UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN
INTRODUCCIÓN A LA CIENCIA DE DATOS

Tarea Opcional 1: Big Data

José Manuel Alvarez García

CI 25038805

Caracas, Junio 2016.

Índice general

Índice de figuras

Capítulo 1

Conceptos Básicos

1.1. Ciencia de los Datos y Big Data

Figura 1.1: Nube de palabras de Big Data.

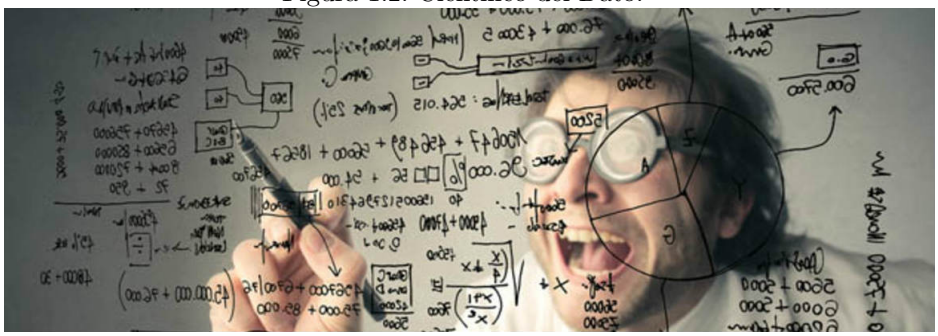
Ciencia de los datos consiste en la generación de conocimiento a partir de grandes volúmenes de datos, aplicando técnicas de procesamiento paralelo y distribuido, para así implementar algoritmos que permitan predecir o detectar patrones sobre los datos almacenados. A partir de los resultados obtenidos, se podrán construir herramientas que permitan analizar los resultados y apoyar los procesos de toma de decisiones.

Big Data es una plataforma tecnológica que permite almacenar de manera distribuida y procesar de manera paralela y distribuida conjuntos de datos, que, por su gran volumen, superan las capacidades de las plataformas de tecnología de la información tradicionales, ya sea porque tomaría demasiado tiempo procesar dichos datos, o porque sería muy costoso implementar una arquitectura que soporte tal cantidad de datos.

Estos términos se encuentran en estrecha relación. No todos los problemas de volumen de datos se pueden solucionar con base en el empleo de tecnología de la información y análisis de datos, por ello, se debe hablar del científico del dato. El objetivo de éste, es saber cuáles son los problemas que se pueden resolver mediante el empleo de estos recursos. Sin embargo, no es posible abarcar en un único perfil todo el conocimiento que requiere un departamento de Ciencia de los datos o que trabaje con Big Data. Por ello, este tipo de trabajos debe estar centrado en un equipo multidisciplinario. Así, la relación vendría dada de que la Ciencia de los datos consiste en un conjunto de conocimientos estructurados que permiten la toma de decisiones basadas en el análisis (grupal) de grandes cantidades de datos.

1.2. Conocimientos y habilidades de un Científico del Dato

Figura 1.2: Científico del Dato.



1.2.1. Aprendizaje Automático

Es una rama de la inteligencia artificial que consiste en el diseño y construcción de aplicaciones que son capaces de aprender mediante entradas y salidas de datos informáticos.

Una de las principales características, es que se aprende por la experiencia de los datos, lo cual es importante para tomar decisiones en las grandes empresas. El aprendizaje automático incorpora disciplinas como el análisis predictivo y la minería de datos, para el reconocimiento de patrones. El crecimiento de los datos actualmente ha hecho que tenga una gran importancia el aprendizaje automático relacionado con Big Data, al igual que está vinculado a modelos de predicción.[1]

1.2.2. Minería de Datos

Es el análisis de grandes datos que se establece para encontrar relaciones insospechadas, intentando descubrir patrones para resumir los grandes volúmenes de éstos en formas novedosas que sean comprensibles y útiles para el dueño del dato.

Por su parte, Big Data hace referencia a grandes cantidades de datos que superan la capacidad de procesamiento habitual del software informático existente. Por lo que Big Data es la tecnología capaz de capturar, gestionar y procesar en un tiempo razonable y de forma veraz estos datos. El crecimiento de estos es el principal motor de la popularidad de la Minería de datos y Big Data. Ambas, pero sobre todo la minería de datos han venido madurando caracterizadas por métodos científicos sólidos y muchas aplicaciones prácticas.[2]

1.2.3. Inteligencia Artificial

Es una rama de la computación que relaciona fenómenos naturales con una analogía artificial a través de programas algorítmicos. A través de la inteligencia artificial, se han desarrollado diversos sistemas expertos que pueden imitar la capacidad mental del ser humano y relacionan reglas de sintaxis del lenguaje hablado y escrito sobre la base de la experiencia, para luego hacer juicios acerca de un problema, cuya solución se logra con mejores juicios y más rápidamente que el ser humano, y por lo cual se puede relacionar fuertemente con el Aprendizaje automático.[3]

Actualmente, se está trabajando en nuevo software que combina tecnologías de análisis sintáctico y semántico de datos, con lo cual es capaz de reconocer y analizar en tiempo real las alertas alimentarias que se producen en el mundo. La herramienta, desarrollada por el centro tecnológico Ainia, en Valencia (España), monitoriza las grandes bases de datos oficiales y filtra la información relevante de forma automática, lo que supone un gran ahorro de tiempo y esfuerzo.[4]

1.2.4. Inteligencia de Negocios

Es una estrategia empresarial que persigue incrementar el rendimiento de la empresa o la competitividad del negocio, a través del estudio de los datos históricos de una organización, tales como transacciones u operaciones diarias, que usualmente se encuentran en grandes almacenes de datos (*Data Warehouses*).[5] Mientras que la Inteligencia de negocios se encarga de analizar datos consolidados e históricos, Big Data es capaz de reforzar este análisis en un entorno más amplio, con más dinamismo y multiplicidad. Permitiendo el análisis de una cantidad más grande de datos no estructurados y de gran importancia para la organización, Big Data aumenta todavía más la relevancia y la utilidad de la Inteligencia de negocios. Por ello, se puede decir que esta última es complementada por Big Data.[6]

1.3. Las V's de Big Data

1.3.1. Volumen

Big Data debe ser capaz de gestionar grandes volúmenes de datos que se generan diariamente por las empresas y organizaciones de todo el mundo. Por ejemplo, la cadena de supermercados americana Wallmart, almacena más de 1 millón de transacciones comerciales cada hora identificando los productos que compran sus clientes, más de 100.000GB de información almacena la red social Facebook diariamente. Toda esta información debe ser almacenada para su futura gestión.[7]

1.3.2. Variedad

Big Data debe tener la capacidad de combinar una gran variedad de información digital en los diferentes formatos en las que se puedan presentar, ya sean en formato video, audio o texto. Diferentes fuentes de información como las nuevas tecnologías que monitorizan nuestra actividad física, el internet de las cosas que conectará los dispositivos y máquinas entre sí, millones de mensajes escritos en redes sociales como Facebook o Twitter, millones de videos subidos a Youtube cada día son ejemplos entre otros de fuentes generadoras de diferentes tipos de información.[7]

1.3.3. Velocidad

Big Data debe ser capaz de almacenar y trabajar en tiempo real con las fuentes generadoras de información como sensores, cámaras de videos, redes sociales, *blogs*, páginas *web* y fuentes que generan millones y millones de datos cada segundo. Además, la capacidad de análisis de dichos datos tienen que ser rápidos reduciendo los largos tiempos de procesamiento a comparación de aquellos que presentan las herramientas tradicionales de análisis.[7]

1.3.4. Veracidad

Big Data debe ser capaz de tratar y analizar inteligentemente este vasto volumen de datos con la finalidad de obtener una información verídica y útil que permita mejorar la toma de decisiones.[7]

1.3.5. Valor

Big Data debe ser capaz de saber qué datos se deben de recolectar, ya que todas las empresas generan, trabajan y gestionan multitud de datos; pero la clave está en cómo obtener la mejor información, el mejor valor y conocimiento para así obtener la mayor rentabilidad posible, saber qué soluciones analíticas usar y saber cómo convertir los datos en conocimiento.[8]

1.3.6. Variabilidad

Big Data debe ser flexible a la hora de adaptarse a nuevos cambios en el formato de los datos, tanto en la obtención como en el almacenamiento y su procesado. Variabilidad sobre el tipo y origen de los datos, ya que estos pueden ser estructurados o no estructurados.[9]

1.3.7. Visualización

Big Data debe representar de manera comprensible y medible los datos obtenidos para encontrar patrones y claves ocultas en el tema a investigar. Es importante conectar la visualización de datos, porque aunque Big Data es en principio algo que interesa a empresas, a nivel individual también puede tener un gran papel, sobre todo usado como método de extracción de conclusiones en estudios sociales.[10]

Capítulo 2

Apache

2.1. Apache Hadoop

Figura 2.1: Logotipo de Apache Hadoop.

Es un *framework* de *software* que soporta aplicaciones distribuidas bajo una licencia libre. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en los documentos Google para MapReduce y Google File System.

Hadoop es un proyecto de alto nivel Apache que está siendo construido y usado por una comunidad global de contribuyentes, mediante el lenguaje de programación Java. Yahoo! ha sido el mayor contribuyente al proyecto, y usa Hadoop extensivamente en su negocio.[11]

2.2. Herramientas del ecosistema Hadoop

- **HDFS** (*Hadoop Distributed File System*). Ofrece un marco básico para la división de colecciones de datos entre varios nodos al utilizar la replicación para recuperarse de la falla del nodo. Los archivos grandes se dividen en bloques, y varios nodos pueden contener todos los bloques de un archivo.

El sistema de archivos está diseñado para mezclar la tolerancia a fallas con un alto rendimiento. Se cargan los bloques para mantener la transmisión constante y generalmente no se almacenan en memoria caché para minimizar la latencia.[12]

- **Hive**. Está diseñado para regularizar el proceso de extracción de los bits de todos los archivos en las bases de datos. Ofrece un lenguaje similar a SQL que analiza los archivos y extrae los fragmentos que su código fuente necesita. Los datos llegan en formatos estándar y Hive los convierte en un algo que se puede gestionar.[12]

- **Pig**. Ejecuta a través de los datos un código escrito en un lenguaje propio, llamado Pig Latin, lleno de abstracciones para la manipulación de los datos. Esta estructura dirige a los usuarios hacia algoritmos que son fáciles de ejecutar en paralelo a través del clúster. Pig viene con funciones estándar para tareas comunes como sacar un promedio de los datos, trabajar con fechas, o buscar diferencias entre las cadenas.[12]

- **YARN**. Combina un administrador central de recursos que reconcilia la forma en que las aplicaciones utilizan los recursos del sistema de Hadoop con los agentes de administración de nodo que monitorean las operaciones de procesamiento de nodos individuales del clúster. Ejecutándose en clústeres de hardware básicos, Hadoop ha atraído un interés particular como zona de espera y de almacenamiento de datos para grandes volúmenes de datos estructurados y no estructurados destinados al uso en aplicaciones de analítica.[13]

- **MapReduce**. Es una técnica de procesamiento y un programa modelo de computación distribuida basada en Java que contiene dos tareas importantes, mapear y reducir. La tarea de mapear, toma un conjunto de datos y se convierte en otro, en el que los elementos se dividen en tuplas (pares clave/valor). Seguidamente la tarea de reducir, toma la salida de un mapeo o representación como entrada y combina los datos tuplas en un conjunto más pequeño de tuplas.[14]

2.3. Hadoop MapReduce y Apache Spark

Figura 2.2: Logotipo de Apache Spark.

Apache Spark combina un sistema de computación distribuida a través de clústeres de computadores con una manera sencilla y elegante de escribir programas. No es complicado entender Spark si se le compara con su predecesor, MapReduce, el cual revolucionó la manera de trabajar con grandes conjuntos de datos ofreciendo un modelo relativamente simple para escribir programas que se podían ejecutar paralelamente en cientos y miles de máquinas al mismo tiempo.

Gracias a su arquitectura, MapReduce logra prácticamente una relación lineal de escalabilidad, ya que si los datos crecen, es posible añadir más computadores y ocupar el mismo tiempo de cómputo. Spark mantiene la escalabilidad lineal y la tolerancia a fallos de MapReduce, pero amplía sus bondades gracias a diversas funcionalidades:

Spark soporta el flujo de datos acíclico. Cada tarea de Spark crea un grafo acíclico dirigido de etapas de trabajo para que se ejecuten en un determinado clúster. En comparación con MapReduce, que crea un grafo con dos estados predefinidos (mapear y reducir), los grafos de Spark pueden tener cualquier número de etapas. Además, Apache Spark mejora con respecto a los demás sistemas en cuanto a la computación en memoria. Sus conjuntos flexibles de datos permiten a los programadores realizar operaciones sobre grandes cantidades de datos en clústers de una manera rápida y tolerante a fallos.[15]

Bibliografía

[1] Aprendizaje automático: el big data para el crecimiento de las grandes empresas. http://reportedigital.com/e-learning/aprendizaje-automatico-big-data-crecimiento-grandes-empresas/.

[2] Minería de datos (data mining) vs grandes datos (big data). http://www.tuataratech.com/2015/06/mineria-de-datos-data-mining-vs-grandes.html.

[3] Qué es la inteligencia artificial. http://bvs.sld.cu/revistas/san/vol2_2_98/san15298.htm.

[4] http://www.agenciasinc.es/noticias/inteligencia-artificial-y-big-data-para-analizar-y-gestionar-alertas-alimentarias. http://www.agenciasinc.es/Noticias/Inteligencia-artificial-y-big-data-para-analizar-y-gestionar-alertas-alimentarias.

[5] Qué es inteligencia de negocios (business intelligence). http://www.itmadrid.com/que-es-inteligencia-de-negocios-business-intelligence/.

[6] Big data y su fuerza para business intelligence. https://stefanini.com/es/2013/10/big-data-y-su-fuerza-para-business-intelligence/.

[7] Qué es big data. http://www.quees.info/que-es-big-data.html.

[8] Las 5 vs del big data en el marketing digital. http://blogueandoalos50.com/las-5-vs-del-big-data-en-el-marketing-digital/.

[9] Big data. https://prezi.com/u-cvki9mnfbk/big-data/.

[10] Las mejores herramientas para visualizar grandes cantidades de datos. https://hipertextual.com/presentado-por/bbva/visualizacion-de-datos.

[11] Hadoop. https://es.wikipedia.org/wiki/Hadoop.

[12] 18 herramientas de hadoop para procesar big data. http://cioperu.pe/fotoreportaje/14938/18-herramientas-de-hadoop-para-procesar-big-data/.

[13] Apache hadoop yarn (yet another resource negotiator). http://searchdatacenter.techtarget.com/es/definicion/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator.

[14] Hadoop - mapreduce. http://www.tutorialspoint.com/es/hadoop/hadoop_mapreduce.htm.

[15] Apache spark: qué es y cómo funciona. https://geekytheory.com/apache-spark-que-es-y-como-funciona/.