Actas de XIV Jornadas de ARCA
Sistemas Cualitativos y sus Aplicaciones en
Diagnosis, Robótica e Inteligencia Ambiental
F. J. Ruiz, N. Agell, J. A. Ortega (Eds.)
(c) 2012

Big Data: Un nuevo problema computacional

D. López, P. Gallego, A. Fernández-Montes, J. I. Sánchez-Venzalá, J. Torres
University of Seville
Computer Languages and Systems Dept.,
41012, Seville, Spain
dlopez@cica.es, pgallego@cica.es, afdez@us.es, jisanchez@us.es, jortega@us.es

Abstract

El aumento de la capacidad de procesamiento en los computadores permite llevar a cabo tareas que hasta ahora no eran viables: simulación de procesos naturales, almacenamiento de datos geográficos, económicos, multimedia, inteligencia social, etc. Como contrapartida, el volumen de datos que se genera en este tipo de aplicaciones puede llegar a provocar que el coste computacional para su tratamiento y análisis, con las herramientas actuales, sea tan alto que vuelva a convertirse en un problema inabordable. Este problema es conocido por el término 'Big Data'.

Según afirma IBM en un estudio reciente: "Se producen más de 2,5 quintillones de bytes al día, hasta el punto de que el 90% de los datos del mundo han sido creados durante los 2 últimos años". De seguir con este ritmo de crecimiento en breve se generará más volumen de datos del que se puede analizar, disminuyendo su utilidad al tiempo que el coste y el riesgo de pérdida de información aumenta.

El problema se acrecienta todavía más en aplicaciones con tratamiento de información en tiempo real, donde el valor de los datos reside en la actualidad de los mismos. En este tipo de aplicaciones una gestión y tratamiento eficientes de los datos es de vital importancia.
Las dificultades más habituales en el tratamiento radican en la captura, análisis, almacenamiento, búsqueda, compartición, y visualización. Las técnicas existentes en la actualidad para abordar este problema no son efectivas, por lo que el problema 'Big Data' sigue abierto.

Este problema tiene una gran repercusión en diversos ámbitos: redes sociales, aplicaciones de video, dispositivos móviles, webs, laboratorios astrofísicos, simulación científica, captura de datos de clientes (compañías eléctricas, telefónicas, etc).

Por tanto, los beneficios potenciales de un tratamiento y gestión eficientes de grandes volúmenes de datos, y su gran aplicabilidad sobre múltiples campos, hace que sea un tema muy atractivo, en el que actualmente trabajan numerosas empresas y grupos de investigación.

Desde el punto de vista de la investigación es un problema relativamente reciente, por lo que el desarrollo de soluciones y la bibliografía son todavía escasos en comparación con otros campos. En este artículo se abordarán las técnicas utilizadas actualmente y las principales líneas que se siguen para desarrollar soluciones futuras.

1 Introducción

Llamada a ser la materia que alimentará la siguiente revolución social y económica de nuestro siglo, el sector de las tecnologías de la información y la comunicación produce grandes cantidades de datos que por la velocidad a la que se generan, junto con las múltiples fuentes y formatos y la capacidad de procesamiento actual, hacen realmente complicado trabajar con ellos, siendo necesario al mismo tiempo equilibrar tamaño, velocidad, diversidad y capacidad de proceso.

En los próximos años, el término Big Data adquirirá importancia progresivamente, ya que representa un reto inmediato en innovación tecnológica: capacidad para el tratamiento y análisis de repositorios de datos de tamaño tan desproporcionado, que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales.

El paradigma de Big Data no sólo representa la capacidad de almacenamiento de la información generada por múltiples fuentes y en múltiples formatos a gran velocidad, sino la capacidad de interconectarlos y generar una inteligencia colectiva, aplicable a cualquier ámbito.

Big Data es un reto tecnológico que implica numerosos ámbitos, donde billones de bytes de información acerca de toda clase de fenómenos y actividades se producen, se al-

macenan y se difunden a través de diversos canales, como el teléfono móvil, una red social o la memoria de cualquier máquina.

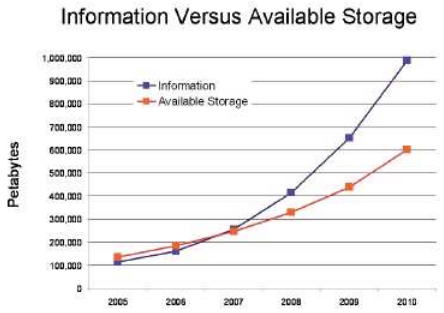

Figura 1 - Información vs. Almacenamiento disponible

Basta echar un vistazo a la sociedad de la información actual, para percibir la existencia de este fenómeno: proliferación de páginas web, aplicaciones multimedia (sonido, imagen y vídeo), redes sociales, dispositivos móviles, apps, sensores y redes de sensores, etc. Según IBM, se generan más de 2.5 quintillones de bytes al día, y se estima que el conjunto de soporte de almacenamiento existente en nuestro planeta contiene 2,5 zetabytes de información.

Sin embargo, el problema radica en que la mayor parte de dicha cantidad de datos se ha producido en apenas los dos últimos años y que cada anualidad, además, crece a un ritmo aproximado del 150 por ciento. Esta cifra pone de manifiesto la necesidad no solo de ser capaces de almacenar y organizar los datos que se generan, sino también de analizarlos y utilizarlos.

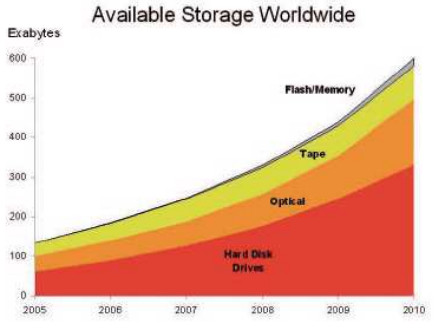

Figura 2 - Evolución del almacenamiento disponible

Para llegar a esta situación, la sociedad de la información ha evolucionado, equipándose con máquinas progresivamente más potentes; desde teléfonos inteligentes, tabletas y ordenadores portátiles, hasta televisores de alta definición, junto con la gran revolución social que ha supuesto internet y todo lo que le rodea, especialmente las redes sociales y la denominada Web 2.0.

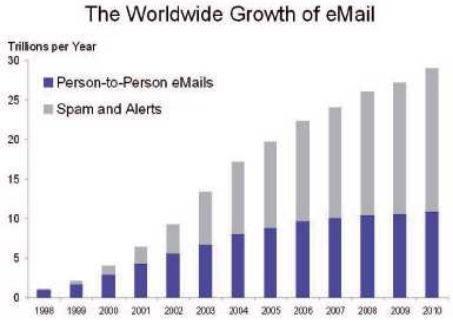

Figura 3 - Evolución del envío de e-mails y SPAM

Sin embargo, se espera un impacto todavía mayor sobre los llamados *macrodatos*, con la evolución del *Internet of Things* o *Internet de las Cosas,* donde cada objeto o elemento cotidiano dispone de un identificador y está conectado en red, adquiriendo, creando y comunicando información.

En estos momentos existen aproximadamente unos 30 millones de sensores en todo el planeta que captan datos y distribuyen la información por la red mediante alguna tecnología de comunicación. Estos sensores están presentes en terminales móviles, contadores de energía, vehículos, maquinaria industrial e incluso sencillos artículos de consumo como prendas de vestir o entradas para el cine.

2 Beneficios de Big Data

El gran atractivo de los datos masivos es su potencial aparente para predecir comportamientos y fenómenos. La estadística funciona mejor, si en lugar de 6 disponen de hasta 300 factores de cálculo, pero superan todas las expectativas cuando se combina el estudio convencional de conjuntos de información tradicionales, como los pedidos de los clientes o el flujo de inventario, con otros menos convencionales y generalmente de límites difusos, como por ejemplo los datos extraídos de aplicaciones webs sociales o correo electrónico.

Conocer en detalle a un usuario particular, a un consumidor o a un ciudadano, ayuda a anticipar sus comportamientos, expectativas y necesidades concretas. Permitiendo, en resumen, ahorrar tiempo, hacer más eficaces determinados servicios y tomar decisiones más precisas.

En general, Big Data aporta beneficios en: toma de decisiones más rápida y eficaz, análisis predictivo y optimización continua de los sistemas de trabajo y mejora de la eficiencia.

A continuación se enumeran diversas aplicaciones reales de Big Data, explicando qué puede ofrecer, diferenciando además entre ámbito empresarial o público.

A nivel Empresarial:

- **Ofrece una descripción cada vez más precisa y detallada de las fluctuaciones y rendimientos de todo tipo de recursos**. Procter & Gamble, empresa de productos de higiene, es capaz de integrar en una única herramienta la manera en la que los consumidores de 80 países distintos utilizan diariamente unas 4.000 millones de dosis de sus productos.

- **Permite realizar adaptaciones "experimentales" a cualquier escala de un proceso y conocer su impacto en tiempo casi real sobre un bien o servicio concreto**. Wal-Mart y Coca-Cola son dos compañías que combinan ya sus bases de datos y sus plataformas de análisis para estudiar la información en tiempo real que obtienen masivamente de sus respectivos clientes (por ejemplo, a través de las máquinas de vending).

- **Ayuda a conocer mejor la demanda y a realizar una segmentación más ajustada de la oferta.** El Financial Times utiliza el análisis de los datos masivos para optimizar las tarifas de sus anuncios según la demanda inmediata de sus lectores: qué leen, a qué hora, de qué sección, desde qué localidad… Sus ventas son hoy mayores debido no sólo a un mejor conocimiento del producto, sino, ante todo, a una capacidad más elevada de sus profesionales para detectar los nichos de la publicación insuficientemente explotados.

- **Acelera el desarrollo de prestaciones y productos cada vez más innovadores y eficientes.** El servicio 1004 de atención al cliente de Telefónica utiliza modelos predictivos para determinar el número de llamadas que recibirá en fechas muy concretas. De esta forma, por ejemplo, el call center más grande de Europa (14 millones de llamadas al mes) ha conseguido mejorar en un 50% su eficiencia.

A nivel público:

- **Eficiencia, a través de decisiones más inteligentes en torno a la organización de los distintos departamentos, la priorización de las tareas internas y la reducción de los costes de funcionamiento operativo**. El Instituto Global Mckinsey estima que la explotación de los conjuntos de datos masivos alberga un potencial anual de 240.000 millones de euros para la sanidad estadounidense (más del doble de la inversión española en este sector) y un valor de 200.000 millones de euros para la administración pública europea (casi el equivalente al PIB de Grecia).

- **Lucha contra el fraude y errores no detectados**. Gracias a la gestión de conjuntos masivos de datos, la Oficina Federal de Investigación de los Estados Unidos, culminó en 2011 la mayor operación de su historia contra el fraude en el sistema de cobertura médica del Gobierno. En concreto, destapó una red de empresas y particulares que facturaron ilegalmente a cuenta del programa de asistencia pública Medicare, destinado a personas mayores de 65 años, unos 4.100 millones de euros. Esta cifra representa casi el 1 por ciento de la dotación económica de dicho programa en 2010.

- **Mejoras en la recaudación de impuestos**. Se considera que el tratamiento a gran escala de la información que atesora la Hacienda del Reino Unido podría ahorrar a los contribuyentes de ese país entre 20.000 y 41.000 millones de euros, es decir, una media de 470 euros por persona. Un 6,25 por ciento de esa cantidad se obtendría gracias a una reducción significativa del fraude; un 12,5 por ciento a la mejora del sistema de recaudación de impuestos; y un 81,25 por ciento a un incremento de la eficacia operativa.

En resumen, recogidos y analizados con una tecnología adecuada, los conjuntos masivos de información pronto ayudarán a transformar el mundo, consiguiendo más beneficios empresariales y reduciendo el consumo público.

A día de hoy, una empresa que prescinda de servicios de análisis de grandes cantidades de datos, la situará en clara desventaja frente a sus competidores.

3 Implementación de Big Data

Están surgiendo multitud de herramientas relacionadas con el Big Data, y se esperan que sigan apareciendo en los próx-

imos años muchas más. A continuación se enumeran los distintos programas o tecnologías más relevantes.

NoSQL: Los sistemas tradicionales de BBDD, tienen una serie de problemas para desenvolverse en el nuevo contexto que se presenta. Es por ello que han surgido alternativas para intentar paliar los efectos de la gran cantidad de datos. Entre ellas tenemos.

- **Leer datos es costoso:** En el modelo relacional los datos se representan mediante conjuntos relacionados entre sí. Realizar una consulta en el modelo relacional implica juntar grandes conjuntos de datos con operaciones algebraicas (como el producto cartesiano) y luego filtrar todo el conjunto resultante lo cual requiere de una gran complejidad computacional.

- **Transaccionalidad innecesaria:** El objetivo de las transacciones es asegurar la integridad de los datos. A veces estas transacciones no son necesarias completamente. Podemos pensar en modelos de negocio en los que las transacciones sean necesarias, y otros que no, en ese caso, se podrán evitar los motores relacionales.

- **Escalabilidad:** El gran problema de las bases relacionales es la escalabilidad. Estas fueron pensadas para correr en un solo servidor con mucha potencia, o como mucho tener replicaciones y balanceo de carga. Estas bases se basan en la idea de escalabilidad vertical, o sea, cuando el servidor se queda pequeño, se escala para tener más potencia (mas CPU, RAM, disco…). El costo de este método aumenta exponencialmente y tiene un límite en que no se puede seguir escalando. La escalabilidad horizontal por otra parte, se basa en poner más servidores de forma paralela y tener la base de datos distribuida entre todos esos servidores. En un RDBMS partir la base en muchos servidores conlleva muchos problemas de rendimiento y gestión, ya que se rompe la integridad del modelo al estar dividido. Hacer una consulta puede involucrar traer datos de todos los servidores con el alto coste computacional que esto implica.

- **Representación del modelo en una RDBMS**. Si bien es posible representar la mayoría de modelos usando el modelo relacional, no siempre resulta la mejor opción. De hecho, actualmente se programa siguiendo el paradigma orientado a objetos, lo que provoca el problema de traducir los objetos a un modelo relacional.

Por otro lado, se ha de tener en cuenta el teorema de CAP, que dice que una BBDD solo puede cumplir dos de los siguientes tres principios: consistencia, disponibilidad y tolerancia al particionamiento. En el caso de los RDBMS, se le da más importancia a la consistencia y a la disponibilidad, en detrimento de la tolerancia al particionamiento. Por otro lado, las diferentes opciones de NoSQL dan mayor prioridad a la tolerancia y en ocasiones a la disponibilidad.

Si tuviésemos que enumerar las principales características de este modelo de sistemas destacaríamos, la ausencia de esquemas en los registros de datos, escalabilidad horizontal sencillez y alta velocidad.

La forma mas usada para su implementación es la de key-value. Éstas son las bases de datos más simples en cuanto su uso, donde el tipo de contenido no es importante para la base de datos, solo la clave y el valor que tiene asociado. A diferencia de un RDBMS no necesita definir un esquema (columnas, tipos de datos) para almacenar la información. Los almacenes key-value son muy eficientes para lecturas y escrituras, además de que pueden escalar fácilmente particionando los valores de acuerdo a su clave. Por ejemplo, aquellos cuya clave está entre 1 y 1000 van a un server, los de 1001 a 2000 a otro, etc. Esto los hace ideales para entornos altamente distribuidos y aplicaciones que necesitan escalado horizontal.

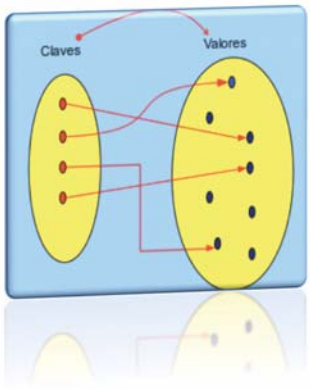

Figura 4 - Esquema de mapeo Clave-Valor

Htable es una implementación concreta de NoSQL. Es de Software libre, y fue modelada en java siguiendo el ejemplo de BIG Table creado por Google. Lo desarrolla la Fundación Apache como parte del proyecto Apache Hadoop, por lo que para su funcionamiento necesita del sistema de ficheros distribuidos HDFS.

Para su implementación se usa un *hash-map* de datos persistente, multidimensional, ordenado, poco denso y distribuido. El *hash-map* está indexado por una clave de fila, una clave de columna y una timestamp.

Con persistencia, se refiere a que al terminar de usar los datos éstos siguen existiendo, de la misma forma que un archivo sigue existiendo en un sistema de archivos.

En cuanto a distribuido, se refiere a que gracias a las características de su sistema de archivos el almacenamiento puede ser repartido entre un conjunto de máquinas independientes. Los datos se replican a través de una serie de nodos que participan de una manera análoga a cómo los datos están almacenados en un RAID. Esto genera una capa de protección contra fallos del sistema.

A diferencia de la mayor parte de las implementaciones de *hash-map*, en Hbase los pares clave/valor se guardan en un estricto órden alfabético. Es decir, la clave "aaaa" debe estar al lado de la fila con la clave "aaab" y muy lejos de la "zzzz". Esta caracterísitca es muy importante ya que por lo general, este tipo de sistemas tienden a ser muy grandes y distribuidos. La proximidad espacial de filas con llaves similares asegura que, cuando hay que recorrer la tabla, los elementos de mayor interés para la búsqueda están cerca unos de otros, mejorando sensiblemente el acceso a los datos.

Como ya hemos explicado en este modelo se sustituye el concepto de "columnas", por lo que Hbase se puede entender como un mapa multidimensional, es decir, un mapa de mapas.

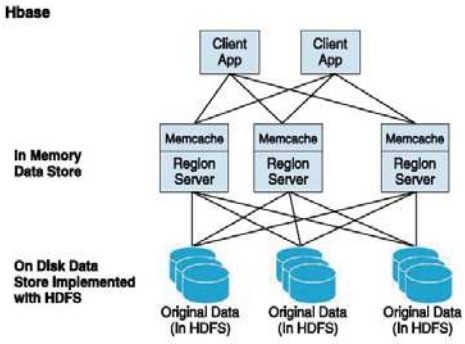

Figura 5 - Esquema de Hbase

Hadoop, como ya se ha mencionado anteriormente, está muy vinculado con HBase, ya que forma parte del mismo proyecto. Hadoop se inspiró en Map Reduce y Google File System.

Una funcionalidad clave consiste en que para la programación efectiva de trabajo, cada sistema de archivos debe conocer y proporcionar su ubicación. Las aplicaciones Hadoop pueden usar esta información para ejecutar trabajos en el nodo donde están los datos o, en su defecto, en el mismo rack/switch, reduciendo así el tráfico de red. El sistema de archivos HDFS usa esta técnica cuando replica datos, para intentar conservar copias diferentes de los datos en racks diferentes. El objetivo es reducir el impacto de un corte de energía en un rack o de fallo de interruptor de modo que incluso si se producen estos eventos, los datos todavía puedan ser legibles.

Un clúster típico Hadoop incluye un nodo maestro y múltiples nodos esclavo. El nodo maestro consiste en *jobtracker* (buscador de trabajos), *tasktracker* (buscador de tareas), *namenode* (nodo de nombres), y *datanode* (nodo de datos). Un esclavo consiste en un nodo de datos y un rastreador de tareas. Los diferentes envíos de datos se realizan mediante SSH.

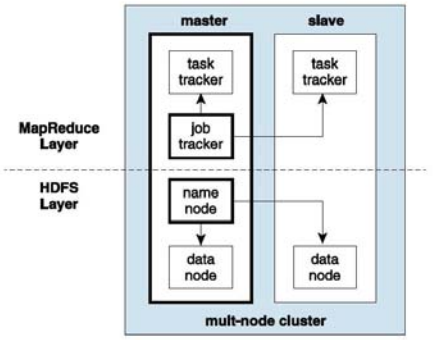

Figura 6 - Esquema de Hadoop

Por su parte, el Hadoop Distributed File System (HDFS) es un sistema de archivos distribuido, escalable y portátil desarrollado en Java para el framework Hadoop. El sistema de archivos usa la capa TCP/IP para la comunicación, mientras que los clientes usan RPC para comunicarse entre ellos. El HDFS almacena archivos grandes (el tamaño ideal es de 64 MB), a través de múltiples máquinas. Consigue fiabilidad mediante replicado de datos a través de múltiples hosts, y no requiere almacenamiento RAID en ellos.

Con el valor de replicación por defecto, los datos se almacenan en 3 nodos: dos en el mismo rack, y otro en un rack distinto. Los nodos de datos pueden hablar entre ellos para reequilibrar datos, mover copias, y conservar la alta replicación de datos.

HDFS no cumple totalmente con el estándar POSIX porque los requerimientos de un sistema de archivos POSIX difieren de los objetivos de una aplicación Hadoop, ya que en este caso el objetivo no es tanto cumplir con los estándares

POSIX sino buscar la máxima eficacia y rendimiento de datos.

4 Conclusiones

Big Data es un paradigma de gestión de grandes volúmenes de datos que suscitará un interés cada vez mayor debido al aumento progresivo de la generación de datos que provocan la evolución de las nuevas tecnologías junto con el desarrollo de la sociedad de la información: redes sociales, multimedia, sensores, Internet 2.0, Internet of things, etc.

Las tecnologías que han sido válidas hasta el momento no lo son para manejar estos volúmenes de datos. Al mismo tiempo que se complica el almacenamiento, debido a la imposibilidad de analizar y relacionar volúmenes tan grandes de datos, estos pierden valor.

Hasta ahora, la información se encontraba estructurada y tenía un ámbito o extensión bien definidos, lo que permitía analizarla de manera conveniente para extraer conocimiento. Sin embargo, el crecimiento del volumen de datos y la no escalabilidad de los métodos y técnicas utilizadas para el análisis, impiden la aplicación a volúmenes más extensos. Como consecuencia, al disponer de una cantidad mayor de información, su análisis debería ser más productivo. Sin embargo, esto no es así, debido al aumento exponencial de la complejidad del análisis.

Como se ha mostrado, ya existen diversas técnicas y productos que utilizan este paradigma, y con toda probabilidad aparecerán muchos más en los próximos años.

Desde el punto de vista empresarial, el Big Data, es un gran problema, debido el coste que supone el manejo de tan ingente cantidad de datos, al tiempo que constituye una nueva línea de negocio donde ofrecer soluciones y productos.

Desde el punto de vista de la investigación, el campo del Big Data, se encuentra en una fase muy temprana de desarrollo, y ofrece una gran cantidad de líneas abiertas y problemas a resolver.

Lo que en un principio comenzó como un problema, se ha convertido gracias a las nuevas técnicas de almacenamiento y análisis en una gran oportunidad y un gran negocio para las empresas. Y no solo para las compañías que se encargan de almacenar, gestionar y analizar los datos, sino para todas esas corporaciones que pueden sacar valor de los mismos. Hasta ahora, la información que se manejaba provenía de fuentes conocidas, eran datos estructurados y flexibles, cuyo análisis tenía unos costes razonables. Pero la llegada y popularización de las redes sociales, vídeos, imágenes, etc… ha provocado una avalancha de datos denominados desestructurados que son más difíciles de medir y analizar, pero cuyo análisis puede ser de suma utilidad para las empresas a la hora de tomar decisiones. Ahora existe un volumen de información mayor que el que había antes, y no tiene por qué ser conocida a priori. Ante este reto, la tecnología actual está preparada y disponible y ya está empezando a gestionarlo. Nos encontramos en una fase que consiste en acceder a esa información de forma inteligente y sacar patrones de conocimiento, es decir, sacar el valor del dato y esto se distingue del dato clásico: el verdadero valor no se conoce a priori, en el momento de la adquisición, sino más adelante, siendo también importante hay que saber integrar esos datos no estructurados con los racionales y cruzarlos.

Referencias

[ABI Research, 2010] ABI research, Consumer technology barometer: Mobile, 2010.

[Bohn et al., 2010] Bohn, Roger, James Short, and Chaitanya Baru, How much information? 2010: Report on enterprise server information, University of California, San Diego, Global Information Industry Center, January 2011.

[Bollier et al., 2010] Bollier, David, The promise and peril of big data, Aspen Institute, 2010.

[Bracht et al., 2005] Bracht, U., and T. Masurat, "The Digital Factory between vision and reality," Computers in Industry 56(4), May 2005.

[Chui et al., 2010] Chui, Michael, Markus Löffler, and Roger Roberts, "The Internet of things," McKinsey Quarterly, March 2010.

[Davenport et al., 2010] Davenport, Thomas H., and Jeanne G. Harris, Analytics at work: Smarter decisions, better results. Cambridge, MA: Harvard Business Press, 2010.

[Davenport et al., 2007] Davenport, Thomas H., and Jeanne G. Harris, Competing on analytics: The new science of winning. Cambridge, MA: Harvard Business Press, 2007.

[Gleick et al., 2011] Gleick, James, The information: A history. A theory. A flood. New York: Pantheon Books, 2011.

[Hubbard et al., 2010] Hubbard, Douglas W., How to measure anything: Finding the value of intangibles in business, New York, Wiley, 2010.