DOCUMENTO DE INVESTIGACIÓN 03/2013

BIG DATA EN LOS ENTORNOS DE DEFENSA Y SEGURIDAD

DOCUMENTO RESULTADO DEL GRUPO DE TRABAJO SOBRE BIG DATA, DE LA COMISIÓN DE INVESTIGACIÓN DE NUEVAS TECNOLOGÍAS DEL CENTRO SUPERIOR DE ESTUDIOS DE LA DEFENSA NACIONAL. (CESEDEN)

Tcol. D. José Antonio Carrillo Ruiz
Jefe del Área de Información y Ayuda a la Decisión (AIAD)
Vicesecretaría General Técnica
Doctor en Ciencias Matemáticas
Profesor de la asignatura Estadística (Grado en Ingeniería en Tecnologías de la Información)-UNED,
Diplomas Militares en Estadística e Investigación Operativa

Dr. D. Jesús E. Marco De Lucas
Instituto de Física de Cantabria (IFCA) (Director de 2003 a 2007)
Centro mixto de la Universidad de Cantabria (UC) y el Consejo Superior de Investigaciones Científicas (CSIC).
Doctor en Física y profesor de investigación del CSIC en el IFCA.

Dr. D. Juan Carlos Dueñas López
Catedrático de Ingeniera telemática ETSI Telecomunicación
Actual director del Centro de "Open Middleware" de la Universidad Politécnica de Madrid.
Presidente del capítulo español de IEEE Computer.
Doctor Ingeniero de Telecomunicación

D. Fernando Cases Vega
Gerencia de Planeamiento Tecnológico, Área de Prospectiva Tecnológica del Ministerio de Defensa.
Ingeniero Industrial especialista en Electrónica y Automática.

D. José Cristino Fernández
Ingeniero Jefe del Área de Desarrollo.
Future Space S.A.

D. Guillermo González Muñoz de Morales
Unidad de Observatorios Tecnológicos
Sistema de Observación y Prospectiva Tecnológica
Subdirección General de Tecnología e innovación
Dirección General de Armamento y Material

Tcol. D. Luis Fernando Pereda Laredo
División de Sistemas de Información y Telecomunicaciones
Estado Mayor Conjunto de la Defensa
Diplomado en Telecomunicaciones Militares y Guerra Electrónica

Contenido

1. Introducción al concepto de Big Data..... ...7

1.1.- Definiciones comunes..... ...7

1.2.- Contexto general de aplicación.....9

2. Capacidades técnicas que comprende.....9

2.1.- Adquisición.....9

2.2.- Transmisión.....13

2.3.- Almacenamiento y procesado.....15

2.4.- Presentación.....17

2.5.- Tratamiento automático.....19

2.6.- Explotación de datos. Aprendizaje y toma de decisión automáticos..... 23

2.6.1.- Depuración, exploración y visualización.....25

2.6.2.- Aprendizaje supervisado.....27

2.6.3.- Aprendizaje no supervisado.....40

3. Estudio de aplicaciones de Big Data en Defensa y Seguridad.....43

3.1.- Introducción.....43

3.2.- Aplicaciones específicas.....46

3.2.1.- Detección de intrusión física en grandes espacios o infraestructuras abiertas 46

3.2.2.- Computación sobre información cifrada.....47

3.2.3.- Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos).....49

3.2.4.- Criminología Computacional.....51

3.2.5.- Uso fraudulento de recursos corporativos y/o sensibles.....52

3.2.6.- Análisis de vídeo en tiempo real / búsqueda y recuperación rápida en librerías de vídeo.....53

3.2.7.- Inteligencia Visual en Máquinas.....54

3.2.8.- Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos.....55

3.2.9.- Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos.....57

3.2.10.- Consciencia situacional.....59

3.2.11.-Traducción automática gran escala (en número de idiomas y en volumen).....60

3.2.12.- Predicción de eventos.....62

3.3.- Factores impulsores y limitadores en la aplicación de Big Data en Seguridad y Defensa.....63

3.3.1.- Listado de Factores impulsores.....64

3.3.2.- Relevancia de cada uno de los factores impulsores en las aplicaciones específicas.....69

3.3.3 Listado de factores limitadores / desafíos existentes.....71

3.3.4.-Relevancia de cada uno de los factores limitadores/desafíos en las aplicaciones específicas.....77

4.- Casos concretos, aplicaciones.....79

4.1.- Programa del combatiente del futuro (COMFUT).....79

4.2.- Inteligencia de imágenes.....86

4.2.1.- Justificación.....87

4.2.2.- La tecnología del procesado de imágenes.....87

4.2.3.-Necesidades en Defensa.....100

4.2.4.-Conclusiones.....105

4.2.5 Listado de figuras.....106

4.3.- Guerra electrónica (EW).....106

4.3.1.- Una aproximación a la Guerra Electrónica.....106

4.3.2.- Definición de Guerra Electrónica (EW).....107
86

4.3.3.- Un par de ejemplos ilustrativos.....107

4.3.4.- Acciones de EW.....108

4.3.5.- Objetivos de las acciones.....108

4.3.6.- Tipos de acciones ESM.....109

4.3.7.- ¿Qué podría aportar Big Data?.....110

5.- Análisis.....112

5.1.- Aplicación de Big Data.....112

5.2.- Posible evolución.....117

6.- Conclusiones del estudio.....120

7.- Referencias.....123

1. Introducción al concepto de Big Data

En estos últimos años, los ámbitos empresarial, académico, investigador y de la administración han estado haciendo frente a la avalancha de datos con un nuevo concepto que ha dado en denominarse **Big Data**. Por la simple denominación usada se entiende que se trata de grandes volúmenes de información que no es sencillo tratar con las herramientas y procedimientos tradicionales. Encierra esta idea el tratamiento de información que hace evolucionar los métodos y recursos habituales para hacerse cargo de grandes **volúmenes** de datos (de terabytes pasamos a zetabytes). Estos se generan a gran **velocidad** (pasamos de datos en lotes/archivos a datos en "streaming") y además se añade una posible componente de complejidad y **variabilidad** en el formato de esos datos (pasamos de datos estructurados a datos semi-estructurados o no estructurados). Todo ello requiere de técnicas y tecnologías específicas para su captura, almacenamiento, distribución, gestión y análisis de la información.

También recientemente se añade una nueva "v" de valor: los datos por sí mismos, aun siendo muchos, no proporcionan valor a una empresa u organización. Es su tratamiento, a través de un proceso de planteamiento de hipótesis, creación de modelos estadísticos y semánticos, y definición de algoritmos de corta o larga duración, lo que permite descubrir el significado oculto en esos grandes volúmenes de datos.

1.1.- Definiciones comunes

ALGUNAS DE LAS ENTIDADES Y ORGANIZACIONES HAN DADO SU PROPIA DEFINICIÓN DE LO QUE ENTIENDEN POR BIG DATA Y PUEDE AYUDAR A ENTENDER ESTE CONCEPTO ASÍ SE TIENE:

WIKIPEDIA

"Big Data" es un término aplicado a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable. Los tamaños del "Big Data" se encuentran constantemente en movimiento creciente, de esta forma en 2012 se encontraba dimensionada en un tamaño de una docena de terabytes hasta varios petabytes de datos en un único data set. En la metodología MIKE2.0 dedicada a investigar temas relacionados con la gestión de información, definen Big Data en términos de permutaciones útiles, complejidad y dificultad para borrar registros individuales.

O'REILLY RADAR

"Se considera Big Data cuando el volumen de los datos se convierte en sí mismo

parte del problema a solventar" ().

EMC/IDC

"Las tecnologías de Big Data describen un nuevo conjunto de tecnologías y arquitecturas, diseñadas para extraer valor y beneficio de grandes volúmenes de datos con una amplia variedad en su naturaleza, mediante procesos que permitan capturar, descubrir y analizar información a alta velocidad y con un coste reducido"

McKinsey Global Institute (MGI) en Junio de 2011,

"conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenado, gestión y análisis de las herramientas de base de datos".

The IBM Big Data Platform

Big Data represents a new era of computing – an inflection point of opportunity where data in any format may be explored and utilized for breakthrough insights - whether that data is in-place, in-motion, or at-rest. IBM is uniquely positioned to help clients navigate this transformation.

IBM, considera que hay "Big Data", si el conjunto de información supera el terabyte de información, es sensible al tiempo, y mezcla información estructurada con no estructurada. Así, su enfoque trata de buscar la forma mejor de aprovechar estos datos, su gestión, su combinación (datos estructurados con los que no lo son), la aplicación de algoritmos predictivos de comportamiento, y con todo ello, permitir la toma de decisiones que añadan valor al negocio.

"In a 2001 research report, *META Group (now Gartner)* analyst Doug Laney defined data growth challenges and opportunities as being three dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources) (3V).

In 2012, Gartner updated its definition as follows: "Big Data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Un reciente estudio realizado por la consultora IDC Digital Universe, revela que el tamaño del universo digital alcanzó los 2,8 zettabytes (ZB) en 2012, y predice que para 2020 se alcancen los 40 ZB, a partir de los datos generados por personas y dispositivos, aunque sólo el 0,5% de la información se analiza.

Las estimaciones indican que sólo un 20% del universo digital cuenta actualmente con protecciones frente a robo digital, medidas de privacidad o cumplimiento de reglamentaciones. Y el volumen de información va a crecer mucho más rápido que la inversión en hardware, software, servicios, telecomunicaciones y personal (un 40% entre 2012 y 2020); como consecuencia, la inversión por gigabyte se reducirá en un 90%. Con todo, se estima que el crecimiento de sectores emergentes del universo digital

(gestión de almacenamiento, seguridad, Big Data, Cloud Computing) puede variar entre el 35% y el 65%.

1.2.- Contexto general de aplicación

Big Data no es una tecnología en sí misma, sino más bien un planteamiento de trabajo para la obtención de valor y beneficios de los grandes volúmenes de datos que se están generando hoy en día. Se deben contemplar aspectos como los siguientes:

- Cómo capturar, gestionar y explotar todos estos datos.
- Cómo asegurar estos datos y sus derivados, así como su validez y fiabilidad.
- Cómo disponer la compartición de estos datos y sus derivados en la organización para la obtener mejoras y beneficios.
- Cómo comunicar estos datos y sus derivados (técnicas de visualización, herramientas, y formatos) para facilitar la toma de decisión y posteriores análisis.

2. Capacidades técnicas que comprende

2.1.- Adquisición

Una de las principales razones de la actual "explosión de la información digital" es la evolución exponencial en el número y capacidad de los sistemas de adquisición de datos. Esta evolución tiene una clara raíz tecnológica: la mejora, abaratamiento y miniaturización de estos sistemas, tanto de los sensores cómo de la electrónica asociada, y en particular de su conexión a la red Internet.

Un ejemplo que muestra esta evolución tecnológica son los teléfonos móviles, que pueden integrar sensores relativamente sofisticados (cómo una cámara de alta resolución o un receptor GPS) junto a otros sencillos (sensores de aceleración o el propio micrófono) y son capaces de digitalizar y transmitir la información correspondiente a la red, todo ello con un coste reducido.

La evolución de la microelectrónica, y actualmente de la nanotecnología, ha permitido diseñar todo tipo de sensores para medidas físicas y químicas, y cada vez más también biológicas, con buena resolución y fiabilidad y a un coste cada vez menor. Sensores ideados en un principio para una función específica, cómo los sensores de gases que incorporan los motores de los coches para mejorar su rendimiento, o los sensores de consumo de electricidad de un contador doméstico, pueden integrar su información fácilmente y casi en tiempo real en la red. En la idea inicial de un Internet of Things todos estos sensores ubicuos se autoidentificarían a la red y proporcionarían

de forma autónoma un gran volumen de datos, útil para entender diferentes modelos físicos, sociales o económicos, para sistemas complejos de modelar. En iniciativas cómo las Smart Cities se integran este tipo de sensores y otros similares cómo sensores de paso, de presencia, o estaciones medioambientales, con otros más complejos cómo cámaras, con el propósito de optimizar un modelo de uso de recursos urbanos. El flujo de información proporcionado por estos sensores al sistema distribuido o central de monitorización y almacenamiento de los mismos viene limitado por el ancho de banda de la red de comunicación disponible, típicamente del orden de Kilobits/s o Megabits/s si se usa tecnología sin cables (GSM/GPRS/UTMS o wifi/wimax) hasta Gigabit/s para conexiones por cable o fibra (reservado normalmente a la conexión de hubs de sensores o de cámaras de alta resolución). El reto viene de la integración de un gran número de sensores: pensemos por ejemplo en la integración de millones de medidas en tiempo casi real de consumo eléctrico en hogares proporcionado por contadores inteligentes.

Por otra parte, y especialmente en el ámbito de la investigación, continúa el desarrollo de detectores y sensores cada vez más complejos para lograr medidas de gran precisión. Destacan detectores cómo los empleados en grandes experimentos científicos: en el Large Hadron Collider (LHC) del CERN (Laboratorio Europeo de Física de Partículas) el detector denominado CMS http://en.wikipedia.org/wiki/Compact_Muon_Solenoid , que mide las trazas de las partículas resultado de la colisión de dos protones a una energía 8000 veces su propia masa, tiene más de 15 metros de diámetro, pesa 12.000 toneladas, y cuenta con más de 75 millones de canales de lectura. Este detector en operación registra 1000 colisiones por segundo, lo que resulta en un volumen de datos muy elevado (se han superado recientemente los 100 Petabytes de datos grabados correspondientes a dos años de operación del LHC). Tanto en Astronomía cómo en el área conocida cómo Observación de la Tierra, el despliegue de satélites y aviones de observación con cámaras multiespectrales en un sentido amplio, ha dado lugar igualmente a un gran volumen de datos que cubren prácticamente todo el espacio y espectro, y pretenden hacerlo con una alta resolución tanto espacial cómo temporal. Nuevas técnicas de alta precisión, como los sistemas LIDAR (o LADAR) avanzan en la misma dirección: mejorar la precisión y rango de variables que conocemos sobre los sistemas físicos y humanos para un entorno dado de interés.

La llegada de la web 2.0 ha supuesto un cambio en el modo de comunicación de los usuarios en Internet, de esta forma los usuarios han dejado de ser meros receptores de información y han comenzado a ser generadores de la misma. Hoy en día, la mayor parte de los usuarios forman parte de las redes sociales, disponen de sus propios blogs o comentan en ellos, participan en foros, comentan noticias,… lo que provoca que el volumen de información disponible haya crecido de forma exponencial en los últimos años, como muestran algunos datos relevantes:

- Durante el 2012, los usuarios de YouTube subieron 48 horas de vídeos nuevos cada minuto.

- Se crearon 571 sitios web cada minuto durante el pasado año
- Cada día Facebook procesa 100 terabytes de información procedente de sus usuarios.
- Según datos de Twitter a principios de 2012, se escribieron 175 millones de tweets al día.
- Se prevé que la producción de datos sea 44 veces mayor en 2020 que en 2009.

De hecho, la necesidad de poder manejar estos volúmenes de información que se producen en Internet es uno de los pilares básicos del nacimiento del término Big Data.

Esta fuente de información es fundamental a la hora de plantear técnicas de adquisición de datos en el sector de la Seguridad y la Defensa, de hecho, una de las disciplinas que engloban la inteligencia es lo que se conoce como OSINT (Open Source Intelligence) que bebe de las fuentes abiertas para la generación de inteligencia, entre ellas la información de Internet. De esta forma, algunos países como Estados Unidos han creado órganos específicos como el National Intelligence Open Source Center (OSC), o Australia con su National Open Source Intelligence Centre (NOSIC). A nivel nacional algunas publicaciones como "Como explotar OSINT eficazmente"(http://www.defensa.gob.es/ceseden/Galerias/esfas/destacados/en_portada/COMOx20EXPLOTARx20OSINTx20EFICAZMENTE.pdf) resaltan la importancia de esta fuente de información, en ella su autor, el Cte. del ET. D. José Raúl Martín Martín, identifica como desventaja de esta disciplina uno de los problemas que tratan de resolver las técnicas de BigData: "El volumen de información es tan brutal, que requiere de organizaciones especializadas en la búsqueda y tratamiento de fuentes abiertas, centralizada al más alto nivel, incluso a nivel político, poniendo la información a disposición del nivel militar estratégico y operacional".

En el mismo contexto de fuentes abiertas, se está produciendo un gran movimiento alrededor de lo que se conoce como Open Data. Open Data es un término que denomina a una serie de iniciativas cuyo objetivo es proporcionar un acceso a ciertos datos de manera universal, de tal forma que dichos datos puedan ser consumidos sin restricciones de patentes ni licencia. La definición oficial de Open Data implica que los datos pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen. De esta forma, sus características fundamentales más importantes son:

Disponibilidad y acceso: la información debe estar disponible como un todo y a un costo razonable de reproducción, preferiblemente descargándola de internet. Además, la información debe estar disponible en una forma conveniente y modificable.

Reutilización y redistribución: los datos deben ser provistos bajo términos que permitan reutilizarlos y redistribuirlos, e incluso integrarlos con otros conjuntos de

datos.

Participación universal: todos deben poder utilizar, reutilizar y redistribuir la información.

Este término ha ganado popularidad en los últimos tiempos gracias al lanzamiento de iniciativas como OpenDataGoverment u OpenScienceData. Las iniciativas de OpenDataGoverment promueven que las Administraciones Públicas proporcionen un acceso a los datos públicos de manera universal. Un ejemplo de estas iniciativas es la lanzada en el portal http://datos.gob.es que gestiona el catálogo de información pública de la Administración General del Estado. Por otro lado, "open science data" es un tipo de OpenData focalizado en la publicación de las observaciones y resultados de las actividades científicas disponibles para su análisis y reutilización de forma pública. Si bien, esta idea fue promovida activamente en los años 50, el auge de internet ha provocado una disminución significativa de los costes y tiempos requeridos para la publicación y obtención de dichos datos.

Pero quizás la mayor avalancha de datos experimentales va a venir unida a nosotros mismos. Dejando a un lado la "traza digital" de nuestra actividad en internet, en nuestro móvil y en nuestras tarjetas de pago, que pueden considerarse sistemas "heterodoxos y no suficientemente controlados" de adquisición de datos, la otra clave va a ser el desarrollo de la medicina personalizada, gracias al desarrollo de los nuevos chips genómicos, el uso cada más extendido y sistemático de equipos de adquisición y análisis de imagen médica, y la incorporación a la vida diaria de los sistemas de tele-asistencia. Los sistemas de adquisición para este caso son muy diferentes. Los equipos de imagen médica, normalmente manejados en hospitales y por personal especializado, son cada vez más precisos y complejos, aportando en un solo examen una gran cantidad de información (por ejemplo en un TAC). Los sistemas de secuenciación genómica, hasta ahora relativamente complejos, están en plena evolución: empiezan a estar disponibles chips de bajo coste conectables directamente a un sistema sencillo de adquisición, que podría ser un teléfono inteligente o una tableta. De la misma forma los sistemas de tele-asistencia incorporan no sólo cámaras y micrófonos de conversación del paciente con el personal sanitario, sino también conexión para sensores básicos (tensión, temperatura, pulsaciones, ritmo cardíaco) y no tan básicos (cámaras para detección de melanomas). Aunque se trata de medidas sencillas, la monitorización de cientos de miles o millones de pacientes, y el interés de analizar correlaciones y contar con sistemas de alarma ante epidemias, hace de este campo uno de los mayores retos en el campo de Big Data. En conjunto el diseño, operación, explotación, protección y preservación de historias médicas es un proyecto muy complejo por la importancia de la privacidad de dichos datos.

Técnicamente debe indicarse la gran variedad de técnicas y protocolos empleados en la adquisición de datos, formato de los mismos, y forma de publicación a la red. Existen algunos estándares, principalmente de-facto, dentro de comunidades organizadas, incluyendo en muy pocos casos la definición formal de metadatos. Este esfuerzo es

especialmente crítico para desarrollar aplicaciones interdisciplinares, como por ejemplo aborda el Open Geospatial Consortium en la iniciativa Sensor Web Enablement (SWE), y el estándar IEEE 1451.

Concluir este apartado indicando que esta explosión en el campo de la adquisición de datos conlleva muy importantes implicaciones sociales, culturales y filosóficas, pendientes de abordarse.

2.2.- Transmisión

En general, se distinguen varios tipos de tráfico de datos que imponen requisitos sobre la red de transporte: para tráfico de voz, el ancho de banda puede variar entre 21 y 300Kbps, dependiendo del codificador de VoIP; de acuerdo a la norma ITU G.114 el retardo máximo debe ser de 150ms, y se permite un desplazamiento (jitter) medio de menos de 30ms, con pérdidas de menos de 1%. Si el tráfico corresponde a conferencia vídeo interactivo: 480 Kbps, retardo máximo de 150ms, y desplazamiento y pérdidas como para voz. También, para flujos de vídeo: 1.5 Gbps para HDTV sin comprimir, 25-60 Mbps para HDTV MPEG2, 4-6 Mbps para MPEG2 TV, con un retardo de 5 segundos, y pérdidas de menos del 5%. El desplazamiento dependerá de la capacidad de los buffer de aplicación. A estos tipos de tráfico hay que añadir los flujos de datos o de eventos de tiempo real que producen las redes de sensores o los sistemas financieros. Este tráfico es enormemente dependiente del tipo de aplicación; la gestión energética de edificios no requiere tasas altas de transferencia, pero un sistema de almacenamiento remoto de imágenes médicas sí las necesita. Las aplicaciones de Big Data que se apoyan en recogida de datos de campo necesitarán el máximo ancho de banda disponible.

El tráfico interno en un centro de proceso de datos tiene unos requisitos especiales de velocidad; como ya se ha comentado, los centros de proceso para soporte a la computación en la nube se configuran en grupos (clúster) de servidores, almacenamiento y conectividad, de manera que la mayor parte del tráfico interno (hasta el 90%) viaja entre servidores del mismo clúster, y el restante entre grupos. La situación actual es de cierto bloqueo de proveedor, de forma que la empresa que provee los servidores para un centro suele proporcionar también la conectividad interna. En este contexto las necesidades de estandarización son muy altas (se estima que permitirían ahorros de cerca de un 25%), y hay en marcha trabajos por parte de IETF, IEEE y otras organizaciones para definirlos, como el protocolo Multiple STP, IEEE 802.3ad (link aggregation group), Transparent Interconnection of Lots of Links (TRILL), o Shortest Path Bridging (SPB), algunos de los cuales ya cuentan con implementaciones comerciales, como FabricPath (capaz de soportar 16000 direcciones MAC y 128000 direcciones IP), Brocade (hasta 32000 direcciones MAC), o QFabric (96000 direcciones MAC, 24000 direcciones IP sobre 10Gb Ethernet).

Las aplicaciones de tratamiento masivo de datos (Big Data), al igual que las de

transferencia de datos entre centros de cómputo, van a exigir mayores capacidades, lo que en la mayor parte de los casos se traduce en la necesidad de establecer redes privadas, enlaces dedicados u otros recursos de uso específico. En general y dependiendo del alcance geográfico, las redes de comunicaciones se pueden clasificar en:

Núcleo de red, con alcances de cientos y miles de kilómetros, capaces de proporcionar conectividad entre puntos de presencia, para lo cual hacen uso de tendidos de fibra óptica propiedad de grandes empresas, consorcios, y estados. Normalmente cursan tráfico entre puntos de presencia de la red. La capacidad actual varía entre 2.5 y los 10 Gbps.

- Redes metropolitanas, que alcanzan hasta los 100 Km. sobre áreas urbanas; en muchos casos se trata de anillos SONET/SDH.
- Redes de acceso, que proporcionan la conectividad a los usuarios finales mediante tecnologías como DSL, CATV, PON, inalámbricas, etc.
- Redes de área personal, con un alcance de decenas de metros, suelen ser inalámbricas y recogen información de sensores.
- Redes de satélite, de propósito particular en las que los satélites pueden ser transmisores de información y también generadores de ésta.

La aparición de nuevos paradigmas de computación, tales como la computación en la nube, o el tratamiento masivo de datos, imponen nuevas exigencias a las redes de datos: aquellas que comunican los ordenadores que forman los clúster o centros de proceso de datos; y sobre todo para nuestro estudio, las redes que son capaces de mover los datos desde su origen (redes de sensores, redes de área personal), hasta el lugar en el que se va a producir el procesado. Es importante hacer notar que este movimiento de datos no es simétrico: los datos que se recogen en el campo de operaciones se mueven hacia el lugar de procesado (y este es el movimiento masivo de datos); en algunas ocasiones se producen resultados de la toma de decisiones que vuelven al campo de operaciones -pero en este segundo momento no se puede hablar de datos masivos-. La seguridad en el movimiento de datos también es una característica crucial en el ámbito de la defensa, pero que impone una sobrecarga sobre la tasa de transferencia de información.

De aquí se deriva una consecuencia importante para la arquitectura técnica de un sistema de tratamiento masivo de datos en el ámbito de la seguridad y la defensa, que opere sobre datos recogidos en un área de operaciones: los datos que se obtienen de las diferentes redes de área personal o de corto alcance -si tienen el volumen suficiente para ser considerados masivos- han de concentrarse en una pasarela o punto de acceso, que estará conectado a las redes de área extensa para alcanzar la infraestructura de procesado masivo de datos.

En el mundo civil aparece también la necesidad de enviar grandes volúmenes de datos a largas distancias; el mundo de las finanzas es un buen ejemplo de movimiento

de gran cantidad de información bajo restricciones de tiempo de transferencia. Es por este motivo por el que se han desarrollado una serie de técnicas que se denominan "optimización WAN", que tratan de mejorar la eficiencia en la transferencia de datos sobre estas redes de área amplia. Dentro de la optimización WAN, tiene interés particular el enfoque de las redes WAN intercentros de procesado de datos, cuyo tráfico va desde los 100Mbits/s hasta 1GBits/s). Existen diferentes técnicas para mejorar el caudal, latencia, tasa de éxito y otros parámetros; entre ellas se pueden mencionar las técnicas de compresión, optimización de latencia, perfilado de tráfico, etc. Los principales fabricantes de equipamiento de red tienen productos para este segmento de mercado, aunque el siguiente reto al que se enfrentan es el de transmitir flujos de datos ("data streams") en tiempo real –según se producen- para lo cual es preciso no sólo aumentar el caudal efectivo sino mantener o mejorar la latencia en las comunicaciones.

2.3.- Almacenamiento y procesado

"Hace una década, la escalabilidad del almacenamiento de datos se consideraba uno de los mayores problemas técnicos a los que se enfrentaban los propietarios de dichos datos. Sin embargo, la incorporación de nueva tecnología eficiente y escalable en gestión de datos y almacenamiento ha hecho que este deje de ser el problema que era" [cita de Big Data, a New World of Opportunities, NESSI White Paper, Diciembre 2012]. Efectivamente, mientras que en el año 2000 la instalación de un sistema de almacenamiento de capacidad 10 Terabytes (10.000 Gigabytes) era un pequeño reto técnico y económico, en 2013 la instalación de un sistema de almacenamiento de 1 Petabyte (1000 TB), que ocupa menos de un rack standard, no conlleva ninguna dificultad. La razón hay que buscarla por un lado en la evolución de la capacidad y precio de los discos magnéticos (actualmente alcanzan 3TB en formato SAS 3.5") y por otra parte en la mejora de los sistemas de conexión y ficheros distribuidos. Tanto los sistemas SAN (Storage Area Network) en los que cada computador conectado a la red de almacenamiento puede acceder directamente a cada bloque de información almacenado, conectado usualmente por fibra mediante el protocolo FCP (Fiber Channel Protocol, SCSI sobre Fibra) a velocidades de 8Gbit/s, o NAS (Network Attached Storage), en los que se accede a los sistemas de disco a nivel de fichero, normalmente por protocolo IP, a velocidad de hasta 10 Gbit/s, permiten implementar soluciones distribuidas, con redundancia, y actualmente a un coste asumible. El desarrollo de soluciones más flexibles y de menor coste y complejidad, cómo ISCSI, y la lenta pero imparable introducción de los discos de estado sólido (SSD), con mucho menor consumo de energía y mejor rendimiento especialmente en acceso aleatorio, permite esperar una evolución muy positiva en los próximos años. Los sistemas de ficheros de alto rendimiento, cómo GPFS o Lustre, ofrecen al usuario, a través de servidores dedicados, volúmenes de trabajo de varios Petabytes con un rendimiento igual o superior al proporcionado por sistemas de disco local, limitado solamente por la conexión de red y la potencia del propio ordenador. Esta solución es la que implementan actualmente la mayoría de los centros de supercomputación, utilizando incluso conexiones Infini-

band que permiten velocidades de conexión máxima de 40Gb/s y mínima latencia.

Por otra parte el análisis o procesado de estos grandes volúmenes de datos se realiza normalmente en clústeres formados por servidores de alto rendimiento, interconectados entre sí y conectados a los sistemas de almacenamiento descritos. La potencia de los sistemas de computación ha crecido de forma exponencial desde los años 70, cómo intuitivamente predijo en 1965 Gordon Moore, cofundador de Intel: cada dos años aproximadamente se duplica el número de transistores por circuito integrado y por tanto la potencia del procesador. Esta evolución se refleja en la potencia de los servidores actuales, con varios procesadores y múltiples núcleos por procesador, que superan fácilmente los 300 Gigaflops (300.000 millones de operaciones por segundo). Del mismo modo han aparecido los procesadores basados en aceleradores de gráficos o GPUs, con más de 500 núcleos cada uno, que permiten superar la potencia de 1 Teraflop (1000 Gigaflops) en un solo servidor. Así la mayoría de los supercomputadores incluidos en la lista top500 [top500.org] supera los 100 Teraflops y los más potentes llegan a superar los 10 Petaflops (10.000 Teraflops) y los 500.000 núcleos.

En los clústeres los nodos pueden operar de modo "individual", procesando cada uno de ellos información de forma independiente sin necesidad de comunicación con los demás nodos, o de forma colectiva en "paralelo", cuando el procesado requiere comunicación entre los nodos. Un ejemplo de este segundo caso es el procesado segmentado de una imagen muy grande, o el entrenamiento de una red neuronal sobre un gran volumen de datos. La "paralelización" y optimización de las aplicaciones para que hagan uso de muchos núcleos es un paso crítico para mejorar el rendimiento de muchas aplicaciones en Big Data.

Frente a esta infraestructura "clásica" de almacenamiento y procesado de datos, que funciona de modo muy eficaz para muchos problemas clásicos, incluyendo la implementación de bases de datos SQL, han aparecido en relación con ciertos problemas de Big Data nuevas soluciones en principio más económicas y eficientes para el almacenamiento y posterior procesado de datos. Probablemente la más conocida es la implementada por Google para el procesado de las búsquedas, basada en el denominado Google File System (GFS) que proporciona acceso a datos distribuidos sobre grandes clústeres de servidores de muy bajo coste, junto con la técnica MapReduce que distribuye la tarea de procesado de la información entre los nodos en los que está almacenada de forma local. La plataforma abierta HADOOP permite implementar esta solución de modo eficiente y escalable, empleando el sistema denominado HDFS (Hadoop Data File System).

La combinación de este tipo de sistema de ficheros distribuidos junto con nuevas técnicas de bases de datos NoSQL, permite abordar de forma muy eficiente y extremadamente escalable problemas de Big Data del tipo almacenamiento y análisis de un flujo de millones de "mensajes", sean mensajes intercambiados en una red social, búsquedas en Google, registros de millones de sensores, o trazas de la actividad de cientos de miles de personas. Además muchas de estas bases de datos NoSQL, renunciando

a algunos de los requerimientos de las bases de datos SQL, permiten extensiones en columnas, incluir información heterogénea, mejor escalabilidad, etc.

La evolución de las plataformas de almacenamiento y procesado "hacia" la red en los últimos años ha sido imparable. En primer lugar por razones de escalabilidad y posibilidad de compartir información y recursos. Este es el principal motivo de aparición del "GRID", un sistema en el que el usuario accede a los recursos de computación y almacenamiento independientemente de donde estén ubicados, utilizando un certificado digital para definir su identidad y permisos de uso. Así el Worldwide LHC Computing Grid [http://wlcg.web.cern.ch/] integra recursos de 170 centros de computación de todo el mundo conectados en su mayoría por red de fibra oscura a 10Gb/s, para almacenar y procesar un total de más de 100 Petabytes de datos de colisiones en LHC.

El segundo motivo de evolución hacia la red es a la vez económico y tecnológico: las soluciones "CLOUD" permiten aprovechar recursos muy grandes, por tanto con gran beneficio de escala, para ofrecer a todo tipo de clientes servicios de computación basados en máquinas virtuales, escalables bajo demanda, accesibles a través de la red. La propuesta es en general atractiva para los clientes potenciales, ya que elimina la necesidad de recursos locales, y permite acceder a gran cantidad de recursos si es necesario, aunque a un coste final que puede ser elevado, no solo por horas de computación sino por transferencia y almacenamiento de los datos. Sistemas cómo las denominadas EC-2 "clúster compute instances" de Amazon permiten implementar soluciones de alto rendimiento en un entorno Cloud, aunque el coste es aún el doble que el correspondiente a realizar el mismo proceso en un centro de computación de alto rendimiento ["Computing e-Infrastructure cost estimation and analysis – Pricing and Business models" proyecto europeo e-Fiscal].

2.4.- Presentación

La propia definición de Big Data lleva implícita la dificultad de poder analizar, entender y presentar los datos y posibles modelos asociados.

En primer lugar dado el volumen de los mismos, en general habrá que recurrir a la presentación bien de una muestra estadísticamente significativa o de valores estadísticos representativos de la muestra. Por ejemplo en el análisis de millones de colisiones de partículas en CMS podemos presentar los gráficos cinemáticos de las 25 colisiones que hemos seleccionado cómo procedentes de la producción de un bosón de Higgs, o bien mostrar la distribución de la masa reconstruida de las partículas mostrando el pico identificativo de la producción del bosón. Más interesante puede ser el análisis de correlaciones entre las diferentes variables, y su posible asociación con un modelo. Por ejemplo la definición y entrenamiento de una red neuronal, el análisis de los pesos de los diferentes nodos, y las correlaciones entre variables para las muestras procesadas.

Dejando aparte casos especiales, en los que se puede usar por ejemplo audio o

sonido para la presentación, lo usual es recurrir a imágenes, secuencias de imágenes o video, para la presentación. (Plena validez del conocido dicho: una imagen vale más que mil palabras)

En el caso de imágenes destacar la importancia de que las herramientas analíticas cuenten con su correspondiente versión de presentación, por ejemplo en identificación de patrones en imágenes, y permitan una interacción ágil y recordable, de modo que las imágenes se puedan "informar" pasando a enriquecer la base de conocimiento y dando valor añadido a las mismas. Es un caso de uso claro por ejemplo en aplicaciones médicas.

La integración de múltiples formatos de información se suele realizar a través de los denominados "dashboards" o paneles de control, que muestran múltiples gráficos, independientes o no, a una o varias personas. Estos paneles pueden mostrar información en tiempo real o casi real, superponer información de modelos ejecutados en tiempo real o casi real, implementar alarmas sonoras y/o visuales, etc. El abaratamiento del hardware necesario, incluyendo tarjetas que gestionan hasta 16 monitores desde un solo servidor, permite implementar estos dashboards de modo económico y flexible. Una situación especialmente interesante es el despliegue de paneles 2D para mostrar la evolución de diferentes variables en una misma zona geográfica, o usar múltiples paneles para cubrir una zona muy extensa. La integración con sistemas de información geográfica (GIS) o similares (GoogleEarth, GoogleMap, etc.) es una posibilidad cada vez más extendida. Se suele integrar la representación de datos de otras fuentes, por ejemplo satélites, en tiempo casi-real.

No es tan evidente sin embargo la integración de la dimensión temporal, siendo en muchos casos muy relevante, que se suele limitar a un cursor de desplazamiento temporal, sin permitir el marcado de la evolución de fenómenos de forma sencilla.

El paso de gráficos en 2D a 3D permite realizar en primer lugar inspecciones sobre estructuras 3D, y en segundo lugar análisis de correlaciones más complejos. A pesar de la popularidad actual del 3D, su uso profesional no está aún muy extendido, quizás en parte por requerir el uso de hardware personal específico (normalmente gafas 3D y un monitor o un proyector).

Más interesante, pero aún menos popular, es el uso de sistemas inmersivos, cuyo máximo exponente son los sistemas CAVE. Estas habitaciones están dotadas con pantallas o paneles que proyectan imágenes 3D rodeando al usuario, que puede interaccionar virtualmente con la información mostrada, bien para entender una estructura en 3D, bien para analizar información dispersa espacialmente, o bien para entrenarse o tomar un decisión frente a una situación real o simulada con la máxima concentración y realismo posible. Existen versiones mucho más limitadas y sencillas pero de coste mucho menor cómo los cascos personales 3D que pueden proporcionar en muchos casos una buena sensación de inmersión, aunque con la limitación en principio de un uso personal, de no estar preparados para compartir la experiencia.

Por último insistir en la importancia de la presentación en situaciones en tiempo real, que pueden requerir toma de decisiones casi inmediatas. Un buen ejemplo puede ser un dashboard, inmersivo o no, de una situación de emergencia, en la que el equipo de mando puede analizar en tiempo real sobre un GIS la posición del personal desplegado, incluso la visión local de la situación transmitida por los mismos, la predicción de avance del riesgo en base a modelos, y las mejores posibilidades de actuación o de evacuación. En el caso por ejemplo de un incendio, puede ser necesario el procesado remoto del stream de datos del video para detectar patrones de posibles fuentes de peligro adicional en la zona aun a través del humo, cómo puede ser un material inflamable

2.5.- Tratamiento automático

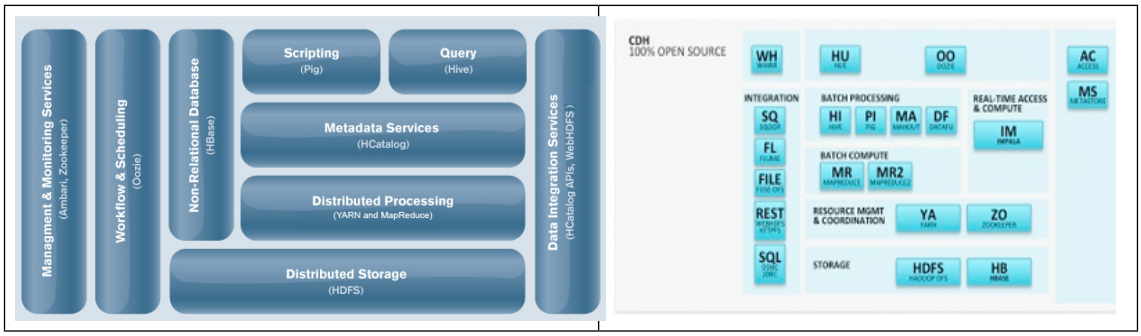
La amplia disponibilidad de datos sólo resulta efectiva si de ellos se puede obtener más información. Recordemos que el avance que se espera avanza en las tres líneas de volumen de datos, variabilidad de éstos y velocidad de procesado; y que en este momento es posible abordar el tratamiento automático de grandes volúmenes de datos porque los ordenadores son más potentes, se han propuesto y usado con éxito arquitecturas altamente distribuidas, y se han creado las plataformas de software que hacen un uso efectivo de estas facilidades. El hecho de que este software esté a disposición de los potenciales usuarios sin coste de licencia (código abierto), en modo de acceso a servicio (sin necesidad de disponer de una infraestructura costosa) y utilizando modelos y lenguajes más difundidos que los tradicionales del mundo de la computación de altas prestaciones, han provocado el aumento de las expectativas sobre Big Data.

Con respecto al software que permite tratar estos datos no existe un modelo de referencia que haya sido aceptado universalmente, como ocurre con las tecnologías en sus primeras fases de vida. Sí que hay propuestas industriales y del mundo académico que tratan de definir cuáles son los componentes que conforman un sistema de Big Data, incluyendo sus relaciones y propiedades. En muchas de estas propuestas de arquitectura aparece el concepto de computación en la nube como ligado al de Big Data. Sin entrar en detalles, la computación en la nube es un paradigma de cómputo en el cual los recursos (de almacenamiento, proceso o comunicación) configurados como un sistema distribuido, se ofrecen como servicios a través de la red, que serán consumidos por los usuarios de forma transparente, elástica y bloqueando para su uso sólo los recursos necesarios. Se puede decir que la computación en la nube, a través de sus modos de uso: IaaS (Infrastructure as a Service), PaaS (Platform as a Service) y SaaS (Service as a Service) ofrecen máquinas virtuales, plataformas de ejecución y aplicaciones o servicios, respectivamente, para ser utilizados en el tratamiento de grandes volúmenes de datos.

Así pues, tratando de referirnos al modelo de arquitectura más sencillo, se podría decir que los sistemas de tratamiento de Big Data constan de tres niveles:

- El nivel de persistencia, almacenamiento y/o provisión de los datos, en donde se encuentran desde las bases de datos relacionales y no relacionales, los sistemas de ficheros adaptados, hasta los procesadores de flujos de datos en tiempo real o los procesadores complejos de eventos, pasando por las herramientas de extracción, transformación y carga para obtener y adaptar los datos del exterior.
- El nivel de análisis, que acoge plataformas de tratamiento distribuido masivo de datos como Hadoop o MapReduce, sobre las cuales se programan las aplicaciones de análisis. Muchas de ellas se han venido ensayando en el mundo de la Minería de Datos, que combina técnicas estadísticas y de aprendizaje automático, aunque también aparecen técnicas de procesado de señal, reconocimiento de patrones, optimización, análisis de grafos, etc.
- El nivel de aplicación y presentación, que corresponde al ámbito de la visualización de datos, o a su elaboración mediante técnicas de inteligencia de negocio (Business intelligence) para presentarlo a los niveles directivos de la organización; o incluso a la automatización de la toma de decisiones. Aunque el ideal es la toma autónoma de decisiones, a menos que haya razones que lo imposibiliten, se mantiene el elemento humano como último elemento. La competición que organiza la agencia DARPA norteamericana de creación de sistemas autónomos de conducción que se prueban sobre vehículos reales que han de conducirse a través del desierto durante decenas de kilómetros es un buen ejemplo de toma de decisiones autónoma.

La siguiente figura presenta la arquitectura software de dos de las plataformas de Big Data más utilizadas actualmente, ambas de código abierto. A la derecha se muestra Cloudera (versión 4), y a la izquierda la plataforma de HortonWorks (versión 2) a principios de 2013.



Las dos incorporan HDFS (Hadoop File System) como servicio básico de almacenamiento; es un sistema de ficheros distribuido y escalable escrito en Java, que se comunica con el resto del sistema usando TCP/IP, y puede almacenar ficheros grandes (tamaño típico 64MB). Otros sistemas parecidos son HBase, que a veces se cataloga como base de datos no relacional. El movimiento NoSQL que ha ido ganando adeptos sobre todo en el contexto de Big Data, propone no usar el modelo relacional ni ofrecer interfaces de acceso y consultas SQL. Por el contrario, usan modelos de

clave-valor y acceso usando métodos de un lenguaje de programación convencional como Java, C# o C++, con lo que consiguen manejar grandes cantidades de datos en tiempos reducidos, a expensas de no poder realizar operaciones complejas de relaciones entre datos (join) ni cumplir siempre las capacidades ACID (atomicity, consistency, isolation, durability).

Para el análisis, sin duda el componente más famoso es MapReduce: una biblioteca para cómputo paralelo sobre grandes colecciones de datos. Su primer uso fue el cálculo del algoritmo PageRank de Google por el cual se puntúa la calidad de una página web a la hora de indexarla. En realidad MapReduce es un modelo de programación distribuida que indica que los algoritmos deben dividirse en dos tipos de tareas: la función mapa que toma cada dato de la entrada en paralelo y lo transforma, asociando el resultado a un grupo; y la función reduce que toma cada grupo en paralelo y proporciona un resultado final. Su implementación más importante es Apache Hadoop, que permite aplicar el modelo MapReduce sobre conjuntos de petabytes de datos y sobre miles de nodos de cómputo. Sobre Hadoop se puede programar en Pig; o utilizar Mahout, una biblioteca abierta de implementaciones distribuidas de algoritmos de aprendizaje automático (clasificación, clustering, regresión, reducción de dimensiones, filtros colaborativos, similitud de vectores y otros).

En lo relativo al software de almacenamiento, en los últimos años se popularizan varios tipos de sistemas de gestión de bases de datos que difieren de los relacionales en varios aspectos: no soportan SQL como principal lenguaje de consultas, no utilizan estructuras fijas como tablas, no soportan operaciones tipo Join, ni garantizan las propiedades ACID (atomicidad, coherencia, aislamiento y durabilidad) por completo. A continuación se describen los tipos principales de bases de datos NoSQL.

Tipo	Descripción	Sistemas de código abierto
Big Table	Partición de una tabla en columnas situadas en diferentes servidores	Cassandra, Hbase, HyperTable
Atributo-valor	Objeto binario (BLOB) con clave	Voldemort, Riak, Amazon S3
Almacén documental	Conjuntos de datos identificados por etiquetas	CouchDB, MongoDB
Grafos	Información estructurada en forma de red o grafos	Neo4j, Titan, Infinite Graph
Índices Full-text	Sistemas que indexan sobre textos no estructurados	Apache Lucene, Zoie
Almacén XML	Bases de datos jerárquicas conformes a esquemas XSD-XML, con XPath como lenguaje de consulta	BaseX

En un intento por dar solución a la característica de Velocidad de las soluciones Big Data, surgen también los sistemas de procesado masivo de datos en tiempo real. En el mundo del código abierto, la plataforma Storm aparece a partir de Twitter, que permite la computación distribuida de flujos de datos en tiempo real (procesado de más de un millón de tuplas por segundo), a lo que añade características de interoperabilidad,

rendimiento, escalabilidad y tolerancia a fallos.

Nadie duda de la potencia de las soluciones basadas en código abierto; incluso las principales soluciones comerciales usan algunos de los sistemas abiertos como base para sus productos. Entre estas soluciones de grandes empresas se pueden mencionar: Oracle Big Data Appliance, IBM Big Data Platform, Microsoft Big Data, Big Data on Amazon Web Services (AWS) o Google Big Query. A día de hoy, las cuatro primeras utilizan Apache Hadoop como base, y lo enriquecen con conectores, diferentes mecanismos de almacenamiento NoSQL, herramientas, soporte en la nube, etc. La última funciona como servicio en la nube, aportando también diferentes herramientas auxiliares, teniendo su fuerte en Dremel-BigQuery, un servicio de consulta que permite ejecutar consultas SQL sobre bloques muy grandes de información.

El lector interesado puede encontrar más información sobre estos productos en las páginas web de estas empresas. El tratamiento masivo de datos se reconoce como un área en fuerte expansión en los próximos años, por lo que estas grandes empresas del software y los servicios tratan de posicionarse en el mercado utilizando sistemas de código abierto como base de sus productos, ofreciendo mejoras, añadidos, servicios y consultoría a partir de ellos.

Este puede ser un buen punto para la reflexión sobre los modelos de código abierto; en su definición básica, el código abierto es software que se desarrolla y distribuye de forma abierta y gratuita, lo que significa que no se cobra licencia por su uso, y el código fuente no se mantiene en secreto, de forma que cualquiera puede cambiarlo y publicarlo de nuevo. Aunque en sus inicios en la década de los 80 y 90 se generó una gran discusión sobre sus posibilidades y amenazas para la industria del software, se reconoce a partir de los 2000 que el código abierto permite soportar la colaboración de empresas grandes, pequeñas, comunidades de desarrollo sin ánimo de lucro, desarrolladores independientes; todo ello para la creación de ecosistemas de valor sostenibles. Aparte de los muy conocidos éxitos derivados de los sistemas operativos derivados de Linux (incluyendo Android), o navegadores Web, el código abierto ha logrado alcanzar su pleno potencial a través de la creación de fundaciones como Apache o Eclipse, que agrupan grandes comunidades con multitud de proyectos en los que colaboran grandes empresas como las que se mencionan un par de párrafos más arriba. Precisamente es la fundación Apache la que da soporte al proyecto Apache Hadoop, piedra angular del tratamiento masivo de datos; y empresas que tradicionalmente eran reacias a participar en estas iniciativas no sólo aportan conocimientos y personal, sino que colaboran en establecer la evolución de estos sistemas, una vez superadas las reticencias respecto al modelo de negocio y la seguridad.

En cuanto a los casos de éxito del tratamiento masivo de datos en el mundo civil, un estudio del MIT Technology Review muestra que es la clave de algunos modelos de negocio emergentes, como los anuncios sociales de Facebook, el motor de recomendaciones de Netflix o el esfuerzo de análisis comercial de Walmart. Esta empresa es pionera en el uso de Hadoop para la explotación de datos comerciales; ha

creado herramientas que analizan las compras en sus tiendas y las cruzan con datos públicos de la web y de redes sociales para hacer recomendaciones en relación con los intereses de cada usuario, ofreciendo información de localización de la tienda más cercana que tiene el producto. También esta empresa ha sido pionera en soluciones crowdsourcing para el lanzamiento de productos frente a grandes audiencias de Internet.

También se prevé el uso de Big Data en los sectores energético, aeronáutico, aeroespacial y automotriz: uso de información recopilada de usuarios para optimizar la cantidad de energía que se produce, optimización de costes en la cadena de suministro (reducción de tiempo de fabricación entre un 30% y un 50% en Fiat, Toyota y Nissan). La empresa General Electric tiene previsto monitorizar las turbinas de avión que fabrica y podrá procesar hasta 30Tb por vuelo, con lo que podrá conocer el estado de los motores en tiempo real.

2.6.- Explotación de datos. Aprendizaje y toma de decisión automáticos.

En la mención de las herramientas y de productos hecha hasta aquí se ha obviado el hecho de que al fin y al cabo lo que se va a manejar es información. Todos somos conscientes de la importancia que la información, en su acepción más amplia, ha ido adquiriendo en las últimas décadas. Incluso, algunos estudiosos van un paso más allá y caracterizan este periodo de tiempo, precisamente, por este concepto, y así nos encontramos ante la "era de la información".

Sin embargo, conviene profundizar en qué debemos entender por información, estableciendo las diferencias entre "dato" e "información". Si bien es posible encontrar varias definiciones de estos conceptos dependiendo de la aproximación a los mismos, de forma genérica podemos establecer que dato es la representación, en cualquier formato, de un atributo o característica de una entidad, proporcionando una descripción de un aspecto de la misma. Por otro lado, información es el resultado de un proceso de organización y procesado de un conjunto de datos, del que se obtiene como resultado un mensaje que permiten ampliar el nivel de conocimientos del sujeto o sistema que lo recibe.

Desde la aparición de los primeros computadores hasta nuestros días, la recogida de datos ha evolucionado de forma exponencial. La aparición de diferentes dispositivos para la recogida de datos como escáneres y videocámaras y la multiplicación de soportes de datos, como los diferentes tipos de tarjetas (tenemos una tarjeta prácticamente para cada tipo de transacción que realizamos), ha hecho posible que en un periodo relativamente corto de tiempo hayamos pasado de recoger información basada en caracteres alfanuméricos y por medio de un terminal informático, a estar generando y recogiendo datos, en varios formatos, prácticamente de forma continua en acciones tan cotidianas como comunicarnos por medio de terminales móviles, ordenar transacciones financieras, realizar compras en supermercados, etc. Según la información disponible en la web

oficial de IBM, en la que se cuantifica esta recogida de datos, cada día se recogen 2.5 trillones de bytes de datos, lo que es equivalente a que el 90% de los datos almacenados a día de hoy han sido recogidos en los dos últimos dos años.

Ya que el acopio de datos por sí mismo no proporciona ningún avance en el nivel de conocimiento de quien lo realiza, paralelamente a este almacenamiento se han desarrollado un conjunto de procedimientos con el objetivo de explotar los mismos. Esta explotación debe ser entendida fundamentalmente como la obtención de información que permita aumentar el nivel de conocimiento, como elemento básico para la toma de decisiones, en el ámbito en el que se desenvuelve la organización que tiene acceso a estos conjuntos de datos y los explota. El desarrollo de estos procedimientos, entre otros conjuntos de técnicas con objetivos similares, se ha realizado por investigadores con perfiles profesionales diferentes, aunque la mayoría de ellos se encuadran en disciplinas como las matemáticas (estadística e investigación operativa), ingeniería e informática, dando lugar a una nueva área de conocimiento conocida como inteligencia artificial (existen otras denominaciones también admitidas con carácter general, dependiendo de la aproximación a este tipo de procedimientos, como aprendizaje estadístico). Algunos de los trabajos teóricos que sirven de base a estos procedimientos son anteriores al desarrollo de la capacidad de cálculo proporcionada por la herramienta computacional, y la potencia y aplicabilidad de los mismos no se ha desarrollado hasta contar con este apoyo imprescindible.

Dentro de este ámbito de la explotación de los conjuntos de datos con el fin de extraer conocimiento, es posible distinguir diferentes tipos de tareas, cada una de las cuales constituye un tipo de problema a ser resuelto por técnicas normalmente distintas. Si bien todas estas tareas se centran en trabajos a partir un conjunto de datos o *dataset*, cada una de ellas tiene sus propios requisitos y procedimientos. Como consecuencia, el tipo de información a extraer de cada una de estas tareas puede diferir de la extraída del resto.

Se puede establecer una primera clasificación entre estas tareas, diferenciando aquellas que son de tipo predictivo (clasificación y regresión) de las que son descriptivas (clustering y asociación). A continuación se propone una definición con carácter introductorio de las mismas, puesto que serán presentadas con mayor detalle en las páginas siguientes:

- Clasificación: En ella, cada registro del conjunto de datos pertenece a una clase, indicada mediante el valor de un atributo llamado clase del registro. Este atributo puede tomar varios valores discretos, cada uno de los cuales caracteriza a cada clase. El objetivo de los algoritmos empleados en esta tarea es asignar los valores de clase a registros no clasificados, minimizando los errores de clasificación, y valiéndose para ello de aquellos registros del *dataset* (forman el denominado conjunto de entrenamiento o aprendizaje) en los que sí aparecen todos los atributos.

- Regresión: Tarea predictiva consistente en asignar un valor real a un atributo de un registro a partir de los valores, también reales en su versión más general, del resto de atributos del registro. La asignación se realiza mediante una función real, construida a partir de los registros del *dataset* que sí contienen todos los atributos, tanto el que se ha de predecir como el resto, denominados predictores. El algoritmo de construcción de esta función real se basa en minimizar el error entre el valor predicho y el real de estos registros, que forman el conjunto de entrenamiento de esta tarea.
- Clustering: Esta tarea consiste en dividir o segmentar el conjunto de registros. El objetivo de este tipo de algoritmos es diseñar los clústeres o grupos de tal forma que los registros incluidos en uno de ellos se caracterizan por su gran similitud o cercanía (en el sentido más amplio del término), mientras que entre los grupos (considerados de forma global) no existen similitudes o se encuentran suficientemente alejados.
- Asociación: El objetivo de esta tarea es identificar relaciones no explícitas entre atributos categóricos. Mediante las reglas de asociación no se establece ningún tipo de relación causa efecto, sino que el objetivo de esta tarea son aquellas relaciones que se establecen en función de la ocurrencia conjunta de determinados atributos.

2.6.1.- Depuración, exploración y visualización

Si bien los avances técnicos permiten la adquisición de datos de una forma ágil, estos no están exentos de imprecisiones, incoherencias y errores. Por lo tanto, como paso previo a la aplicación de cualquier procedimiento de obtención de información de un conjunto de datos es imprescindible someterlos a un proceso de depuración para minimizar los efectos perversos de estas disfunciones. En general, este preproceso precisa de recursos y tiempo, pero permiten añadir mayor precisión a la información y conocimiento posteriores.

También resulta muy aconsejable una exploración de los mismos una vez depurados. Son varios los objetivos que se pueden conseguir con estas técnicas. En primer lugar la exploración de los datos permite la toma de contacto y la familiarización con los mismos. Esta exploración se puede realizar de dos formas principalmente, siendo éstas complementarias. En primer lugar la visualización de los datos mediante el empleo de procedimientos gráficos también van a permitir profundizar en el conocimiento general del conjunto de datos. En segundo lugar. Es aconsejable realizar una exploración mediante el empleo de procedimientos numéricos descriptivos simples y de aplicación directa al conjunto de datos.

Si en el análisis exploratorio de datos se perciben evidencias de información redundante o si bien el volumen del conjunto de datos resulta excesivo para las capacidades de los posteriores procesos de tratamiento, se hace aconsejable cuando no

necesario reducir la dimensión de este conjunto. Entre las principales técnicas para llevar a cabo esta reducción, podemos citar las siguientes.

- **Análisis de componentes principales**

Posiblemente se trate de la técnica más empleada para reducir la dimensión de un conjunto de datos y el objetivo de la misma es reducir los atributos $x_1,x_2,\ldots,x_m$ de los registros, entre las que existe cierto nivel de correlación, en un segundo conjunto de atributos $z_1,z_2,\ldots,z_p$ incorreladas, de tal forma que $p < m$. Desde un punto de vista geométrico, esta transformación se corresponde con una proyección. Como es lógico existe un gran número de trasformaciones posibles, pero la que permite una mayor reducción con una pérdida menor de información (en todo proceso de reducción hay que asumir cierto coste en este sentido) es aquella en la que los nuevos atributos z son independientes entre sí, y están construidos de tal manera que z_1 es más relevante que z_2, a la vez z_2 que es más relevante que z_3 y así sucesivamente hasta llegar a que el atributo z_{p-1} es más relevante que z_p.

Más formalmente se considera un conjunto de registros de la forma $(x_1,x_2,\ldots,x_n)$ y tratamos de calcular, a partir de ellos un nuevo conjunto de atributos $(y_1,y_2,\ldots,y_n)$ incorreladas entre sí, con varianzas decrecientes y de tal forma que cada atributo sea combinación lineal de los atributos x , es decir, , $y_j=a_{j1}x_1+a_{j2}x_2+\ldots+a_{jn}x_n$, o expresado en formato matricial $y=A.x$ donde A es una matriz de dimensión $n \times n$. Luego el problema, en este punto pasa por determinar los valores de los elementos de la matriz A de tal forma que la trasformación verifique los requisitos ya mencionados. Para ello se exigen criterios de ortogonalidad en la transformación, y mediante el empleo de multiplicadores de Lagrange, se llega a que los valores de los elementos de A se obtienen mediante los autovalores de la matriz de covarianzas de la transformación.

Tras resolver este problema, aún no se ha efectuado la reducción del número de atributos del dataset, que sigue siendo n. Esta reducción se realiza mediante la selección de las p primeras componentes del nuevo vector de atributos obtenido tras aplicar la transformación anterior. Tras este proceso de selección, es muy importante evaluar la "calidad" de la reducción de la transformación, entendida cómo perdida de información respecto al conjunto original de datos. Para ello se emplea una técnica basada en el conjunto de autovalores de la transformación $e=(e_1,e_2,\ldots,e_n)$ donde el valor de cada uno de ellos representa la relevancia de cada uno de los nuevos atributos. La suma de los componentes de este vector coincide con la dimensión del conjunto de atributos inicial, n, por lo que es muy fácil pasar de valores absolutos a porcentajes para una mayor comprensión de la importancia relativa de los mismos. A partir de este punto, la selección de los p atributos de interés ya es decisión del analista, que de esta forma conoce la relevancia del nuevo conjunto de atributos que decida seleccionar. Aunque igualmente, se puede introducir como parámetro un valor mínimo de varianza explicada por la selección del nuevo conjunto de atributos, consiguiendo así un mayor grado de automatización del proceso.

- **Análisis factorial**

El objetivo principal de las técnicas de análisis factorial es analizar la estructura de las interrelaciones entre un gran número de atributos de un conjunto de datos. Utilizando esta información se calcula un conjunto de dimensiones latentes, conocidas como factores, que tratan de explicar estas interrelaciones. Al ser el número de factores sensiblemente menor que el de atributos originales del conjunto de datos, se consigue la reducción de la dimensión del mismo.

Para conseguir este objetivo, los factores comunes han de explicar las covarianzas o correlaciones entre los datos, y cualquier porción de la varianza inexplicada por estos se asigna a términos de error residuales, llamados factores únicos o específicos. El análisis factorial puede ser exploratorio o confirmatorio. El primero de ellos, que resulta ser el más interesante desde la perspectiva de Big Data, se caracteriza porque, a priori, no se conoce el número de factores. Aunque pueden realizarse análisis factoriales con variables discretas y/u ordinales, donde mejores resultados se obtienen de esta técnica es en el caso de variables cuantitativas continuas. Los resultados del análisis factorial no son invariantes a cambios de origen y escala, por lo que es recomendable estandarizar los datos antes del comenzar este análisis.

Desde un punto de vista más formal, la expresión por la que se rige este tipo de análisis es del tipo $x_i=a_{i1}F_1+a_{i2}F\_2+\ldots+a_{ik}F_k+u_i$, donde $F_1,F_2,\ldots,F_k$ con $(k<<n)$, donde n representa el número de atributos del dataset, son los factores comunes u_i es el factor único o específico asociado al atributo x_i y los coeficientes a_{ij} reciben el nombre de cargas factoriales. Por lo que respecta a la varianza de cada uno de los atributos, se descompone en dos partes: la primera, denominada comunalidad, representa la varianza explicada por los factores comunes, mientras que el segundo elemento de la descomposición representa la parte de la varianza de los atributos explicados por los factores únicos o específicos, y recibe el nombre de especificidad.

La determinación de los factores es, sin duda alguna, la principal cuestión a resolver en este tipo de análisis. Se pueden citar, entre otros, los siguientes procedimientos: método de componentes principales, de ejes principales, de máxima verosimilitud y de mínimo cuadrados, ya sean ponderados o generalizados. Entre estos, el método de las componentes principales, que resulta ser el empleado con mayor frecuencia, consiste en estimar las puntuaciones factoriales mediante las puntuaciones tipificadas de las primeras componentes principales y la matriz de cargas factoriales mediante las correlaciones de los atributos originales con dichas componentes.

2.6.2.- Aprendizaje supervisado

El objetivo del aprendizaje supervisado se concentra en el estudio de los modelos que permiten predecir el comportamiento de un atributo, denominado de salida, respuesta o output, como función de uno a varios atributos, denominados explicativos, entrada o input.

Cuando el atributo de salida es de tipo numérico, generalmente de carácter continuo, se habla de modelos de regresión, mientras que en el caso de atributos de salida de naturaleza categórica se hablad e modelos de clasificación.

Para construir la función que permite establecer el valor de los atributos respuesta, a partir de los atributos de entrada, y que según el caso recibe el nombre de función de regresión o clasificación, es necesario disponer de un conjunto de datos en el que se incluyan tanto los valores de los atributos explicativos, como el valor del correspondiente atributo de salida, y que recibe el nombre de conjunto de entrenamiento. Este conjunto de datos se divide en dos grupos, normalmente en la proporción 2/3 y 1/3. El primero de ellos, que recibe el nombre de conjunto de aprendizaje, tiene por finalidad la construcción del modelo predictivo, mientras que el segundo grupo de datos recibe el nombre de conjunto de prueba o test, y tiene por objetivo comprobar el grado de eficiencia o calidad del modelo.

- **Modelos de Clasificación**

Tal y como ya se mencionó en el apartado anterior, en el caso del aprendizaje supervisado los registros del conjunto de aprendizaje son de la forma , donde x es un conjunto de atributos e y es el atributo respuesta, que en el caso específico de los modelos de clasificación, se suele denominar "etiqueta" de la clase o categoría. Esta etiqueta de clase debe ser un atributo categórico.

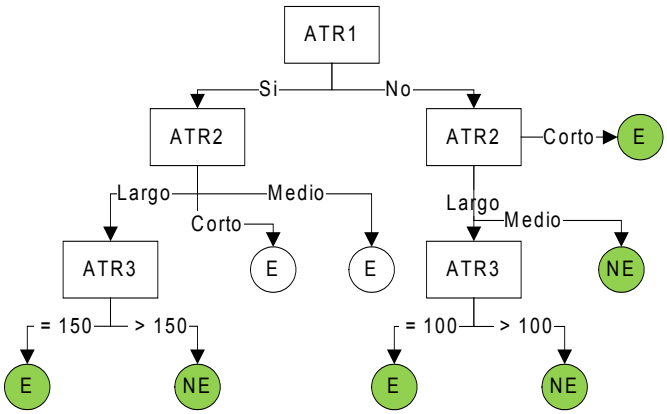
La primera tarea a desarrollar en un procedimiento de clasificación, tras la partición del dataset disponible en conjunto de aprendizaje y de prueba, es el propio proceso de aprendizaje. De forma general, el resultado de este proceso se sustancia, de forma general en la denominada función de clasificación o clasificador. Tras esta fase de aprendizaje, sigue una segunda, cuya finalidad es comprobar la eficiencia, "calidad" o "bondad" del clasificador. Para ello este clasificador se aplica a los atributos de entrada del conjunto de prueba, asignando a cada uno de los registros el valor de clase y comprobando, finalmente la coincidencia entre los valores reales, conocidos, con los asignados por el clasificador, mediante la conocida como matriz de confusión.

El esquema general del modelo de clasificación se puede ver de forma esquemática en la Ilustración 1.

Con este esquema general, son varias las técnicas que se pueden emplear en el proceso de aprendizaje, obteniendo en cada caso un clasificador con formatos diferentes. Entre las más empleadas podemos citar las siguientes:

- **Árboles de clasificación**

Sin duda alguna es una de las técnicas más empleadas para diseñar clasificadores, junto a los clasificadores basados en reglas. A ello contribuye fundamentalmente la facilidad de interpretación, ya que el clasificador se materializa en un diagrama de árbol. Un árbol se compone de nodos y arcos. Los nodos pueden ser raíz, que no tienen ningún arco de entrada y varios arcos de salida, nodos intermedios que tienen un único arco de entrada y varios de salida y nodo terminales u hojas, que tiene un único nodo de entrada y ninguno de salida. Por su parte el papel de los arcos es unir nodos. En un árbol, cada nodo terminal está asignado a un valor de la variable categórica que define el atributo de clasificación.



En la imagen lateral se puede observar la estructura de un árbol de clasificación, con

la variable de clasificación de dos etiquetas: "E" y "NE" y tres atributos.

En principio, el número de árboles de decisión que se pueden construir crece de forma exponencial conforme lo hace el número de atributos del dataset, y como es natural, existen unos más eficientes que otros. El empleo de procedimientos exhaustivos para la determinación del árbol de máxima eficiencia en el ámbito de Big Data es impracticable debido al grado de complejidad. Por ellos se han desarrollado con un grado de eficiencia más que razonable y con un coste computacional razonable que proporcionan árboles subóptimos. Estos algoritmos emplean, entre otros procedimientos, los de tipo greedy con decisiones localmente óptimas respecto al atributo y el valor del mismo que se ha de emplear para particionar los datos recogidos en cada nodo no final. Entre estos algoritmos destaca el de Hunt.

Posiblemente, la decisión más importante que hay que adoptar en la aplicación de estos procedimientos es la determinación del mejor criterio de separación en cada nodo no final, que determinarán la caracterización de cada arco que conduce a sus arcos de nivel inferior, denominados nodos hijos. Uno de los criterios más empleado para la adopción de esta decisión es la valoración del denominado grado de impureza de un nodo. Sí, se considera un clasificador bicategórico, el grado de impureza se representa por un par de la forma $(p_i;p_j)$, donde p_i representa los registros catalogados con la primera etiqueta de la clase, mientras que p_j representa los catalogados con la segunda etiqueta. Obiamente, se verifica $p_i;p_j=1$. Cuanto menor sea el grado de impureza, el nodo, desde el punto de vista del clasificador, tendrá un comportamiento mejor. Así, un nodo con un grado de impureza (0;1) tiene un grado de impureza nulo, mientras que un valor de la forma (0.5;0.5) representa el mayor grado de impureza posible.

Si bien esta metodología focaliza su entorno en variables categóricas, es posible aplicarla al caso de otro tipo de atributos, como los numéricos continuos, mediante procedimientos de discretización.

Igualmente, cuando el grado de complejidad de los árboles crece de forma desmesurada en detrimento de la eficiencia del modelo, existen ciertos procedimientos para disminuir esta complejidad, normalmente a costa de cierto grado de imprecisión en la clasificación. Estos procedimientos se basan en técnicas de acotación.

- **Clasificadores basados en reglas**

Los clasificadores basados en reglas son empleados con gran frecuencia en multitud de contextos. También se trata de un clasificador que goza de una fácil interpretación y, como consecuencia, la aplicación del mismo es inmediata. Un clasificador de esta naturaleza está formado por un conjunto de expresiones, que responden a un esquema básico del tipo "Si….entonces…". Más formalmente, una regla de las que conforman este tipo de clasificadores adopta la forma $(A_1\ op\ v_1) /\backslash (A_2\ op\ v_2) /\backslash \ldots /\backslash (A_k\ op\ v_k) \dashrightarrow y_i$, d, donde $(A_j;v_j)$ representa que el atributo A_j toma en valor v_j, op es un operador lógico seleccionado entre los que forman el conjunto $\{=;\neq;<;>;\leq;\geq\}$ e y_1 que representa el valor de la etiqueta de clasificación. De esto modo, a continua-

ción, se presenta un clasificador basado en reglas, que tal como se puede apreciar, es exactamente el mismo clasificador que se presentó a modo de ejemplo de árbol de clasificación:

R1: **Si** *(Atr2=Largo* ∧ *Atr1=Si* ∧ *Atr3150)* **entonces** *Clase=E*

R2: **Si** *(Atr2=Largo* ∧ *Atr1=Si* ∧ *Atr3150)* **entonces** *Clase=NE*

R3: **Si** *(Atr2=Largo* ∧ *Atr1=No* ∧ *Atr3100)* **entonces** *Clase=E*

R4: **Si** *(Atr2=Largo* ∧ *Atr1=No* ∧ *Atr3100)* **entonces** *Clase=NE*

R5: **Si** *(Atr2=Medio* ∧ *Atr1=Si)* **entonces** *Clase=E*

R6: **Si** *(Atr2=Medio* ∧ *Atr1=No)* **entonces** *Clase=NE*

R7: **Si** *(Atr2=Corto)* **entonces** *Clase=E*

Las estrategias para la extracción de reglas pueden ser de dos tipos: los métodos directos, que se caracterizan por la extracción de las reglas de clasificación directamente del conjunto de datos, donde el algoritmo de cubrimiento secuencia es de los más empleados de este tipo de métodos, y los denominados métodos indirectos, que extraen estas reglas a partir de técnicas más propias de otros tipos de clasificadores, como los árboles de decisión o las redes neuronales.

La calidad de los clasificadores basados en reglas se puede evaluar mediante varios tipos de medidas, entre las que destacan el cubrimiento y la precisión. Dado un conjunto de datos ***D*** , y una regla de clasificación ***r:A--->y***, se define el cubrimiento de la regla ***r*** como la proporción de registros del conjunto que verifican el antecedente de la regla, mientras que la precisión de la regla se define cómo la fracción de aquellos registros de que satisfacen la regla entre aquellos otros registros que solo satisfacen el antecedente de la regla .

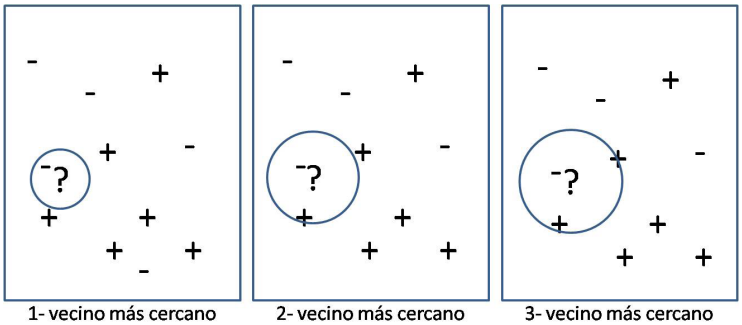
Formalmente:

$$\text{Cubrimiento } \frac{|A|}{D} \qquad \text{Precisión } \frac{|Any|}{|A|}$$

- **Clasificadores basados en el vecino más cercano**

Este procedimiento de clasificación es muy diferente a los dos anteriores. En concreto, el clasificador es un algoritmo que debe ejecutarse cada vez que sea preciso realizar una tarea de clasificación, dando como resultado de esta ejecución el valor de la etiqueta de la clase que corresponde a un nuevo registro sin clasificar. Es decir, No existe una función clasificadora obtenida como resultado de un proceso de aprendizaje. El principio de este procedimiento de clasificación aparece, no sin cierta sorna, en la literatura bajo el principio enunciado explícitamente como "si algo anda como un

pato, grazna como un pato y parece un pato, probablemente es un pato". Es decir, se trata de evaluar las similitudes entre el conjunto de atributos del registro a clasificar y los correspondientes a los atributos ya clasificados, asignando al nuevo registro la etiqueta de clase correspondiente a aquellos registros similares al mismo. La primera cuestión que surge de forma inmediata es: ¿cuántos registros similares o vecinos para continuar con la terminología propia de este tipo de algoritmos hay que considerar? Se trata de uno de los parámetros de ejecución del algoritmo, que ha de definir el decisor con anterioridad a la ejecución del mismo, además de definir criterios de actuación en caso de que se produzca un empate en la definición de la etiqueta de clase del nuevo registro.



Formalmente y una vez obtenido el listado de los vecinos más próximos, D_z al registro a clasificar x, éste último se clasificará con la con la etiqueta y' si:

$$y' = arg\ \max_{v} \sum_{(x_i,y_i)ED_z} I(v=y_i)$$

Donde v es el conjunto de etiquetas de la clase, y_i es la etiqueta correspondiente al vecino más próximo asociado con cada sumando e I es la función indicatriz.

- **Clasificadores bayesianos**

En algunos problemas de clasificación, la relación entre el conjunto de atributos y la categoría de la clase no es determinista, aunque sea posible encontrar casos en idénticos en el conjunto de datos de entrenamiento. Esta situación emerge por la presencia de los denominados datos "ruido" o por la propia naturaleza de los valores que pueden adoptar algunos atributos del dataset. Por lo tanto, el elemento distintivo de este procedimiento de clasificación es la existencia de un carácter probabilístico en el patrón que relaciona el conjunto de atributos y la etiqueta de clase de los registro. Los principios teóricos de esta técnica de clasificación se encuentran el conocido teorema de Bayes.

En este escenario de aleatoriedad, tanto el conjunto de atributos x como las etiquetas de clase y son tratados como variables aleatorias y, por lo tanto, es posible determinar el valor de la probabilidad de la variable aleatoria y condicionada por el valor que adopte la variable aleatoria x, denotado por $P(y/x)$, y conocido como probabilidad a

posterior de ***y***, en contraposición a la probabilidad a priori de ***y***, denotado por ***P(y)***. Precisamente, el objetivo de la fase de entrenamiento del clasificador es la determinación de los valores ***P(y/x)*** para todas las posibles combinaciones de los valores de estas variables, aunque en ocasiones esto no será factible en su totalidad, ya que es posible que en el conjunto de datos de aprendizaje en particular y de entrenamiento en general, exista alguna combinación que no esté presente y, en consecuencia, no se podrá evaluar el valor de la probabilidad de éstas.

Una vez conocidas estas probabilidades, un nuevo registro sin clasificar ***x´*** será clasificado en función de la maximización de los valores de las probabilidades condicionadas ***P(y/x')***, es decir, será clasificado con el valor de la etiqueta de clase ***y´*** que satisfaga la relación:

$$y' = \underset{(y_i)}{arg\ max}\ P(y_i/x')$$

donde y_i representa a cada uno delos valores posibles de la variable categórica de clasificación ***y***.

Tal y como se mencionó anteriormente, tras la fase de entrenamiento es posible que aparezcan determinadas combinaciones de valores de atributos ***x´´*** y de la categoría de clase ***y´´***, cuyo valor de probabilidad condicionada ***P(y'/x")*** no se haya podido determinar al no haber aparecido ningún caso en el conjunto de entrenamiento. Esta circunstancia no supone la imposibilidad de construcción de este tipo de clasificadores. En este caso, es posible la aplicación de técnicas basadas en los principios por los que se rige la teoría de muestras y la inferencia estadística, para realizar una estimación de dichos valores, habilitando de nuevo la construcción del clasificador.

- **Redes neuronales**

Una red neuronal es una abstracción lógica-matemática inspirada en algunos aspectos de la fisiología y del comportamiento del cerebro humano en la transformación de uno o varios estímulos de entrada en una o más repuestas. Considerando el punto de vista de este procedimiento de clasificación, y con el propósito de emular los aspectos más relevantes de la estructura del cerebro humano, formado por neuronas interconectadas, una red neuronal está formada por sendos conjuntos de nodos y arcos dirigidos, todos ellos igualmente interconectados.

La arquitectura de una red neuronal en su versión más simple es conocida como perceptrón. Este tipo de perceptrones está formado por dos tipos de nodos: nodos de entrada y nodos de salida. Los nodos de entrada se emplean para representar los atributos de entrada al perceptrón, mientras que el nodo de salida sirve de soporte al modelo de salida que caracteriza al perceptrón. Estos nodos, que son los elementos que conforman las redes neuronales más complejas, reciben el nombre de neuronas o unidades.

En un perceptrón, cada nodo de entrada está conectado, mediante un arco ponderado, al nodo de salida. La ponderación de los arcos simula el valor de la intensidad de la conexión sinóptica entre las neuronas que conforman un cerebro real. La fase de aprendizaje de un perceptrón consiste, precisamente, en determinar los valores óptimos de las ponderaciones de los arcos, con el objetivo de conseguir el ajuste máximo en el patrón existente entre los atributos de entrada y la etiqueta de salida. En general, un perceptrón determina el valor de la etiqueta de salida correspondiente a un determinado conjunto de valores de los atributos de entrada, realizando la suma ponderada de estos valores de entrada y, en algunos casos, detrayendo de esta suma un factor de sesgo denotado por .

La fase de aprendizaje del perceptrón se realiza mediante un proceso iterativo, que se rige por la expresión $w_j^{(k+1)} = w_j^{(k)} + \lambda (y - \hat{y}_i^{(k)}) x_{ij}$ donde $w_j^{(k)}$ es el peso asociado con el arco correspondiente a la entrada j-ésima del perceptrón en la iteración k-ésima de la ejecución del algoritmo de aprendizaje, es un parámetro conocido como índice de aprendizaje y x_{ij} es el valor del atributo j-ésimo del registro de entrenamiento x_i. El índice de aprendizaje es un parámetro que toma valores entre 0 y 1, y se emplea para controlar la intensidad del ajuste permitido en cada iteración del algoritmo. Cuando los valores de este índice se encuentren próximos a cero, el nuevo peso estará fuertemente influenciado por el valor anterior, mientras que valores del índice próximos a 1 harán que el nuevo valor del peso sea especialmente sensible al ajuste de la iteración en ejecución.

Una red neuronal es una estructura sensiblemente más compleja que la correspondiente a un perceptrón, elemento fundamental para la construcción de la propia red. Este nivel de complejidad se puede alcanzar mediante dos vías. En primer lugar, la red neuronal puede estar formada por varias capas de perceptrones, situadas entre la capa de entrada y la de salida de la propia red neuronal. Estas capas intermedias reciben el nombre de ocultas. Este tipo de redes se denominan redes multicapas. En una red neuronal con control de entrada, los nodos de una capa solamente se conectan con los nodos de otra capa, mientras que cuando no existe esta restricción, se habla de redes recurrentes. En segundo lugar, el nivel de complejidad de una red neuronal puede aumentar cuando ésta emplea funciones de activación, además de la señal de entrada. Los ejemplos más empleados de este tipo de funciones incluyen a las lineales, sigmoidales y tangentes hiperbólicas. Estas funciones permiten a los nodos de capas ocultas y de salida producir valores de salida no lineales.

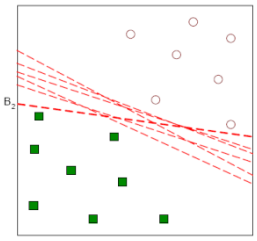
El aprendizaje de una red neuronal, que como ya se mencionó, trata de determinar los valores correspondientes del juego de pesos , emplean como estrategia para conseguir este objetivo la determinación de aquellos valores de que minimicen el error cuadrático medio, es decir, que minimizan la expresión

$1/2 \sum i=1 (y_i - \hat{y}_i)^2$, siendo una función del conjunto de valores w.

- *Máquinas de soporte vectorial*

Esta técnica de clasificación ha recibido una considerable atención en los últimos años, como consecuencia de haber mostrado resultados muy prometedores en muchas situaciones prácticas, especialmente en aquellas en las que el conjunto de datos presenta una dimensión muy alta. Otro aspecto importante de este procedimiento de clasificación, por el que también supera a otras alternativas, es que el clasificador resultante se basa en un conjunto reducido de registros del conjunto de aprendizaje, que reciben el nombre de vectores de apoyo.

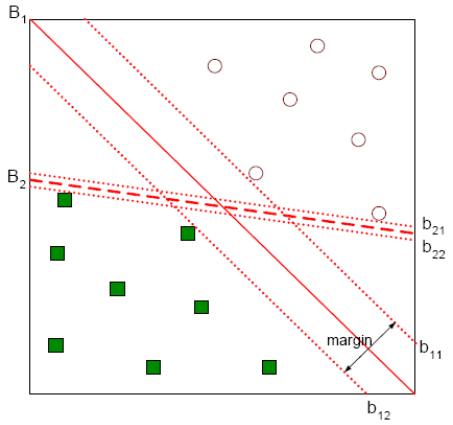
Uno de los elementos más característico de este procedimiento es el denominado hiperplano de margen máximo, alrededor del cual giran el desarrollo del mismo. Si se supone un conjunto de datos, cuyos registros están clasificados por medio de las etiquetas de un atributo categórico dicotómico y, además, son separables de forma lineal, es decir, es posible establecer un hiperplano de tal forma, que todos los elementos de una clase queden a un lado de este hiperplano mientras que los elementos de la segunda clase queden al lado contrario del mismo hiperplano.



Lo normal es que sea posible encontrar un número muy elevado de hiperplanos que cumplan con esta condición de separabilidad del conjunto de datos según el atributo de la clase (observar gráfico). A pesar de este número tan elevado de posibles hiperplanos, se espera que algunos de ellos se comporten mejor que otros en el momento de clasificar nuevos registros, tras la fase de entrenamiento. El clasificador se construye mediante la selección de uno de estos hiperplanos como umbral de separación entre clases para registros futuros.

Para facilitar la comprensión de este concepto se propone la figura siguiente. De entre todos los potenciales hiperplano que se muestran en la figura, se consideran dos de ellos como candidatos a clasificador, denotados como B1 y B2. Como se puede apreciar de forma gráfica, ambos hiperplano pueden separar según sus clases a todos los registros del conjunto de datos sin cometer ningún error en estas clasificaciones. Asociado a cada uno de estos hiperplanos existen otros dos hiperplanos b11 y b12 en el caso de B1, y b21 y b22 en el caso del B2, generados a partir de sus hiperplanos de referencia, alejándose por cada lado del mismo hasta tocar al primero de los registros del conjunto de datos que aparecen en su trayectoria. Para cada hiperplano de referencia, B1 y B2, se define como hiperplano de margen máximo a aquel cuya distancia entre los hiperplanos subordinados (b11 y b12 en el caso de B1 y b21 y b22 en el caso de B2) sea máxima. En este caso, es obvio que el hiperplano que verifica esta condición es B1. Los

hiperplanos con esta característica tienden a clasificar con menos errores los nuevos registros que aquellos con márgenes menores. De una manera más formal, decimos que este clasificador se basa en la minimización del riesgo estructural.



Consideremos un clasificador de tipo binario, para lo cual se cuenta con un conjunto de entrenamiento formado por N registros. Cada uno de ellos de la forma (x_i,y_i) de donde x_i es igual a $x_i=(x_{i1},x_{i2},\ldots,x_{ik})^T$, donde es el conjunto de k atributos del registro con ordinal i. Por convención $y_i \in \{1,-1\}$, denota la clase correspondiente al registro con ordinal i. El umbral de decisión, es decir, la frontera de clasificación de un clasificador lineal es de la forma $w.x+b=0$ donde w y b son parámetros del modelo cuyo valor es necesario determinar durante la fase de entrenamiento. Para ello, se plantea el siguiente problema de optimización con restricciones $\min_w \frac{\|w\|^2}{2}$

sujeto al conjunto de restricciones $y_i\,(w.x_i+b)\geq 1 \qquad i=1,2,\ldots,N$

Al tratarse de una función objetivo cuadrática con restricciones lineales, puede ser resuelto mediante el problema dual del modelo de optimización basado en los multiplicadores de Lagrange. Una vez resuelto este problema, y los parámetros del hiperplano son conocidos, el clasificador es una función f, que ante la presencia de un nuevo registro lo clasifica según la valoración de la siguiente función: $f(z)=sign(w.z+b)$

En el caso de datos no separables, es necesario reformular el procedimiento, permitiendo algunas violaciones de las restricciones anteriores. En concreto, es posible establecer un equilibrio entre la anchura del margen y el número de errores permisibles en la fase de entrenamiento. Esto se consigue mediante el empleo de variables de holgura E_i, modificando el planteamiento original del problema original, dando lugar al siguiente: $\min_w \frac{\|w\|^2}{2} + C\left(\sum_{i=1}^{n} E_i\right)^k$

sujeto al conjunto de restricciones donde

$(w.x_i+b) \geq 1-E_i$ "Si" $y_i=1$

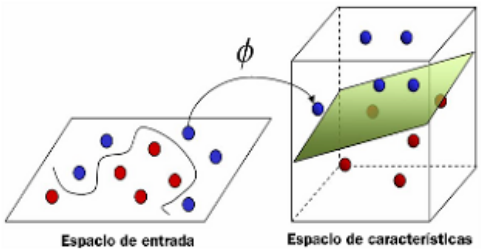
$(w.x_i+b) \leq -1+E_i$ "Si" $y_i=-1$

donde C y k son parámetros de penalización por los errores en la clasificación de los registros de entrenamiento.

La formulación descrita en los párrafos anteriores permite la determinación de un clasificador lineal. Sin embargo, no todos los conjuntos de datos admiten este tipo de clasificador, y será necesario aplicar un procedimiento para el caso, que es además el más general, de no linealidad del umbral de separación de los datos. La clave para dar respuesta a problemas de esta naturaleza se basa la transformación de los datos del conjunto original X en un nuevo espacio $P(X)$, de dimensión mayor que el de origen, de tal manera que en esta nueva dimensión sea posible emplear un límite de tipo lineal para separar los registros. Esta transformación se realiza mediante el empleo de las denominadas funciones kernel, que se aplican a los pares de registros $x,y \; E \; X$. Las funciones kernel, a la que denotamos como k, más empleadas son las siguientes:

- $k(x,y) = exp\left(\frac{-\|x-y\|^2}{c}\right)$
- $k(x,y) = \left((x.y)+\theta\right)^d$
- $k(x,y) = \tanh(k(x.y)+\theta)$
- $k(x,y) = \frac{1}{\sqrt{\|x-y\|^2+c^2}}$

donde $c, \theta, k \in \mathbb{R}$ y $d \in \mathbb{N}$.



Tras esta transformación se puede emplear el procedimiento descrito anteriormente para la construcción del clasificador.

Así, en esta imagen queda reflejado la forma en que se puede pasar de un conjunto

de datos en dimensión 2, que no son separables linealmente, a una representación de los mismos en dimensión 3, en la que sí es posible establecer una frontera de clasificación lineal de los mismos.

- **Modelos de regresión**

Los modelos de regresión constituyen un procedimiento predictivo empleado fundamentalmente cuando tanto el atributo respuesta como los atributos explicativos son de naturaleza cuantitativa. En el caso de que el atributo explicativo sea único, se habla de regresión simple, mientras que si se consideran varios de estos, se habla entonces de regresión múltiple.

La forma general de este tipo de modelo descompone los valores que toma el atributo a explicar o atributo respuesta de cada registro en dos componentes, uno función de los atributos explicativos del registro, y un segundo componente, específico del registro. Así, la expresión general de los modelos de regresión es $y_i=r(x_{i1},x_{i2},\ldots,x_{ip})+E_i$, donde r es la función que relaciona los atributos explicativos con el atributo explicado, y E_I es la parte específica del registro que no puede ser explicado mediante el vector de atributos x_i. En el caso de los modelos de regresión, la estimación de la función , se realiza de tal forma que, en promedio, minimice el error cuadrático medio de las desviaciones de los registros reales respecto a los estimados en la fase de entrenamiento.

El modelo más simple de estos modelos es la denominada regresión lineal, que se puede expresar mediante la expresión

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

Por lo tanto, la construcción del modelo exige la determinación de los valores $B_0,B_1,\ldots,B_p$. Para ello, el punto de partida es el conjunto de entrenamiento, en la que aparecen asociados en un mismo registro los valores del atributo explicado y los atributos explicativos . A partir de estos registros se plantea el siguiente problema de minimización:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Además del modelo lineal descrito anteriormente, es posible construir los denominados modelos no lineales. Estos últimos se caracterizan por relajar la hipótesis de linealidad de la función en la expresión general del modelo de regresión $y_i=r(x_{i1},x_{i2},\ldots,x_{ip})+E\_i$. Como consecuencia de esta relajación, que permite a r adoptar cualquier forma, el número de variantes es prácticamente infinito. No obstante, la complejidad de los modelos crece de forma muy considerable, dejando de ser eficientes, especial-

mente en el caso valores de p que sean distintos de varias unidades y de grandes volúmenes de datos, escenario natural donde se encuentran los problemas a resolver en el marco del Big Data. A modo de ejemplo, y por ser uno de los modelos no lineales más empleados, se presenta el caso polinomial de grado n. En esta variante, la función que predice el valor de la variable y_i en función del vector de atributos x_i es de la forma:

$$y_i = \beta_0 + \beta_1 x_{i1}^l + \beta_2 x_{i2}^l + \cdots + \beta_p x_{ip}^l, \text{ donde } l \leq n.$$

En el caso particular de ser $n=2$ se habla de regresión cuadrática o parabólica y en el caso de ser $n=3$ se habla de regresión cúbica.

En los últimos años también ha despertado el interés de los investigadores los modelos de regresión en el caso de atributos respuesta de tipo categórico, mientras que en el caso de los atributos explicativos se abren posibilidades tanto a variables categóricas, además de las cuantitativas ya mencionadas. Sin duda alguna, uno de los modelos más relevantes es el conocido como modelo de regresión logística.

En este modelo logístico se considera que el atributo a explicar es de tipo categórico, siendo el caso con dos categorías el más importante, debido fundamentalmente a lo conveniente de sus propiedades matemáticas, su funcional y su fácil interpretación en término de probabilidad asociada al evento a explicar, aunque es justo mencionar que el modelo probit ha gozado de una implantación práctica muy similar.

Si inicialmente consideramos un modelo lineal con registros con un único atributo x_i, tendríamos una expresión del modelo (prescindiendo del factor no explicado por el atributo) de la forma $y_i = \beta_0 + \beta_1 x_{i1}$, que al representar gráficamente nos permitiría observar que la línea de regresión no está acotada en el intervalo **[0.1]**, y por lo tanto no va a representar una probabilidad. Una de las formas de soslayar este inconveniente, es realizar una transformación de la función anterior a una función de distribución probabilística, F, que sí que goza de esta propiedad, ya que para cualquier función de distribución se verifica, entre otras, la propiedad $0 \leq F(x) \leq 1$. En el caso de emplear la función de distribución logística, el modelo se denomina de regresión logística, y si se emplea la función de distribución normal se tiene el modelo de regresión probit.

El modelo de regresión logística parte de la hipótesis de que los datos siguen el siguiente modelo $\ln\left(\frac{y_i}{1-y_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = z.$.

Operando sobre esta expresión se llega a $y_i = \frac{e^z}{1+e^z}$ que evidentemente recuerda a la función de distribución logística. Es inmediato comprobar cómo este

modelo no se corresponde con los modelos lineales de regresión, pero sí lo es a escala logarítmica, es decir, la diferencia de probabilidad de que ocurra un suceso respecto de que no ocurra es lineal pero en escala logarítmica. En cuanto a la resolución del problema de la determinación de los parámetros, se suelen emplear varios procedimientos, aunque uno de los más habituales es el denominado método de Newton-Raphson.

2.6.3.- Aprendizaje no supervisado

En este tipo de aprendizaje se pretende la extracción de correlaciones interesantes, patrones frecuentes, asociaciones o estructuras casuales entre registros de un conjunto de datos. En este tipo de procedimientos el elemento predictivo del aprendizaje supervisado deja paso a un elemento descriptivo (aunque en ciertos casos la frontera entre estos elementos no está claramente definida), que siga el principio de "dejar que los datos hablen solos".

- **Asociación**

El objetivo de las denominadas reglas de asociación es el establecimiento de relaciones no explícitas basadas en la aparición conjunta de determinados grupos de atributos del conjunto de datos. Esta tarea se descompone normalmente en dos subproblemas. El primero de ellos consiste en encontrar los grupos de atributos cuyas ocurrencias simultáneas exceden un umbral predeterminado, que son calificados como frecuentes. El segundo problema consiste en generar las reglas de asociación propiamente dichas, a partir de los conjuntos de atributos frecuentes establecidos en el primer subproblema y que cumplen la condición de superar un nivel de confianza mínimo.

Formalmente, y empleando la nomenclatura y argot propio de este tipo de tareas, se denota por $I=I_1,I_2,\ldots,I_m$ al conjunto de atributos incluidos en el dataset D. Una regla de asociación es una implicación de la forma $X=>Y$ donde $X,Y \subset I$ son grupos de atributos que, en éste ámbito, reciben el nombre de itemsets, y además $X \cap Y = \emptyset$ .El itemset X recibe el nombre de antecedente, mientras que Y recibe el nombre de consecuente. La regla ha de entenderse en el sentido de que la presencia de los atributos incluidos en el itemset X en un determinado registro del dataset, implica la presencia de los atributos incluidos en el itemset Y del mismo registro.

Existen dos medidas fundamentales para caracterizar estas reglas de asociación, de forma que puedan establecerse como patrones o relaciones dentro del *dataset* y no resulten ser meras coincidencias no generalizables. Estas medidas son el denominado soporte mínimo, denotado por s , y mínima confianza c , y el umbral de las mismas son fijadas por el analista como parámetro del procedimiento.

El soporte de la regla de asociación $X=>Y$ se define como el porcentaje o fracción de registros que contienen $X \cup Y$ respecto al número total de registros del conjunto

de datos. Por otro lado, la confianza de la regla de asociación ***X=>Y*** se define como el porcentaje o fracción de registros que contienen ***X U Y*** respecto al número de registros que contienen a ***X***. Se trata de una medida de la robustez de la regla de asociación, así si la regla ***X=>Y*** tiene un nivel de confianza del 80%, implica que en el 80% de los registros en los que aparece ***X*** también aparece ***Y***. Estos conceptos se expresan mediante las siguientes expresiones:

$$s(X \Rightarrow Y) = \frac{Card(X \cup Y)}{m} \qquad s(X \Rightarrow Y) = \frac{Card(X \cup Y)}{Card(X)}$$

donde ***m*** denota el número de registros del *dataset.*

Siguiendo con el argot propio de las reglas de asociación, un conjunto de ítems recibe el nombre de itemset. El número de ítems que componen un itemset se denomina longitud del itemset, de tal forma que los itemsets de longitud ***k*** reciben el nombre de k-itemset. De forma general, el algoritmo para la extracción de las reglas de asociación, que recibe el nombre de algoritmo a priori y se basa en el principio del mismo nombre, es un proceso iterativo, cuyo criterio de parada viene marcado por la imposibilidad de encontrar itemsets de mayor tamaño, y cuyas etapas principales son:

1. El conjunto de candidatos a k-itemset se genera con extensiones de longitud ***1*** a partir de los (k-1)-itemsets de la iteración anterior.
2. Se comprueba que los candidatos a k-itemset superan el umbral fijado.
3. Se desechan los itemset que no superan este umbral. Los restantes se confirman como k-itemset.

Se trata de un algoritmo simple, pero su eficiencia es mejorable, ya que presenta dos grandes debilidades. En primer lugar solo permite la construcción de reglas de asociación en el que sólo exista un itemset en el consecuente y, en segundo lugar, presenta un nivel de complejidad exponencial que puede llegar a dificultar la extracción de reglas de asociación relevantes en un tiempo razonable Esto se debe fundamentalmente, y en primer lugar, a la generación de itemsets candidatos que es el proceso más exigente desde el punto de vista del tiempo de computación, del empleo de espacio y memoria del computador así como la necesidad efectuar un excesivo número de procesos de lectura de la base de datos. Las vías de mejora de la eficiencia de este algoritmo pasan, entre otras estrategias menos estudiadas, por las cinco siguientes:

- Reducción del número de consultas a la base de datos.
- Realizar un proceso de muestreo sobre el conjunto de datos, reduciendo la dimensión del dataset de trabajo.
- Añadir restricciones adicionales a la estructura de los patrones objetivo.
- A través de procesos de paralelización computacional.

Inicialmente, los conjuntos de datos a los que era posible aplicar el algoritmo a priori debían estar formados por atributos binarios. Cabe recordar que el esquema inicial de este problema surge de la denominada "cesta de la compra", cuyas transacciones se modelan mediante una matriz (dataset) cuyas columnas se asocian con los artículos susceptibles de ser adquiridos y las filas con los clientes que los adquieren. Si el elemento ***(i,j)*** de la matriz toma el valor "***1***" se interpreta como que el cliente i-ésimo ha adquirido el producto j-ésimo, mientras que si toma el valor "***0***" el sentido es el contrario.

Con posterioridad, esta restricción respecto al carácter de los atributos, se ha ido relajando. Mediante los procesos de transformación adecuados, se pueden establecer reglas de asociación cuando estos atributos son categóricos con más de dos categorías. Igualmente sucede con los conjuntos de datos en los que existen atributos continuos, que tras un pertinente proceso de discretización pueden ser sometidos a una tarea de extracción de reglas de asociación. Por último, también es necesario destacar la posibilidad de extracción patrones de asociación considerando el aspecto secuencial de los atributos, es decir, cuando el factor tiempo interviene en este proceso y, por lo tanto, además de apariciones conjuntas de atributos, es posible extender estos patrones a otros en los que intervenga, además, el orden de aparición de los mismos.

- **Análisis clúster**

El análisis clúster es la segunda tarea más frecuente de las encuadradas dentro del aprendizaje no supervisado. La finalidad de este tipo de análisis es la segmentación de un conjunto de datos en varios grupos o clústeres, de tal forma, que la distancia (entendida en el sentido más amplio del término) entre los distintos registros que componen cada clúster sea mínima a la vez que la distancia existente entre los diferentes clústeres sea máxima. A su vez, cada clúster puede ser subdividido, con este mismo tipo de técnicas, en subclústeres, obteniéndose de esta forma una separación jerárquica de los diferentes registros que componen el conjunto de datos, y cuya representación gráfica se conoce como dendograma. Este proceso de segmentación está basado de forma exclusivas en los atributos que describen a cada registro. En algunas ocasiones, este tipo de procedimientos suele ser un paso previo a otros tipos de análisis a realizar con los diferentes grupos así formados.

Desde un punto de vista estratégico, se pueden considerar varias formas de establecer los clústeres:

1. Bien separados: Según este principio, un clúster es un conjunto de registros en el que cada registro está más próximo (es más similar) a cualquier otro del mismo clúster que a registros de clústeres diferentes. En este tipo de estrategias suele ser habitual la definición de un determinado umbral de cercanía o similitud para fijar un criterio de cercanía de registros en un clúster. En general, esta estrategia, muy ideal, sólo se verifica cuando la naturaleza de los registros, por sí misma, favorece la cercanía de los registros de un clúster frente a los correspondientes

a clústeres muy alejados.

2. Basados en prototipos: En esta estrategia de construcción de clústeres, cada clúster es un conjunto de registros en el que todos son más similares a uno de ellos, definido como prototipo que define el clúster, que a cualquier otro prototipo de los restantes clústeres. En el caso de atributos continuos, el prototipo de un clúster es, a menudo, el centroide del mismo, es decir, la "media" de todos los registros que componen el clúster. En el caso de atributos categóricos, el centroide suele ser el valor más frecuente de este atributo en el clúster. En general, los clústeres así formados también reciben el nombre de basados en el centro.

3. Basados en grafos: Se trata de una estrategia que descansa en la representación de los registros mediante grafos. Así, cada registro está representado por un nodo del grafo, mientras que los arcos del mismo representan las relaciones entre los diferentes nodos (registros), por lo que cada clúster queda formado por aquellos registros conectados por el grafo de referencia. Es posible dotar de una métrica al grafo, de tal forma, que cada arco tenga asociado un valor que permita establecer la intensidad de la relación entre los nodos (registros) que une. Así, el diseño de los clústeres no sólo depende de la existencia o no de un nodo en el grafo, sino que además es posible añadirla condición adicional de que la intensidad del mismo supere un determinado umbral

4. Basados en la densidad: Siguiendo esta estrategia, un clúster es una región de alta densidad de registros rodeada de una región de baja densidad. Este tipo de estrategias es muy empleada cuando los clústeres son muy irregulares o se encuentran muy entrelazados, o cuando se ha detectado la presencia de ruido o outliers.

5. Basados en propiedades compartidas: De forma general, un clúster de esta naturaleza se define mediante una propiedad de la que gozan los diferentes registros que lo forman. Evidentemente, algunas de las estrategias anteriores puede ser clasificada dentro de este grupo (cómo por ejemplo el basado en prototipos que comparten la propiedad de mínima distancia al centro). Sin embargo, la estrategia basada en propiedades compartidas va más allá, y se centra en otro tipo de propiedades, que suelen derivar del conjunto integro de objetos.

3. Estudio de aplicaciones de Big Data en Defensa y Seguridad

3.1.- Introducción

Aunque la generación de la inmensa cantidad de datos, así como la necesidad de

procesarlos y explotarlos de manera eficaz y eficiente es transversal a toda nuestra sociedad, qué duda cabe que el ámbito de la defensa y la seguridad son dos de los entornos donde resulta interesante analizar cómo el Big Data puede aplicarse y ofrecer beneficios.

Por ello, en los próximos apartados se realiza este análisis que tiene por objetivo último la identificación de qué actividades, presentes y futuras, se pueden beneficiar del Big Data. Para comenzar, se refleja a continuación qué aspectos generales del contexto actual y futuro en defensa y seguridad, presentan un potencial interés desde el punto de vista de Big Data.

Tanto el ámbito de la defensa como el de la seguridad están marcados por un enfoque prioritario hacia la prevención. Prevenir siempre es mejor y menos costoso que curar. Sin embargo, la prevención requiere decisiones en ventanas de tiempo muy definidas y con un alto nivel de síntesis de la inmensidad de datos y factores involucrados. Por otro lado, los conflictos y/o crisis recientes y actuales han visto crecer su grado de complejidad, y todo hace prever que esta tendencia continuará. Ejemplos de esta complejidad son el mayor número e interconectividad de los actores y acciones, derivada de la globalización; los nuevos escenarios, con líneas divisorias muy difusas entre lo civil y lo militar; entornos intensivos en información con creciente mezcla de escenarios virtuales (ciberdefensa, económicos, etc.) y reales etc.

Para poder trabajar con la creciente complejidad y abundancia de datos, es necesario un mayor enfoque en la comprensión de la situación, especialmente en aquellos ámbitos donde los objetivos (blancos, enemigos, criminales, etc.) son en apariencia de pequeña escala y/o de carácter ambiguo. Esta difusión de los objetivos impone un desafío creciente para las capacidades en defensa y seguridad occidentales, ya que éstas se fundamentan en la habilidad de encontrar, prevenir y golpear en la fuerza enemiga (o elementos fuera de la ley, en el caso de la seguridad interior). Sin embargo, las amenazas presentes (y futuras) procuran utilizar el entorno actual, caracterizado por la congestión (de información, de alternativas, de actores, etc.) y por el ruido (falsas alarmas, indistinción civil-militar, entornos urbanos, etc.) para esconderse en el mejor de los casos, o para utilizarlo en nuestra contra de manera directa, en el caso peor.

Y es en estos desafíos donde la utilización de Big Data puede ofrecer mejoras en las capacidades actuales y soluciones a problemas, ya sea existentes o bien emergentes. Así, de manera genérica podemos decir que la aplicación de "Big Data" a defensa y seguridad persigue capturar y utilizar grandes cantidades de datos para poder aunar sensores, percepción y decisión en sistemas autónomos, y para incrementar significativamente el que el entendimiento de la situación y contexto del analista y el combatiente/agente del orden.

Por ello, no debería ser una sorpresa para el lector que en el ámbito de la seguridad y defensa Big Data se pueda aplicar (potencialmente) en áreas como las siguientes:

- Vigilancia y Seguridad de fronteras
- Ciberdefensa / Ciberseguridad
- Lucha contraterrorista y contra crimen organizado
- Lucha contra el fraude
- Seguridad ciudadana
- Inteligencia militar
- Planeamiento táctico de misiones.
- …

Desde este punto de partida, el estudio ha pretendido profundizar en la aplicación de Big Data a cada una de estas áreas. Para ello se han identificado 12 aplicaciones específicas directamente asociadas bien a necesidades/problemas en los que Big Data puede arrojar ventajas frente a otras soluciones tecnológicas, o bien asociadas a nuevas oportunidades que Big Data abre para la exploración de nuevas capacidades en seguridad y defensa.

Las aplicaciones específicas identificadas son:

1. Detección de intrusión física en grandes espacios o infraestructuras abiertas
2. Computación sobre información encriptada
3. Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos)
4. Criminología computacional
5. Uso fraudulento de recursos corporativos y/o sensibles
6. Análisis de video en tiempo real / Búsqueda y recuperación rápida en librerías de video.
7. Inteligencia visual en máquinas
8. Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos.
9. Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos.
10. Consciencia situacional
11. Traducción automática a gran escala (en número de idiomas y en volumen)
12. Predicción de eventos

De manera complementaria a estas perspectivas de aplicación, se ha realizado un análisis de los factores, tanto inherentes a Big Data como contextuales, que de manera prospectiva se entiende que suponen o bien un freno o bien un empuje al desarrollo y aplicación de Big Data a Seguridad y defensa.

3.2.- Aplicaciones específicas

3.2.1.- Detección de intrusión física en grandes espacios o infraestructuras abiertas

Las técnicas y planteamientos de Big Data son aplicables a la seguridad física de instalaciones o infraestructuras. De especial interés por las elevadas dificultadas técnicas, o de costes es la detección de intrusión física en grandes espacios o infraestructuras que cubre una gran extensión. Esto conlleva la detección de posibles amenazas, su clasificación, y en el caso necesario, su localización y seguimiento. El incremento del área implica o bien más sensores o bien mejores resoluciones en estos sensores, o generalmente una combinación de ambos, lo que en cualquier modo se traduce en una explosión de datos. Esta cantidad de datos, y la necesidad de de monitorizar en tiempo real de manera automática hace que este problema sea un candidato para Big Data.

De hecho, así lo pensaron en IBM cuando pusieron en marcha el proyecto Terraechos para el Laboratorio Nacional del Departamento de Energía de EE.UU. El sistema resultante de este proyecto debe consumir y analizar de manera continua grandes cantidades de datos acústicos provenientes de los objetos en movimiento del entorno del perímetro y área a vigilar, tanto sobre la superficie como debajo de la superficie. Para cumplir con su cometido el sistema debe discriminar en tiempo real sonidos provenientes de animales, arboles, climatología, de posibles intrusos. Para ello, implementaron una red de sensores acústicos provenientes de la US Navy, que alimentaban continuamente con terabytes de información un procesador (IBM InfoSphere Streams) que analiza los datos acústicos.

Está previsto que la demanda y oferta de este tipo de aplicaciones aumente en los próximos años. Hay que tener en cuenta que el coste de la adquisición de datos en lo que respecta al sensor se reduce de manera constante, no así el coste del procesamiento, debido a la escalada de la complejidad con el número de sensores. En cualquier caso, la complejidad de las soluciones variará, y no en todas ellas será necesario sensores de última generación, ni sistemas punteros de procesamiento como el propuesto por IBM. En lo que respecta a la demanda, tanto en el ámbito de la seguridad de los ciudadanos con el empuje de la Comisión Europea a la protección de infraestructuras críticas, como en el ámbito de la protección de las fuerzas desplegadas, con las operaciones en entornos mixtos que demandan no solo la detección sino también la identificación (civil-militar, amigo-enemigo), se prevén necesidades en el ámbito de la vigilancia de grandes áreas.

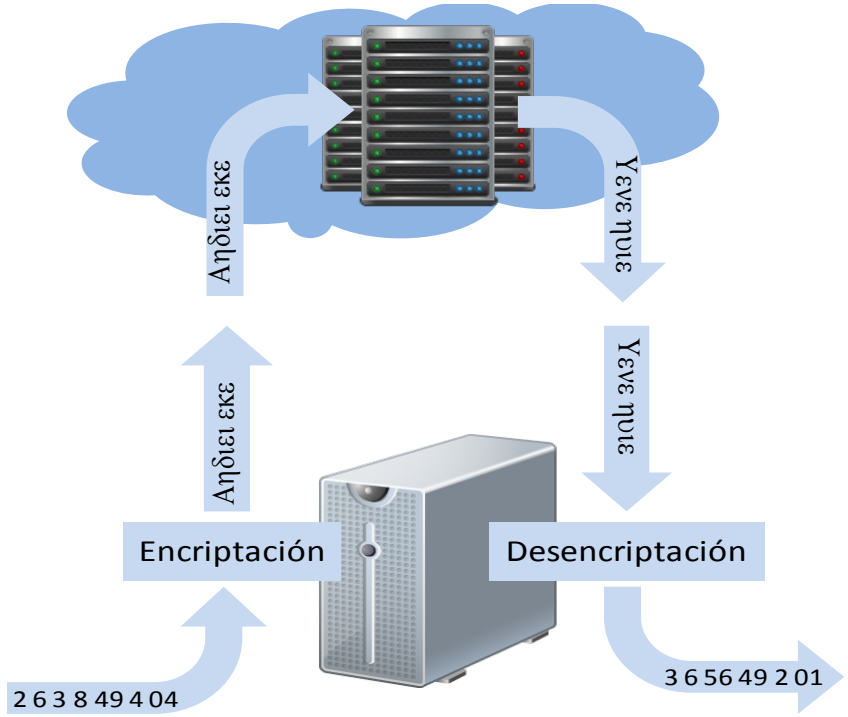
Áreas de aplicación relacionadas	Vigilancia y Seguridad perimetral. Vigilancia y Seguridad de fronteras Seguridad física de infraestructuras críticas.

3.2.2.- Computación sobre información cifrada

Hoy en día muchos de los datos de interés para una aplicación residen en servidores deslocalizados físicamente de donde se está accediendo y necesitando esos datos e información. Un ejemplo muy actual son los servidores en la "nube". En el ámbito de seguridad y defensa también se da esta configuración, y en ellos como es lógico, la seguridad de la información es un tema especialmente relevante.

Una manera de abordar este desafío, es poder operar, computar, sobre esos datos sin necesidad de desencriptarlos previamente. Mediante este planteamiento, los datos por lo tanto permanecen cifrados, quedando la seguridad de estos dependiendo de la robustez del cifrado, pero en el acceso y operación a los datos no se generan debilidades en la seguridad potencialmente utilizables para la intercepción de la información por un tercero.

En la actualidad existen principalmente dos aproximaciones para implementar este planteamiento. La primera es la denominada encriptación homomórfica (FHE, fully homomorphic encryption). Los datos se cifran antes de ser transferidos a los servidores, adicionalmente se incorpora a los servidores un interpretador software que permite trabajar con los datos sin necesidad de descifrarlos. Los resultados se mandan a los usuarios cifrados, solo los usuarios con la clave adecuada pueden descifrarlos. Sin embargo, la encriptación homomórfica presenta un inconveniente que la hace prácticamente inviable, enlentece la computación / operación con estos datos alrededor de diez órdenes de magnitud.



Otra posible aproximación menos exigente sobre la capacidad computacional, en este caso estamos hablando de dos órdenes de magnitud, es la denominada computación multi-parte (SMC), en la cual diferentes entidades/elementos computacionales se unen para ejecutar operaciones cada uno sobre una parte de los datos, manteniendo así la privacidad del conjunto total de los datos.

Estas dos aproximaciones demandan altas capacidades de computación, algo que las hace inviables con las soluciones computaciones empleadas hasta la fecha. Sin embargo, se está estudiando la aplicación de las técnicas de Big Data, aprovechando una escalabilidad en el procesamiento de datos impensable en otras circunstancias, con el objetivo de conseguir que la computación sobre información cifrada sea práctica. En este sentido se están abordando aspectos asociados a la mejora de la computación multi-parte segura, a la implementación de hardware y software optimizada, y a diferentes lenguajes de programación y tipos de datos y algoritmos.

Un ejemplo de trabajo en esta área es el proyecto PROCEED (Programming Computation on Encrypted Data) de la agencia Darpa de EE.UU., que de tener éxito podría significar un avance significativo en como diseñar arquitecturas computacionales en entornos no seguros.

Áreas de aplicación relacionadas	Comunicaciones seguras Bancos de datos para los ámbitos financieros, seguridad interior, inteligencia, defensa.

3.2.3.- Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos)

La aplicación de Big Data al mundo de la ciberseguridad / ciberdefensa, persigue fundamentalmente dos objetivos:

- Reducir la cantidad de tiempo que los analistas pasan descubriendo ciberataques al agrupar y correlacionar fuentes de redes de datos dispares e,
- incrementar la precisión, tasa y velocidad de detección de ciber-amenazas a redes de ordenadores.

Para ello es necesario cambiar el modo en que la información relativa a las redes es adquirida, procesada y puesta a disposición de los ciber-defensores. Procurando proporcionarles datos conectados y correlacionados, para que puedan abordar directamente el problema del orden de escala de los datos asociados a la seguridad de las redes. A este desafío hay que añadir la tendencia creciente a integrar en esas redes diversos dispositivos de IT no corporativos.

Las soluciones necesitaran:

- Indexar las fuentes de datos de la red automáticamente o con una mínima intervención humana
- Integrar estructuras de datos dispares (que no presentan consistencia en su una estructura de datos)
- Permitir a los analistas realizar inferencias a partir de la base de datos de grupos y correlaciones (es decir, buscar relaciones entre cualquier campo de datos de la red)

Dos ejemplos de estos programas en el mundo de la Defensa son el Cyber Targeted-Attack Analyzer y el Cyber-Insider Threat (CINDER), ambos de DARPA.

El mundo de la seguridad de redes en organizaciones y en infraestructuras críticas, tiene ya un largo recorrido tanto en investigación como en productos existentes, sin embargo los nuevos requisitos impuestos por las dimensiones volumen de datos, y prestaciones demandadas hacen que los productos actuales queden en desventaja frente a las amenazas emergentes.

Así las nuevas capacidades de Big Data surgen en un momento en el que las organizaciones hacen frente a nuevos riesgos originados por dos desafíos:

1. Disolución de los límites de las redes: La extensión y apertura de las redes de datos de las organizaciones, permitiendo a socios, suministradores y clientes acceder a información corporativa mediante nuevas formas dinámicas para impulsar la innovación y la colaboración, hace que estas redes se vuelven más vulnerables al mal uso y el robo de datos. Las aplicaciones y datos corporativos son cada vez más accesibles mediante servicios en la nube y dispositivos móvi-

les, rompiendo los últimos límites de la red corporativa e introduciendo nuevos riesgos en la información y vectores de amenaza.

2. Adversarios más sofisticados: Los ciberatacantes se han vuelto más adeptos a realizar ataques complejos y muy específicos que evitan las defensas tradicionales, las medidas estáticas de detección de amenazas y las herramientas basadas en firma. A menudo, los ataques o fraudes no son detectados hasta que el daño se ha realizado.

Por esto, es necesario soluciones más ágiles basadas en evaluaciones dinámicas de riesgo. El análisis de grandes volúmenes de datos y operaciones de seguridad en tiempo real serán esenciales para proporcionar una seguridad significativa.

Las soluciones de seguridad analítica de Big Data se caracterizan básicamente por:

- Escala. Las soluciones de seguridad analítica de Big Data deben ser capaces de recoger, procesar y almacenar entre terabytes y petabytes de datos para una gran variedad de actividades de seguridad analítica
- Flexibilidad analítica. Las soluciones de seguridad analítica de Big Data deben proporcionar a los usuarios la habilidad de interactuar, buscar y visualizar este volumen de datos de múltiples formas
- Rendimiento. Las soluciones de seguridad analítica de Big Data deben construirse con una arquitectura computacional adecuada capaz de procesar algoritmos y búsquedas complejas y mostrar los resultados en un tiempo aceptable.

En las primeras fases de este mercado, se pueden encontrar dos visiones y por tanto dos tipos de soluciones de seguridad analítica de Big Data:

1. Soluciones de seguridad analítica de Big Data en tiempo real. Las soluciones de seguridad analítica de Big Data en tiempo real son una evolución de las soluciones de gestión de registros construidas para los requisitos de rendimiento y escalas actuales. Estas soluciones se construyen alrededor de una arquitectura distribuida, formada por dispositivos diseñados para el procesado continuo local y procesado paralelo conjunto. Algunos ejemplos de soluciones de seguridad analítica de Big Data en tiempo real incluyen Click Security, Lancope, y Solera Networks.

2. Soluciones de seguridad analítica de Big Data asimétricas. Se trata de en una categoría de soluciones relativamente nueva, diseñada para las necesidades no lineales de los analistas de seguridad que habitualmente pasan de búsqueda en búsqueda cuando investigan eventos de seguridad individuales y/o comportamientos anómalos de sistemas, redes, actividades de usuario ,etc. Las soluciones de seguridad analítica de Big Data asimétricas pueden basarse en repositorios de datos propietarios, pero es probable que todos los productos acaben basándose en tecnologías de Big Data como Cassandra, Hadoop, y NoSQL. La idea es que

los analistas de seguridad alimentaren estas soluciones con grupos de actualizaciones que contengan terabytes de datos estructurados y no estructurados con el fin de observar las tendencias históricas de seguridad en grandes períodos de tiempo. Las soluciones asimétricas estarán basadas en algoritmos de aprendizaje artificial, análisis de clústeres y visualización avanzada. Las primeras soluciones en este ámbito vienen de empresas como LexisNexis, PacketLoop, y RedLambda

Algunas empresas, como IBM, LogRhythm, Narus, RSA, y Splunk ofrecen capacidades en ambas áreas, al integrar habitualmente varios productos en una sola arquitectura de seguridad.

En los próximos años se prevé que la analítica de Big Data provoque una disrupción en el status quo en la mayoría de segmentos de productos de seguridad en redes, incluyendo monitorización de redes, autenticación y autorización de usuarios, gestión de identidades, detección de fraudes y gobernanza, riesgos y conformidad. En un medio plazo, se prevé que las herramientas evolucionen hasta permitir un abanico de capacidades de predicción avanzadas y controles automatizados en tiempo real.

Áreas de aplicación relacionadas	Protección de redes de telecomunicaciones Ciberdefensa Ciberseguridad Protección de Infraestructuras críticas

3.2.4.- Criminología Computacional

Una de las áreas donde los conceptos de Big Data encuentran una aplicación directa es la Criminología Computacional, donde la capacidad de analizar grandes volúmenes de datos relacionados con actividades criminales multiplica las posibilidades de neutralización de las amenazas relacionadas con dichas actividades. En este contexto, Big Data estaría relacionado con técnicas como el minado de datos criminales, el análisis de agrupaciones y el aprendizaje de reglas de asociación para la predicción del crimen, análisis de redes criminales, análisis de textos multilingües, análisis de opiniones y sentimientos, etc.

Programas de investigación como el COPLINK o el Dark Web Research de la Universidad de Arizona ofrecen un excelente ejemplo del potencial de estas tecnologías. COPLINK, desarrollado inicialmente con fondos de la National Science Foundation y el Departamento de Justicia de Estados Unidos, es un sistema de compartición de información y de minado de datos criminales utilizado por más de 4500 departamentos de policía en los Estados Unidos y por 25 países OTAN. El sistema COPLINK fue adquirido por IBM en 2011. Dark Web, financiado por la National Science Foundation y el Departamento de Defensa, ha generado una de las mayores bases de datos existentes para la investigación del terrorismo, con cerca de 20 terabytes de información sobre

sitios web y contenidos de redes sociales relacionados con terrorismo.

Otro punto de vista de la criminología computacional donde Big Data es de aplicación es en el seguimiento de actividades sospechosas en diferentes redes (tanto internet como de una organizacional). Siendo de interés, identificar qué información es recogida como soporte para la actividad criminal.

Desde el punto de vista retrospectivo, Big Data también puede ser utilizado para documentar y recopilar pruebas de actividades ilícitas (realizadas tanto dentro como fuera de Internet o red IT corporativa). Aprovechando la capacidad de almacenar y procesar gran cantidad de información.

Áreas de aplicación relacionadas	Seguridad en general, pero especialmente en: Lucha contraterrorista y crimen organizado Lucha contra el fraude Lucha contra usurpación de identidad Lucha contra pedofilia y redes de explotación infantil

3.2.5.- Uso fraudulento de recursos corporativos y/o sensibles

En las grandes organizaciones el número de usuarios y de recursos, ya sean recursos sensibles o no, dan lugar a numerosas reglas y protocolos de acceso, y sobre todo a un número muy grande sesiones concurrentes de acceso a estos recursos.

En este sentido, y gracias a las técnicas de Big Data, es posible afrontar el análisis en tiempo de real de los datos de estas sesiones de acceso, identificando patrones de comportamiento y distinguiendo usos legítimos de los recursos frente a otros, ya sean relacionados con el abuso de los recursos de información de la organización, o relacionados con ataques malintencionados. Es importante recordar que muchos de los ataques a las redes de información de las organizaciones se realizan desde dentro, o aprovechando fallos de seguridad de la red interna de la organización.

Por otro lado, existen organizaciones donde es necesaria una flexibilidad en las reglas y protocolos de acceso. En la actualidad el control de acceso a los sistemas suele estar sujeto a unas políticas muy rígidas, basadas siempre en perfiles prefijados o demasiado simples para la complejidad requerida, con dificultades para adaptarse a la situación real del momento y con estrictas reglas que impiden su adecuación en tiempo real.

Las técnicas de Big Data favorece el desarrollo de todas estas herramientas.

Áreas de aplicación relacionadas	Seguridad de informática Gestión del conocimiento en grandes organizaciones

3.2.6.- Análisis de vídeo en tiempo real / búsqueda y recuperación rápida en librerías de vídeo

En las operaciones militares actuales, o en el caso del ámbito de seguridad con la proliferación de sistemas de video vigilancia instalados en las ciudades, se recogen una cantidad ingente de datos de video. Esta cantidad es especialmente masiva en labores de inteligencia, vigilancia, adquisición de objetivos y reconocimiento debido al incremento del uso de UAVs.

En esta situación surge un problema importante al no disponer de suficiente capacidad de análisis o incluso de tiempo material para la revisión de tanta cantidad de información. Así se identifican dos aplicaciones donde Big Data puede ser de utilidad:

1. El análisis de vídeo en tiempo real.
2. La búsqueda y recuperación rápida en librerías de vídeo.

Como se indicaba, un área de interés para la primera aplicación es la de ISTAR (de inteligencia, vigilancia, adquisición de objetivos y reconocimiento), ofreciendo a los analistas de imágenes militares la capacidad de explotar la gran cantidad de imágenes y video capturados. De esta manera es posible para los analistas establecer alertas asociadas a diferentes actividades y sucesos de interés mientras estos están ocurriendo (por ejemplo alertas del tipo "una persona acaba de entrar en el edificio"). O incluso de manera predictiva en base a patrones de hechos ya conocidos, que permitan adelantarse a los acontecimientos en un tiempo suficiente para poder reaccionar frente a las amenazas.

En el caso de la segunda aplicación, se persigue el desarrollo de agentes inteligentes ("búsqueda basada en contenidos") que permitan encontrar contenidos de vídeo de interés en librerías con en miles de horas de grabaciones de vídeo. Hasta ahora, la mayoría de las búsquedas en librerías de video requieren una gran cantidad de intervención humana, bien mediante visionado rápido del video, o con búsquedas mediante metadatos y/o anotaciones realizadas anteriormente. El objetivo por tanto, es detectar de manera rápida y precisa en estas grandes librearías de video actividades potencialmente sospechosas, o trabajar en la detección de objetos (por ejemplo vehículos: acelerando, girando, parando, adelantando, explotando o en llamas, formando convoyes o manteniendo una distancia con otro, entre otras posibilidades), o también identificar comportamientos extraños de una persona (excavando, dejando abandonado un objeto, etc.), o en interacciones hombre-hombre (siguiendo, reuniéndose,

moviéndose conjuntamente, intercambiando objetos, etc.) que puedan resultar de interés para una investigación.

Ejemplos de programas en esta línea son el VIRAT (Video and Image Retrieval and Analysis Tool) de DARPA y el Intelligent Video Analytics de IBM.

Áreas de aplicación relacionadas	Vigilancia y Seguridad perimetral. Vigilancia y Seguridad de fronteras Seguridad física de infraestructuras críticas. Seguridad ciudadana Inteligencia militar

3.2.7.- Inteligencia Visual en Máquinas

Desde su concepción inicial, los sistemas de visión artificial han experimentado un progreso continuo en la capacidad de reconocer los distintos objetos de una escena y sus propiedades. Sin embargo, para que un sistema de visión artificial sea verdaderamente "inteligente", debería ser capaz no sólo de reconocer los objetos, sino también de identificar y razonar sobre las acciones y relaciones que se establezcan entre dichos objetos. De esta manera, un sistema de estas características ("*smart camera*") sería capaz no sólo de identificar los objetos, sino de describir lo que está sucediendo en la escena, y procesar alertas en caso de ser necesario.

Una de las principales limitaciones para el desarrollo de estas cámaras inteligentes es el elevado número de datos a procesar por parte de la algoritmia de inteligencia visual, lo que imposibilita actualmente su utilización en tiempo real en un entorno operativo. En este sentido, los conceptos de Big Data se revelan como una prometedora solución para potenciar esta capacidad de procesado y avanzar hacia una inteligencia visual completa.

Los desarrollos en marchan persiguen añadir a los actuales sistemas (con capacidad de reconocimiento de objetos) los fundamentos perceptuales y cognitivos necesarios para el razonamiento de las acciones que se producen en la escena. La idea es conseguir automatizar la habilidad para aprender "representaciones" de acciones entre objetos a partir de los datos visuales de la escena, de manera que se pueda razonar posteriormente sobre lo que está sucediendo en dicha escena. En paralelo se pretende reducir también los requisitos computacionales de los sistemas de inteligencia visual, de manera que se consiga en un futuro superar las limitaciones para su uso operativo.

Entre las iniciativas destacadas en este campo se puede mencionar el programa PERSEAS (Persistent Stare Exploitation and Analysis System) y el *Mind's Eye ambos* de la agencia DARPA.

En los próximos años se espera que las tecnologías de inteligencia visual permitan construir cámaras inteligentes capaces de responder a preguntas del tipo "¿Qué acaba de pasar en la escena?". En esta evolución, se espera que las tecnologías de Big Data jueguen un papel determinante.

Áreas de aplicación relacionadas	Vigilancia y Seguridad perimetral. Vigilancia y Seguridad de fronteras Seguridad física de infraestructuras críticas. Seguridad ciudadana Inteligencia militar

3.2.8.- Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos

Actualmente, los analistas de inteligencia militar se enfrentan con la tarea de manejar los enormes y crecientes volúmenes de datos complejos provenientes de múltiples fuentes y tipos de inteligencia. El objetivo es fusionar todos estos datos para la identificación automática de posibles amenazas y operaciones a través del análisis de la información, como por ejemplo la originada en sensores de imagen, sensores radar e interceptación de comunicaciones.

Todos estos datos deben ser evaluados, relacionados y finalmente usados en el apoyo a la toma de decisión y acciones en una ventana de tiempo crítico. El uso de correlación de diversos tipos de información sobre personas, eventos, fechas, detección de actividad y seguimiento, etc., permite mejorar la habilidad de los analistas para procesar información de forma más efectiva y eficiente. Por ejemplo, se pretende proporcionar capacidades de seguimiento de estos elementos (de personas, eventos, etc.) en tiempo real o cercanas a tiempo real para el apoyo directo a usuarios tácticos en el campo de batalla, mediante el desarrollo de tecnologías semi- o completamente automáticas. Para ello es necesario la:

- Combinación, análisis y explotación de datos e información originada en múltiples fuentes, incluyendo sensores de imagen, otros sensores y otras fuentes;
- Gestión eficiente de las tareas asignadas a los sensores
- Detección e identificación de amenazas mediante el uso de algoritmos de descubrimiento y predicción del comportamiento.

De manera general se persiguen los siguientes objetivos:

- reemplazar los silos de información existentes por un sistema integrado que opere a nivel nacional, de teatro de operaciones e incluso en sistemas de

inteligencia táctica de menor nivel;

- optimizar el manejo de datos para lograr hacer uso de forma efectiva de tecnologías ISR existentes y emergentes
- independencia respecto a la misión y al sensor y tener aplicabilidad en teatros de operación que cambian de forma dinámica;
- estar basados en estándares para permitir añadir eliminar, sustituir y modificar componentes de hardware y software según se vayan desarrollando y están disponibles para su integración y
- promocionar la colaboración eficiente entre los analistas de inteligencia e incrementar la eficiencia y eficacia de los analistas individuales a través de un cuadro ISR global y unificado.

Uno de los programas más destacados en este ámbito es el Insight de Darpa, que ya se ha probado en escenarios de guerra asimétrica, con énfasis en operaciones de contra-insurgencia donde ha demostrado funcionalidades en el ciclo completo, es decir, de la fuente-a-analista y del analista al operativo. Para ello, se priorizaron sensores, potencia de computación y capacidad analítica para el apoyo directo a brigadas tácticas y batallones en el marco de una misión de seguridad en un área amplia y para la obtención de capacidades de:

- capturar información de los sistemas de mando de batalla,
- capturar información y de interactuar dinámicamente con fuentes espaciales, aéreas y terrestres
- fusionar datos de diferentes fuentes de inteligencia.
- almacenar, indexar, buscar y entregar de forma persistente información recibida de fuentes múltiples, y presentarla al analista
- detectar redes enemigas, seguir a múltiples vehículos en tiempo real
- proporcionar información relevante y a tiempo a los operativos
- repetición virtual y simulación de sensores como herramientas para la integración y ensayo del sistema.

Otros programas destacables de Darpa en este ámbito, son el programa Anomaly Detection at Multiple Scales (ADAMS) que aborda el problema de detección de anomalías y la caracterización de grandes conjuntos de datos y el programa Persistent Stare Exploitation and Analysis System (PerSEAS). Este último está desarrollando capacidades de identificación de amenazas de forma automática e interactiva basadas en la correlación de múltiples actividades y eventos dispares en *wide area motion imagery* (WAMI) y datos de múltiples fuentes de inteligencia. PerSEAS permitirá nuevos métodos de adjudicación de hipótesis de amenazas y análisis forense a través de modelos

basados en actividad y capacidades de inferencia.

Se espera que las investigaciones en marcha ofrezcan en el medio plazo:

- Clasificación robusta para detectar, geo-registrar e identificar objetos de la superficie de forma precisa a pesar de las dificultades del entorno, configuraciones y emplazamientos.
- Herramientas robustas de automatización para identificar relaciones, patrones vitales y actividades de vehículos terrestres
- Herramientas robustas para capturar, almacenar y recuperar información basada en HUMINT para identificar e impulsar el apoyo local contra los insurgentes
- Herramientas específicas de dominio para capturar, buscar y explotar información explícita de redes insurgentes a partir de fuentes de datos textuales no estructuradas.

Y en plazo temporal cercano a los 10 años:

- Clasificación robusta para detectar, geo-registrar e identificar todos los objetos de la superficie de forma precisa a pesar de las dificultades del entorno, configuraciones y emplazamientos.
- Herramientas de automatización robusta para identificar relaciones, patrones vitales y actividades de soldados a pie
- Herramientas robustas para buscar, minar, y explotar datos de fuentes abiertas para identificar todos los aspectos de redes insurgentes.

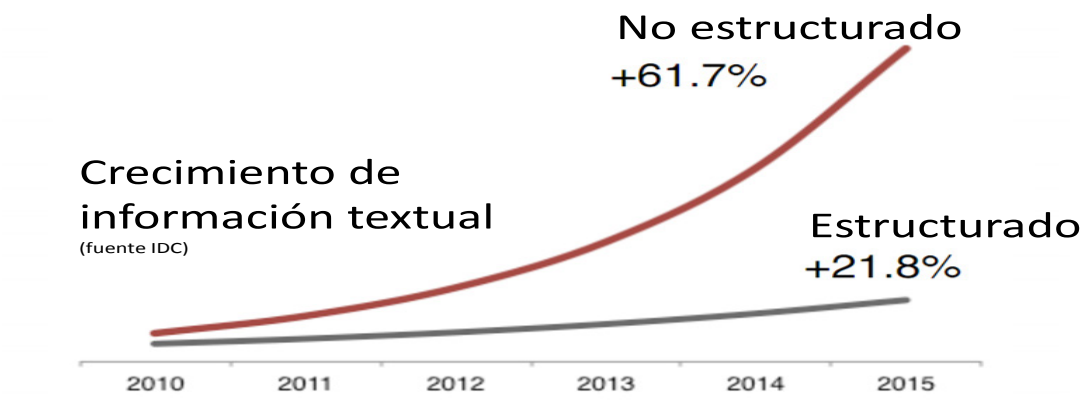
Áreas de aplicación relacionadas	Planeamiento táctico de misiones. Toma de decisión en tiempo real para operaciones (Defensa/seguridad).

3.2.9.- Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos

En un gran número de operaciones militares, así como operaciones de seguridad, los datos en formato de texto proporcionan información esencial sobre actitudes, sucesos o relaciones entre los distintos actores implicados en la operación.

En las últimas décadas, el acelerado ritmo de avance de las tecnologías de las telecomunicaciones (internet, telefonía móvil, etc.) ha hecho posible que se disponga de cantidades ingentes de textos en una gran variedad de formatos. La aplicación de las técnicas de análisis textual (*text analytics*) a este gran volumen de datos ofrece nuevas posibilidades para la extracción de las informaciones relevantes para la toma de decisión. Las técnicas de Big Data pueden apoyar a que este análisis textual que se realice en

tiempo real y proporcione resultados exhaustivos y precisos, apoyando eficazmente a la toma de decisión en las ventanas de tiempo marcadas por las misiones.



En este sentido, merece la pena señalar que el Departamento de Defensa de Estados Unidos (DoD) ha reconocido que las técnicas de análisis de textos pueden tener un papel esencial en las futuras capacidades militares.

Desafíos abiertos para la aplicación eficaz de Big Data en este ámbito son el desarrollo de algoritmos escalables para el procesamiento de datos imperfectos situados en almacenes de datos distribuidos; y el desarrollo de herramientas efectivas para la interacción hombre-máquina que sean rápidamente personalizables para facilitar el razonamiento visual para diversas misiones.

Ejemplos de estos programas son el *Data to Decisions* (D2D), promovido por la OSD (*Office of the Secretary of Defense*) o el programa XDATA de la agencia DARPA.

En los próximos años se espera que las técnicas de análisis textual adquieran una importancia cada vez mayor en ámbitos muy diversos, no sólo en el de defensa y seguridad. En este proceso, las tecnologías asociadas al concepto de Big Data tendrán un papel fundamental. Aspectos como la representación eficiente de la información (estructurada o no estructurada), la capacidad de alcanzar altos ratios de compresión de la información sin pérdida de fidelidad o el comportamiento anticipatorio de los recursos computacionales, de manera que se pueda atender al elevado número de peticiones de análisis por parte de los usuarios, serán objeto de investigación en las futuras iniciativas en este ámbito.

Áreas de aplicación relacionadas	Planeamiento táctico de misiones. Toma de decisión en tiempo real para operaciones (Defensa/seguridad).

3.2.10.- Consciencia situacional

Al comienzo del capítulo, se indicaba que el entendimiento de la situación y contexto del analista y el combatiente/agente del orden podía ser considerada como la aplicación de "Big Data" a defensa y seguridad por antonomasia.

La necesidad de identificar las amenazas en entornos complejos, inciertos e incluso contradictorios en los que existen gran cantidad de datos disponibles de fuentes abiertas, es de vital importancia tanto para la detección de la amenaza como para su neutralización.

La comprensión o entendimiento contextual se consigue a través de la combinación de técnicas del procesado hombre /máquina en la que se aprovecha la capacidad cognitiva de la persona para fundir y asimilar múltiples fuentes y tipos de información, y así alcanzar nuevas perspectivas. Las técnicas de Big Data se aplican en dicho entendimiento contextual facilitando el análisis y procesado de múltiples datos.

Más específicamente, Big Data puede contribuir mediante:

- El desarrollo de sistemas de aprendizaje automático que procesen lenguaje natural e inserten la representación semántica resultante en una base de conocimiento en lugar de basarse en los procesos actuales, costosos y que requieren mucho tiempo para la representación del conocimiento ya que requieren de personas con diferentes áreas de expertise que organizan el conocimiento.
- Enlazar estas bases de conocimiento con capacidades de "inteligencia visual" en máquinas, en las que se identifican no sólo los objetos existentes en diferentes imágenes/videos, sino también las relaciones (acciones) entre esos objetos, permitiendo interpretar de manera automática las imágenes, y construir una narrativa de la información visual.
- Añadiendo los resultados provenientes de la correlación y agregación de datos OSINT (demográficas, económicas, patrones sociales, etc...).

Ejemplos de los aspectos sobre los que se espera poder obtener información son:

- El descubrimiento de sucesos específicos, tanto en planeamiento como que hayan ocurrido.
- Creencias y valores que motivan ciertos comportamiento de interés
- El descubrimiento de temas y conceptos desarrollados colaborativamente en el

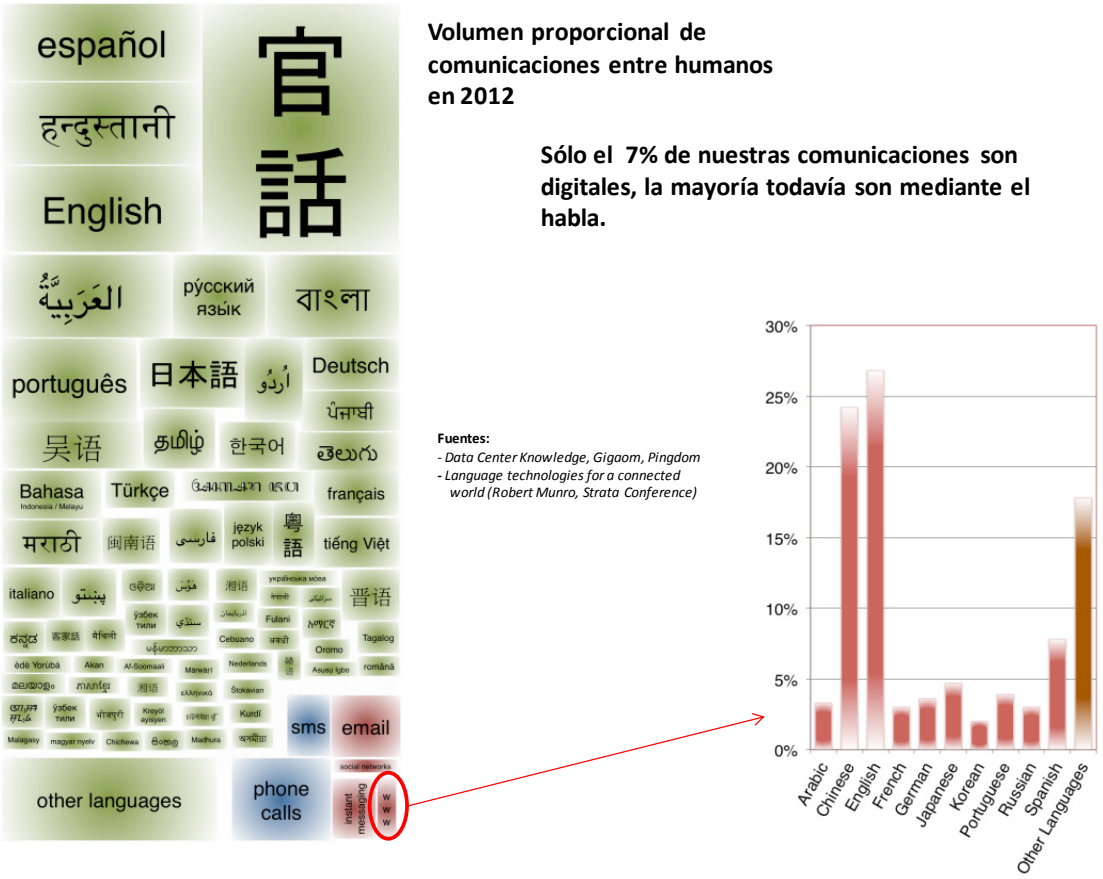
seno de grupos y organizaciones.

- El análisis de relaciones semánticas existentes en un grupo.
- Fortaleza y dependencia de las relaciones existentes en un grupo.
- Análisis semántico y de tendencias en el apoyo a grupos, temas y personas.

Áreas de aplicación relacionadas	Planeamiento táctico de misiones. Toma de decisión en tiempo real para operaciones (Defensa/seguridad).

3.2.11.- Traducción automática gran escala (en número de idiomas y en volumen)

Aunque en los últimos años han producido notables avances en la recuperación y tratamiento de la información (un ejemplo son los motores de búsqueda inteligentes y adaptativos basados en basados en ontologías), el progreso en la "digestión" o la comprensión de la información no ha avanzado con el mismo ritmo. Una de las razones para esta divergencia, es el enorme aumento de la información no estructurada relacionada tanto con el teatro de operaciones y el entorno directo del combatiente, como con los escenarios de seguridad.



Este hecho se complica al considerar que muchas de las áreas donde se realizan las operaciones son muy ricas en idiomas y dialectos, o que la globalización requiera una concepción de la seguridad no circunscrita a las fronteras, y por lo tanto con gran diversidad lingüística.

Por estas razones, las capacidades de traducción automática a gran escala (en número de idiomas y en volumen de texto/audio procesable) se convierten en fundamentales para las necesidades operativas de muchas misiones de defensa y de seguridad. Esto es especialmente crítico en el caso de las misiones internacionales, donde el lenguaje proporciona muchos de los elementos clave para entender la cultura, el carácter, las opiniones y apreciaciones de los naturales del país donde se desarrolla la operación.

En este sentido, el desarrollo de algoritmos computacionalmente eficientes para el procesamiento (ontologías y minería de datos), la extracción de información a gran escala, el reconocimiento de voz y el resumen automático, representan una oportunidad para la aplicación de Big Data a las necesidades de traducción en defensa y seguridad.

El objetivo es pasar de un concepto de traducción de un documento cada vez a traducciones automáticas de muchos documentos en bloque.

No podemos finalizar este factor sin considerar el efecto globalizador y transformacional

en numerosos ámbitos (negocios, seguridad, defensa, creación cultural, educación, etc.) derivado de la posibilidad de trasladar estas capacidades de traducción casi automática a gran escala (en número de idiomas y volumen de información) a la "nube" al estilo de un servicio como el de Google.

Áreas de aplicación relacionadas	Toma de decisión en tiempo real para operaciones (Defensa/seguridad). Lucha contraterrorista y crimen organizado Inteligencia industrial

3.2.12.- Predicción de eventos

Los analistas en los servicios de inteligencia necesitan de la capacidad de monitorizar y analizar rápidamente la información de eventos formada por grandes volúmenes de datos de texto no estructurados con el fin de lograr y mantener un entendimiento de los acontecimientos y poder formular sobre ellos estimaciones y predicciones a futuro. La cantidad de datos de texto no estructurados disponibles va mucho más allá de lo que se puede leer y procesar en un tiempo determinado. Es aquí donde el Big Data puede contribuir y facilitar el manejo y análisis de toda esta cantidad enorme de datos.

Uno de los grandes retos tecnológicos en este campo es avanzar en la extracción de eventos con sus atributos de modalidad, polaridad, especificidad y tiempo. La modalidad de un evento indica si se trata de un evento real o no. Ejemplos de modalidad de un evento son asertiva "la bomba explotó el domingo", creencia "se cree que será condenado", hipotética "si fuese arrestado, se le acusaría de asesinato", etc. La polaridad de un evento indica si el evento ocurrió o no.

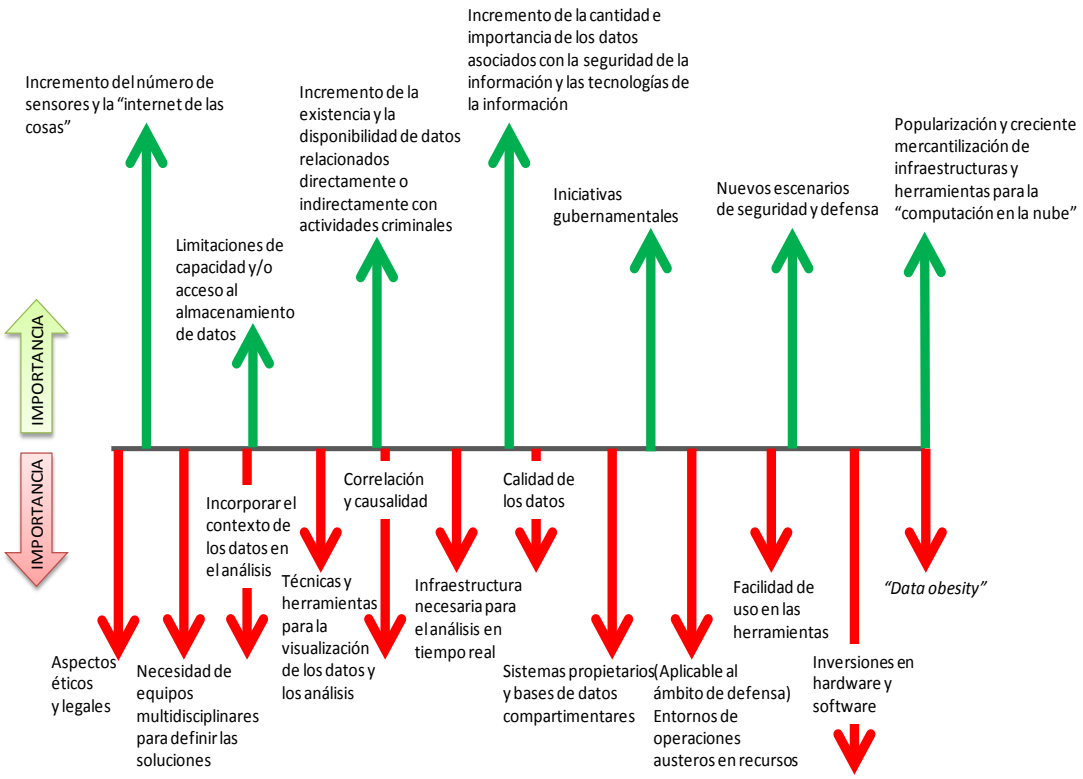
Para lograr estos retos se investigan nuevos métodos innovadores que puedan apoyar la predicción de eventos, entre ellos destacan:

- Mediante análisis textual, la extracción, caracterización y monitorización continua de redes sociales y grupos de interés. Extracción de las tendencias temporales, es decir, la frecuencia de los contactos entre los nodos o grupos, tiempo con el contacto, contactos periódicos, ruta de retraso en la difusión de la información, etc.
- Detección de la presencia de nodos puente que permiten descubrir subredes ocultas, y determinar el flujo de recursos (información, dinero, influencia) dentro de la red social.

Áreas de aplicación relacionadas	En ámbito militar en humint/operaciones en entornos urbanos. Control y comportamientos de multitudes. Seguridad ciudadana. Preparación de seguridad de eventos singulares (deportivos, políticos, etc.)

3.3.- Factores impulsores y limitadores en la aplicación de Big Data en Seguridad y Defensa.

Con objeto de poder valorar la evolución del Big Data en general, y sobre todo en lo referente a su aplicación a Seguridad y Defensa, se ha realizado un análisis de cuáles son los factores más influyentes en el desarrollo de estas aplicaciones de Big Data. Estos factores responden a varios aspectos como pueden ser las características técnicas de Big Data, el entorno de seguridad y defensa, el contexto social-económico, etc. Se han dividido entre factores impulsores del desarrollo, es decir que demandan o favorecen el desarrollo y aplicación de Big Data, y factores limitadores, que por el contrario, frenan o presentan desafíos destacables para el desarrollo y aplicación de Big Data.



A modo de resumen y de visión global se presenta en la siguiente figura el conjunto total de estos factores, quedando en la parte superior los favorables en la aplicación de Big Data, y en la parte inferior, aquellos que suponen un desafío o una limitación en la aplicación de Big Data en seguridad y defensa (y posiblemente en también en otros ámbitos).

Adicionalmente, se ha estimado la importancia de cada uno de estos factores, tanto para el favorecimiento como para la limitación de Big Data, reflejando esta importancia con el tamaño de cada una de las flechas.

3.3.1.- Listado de Factores impulsores

- Incremento del número de sensores y la "internet de las cosas".

Las técnicas y tecnologías de Big Data que nos facilitan el análisis y explotación de datos no servirían de mucha utilidad sin la existencia de éstos. Para bien o para mal, las estadísticas y previsiones en este ámbito no dejan lugar a dudas de que avanzamos hacia un mundo altamente conectado y con una proliferación de datos con un crecimiento exponencial.

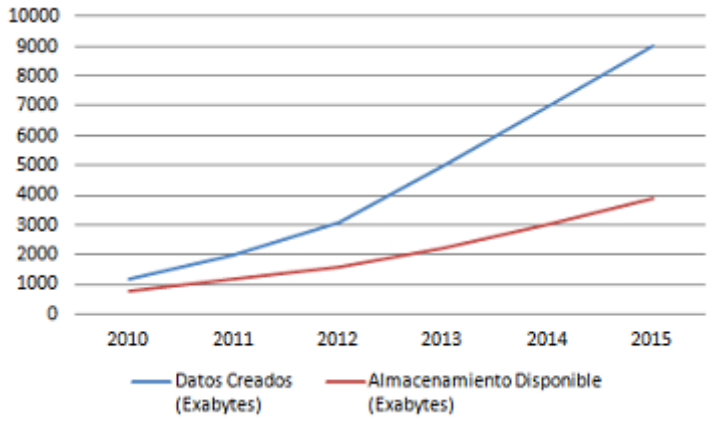
Esta proliferación está alimentada principalmente por tres tendencias:

- Proliferación generalizada de fotos, audio y vídeo debido a los medios de comunicación y la emergencia de contenidos compartidos en las denominadas redes sociales.

- La emergencia del "todo como sensor" y del "internet de las cosas". A la existencia de diferentes dispositivos comunes como los teléfonos inteligentes generando datos de forma continua, hay que añadir nuevas categorías de sensores generadores de datos. De tal manera que es posible decir que evolucionamos hacia un mundo donde una gran mayoría de los objetos sean sensores, y lo que es aún más disruptivo, todos ellos potencialmente conectados al denominado "internet de las cosas". El concepto "internet de las cosas" hace referencia a la evolución, ya en marcha desde años, donde internet está comunicando entre sí objetos físicos en mayor proporción que conectando a personas. Se estima que en 2015, no sólo el 75% de la población mundial tendrá acceso a internet, sino que también lo tendrán 6 mil millones de dispositivos.

- Limitaciones de capacidad y/o acceso al almacenamiento de datos.

Añadido sobre el factor anterior hay que destacar que la velocidad, resolución y tasa de datos de los sensores se ha incrementado. Sin embargo, el ancho de banda de los canales de comunicación no es capaz de seguir este ritmo de incremento de la cantidad de datos. Este desafío ha creado la necesidad de, o bien procesar los datos localmente en el sensor, o bien almacenar los datos para su posterior descarga y post-procesado después de completar la misión. Sin embargo, tal y como se presenta en la figura, existe una divergencia entre la cantidad de datos generados y el almacenamiento disponible.



Así, el hecho de que más sensores adquieran más datos, con una mayor calidad y cantidad, está impulsando la necesidad tanto del procesado "in situ". Mientras que el almacenamiento de datos en bruto es esencial en muchos sistemas militares, la siempre creciente avalancha de datos está haciendo poco práctico el continuar simplemente capturando y almacenando datos.

En este sentido, se persigue encontrar técnicas que puedan capturar solamente los datos relevantes y, o almacenar una cantidad significativamente reducida de datos, o tomar decisiones inmediatas e inteligentes basados en datos en tiempo real.

En ocasiones, se favorece el procesado local de datos en lugar del almacenamiento debido al retraso que implica el almacenamiento y el post-procesado y el efecto que esta latencia pueda tener en las funciones de inteligencia. En estas circunstancias, el

procesado de datos en el sensor se centra normalmente en técnicas de reducción de datos donde el ancho de banda resultante es suficientemente bajo para su transmisión sobre los canales de comunicación.

Además de poder aplicar las técnicas de Big Data al volumen de datos generados por cada sensor, también surge una nueva posibilidad, el denominado 'Near Sensor Computing' (NSC). Utilizando un ejemplo para su explicación, el NSC encaja con el equivalente de acumular todos los datos generados por cada miembro de un pelotón (el tratamiento de esa cantidad de datos en tiempo real ya podría considerarse en sí mismo como un problema de Big Data) y procesarlos conjuntamente a nivel de pelotón, revertiendo inteligencia tanto al propio pelotón como a niveles organizativos/tácticos superiores.

El NSC junto a pequeños sensores permite la computación en tiempo real de gran cantidad de datos desde el propio sensor sin necesitar sistemas de comunicaciones costosos y con elevados requisitos de potencia o aparatosos sistemas de almacenamiento. Estos conjuntos de sensores con dispositivos NSC se están utilizando en la construcción de capacidades "sense and avoid" de UAVs.

- Incremento de la existencia y la disponibilidad de datos relacionados directamente o indirectamente con actividades criminales.

Los diferentes gobiernos y servicios de inteligencia recopilan y almacenan una creciente cantidad de datos relativos a delitos, fichas criminales, atentados terroristas, ataques de ciberseguridad, redes criminales, etc. Al mismo tiempo y dado el carácter global de una parte de las actividades criminales, en los últimos años se han realizado importantes esfuerzos en la estandarización e interconexión de mucha de esta información.

La existencia y disponibilidad de todos estos conjuntos de datos supone una oportunidad, para aquellos usuarios autorizados, como fuente de nuevos datos y evidencias derivadas del cruce de datos entre estas bases de datos y/o con otras externas no asociadas directamente con la actividad criminal (acceso a datos bancarios, de compañías aéreas, etc..).

- Incremento de la cantidad e importancia de los datos asociados con la seguridad de la información y las tecnologías de la información.

Hoy en día, uno de los desafíos más preeminentes en el ámbito de la seguridad en TICS, incluyendo la ciberseguridad/ciberdefensa, es el procesado y análisis de la ingente cantidad de datos e información asociados a la seguridad de la información y las comunicaciones. La importancia de la seguridad de redes IT y la creación de unidades de ciberseguridad y ciberdefensa, derivadas de la emergencia en cantidad, dificultad y relevancia de los ataques cibernéticos, demandan más y mejores herramientas de apoyo al análisis forense y de vulnerabilidades asociadas con estos

- Iniciativas gubernamentales

Toda esta proliferación de datos y el hecho de que muchos de estos datos se generen, almacenen y potencialmente se exploten en el ámbito gubernamental (seguridad, defensa, pero también sanidad, energía, gestión local, etc.) ha hecho que diferentes gobiernos nacionales y administraciones tanto regionales como supranacionales, estén promocionando diferentes iniciativas relacionadas directa e indirectamente con Big Data.

En el ámbito nacional, hay que destacar la puesta en marcha del portal de datos públicos (http://datos.gob.es/datos/). De manera similar la Unión Europea ha lanzado un portal a nivel europeo (http://open-data.europa.eu/).

Además, en el marco de la investigación europea, existen diferentes proyectos sobre Big Data y sus aplicaciones, entre los que destacan:

1. BIG - Big Data Public Private Forum

Este proyecto se centra en la definición de una estrategia que aborde los esfuerzos necesarios, en el ámbito de la investigación y la innovación, para la implementación exitosa del potencial y beneficios de Big Data en la economía europea.

2. Scalable data analytics

Herramientas y capacidades para desplegar y gestionar procesos de análisis de datos de forma robusta y con gran rendimiento sobre cantidades de datos extremadamente grandes. Centrado en algoritmos escalables, marcos de software, visualización, y optimización de redes de Big Data y de hardware.

3. Content analytics and language technologies: Cross-media content analytics

Métodos y herramientas innovadores para minería de información no estructurada embebida en texto, voz, audio y vídeo, para los propósitos de interpretación de la conciencia contextual, correlación, agregación y resumen, transformación de la información en conocimiento usable y procesable. Se pone especial énfasis en la inteligencia social y colectiva de fuentes multilingües. Los proyectos deberán lograr una amplia cobertura con una interpretación semántica eficiente. Es de gran interés

la habilidad de capturar sentimientos y representar conceptos y eventos, identificar relaciones y similitudes, interpretar tiempo y espacio, dentro y a través de medios individuales, y de incrementar nuestra habilidad para detectar y explotar significados que de otra forma permanecerían ocultos a través de diversas aplicaciones.

4. IMPART - Intelligent Management Platform for Advanced Real-Time media processes

IMPART investigará, desarrollará y evaluará soluciones inteligentes de gestión de la información para problemas de Big Data en el campo de la producción cinematográfica digital.

5. OPTIQUE - Scalable End-user Access to Big Data –

El acceso escalable para el usuario final al Big Data es crítico para conseguir un análisis de datos efectivo y la creación de valor. OPTIQUE traerá un cambio de paradigma para el acceso a datos reduciendo el tiempo de los plazos de entrega de las peticiones de información a minutos, en lugar de días.

Por otro lado, también es necesario destacar la iniciativa estadounidense de promoción del Big Data donde el Departamento de Estado de EE.UU. y el de Defensa han identificado Big Data como "una gran apuesta" transformadora, tanto a nivel federal como en el Departamento de Defensa (DoD). Para ello el DoD ha comprometido 250 M$ anuales, de los cuales 60 M$ se dedican a Investigación, a inversiones en relacionadas con Big Data en las áreas de transformación de datos en decisiones, autonomía, y sistemas de interfaz hombre-máquina.

Hay que destacar que estas áreas coinciden con tres áreas "multiplicadoras de la fuerza" de las siete áreas prioritarias en Ciencia y Tecnología identificadas por el DoD para 2013-2017.

También es necesario destacar que la DISA (Agencia de Sistemas de Información de Defensa) ha incluido las tecnologías para Big Data como estratégicas dentro de su Plan Estratégico 2013-2018.

- Nuevos escenarios de seguridad y defensa

En los nuevos escenarios de seguridad y defensa la complejidad y la adaptabilidad de las amenazas han sobrepasado nuestra capacidad para encontrar sus trazas en grandes volúmenes de datos dentro del tiempo de la misión.

En el caso de la defensa, hay que tener en cuenta que el estado final del escenario global supone que un grupo de operadores móviles desplazándose por el campo de batalla con el apoyo de una mezcla de sensores fijos y móviles, datos de inteligencia y con conectividad a una infraestructura de apoyo al teatro de operaciones, con recursos mínimos en personal, capacidad de computación y ancho de banda. Las amenazas modernas son asimétricas y a menudo no representan a estados nación. El campo de batalla es más amplio y los oponentes se mezclan con el entorno cultural y

empleando tácticas y procedimientos no tradicionales. Las actividades asociadas a la amenaza ocurren en escalas de tiempo variables, además las firmas de estas actividades son en general parcialmente desconocidas. El desarrollo de capacidades para identificar rápidamente la firma de estas tácticas y procedimientos de estas amenazas es esencial.

- Popularización y creciente mercantilización de infraestructuras y herramientas para la "computación en la nube".

Hasta hace unos años, la computación de altas prestaciones (High Performance Computing - HPC) sólo estaba disponible para las universidades y grandes instituciones debido al alto coste de los equipos. Actualmente, la reducción del coste de servidores y almacenamiento permite que la HPC sea más barata. Por otro lado, la proliferación de servicios de HPC en la nube, ofertados por empresas como Amazon, Microsoft y Google, entre otras, hace aún más sencillo el acceso a una elevada capacidad de computación y su aplicación a diferentes volúmenes de datos y problemas, permitiendo la generación de soluciones innovadoras y la personalización de aplicaciones asociadas a problemas concretos.

3.3.2.- Relevancia de cada uno de los factores impulsores en las aplicaciones específicas

Para cada una de las 12 aplicaciones específicas se ha estimado la relevancia de cada uno de estos factores impulsores. Utilizando una clasificación de si/no, se han marcado en gris aquellas aplicaciones y factores donde existe un cierto grado, en mayor o menor medida, de influencia del factor en el desarrollo de la aplicación.

EL cuadro resultante puede verse en la figura de la página siguiente.

	Factores impulsores						
	Incremento del número de sensores y la "internet de las cosas"	Limitaciones de capacidad y/o acceso al almacenamiento de datos	Incremento de la existencia y la disponibilidad de datos relacionados directamente o indirectamente con actividades criminales	Incremento de la cantidad e importancia de los datos asociados con la seguridad de la información y las tecnologías de la información	Iniciativas gubernamentales	Nuevos escenarios de seguridad y defensa	Popularización y creciente mercantilización de infraestructuras y herramientas para la "computación en la nube"
Detección de intrusión física en grandes espacios o infraestructuras abiertas	■	■					
Computación sobre información encriptada			■	■			■
Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos)				■	■	■	■
Criminología computacional			■	■	■	■	
Uso fraudulento de recursos corporativos y/o sensibles				■			■
Análisis de video en tiempo real / Búsqueda y recuperación rápida en librerías de video.	■	■			■	■	
Inteligencia visual en máquinas	■	■			■	■	
Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos.	■	■		■	■	■	■
Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos.	■	■	■		■	■	■
Consciencia situacional	■	■	■	■	■	■	
Traducción automática a gran escala (en número de idiomas y en volumen)			■		■	■	■
Predicción de eventos	■		■	■	■	■	■

-

3.3.3 Listado de factores limitadores / desafíos existentes

- Aspectos éticos y legales

Uno de los mayores desafíos a los que se enfrenta la aplicación de Big Data es el del cumplimiento de los aspectos éticos y legales en la utilización y explotación de los datos. Este desafío es especialmente relevante en el área de la seguridad, aunque otras áreas de potencial interés para Big Data como la de sanidad también se ven afectadas muy directamente por este desafío.

Las organizaciones en general, pero en especial las públicas, que reciben y acceden a multitud de datos personales y de carácter sensible, deben desarrollar y adoptar códigos de conducta de responsabilidad en el análisis y manejo de estos datos y resultados. Es posible que las leyes al respecto existentes hasta la fecha no cubran todo el abanico de posibilidades y usos en la captación, análisis y explotación de datos que Big Data está previsto que ofrezca, y que por lo tanto sea necesario actualizarlas.

- Necesidad de equipos multidisciplinares para definir las soluciones

Big Data no es un software que se ejecuta y produce informes de manera automática, tampoco consiste en un conjunto de sistemas informáticos que una vez instalados y configurados, empiezan a ofrecer aplicaciones y soluciones, como las identificadas en el anterior apartado.

Para ello, es necesario un correcto enfoque en el planteamiento del problema, y un análisis de cómo las técnicas y herramientas de Big Data pueden aplicarse a la solución del problema.

En la literatura sobre Big Data generada en los dos últimos años, se identifica para la definición de las soluciones la necesidad de un perfil que se ha venido a denominar como "científico de datos" ("*data scientist*"). Este perfil deberá aunar habilidades y capacidades en los ámbitos de:

- Conocimiento detallado de las fuentes de datos, así como la tipología de estos. El conocimiento de la calidad y naturaleza de las fuentes y datos es esencial para la identificación y análisis de patrones en el conjunto de datos.
- Algoritmia estadística y de aprendizaje automático, desde el punto de vista teórico y práctico.

En algunas ocasiones, a estos dos aspectos se añade que estos "científicos de datos" deban conocer y entender bien el contexto y complejidad del entorno de aplicación, que en el caso de este estudio, sería cualquiera de las aplicaciones en seguridad y defensa. Si bien esto siempre es deseable, hay que reconocer la dificultad de encontrar perfiles que aúnen estos tres aspectos con la profundidad suficiente.

Por ello, se entiende como un planteamiento mucho más práctico, la formación de

equipos multidisciplinares que en conjunto sumen estas tres perspectivas.

- Incorporar el contexto de los datos en el análisis

De manera muy ligada a la necesidad de equipos multidisciplinares, para el correcto análisis y explotación de los datos es necesario incorporar el contexto de los datos, así como las limitaciones del propio conjunto de datos.

Aunque existen referencias[12] a la no necesidad del contexto y la suficiencia de una gran cantidad de datos para la identificación de correlaciones relevantes, la realidad es que en la gran mayoría de las situaciones, este planteamiento lleva a errores.

Hay que tener en cuenta que algunos datos son intrínsecamente inciertos, por ejemplo las señales GPS que rebotan entre los edificios. Así en este ejemplo, el contexto puede aportar criterios y nuevas fuentes, como comentarios sociales añadidos a la información acerca de una ubicación geoespacial. La combinación de múltiples fuentes menos fiables puede dar lugar a un punto de datos más preciso y útil.

Además de necesario, incorporar el contexto de los datos al análisis, permite reducir la utilización de los recursos (computación, almacenamiento, etc.), al permitir enfocar el planteamiento de Big Data solamente en aquellas fuentes que son más relevantes para dicho contexto, o en diferentes subconjuntos de datos de acuerdo a diferentes criterios de filtrado en ese contexto.

Finalmente, el contexto resulta muy útil para que en el análisis de los datos se caiga en la trampa de la apofenia, que consistente en ver patrones, conexiones o ambos en conjuntos de datos que en realidad no poseen ningún sentido ni relación entre ellos. Nuestros cerebros están predispuestos a identificar patrones en conjuntos de datos. Disponer de un contexto que aporte guías y criterios, ayuda a evitar identificar conclusiones erróneas en el análisis de volúmenes de datos.

- Técnicas y herramientas para la visualización de los datos y los análisis

Con la emergencia del tratamiento de grandes volúmenes de datos, en tiempo real y con posibilidad de que estos no estén estructurados, se ha incrementado el interés en las técnicas y herramientas para la visualización de los datos y los análisis, tanto en el ámbito académico como en el empresarial. Esta visualización en el ámbito de Big Data pretende combinar las fortalezas humanas y de los ordenadores para el procesado de datos.

El objetivo es que esta visualización permita la participación activa de una persona en el proceso de análisis a través de interfaces visuales interactivos, y apoyar en la comunicación de resultados del análisis a los destinatarios relevantes.

1 *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* por Chris Anderson

2 Butte Lab del Stanford Center for Biomedical Informatics Research

En la actualidad existen numerosas y diversas herramientas informáticas, la mayoría son evolución de plataformas de Business Intelligence, y son una primera aproximación a la complejidad de estas visualizaciones. En el ámbito de la visualización en Big Data, los ámbitos académicos y empresariales se enfrentan, en general, a diversos desafíos, entre los que destacan:

- Resolver la variedad de datos multidimensionales y aspectos espacio-temporales.
- Énfasis en el análisis de datos, la resolución de problemas y/o toma de decisiones;
- Como la mayoría del Big Data es dinámico y referenciado temporalmente, es necesario tener en cuenta los aspectos específicos del tiempo. En contraste con las dimensiones normales de los datos, que suelen ser "planos", el tiempo tiene una estructura semántica inherente.
- Otro asunto clave es el apoyo al análisis en escalas múltiples. Queda mucho por hacer en análisis visual para lograr cambiar. Las escalas apropiadas del análisis no siempre están claras a priori. Las interfaces visuales interactivas tienen un gran potencial para facilitar la búsqueda empírica de escalas de análisis aceptables y para la verificación de resultados modificando la escala y los medios de cualquier agregación.
- Correlación y causalidad

La correlación entre datos es una de las técnicas fundamentales para la búsqueda de información en grandes volúmenes de datos. Partiendo de esta afirmación, es importante realizar algunas apreciaciones asociadas a los conceptos de correlación y de información, y que son de importancia a la hora de aplicar Big Data, en especial en lo que se refiere al análisis.

En relación a la primera, la correlación, es una de las técnicas más utilizadas para encontrar trazas de información en datos en bruto. Pero en el momento en que estas correlaciones se utilizan como base para la formulación de modelos y teorías, se empieza a entrar en terreno resbaladizo. La correlación no establece la causalidad. Al contrario que con la deducción, que uno puede demostrar que funciona de manera general, la inducción, solo demuestra que esa relación ha funcionado en el pasado, sin establecer ningún fundamento para que esas correlaciones puedan servir de base para modelos y teorías acerca de esos datos. Este hecho tiene especial importancia en el área de la seguridad y la inteligencia. Así puede pasarnos como al pavo inductivista en la paradoja de Bertrand Russell, en el que se infieren hechos no sólo equivocados, sino que cuando mayor es el nivel de certeza en nuestras inferencias, debido al mayor número de observaciones/correlaciones, es precisamente cuando más equivocados y expuestos estamos a la realidad no representada por los datos/correlaciones. Por otro lado, también es importante tener en cuenta que las técnicas de Big Data (y algunos de los volúmenes de datos) también puede ser accesibles y explotados por el delincuente/adversario para identificar cuáles son las correlaciones más relevantes y actuar en

consecuencia de manera divergente a estas correlaciones.

No obstante, en algunas aplicaciones específicas las correlaciones son información suficiente para inferir conocimientos y perspectivas. Ejemplos de estas aplicaciones específicas son el filtrado colaborativo y las asociaciones tiempo-lugar en el ámbito social.

El segundo aspecto a destacar tiene que ver con el concepto de información y su trasposición por conocimiento e inteligencia. Así en algunos casos, el valor de Big Data se exagera al no tener en cuenta que el conocimiento e inteligencia extractable será muchísimo menor que el volumen de información extractable, y este a su vez muchísimo menor que el volumen de los datos.

Para poder recorrer el camino desde los datos hasta la inteligencia, es necesario que:

- Los datos sean interpretables: hay que tener en cuenta que dado que muchos de los datos pueden ser no estructurados, la interpretación de esos datos no estructurados pueda no ser posible, o por lo menos al coste adecuado.
- La información extractada sea relevante. La información extraída del volumen de datos debe ser relevante para poder ser útil en la generación de conocimiento e inteligencia. Cuidado, la relevancia es un aspecto subjetivo y contextual.
- La información debe ofrecer conocimientos o bien no disponibles o bien disponibles pero no identificados. Al igual que la relevancia, este un aspecto que es subjetivo.
- Infraestructura necesaria para el análisis en tiempo real

Aunque una de las características asociadas a Big Data es el análisis en tiempo real de "streams" de datos, las capacidades de sistemas existentes para procesar tal flujo de información y responder a búsquedas en tiempo real (con la posibilidad de realizarlo para miles de usuarios concurrentes) son limitadas.

Visto de otra manera, la infraestructura y medios necesarios para realizar los análisis en tiempo real en muchas de las aplicaciones son muy costosos, y pueden desafiar la viabilidad o lógica del proyecto entero.

Además, no hay que olvidar que las tareas analíticas requeridas para el procesado de las tareas analíticas que se requieren para el procesado de flujos de datos son tan intensivas en conocimiento que se necesitan tareas de razonamiento automatizado.

En conclusión aunque existen soluciones ya plenamente funcionales (como por ejemplo el proyecto de IBM Terraechos) el problema del procesado eficaz y eficiente de flujos de datos en tiempo real está lejos de ser resuelto, incluso si se tienen en cuenta los avances recientes en bases de datos NoSQL y tecnologías de procesado en paralelo.

- Calidad de los datos

Charles Babbage, originador del concepto de ordenador programable y al que se atribuye el invento del primer ordenador mecánico que condujo a diseños más complejos, en su libro "Pasajes de la vida de un filósofo" reproduce esta pregunta que dice que le fue hecha por un miembro del parlamento al respecto de su ordenador mecánico,

"Mr. Babbage, ¿si pone en la máquina datos erróneos, se obtendrán las respuestas correctas?" A lo que Babbage contestó *"No soy capaz de comprender correctamente el tipo de confusión de ideas que puede provocar tal pregunta."*

Las ideas de Babbage sobre el asunto de la calidad de datos, fueron expuestas de forma nítida por George Fuechsel, un programador e instructor de IBM, un siglo después como "garbage in, garbage out," or "GIGO", al que se le da habitualmente crédito por acuñar el término. Este término incluso ha evolucionado hacia una variación del término, incluso más acorde con el desafío de la calidad de los datos en Big Data, que es "garbage in, gospel out," refiriéndose a la tendencia a poner una fe injustificada en la precisión de los datos generados por un ordenador.

Pues bien, aunque en estos ejemplos el problema de la calidad de los datos parece necesitar poco esfuerzo de comprensión, el hecho es, que contemplar la calidad de los datos es uno de los factores determinantes a la hora de aplicar Big Data, y no siempre es tenido en cuenta.

La comprensión de la calidad de los datos en un mundo de datos estructurados ha alcanzado un elevado grado de madurez. Desafortunadamente, no puede decirse lo mismo sobre los datos no estructurados. Asegurar el mismo nivel de calidad en los datos no estructurados de forma que no distorsionen el análisis subsecuente es mucho más difícil de aplicar.

Validar la enorme cantidad de información es un enorme desafío, dado que existe una gran variedad de tipos de fuentes y de contenidos. Además, la complejidad del lenguaje humano es tal que no es sencillo derivar reglas de validación que apliquen a todos las áreas.

La gran cantidad de datos que existe actualmente, y la frecuencia con la que es actualizada, hacen que la tarea de validación sea muy compleja y hacen que las reglas complejas que requieren enormes recursos computacionales sean inviables. Otra posibilidad es aplicar el contexto de los datos, así como propiedades intrínsecas de esos volúmenes de datos, para filtrar y depurarlos. En este sentido se utilizan técnicas de fusión de datos y de matemáticas avanzadas, como sólidas técnicas de optimización y planteamientos de lógica difusa.

- Sistemas propietarios y bases de datos compartimentares.

Uno de los desafíos a los que se enfrenta Big Data es la propia filosofía de gestión de la información y el conocimiento existente en muchas de las organizaciones, tanto públicas como privadas. Diferentes planteamientos y direcciones en la gestión

en muchas organizaciones provoca la existencia de diferentes bases de datos estancas a la utilización de otros departamentos en la organización. Esta estanqueidad entre las diferentes bases de datos, puede deberse a problemas técnicos (estructura de datos, diferente software, o aspectos estratégicos o legales, etc.), que en mayor o menor grado podrán ser solucionables desde el punto de vista de Big Data, o deberse a razones culturales dentro de la organización, que supondrán un desafío mayor para la aplicación de Big Data.

Por otro lado, también hay que tener en cuenta la existencia en muchas de las organizaciones de sistemas propietarios en los que la utilización de la mayoría de los datos queda velada e inaccesible al personal de la organización. Esta circunstancia tiene implicaciones en el aprovechamiento y explotación de los datos propios de la organización (como sucede en la estanqueidad de las bases de datos) y también la integración de estas herramientas propietarias con bases de datos externas a la organización.

Estas dos circunstancias concurren con cierta frecuencia en el ámbito de las administraciones ligadas a la defensa y a la seguridad (p.e. Ministerio de Defensa y Ministerio de Interior).

- Entornos de operaciones austeros en recursos (habitual ámbito de defensa)

Los entornos geográficos donde se realizan las operaciones militares, presentan por lo general recursos muy limitados en lo que respecta a energía eléctrica, espacio y capacidad de refrigeración. Estas limitaciones podrían ser soslayadas transmitiendo los datos a otras ubicaciones menos exigentes en lo que respecta a estos aspectos. Sin embargo, el ancho de banda disponible tampoco suele ser muy abundante. De hecho, en la actualidad a nivel del combatiente, existe una significada divergencia entre las posibilidades de generación de datos (video, imágenes VIS/IR, etc.) y las capacidades de transmisión de datos hacia otros miembros de la unidad y/o otras unidades / niveles jerárquicos superiores.

Una posible solución a esta divergencia es el 'Near Sensor Computing' (NSC) indicado anteriormente.

- Facilidad de uso en las herramientas

Se indicaba anteriormente como uno de los factores impulsores para Big Data está siendo la popularización y creciente mercantilización de infraestructuras y herramientas para la "computación en la nube". Sin embargo, son necesarios avances en la facilidad de uso de las herramientas para la captación, análisis, visualización y explotación.

El objetivo último es que los expertos en cada ámbito, como por ejemplo inteligencia militar, puedan aplicar sus conocimientos a soluciones Big Data con plataformas informáticas de fácil instalación y manejo.

- Inversiones en hardware y software

Dependiendo de cada organización y/o país este factor presenta mayor o menor

incidencia. En cualquier caso, está claro que para poder aplicar y aprovechar el potencial de Big Data, las organizaciones deben invertir recursos en la obtención infraestructuras (hardware y software) y *know-how*, a la vez que, posiblemente, actualizar o abandonar algunos de los sistemas de información actuales.

Según el tipo de aplicaciones de Big Data perseguidas y el tipo de organización, la dimensión de la inversión variará, desde la infraestructura computacional necesaria para poder capturar, procesar, buscar y recuperar millones de objetos por segundo, hasta infraestructuras de supercomputación que es posible subcontratar a diferentes proveedores de servicios.

- "Data obesity"

Contando con el factor impulsor del incremento del número de sensores y la "internet de las cosas" y con el constante decremento del precio por TB de almacenamiento, puede parecer que almacenar una gran cantidad de datos, no sólo es fácil, sino que también es barato, y que por lo tanto cuántos más datos se almacenen y/o se tenga acceso mejor para una organización. Sin embargo, este planteamiento, que en la literatura inglesa se ha acuñado con el término "*data obesity*", presentará problemas a medio plazo para las organizaciones.

Data obesity puede definirse como la acumulación indiscriminada de datos sin una estrategia apropiada para descartar aquellos datos innecesarios o no deseados. Según se avance en un mundo caracterizado por la abundancia de datos, la tasa de adquisición de datos aumentará exponencialmente. Incluso si el coste de almacenamiento de todos los datos fuera asequible, el coste de procesado, limpiado y análisis de los datos será prohibitivo. Una vez alcanzada esta fase, la "obesidad de datos" será un gran problema al que se tendrán que enfrentar las organizaciones en años venideros.

Dado que el mercado aún no está maduro y los problemas de *data obesity* no están entre las principales preocupaciones de las organizaciones, no se está viendo una proliferación de herramientas para gestionar la obesidad en los datos. Es previsible que esta situación cambie cuando las organizaciones entiendan los riesgos asociados a la *data obesity*.

En cualquier caso, más que el hecho de tener las herramientas necesarias para hacer frente a la *data obesity*, es importante que las organizaciones entiendan que esta abundancia superflua de datos puede prevenirse mediante un conjunto de buenas prácticas. Actualmente no hay estándares en la industria para evitarla, y cada organización acumula y genera datos de una forma distinta.

3.3.4.- Relevancia de cada uno de los factores limitadores/desafíos en las aplicaciones específicas

Para cada una de las 12 aplicaciones específicas se ha estimado la relevancia de cada uno de estos factores limitadores. Utilizando una clasificación de si/no, se han marcado

en gris aquellas aplicaciones y factores donde existe un cierto grado, en mayor o menor medida, de influencia del factor en el desarrollo de la aplicación.

El cuadro resultante puede verse en la figura de la página siguiente.

	Factores limitadores / desafíos existentes											
	Aspectos éticos y legales	Necesidad de equipos multidisciplinares para definir las soluciones	Incorporar el contexto de los datos en el análisis	Técnicas y herramientas para la visualización de los datos y los análisis	Correlación y causalidad	Infraestructura necesaria para el análisis en tiempo real	Calidad de los datos	Sistemas propietarios y bases de datos compartimentares	Entornos de operaciones austeros en recursos	Facilidad de uso en las herramientas	Inversiones en hardware y software	*"Data obesity"*
Detección de intrusión física en grandes espacios o infraestructuras abiertas				■	■	■	■		■		■	
Computación sobre información encriptada	■					■					■	
Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos)	■	■	■	■	■	■	■	■		■	■	
Criminología computacional	■	■	■	■	■		■	■		■		■
Uso fraudulento de recursos corporativos y/o sensibles	■		■	■	■			■				
Análisis de video en tiempo real / Búsqueda y recuperación rápida en librerías de video.			■	■		■			■		■	■
Inteligencia visual en máquinas		■	■	■		■			■		■	■
Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos.	■	■	■	■	■	■	■	■	■		■	■
Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos.		■	■	■	■	■	■	■	■		■	■
Consciencia situacional		■	■	■	■	■	■		■			■
Traducción automática a gran escala (en número de idiomas y en volumen)			■	■		■				■		
Predicción de eventos	■	■	■	■	■		■		■			■

4.- Casos concretos, aplicaciones

4.1.- Programa del combatiente del futuro (COMFUT)

Los numerosos avances tecnológicos producidos en el siglo XX, sobre todo en su último cuarto en las áreas de tratamiento de la información, han sido de los principales motores de desarrollo de la industria y de la sociedad en su conjunto. Se ha acuñado el concepto de *sociedad de la información* comprobándose que, al conocido dicho de que la información es poder, ahora se le puede dar un rasgo complementario: no basta con disponer de información, esta debe ser tratada y procesada para que permita tomar decisiones. Por tanto más que el poder venga de la propia información tal cual, dicho poder viene de tratar esta información, de extraer consecuencias cuando es procesada.

El procesamiento incluye múltiples funciones diferentes como la transmisión, el análisis, la correlación, el almacenamiento para poder volver a tratar la información y los resultados de los análisis anteriores cuando existan más elementos, etc. Unido a este concepto de procesado de información se ha producido un avance muy acusado en la miniaturización y además en materiales, equipos, etc. de muchos de los elementos que intervienen en el propio procesado. Con esta situación es obvio intentar aplicar estos avances a cualquier actividad, incluida la que realiza un combatiente. Algo similar ocurrió al comienzo del siglo veinte con la aparición del carro de combate que se basaba fundamentalmente en los avances que la industrialización permitía en aquellos momentos; si bien esto no era tan novedoso pues Leonardo Da Vinci ya había esbozado un modelo de "tanque" algún que otro siglo antes.

La política de armamento del Ministerio junto con la participación en foros y organizaciones internacionales dieron pié a la creación de un programa de I+D. Este programa ha desarrollado varios proyectos dentro de uno solo global aplicando en su concepción los siguientes principios:

- **Potenciación de las capacidades individuales del combatiente**. La forma en que de desarrollan las actividades de una unidad militar tienen en común que el soldado debe realizar una serie de actividades fuera de su medio de transporte y de la protección de una posición estable. Dependiendo de las misiones de la propia unidad, y de las que específicamente deba desarrollar cada combatiente, es necesario contemplar que cada actuación individual es la base de la acción de toda la unidad llegando en muchos casos a ser esta una consecuencia directa, sino una simple suma, de las acciones individuales. Estas deben ser coordinadas de modo que cuanto más efectiva sea esta acción individual mejor se cumple la misión, con menos desgaste y mayor rapidez.
- **Integración en la unidad superior y en su medio de transporte**. La necesidad de movimiento hoy en día es indiscutible de hecho la denominación de la

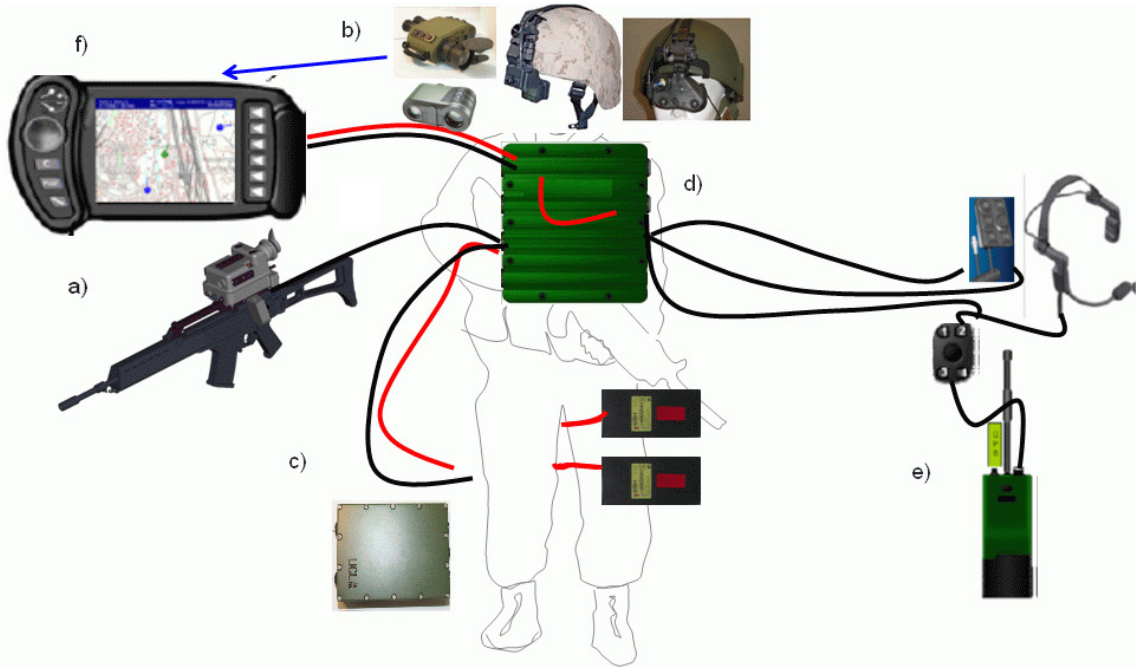
actuación de las unidades y de sus componentes es la "acción" que de modo implícito ya sugiere movimiento. Consecuentemente no se concibe un combatiente aislado sino como parte de un todo mucho más amplio. Tradicionalmente se hablaba del "binomio" cada soldado se "asociaba" con otro para mejorar todas sus capacidades. Actualmente este aspecto se ha generalizado y se entiende que la acción individual se desarrolla en el seno de la unidad a la que pertenece. Más aún no solo se le ve como combatiente sino también como elemento de inteligencia, sensor, la información que se puede obtener de la situación privilegiada de cada combatiente es de un valor difícil de desdeñar.

- **Condiciones de seguridad y protección individual y colectiva.** La eficacia del trabajo de una persona depende en gran medida de las condiciones de seguridad, facilidad para respaldar su acción y atención que se le puede dar en caso de accidente o baja, lo que ya por si supone un refuerzo moral. Se trata de un factor eminentemente humano que se concreta en proporcionar cobertura, protección y conocimiento del estado y situación en que se encuentra tanto de modo individual como en su equipo. Una de las circunstancias que más mina la moral es el fratricidio, es fundamental que cada uno se vea y sienta que está a salvo de ese error del compañero.

Con estas premisas se ha agrupado el equipamiento con que se pretende dotar al soldado en varios Subsistemas componentes:

- Armamento
- Fuentes de Alimentación
- Eficacia de Fuego
- Información y Comunicación (C4I)
- Supervivencia y protección
- Sostenimiento
- Integración en vehículo y Adiestramiento

De modo gráfico se aprecian los componentes de algunos de estos subsistemas en la siguiente figura:



a) Arma de fuego con diferentes visores

b) Binoculares con designador y telémetro

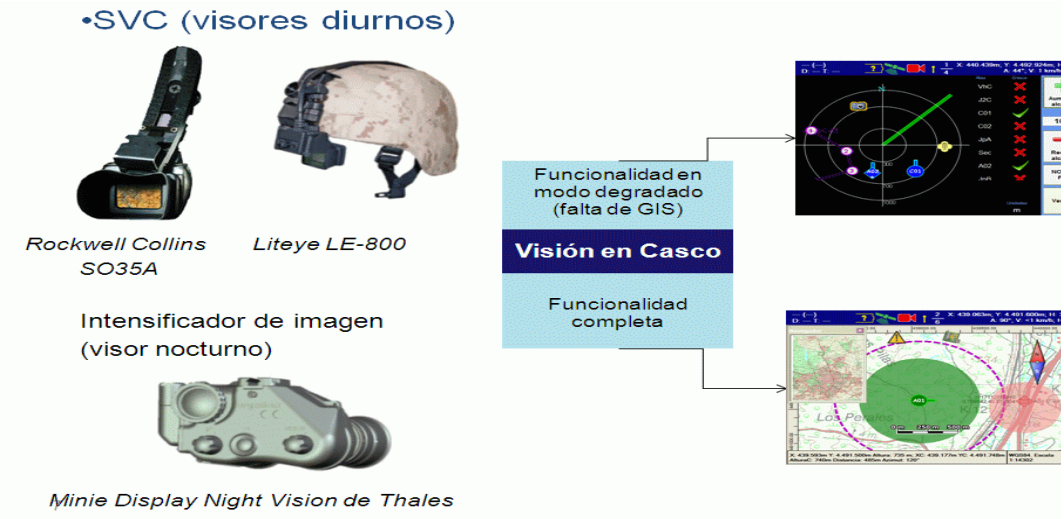
c) Conjunto de alimentación con diferentes baterías

d) Panel de interconexiones

e) Radio personal

f) Asistente digital

Un detalle de cómo se puede presentar la visión que se ofrece al propio soldado se ve a continuación bien a través del asistente digital o bien con el visor del arma o casco.



Es evidente que una presentación de los elementos que tiene alrededor para cualquier soldado es muy importante, sobre todo en los momentos de baja visibilidad, que suelen ser la mayoría. Por otra parte el asistente digital le permite una cierta capacidad de almacenamiento y procesado "in situ". La figura siguiente muestra algunas de las facilidades que puede aportar. Hasta ahora el único modo de disponer de algo parecido era fiarlo a la transmisión por radio, o del elemento de comunicación que tuviera, y hacerlo en remoto. Este mismo elemento permite disponer en línea de alguna facilidad de mapas o croquis del entorno y por supuesto de la forma de representar al resto de individuos presentes en la acción.

		ComFutv1
Software	Sistema operativo	Windows XP
	Motor GIS	SIGMIL
	Modelo de terreno	SIGMIL
	SW de desarrollo	C#
	Arquitectura SW	4-tier
	GUI	Tipo teléfono móvil
	Modelo de datos	Basado en JC3IEDM
	Modos de visualización SA	Basada en GIS y Radar
	Simbología	APP6-a Iconos personalizados

La radio tradicionalmente se concebía como un medio de dar y recibir órdenes por lo que su empleo se orientaba más al jefe del equipo (Pelotón, Sección, etc.) que a los componentes; actualmente el empleo de este tipo de medios se entiende que potencian las capacidades tanto de modo individual como de coordinación de toda la unidad.

En esta versión se ha elegido la radio SpearNet de ITT que como características más revelantes se tiene:

Permite el intercambio de información (Voz-Datos-Vídeo) entre los combatientes con tasas entre 100 y 1.500 Kbps.

Interfaz Hombre – Máquina sencillo

Banda de Frecuencias Sintonizables (1.2-1.4GHz)

Espectro Ensanchado:

- Mejora fiabilidad y seguridad para voz y datos
- Protección frente a interferencias y perturbaciones.
- Baja probabilidad de interceptación.

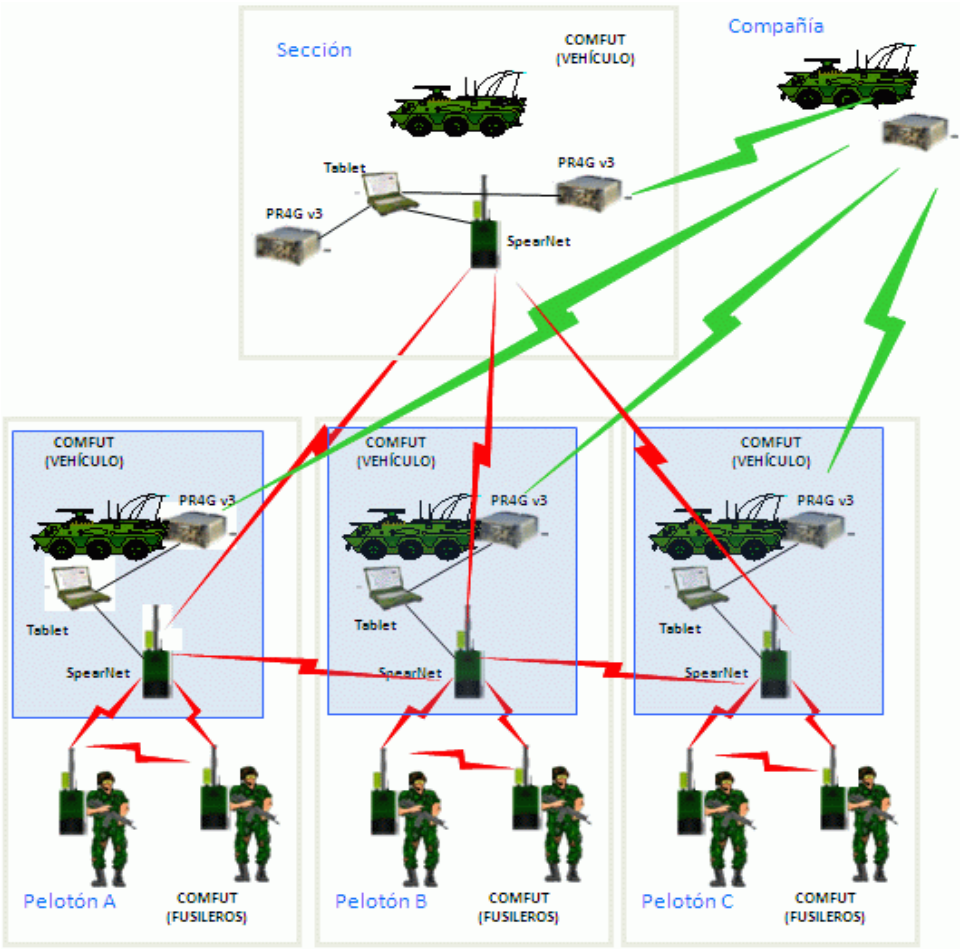
Interfaces: Ethernet, USB, RS 232.

Direccionamiento 100 % IP "multihop"

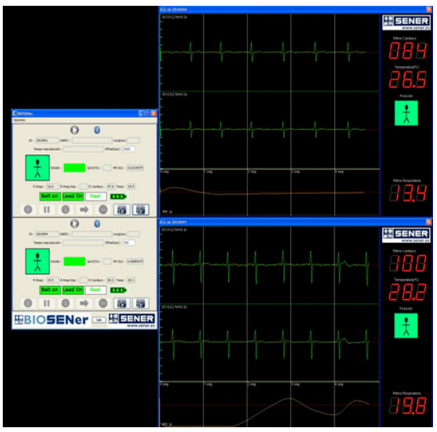
Potencia de salida de 600 mW.

Peso < 700 gr.

La configuración de redes se muestra en el siguiente gráfico. Se trata de una configuración de una **RED MESH** dentro del pelotón de modo que exista un acceso a la red general de todos sus componentes bien directamente o bien mediante algún salto intermedio. Esa red general es la red radio de combate que actualmente está formada por equipos PRG4v3.



En color rojo se muestra la formación de una o varias redes Mesh de modo interno en cada pelotón o entre varios de ellos de modo que puedan llegar a nivel de sección. Su integración con la Red Radio de Combate se puede realizar en el vehículo del pelotón o bien en el nivel de Sección o incluso superior. Esta configuración puede ser cambiante según la situación. En todo caso una de las primeras limitaciones se manifiesta en este punto, por el bajo volumen de tráfico que puede cursar la Red Radio de Combate frente al que puede conseguirse en estas redes Mesh.



Como último elemento del equipo individual se pensó en un sensor que monitorizara algunos parámetros fisiológicos del combatiente con objeto de poder ayudarle o tener una idea más rica de la situación que tiene. Este elemento ha sido descartado para la actual versión del programa pero aquí podemos contar con el ya que será normal que se incorpore en versiones futuras.

Para más información sobre este programa se puede consultar el Boletín Tierra nº 182 de 21 de diciembre de 2010 o el Boletín de Observación Tecnológica nº 30 del primer trimestre de 2011. En ambas publicaciones se amplía la información aquí reflejada.

En resumen en un despliegue típico de una unidad tipo Sección se compone de tres o cuatro Pelotones con un número variable de más de treinta combatientes. El equipamiento de cada combatiente no es uniforme ya que, como se mencionó anteriormente, las capacidades a disponer son de la Unidad en conjunto no de cada combatiente. Así se puede considerar que esta unidad típica dispondría de lo siguiente:

Fuentes de información

Concepto	Cantidad	Total	Información a tratar
Binoculares con telémetro	1 / Pelotón	4	Vídeo
Visores de arma	2 / Pelotón	8	Secuencia de imágenes fijas
Sensores Biométricos	1 / combatiente	35	Datos de baja velocidad
Sensores de posición (GPS –Galileo)	1 / combatiente	35	Datos de baja velocidad
Voz	1 / Pelotón	4	Voz no continua

Comunicaciones

Estaciones y redes	Cantidad	Tasa de datos
Estación de la Red Radio de combate	1 / Pelotón y 1 de la Sección	128 kbps.
Estación individual de cada red de intercambio local	1 / combatiente	Hasta 1,5 Mbps

Computación

Elemento	Cantidad	Total
Asistente digital	1 / combatiente	35
Computador tipo PC	1 / vehículo	5

Almacenamiento

No se dispone de ningún equipamiento específico para esta función así que se debería emplear los recursos de los asistentes digitales individuales así como los disponibles en cada computador de vehículo. Lógicamente la posibilidad de añadir algún dispositivo

en el propio vehículo para poder descargar información de los asistentes digitales individuales está abierta.

4.2.- Inteligencia de imágenes

Uno de los objetivos tradicionales de la tecnología consiste en conseguir que las máquinas trabajen con la misma potencia que el cerebro humano. Hasta ahora las máquinas ya habían batido a los humanos en tareas repetitivas o en cálculos matemáticos complejos. Pero hasta ahora no han conseguido realizar razonamientos avanzados.

El cerebro humano es capaz de procesar millones de imágenes, movimiento, sonido y otra información de múltiples fuentes. El cerebro es excepcionalmente eficiente y eficaz en su capacidad para dirigir un curso de acción y, en este sentido, se encuentra por encima de cualquier ordenador disponible en la actualidad.

Uno de los sentidos humanos más complejos es la vista y por ello, el desarrollo de un sistema equivalente se convierte en un reto que va evolucionando a medida que la tecnología evoluciona. La tecnología de procesamiento de imagen y de video ha crecido con los avances en múltiples campos como en visión artificial (visión por computador), en el reconocimiento del habla, en el desarrollo de sistemas basados en reglas (y sus motores de decisión) etc. El objetivo consiste en aunar todas estas tecnologías para el procesamiento de video (flujos de imagen o *streams*) en tiempo real.

La demanda de este tipo de tecnologías viene incrementándose año tras año por la facilidad de acceso a la producción de video. Hoy día los teléfonos móviles inteligentes poseen una cámara de video y capacidad de cálculo suficiente para ofrecer servicios de video que, o bien se procesan en el propio dispositivo, o bien se procesan en la nube (facilitando el acceso al procesamiento de video). Estamos en la era multimedia y en estos dispositivos, el ancho de banda empieza a no ser un problema.

Pero esa demanda no proviene exclusivamente del entorno de los usuarios privados individuales, sino que además también proviene del entorno empresarial y del entorno de las administraciones públicas, donde se encuentran los ámbitos de Defensa y Seguridad. Las imágenes y secuencias de imágenes (vídeos) representan alrededor del 80 por ciento de todos los grandes volúmenes de datos no estructurados corporativos y públicos. Parece lógico que exista una demanda elevada de extraer la información contenida en esos vídeos. Estas grandes cantidades de datos no estructurados, los grandes sistemas de procesamiento con muchas tecnologías involucradas y las grandes posibilidades de extraer información adicional son las bases para poder identificar aplicaciones de Big Data en el procesamiento de imágenes.

Este escrito tratará de describir los sistemas de procesamiento de imágenes, la tecnología subyacente que permite usarlos y las demandas y necesidades de Defensa. En paralelo se tratará de explicar cómo las tecnologías de Big Data pueden ser una herramienta de la que estos sistemas puedan beneficiarse.

4.2.1.- Justificación

Con el entorno descrito en la introducción, el procesamiento de imágenes o análisis de imágenes, aparece como una posible solución a las necesidades de incorporar información no visual pero que procede de esas imágenes de vídeo. Más formalmente, el análisis de imágenes consiste en la extracción automática de la información encontrada en las imágenes digitales a través de algoritmos y/o procesamiento lógico. El reconocimiento de caracteres (OCR en sus siglas en inglés) fue uno de los primeros problemas en ser abordado dentro del procesamiento de imágenes y de encontrar soluciones de uso aceptables. Actualmente el uso de códigos QR es otro ejemplo simple pero interesante. A partir de este punto, se pueden encontrar otros ejemplos más complejos, como el reconocimiento facial y la deducción de la posición o el análisis del movimiento.

Volviendo a la comparativa biológica, una imagen es un conjunto de señales detectadas por el ojo humano y procesado por la corteza visual del cerebro creando una experiencia viva de una escena que se asocia inmediatamente con conceptos y objetos anteriormente percibidos y grabados en la memoria.

Para un ordenador, las imágenes sólo son o una matriz bidimensional de datos donde se almacena en cada elemento la información sobre la iluminación de un punto de la escena visualizada (esta es la versión en blanco y negro de una imagen, aunque extrapolable a una imagen en color). La iluminación máxima en un punto se correspondería con un valor máximo (255 en la mayoría de los casos), y una ausencia de iluminación se correspondería con un valor mínimo (un cero). En esencia, para un ordenador, la escena visualizada no es más que esta matriz rellenada de este modo.

Pero la información lumínica sólo se corresponde con los datos "crudos" obtenidos de un sensor (la cámara). El cerebro humano obtiene de un modo más o menos inmediato información adicional, por lo que un sistema de visión deberá analizar estos datos. Para llevar a cabo el análisis de imágenes o vídeos, la codificación lumínica debe transformarse en construcciones que representan objetos, rasgos físicos y movimiento de los objetos de la escena. Esta información, consistente en la información de los objetos y del movimiento, es la que los ordenadores utilizan para extraer conocimiento de la escena.

4.2.2.- La tecnología del procesado de imágenes

Una definición simple de procesado de imagen podría ser la siguiente: es la transformación de unas imágenes y/o flujo de imágenes (vídeo) como elementos de entrada para la obtención de datos de interés. El interés de los datos obtenidos no es más que la finalidad de un sistema de visión, y en gran medida, este interés será uno de los factores más importantes que determinen la arquitectura de un sistema de procesado de imágenes.

No obstante, los sistemas de visión artificial[3] llevan desarrollándose durante mucho tiempo y, aunque en gran medida el proceso sigue siendo un arte, el camino recorrido es largo, habiéndose conseguido avances significativos en muchas y variadas aplicaciones. La diversidad de aplicaciones es elevada, pero los sistemas de visión artificial suelen tener una base tecnológica similar. Desde el punto de vista físico, los sistemas se pueden subdividir en el los elementos físicos y los elementos lógicos:

- ***Elementos físicos***

- Iluminación y captura de imágenes.

Estos dos elementos se corresponden con la parte más sensorial de un sistema de visión, y están ligados íntimamente. Las escenas donde se necesite que trabajen los sistemas de visión pueden ser muy variadas y la elección de estos componentes debe ser función de esta necesidad. Los componentes de iluminación y captura para una misión de vigilancia nocturna deben ser diferentes que para una aplicación de detección de defectos en un entorno industrial en una fábrica.

Con respecto a la iluminación hay que reseñar que no siempre se puede elegir el tipo de iluminación. Habrá necesidades en las que se pueda imponer una iluminación artificial (focos, lámparas, iluminadores,…) y habrá otras en las que no se pueda optar más que por la iluminación natural del entorno. Incluso más, en otras ocasiones, la luz natural será un impedimento y habrá que filtrarla de algún modo.

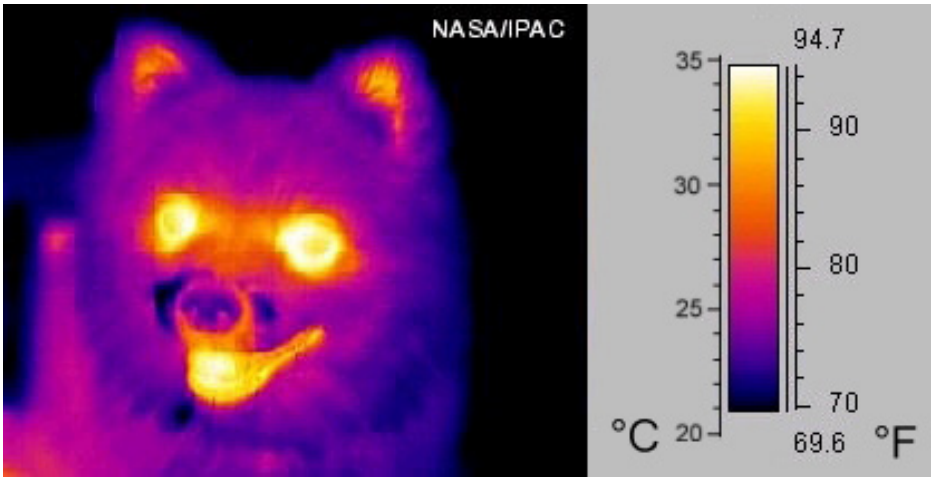


Figura 1: Imagen de un perro tomada con una cámara de infrarrojos.

Los sistemas de captura de imágenes son muy variados. Sin ser exhaustivos se pueden considerar como sistemas de captura a las cámaras CCD sensibles al espectro visible, cámaras infrarrojas, sistemas LIDAR donde la iluminación se realiza mediante láser,

3 La literatura en castellano es muy rica en la denominación de estos sistemas: sistemas de visión artificial, de visión por computador, de procesamiento de imágenes, o de procesado de imágenes. Aunque existen diferencias, se usarán indistintamente en este escrito.

o sistemas SAR donde la "iluminación" se realiza mediante una señal de radar que posteriormente se reconstruye en forma de imagen.

Figura 2: soldado con visor y recreación de la escena.

Adicionalmente los sensores pueden tener características adicionales. Por ejemplo, existen cámaras que capturan la imagen en dos tiempos, y reconstruyen una imagen con las filas pares de la primera captura y las impares de la segunda (video entrelazado, en contraposición con el video progresivo). Las cámaras de captura de imágenes a gran velocidad son otro ejemplo de características adicionales.

Un apunte adicional con respecto a las cámaras: no hay que olvidar que las ópticas que acompañan a las cámaras también son parte de la solución y que cada una se escoge en función del problema a resolver.

Figura 3: Inspección de un sensor térmico montado en un helicóptero OH-58D Kiowa Warrior.

- Tarjetas de adquisición y sistemas de procesado.

Estos componentes, siendo meramente funcionales, pueden encontrarse como dispositivos independientes o como partes integradas en otros. El primero se caracteriza por transformar las señales eléctricas provenientes de una cámara, en unos y ceros (o valores discretos entre 0 y 255). Los diferentes tipos de estándares de imagen hacen que estos sistemas se especialicen en función del tipo de sensor. Las cámaras CCD suelen dar como resultado una señal de vídeo similar a la antigua televisión terrestre (estándares PAL, NTSC, SECAM,…). Y esta señal debe ser transformada en valores útiles para su procesamiento por un ordenador (los ordenadores procesan números, no señales eléctricas).

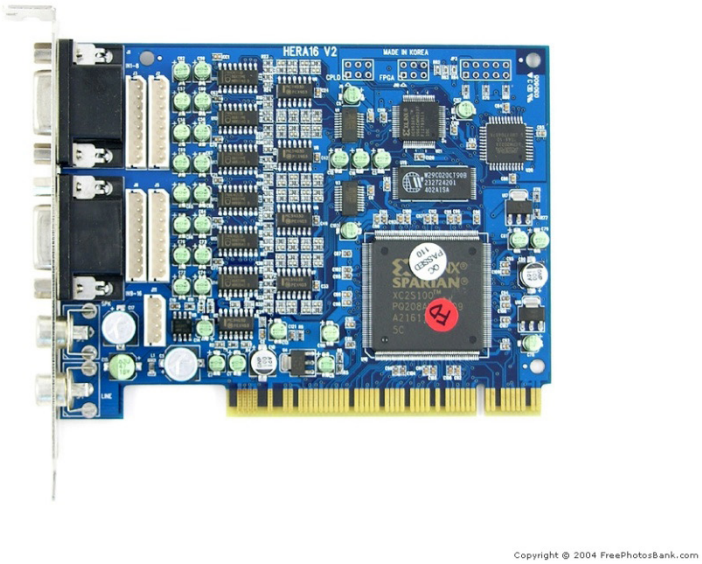


Figura 4: Tarjeta de adquisición de imágenes y/o video.

Los sistemas SAR (después de un complejo procesado interno) dan como resultado una imagen directamente (la matriz bidimensional de números) por lo que no necesitan este componente.

El segundo componente funcional es el sistema de procesado. Este componente es el que toma como entrada las imágenes como entrada y dará como resultado la información que se quería obtener. Será el que albergue los componentes lógicos del sistema. Físicamente se puede encontrar en forma de tarjeta electrónica que se incorpora a un ordenador, o un mismo ordenador, o incluso un micro-controlador (si el procesamiento es ligero).

La evolución de estos componentes viene de la mano de la evolución de la electrónica. La rapidez de procesamiento suele ser un requisito en la mayoría de estos sistemas. Aunque procesar una imagen puede resultar rápido en un ordenador moderno, el procesado de una señal de video de alta resolución que se presenta a 50 imágenes por segundo puede resultar computacionalmente hablando, bastante pesado.

Tecnológicamente, tanto las tarjetas capturadoras como los sistemas de procesamiento,

tiene en sus tripas componentes electrónicos de alto rendimiento. Tecnologías como los DSP, FPGA, CPLD[4] suelen ser habituales. Como curiosidad se puede mencionar la existencia de una tendencia a utilizar componentes de visualización (tarjetas gráficas 3D de los ordenadores comunes) para el procesamiento de datos, en este caso de imágenes. Esta tecnología aprovecha al máximo la posible paralelización de cálculos complejos.

- Sistemas integrados.

Se podría decir que los elementos mencionados anteriormente son elementos funcionales e, históricamente, aparecieron como componentes diferenciados. La tecnología microelectrónica evoluciona reduciendo cada vez más el tamaño de los componentes, permitiendo que en un circuito integrado puedan existir los componentes que anteriormente aparecían como elementos discretos. En el ámbito de la visión artificial también se ha dado este proceso. De este modo se pueden encontrar cámaras que incorporan internamente DSP o FPGA, o componentes de procesamiento que incorporan cámaras.

Una mención especial de este tipo es el desarrollo de la Universidad Carnegie Mellon[5] consistente en mini sistema de visión artificial. La cuarta versión del producto integra en una placa todos los componentes necesarios (menos la iluminación) para realizar complejos análisis de imágenes, además de presentarlos en un factor de forma compatible con la plataforma Arduino[6]. Esta plataforma es impulsora de una iniciativa de hardware abierto (los esquemas y programas son de libre acceso) y ha popularizado el uso de micro-procesadores, permitiendo que muchos profesionales y aficionados realicen desarrollos de una forma sencilla y relativamente barata. Esta "ola" favorece el intercambio de ideas y la colaboración entre grupos, facilitando las nuevas aplicaciones en visión artificial. La iniciativa también viene impulsada por otro fenómeno conocido popularmente como "el internet de las cosas", *grosso modo* los micro-procesadores incorporan capacidad de comunicación en internet, y la visión artificial también participa de esta "moda".

4 DSP: *Digital Signal Processor*; FPGA: *Field Programmable Gate Array*, CPLD: *Complex Programmable Logic Device*. Son tecnologías electrónicas de circuitos integrados programables y complejos.

5 http://www.cmucam.org/

6 *http://arduino.cc/*

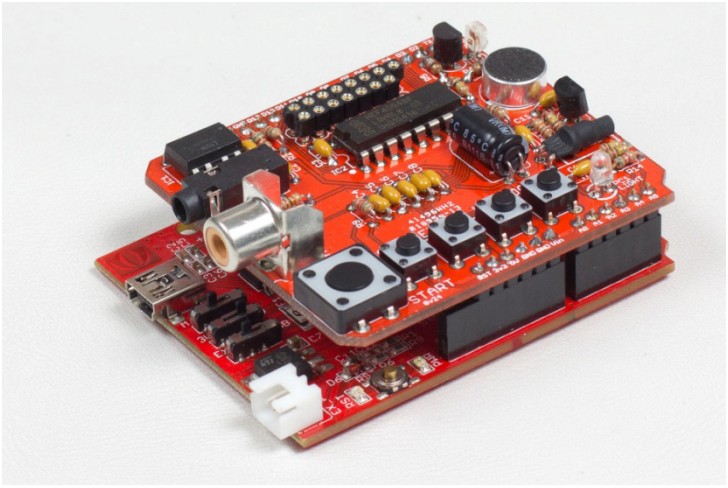


Figura 5: tarjeta (clon de Arduino) con controladora de video acoplada.

Un ejemplo más simple es la aparición de pequeñas placas[7] de no más de 3 cm x 3 cm donde se incorpora un CCD, una pequeña óptica y un mini-procesador. El mini-procesador se encarga de ordenar la captura de la imagen, y transformarla a un formato comprimido tipo JPEG y enviarla a través de un puerto serie interno a otro sistema para su posterior análisis. El sensor CCD, el mini-procesador y la unidad de control del puerto serie están incorporadas en la misma mini-placa.

- Grandes sistemas de procesado en paralelo.

El salto del procesamiento de imágenes en un sistema al procesado masivo del mismo pasa por las tecnologías de procesado masivamente paralelo. Hoy en día existen compañías y centros de cálculo que disponen de una infraestructura inmensa de potencia de cálculo. Esta potencia se consigue mediante la acumulación de múltiples ordenadores (o nodos). El *Grid Computing* surgió como una solución facilitadora en la gestión y ejecución de las aplicaciones. El ***grid computing*** consta de herramientas de varios niveles funcionales:

- Los primeros componentes surgieron como herramientas que facilitaban la comunicación entre los diferentes nodos de cálculo, de tal forma que los procesos se subdividían en tareas y el *grid computing* ayudaba en la tarea de coordinación mediante librerías de comunicaciones entre nodos.
- El siguiente nivel facilitador fue el correspondiente al almacenamiento de los datos. Los nodos del sistema no son necesariamente homogéneos, pudiéndose especializar en diferentes tipos. Este fue el caso de los servidores de datos, que son nodos especializados que acceden a sistemas de ficheros distribuidos o sistemas de gestión de bases de datos. El acceso a estos nodos facilita actividades de búsqueda en bases de datos y de minería de datos.

7 Tamaños de mayo de 2013, la miniaturización evoluciona día a día.

- *Grid computing* añadió la capacidad de gestión de sus aplicaciones: Cada aplicación podía constar de múltiples aplicaciones y, de algún modo, se debería facilitar su instalación, puesta en marcha, actualización y puesta fuera de servicio.
- En el nivel superior, el *grid computing* facilitaba la labor de gestión de servicios que constaban de una composición de los niveles anteriores, permitiendo su comercialización por profesionales como servicios integrados.

De forma concurrente varias iniciativas abiertas iban desarrollando aplicaciones en cada uno de esos niveles, e incluso con varias funcionalidades. Las empresas del sector también fueron evolucionando productos similares con sus estándares y sus aplicaciones previas. De este modo el mercado se encontró con una amalgama de productos orientados a facilitar el procesado masivamente en paralelo con todas las connotaciones técnicas que implica. En este punto se puede empezar a hablar de tecnologías (múltiples y variadas) de Big Data.

Figura 6: superordenador Cray XE6 del centro de cálculo de altas prestaciones de Stuttgart.

Pero las tecnologías de la información son muy dinámicas y siempre intentan evolucionar apoyándose en unos procesos de marketing a nivel mundial y acuñando nuevos términos. Así aparece el ***Cloud Computing***, nombre que se emplea para cualquier servicio cliente-servidor[8] que se oferte en internet a través del protocolo HTTP, bien

8 En este contexto el término cliente-servidor hace referencia al modelo de comunicaciones más habitual del protocolo TCP-IP, nivel de comunicaciones de transporte de internet sobre el que se ejecutan otros como el HTTP o protocolo web.

en forma de página/aplicación web, o bien mediante los llamados *web-services*. De esta forma, todas las tecnologías y productos que surgieron del incipiente mundo del Big Data (*grid*, comunicaciones entre nodos, almacenamiento, …) fueron provisionadas de una capa exterior que permitía su acceso a través de la web, es decir a través del mencionado protocolo HTTP.

Otra tecnología facilitadora del procesado masivamente paralelo fue la **virtualización de máquinas**. Los ordenadores actuales poseen tal capacidad de cálculo que son capaces de ejecutar dos o más sistemas operativos a la vez, en lo que se viene a llamar máquina virtual. De este se modo, un único ordenador se desdobla en varios, cada uno de ellos ejecutando su propio sistema operativo y su propia aplicación. La ventaja es la facilidad de instalación y la independencia de las aplicaciones (si falla una aplicación no afecta al resto del sistema). Esta propiedad es utilizada para ejecutar diferentes aplicaciones a la vez, cada una con sus datos y sus algoritmos.

El escenario actual es que existen grandes empresas como Google, Microsoft y Amazon que ofrecen sus servicios de procesado y almacenado de grandes volúmenes de datos a la carta a través de internet. Detrás de esos servicios están todas las tecnologías antes citadas y que su aplicación directa es el procesado masivamente paralelo que puede ser de aplicación en Big Data.

- Aplicaciones de visión artificial mediante grandes sistemas de procesado en paralelo.

En cuanto a las aplicaciones de visión artificial mediante Big Data en el ámbito civil se pueden mencionar varios ejemplos. Un primer ejemplo académico viene de la Universidad de Catania[9], proponiendo una infraestructura de *grid* para el procesamiento masivo de imágenes. La universidad plantea unos servicios de visión artificial soportados por una infraestructura grid. La propuesta incluye la estandarización de los servicios de visión artificial, de este modo se permite un acceso sencillo a las operaciones de visión artificial para los desarrolladores que no dispongan de un *curriculum* en *grid computing*.

Dejando atrás las herramientas y pasando a las aplicaciones, se puede mencionar la propuesta de crear una red *grid* que mediante el procesado masivo de las imágenes de un sistema de tráfico sea capaz de detectar el flujo del tráfico, la detección de vehículos, su clasificación por categorías, velocidad y detección de accidentes. Un sistema así proponen los autores[10] del artículo del libro "*Advances and Innovations in Systems, Computing*

9 A. Crisafi, D. Giordano, C. Spampinato, "GRIPLAB 1.0: Grid Image Processing Laboratory for Distributed Machine Vision Applications", Dept. of Inf. & Telecommun. Eng., Univ. of Catania, Catania. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4806916&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4806916.

10 Ivica Dimitrovski, Gorgi Kakasevski, Aneta Buckovska, Suzana Loskovska, Bozidar Proevski "Grid Enabled Computer Vision System for Measuring Traffic Parameters", 2007, pp 561-565.

Sciences and Software Engineering".

Por ser de interés en cuanto al I+D y al futuro a medio plazo, se incluye una referencia al proyecto "*Realtime Stream Processing*"[11] alojado en el sitio web "Future Grid Portal". Proyecto orientado al procesamiento de video en tiempo real mediante *grid computing* y con software de código abierto.

- ***Elementos lógicos***

Volviendo al objetivo del análisis de imágenes y centrándose, no en el análisis de una sola, sino en el flujo de imágenes o vídeo, se trata de obtener una interpretación de una información no estructurada de la realidad (el vídeo) teniendo como resultado un conjunto de variables analizables por un ordenador. Además, esas variables podrán ser referenciadas en el tiempo, obteniendo series temporales.

Para la obtención de esa información se debe tratar las imágenes en diferentes procesos o fases. Las fases clásicas del procesamiento de imágenes son un preprocesado, una segmentación (incluyendo la extracción de características) y el procesado propiamente dicho. El procesado se puede subdividir en la identificación de relaciones entre las características extraídas, las variables objetivo y el tiempo. Y la segunda subdivisión consistirá en el marcado temporal de las variables.

- Preprocesado

La primera de las fases del análisis de imágenes consiste en el pre-procesado de las mismas. Se trata de intentar mejorar la imagen capturada mejorando la iluminación, el contraste, o el realzado de figuras, etc. Estás técnicas se suelen basar en la aplicación de filtros digitales a la imagen o mediante tablas de transformación (*Look Up Tables*).

11 https://portal.futuregrid.org/node/1190/project-details

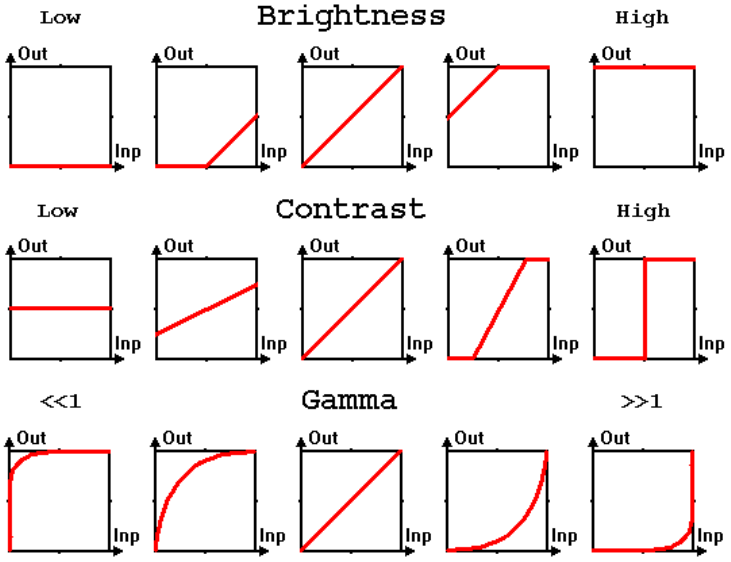


Figura 7: Ejemplo de filtros de preprocesado.

Cada caso requerirá de unas técnicas determinadas y dependerá de la iluminación con la que se hayan capturado las imágenes, así como del propio sensor. Todas estas circunstancias influyen en la elección de unas transformaciones u otras.

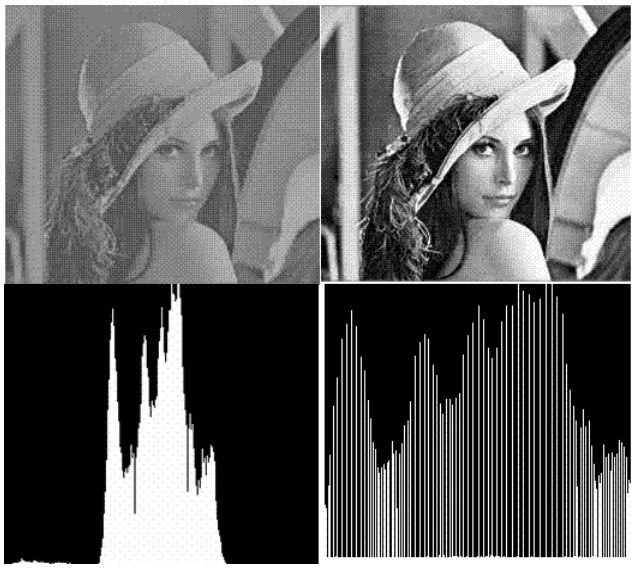


Figura 8: Ejemplo clásico de realzado de una imagen mediante la técnica de ecualización.

- Segmentación y extracción de características

La segmentación es la fase en la que trata de hallar regiones o "segmentos" de la escena que tienen un conjunto común de características. Su obtención se puede realizar mediante técnicas de estudio de la distribución del color, de niveles de intensidad, de textura, buscando zonas móviles y estacionarias de una escena. Se pueden detectar bordes en general o en una dirección en particular.

Figura 9: Ejemplos del pre-procesamiento mediante técnicas de detección de bordes.

Existen numerosas publicaciones de algoritmos de segmentación de imágenes, cada uno con un propósito específico. Las técnicas más usadas para ello pueden enumerarse en:

- detección de bordes, curvas y esquinas
- detección de gradientes de luminosidad
- identificación de texturas
- detección de gradientes de color
- matrices de movimiento (gradientes de velocidad)
- …

Con la segmentación, la imagen ha pasado de una imagen bidimensional a un conjunto de objetos a los que se les puede extraer características como la posición de su centro de gravedad, las dimensiones, su iluminación media, distancia a sus vecinos, etc. Aunque esas características pueden ser tan complejas como se necesite y puedan incluir información adicional (como la geo referencia). Al final del proceso se obtendrán los valores de las variables objetivo, que serán o bien alguna de estas características o bien una combinación de las mismas junto con información adicional (información externa a la visión artificial, p.e. posición de un GPS).

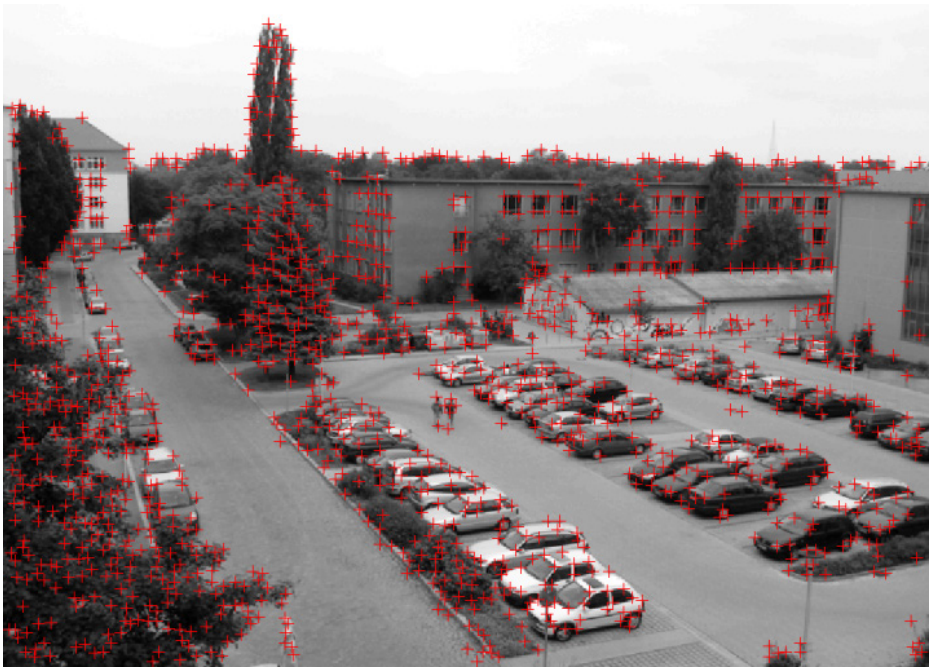


Figura 10: ejemplo del resultado de una segmentación, identificando los puntos de interés.

- Obtención de las variables objetivo

Existe un caso particular que se corresponde con el análisis de video. Como se dijo previamente, desde el punto de vista del análisis de imágenes, un video no es más que una secuencia de fotogramas. La segmentación y extracción de características se realizará en este caso por cada una de las imágenes que forman la secuencia. El problema añadido es que se debe de relacionar los objetos segmentados en un fotograma con los objetos segmentados en los fotogramas anteriores.

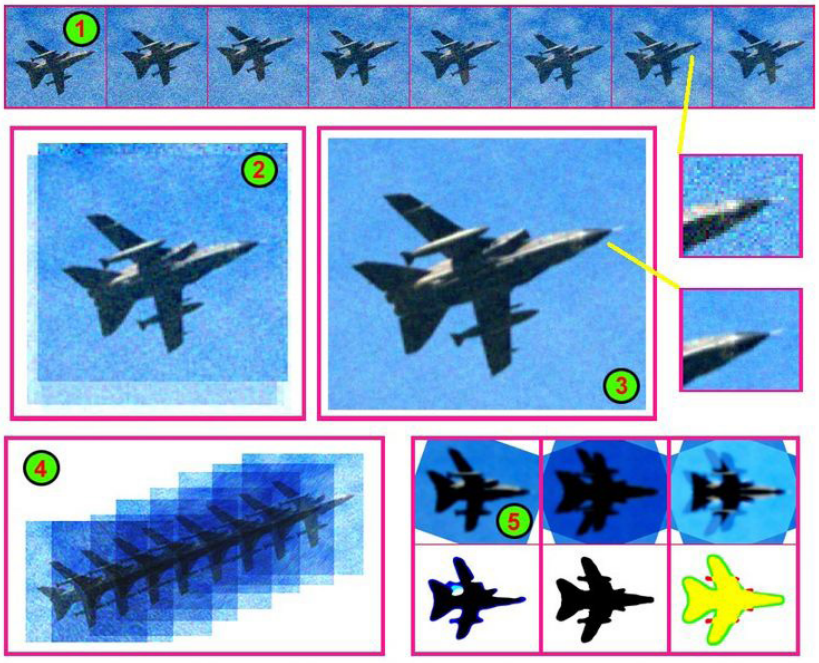


Figura 11: Segmentación y etiquetación en de una secuencia de video.

En ciertos momentos de la secuencia aparecerán nuevos objetos, mientras que en otras ocasiones habrá otros objetos que desaparezcan de la escena. Esta fase requiere el uso de tecnologías o algoritmos inteligentes (o *machine learning*) que permitan describir la escena. Suele ser habitual la utilización de técnicas basadas en modelos predictivos o en clasificadores. Las tecnologías inteligentes se usan en función del problema, pudiéndose encontrar sistemas que incluyen las tecnologías inteligentes clásicas:

- Sistemas basados en reglas con sus motores de inferencia,
- Redes neuronales con entrenamiento o con aprendizaje automático,
- Lógica borrosa,
- Algoritmos inteligentes de búsquedas de caminos óptimos o de minimización de funciones complejas,
- Etc.

Uno de los resultados de aplicar las técnicas inteligentes suele consistir en la elección de una imagen que es la que contiene la información que se necesita (identificación de un objetivo). En otras ocasiones no será una única imagen sino un subconjunto de las mismas o una sub-escena del video inicial. Y estas imágenes tendrán anotaciones que puedan servir al objetivo final.

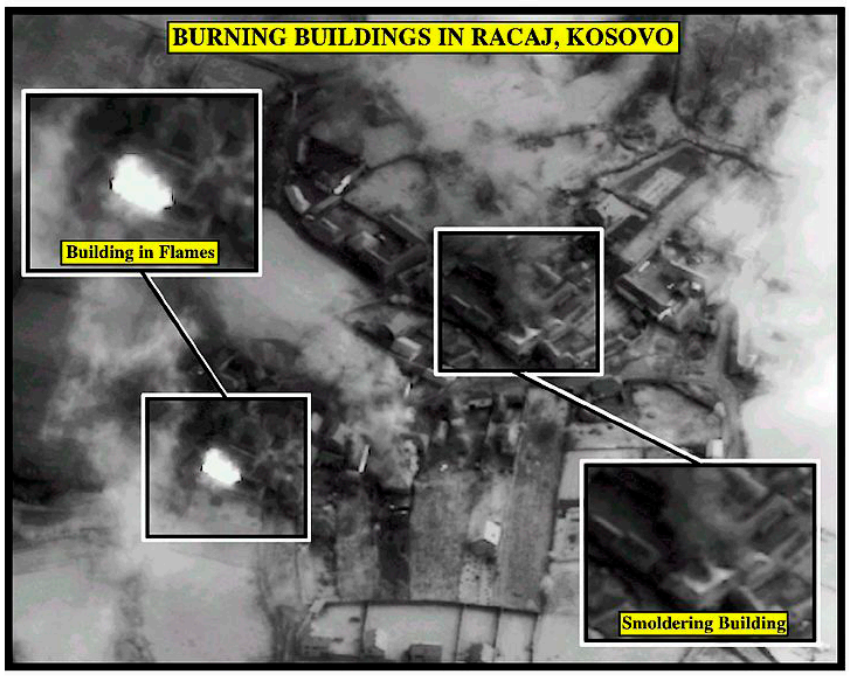


Figura 12: Imagen etiquetada proveniente de un vehículo aéreo.

- **Procesamiento de grandes volúmenes de datos**

El proceso de transformación de grandes volúmenes de datos (incluidos los datos provenientes de imágenes o de videos) en arquitecturas de alto nivel, se organiza en

etapas progresivas donde en cada etapa se agrega valor. Se trata de aprovechar la aparición de las tecnologías de Big Data y los avances computacionales y científicos en los campos de la estadística, matemáticas, investigación de operaciones, reglas de negocio y de aprendizaje, al análisis de imágenes de forma industrial (véase -)-. Es decir, las transformaciones se realizan masivamente como si las imágenes llegasen a una fábrica y fuesen procesadas en diferentes zonas mediante una cadena de transformaciones.

Pero para poder clasificar este proceso dentro de las aplicaciones de Big Data es necesario que se den condiciones adicionales. Considerarlo como Big Data solo por tener un gran volumen de video no es suficiente, es necesario que esa infraestructura esté enmarcada dentro de una arquitectura donde se pueda disponer de información adicional, proveniente de bases de datos de información previa, bases de datos no estructuradas, información de fuentes abiertas y que las conclusiones que se obtengan del procesamiento de imágenes tengan en cuenta todas estas fuentes. Simplificando, la infraestructura de Big Data debe formar parte de los sistemas de Fusión de Datos.

Hay que tener en cuenta que, como se ha venido expresando en los puntos anteriores, el procesado depende mucho de la escena, de la captura y de los objetivos que se quiera obtener. La infraestructura de Big Data sólo es una herramienta que permite flexibilizar soluciones e incorporar información adicional, aunque estas soluciones estén muy restringidas a un problema muy particular. La flexibilidad de las soluciones de Big Data permitirá cambiar la configuración de la infraestructura para particularizar soluciones.

Hay que recalcar que no existe un sistema de procesamiento de imágenes del que se obtenga cualquier información que se necesite; ni si quiera obtener información fiable al cien por cien. En la mayoría de los casos es necesaria la supervisión de un experto. Como triste ejemplo de lo que se está exponiendo, valga el caso de los atentados de 2013 en la maratón de Boston. Las autoridades tratan de estudiar las imágenes del atentado con todo su potencial tecnológico. Y a pesar de que los terroristas eran residentes legales y, por tanto, las autoridades disponían de fotos de ellos (carnet de conducir), no se logra identificarles hasta que los medios de comunicación publican las fotografías de los sospechosos. El potencial tecnológico no fue suficiente porque el reconocimiento de caras en un entorno no determinado no es un problema resuelto.

4.2.3.- Necesidades en Defensa

La necesidad de analizar datos y de prescribir acciones es un fenómeno generalizado y en el entorno de las administraciones públicas, aunque sigue siendo una tarea que se realiza de forma poco automática, más bien artesanal. Este fenómeno es bastante significativo en el entorno de Defensa. Un ejemplo claro de esta afirmación la da la consultora GovWin Networks quienes afirman que el Departamento de Defensa

estadounidense gasta el 58,4% [12]de todo el gasto federal en almacenamiento de datos, de los cuales la mayoría provienen de la necesidad de almacenar videos. Las fuentes de estos videos son:

1. UAV (o *drones*) como el Predator recogen una ingente cantidad de videos para el reconocimiento de imágenes en escenarios hostiles.
2. Imágenes y videos procedentes de los vehículos de exploración y reconocimiento terrestres, bien en el espectro visible o bien en el infrarrojo.
3. Imágenes obtenidas de satélites de vigilancia.
4. Cámaras de vigilancia en lugares públicos gestionados por las diferentes administraciones estatales y locales.
5. Cámaras de vigilancia en los entornos de lugares privados como hospitales, colegios, y empresas.Videos publicados y compartidos en las diferentes redes sociales, tales como YouTube, Facebook, Twitter, blogs u otros lugares del ciberespacio.

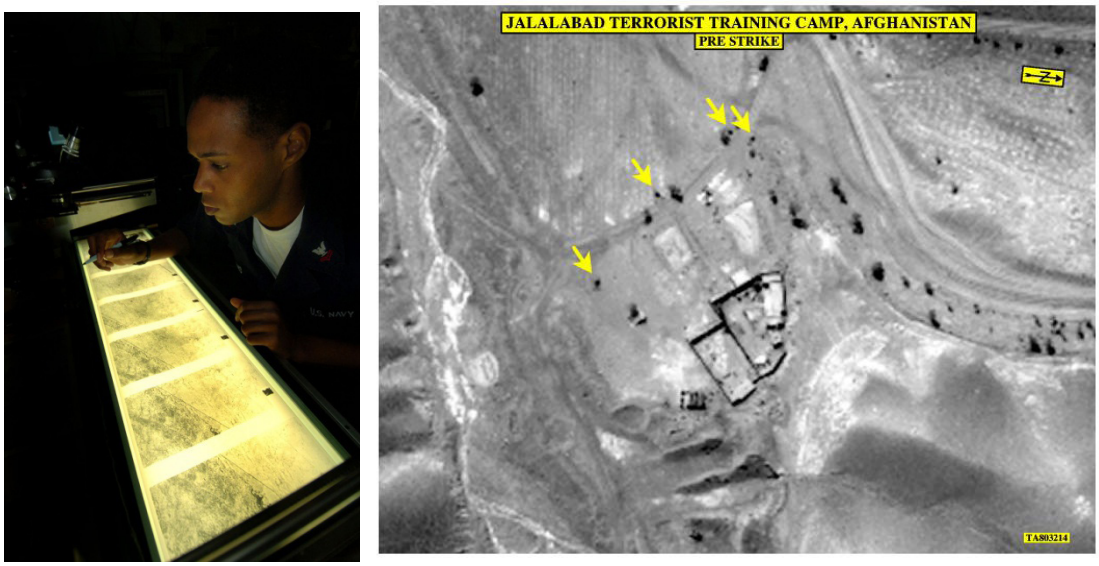


Figura 13: Ejemplo de la actividad de supervisión de imágenes y de una imagen etiquetada.

Reconociendo que las infraestructuras de Big Data todavía no se aplican masivamente en el procesamiento de imágenes, se deberían identificar posibles aplicaciones partiendo de las ya existentes. Por tanto ¿qué capacidades pueden ofrecer estos sistemas de utilidad para Defensa?:

- Detección de movimientos.
- Detección de accesos no permitidos en zonas de exclusión (en vigilancia de

12 http://govwin.com/bernaspi_blog/big-data-cloud-and-video/713070

infraestructuras críticas).

- Reconocimientos faciales (en determinados entornos).
- Seguimiento y reconocimiento de objetivos de imágenes (bien de UAV o de otras plataformas y fuentes).
- Reconocimiento de comportamientos sospechosos en lugares públicos.
- Cálculos de altura de edificios en imágenes aéreas para zonas urbanas de conflicto.
- Identificación de actividades económicas o de cosechas en zonas en conflicto.
- Identificación de objetos abandonados sospechosos de ser IED.
- Como se puede apreciar estas capacidades tienen gran potencialidad en el mundo militar, y de hecho, ya se hace uso de muchas de ellas en los diferentes sistemas existentes y desplegados. Las aplicaciones donde se pueden encontrar se pueden agrupar en las siguientes:
- Herramientas de Inteligencia, reconocimiento y seguimiento de objetivos (ISTAR).
- Herramientas de visualización del estado operacional del campo de batalla.
- Herramientas de Ayuda a la toma de decisiones.

Aunque las dos últimas pueden ser una consecuencia de la primera de las aplicaciones, con arquitecturas de Big Data las tareas se facilitan pudiéndose llegar a la integración de esas herramientas.

- **I+D**

Como se ha descrito, Defensa demanda ciertas capacidades que se podrían adquirir mediante la combinación de tecnologías de procesamiento de imágenes e infraestructura de Big Data. Por una parte, varios fabricantes ofrecen productos de Big Data, y por otra las empresas de ingeniería ofrecen sistemas de procesamiento de imágenes cada vez más potentes e inteligentes. La convergencia de ambos mundos para proporcionar la capacidad demandada no es sencilla y el camino a recorrer pasa por programas de I+D que minimicen los riesgos.

La infraestructura necesaria para desarrollar programas de I+D en este tema implica que, hoy día, el I+D esté liderado por organizaciones internacionales o por EE.UU. En el ámbito de defensa esta afirmación se traduce en que los que lideran la investigación sean la *Science and Technology Organization* (STO) de la OTAN y la agencia DARPA del Departamento de Defensa de EE.UU.

- STO

La *Science & Technology Organization* constituye la organización sobre la que pivota

el apoyo en investigación que la OTAN necesita para el cumplimiento de sus objetivos. Desde el punto de vista de la I+D, la STO se organiza entorno a paneles de áreas temáticas donde se realizan actividades de investigación o de difusión. Los países que forman parte de la STO tienen el deber de aportar sus recursos para definir, conducir y promover la cooperación en materia de investigación e intercambiando información.

Tan lejano como en el año 2000, la STO organizó un simposio en Canadá con el nombre IST-020/RWS-002 "*Multimedia Visualization of Massive Military Datasets*"[13], que aunque se centró en la visualización, tuvo presentaciones sobre vigilancia, fusión de grandes cantidades de datos, y técnicas matemáticas de procesamiento.

Figura 14: Logotipo de la *Science and technology Organization* de la OTAN.

Posteriormente se han seguido realizando esfuerzos en esta dirección, de esta forma, la STO ha puesto en marcha un taller y un simposio para octubre de 2013 denominados IST-116 "*Visual Analytics*", y IST-117 "*Visualization for Analysis*", cuyos objetivos consisten en "discutir métodos a través de los que se puede dar soporte a la decisión mediante análisis de visualización en diferentes problemáticas, y explorar las posibilidades o encontrar necesidades de utilización para los masivos datos heterogéneos que se disponen". Estas actividades son la consecuencia de un grupo de trabajo que está en marcha durante el presente 2013.

Adicionalmente, la STO a través de su panel IST (*Information, Systems Technologies*) ha puesto en marcha varios grupos de trabajo entorno a la temática de la fusión de datos. El tema principal no es el procesamiento de las imágenes pero dentro de las fuentes heterogéneas que analizan estos estudios se encuentran las grandes bases de datos de imágenes y videos disponibles en OTAN.

- *DARPA*

La *Defense Advanced Research Projects Agency* (DARPA), es una agencia gubernamental estadounidense que tiene el objetivo de prevenir los impactos estratégicos de los posibles enemigos de EE.UU. mediante el desarrollo de actividades que permitan mantener la superioridad tecnológica militar.

13 http://ftp.rta.nato.int/public//PubFullText/RTO/MP/RTO-MP-050///MP-050-$$ALL.pdf

En abril de 2012 el gobierno de Obama puso en marcha una iniciativa nacional orientada a mejorar las herramientas y técnicas necesarias para acceder, organizar y descubrir información útil de los grandes volúmenes de datos de los que disponen su Administración. Parte de los recursos financieros de este programa se han destinado a Defensa, hecho que ha implicado que DARPA ponga en marcha varios programas de investigación en esta dirección. Entre ellos, hay dos programas relacionados con el procesamiento masivo de imágenes, el *Mind's Eye* y el programa *Insight*.

- *Mind's Eye*

El objetivo de este programa[14] consiste en dotar a un sistema autónomo terrestre (UGV) de una cámara inteligente que sea capaz de describir la escena que está visualizando. La descripción se realizará mediante la etiquetación de la escena con los verbos que la pueden describir. Adicionalmente, el sistema será capaz de aprender nuevos conceptos directamente de su experiencia visual.

Las imágenes que se obtengan desde el UGV serán procesadas mediante una infraestructura Big Data donde se procesen y obtengan las variables necesarias para definir la escena, pero además contará con la información adicional que se disponga desde otras fuentes de la inteligencia militar.

El resultado de este proyecto impactará directamente a la forma en que se pueda obtener el estado operacional de un escenario sin necesidad de enviar combatientes directamente, con la consiguiente minimización de riesgos.

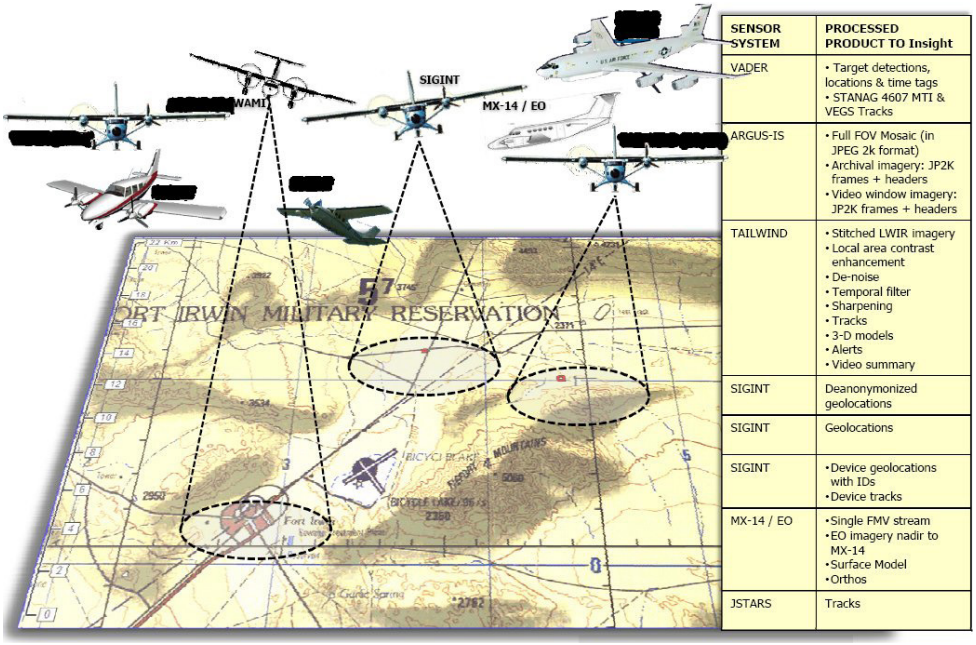
- *Insight*

El programa Insight[15] trata de mitigar los costes de supervisar las oleadas de datos que pueden llegar a los combatientes en un momento dado y con el objetivo de que pueda conseguirse en tiempo casi-real. Para ello el programa está trabajando en el futuro sistema de inteligencia, reconocimiento y seguimiento (ISR) en el que aúna diversas tecnologías incluyendo las relacionadas con Big Data y con el procesamiento de imágenes provenientes de los sensores ISR desplegados en el campo de batalla (sensores complejos, texto, imágenes, video, …)

Las primeras fases de este programa demostraron que era posible (según DARPA) producir información fiable única (no duplicada) proveniente de los datos ISR que disponían los investigadores que realizaron el estudio. El conjunto con el que contaron consistía en 135 Tera bytes procedentes de 240 usuarios de la Administración, de la industria o del entorno académico.

14 http://www.darpa.mil/Our_Work/I2O/Programs/Minds_Eye.aspx

15 http://www.darpa.mil/Our_Work/I2O/Programs/Insight.aspx



SENSOR SYSTEM	PROCESSED PRODUCT TO Insight
VADER	• Target detections, locations & time tags • STANAG 4607 MTI & VEGS Tracks
ARGUS-IS	• Full FOV Mosaic (in JPEG 2k format) • Archival imagery: JP2K frames + headers • Video window imagery: JP2K frames + headers
TAILWIND	• Stitched LWIR imagery • Local area contrast enhancement • De-noise • Temporal filter • Sharpening • Tracks • 3-D models • Alerts • Video summary
SIGINT	Deanonymonized geolocations
SIGINT	Geolocations
SIGINT	•Device geolocations with IDs •Device tracks
MX-14 / EO	•Single FMV stream •EO imagery nadir to MX-14 •Surface Model •Orthos
JSTARS	Tracks

Figura 15: Representación de los medios aéreos de los que dispondrá el programa Insight.

4.2.4.- Conclusiones

Las técnicas de procesamiento de imágenes son muy demandadas en aplicaciones militares, aunque las soluciones actuales se centran en entornos muy controlados que faciliten la obtención de información útil. La evolución de las mismas se orienta a que estos sistemas sean más efectivos y sean menos dependientes de la iluminación, de la propia escena, de la captura y de tantos otros factores externos.

Las tecnologías y soluciones Big Data se presentan como una herramienta de ayuda que permite mejorar los sistemas de procesamiento de imágenes en varios aspectos:

- Flexibilizar las infraestructuras para poder poner en marcha soluciones a medida fácilmente
- Facilitar el procesado masivo en paralelo dando potencia de cálculo a los sistemas
- Proveer de información externa adicional con la que poder fusionar información, mejorando la calidad de la misma (información más fiable).

No obstante, las infraestructuras necesarias y los grandes volúmenes de datos necesarios como fuente implican que sólo grandes organizaciones y/o naciones podrán acceder a las mejoras que estas soluciones pueden proporcionar. A medida que los costes vayan reduciéndose y/o la generación de nuevos contenidos vaya aumentando, estas soluciones podrán ir ganando terreno a los sistemas individuales que no puedan ser considerados como Big Data.

Por último concluir que las principales soluciones a las que se pueden aplicar estas tecnologías dentro del ámbito de Defensa se circunscriben a la inteligencia militar, o a los sistemas ISTAR.

4.2.5 Listado de figuras

Todas las imágenes excepto la 14 y la 15 (orígenes en la STO y DARPA) provienen de Wikimedia: http://commons.wikimedia.org

Figura 1: Imagen de un perro tomada con una cámara de infrarrojos. 88

Figura 2: soldado con visor y 89

Figura 3: cámara infrarrojos en helicóptero 89

Figura 4: tarjeta de adquisición de imágenes y/o video. 90

Figura 5: tarjeta (clon de Arduino) con controladora de video acoplada. 92

Figura 6: superordenador Cray XE6 del centro de cálculo de altas prestaciones de Stuttgart. 93

Figura 7: Ejemplo de filtros de preprocesado. 96

Figura 8: Ejemplo clásico de realzado de una imagen mediante la técnica de ecualización. 96

Figura 9: Ejemplos del pre-procesamiento mediante técnicas de detección de bordes. 97

Figura 10: ejemplo del resultado de una segmentación, identificando los puntos de interés. 98

Figura 11: Segmentación y etiquetación en de una secuencia de video. 98

Figura 12: Imagen etiquetada proveniente de un vehículo aéreo. 99

Figura 13: Ejempo de la actividad de supervision de imágenes y de una imagen etiquetada. 101

Figura 14: Logotipo de la Science and technology Organization de la OTAN. 103

Figura 15: Representacion de los medios aéreos de los que dispondrá el programa Insight. 105

4.3.- Guerra electrónica (EW)

4.3.1.- Una aproximación a la Guerra Electrónica

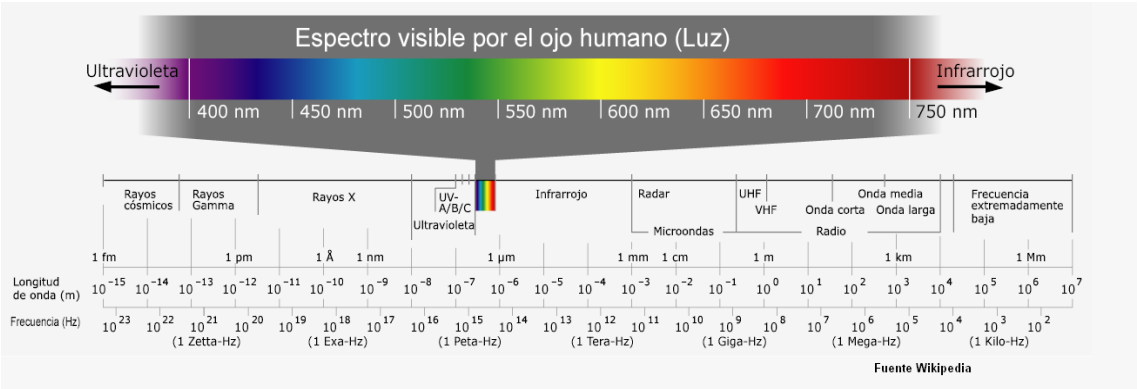
Como primera medida para entender en qué se basa y cómo actúan las Unidades y equipos de guerra electrónica es necesario introducir, desde una perspectiva eminentemente doctrinal, una serie de definiciones, tareas, etc. que se manejan en este campo. Así se puede enmarcar la aplicación de las técnicas asociadas a Big Data. Información más extensa puede consultarse la monografía del Sistema de Observación y Prospectiva

Tecnológica (SOPT), que se puede descargar de http://www.tecnologiaeinnovacion.defensa.gob.es o bien adquirir en el servicio de publicaciones del Ministerio de Defensa.

4.3.2.- Definición de Guerra Electrónica (EW)

Es el conjunto de actividades que tienen como objetivo reducir o impedir el empleo del espectro electromagnético por parte del adversario y a la vez conservar y permitir la utilización de dicho espectro en beneficio propio.

Esto es que el "campo de batalla" donde se librarán las acciones de esta guerra se concreta en el espectro electromagnético con las posibilidades que ofrece el empleo de la radiación electromagnética en caso de conflicto o crisis. El factor determinante de este posible aprovechamiento es la propagación radioeléctrica. Hoy en día es de constante aplicación por el gran desarrollo que ha tenido tanto la telefonía celular, "telefonía móvil" como habitualmente se conoce, y la redes WIFI. En ambos casos no sorprende en absoluto hablar de si "hay cobertura o no". Esta situación se nos presenta cada día y en ella subyacen estos fenómenos de propagación radioeléctrica. Ahora bien, si además de emplear estos medios y fenómenos de modo directo tal y como se ha mencionado o como puede ser el empleo del radar; se pretende extraer información de las emisiones ajenas, nos situamos en una posición diferente que a continuación se explicará de modo somero. A priori puede parecer una idea extravagante o de escaso beneficio sin embargo veamos un par de ejemplos, no precisamente modernos, para reconocer el valor de estas acciones.



4.3.3.- Un par de ejemplos ilustrativos

Acorazado Bismark: Durante la Segunda guerra mundial tras una persecución por el Atlántico este acorazado rompe el contacto con sus perseguidores, sin saberlo, emite dos mensajes informando al Alto Mando alemán. Estas emisiones son captadas por los ingleses y derivan en volver a encontrarlo, reanudan la persecución y acaban con su hundimiento.

Operación de supresión de defensa aérea: En 1982 el valle de la Bekaa está defendido

por diferentes baterías sirias, Israel provoca su activación con aviones no tripulados y graba las características técnicas de las emisiones que emplean las defensas a lo largo de varias semanas. El primero de junio lanza un simulacro amplio de ataque con aviones no tripulados, estos provocan la reacción de las defensas sirias obligándoles a emplear sus misiles y activando sus contramedidas. Con esta información los israelíes programan sus equipos de ataque electrónico y en el ataque real consiguen la superioridad aérea mediante la anulación de las medidas de las baterías SAM sirias. Además impiden la comunicación de los aviones de ataque sirio con sus bases por lo que estos actúan "a ciegas". En tres días el balance es de más de 80 aviones sirios derribados por ninguno de Israel y anuladas las más de veinte baterías de tierra.

4.3.4.- Acciones de EW

Las actividades de EW son de diferente naturaleza y se dividen en:

- ESM (electronic support measures) Medidas de apoyo electrónico. Son aquellas acciones basadas en el empleo directo de las emisiones propias o no, como puede ser el radar o descubrir la posición de un emisor por captar sus señales desde varios puntos diferentes.
- ECM (electronic countermeasures) Contramedidas electrónicas o EA (electronic attack) Ataque electrónico. Son acciones que procuran impedir, negar o engañar al adversario mediante el empleo de emisiones propias o las del mismo adversario.
- EPM (electronic protective measures) Medidas de protección electrónica. Se trata de acciones que tratan de emplear los sistemas y emisiones propios a pesar del empleo por parte del adversario de sus acciones ECM. Estas medidas suelen ser de dos tipos: puramente técnicas, que van embebidas en los sistemas o procedimentales, que se basan en la forma de empleo de los sistemas y equipos.

4.3.5.- Objetivos de las acciones

Los sistemas emplean las emisiones radioeléctricas de forma distinta según la finalidad buscada. Por lo que las acciones asociadas entonces también se denominan de modo diferente:

- COMINT (communications intelligence): son las acciones sobre emisiones cuyo objeto son las comunicaciones (redes radio de voz, datos, vídeo) la característica fundamental de estas es que transportan información.
- ELINT (electronic intelligence): son acciones sobre emisiones de no- comunicaciones (radar, ayudas a la navegación, etc.) estas no suelen transportar información sino que esta se extrae del comportamiento de la emisión, situación del emisor, dirección en la que emite, código de identificación, etc.

Se diferencian fundamentalmente en la mayor complejidad que suelen tener las modulaciones de emisiones de comunicaciones además de que suelen emplear bandas de frecuencias distintas. Sin embargo, y fundamentalmente por abreviar, ambos tipos de acciones se reúnen en un término, SIGINT: inteligencia de señales, cuando no se pretende distinguir entre unas y otras, o sea irrelevante tal distinción.

El desarrollo de cualquiera de cualquier acción tiene dos partes, por un lado la acción propiamente dicha de la incidencia que crea o resuelve a nivel técnico y por otro el marco global en el que se ha diseñado. Esto último puede estar dirigido a muy distintos fines: obtención de información y generación de inteligencia o bien un fin muy determinado como la autoprotección, que trata de evitar un daño en la plataforma en la que se lleva instalado el sistema. Ejemplos de estas últimas serían las medidas y alertas de un alertador radar (Radar Warning Receiver – RWR) que se complementan con lanzadores de chaff, bengalas, etc. Algunas de estas acciones se complementan unas a otras de modo que la función final es la combinación de varias de ellas, como el ejemplo mencionado.

Las acciones en las cuales se maneja información como objeto son las asociadas a las medidas de apoyo electrónico (ESM). En estas se puede distinguir entre la información que de modo nativo puede transportar o la meta información asociada al punto de emisión, momento de la misma, parámetros técnicos que la componen: frecuencia, ancho de banda modulación y parámetros de la misma, etc. Todas esas medidas pueden ser útiles para extraer conclusiones.

4.3.6.- Tipos de acciones ESM

Las acciones anteriores se concretan en:

- Búsqueda: sintonización de cada frecuencia en la que no solo existe ruido.
- Interceptación: mantenimiento de la frecuencia que emite para conocer sus parámetros
- Identificación: cotejo de dichos parámetros con emisores esperados o conocidos.
- Localización: sincronización del momento de emisión con otros receptores situados en diferentes lugares para localizar el emisor (Angle of arrival - AOA).

Estas tareas se describen desde una perspectiva individual de cada receptor y los métodos de trabajo tradicionales deben adecuarse a sistemas con capacidad de computación aboliendo métodos manuales e individuales. Para aplicar las técnicas de Big Data el enfoque debe estar en el conjunto del sistema más que en cada receptor, aunque esto no debe olvidarse.

La organización habitual de un sistema que trate de obtener estas medidas consiste en desplegar un conjunto de receptores en el área de interés. A su vez se divide el espectro en bandas que comprendan las emisiones objetivos, cada ubicación debe tener receptores / sensores de las emisiones que pretenda recibir. El hecho de dividir

el espectro en bandas con objeto de "atender" cada posible emisor y desplegar los receptores para cubrir zonas geográficas diferentes implica la entrada en juego del fenómeno de la propagación ya mencionado. Este medio se comporta diferentemente según la frecuencia considerada, aunque en caso de hablar de bandas de frecuencias el comportamiento es muy similar en cada banda, como sucede con todo fenómeno físico. En los puntos de recepción las condiciones también varían según la dirección de llegada, las alturas de las antenas, la composición del suelo, etc. Esto quiere decir que el conjunto de factores a tener en cuenta en el momento de recibir una señal, se multiplica y que estos van a depender de elementos con unos componentes aleatorios.

Una vez que cada receptor ha tomado cada "muestra" de señal individualmente es necesario reunir y correlacionar las características de las señales para poder analizar su situación, tráfico o actividad etc. Esta comparación o correlación se realiza en sus aspectos técnicos, de espacio y tiempo. Como resultado se obtiene la situación geográfica, sus parámetros técnico la actividad de conjunto, su posible relación, etc. De todo ello se pueden obtener conclusiones acerca de la función que desarrolla el emisor o red de emisores, los eventos o situaciones que predicen, etc.

El enfoque descrito implica analizar en cada receptor todas las señales captadas para identificarlas, posteriormente se centralizan los parámetros de cada análisis y se cruzan entre los emisores y bandas correspondientes. Este modo de actuar es el habitual y proporciona información sobre estas emisiones. Tiene el gran inconveniente del propio planteamiento al hablar de "señales captadas" porque ¿qué ocurre con las que no se captan? Actualmente es conocida la posibilidad de emplear métodos de modulación o codificación que su densidad espectral puede estar por debajo del nivel del ruido. Lógicamente esas señales no "se captan", luego no se analizan, luego la información que pueden proporcionar pasa inadvertida. Si además se elucubra sobre quién va a emplear estos métodos y qué tipo de información puede manejar… Se puede entrever que tal información que no se aprovecha puede ser muy interesante. Aquí pueden entrar a desempeñar un papel relevante las técnicas de tratamiento de información masiva e intentar un análisis con un enfoque diferente.

4.3.7.- ¿Qué podría aportar Big Data?

- **Una primera idea**

El método descrito hasta aquí en el tratamiento de las señales no es el adecuado, no proporciona un volumen de información que pudiéramos considerar "grande" como para aplicarle las herramientas de Big Data. Este volumen lo va a proporcionar una asunción inicial "en cada banda considerada existe ruido sumado a una señal que se pretende analizar". Por tanto de lo que se trata es de reproducir centralizadamente las "pseudoseñales" procedentes de las áreas de interés (correlación espacial) y en los momentos que se van correspondiendo (correlación temporal). Por tanto las tareas a realizar en este sistema serían:

1. Traslado del "ruido y señal" que existe en las bandas del espectro que son de interés. Dependiendo de la capacidad disponible de transmisión y de la resolución de ancho de banda o velocidad de datos la digitalización que se necesita en cada ubicación de un receptor variará. No es preciso realizar el almacenamiento de ese volumen de información de modo local, hay que ser consciente de que la inmensa mayoría de esa digitalización es solo ruido.

2. Centralización de los productos digitalizados de varias ubicaciones, estos deben estar suficientemente identificados en tiempo, punto de recepción y ángulo de llegada. Para:

a) Comprobar que responde a la zona apropiada (correlación espacial)

a) Y en los mismos momentos (correlación temporal)

3. Correlación de las "pseudoseñales" que responden a la zona de interés (correlación espectral).

Este análisis revelaría la existencia de señales que hubieran pasado inadvertidas por baja potencia o baja densidad espectral permitiendo el análisis en tiempo real del tipo de emisión y su contenido. A partir de aquí el tratamiento más tradicional en EW y es en ese momento cuando existe información para ser almacenada.

- **Una segunda opción**

Actualmente el método más utilizado para el intercambio de información es la voz y según estudios de futuro de operadores de telecomunicaciones de ámbito global seguirá siendo así en las próximas décadas. Esta situación se da tanto en redes fijas, móviles, celulares, etc. En ellas la utilización de la voz es una constante hoy en día y, a pesar de la existencia de múltiples formas de realizar este intercambio, la voz y el vídeo son los contenidos principales. Igualmente la situación en las redes radio es la misma y además es habitual que exista la barrera del idioma. Así las emisiones que correspondan a voz pueden ser sometidas a un tratamiento de análisis semántico, traducción automática, transcripción, indexación, etc.

Solo hay que recordar algún que otro aplicativo de no hace mucho tiempo que era capaz de ir transcribiendo lo que recibía por el micrófono, IBM Via Voice por ejemplo. Si a esto se le suma la traducción de esta transcripción y posteriormente una síntesis de voz con las adaptaciones necesarias de idiomas, acentos, etc. No encontraríamos con un "sistema" capaz de traducir y dejar por escrito las conversaciones que haya captado. Quizá el acierto o aprovechamiento de algo así será, probablemente, muy bajo; pero hay que compararlo con el hecho de que esas conversaciones se desperdician, si se llegan a grabar no siempre se dispone de personas que las puedan revisar. Es decir, que de no tener nada, se puede llegar a tener… y una vez que el sistema evolucionara el aprovechamiento sería mucho mayor.

5.- Análisis

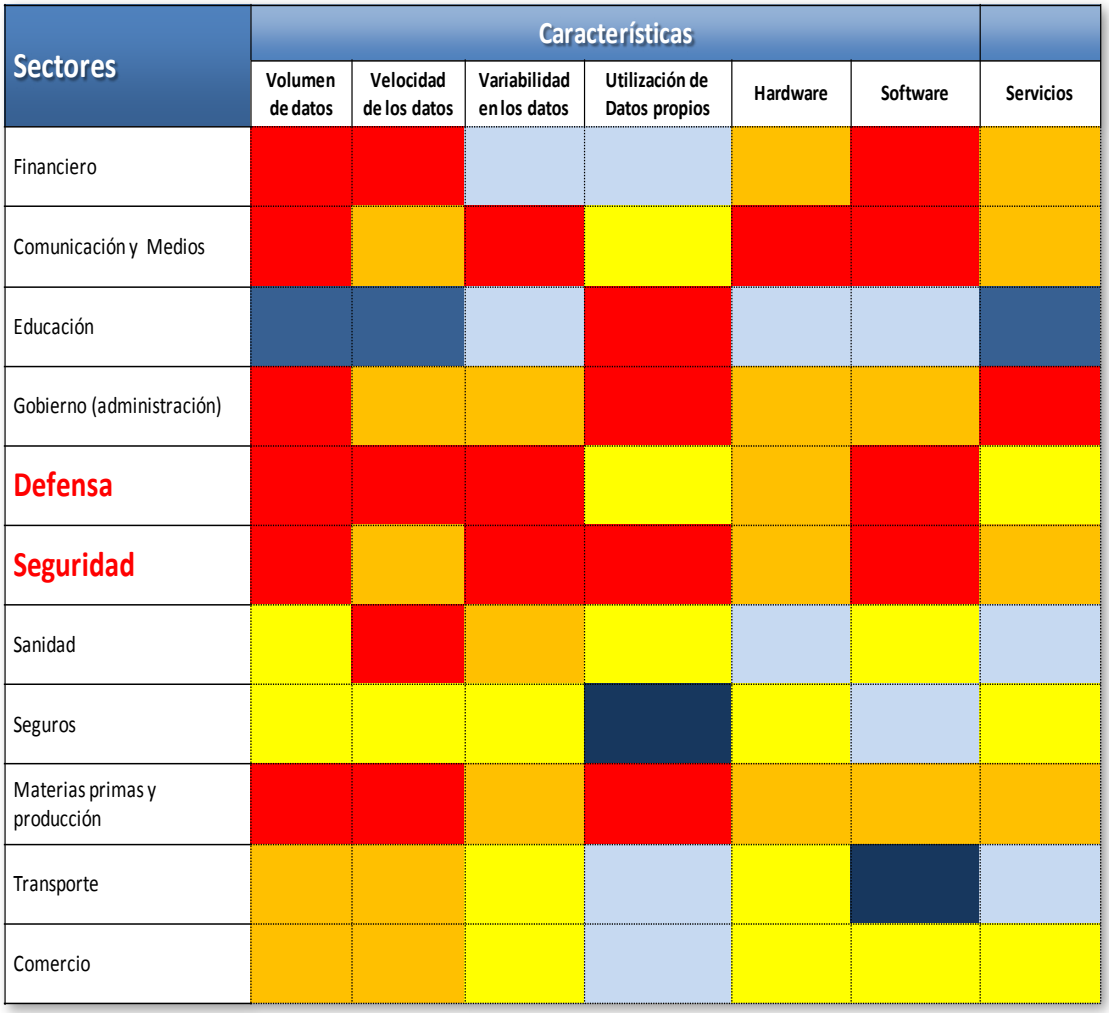
A partir de la información expuesta en los apartados anteriores y haciendo uso de otras fuentes de reconocido prestigio se han elaborado un conjunto de análisis, recogidos en tablas e infografías para facilitar su lectura, que pretenden guiar al lector respecto a la aplicación de Big Data y su posible evolución en el futuro.

5.1.- Aplicación de Big Data

En julio de 2012, Gartner publicó un análisis que reflejaba la potencial aplicabilidad de diferentes sectores de la actividad económica-social con respecto a Big Data. Entre estos sectores, financiero, educación, sanidad, producción, etc., no se especificaba ni el sector de seguridad ni el de defensa, aunque podría argumentarse que estaban incluidos en el apartado designado para el sector gubernamental. En cualquier caso, en el ámbito de este estudio hemos querido hacer uso de este análisis y añadir apartados específicos para los sectores de seguridad y defensa, conservando el de gobierno, si bien entiendo este ya desde una perspectiva puramente administrativa. Esta ampliación al análisis de Gartner nos permite comparar la aplicabilidad de Big Data con respecto a otros sectores. Para ello, se contrasta cada sector desde las perspectivas de:

- Volumen de datos manejado en ese sector.
- Velocidad de los datos, es decir, cómo de rápido se generan los datos y tiempo de vida útil de esos datos.
- Variabilidad en los datos, indica si los datos presentan un carácter homogéneo en su estructura (p.ej. bases de datos estructurales) o no (textos no estructurados, imágenes, videos, etc.)
- Utilización de datos propios. Esta perspectiva analiza el nivel de conocimiento, aprovechamiento y explotación de la información propia que la organización posee distribuida entre los diferentes elementos propios de la estructura (ordenadores, bases de datos, servidores de almacenamiento, etc.).
- Hardware, indica el nivel de demanda de elementos hardware, asociados directamente o indirectamente con Big Data, para la ejecución de sus funciones.
- Software, indica el nivel de demanda de elementos software, asociados directamente o indirectamente con Big Data, para la ejecución de sus funciones
- Servicios, indica el nivel de demanda de servicios, asociados directamente o indirectamente con Big Data, para la ejecución de sus funciones.

Cada una de estas perspectivas se ha clasificado desde muy alto a muy bajo, para cada uno de los sectores. Todo ello se refleja en la tabla siguiente:



Característica dominante o potencial aplicabilidad en el caso de servicios, software, hardware.

- Muy alta
- Alta
- Media
- Baja
- Muy baja

Como se puede apreciar en la tabla, el sector de seguridad y defensa, parecen a priori, dos sectores donde se dan las condiciones iniciales de Big Data, volumen, velocidad, variedad, y además son consumidores tradicionales de soluciones hardware, software y servicios.

A partir de la información identificada en cada una de las aplicaciones específicas estudiadas en los apartados anteriores, es posible recopilar un conjunto de áreas de aplicación en el ámbito de seguridad y defensa. Tal y como se presentan a continuación:

Aplicaciones Big Data en Defensa y Seguridad
Detección de intrusión física en grandes espacios o infraestructuras abiertas
Computación sobre información encriptada
Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos)
Criminología computacional
Uso fraudulento de recursos corporativos y/o sensibles
Análisis de video en tiempo real / Búsqueda y recuperación rápida en librerías de video.
Inteligencia visual en máquinas
Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos.
Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos.
Consciencia situacional
Traducción automática a gran escala (en número de idiomas y en volumen)
Predicción de eventos

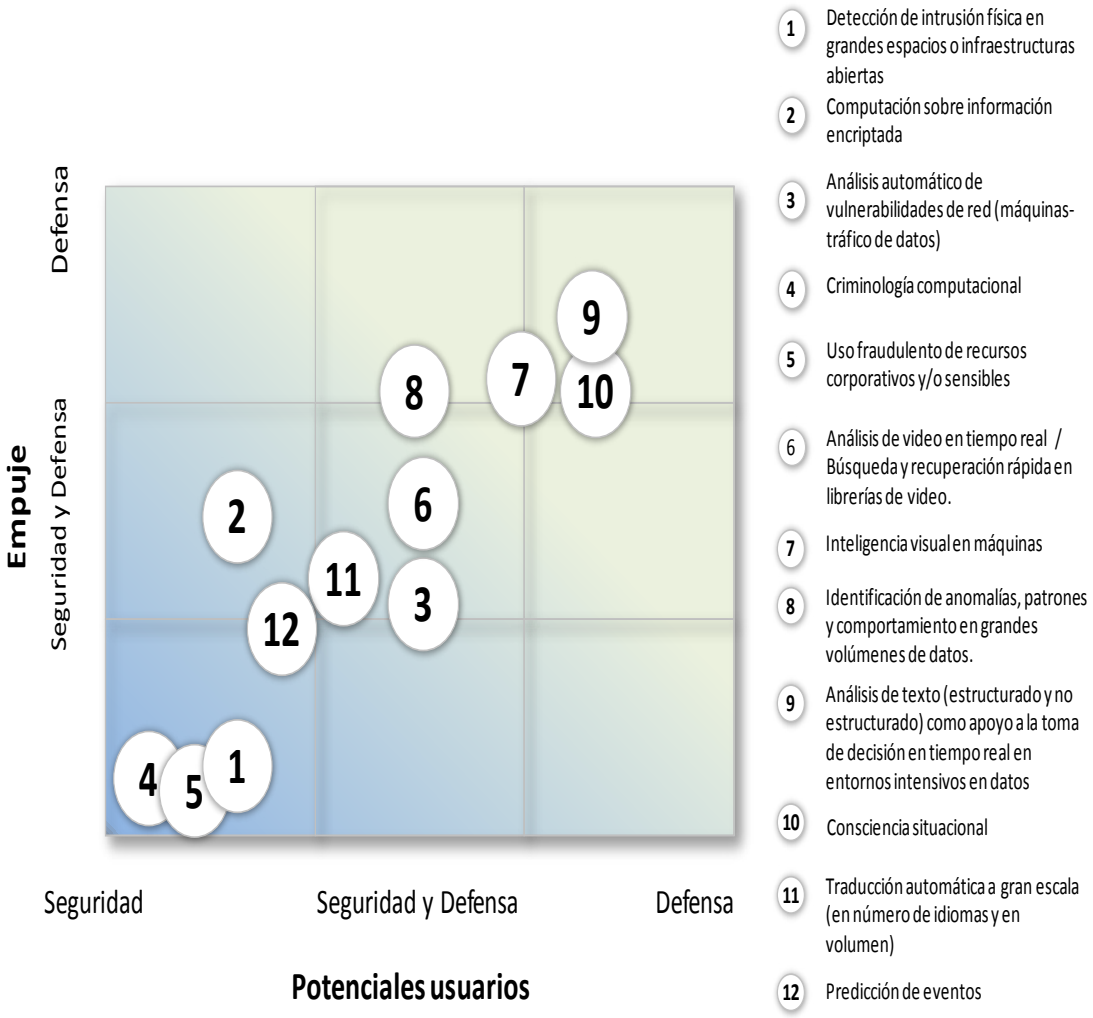
- Vigilancia y Seguridad perimetral.
- Vigilancia y Seguridad de fronteras
- Seguridad física de infraestructuras críticas.
- Comunicaciones y redes seguras
- Bancos de datos para los ámbitos financiero, seguridad interior, inteligencia, defensa.
- Protección (redes IT) de Infraestructuras críticas
- Ciberdefensa / Ciberseguridad
- Lucha contraterrorista y contra crimen organizado
- Lucha contra el fraude
- Control y seguridad de recursos informáticos y datos en organizaciones
- Gestión del conocimiento en grandes organizaciones
- Seguridad ciudadana
- Inteligencia militar
- Planeamiento táctico de misiones.
- Toma de decisión en tiempo real para operaciones (Defensa/seguridad).
- Inteligencia industrial
- En ámbito militar en HUMINT/operaciones en entornos urbanos.
- Preparación de seguridad de eventos singulares (deportivos, políticos, etc.)
- Control y comportamientos de multitudes
- ...

Por otro lado, y desde un punto de vista más técnico, cada una de estas aplicaciones específicas en seguridad y defensa, van a demandar un perfil diferente para su implementación. Así, por ejemplo algunas serán más intensivas en tecnologías y técnicas asociadas a la adquisición de datos, y otras al análisis. En este sentido, y a partir de la clasificación identificada en las fichas de cada una de estas aplicaciones, se ha elaborado la siguiente tabla, que permite obtener una imagen sobre qué funciones se debe hacer más esfuerzo según qué tipo de aplicaciones o en general según seguridad/defensa. Para el caso de defensa, se deduce que de manera general las tecnologías y técnicas más relevantes serán aquellas que soporten las funciones de adquisición, búsqueda, análisis y visualización de datos. Mientras que para el caso de seguridad, el foco se centra más en las funciones de almacenamiento, búsqueda y análisis. Lógicamente, con un mayor número de aplicaciones específicas o con otras diferentes a las en este estudio identificadas, esta distribución de relevancia en las funciones puede variar.

Aplicaciones Big Data en Defensa y Seguridad	Funciones predominantes						
	Adquisición	Almacenamiento	Búsqueda de datos	Seguridad de datos	Análisis	Visualización	Gestión
Detección de intrusión física en grandes espacios o infraestructuras abiertas	■				■		
Computación sobre información encriptada		■	■	■			
Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos)				■	■		■
Criminología computacional	■	■	■		■		
Uso fraudulento de recursos corporativos y/o sensibles				■	■		■
Análisis de video en tiempo real / Búsqueda y recuperación rápida en librerías de video.		■	■		■		
Inteligencia visual en máquinas	■	■	■		■		
Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos.	■				■	■	
Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos.	■		■		■	■	
Consciencia situacional	■		■		■	■	
Traducción automática a gran escala (en número de idiomas y en volumen)	■		■		■	■	
Predicción de eventos			■		■	■	

Para finalizar el análisis de la aplicación de Big Data, se ha pretendido reflejar el carácter predominante (seguridad/defensa) de cada una de las aplicaciones específicas identificadas. Así como qué sector (seguridad/defensa/ambos) actúa como promotor y financiador principal en cada aplicación, y qué sector (seguridad/defensa/ambos) es potencialmente el usuario principal de esa aplicación. Los resultados, mostrados en la figura siguiente, identifican oportunidades de aprovechamiento de los esfuerzos del

otro sector (como por ejemplo en las aplicaciones identificadas como 2, 3, 8, 11, y 7). No se identifican anomalías, como ocurre en otras áreas tecnológicas, donde un sector presente alto interés y potencial uso de los resultados del otro sector, sin tener un nivel de esfuerzo comparable, como sería el caso de los cuadrantes superior izquierdo, e inferior derecho.



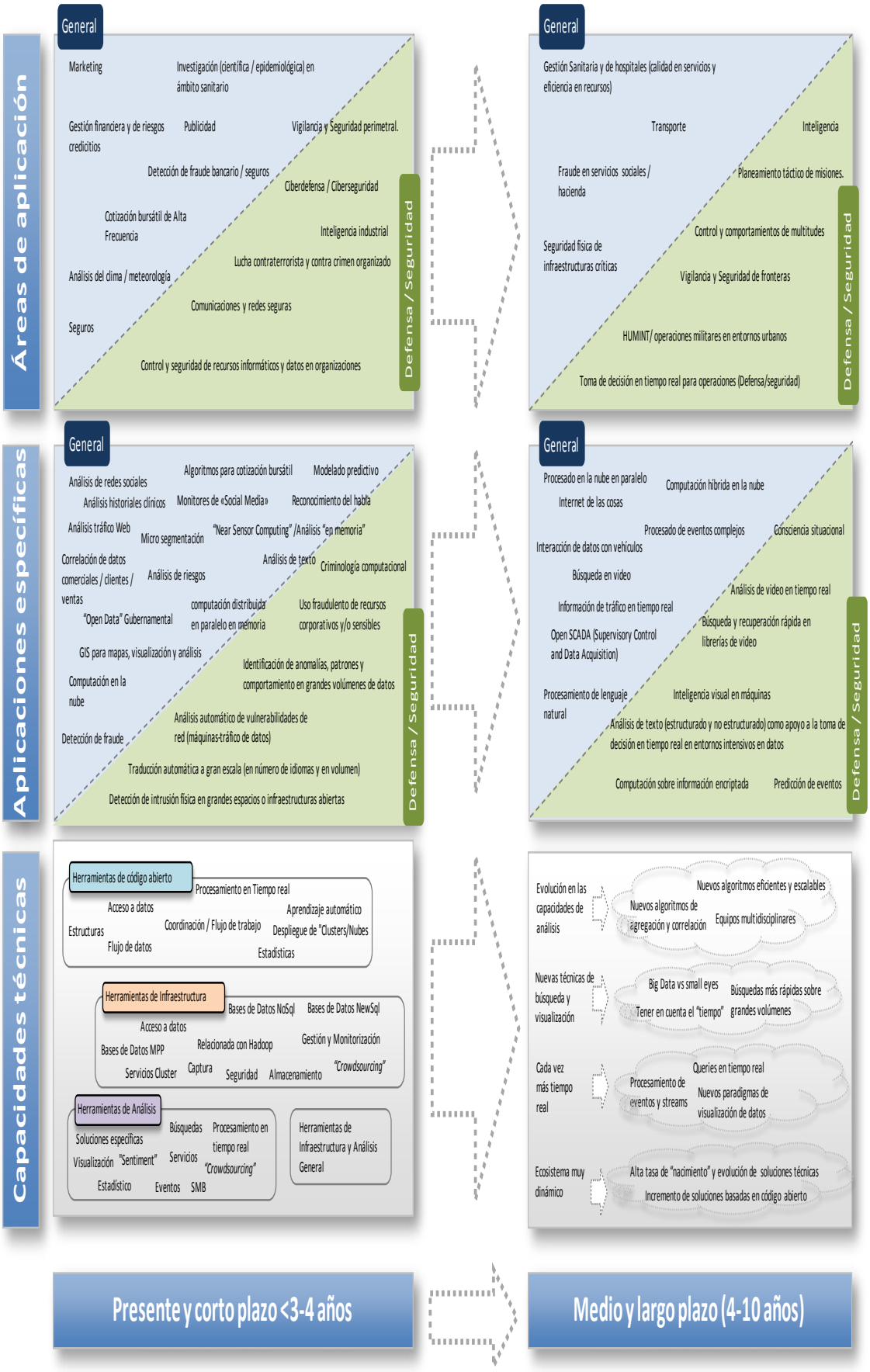
5.2.- Posible evolución.

Se presenta a continuación un posible mapa de evolución de Big Data y sus aplicaciones, tanto desde el punto de vista de seguridad y defensa como desde el de otros sectores generales, como el de educación, financiero, o el de sanidad por ejemplo.

Este mapa se ha realizado combinando la información sobre las aplicaciones específicas con la producida por Gartner (http://www.gartner.com/id=2100215) para su *HypeCycle* sobre Big Data, así como las tendencias sobre las capacidades técnicas identificadas en el proyecto FP7 *BIG - Big Data Public Private Forum* (http://www.big-project.eu/).

En el mapa se presentan tres vistas, tanto para el momento presente y corto plazo, como para el medio y largo plazo. Estas tres vistas, correspondientes a las capacidades técnicas, aplicaciones específicas y diferentes áreas de aplicación, se dividen en el ámbito general, parte izquierda de los cuadrantes, y en ámbito de seguridad y defensa, en la parte derecha de los cuadrantes.

El cuadrante, de capacidades técnicas se desglosa a su vez en un listado (gráfica adjunta) con los diferentes actores técnicos más relevantes para cada una de las tecnologías y sistemas vigentes en la actualidad.


Áreas de aplicación
General
Marketing
Investigación (científica / epidemiológica) en ámbito sanitario
Gestión financiera y de riesgos crediticios
Publicidad
Vigilancia y Seguridad perimetral.
Detección de fraude bancario / seguros
Ciberdefensa / Ciberseguridad
Cotización bursátil de Alta Frecuencia
Inteligencia industrial
Análisis del clima / meteorología
Lucha contraterrorista y contra crimen organizado
Comunicaciones y redes seguras
Seguros
Control y seguridad de recursos informáticos y datos en organizaciones
Defensa / Seguridad
General
Gestión Sanitaria y de hospitales (calidad en servicios y eficiencia en recursos)
Transporte
Inteligencia
Fraude en servicios sociales / hacienda
Planeamiento táctico de misiones.
Seguridad física de infraestructuras críticas
Control y comportamientos de multitudes
Vigilancia y Seguridad de fronteras
HUMINT/ operaciones militares en entornos urbanos
Toma de decisión en tiempo real para operaciones (Defensa/seguridad)
Defensa / Seguridad
Aplicaciones específicas
General
Algoritmos para cotización bursátil
Modelado predictivo
Análisis de redes sociales
Análisis historiales clínicos
Monitores de «Social Media»
Reconocimiento del habla
Análisis tráfico Web
Micro segmentación
"Near Sensor Computing" /Análisis "en memoria"
Correlación de datos comerciales / clientes / ventas
Análisis de riesgos
Análisis de texto
Criminología computacional
"Open Data" Gubernamental
computación distribuida en paralelo en memoria
Uso fraudulento de recursos corporativos y/o sensibles
GIS para mapas, visualización y análisis
Identificación de anomalías, patrones y comportamiento en grandes volúmenes de datos
Computación en la nube
Análisis automático de vulnerabilidades de red (máquinas-tráfico de datos)
Detección de fraude
Traducción automática a gran escala (en número de idiomas y en volumen)
Detección de intrusión física en grandes espacios o infraestructuras abiertas
Defensa / Seguridad
General
Procesado en la nube en paralelo
Computación híbrida en la nube
Internet de las cosas
Procesado de eventos complejos
Consciencia situacional
Interacción de datos con vehículos
Búsqueda en video
Análisis de video en tiempo real
Información de tráfico en tiempo real
Búsqueda y recuperación rápida en librerías de video
Open SCADA (Supervisory Control and Data Acquisition)
Procesamiento de lenguaje natural
Inteligencia visual en máquinas
Análisis de texto (estructurado y no estructurado) como apoyo a la toma de decisión en tiempo real en entornos intensivos en datos
Computación sobre información encriptada
Predicción de eventos
Defensa / Seguridad
Capacidades técnicas
Herramientas de código abierto
Procesamiento en Tiempo real
Acceso a datos
Aprendizaje automático
Estructuras
Coordinación / Flujo de trabajo
Despliegue de "Clusters/Nubes
Flujo de datos
Estadísticas
Herramientas de Infraestructura
Bases de Datos NoSql
Bases de Datos NewSql
Acceso a datos
Bases de Datos MPP
Relacionada con Hadoop
Gestión y Monitorización
Servicios Cluster
Captura
Seguridad
Almacenamiento
"Crowdsourcing"
Herramientas de Análisis
Soluciones específicas
Búsquedas
Procesamiento en tiempo real
Visualización
"Sentiment"
Servicios
"Crowdsourcing"
Estadístico
Eventos
SMB
Herramientas de Infraestructura y Análisis General
Evolución en las capacidades de análisis
Nuevos algoritmos eficientes y escalables
Nuevos algoritmos de agregación y correlación
Equipos multidisciplinares
Nuevas técnicas de búsqueda y visualización
Big Data vs small eyes
Búsquedas más rápidas sobre grandes volúmenes
Tener en cuenta el "tiempo"
Cada vez más tiempo real
Queries en tiempo real
Procesamiento de eventos y streams
Nuevos paradigmas de visualización de datos
Ecosistema muy dinámico
Alta tasa de "nacimiento" y evolución de soluciones técnicas
Incremento de soluciones basadas en código abierto
Presente y corto plazo <3-4 años
Medio y largo plazo (4-10 años)

Infraestructura
Bases de Datos NoSql
10GEN
COUCHBASE
HYPERTABLE
NEO4J
BASHO
CLOUDANT
Bases de Datos NewSql
MARKLOGIC
PARADIGM4
MEMSQL
SQLFIRE
VOLTDB
DRAWNSCALE
Bases de Datos MPP
VERTICA
KOGNITIO
PARACCEL
GREENPLUM
TERADATA
NETEZZA
INFINIDB
MICROSOFT SQL SERVER
Relacionada con Hadoop
CLOUDERA
HORTONWORKS
MAPR
IBM INFOSPHERE
GREENPLUM
AMAZON
HADAPT
INFOCHIMPS
HSTREAMING
ZETTASET
MORTAR
QUBOLE
SQRRL

Infraestructura
Servicios Cluster
LEXISNEXIS
HPCC SYSTEMS
ACUNU
Gestión y Monitorización
OUTER THOUGHT
OCEANSYNC
STACKIQ
DATADOG
Seguridad
STORMPATH
IMPERVA
TRACE VECTOR
CODEFORTYTWO
DATAGUISE
Captura
ASPERA
NODEABLE
Almacenamiento
CLEVERSAFE
PANASAS
NIMBLESTORAGE
AMPLIDATA
COMPUVERDE
Crowdsourcing
CROWD COMPUTING
SROWDFLOWER
AMAZON MECHANICAL TURK

Fuentes de Datos
FACTUAL
DATAMARKET
WINDOWS AZURE
PREMISE
KNOEMA
INFOCHIMPS
DATASIFT
GNIP
SPACE CURVE

Herramientas de código abierto
Estructuras
HADOOP MAP REDUCE
HDFS
Flujo de datos
HIVE
PIG
Acceso a datos
CASSANDRA
APACHE HIBASE
COUCHDB
SCIDB
SQOOP
MONGODB
FLUME
Coordinación / Flujo de trabajo
ZOOKEEPER
TALEND
OOZIE
Tiempo real
STORM
Herramientas Estadísticas
SCIPY
R
Aprendizaje en máquinas
MAHOUT
Despliegue de "Clusters/Nubes"
WHIRR

Análisis
Soluciones
PALANTIR
PERVASIVE
PLATFORA
DATAMEER
KARMASPHERE
DATAHERO
DIGITALREASONING
DATASPORA
Visualización
QUID
TABLEAU
CENTRIFUGE
ACTUATE
METALAYER
AYASDI
CLEARSTORY
ISS
QUANTUM4D
Estadístico
P(K)
PRIOR KNOWLEDGE
REVOLUTION ANALYTICS
SAS
SPSS
MATLAB
"Sentiment"
CRIMSOM HEXAGON
GENERAL SENTIMENT
Servicios
THINK BIG
MU SIGMA
OPERA
Búsquedas
ELASTICSEARCH
AUTONOMY

Análisis
Eventos
RAPLEAF
FLIPTOP
RECORDED FUTURE
PLACEIQ
RADIUS
Procesamiento en tiempo real
CONTINUITY
PARSTREAM
FEEDZAI
Crowdsourcing
DATAKIND
KAGGLE
SMb
SUMAIL
RJMETRICS
CUSTORA

Infraestructura / Análisis General
SAP
SAS
IBM
GOOGLE
ORACLE
MICROSOFT
VMWARE
AMAZON
1010DATA
METAMARKETS
TERADATA
AUTONOMY
NETAPP

6.- Conclusiones del estudio

Hasta aquí se ha intentado dar una visión de lo que se puede entender por Big Data, qué funciones abarca y cuáles son las herramientas y recursos disponibles. Acorde con la gran cantidad de información que hoy en día se maneja en Internet, la imagen es el principal medio de comunicación, ya sea fija o en movimiento, todo lo envuelve, y así se ha incidido en las técnicas más básicas del tratamiento de imágenes y la vertiente matemática necesaria. Ciñéndonos a los entornos de defensa y seguridad se han incluido un conjunto de aplicaciones posibles así como un análisis de cómo pueden evolucionar estas aplicaciones.

Una vez concluidas tales descripciones se puede tomar un punto de vista más distante e intentar reorganizar mentalmente qué es lo que se tiene entre manos. Como primera medida y con objeto de explicar esta situación, hay que tener en cuenta un par hechos insoslayables:

- La potencia de cálculo y almacenamiento que se encuentra actualmente en un equipo personal es equivalente a servidores corporativos de hace no demasiados años.
- Las velocidades de transmisión datos manejadas en el mercado doméstico, y es una muy buena medida en relación a lo que se pretende exponer, han pasado de algunas decenas de kbps a finales del siglo pasado a decenas de Mbps actualmente.

Del mismo modo el software ha ido aprovechamiento esas capacidades, de modo paulatino en unos casos y siendo el motor en otros. Esta situación ha permitido enfocar los problemas con planteamientos diferentes. En todo ello es de reconocer la gran influencia que el mundo del juego digital ha tenido, lo que en algún momento se ha denominado despectivamente como "matar marcianitos". Las posibilidades del equipamiento han ido siendo asumidas por los desarrolladores que a su vez han ido reclamando mejores recursos para mejorar aspectos de velocidad, dimensión, resolución, etc. Igual que la lucha entre el proyectil y la coraza, cada evolución de uno de ellos se responde a una nueva evolución del otro.

Desde un punto de vista meramente tecnológico no se puede decir que estemos ante una revolución de una u otra capacidad. Muchas de las técnicas a emplear son ampliamente conocidas y aplicadas desde hace tiempo, tanto en los campos de la computación, distribuida o centralizada, como de la transmisión o almacenamiento. Sin embargo lo que sí se puede calificar como novedoso, dentro de un orden, es el hecho de haber reunido todas esas técnicas, llevarlas al límite y aplicarlas a la resolución de nuevos retos.

Estamos acostumbrados a disponer de información estructurada que se guarda en una base de datos y que responde a las características de la naturaleza de esa información.

Sin embargo la mente humana funciona de otra forma, somos capaces de relacionar situaciones, hechos o cualquier tipo de información aplicando unas condiciones más sutiles, menos definidas. Con informaciones parciales que no presentan una estructuración aparente se llega a entender que se relacionan, que existe una conexión que a veces calificamos de intuición. Esta información no tiene porqué responder a un esquema definido y sin embargo una vez relacionada entre sí nos parece lo más lógico ese enlace y relación. Y aquí es donde se manifiesta el avance que se puede tener al emplear y aplicar estas técnicas, nos ayudan a relacionar o calcular datos de informaciones aparentemente inconexos. Esto es, acercan el comportamiento de las máquinas al funcionamiento de la mente humana, tomando con suficiente cautela tal afirmación. En este ámbito la función que necesita un mayor impulso es la de **presentación.**

En general la utilización de medios que compendien las informaciones disponibles se limitan a paneles más o menos grandes donde se intenta reproducir directamente esa información, dato o relación que se ha puesto en evidencia. Pero esto no es suficiente, es necesario que la representación sea no simplemente del dato sino de la tendencia, de los factores que han intervenido en su evolución y de cuanto han influido. Un ejemplo claro de esto es el interés de una parte de la comunidad científica en demostrar la influencia humana en "el cambio climático" y como desde opiniones contrarias se les acusa de manipulación de esas informaciones para obtener las conclusiones propuestas inicialmente.

Por otra parte la disponibilidad de la información hasta no hace mucho estaba al alcance y era "propiedad" de algunas corporaciones, siendo las más representativas los estados. Sin embargo ahora existe gran cantidad de información disponible y que no es propiedad de esas corporaciones. Los individuos o entidades mucho menores pueden manejarla y tienen la ocasión de obtener resultados con sus análisis. Caso claro es la profusión de las redes sociales. Esta generación de información se ha disparado necesitando nuevas formas de enfrentar los problemas que resultan al intentar tratarla, por eso se habla de **nuevos retos** actualmente se puede intentar manejar una parte importante de toda esa información para sacar conclusiones. Es posible identificar soluciones de problemas o planteamientos que hasta hace poco eran inimaginables. Precisamente en esa fuente de propuestas, de nuevas actividades o de planteamientos novedosos es donde reside la ventaja que una corporación puede tener frente a otras: puede generar actividad donde ni siquiera se pensaba que podía haberla. Esta diferenciación es la que le permitiría tener un objeto de negocio propio.

En definitiva, desde un punto de vista eminentemente técnico la evolución de estas herramientas no es simplemente una progresión más o menos prevista de versiones o de metas alcanzables. No es simplemente una mejora en la implementación de un algoritmo o la disponibilidad de máquinas más potentes. Probablemente se trate de incluir como parte de las soluciones, además del dominio de la propia técnica, otros elementos no propios de un enfoque técnico o funcional. Intentando buscar un símil es como hizo Einstein, que se planteó la duda sobre hechos y afirmaciones hasta

entonces axiomáticos y así pudo desarrollar una trayectoria de pensamiento totalmente distinta.

Otro factor determinante ha sido la aportación individualizada de muchas personas u organizaciones, no en vano la mayoría de las herramientas son de software de fuentes abiertas (open source). Con este tipo de colaboración es posible que la aplicación de estas capacidades tenga muchas más posibilidades de desarrollarse en entornos abiertos donde estas aportaciones individuales van sumándose unas a otras. Donde el límite que uno llega a ver insalvable por su propio condicionamiento es fácilmente superado por otros. Hay que tener en cuenta que se trata de cientos de miles de personas que colaboran, de modo anárquico cierto, sin una sola dirección pero precisamente por ello las soluciones, las ideas pueden ser múltiples y diferentes.

Por último dos aspectos que no se pueden olvidar y que ya están en las puertas de la esfera social de Internet. Por un lado "**la Internet de las cosas**". El volumen de información que se va a volcar y a tener disponible desde todo tipo de elementos va a permitir un sinfín de posibilidades para quien tenga una buena idea y llegue a desarrollarla. Y precisamente esto nos lleva a la segunda y última faceta de todo esto: ¿hasta qué punto toda esa información disponible en la red puede volverse en contra de cada individuo? Este es un **enfoque ético y legal**, al final las cosas se conectan a la red porque son de alguien y por una razón de quien lo hace, es decir, estarán ligadas a las personas, a las organizaciones y revelarán mucha información de sus dueños, de quienes actúen sobre ellas; como ya está ocurriendo. Esto lo que plantea es la necesidad de desarrollo de un marco regulatorio que proteja a las personas del aprovechamiento fraudulento de esa información. En sí misma esta idea es muy compleja actualmente no es posible hacer valer las normas más allá de los estados que las promulguen y con otros estados que se han comprometido en su cumplimiento. Y el hecho que echa por tierra lo anterior es cuando son los propios estados, o las organizaciones multinacionales, los interesados en aprovechar esa cantidad de información. No hay que hacer muchas disquisiciones para recordar algunos hechos recientes con nombre propio: Wikileaks, Snowden, NSA …

7.- Referencias

1. Michael S., Rebecca S., Janet S., otros, Analytics: el uso de Big Data en el mundo real, Informe ejecutivo de IBM Institute for Business Value, IBM Corporation 2012.

2. Jean-Pierre D., Oracle Big Data for the Enterprise, Oracle Whitepaper, 2012.

3. Ivan P., Jaime G., Big Data: Cómo la avalancha de datos se ha convertido en un importan-te beneficio, informe de TICbeat, 2012.

4. John G., David R., The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, Sponsored by EMC Corporation, 2012.

5. Helen S., Peter H., Oracle Information Architecture: An Architect's Guide to Big Data, Oracle Whitepaper, 2012.

6. No SQL para No programadores, http://lapastillaroja.net/2012/02/nosql-for-non-programmers/.

7. Tom W.: Hadoop: The Definitive Guide, Third Edition. ISBN: 978-1-449-31152-0 1327616795, 2012.

8. Platforms for Big Data, http://www.cubrid.org/blog/textyle/369047.

9. Tutorial Storm, https://github.com/nathanmarz/storm/wiki/Tutorial.

10. Amr A., Webster M., The Plataform for Big Data: a Blueprint for Next-Generation Data Management, Cloudera White Paper, 2013.

11. Hortonworks, Apache Hadoop: The Big Data Refinery, Hortonworks White Paper, 2012.

12. MapR, The MapR Distribution for Apache Hadoop: Easy, Dependable & Fast Hadoop, WhitePaper, 2011.

13. IBM, The IBM Big Data Platform, IBM Solution Brief, 2012.

14. Sergey M., Andrey G., Jing L. Geoffrey R., Shiva S., Matt T., Theo V, Dremel: Interactive Analysis of Web-Scale *Datasets*, 36th International Conference on Very Large Data Ba-ses, Singapore, 2010.

15. Kasunory S., An Inside Look at Google BigQuery, Google BigQuery WhitePaper, 2012.

16. Jinesh V., Sajee M., Overview of Amazon Web Services, March 2013.

17. Philip R., Big Data Analytics, TDWI Best Practices Report, 2011.

18. Srinath P. Thilina G., Hadoop MapReduce Cookbook, ISBN: 978-1-84951-728-7, 2013.

19. Mark van R., Walmart Makes Big Data Part of Its DNA, http://smartdatacollective.com/bigdatastartups/111681/walmart-makes-big-data-part-its-social-media.

20. Fidelity Worldwide Investment, Big Data: una "revolución industrial" en la gestión de los datos industriales, In Perspective, Ep. 5, 2012.

21. Tech America Foundation, Demystifyng Big Data: A Practical Guide To Transforming The Business of Government, 2012