FACULTAD DE INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

PROCESAMIENTO PARALELO PARA BIG DATA

Año 2018

Carrera: Especialización en Inteligencia de Datos orientada a Big Data
Carga Horaria: 64 Hs.

Profesor a Cargo:
Ing. A. De Giusti - Dr. Enzo Rucci

OBJETIVO

Aplicar los conceptos fundamentales del procesamiento paralelo al caso de grandes volúmenes de datos (Big Data).
Presentar sistemas de almacenamiento de Big Data y herramientas de procesamiento paralelo sobre los mismos.
Analizar el empleo de técnicas y herramientas de HPC (High Performance Computing) en tratamiento de Big Data, desde el punto de vista rendimiento y eficiencia.

MODALIDAD DE EVALUACION

Evaluación escrita individual.
Alternativamente definición de trabajos aplicados a defender en forma individual, dentro de un plazo determinado (en principio 3 a 6 meses).

PROGRAMA

Unidad 1: Conceptos básicos de paralelismo

Procesamiento paralelo. Arquitecturas paralelas. Servidores, Clusters y Cloud. Modelos de programación. Métricas. Herramientas.

Unidad 2: Introducción a Big data. Relación con Paralelismo

Fundamentos. Objetivos. Modelos de datos y modelos de procesamiento. Paradigma Map-Reduce. Apache Hadoop. Por qué paralelismo sobre Big Data?

Unidad 3: Sistemas de almacenamiento para Big Data

Sistemas de archivos distribuidos. Clasificación. Apache HDFS.
Bases de datos relacionales. Bases de datos NoSQL. Hive, Shark, MongoDB, Cassandra.

Unidad 4: Procesamiento paralelo en Big Data

Variantes y extensiones de Map-Reduce.
Procesamiento por lotes y stream. Hadoop, Spark, Storm.
Técnicas de programación.
Evaluación de rendimiento. Optimización en el acceso y la comunicación de datos.
Tuning de parámetros.
Análisis de la eficiencia energética en algoritmos paralelos sobre Big Data.

Unidad 5: Casos de estudio

Análisis de problemas sobre grandes volúmenes de datos de diferentes clases (texto, señales, secuencias, grafos, imágenes) y su tratamiento con herramientas clásicas de Big Data y posibles optimizaciones con algoritmos paralelos.

BIBLIOGRAFIA

G. Hager and G. Wellein (2011) Introduction to High Performance Computing for Scientists and Engineers, H. Simon, Ed. CRC Press.

M. Giles and I. Reguly (2014) Trends in high-performance computing for engineering calculations, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 372, no. 2022, p. 20130319.

T. Rauber and G. Rünger (2010) Parallel Programming for Multicore and Cluster Systems. Springer-Verlag, Berlin Heidelberg.

I. Foster and D. B. Gannon (2017) Cloud Computing for Science and Engineering (Scientific and Engineering Computation).

Leskovec, J. et al (2014) Mining of massive datasets. Cambridge University Press

Shanmuganathan, S. (2014) From data mining and knowledge discovery to big data analytics and knowledge extraction for applications in science

Kakhani, M. K. et al. (2013). Research Issues in Big Data Analytics. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2(8).

Kaisler, S. et al. (2013). Big data: Issues and challenges moving forward. In System Sciences (HICSS), 46th Hawaii International Conference on (pp. 9951004). IEEE.

Chih-Fong Tsai, Wei-Chao Lin, Shih-Wen Ke (2016) Big data mining with parallel computing: A comparison of distributed and MapReduce methodologies, Journal of Systems and Software, Volume 122, 2016, Pages 83-92, ISSN 0164-1212, http://dx.doi.org/10.1016/j.jss.2016.09.007.

Y. Zhang *et al*. (2016) Parallel Processing Systems for Big Data: A Survey, in *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2114-2136, Nov. 2016. doi: 10.1109/JPROC.2016.2591592

Senger, H.and Geyer, C (2016) Parallel and distributed computing for Big Data applications, Concurrency and Computation: Practice and Experience, 28 (8), http://dx.doi.org/10.1002/cpe.3813

J. Dean and S. Ghemawat (2004) “MapReduce: Simplified data processing on large clusters,” in Proc. 6th Symp. Operating Syst. Design Implementation, San Francisco, CA, USA, Dec. 6–8, 2004, pp. 137–150. USENIX Association