LA MINERÍA DE DATOS Y SU EVOLUCIÓN A LA GRID

ANDRÉS SANTIAGO SERNA TANGARIFE

UNIVERSIDAD EAFIT

DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS

MEDELLÍN

2010

LA MINERÍA DE DATOS Y SU EVOLUCIÓN A LA GRID

ANDRÉS SANTIAGO SERNA TANGARIFE

Trabajo de grado para optar al título de
Ingeniero de Sistemas

Asesora
SONIA CARDONA RIOS
COORDINADORA ESPECIALIZACIÓN EN SISTEMAS DE INFORMACIÓN

UNIVERSIDAD EAFIT
DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS
MEDELLÍN
2010

CONTENIDO

1. INTRODUCCIÓN

[DUBITZKY 2008] dice que debido al incremento en el uso de sistemas computarizados en las organizaciones, la cantidad disponible de datos digitales ha crecido en gran manera y se han vuelto incluso un activo esencial para el funcionamiento del negocio. Un artículo de la International Journal of Grid Computing and eScience [STANKOVSKI 2007] habla acerca de la Minería de Datos dando a entender que el uso efectivo y eficiente de los datos y la transformación de estos en información y conocimiento, se ha convertido en la base fundamental de las organizaciones para alcanzar el éxito.

En este mismo artículo, se habla acerca de cómo la complejidad y cantidad de datos que debe manejar una organización crecen constantemente. Es así que el explorar, analizar e interpretar dichos datos se ha convertido en un reto y además en una tarea esencial en los procesos que se llevan día a día en las organizaciones. [DUBITZKY 2008] enfatiza sobre la Minería de Datos como la tecnología que hoy en día habilita realizar dichas tareas, por esto su uso puede extenderse a muchos usuarios en la organización ya que puede ser utilizada desde la simple extracción y procesamiento de datos, hasta la construcción de sistemas de algoritmos y modelos para solucionar problemas complejos como detección de fraude, detección de intrusos, monitoreo, automatización de procesos y todo aquello donde los datos representen una clave esencial del negocio y cuyo procesamiento para obtener información y conocimiento no es una tarea tan sencilla.

Por esta razón, la Minería de Datos ha tomado hoy gran importancia, sin embargo, han comenzado a surgir nuevas necesidades. [KURMAN 2008],

[DUBITZKY 2008], [STANKOVSKI 2007] y [KARGUPYAM 2005] se expresan acerca de cómo cada vez se requieren soluciones de Minería de Datos más sofisticadas y cada vez nos damos cuenta que los sistemas de datos y computadores convencionales son muy limitados para ofrecer un buen rendimiento a los procesos de minería. Los datos día a día crecen en dimensiones descomunales y cada vez se hace más difícil realizar procesos de minería porque toman mucho tiempo para arrojar resultados. Es aquí, donde aparece una posible solución que es la Computación Grid, la cual busca solucionar problemas que no pueden ser resueltos en un tiempo razonable con computadores convencionales, mediante el uso de diferentes procesadores y/o máquinas que se distribuyan las tareas (divide y vencerás) y finalmente se obtengan resultados más rápida y eficientemente. La Minería de Datos en Grid, permite el aprovechar diferentes recursos computacionales conectados en una red, con el fin de brindar soluciones a problemas de minería más complejos, con datos a gran escala y con un mayor rendimiento que el que ofrece una herramienta de Minería de Datos convencional.

El objetivo de este trabajo es analizar la evolución del concepto de Minería de Datos a Computación Grid y determinar algunos factores que se deben tener en cuenta para evaluar la viabilidad de usar esta tecnología en organizaciones que cuentan con procesos de Minería de Datos convencional. Para esto, se expone el estado del arte de la Minería de Datos, el estado de arte de la Computación Grid y el concepto de Minería de Datos Grid, los beneficios que ésta puede brindar a la organización y buenas prácticas que se deben tener en cuenta para el uso de la Minería de Datos.

El capítulo 2, busca describir cuál es el significado de la Minería de Datos y su importancia hoy en día como factor competitivo en las organizaciones, buenas prácticas, técnicas, herramientas, aplicaciones, oportunidades y retos. En el capítulo 3 expone los conceptos y aplicaciones de la Computación Grid. En el capítulo 4 se analiza cómo el surgimiento de la Minería de Datos Grid busca

solucionar las dificultades que comienzan a presentarse debido al crecimiento de los negocios y la necesidad cada vez más importante de analizar sus datos, aprovechando los recursos computacionales con los que cuenta la organización y se presentan algunos algoritmos de Minería de Datos distribuidos junto con su análisis de rendimiento realizado por algunos autores. El capítulo 5, lista algunos factores que se deben tener en cuenta a la hora de analizar la viabilidad de implementar Minería de Datos Grid, extraídos de las experiencias que varios de los autores citados en este trabajo han tenido frente al tema.

2. CONCEPTOS Y APLICACIONES DE LA MINERÍA DE DATOS

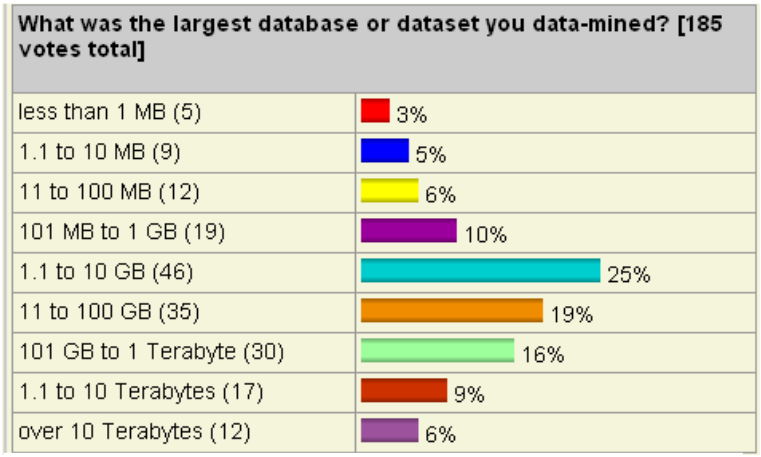
2.1. Conceptos generales de la Minería de Datos

[KANTARDZIC 2003] hace mención acerca del proceso de entrar en la era digital que se ha venido viviendo en los últimos años. Cada día crece el problema de poder controlar e interpretar las grandes cantidades de datos que constantemente generan los negocios. Nuestra capacidad para entender y analizar estas grandes cantidades de datos está muy por debajo de nuestra habilidad para obtener y almacenar los datos. Hoy en día, casi que cualquier actividad que realizamos en nuestro entorno genera datos que son un muy importante insumo para la toma de decisiones.

[Ídem] da a entender que rápidamente ha venido creciendo la abertura entre la capacidad de obtención, organización de los datos y la capacidad de análisis de los datos. El hardware y las tecnologías de bases de datos actuales permiten el almacenamiento de los datos de una forma eficiente, confiable y barata. Sin embargo, cuando se trata de negocios, medicina, ciencia o gobierno los datos por sí solos generan muy poco valor para las empresas. Lo que verdaderamente vale es el conocimiento que pueda ser extraído y utilizado a partir de los datos. Un ejemplo mencionado por Kantardzic es el de un supermercado, este probablemente quiera utilizar sus datos sobre las ventas de mercancías para extraer conocimiento acerca de cómo se relaciona la venta de algunos de sus productos y ciertos grupos demográficos. Este conocimiento podría ser utilizado para introducir nuevas campañas dirigidas con un retorno económico predecible, lo que no se podría lograr mediante una campaña sin ningún enfoque.

Como se describe en [ibídem], el principal problema que surge a la hora de tratar de analizar los datos para generar este conocimiento, es que el tamaño y dimensión de los datos puede llegar a ser tan enorme, que se hace difícil, sino imposible, la interpretación y análisis de los datos de forma manual e incluso para algunos análisis semi automáticos por medio de computadores. Se puede trabajar fácilmente con unos cientos o miles de registros, pero cuando llegamos al orden de millones de datos que además poseen cientos o miles de características, se vuelve algo prácticamente inmanejable. En teoría, cualquiera pudiera pensar que mientras más datos se posean del negocio se puede llegar a mejores conclusiones, pero en la práctica muchas dificultades pueden surgir.

Ilustración 1: Encuesta: ¿Cuál fue la base de datos más grande que haya minado?



Extraído de www.kdnuggets.com - 2009

¿Cuál es la solución? Mayor trabajo y esfuerzo quizá, pero llegará el momento en que no se pueda seguir porque siempre habrá límites. Contratar más empleados si puede hacerlo económicamente, o simplemente se pueden ignorar los datos pero dejaría de ser competitivo en el mercado. La única solución, es reemplazar el proceso clásico de analizar e interpretar los datos por una tecnología llamada Minería de Datos.

2.1.1. Significado e importancia de la Minería de Datos. [OLSON 2008] menciona cómo hoy en día, con la evolución de los computadores y los sistemas digitales, es normal encontrarse con grandes volúmenes de datos en nuestros computadores, redes y vidas. Incluso, las instituciones, gobiernos, negocios, y en general todas las organizaciones dedican continuamente gran cantidad de recursos para recolectar y almacenar datos. Grandes cantidades de datos son generadas desde las máquinas registradoras, escáneres, bases de datos de la compañía y todos estos datos son explorados, analizados, procesados y reusados para toda clase de actividades que realiza la organización.

Sin embargo, generalmente pequeñas cantidades de estos datos son utilizados porque en muchos casos los volúmenes son demasiado grandes para ser manipulados o la estructura de los datos es muy compleja para ser analizada efectivamente. Esto generalmente limita las decisiones que la organización debe tomar día a día, pues en el mundo altamente competitivo de hoy, el buen entendimiento de los datos y la extracción de conocimiento de estos son un insumo vital para poder tomar las mejores decisiones. Por esto, [Ídem] menciona que la necesidad de entender conjuntos de datos muy grandes, complejos y llenos de información es muy común en todos los campos de negocios, ciencia e ingeniería. Y además, la habilidad para extraer conocimiento escondido en los datos y tomar decisiones frente a esos conocimientos, se ha vuelto muy importante para todas las organizaciones que quieran ser competitivas. Incluso, [HORNICK 2007], con el fin de resaltar la importancia de la Minería de Datos hoy en día, menciona que existen reglamentaciones (como Sarbanes, Oxley) que requieren el almacenamiento de grandes cantidades de datos históricos y muchas compañías han hecho grandes esfuerzos por recolectar prácticamente todo sobre sus negocios y asegurar que los datos sean confiables, limpios y accesibles. En consecuencia, los ejecutivos quieren utilizar este costoso recurso en un buen uso. Entonces es allí donde surge la pregunta: ¿Cómo poder aprovechar estas grandes cantidades de datos? ¿Cómo pueden las organizaciones extraer valor de esas

grandes cantidades de registros que existen acerca de sus clientes? La respuesta es Minería de Datos. [Ídem] da una definición general de la Minería de Datos como el "proceso de aplicar metodologías y técnicas utilizadas para descubrir en los datos conocimiento representado en patrones y relaciones entre los datos. Minería de Datos, consiste en construir un modelo, que es una representación de los patrones encontrados utilizando datos históricos, y aplicar ese modelo en nuevos datos para verificar si cumplen con los mismos patrones".

[ibídem] hace énfasis en que actualmente, las compañías que no utilizan Minería de Datos en sus procesos de negocios probablemente no van a poder sacar el mayor provecho de sus datos para mejorar sus ingresos y obtener mejores ganancias. La experiencia de sus clientes puede ser inferior al realizar ofertas irrelevantes y solicitudes al azar. Las compañías pueden estar confundidas a la hora de tratar de determinar por qué sus clientes los están abandonando o cuál es el perfil de sus clientes que pueden generar mayor valor a la organización. Por esta razón, la Minería de Datos es un factor competitivo para las organizaciones.

Profundizando más en el concepto de Minería de Datos, [KANTARDZIC 2003] describe cómo en la práctica, las dos principales metas de la Minería de Datos son predicción y descripción. Predicción involucra el utilizar algunas variables o campos en el conjunto de datos para predecir valores futuros desconocidos de otras variables de interés. Descripción, por el otro lado, se centra en encontrar patrones que describan los datos y que además puedan ser interpretados por los humanos. Para predicción, la meta de Minería de Datos es el producir un modelo expresado como un código ejecutable que pueda ser utilizado para realizar clasificación, predicción, estimación u otras tareas similares. Para descripción, la meta de Minería de Datos es el obtener entendimiento del sistema analizado descubriendo patrones y relaciones en grandes conjuntos de datos. Las metas de predicción y descripción se alcanzan utilizando técnicas de Minería de Datos para predecir comportamiento individual (clasificación y regresión), segmentar una

población (clustering), como también para identificar características que puedan impactar a un resultado en particular (importancia de atributos).

Por ejemplo, La Minería de Datos es utilizada para conocer a qué clientes seleccionar para una campaña en específico. Las empresas pueden determinar qué aspectos de sus procesos están arrojando menores resultados y por qué. Los proveedores de servicios financieros, como bancos, pueden conocer qué clientes presentan un mayor riesgo para un préstamo y conocer el comportamiento transaccional de sus clientes para determinar si están en peligro de fraude.

Los más importantes orígenes de la Minería de Datos, según [Ídem], son la estadística y el aprendizaje automático. Las bases de estadística son la matemática y como contraste el aprendizaje automático tiene sus orígenes en la computación. La estadística moderna está muy orientada hacia la noción de modelos, es decir, hacia la estructura o aproximación de estructura que podría tener los datos. Por otro lado, aprendizaje automático, enfatiza en los algoritmos, es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Esto lleva a la Minería de Datos a una orientación práctica, a la motivación de probar algo para analizar cómo es su rendimiento sin la necesidad de esperar pruebas formales o efectividad.

En la actualidad existe una gran variedad de herramientas software que facilitan realizar procesos de Minería de Datos. Incluso hoy en día, los usuarios no necesitan ni siquiera tener conocimientos avanzados acerca de las técnicas de Minería de Datos para aprovechar sus beneficios. Los algoritmos y el proceso de Minería de Datos han sido ocultos por interfaces de usuario amigables, que presentan los resultados de Minería de Datos sin la necesidad de conocer los complicados mecanismos algorítmicos que se requieren para obtener los resultados. Por esta razón, la Minería de Datos se ha vuelto mucho más accesible

a mucha clase de usuarios. Incluso, los no expertos en análisis de datos pueden sacar grandes provechos de su uso.

La Minería de Datos permite reducir costos, incrementar ganancias, generar nuevos descubrimientos, automatizar tareas muy laboriosas, identificar fraude, mejorar la experiencia del usuario o cliente y muchos otros beneficios para las organizaciones. Es así que la Minería de Datos se ha convertido en una estrategia competitiva para todas las industrias y proyectos.

2.1.2. Proceso general de la Minería de Datos. Existen muy buenos esfuerzos por estandarizar los procesos de Minería de Datos. [HORNICK 2007] y [OLSON 2008] describen dos de los principales estándares de Minería de Datos que actualmente existen llamados CRISP-DM (Cross-Industry Standard Process for Data Mining) y SEMMA (sample, explore, modify, model, assess), los cuales definen el proceso que se debe seguir para realizar Minería de Datos.

[KANTARDZIC 2003] y [Two Crows 2007] se refieren al proceso de Minería de Datos de forma general el cual puede ser definido como un proceso iterativo. En primera instancia, los datos son examinados con algún tipo de técnica analítica, para luego mirar los datos desde otra perspectiva quizá realizando modificaciones sobre los datos y luego comenzar nuevamente desde el principio utilizando otra herramienta de análisis de datos obteniendo los mismos o mejores resultados. Esto puede repetirse varias veces, cada técnica es utilizada para probar los diferentes aspectos de los datos y hacer diferentes preguntas acerca de los datos. Es un proceso que si se realiza de forma bien planeada puede llevar a conclusiones sobre los datos acerca de qué es más prometedor y revelador. En forma resumida, el proceso en general involucra los siguientes pasos:

- **Conocer el negocio y el problema:** Hay quienes tienden a enfocarse en la técnica de Minería de Datos sin tener en cuenta una clara definición del

negocio y el problema que se pretende resolver utilizando Minería de Datos. Pueden formularse muchas hipótesis iniciales en esta etapa del proceso de Minería de Datos y realmente se requiere de experticia tanto en el dominio de aplicación de los datos, como en el campo de Minería de Datos para poder obtener buenos resultados. En la práctica, esto significa una interacción muy cercana entre el experto en Minería de Datos y el experto en el problema y esta interacción debe continuar durante todo el proceso de Minería de Datos.

- **Reunir los datos:** Esta etapa se enfoca en cómo los datos son generados, reunidos y cómo asegurar la calidad de las muestras de datos para poder así obtener resultados satisfactorios a la hora de realizar los procesos de análisis.

- **Pre procesar los datos:** Una vez recolectados los datos, es necesario detectar y eliminar valores inusuales en los datos, errores de codificación, errores de guardado. Estos datos no representativos, pueden afectar seriamente el modelo producido posteriormente. En esta etapa también se busca recodificar los datos para que todos cumplan con los mismos estándares de medida y dimensión, puesto que estas diferencias también pueden alterar dramáticamente los resultados de los análisis.

- **Estimar el modelo:** La principal tarea en esta etapa del proceso, es lograr seleccionar e implementar la técnica de Minería de Datos más adecuada según la definición del problema. Generalmente en la práctica, la implementación está basada en la generación de muchos modelos para ser evaluados y finalmente seleccionar el mejor.

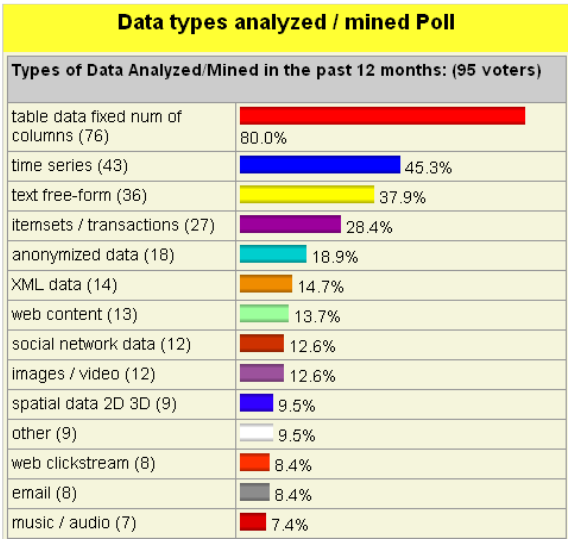
- **Interpretación y despliegue del modelo:** En la mayoría de los casos, los modelos de Minería de Datos deben ayudar a tomar decisiones. Por esta

razón, los modelos generados deben ser fáciles de interpretar e integrar con el negocio.

2.1.3. Característica de los datos. [KANTARDZIC 2003] Habla acerca de la clasificación de los datos que son la fuente esencial para realizar Minería de Datos. pueden ser clasificados en datos estructurados, datos semi estructurados y datos no estructurados. Los datos estructurados consisten en campos muy bien definidos con valores numéricos o alfanuméricos, mientras los semi estructurados pueden ser imágenes, documentos, reportes, etc. Un ejemplo de datos no estructurados puede ser un video grabado por una cámara de vigilancia. Estos tipos de datos no estructurados requieren un mayor esfuerzo de procesamiento para poder extraer y estructurar la información contenida en ellos.

Los datos estructurados son generalmente llamados datos tradicionales, mientras los datos semi estructurados y no estructurados son llamados datos no tradicionales o datos multimedia. La mayoría de las herramientas de Minería de Datos existentes están enfocadas hacia los datos tradicionales, sin embargo, el desarrollo de herramientas de Minería de Datos para datos no tradicionales también ha venido progresando rápidamente.

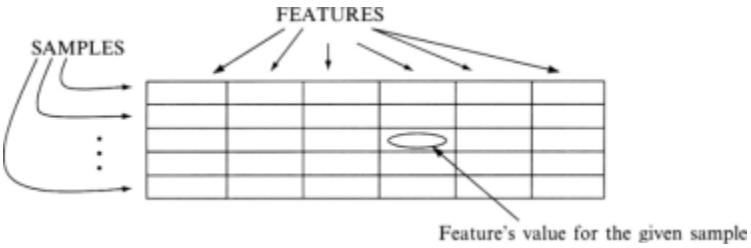
Ilustración 2: Encuesta: Tipos de datos analizados



Extraído de www.kdnuggets.com - 2008

El modelo estándar de datos estructurados puede verse como una colección de casos (también llamado muestras) los cuales poseen un número de características. Usualmente, la forma de representar los datos estructurados para problemas de Minería de Datos es de forma tabular, donde las columnas son las características de los objetos almacenados en la tabla y las filas las muestras o casos que poseen dichas características.

Ilustración 3: Representación tabular de los datos



Extraído de [KANTARDZIC 2003]

2.1.4. Calidad de los datos. Este es un aspecto muy importante a tener en cuenta si verdaderamente se quiere obtener modelos de Minería de Datos confiables, por esta razón, no se puede entrar inmediatamente de lleno a aplicar las técnicas de Minería de Datos sin antes asegurar que los datos sean confiables, es importante realizar primero un análisis de la calidad de este insumo principal. Es obvio que la calidad de los datos tiene un profundo efecto en los resultados que los análisis puedan arrojar, si no se asegura la calidad, los resultados de las decisiones tomadas pueden llegar a tener efectos catastróficos. Por esta razón, algunos autores mencionan pautas que se deben seguir para asegurar la calidad de los datos:

- Los datos deben ser exactos. Se debe verificar que los nombres de los campos están escritos correctamente, que los rangos estén bien definidos, que los valores estén completos, etc.

- Los datos deben estar almacenados de acuerdo a su tipo de dato. Se debe verificar que los datos numéricos no estén representados como datos de caracteres, que los enteros no estén en forma de números reales, etc.
- Los datos deben tener integridad.
- Los datos deben ser consistentes. La forma y el contenido deben seguir siendo los mismos aún después de ser integrados con otros conjuntos de datos de otras fuentes.
- Los datos no deben ser redundantes ni duplicados.
- Los datos deben estar asociados al tiempo.
- Los datos deben ser entendibles, es bueno utilizar nombres estándares que permitan al usuario identificar a qué dominio pertenecen.
- El conjunto de datos debe estar completo. Se debe reducir a lo máximo la existencia de datos perdidos.

2.1.5. ¿Qué se necesita para la Minería de Datos? Antes de realizar cualquier proceso de Minería de Datos, se debe estudiar si realmente se cuenta con todos los elementos necesarios para lograr obtener resultados exitosos. [OLSON 2008] enseña que la Minería de Datos requiere identificar un problema junto con una colección de datos que pueda dar un mejor entendimiento acerca del problema. Además se debe contar con herramientas de visualización que muestren los datos y análisis estadísticos fundamentales, y herramientas de Minería de Datos que cuenten con los algoritmos apropiados para dar solución al problema. Las herramientas de Minería de Datos deben tener la capacidad de aplicar gran cantidad de modelos escalables, capaces de predecir con exactitud respuestas entre acciones y resultados, capaces de implementación de automatizaciones. Se requiere una selección apropiada de los datos y la transformación necesaria de estos para generar resultados confiables. Se requiere seleccionar la cantidad de variables apropiadas y un entendimiento fundamental acerca de los conceptos estadísticos que permitan comprender los resultados.

2.1.6. ¿Qué es realmente la Minería de Datos? Hay que tener mucho cuidado a la hora de definir qué es la Minería de Datos, pues hay quienes mal interpretan su funcionalidad y no logran sacar mayor provecho a lo que realmente Minería de Datos es. Es así, que [HORNICK 2007] nos ejemplifica y dice que cuando se le pregunta a algunos usuarios de Minería de Datos si ellos realmente minan sus datos, es común escuchar: "claro que sí". Pero, cuando se les pregunta cómo minan sus datos o qué herramientas utilizan, dicen tener una base de datos muy grande y la analizan con consultas SQL complejas. Para algunos, esto es Minería de Datos.

[Ídem] enfatiza diciendo que el tema de consultas y reportes es simplemente un análisis deductivo, esto es, extracción de los detalles y sumarizar los datos basados en preguntas formuladas por humanos. Por ejemplo, responder preguntas como "qué tiendas vendieron reproductores DVD el trimestre pasado" y "cuanto se hizo en cada venta" es común. Es importante entender que este tipo de consultas y preguntas no es Minería de Datos en el sentido de aprendizaje automático. Responder estas preguntas puede hacerse a través de consultas SQL de manera muy sencilla.

En contraste, como varios de los anteriores autores mencionados afirman, el verdadero sentido de la Minería de Datos es el lograr descubrir conocimiento de patrones ocultos en los datos. Esto arroja un aprovechamiento inductivo del análisis de los datos obteniendo así resultados al analizar cada uno de los potenciales millones de registros. Minería de Datos permite responder aquellas preguntas de cuantos ingresos generará cada tienda de la venta de algún producto en específico durante los siguientes meses, o qué clientes van a comprar ciertos productos y por qué, o incluso predecir qué clientes se van a cambiar a la competencia y de esta forma poder enfocar de una manera más efectiva una campaña de mercado.

2.1.7. ¿Qué no es la Minería de Datos? Una vez entendida la importancia de la Minería de Datos, también hay que reconocer sus límites. [Two Crows 2007] nos da aspectos claves a tener en cuenta para no caer en errores referentes a la Minería de datos: “La Minería de Datos es una herramienta, no una varita mágica. No se sentará a observar qué sucede en su base de datos y enviarle un e-mail para informarle cuando halle un patrón interesante. No elimina la necesidad de conocer su negocio, de comprender sus datos, o de entender los métodos analíticos. La Minería de Datos ayuda a los analistas de negocio con la búsqueda de patrones y relaciones en los datos, no a valorar los patrones de la organización. Además, los patrones descubiertos por la Minería de Datos deben ser verificados en el mundo real”.

Como lo explica [Ídem], los resultados obtenidos utilizando la Minería de Datos, no son necesariamente las causas exactas de una acción o comportamiento. Por ejemplo, la Minería de Datos podría determinar que los hombres que están en cierto rango de edad con cierta cantidad de ingresos que se suscriben a ciertas revistas, es probable que sean los compradores de un producto que se desea vender. Mientras que usted puede tomar ventaja de este modelo, por ejemplo con sus campañas de marketing dirigidas a personas que encajan en el patrón, usted no debe suponer que alguno de estos factores haga que los clientes vayan a comprar su producto.

Para asegurar resultados significativos, es vital entender los datos y comprender muy bien cómo funciona el negocio. Hay que recordar que la calidad de los resultados es sensible a los valores atípicos (valores en los datos que son muy diferentes de los valores típicos de la base de datos), las columnas irrelevantes o columnas que varían entre sí (como la edad y fecha de nacimiento), la forma de codificar los datos, los datos que se dejan y los datos que se excluyen. Es allí donde juega un papel muy importante la calidad de los datos.

Los algoritmos pueden variar en su sensibilidad de acuerdo a las características de los datos. Además, no es conveniente depender de una única herramienta de Minería de Datos para tomar todas las decisiones. [ibídem] lo explica diciendo que la Minería de Datos no puede de forma automática encontrar soluciones sin orientación y un análisis y pruebas para verificar que los resultados sean correctos. En lugar de establecer el objetivo, "Contribuir a mejorar la respuesta a mi solicitud por correo directo," se podría utilizar la Minería de Datos para encontrar las características de las personas que: responden a una solicitud, o responden y además hacen una gran compra. Los patrones que la a Minería de Datos encuentra en esos dos objetivos pueden ser muy diferentes.

Una buena recomendación, es que aunque una buena herramienta de Minería de Datos nos libra de las complejidades de las técnicas estadísticas y los algoritmos de minería, sí es necesario entender el funcionamiento de las herramientas que sean elegidas y los algoritmos en los que se basan, puesto que las decisiones que se tomen en la configuración de la herramienta de Minería de Datos y la parametrización de sus algoritmos pueden afectar a la precisión y la velocidad de los modelos.

Otro punto importante mencionado por el mismo autor, es que la Minería de Datos no sustituye a los analistas especializados en el negocio o a los administradores, sino más bien les da una poderosa herramienta para mejorar el trabajo que están haciendo. Cualquier empresa que conoce su negocio y sus clientes ya es consciente de muchos patrones importantes de comportamientos que sus empleados han observado en los clientes a través de los años. Lo que puede hacer la Minería de Datos es confirmar tales observaciones empíricas y encontrar nuevos patrones sutiles que pueden mejorar incrementalmente el rendimiento del negocio.

La Minería de Datos no reemplaza las técnicas tradicionales de estadística. El desarrollo de la mayoría de las técnicas estadísticas, hasta hace poco, estaba basado en la teoría y métodos analíticos que funcionaron bastante bien en las pequeñas cantidades de datos que se analizan. El aumento de la capacidad de los ordenadores y su menor costo, junto con la necesidad de analizar enormes conjuntos de datos con millones de registros, han permitido el desarrollo de nuevas técnicas y herramientas que facilitan el análisis.

2.1.8. Confiabilidad de la Minería de Datos. Muchos pueden llegar a preguntarse ¿qué tan confiable es la Minería de Datos? [HORNICK 2007] nos da la respuesta asegurando que la confiabilidad de los resultados generalmente depende de la disponibilidad de suficientes datos, la calidad de los datos, la técnica seleccionada y las habilidades de aquellos que preparan los datos, seleccionan los parámetros para los algoritmos y analizan los resultados. Si los datos contienen valores erróneos o perdidos, los algoritmos de Minería de Datos pueden tener dificultades para descubrir patrones significativos en los datos. Por tal motivo no basta únicamente con tener la herramienta sino también se requiere personal experto que sepa explotar su potencial.

Cuando un modelo de Minería de Datos es generado por primera vez, este puede ser bastante confiable en términos de exactitud en sus predicciones sobre nuevos datos. Sin embargo, la calidad de un modelo puede ser variable y la exactitud del modelo no es constante en el tiempo. Pocas cosas permanecen constantes, especialmente cuando los seres humanos están involucrados. Los gustos cambian, las necesidades cambian, la tecnología cambia, existen muchos eventos en la vida fuerzan el cambio. Por ejemplo, un modelo que puede haber sido excelente para predecir riesgos en tarjeta de crédito, pero para un mes en específico puede comenzar a mostrar signos de mal rendimiento. Para este caso, el modelo necesita ser reconstruido utilizando datos más recientes. La Minería de

Datos y sus artefactos requieren revisión periódica y mantenimiento para mantener resultados confiables.

2.2. Técnicas y Herramientas de la Minería de Datos

Para entender de mejor manera cuál es el alcance y el potencial que puede brindar la Minería de Datos frente al análisis de los datos, es importante conocer las técnicas y herramientas que actualmente existen y para qué sirven. La Minería de Datos ha evolucionado en una gran cantidad de algoritmos que permiten análisis específicos de los datos y generan descripción de patrones en los datos que permiten tomar decisiones más acertadas y así generar valor para la organización.

2.2.1. Aprendiendo de los datos. [KANTARDZIC 2003] hace una comparación acerca del desarrollo de los modelos de Minería de Datos con la capacidad de los sistemas biológicos de aprender, en particular, la de los humanos. Los sistemas biológicos tienen la capacidad de aprender de lo desconocido. Los humanos y animales tienen capacidades superiores para reconocer patrones y así poder realizar tareas como identificar rostros, voces, olores, etc. Las personas no nacen con dichas capacidades sino que las aprenden a través de los datos obtenidos de la interacción con el ambiente.

Todo proceso de aprendizaje consiste en 2 fases principales:

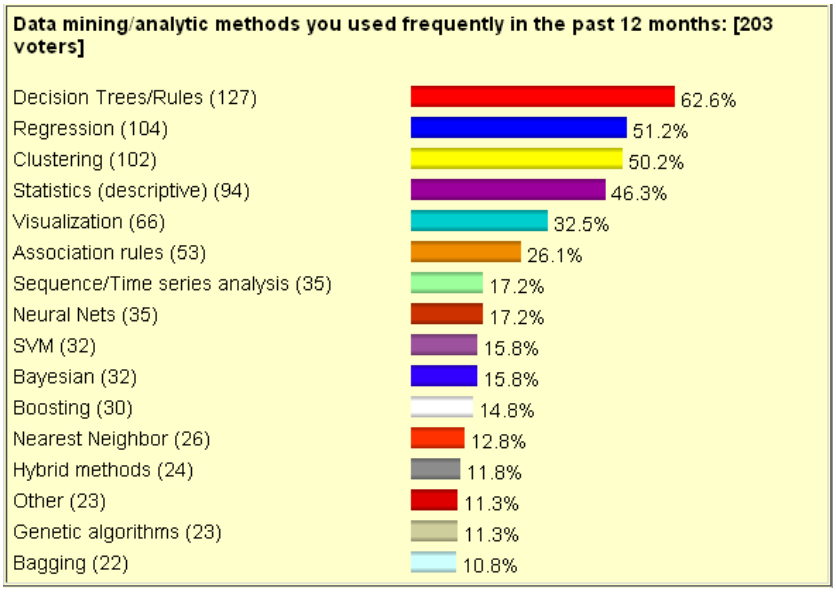
- Aprender o estimar dependencias desconocidas en el sistema a través de un conjunto o muestra de datos dados.
- Usar las dependencias estimadas para predecir nuevas salidas para futuros valores de entrada en el sistema.

Básicamente, este es el proceso utilizado por los algoritmos de Minería de Datos para generar modelos que permitan el descubrimiento de patrones sobre los datos o predecir futuros comportamientos.

Los algoritmos de Minería de Datos se pueden clasificar en varios tipos: estadísticos, clasificación, asociación, segmentación, algoritmos genéticos, lógica difusa y métodos de visualización.

A continuación, se presenta una encuesta sobre los métodos de Minería de Datos más utilizados:

Ilustración 4: Encuesta: Técnicas más usadas en la Minería de Datos



Extraído de www.kdnuggets.com - 2007

A continuación, se describirán algunos algoritmos ampliamente utilizados en la Minería de Datos.

2.2.1.1. Algoritmos estadísticos. La estadística es la ciencia de recolectar y organizar datos y sacar conclusiones de los conjuntos de datos. La descripción de las características generales de los conjuntos de datos es el objetivo de la estadística descriptica (Extraído de [SPSS 2008] y [KANTARDZIC 2003]).

METODO	DESCRIPCIÓN
Inferencia estadística	Comprende los métodos y procedimientos para deducir propiedades (hacer inferencias) de una población, a partir de una pequeña parte de la misma (muestra).
Evaluación de diferencias entre conjuntos de datos	Consiste en estadísticos como son la tendencia central y la dispersión de los datos, que permiten evaluar las diferencias en los conjuntos de datos. Los más comunes son la media, la mediana y la varianza.
Regresión logística	Es una técnica estadística para clasificar los registros a partir de los valores de los campos de entrada. La regresión logística trabaja creando un conjunto de ecuaciones que relacionan los valores de los campos de entrada con las probabilidades asociadas a cada una de las categorías de los campos de salida. Una vez se ha generado el modelo, se puede utilizar para calcular las probabilidades de datos nuevos. Para cada registro, se calcula una probabilidad de pertenencia a cada categoría posible de salida. La categoría objetivo con la probabilidad más alta se asigna como el valor de salida pronosticado para cada registro.
Análisis lineal de discriminante	Busca generar una función discriminante (o, para más de dos grupos, un conjunto de funciones

	discriminantes) basada en combinaciones lineales de las variables de predictor que ofrecen la mejor discriminación entre los grupos. Las funciones se generan a partir de una muestra de casos cuya pertenencia al grupo se conoce; las funciones se pueden aplicar entonces a nuevos casos con mediciones para las variables de predictor pero con una pertenencia al grupo desconocida.

2.2.1.2. Algoritmos de clasificación. Este tipo de modelos permiten pronosticar un resultado conocido, como saber si un cliente comprará o se irá, o si una transacción se ajusta a un patrón conocido de fraude. Los siguientes son algunos algoritmos de clasificación (Extraído de [SPSS 2008]):

METODO	DESCRIPCIÓN
Clasificación y regresión (C&R)	Es un método de pronóstico y clasificación basado en árboles, genera un árbol de decisión que permite pronosticar o clasificar observaciones futuras. C&RT comienza por realizar un examen de los campos de entrada para buscar la mejor división, que se ha medido mediante la reducción del índice de impureza resultado de la división. La división define dos subgrupos, que se siguen dividiendo en otros dos subgrupos sucesivamente hasta que se activa un criterio de parada. Todas las divisiones son binarias (sólo se crean dos subgrupos).
QUEST (árbol estadístico eficiente	Es un método de clasificación binario para generar árboles de decisión. Consiste en reducir la tendencia

insesgado y rápido)	de los métodos de clasificación de árboles para favorecer a los predictores que permiten realizar más divisiones, es decir, las variables predictoras continuas o las correspondientes a varias categorías
CHAID	Detección automática de interacciones mediante chi-cuadrado (del inglés Chi-squared Automatic Interaction Detection), es un método de clasificación para generar árboles de decisión mediante estadísticos de chi-cuadrado para identificar divisiones óptimas
Listas de decisiones	Identifican subgrupos o segmentos que muestran una mayor o menor posibilidad de proporcionar un resultado binario (sí o no) relacionado con la muestra global. Por ejemplo, puede buscar clientes con menos posibilidades de pérdida o con más posibilidades de decir sí a una campaña u oferta determinada.
Regresión lineal	Es una técnica de estadístico común utilizada para resumir datos y realizar pronósticos ajustando una superficie o línea recta que minimice las discrepancias existentes entre los valores de salida reales y los pronosticados.
PCA/Factorial	Proporciona técnicas eficaces de reducción de datos para reducir la complejidad de los datos. El objetivo es encontrar un número pequeño de campos derivados que resuman de forma eficaz la información del conjunto original de campos.
Red neuronal	Utiliza un modelo simplificado que emula el modo en que el cerebro humano procesa la información. Las

	unidades básicas son las neuronas, que generalmente se organizan en capas. La red aprende examinando los registros individuales, generando un pronóstico para cada registro y realizando ajustes a las ponderaciones cuando realiza un pronóstico incorrecto. Este proceso se repite muchas veces y la red sigue mejorando sus pronósticos hasta haber alcanzado uno o varios criterios de parada.
C5.0	Genera un árbol de decisión o un conjunto de reglas. Los modelos C5.0 dividen la muestra en función del campo que ofrece la máxima ganancia de información. Las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible dividir las submuestras de nuevo. Por último se vuelven a examinar las divisiones del nivel inferior, y se eliminan o podan las que no contribuyen significativamente con el valor del modelo
Selección de características	Puede que los problemas relacionados con la minería de datos impliquen cientos, o incluso miles, de campos que se pueden utilizar potencialmente como predictores. Por consiguiente, puede que se invierta mucho tiempo y esfuerzo en examinar qué campos o variables se incluirán en el modelo. Para limitar las opciones, se puede utilizar el algoritmo Selección de características para identificar los campos que son más importantes para un análisis específico
Modelo lineal	Amplía el modelo lineal general para que la variable

generalizado	dependiente esté linealmente relacionada con los factores y covariables a través de una función de enlace especificada.
Red bayesiana	Permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real para establecer la probabilidad de instancias.
Regresión de Cox	La regresión de Cox crea un modelo predictivo para datos de tiempo hasta el evento. El modelo produce una función de supervivencia que pronostica la probabilidad de que el evento de interés se haya producido en el momento dado t para valores determinados de las variables del predictor. La forma de la función de supervivencia y los coeficientes de regresión para los predictores se calculan a partir de los sujetos observados; a continuación, el modelo puede aplicarse a nuevos casos que tengan medidas para las variables del predictor.
Máquina de vectores de soporte (SVM)	Permite utilizar una máquina de vectores de soporte para clasificar los datos. SVM es ideal para conjuntos de datos grandes, es decir, con un gran número de campos predictores. Puede utilizar la configuración por defecto en el nodo para producir un modelo básico relativamente rápido o puede utilizar la configuración de Experto para experimentar con tipos diferentes del modelo SVM
Modelo de respuesta de auto aprendizaje (SLRM)	Permite crear un modelo en el que un solo caso nuevo o un pequeño número de casos nuevos se pueden utilizar para volver a calcular el modelo sin

	tener que entrenar de nuevo el modelo utilizando todos los datos.
Serie temporal	Estima modelos de suavizado exponencial, modelos autor regresivos integrados de media móvil (ARIMA) univariados y modelos ARIMA (o de función de transferencia) multivariados para series temporales y genera datos de predicciones.

2.2.1.3. Algoritmos de asociación. Permiten pronosticar varios resultados; por ejemplo, los clientes que adquirieron el producto X también adquirieron Y y Z. Los siguientes son algunos algoritmos de asociación (Extraído de [SPSS 2008]):

METODO	DESCRIPCIÓN
Inducción de reglas generalizado (GRI)	Es capaz de encontrar las reglas de asociación existentes en los datos. Por ejemplo, si un cliente compra una cuchilla y una loción para después del afeitado, hay un 80% de probabilidades de que el cliente también compre la crema de afeitado.
A priori	Extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. Por ejemplo, "si un cliente compra una cuchilla y una loción para después del afeitado, hay un 80% de posibilidades de que el cliente compre también crema de afeitado". A priori extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. A priori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema

	de indización para procesar conjuntos de datos de gran tamaño de forma eficaz.
CARMA	Utiliza un algoritmo de descubrimiento de reglas de asociación para encontrar reglas de asociación existentes en los datos. Por ejemplo, si un cliente del sitio Web adquiere una tarjeta y un enrutador de gama alta inalámbricos, es muy probable que también adquiera un servidor de música inalámbrico si se le ofrece. El modelo CARMA extrae un conjunto de reglas de los datos sin necesidad de especificar campos de Entrada (predictor) ni de Salida (objetivo). Esto significa que las reglas generadas se pueden utilizar en una variedad de aplicaciones mucho más amplia.
Secuencia	Descubre patrones, en datos secuenciales u ordenados en el tiempo, con el formato pan > queso. Los elementos de una secuencia son conjuntos de elementos que constituyen una única transacción. Por ejemplo, si una persona va a la tienda y compra pan y leche y, varios días después, vuelve a la tienda para comprar un poco de queso, la actividad de compras de esa persona se puede representar como dos conjuntos de elementos. El primer conjunto de elementos contiene pan y leche y el segundo contiene queso. Una secuencia es una lista de conjuntos de elementos que tiende a producirse en un orden previsible. Este algoritmo detecta secuencias frecuentes y crea un modelo que se puede utilizar para realizar pronósticos.

2.2.1.4. Algoritmos de segmentación: Se centran en la identificación de grupos de registros similares y en el etiquetado de registros según el grupo al que pertenecen. Esto se lleva a cabo sin la ventaja que ofrece el conocimiento previo sobre los grupos y sus características, y diferencia a los modelos de conglomerados de otras técnicas de modelado en que no hay campos de salida u objetivo predefinidos para el modelo que se va a pronosticar. No hay respuestas correctas o incorrectas para estos modelos. Su valor viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones. Los siguientes son algunos algoritmos de segmentación (Extraído de [SPSS 2008]):

METODO	DESCRIPCIÓN
K-medias	Ofrece un método de análisis de conglomerados. Se puede utilizar para conglomerar el conjunto de datos en distintos grupos cuando no se sabe lo que son al principio. Los modelos de K-medias no utilizan un campo objetivo. Este tipo de aprendizaje, sin campo objetivo, se denomina aprendizaje no supervisado. En lugar de intentar predecir un resultado, los modelos de K-medias intentan revelar los patrones en el conjunto de campos de entrada. Los registros se agrupan de manera que los de un mismo grupo o conglomerado tiendan a ser similares entre ellos, y que los de otros grupos sean distintos.
Kohonen	Son un tipo de red neuronal que realiza conglomerados, también conocidos como knet o como un mapa auto organizativo. Este tipo de redes se puede utilizar para conglomerar el conjunto de datos en distintos grupos cuando no se sabe lo que

	son al principio. Los registros se agrupan de manera que los de un mismo grupo o conglomerado tiendan a ser similares entre ellos y que los de otros grupos sean distintos.
Bietápico	Se puede utilizar para conglomerar el conjunto de datos en distintos grupos cuando no se sabe lo que son al principio. Al igual que Kohonen y K-medias, los modelos de conglomerados bietápicos no utilizan un campo objetivo. En lugar de intentar predecir un resultado, el conglomerado Bietápico intenta revelar los patrones en el conjunto de campos de entrada. Los registros se agrupan de manera que los de un mismo grupo o conglomerado tiendan a ser similares entre ellos, y que los de otros grupos sean distintos.
Detección de anomalías	Se utilizan para identificar valores atípicos, o casos extraños, en los datos. A diferencia de otros métodos de modelado que almacenan reglas acerca de casos extraños, los modelos de detección de anomalías almacenan información sobre el patrón de comportamiento normal. Esto permite identificar valores atípicos, incluso si no se ajustan a ningún patrón conocido, y puede ser especialmente útil en aplicaciones, como detección de fraudes, donde pueden surgir patrones nuevos constantemente. La detección de anomalías es un método no supervisado, lo que significa que no requiere un conjunto de datos de entrenamiento que contenga casos conocidos de fraudes para utilizarlos como punto de partida.

2.2.1.5. Algoritmos genéticos. [LAROSE 2006] da a entender el concepto de los algoritmos genéticos comparándolos con la evolución. Los algoritmos genéticos buscan imitar el proceso de cómo actúa la selección natural y aplicarlo a la solución de problemas de investigación y negocios. [KANTARDZIC 2003] explica cómo los problemas que las especies biológicas han resuelto tienen que ver con el caos, la suerte, temporalidad, y otros tipos de interactividad no lineal. Estas son las características de los problemas que han demostrado ser especialmente intratables utilizando métodos clásicos de optimización. Los algoritmos genéticos aproximan una solución óptima al problema aplicando métodos de búsqueda modelando algunos fenómenos naturales como herencia genética y la lucha por supervivencia. Los algoritmos genéticos han sido aplicados exitosamente en problemas como rutas, horarios, juegos, problemas de transporte, optimización de consultas a bases de datos, aprendizaje automático, etc.

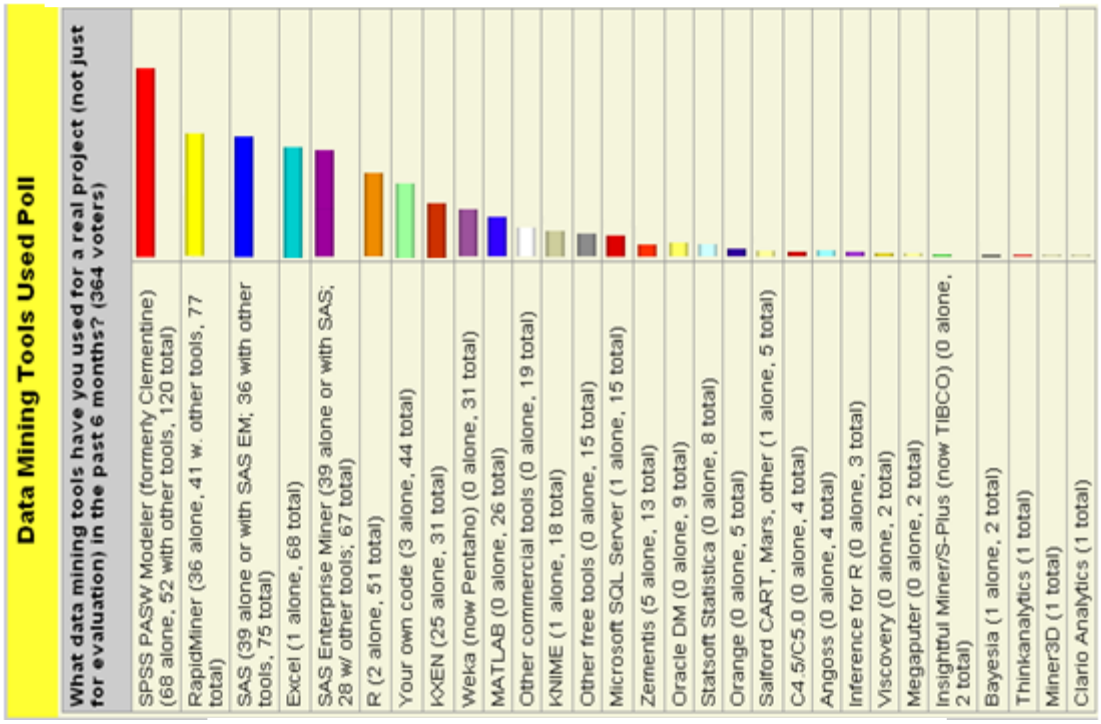
2.2.1.6. Conjuntos difusos y lógica difusa. [KANTARDZIC 2003] explica el concepto de difuso como un derivado de los fenómenos que comúnmente ocurren en el mundo real. Por ejemplo, la lluvia es un fenómeno natural común que es difícil de describir precisamente porque puede llover con intensidad variable, desde una tormenta hasta una llovizna. Como la palabra lluvia no describe adecuada o precisamente todos los tipos de variaciones y la cantidad intensidad en la lluvia, entonces, “lluvia” es considerado como un fenómeno difuso. También los conceptos formados en el cerebro humano para percibir, reconocer, y categorizar fenómenos naturales son difusos. Los conjuntos difusos y la lógica difusa permiten manejar lo incierto de una manera muy intuitiva y lógica, abriendo caminos para resolver problemas donde no existen modelos matemáticos precisos y donde las definiciones y conocimientos acerca del problema son imprecisos.

2.2.1.7. Métodos de visualización de datos. Este método consiste en el uso de herramientas para ilustrar gráficamente las relaciones entre los datos.

[HAND 2007] menciona que resulta muy útil incluir la capacidad humana en el proceso de exploración de los datos y así combinar la flexibilidad, creatividad y el conocimiento general de los humanos con la gran capacidad de almacenamiento y procesamiento de los computadores de hoy en día. Los seres humanos normalmente no piensan en términos de datos, ellos son inspirados y piensan en términos de imágenes mentales acerca de una situación y asimilan la información más rápida y eficientemente utilizando imágenes visuales en vez de texto o formatos tabulares. La visión humana es el medio más poderoso para deshacerse de información irrelevante y detectar patrones significativos. Los científicos han descubierto que ver y entender juntamente permite a los seres humanos el descubrir nuevo conocimiento con una comprensión más profunda de las grandes cantidades de datos.

2.2.2. Herramientas de Minería de Datos. Actualmente, en el mercado existen varias herramientas de Minería de Datos tanto comerciales como gratuitas. Las herramientas de Minería de Datos facilitan tareas como: limpieza de datos, cruce de datos, visualización de datos, ejecución de modelos de minería y visualización de resultados para análisis y toma de decisiones. Todo lo anterior mediante interfaces de usuario amigables que permiten parametrizar o modelar fácilmente el flujo de datos para su transformación y posterior análisis. El siguiente cuadro muestra una encuesta realizada sobre las herramientas de Minería de Datos más utilizadas:

Ilustración 5: Encuesta: Herramientas de Minería de Datos más usadas



Extraído de www.kdnuggets.com - 2009

A continuación, se presenta una descripción general de algunas de las herramientas de Minería de Datos más conocidas en el mercado:

Herramienta	Tipo	Propiedades	Algoritmos
SPSS PASW Modeler	Comercial	Interfaz visual fácil de usar Administración y colaboración: proyectos enfocados en CRSP-DM; Facilidad para preparar los datos y automatizar modelos Tipos de datos: Estructurados, no	C&RT, CHAID and QUEST, Decision List, K-means, GRI, Factor/PCA, Linear Regression, Nearest Neighbor, Binary classifier and numeric predictor, Self-learning response model, Time-series forecast models, C5.0 decision tree, Neural Networks, Support Vector Machines, Bayesian Networks, Cox Regression, Binomial and

		estructurados, Web Site Manipulación fácil de datos. Exportación de modelos	multinomial logistic regression, Discriminate analysis, General linear models (GLM), Auto Cluster, Kohonen Network, Two-Step Clustering, Anomaly Detection, A priori, CARMA, Sequential association algorithm
Rapid Miner	Libre	Corre en varios sistemas operativos. Diseño de procesos muy intuitivo. Manejo eficiente de los datos Interfaz gráfica. Mecanismo de extensión. Esquemas de optimización automáticos Acceso a diferentes fuentes de datos como Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, mySWL, Postgres, SPSS, dBase	Lazy Modeling, Bayensian, Tree induction, Rule induction, Redes neuronales, Funtion fitting, Regreción lógistica, Support Vector, Analisis discrimitante, Meta modeling, Data transformation, Attribute Weighting, Kmeans, DBSCAN, Random Clustering, Agglomerative clustering, Top Down Clustering, Flatten clurstering, Aprorie, Patrones secuenciales, Tertius, Correlación y dependencia, Similariry computation, Text processing
SAS Enterprise Miner	Comercial	Interfaz gráfica fácil de usar Procesamiento escalable Preparación de los datos, sumarización y exportación Modelos avanzados predictivos y descriptivos Modelos basados en el negocio, generación de reportes. Procesos de valoración automáticos Procesamiento basado	Clustering and self-organizing maps, Market basket analysis, Sequence and Web path analysis, Dimension reduction techniques, Variable selection, LARS (Least Angle Regression) variable selection, Principal components, Variable clustering, Time series mining, Manage time metrics with descriptive data, Linear and logistic regression, Decision trees, Gradient boosting, Neural networks, Partial least squares regression, Two-stage

		en servidor. Procesamiento paralelo y en grid	modeling, Memory-based reasoning, Model ensembles, including bagging and boosting
R	Libre	Múltiples fuentes: CSV, TXT, ODBC. Fácil exploración de los datos. Gráficas. Métodos de transformación de los datos.	Cluster: KMeans; Hierarchical (hclust) with Dendrogram and Seriation Plots; Associations: Apriori (arules) Market Basket; Modelling: Decision Tress (rpart); Generalised Linear Models; Boosting Random Forests; Support Vector Machines (kernlab); Evaluation: Confusion Matrix; Risk Chart; Lift Charts; ROC Curve and AUC (ROCR); Precision; Sensitivity; Interoperability: Export of models as PMML
WEKA	Libre	Implementado en Java, puede correr en casi cualquier plataforma. Extensa colección de técnicas para pre procesamiento de datos y modelado. Fácil de utilizar gracias a su interfaz gráfica de usuario.	Bayesian classifiers, Trees, Rules, Functions, Lazy classifiers, Miscellaneous classifiers, Bagging and randomization, Boosting, Combining classifiers, Cost-sensitive learning, Optimizing performance, Retargeting classifiers for different tasks, Clustering, Association-rule learners, Attribute selection
KMINE	Libre	Escalabilidad y manejo sofisticado de los datos, extensible, interfaz de usuario intuitiva, métodos de importación y exportación, ejecución paralela, múltiples fuentes de datos, manipulación fácil de los datos	Association Rules, Bayes, Clustering, Rule Induction, Neural Network, Decision Tree, K Nearest Neighbor, Multi dimensional scaling, Principal component analysis, SVM, Scoring

2.3. Aplicaciones de la Minería de Datos

En la actualidad, la Minería de Datos es ampliamente utilizada en muchos sectores de la industria. Son muchas las aplicaciones que se han encontrado a la capacidad de procesamiento y análisis que brindan las herramientas y los algoritmos de Minería de Datos. A continuación se hará una breve descripción de algunas de las aplicaciones más importantes de la Minería de Datos. Estas descripciones fueron extraídas de [KANTARDZIC 2003], [HAN 2006] y [HORNICK 2007].

Ilustración 6: Encuesta: Aplicaciones de Minería de Datos

Data Mining Applications 2009

Industries / Fields where you applied Data Mining in 2009 [180 votes total] See Comparison with 2008 results below	
CRM/ consumer analytics (59)	32.8%
Banking (44)	24.4%
Direct Marketing/ Fundraising (29)	16.1%
Credit Scoring (28)	15.6%
Telecom / Cable (26)	14.4%
Fraud Detection (25)	13.9%
Retail (21)	11.7%
Health care/ HR (21)	11.7%
Finance (20)	11.1%
Science (19)	10.6%
Advertising (19)	10.6%
e-Commerce (18)	10.0%
Insurance (18)	10.0%
Web usage mining (15)	8.3%
Social Networks (14)	7.8%
Medical/ Pharma (14)	7.8%
Biotech/Genomics (14)	7.8%
Search / Web content mining (12)	6.7%
Investment / Stocks (12)	6.7%
Security / Anti-terrorism (9)	5.0%
Education (8)	4.4%
Government/Military (7)	3.9%
Manufacturing (6)	3.3%
Travel / Hospitality (5)	2.8%
Social Policy/Survey analysis (3)	1.7%
Entertainment/ Music (3)	1.7%
Junk email / Anti-spam (1)	0.6%
None (2)	1.1%
Other (14)	7.8%

Extraído de www.kdnuggets.com - 2009

2.3.1. Análisis financiero. La mayoría de los bancos e instituciones financieras ofrecen una gran cantidad de servicios, como cheques, ahorros, negocios, transacciones de clientes, servicios de inversión, créditos, préstamos, entre otros. Los datos financieros recolectados por los bancos y la industria financiera son muy completos, confiables y de alta calidad, esto facilita grandemente el análisis de datos y la Minería de Datos para así lograr aumentar la competitividad de la compañía.

En los bancos, generalmente la Minería de Datos es usada para modelar y predecir fraude, evaluar riesgos, analizar tendencias, analizar ganancias, campañas de mercadeo, etc. Generalmente se utilizan redes neuronales para realizar pronósticos de precios, de canje, análisis de tasas, predecir desastres financieros, etc.

2.3.2. Telecomunicaciones. La industria de las telecomunicaciones rápidamente ha evolucionado desde ofrecer servicios telefónicos locales y de larga distancia, hasta ofrecer muchos otros servicios como fax, móvil, imágenes, email, internet, etc. Todo esto gracias a la integración de las telecomunicaciones, las redes de computadores, Internet, y otros tipos de comunicación.

Esto ha convertido esta industria en un negocio altamente competitivo, generando así la necesidad de entender mejor a los clientes, asegurarlos y generar estrategias para crear nuevos productos, nuevas ofertas. Esto crea una gran demanda de Minería de Datos para ayudar a entender el nuevo negocio, identificar patrones en las telecomunicaciones, capturar actividades fraudulentas, hacer un uso mejor de los recursos, mejorar la calidad de los servicios. En general, la industria de las telecomunicaciones está interesada en utilizar Minería de Datos para lograr responder algunas preguntas estratégicas como:

- Como retener a los clientes y mantenerlos leales a pesar de las ofertas de la competencia, las ofertas especiales y la reducción de los precios.

- Qué clientes van a quejarse
- Cuándo una inversión riesgosa, como fibra óptica, es aceptable
- Cómo predecir si un cliente comprará productos adicionales
- ¿Qué características diferencian nuestros productos a los de los competidores?

2.3.3. Industria Minorista. Los minoristas, han mejorado el proceso de toma de decisiones gracias la mejora de eficiencia en la administración de sus inventarios y pronósticos financieros. La necesidad que tienen hoy en día de almacenar y analizar sus datos les permite tener una gran oportunidad para sacar ventaja de la Minería de Datos. La industria minorista es el área que más ha aplicado Minería de Datos desde que almacenan grandes cantidades de datos de sus ventas, historia de compras de sus clientes, transportes de mercancía, patronos de consumo, etc. La cantidad de datos recolectados continuamente se expande rápidamente incluso gracias a la WEB. Este tipo de datos es un excelente recurso para Minería de Datos.

Para este tipo de industria, la Minería de Datos puede identificar comportamientos, patrones y tendencias en las compras de los clientes, y así poder mejorar altamente la calidad del servicio al cliente, asegurar la retención y satisfacción de los clientes, mejorar la distribución, transporte y políticas de la mercancía, y en general, reducir el costo del negocio e incrementar las ganancias.

2.3.4. Cuidado de la salud. Con la gran cantidad de información y situaciones en la industria del cuidado de la salud y la industria farmacéutica, existen muchas oportunidades para la Minería de Datos y los beneficios que se obtienen de los resultados son enormes. El almacenamiento de los registros de los pacientes en formato electrónico y el desarrollo de sistemas de información médicos, han permitido que gran cantidad de información clínica esté disponible para realizar análisis. Regularidades, tendencias, eventos extraídos de esos datos por los

métodos de Minería de Datos, pueden ser utilizados para mejorar los servicios en temas de salud, o incluso, pueden ser de gran importancia para los clínicos con el fin de que estén informados y puedan tomar decisiones correctas.

Los clínicos evalúan la condición de los pacientes todo el tiempo. El análisis de grandes cantidades de este tipo de datos puede proveer a los médicos información importante acerca del proceso de las enfermedades. Minería de Datos ha sido usada exitosamente en muchas aplicaciones médicas, como monitoreo en crecimiento de los niños, validaciones en cuidados intensivos, análisis en pacientes diabéticos, monitoreo de anestesia inteligente, etc. Visualización de datos y redes neuronales son áreas importantes de la Minería de Datos aplicadas al campo de la medicina.

2.3.5. Ciencia e Ingeniería. Enormes cantidades de datos han sido generadas en ciencia e ingeniería, por ejemplo en cosmología, biología molecular, ingeniería química. En cosmología se necesitan herramientas de computación avanzadas para ayudar a los astrónomos a entender el origen de estructuras cosmológicas de gran escala, como también la formación y evolución de sus componentes astrofísicos. Hoy en día se han recolectado terabytes de datos de imágenes que contienen alrededor de 2 billones de objetos celestiales. Esto representa un reto enorme para los astrónomos al tratar de catalogar completamente este conjunto de datos.

En biología molecular, Minería de Datos es aplicada en áreas como genética molecular, secuencias de proteínas, determinación estructuras macro moleculares, etc. En ingeniería química, se han utilizado modelos avanzados para describir la interacción entre varios procesos químicos, e incluso se han desarrollado nuevas herramientas para obtener una visualización de esas estructuras y procesos.

2.3.6. Adquisición de clientes. Obtener nuevos clientes es el objetivo más importante para el crecimiento de muchas compañías. Sin embargo, no todos los clientes ofrecen las mismas ganancias. Mercadeo probablemente quiera seleccionar un sub conjunto de clientes utilizando algún tipo de criterio, como ingresos, edad, u otros criterios que permitan identificar si un cliente no será leal a la compañía, es decir, el cliente no estará mucho tiempo con la compañía o comprará productos exclusivos de la compañía. El realizar grandes ofertas que atraigan ese tipo de clientes resulta en costos muy altos pues no traería benecitos a largo plazo.

Otros clientes pueden ser leales, pero al mismo tiempo ser muy poco frecuentes con sus compras o comprar productos de bajo precio. Enfocarse en estos tipos de clientes con ofertas generosas puede resultar improductivo. La meta de adquirir clientes es lograr enfocarse en aquellos clientes que tienen la mayor probabilidad de respuesta y lealtad.

Minería de Datos es un competente clave a la hora de buscar estrategias para adquirir nuevos clientes ofreciendo gran variedad de técnicas. Algoritmos de clasificación y clúster pueden ser usados para identificar varios segmentos que existen en los clientes actuales utilizando los datos históricos de los productos que los clientes han comprado y los atributos que poseen dichos clientes. Con esta información, se puede analizar cuáles clientes con los que ofrecen mejores ganancias a la compañía. Luego, se pueden utilizar los algoritmos para analizar cada uno de los segmentos de los clientes, se puede determinar cuáles son los aspectos que existen en cada segmento y en cada cliente individual que los llevan a comprar ciertos productos. Una vez se conoce cuáles son las características de los clientes que más compran o cuáles son leales, con esta información se pueden realizar campañas dirigidas a estos tipos de clientes.

2.3.7. Retención de clientes. Un gran problema en los negocios de la industria hoy en día, es lograr retener los clientes. La pregunta que todos quisieran responder es ¿por qué mis clientes me dejan? Los clientes dejan las compañías por varias razones, por ejemplo, insatisfacción con el servicio, cambio de localidad, encuentran mejores ofertas. Sin embargo, estas razones no siempre son obvias ante los hechos.

Un esfuerzo efectivo en la retención de clientes requiere lograr identificar los clientes que se van a ir incluso antes de que se vayan y así tomar acciones al respecto con el fin de retener a los clientes. Minería de Datos puede ser aplicada para identificar características de los clientes y comportamiento pasado y presente con el fin de determinar factores que indiquen por qué la pérdida de los clientes, por ejemplo según la edad, grupo étnico, región geográfica, tipos de productos que consumen, etc. Al identificar estos factores mediante Minería de Datos, es posible enfocar mejor campañas y estrategias con el fin de realizar acciones que minimicen el riesgo de perder a los clientes.

2.3.8. Detección de fraude. Donde quiera que exista dinero, existe potencial de fraude. Todas las industrias son vulnerables a los individuos que quieren obtener ganancias personales de forma ilegal. Minería de Datos puede ayudar a la detección de fraude mediante modelos de clúster. El objetivo en primer lugar es lograr agrupar los datos del clúster. Luego, se puede revisar cada uno de los clúster para verificar si hay concentración de fraude conocido en alguno de los clúster, indicando que el fraude tiende más a estar en un clúster dado que en otros. Luego se pueden buscar casos que no estén en ningún clúster y estos son los que se convierten en principales candidatos de investigación. Un segundo método que se puede utilizar es la clasificación. Primero se identifican los fraudes manualmente en los datos históricos, luego, la meta es aprender a distinguir entre el comportamiento fraudulento y no fraudulento. Un algoritmo de clasificación como árboles de decisiones, aplicado a este conjunto de datos con campos que

indiquen las características de los clientes como sexo, edad, ingresos, etc, puede lograr predecir fraude en nuevos datos. Los nuevos casos resultantes con una alta probabilidad predicha por el algoritmo pueden ser candidatos potenciales a ser investigados por fraude.

2.3.9. Calificación de créditos. Cada vez que alguien aplica para un préstamo, hipoteca, o una tarjeta de crédito, su historia crediticia es verificada y su situación financiera actual es evaluada para determinar una calificación de crédito. Esta calificación indica el tipo de riesgo que representa a la entidad financiera encargada de manejar el crédito. La calificación de crédito tiene en cuenta varios datos como demografía del cliente, historia crediticia, activos, deudas pendientes, etc. Mientras más exacta sea la calificación, la entidad financiera podrá tomar la mejor decisión frente a un préstamo.

Históricamente, métodos estadísticos eran empleados para calificar créditos, sin embargo, hoy en día Minería de Datos juega un papel muy importante a la hora de valorar un crédito gracias a la gran cantidad de atributos predictores que existen en los clientes.

2.3.10. Sector público. Hay muchas posibilidades en el sector público para aplicar Minería de Datos, desde análisis de crímenes hasta sistemas de loterías. En análisis de crímenes, la ley se ha esforzado en adquirir y manejar colecciones de datos de una forma sofisticada, aprovechando estos datos para realizar análisis tácticos del crimen, comportamientos de violencia, análisis de grabaciones telefónicas y monitoreo de Internet, permitiendo así crear mejores estrategias para enfrentar el delito. Minería de Datos puede ser utilizada para identificar terroristas y analizar volúmenes grandes de documentos de texto, incluyendo WEB y Email, buscando posibles brechas en la seguridad nacional. En los sistemas de lotería, Minería de Datos es empleada para incrementar las ganancias prediciendo qué juegos prefieren los clientes.

2.4. Oportunidades y retos de la Minería de Datos

Hoy en día, los sistemas de bases de datos de las organizaciones se han vuelto en un elemento crítico del negocio, los costos de almacenamiento se han reducido drásticamente y las presiones competitivas aumentan cada momento. La Minería de Datos puede extraer gran valor de estos inmensos repositorios de información para permitir a las compañías tomar mejores decisiones y realizar mejores estrategias que permitan obtener mayores ganancias a partir de sus datos. Sin embargo, constantemente surgen nuevos retos y aspectos claves que deben ser tenidos en cuenta si se desea realizar una Minería de Datos exitosa.

[WANG 2003] dice que aunque es fácil reconocer la importancia de la Minería de Datos, también es importante ver la otra cara de la moneda y examinar las dificultades que se pueden presentar a la hora de realizar Minería de Datos. Constantemente surgen nuevas oportunidades y retos para nuevos desarrollos y mejoras que los distribuidores de aplicaciones de Minería de Datos pueden realizar para resolver estos problemas y hacer de la Minería de Datos un proceso más eficiente y seguro para las organizaciones.

Es importante tener en cuenta que cuando no se usa apropiadamente, la Minería de Datos puede generar grandes cantidades de basura. Si no se tiene suficiente experiencia se puede caer fácilmente en algunas trampas de la Minería de Datos. El proceso de Minería de Datos puede llevar mucho tiempo de planeación y preparación. Simplemente el tener una gran cantidad de datos no es garantía para que un proyecto de Minería de Datos sea exitoso. Minería de Datos puede producir correlaciones falsas y generar interpretaciones erradas si no se realiza correctamente. Si no se tiene los conocimientos suficientes, no se eligen correctamente las bases de datos y se utilizan demasiadas variables durante el proceso, fácilmente el proyecto puede fracasar y no se logrará sacar ningún beneficio de la Minería de Datos.

El éxito en Minería de Datos depende de lograr orquestar tanto los beneficios como los inconvenientes que pueden surgir. Finalmente, los beneficios se obtienen de acuerdo a las interpretaciones que se le den a los resultados arrojados. Por esta razón, el sólo hecho de contar con datos no es el tema central de Minería de Datos, hay que recordar siempre que ésta puede ser malinterpretada y así llevar a errores costosos.

A continuación, se nombrarán algunos aspectos que retan a las organizaciones frente a los procesos de Minería de Datos desde el punto de vista tecnológico y desde el surgimiento de nuevas necesidades de crear nuevas técnicas de Minería de Datos. Todo esto genera nuevas oportunidades para que la industria cree nuevas estrategias y desarrollos frente al tema de Minería de Datos. [Ídem] describe en detalle estas oportunidades y a continuación se presenta un breve resumen de algunas de ellas:

2.4.1. Aspecto tecnológico

- **Soporte del personal de TI:** Minería de Datos requiere que los datos sean accesibles las 24 horas del día, los 7 días de la semana y que además estén adecuadamente protegidos todo el tiempo. Si no hay datos disponibles, simplemente no se puede hacer Minería de Datos. Para esto se requiere que el personal de TI brinde el soporte adecuado y que tanto el Hardware como el Software estén preparados para las grandes cantidades de datos que se procesarán y que serán extraídas de los servidores.

- **Requerimientos de infraestructura de TI:** En general, las grandes empresas no tienen muchos problemas con este tema, pues ya cuentan con la infraestructura de TI Hardware/Software que se requiere para poder ejecutar los algoritmos de Minería de Datos. Sin embargo, hay que tener en cuenta que mientras más datos vayan a ser procesados, más poderosa

debe ser la infraestructura de TI. Esto puede ser un problema para empresas medianas o pequeñas que no cuenten con los recursos suficientes para soportar la capacidad de procesamiento que se requiere para poder extraer conocimiento de los datos.

- **Accesibilidad y usabilidad:** Muchas organizaciones han experimentado serios problemas al intentar implementar un proyecto de Minería de Datos estándar. Muchos de estos problemas no recaen en la tecnología sino más bien en el personal que la utiliza. Para generar un impacto exitoso en la organización, el sistema de Minería de Datos debe ser rápido, accesible y amigable. Hoy en día esto se logra con las avanzadas herramientas que existen para realizar minería y procesar los datos, sin embargo es importante que los directivos puedan interpretar los resultados finales. Los resultados deben fluir rápidamente, si los usuarios finales tienen problemas para acceder a la información en corto tiempo y así poder resolver las necesidades y tomar decisiones, simplemente no se alcanza ningún tipo de beneficio.

- **Asequibilidad y eficiencia:** Hay que reconocer que los procesos de Minería de Datos son costosos y requieren gran inversión tanto en Hardware como Software. Implementar un sistema efectivo de Minería de Datos puede ser complicado y requerir esfuerzos muy costosos para la organización. Por esta razón, aquellas organizaciones que están buscando recortar costos y contener las pérdidas, no quieran invertir recursos de este tipo de proyectos. Además, levantar un proyecto de Minería de Datos exitoso puede tomar mucho tiempo y de seguro no arrojará impactos inmediatos sobre los ingresos de la organización. Y no sólo eso, sino que lograr calcular el ROI resultante de este tipo de proyectos resulta muy difícil.

- **Escalabilidad y Adaptabilidad:** Escalabilidad se refiere a qué tan bien los sistemas de computadores Hardware o Software pueden adaptarse a las demandas crecientes. La Minería de Datos tiende a trabajar con grandes cantidades de datos que continuamente crecen y requieren mayor capacidad de procesamiento, por lo cual la escalabilidad de los sistemas computacionales se convierte en un serio problema para la Minería de Datos. Incluso, en la más simple forma de análisis de datos, la velocidad y memoria se convierten en un punto muy importante.

- **Visión empresarial:** Otro factor que puede ser un problema para la Minería de Datos, es que los proyectos no sean vistos como algo que requiere el esfuerzo de toda la organización. En muchos proyectos fracasados de Minería de Datos, el factor común fue que las compañías los veían como un proyecto sólo del área de TI. Si el proyecto no tiene una perspectiva organizacional fácilmente va a fracasar.

- **Problemas con redes neuronales:** El principal problema con redes neuronales es que el tiempo de aprendizaje puede llevar mucho tiempo y recursos para completarse, mientras que los resultados de la Minería de Datos son críticos para los usuarios con el fin de poder tomar decisiones. Además, las redes neuronales han probado ser muy buenas sobre conjuntos de datos pequeños, pero cuando éstos crecen tienden a volverse muy ineficientes.

- **Problemas con árboles de decisión:** Los problemas que existen con los árboles de decisión pueden dividirse en dos categorías: problemas algorítmicos que dificultan el lograr árboles pequeños, y problemas de representación. Los árboles son buenos para problemas pequeños pero se vuelven difíciles de manejar cuando estos problemas comienzan a crecer.

Además, se requiere software especial y algoritmos complejos para poder dibujar los árboles.

- **Problemas con visualización de datos:** Este método permite a los usuarios obtener un entendimiento más intuitivo de los datos. Generalmente, se utilizan herramientas de gráficos para ilustrar de mejor manera las relaciones entre los datos. La visualización de datos permite de mejor manera y de forma más fácil visualizar los patrones y tendencias en los datos. Sin embargo, la mayor dificultad es que los volúmenes de datos crecen constantemente, haciendo así más difícil el poder discernir patrones exactos de los conjuntos de datos y genera la necesidad de mayores requerimientos técnicos para poder generar modelos visuales de los datos. Aún hay que trabajar en mejores herramientas que permitan manipular y visualizar de una manera más eficiente los datos.

2.4.2. Oportunidades y nuevas herramientas

- **Minería de datos de Texto**: La idea con Minería de Datos de texto procesar la estructura del texto para analizar el contenido de este y mediante la aplicación de algunos métodos lingüísticos descubrir patrones y tendencias sobre los textos. Hoy en día están surgiendo muchas aplicaciones para este tipo de Minería de Datos, especialmente en áreas de seguridad.

- **Minería de datos ubicua**: Consiste la utilización de algoritmos de Minería de Datos avanzados para aplicaciones móviles y distribuidas. Este tipo de Minería de Datos tiene grandes retos técnicos como arquitectura, control, seguridad y comunicación.

- **Minería de datos de hyper texto e hyper medios**: Este tipo de Minería de Datos, puede ser categorizada por minar datos que incluye texto, vínculos, marcas de texto, XML, y otros tipos de información hyper medio. Por esta razón está muy vinculado con la Minería Web y la Minería multimedia.

- **Minería de datos multimedia:** Es el análisis de varios tipos de datos, incluyendo imágenes, video, audio y animación. Por ejemplo, en Minería de Datos de audio, la idea es básicamente utilizar señales de audio para indicar patrones o para representar características de los resultados de minería basados en el tono, tiempo, instrumentos musicales, etc. Con videos, se puede realizar Minería de Datos para identificar patrones de comportamientos que ayuden a tomar decisiones frente a lo que está ocurriendo en una cámara de video en vivo y en directo.

- **Minería de Datos espacial y geográfica:** Además de los datos estadísticos o numéricos, también es importante considerar otro tipo de información completamente diferente, como la información espacial y geográfica que puede contener información acerca de datos astronómicos, recursos naturales, etc. La gran mayoría de estos datos están orientados hacia las imágenes y pueden representar información muy valiosa si son propiamente analizados y minados. Analizar este tipo de datos puede ser útil en áreas como teledetección, imágenes médicas, navegación y temas relacionados a estos.

3. CONCEPTOS Y APLICACIONES DE LA COMPUTACIÓN GRID

3.1. Conceptos generales de la Computación Grid

La Computación Grid, es una tecnología que permite explotar de una forma muy efectiva los variables recursos computacionales con los cuales pueda contar una organización. [ABBAS 2004] habla sobre cómo los sistemas de computación, almacenamiento y redes han crecido exponencialmente en capacidad en los últimos años, mientras que sus costos se han reducido igualmente. La Computación Grid ofrece una gran variedad de tecnologías que aprovechan estos avances económicos relacionados con los computadores y redes, y provee herramientas a las organizaciones las cuales pueden usar para reducir drásticamente los costos en tecnología e incrementar la productividad de sus empleados y activos tecnológicos, generando así grandes beneficios para la organización.

3.1.1. Significado de la Computación Grid. [ALKADI 2007], describe la Computación Grid como un término que puede ser aplicado similarmente a un gran número de computadores los cuales están conectados entre sí para resolver un problema de forma colectiva o problemas de gran complejidad y magnitud. La idea fundamental de construir una grid de computadores es el utilizar el tiempo ocioso de ciclos de procesamiento. El rol que cada procesador juega está cuidadosamente definido y hay casi una completa transparencia en la forma de trabajar de cada procesador/computador en la grid. Esto es llamado la “división de trabajo”, en términos humanos es equivalente a un estudiante y su grupo de amigos que colectivamente resuelven una tarea que contiene más de un problema simple. La solución es trivial pero el esfuerzo es colectivo. [JACOB 2005], complementa esta descripción y habla de la Computación Grid como el medio por

el cual usuarios o aplicaciones obtienen acceso a recursos computacionales (procesadores, almacenamiento, datos, aplicaciones, etc) según sus necesidades, con poco o ningún conocimiento acerca de dónde están ubicados esos recursos o cuál es su tecnología subyacente, hardware, sistema operativo, etc. Según [ORACLE 2009], estos recursos de TI son virtualizados y utilizados como un único conjunto de servicios compartidos que pueden ser provisionados y distribuidos, y que pueden redistribuirse de acuerdo a las necesidades.

[ABBAS 2004] se expresa acerca de la Computación Grid como una tecnología que saca ventaja de todos los avances que se han logrado en los últimos años relacionados con la velocidad de los microprocesadores, comunicaciones, la capacidad de almacenamiento, la WEB e Internet. Se han desarrollado una serie de estándares y protocolos que permiten desagregar las plataformas de cómputo y distribuirlas a través de redes como recursos que pueden ser utilizados por cualquier usuario (persona o máquina) en cualquier momento y lugar. El autor menciona cómo una compañía con una Grid de 1000 computadores de escritorio, puede utilizarlos todos en conjunto como una sola plataforma de cómputo y así proveerse de capacidad computacional comparable a la mayor supercomputadora en el mundo. Esta tremenda capacidad puede ser usada por las compañías en cientos de formas y a mucho menor costo comparado con lo que puede valer una supercomputadora.

[ORACLE 2009], explica que la Computación grid opera en los siguientes principios tecnológicos:

- **Estandarización:** Los departamento de TI han logrado mayor interoperabilidad y reducido la sobrecarga en la administración de los sistemas estandarizando en sistemas operativos, servidores, almacenamiento, hardware, componentes middleware y componentes de redes. Estandarizar, también ayuda a reducir la complejidad operacional en

los centros de datos simplificando el desarrollo de aplicaciones, configuración e integración.

- **Virtualización:** Virtualizar los recursos de TI, significa que las aplicaciones no están atadas un servidor en específico, almacenamiento, o componentes de servidores y pueden utilizar cualquier recurso TI virtualizado. La virtualización se realiza mediante una capa de software sofisticada que oculta la complejidad de los recursos de TI y presenta una interfaz simplificada y coherente utilizada por las aplicaciones y otros recursos de TI.

- **Automatización:** Debido al gran número potencial de componentes, tanto virtuales como físicos, la Computación Grid demanda la automatización a gran escala de las operaciones de TI. Cada componente requiere administración de la configuración, aprovisionamiento en demanda, monitoreo, y otras tareas administrativas. Una solución para administrar la grid debe asegurar que los ahorros de costos en infraestructura, no se evaporen como resultado de contratar personal adicional para administrar la grid.

3.1.2. ¿Qué hace posible la Computación Grid? [ABBAS 2004] menciona algunos aspectos que han permitido la evolución de la Computación Grid y que han motivado su implementación en las organizaciones. A continuación se presenta un resumen de dichos aspectos:

- **Evolución de la infraestructura de TI**: Crecimiento en velocidad de microprocesadores, almacenamiento y capacidades ópticas. Anteriormente se consideraban como desarrollos por separado, hoy en día se han juntado para generar poderosas tecnologías como Computación Grid.

- **Tecnología microprocesador**: Aumento de la capacidad de procesamiento en los computadores personales de escritorio.

- **Tecnología de redes óptica**: Mayor tráfico de red en una sola fibra.

- **Tecnología de almacenamiento**: Aumentos en capacidad de almacenamiento y velocidad de extracción.

- **Tecnología inalámbrica**: La tecnología inalámbrica hoy en día es barata y muy accesible. Se pueden encontrar miles de lugares con Wi-Fi en todo el planeta en universidades, aeropuertos, cafés, y otros sitios públicos.

- **Tecnología de sensores**: Un avance clave en sensores, es la habilidad de no únicamente recolectar información a su alrededor, sino pasar la información a través de sistemas para ser analizada y procesada por varios canales de comunicación integrados. La meta, es lograr tener todo conectado en una cadena de suministros dinámica y automática que une los negocios y consumidores en una relación de beneficios.

- **Infraestructura global de Internet:** Gracias a Internet, hoy en día se cuenta con una gran infraestructura interconectada al rededor de todo el mundo lista para ser utilizada por la Computación Grid. Además, un gran beneficio es la gran cantidad de investigación y desarrollo que está al alcance de todos a través de este medio. Actualmente Internet es el principal conducto para el comercio y los negocios.

- **World Wide Web y Web Services**: Los Web Services, proveen un framework para interactuar entre aplicaciones basado en protocolos WEB. En vez de los modelos tradicionales cliente / servidor, este tipo de servicios no proveen a los usuarios con una interfaz gráfica, sino que permiten

compartir fácilmente la lógica de negocio, datos y procesos a través de Internet, no están atados a ningún sistema operativo ni lenguaje de programación. Los Web Services brindan un sistema ideal para la Computación en Grid.

- **Movimiento de código libre**: Internet y la WWW, han impulsado un movimiento poderoso para el desarrollo de software de código libre, el cual ha permitido que tecnologías como Computación en Grid hayan avanzado enormemente.

3.1.3. Beneficios de los ambientes Grid. [MAGOULES 2009], [ORACLE 2009] y [JACOB 2005], mencionan algunos beneficios que provee la tecnología Grid frente al tema de procesamiento y aprovechamiento de los recursos con los que cuenta la organización. A continuación, se nombran algunos de los más relevantes y se da un resumen de su descripción según los autores:

- **Aprovechar las inversiones realizadas en recursos de Hardware:** En la empresa típica existe una alta cantidad de capacidad sin usar en la infraestructura de TI. Las grids pueden ser utilizadas en la infraestructura empresarial existente (incluyendo la multitud de computadores de escritorio y servidores existentes), y así mitigar la necesidad de invertir en nuevo Hardware. Los costos no sólo disminuirían en la adquisición de Hardware y software, sino también al eliminar gastos en aire acondicionado, electricidad, etc.

- **Reducir gastos operacionales:** La Computación Grid permite un nivel de automatización y facilidades previamente no identificadas en los ambientes de TI. Las capacidades que brinda la Computación Grid liberan a los administradores de tareas rutinarias y les permite enfocarse en generar

mayor valor a la organización. Las capacidades de las grids de cruzar los límites departamentales y geográficos, incrementan la capacidad computacional para toda la empresa y mejora los niveles de redundancia en la infraestructura.

- **Crear una infraestructura de TI empresarial escalable y flexible:** La Computación Grid permite a las compañías añadir recursos linealmente basados en los requerimientos del negocio en tiempo real. Los proyectos ya no tienen que ser suspendidos por la falta de capacidad computacional, espacio en el centro de datos, o prioridades del sistema. La infraestructura completa de la empresa está disponible para ser aprovechada.

- **Acelerar el desarrollo de productos, mejorar el tiempo en ventas e incrementar la satisfacción del cliente:** La Computación Grid permite acelerar el desarrollo de productos y ayudar a llevar los productos al mercado más rápidamente. La reducción considerable, por ejemplo, en tiempos de simulación puede hacer que los productos estén listos en menor tiempo. También provee la capacidad para realizar diseño de productos más exhaustivos y detallados, con los recursos computacionales a través de la grid se puede trabajar rápidamente con modelos y escenarios complejos para detectar fallas de diseño.

- **Incrementar productividad**: A través de la Computación Grid se logra la reducción en tiempos de procesamiento gracias al incremento computacional ofrecido y así liberar tiempo a sus empleados para mejores usos.

- **Tolerancia a fallos y fiabilidad**: Supongamos que un usuario envía un proceso para ser ejecutado en un nodo particular en la grid. El proceso asigna los recursos apropiados, basado en la disponibilidad y la política de

planificación en la grid. Ahora, supongamos que ese nodo donde se está ejecutando el proceso por alguna razón falla, la grid realiza una provisión automática para reenviar el proceso a otro recurso disponible cuando es detectada alguna falla.

- **Equilibrio y uso compartido de recursos variados:** El balancear y compartir recursos son un aspecto importante en las grid, las cuales proveen las características adecuadas de administración de recursos. Este aspecto, habilita a la grid para distribuir uniformemente las tareas a los recursos disponibles. Suponga que el sistema en la grid está sobrecargado, el algoritmo de planificación de la grid puede re planificar algunas tareas de otro sistema que están retrasadas o menos cargadas haciendo uso de los recursos subutilizados.

- **Procesamiento en paralelo**: Muchas tareas pueden ser divididas en múltiples tareas y cada una de estas puede ser ejecutada en diferentes máquinas. Por ejemplo, el modelado matemático, animación en 3D, carga de imágenes, etc. Este tipo de aplicaciones pueden ser desarrolladas para ser ejecutadas en sub tareas, y el resultado de cada una de esas sub tareas puede ser recombinado para producir el resultado deseado. La computación en grid es el ambiente ideal para este tipo de tareas brindando su capacidad computacional para obtener resultados de forma más rápida y eficiente.

- **Calidad del servicio:** La Computación Grid al proveer tolerancia a fallos, fiabilidad y capacidad de procesamiento paralelo en algunas tareas, hace un uso más eficiente de los recursos de procesamiento brindando a los usuarios un mejor servicio a la hora de ejecutar sus procesos y entregar sus resultados.

- **Responder en tiempo real a cargas de trabajo dinámicas:** La mayoría de las aplicaciones actuales están vinculados a programas informáticos específicos y silos de hardware, lo que limita su capacidad para adaptarse a las nuevas cargas de trabajo. Este uso costoso e ineficiente de los recursos de TI significa que los departamentos de TI deben sobre provisionar su hardware para que cada aplicación pueda manejar picos de altas cargas de trabajo. La Computación Grid permite asignar y desasignar dinámicamente los recursos de TI, según sea necesario, proporcionando una capacidad de respuesta mucho mejor a altas cargas de trabajo a una escala global.

- **Recursos y organizaciones virtuales para colaboración:** La Computación Grid proporciona un entorno para la colaboración entre un público más amplio mediante estándares que permiten a sistemas muy heterogéneos trabajar juntos para formar la imagen de un sistema informático virtual que ofrece una gran variedad de recursos. Los usuarios de la red se pueden organizar de forma dinámica en una serie de organizaciones virtuales, cada uno con diferentes exigencias políticas. Estas organizaciones virtuales pueden compartir sus recursos en conjunto como una grid más grande.

- **Acceso a recursos adicionales:** además de la los recursos de procesamiento y almacenamiento, una grid puede proporcionar también acceso a otros recursos. Por ejemplo, si un usuario necesita aumentar su ancho de banda total a Internet para implementar un motor de búsqueda de Minería de Datos, el trabajo se puede dividir en la grid entre las máquinas que tienen conexión independiente a Internet. De esta manera, la capacidad total de búsqueda se multiplica, ya que cada máquina tiene una conexión independiente a Internet. Algunas máquinas pueden tener instalado software licenciado muy costoso que los usuarios requieren. Los procesos de los usuarios se pueden enviar a estas máquinas y explotar así

las licencias de software. Algunas máquinas de la red pueden tener dispositivos especiales. La mayoría hemos utilizado impresoras remotas, tal vez con capacidad de color avanzada o velocidades más rápidas. Del mismo modo, una grid se puede utilizar para hacer uso de otros equipos especiales.

3.2. Estándares de la Computación Grid.

[MAGOULES 2009] y [JACOB 2005], resumen algunos de los estándares abiertos utilizados para implementar computación en grid:

3.2.1. Web Services. Los servicios grid, definidos por la OGSA, son una extensión de los servicios WEB. Las 4 especificaciones de Web Service son:

1. Extensible Markup Language (XML): XML es un lenguaje de marcas, cuyo propósito es facilitar compartir datos a través de diferentes interfaces utilizando un formato común. Todos los mensajes intercambiados a través de los Web Services son en formato XML.

2. Simple Object Access Protocol (SOAP): SOAP es un protocolo basado en mensajes, el cual puede ser utilizado para comunicaciones en Internet. Los mensajes SOAP son basados en XML y por esto son independientes de la plataforma. Los mensajes SOAP son transmitidos a través de HTTP. Por lo tanto, contrario a tecnologías como RPC o CORBA, los mensajes SOAP pueden atravesar un corta fuegos. Los mensajes SOAP son adecuados cuando se envían mensajes pequeños. Cuando el tamaño de los mensajes se incrementa, se aumenta la sobrecarga en el transporte del mensaje y por lo tanto la eficiencia de la comunicación se disminuye.

3. Web Service Definition Language (WSDL): WSDL es un documento XML utilizado para describir la interfaz del Web Service, utilizando elementos como: tipo de puerto, mensaje, tipos, protocolo de comunicación y servicio.

4. Universal Description, Discovery and Integration (UDDI): UDDI es un registro basado en XML utilizado para encontrar un Web Service en Internet. Es una especificación que permita a un negocio publicar información acerca de si mismo y su Web Service permitiendo así que otro Web Service pueda localizar esta información.

3.2.2. Open Grid Services Architecture (OGSA). Define un marco de trabajo basado en Web Services para implementar una grid. Busca estandarizar el servicio que provee una grid como descubrimiento de recursos, administración de recursos, seguridad, etc, a través de una interfaz estándar de Web Service. También define aquellas características que no son necesarias para la implementación de una grid, pero son deseables. OGSA está basado en las especificaciones existentes de Web Services, y añada características para volver los Web Services adecuados para el ambiente grid.

3.2.3. Open Grid Services Infrastructure (OGSI). OGSA describe las características que son necesarias para la implementación de servicios ofrecidos por la grid, como un Web Service. Sin embargo, no provee los detalles de implementación. OOGSI, prevé una especificación técnica formal necesaria para la implementación de servicios grid. También provee los mecanismos para la creación, administración e interacción a través de servicios grid.

3.2.4. Web Services Resource Framework. Define las convenciones para el manejo de estados permitiendo a las aplicaciones descubrir e interactuar con Web Servicies con estados de una manera estándar. Los Web Services estándares no tienen la noción de estado. Las aplicaciones basadas en grid, necesitan la noción de estado porque generalmente realizan una serie de peticiones donde la salida de una operación puede depender del resultado de operaciones previas. WSRF, puede ser usado para desarrollar un servicio grid con estados.

3.2.5. Open Grid Services Architecture-Data Access and Integration. OGSA-DAI es un middleware que permite el acceso e integración de fuentes de datos distribuidas utilizando una grid. Provee soporte para varias fuentes de datos como bases de datos relacionales y XML. Estas fuentes de datos pueden ser consultadas, actualizadas y transformadas vía un Web Service OGSA-DAI. Estos Web Services pueden ser desplegados por una grid, volviendo así las fuentes de datos habilitadas para la grid.

3.3. Técnicas y Herramientas

3.3.1. Componentes software de la Grid. Según [JACOB 2005], hay muchos aspectos de la Computación Grid que normalmente son controlados a través de software. Estas funciones pueden ser manejadas a través de procesos muy manuales hasta procesos que son manejados autonómicamente mediante software sofisticado. Los componentes principales de la grid de acuerdo a [Ídem] son:

- **Componentes administrativos:** Cualquier sistema Grid debe tener componentes administrativos. Primero, debe existir un componente que realice un seguimiento a los recursos disponibles en la grid y qué usuarios son miembros de la grid. Esta información es utilizada principalmente para

decidir dónde deben ser asignados los procesos. Segundo, deben haber componentes de medición que determinen tanto la capacidad de los nodos en la grid y su tasa de utilización actual en cualquier momento dado. Esta información se utiliza para programar procesos en la grid. Dicha información también se utiliza para determinar la salud de la grid, alertando al personal de problemas como cortes o congestión. Esta información también se utiliza para determinar los patrones de uso general y estadísticas, así como registro y conteos del uso de los recursos grid. En tercer lugar, un software avanzado de administración de la grid debe, de forma automática, manejar muchos aspectos de la grid. Esto se conoce como computación orientada a la recuperación. Este software puede automáticamente recuperarse de diversos tipos de fallos e interrupciones en la grid, buscando formas alternativas de realizar los procesos.

- **Gestión distribuida de la Grid:** Las grids grandes pueden tener una organización jerárquica u otro tipo de topología usualmente de acuerdo de la topología de conectividad. Esto es, las máquinas que están localmente conectadas a una LAN forman un clúster. La grid puede estar organizada de forma jerárquica compuesta por clústeres de clústeres. La labor de gestión de la grid es distribuida con el fin de aumentar la escalabilidad de la red. La recolección y operación de la grid, y los recursos de datos, así como la planificación de procesos se distribuye para que coincida con la topología de la grid. Por ejemplo, un planificador de procesos no programará un proceso presentada directamente a la máquina que la ejecutará. En cambio, el proceso es enviado a un planificador de nivel inferior que se ocupa de un conjunto de máquinas. El planificador de nivel inferior se encarga de la asignación a la máquina específica. Del mismo modo, la recopilación de información estadística se distribuye. Los clústeres de bajo nivel reciben la información de la actividad de las maquinas individuales, la

agregan, y la envían directamente a los nodos administrativos de alto nivel en la jerarquía.

- **Software de donante.** Cada máquina que aporta recursos, normalmente necesita inscribirse como miembro de la grid e instalar algún software que administra el uso de sus recursos en la grid. Por lo general, algún tipo de identificación y procedimiento de autenticación se debe realizar antes que una máquina pueda unirse a la grid. A menudo, los certificados, como los disponibles a través de las autoridades de certificación, se puede utilizar para establecer y garantizar la identidad de la máquina de los donantes como de los usuarios y la propia grid. El sistema grid recolecta información acerca de los recursos nuevos adheridos a la a través de la grid. Las máquinas donantes, suelen tener algún tipo de monitor que determina o mide la disponibilidad de la máquina y el tipo o cantidad de recursos utilizados.

- **Presentación del software:** Por lo general, cualquier máquina miembro de la grid se puede utilizar para enviar los procesos a la grid e iniciar las consultas en la grid. Sin embargo, en algunos sistemas grid, esta función se implementa como un componente independiente instalado en los nodos o clientes. Cuando una grid se construye usando recursos dedicados, se suele instalar un software de presentación diferente en el escritorio del usuario o estación de trabajo.

- **Planificadores:** La mayoría de sistemas grid incluyen algún tipo de software de planificación de procesos. Este software localiza una máquina en la cual ejecutar el proceso grid que ha sido enviado por un usuario. Algunos planificadores implementan un sistema de prioridad de los procesos. A veces, esto se hace usando varias colas de trabajo, cada una con una prioridad distinta. Cuando hay máquinas disponibles en la grid para

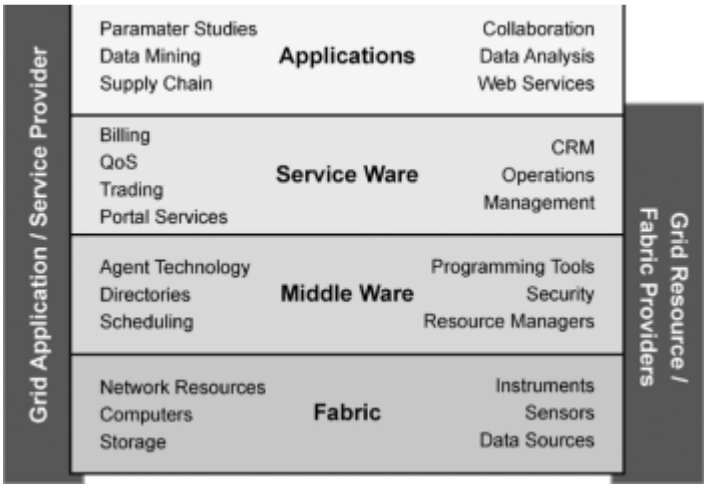
ejecutar trabajos, los procesos son enviados a las colas de más alta prioridad. Políticas de diversa índole también se aplican los planificadores. Las políticas pueden incluir diversos tipos de restricciones en los procesos, usuarios y recursos. Por ejemplo, puede haber una directiva que restringe los procesos de la grid para ser ejecutados en determinados momentos del día.

- **Comunicaciones:** Un sistema grid puede incluir software que permita a los procesos comunicarse entre sí. Por ejemplo, una aplicación puede dividirse a sí misma en un número grande de subprocesos. Cada uno de esos subprocesos es un proceso separado en la grid. Sin embargo, la aplicación puede implementar un algoritmo que requiere que los subprocesos comuniquen alguna información entre ellos. Los subprocesos necesitan poder localizar otros subprocesos específicos, establecer conexiones de comunicación con ellos y enviar los datos apropiados.

3.3.2. Tecnología de la Computación Grid. [ABBAS 2004] La Computación Grid es la fusión de numerosas tecnologías, utilizadas en armonía para proveer un servicio valioso y significativo a los usuarios. A continuación, se presenta una descripción de las diferentes tecnologías, según el autor, categorizadas en grupos basados en su rol y función.

Los productos requeridos para crear una infraestructura de computación en grid, se pueden definir en cuatro categorías y dos categorías adicionales para servicios:

Ilustración 7: Taxonomía computación grid



Extraído de [ABBAS 2004]

- **Fábrica:** Es una capa en la taxonomía que se refiere a los productos hardware utilizados para construir una infraestructura grid, como son computadores personales, clústeres, computadores de alto rendimiento y dispositivos de almacenamiento, como también dispositivos de red como enrutadores, switches, balanceadores de carga, sistemas caché, etc. La capa de fábrica, también consiste en varios tipos de sensores y otros instrumentos científicos.

- **Software intermedio:** Esta capa consiste en proveedores de software y productos que administran y facilitan el acceso a los recursos disponibles en la capa de fábrica. Estos productos realizan la función de administradores de recursos, seguridad, planificación y ejecución de tareas. La capa de Software Intermedio también incluye las utilidades desarrolladas para facilitar el uso de recursos por varias aplicaciones. Una herramienta de paralelización o compilador, puede residir en esta capa.

- **Servicios:** Esta capa consiste en proveedores que ofrecen soporte operacional para sistemas de computación grid, o ayudan a integrarlos con los sistemas existentes, como facturación, administración de cuentas, administración de software, etc.

- **Aplicaciones:** Esta capa consiste en todas las aplicaciones software y proveedores que utilizan la infraestructura de computación en grid. Hay miles de compañías que eventualmente adaptan sus aplicaciones para la grid.

- **Proveedores de recursos grid:** Consiste en las compañías que proveen varios recursos, listados en la capa de fábrica, como servicios que pueden ser arrendados en diferentes términos. Por ejemplo, un cliente renta ciclos de CPU de un proveedor de recursos grid para ejecutar una aplicación especial de computación intensiva, y al mismo tiempo puede querer rentar capacidad adicional de almacenamiento. Los proveedores no se preocupan por las aplicaciones que son ejecutadas por el cliente en sus plataformas. Los proveedores de recursos grid, compran equipos y software de los proveedores de fábrica, Software Intermedio y servicios.

- **Proveedores de aplicaciones grid:** Provee servicios de computación en grid a usuarios de una aplicación en particular o varias aplicaciones. El cliente, en este caso, compra tiempo al proveedor para utilizar la aplicación. Los proveedores de aplicaciones pueden elegir si comprar servicios de los proveedores de recursos grid, o construir su propia infraestructura grid.

- **Consultores grid:** Existen compañías consultoras especializadas en proveer servicios a las compañías proveedoras de computación en grid como también a empresas que quieran habilitar sus aplicaciones e infraestructura para la grid.

3.3.3. Conceptos básicos de la arquitectura Grid. [MAGOULES 2009] describe algunos aspectos que se deben tener en cuenta frente a la arquitectura de la Computación Grid a la hora de diseñar e implementar un sistema Grid, y que deben considerarse igualmente a la hora de buscar implementar Minería de Datos Grid.

- **Seguridad:** Como cualquier sistema en el mundo, la seguridad es un aspecto vital de la computación en grid. Las tres características de seguridad más deseables que la grid debe tener son: inicio de sesión único, autenticación y autorización. Inicio de sesión único significa que el usuario puede iniciar sesión una vez utilizando sus credenciales de seguridad y así puede acceder al servicio de la grid por cierto tiempo. Autenticación se refiere a proveer las pruebas necesarias para establecer la identidad de alguien, como es la contraseña para autenticarse en el servidor al iniciar sesión en la cuenta de correo. Autorización es el proceso que verifica los privilegios asignados a un usuario. La autorización se realiza después de que la identidad de un usuario ha sido establecida a través de la autenticación.

- **Administración de recursos:** La grid debe optimizar los recursos a su disposición para asegurar el máximo rendimiento posible. La administración de recursos incluye la entrega de un trabajo de forma remota, verificar su estado mientras está en proceso de obtener una salida cuando este ha terminado su ejecución. Cuando un proceso es entregado, los recursos disponibles son descubiertos a través de un servicio de directorio. Luego, los recursos son seleccionados para ejecutar un proceso individual. Esta decisión es tomada por otro componente administrador de recursos que pertenece a la red, llamado, el Planificador. La decisión de planificación puede ser basada en varios factores, por ejemplo, si una aplicación consiste en varios procesos que requieren ejecución secuencial porque el

resultado de uno de los procesos es necesario por otro proceso, entonces, el planificador puede planificar esos procesos secuencialmente. La decisión de planificación puede ser también basada en la prioridad que posea el proceso del usuario.

- **Administración de datos:** La administración de datos en grids involucra una gran variedad de aspectos necesarios para administrar grandes cantidades de datos. Esto incluye acceso de datos seguro, replicación y migración de los datos, administración de metadatos, indexación, planificación consciente de datos, uso de caché, etc. Planificación consciente de datos, se refiere a que las decisiones de planificación deben ser tomadas en el lugar donde se encuentran los datos. Por ejemplo, el Planificador puede asignar un proceso a un recurso ubicado cerca de los datos en vez de transferir grandes cantidades de datos a través de la red lo cual puede tener grandes consecuencias en el rendimiento de la grid. Supóngase que el proceso ha sido planificado para ser ejecutado en un sistema que no posee los datos necesarios para el proceso. Estos datos deben ser transferidos al sistema donde el proceso va a ser ejecutado. Por lo tano, un módulo administrador de datos en grid debe proveer una forma segura y confiable de transferir los datos a través de la grid.

- **Descubrimiento y monitoreo de información:** El servicio de descubrimiento de la grid, permite localizar los recursos necesarios para ejecutar un proceso. Este contiene una lista de recursos disponibles para ser utilizados por la grid y el estado en que estos se encuentran. Cuando el Planificador consulta el servicio de información para determinar qué recursos están disponibles, este puede indicar restricciones como encontrar aquellos recursos que son relevantes y son más adecuados para cierto proceso. Si estamos hablando acerca de capacidad computacional necesaria para un proceso, y el proceso requiere CPUs rápidas para su

ejecución, solo son seleccionadas aquellas máquinas que son suficientemente rápidas para terminar la ejecución del proceso a tiempo. El servicio de descubrimiento de información puede funcionar de dos maneras. Este puede publicar el estado de los recursos disponibles a través de una interfaz definida (Web Services), o puede ser consultado para entregar la lista disponible de recursos.

3.3.4. Tipos de recursos en la Grid. [JACOB 2005] define la grid como una colección de máquinas, generalmente también llamadas nodos, recursos, miembros, donantes, clientes, y muchos otros términos. Todos contribuyendo cualquier combinación de recursos en la grid como un todo. Algunos recursos pueden ser utilizados por todos los usuarios de la grid, mientras otros pueden tener restricciones específicas. Según el autor, los recursos que generalmente se comparten en un grid son los siguientes:

- **Computación:** El recurso más común disponible en una grid es los ciclos de procesamiento proporcionados por los procesadores de las máquinas en la grid. Los procesadores pueden variar en velocidad, arquitectura, plataforma de software, y otros factores asociados, como la memoria, almacenamiento y conectividad. Hay tres formas principales para explotar los recursos de computación en una grid. La primera y más simple es usarla para ejecutar una aplicación existente en una máquina disponible en la grid en vez de forma local. La segunda es utilizar una aplicación diseñada para dividir sus procesos de tal manera que las partes separadas puedan ser ejecutadas en paralelo en diferentes procesadores. Y la tercera es ejecutar una aplicación, que debe ser ejecutada muchas veces, en muchas máquinas diferentes en la grid. La escalabilidad es una medida de qué tan eficientemente son usados los múltiples procesadores en una grid. Si el doble de procesadores hacen que una aplicación sea completada en la

mitad del tiempo, entonces se puede decir que es perfectamente escalable. Sin embargo, puede haber límites de escalabilidad cuando las aplicaciones sólo se pueden dividir en un número limitado de partes o si esas partes experimentan algunas otras interdependencias como competencia por recursos de algún tipo.

- **Almacenamiento:** El segundo recurso más común en la grid es el almacenamiento de datos. La grid provee una vista integral del almacenamiento de datos generalmente llamado grid de datos. Cada máquina en la grid usualmente provee alguna cantidad de almacenamiento para ser usada en la grid, aún si es temporalmente. El almacenamiento puede ser memoria principal o memoria secundaria, utilizando discos duros o cualquier otro tipo de medio de almacenamiento permanente. La memoria principal tiene un rápido acceso pero es volátil. Esta puede ser usada para almacenar datos de caché o servir como almacenamiento temporal para ejecutar aplicaciones. La capacidad puede ser incrementada utilizando el almacenamiento de múltiples máquinas con un sistema de archivos unificado, eliminando la restricción de tamaño máximo que generalmente imponen los sistemas de archivos ofrecidos con los sistemas operativos. Un sistema de archivos unificado puede también tener un único nombre para almacenamiento en grid. Esto hace que los usuarios puedan referenciarse más fácilmente a los datos almacenados en la grid, sin la necesidad de conocer su ubicación exacta. De manera similar, un software de base de datos especial, puede utilizar una variedad de bases de datos individuales y archivos parar formar una base de datos más grande y comprensiva, accesible utilizando funciones de consultas en base de datos.

- **Comunicaciones:** El rápido crecimiento de la capacidad de comunicación entre las máquinas hoy en día hace que la computación grid sea práctica, comparada con el límite de ancho de banda disponible cuando esta apenas

estaba emergiendo. Por esta razón, no es sorpresa que otro recurso importante en una grid sea la capacidad de comunicación de datos. Esto incluye las comunicaciones entre la grid y agentes externos a la grid. Las comunicaciones con la grid son importantes para enviar procesos y sus datos requeridos a diferentes puntos en la grid. Algunos procesos requieren que una gran cantidad de datos sea procesada, y quizá no siempre residan en la máquina donde se está ejecutando el proceso. El ancho de banda disponible para tales comunicaciones a menudo puede ser un recurso crítico que puede limitar la utilización de la grid. Rutas redundantes de comunicación son a veces necesarias para manejar mejor los posibles fallos de la grid y el tráfico de datos excesivo.

- **Software y licencias:** La grid puede tener instalado software que puede ser demasiado costoso de instalar en cada una de las máquinas de la grid. Usando una grid, los procesos que requieren este software se envían a las máquinas particulares donde se encuentra instalado. Cuando los costos de licencia son significativos, este enfoque puede ahorrar gastos importantes para una organización. Algunos acuerdos de licencia de software permiten que el software sea instalado en todas las máquinas de una grid, pero puede limitar el número de instalaciones que pueden ser utilizados simultáneamente en un momento dado.

- **Equipo especial, capacidades, arquitecturas y políticas:** Las plataformas en la grid pueden tener diferentes arquitecturas, sistemas operativos, dispositivos, capacidades y equipos. Cada uno de estos elementos representa un tipo diferente de recurso que puede ser utilizado como un criterio de asignación de procesos a las máquinas. Mientras algún software puede estar disponible para muchas arquitecturas, por ejemplo, PowerPC y X86, otro software generalmente está diseñado para ser ejecutado sólo en un tipo particular de hardware y sistema operativo. Estos

atributos deben ser considerados a la hora de asignar los procesos a recursos en la grid. En algunos casos, el administrador de la grid puede crear un nuevo tipo de recurso artificial que es utilizado por los planificadores para asignar el trabajo de acuerdo a políticas u otras restricciones. Por ejemplo, algunas máquinas pueden ser designadas para ser utilizadas únicamente en investigaciones médicas. Estas serían identificadas por un atributo de investigación médica y el planificador puede ser configurado para asignar sólo los procesos que requieren las máquinas de los recursos de investigación médica. Otros pueden participar en la grid sólo si no se utilizan con fines militares. En esta situación, los procesos que requieren un recurso militar no se asignarían a dichas máquinas. Por supuesto, los administradores tendrían que imponer una clasificación en cada tipo de proceso a través de algún procedimiento de certificación para utilizar este tipo de enfoque.

3.3.5. Clasificación de las grids. [STANOEVSKA 2009], presenta dos formas como pueden clasificarse las grids:

3.3.5.1. Clasificación de las grids de acuerdo a los recursos en los que están enfocadas. Aunque la meta final de la Computación Grid es proveer recursos compartidos de cualquier tipo, históricamente los middleware emergieron con enfoques en tipos recursos específicos. Por lo tanto, de acuerdo a los recursos en que están enfocados, los middleware de grid se pueden distinguir en:

- Grids enfocadas en compartir recursos de computación.
- Grids de datos, enfocadas en almacenamiento controlado, administrar y compartir datos heterogéneos distribuidos a gran escala
- Grids de aplicaciones, enfocadas en administración de aplicaciones y el proveer acceso remoto a software y librerías de forma transparente.

- Grids de servicios, resultado de la convergencia entre Grid y orientación al servicio.
- Procesar y apoyar el intercambio eficiente de los servicios.

3.3.5.2. Clasificación de las grids de acuerdo al alcance de los recursos compartidos involucrados. De acuerdo al alcance de los recursos compartidos involucrados en la grid, se pueden distinguir los siguientes usos dados a la grid en las compañías:

- **Cluster de grids:** son una colección de computadores conectados por una LAN de alta velocidad y diseñados para ser usados como un recurso de computación o procesamiento de datos. Un clúster es una identidad homogénea. Sus componentes difieren principalmente en configuración, mas no en arquitectura. Los clúster de grids son recursos locales que operan bajo una firewall y son controlados por una sola entidad administrativa que tiene completo control sobre cada componente. Por esta razón, los clúster no están relacionados con el compartir recursos y no son considerados grids en el más estricto de los sentidos. Sin embargo, son usualmente puntos de partida para construir grids y el primer paso para construir Computación Grid. Los clúster de grids mejoran la capacidad de cómputo y procesamiento en las empresas.

- **Grid empresarial:** El término grid empresarial, se utiliza para referirse a la aplicación de Computación Grid para compartir recursos dentro de los límites de una sola empresa. Todos los componentes de una grid empresarial operan dentro del firewall de la empresa, pero puede ser heterogénea y físicamente distribuida a través de múltiples ubicaciones geográficas de la compañía y pueden pertenecer a diferentes dominios administrativos. Estos sistemas poseen control básico de los recursos y

mecanismos de tolerancia a fallos, así como herramientas de análisis de rendimiento y depuración.

- **Utility Grid:** Una grid que es propiedad y está desplegada por un proveedor, se llama una Utility Grid. Los servicios ofrecidos a través de una Utility Grid generalmente son capacidad de cómputo y/o almacenamiento en forma de pago por uso. Un Utility Grid opera fuera del firewall del usuario, el usuario no es dueño de la Utility Grid y no tiene control sobre su funcionamiento. Esto significa que la empresa usuaria tiene que transmitir los datos y necesidades de cómputo a la Utility Grid, y recoger los resultados también de esta. Es así, que utilizando Utility Grids, los riesgos de seguridad y privacidad, así como preocupaciones con respecto a la fiabilidad se incrementan. Esto tiene un fuerte impacto frente a la necesidad de utilizar o no un Utility Grid, o qué datos exponer a la Utility Grid y qué datos mantener bajo la firewall. Por el otro lado, de forma positiva, Utility Computing no requiere ninguna inversión inicial en infraestructura de TI y permite transformar los costos de inversión de capital en costos variables. Utility Computing, además, ofrece escalabilidad y flexibilidad de los recursos de TI.

- **Grids comunitarias:** el origen de la idea de las grids comunitarias es la e-ciencia. Muchos esfuerzos de investigación, especialmente en las ciencias naturales, requieren de esfuerzos conjuntos de investigación de los científicos y el intercambio de las infraestructuras de las instituciones de investigación de todo el mundo. La cooperación suele dar lugar a una Organización Virtual, donde el intercambio de recursos se lleva a cabo. Hoy en día, la necesidad de cooperación es cada vez mayor también en el mundo de los negocios. Debido a la globalización, las empresas están cada vez más involucradas en las cadenas de suministro globales y el éxito de una empresa depende también cada vez más en la colaboración eficaz con

otras organizaciones. La mayor necesidad de colaboración eficiente también es necesaria en otros procesos de negocio, por ejemplo el diseño colaborativo de productos, los sitios de colaboración en línea y similares.

3.3.6. Topologías Grid. [JACOB 2005], describe las diferentes topologías que pueden adoptar las grids, a continuación se menciona la descripción del autor:

- **Intragrid:** Una topología intragrid típica, existe dentro de una sola organización, proporcionando un conjunto básico de servicios Grid. La organización podría estar integrada por un número de máquinas que comparten un dominio de seguridad común, y comparten los datos internamente en una red privada. Las principales características de una intragrid es un proveedor único de seguridad, el ancho de banda de la red privada es alto y siempre disponible, y hay un ambiente único en una sola red. Dentro de un intragrid, es más fácil de diseñar y operar redes de cómputo y de datos. Una intragrid proporciona un conjunto relativamente estable de los recursos informáticos y la capacidad de compartir datos fácilmente entre sistemas de la red.

- **Extragrid:** Basado en una sola organización, la extragrid amplía el concepto al reunir a dos o más intragrids. Una extragrid, por lo general involucra a más de un proveedor de seguridad, y aumenta el nivel de complejidad de la administración. Las principales características de una extragrid son seguridad dispersa, múltiples organizaciones y conectividad remota/WAN. Dentro de una extragrid, los recursos se vuelven más dinámicos y la grid debe ser más reactiva a los recursos y los componentes que fallan. El diseño se hace más complicado y los servicios de información se vuelven relevantes con el fin de garantizar que los recursos de la grid tienen acceso a la administración de carga de trabajo de gestión en tiempo

real. Una empresa se beneficiaría de una extragrid siempre y cuando exista una iniciativa empresarial para la integración con socios de negocios externos de confianza.

- **Intergrid:** Una intergrid requiere la integración dinámica de aplicaciones, recursos y servicios con patrones, los clientes, y cualquier otra entidad autorizada que tendrán acceso a la grid a través de Internet/WAN. Una topología intergrid, se utiliza principalmente por empresas de ingeniería, industrias de ciencias de la vida, los fabricantes y las empresas del sector financiero. Las principales características de una intergrid son seguridad dispersa, múltiples organizaciones y conectividad remota/WAN. Los datos de una intergrid son datos públicos globales, y las aplicaciones (tanto verticales como horizontales) deben ser modificadas para una audiencia global.

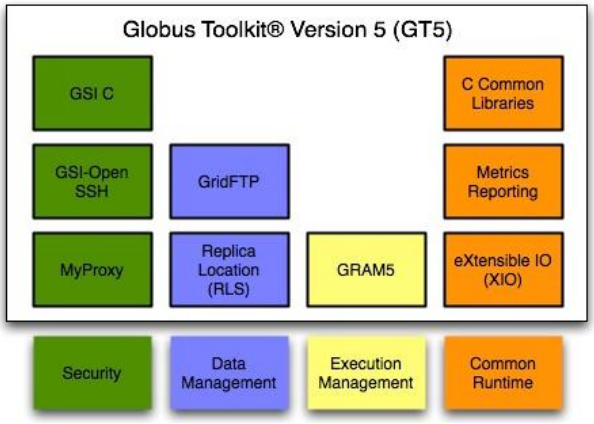
3.3.7. Herramientas de Computación Grid. Se han venido desarrollando muchas soluciones Software que abstraen la complejidad de la grid y brindan este poder computacional mediante herramientas de desarrollo que permiten crear sistemas grid de forma fácil y ágil. A continuación, se realizará una pequeña descripción de algunas plataformas Software que permiten la construcción de Computación Grid:

3.3.5.3. Globus ToolKit. Globus Toolkit (GT) es proporcionado por Globus Alliance y es una herramienta de código libre para construir aplicaciones grid. GT incluye servicios de monitoreo y descubrimiento de recursos, infraestructura para envío de procesos, infraestructura de seguridad y servicios de administración de datos. Permite el desarrollo de nuevos Web Services utilizando Java, C y Phython.

Permite compartir poder computacional, bases de datos y otras herramientas de manera segura a través de los límites corporativos, institucionales y geográficos sin la necesidad de sacrificar autonomía. Globus incluye servicios de software y librerías para monitorear los recursos, descubrimiento, seguridad y administración de archivos. Adicionalmente, es ampliamente utilizado por proyectos de ciencia e ingeniería y es la base fundamental donde compañías líderes en TI están construyendo productos comerciales Grid significativos.

Más información acerca de la herramienta puede ser encontrada en la página Web http://www.globus.org/toolkit/.

Ilustración 8: Arquitectura Globus



Extraído de http://www.globus.org/toolkit/

3.3.5.4. Glite. Glite es un conjunto de componentes diseñados para compartir recursos. En otras palabras, es un Software intermedio para construir grids. Glite es desarrollado el proyecto EGEE (Enabling Grids for E-sciencE) y además combina el esfuerzo de científicos e ingenieros de alrededor de 32 países en el mundo, con el fin de crear una infraestructura grid que sea robusta y altamente accesible. Con la colaboración de 30.000 CPUs y alrededor de 5 Petabytes de almacenamiento, gLite permite cumplir con las demandas de gran

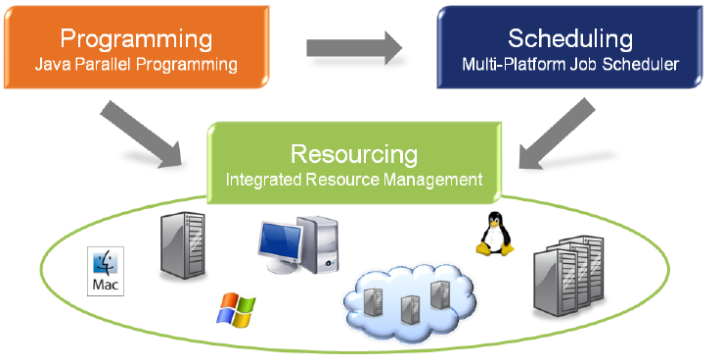
poder computacional en varios campos científicos de geología, finanzas, física y climatología.

Actualmente se encuentra en la versión 3.5 y se puede conseguir más información acerca de esta herramienta en la página Web http://glite.web.cern.ch/glite/

3.3.5.5. Proactive. Proactive es una solución de código abierto para computación en paralelo, distribuida y de múltiples núcleos. Proactive ofrece programación paralela en Java perfectamente integrada con planificación y administración de recursos.

ProActive simplifica la programación y ejecución de aplicaciones en paralelo en Linux, Windows y Mac, junto con la administración de recursos como computadores personales, servidores, clústeres, grids empresariales y nubes.

Ilustración 9: Arquitectura ProActive



Extraído de http://proactive.inria.fr/

Se puede encontrar más información sobre la herramienta ProActive, en la página Web: http://proactive.inria.fr/

3.4. Aplicaciones de la Computación Grid. [ABBAS 2004], menciona algunas aplicaciones donde la Computación Grid ha tenido gran relevancia en diferentes sectores de la industrias:

3.4.1. Método Monte Carlo. El método Monte Carlo es ampliamente utilizado en varías áreas de investigación científica. Este método es utilizado para resolver sistemas con soluciones analíticas desconocidas mediante la aproximación de expresiones matemáticas complejas y difíciles de evaluar. El método se basa en la generación de números seudoaleatorios con el fin de aplicarlos a cualquier tipo de problema y aproximar así una solución.

3.4.2. Servicios financieros. Las aplicaciones financieras son desafiadas a lo largo de dos dimensiones. Primero, deben procesar mucha información, como un modelo de riesgo de cartera, antes de que la información pierda su valor; segundo, el modelo debe ser ejecutado con un grado de rigor para que genere resultados creíbles y recurribles. Casi todas las aplicaciones de modelado financieras hoy en día están basadas en simulaciones Monte Carlo, estas aplicaciones son paralelas, pero la calidad de los resultados se basa en gran medida de la calidad de los flujos de números aleatorios empleados por la aplicación. Los generadores escalables paralelos de números aleatorios, aumentan la disponibilidad de flujos estadísticamente independientes de números aleatorios y aumentan la confianza asociada con los resultados. Es por esto, que las simulaciones Monte Carlo se ajustan muy bien para ser desplegadas en computación grid.

3.4.3. Fábricas. Hay muchas aplicaciones en este sector que adquieren ventajas de la computación en grid hoy en día. Grandes fabricantes de automóviles utilizan computación grid para experimentos de simulación de choques, mientras los fabricantes de aviones lo utilizan para simular túneles de viento. La dinámica de fluidos computacional es un área que también se beneficia enormemente de la computación grid. La dinámica de fluidos computacional predice el

comportamiento del flujo del fluido, transferencia de calor, masa, cambios de fases, relaciones químicas, movimientos mecánicos y la deformación de estructuras sólidas. Millones de dólares pueden ser ahorrados en los costos de construcción de prototipos y potenciales retrasos en los productos utilizando modelos computacionales para estudiar el comportamiento del sistema bajo distintas condiciones utilizando computación grid.

3.4.4. Medios y entretenimiento. Procesamiento de gráficos, compresión de contenidos digitales, y codificación son algunas de las áreas del sector de medios y entreteniendo donde la computación grid está teniendo grandes impactos. Tanto grandes como pequeños estudios obtienen grandes ventajas de esta tecnología para hacer frente al contenido digital que siempre está en incremento. Otro enfoque es la distribución de contenidos digitales. Las compañías radiodifusoras pueden querer compartir archivos multimedia y otros recursos técnicos distribuidos a través de soluciones técnicas ya desarrolladas para computación grid. Por ejemplo, una compañía puede tener una red distribuida procesando video en muchos nodos diferentes al mismo tiempo para mezclar video en vivo con eventos almacenados, lo cual requiere mucho ancho de banda y alto procesamiento.

3.4.5. Ciencias químicas y de materiales. Los científicos de materiales, químicos y bioquímicos, físicos, e ingenieros químicos que realizan simulación y modelado, pueden beneficiarse de un incremento en el poder de cómputo para un procesamiento más rápido. Algunos químicos, por ejemplo, están ocupados en crear y estructurar bases de datos de compuestos y materiales. Estadística combinatoria puede ser aplicada a estos modelos para crear y descubrir nuevos compuestos y materiales que cumplan con las características deseadas. Las grid de datos son utilizadas para manejar este volumen de datos mientras que las grids de cómputo son utilizadas para análisis y simulaciones.

3.4.6. Juegos. Los juegos multi jugadores se han vuelto muy populares en Internet en los últimos años. Mucha gente incluso paga hoy en día para jugar estos juegos. La animación digital y el procesamiento de imágenes necesarios para crear estos tipos de juegos, y el ponerlos a disposición de cientos o miles de usuarios para jugar al mismo tiempo, requieren una enorme capacidad computacional. La tecnología de computación grid está siendo utilizada hoy en día para crear infraestructura escalable que permita el funcionamiento de dichos juegos.

3.4.7. Sensores. Los sensores, puede ser ubicados en una amplia variedad de máquinas u objetos inanimados, incluso en la tierra con el fin de verificar su estado. Algunos de estos sensores, como dispositivos de monitoreo médico, pueden ser puestos en humanos. Estos sensores organizados en un sistema de computación grid, e interconectados con máquinas a través de tecnologías inalámbricas, tienen amplias variedades de usos y grandes beneficios.

3.5. Oportunidades y retos de la Computación Grid

[Dr. BURKE 2008] habla sobre cómo los beneficios que puede traer la Computación Grid giran en torno a los costos y el acceso a los recursos computacionales que muchas instituciones tienen subutilizados, sin embargo, una vez que la grid ha sido desplegada, surgen costos del cambio que son el esfuerzo requerido para poder integrar y habilitar las aplicaciones empresariales para que trabajen en cualquier infraestructura de grid que haya sido seleccionada. Esto, generalmente es realizado utilizando herramientas de desarrollo y generalmente no es muy significante, pero debe ser cuidadosamente revisado. Una forma de mitigar estos costos es introducir nuevo software grid en la empresa que soporte nuevas aplicaciones habilitadas para la grid, y dejar sin cambios el software existente y su integración con el software de la grid.

Otro reto que debe enfrentar la Minería de Datos en los ambientes grid es mencionado por [Ídem], y es que la necesidad de poder separar los problemas de tal forma que sean distribuibles a través de la grid puede ser un cuello de botella a la hora de tratar de implementar los nuevos sistemas. La Computación Grid también depende de una comunicación de datos confiable y de bajo costo para ser exitosa. El costo de mejorar la velocidad de la transmisión de datos debe ser comparado con el costo de mejorar los recursos computacionales. Los datos deben ser procesados lo más cerca posible de donde están almacenados para así poder reducir costos de la red.

Otro problema es la estabilidad de los datos. Lo ideal es trabajar con datos que no son frecuentemente ingresados, actualizados o corregidos. Estos cambios pueden causar la necesidad de volver a ser computados y así incrementar el tiempo de cómputo, el costo de transmitir datos en la red, y causar un incremento en la congestión de la red. Los problemas que utilizan muchas cantidades de datos son mejor manejados cuando el tiempo de computación y de red son mínimos.

[JACOB 2005], explica las barreras que frecuentemente existen para obtener la escalabilidad perfecta a la hora de querer utilizar la Computación Grid para realizar procesamiento en paralelo. La primera barrera, depende de los algoritmos utilizados para dividir la aplicación entre varias CPUs. Si el algoritmo puede ser únicamente separado en un número limitado de partes independientes, entonces, ahí hay una barrera de escalamiento. La segunda barrera aparece si las partes no son completamente independientes, esto definitivamente limita la escalabilidad. Por ejemplo, si todos los subprocesos necesitan leer y escribir sobre un mismo archivo o base de datos, el acceso limitado de ese archivo o base de datos se convertirá en un factor limitante en la escalabilidad de la aplicación. Otro problema que puede surgir es la latencia de las comunicaciones entre los procesos, las capacidades de la red de comunicaciones, los protocolos de sincronización, el ancho de banda, retrasos que interfieran en las necesidades de tiempo real, etc.

Un aspecto muy importante a tener en cuenta mencionado en [DUBITZKY 2008], es que los ambientes de Computación Grid pueden ser muy complejos y su complejidad nace de la heterogeneidad de los propios recursos Software y Hardware que componen la grid. [ALKADI 2007], explica cómo a medida que la escalabilidad de la grid incrementa, los problemas asociados con su buen funcionamiento tienden a crecer proporcionalmente. Las principales áreas de preocupación son la seguridad, la asignación de recursos, la programación de tareas, el acceso a datos, gestión de políticas, tolerancia a fallos, recuperación de errores, autonomía y calidad del servicio.

Sin duda alguna, todos estos son aspectos que retan el uso de la Computación Grid y surgen amplias oportunidades de investigación y desarrollo para la mejora de las herramientas disponibles y así poder explotar de forma efectiva el poder computacional que brinda la grid.

4. MINERÍA DE DATOS GRID

4.1. Significado de la Minería de Datos Grid

Actualmente, las organizaciones están generando más y más datos a una tasa exponencial y por esta razón cada vez es necesario contar con más poder computacional y facilidades de almacenamiento para realizar procesos de Minería de Datos. [KURMAN 2008] hace este análisis y menciona cómo datos en el orden de Terabytes a Pentabytes están siendo generados por muchos proyectos científicos incluyendo la astronomía, la física, la bioinformática, etc. Y por su parte las organizaciones de los diferentes sectores no se quedan atrás a la hora de generar grandes cantidades de datos como insumo principal para la toma de decisiones. Minería de Datos es el proceso de extraer información y conocimiento de grandes volúmenes de datos, y por su naturaleza, la Minería de Datos es un proceso que consume muchos recursos computacionales.

[DEPOUTOVITCH 2005] describe cómo cada vez se dificulta más incluso alcanzar pequeñas mejoras de rendimiento sólo ajustando los algoritmos de Minería de Datos. [Ídem] explica que el muestrear los datos para realizar los cálculos puede ayudar, sin embargo, hay un alto precio al reducir la exactitud de los resultados lo cual es inaceptable a la hora de tener que tomar decisiones críticas frente a estos. El incrementar la capacidad de Hardware no ayuda mucho ya que éste siempre tendrá sus límites y no importa cuánto dinero se invierta no se pueden superar estos límites. Cuando no se puede mejorar la capacidad de un CPU, se puede incrementar su número. Es aquí donde los procesadores multiprocesos y los servidores se vuelven comunes. Aunque estos ofrecen mejor rendimiento, el número total de procesadores en un servidor sigue siendo limitado y además son excesivamente costosos.

Por estas razones, hay que buscar una solución, pues siendo la Minería de Datos un factor competitivo hoy en día para las organizaciones, existe la urgente necesidad de lograr superar los problemas de la necesidad de contar con poder computacional suficiente para que los procesos de Minería de Datos sean eficientes y efectivos para la toma de decisiones. Es allí cuando surge en rescate de la Minería de Datos la Computación Grid. Para enfrentar los problemas de rendimiento, la Computación Grid es una de las mejores alternativas basada en la red que permite obtener un alto rendimiento. La Computación Grid posibilita utilizar procesamiento cooperativo de una o varias tareas utilizando diferentes computadores.

[DUBITZKY 2008] habla sobre este tema y explica cómo muchos proyectos de investigación se han enfocado en proveer sistemas que faciliten la Minería de Datos en sistemas basados en grid. Cada vez más, se necesitan procesos de Minería de Datos en los campos de mercadeo, negocios, inteligencia, multimedia y muchas otras áreas donde la información y las necesidades de procesamiento crecen cada día y requieren ser procesadas para la toma de decisiones, y por lo tanto los tiempos de procesamiento deben ser rápidos y se requieren mayores recursos de computación. Es allí donde la Computación Grid tiene su alcance, y por lo tanto, hay una gran necesidad de algoritmos distribuidos de Minería de Datos que trabajen bien en la grid.

4.2. Técnicas y Herramientas de Minería de Datos Grid

Una de las grandes ventajas de las plataformas de Computación Grid, es que extraen la complejidad de la grid, es decir, para el usuario final que ejecuta los procesos la grid debe ser transparente. Por esta razón, al utilizar Minería de Datos Grid, los procesos generales de Minería de Datos no cambian, las aplicaciones deben permitir que estos sean completamente transparentes frente al ambiente

grid. Igualmente las técnicas (algoritmos) de Minería de Datos teóricamente deben generar los mismos resultados, sin embargo, pueden requerir parametrización adicional de acuerdo a su nueva naturaleza distribuida y las consideraciones del autor de la solución. Es el funcionamiento como tal de los algoritmos que cambia a su interior y la forma como las herramientas tienen que gestionarlos.

4.2.1. Creando Minería de Datos Grid. [DEPOUTOVITCH 2005] en su artículo habla acerca de la Minería de Datos Grid y brinda una descripción general del proceso de generación de modelos de Minería de Datos en Grid. Lo siguiente es un extracto del artículo que nos permite entender cómo es este proceso.

En la Minería de Datos, la predicción es la tarea de producir conocimiento basado en los datos. Los datos de entrada para la predicción están divididos en dos partes: los datos de entrenamiento y los datos objetivo para producir la predicción. Ambas partes tienen las mismas variables. Las variables se dividen en dos grupos: independientes y dependientes. Las variables independientes son utilizadas tanto para entrenamiento como predicción, mientras las dependientes son utilizadas sólo en la parte de predicción. La meta es predecir variables dependientes en los datos.

Existen muchos métodos de predicción disponibles, entre estos están:

- Árboles de decisión, que permiten generar conjuntos de reglas
- Regresiones lineales, que están basadas en los datos de entrenamiento con el fin de generar fórmulas para calcular las variables dependientes en los datos objetivos utilizando las variables independientes
- Método del vecino más cercano, que busca los registros más similares entre sí y utiliza sus vecinos para clasificación y predicción.

Antes de generar algoritmos de Minería de Datos distribuidos, primero se requiere verificar si es posible desarrollar su versión de procesamiento en paralelo. Los dos

primeros métodos requieren mucho tiempo de preparación y requieren muchos cambios para poder paralelizar los algoritmos. Sin embargo, el tercer método es atractivo porque se puede realizar con el mínimo pre procesamiento y además se puede predecir cada registro independientemente de los otros, por lo tanto su algoritmo no requiere ningún tipo de modificación.

El algoritmo de Minería de Datos distribuido, se puede describir de la siguiente manera:

1. Dividir el conjunto de datos objetivos en el número de subconjuntos deseado. El número de subconjuntos debe ser suficiente para lograr obtener ventajas de los nodos computacionales disponibles. Por otro lado, los subconjuntos deben tener el tamaño ideal de tal forma que el consumo en el proceso de planificación y pre procesamiento no sea significante comparable con el tiempo de procesamiento del conjunto de datos completo.
2. Copiar los datos de entrenamiento en cada nodo de procesamiento
3. Procesar el entrenamiento de los datos en cada nodo
4. Asignar un subconjunto de los datos a cada nodo para realizar las predicciones vs. el sub conjunto de datos asignado.
5. Obtener los resultados de los diferentes nodos de procesamiento.

4.2.2. Algoritmos distribuidos de Minería de Datos. Los algoritmos para Minería de Datos Grid aun requieren amplia investigación y desarrollo. Aunque ya se han creado algunas herramientas que ofrecen Minería de Datos Grid, realmente son pocas las que se pueden encontrar en el mercado y es un área en la que aún se continua investigando para lograr desarrollar algoritmos y métodos de Minería de Datos adecuados para funcionar eficientemente en la grid.
Hay autores que mencionan propuestas de algunos algoritmos de Minería de Datos en su versión distribuida. [KURMAN 2008] menciona algunos algoritmos

como árboles de decisión, reglas de asociación, apriori, k-Means y algunas otras propuestas para realizar segmentación y clasificación. También [ZAKI 2000], expone significativos avances en algoritmos de Minería de Datos en su versión paralela para realizar clasificación, segmentación y asociación, lo cual los hace buenos candidatos para ser adaptados a versión distribuida.

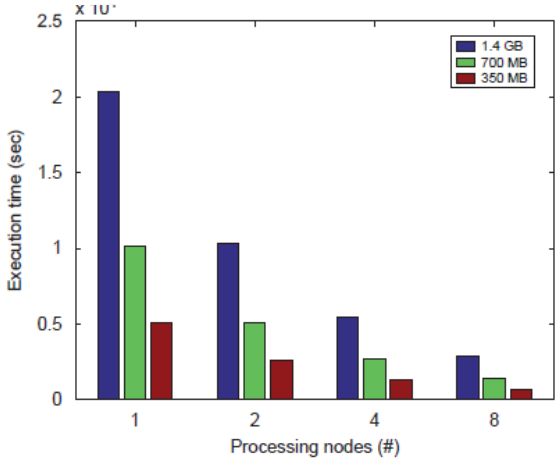
[GLIMCHER 2007], profundiza más acerca del tema cuando habla acerca de un sistema middleware que se ha venido desarrollando desde hace algunos años llamado FREERIDE-G. Éste está enfocado en los retos involucrados en el desarrollo de la implementación de algoritmos de Minería de Datos paralelos y distribuidos. [Ídem] menciona que éste sistema "se basa en la observación de que las versiones paralelas de varias técnicas de Minería de Datos bien conocidas, comparten una estructura similar relativa, y pueden ser paralelizadas dividiendo los datos a través de nodos. La computación en cada nodo involucra leer los datos en un orden arbitrario, procesar cada dato y realizar cálculos de manera local. Los cálculos involucran únicamente operaciones conmutativas y asociativas, lo cual significa que el resultado es independiente del orden en que los datos son procesados. Luego de los cálculos locales en cada nodo, se realiza un cálculo global para obtener finalmente el modelo".

Según [Ídem], algunas de las técnicas de Minería de Datos en su versión paralela y distribuida que han sido cuidadosamente estudiadas son: Asociación a priori, redes bayesianas para clasificación, segmentación k-Means, clasificación k-vecino más cercano y redes neuronales artificiales. Algunos otros autores, mencionan versiones de algoritmos de Minería de Datos en su versión paralela y distribuida, tales como: reglas de asociación [SAKTHI 2008], máquinas de vectores [MELIGY 2009], minería de texto [SARNOVSKÝ 2009].

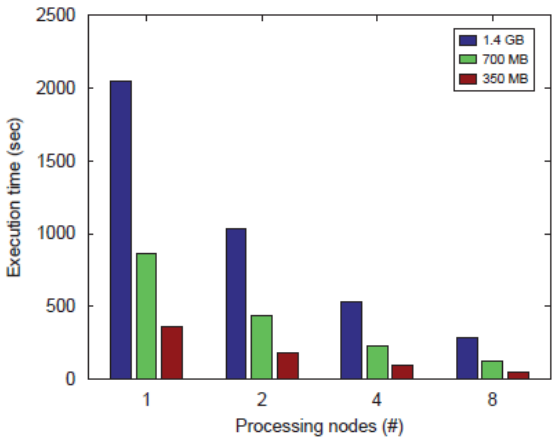
[GLIMCHER 2007] hace un estudio de FREERIDE-G con el fin de evaluar el rendimiento de los algoritmos paralelos de Minería de Datos en Grid: k-means y k-

vecino más cercano. Los algoritmos fueron evaluados con 3 conjuntos de datos, de tamaño de 350 MB, 700 MB y 1.4 GB, respectivamente. La prueba mide el tiempo de ejecución de estos tres conjuntos de datos utilizando 1, 2, 4 y 8 nodos. Las siguientes gráficas muestran los resultados obtenidos:

Ilustración 10: Rendimiento algoritmos k-means y k-vecino más cercano



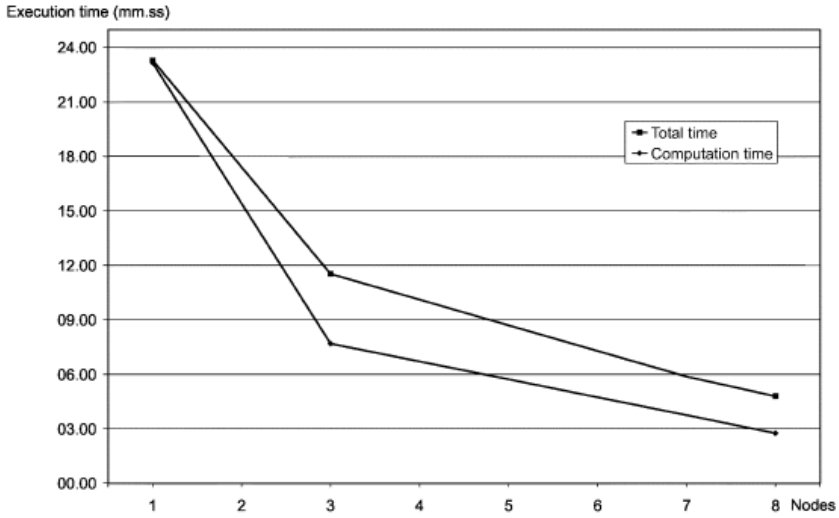
Rendimiento algoritmo k-means [GLIMCHER 2007]



Rendimiento algoritmo k-vecino más cercano [GLIMCHER 2007]

[CANNATARO 2004], realiza un análisis similar para el algoritmo de Minería de Datos C4.5 (árboles de decisión), en su versión distribuida, utilizando el framework de Minería de Datos Grid Knowledge Grid. Para el análisis se utilizó una base datos de un tamaño de 712 MB, con aproximadamente 5 millones de registros. El experimento se realizó con estaciones de trabajo de especificaciones hardware desde Pentium III 800 MHz hasta Pentium 4 1500 MHz. El resultado fue el siguiente:

Ilustración 11: Rendimiento algoritmo C4.5



Rendimiento algoritmo C4.5 [CANNATARO 2004]

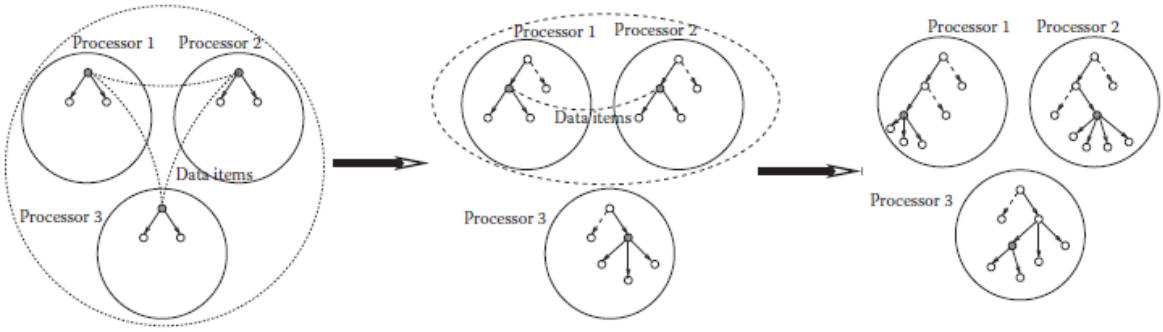
A continuación, se presentará un resumen de algunas propuestas de algoritmos de Minería de Datos desarrollados para funcionar en ambientes distribuidos.

4.2.2.1. Árboles de Decisión. [KURMAN 2008], explica cómo los árboles de decisión son muy populares en las herramientas de Minería de Datos, ya que éstos obtienen exactitud razonable y son relativamente menos costosos para construir y usar computacionalmente.

Un árbol de decisión consiste en un conjunto de nodos y hojas. Cada nodo tiene una decisión de división y un atributo de división asociado a él. Las hojas tienen una etiqueta de clase asignada. Una vez se construye un árbol de decisión, el proceso de predicción es relativamente sencillo: el proceso de clasificación comienza en la raíz, y una ruta hacia alguna hoja es seguida de acuerdo a las decisiones de división en cada nodo. La etiqueta de clase adjunta a la hoja es finalmente asignada al registro de entrada.

Hay un paralelismo inherente en la construcción de algoritmos de árboles de decisión, ya que todos los hijos de un nodo pueden ser procesados concurrentemente. El proceso comienza con todos los procesadores cooperando para expandir el nodo raíz. Los procesadores son luego divididos a través de los nodos hijo y cada división trabaja en su parte del subárbol. Este proceso es repetido recursivamente hasta que cada procesador tiene su propio subárbol.

Ilustración 12: Generación árbol de decisión

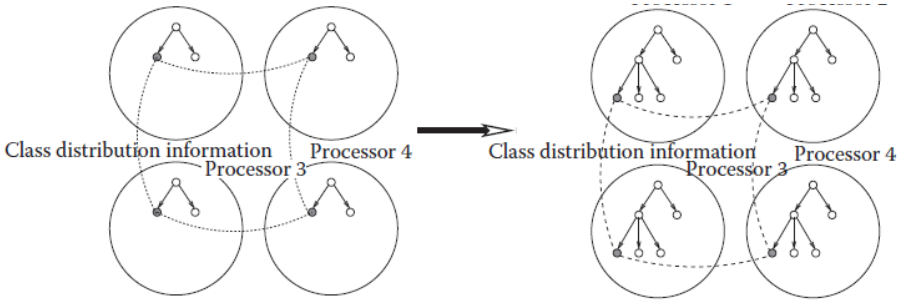


Extraído de [KURMAN 2008]

La imagen muestra un ejemplo. Inicialmente todos los tres procesadores trabajan juntos para extender el nodo raíz y obtener dos nodos hijos. Los procesadores 1 y 2 son asignados al hijo izquierdo, mientras el procesador 3 es asignado al hijo derecho. Luego, el hijo izquierdo es expandido para obtener 2 hijos que son distribuidos entre los Procesadores 1 y 2. Cada procesador ahora construye su propio subárbol utilizando un algoritmo secuencial.

Otro método, también llamado construcción de árbol síncrona, construye el árbol concurrentemente en todos los procesadores. Para cada nodo del árbol, todos los procesadores intercambian información con el fin de seleccionar el atributo más apropiado. La siguiente imagen muestra un ejemplo donde el nodo actual es el hijo más a la izquierda de la raíz. Todos los procesadores colaboran para expandir este nodo y obtener tres hijos. Nuevamente, el nodo más a la izquierda de esos nodos hijos es el nodo actual y los procesadores trabajan juntos para expandirlo.

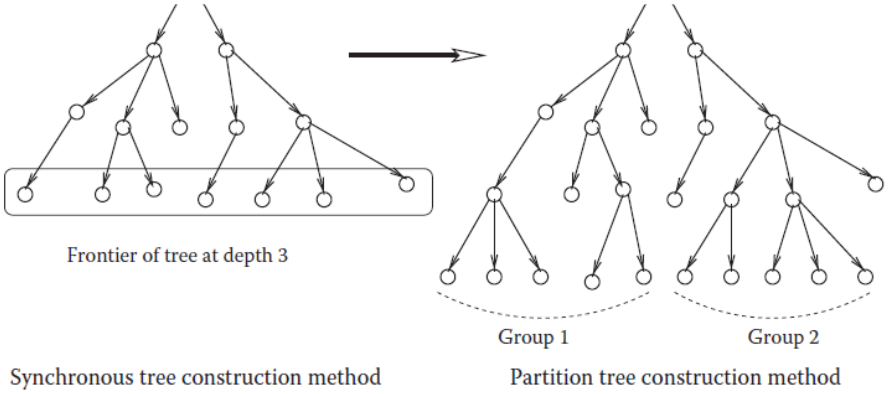
Ilustración 13: Generación árbol de decisión síncrona



Extraído de [KURMAN 2008]

Otro método consiste en la generación híbrida del árbol, este método combina las ventajas de los dos algoritmos anteriores para obtener una técnica más eficiente. El costo de comunicación de la construcción del árbol de forma síncrona se vuelve muy alto si la frontera del árbol crece mucho. Por otro lado, el método de construcción del árbol de forma particionada tiene altos costos al mover los datos correspondientes a los diferentes grupos de procesadores cuando se realiza la partición. La formulación híbrida del algoritmo aplica la construcción síncrona del árbol siempre y cuando los costos de comunicación estén dentro de los límites. Cuando el costo de comunicación alcanza el límite asignado, la frontera actual del árbol es particionada en dos partes; una parte es asignada a la mitad de los procesadores y la otra es asignada a la otra mitad como se muestra en la siguiente imagen.

Ilustración 14: Generación árbol de decisión híbrida



Extraído de [DUBITZKY 2008]

Los algoritmos anteriores, fuero inicialmente diseñados para implementaciones en paralelo, pero han sido muy bien adaptados para trabajar en sistemas distribuidos.

4.2.2.2. k-Means. Según [Ídem]**,** k-means es uno de los algoritmos desarrollados más simples para generar clústeres. El algoritmo selecciona arbitrariamente k centroides de clústeres iniciales e iterativamente refina la ubicación de los datos en los clústeres basado en qué tan bien se ajustan a la distribución natural del conjunto de datos de entrada. Una versión distribuida de este algoritmo está basada en la premisa de que se puede realizar una aproximación decente de las posiciones de los clústeres a partir de una muestra pequeña del conjunto de datos entero, y las posiciones pueden ser distribuidas iterativamente a los nodos de procesamiento y allí mismo refinadas para obtener el mismo resultado de k-means.

Con el fin de generar las ubicaciones aproximadas del centroide, cada nodo envía una muestra de sus datos al nodo central y las muestras son agrupadas con k-means. El proceso refinado requiere que cada nodo haga por lo menos un recorrido a través de los datos y rastree todos los puntos que caen cerca de los límites del clúster. Como las operaciones de muestro y ajuste requieren transmisión individual de registros, este algoritmo es adecuado sólo para las aplicaciones que no requieren privacidad de los datos.

4.2.2.3. Máquinas de Vectores. [MELIGY 2009], expone una versión distribuida del algoritmo de Minería de Datos máquina de vectores, el cual puede ser programado en C y MPI (interface de envío de mensajes). El algoritmo se divide en dos tipos de procesos, proceso maestro y proceso esclavo, los cuales funcionan de la siguiente manera:

Proceso Maestro:

1. Leer los datos de entrada “muestra de entrenamiento” y almacenarlos en un array.
2. Enviar el array de datos de entrenamiento con su tamaño a cada nodo de procesamiento.
3. Recibir la los resultados de procesamiento desde cada nodo.
4. Almacenar los valores del array y encontrar el valor máximo entre los n nodos, el cual será la etiqueta de la clase.

Proceso esclavo:

1. Recibir los datos de entrenamiento del proceso maestro y almacenarlos en un array de datos.
2. Read Clase 1 "datos de prueba" del archivo y almacenarlo en un array.
3. Combinar los datos de aprendizaje con la Clase 1 y guardar el resultado en un array.
4. Encontrar la suma de la combinación del array.
5. Enviar los resultados de la suma de vuelta al proceso maestro.

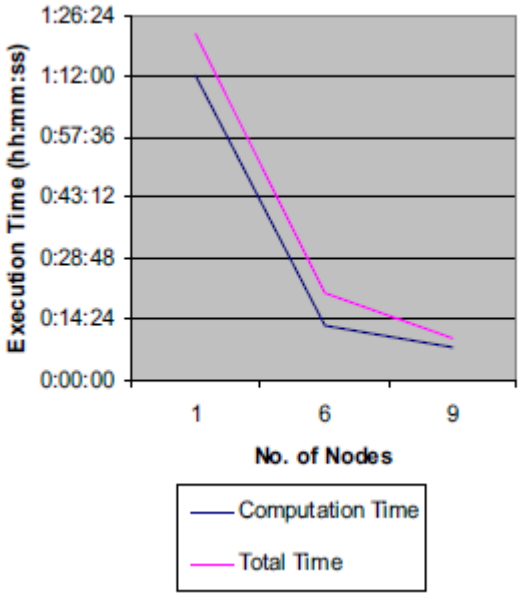
4.2.2.4. Reglas de asociación. [SAKTHI 2008], explica una propuesta de algoritmo distribuido para reglas de asociación. Una regla de asociación es una regla que implica cierta asociación entre un conjunto de datos (que pueden ocurrir juntos o que uno implica el otro) en una base de datos. Apriori es el algoritmo más frecuentemente usado para reglas de asociación.

Minar reglas de asociación en una única gran base de datos requiere mucho poder de procesamiento. Dadas las propiedades de los ambientes distribuidos, la tecnología convencional para utilizar Minería de Datos centralizada ya no sigue siendo adecuada para los nuevos sistemas. El algoritmo Apriori puede ser ejecutado en una grid de múltiples nodos en paralelo.

En cada nodo de la grid, el algoritmo distribuido realiza conteos de soporte locales y elimina los elementos no frecuentes. Luego de completar la eliminación, cada nodo emite mensajes que contienen todos los conjuntos remanentes a los otros nodos de la grid para solicitar sus conteos de soporte. Subsecuentemente, todos los nodos encuentran los conjuntos de elementos globales para esa iteración, y luego proceden a la siguiente iteración.

[Ídem] realiza pruebas del algoritmo. Para la prueba se utilizaron algunos nodos interconectados en una LAN y otros en una WAN. En cada nodo se instaló Globus 3 toolkit y el software necesario para calcular reglas de asociación de forma distribuida sobre los datos. El hardware de las máquinas es P4 2.4 MHz, 1 Gb RAM y con los sistemas operativos WindowsXP y Linux. Los resultados son los siguientes:

Ilustración 15: Rendimiento algoritmo Apriori



Rendimiento algoritmo Apriori [SAKTHI 2008]

4.2.2.5. Redes neuronales. [PETHICK 2003], propone 2 estrategias para implementar el algoritmo de redes neuronales de forma distribuida, las cuales denomina paralelismo ejemplar y paralelismo neural.

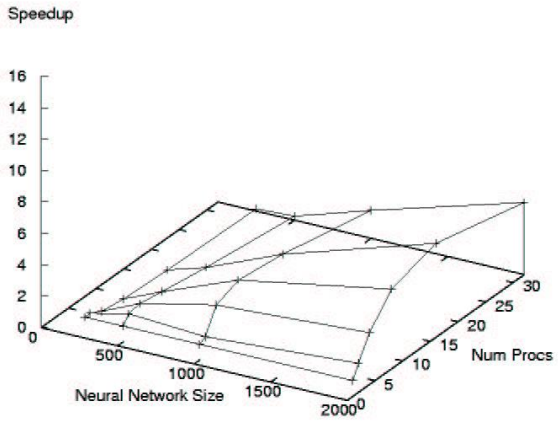
Paralelismo ejemplar: Esta versión distribuida del algoritmo, se diferencia de su versión centralizada únicamente en la adición de un proceso de inicialización y un paso de sincronización al final de cada epoch. El proceso de inicialización consiste en la distribución de los datos de entrenamiento a todos los procesos y la sincronización de todas las matrices de peso iniciales. La sincronización al final de cada epoch, involucra de cada proceso el envío de sus matrices de cambios de peso parciales al nodo central, quien las sumariza y comunica un conjunto actualizado de matrices de peso a todos los procesos como preparación para el siguiente epoch. La parte esencial del algoritmo centralizado, el cual calcula la salida de la red neuronal de un patrón de entrada dado y determina los cambios de peso para corregir cualquier error, no tienen ningún cambio, y es ejecutado por todos los procesos en una copia idéntica de las matrices de peso.

Paralelismo neural: el algoritmo utiliza el paralelismo natural que implica la naturaleza distribuida de una red neuronal artificial. En el caso más puro, cada procesador en la grid es responsable por calcular la activación de una sola neurona, aunque esto generalmente no es práctico ni conveniente. En su lugar, se debe determinar un buen mapeo topológico de las neuronas para cada nodo. Si se requiere una actualización del patrón, sólo las neuronas de una sola capa pueden ser evaluadas en paralelo mientras que las neuronas de las capas subsiguientes utilizan esas activaciones como sus entradas.

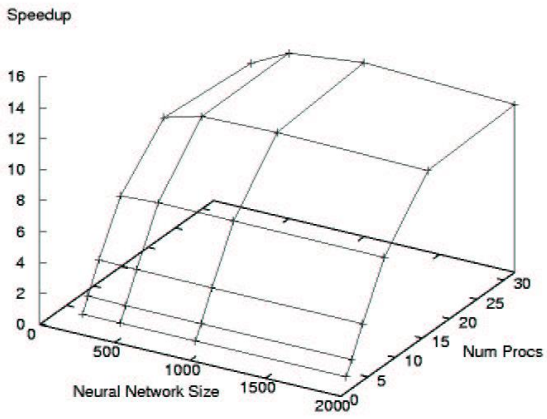
[Ídem] realiza pruebas del algoritmo. Para las pruebas, se utilizaron cuatro conjuntos de datos de diferente tamaño, cada uno con 100, 1000, 10000, y 20000 pares de entrenamiento. Cada prueba se realizó en una grid compuesta por 1, 2, 4, 8, 16, o 32 nodos. Las pruebas se realizaron utilizando 32 máquinas Red Hat

GNU-Linux. El hardware de cada máquina fue Intel Pentium II de 350 MHz y 192MB de memoria. La red fue una Ethernet LAN de 100MB.

Ilustración 16: Rendimiento algoritmo Redes Neuronales



Rendimiento algoritmo paralelismo neural [PETHICK 2003]

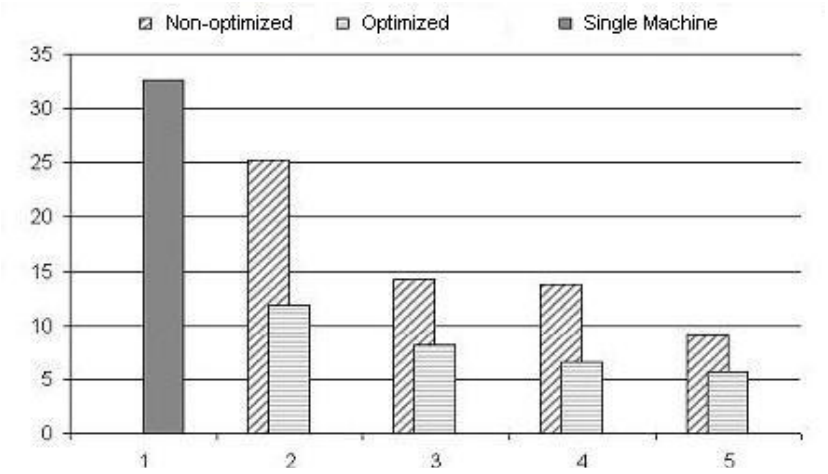


Rendimiento algoritmo paralelismo ejemplar [PETHICK 2003]

4.2.2.6. Minería de Texto. [SARNOVSKÝ 2009] define tres posibilidades para realizar minería de texto distribuida. Una de ellas es clasificación mediante árboles de decisión: la clasificación de texto es el problema de asignar un documento de texto en uno o más categorías o clases basado en su contenido. La segunda posibilidad es segmentación usando mapas auto-organizativos y la tercera es el uso de análisis conceptual formal en análisis de texto

- Pruebas con árboles de decisión: Para la prueba se utilizaron cinco estaciones de trabajo Sun Blade 1500, 1062 MHz Sparc CPU, 1.5 GB RAM conectadas a una red de 100 MBit.

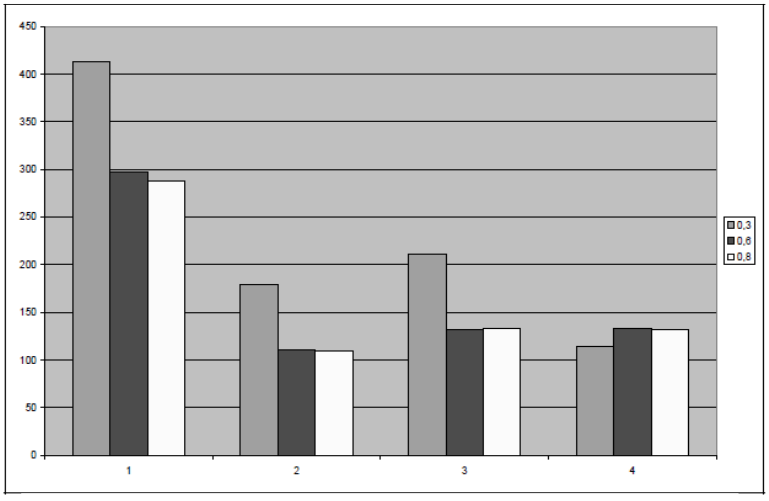
Ilustración 17: Pruebas utilizando árboles de decisión



Extraído de [SARNOVSKÝ 2009]

- Pruebas con segmentación usando mapas auto-organizativos: Para la prueba se utilizó un servidor central (4 x UltraSPARC-III 750 MHz, 8 GB RAM) y seis estaciones de trabajo SUN, una red de 100 Mbit/s, colecciones de datos Times60 (420 documentos) y Reuters-21578 (12902 documentos).

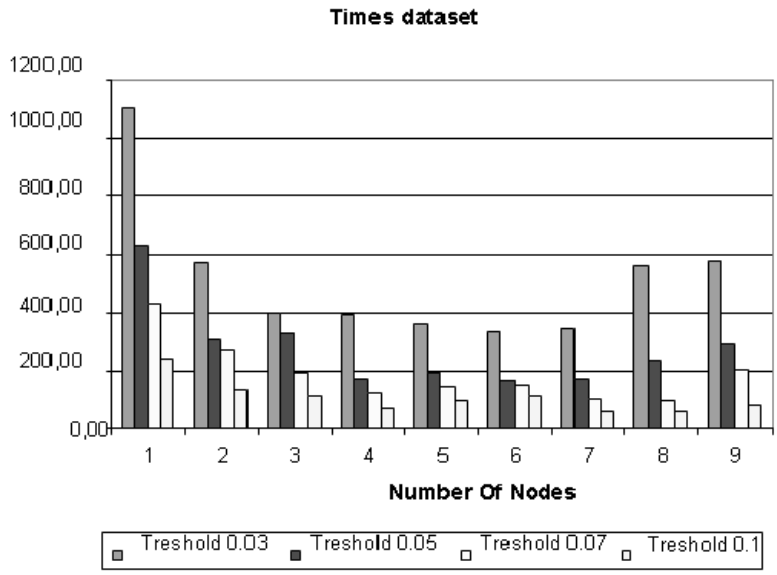
Ilustración 18: Pruebas con segmentación usando mapas auto-organizativos



Extraído de [SARNOVSKÝ 2009]

- Pruebas con análisis conceptual formal: Para la prueba se utilizaron nueve estaciones de trabajo conectadas en la grid se utilizaron varios valores para el parámetro threshold de 0.03, 0.05, 0.07 y 0.1. Para cada valor del parámetro threshold se ejecutó el algoritmo 3 veces y se realizó un promedio del tiempo de las ejecuciones.

Ilustración 19: Pruebas con análisis conceptual formal



Extraído de [SARNOVSKÝ 2009]

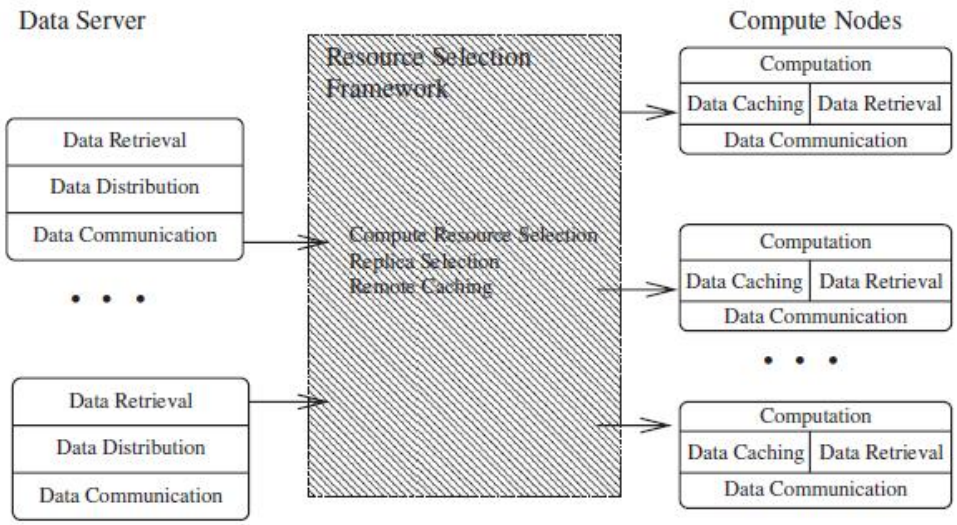
4.2.3. Herramientas de Minería de Datos Grid

4.2.3.1. FREERIDE-G. [GLIMCHER 2007], describe detalladamente este framework para Minería de Datos Grid. La función básica del sistema es automatizar la obtención de datos de repositorios remotos y coordinar el análisis paralelo de dichos datos a través de recursos computacionales de usuarios finales. El sistema espera que los datos estén almacenados por trozos, cuyo tamaño sea manejable por los nodos.

Este middleaware está desarrollado como un sistema cliente/servidor. Los tres componentes principales son el servidor de datos, el nodo computacional cliente, y un framework de selección de recursos. El servidor de datos corre en cada nodo de repositorio de datos en línea, con el fin de automatizar el envío de datos a los nodos de procesamiento de usuario final. Un servidor de procesamiento corre en cada nodo de procesamiento de usuario final con el fin de recibir los datos del repositorio en línea y realizar análisis específicos de aplicación en estos.

El framework es configurable para acomodar n nodos de servidor de datos y m nodos de procesamiento de usuario. A continuación, se presenta un esquema de la arquitectura FREERIDE-G.

Ilustración 20: Arquitectura FREERIDE-G



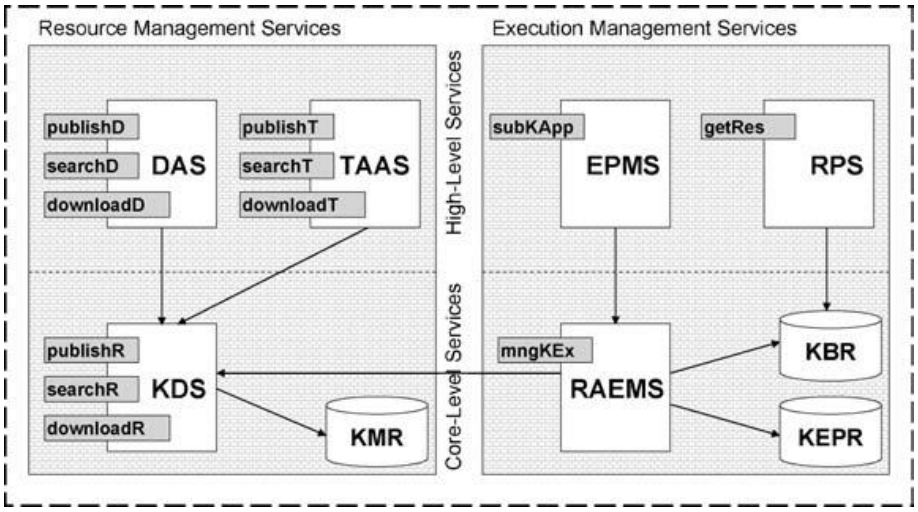
Extraído de [GLIMCHER 2007]

4.2.3.2. The Knowledge Grid. [CANNATARO 2004] y [DUBITZKY 2008], mencionan un framework para Minería de Datos Grid llamado Knowledge Grid. Es una arquitectura de software paralela y distribuida que integra técnicas de Minería de Datos y tecnologías Grid. En la arquitectura de Knowledge Grid, las

herramientas de Minería de Datos están integradas con mecanismos de datos grid genéricos y servicios. Por esta razón, Knowledge Grid, puede ser utilizada para realizar Minería de Datos en conjuntos de datos muy grandes a través de grids.

La arquitectura de Knowledge Grid, utiliza mecanismos de Grid básicos para construir servicios de descubrimiento de conocimiento utilizando herramientas y servicios Grid. Estos servicios pueden ser desarrollados de diferentes maneras utilizando ambientes habilitados para la grid. El sistema está implementado basado en Globus Toolkit. A continuación, se presenta un esquema de la arquitectura de Knowledge Grid.

Ilustración 21: Arquitectura Knowledge Grid



Extraído de [DUBITZKY 2008]

4.2.3.3. Grid Weka. Grid Weka, es una modificación de la herramienta de Minería de Datos conocida como Weka. Esta modificación habilita Weka para utilizar los recursos de varios computadores a la hora de realizar diferentes funciones.

Grid Weka consiste en dos componentes: Servidor Weka y Cliente Weka. Cada computador participando en la Grid debe ejecutar el Servidor Weka.

La herramienta se puede utilizar en los siguientes escenarios:

- Construir un clasificador en una máquina remota
- Probar un clasificador previamente construido, en varias máquinas de forma paralela
- Predicción en varias máquinas de forma paralela, sobre un conjunto de datos utilizando un clasificador previamente construido.
- Utilizar varias máquinas para realizar validaciones en paralelo

Se puede encontrar más información acerca de la herramienta en la página web http://userweb.port.ac.uk/~khusainr/weka/

4.2.3.4. Weka4Ws. Weka4Ws, es un marco de trabajo desarrollado para extender la funcionalidad de la herramienta de Minería de Datos Weka para soportar Minería de Datos distribuida en ambientes Grid.

La meta de Weka4WS, es extender Weka para soportar la ejecución remota de algoritmos de Minería de Datos a través de WSRF Web Services. De esa manera, se pueden ejecutar tareas de Minería de Datos distribuida de forma concurrente en nodos de la grid descentralizados explotando así la distribución de los datos y mejorando el rendimiento de la herramienta.

En Weka4WS, se pueden ejecutar de forma remota en recursos grid, algoritmos de Minería de Datos de clasificación, clúster y reglas de asociación. Para habilitar la ejecución remota, los algoritmos disponibles en la librería Weka son expuestos a través de un Web Service, permitiendo que sean fácilmente desplegados en nodos de la Grid. Weka4WS también extiende la interfaz de usuario de Weka con

el fin de permitir la invocación de los algoritmos de Minería de Datos que están expuestos como un Web Service en nodos remotos de la Grid.

Con el fin de asegurar la interoperabilidad con ambientes grid estándares, Weka4WS ha sido diseñado utilizando WSRF (Web Services Resource Framework) y además ha sido desarrollado utilizando la librería WSRF que provee Globus Toolkit.

Weka es distribuido en dos paquetes separados:

1. Weka4WS-client, que contiene el software cliente que debe ser instalado en las máquinas de los usuarios.
2. Weka4WS-service, que contiene el WSRF que debe ser instalado en los nodos de procesamiento

Weka4WS requiere la instalación completa de Globus Toolkit en los nodos de procesamiento y solo el Java WS Core (subconjunto del Globus Toolkit) en los nodos de usuario. La versión completa del Globus Toolkit funciona sólo en plataformas Unix, por lo tanto, Weka4WS-service solamente puede ser instalado en este tipo de sistemas, mientras que Weka4WS-client puede ser instalado en plataformas Unix o Windows.

Se puede encontrar más información acerca de la herramienta en la página web http://grid.deis.unical.it/weka4ws/

4.2.3.5. DataMiningGrid. DataMiningGrid es una herramienta desarrollada para proveer un ambiente adecuado para ejecutar análisis de datos y descubrimiento de conocimiento, utilizando Minería de Datos basada en Grid. DataMiningGrid es una herramienta de código libre que provee funcionalidades de Minería de Datos especializadas como manipulación de datos, intermediación de

datos, y otras aplicaciones de acuerdo a las diferentes tareas y metodologías de Minería de Datos.

Algunos aspectos clave en el desarrollo de la herramienta DataMiningGrid son:

1. Transparencia de la grid: Los usuarios finales pueden realizar tareas de Minería de Datos sin la necesidad de entender los aspectos tecnológicos de los sistemas grid.

2. Soporte para desarrollo de aplicaciones: Los desarrolladores de soluciones de Minería de datos, pueden habilitar aplicaciones de Minería de Datos existentes para la Grid con poca intervención en el código fuente existente de las aplicaciones.

3. Arquitectura orientada al servicio e interoperabilidad: La herramienta se adhiere a las tecnologías grid y estándares emergentes como WSRF, y está basado ampliamente en la tecnología de código abierto.

Algunas de sus características son:

- Soporta estándares de Minería de datos (CRISP-DM, PMML)
- Interfaz gráfica diferente de acuerdo al usuario (usuario final, desarrollador, administrador)
- Soporte de flujos de trabajo
- Arquitectura orientada al servicio (SOAP)
- Administración de los datos distribuida
- Soporta estándares Grid
- Middleware, basado en Globus Toolkit, Unicore, entre otros
- Paralelismo

Se puede encontrar más información acerca de la herramienta en la página web http://www.datamininggrid.org

4.2.3.6. GridMiner. La meta de la herramienta es hacer frente a todas las tareas del proceso de descubrimiento de conocimiento en la grid, e integrarlas en una aplicación avanzada orientada al servicio.

La arquitectura del sistema GridMiner está basada en CRISP-DM. Grid Miner provee un ambiente confiable de alto rendimiento de Minería de Datos y OLAP, y el sistema resalta la importancia de las aplicaciones habilitadas para Grid en términos de e-ciencia y análisis detallados de grandes conjuntos de datos científicos.

Características:

- GridMiner permite interoperabilidad con diferentes tecnologías de acceso y administración de los datos, desarrolladas por otras comunidades. Sobre esas tecnologías soporta servicios para acceso a los datos, integración de datos, pre procesamiento de datos y computación estadística.

- GridMiner implementa un número de algoritmos de Minería de Datos más comunes, algunos habilitados para ejecución en paralelo, también soporta tareas de minería de texto. La arquitectura del sistema está orientada al servicio y posee un sistema que facilita la integración y ejecución de los servicios como un flujo de trabajo. GridMiner permite la integración de bases de datos distribuidas. Algunos de los servicios de Minería de Datos dados por la herramienta son: Árboles de decisión en su versión distribuida, patrones de secuencia, clasificación de texto en forma paralela utilizando árboles de clasificación, Clústeres utilizando el algoritmo k-means, versiones paralelas y distribuidas de redes neuronales, reglas de asociación, entre otros.

- Administración de seguridad para manipulación y acceso distribuido a los datos y acceso a los servicios.

- Interfaz gráfica diseñada en Java que puede ser ejecutada remotamente por “Java Web Start”, además brinda facilidades de exploración, parametrización, configuración, interacción y ejecución de las tareas de Minería de Datos.

Se puede encontrar más información acerca de la herramienta en la página web http://www.gridminer.org/

4.2.3.7. ADaM toolkit. ADaM (Algorithm Development and Mining) es un grupo de herramientas de minería y procesamiento de imágenes que consisten en componentes interoperables que pueden ser vinculados entre sí en variedades de formas para resolver problemas de diversos dominios. ADaM tiene alrededor de 100 componentes que pueden ser configurados para crear procesos personalizados de minería. Utilidades de pre procesamiento y análisis, permiten a los usuarios aplicar Minería de Datos en problemas específicos. Se pueden añadir fácilmente nuevos componentes para adaptar el sistema a diferentes problemas científicos.

Permite la fácil integración de algoritmos desarrollados por terceros y reutilizar los componentes de ADaM en otros sistemas. ADAM proporciona esto a través de la utilización de componentes autónomos en una arquitectura distribuida. Cada componente cuenta con un programa en C, C++, o interfaz de programación de aplicaciones (API), un archivo ejecutable en apoyo de herramientas genéricas de secuencias de comandos (Perl, Python, Shell Scripts), y eventualmente interfaces Web Service para soportar WEB y aplicaciones Grid. Los componentes de ADaM son módulos de propósito general de minería y procesamiento de imágenes, que pueden ser reutilizados por múltiples soluciones y disciplinas. Estos componentes

están bien posicionados para satisfacer las necesidades de la Minería distribuida y servicios de procesamiento de imágenes en Web y aplicaciones Grid.

Se puede encontrar más información acerca de la herramienta en la página web http://datamining.itsc.uah.edu/adam

4.3. Aplicaciones de la Minería de Datos Grid

Además de las aplicaciones que pueden ser dadas a la Minería de Datos convencional, la Minería de Datos Grid brinda más posibilidades inherentes a la naturaleza de la computación Grid. A continuación, se describen algunas aplicaciones para las cuales puede ser utilizada la Minería de Datos Grid, extraído de los textos de diferentes autores.

4.3.1. Escenarios Empresariales. [BOLLINGER 2004], brinda un buen ejemplo sobre un escenario empresarial donde puede ser aplicada la Minería de Datos Grid. El autor, habla sobre un supermercado de franquicias conformado por su sede principal, regionales y tiendas miembros distribuidas. Cada tienda obtiene sus datos transaccionales mediante el escaneo de códigos de barras cuando los clientes compran los productos. Algunas tiendas dependen de su rama regional para almacenar sus datos transaccionales, mientras otras tiendas tienen sus propias bases de datos locales. Cada regional o tienda miembro posee su propia herramienta de Minería de Datos y la utiliza de forma local. La sede principal, utiliza Minería de Datos para analizar los datos transaccionales de cada tienda sin la necesidad de tener que transmitir y reprocesar todo el conjunto de datos completo.

Este es un caso típico de Minería de Datos basada en la Grid, donde se necesitan compartir los patrones relevantes de las diferentes tareas Minería de Datos para

finalmente generar una estimación global total. Para este enfoque de Minería de Datos, el autor basa su solución utilizando reglas de asociación, una de las técnicas de Minería de Datos. Utilizando algoritmos como Apriori, esta técnica de Minería de Datos es paralelizable, las tareas pueden ser divididas en pequeños trabajos independientes. Estos trabajos independientes pueden ser distribuidos sobre una granja de servidores siendo asignados por un planificador.

4.3.2. Astronomía. [KARGUPYAM 2005] menciona cómo la Minería de Datos distribuida se ha venido convirtiendo en la norma en astronomía. A medida que los datos provenientes de todos los estudios que se hacen del cielo han venido creciendo en gran cantidad, estos ya no pueden ser descargados a un sitio central para realizar análisis mediante Minería de Datos. Por esta razón, los algoritmos de Minería de Datos distribuidos se han convertido en una herramienta esencial para permitir descubrir conocimiento escondido en todas las bases de datos heterogéneas geográficamente distribuidas que poseen información astronómica.

Por ejemplo, las diferentes misiones de la NASA (como 2MASS, WISE, GALEX, SDSS, LSST) han generado millones de datos, los cuales se encuentran distribuidos en diferentes bases de datos heterogéneas. Todos estos datos se requieren caracterizar, clasificar, analizar e interpretar conjuntamente para poder extraer conocimiento escondido relevante sobre el cosmos. Es imposible acceder, minar, navegar, explorar y analizar todos estos datos en su condición distribuida y no es posible generar una única base de datos dada la gran cantidad de información que estas poseen. Por esta razón, los algoritmos de Minería de Datos distribuidos capaces de procesar estos registros de forma paralela y distribuida sobre la grid, tienen gran relevancia hoy en día.

Para enfrentar este problema, existe actualmente un proyecto llamado GRIST (http://grist.caltech.edu/), cuyo objetivo es investigar e implementar maneras para que los astrónomos, científicos, y el público en general puedan contar con

capacidad computacional fácil, potente y distribuida para la diferente información astronómica que manejan. Uno de los servicios que GRIST soporta es el de clasificación k-means. Éste ha sido desarrollado para clasificar grandes catálogos de imágenes, basado en una propiedad dada.

4.3.3. Medicina. [DUBITZKY 2008] menciona una aplicación de la Minería de Datos Grid en la medicina para el manejo de pacientes con traumas cerebrales. Un trauma cerebral generalmente resulta de un accidente donde la cabeza es golpeada por un objeto. Los sobrevivientes de un trauma cerebral pueden quedar significativamente afectados por pérdidas en sus funciones cognitivas, sicológicas o físicas.

En el primer examen de un paciente con trauma cerebral, es muy común asignar al paciente a una categoría, la cual permite planear el tratamiento para el paciente y además permite predecir el resultado final del tratamiento. Hay cinco categorías para el resultado final que pueden ser: muerte, vegetativo, discapacidad grave, discapacidad moderada y buena recuperación. Es obvio que el resultado final está influenciado por varios factores, los cuales son usualmente conocidos y generalmente son monitoreados y almacenados en un almacén de datos por los hospitales.

Es evidente que si se quiere categorizar al paciente, se debe tener un conocimiento previo basado en casos y resultados finales de otros pacientes con el mismo tipo de herida. Este conocimiento puede ser minado de datos históricos y representado como un modelo de clasificación. El modelo puede ser usado para asignar al paciente a una de las categorías de resultado.

Uno de los supuestos básicos en la clasificación es que al considerar un mayor número de casos, la exactitud del modelo final puede ser mejorada. Por esta razón, acceder a los datos de casos similares de trauma cerebral almacenado en

otros hospitales, podría ayudar a crear un modelo de clasificación más exacto. En un ambiente grid, un grupo de hospitales puede compartir sus recursos de datos como registros anónimos de pacientes o algunos datos estadísticos relacionados con la administración de los hospitales.

En este escenario, es necesario enfrentar otros retos como seguridad de acceso a los datos distribuidos, limpieza e integración de los datos y su transformación en un formato adecuado para Minería de Datos. También hay otros aspectos relacionados a este problema como aspectos legales o privacidad de los datos que pueden ser resueltos, especialmente para aquellos registros que contienen datos sensibles de los pacientes.

4.3.4. Industria del turismo. [DANUBIANU 2009], menciona la importancia de la Minería de Datos en el turismo, pues este es uno de los mayores generadores de ingresos aún incluso en épocas de crisis, gracias a la demanda de transporte, alojamiento, comida y bebidas que este genera. Por esta razón, es muy importante tener las estrategias adecuadas las cuales están basadas en las diferentes tendencias que únicamente pueden ser obtenidas a través de análisis sofisticados. Es así, que el autor propone dos modelos de Minería de Datos distribuida que podrían beneficiar grandemente a este sector, al aprovechar las diferentes fuentes de datos con las que cuentan los establecimientos de alojamiento y que pueden ser analizadas en conjunto.

Primero, se parte de la realidad de que cada establecimiento de alojamiento gestiona sus propios datos. Cada base de datos contiene datos sobre los clientes, sobre los servicios solicitados, sobre el dinero gastado, etc. Si cada uno de estos sistemas permite, en termitos de privacidad de los datos, que parte de sus datos sean utilizados para análisis, interconectados a través de una red creando un sistema de base de datos distribuido heterogéneo, es posible realizar proyecciones sobre todos estos datos y extraer conocimiento de forma distribuida aprovechando así los datos de toda la industria del lugar.

Para las pruebas, [Ídem] realizó una instalación en cada sitio de servicios e interfaces adecuadas para Minería de Datos. De esa forma, es posible aplicar métodos de Minería de Datos local en cada sitio (por ejemplo reglas de asociación). Los resultados locales, son replicados en un único nodo donde son combinados para obtener una solución global. Una de las desventajas encontradas en este sistema por el autor, está relacionada con volúmenes de datos pequeños procesados en algunos sitios locales, lo cual puede llevar a resultados parciales inconclusos. La segunda desventaja, es la necesidad de realizar futuras operaciones en el nodo donde los resultados son recolectados para luego calcular el resultado final.

4.3.5. Apoyo avanzado en analítica para e-ciencia. [DUBITZKY 2008] expone la importancia que puede tener hoy en día la Minería de Datos Grid para la e-Ciencia. El término e-Ciencia se refiere a la ciencia a gran escala que cada vez se lleva a cabo a través de colaboraciones globales distribuidas facilitada por Internet. Este fenómeno es una fuerza importante en los temas de investigación actual y programas de desarrollo en general, cada científico como usuario individual requiere que sus empresas científicas posean características tales como el acceso a las colecciones de datos muy grandes y los recursos de computación a escala muy grande. Un componente clave de este desarrollo es la e-ciencia analítica, que es un campo de investigación muy dinámico que incluye métodos científicos rigurosos y sofisticados para pre-procesamiento de los datos, integración, análisis, minería de datos y visualización asociada con la extracción de información y descubrimiento de conocimiento a partir de datos científicos conjuntos.

A diferencia de la analítica de negocio tradicional, la analítica de e-ciencia tiene que tratar con conjuntos de datos grandes, complejos y heterogéneos y a menudo dispersos geográficamente, que contienen volúmenes medidos en terabytes y hasta petabytes. Debido al gran volumen y alta dimensionalidad de los datos, las tareas asociadas a la analítica a menudo son tanto de entrada/salida como de

cálculos intensivos y por lo tanto dependen de la disponibilidad de almacenamiento de alto rendimiento, hardware, recursos software y soluciones de software.

El proceso de descubrimiento de conocimiento en ámbitos científicos, se realiza haciendo uso de datos y recursos distribuidos. Las principales características de este proceso son las siguientes:

- En las diferentes etapas del proceso de descubrimiento de conocimiento, los investigadores necesitan acceder, integrar y analizar datos de fuentes dispersas, con el in de utilizar estos datos para hallar patrones y modelos, y alimentar estos modelos para futuras etapas en el proceso. Existen muchos componentes de software para análisis de datos que pueden ser utilizados para analizar los datos. Algunos componentes software pueden ser utilizados en la máquina local del usuario, mientras otros pueden ser utilizados para ejecución remota en servidores.

- El proceso de descubrimiento en sí mismo es casi siempre realizado por equipos de investigadores que colaboran entre sí y que necesitan compartir los conjuntos de datos, los resultados derivados de estos conjuntos de datos, y, más importante, los detalles acerca de cómo estos resultados se obtuvieron. Dado que todo el proceso de descubrimiento es ejecutable, el usuario final si lo desea, puede empaquetarlo como un programa ejecutable (o componente de software) para el acceso y uso por otros investigadores

- Por último, con un gran número de procesos de descubrimiento que se generan por diferentes grupos de investigación, es esencial ser capaz de almacenar tales procesos dentro de un almacén de procesos, donde los científicos puedan buscar, recuperar y reutilizar los procedimientos desarrollados a partir de un escenario en otros escenarios similares

La Herramienta GridMiner, ha sido desarrollada dentro de un proyecto de investigación en la Universidad of Vienna. El objetivo del proyecto es hacer frente a todas las tareas del proceso de descubrimiento de conocimiento en la grid e integrarlos en una aplicación grid avanzada orientada a servicios. GridMiner consta de dos componentes principales: tecnologías y herramientas, y casos de uso que muestran cómo las tecnologías y herramientas pueden trabajar en conjunto y cómo pueden utilizarse en situaciones reales. La herramienta está basada en el estándar de Minería de Datos CRISP-DM y sus fases como pasos esenciales en los flujos de trabajo científicos orientados a servicios.

4.3.6. Lucha contra desastres naturales. [GUO 2004] habla sobre el sistema Discovery Net como ejemplo de una plataforma para e-Ciencia y que ha sido utilizada para diferentes aplicaciones en campos como ciencias de la vida, monitoreo ambiental y modelado de riesgo geológico. En estos campos pueden encontrarse diversas fuentes de información como diferentes dispositivos, sensores, bases de datos, componentes de análisis y recursos computacionales accesibles a través de Internet o la grid.

Los requisitos de entornos de Minería de Datos para la e-Ciencia hacen que sea imposible el uso tradicional de sistemas cerrados de Minería de Datos que suponen una base de datos centralizada o un almacén de datos. Por esta razón, se requieren plataformas para Minería de datos que permitan la integración de fuentes de datos distribuidas y herramientas distribuidas para las actividades de descubrimiento de conocimiento.

[Ídem], realiza un análisis sobre los siguientes escenarios, los cuales son ejemplos donde la Minería de Datos Distribuida trae grandes beneficios:

a. Científicos colaborando en el análisis diario de nuevos gnomas virales como el virus SARS y estudio de su evolución.
b. Científicos colaborando en el análisis de la contaminación ambiental y correlacionándolos con registros médicos disponibles
c. Científicos colaborando en el análisis de imágenes satelitales para modelar posibles efectos de los terremotos en regiones pobladas

4.3.7. Minería para máquinas mal configuradas en sistemas grid. Dada la heterogeneidad y escalabilidad de los sistemas grid, se hace muy difícil e ineficiente el análisis de su funcionamiento de forma manual. Por esta razón, [DUBITZKY 2008] describe una propuesta para automatizar estos análisis mediante descubrimiento de conocimiento utilizando algoritmos distribuidos.

Los sistemas de Grid son difíciles de manejar. En primer lugar, a menudo sufren más fallas que los otros grandes sistemas: el hardware suele ser más heterogéneo, al igual que las aplicaciones que se ejecutan, y ya que suelen reunir los recursos que pertenecen a varios dominios administrativos, no existe una autoridad única capaz de hacer cumplir las normas de mantenimiento (por ejemplo, con respecto a las actualizaciones de software). Una vez que surgen problemas, la gran complejidad del sistema hace que sea muy difícil seguirles la pista y explicar por qué suceden. El mantenimiento, por lo tanto, se realiza cuando ocurren las fallas y raras veces es preventivo.

Existen dos maneras posibles de automatizar el mantenimiento y asistencia al sistema grid. Una es mediante la construcción de un sistema experto, el cual detecta patrones previamente estudiados de fallas. Por otro lado, se puede aplicar descubrimiento de conocimiento a los datos, tratando de encontrar comportamientos anormales en las máquinas, lo cual puede encontrar errores que aun no han sido estudiados por el sistema experto y que también pueden tener miles de máquinas las cuales pueden estar produciendo megabytes de datos de

monitoreo diariamente. Si todos estos datos se centralizan para ser procesados, probamente sofocarían al sistema y requerirían un almacenamiento más especializado.

El sistema adopta el enfoque de Minería de Datos distribuida para la detección de máquinas mal configuradas. El algoritmo obtiene datos de las fuentes disponibles en el sistema grid. Convierte los datos a un significado semántico y los almacena en la misma máquina de donde provienen. Cuando se requiere un análisis, un algoritmo distribuido de detección de anomalías se emplea para identificar las máquinas mal configuradas. El propio algoritmo se implementa como un flujo de trabajo recursivo de procesos grid y se adapta especialmente a los sistemas grid en los que las máquinas podrían no estar disponibles la mayoría del tiempo, o que a menudo todas las máquinas tienen fallas.

4.4. Oportunidades y retos de la Minería de Datos Grid

Además de todos los retos que debe enfrentar la Minería de Datos como tal y los retos que debe enfrentar la Computación Grid, también surgen algunas oportunidades y retos a los cuales se debe hacer frente cuando se intentan mezclar ambas tecnologías. [KURMAN 2008], explica que hay una pequeña pero muy importante diferencia entre los algoritmos diseñados para sistemas paralelos y algoritmos los diseñados para sistemas distribuidos. Generalmente, los algoritmos de Minería de Datos paralelos hacen frente a sistemas de memoria compartida o sistemas de memoria distribuida con interconexiones rápidas. La Minería de Datos distribuida generalmente hace frente a sistemas conectados a través de redes lentas Ethernet LAN o WAN. La principal diferencia entre los sistemas paralelos y distribuidos son los costos de comunicación, velocidad de interconexión y distribución de los datos. Estas limitaciones, han generado nuevos retos de investigación y desarrollo frente a nuevos sistemas de Minería de Datos.

Los algoritmos de Minería de Datos en grid deben soportar el proceso completo de Minería de Datos (pre procesamiento, Minería de Datos y pos procesamiento) en manera similar que las versiones centralizadas lo hacen. Esto significa, que todas las tareas de Minería de Datos, incluyendo limpieza, selección de atributos, etc, deben poder ser ejecutados de forma distribuida para que esta brinde un verdadero beneficio.

Aunque ya se ha hecho algunos esfuerzos por desarrollar algoritmos distribuidos eficientes de Minería de Datos, los aspectos de los ambientes distribuidos, como planificación y manejo de recursos, son aspectos críticos para el éxito de la Minería de Datos en este tipo de ambientes. Por lo tanto, el desarrollo de herramientas de Minería de Datos, dentro de infraestructuras de alto rendimiento y computación distribuida, es un gran reto para futuros desarrollos en el campo de la Minería de Datos.

[ZAKI 2000], menciona una serie de oportunidades y retos que han venido enfrentando los algoritmos de Minería de Datos en sistemas distribuidos:

- **Alta dimensionalidad**: Los algoritmos de Minería de Datos tienen diferente complejidad y pueden no ser lineales impidiendo así su paralelización y respectiva distribución sobre la grid.

- **Gran tamaño de los datos**: La bases de datos continúan creciendo en tamaño constantemente. La gran mayoría de los algoritmos de Minería de Datos son iterativos y procesan los datos varias veces lo cual puede dificultar la escalabilidad de los algoritmos.

- **Ubicación de los datos**: Hoy en día las bases de datos usualmente están lógica y físicamente distribuida, lo cual requiere un enfoque descentralizado

de la Minería de Datos. La base de datos puede estar particionada horizontalmente donde las transacciones están en diferentes lugares, o tal vez pueda estar particionada verticalmente, con diferentes atributos en diferentes lugares. Los algoritmos de Minería de Datos distribuida generalmente están enfocados para lidiar con un enfoque de partición horizontal y no vertical.

- **Tipo de datos**: La mayoría de investigaciones de algoritmos distribuidos de Minería de Datos se ha enfocado en datos estructurados. Sin embargo, el soporte para otro tipo de datos semi estructurados y no estructurados también es crucial.

- **Sesgo en los datos**: Uno de los problemas que afectan negativamente el equilibrio de carga en los algoritmos de minería de datos en paralelo es la sensibilidad al sesgo de los datos. La mayoría de los métodos consisten en particionar la base de datos horizontalmente en bloques de igual tamaño. Sin embargo, el número de patrones existes en cada bloque puede estar altamente sesgado. Generar los bloques de forma aleatoria puede ser una solución, pero sigue siendo inadecuado.

- **Equilibrio de carga dinámica**: Los algoritmos deben enfrentar el problema de distribuir la carga de procesamiento adecuada a los diferentes nodos de procesamiento, pues en estos puede haber diferentes tipos de usuarios realizando diversos tipos de procesos.

- **Métodos incrementales**: Los datos son constantemente recolectados y actualizados, además, los patrones en estos pueden cambiar constantemente. Hoy en día no existen algoritmos de Minería de Datos incrementales por naturaleza, es decir, que puedan manejar las

actualizaciones sin tener que volver a calcular los patrones o reglas sobre la base de datos.

- **Minería Multi-tabla, diseño de los datos y sistemas de indexación**: Existen un gran reto para realizar Minería de Datos a través de múltiples tablas o bases de datos distribuidas con diferentes estructuras. Tradicionalmente, el realizar minería sobre estas múltiples tablas requiere crear una única gran tabla que contenga todas las tablas. Se requieren mejores métodos que permitan procesar este tipo de datos.

- **Interacción, Gestión de Patrones, Minería meta-nivel**: El proceso de descubrimiento de conocimiento sobre los datos es altamente interactivo, empezando por que el ser humano participa en casi todos los pasos. Por ejemplo, el usuario está altamente involucrado en las fases de entendimiento inicial de los datos, selección, limpieza, transformación. Estos pasos por defecto consumen más tiempo que la misma minería. Además, dependiendo de los parámetros asignados por el usuario, los métodos de minería pueden generar muchos patrones diferentes a ser analizados. Por esta razón, se requieren métodos que permitan realizar consultas multi nivel en los resultados, imponer limitaciones que se enfoquen en los patrones de interés, refinar o generalizar reglas, etc. Los métodos de minería de datos paralela y distribuida pueden ser exitosos en proporcionar la rápida respuesta deseada en todos los pasos anteriores.

5. FACTORES A TENER EN CUENTA PARA DETERMINAR LA VIABILIDAD DE UTILIZAR MINERÍA DE DATOS GRID

Los sistemas de Computación Grid, brindan un potencial muy grande para lograr aprovechar los recursos computacionales que se poseen, con el fin de obtener grandes capacidades de procesamiento y lograr optimizar así los procesos de Minería de Datos. Sin embargo, son pocas las herramientas de Minería de Datos Grid que se pueden conseguir en el mercado que brinden la funcionalidad y usabilidad que se requiere alcanzar en las organizaciones, e incluso, los proveedores de herramientas de Minería de Datos más reconocidos aún no ofrecen técnicas de Minería de Datos Grid en sus soluciones.

Aunque se puedan encontrar algunas herramientas de Minería de Datos que brindan esta funcionalidad grid, aún son muy limitadas y poco robustas, y por lo tanto, es una área que aún requiere mucho trabajo para lograr obtener finalmente resultados tangibles sobre los beneficios que teóricamente se pueden alcanzar, y las facilidades de uso que se requieren para poder obtener un verdadero provecho de esta tecnología.

Sin embargo, las pruebas y algoritmos sobre Minería de Datos Grid que presentan algunos autores son prometedoras, y por esa razón vale la pena determinar y analizar los factores clave que se deben tener en cuenta para implantar o comenzar a realizar pruebas con esta tecnología. A continuación, se realiza un análisis de algunos factores que se deben tener en cuenta para determinar la viabilidad de utilizar Minería de Datos Grid. El análisis, está orientado hacia la evolución de la Minería de Datos a la Computación Grid y los nuevos elementos que surgen cuando ambas tecnologías convergen, y que además se deben tener

en cuenta para poder así realizar una implantación exitosa de esta tecnología. Para esto, hay que analizar tanto aspectos de la Minería de Datos convencional como aspectos de la Computación Grid, con el fin de tener en cuenta los por menores de ambas tecnologías.

Los factores, están basados en [STANKOVSKI 2007], según la perspectiva de arquitectura de la herramienta de Minería de Datos Grid DataMininGrid para implementar aplicaciones de Minería de Datos Grid; [DUBITZKY 2008], según su concepto de Minería de Datos Grid, algunas herramientas de Minería de Datos Grid y algunas aplicaciones expuestas en el libro del autor; autores frente al tema de aplicaciones de Minería de Datos Grid como [DANUBIANU 2009], [KARGUPYAM 2005], [BOLLINGER 2004], [GUO 2004]; autores frente al tema de técnicas de Minería de Datos Grid como [SARNOVSKÝ 2009], [MELIGY 2009], [KURMAN 2008], [SAKTHI 2008], [GLIMCHER 2007], [DEPOUTOVITCH 2005], [CANNATARO 2004], [PETHICK 2003]; autores frente al tema de Computación Grid como [WEGENER 2008], [STANKIVSKY 2008], [GENTZSCH 2007], [KRAVTSOV 2006], [OPIYO 2005], [JACOB 2005]; y en general, la mayoría del material de los autores utilizados para construir el estado del arte de la Minería de Datos, Computación Grid y Minería de Datos Grid expuesto en este documento.

El análisis de los factores, se da teniendo en cuenta las experiencias y sugerencias dadas por los autores evaluando algoritmos de Minería de Datos distribuidos, aplicaciones, frameworks, middlewares y herramientas tanto de Minería de Datos Grid como de Computación Grid; y además, mis propias experiencias en el uso de la tecnología de Minería de Datos en una entidad financiera. Todo, frente a los elementos que se necesitan para poder realizar una Minería de Datos exitosa.

5.1. Factores frente al proceso general de la Minería de Datos

Hoy en día, existen algunos estándares que definen el proceso que se debe seguir para realizar Minería de Datos. Los estándares más reconocidos son CRISP-DM y SEMMA, y sobre estos están construidas las herramientas de Minería de Datos más reconocidas en el mercado que son Clementine (CRISP-DM) y Sas Enterprese Miner (SEMMA). Las organizaciones que tienen procesos de Minería de Datos maduros, suelen trabajar bajos estos estándares, por este motivo, uno de los factores muy importantes a la hora de implantar Minería de Datos Grid, es que *se debe asegurar que la herramienta de Minería de Datos Grid a utilizar, también cumpla con algún estándar de Minería de Datos*, especialmente aquel sobre el cual la organización ha venido trabajando.

Otro factor muy importante expuesto por algunos autores, es que si se quiere poder obtener un verdadero provecho de la Minería de Datos Grid, l*os usuarios técnicos y finales deben poder realizar las tareas de Minería de Datos sin la necesidad de entender el detalle de los aspectos de la tecnología Grid sobre la cual corren los procesos.* Es decir, deben poder ejecutar cada uno de los pasos del proceso de Minería de Datos bajo cualquier estándar, sin la necesidad de conocer a fondo la tecnología grid que soporte los procesos. Si la herramienta de Minería de Datos Grid no abstrae lo suficientemente bien los aspectos de la grid, pueden surgir nuevas dificultades y complejidades de uso que impidan obtener verdaderos beneficios.

Sin embargo, por muy transparente que logre ser el sistema grid para el usuario de la herramienta, sí hay algunos factores muy importantes que se deben tener en cuenta en cada una de las etapas de la Minería de Datos, para lograr tener un control efectivo y obtener resultados satisfactorios a la hora de utilizar Minería de Datos Grid:

- **Conocer el negocio y el problema:** Algunos autores que explican diferentes algoritmos de Minería de Datos Grid, mencionan que *al distribuir la estimación de los modelos de Minería de Datos se puede perder algo de confiabilidad en sus resultados; aunque se han desarrollado técnicas para ajustar los resultados finales de los algoritmos de Minería de Datos distribuidos luego de haber realizado la estimación global, aun existen riesgos de perder exactitud*.

  Por esta razón, al tratar de definir si utilizar Minería de Datos centralizada o grid, surgen cuestiones sobre calidad de los resultados vs eficiencia de ejecución. Por este motivo, es importante tener muy claro cuál es el problema que se quiere solucionar, cuál es el alcance de la Minería de Datos frente a esta solución y cuál es la importancia de obtener resultados a tiempo frente a las decisiones finales que se tomarán frente a estos.

- **Reunir los datos:** Otro factor clave para determinar si utilizar Minería de Datos centralizada o grid, es la dimensión del problema (cantidad de datos a procesar). *Si la cantidad de datos a procesar que requiere la solución al problema no es lo suficientemente grande y éste se distribuye a través de diversos nodos en la grid, no habrá suficientes datos en cada nodo para realizar una estimación de patrones adecuada sobre el conjunto de datos*.

  En esta etapa, es muy importante determinar factores de cómo son generados los datos, cuál es su calidad, cómo pueden reunirse para ser utilizados, su ubicación (si están distribuidos o no en diferentes bases de datos u ubicación geográfica) y la cantidad promedio de datos a procesar para generar y ejecutar los modelos. Todo esto, con el fin de

dimensionar claramente el problema y determinar cuál solución de Minería de Datos sería la más adecuada, si centralizada o grid.

- **Pre procesar los datos:** Esta es una etapa esencial en la Minería de Datos, pues en esta se asegura la calidad de los datos. Aunque hoy en día existen diversas técnicas para limpieza de los datos, si la cantidad de datos a procesar es excesivamente grande, también se puede dificultar generar procesos que aseguren datos confiables. Frente a esto, el poder computación del sistema grid también puede ser utilizado para realizar una limpieza de datos colaborativa, acelerando así en cierta medida este proceso que puede consumir mucho tiempo de procesamiento. Frente a este tema, es bueno investigar algoritmos y técnicas de limpieza de datos paralelizables que puedan ser utilizados en la grid e integrados con la herramienta de Minería de Datos Grid que se posee, si esta no cuenta con esta capacidad.

- **Estimar el modelo:** En esta etapa se seleccionan los algoritmos de Minería de Datos a utilizar más apropiados que respondan a las necesidades y complejidades del problema. Algo muy importante a tener en cuenta, es que generalmente los algoritmos de Minería de Datos convencionales poseen varios parámetros que afectan su rendimiento y exactitud frente a las estimaciones que el algoritmo realice, y además se deben realizar diferentes pruebas con diferentes parámetros para lograr así determinar de qué forma el algoritmo arroja mejores resultados.

  Al trabajar en un sistema Grid, surgen nuevas necesidades de parametrización de acuerdo a la naturaleza de los datos, de la grid y del problema a enfrentar. Por ejemplo, se puede tener la necesidad de parametrizar el algoritmo para que únicamente se ejecute en determinadas máquinas en la grid, o se puede querer realizar pruebas

para analizar con qué número de recursos el algoritmo ofrece mayor eficiencia frente a rendimiento y exactitud en sus resultados, el algoritmo puede recibir parámetros para utilizar diferentes técnicas con el fin ajustar el modelo global final y mejorar su exactitud.

Es así que, por más transparente que sea el sistema grid frente al proceso de Minería de Datos utilizando una herramienta de Minería de Datos grid, un factor muy importante a tener en cuenta es que *es bueno contar con usuarios que tengan conocimiento tanto de Minería de Datos como de tecnología Grid, pues estos pueden realizar mejores elecciones frente a algoritmos, parámetros, flujos de trabajo y preferencia sobre recursos Grid para ejecutar algunos modelos particulares de Minería de Datos.*

- **Interpretación y despliegue del modelo:** Una vez construido el modelo de Minería de Datos, hay que determinar cómo será este desplegado en el negocio y cómo mostrará sus resultados a los usuarios finales para que éstos puedan interpretarlos. Una cosa es construir el modelo mediante técnicas de Minería de Datos Grid (etapa de estimar el modelo), y otro asunto es habilitar el modelo para que este sea ejecutado y realice las estimaciones sobre los registros sobre los cuales realizará la estimación y genere los resultados. Este proceso es paralelizable, puesto que el modelo realiza la estimación sobre un único registro, sin la necesidad de conocer el conjunto de datos completo, por esto, esta tarea puede ser distribuible en la grid.

  Antes de migrar a Minería de Datos Grid, es importante revisar todos los modelos actuales de Minería de Datos con los cuales se cuenta, con el fin de determinar cuáles son susceptibles de utilizar en la grid a miras de optimizar su funcionamiento y cuáles es necesario construir nuevamente

para habilitarlos en la herramienta de Minería de Datos Grid Por esta razón, un factor importante es que *los usuarios técnicos deben estar en la capacidad de habilitar los modelos existentes en la organización para funcionar en la Grid (de manera práctica, los modelos actualmente existentes desarrollados con Minería de Datos convencional son fácilmente distribuibles. Se aplicaría la estimación de manera distribuida para la necesidad de crear nuevos modelos o reentrenar los existentes).*

Un aspecto muy importante frente a la generación y despliegue del modelo de Minería de Datos en un ambiente grid, es que los datos que se utilizarán para generar o ejecutar el modelo se distribuirán en varias máquinas pertenecientes a la grid, por esta razón, un factor importante es *asegurar control de acceso y políticas de seguridad adecuadas sobre los datos que se acomoden a la infraestructura Grid implementada y a la confidencialidad requerida de la información de cada una de las áreas.* Existen áreas que manejan información muy confidencial de sus procesos de negocio y que probablemente no quieran que sus datos viajen y sean procesados en recursos pertenecientes a otras áreas que también hagan parte de la grid. Es muy probable, que este tipo de áreas prefieran que algunos de sus modelos que manejan información más crítica y confidencial, sean ejecutados únicamente en recursos de su misma área pertenecientes a la infraestructura grid que se posea para realizar Minería de Datos.

5.2. Factores frente a las herramientas de la Minería de Datos Grid

Actualmente, es difícil encontrar herramientas de Minería de Datos que ofrezcan la funcionalidad de sus algoritmos sobre la grid. Aunque se pueden encontrar algunas herramientas y frameworks de Minería de Datos Grid que han venido

siendo desarrollados y probados, aún se requiere mayor trabajo para que la Minería de Datos Grid sea tan intuitiva y fácil de utilizar como lo han logrado algunos proveedores de herramientas de Minería de Datos convencionales. Sin embargo, es bueno analizar algunos aspectos que han permitido que algunas herramientas de Minería de Datos estén teniendo tanto éxito en las organizaciones y qué factores surgen cuando la Minería de Datos se sumerge en la grid y que también deben ser tenidos en cuenta para poder utilizar satisfactoriamente Minería de Datos en este ambiente.

Un factor muy importante es, que si se quiere obtener un verdadero provecho de la Minería de Datos, *la herramienta a utilizar debe brindar posibilidades para crear flujos de trabajo que cubran todas las etapas de la Minería de Datos.* La herramienta debe ajustarse a algún estándar de Minería de Datos, permitiendo ser utilizada desde la etapa de conocimiento del problema y del negocio hasta despliegue de los modelos y obtención e interpretación de los resultados. *También es importante, que la herramienta brinde una amplia variedad de algoritmos de Minería de Datos en su versión distribuida, que puedan utilizarse tanto de forma distribuida como centralizada, y que además sea fácilmente extensible para la implementación de nuevos algoritmos.*

La Minería de Datos, tiene una alta complejidad debido a sus algoritmos y conocimientos que se requieren para utilizarlos. Sin embargo, hoy en día existen herramientas como Clementine y Sas Enterprise Miner, que han logrado abstraer la complejidad de la Minería de Datos y la han vuelto más intuitiva y manejable para los usuarios, e incluso ajustándose a estándares de Minería de Datos, permitiendo así obtener mejores beneficios de esta. Por este motivo, un factor clave frente a la Minería de Datos Grid, es que *definir y realizar los procesos de Minería de Datos en la grid debe ser amigable y sencillo.* La herramienta de Minería de Datos Grid a utilizar, debe lograr abstraer de forma satisfactoria para los usuarios, tanto la complejidad de la grid como la complejidad de los algoritmos,

manipulación y transformación de los datos que se requieren para realizar Minería de Datos. Si diseñar los modelos en la herramienta es difícil, poco intuitivo y limitado, y además el usuario tiene que enfrentarse a las complejidades que surgen en la grid, los procesos de Minería de Datos pueden volverse tediosos opacando así los beneficios que puede brindar esta tecnología.

Algunos otros factores que se deben tener en cuenta en la herramienta, frente a la abstracción de la complejidad de la grid, son:

- *El sistema grid debe ser extensible sin la necesidad de forzar a todos los usuarios a instalar continuamente actualizaciones.* Además, se debe tener especial cuidado de que dichas actualizaciones no afecten el funcionamiento de los modelos que ya han sido creados y que se encuentran actualmente desplegados en la grid.

- *El sistema Grid debe ser escalable, debe permitir añadir sin ningún problema nuevos recursos que se acomoden al incremento de usuarios y demanda de procesamiento sin perder rendimiento.* Inevitablemente los negocios siguen creciendo y la necesidad de tener más poder computacional no se queda atrás. Si el sistema Grid no es escalable, con toda seguridad surgirán nuevamente problemas de rendimiento y cuellos de botella en los procesos de Minería de Datos.

- *La herramienta debe permitir utilizar fácilmente el conocimiento técnico de algunos usuarios para definir, configurar y parametrizar los detalles de la herramienta frente a la Minería de Datos y la Grid.* Si la herramienta no es lo suficientemente flexible frente a su configuración, difícilmente podrá ser adaptada a las necesidades del negocio.

- *Los procesos de Minería de Datos deben ser ejecutados en las máquinas adecuadas en la grid sin la necesidad de interacción directa con el usuario.* Por esta razón, sería bueno que la herramienta brinde opciones para definir procesos de Minería de Datos críticos y confidenciales, para que estos sean ejecutados automáticamente en segmentos más seguros de la grid.

Otro factor clave que se debe tener en cuenta es que *la herramienta de Minería de Datos Grid debe ser la apropiada, esta no debe requerir modificaciones de su código fuente para ser adaptada a las necesidades del negocio específicas.* Principalmente, porque las necesidades del negocio constantemente pueden cambiar, y si no se cuenta con una herramienta de Minería de Datos lo suficientemente flexible, que sea capaz de adaptarse fácilmente a las necesidades del negocio sin la necesidad de cambiar su estructura y diseño, seguramente será muy costosa de mantener y actualizar en el tiempo.

Por esta razón, otro factor muy importante es que se debe *evaluar cuidadosamente todas las restricciones que se presenten con las herramientas de Minería de Datos Grid disponibles en el mercado, para lograr determinar cuál es la más adecuada para su implantación*. No quedarse con la primera que se encuentre, sino, tomarse el tiempo para mirar los pros y contras de cada una de las herramientas y lograr determinar así cuál es la que mejor se ajusta a las características propias del negocio. Otro factor clave frente a esto, es *tener en cuenta a todos los usuarios de Minería de Datos para evaluar si la solución que se vaya a implantar de Minería de Datos Grid sí cumple con todas las necesidades específicas de cada usuario.* Especialmente, si la organización ya cuenta con procesos de Minería de Datos maduros y cuenta con diversos modelos y aplicaciones de minería inmersas en sus procesos de negocio.

Hay algunos otros factores adicionales importantes para evaluar la viabilidad de utilizar o no una herramienta de Minería de Datos Grid, que se deben tener en

cuenta y que son mencionados por algunos autores. Estos factores surgen dada la naturaleza de la tecnología grid:

- *La Herramienta debe ser evaluada para asegurarse que provea un uso adecuado de los recursos grid, buen rendimiento en los algoritmos y alta escalabilidad.* La herramienta no solo debe enfrentar los aspectos subyacentes de la Minería de Datos, sino que también debe ser robusta frente a todos los aspectos existentes en la Computación Grid.

- *El sistema de Minería de Datos Grid debe acoplarse satisfactoriamente a la heterogeneidad de todos los recursos computacionales, sistemas operativos, redes y diferentes fuentes de datos que se utilizan para realizar Minería de Datos.* Si no hay un buen acoplamiento, no se podrá extraer ningún beneficio adicional de rendimiento utilizando los recursos computacionales que posee la organización.

- *Se debe asegurar que los usuarios tengan la capacidad de monitorear el progreso de sus modelos en la Grid para poder responder con acciones apropiadas.* Si los procesos de Minería de Datos fallan en la grid, es muy importante que los usuarios puedan determinar fácilmente cuáles son los errores, para reportarlos y poder tener en funcionamiento los modelos de Minería de Datos sin ningún contratiempo.

- *La solución Grid debe estar basada en estándares relacionados a Computación Grid y en tecnología abierta ampliamente usada.* Actualmente, existen tecnologías de Computación Grid ampliamente reconocidas y utilizadas. Es bueno que la herramienta de Minería de Datos Grid, utilice este tipo de tecnologías como soporte para su funcionamiento en la grid.

5.3. Factores frente a las técnicas de la Minería de Datos Grid

Hasta ahora, se han desarrollado y evaluado una amplia gama de algoritmos de Minería de Datos distribuidos obteniendo muy buenos resultados. Sin embargo, dada la naturaleza heterogénea de los componentes que conforman una grid, *inicialmente se recomienda realizar pruebas sobre los algoritmos de Minería de Datos Grid vs su versión centralizada, con el fin de medir su eficiencia y efectividad y así lograr determinar cuál es la mejor solución para cada modelo en específico*. Existen muchos factores tecnológicos que pueden limitar la eficiencia de algunos algoritmos en la grid y que pueden afectar los resultados finales arrojados por los modelos. Por esta razón, en un principio, es bueno realizar pruebas y para determinar que ajustes se requieren hacer tanto sobre la herramienta de Minería de Datos como en los componentes del sistema grid como planificadores, recursos computacionales, comunicación entre los recursos y topología de la grid, todo con el fin de generar un ambiente óptimo donde los algoritmos de Minería de Datos distribuidos ofrezcan mayor rendimiento.

También se debe asegurar que *la infraestructura grid esté adecuadamente construida, de tal forma que se acople bien a los procesos distribuidos de los diferentes algoritmos de Minería de Datos que requieran comunicación e interdependencia para generar resultados.* Los algoritmos pueden generar diferentes cargas de procesamiento y tener diferentes necesidades de comunicación entre los nodos para realizar sus resultados de manera distribuida. Si no se asegura una infraestructura adecuada de la grid, seguramente no se obtendrá el crecimiento deseado en cuestiones de rendimiento.

Otro factor importante que surge al utilizar tecnología grid para Minería de Datos, es que también se deben *generar estrategias que permitan evaluar el rendimiento de los algoritmos de planificación utilizados por el sistema grid y determinar la mejor configuración que brinde mayor eficiencia a los algoritmos de Minería de*

Datos distribuidos. Dado que el sistema grid es escalable y las necesidades de realizar Minería de Datos crecen constantemente, es importante realizar seguimiento continuo al rendimiento de la grid y definir si esta requiere mantenimiento o mejoras que permitan mantener el buen rendimiento de los algoritmos de Minería de Datos distribuidos.

5.4. Factores frente a las aplicaciones de la Minería de Datos Grid

Uno de los grandes beneficios frente a las aplicaciones, que se obtienen al utilizar Computación Grid para realizar Minería de Datos distribuida, es que en la Minería de Datos centralizada las aplicaciones están limitadas por las grandes cantidades de datos y capacidad de procesamiento de las máquinas, lo cual limita enormemente la escalabilidad de las aplicaciones. Con grid, se obtiene mayor escalabilidad frente a los datos y capacidad de procesamiento, y permite la estimación de modelos de forma colaborativa, e incluso, inter organizacional.

Con Minería de Datos Grid, diferentes organizaciones pueden compartir sus modelos de Minería de Datos sin la necesidad de compartir sus datos sensibles, con el fin de realizar estimaciones globales y ajustes sobre sus modelos particulares y así obtener modelos de Minería de Datos más ajustados a lo que realmente está ocurriendo en el entorno.

Sin embargo, *es importante entender que no todas las aplicaciones pueden ser transformadas para ejecutar en paralelo en una grid y alcanzar esa escalabilidad deseada.* Por esta razón, puede haber algunos algoritmos y aplicaciones de Minería de Datos con los que cuente la organización que no podrán ser distribuibles a través de la grid, y por lo tanto, tendrán que seguir siendo utilizados de forma centralizada.

Esta limitación, para nada quita los beneficios que puede brindar la Minería de Datos Grid a la organización. Siempre y cuando en el mundo de algoritmos de Minería de Datos que la organización usa, ésta utilice algoritmos de Minería de Datos que puedan ser distribuibles en la grid, y que además necesiten de alto rendimiento en tiempo para arrojar los resultados, la Minería de Datos Grid sigue siendo una buena opción.

5.5. Factores técnicos de la Minería de Datos Grid

Además de todos los factores anteriormente mencionados, existen algunos aspectos técnicos que son importantes tener en cuenta a la hora de querer utilizar Minería de Datos en la grid, con el fin de lograr obtener recursos y procesos óptimos que permitan asegurar la permanencia y buen rendimiento del servicio. Al migrar la Minería de Datos a la grid, ya no solamente están los aspectos técnicos referentes a la Minería de Datos, sino también todos los detalles técnicos de la Computación Grid.

A continuación, se describen algunos de estos factores técnicos:

- *Se debe contar con los recursos computacionales adecuados, tanto para los planificadores de la Grid como para sus correspondientes nodos de procesamiento.* Sin duda alguna, este es un factor importante para determinar si es viable o no implementar esta tecnología, pues si los nodos de procesamiento son máquinas de usuarios que están siendo utilizadas continuamente, y estas máquinas en ningún momento poseen recursos ociosos de procesamiento; o al enviar los procesos por medio de la grid a las máquinas, estos afectan el desempeño de los usuarios que las utilizan, definitivamente no se cuenta con los recursos adecuados.

- *Se debe contar con buena conectividad y buen rendimiento de la red que interconecta los nodos en la Grid. La infraestructura de red debe ser óptima para que permita obtener un mejor rendimiento en la velocidad de las comunicaciones.* La Computación Grid, realiza un uso intensivo de la red para enviar los diferentes procesos a los diferentes nodos de procesamiento, y de igual manera enviar el resultado final de los nodos de procesamiento hacia el nodo central para unificar los resultados. También, pueden existir procesos que requieren comunicarse entre sí en los diferentes nodos de procesamiento para poder realizar así sus cálculos. Por tal motivo, si la red no es óptima, finalmente no se verán buenos resultados de rendimiento. Por este motivo, se debe *realizar pruebas para determinar si la red puede soportar el procesamiento intensivo de datos que realizan los algoritmos de Minería de Datos y que no vayan a resultar problemas de eficiencia por causa de la red.* Una posible estrategia que se puede utilizar para optimizar las comunicaciones en la red, es *buscar que los nodos en la grid se encuentren lo más cerca posible de los datos para eliminar costos en tiempo de transferencia.*

- También, se deben *definir la configuración y políticas más adecuadas sobre los planificadores, para que los procesos se ejecuten en el lugar más adecuado de acuerdo a las solicitudes de procesamiento.* Como se mencionaba anteriormente, pueden existir algunos procesos que requieran ser ejecutados en secciones de la grid específicas y no en cualquier máquina de la grid, por razones de seguridad o confidencialidad de la información. O pueden existir procesos que requieran una configuración de hardware específica para procesar los datos y arrojar sus resultados. Pueden existir muchos aspectos diferentes que deben ser evaluados detalladamente para lograr determinar así las configuraciones más adecuadas.

- Dada la criticidad que pueden llegar a tener los procesos de Minería de Datos en la organización, es importante asegurar que los servicios estén continuamente funcionando y sin problemas. Por esta razón, *debe existir un monitoreo adecuado de la Grid que permita conocer quién la está usando y cómo, con el fin de medir su rendimiento para solucionar los posibles problemas a tiempo.*

- Dada la complejidad técnica de la grid, es importante que este componente sea administrado principalmente por TI, como soporte tecnológico para optimizar los procesos de Minería de Datos en las diferentes áreas. Por esta razón, *Se debe garantizar el servicio y soporte por parte de TI para realizar ajustes y nuevos desarrollos sobre el sistema Grid y solucionar problemas eventuales que se puedan presentar.* Además, *Se recomienda buscar personas con experiencia en tecnologías Grid, que ayuden a estructurar la infraestructura más adecuada y dar soporte a las necesidades específicas de Minería de Datos.* También, *se debe contar con recursos que realicen periódicamente mantenimientos a la grid y recursos compartidos en esta, de tal forma que se asegure su buen rendimiento*

5.6. Factores frente a los retos de la Minería de Datos Grid

Finalmente, otros factores muy importantes para determinar si es viable o no utilizar Minería de Datos Grid en la organización, son todos los retos que se deben enfrentar y que se presentan en la Minería de Datos, la Computación Grid y cuando ambas tecnologías convergen en la Minería de Datos Grid, los cuales son mencionados en los numerales 2.4, 3.5 y 4.4 de este documento. Se debe evaluar hasta dónde la organización está preparada para enfrentar todos estos factores, qué tan maduros son sus procesos de Minería de Datos y sus procesos

tecnológicos, con el fin de estructurar adecuadamente la infraestructura que se requiere para Minería de Datos Grid y lograr obtener verdaderos beneficios.

6. CONCLUSIONES Y TRABAJOS FUTUROS

La idea para este proyecto de grado nació dada la experiencia que he obtenido al trabajar con herramientas de Minería de Datos, especialmente Clementine, en una entidad financiera. Desde allí, pude observar cómo al interior de la organización, aquellas personas que no cuentan con herramientas de Minería de Datos y que requieren analizar los datos y tomar decisiones, se les dificulta en gran manera el realizar esta labor. Desde ese instante reconocí la importancia de la Minería de Datos.

En la actualidad, las organizaciones para sobrevivir en el mercado deben ser altamente competitivas. Para ser competitivas, deben saber obtener el mejor provecho de sus recursos y ser muy ágiles a la hora de tomar decisiones frente a todo lo que les ocurre día a día y minuto a minuto. Para poder tomar decisiones certeras, es necesario contar con datos oportunos acerca del negocio y con la capacidad de analizar estos datos para extraer conocimiento que no es deducible a simple vista. Con la gran cantidad de datos que generan los negocios, cada vez es más difícil procesar y analizar estos datos de forma manual. Por esta razón, las técnicas de Minería de Datos han adquirido gran importancia hoy en día como factor competitivo para las organizaciones.

Existen gran cantidad de técnicas de Minería de Datos que permiten mejorar e incluso automatizar algunos análisis deductivos o predictivos que cualquier ser humano pueda hacer. Existen muchas herramientas licenciadas y de código libre para realizar Minería de Datos, la gran mayoría de ellas abstrae la complejidad matemática y estadística de los algoritmos permitiendo así a los usuarios realizar análisis avanzados sobre los datos sin la necesidad de contar con estos

conocimientos. También hay herramientas de Minería de Datos que ofrecen interfaces gráficas de modelado de los datos que permiten manipular enormes cantidades de datos de una manera muy sencilla y amigable. Por estas razones, la Minería de Datos hoy en día ha obtenido gran popularidad en las empresas.

Sin embargo, la labor de Minería de Datos requiere disciplina y buen entendimiento del negocio si se quiere obtener buen provecho de ella, de lo contrario, si no se adquieren buenas prácticas y no se dedica el tiempo necesario a planear y estudiar las técnicas que brinda esta tecnología, la Minería de Datos puede generar gran cantidad de basura que finalmente no dirá nada relevante acerca del negocio y por el contrario hará más difícil la toma de decisiones. Si se conocen las técnicas y el negocio, será más fácil sacar provecho y es allí donde se nota la verdadera importancia y relevancia de la Minería de Datos. Por este motivo, el capítulo dos: “Conceptos y aplicaciones de Minería de Datos”, tiene gran relevancia, pues para entender la importancia de solucionar algunos problemas y limitaciones que comienzan a surgir para la Minería de Datos, primero es importante entender muy bien cuál es la importancia de la Minería de Datos hoy en día y por qué vale la pena profundizar e investigar aún más en este tema.

Uno de los principales problemas que comienzan a tener que enfrentar los algoritmos de Minería de Datos convencionales, es la labor de procesar la gran cantidad de datos que crece exponencialmente en las organizaciones. Esta gran cantidad de datos hace que el rendimiento de los algoritmos decaiga notablemente. Es un problema que he tenido que enfrentar personalmente en mi experiencia, especialmente en ciertos días críticos del mes donde se aumenta mucho las operaciones de los clientes haciendo que los equipos donde se ejecutan los procesos de Minería de Datos, casi que literalmente, revienten por causa de la alta cantidad de registros que se requieren procesar. Esta dificultad me motivó a investigar y a encontrar técnicas de Minería de Datos colaborativa utilizando la tecnología de Computación Grid, la cual permite aprovechar diversos

recursos computacionales para procesar los datos de una manera más rápida y eficiente.

En un principio, fue difícil encontrar información relevante acerca de Minería de Datos en ambientes Grid. Se puede notar que aún es un área que requiere mucha investigación, pues la labor de modificar los algoritmos de Minería de Datos actuales para que funcionen en versión distribuida es una tarea que se ha venido desarrollando desde hace poco tiempo. Aunque hay varias propuestas de algoritmos distribuidos para Minería de Datos, son escazas las herramientas que se pueden conseguir que brinden esta funcionalidad.

Aunque actualmente existen varias herramientas de Minería de Datos que desarrollan el concepto de la grid, algunas ya no podían ser descargadas y otras funcionan únicamente en ambientes tipo UNIX y aún el concepto de grid que manejan tanto para generación y ejecución de los modelos no está desarrollado completamente. Esta es una de las mayores pruebas de que aún se requiere más investigación en el tema y sería un área interesante en la cual invertir recursos dada la importancia que tiene hoy la Minería de Datos. Es importante también, analizar detalladamente las propuestas de algoritmos de Minería de Datos distribuidos existentes y verificar sus posibilidades de implementación para realizar pruebas y así poder analizar su eficiencia frente a los algoritmos de Minería de Datos convencionales que se utilizan actualmente en las organizaciones.

Se podría pensar que con la evolución del Hardware, estos problemas de rendimiento se irán reduciendo con el tiempo y en el futuro éste estará preparado para recibir el procesamiento intensivo que requiere la Minería de Datos y para enfrentar el crecimiento exponencial de la cantidad de datos que generan y recogen los negocios por su crecimiento. Sin embargo, es importante recordar que este tipo de Hardware generalmente es muy costoso y siempre tendrá sus límites de capacidad.

Hoy en día, las tecnologías emergentes de Computación Distribuida están obteniendo muy buen renombre en las organizaciones, gracias a la capacidad que tienen de aprovechar todos los recursos tecnológicos con los que cuenta la organización (incluso sin importar su distribución geográfica) para mejorar la eficiencia en los procesos que requieren recursos de procesamiento reduciendo así los costos en Hardware.

La Minería de Datos Grid, es una buena solución para superar los problemas de rendimiento que se pueden presentar con los algoritmos de Minería de Datos convencionales que existen. Teóricamente, hay buenas posibilidades para su implementación y desarrollo, aunque aún requiere mucha investigación y la generación de herramientas que permitan extraer verdadero valor para la organización.

Se propone como desarrollos futuros a este trabajo, realizar investigaciones más profundas sobre métodos para generar algoritmos de Minería de Datos distribuidos eficientes, que realmente hagan un buen provecho de la Computación Grid, u otras técnicas de Sistemas Distribuidos que permitan enfrentar también este problema del procesamiento de una forma efectiva y eficiente. También, se propone comenzar a estudiar el desarrollo de interfaces de usuario amigables para el desarrollo de Modelos de Minería de Datos Grid que soporten todo el proceso de Minería de Datos de forma distribuida y transparente para los usuarios. Además, validar la opción de implementar extensiones que brinden funcionalidad de algoritmos de Minería de Datos en Grid para las herramientas comerciales existentes. También, se utilizará este trabajo con el fin de construir una propuesta para una entidad financiera que cuenta con procesos de Minería de Datos convencional en varias de sus áreas, con el fin realizar un análisis de viabilidad para evaluar qué tan factible es incorporar esta tecnología en la organización, con el fin de optimizar los procesos de Minería de datos con los que cuenta.

7. BIBLIOGRAFÍA

[ABBAS 2004] ABBAS, Ahmar. Grid Computing: A Practical Guide to Technology and Applications. 1ª Edición. Charles River Media. 2004. ISBN: 1584502762.

[ALKADI 2007] ALKADI, Ihssan. Grid Computing: The Trend Of The Millenium. University of Louisiana at Lafayette. 2007.

[AWAD 2009] AWAD, M. Design and Implementation of Data Mining Tools. 1ª Edición. Boca Raton, FL. Auerbach Publications. 2009. ISBN 9781420045901.

[CANNATARO 2004] CANNATARO, Mario. Distributed Data Mining on Grids: Services, Tools, and Applications. 2004.

[CUNHA 2006] CUNHA, Jose C. Grid computing: software environments and tools. 1ª Edición. London. Sprinter. 2006. ISBN: 9781852339982.

[DEPOUTOVITCH 2005] DEPOUTOVITCH, Alex, WAINSTEIN, Alex. Building Grid-Enabled Data-Mining Applications. Consultado Febrero 2010. http://www.ddj.com/database/184406345?pgno=1

[Dr. BURKE 2008] Dr. BURKE, James. The Emergence of Grid Computing. Consultado Febrero 2010. http://www.prudens.com/patens/compswinfra/gridcomp.html

[DUBITZKY 2008] DUBITZKY, Werner. Data Mining Techniques in Grid Computing Environments. 1ª Edición. 2008. River Street, Hoboken, NJ. ISBN: 9780470512586.

[GENTZSCH 2007] GENTZSCH, Wolfgang. Grid Initiatives: Lessons Learned and Recommendations. Consultado Junio 15 de 2010. http://.ogf.org/UnderstandingGrids/documents/Grid_Initiatives_July_12_2007.pdf

[GRAY 2005] GRAY, JIM. Data Mining Practical Machine Learning Tools and Techniques. Microsoft Research, 2ª Edición. San Francisco, CA. 2005. ISBN: 0120884070.

[HAN 2006] HAN, Jiawei. Data Mining Concepts and Techniques. 2ª Edición. Morgan Kaufmann. 2006. ISBN: 9781558609013.

[HAND 2007] HAND, David. Intelligent Data Analysis. 2ª Edición. Springer. 2007. ISBN 9783540430605.

[HORNICK 2007] HORNICK F. Mark. Java Data Mining. 1ª Edición. Jim Gray, Microsoft Research. ISBN: 9780123704528.

[JACOB 2005] JACOB, Bart, BROWN, Michael, FUKUI, Kentaro, TRIVEDI, Nihar. Introduction to Grid Computing. IBM. 2005.

[KANTARDZIC 2003] KANTARDZIC, Mehmed. Data Mining Concepts, Models, Methods. 1ª Edición. John Wiley & Sons. 2003. ISBN: 0471228524.

[KARGUPYAM 2005] KARGUPYAM, Hillol. Data Mining: Next Generation Challenges and Future Directions. 1ª Edición. MIT Press. 2005. ISBN: 9780262612036.

[KRAVTSOV 2006] KRAVTSOV, Valentin. Service-based Resource Brokering for Grid-Based Data Mining. 2006.

[KURMAN 2008] KURMAN, Vipin. Next Generation of Data Mining: High-Performance Distributed Data Mining. 1ª Edición. Boca Ratón: Chapman & Hall/CRC. 2008. Capítulo 8 parte 2 pág 152-168. Serie: 9781420085.

[KURMAN 2009] KURMAN, Vipin. Top Ten Algorithms in Data Mining. 1ª Edición. Chapman & Hall/CRC. Boca Raton, FL. 2009. Serie: 9781420089646.

[LAROSE 2006] LAROSE, Daniel. Data Mining Methods and Models. 1ª Edición. John Wiley & Sons. 2006. ISBN: 9780471666561.

[MAGOULES 2009] MAGOULES, Frederic. Introduction to Grid Computing. Applications. 1ª Edición. Chapman & Hall. 2009. ISBN: 9781420074062.

[MELIGY 2009] MELIGY, Ali. A Grid-Based Distributed SVM Data Mining Algorithm. European Journal of Scientific Research. Middle East University. 2009. ISSN 1450-216X Vol.27 No.3 (2009), pp.313-321.

[OLSON 2008] L. OLSON, David. Advanced Data Mining Techniques, Springer. 1ª Edición. Lincoln, NE. Sprinter. 2008. ISBN: 9783540769163.

[OPIYO 2005] OPIYO, Elisha. Computing Research Challenges and Opportunities with Grid Computing. 2005.

[ORACLE 2009] ORACLE. Oracle Grid Computing. 2009.

[PETHICK 2003] PETHICK, Mark, LIDDLE, Michael, WERSTEIN, Paul, y HUANG, Zhiyi. Parallelization of a Backpropagation Neural Network on a Cluster Computer. University of Otago. 2003.

[SAKTHI 2008] SAKTHI, U. Parallel and Distributed Mining of Association Rule on Knowledge Grid. World Academy of Science, Engineering and Technology. 2008.

[SARNOVSKÝ 2009] SARNOVSKÝ, Martin. Grid-based Support for Different Text Mining Tasks. Centre for Information Technologies. Technical University of Košice. 2009.

[SPSS 2008] SPSS Inc. Documentación Clementine 12: Algorithms Guide, SPSS Inc. Versión 12. Chicago. Integral Solutions Limited. 2007.

[STANKIVSKY 2008] STANKIVSKY, Vlado. Digging Deep into the Data Mine with DataMiningGrid. IEEE. 2008.

[STANKOVSKI 2007] STANKOVSKI, Vlado. Grid-enabling data mining applications with DataMiningGrid. 1ª Edición. ScienceDirect. 2007.

[STANOEVSKA 2009] STANOEVSKA, Katarina, WOZNIAK, Thomas, RISTOL, Santi. Grid and Cloud Computing, A Bussines Perspective on Technology and Applications. Springer. 2009. ISBN 9783642051920.

[Two Crows 2007] CORPORATION, Two Crows. Introduction to Data Mining and Knowledge Discovery, 3ª Edición. Potomac, MD. 2007. ISBN: 1892095025.

[WANG 2003] WANG, John. Data Mining Opportunities and Challenges. Hershey PA. 1ª Edición. Idea Group Publishing. 2003. ISBN 1591400511.

[WEGENER 2008] WEGENER, Dennis. GridR: An R-based grid-enabled tool for data analysis in ACGT clinicogenomics trials, 2008.

[ZAKI 2000] ZAKI, Mohammed. Large-Scale Parallel Data Mining. Springer. 2000. ISBN: 3540671943.