**Asignatura: Programación Paralela con R y Rstudio. Una introducción**

**Profesor de la asignatura: Rina Surós**

**Objetivos:**  El objetivo de este curso es el desarrollo y optimización de algoritmos altamente escalables a ser ejecutados en plataformas paralelas para el análisis de datos. Utilizaremos para ello las herramientas R y Rstudio de programación paralela, para computadoras personales. Desde el punto de vista de la informática, el análisis de datos requiere del paralelismo para almacenar, recuperar y procesar grandes cantidades de datos, más grandes de las que normalmente se pueden tratar en una computadora personal. La implementación eficiente a gran escala requiere enfoques y algoritmos completamente nuevos. Usaremos las técnicas de procesamiento de datos conocidas, independientemente del campo de aplicación.

Al final del curso revisaremos las últimas técnicas para problemas realmente complejos que requieren combinar Algoritmos Paralelos y Big Data como el paquete paquete ff y bigmemory (The BigMemory Project ([www.bigmemory.org](http://www.bigmemory.org/%22%20%5Ct%20%22_blank))).

**Conocimientos previos recomendados:**  Estadística básica. Manejo instrumental de R (Rstudio)

**Metodología de enseñanza:**

Previo a cada clase se entrega un documento powerpoint con el contenido de los puntos correspondientes según el programa del curso, casos de estudio, códigos y resultados de la ejecución, tal como se presentará en a los estudiantes. Cada sesión tendrá el contenido teórico correspondiente, se plantea un problema tipo y su código secuencial, se presentan las técnicas de paralelización adecuadas y las funciones de R (Rstudio) que permiten implantar el código paralelo. Se desarrolla un código paralelo y se ejecuta para un número de datos que permita el efecto demostración. En los trabajos el estudiante mostrará resultados para grandes cantidades de datos.

Las clases se graban y se publican para que estén a disposición de los cursantes.

* Horas de clase (teórico): 3 horas de clases teóricas por semana
* Horas de consulta: 1 hora por semana

# Forma de evaluación:

A lo largo del curso los participantes reciben 5 tareas de contenido teórico y de programación y una trabajo final. El trabajo final consiste en paralelizar una aplicación en todas sus etapas: análisis y descomposición del problema, planificación del proceso de ejecución sobre la máquina disponible para cada estudiante, análisis de los resultados (caso secuencial contra caso paralelo). Se utilizarán las herramientas más recientemente desarrolladas para grandes bases de datos.

**Temario:**

1. Sistemas paralelos y distribuidos. Estructura de los sistemas paralelos. El sistema de comunicaciones: paso de mensajes, protocolos de comunicación. Sincronización de procesos, modelo de consistencia de memoria. Modelos de paralelismo. Evaluación de rendimiento. Ley de Amdahl.
2. Arquitecturas de máquinas paralelas. Tipos de arquitectura y clasificación. Entorno de programación. Escalabilidad. Restricciones del cálculo. Limitaciones de CPU y restricciones de la memoria. Dificultades de Entrada/Salida de datos. Conflictos de la transferencia de datos.
3. Desarrollo de aplicaciones paralelas. Paradigmas de la computación paralela. El paradigma maestro-esclavo. Escalabilidad y balanceo de carga. Cuando paralelizar. Como paralelizar.
4. Análisis de problemas y algoritmos tipo. Ejemplos de uso paralelo.
5. Calculo paralelo con R. Los paquetes de programación paralela. Estructura de un programa tipo. Paradigmas y desafíos de programación. Análisis y paralelización de problemas tipo. Selección de problemas. Análisis de potencialidad de paralelismo. Análisis de proyección del algoritmo paralelo sobre la máquina. Ejecución. Análisis de rendimiento de la solución implementada.

**Referencias**

* Big Data Analytics with R and Hadoop. Vignesh Prajapati. Published by Packt Publishing Ltd. 2013. ISBN 978-1-78216-328-2.
* Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Published by John Wiley & Sons, Inc. 2015. ISBN: 978-1-118-87613-8.
* Parallel R: Data Analysis in the Distributed World. Q. Ethan

McCallum Stephen Weston. ISBN-13: 978-1449309923. ISBN-10: 1449309925

* Mastering Parallel Programming with R. Simon R. Chapple, Eilidh Troup, Thorsten Forster, Terence Sloan. EditorialO’Reilly. ISBN-13: 978-1784394004. ISBN-10: 1784394009

**Referencias WEB**

* <https://cran.r-project.org/web/views/HighPerformanceComputing.html>
* The BigMemory : www.bigmemory.org
* The Package ‘ff’ <https://cran.r-project.org/web/packages/ff/ff>
* Presentación de la RedClara.: