EXTRACCIÓN DE CONOCIMIENTO EN REDES SOCIALES MEDIANTE HERRAMIENTAS DE SOFTWARE LIBRE Y PLATAFORMAS DE HARDWARE PARALELO-DISTRIBUIDAS

Gouiric, Guillermo Adrián; Ortega, Manuel Oscar; Klenzi, Raúl Oscar
Instituto de Informática / Departamento Informática / Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan
Domicilio: Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan, CPA: J5402DCS, 0264-260353 0264-4260355
{guillegouiric;manuel.ortega;rauloscarklenzi}@gmail.com

RESUMEN

El presente trabajo pretende realizar tareas de extracción de conocimiento en grandes colecciones de datos mediante la aplicación de la herramienta de software libre de aprendizaje de máquina KNIME ANALYTYCS procesando datos provenientes de la Red Social (Twiter) y ejecutándose en plataformas paralelo distribuidas tratando de cotejar las mejoras de performance respecto de las aplicaciones secuenciales en la caracterización de perfiles de usuario. A tal efecto se habrá de trabajar con la versión de KNIME ANALYTICS 3.5.2, ejecutándose sobre un cluster de cuatro terminales de cómputo constituyendo el paradigma de computación distribuida cada una de las cuales cuenta con placas GPU computing sobre una de las cuales se ejecutará la instancia paralela del análisis.

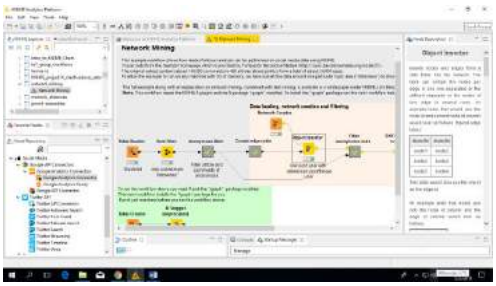
Palabras clave: Minería de Datos, Redes Sociales, Paralelismo, HPC, GPU.

CONTEXTO

En el ámbito del proyecto CICITCA_UNSJ “Ciencia de los Datos aplicada a grandes colecciones de datos” ejecutado en el bienio 2016-2017 y la continuidad buscada por el grupo de investigadores que lo conformaban, en la presentación de un proyecto para el bienio 2018-2019 actualmente en evaluación “Visualización y DeepLearning en Ciencia de los Datos” se ha conformado un cluster de cuatro?? Computadoras cada una de las cuales cuenta con unidades NVIDIA de GPU computing. Con el hardware citado y desde la realización de trabajos finales de grado y becas de investigación de alumnos avanzados se lleva adelante la determinación de perfiles de usuarios de TWITER y Redes sociales en general, mediante la utilización de software libre de aprendizaje de máquina y cómputo paralelo distribuido.

La elección del entorno de software KNIME ANALYTICS 3.5.2 se centra en lo expresado en [1] KNIME es una plataforma cohesionada para científicos de datos de todos los niveles de habilidades, que proporciona un marco de ciencia de datos único y consistente. Ofrece capacidades de acceso y manipulación de datos de alta calificación, una amplia y completa gama de algoritmos y herramientas de aprendizaje automático adecuadas tanto

para principiantes como para científicos de datos experimentados. La plataforma de KNIME se integra con otras herramientas y plataformas, como R, Python, Spark, H2O.ai, Weka, DL4J y Keras, a la vez que permite su ejecución considerando procesadores multinúcleos y/o GPGPU computing. La ayuda contextual de KNIME es más flexible que los "asistentes" fijos. La interfaz de usuario y los extensos ejemplos proporcionados con la plataforma atraen a la comunidad de científicos de datos.



Interfaz de KNIME ANALYTICS 3.5.2

1. INTRODUCCIÓN

Para explicitar conceptos brindados en el contexto de la propuesta, habrán de definirse aspectos relevantes a la misma:

Big Data hace referencia a una colección de datos de considerable dimensión que hacen imposible el procesamiento con aplicaciones tradicionales de base de datos.

En [2] definen *Big Data* usando las tres V's: Volumen, Velocidad y Variedad.

Volumen se refiere a la cantidad de datos, desde Terabytes (TB) a Petabytes (PB), relacionado con la estructura de estos datos incluyendo registros, transacciones, archivos, y tablas. El volumen de estos datos, se prevé que crecerá 50 veces para el 2020.

Velocidadse refiere a la forma de transferir los grandes volúmenes de datos incluyendo transmisión en batch, tiempo real y flujos. La velocidad, incluye tiempo y latencia, características propias del manejo de datos. Los datos pueden ser analizados, procesados, almacenados y manipulados en forma rápida, o con un retardo entre eventos.

Variedad de los grandes volúmenes de datos, se refiere a los diferentes formatos que pueden adoptar los datos, incluyendo estructurados, semi estructurados, desestructurados y todas las combinaciones de estos tres. El formato de los datos incluye: documentos, mails, mensajes de texto, audio, imágenes, video, gráficos, entre otros.

High-Performance Computing (HPC) según [3] se usa para describir ambientes de cómputo que utilizan supercomputadoras o clusters de computadoras para atender requerimientos de cómputo complejos o aplicaciones con requerimientos altos de tiempo o que requieren procesamiento de grandes volúmenes de datos.

La tecnología HPC es apropiada tanto para las aplicaciones de cálculo intensivo, como las de procesamiento intensivo de datos. Las plataformas HPC utilizan un alto grado de paralelismo que tiende a usar multiprocesadores especializados con arquitecturas de memoria que han sido altamente optimizadas para cálculos numéricos.

Las PC's actuales tienen más poder de cómputo que las supercomputadoras de hace una década. Estas PC's, poseen procesadores con múltiples cores (multicore), con poderosas placas de video (GPGPU), inicialmente concebidas para abordar tareas gráficas que requieren gran cantidad de cómputo en paralelo. Las PC´s con procesadores multicore, y placas gráficas GPU, constituyen lo que en [4] define como computación heterogénea, ya que combina más de un tipo de procesador. Este tipo de configuración de hardware, constituye una herramienta nueva para la computación HPC.

Redes sociales

En las redes sociales son sitios de Internet formados por comunidades de individuos con intereses o actividades en común (como amistad, parentesco, trabajo) y que permiten el contacto entre estos, de manera que se puedan comunicar e intercambiar información.

Minería de Datos

La minería de datos, Data Mining, según [6] es un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining. Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos.

Sistemas Distribuidos

Los sistemas distribuidos intentan hacer que un conjunto de computadoras físicamente separadas (cada una con sus propios recursos como procesador, memoria, buses) trabajen cooperativamente para resolver un problema grande, y que externamente se vean como una sola unidad con un gran poder de procesamiento.

Se define en [7] un Sistema Distribuido como un sistema en el cual componentes de hardware y software, localizadas en computadores de red, se comunican y coordinan sus acciones sólo por paso de mensajes

Las computadoras conectadas mediante red pueden estar físicamente a cualquier distancia, inclusive separadas por continentes, pero también en el mismo edificio o habitación.

CUDA-Paralelismo

Según [8] a través del uso de CUDA es posible construir aplicaciones paralelas capaces de aprovechar los múltiples cores de una GPU. Hacer programas para ser ejecutados sobre una GPU implica comprender la interacción de datos y de control que se produce entre la CPU y la GPU, dando lugar a una nueva modelo de programación paralela. Este modelo combina un enfoque de trabajo distribuido con uno de trabajo paralelo. El enfoque distribuido es consecuencia del hecho de que la CPU y la GPU tienen memorias disjuntas y arquitecturas de hardware diferentes. Además, ejecutan tareas que, si bien están relacionadas, no son las mismas, y por lo tanto no ejecutan las mismas instrucciones. El enfoque paralelo se da por el uso de los múltiples cores de la GPU, de modo que todos pueden ejecutar la misma instrucción al mismo tiempo, pero sobre diferentes datos

Rendimiento o performance

El rendimiento es la rapidez con la que el computador puede ejecutar programas, es la inversa del tiempo requerido por una computación, el cual se calcula con la siguiente expresión:

$$\frac{1}{Tiempo\ de\ ejecución}$$

Speed Up

Es una métrica usada para medir la mejora relativa de la performance de un programa cuando a este se le hace alguna modificación. La forma básica para calcular el Speed Up es:

$$\frac{\text{Tiempo de ejecución sin la mejora}}{\text{Tiempo de ejecución con la mejora}}$$

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

En el marco de la investigación se pretende:

- Analizar y describir el conjunto de estudios y prácticas requeridos en Ciencia de Datos.
- Analizar diferentes API's (Application Programming Interface) de aplicación para el relevamiento de datos Redes Sociales y su análisis e interpretación, sobre plataformas paralelo-distribuidas, con las formas de representación que las herramientas a utilizar poseen.

3. RESULTADOS ESPERADOS

Se espera comprobar mejoras en la performance, de los algoritmos paralelo distribuidos respecto de las alternativas secuenciales. Especialmente se espera una interesante mejora en el SpeedUp de las aplicaciones distribuidas y fundamentalmente paralela. Se trabajará centralizando las actividades en la interfaz de KNIME dado que la misma tiene capacidades de interactuar con diferentes entornos de software y trabajar en plataformas multinúcleos y/ GPGPU computing.

4. FORMACIÓN DE RECURSOS HUMANOS

Esta propuesta, originalmente contenida en el proyecto "Ciencia de los Datos en Grandes colecciones de Datos", permite continuarse en el proyecto actualmente en evaluación "Visualización y DeepLearning en Ciencia de Datos". En este contexto la profundización del conocimiento adquirido por los integrantes del proyecto, permite formar nóveles docentes-investigadores del departamento informática, así como la dirección y defensa de diferentes tesis de maestría y trabajos finales de grado, Simultáneamente, permite proponer alumnos de grado a becas de alumnos avanzados e iniciación. La significativa sinergia alcanzada por el grupo de trabajo ha permitido extender las aplicaciones del Data Science a otros proyectos a la vez que proponer diferentes instancias de propagación del conocimiento que se vuelca a las cátedras de las carreras del Departamento Informática, así como en charlas, y/o cursos de posgrado.

5. BIBLIOGRAFÍA

[1] «Gartner 2018 Magic Quadrant for Data Science and Machine Learning – Read the report». [En línea]. Disponible en: https://www.kdnuggets.com/2018/02/domino-gartner-mq-data-science-machine-learning.html. [Accedido: 16-mar-2018].

[2] B. Furht y F. Villanustre, *Big Data Technologies and Applications*. Springer, 2016.

[3] B. Furht y A. Escalante, *Handbook of Data Intensive Computing*.

Springer Science & Business Media, 2011.

[4] Y. Tan, *GPU-based Parallel Implementation of Swarm Intelligence Algorithms*. Morgan Kaufmann, 2016.

[5] Anonimo.(s.F). *Concepto de Redes Sociales.* Recuperado de http://concepto.de/redes-sociales/

[6] GestioPolis.com Experto. (2001*)¿Qué es Data Mining?* Recuperado de www.gestiopolis.com/que-es-data-mining/

[7] G. Coulouris*, Sistemas distribuidos, conceptos y diseño*, Addison Wesley, 2002.

[8] N. Wilt, *The CUDA Handbook*, Addison-Wesley, 2013.

.