *machine learning* como **TensorFlow**.

**Objetivos:**

Al finalizar el curso el participante debe ser capaz de:

* Conocer los fundamentos teóricos y aspectos tecnológicos asociados a la Ciencia de los Datos
* Utilizar métodos y herramientas para el análisis y procesamiento de grandes volúmenes de datos.
* Construir programas en Python para solucionar problemas emblemáticos en la Ciencia de Datos.

**Perfil del estudiante:**

El curso está dirigido a profesionales con conocimientos básicos/intermedios en programación, así como estudiantes de postgrado en las áreas de ciencias de la computación.

**Contenido Temático:**

* **Tema 1: Introducción a la Ciencia de Datos**. Definiciones y conceptos básicos sobre Ciencia de Datos (Data Science). Grandes Volúmenes de Datos (Big Data). Campos de aplicación de la Ciencia de Datos. Áreas afines.
* **Tema 2: Grandes Volúmenes de Datos**. Definiciones y conceptos básicos en el área de Grandes Volúmenes de Datos. Evolución de los Grandes Volúmenes de Datos. Fuentes de Grandes Volúmenes de Datos: Fuentes de datos privadas. Fuentes de datos públicas (Open Data). Concepto de almacenes de datos e inteligencia de negocios. Sistemas de Bases de Datos relacionales y Bases de Datos no relacionales. Interfaces de Programación de aplicaciones (API’s) y Servicios Web.
* **Tema 3: Sistemas de Computación para Grandes Volúmenes de Datos**. Definiciones y conceptos básicos de paralelismo. 3.2. Modelos de Computación Paralela: Divide y vencerás (Divide&Conquer), Cola de Tareas (TasksQueue), Repartir y Reducir (Map and Reduce), Segmentación (Pipeline) etc. Sistemas de Computación Paralela como plataforma de hardware para el procesamiento de Grandes Volúmenes de Datos. Caso de estudio: a) Clúster de Computación. Arquitectura. Funcionamiento. b) Clúster de Datos y Centros de Datos. Arquitectura. Funcionamiento. (Webservices). Uso de la herramienta de **Python Pandas** para la Preparación de Datos: extracción, transformación, carga y limpieza de datos (ETL-C). Ejemplos prácticos.
* **Tema 4: Software para el Procesamiento de Grandes Volúmenes de Datos.** Uso de **Docker** para construir un cluster *Beowulf*. Fundamentos del Modelo de Cómputo MapReduce. Uso de la herramienta de **Python MPI4y** y **Numpy** para implementar un MapReduce.
* **Tema 5: Introducción a la Minería de Grandes Volúmenes de Datos**. Concepto de Minería de Datos (Data Mining). Aplicación de la Minería al procesamiento de Grandes Volúmenes de Datos. Algoritmos de Minería de Datos usados en el procesamiento de Grandes Volúmenes de Datos: Algoritmos predictivos. Algoritmos descriptivos. Metodología de Minería de Datos para Grandes Volúmenes de Datos. Uso de la herramienta de Python TensorFlow para implementar un modelo predictivo básico basado en datos.

**Evaluación**

* 3 quices 10% c/u
* 3 talleres 10% c/u
* 1 proyecto 40%

**Observaciones finales**:

* El curso tiene una duración de 70hrs.
* La modalidad es bimodal
* Se utilizaran herramientas de comunicación como Google Meet y el Laboratorio del Postgrado en Ciencias de la Computación.

|  |  |  |
| --- | --- | --- |
| **Docente** | **Coordinador** | **Personal Administrativo** |
| Dr. Andres Sanoja | Dr. Andres Sanoja | Rosiris Morales |

**Bibliografia**

* Anand Rajaraman, Jure Leskovec and Jeffrey D. Ullman. Mining of Massive Datasets. Stanford University 2013.
* Ian H. Witten, Eibe Frank, and Mark A. Hall. Data Mining Practical Machine Learning Tools and Techniques. 3rd Edition. Morgan Kaufmann 2011.
* Peter Pacheco. An Introduction to Parallel Programming. Morgan Kaufmann Publishers. 2011.
* Donald Miner and Adam Shook. MapReduce Design Patterns. O’Reilly. 2013.
* Pete Warden. Big Data Glossary. O’Reilly. 2011.
* Python Data Science Handbook. Jake VanderPlas. ‎ Shroff/O'Reilly; First edition (1 January 2016); Shroff Publishers & Distributors Pvt. Ltd., B-103, First Floor, Railway Commercial Complex, Sector 3, Sanpada (E), Navi Mumbai 400705,
* Data Science from Scratch: First Principles with Python. Joel Grus. Shroff/O'Reilly; Second edition (5 May 2019); Shroff Publishers & Distributors Pvt. Ltd., B-103, First Floor, Railway Commercial Complex, Sector 3, Sanpada (E), Navi Mumbai 400705
* Data Science Projects with Python. Stephen Klosterman. Packt Publishing Limited (30 April 2019).
* Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd Edition. Wes McKinney. O'Reilly Media; 2nd edition (November 14, 2017)
* Python For Data Science. Ethan Williams. Independently Published (18 August 2019)
* Python for Data Science For Dummies. John Mueller. Wiley (1 January 2019)