Temas aprobados por Consejo Directivo:

1. Implementación y evaluación utilizando Hadoop de un mecanismo para anonimizar tráfico de red en formato netflow.
   1. Descripción: En seguridad de redes de datos, se suele utilizar los registros en formato netflow (generados por routers comerciales) con propósitos de análisis forense. En ocasiones, se subcontrata el servicio de consultores externos para que realicen este análisis, surgiendo el problema de la privacidad de la información contenida en los netflows. El proyecto FLAIM ha planteado varios mecanismos que pueden ser utilizados para anonimizar (a diversos niveles) la información contenida en netflows. Si la cantidad de registros a anonimizar es grande, el proceso de anonimización puede tomar horas, días o hasta semanas. Es por esto que surge la necesidad de disponer de una manera de realizar dicho proceso, en un ambiente distribuido que permita reducir considerablemente los tiempos de procesamiento.
   2. Observación: El grupo seleccionará uno de los mecanismos del proyecto FLAIM (Framework for Log Anonymization and Information Management).
2. Estudio de la eficiencia de la plataforma Hadoop para la ejecución distribuida y paralela de simulaciones FDTD (finite-difference time-domain) con el paquete Meep.
   1. Descripción: Investigadores en las áreas de física, electrónica y otras, suelen utilizar el paquete Meep para realizar simulaciones FDTD. Debido a su complejidad y cantidad de simulaciones a realizar, usualmente trabajan con clústeres de computadoras. Debido a la acogida que la plataforma Hadoop para procesamiento distribuido tiene en la actualidad, sería conveniente poder utilizar clústeres Hadoop para realizar simulaciones FDTD, aprovechando así la infraestructura que muchas universidades están montando para esta plataforma. En este trabajo, se estudiará y evaluará la factibilidad de utilizar esta clústeres Hadoop para la ejecución distribuida y paralela de simulaciones FDTD (finite-difference time-domain) con el paquete Meep.
   2. Observación: Tema en colaboración con Ing. German Vargas.
3. Desarrollo de una herramienta para la creación y administración de clústeres computacionales para simulaciones FDTD (finite-difference time-domain) con el paquete Meep, sobre el servicio Elastic Compute Cloud (EC2) de los Amazon Web Services (AWS).
   1. Descripción: Investigadores en las áreas de física, electrónica y otras, suelen utilizar el paquete Meep para realizar simulaciones FDTD. Debido a su complejidad y cantidad de simulaciones a realizar, usualmente trabajan con clústeres de computadoras. Debido a la acogida que la plataforma EC2 de AWS ha tenido en la actualidad, investigadores ya la han empezado a usar para levantar clústeres de alto rendimiento. StarCluster es una herramienta que permite levantar clústeres de alto rendimiento (High Performance Computing) sobre EC2. Lastimosamente, los clústeres levantados por StarCluster no tienen el paquete Meep instalado. En este trabajo, se desarrollará una herramienta para la creación y administración de clústeres computacionales de propósito específico (simulaciones FDTD con el paquete Meep) sobre el servicio EC2 de AWS.
   2. Observación: para este proyecto los estudiantes adaptarán o extenderán la herramienta StarCluster. Tema en colaboración con Ing. German Vargas.
4. Utilización de la plataforma Hadoop para implementar un programa distribuido que permita encontrar las paredes de células de la epidermis de plantas modificadas genéticamente.
   1. Descripción: Encontrar las paredes de células de la epidermis de plantas modificadas genéticamente. La meta consiste en saber cuántas hay. Se usarán dos técnicas: una propuesta por Stger para detección estructuras lineales (la salida son puntos localizados sobre la línea) y otra por Sappa para integrar puntos en segmentos.
   2. Observación: Tema en colaboración con Ing. Daniel Ochoa.
5. Evaluación de MapReduce, Pig y Hive, sobre la plataforma Hadoop, para la implementación de tres programas de procesamiento masivo de datos.
   1. Descripción: Se implementará WordCount y tres programas más seleccionados por los estudiantes sobre estos tres esquemas de programación soportados por Hadoop, y se realizará un análisis y comparación de rendimiento en los tres esquemas, para clústeres de varios tamaños (2, 4, 6, 10, 15 y 20 nodos).
6. Planteamiento y validación de un modelo matemático que permita estimar el correcto número de mappers y reducers a utilizar en un proceso MapReduce
   1. Descripción: Uno de los problemas que se presenta a programadores al momento de implementar programas MapReduce es el determinar el número de mappers y reducers que, para un problema en particular, produzca los mejores tiempos de rendimiento. Utilizando estadística y teoría de colas, se puede plantear un modelo matemático que permita estimar el número adecuado de mapers y reducers a utilizar. El modelo planteado será validado a través de pruebas reales con Hadoop.

Referencias:

* Dean, J. y Ghemawat, S. “MapReduce: Simplified Data Processing on Large Clusters“. En Memorias del Sixth Symposium on Operating System Design and Implementation (OSDI 2004). San Francisco, CA-EE.UU. Diciembre, 2004. Disponible en línea en: <http://labs.google.com/papers/mapreduce.html>.
* Ghemawat, S., Gobioff, H., y Leung, S. “The Google File System“. En Memorias del 19th ACM Symposium on Operating Systems Principles. Lake George, NY-EE.UU., Octubre, 2003. Disponible en: <http://labs.google.com/papers/gfs.html>.
* Meep: <http://ab-initio.mit.edu/wiki/index.php/Meep>.
* FLAIM: <http://flaim.ncsa.illinois.edu/>.
* StarCluster: <http://web.mit.edu/stardev/cluster/>.

Criterio de selección que sera utilizado:

* No se ha considerado a ningún estudiante que le falten más de tres materias para egresar (exceptuando inglés), debido a que la carga académica para el estudiante sería excesiva.
* De los estudiantes que no fueron eliminados según el criterio anterior, se seleccionó a los 22 estudiantes que obtuvieron la mejor puntuación, usando la siguiente fórmula:

Puntaje = Promedio general – R \* 0.02 – E \* 0.03 + M \* 0.075

donde,

R: # materias reprobadas

E: # materias que le faltan para egresar (exceptuando inglés)

M: motivación para tomar la materia, evaluado del 1 (mínimo) al 5 (máximo):